An LLM-Based Simulation Framework for Embodied Conversational Agents in Psychological Counseling

Lixiu Wu[*1], **Yuanrong Tang**[*1], **Qisen Pan**[1], **Xianyang Zhan**[1], **Yuchen Han**[1], **Lanxi Xiao**[1], **Tianhong Wang**[1], **Chen Zhong**[1], **Jiangtao Gong**[1†]

[1]AI Industry Research, Tsinghua University, Beijing, China
wulx1102@gmail.com, tangxtong2022@gmail.com, gongjiangtao@air.tsinghua.edu.cn

Abstract

Due to privacy concerns, open dialogue datasets for mental health are primarily generated through human or AI synthesis methods. However, the inherent implicit nature of psychological processes, particularly those of clients, poses challenges to the authenticity and diversity of synthetic data. In this paper, we propose ECAs (short for Embodied Conversational Agents), a framework for embodied agent simulation based on Large Language Models (LLMs) that incorporates multiple psychological theoretical principles. Using simulation, we expand real counseling case data into a nuanced embodied cognitive memory space and generate dialogue data based on high-frequency counseling questions. We validated our framework using the D4 dataset. First, we created a public ECAs dataset through batch simulations based on D4. Licensed counselors evaluated our method, demonstrating that it significantly outperforms baselines in simulation authenticity and necessity. Additionally, two LLM-based automated evaluation methods were employed to confirm the higher quality of the generated dialogues compared to the baselines.

Code —
https://github.com/AIR-DISCOVER/ECAs-Dataset

Introduction

Mental health counseling data plays a crucial role in multiple applications: training novice counselors (Kuehne et al. 2018), developing AI-assisted counseling systems (Furlan et al. 2021; Tang et al. 2025), and automating mental health diagnosis (Ping 2024). However, the highly private and sensitive nature of psychological counseling creates significant barriers to collecting and sharing real counseling dialogue data (Miner et al. 2019), hindering both AI development and professional training in this field.

Current approaches to data scarcity fall into two categories. The first relies on simulated dialogues created by human, which offer high authenticity but are costly and may introduce sampling biases due to limited expert availability and individual perspectives (Schatzmann, Georgila, and Young 2005). The second leverages AI-based data synthesis and augmentation, as demonstrated by domain-specific models like HEAL (Yuan et al. 2024). Although these AI-based approaches provide better scalability, they typically capture only surface-level language patterns rather than the underlying psychological complexity (Kjell, Kjell, and Schwartz 2024). This limitation often results in distribution shift and pattern collapse during self-training loops, leading to progressive model degradation (Shumailov et al. 2024).

Recent advances in LLM-based agent simulation show promising results in generating high-fidelity data for social science research, particularly through embodied agents capable of simulating social experiences and interactions (Liang et al. 2024; Wang et al. 2024b; Wang, Chiu, and Chiu 2023; Zheng et al. 2024). Notable examples include generative agents for social behavior modeling (Park et al. 2023) and socially aligned language models (Liu et al. 2024).

However, simulating psychological counseling presents unique challenges beyond general social simulation. Mental processes, especially those of clients with psychological conditions, are inherently complex and hidden beneath observable behaviors (Sircova et al. 2015). While recent works like patient-Psi (Wang et al. 2024a) and Roleplay-doh (Louie et al. 2024) have made initial progress, effective simulation requires deep integration with psychological theories and counseling principles. Therefore, our core research question is: How can we develop LLM-based simulations grounded in psychological and counseling theories to generate authentic, rich counseling dialogue data that facilitates research and development in this field?

To address these challenges and advance the field of psychological counseling simulation, we introduce and implement a novel framework called ECAs designed to simulate the **E**mbodied **C**onversational **A**gent**s** in psychological counseling, integrating counseling theories of Cognitive Behavioral Therapy (CBT) (Beck 2021) with diagnostic frameworks. Our methodology begins with an extensive review of psychological counseling theories, culminating in the formulation of six essential principles and preliminaries for simulation. Leveraging the capabilities of LLMs, we then develop a sophisticated method to expand real counseling case data, creating a highly realistic and nuanced embodied cognitive memory space. Based on this, we construct counselor and client agents that generate dialogue data simulat-

[*]These authors contributed equally.
[†]Corresponding author.

ing interactions between counselors and clients, carefully crafted based on high-frequency counseling questions. To rigorously assess the efficacy of our ECAs framework, we employ the D^4 (Yao et al. 2022) dataset as a benchmark and involve licensed human counselors in the evaluation process.

Our key contributions include: i) We introduce a novel LLM-based social simulation framework for nuanced embodied conversational agents in psychological counseling. ii) By reviewing psychological theories, we derive six simulation principles for the ECAs framework. iii) The quality and authenticity of our simulations are validated through both human expert evaluation and the D^4 benchmark. iv) We generate a public ECAs dataset to support future research, along with two LLM-based automated methods for evaluating dialogue quality.

Principles and Preliminaries for Simulation

The complexity of psychological counseling simulation necessitates a multi-faceted theoretical foundation, as it requires modeling both intricate human characteristics and professional therapeutic interactions. In this section, we introduce six Simulation Principles and Preliminaries (SPs) that enable ECAs to generate high-fidelity counseling conversations grounded in established psychology and psychotherapy theories to address key simulation challenges.

SP1: Represent Comprehensive Life-Stage Experiences. The primary principle of designing ECAs that authentically represent a client's life experiences across various stages. This involves creating a comprehensive framework for factual memories that incorporates both subjective experiences and event data. The approach aims to capture key long-term memories as well as recent memories, enhancing the understanding of the client's life trajectory. This goal is grounded in embodied cognition theory and research on situational cues in memory recall, which emphasize the importance of comprehensive and authentic representations in psychological counseling

Formally, let $L = l_1, \ldots, l_n$ be the set of life experiences, each with temporal marker t_i and subjective significance s_i. The life-experience representation is $M(c) = f(L, T, S)$, where T and S are the temporal and significance mappings. Experiences are categorized as past or recent; past experiences are further divided by age into Childhood, Adolescence, Youth, and Middle Age. Let L_r, T_r, and S_r denote recent experiences and their mappings. $M(c)$ is represented:

$$M(c) = \sum_{i=1}^{4} \alpha_i S_i \cdot f(L_i, T_i) + \beta S_{\mathrm{r}} \cdot f(L_{\mathrm{r}}, T_{\mathrm{r}}) \quad (1)$$

Here, α weights the four past stages, reflecting their stronger lasting impact on the client's psychological state, while β assigns a comparatively lower weight to recent experiences, reflecting their relatively smaller impact.

SP2: Simulate Client's Cognitive Processes. Mental disorders manifest through distinct patterns of thinking and behavior (Beck et al. 2024; Ellis 1962). Counseling theories provide systematic frameworks to understand these patterns, with Cognitive Behavioral Therapy (CBT) being particularly influential in modeling thought structures (Beck 2020). Formally, let $B_c = b_1, \ldots, b_n$ be core beliefs, $B_i = i_1, \ldots, i_m$ intermediate beliefs, and $A = a_1, \ldots, a_k$ automatic thoughts, jointly shaping thought patterns P. The cognitive process is $C(c) = g(B_c, B_i, A, P)$, where P maps beliefs and experiences to automatic thoughts:

$$(b_n, i_m, l_n) \rightarrow a_k \quad (2)$$

It models how core beliefs, intermediate beliefs and experiences jointly shape automatic thoughts.

SP3: Integrate Detailed Perceptual Memories. Subjective emotions are crucial in psychological counseling. Ianì notes that events consist of perceptual information, reactivating sensorimotor circuits during recall (Ianì 2019). Culbertson highlights 'deep memory' as bodily recollection of trauma, revealing responses beyond verbal expression (Culbertson 1995). We therefore integrate perceptual memories of clients' subjective experiences. This enhanced emotional authenticity will enable more effective and true-to-life counseling simulations. Let $V(c) = h(\xi_e, \xi_b, \xi_p, R)$ denote perceptual memory, where ξ_e, ξ_b, ξ_p are emotional, behavioral, and physiological responses. R is represented as:

$$(\sum_{n=1}^{n} b_n, \sum_{m=1}^{m} i_m, \sum_{k=1}^{k} a_k) \rightarrow \xi_e, \xi_b, \xi_p \quad (3)$$

It models how core beliefs, intermediate beliefs and automatic thoughts jointly shape these responses.

SP4: Model Social Interactions and Relationships. The principle is to simulate the complex web of social interactions and relationships that shape a client's experiences and behaviors. This framework aims to capture the nuances of interpersonal dynamics, including familial bonds, friendships, professional relationships, and broader social connections. By incorporating theories of social psychology and attachment (Bowlby 1980), the goal is to create a realistic representation of how social relationships influence a client's thoughts, emotions, and actions. Let $R(c) = r_1, \ldots, r_n$ denote social relationships across five developmental stages. Each r_i encodes relationship networks and interaction patterns for a developmental period, capturing the evolution of social connections and their influences. $R(c)$ is defined as:

$$R(c) = \sum_{i=1}^{5} w_i \left(Density(r_i) + Familiarity(r_i) \right) \quad (4)$$

Where w_i are coefficients that weigh the importance of relational Density and Familiarity across different stages.

SP5: Maintain Consistency in Data Synthesis. This principle aims to preserve accuracy and coherence in client portraits by aligning simulated social environments, cultural backgrounds, and behavioral traits with real-world client profiles. We seek to ensure temporal consistency in the sequence of simulated events and memories, while maintaining contextual accuracy in generated physical environments and social interactions. A key objective is to address limitations of LLM-based data augmentation in complex and sensitive counseling scenarios, as highlighted by recent research (Fei et al. 2023; Navigli, Conia, and Ross 2023; Cilliers 2020), aiming to mitigate issues of data bias and inconsistency often encountered in sophisticated NLP tasks, particularly in sensitive domains like psychological counseling.

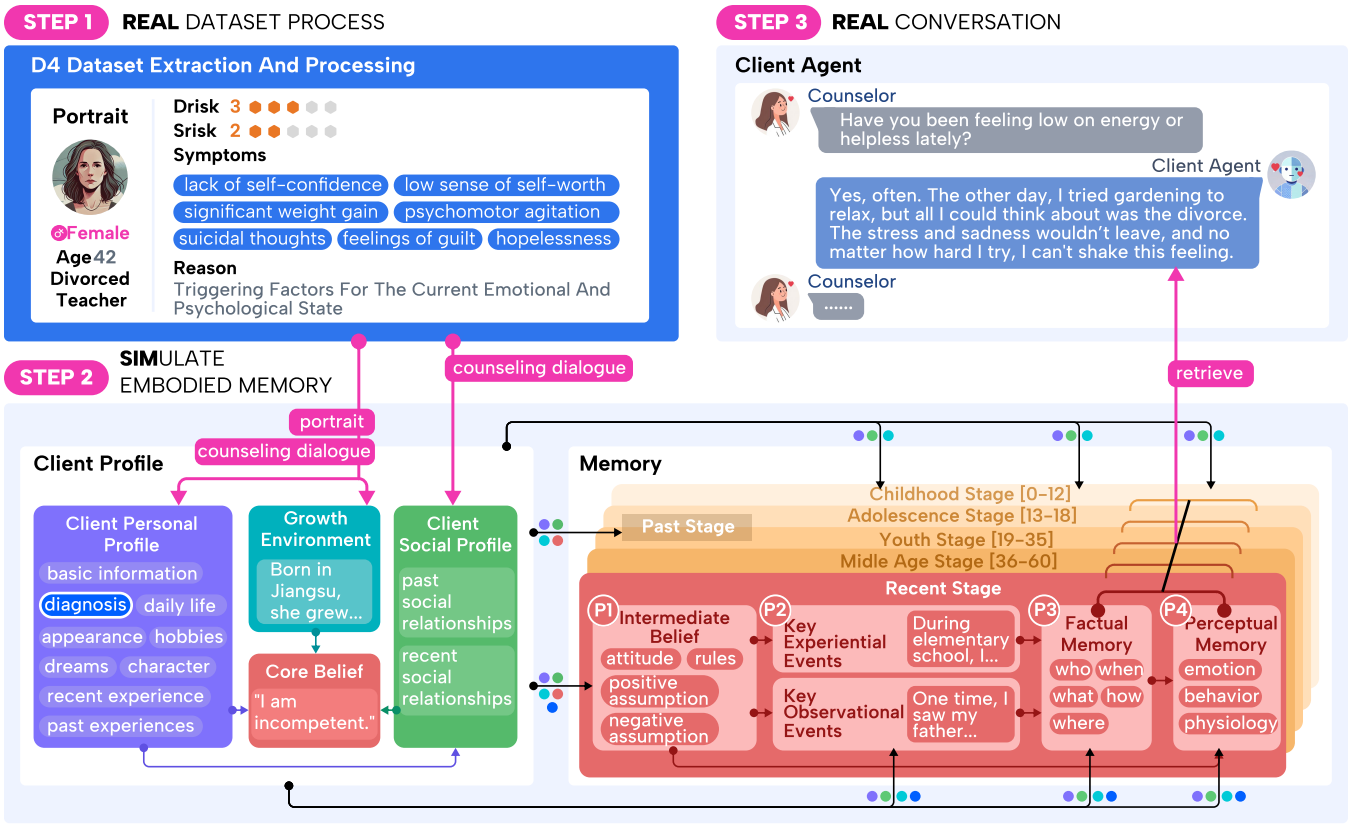

Figure 1: ECAs Framework Overview. The process consists of three steps: Step 1, extracting base information for the *Client Agent* from real datasets; Step 2, expanding the agent's two profiles during memory simulation to form a complete Client Profile, and generating a embodied memory space, including beliefs, cognitive processes, and memories, based on this comprehensive profile; Step 3, dynamically retrieving context-relevant memories during real conversations to ensure realism and consistency.

PS6: Enable Context-Driven Memory Retrieval. AI agents for simulated clients require context-driven memory retrieval to provide realistic, adaptive responses. Systems like Patient-Psi (Wang et al. 2024a) and Roleplay-doh (Louie et al. 2024) demonstrate this need. Effective memory retrieval enables AI agents to provide authentic interactions in healthcare training (Li et al. 2024). Barsalou's work demonstrates that situational triggers activate embodied experiences through pattern completion and inference (Barsalou 2008), enhancing contextual relevance and memory retrieval in simulated counseling sessions. Our ECAs framework employs a context-aware dynamic retrieval mechanism to automatically access related memories, enhancing dialogue depth and authenticity in simulated client interactions. For counseling questions from D^4, at dialogue time step t, the retrieval function $m_t = f(q_t, M(c), H_t)$ selects relevant memories from the client memory space $M(c)$ based on the counselor's question q_t and dialogue history H_t to guide high-quality response generation.

ECAs Framework

This section introduces our ECAs framework for generating psychological counseling interactions with embodied *Client Agents*, covering memory space construction and high-quality dialogues generation. Section extracts base information in real client data from D^4 dataset. Section simulates personal profile, social profile, and embodied memory. Section describes memory retrieval for realistic counseling. Section reports a pilot study and design iteration based on therapist feedback to refine memory depth and relevance.

Real Dataset Process

To ensure that the generated embodied memories are grounded in credible real-life data and are close to the memories of clients with psychological problems [SP5], we selected the D^4 Chinese dialogue dataset, which includes real client data and depression diagnosis information, as the key basis for data generation. The portrait of the client collected during the Portrait Collection phase of D^4 is used as the foundational information for the *Client Agent*. The real counseling dialogue data between the client and counselor is extracted to provide context for profile generation, and the summary from the Professional Diagnosis serves as the description of the *Client Agent*'s current status.

Simulate Embodied Memory

Client Profile Generation Using real data extracted from D^4, the *Client Agent*'s profile is generated in two parts:

Client Personal Profile (see Figure 2) and Client Social Profile (see Figure 3), ensuring a high degree of consistency between persona and social relationships [SP1, SP4, SP5].

Figure 2: Client Personal Profile. The persona describes the *Client Agent*'s basic information such as name, personality, and appearance, along with background and past experiences to form a complete psychological trajectory.

Part 1, ECAs utilize the extracted portrait of the client and counseling dialogue as inputs to generate the Client Personal Profile via LLMs. This profile encompasses the *Client Agent*'s recent stage, personality, appearance, hobbies, dreams, daily habits, and recent experience L_{recent}. Additionally, it backtracks the client's growth environment and summarizes past experiences L_{past}, constructing a coherent psychological trajectory, and aligning personality, behavior, and emotional states with life history for accurate simulations of client's emotions and behaviors. **Part 2**, to align with Part 1, ECAs simulate the *Client Agent*'s social networks $R(c)$ across past and recent stages based on the Client Personal Profile and counseling dialogue. It reflects changes in the number and familiarity of social connections over time, providing external insights into how social environments trigger and sustain depressive symptoms.

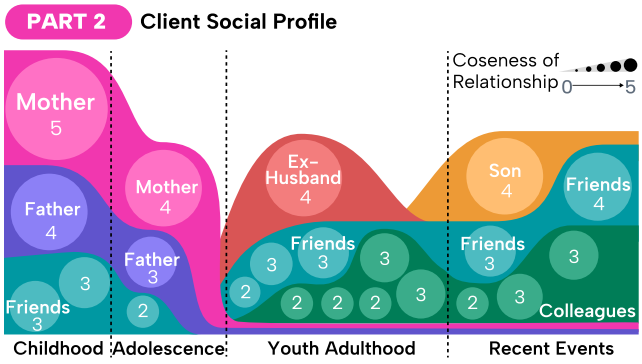

Figure 3: Client Social Profile. The evolution of the *Client Agent*'s social network and relationships over time is reflected, reinforcing the consistency between the social interaction memories and personal profile.

Client Embodied Memory Generation To address the lack of task-specific information in data generated by LLMs, we designed a 4-phase LLM-based generation paradigm (see Figure 1) grounded in CBT theory. This paradigm generates the *Client Agent*'s core beliefs B_c, intermediate beliefs B_i, factual memories from experiences L, automatic thoughts A and perceptual memories[1] $V(c)$ across different life stages based on the two profiles [SP2, SP3, SP5].

Phase 1: Generating beliefs. Using Client Personal Profile and Client Social Profile, we first generate recent-stage core beliefs B_c and intermediate beliefs B_i, with B_c guiding memory simulation. Then, by backtracking through descriptions of past stages, environments, and the Client Social Profile, we derive B_i for each past stage. Each intermediate belief i_m within B_i covers attitudes toward the self, others, and the world, the rules the client follows in specific problem areas, and associated positive and negative assumptions.

Phase 2: Identifying key events. The generation process is consistent across both recent and past stages, as well as in Phase 3 and Phase 4, with all stages using the Client Profile as input. Notably, 'diagnosis' influences only the recent stage, as illustrated in Algorithm 1. The LLMs analyze key experiential and observational events, which form the basis of the *Client Agent*'s experiences L and cause shifts or cognitive changes at each stage with a timeframe of occurrence. Each event is refined into 1-3 descriptions, establishing cause-effect relationships among them.

Phase 3: Forming factual memories. Both recent and past key events from Phase 2 are reviewed and optimized for realism. The focus is on enriching factual details and ensuring emotional authenticity while avoiding abstract descrip-

[1]The concepts of core beliefs, intermediate beliefs, and automatic thoughts are derived CBT theory. Factual memories and perceptual memories are novel concepts introduced in this framework to enhance the simulation of client experiences.

Algorithm 1: Event and Memory Generation Across Stages

Input: P: client personal profile without diagnosis; P^*: client personal profile with diagnosis; G: growth environment; S: client social profile; B: core and intermediate beliefs; List = [1, 2, 3, 4, 5] {1: Childhood, 2: Adolescence, 3: Youth, 4: Middle Age, 5: Recent Stage}

Output: E: Can be key events, factual memories, and perceptual memories

$Num \leftarrow$ calculate_number_of_past_stages($client.age$)
for each $i \in$ List **do**
  if $i \leq Num$ **then**
    $E \leftarrow P[i] + G + S[i] + B[i]$
  else
    $i \leftarrow 5$ {Force i to be 5 for the recent stage}
    $E \leftarrow P^*[i] + G + S[i] + B[i]$
  end if
end for

tions. Each key event follows the 4W1H (Who, What, When, Where, How) format and integrates into a complete event description, which serves as *Client Agent*'s factual memory.

Phase 4: Developing perceptual memories. Automatic thoughts A, being rapid responses to situations, stem directly from B_c and B_i. Each automatic thought a_k links its corresponding core belief b_n and i_m. These thoughts and beliefs are further analyzed to generate likely emotional responses ξ_e, behavioral responses ξ_b, and physiological responses ξ_p, forming the *Client Agent*'s perceptual memory $V(c)$.

Real Conversation

To simulate realistic counseling interactions, the *Client Agent* employs a scenario-driven dynamic retrieval mechanism to extract memories most relevant to the dialogue history H_t and counselor's questions q_t before responding [PS6]. As real clients do not explicitly and actively mention general B_c or B_i during counseling, the retrieved memory types are limited to factual memory from L, $V(c)$, and A.

First, LLMs analyze the H_t between the *Client Agent* and the counselor, determine the required memory type for the response, and memories containing matching keywords kw are retrieved from $M(c)$, and m is the *memory* entry:

$$m_{matched} = \{m \in M(c) \mid kw(m) \cap kw(q_t \cup H_t) \neq \emptyset\} \quad (5)$$

Next, cosine similarity is calculated between the vector representations of the conversation context and each memory, selecting the top-3 most relevant memories:

$$m_t = sort\left(\left\{\frac{v_c \cdot v_m}{\|v_c\|\|v_m\|} \mid v_m \in m_{matched}\right\}\right)[:3] \quad (6)$$

Here, v_c is the vector representation of c, and v_m is the vector representation of each memory m. This process ensures that the most semantically relevant memories are retrieved and ranked for the response.

Pilot Study and Iteration

Refinement Process The pilot study underwent three iterations to refine the *Client Agent*'s embodied memory generation and its use in counseling scenarios. In the first iteration, we consulted a therapist to align the memory construction approach with psychological principles. In the second iteration, two therapists evaluated the consistency of the optimized memory scripts with the client profile. However, feedback revealed that script evaluation alone did not fully capture the memory's function in real conversations and could affect assessments of depression and suicide risk, highlighting the need for a more interactive method. In the third iteration, we generated dialogue data using therapist questions to assess the memory's application in counseling. An LLM extracted highly relevant memory-related questions from the D^4 dataset. While the memory was reasonable and aligned with the client profile, it lacked depth in addressing suicide risk, eating behaviors, and probing questions.

Result Based on the findings, we refined the depth and completeness of the extracted questions. We carefully reselected questions from the D^4 dataset, focusing on five key areas: depression risk, eating behavior, sleep patterns, suicide risk, and social life. These questions were further expanded to explore onset, frequency, intensity, duration, and context, forming a comprehensive set of fourteen questions aligned with real counseling needs.

Evaluation

Dataset Construction

To support expert and automated evaluations, we construct a dataset using the ECAs framework based on D^4 real data.

Dataset	Group	AP	Memory				LM	Avg.MN
			FM	PM	CB	IB		
D^4 (Yao et al. 2022)	Depression clients	8	-	-	-	-	-	-
CharacterDial (Zhou et al. 2024)	Social Characters	6	✔	-	-	-	-	1
PATIENT-Ψ-CM (Wang et al. 2024a)	Mental clients	4	-	✔	✔	✔	-	1
ECAs-dataset (ours)	**Depression clients**	**25**	✔	✔	✔	✔	✔	**134.6**

Table 1: Comparison with related datasets. AP: Attributes of Profiles, FM: Fact Memory, PM: Perceptual Memory, CB: Core Belief, IB: Immediate Belief, LM: Life-stage experience and Memory retrieval, MN: Memory Nodes.

It includes two key components: (1) detailed personal profile and social profile for 451 *Client Agent*s, (2) a comprehensive and substantial embodied memory space for 100 of these *Client Agent*s. Each memory space consists of approximately 400 to 1,500 individual memories, which are grouped into memory nodes. Each memory node encapsulates a complete set of factual memories, perceptual memories, and cognitive processes, capturing diverse stages and experiences across the client's lifetime.

Our dataset was compared with related datasets and demonstrated its distinctive feature of containing a greater quantity and diversity of embodied thinking information (see Table 1). Notably, each *Client Agent* features a richer

profile with more attributes and a more extensive embodied memory space containing a greater number of memory nodes. In addition to supporting explicit dialogue generation, embodied thinking information can be broadly applied to various other forms of explicit language behavior.

Experimental Design and Settings

To assess the quality of the embodied memory generated by ECAs, we conducted a two-pronged evaluation of dialogues, involving Expert Evaluation by human counseling professionals and automated evaluation utilizing LLM. Dialogues were generated through three methods. Control Group 1 and Control Group 2 both used the original, unextended persona data from the D^4 dataset, aligned with the Experimental Group's personas before ECAs-specific memory extension:

(1) Experimental Group (ECAs): Dialogues generated by GPT-4o, based on the personas and memory settings from our ECAs-dataset, using 14 high-frequency counseling questions, with 3 repeated for consistency assessment.

(2) Control Group 1 (GPT-4o): Dialogues generated by GPT-4o with the same questions as the Experimental Group.

(3) Control Group 2 (D^4): Original dialogue data collected from human-to-human interactions in the D^4 dataset.

For expert evaluation, we recruited five qualified counseling professionals (with certifications, 200+ hours of counseling experience, and advanced degrees in Applied Psychology) to analyze the quality and efficacy of the dialogues. Each professional evaluated the same six randomly selected *Client Agents* with embodied memory space from the ECAs-dataset, assessing necessity, sufficiency, fidelity, and consistency of their responses. Expert comments were provided during scoring as justification and categorized as positive (P_) or negative (N_) based on the same four dimensions.

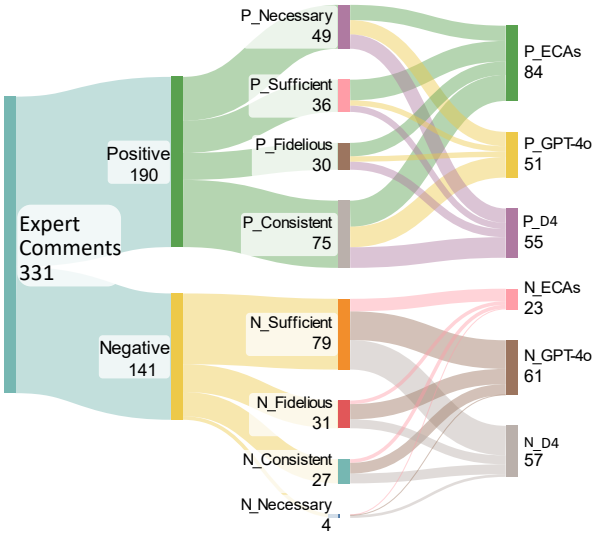

Figure 4: Classification of Positive and Negative Expert Comments. Expert comments are categorized as positive (P_) or negative (N_) based on four dimensions, and grouped into ECAs (Ours), GPT-4o, and D^4.

Concurrently, we used 100 *Client Agents* with embodied memory space from the ECAs-dataset and implemented two automated evaluation methods using GPT-4o as the evaluator. The first method assesses dialogue utility in supporting diagnostic decision-making by classifying depression risk and suicide risk into four severity levels (no risk, mild, moderate, and severe). The second method directly assesses dialogue quality based on fidelity, comprehensiveness, consistency, plausibility, and specificity. This dual approach aimed to offer both nuanced human insights and consistent, large-scale automated analysis, providing a comprehensive evaluation of the embodied memory quality generated by our ECAs framework compared to the baseline approaches.

Human Expert Evaluation

Specifically, experts assessed the necessity and sufficiency of facts and emotional details provided by the client for counseling evaluation, the authenticity of the client's reported experiences and feelings in relation to real clients, and the consistency of the client's responses with their character profile and across similar events.

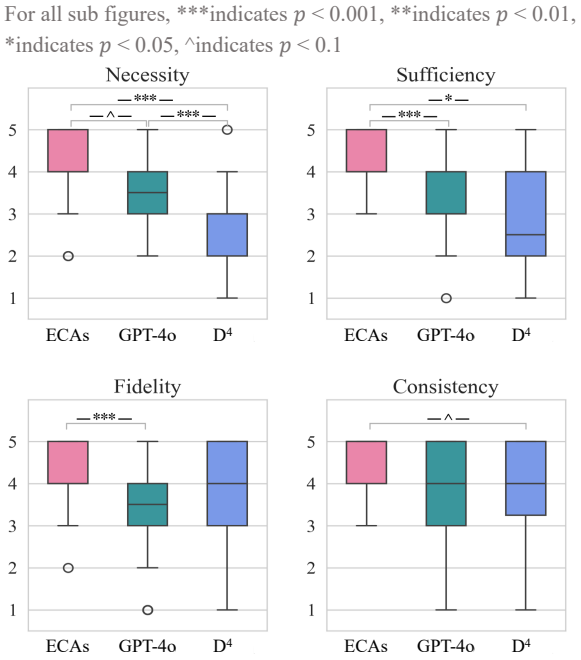

Figure 5: Comparison of ECAs (Ours), GPT-4o, and D^4 Performance Across Four Dimensions. Box plots show the distribution of scores for Necessity, Sufficiency, Fidelity, and Consistency.

Results As shown in Figure 4, dialogues generated by ECAs achieved the highest positive and lowest negative expert evaluations, outperforming GPT-4o and D^4. Experts regarded the dialogues generated by our method as more sufficient, necessary, fidelious, and consistent. Notably, ECAs received three times more positive and one-third fewer negative evaluations for sufficiency. ECAs dialogues demonstrated stronger relevance to depression diagnosis and significantly higher quality and reliability than the benchmarks.

Figure 5 further highlights the significant improvements demonstrated by our framework across multiple dimensions compared to both GPT-4o and human-generated responses.

In terms of **Necessity**, a repeated measures ANOVA revealed significant differences among the three conditions, $F(2, 87) = 16.173$, $p < 0.001$, $\eta^2 = 0.271$. Post-hoc tests using Bonferroni correction showed that ECAs marginally outperformed GPT-4o (mean difference $= 0.56667$, $p = 0.074 < 0.1$). ECAs significantly outperformed human responses (mean difference $= 1.40000$, $p < 0.001$), and GPT-4o also significantly outperformed human responses (mean difference $= 0.83333$, $p = 0.003 < 0.01$). Expert E4 praised the ECAs' performance, noting that *'the additional emotional details provided by ECAs enriched the dialogue, making it more comprehensive'*. Expert E5 emphasized that the ECAs' responses *'contained numerous specific details that* ***significantly aid in assessing the client's precise situation and emotional state****'*.

For **Sufficiency**, the ANOVA again showed significant differences, $F(2, 87) = 15.796$, $p < 0.001$, $\eta^2 = 0.266$. Post-hoc comparisons indicated that ECAs significantly outperformed both GPT-4o (mean difference $= 1.06667$, $p = 0.001$) and human responses (mean difference $= 1.50000$, $p < 0.001$). Expert E3 noted ECAs' output as *"very much aligns with a client immersed in grief"*, while Expert E5 noted the *'reasonable grief reactions to the loss of a loved one'*. Expert E4 described the responses as ***'delicate and comprehensive'***, with Expert E3 further praising the *'vivid examples and nuanced thoughts and emotions'* presented.

Regarding **Fidelity**, the ANOVA revealed significant differences, $F(2, 87) = 5.188$, $p = 0.007 < 0.01$, $\eta^2 = 0.107$. Post-hoc tests showed that ECAs significantly outperformed GPT-4o (mean difference $= 1.00000$, $p = 0.005 < 0.01$) Expert E3 highlighted that the ECAs' responses *'consistently addressed the counselor's questions accurately, without deviating from the topic'* and *"****effectively matched and addressed the questions*** *at hand"*.

For **Consistency**, the ANOVA showed significant differences, $F(2, 87) = 3.217$, $p = 0.045 < 0.05$, $\eta^2 = 0.069$. There was a marginally significant difference between ECAs and human responses (mean difference $= 0.56667$, $p = 0.079 < 0.1$). Experts noted that the responses were *'consistent with the client's identity and experiences'* (E5), with E4 adding that *"the client's efforts, lack of self-confidence, concern for parental opinions, and the series of depressive symptoms stemming from graduate school pressure were all highly congruent"* with the expected profile.

These results indicate that the ECAs framework consistently produced higher quality responses across all measured dimensions, with particularly strong improvements in necessity and sufficiency compared to both GPT-4o and human-generated responses.

Automated Evaluation

For the automated evaluation, the first automated evaluation focused on classifying depression risk (drisk) and suicide risk (srisk) based on dialogue data. To ensure fairness and mitigate the potential impact of uneven sample distribution on the fairness of results, the four-class classification is evaluated by macro-averaged Precision, Recall, and F1 by sklearn[2]. As shown in Table 2, dialogues generated by our framework achieve the highest performance across both tasks, demonstrating that embodied memory component provides more vital information and significantly contributes to the diagnostic process.

Task	Method	Precision	Recall	F1
drisk	GPT-4o	0.44±0.02	0.41±0.05	0.35±0.04
	D^4	0.49±0.03	0.44±0.04	0.40±0.05
	Ours	**0.51**±0.04	**0.50**±0.08	**0.42**±0.04
srisk	GPT-4o	0.64±0.05	0.59±0.03	0.59±0.04
	D^4	0.67±0.04	0.55±0.05	0.59±0.05
	Ours	**0.71**±0.08	**0.66**±0.05	**0.67**±0.06

Table 2: Depression and Suicide Severity Classification

Method	F	C1	C2	P	S	Total
D^4	2.40/5	2.91/7	1.03/2	1.98/3	0.96/2	9.28/20
Ours	**4.66**/5	**6.26**/7	**1.99**/2	**2.98**/3	**1.97**/2	**17.90**/20

Table 3: Auto Qulity Evaluation Across Five Dimensions

For the second automated evaluation, we compared the dialogue quality generated by our method with that from human role-playing in Control Group 2. Evaluation covered five key metrics: fidelity (F), comprehensiveness (C1), consistency (C2), plausibility (P), and specificity (S), with weighted scores reflecting each metric's importance. Results in Tabl 3 show that ECAs-generated dialogues performed strongly across all metrics, especially in fidelity and comprehensiveness, significantly outperforming human-generated dialogues in D^4. From the LLM perspective, this indicates that ECAs-generated dialogues provide deep, authentic emotional responses across diverse life experiences while maintaining high consistency and plausibility.

Conclusion

This paper presents ECAs, a novel framework for simulating embodied client agents in psychological counseling that generates high-fidelity counseling conversational data. Integrating counseling theories By integrating counseling theories with LLMs, we expand real counseling data into a nuanced cognitive memory space to generate realistic dialogues. Guided by six principles for simulation derived from counseling theories, our framework was validated using the D^4 dataset and evaluated by licensed counselors, demonstrating significant improvements in simulation authenticity and necessity. Additionally, automated evaluation methods validated the higher-quality dialogues generated by our framework. Looking ahead, the ECAs holds immense potential for counselors to generate custom scenarios on-demand, enhancing both training and research capabilities.

[2]https://scikit-learn.org

Acknowledgments

This work was supported by the Beijing Municipal Science and Technology Project (Nos. Z231100010323005) and China Natural Science Foundation Youth Fund 62202267.

References

Barsalou, L. W. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1): 617–645.

Beck, A. T.; Rush, A. J.; Shaw, B. F.; Emery, G.; DeRubeis, R. J.; and Hollon, S. D. 2024. *Cognitive therapy of depression*. Guilford Publications.

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

Beck, J. S. 2021. *Cognitive behavior therapy: Basics and beyond, 3rd ed.* The Guilford Press.

Bowlby, J. 1980. Attachmentandloss: Vol. 3. *Loss. New*.

Cilliers, L. 2020. Wearable devices in healthcare: Privacy and information security issues. *Health information management journal*, 49(2-3): 150–156.

Culbertson, R. 1995. Embodied memory, transcendence, and telling: Recounting trauma, re-establishing the self. *New Literary History*, 26(1): 169–195.

Ellis, A. 1962. *Reason and emotion in psychotherapy.* Lyle Stuart.

Fei, Y.; Hou, Y.; Chen, Z.; and Bosselut, A. 2023. Mitigating Label Biases for In-context Learning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14014–14031. Toronto, Canada: Association for Computational Linguistics.

Furlan, R.; Gatti, M.; Menè, R.; Shiffer, D.; Marchiori, C.; Giaj Levra, A.; Saturnino, V.; Brunetta, E.; and Dipaola, F. 2021. A natural language processing–based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study. *JMIR medical informatics*, 9(4): e24073.

Ianì, F. 2019. Embodied memories: Reviewing the role of the body in memory processes. *Psychonomic bulletin & review*, 26(6): 1747–1766.

Kjell, O. N.; Kjell, K.; and Schwartz, H. A. 2024. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333: 115667.

Kuehne, F.; Ay, D. S.; Otterbeck, M. J.; and Weck, F. 2018. Standardized patients in clinical psychology and psychotherapy: A scoping review of barriers and facilitators for implementation. *Academic Psychiatry*, 42: 773–781.

Li, Y.; Zeng, C.; Zhong, J.; Zhang, R.; Zhang, M.; and Zou, L. 2024. Leveraging Large Language Model as Simulated Patients for Clinical Education. arXiv:2404.13066.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904. Miami, Florida, USA: Association for Computational Linguistics.

Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Yang, D.; and Vosoughi, S. 2024. Training Socially Aligned Language Models on Simulated Social Interactions. In *The Twelfth International Conference on Learning Representations*.

Louie, R.; Nandi, A.; Fang, W.; Chang, C.; Brunskill, E.; and Yang, D. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10570–10603. Miami, Florida, USA: Association for Computational Linguistics.

Miner, A. S.; Shah, N.; Bullock, K. D.; Arnow, B. A.; Bailenson, J.; and Hancock, J. 2019. Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in psychiatry*, 10: 746.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality*, 15(2).

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Ping, Y. 2024. Experience in psychological counseling supported by artificial intelligence technology. *Technology and Health Care*, 1–18.

Schatzmann, J.; Georgila, K.; and Young, S. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 45–54.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759.

Sircova, A.; Karimi, F.; Osin, E. N.; Lee, S.; Holme, P.; and Strömbom, D. 2015. Simulating irrational human behavior to prevent resource depletion. *PloS one*, 10(3): e0117612.

Tang, Y.; Kang, Y.; Wang, Y.; Wang, T.; Zhong, C.; and Gong, J. 2025. CA+: Cognition Augmented Counselor Agent Framework for Long-term Dynamic Client Engagement. *arXiv preprint arXiv:2503.21365*.

Wang, R.; Milani, S.; Chiu, J. C.; Zhi, J.; Eack, S. M.; Labrum, T.; Murphy, S. M.; Jones, N.; Hardy, K. V.; Shen, H.; Fang, F.; and Chen, Z. 2024a. PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12772–12797. Miami, Florida, USA: Association for Computational Linguistics.

Wang, Z.; Chiu, Y. Y.; and Chiu, Y. C. 2023. Humanoid Agents: Platform for Simulating Human-like Generative

Agents. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 167–176. Singapore: Association for Computational Linguistics.

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2024b. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 257–279. Mexico City, Mexico: Association for Computational Linguistics.

Yao, B.; Shi, C.; Zou, L.; Dai, L.; Wu, M.; Chen, L.; Wang, Z.; and Yu, K. 2022. D4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2438–2459. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yuan, D.; Rastogi, E.; Naik, G.; Rajagopal, S. P.; Goyal, S.; Zhao, F.; Chintagunta, B.; and Ward, J. 2024. A Continued Pretrained LLM Approach for Automatic Medical Note Generation. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 565–571. Mexico City, Mexico: Association for Computational Linguistics.

Zheng, X.; Wu, L.; Yan, Z.; Tang, Y.; Zhao, H.; Zhong, C.; Chen, B.; and Gong, J. 2024. Large language models powered context-aware motion prediction in autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 980–985. IEEE.

Zhou, J.; Chen, Z.; Wan, D.; Wen, B.; Song, Y.; Yu, J.; Huang, Y.; Ke, P.; Bi, G.; Peng, L.; Yang, J.; Xiao, X.; Sabour, S.; Zhang, X.; Hou, W.; Zhang, Y.; Dong, Y.; Wang, H.; Tang, J.; and Huang, M. 2024. CharacterGLM: Customizing Social Characters with Large Language Models. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1457–1476. Miami, Florida, US: Association for Computational Linguistics.