

SimCounsel: Emotion-Aware Simulation of Multi-Agent Counseling

Anonymous ACL submission

Abstract

While counseling is a promising domain for applying large language model (LLM) based multi-agent simulations, research in this area remains in its early stages. In this study, we propose SimCounsel, a multi-agent simulation framework designed to explore the effectiveness of counseling interventions. SimCounsel incorporates emotion-aware memory retrieval mechanisms and persona update strategies based on real psychological theories, enabling more human-like and context-sensitive agent behavior. Both automated evaluations and expert evaluations confirmed that the proposed framework not only realistically simulates the behavior of real clients and therapists, but also effectively facilitates analysis of counseling outcomes. These findings suggest that SimCounsel holds significant promise as a scalable and research-friendly tool for counseling-related simulations. All the code for the simulation is presented on GitHub.¹

1 Introduction

The remarkable performance of LLMs has demonstrated their ability not only to generate human-like text, but also to emulate complex patterns of thought and behavior (Yao et al., 2023; Shinn et al., 2023; Wei et al., 2022). This capability has driven significant progress in multi-agent simulations, a research area that seeks to replicate human social phenomena by deploying autonomous agents capable of reasoning and interacting within virtual environments. For example, the Generative Agents (Park et al., 2023) introduced how LLM-driven agents, when placed in a simulated village, can engage in realistic social behaviors such as information diffusion, relationship formation, and cooperative problem-solving, closely mirroring real-world dynamics. Building on this foundation, subsequent multi-agent simulation studies

have extended these approaches to domain-specific applications (Li et al., 2025; Zhang et al., 2024; Li et al., 2023b), addressed technical limitations (Yan et al., 2025), and incorporated insights from psychology and social science to enhance the realism and complexity of agent interactions (Mitsopoulos et al., 2024; Wang et al., 2023; Sun et al., 2024a). These advances highlight the unique strengths of multi-agent simulations, including cost efficiency, scalability, and the ability to control experimental conditions without the logistical and ethical challenges associated with human subjects.

Nevertheless, although these simulations increasingly capture the complexity of human behavior, most agent memory models still fall short of adequately incorporating the influence of emotion. Psychological studies have shown that emotionally intense experiences are more likely to be retained over long periods (Yonelinas and Ritchey, 2015) and that the likelihood of recalling a memory can vary substantially depending on the individual’s emotional state at the time of recall (Kuiken, 1991; Blaney, 1986). This gap underscores the need for a more human-like memory model that incorporates emotional context alongside traditional cognitive attributes.

Moreover, while multi-agent simulations offer valuable insights into human behavior, their application in the mental health domain remains under-explored. Current research in this area primarily focuses on replicating psychotherapy sessions (Lee et al., 2024; Qiu and Lan, 2024), often neglecting the broader, real-life contexts in which these interactions occur. As a result, while these studies can assess the quality of therapeutic dialogues, they struggle to capture the long-term, real-world impact of therapy on the overall behavior, cognition, and emotional well-being of the client agents.

To address these limitations, we propose SimCounsel, a novel multi-agent simulation framework designed to evaluate the real-world impact of psy-

¹Link to be provided after review for anonymity.

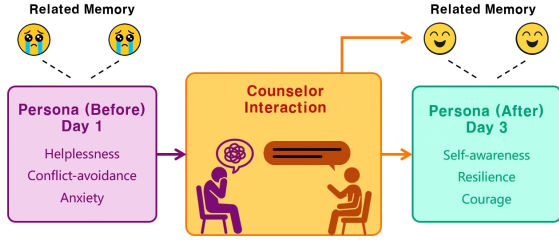


Figure 1: Overview of the SimCouncil Framework

chotherapy. Our framework bridges the gap between isolated therapy sessions and the broader life contexts of clients by integrating realistic client agents modeled on actual behavioral data. It also introduces a memory model that accounts for emotional context, providing a more human-like approach to memory retrieval, and allows for continuous persona updates, enabling real-time observation of therapeutic effects. Figure 1 presents an overview of the proposed framework, illustrating key components and interactions within the system.

Our main contributions are as follows:

- We introduce an emotion-aware memory retrieval that allows agents to recall experiences in a manner more closely aligned with human memory processes, capturing both cognitive and emotional contexts.
- We develop a realistic client simulation based on real-world data, supporting more authentic therapeutic interactions.
- We implement a dynamic persona update mechanism that allows continuous tracking of client agents, providing a more comprehensive view of the long-term impact of therapeutic interventions.

2 Related Works

2.1 Psychological Foundations of Memory, Emotion, and Behavior

According to the Adaptive Control of Thought-Rational (ACT-R) model (Anderson, 2014; Anderson et al., 2004) and Instance-Based Learning (IBL) theory (Gonzalez et al., 2003), human memory is shaped by external stimuli and experiences, which significantly influence decision-making and behavior. However, emotion also plays a critical role, as emotionally charged events tend to form stronger and more enduring memories compared to neutral ones (LaBar and Cabeza, 2006; Hogeveen et al., 2016; Yonelinas and Ritchey, 2015). The Emotional Binding Model suggests that the amygdala

forms stronger associations between items and emotions than between items and context, leading to longer retention of emotionally intense experiences (Yonelinas and Ritchey, 2015). This effect is further amplified as emotional arousal increases, promoting more robust memory consolidation (LaBar and Cabeza, 2006). Additionally, memory and mood interact through multiple mechanisms, including mood dependence and mood congruence. Mood dependence refers to the tendency for information learned in a particular mood to be more easily recalled when the same mood is experienced again (Kuiken, 1991). In contrast, mood congruence describes the tendency for emotionally congruent information to be more readily remembered when it aligns with the current mood (Blaney, 1986). These mood-related memory biases can reinforce negative memory schemas in mood disorders, contributing to the persistence of symptoms (Tyng et al., 2017; Lewis and Critchley, 2003). Recently, there has been growing interest in computational approaches that integrate emotional factors into human-like memory systems, aiming to capture the nuanced ways in which memory, emotion, and behavior interact (Huang et al., 2024).

2.2 Persona and Memory Systems in Autonomous Agents

Autonomous agents are self-directed systems that sense their environment and act accordingly based on the information they perceive (Franklin and Graesser, 1997). They typically consist of four main modules: Profile, Memory, Planning, and Action (Wang et al., 2024a). Among these, the Profile and background information collectively form the agent’s Persona, which defines its unique identity and plays a critical role in creating realistic, human-like simulations that enhance user immersion (Chen et al., 2023; Sun et al., 2024b). Persona has been utilized in various contexts, including the recreation of historical figures (Shao et al., 2023), the generation of fictional characters (Li et al., 2023a; Ahn et al., 2024), and the modeling of real-world social dynamics based on demographic data (Mitsopoulos et al., 2024; Chuang et al., 2024b; Wang et al., 2023). However, maintaining long-term consistency in persona remains a significant challenge, and various approaches have been proposed to address this issue (Madotto et al., 2019; Chen et al., 2023; Ahn et al., 2024).

Human-like Memory modules such as short-term and long-term memory structures (Zheng

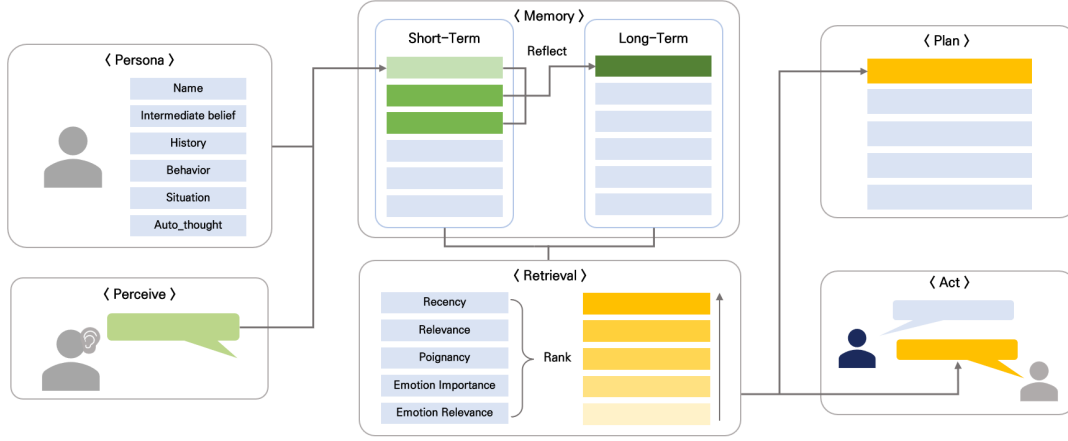


Figure 2: Architecture of the Agent Module

et al., 2024) are also crucial for making agents behave more like humans. Furthermore, there have been attempts to enhance memory retrieval performance through emulating human memory mechanisms (Hou et al., 2024; Park et al., 2023). In addition to Persona and Memory, Planning and Action modules are essential for autonomous agents to set goals and execute actions independently. There are also ongoing efforts to improve their performance through reinforcement learning with natural language (Shinn et al., 2023) and external knowledge (Yao et al., 2023).

2.3 Multi-Agent Interaction and Social Dynamics

Agent-Based Modeling (ABM) has been widely used to model collective opinion formation, behavioral change, and social dynamics by representing individual agents as quantified state values interacting through predefined rules (Terna, 2000; Flache et al., 2017). However, traditional ABM approaches often struggle to capture the rich complexity of human communication, including linguistic nuance, emotional responses, and identity-based interactions (Flache et al., 2017). To address these limitations, recent research has focused on LLM-based multi-agent simulation, which significantly extends the expressive power of traditional ABM by leveraging natural language understanding and generation capabilities (Gao et al., 2023). Building on the foundation of Generative Agents (Park et al., 2023), recent studies have developed agents that evolve emotions, cognition, and personality over time (Li et al., 2024), incorporate internal drives and social connectedness (Wang et al., 2023), and simulate social dynamics by modulating cogni-

tive biases and memory structures (Chuang et al., 2024a).

2.4 Realistic Client Simulation for Counseling

Counseling using LLMs is a recent research trend, but existing datasets face challenges due to privacy concerns and the lack of high-quality, realistic data (Lee et al., 2024; Qiu and Lan, 2024). To overcome these challenges, (Lee et al., 2024) and (Qiu and Lan, 2024) proposed a method to create synthetic data using LLMs to generate virtual counseling data. Client simulation for counselor training is another related research trend, which offers advantages in terms of cost and safety compared to traditional training methods (Demasi et al., 2019; Wang et al., 2024b; Yang et al., 2025). Beyond these trends, other approaches include frameworks that transform expert feedback into prompts for more accurate client responses (Louie et al., 2024), and studies focused on building clients to evaluate the effectiveness of Digital Therapeutics (DTx) (Reichenpfader and Denecke, 2024), highlighting the growing interest in creating realistic, data-driven training environments for counseling.

3 Method

The proposed framework consists of an **Agent Module** and a **Simulation Module**, which together manage the overall simulation environment. The Agent Module is composed of a **Memory System** and a **Persona System**, allowing each agent to realistically emulate predefined character traits and behaviors. In contrast, the Simulation Module coordinates multiple Agent Modules, as well as **Location and Conversation Systems**, in parallel, ensuring that agent behaviors and interactions are seam-

lessly integrated based on their personas, resulting in more natural and contextually appropriate simulations.

3.1 Agent Module

The Agent Module serves as the core component that enables agents to interact with their environment, learn from their experiences, and adapt their behavior over time. The overall structure of the Agent Module is illustrated in Figure 2.

3.1.1 Memory System

The Memory System is responsible for storing and managing an agent’s experiences, both in the short and long term.

- **Short-Term Memory (STM)** temporarily stores recent interactions, events, and plans, allowing agents to quickly recall the recent contexts. It consists of four main types: **Event**, **Chat**, **Day Plan**, and **15-Minute Plan**, each designed to handle specific types of short-term information.

- **Long-Term Memory (LTM)** is designed to preserve experiences that have long-term significance, such as emotional changes, cognitive shifts, and personal growth. LTM contains only a single type, **Thought**, which is generated through the Reflection of recent experiences in STM and the summary of conversations.

- **Memory Retrieval (STM+LTM)** integrates various factors like **Recency**, **Poignancy**, **Relevance**, **Emotion Score**, and **Emotion Relevance** to identify the most relevant memories from both STM and LTM. Typically, it selects the top 5 memories from them, resulting in a final set of 10 memories that can be used for planning, decision-making or further interaction.

Detailed information about the memory system can be found in Appendix A.1.

3.1.2 Persona System

The Persona System is responsible for defining the core traits, current emotional state, and history of each agent. These personas are generated using LLMs based on real-world experience data and serve to simulate diverse characters within the virtual environment. Personas form the foundation for planning, reflection, and conversation, guiding how agents respond to different situations and interact with others. Additionally, these personas evolve dynamically based on the agent’s accumulated experiences within the simulation, allowing for realistic personality changes over time.

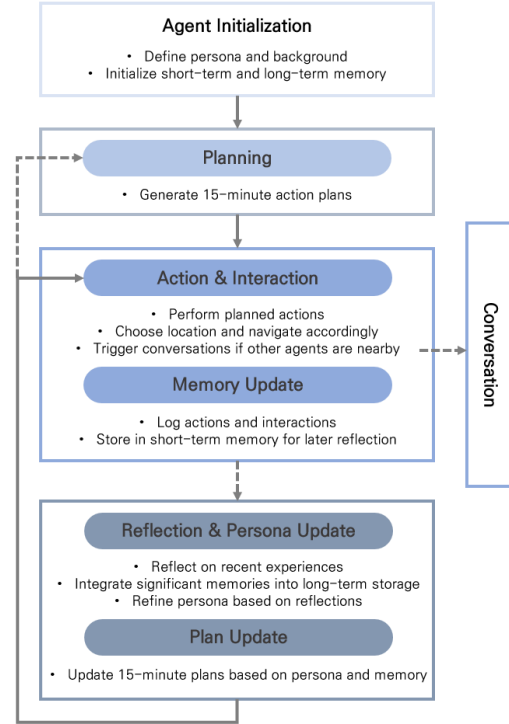


Figure 3: Simulation cycle through Planning, Action, Memory Update, Reflection, and Plan Update.

3.2 Simulation Module

The Simulation Module is a core component that manages the daily behaviors, interactions, and self-reflection processes of agents. It operates in a cyclical structure with multiple stages, including planning, interaction, memory management, and persona updates. The details of each stage are provided in the following sections, and the overall structure is illustrated in Figure 3.

(1) **Planning.** In this stage, the agent formulates a plan for the next 15 minutes based on the recursively decomposed planning approach proposed in previous studies (Park et al., 2023; Wang et al., 2023). The agent generates context-appropriate actions by considering its current persona, surrounding environment, and short-term and long-term memories.

(2) **Action & Interaction.** During this stage, the agent executes the planned activities by moving to the appropriate location and initiating predefined actions. If another agent is present at the selected location, the conversation module is triggered, allowing spontaneous interactions to take place.

(3) **Memory Update.** In this stage, the agent stores the actions performed and conversations it has participated in within the STM. These memories play a critical role in the subsequent Reflection

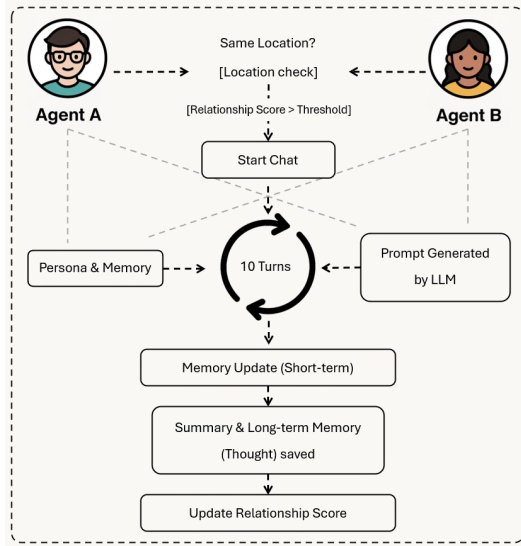


Figure 4: Conversation System Pipeline

& Persona Update stage, contributing to the agent’s short-term experience accumulation.

(4) Reflection & Persona Update. When a pre-defined threshold of memory generation is reached, the agent enters the reflection phase, evaluating recent actions and conversations. This process generates reflections that are stored in LTM and used to update the agent’s persona. This stage captures long-term personality development based on real experiences, enabling agents to evolve beyond simple repetitive behaviors.

(5) Plan Update. In this final stage, the agent reviews and potentially revises its existing plans based on the updated persona, recent behavioral patterns, and environmental context. This allows the agent to adapt to changing circumstances.

3.2.1 Location System

This simulation framework models the physical environment in which agents operate as a hierarchical graph, allowing for more realistic spatial interactions. Agents determine their next location based on their current position and plan using natural language responses generated by an LLM. This approach reflects the intentional planning observed in human behavior. For a detailed overview of the specific conditions used in this framework, please refer to the Appendix A.2.

3.2.2 Conversation System

The Conversation System is a core component designed to simulate natural interactions between agents. As shown in Figure 4, a conversation is initiated probabilistically based on the relationship

score when agents occupy the same physical space. When a conversation begins, each agent generates responses based on its persona and memory, allowing for contextually appropriate interactions. These responses are stored in the agent’s STM. The dialogue continues for up to 10 turns to prevent excessive memory accumulation and to ensure balanced resource usage across agents. After the conversation ends, each agent summarizes the interaction based on its persona and stores it in the LTM, while also updating the relationship score with the other agent based on the conversation content. This structure is essential for modeling realistic social interactions and the dynamic evolution of relationships within the simulation.

4 Experiment

4.1 Agent Information

Client Agents. For realistic client simulations, this study utilized the Patient-Psi Dataset (Wang et al., 2024b), which is derived from actual therapy sessions summarized by GPT-4 Turbo and manually structured by expert clinicians into detailed cognitive models. This dataset includes 106 cognitive models in English, each containing key client information such as intermediate beliefs, histories, behaviors, situations, and automatic thoughts. In this study, we used the behavior sections of these cognitive models to initialize the memories of client agents. Their initial personas were constructed based on their histories and generated memories. To generate initial plans, we considered not only histories and behaviors but also situations and automatic thoughts, ensuring a comprehensive representation of each client’s cognitive context. For consistency and ease of evaluation, we selected only cognitive models specifically related to insomnia, resulting in a final set of three client agents for the simulations.

Counselor and Layperson Agents. To enable meaningful social interactions, we also implemented Counselor Agents and Layperson Agents. These agents were designed to promote persona updates in client agents through repeated interactions. Unlike client agents, they were manually configured using predefined memories and situations, rather than real-world data. Layperson Agents primarily engage in casual conversations without specialized therapeutic skills, providing only surface-level empathy. Their interactions are generally supportive but lack the deeper cognitive

insights required for formal counseling. In contrast, Counselor Agents are specifically designed to use Socratic questioning techniques (Paul and Elder, 2019; Paul and Binker, 1990) to identify and challenge distorted cognitions in client agents’ speech.

4.2 Experimental Conditions

To examine the impact of counseling within the simulation, each client agent was tested under three experimental conditions:

(1) Client-only Condition. Client agents run the simulation alone without interacting with other agents, relying solely on their persona and memory to respond to events and progress through the simulation.

(2) Client-Layperson Condition. This condition includes a layperson agent within the environment, allowing agents to form relationships through conversations and interactions, thus providing a more realistic representation of social interactions.

(3) Client-Counselor Condition. This condition introduces a counselor agent capable of Socratic questioning to guide psychological insight. It extends beyond simple emotional interactions, creating an environment where external support can influence cognitive restructuring, emotional shifts, and self-awareness in the client agents.

Detailed parameter settings for each condition are provided in Appendix A.3 and A.4.

4.3 Evaluation Methods

To assess the realism and effectiveness of this framework, both automated and human evaluations were conducted.

4.3.1 Automated Evaluation

SRS Evaluation. Counseling effectiveness was assessed using the Session Rating Scale (SRS) (Campbell and Hemsley, 2009; Duncan et al., 2003) and G-eval (Semnani et al., 2023; Liu et al., 2023; Chiang and yi Lee, 2023) methods. The SRS measures four key aspects of the counseling relationship: therapeutic alliance, goals and topics, approach and method, and overall satisfaction. It provides a direct assessment of the client-counselor relationship quality. Additionally, G-eval uses a large language model as the evaluator, capable of automated scoring with accuracy comparable to human raters. After each conversation, G-eval was applied to evaluate SRS scores between agents, providing a 1-to-10-point rating based on context.

Activity Pattern Analysis. Following the methodology of (Wang et al., 2023), plan data from pre- and post-simulation runs for all three client agents were collected and segmented into 15-minute behavioral sequences for K-means clustering analysis. A total of 722 action data points (before: 278, after: 494) were analyzed. The optimal number of clusters was determined using the Elbow Method, allowing for a comparative analysis of changes in activity patterns before and after counseling sessions.

4.3.2 Human Evaluation

The study involved 60 participants, including 30 general participants and 30 experts from psychology and medical fields, to evaluate three aspects of simulation. The general participant group focused on evaluating agent’s memory retrieval while the expert group assessed persona consistency and persona update. All evaluations used a 1-to-5 Likert scale, where higher scores indicate better performance. Further details regarding the human evaluation can be found in Appendix A.5.

Memory Retrieval Evaluation. General participants assessed the impact of emotion-based memory retrieval by comparing responses generated by the emotion-aware memory agent with those generated using the memory retrieval function from prior studies (Park et al., 2023). Evaluation criteria included Response Appropriateness and Memory Relevance, aiming to determine whether emotion-aware memory retrieval leads to more human-like thought processes.

Persona Consistency Assessment. Experts evaluated the degree to which generated client and counselor agents resembled real-world clients and counselors. They reviewed persona profiles and conversation transcripts from client agents, rating them across four dimensions: Client Data Consistency (*alignment with initial persona data*), Similarity to Real Clients (*resemblance to real client responses*), Client Coherence (*internal consistency in client replies*), and Similarity to Real Counselors (*alignment with real counselor speech patterns*). This assessment aimed to evaluate how closely the agents’ thought patterns and conversational behaviors mirrored those of real clients and counselors, considering both initial persona settings and ongoing dialogue consistency.

Persona Update Evaluation. In this session, experts evaluated the persona shifts of simulated clients before and after counseling. This evaluation examined the impact of counseling on persona updates by comparing the three primary simulation conditions: **Client-only**, **Client-Layperson**, and **Client-Counselor**.

5 Result & Discussion

5.1 Automated Evaluation

SRS Evaluation. Conversations between client agents and counselor agents (n=55) or layperson agents (n=66) were analyzed using the SAS, which evaluates four key dimensions: relationship, goals, approach, and overall satisfaction. Significant differences ($p < .01$) were observed across all dimensions, with average scores of 9.71, 9.90, 8.57, and 4.39 for counselor interactions, compared to 5.87, 1.96, 4.13, and 1.66 for layperson interactions. These results indicate that counselor agents facilitated more effective therapeutic dialogues and established stronger rapport with client agents, despite both agent types attempting empathetic engagements.

However, interpreting SRS scores requires caution. A total SRS score below 36, or an individual item score below 9, may indicate inadequate relationship formation within the session (Duncan et al., 2003). While counselor agents generally received higher scores, some sessions still fell below these thresholds, particularly in the overall satisfaction dimension, which had the lowest average score. This pattern may be partially attributed to the experimental constraint that limited conversation length to a maximum of 10 turns, potentially cutting off critical discussions prematurely.

Activity Pattern Analysis. Clustering analysis of 722 action data points identified six behavior clusters, reordered based on semantic similarity across agents. Cluster proportions were normalized to compare pre- and post-simulation behaviors. Notably, C0 (*emotion and thought immersion*) and C3 (*routine recovery and emotional regulation*) increased from 16.0% to 30.3% and 11.5% to 14.8%, respectively, indicating more reflective, recovery-oriented behaviors. In contrast, the C1 (*sleep failure and fatigue*) cluster decreased from 23.6% to 19.3%, suggesting a reduction in negative emotional states.

Clustering analysis further highlighted the counselor's impact. In one case, the C5 (*asleep*) cluster

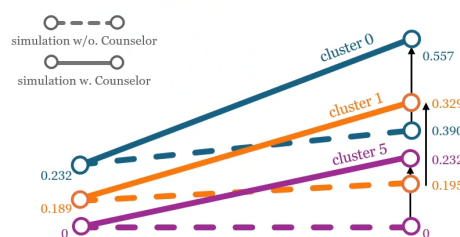


Figure 5: Effect of Counselor Agent on Cluster Transitions of Client Agent

was activated only during counselor interactions, reflecting improved sleep-related behavior. Another case showed a notable increase in relationship-focused cognition (15.8% to 28.7%) after counselor engagement, indicating a greater willingness to confront sensitive social issues. Overall, simulations with a counselor agent promoted recovery-oriented behaviors and reduced negative affect, emphasizing the importance of psychotherapy interventions. Figure 5 visualizes these cluster changes before and after the simulation.

5.2 Human Evaluation

Memory Retrieval Evaluation. The impact of emotion-based memory retrieval was evaluated by 30 general participants. While the emotion-based approach received slightly higher average scores, the differences were not statistically significant compared to the baseline method. Specifically, Response Appropriateness scores were 4.39 vs. 4.28 ($p = .255$), and Memory Relevance scores were 4.23 vs. 4.13 ($p = .239$). The minimal difference in scores may be due to the generally high performance of both methods, making it challenging for participants to perceive a clear distinction. This is supported by the fact that both methods achieved average scores above 4 on a 5-point scale. Additionally, open-ended survey responses revealed that 10 out of 30 participants reported no noticeable difference between the two approaches, while 9 participants noted that the emotion-based method produced responses that felt more empathetic. These findings suggest that emotional memory can contribute to more human-like responses, aligning with the goal of creating contextually appropriate and emotionally resonant interactions.

Persona Consistency Assessment. A separate evaluation by 30 experts assessed dialogues between three client agents and a counselor agent across four key metrics. The average scores were

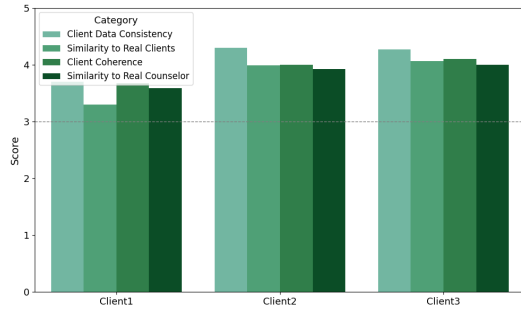


Figure 6: Evaluation of Persona Consistency in Client and Counselor Agents

4.09, 3.79, 3.92, and 3.84, indicating that, while the generated agents do not perfectly replicate real clients and counselors, they demonstrate a reasonable level of realism from the perspective of human evaluators. These scores also highlight the broader challenge of defining and evaluating what constitutes "client-like" or "counselor-like" behavior, given the inherent diversity in real-world clinical interactions. Figure 6 presents a visualization of the average scores for the three client and counselor agents, illustrating their performance across the four evaluation metrics.

Persona Update Evaluation. An evaluation by 30 experts assessed persona changes in three agents across three different simulation environments. The average persona change scores were 2.97, 3.20, and 4.10, indicating a statistically significant difference among the three simulation environments ($F(2, 87) = 22.01, p < .001$). Post-hoc analysis using Tukey’s HSD showed that the difference between the agent-only and agent-layperson conditions was not statistically significant ($q = 1.83, p > .05$). However, the agent-counselor condition demonstrated a statistically significant improvement compared to both the agent-only ($q = 8.88, p < .001$) and agent-layperson ($q = 7.05, p < .001$) conditions. Detailed analyses of individual agents and evaluations of persona update are provided in Appendix C.1. Figure 7 presents a summary of expert-rated persona changes before and after the simulations, where scores above 3 indicate positive persona changes, while scores below 3 reflect negative changes.

These findings indicate that simulations can effectively induce meaningful persona changes in client agents, particularly in mental health contexts. The significantly larger improvements observed in the agent-counselor condition suggest that substantial persona shifts are more likely to occur when

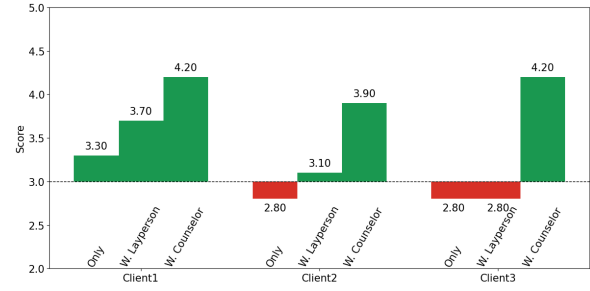


Figure 7: Evaluation of Persona Shift in Client Agents

meaningful counseling interventions are integrated into the simulation environment. This underscores the potential of such simulations as practical tools for predicting and validating counseling effects, highlighting their value in both clinical research and therapeutic training.

6 Conclusion

This study presents a multi-agent simulation framework designed to realistically replicate clients’ emotional and cognitive changes induced by psychotherapy. The framework integrates memory structures enriched with emotional context, allowing agents to engage in more human-like interactions. Client agents with insomnia symptoms and counselor agents were developed based on this approach to simulate realistic persona changes through interactive sessions. Experimental results show that memory retrieval incorporating emotional elements not only facilitates the extraction of contextually relevant memories but also generates more empathetic and realistic responses. Additionally, the findings indicate that only interactions involving counselor agents resulted in significant persona and cognitive changes in client agents, emphasizing the critical role of meaningful therapeutic intervention. These outcomes highlight the potential of this framework as a real-time simulation tool for observing and evaluating the effects of counseling on cognitive and emotional states.

Limitations

The proposed framework provides an environment for pre-observing counseling effects through emotion-based agent simulations, but it also has several limitations. First, the maximum number of conversation turns was capped at 10 due to memory accumulation and resource usage. This restriction may limit the framework’s ability to capture the longer, more iterative nature of real-world coun-

selling sessions, which often involve more extensive exchanges. Additionally, the weighting factors used in the memory retrieval function were set to arbitrary values, and it is unclear whether these values represent the most effective settings. This lack of fine-tuning could reduce the accuracy and realism of emotionally contextualized agent responses. Furthermore, as noted in previous studies (Park et al., 2023; Mitsopoulos et al., 2024; Lin et al., 2023), multi-agent simulations still face challenges in fully replicating the complexity of human psychology and behavior, making it difficult to guarantee that simulation outcomes will precisely match real-world interactions. In particular, real clients often require longer periods of intervention to achieve meaningful psychological changes, creating a potential gap between simulated and actual therapeutic outcomes.

To address these limitations, future studies could explore several key improvements. One approach is to remove the turn limit, allowing for longer, more realistic counseling sessions that better reflect the depth and duration of real-world therapeutic interactions. Another important direction is to conduct further experiments to fine-tune the weighting factors used in memory retrieval, potentially improving the precision of emotionally enriched responses. Finally, increasing the number of agents and expanding the diversity of simulated environments could help capture a wider range of social contexts, providing a more comprehensive evaluation of the therapeutic effects observed in simulations.

Ethical Considerations

Safety. This study aims to support individuals experiencing mental health challenges and novice counselors by providing a simulation that enables preliminary observation of therapeutic effects. However, the outcomes of the simulation should not be blindly trusted or treated as universally applicable. The effectiveness of counseling for real clients may vary depending on individual traits, environments, and personal backgrounds. Therefore, we recommend that any therapeutic strategies demonstrated in the simulation be used as a reference under the supervision of trained professionals, rather than being followed uncritically. Furthermore, this study involved human evaluators. To ensure their safety, the research team obtained approval from an Institutional Review Board (IRB) and verified that

the evaluation environment met ethical and safety standards.

Privacy. Given the highly sensitive nature of mental health counseling, ensuring the privacy of related data is paramount. In constructing the client agents, the research team utilized a dataset from prior work that had undergone prior ethical review, thereby minimizing privacy concerns. Furthermore, the research team obtained permission from the dataset owners for its use in this study. Additionally, all agent names used in the simulation were pseudonyms created by the research team to further ensure anonymity and confidentiality.

Bias. While the study is based on GPT-4o, it is important to acknowledge that large language models may carry inherent societal biases and stereotypes present in the data on which they were trained. Users of this system should remain aware that such biases may lead to skewed or unbalanced simulation outcomes, particularly in relation to certain demographic groups.

Acknowledgments

References

- Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. [TimeChara: Evaluating point-in-time character hallucination of role-playing large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.
- John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. [An integrated theory of the mind](#). *Psychological review*, 111 4:1036–60.
- J.R. Anderson. 2014. *Rules of the Mind*. Taylor & Francis.
- Paul Blaney. 1986. [Affect and memory. a review](#). *Psychological bulletin*, 99:229–46.
- Alistair Campbell and Samantha Hemsley. 2009. [Outcome rating scale and session rating scale in psychological practice: Clinical utility of ultra-brief measures](#). *Clinical Psychologist*, 13(1):1–9.
- Ruijun Chen, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2023. [Learning to memorize entailment and discourse relations for persona-consistent dialogues](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’23/IAAI’23/EAAI’23*. AAAI Press.

- Cheng-Han Chiang and Hung yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) *Preprint*, arXiv:2305.01937. 818
- Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024a. [Simulating opinion dynamics with networks of LLM-based agents](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics. 819
- Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024b. [Beyond demographics: Aligning role-playing LLM-based agents using human belief networks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, Miami, Florida, USA. Association for Computational Linguistics. 820
- Orianna Demasi, Marti A. Hearst, and Benjamin Recht. 2019. [Towards augmenting crisis counselor training by improving message retrieval](#). In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics. 821
- Barry Duncan, Scott Miller, Jacqueline Sparks, D.A. Claud, Lisa Reynolds, J. Brown, and LD Johnson. 2003. The session rating scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy*, 3:3–12. 822
- Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. [Models of social influence: Towards the next frontiers](#). *Journal of Artificial Societies and Social Simulation*, 20(4):2. 823
- Stan Franklin and Art Graesser. 1997. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Intelligent Agents III Agent Theories, Architectures, and Languages*, pages 21–35, Berlin, Heidelberg. Springer Berlin Heidelberg. 824
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. [S3: Social-network simulation system with large language model-empowered agents](#). *Preprint*, arXiv:2307.14984. 825
- Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. 2003. [Instance-based learning in dynamic decision making](#). *Cognitive Science*, 27(4):591–635. 826
- J. Hogeveen, C. Salvi, and J. Grafman. 2016. ‘emotional intelligence’: Lessons from lesions. *Trends in Neurosciences*, 39(10):694–705. 827
- Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. [“my agent understands me better”: Integrating dynamic human-like memory recall and consolidation in llm-based agents](#). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI ’24, page 1–7. ACM. 828
- Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. [Emotional rag: Enhancing role-playing agents through emotional retrieval](#). *Preprint*, arXiv:2410.23041. 829
- D. Kuiken. 1991. *Mood and Memory: Theory, Research and Applications*. SAGE Publications. 830
- Kevin LaBar and Roberto Cabeza. 2006. [Labar ks, cabeza r. cognitive neuroscience of emotional memory](#). *nat rev 7*: 54–64. *Nature reviews. Neuroscience*, 7:54–64. 831
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics. 832
- Penelope A. Lewis and Hugo D. Critchley. 2003. [Mood-dependent memory](#). *Trends in Cognitive Sciences*, 7(10):431–433. 833
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023a. [Chatharuhi: Reviving anime character in reality via large language model](#). *Preprint*, arXiv:2308.09597. 834
- Jiale Li, Jiayang Li, Jiahao Chen, Yifan Li, Shijie Wang, Hugo Zhou, Minjun Ye, and Yunsheng Su. 2024. [Evolving agents: Interactive simulation of dynamic and diverse human personalities](#). *Preprint*, arXiv:2404.02718. 835
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. [Agent hospital: A simulacrum of hospital with evolvable medical agents](#). *Preprint*, arXiv:2405.02957. 836
- Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. [Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents](#). *Preprint*, arXiv:2310.06500. 837
- Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiyue Ping, and Qin Chen. 2023. [Agentsims: An open-source sandbox for large language model evaluation](#). *Preprint*, arXiv:2308.04026. 838
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chengguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on* 839

873	<i>Empirical Methods in Natural Language Processing</i> ,	<i>Empirical Methods in Natural Language Process-</i>	928
874	pages 2511–2522, Singapore. Association for Com-	ing, pages 13153–13187, Singapore. Association for	929
875	putational Linguistics.	Computational Linguistics.	930
876	Ryan Louie, Ananjan Nandi, William Fang, Cheng	Noah Shinn, Federico Cassano, Ashwin Gopinath,	931
877	Chang, Emma Brunskill, and Diyi Yang. 2024.	Karthik R Narasimhan, and Shunyu Yao. 2023. Re-	932
878	Roleplay-doh: Enabling domain-experts to create	flexion: language agents with verbal reinforcement	933
879	LLM-simulated patients via eliciting and adhering to	learning . In <i>Thirty-seventh Conference on Neural</i>	934
880	principles . In <i>Proceedings of the 2024 Conference on</i>	<i>Information Processing Systems</i> .	935
881	<i>Empirical Methods in Natural Language Processing</i> ,		
882	pages 10570–10603, Miami, Florida, USA. Associa-	Guangzhi Sun, Xiao Zhan, and Jose Such. 2024a. Build-	936
883	tion for Computational Linguistics.	ing better ai agents: A provocation on the utilisation	937
884		of persona in llm-based conversational agents . In	938
885	Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and	<i>Proceedings of the 6th ACM Conference on Conversa-</i>	939
886	Pascale Fung. 2019. Personalizing dialogue agents	<i>tional User Interfaces</i> , CUI '24, New York, NY,	940
887	via meta-learning . In <i>Proceedings of the 57th Annual</i>	USA. Association for Computing Machinery.	941
888	<i>Meeting of the Association for Computational</i>		
889	<i>Linguistics</i> , pages 5454–5459, Florence, Italy. Asso-	Guangzhi Sun, Xiao Zhan, and Jose Such. 2024b. Build-	942
	ciation for Computational Linguistics.	ing better ai agents: A provocation on the utilisation	943
890		of persona in llm-based conversational agents . In	944
891	Konstantinos Mitsopoulos, Ritwik Bose, Brodie Mather,	<i>Proceedings of the 6th ACM Conference on Conversa-</i>	945
892	Archana Bhatia, Kevin Gluck, Bonnie Dorr, Christian	<i>tional User Interfaces</i> , CUI '24, New York, NY,	946
893	Lebiere, and Peter Pirolli. 2024. Psychologically-	USA. Association for Computing Machinery.	947
894	valid generative agents: A novel approach to agent-		
895	based modeling in social sciences . <i>Proceedings of</i>	Pietro Terna. 2000. How to build and use agent-based	948
	<i>the AAAI Symposium Series</i> , 2(1):340–348.	models in social science . <i>Mind and Society</i> , 1:57 –	949
		72.	950
896	Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Mered-	Chai M. Tyng, Hafeez U. Amin, Mohamad N. M. Saad,	951
897	ith Ringel Morris, Percy Liang, and Michael S. Bern-	and Aamir S. Malik. 2017. The influences of emotion	952
898	stein. 2023. Generative agents: Interactive simulacra	on learning and memory . <i>Frontiers in Psychology</i> ,	953
899	of human behavior . In <i>Proceedings of the 36th Annual</i>	Volume 8 - 2017.	954
900	<i>ACM Symposium on User Interface Software</i>		
901	<i>and Technology</i> , UIST '23, New York, NY, USA.	Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao	955
902	Association for Computing Machinery.	Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang,	956
		Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei,	957
903	Richard Paul and Linda Elder. 2019. <i>The thinker's guide</i>	and Jirong Wen. 2024a. A survey on large language	958
904	<i>to Socratic questioning</i> . Rowman & Littlefield.	model based autonomous agents . <i>Frontiers of Com-</i>	959
		<i>puter Science</i> , 18(6).	960
905	Richard W. Paul and AJA Binker. 1990. <i>Critical think-</i>	Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin	961
906	<i>ing: What every person needs to survive in a rapidly</i>	Zhi, Shaun M. Eack, Travis Labrum, Samuel M Mur-	962
907	<i>changing world</i> . ERIC.	phy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang,	963
		and Zhiyu Chen. 2024b. PATIENT-ψ: Using large	964
908	Huachuan Qiu and Zhenzhong Lan. 2024. Interactive	language models to simulate patients for training	965
909	agents: Simulating counselor-client psychological	mental health professionals . In <i>Proceedings of the</i>	966
910	counseling via role-playing llm-to-llm interactions .	<i>2024 Conference on Empirical Methods in Natural</i>	967
911	<i>Preprint</i> , arXiv:2408.15787.	<i>Language Processing</i> , pages 12772–12797, Miami,	968
		Florida, USA. Association for Computational Lin-	969
912	Daniel Reichenpfader and Kerstin Denecke. 2024. Sim-	guistics.	970
913	ulating diverse patient populations using patient vi-		
914	gnettes and large language models . In <i>Proceedings</i>	Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023.	971
915	<i>of the First Workshop on Patient-Oriented Language</i>	Humanoid agents: Platform for simulating human-	972
916	<i>Processing (CL4Health) @ LREC-COLING 2024</i> ,	like generative agents . <i>Preprint</i> , arXiv:2310.05418.	973
917	pages 20–25, Torino, Italia. ELRA and ICCL.		
918	Sina Semnani, Violet Yao, Heidi Zhang, and Monica	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	974
919	Lam. 2023. WikiChat: Stopping the hallucination of	Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,	975
920	large language model chatbots by few-shot ground-	and Denny Zhou. 2022. Chain-of-thought prompt-	976
921	ing on Wikipedia . In <i>Findings of the Association</i>	ing elicits reasoning in large language models. In	977
922	<i>for Computational Linguistics: EMNLP 2023</i> , pages	<i>Proceedings of the 36th International Conference on</i>	978
923	2387–2413, Singapore. Association for Computa-	<i>Neural Information Processing Systems</i> , NIPS '22,	979
924	tional Linguistics.	Red Hook, NY, USA. Curran Associates Inc.	980
925	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu.	Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe	981
926	2023. Character-LLM: A trainable agent for role-	Yuan, Jun Zhang, Jie Feng, Fengli Xu, and Yong Li.	982
927	playing . In <i>Proceedings of the 2023 Conference on</i>	2025. Opencity: A scalable platform to simulate	983
		urban activities with massive LLM agents .	984

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, John Pinto, Jenny Giam, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee peng Lim. 2025. [Consistent client simulation for motivational interviewing-based counseling](#). *Preprint*, arXiv:2502.02802.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). *Preprint*, arXiv:2210.03629.

Andrew P. Yonelinas and Maureen Ritchey. 2015. [The slow forgetting of emotional episodic memories: an emotional binding account](#). *Trends in Cognitive Sciences*, 19(5):259–267.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. [Simulating classroom education with llm-empowered agents](#). *Preprint*, arXiv:2406.19226.

Shijie Zheng, Keith He, Le Yang, and Jie Xiong. 2024. [Memoryrepository for ai npc](#). *IEEE Access*, 12:62581–62596.

A Additional Description

This section provides detailed descriptions of the modules and systems that compose our simulation.

A.1 Memory System

The memory system is divided into STM and LTM, each optimized for different timescales and types of cognitive processing.

A.1.1 Memory Types

Short-Term Memory (STM) STM is responsible for temporarily storing recent interactions, immediate plans, and short-term goals. It consists of four primary memory types:

- **Event Memory.** Stores discrete, context-specific experiences, including routine observations and significant moments. Each Event Memory is assigned a unique node_id and includes key attributes such as Poignancy (1-10 scale), Emotion, and Emotion Intensity. These attributes allow the agent to assess the psychological impact of each event, supporting more contextually appropriate responses during retrieval.
- **Chat Memory.** Records conversational exchanges, including both the agent’s own utterances and those of conversation partners. Each message is tagged with attributes like Emotion, Poignancy, and Emotion Intensity, providing

a richer context for understanding emotional tone. Messages are grouped by chat_set_id and message_id, facilitating the analysis of conversation dynamics over time.

- **Day Plan Memory.** Stores structured daily plans, focusing on time management and task organization. Unlike Event and Chat Memories, Day Plan Memories do not include emotional or importance scores, reflecting their more utilitarian nature.
- **15-Minute Plan Memory.** Provides more granular, short-term task planning. These memories capture brief, focused goals without emotional context, optimizing efficiency in task management.

Long-Term Memory (LTM) Preserves long-term experiences and significant cognitive insights, capturing broader emotional changes, cognitive shifts, and personal growth.

- **Thought Memory.** Stores high-level reflections, including significant emotional shifts, cognitive insights, and self-discovery moments. These memories are typically generated when short-term memories in STM exhibit high Poignancy or Emotion Score, triggering the creation of more deeply reflective entries. This structure aligns with the high-level cognitive processes described by (Park et al., 2023), supporting long-term narrative consistency.

A.1.2 Memory Retrieval Mechanism

The memory retrieval process integrates both STM and LTM, using a combination of semantic similarity, emotional relevance, and recency to identify the most contextually appropriate memories for a given input. The retrieval process involves several key steps:

- (1) **Sentence Embedding.** Each memory is converted into a text embedding at the time of storage, enabling efficient semantic comparison. Retrieval begins by generating an embedding for the input query, which is then compared to stored embeddings using cosine similarity.
- (2) **Emotion Similarity Evaluation.** Emotional tone is a critical component of memory relevance. This step evaluates the emotional alignment between the input query and stored memories, assigning higher scores for similar emotions (e.g., joy-joy) and lower scores for opposing emotions (e.g., joy-sadness). Neutral or unrelated emotions

receive moderate scores, capturing a broader range of emotional nuances.

(3) Memory Ranking. Final memory scores are calculated based on a weighted combination of five core factors: Recency, Poignancy, Relevance, Emotion Score, and Emotion Relevance. The weights used in this study are as follows:

- Recency: 0.1
- Relevance: 0.25
- Poignancy: 0.40
- Emotion Score: 0.15
- Emotion Relevance: 0.10

(4) Final Memory Selection. The top 5 memories from both STM and LTM are selected based on these scores, resulting in a final set of 10 highly relevant memories. This balanced approach ensures that both recent context and long-term insights are considered, supporting more contextually appropriate responses.

A.1.3 Retrieval Score Calculation

The final retrieval score for each memory is calculated using the following formula:

$$\begin{aligned} \text{Total Score} = & (W_{rc} \times \text{Recency}) + (W_p \times \text{Poignancy}) \\ & + (W_{es} \times \text{Emotion Score}) + (W_{erl} \times \text{Emotion Relevance}) \\ & + (W_{rl} \times \text{Relevance}) \quad (1) \end{aligned}$$

where:

- W_{rc} = Weight for Recency
- W_p = Weight for Poignancy
- W_{es} = Weight for Emotion Score
- W_{erl} = Weight for Emotion Relevance
- W_{rl} = Weight for Relevance

This formula integrates multiple contextual factors, ensuring that memory retrieval is both semantically and emotionally aligned with the input query. By balancing short-term context with long-term insights, this approach supports the generation of responses that are contextually appropriate and emotionally resonant.

A.2 Location System

Based on a previous study (Wang et al., 2023), the location system in this simulation framework models the physical environment as a hierarchical graph, representing each agent’s possible locations. This structure captures the spatial relationships between locations, facilitating more realistic agent

interactions. The following section presents an example of the specific location hierarchy used in one of the simulation scenarios.

world_client_friend.yaml

World Structure

```
World
|-- Ethan's house
|   |-- Ethan's Kitchen
|   |-- Ethan's Living Room
|   |-- Ethan's Bedroom
|   |-- Ethan's Garden
|
|-- Theodore's house
|   |-- Theodore's Kitchen
|   |-- Theodore's Living Room
|   |-- Theodore's Bedroom
|   |-- Theodore's Garden
|
|-- Zane's house
|   |-- Zane's Kitchen
|   |-- Zane's Living Room
|   |-- Zane's Bedroom
|   |-- Zane's Garden
|
|-- Johnson Park
|-- Counseling center
|-- Market
```

Agents

- Ethan: Ethan’s Bedroom
- Zane: Zane’s Bedroom

A.3 Model Settings

In all simulation conditions, each agent primarily utilizes the GPT-4o model for advanced reasoning and natural language processing. For simpler tasks such as emotion classification and intensity measurement, the GPT-4o-mini model is used. Additionally, the text-embedding-ada-002 model is employed for memory embedding, ensuring accurate retrieval of semantically relevant memories based on inputs.

A.4 Simulation Setup

In the simulation, one full cycle of interaction was defined as a "turn." Each client agent underwent 200 turns per condition. Each turn corresponded to 15 minutes of simulated time. The simulation was accelerated at a rate of three times relative to real time, meaning that one turn took approximately 5 minutes in real-world time. Consequently,

each simulation ran for approximately 16.7 hours in real time, corresponding to roughly 50 hours of simulated time.

A.5 Detailed Description of Human Evaluation

The human evaluation consisted of a total of 60 Korean participants, comprising 30 general evaluators and 30 expert evaluators. General evaluators were undergraduate and graduate students from various academic backgrounds, while expert evaluators were graduate students majoring in psychology or counseling, as well as individuals with clinical experience in counseling. The number of evaluators was determined through a priori power analysis, assuming a medium effect size (Cohen’s $d = 0.5$), with a significance level of 0.05 and statistical power of 0.95. Participants were recruited via a survey link and took part in the experiment either online or offline. As compensation for their participation, general evaluators received a gift voucher worth 10,000 KRW, while expert evaluators received a gift voucher worth 30,000 KRW. Participants were informed about the purpose of the study and were given an explanation of how their evaluations would be used.

Each evaluator was presented with utterance logs and persona logs generated from the simulation and was asked to respond to evaluation items using a 5-point Likert scale. In the case of memory retrieval evaluation, additional qualitative feedback was collected through open-ended questions regarding their subjective impressions. To ensure accurate assessments and account for the potential language barriers, all data were translated into Korean and provided in the participants’ native language.

B Simulation Example

Tables 1-4 contain example outcomes from the simulation. All names appearing in the examples are pseudonyms designated by the authors to ensure anonymity.

C Evaluation Result

Tables 5-7 show additional evaluation results for automatic and human evaluation.

C.1 Additional analysis of persona shift

Experts not only evaluated persona shifts of different conditions, but also assessed whether the observed persona shifts in the simulations were

realistically achievable in real-world clients. The average scores for this assessment were 3.62, 3.80, and 3.70 for the agent-only, agent-layperson, and agent-counselor conditions, respectively. These results suggest that the persona shifts observed in the simulations reflect plausible real-world outcomes, supporting the ecological validity of the simulated interventions. However, it is important to note that the extent and nature of persona changes can vary significantly depending on individual client characteristics and contextual factors. This variation indicates that not all clients will respond uniformly to similar interventions, emphasizing the need for personalized approaches in both simulation design and real-world therapeutic settings.

00:15 am: try to get some rest and sleep 00:30 am: continue efforts to sleep
 00:45 am: continue efforts to sleep 01:00 am: continue efforts to sleep
 01:15 am: continue efforts to sleep 01:30 am: continue efforts to sleep
 01:45 am: continue efforts to sleep 02:00 am: restless and still awake
 02:15 am: restless and still awake 02:30 am: restless and still awake
 02:45 am: restless and still awake 03:00 am: restless and still awake
 03:15 am: wake up briefly and struggle with sleep due to anxious thoughts 03:30 am: lie in bed, trying to calm racing thoughts
 03:45 am: lie in bed, trying to calm racing thoughts 04:00 am: manage to fall back asleep
 04:15 am: manage to fall back asleep 04:30 am: sleep
 04:45 am: sleep 05:00 am: sleep
 05:15 am: sleep 05:30 am: sleep
 05:45 am: sleep 06:00 am: still asleep
 06:15 am: still asleep 06:30 am: wake up and start morning routine (shower, breakfast, etc.)
 06:45 am: continue morning routine 07:00 am: finish breakfast and prepare for the day
 07:15 am: begin work from home, focus on easy tasks 07:30 am: continue working
 07:45 am: continue working on easy tasks 08:00 am: continue working
 08:15 am: stay on task with simple work activities 08:30 am: continue working
 08:45 am: take a short break with deep breathing 09:00 am: resume work with simple tasks
 09:15 am: continue working 09:30 am: take a break with a short walk around the neighborhood
 09:45 am: take a break with a short walk around the neighborhood 10:00 am: resume work with a fresh mindset
 10:15 am: continue work, push to tackle slightly more challenging tasks 10:30 am: stay focused on current tasks
 10:45 am: stay focused on current tasks 11:00 am: take a brief break to clear mind
 11:15 am: resume work 11:30 am: continue working
 11:45 am: stay focused on current tasks 12:00 pm: take a lunch break, enjoy a calming hot drink
 12:15 pm: take a lunch break, enjoy a calming hot drink 12:30 pm: unwind briefly before returning to work
 12:45 pm: unwind briefly before returning to work 01:00 pm: resume work, focus on less demanding tasks to manage stress
 01:15 pm: continue working on manageable tasks 01:30 pm: continue working
 01:45 pm: continue working 02:00 pm: brief distraction period, procrastinate
 02:15 pm: brief distraction period, procrastinate 02:30 pm: refocus attention on work
 02:45 pm: refocus attention on work 03:00 pm: return to work with a renewed focus on one task at a time

Table 1: Sample 15-Minute Plan for the Client Agent

Ethan: I've been feeling really overwhelmed lately with how things are in my family. It seems like no matter how hard I try to keep the peace, I end up feeling even more powerless and alone. It's frustrating to feel like speaking up might just lead to more conflict, but staying quiet makes me feel trapped.

Theodore: It sounds like you're in a tough spot, feeling stuck between wanting to speak up and the fear of more conflict. Can you recall moments when you've navigated family tensions in the past that felt slightly more manageable or offered a sense of relief?

Ethan: Honestly, it's hard to remember a time when dealing with family didn't feel overwhelming. I often feel trapped in a cycle of wanting to stand up for myself yet fearing the fallout every time. The few times I tried to voice my feelings, like with Candace, only left me feeling more isolated and powerless.

Theodore: Ethan, it sounds really challenging to feel both isolated and fearful of backlash when you try to voice your feelings. I wonder if there's been any instance, even if small, where you experienced a moment of connection or understanding despite these difficult dynamics? Such moments can sometimes offer a new perspective or a bit of hope.

Ethan: Honestly, Theodore, it's really hard for me to pinpoint any specific moment of connection or understanding in the midst of all this tension. It often feels like I'm trapped in a cycle where standing up for myself only leads to being more isolated and misunderstood. The fear of backlash keeps me from risking those rare moments that might bring some relief.

Theodore: I'm really sorry to hear that it feels so challenging, Ethan. When you think about the idea of risking those moments that might bring some relief, what thoughts or images come to your mind? Sometimes reflecting on what holds us back can be revealing in itself.

Ethan: Honestly, when I think about taking those risks, I mostly feel a sense of dread and fear of causing drama in the family. It's hard to shake off the idea that speaking up will just end in more conflict and isolation like it did with Candace. I guess I see myself feeling trapped, powerless, and like I'm just destined to maintain the peace at my own expense.

Theodore: It sounds like you're really wrestling with the thought of advocating for yourself, only to fear more isolation and conflict. Considering your courage in sharing this, I'm curious if there's something you need or wish you could receive from your family to feel supported in this challenging situation?

Ethan: Honestly, Theodore, I wish I could feel a sense of support and understanding from my family when I try to speak up. It's hard not to fear the backlash and potential isolation that might follow. Sometimes, I just want to know they're willing to listen without assuming the worst of me.

Theodore: Ethan, it seems like you're longing for a space where your voice is heard and validated without fear. What might it look or feel like for you if your family were to offer that understanding and support you're seeking?

Table 2: Sample Dialogue Between Client (Ethan) and Counselor (Theodore)

Characteristic: Based on the statements provided, Ethan's core characteristics can be described as follows:

1. **Conflict Avoidance and Fear of Rejection:** Ethan seems to prioritize maintaining peace and harmony within her family, even to the point of personal detriment. She has a deep fear of conflict leading to abandonment or disrupting family unity, suggesting a strong desire for acceptance and belonging.

2. **Anxiety and Stress:** Her anxiety about upcoming family events and the loss of sleep indicate that Ethan experiences significant stress related to her family dynamics. This anxiety likely stems from her fear of conflict and feeling trapped in her current situation.

3. **Feelings of Powerlessness and Low Self-Worth:** Ethan feels powerless and incapable of enacting change in her family relationships. Her belief that standing up for herself will worsen situations suggests she has low self-esteem and a lack of confidence in her ability to influence outcomes.

4. **Inner Conflict and Growing Frustration:** While she feels powerless, Ethan also experiences mounting frustration, particularly with specific family members like Candace. This indicates an internal conflict between her desire to maintain peace and her need to address personal grievances.

5. **Tendency Toward Avoidance:** Her procrastination at work may be a coping mechanism, where she avoids responsibilities or tasks due to the emotional burden and mental fatigue from her family issues.

Overall, Ethan appears to be a compassionate individual who struggles with assertiveness and self-advocacy, resulting in stress and emotional turmoil.

Recent feeling: Based on the statements provided, Ethan appears to be struggling with feelings of anxiety, powerlessness, and a lack of control over her recent progress in life. She seems to be experiencing significant stress related to family dynamics, leading to symptoms such as procrastination and sleep loss. Her beliefs about her role in the family suggest that she feels trapped and weak, perceiving herself as unable to assert herself or change her situation without negative consequences. This indicates a sense of helplessness and fear of conflict or rejection, which may contribute to her overall anxiety and stress.

Table 3: Client Persona before the simulation (Day1)

Background:

Characteristic: Ethan's core characteristics, as derived from the statements provided, present a multifaceted picture of someone who is deeply introspective and emotionally aware. Here's a summary based on the information you've given:

1. **Introspective and Self-Aware:** Ethan tends to reflect deeply on his personal experiences and understands the influence of his past, especially familial conflicts, on his current emotional and relational dynamics.
2. **Conflict Avoidant:** He shows a strong aversion to confrontation, a trait that stems from previous negative familial interactions and the fear of further discord or abandonment.
3. **Emotionally Vulnerable:** Ethan deals with low self-esteem and carries a significant fear of rejection, impacting his interactions and making it challenging for him to express his emotions or needs assertively.
4. **Harmony-Oriented:** Prioritizing peace is a significant trait for Ethan, often at the expense of his own needs. This drive arises from his fear that any conflict might lead to isolation or family breakdown.
5. **Commitment to Growth:** Despite internal challenges, Ethan is focused on personal development. He strives to be more assertive and resilient, learning from past patterns while attempting to overcome them.
6. **Internal Conflict:** There exists a consistent struggle within Ethan between the desire to advocate for himself and the fear of potential negative repercussions. This conflict complicates his journey toward empowerment.
7. **Hopeful and Open to Change:** Ethan maintains an optimistic outlook and shows a willingness to change, believing in his capacity for emotional and relational improvement despite occasional feelings of being stuck.
8. **Influence of Supportive Figures:** Support from individuals like Theodore plays an essential role in Ethan's progress. These figures encourage him, provide perspective, and help bolster his optimism.

Overall, Ethan is navigating a delicate balance between self-preservation and self-expression. He is working towards healthier relationships and greater emotional maturity by confronting his fears and slowly disentangling himself from past constraints. His journey is marked by an ongoing effort to develop resilience and assertiveness, with the support and encouragement of those around him, as he moves toward a more empowered state of being.

Recent feeling: Ethan's feelings about his recent progress in life are best described as a blend of cautious optimism and internal struggle. He is learning to voice his needs and take steps toward personal empowerment, but these efforts are accompanied by a sense of vulnerability and fear. The relief he feels when he manages to assert himself is often overshadowed by anxiety about potential disapproval and the consequences of his actions. Despite this, Ethan is committed to growth and is hopeful that each small step will build his confidence and resilience. His introspective nature allows him to see how past family conflicts shape his current emotional responses, which include fears of confrontation and abandonment. These fears make it challenging for him to assert his needs, causing an internal conflict between his desire for self-expression and the preservation of familial harmony. Yet, Ethan remains optimistic about his ability to change and improve, buoyed by support from individuals like Theodore.

Overall, Ethan's journey is marked by a careful navigation between maintaining self-preservation and moving towards self-expression. He is striving for healthier relationships and emotional maturity by addressing his fears and gradually breaking free from past limitations, even as he continues to deal with internal complexities.

Table 4: Client Persona after the simulation with counselor-characteristic (Day3)

	With Therapist	With Layperson
Relationship	9.71	5.87
Goals and Topics	9.90	1.96
Approach and Method	8.57	4.13
Overall	4.39	1.66

Table 5: SAS-Based Evaluation of Counseling Quality in Client–Therapist and Client–Layperson Conversations

	With Emotion(ours)	Without Emotion
Response Appropriateness	4.39	4.28
Retrieved Memory Appropriateness	4.23	4.13

Table 6: Human Evaluation Results for Memory Retrieval With and Without Emotion

Client no. (n=30)	ANOVA Result (F, p)	Environment Comparison	Mean Difference	q Statistic	p-value
Client 1 (n=10)	F(2,27) = 4.65, p = 0.18	client only vs client with layperson	0.4	1.95	>.05
		client only vs client with therapist	0.9	4.38	<.01
		client with layperson vs client with therapist	0.5	2.44	>.05
Client 2 (n=10)	F(2,27) = 6.51, p = 0.005	client only vs client with layperson	0.3	1.35	>.05
		client only vs client with therapist	1.1	4.94	<.01
		client with layperson vs client with therapist	0.8	3.59	<.05
Client 3 (n=10)	F(2,27) = 16.33, p < .001	client only vs client with layperson	0.0	1.83	>.05
		client only vs client with therapist	1.4	8.88	<.001
		client with layperson vs client with therapist	1.4	7.05	<0.001
Total (n=30)	F(2,87) = 22.01, p < .001	client only vs client with layperson	0.23	1.83	>.05
		client only vs client with therapist	1.13	8.88	<.001
		client with layperson vs client with therapist	0.9	7.05	<0.001

Table 7: Overall Human Evaluation Results of Agent Persona Changes Across Different Conditions