PSYMEM: ROLE-ASYMMETRIC INFORMATION FLOW FOR LONGITUDINAL COUNSELING DIALOGUE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal counseling-style dialogue research requires data that capture substantive interaction within sessions and continuity across sessions, yet authentic linked records are difficult to obtain and share. We introduce PsyMem, a synthesis framework combining persona-grounded client simulation, reflection-guided utterance generation, and role-asymmetric cross-session state updates. The client retains a fixed private persona and an evolving concern state, while the counselor carries forward memory derived from preceding dialogues. PsyMem produces 150 linked four-session sequences comprising 600 sessions, with an average of 61 turns per session. We assess within-session quality using a 15-item WAI-derived observer rubric and introduce the Memory Continuity Analysis Framework (MCAF) to evaluate cross-session continuity in stored role summaries. Under GPT-4o evaluation, PsyMem scores 6.04 on WAIOS, versus 5.71 for the best evaluated synthetic baseline on this metric, and 4.96/5 on MCAF. Removing cross-session state lowers MCAF by about 3.9 points; among variants retaining such state (MCAF: 4.94–4.95), the role-asymmetric design attains the highest observed WAIOS (6.04 vs. 5.97–6.02). Across six open-source base models, models fine-tuned on PsyMem-generated dialogues perform best in most MentalChat16K comparisons, while WAIOS results vary by model and evaluator. Code and data are available at `https://anonymous.4open.science/r/PsyMem-B34E37666`.

1 INTRODUCTION

Longitudinal counseling-style dialogue research requires data that capture both substantive exchanges within a session and information carried into later sessions. Authentic linked records are difficult to obtain and share because counseling conversations are sensitive (Li et al., 2023), while report-based reconstructions do not preserve the original counselor–client interaction (Zhang et al., 2024). Synthetic dialogue offers a scalable setting for studying these interaction patterns and training dialogue models.

Existing work addresses this problem at different temporal scales. Response-pair and single-session resources support local counseling interaction (Sun et al., 2021; Xu et al., 2025; Qiu & Lan, 2024; Suresh et al., 2025), while recent systems model longitudinal interaction: AnnaAgent tracks evolving help-seeker state with cross-session memory (Wang et al., 2025), MusPsy generates linked sessions using static and dynamic client states (Wang et al., 2026), and PsychEval stages counselor access to client information (Pan et al., 2026). Together, these lines of work motivate a synthesis workflow that combines sustained within-session interaction with cross-session states whose information sources differ by role.

To address these coupled requirements, we introduce PsyMem, which separates private client state from dialogue-derived counselor memory (Figure 1(a)). PsyMem expands an initial client description into a fixed private persona and maintains a separate evolving concern state. We formulate this separation as role-asymmetric information flow: the two roles receive cross-session context with different provenance and visibility. During each session, the client generates turns from those states and the visible dialogue prefix, while the counselor uses dialogue-derived memory and the same

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prefix. Client utterances undergo reflection-based checks for completeness, concrete detail, and persona consistency. After the session, the client concern state is updated using the persona and prior counselor guidance, and counselor memory is updated from the observed dialogue. The resulting states condition the next session.

We assess within-session quality with a 15-item WAI-derived observer rubric (WAIOS) and the MentalChat16K criteria (Horvath & Greenberg, 1989; Bordin, 1979; Xu et al., 2025), and introduce MCAF to evaluate continuity in stored role summaries. Under GPT-4o evaluation, PsyMem attains a WAIOS score of 6.04, compared with 5.71 for the best evaluated synthetic comparison on this metric, and an MCAF score of 4.96/5. Removing cross-session state reduces MCAF by about 3.9 points; among variants retaining such state, the role-asymmetric design attains the highest observed WAIOS (6.04 vs. 5.97–6.02), with MCAF scores ranging from 4.94 to 4.95. Across six base models, fine-tuning on PsyMem-generated dialogues yields the best MentalChat16K results in most model–judge settings, whereas WAIOS results vary by model and evaluator.

Our contributions are threefold:

1. We formulate longitudinal counseling-dialogue synthesis as a role-asymmetric information-flow problem and instantiate it with an evolving client concern state and dialogue-derived counselor memory updated from role-dependent input sources (Section 3.4).
2. We release the PsyMem dataset and evaluation resources, comprising 150 linked, four-session counseling-style dialogue sequences (600 sessions) generated with persona grounding, reflection-guided generation, and cross-session state updates (Appendix A.2).
3. We introduce MCAF as an automatic diagnostic of stored-summary continuity and assess within-session quality and downstream fine-tuning utility (Sections 3.5 and 4).

2 Related Work

Counseling Dialogue Data and Synthesis. Existing resources vary in provenance, domain, and temporal granularity. RealPsyDial provides authentic counseling sessions (Li et al., 2023), whereas CPsyCounD reconstructs multi-turn consultations from counseling reports (Zhang et al., 2024). PsyQA (Sun et al., 2021), MentalChat16K (Xu et al., 2025), and annotated online-support data (Sharma et al., 2020) are organized primarily as question–answer or post–response pairs, while EmpatheticDialogues provides multi-turn but open-domain conversations (Rashkin et al., 2019). Interactive Agents and DiaSynth instead synthesize standalone dialogues (Qiu & Lan, 2024; Suresh et al., 2025). Among these selected resources, the published data unit is therefore a response pair or an individual dialogue or session rather than a client-linked session sequence (Appendix A.1).

Longitudinal Counseling Systems. Recent longitudinal counseling systems model different aspects of state evolution and information access across sessions. AnnaAgent emphasizes dynamic help-seeker simulation supported by cross-session memory (Wang et al., 2025); MusPsy combines static and dynamic client states with structured memory to generate linked sessions (Wang et al., 2026); and PsychEval regulates counselor access to client information through a disclosure timeline (Pan et al., 2026). PsyMem addresses a complementary design question: how role-specific context should be updated from different information sources across sessions. It maintains a fixed private persona and an evolving concern state for the client, derives counselor memory from observed dialogue, and applies turn-level grounding checks to client utterances.

Long-Term Memory and Reflection for Dialogue Agents. Beyond counseling-specific systems, general-purpose long-term memory systems summarize, organize, or retrieve prior interactions to maintain context over extended dialogues (Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; Jiang et al., 2025). LoCoMo provides a persona-grounded, multi-session open-domain benchmark with temporal structure for evaluating long-term conversational memory (Maharana et al., 2024). A complementary line of work uses reflection to revise generated outputs in response to feedback (Shinn et al., 2023; Madaan et al., 2023). PsyMem targets a different level of abstraction: role-asymmetric information flow in longitudinal counseling-dialogue synthesis. It updates the client concern state and counselor memory from different post-session information sources, while using reflection locally to check individual client utterances.

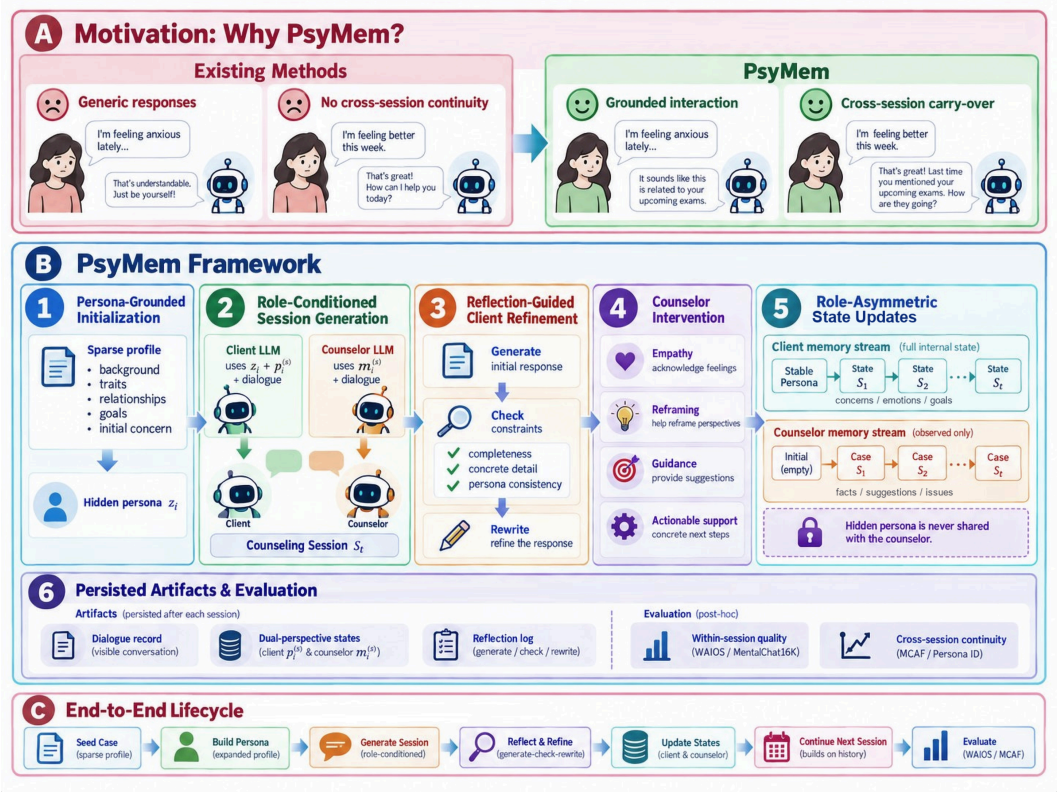

Figure 1: **PsyMem framework.** (A) Schematic motivation contrasting disconnected dialogue with grounded, linked interaction. (B) Persona-grounded initialization, role-conditioned session generation, reflection-guided client-utterance refinement, within-session counselor intervention quality, role-asymmetric information flow through longitudinal state updates. (C) The end-to-end synthesis and post-hoc evaluation lifecycle.

Evaluation of Counseling-Style and Longitudinal Dialogues. WAI-derived observer ratings characterize working alliance through Goal, Task, and Bond, whereas the MentalChat16K criteria capture broader aspects of counseling-dialogue quality (Horvath & Greenberg, 1989; Bordin, 1979; Xu et al., 2025). More generally, automatic dialogue scores are sensitive to the choice of evaluator and prompt design (Liu et al., 2016; Mehri & Eskenazi, 2020; Zheng et al., 2023; Liu et al., 2023). However, these within-session measures do not directly assess whether role-specific records remain coherent across sessions. For this longitudinal setting, we introduce MCAF as a complementary diagnostic of continuity in stored client concern states and counselor memories.

3 Method

3.1 Overview and Task Formulation

Given an initial client description x_i for client i, PsyMem synthesizes an ordered sequence $D_i = (D_i^{(1)}, \ldots, D_i^{(S)})$, where $D_i^{(s)}$ is the client–counselor dialogue in session s. The design aims to sustain informative interaction within a session and carry relevant state across sessions while conditioning each role on its own available information.

Let $H_i^{(s,<t)}$ denote the visible dialogue prefix before turn t of session s. The counselor-turn context consists of prior dialogue-derived memory $m_i^{(s)}$ and $H_i^{(s,<t)}$; the client-turn context additionally contains the private persona z_i and evolving concern $p_i^{(s)}$. The two roles therefore receive different sources of cross-session information during dialogue generation.

Multi-session synthesis loop. PsyMem begins with a fixed persona z_i expanded from x_i and an initial concern state $p_i^{(1)}$; the counselor memory is empty in the first session, $m_i^{(1)} = \emptyset$. Within each session, the client and counselor generate turns using their respective contexts and the visible dialogue prefix. Client utterances pass through a generate–check–rework step before entering the dialogue record. After session s, PsyMem separately updates the client concern state and counselor memory to produce the contexts for session $s + 1$; the process repeats across the linked sequence.

Persona grounding and utterance checking operate within sessions, whereas the two state transitions carry role-specific context between sessions.

3.2 Client Persona and Initial-State Construction

PsyMem expands x_i into a structured persona z_i containing background, interaction style, response tendencies, and initial concern context. Starting from the presenting concern and other context available in x_i, GPT-4o constructs persona fields for demographic and life context and personality-related descriptors, including Big Five traits and attachment-informed interpersonal patterns (Costa & McCrae, 1992; Bowlby, 1969; OpenAI, 2024). Persona fields and generation prompts are documented in Appendix A.3 and Appendix F.1.

The persona z_i is fixed at initialization and does not change across sessions. We denote the initial concern state by $p_i^{(1)}$ and subsequent concern states by $p_i^{(s)}$ for $s > 1$. The counselor's cross-session context is drawn from prior dialogue.

3.3 Role-Conditioned Session Generation and Reflection

Within session s, client turns are generated from z_i, $p_i^{(s)}$, and the visible prefix $H_i^{(s,<t)}$; counselor turns use $m_i^{(s)}$ and the same prefix.

Counselor turns are guided by four prompted response strategies: *empathy* acknowledges and validates disclosed experiences, *reframing* helps examine unhealthy thinking, *guidance* offers at most one personalized suggestion tied to the current scenario, and *actionable support* translates insight into concrete next steps. The generation prompt selects among these strategies according to session phase and modeled client resistance; Figure 1 shows them in the counselor-side generation block.

Both role generators use GPT-4o in the reported synthesis. Extended role-playing can produce vague or incomplete client responses and repetitive exchanges. PsyMem therefore applies a generate–check–rework loop to each client utterance. This use of feedback follows self-refinement methods (Madaan et al., 2023; Shinn et al., 2023), with checks focused on utterance grounding.

For each generated client utterance, a lightweight verifier checks three constraints:

- **Completeness**: the response is semantically complete.
- **Concrete detail**: the response contains specific information, such as an event, time, situation, or interpersonal context.
- **Persona consistency**: the response is compatible with the fixed persona z_i and current concern state $p_i^{(s)}$.

If a check fails, the generator receives targeted feedback and regenerates the client response. The final response and verification metadata, including rework flags and failure reasons, are logged. Sessions end at a natural closing point or when the dialogue becomes excessively long. Appendix F.1 provides the client and counselor generation templates.

3.4 Post-Session Role-Asymmetric State Updates

Role asymmetry arises from differences in information access: the client update uses private client state and accumulated counselor guidance, whereas the counselor update is limited to prior counselor memory and observed dialogue. The resulting states condition the next session.

Client-side state update. After session s, PsyMem updates the client concern state using the previous concern state, accumulated counselor guidance, and the stable persona:

$$p_i^{(s+1)} = F_{\text{client}}\left(p_i^{(s)},\ g_i^{(1:s)},\ z_i\right), \tag{1}$$

where $g_i^{(1:s)}$ denotes counselor suggestions, interaction outcomes, and actionable feedback extracted from previous sessions. The persona state z_i is not modified; only the current concern state is updated.

Counselor-side memory update. The counselor-memory transition is specified in terms of prior counselor memory and observed dialogue:

$$m_i^{(s+1)} = F_{\text{counselor}}\left(m_i^{(s)},\ D_i^{(s)}\right). \tag{2}$$

The counselor memory organizes dialogue-derived facts, prior suggestions, the client's personality characteristics, and the evolving state of the client's concerns:

$$m_i^{(s+1)} = \{\text{facts}_{1:s},\ \text{personality}_{1:s},\ \text{suggestions}_{1:s},\ \text{issues}_{1:s}\}. \tag{3}$$

At the beginning of session $s+1$, the client is conditioned on its hidden persona and updated concern state, while the specified counselor context consists of dialogue-derived memory:

$$\text{Context}_{s+1}^{\text{client}} = [z_i;\ p_i^{(s+1)}], \qquad \text{Context}_{s+1}^{\text{counselor}} = [m_i^{(s+1)}]. \tag{4}$$

Persisted artifacts. After each session, PsyMem stores three artifacts indexed by (client_id, session_id):

- **Dialogue record**: the complete utterance sequence $D_i^{(s)}$.
- **Role-specific states**: the updated client concern state $p_i^{(s+1)}$ and counselor memory $m_i^{(s+1)}$.
- **Reflection log**: per-turn verification metadata, including rework flags and failure reasons.

The dialogue record supplies the observed content for post-session state updates and remains available for analysis; the updated role-specific states provide the role contexts specified in Equation 4 for the next session. Reflection logs are retained for analysis rather than supplied as next-session context. Within the counselor memory, the suggestions component is not injected into the counselor's dialogue context; it serves as reference memory for subsequent memory updates and is used for post-hoc continuity diagnostics.

3.5 Post-hoc Memory Continuity Diagnostic

Within-session dialogue-quality scores do not reveal whether the role-specific information carried across sessions remains coherent as it is updated over time. We therefore introduce the Memory Continuity Analysis Framework (MCAF), a post-hoc diagnostic of the client concern states and counselor memories stored after each session. MCAF operationalizes longitudinal continuity along four complementary dimensions: plausible evolution of the client concern state, retention of relevant counselor-side information, alignment between the two stored states within each session, and consistency of the complete sequence across sessions.

For each four-session sequence, GPT-4o receives the client concern state and counselor memory saved after each session and uses the fixed prompts in Appendix F.3 to perform two complementary diagnostics. For MCAF, the judge assigns 1–5 ratings for client-state continuity, counselor-memory continuity, within-session alignment, and multi-session consistency, whose arithmetic mean is reported as the MCAF average (Appendix C.1; Table 9). The separate longitudinal persona-consistency check determines whether the four sessions plausibly describe the same client based on background details, concerns, interaction patterns, and factual contradictions, returning a Boolean judgment with an associated 0–100 confidence value (Appendix C.2). The identity check is a same-person discrimination task over the four stored records of one sequence, not a multi-candidate identification test; confidence values at or above 80 are read as the same client (Table 10; Appendix C.2).

(a) Average counseling-quality scores across five datasets

	WAIOS GPT-4o	WAIOS Human	MC16K GPT-4o	MC16K Human
PsyQA	4.18	4.66	4.83	6.58
SimPsyDial	5.71	5.05	7.99	7.22
MentalChat16K	5.53	5.13	6.84	7.14
DiaSynth	4.99	5.30	6.14	7.32
PsyMem	**6.04**	**6.49**	**8.76**	**9.38**

(b) Cross-session continuity under MCAF

	Client state	Counselor memory	Within-sess. align.	Multi-sess. consist.	MCAF avg.	Persona ID
PsyQA	1.35	1.40	1.45	1.45	1.41	64
SimPsyDial	1.00	1.00	1.00	1.00	1.00	55
MentalChat16K	1.48	1.55	1.50	1.50	1.51	62
DiaSynth	1.75	1.80	1.70	1.75	1.75	67
PsyMem	**4.96**	**4.96**	**4.95**	**4.95**	**4.96**	95

Figure 2: Evaluation overview. **(a)** Dataset-level WAIOS and MentalChat16K averages under GPT-4o and human evaluation. **(b)** Cross-session MCAF scores and persona identity (MCAF uses a 1–5 scale; the separate identity indicator uses 0–100). Colors are normalized within metric scales; full WAIOS, human-evaluator, and MCAF sub-dimensions are reported in Appendix D.

4 Experiments

4.1 Experimental Setup

Experiment design. We evaluate within-session quality (RQ1), cross-session continuity (RQ2), downstream fine-tuning utility (RQ3), and the contribution of each synthesis component (RQ4).

Datasets. PsyQA provides a real-data reference; SimPsyDial and DiaSynth are synthetic-dialogue comparisons; MentalChat16K provides an established English benchmark. PsyMem is built from 150 help-seeking seed posts sampled from the PsyQA subset used by SimPsyDial (Appendix A.2). Each generated client has four linked sessions, yielding 600 sessions with an average of 61 turns per session. Only PsyMem has native linked sessions among the evaluated corpora.

Evaluation protocol. We assess within-session quality using WAIOS, a 15-item WAI-derived observer rubric covering Goal, Task, and Bond (1–7 per item; prompt in Appendix F.3), and seven MentalChat16K criteria (1–10). Dataset-level results include GPT-4o and separately reported human-evaluator ratings, collected from 11 evaluators without clinical training on the questionnaire in Appendix G (1-10 for the seven MentalChat16K criteria, 1-7 for the three WAI-O dimensions); downstream fine-tuning results use automatic GPT-4o and DeepSeek-V4-Flash judges. Three negatively worded items (WAI-O items 1, 10, and 27) are reverse scored as 8 - x before averaging, and the overall score is the mean of the Goal, Task, and Bond dimension scores.

For cross-session evaluation, we score 50 sampled four-session PsyMem sequences and construct 50 disconnected four-item control sequences for each comparison corpus by grouping independently sampled examples. GPT-4o automatically scores MCAF and the separate identity indicator from stored session summaries; human-evaluator ratings apply only to the within-session WAIOS and MentalChat16K rubrics.

Fine-tuning setup. We fine-tune six baseline models (five general LLMs and one counseling-specific model) on each of the four synthetic datasets and evaluate their generated dialogues with MentalChat16K and WAIOS downstream metrics. We evaluate the generated dialogues with GPT-4o and DeepSeek-V4-Flash as automatic judges. Hyperparameters and generation settings are in Appendix B.

4.2 Dialogue Synthesis Quality (RQ1)

PsyMem receives the highest reported average score in all four within-session evaluation settings. Under GPT-4o WAIOS evaluation, it scores 6.04, compared with 5.71 for the strongest evaluated synthetic comparison, SimPsyDial. The margin spans Goal (5.95 vs. 5.65), Task (5.92 vs. 5.45), and Bond (6.24 vs. 6.03), rather than being confined to one WAIOS dimension. PsyMem also obtains

the highest GPT-4o MentalChat16K average (8.76). The reported human-evaluator ratings show the same ranking, with WAIOS and MentalChat16K averages of 6.49 and 9.38, respectively. Figure 2(a) summarizes these comparisons, and Appendix D reports the complete dimension-level results.

We additionally report judge agreement. Across the 36 model-dataset settings evaluated by each pair of judges, WAIOS overall scores correlate at r = 0.79 (Spearman rho = 0.80) between GPT-4o and DeepSeek-V4-Flash and at r = 0.74 (Spearman rho = 0.88) between GPT-4o and Qwen3.6, and MentalChat16K averages correlate at r = 0.92 (Spearman rho = 0.92) between GPT-4o and DeepSeek-V4-Flash. Automatic and human-evaluator WAIOS scores show moderate agreement (r = 0.58, n = 15 dataset-subscale pairs, p = 0.025). The same evaluators also rated the two lay-perspective comfort items for each dialogue they assessed (Appendix G); averaged over dialogues, PsyMem obtains the highest mean on both (9.0 as client, 9.8 as counselor, on a 1-10 scale), compared with 7.0-7.2 and 7.0-7.4 for the four baseline datasets.

4.3 Cross-Session Continuity (RQ2)

Under the automatic summary-based protocol, PsyMem scores 4.95–4.96 across the four MCAF dimensions, with an average of 4.96/5 and a reported identity indicator of 95/100 (Figure 2(b); Table 14). The constructed disconnected controls, formed by grouping independent items from the four non-longitudinal comparison corpora, obtain MCAF averages of 1.00–1.75 and identity indicators of 55–67.

4.4 Downstream Fine-tuning Utility (RQ3)

To evaluate whether PsyMem-generated dialogues transfer across model families, we fine-tune six models: five general LLMs (LLaMA2-7B (Touvron et al., 2023), DeepSeek-LLM-7B-Chat (DeepSeek-AI et al., 2024), Vicuna-7B-v1.5 (Chiang et al., 2023), Qwen2-7B-Instruct (Yang et al., 2024), and Qwen3.5-9B-Thinking (Qwen Team, 2025)) and one counseling-specific LLM (MeChat (Qiu et al., 2024)). Qwen3.5-9B-Thinking denotes a fine-tuned base model, while the Qwen3.6 evaluator in Appendix E.1 is used only as an additional judge. We fine-tune on four synthetic corpora: PsyMem, SimPsyDial, DiaSynth, and the synthetic subset of MentalChat16K; PsyQA is used only as a static real-world quality reference. Dialogues are evaluated under single-agent autonomous generation and dual-agent role-playing, with model and fine-tuning details in Appendix B.

We evaluate outputs with the MentalChat16K average score, which measures broad counseling quality on a 1–10 scale, and the WAIOS overall score, which measures working alliance on a 1–7 scale.

Figure 3 reports both metrics under GPT-4o and DeepSeek-V4-Flash judges. Detailed MentalChat16K dimensions and WAIOS Goal/Task/Bond scores are provided in Appendix E. Overall, PsyMem-trained variants perform best on most MentalChat16K comparisons and several GPT-4o WAIOS comparisons, whereas DeepSeek-V4-Flash WAIOS results are more mixed, with SimPsyDial remaining competitive for some base models. Additional WAIOS results under Qwen3.6 and GPT-4o are reported in Table 17.

The WAIOS results provide a more conservative, alliance-focused assessment. For DeepSeek-7B-Chat under GPT-4o evaluation, cross-session PsyMem achieves the highest overall WAIOS score (5.46), outperforming SimPsyDial (5.19), MentalChat16K (4.59), and DiaSynth (2.66). Across judges, PsyMem shows more consistent gains on the broader MentalChat16K criteria, whereas its WAIOS improvements vary by base model and judge. Full dimension-level WAIOS results are reported in Appendix E.1.

4.5 Ablation Study (RQ4)

We first ablate three synthesis components: client-utterance reflection, structured persona elicitation, and cross-session memory. We then compare the full role-asymmetric information-flow design with four context variants—*w/o Memory*, *Plain Text*, *Summary*, and *Shared Memory*. Table 1 reports within-session WAIOS, summary-based MCAF, the reported identity indicator, and dialogue statistics.

The component ablations produce modest but consistent WAIOS decreases: removing reflection lowers the overall score by 0.09, and removing structured persona elicitation lowers it by 0.10. The

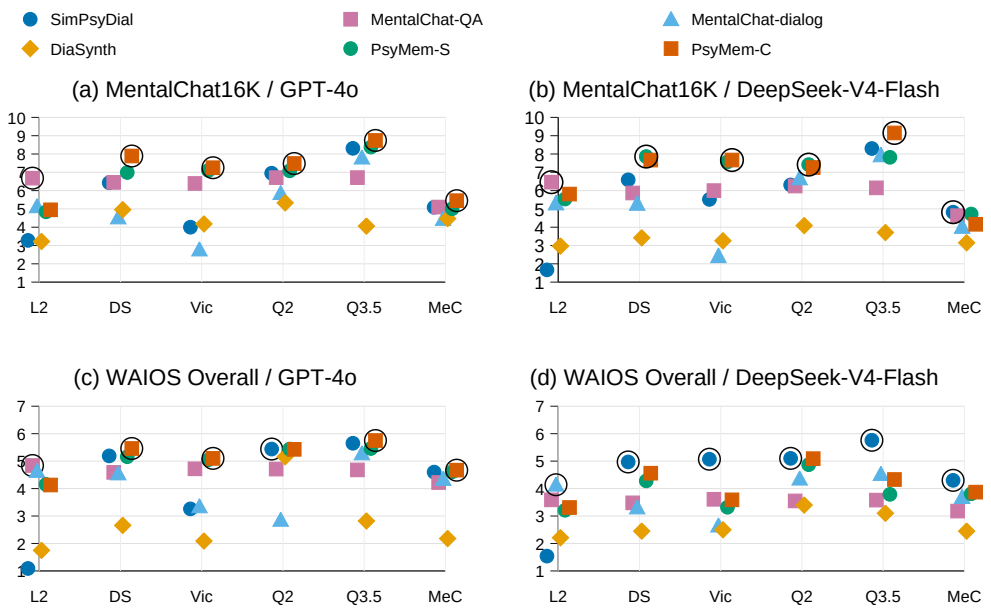

Figure 3: Downstream counseling quality after fine-tuning. Rows show MentalChat16K and WAIOS scores; columns show GPT-4o and DeepSeek-V4-Flash judges. L2, DS, Vic, Q2, Q3.5, and MeC denote the base models; S/C denote single-/cross-session PsyMem. Black rings mark the best training data per base model and judge.

Table 1: Component ablations and cross-session context comparison.
The full PsyMem results are reported as absolute values, while all other variants are shown as relative changes with respect to the full model.

Model Variant	**WAIOS (Intra-session)**				**MCAF (Cross-session)**		**Dialogue Statistics**	
	Goal	Task	Bond	Overall	**MCAF Average**	Identity (0–100)	Avg. Words per Utterance	Avg. Turns
PsyMem (Full)	5.95	5.92	6.24	6.04	4.96	95	105	61
w/o Reflection	↓0.16	↓0.03	↓0.06	↓0.09	↑0.03	±0	↓4	↓16
w/o Char. Elicitation	↓0.13	↓0.05	↓0.12	↓0.10	↑0.05	±0	↓13	↓14
w/o Memory	↓0.11	↓0.05	↓0.12	↓0.10	↓3.91	↓28	↓22	↓23
Plain Text	↓0.06	↓0.08	↓0.08	↓0.07	↓0.02	↑1	↓13	↓38
Summary	↓0.02	0.00	↓0.08	↓0.02	↓0.01	±0	↓18	↓28
Shared Memory	↓0.06	↓0.17	↑0.07	↓0.06	↓0.01	↑1	↓20	↓27

Note: The **MCAF Average** is the arithmetic mean of the four MCAF dimensions (Client-State Continuity, Counselor-Memory Continuity, Within-session Alignment, and Multi-session Consistency); full sub-dimension scores are reported in Appendix D.

w/o Memory condition produces the largest continuity loss, reducing MCAF by 3.91 points, the reported identity indicator by 28 points, average words per utterance by 22, and average turns by 23.

The memory-enabled configurations achieve comparable summary-based MCAF averages (4.94–4.96), while the full role-asymmetric design attains the highest overall WAIOS score (6.04, compared with 5.97 for Plain Text, 6.02 for Summary, and 5.98 for Shared Memory) and produces the longest interactions. Relative to Shared Memory, the full design scores 0.06 higher on Goal, 0.17 higher on Task, and 0.06 higher overall, whereas Shared Memory scores 0.07 higher on Bond and one point higher on the reported identity indicator; the full design also averages 20 more words per utterance and 27 more turns. Thus, within this controlled ablation, the main observed advantage of role asymmetry lies in Goal/Task alignment and overall WAIOS.

5 LIMITATIONS

PsyMem has three principal limitations. First, its fixed persona, simplified concern-state transitions, and text-only setting may underrepresent the variability, multimodality, and non-cooperative dynamics of real counseling. Second, longitudinal evaluation relies on LLM judgments of stored role states and deliberately disconnected controls, without transcript-level or human cross-session validation. Shared Memory also achieves nearly identical MCAF (4.95 vs. 4.96), limiting claims that role asymmetry improves continuity. Third, evaluation validity is constrained by non-clinical, single-rater human judgments, one composite item per WAI-O dimension, and moderate human–automatic agreement; using GPT-4o for both generation and primary automatic evaluation may further introduce judge-specific bias. Future work should add clinically informed, multi-rater, transcript-level longitudinal evaluation with independent judges.

6 CONCLUSION

We introduced PsyMem, a framework for longitudinal counseling-dialogue synthesis that jointly targets substantive within-session interaction and cross-session continuity. Rather than treating longitudinal context as a shared history, PsyMem formulates context management as role-asymmetric information flow, maintaining a fixed private client persona and an evolving concern state while deriving counselor memory from observed dialogue. Under GPT-4o evaluation, PsyMem achieves the highest WAIOS score among the compared datasets and a summary-level MCAF score of 4.96/5; removing cross-session state reduces MCAF by approximately 3.9 points. Among the evaluated memory-enabled configurations, the role-asymmetric design maintains comparable summary-level continuity while attaining the highest observed overall WAIOS. Across six base models, models fine-tuned on PsyMem-generated dialogues perform best in most MentalChat16K comparisons, whereas WAIOS transfer remains model- and judge-dependent. Together, the linked corpus and MCAF provide a controlled setting for jointly studying within-session interaction quality and cross-session continuity in stored role states. More broadly, these findings motivate treating information provenance and role visibility as first-class design choices alongside memory construction and retrieval in longitudinal dialogue synthesis. Future work will extend PsyMem to multimodal longitudinal interactions and evaluate cross-session coherence through multi-rater human assessment of complete dialogue sequences.

AI USE STATEMENT

GPT-4o was used to construct synthetic personas and counseling dialogues and to generate automatic evaluation scores, as described in the method and evaluation sections. A generative AI assistant was also used to assist with manuscript editing and restructuring for clarity and readability. The authors are responsible for the final content of the submission.

ETHICS STATEMENT

PsyMem uses help-seeking posts from the publicly available PsyQA dataset as seed data for synthesizing counseling dialogues. These source data have undergone privacy-preserving processing and de-identification. Generated dialogues are research artifacts, not clinical evidence or deployment-ready systems. Longitudinal coherence may increase over-reliance; systems trained on such data should not substitute for licensed mental-health professionals or crisis services. The human evaluation rated synthetic dialogue artifacts only; it collected no clinical information, and no participants, patients, or vulnerable populations were involved.

REPRODUCIBILITY STATEMENT

Appendix A.2 documents dataset construction, Appendix B.1 reports fine-tuning settings, Appendix C defines the evaluation rubrics, and Appendix F provides the generation and evaluation prompts. The anonymized PsyMem dataset, evaluation resources, and implementation code are available at `https://anonymous.4open.science/r/PsyMem-B34E37666`.

REFERENCES

Edward S. Bordin. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3):252–260, 1979. doi: 10.1037/h0085885. URL https://doi.org/10.1037/h0085885.

John Bowlby. *Attachment and Loss: Vol. 1. Attachment*. Basic Books, New York, 1969.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/, 2023. Blog post, LMSYS Organization.

Paul T. Costa and Robert R. McCrae. Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4(1):5–13, 1992. doi: 10.1037/1040-3590.4.1.5. URL https://doi.org/10.1037/1040-3590.4.1.5.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, et al. DeepSeek LLM: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Adam O. Horvath and Leslie S. Greenberg. Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, 36(2):223–233, 1989. doi: 10.1037/0022-0167.36.2.223. URL https://doi.org/10.1037/0022-0167.36.2.223.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685. Published at ICLR 2022.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13584–13606. Association for Computational Linguistics, 2023.

Pengcheng Jiang, Siru Ouyang, Yizhu Jiao, Ming Zhong, Runchu Tian, and Jiawei Han. A survey on retrieval and structuring augmented generation with large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6032–6042. Association for Computing Machinery, 2025. doi: 10.1145/3711896.3736557. URL https://doi.org/10.1145/3711896.3736557.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10358–10376, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.577. URL https://doi.org/10.18653/v1/2023.acl-long.577.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1230. URL https://doi.org/10.18653/v1/D16-1230.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.17651.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.64. URL https://doi.org/10.18653/v1/2020.acl-main.64.

OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, May 2024. Accessed: 2026-05-02.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, JingYuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, and Liang He. PsychEval: A multi-session and multi-therapy benchmark for high-realism ai psychological counselor. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22192–22212. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1115. URL https://aclanthology.org/2026.findings-acl.1115.pdf.

Huachuan Qiu and Zhenzhong Lan. Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions, 2024. URL https://arxiv.org/abs/2408.15787.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 615–636, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.34.

Qwen Team. Qwen3.5-9B. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026-05-07.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1534. URL https://doi.org/10.18653/v1/P19-1534.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5276. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.425. URL https://doi.org/10.18653/v1/2020.emnlp-main.425.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.11366.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for*

Computational Linguistics: ACL-IJCNLP 2021, pp. 1489–1503. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.130. URL https://doi.org/10.18653/v1/2021.findings-acl.130.

Sathya Krishnan Suresh, Mengjun Wu, Tushar Pranav, and Eng Siong Chng. DiaSynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 673–690. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.40. URL https://doi.org/10.18653/v1/2025.findings-naacl.40.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Bichen Wang, Junzhe Wang, Yixin Sun, Xing Fu, Yanyan Zhao, and Bing Qin. Psychological counseling cannot be achieved overnight: Automated psychological counseling through multi-session conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16593–16609. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.818. URL https://aclanthology.org/2026.findings-acl.818/.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23221–23235. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1192. URL https://aclanthology.org/2025.findings-acl.1192/.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. MentalChat16K: A benchmark dataset for conversational mental health assistance. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 5367–5378. ACM, 2025. doi: 10.1145/3711896.3737393. URL https://doi.org/10.1145/3711896.3737393.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197. Association for Computational Linguistics, 2022.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13947–13966. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.830. URL https://doi.org/10.18653/v1/2024.findings-acl.830.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685. Published at NeurIPS 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731. Association for the Advancement of Artificial Intelligence, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

Table 2: Published data units and longitudinal design of selected counseling and dialogue resources. A dash denotes a capability not reported by the cited resource; sizes follow the sources' published units. **MSC** and **Conversation Chronicles** are included as *general-domain* references for long-term dialogue, not as counseling-specific resources.

Resource	Provenance / unit	Published size	Multi-turn	Linked sessions	State / memory
REALPSYDIAL Li et al. (2023)	Real counseling session	2,382 sessions	Yes	–	–
CPSYCOUND Zhang et al. (2024)	Report-derived reconstructed dialogue	3,134 dialogues	Yes	–	–
PSYQA Sun et al. (2021)	Online question–answer pair	22,346 Q; 56,063 A	–	–	–
MENTALCHAT16K Xu et al. (2025)	Transcript-derived and synthetic QA pair	∼16K pairs	–	–	–
EMPATHETIC DIALOGUES Rashkin et al. (2019)	Crowdsourced open-domain dialogue	∼25K dialogues	Yes	–	–
SHARMA ET AL. Sharma et al. (2020)	Online support post–response pair	10K annotated pairs	–	–	–
SIMPSYDIAL Qiu & Lan (2024)	LLM role-play counseling dialogue	1,000 dialogues	Yes	–	–
DIASYNTH Suresh et al. (2025)	LLM general-domain dialogue	1,035–1,410 per setting	Yes	–	–
General-domain (Non-counseling) Longitudinal Resources					
MSC Xu et al. (2022)	Crowdsourced human-human chat	5,001 episodes	Yes	Yes	Human summaries
CONVERSATION CHRONICLES Jang et al. (2023)	LLM-generated (ChatGPT)	200K episodes; 1M sessions	Yes	Yes	Chronological summaries
PsyMem	LLM client-linked counseling sessions	150 clients; 600 sessions	Yes	Yes	Yes

† MSC and Conversation Chronicles are open-domain datasets included solely for comparison of long-term context modeling. They are not designed for psychological counseling; their memory mechanisms are not directly comparable to counseling-specific state tracking.

A DATASET CONSTRUCTION DETAILS

This appendix details the comparison resources, PsyMem construction pipeline, and stored artifacts used in the experiments.

A.1 COMPARISON DATASETS

Table 2 compares the published data units, sizes, and longitudinal design of eight selected resources alongside PsyMem. PsyQA, SimPsyDial, DiaSynth, and MentalChat16K also serve as experimental comparisons (Section 4).

A.2 PSYMEM CONSTRUCTION PIPELINE

PsyMem is constructed through the four-stage pipeline in Table 3. The pipeline first anchors each client in a real help-seeking post, then expands the initial description into a structured persona, generates four counseling sessions, and finally extracts role-specific state summaries after each session.

Table 3: PsyMem construction pipeline.

Stage	Procedure
Client selection	We sample 150 help-seeking posts from the PsyQA subset used by SimPsyDial, depressive and anxiety-related concerns.
Persona construction	GPT-4o expands each initial description into a structured hidden persona. The client agent uses this persona; the counselor agent uses dialogue-derived context.
Session generation	GPT-4o client and counselor agents conduct reflection-guided role-playing. The reflection mechanism checks client utterances for completeness, concrete detail, and persona consistency. Each session ends at a natural closing point or when it becomes excessively long.
Memory extraction	After each session, we extract client-side and counselor-side memory summaries. These summaries update the current concern state and store dialogue-derived facts, prior suggestions, the client's personality characteristics, and the evolving state of the client's concerns for later sessions.

Table 4: Persona fields used in PsyMem construction.

Category	Fields
Demographics	Age, gender, occupation, education, family status.
Current concern	Main help-seeking issue, onset, severity, recent triggers.
Stable background	Life context, recurring interpersonal patterns, coping tendencies.
Interaction style	Disclosure tendency, response style, resistance level, emotional tone.
Recent state	Current emotional state, recent events, short-term concerns.
Counseling goals	Short-term coping goals and long-term adjustment goals.

Table 5: Structured persona sections used in PsyMem.

Section	Content	Role in synthesis
Core profile	Demographics, current concern, help-seeking motivation, resistance level.	Initial grounding
Background context	Family context, education or work background, important life events.	Long-form consistency
Interaction style	Disclosure tendency, emotional expression, avoidance or engagement patterns.	Response control
Current situation	Recent events, current stressors, immediate goals.	Session initialization
Summary state	Stable traits, recurring concerns, possible entry points for dialogue.	Persona grounding

A.3 Persona Construction Details

Persona construction separates stable identity information from evolving concern state. Stable sections are used for long-form role-playing consistency, whereas the memory module updates session-specific changes over time. Tables 4 and 5 list the corresponding fields.

Corpus statistics. The evaluation corpus contains 600 sessions from 150 generated client personas seeded by sampled help-seeking posts, with four sessions per client. Each session averages 61 turns, resulting in approximately 244 turns per four-session sequence.

B Implementation and Generation Settings

B.1 Fine-tuning Hyperparameters

All baseline models are fine-tuned with LoRA Hu et al. (2021) for parameter-efficient adaptation. Table 6 reports the shared training configuration.

Table 6: Fine-tuning hyperparameters.

Hyperparameter	Value
Adaptation method	LoRA
LoRA rank r	8
LoRA scaling α	16
LoRA dropout	0.1
Epochs	5
Peak learning rate	2×10^{-4}
Scheduler	Cosine
Precision	bf16
Maximum gradient norm	1.0
Cutoff length	1024 tokens
Per-device batch size	2
Gradient accumulation steps	4
Validation split	10%
Maximum training samples	100,000

Table 7: Datasets and generation modes used in experiments. Constructed four-item groups are used only as disconnected controls for the MCAF sanity check.

Dataset	Type	Sessions	Avg. turns/session
PsyQA	Real, single-turn	1 ($4^{\dagger}$)	1
SimPsyDial	Synthetic, multi-turn	1 ($4^{\dagger}$)	13
DiaSynth	Synthetic, multi-turn	1 ($4^{\dagger}$)	4
MentalChat16K	Synthetic / QA	1 ($4^{\dagger}$)	–
PsyMem	Synthetic, cross-session	**4**	**61**

† denotes a constructed four-item disconnected control formed from independent items; it is not a native linked-session sequence.

Table 8: Artifacts stored after each PsyMem session.

Artifact	Content
Dialogue record	Complete utterance sequence for the session.
Role-specific state summaries	Updated client concern state and counselor memory, including facts, prior suggestions, the client's personality characteristics, and the evolving state of the client's concerns.
Reflection log	Verifier outcomes, rework flags, and failure categories for client utterances.

B.2 Generation Modes

During inference, each fine-tuned model generates dialogues according to the native format of its training corpus. PsyMem supports both single-session and cross-session generation, while comparison datasets follow their original data formats. Table 7 summarizes the generation modes used in fine-tuning and MCAF evaluation. For the MCAF sanity check only, we construct four-item disconnected controls by grouping independently sampled examples from each non-longitudinal comparison dataset. PsyQA uses four single-turn counseling pairs, DiaSynth and MentalChat16K use four single-session or QA-style items, and SimPsyDial uses four independently generated multi-turn sessions. These controls are not native linked-session outputs and are not treated as longitudinal baselines.

B.3 Stored Artifacts

For reproducibility, PsyMem stores the three artifacts in Table 8 after each session.

Table 9: Scoring rubric for the Memory Continuity Analysis Framework (MCAF).

Dimension	Aspect	Score 1	Score 3	Score 5
Client-State Continuity	Whether the client concern state remains consistent while showing plausible changes.	Problems or background details are contradictory or unrelated.	Some continuity exists, but changes are weakly motivated or incomplete.	Core concerns are preserved and changes are gradual and well motivated.
Counselor-Memory Continuity	Whether prior disclosures, suggestions, and progress are preserved.	Later summaries omit prior disclosures and suggestions.	Some prior information is recalled, but follow-up is incomplete.	Later summaries consistently preserve prior disclosures, suggestions, and reported progress.
Within-session Alignment	Whether client and counselor states agree within each session.	The two sides contradict each other on key facts or concerns.	The two sides are mostly aligned but miss some details.	Key facts, concerns, and interaction context are well aligned.
Multi-session Consistency	Whether the four sets of session summaries form a coherent sequence.	Session summaries appear unrelated or disconnected.	Some cross-session links exist, but the global sequence is uneven.	The sequence shows stable identity and meaningful carry-over.

Table 10: Scoring rubric for longitudinal persona consistency.

Score range	Interpretation	Criteria
0–20	Clearly different identities	Severe contradictions in background, core concerns, or response style.
21–40	Mostly inconsistent	Some surface similarities exist, but major discrepancies remain.
41–60	Partially consistent	Identity is partly preserved, but notable inconsistencies appear.
61–79	Mostly consistent	The same identity is mostly preserved, with minor gaps.
80–100	Highly consistent	Background, concerns, response style, and interaction patterns are consistently preserved.

C Evaluation Rubrics

C.1 Memory Continuity Analysis Framework

MCAF evaluates the continuity of stored client concern states and counselor memories from a generated four-session sequence. Each dimension is scored on a 1–5 scale, where 1 indicates severe discontinuity and 5 indicates strong continuity. Scores 2 and 4 are reserved for intermediate cases. The overall MCAF score is the arithmetic mean of the four dimensions.All judgments use GPT-4o with temperature 0.1 and a single pass; no ensembling or self-consistency is applied. The MCAF judge uses temperature 0.2.

C.2 Longitudinal Persona Consistency Check

We separately ask a zero-shot LLM judge whether four session memory records appear to belong to the same client. The prompt returns a Boolean classification and a 0–100 confidence value based on the records' background details, concerns, and interaction patterns (Appendix F.3). Table 10 defines descriptive ranges for this study-specific indicator.Corpus-level identity values are the mean judge confidence across the sampled sequences, rescaled to 0-100.

Table 11: GPT-4o WAIOS scores across the five evaluated datasets.

Dataset	Goal	Task	Bond	Average
PsyQA	4.16	3.85	4.53	4.18
SimPsyDial	5.65	5.45	6.03	5.71
MentalChat16K	5.50	5.28	5.81	5.53
DiaSynth	4.67	4.66	5.64	4.99
PsyMem	**5.95**	**5.92**	**6.24**	**6.04**

Table 12: Human-evaluator WAIOS scores across the five evaluated datasets.

Dataset	Goal	Task	Bond	Average
PsyQA	4.58	4.70	4.70	4.66
SimPsyDial	5.47	4.97	4.70	5.05
MentalChat16K	4.83	5.22	5.35	5.13
DiaSynth	5.47	5.47	4.97	5.30
PsyMem	**6.23**	**6.50**	**6.75**	**6.49**

Table 13: Human-evaluator MentalChat16K counseling-quality scores across the five evaluated datasets.

Dataset	Active Listening	Empathy & Validation	Safety & Trustworthiness	Open-mindedness & Non-judgment	Clarity & Encouragement	Boundaries & Ethical	Holistic Approach	Average Score
PsyQA	5.14	6.36	7.10	7.28	6.90	6.54	6.72	6.58
SimPsyDial	6.28	7.28	7.28	7.82	7.46	7.28	7.10	7.22
MentalChat16K	6.86	7.28	7.28	7.46	7.64	7.10	6.36	7.14
DiaSynth	6.86	6.72	7.82	7.82	7.46	7.64	6.90	7.32
PsyMem	**9.42**	**9.46**	**9.10**	**9.28**	**9.46**	**9.28**	**9.64**	**9.38**

Table 14: Full MCAF and persona-consistency scores summarized in Figure 2(b). Identity-consistency indicator is reported on the original 0–100 scale.MCAF average and the identity indicator rank the five corpora identically.

Dataset	Client-State Continuity	Counselor-Memory Continuity	Within-session Alignment	Multi-session Consistency	MCAF Average	Persona ID
PsyQA	1.35	1.40	1.45	1.45	1.41	64
SimPsyDial	1.00	1.00	1.00	1.00	1.00	55
MentalChat16K	1.48	1.55	1.50	1.50	1.51	62
DiaSynth	1.75	1.80	1.70	1.75	1.75	67
PsyMem	**4.96**	**4.96**	**4.95**	**4.95**	**4.96**	**95**

D Dataset-Level Evaluation Details

Tables 11–14 report the WAIOS, human-evaluator, and MCAF sub-dimension scores summarized in Figure 2. To further assess the cross-domain generalizability of the MCAF diagnostic, Table 15 extends this evaluation to two general-domain multi-session datasets under a zero-shot setting. Internal consistency (Cronbach's alpha across the five datasets) is 0.984 for the GPT-4o WAIOS ratings, 0.935 for the human-evaluator WAIOS ratings, 0.982 for the human-evaluator MentalChat16K ratings, and 1.000 for the MCAF dimension scores; for MCAF we therefore also report per-dimension agreement in Table 14.

Table 15: Zero-shot MCAF evaluation on general-domain multi-session datasets. These datasets are open-domain and are not designed for psychological counseling. For these datasets, *counseling coherence* is used as a proxy for *within-session alignment*, as explicit alignment annotations are not available. This score should not be interpreted as a measure of counseling quality.

Dataset	Client-State Continuity	Counselor-Memory Continuity	Within-session Alignment	Multi-session Consistency	MCAF Average	Persona ID
Multi-Session Chat	4.60	4.60	4.60	4.40	4.55	98
CONVERSATION CHRONICLES	3.88	3.56	3.14	3.58	3.54	87.2
PsyMem (Full)	**4.96**	**4.96**	**4.95**	**4.95**	**4.96**	**95**

Note: The *within-session alignment* column for general-domain datasets is approximated by the *counseling coherence* score reported in the original evaluation. This mapping is used only to test the robustness of the MCAF diagnostic across domains and does not imply validity as a counseling measure.

E Full Fine-tuning Evaluation Results

E.1 Full WAIOS Fine-tuning Results

Table 16 reports the full Goal, Task, Bond, and Overall WAIOS scores underlying the main-text downstream analysis.

Table 16: Full WAIOS fine-tuning results across models and datasets under two judges. †: evaluated on single-session multi-turn dialogues; •: evaluated on QA-based single-turn pairs; §: evaluated on cross-session multi-session continuums.

Model	Dataset	GPT-4o judge				DeepSeek-V4-Flash judge			
		Goal	Task	Bond	Overall	Goal	Task	Bond	Overall
LLaMA2-7B	SimPsyDial †	1.09	1.08	1.10	1.09	1.56	1.50	1.55	1.54
	MentalChat16K •	4.72	4.50	5.31	4.84	3.64	2.59	4.53	3.59
	MentalChat16K †	4.75	4.05	5.13	4.64	4.13	2.45	5.84	4.14
	DiaSynth †	1.55	1.54	2.15	1.75	2.22	2.16	2.26	2.21
	PsyMem †	4.36	3.71	4.38	4.15	3.50	2.63	3.51	3.21
	PsyMem §	4.29	3.69	4.42	4.13	3.50	2.73	3.71	3.31
DeepSeek-7B-Chat	SimPsyDial †	4.91	5.05	5.62	5.19	4.94	4.48	5.49	4.97
	MentalChat16K •	4.57	4.28	4.91	4.59	3.60	2.61	4.23	3.48
	MentalChat16K †	4.85	3.98	4.80	4.55	3.85	2.17	3.91	3.31
	DiaSynth †	2.05	2.15	3.77	2.66	2.29	2.18	2.88	2.45
	PsyMem †	5.05	4.97	5.46	5.16	3.98	3.55	**5.31**	4.28
	PsyMem §	**5.33**	**5.16**	**5.89**	**5.46**	4.37	4.00	5.32	4.56
Vicuna-7B-v1.5	SimPsyDial †	3.14	3.01	3.62	3.26	5.19	4.59	5.45	5.07
	MentalChat16K •	4.61	4.41	5.14	4.72	3.75	2.61	4.46	3.61
	MentalChat16K †	3.77	2.80	3.48	3.35	3.18	2.09	2.69	2.65
	DiaSynth †	1.73	1.75	2.78	2.09	2.45	2.36	2.68	2.50
	PsyMem †	4.81	4.79	5.56	5.05	3.11	2.79	4.06	3.32
	PsyMem §	**4.83**	**4.91**	**5.57**	**5.10**	3.35	3.09	4.34	3.59
Qwen2-7B-Instruct	SimPsyDial †	5.23	5.24	5.82	5.44	4.42	4.85	5.86	5.10
	MentalChat16K •	4.63	4.41	5.09	4.71	3.65	2.63	4.37	3.55
	MentalChat16K †	5.07	4.89	5.54	5.17	4.38	3.54	5.17	4.36
	DiaSynth †	2.13	2.31	4.13	2.86	3.30	3.14	3.78	3.40
	PsyMem †	**5.25**	5.16	**5.90**	**5.44**	4.21	4.68	5.70	4.87
	PsyMem §	5.19	**5.25**	5.87	5.43	4.50	4.87	5.91	5.09
Qwen3.5-9B-Thinking	SimPsyDial †	5.38	**5.69**	5.87	5.65	5.75	5.51	6.04	5.76
	MentalChat16K •	4.65	4.33	5.07	4.68	3.74	2.64	4.35	3.58
	MentalChat16K †	5.05	5.17	5.63	5.28	4.43	4.04	5.10	4.53
	DiaSynth †	2.40	2.71	3.34	2.82	3.08	2.92	3.29	3.10
	PsyMem †	5.50	5.09	5.80	5.46	3.73	2.98	4.67	3.79
	PsyMem §	**5.68**	5.57	**6.01**	**5.75**	3.75	4.26	4.97	4.33
MeChat	SimPsyDial †	4.36	4.43	5.03	4.60	4.25	3.66	4.99	4.30
	MentalChat16K •	4.23	3.93	4.52	4.22	3.36	2.47	3.72	3.18
	MentalChat16K †	4.36	3.96	4.73	4.35	3.81	2.94	4.31	3.69
	DiaSynth †	1.67	1.69	3.18	2.18	2.38	2.24	2.72	2.45
	PsyMem †	4.41	**4.45**	4.96	4.61	3.93	3.14	4.36	3.81
	PsyMem §	**4.54**	4.35	**5.12**	**4.67**	3.98	3.06	4.58	3.87

Table 17: Supplementary WAIOS results under Qwen3.6 and GPT-4o evaluators for assessing LLM-judge generalization. †: evaluated on single-session multi-turn dialogues; •: evaluated on QA-based single-turn pairs; §: evaluated on cross-session multi-session continuums.

Model	Dataset	Qwen3.6 evaluator				GPT-4o evaluator			
		Goal	Task	Bond	Overall	Goal	Task	Bond	Overall
LLaMA2-7B	SimPsyDial †	1.31	1.31	1.35	1.32	1.09	1.08	1.10	1.09
	MentalChat16K •	2.64	2.44	3.07	2.72	4.72	4.50	5.31	4.84
	MentalChat16K †	2.95	2.89	3.36	3.07	4.75	4.05	5.13	4.64
	DiaSynth †	1.21	1.22	1.53	1.32	1.55	1.54	2.15	1.75
	PsyMem †	2.19	2.13	2.92	2.41	4.36	3.71	4.38	4.15
	PsyMem §	2.63	2.53	**3.37**	**2.84**	4.29	3.69	4.42	4.13
DeepSeek-7B-Chat	SimPsyDial †	4.93	4.47	5.00	4.80	4.91	5.05	5.62	5.19
	MentalChat16K •	2.54	2.42	3.73	2.90	4.57	4.28	4.91	4.59
	MentalChat16K †	2.51	2.35	2.94	2.60	4.85	3.98	4.80	4.55
	DiaSynth †	1.67	1.65	3.75	2.35	2.05	2.15	3.77	2.66
	PsyMem †	4.98	4.07	5.65	4.90	5.05	4.97	5.46	5.16
	PsyMem §	**5.35**	4.43	**5.86**	**5.21**	**5.33**	**5.16**	**5.89**	**5.46**
Vicuna-7B-v1.5	SimPsyDial †	4.73	5.11	5.39	5.08	3.14	3.01	3.62	3.26
	MentalChat16K •	2.60	2.48	3.88	2.98	4.61	4.41	5.14	4.72
	MentalChat16K †	1.65	1.61	1.77	1.68	3.77	2.80	3.48	3.35
	DiaSynth †	1.67	1.63	3.03	2.11	1.73	1.75	2.78	2.09
	PsyMem †	**4.88**	3.18	5.18	4.41	4.81	4.79	5.56	5.05
	PsyMem §	4.75	3.56	**5.57**	4.63	**4.83**	**4.91**	**5.57**	**5.10**
Qwen2-7B-Instruct	SimPsyDial †	6.66	6.29	6.50	6.48	5.23	5.24	5.82	5.44
	MentalChat16K •	2.61	2.46	3.80	2.96	4.63	4.41	5.09	4.71
	MentalChat16K †	5.80	4.80	5.60	5.40	5.07	4.89	5.54	5.17
	DiaSynth †	1.96	1.94	3.85	2.58	2.13	2.31	4.13	2.86
	PsyMem †	6.55	6.18	**6.59**	6.44	**5.25**	5.16	**5.90**	**5.44**
	PsyMem §	**6.74**	**6.38**	6.53	**6.55**	5.19	**5.25**	5.87	5.43
Qwen3.5-9B-Thinking	SimPsyDial †	6.33	5.75	6.43	6.17	5.38	**5.69**	5.87	5.65
	MentalChat16K •	2.74	2.54	3.96	3.08	4.65	4.33	5.07	4.68
	MentalChat16K †	5.62	4.76	5.86	5.41	5.05	5.17	5.63	5.28
	DiaSynth †	2.39	2.29	3.07	2.58	2.40	2.71	3.34	2.82
	PsyMem †	4.59	3.83	5.78	4.73	5.50	5.09	5.80	5.46
	PsyMem §	**6.65**	4.95	**6.49**	6.03	**5.68**	5.57	**6.01**	**5.75**
MeChat	SimPsyDial †	2.86	2.79	3.28	2.98	4.36	4.43	5.03	4.60
	MentalChat16K •	2.34	2.33	3.39	2.69	4.23	3.93	4.52	4.22
	MentalChat16K †	1.89	1.91	2.46	2.08	4.36	3.96	4.73	4.35
	DiaSynth †	1.15	1.17	3.05	1.79	1.67	1.69	3.18	2.18
	PsyMem †	2.62	2.51	3.27	2.80	4.41	**4.45**	4.96	4.61
	PsyMem §	2.56	2.40	3.12	2.69	**4.54**	4.35	**5.12**	**4.67**

E.2 FULL MENTALCHAT16K FINE-TUNING RESULTS

Tables 18 and 19 expand Figure 3 with all seven MentalChat16K dimensions, reported separately for the GPT-4o and DeepSeek-V4-Flash judges.

Table 18: Full MentalChat16K dimensions across models and datasets, evaluated by GPT-4o. †: single-session multi-turn dialogues; •: QA-based single-turn pairs; §: cross-session multi-session sequences.

Model	Dataset	Active Listening	Empathy & Validation	Safety & Trustworthiness	Open-mindedness & Non-judgment	Clarity & Encouragement	Boundaries & Ethical	Holistic Approach	Average Score
LLaMA2-7B	SimPsyDial †	2.40	2.93	4.57	5.03	2.57	2.00	3.43	3.28
	MentalChat16K •	6.66	7.57	6.44	7.51	6.27	6.20	6.11	6.68
	MentalChat16K †	5.00	5.80	5.20	7.00	4.00	3.60	5.40	5.14
	DiaSynth †	2.56	2.24	4.80	4.25	2.81	3.71	2.14	3.22
	PsyMem †	4.18	4.09	6.55	5.91	3.64	3.09	6.45	4.84
	PsyMem §	4.13	4.63	6.88	6.13	3.50	3.13	6.25	4.95
DeepSeek-7B-Chat	SimPsyDial †	5.88	6.35	7.85	7.74	4.90	5.35	6.96	6.43
	MentalChat16K •	6.35	7.17	6.10	7.42	6.26	6.16	5.70	6.45
	MentalChat16K †	3.27	4.04	6.25	6.40	3.17	3.44	5.15	4.53
	DiaSynth †	4.54	3.58	6.87	6.40	4.78	5.20	3.33	4.96
	PsyMem †	6.69	7.29	8.24	8.06	5.71	**7.65**	5.32	6.99
	PsyMem §	**7.60**	**8.25**	**9.10**	**8.95**	**6.45**	5.95	**8.90**	**7.89**
Vicuna-7B-v1.5	SimPsyDial †	3.00	3.80	5.02	5.87	2.22	5.01	3.07	4.00
	MentalChat16K •	6.48	7.15	6.07	7.51	**6.25**	5.55	5.69	6.39
	MentalChat16K †	2.00	2.56	3.58	4.52	1.91	1.70	3.19	2.78
	DiaSynth †	3.69	2.85	5.77	5.21	3.85	4.93	2.93	4.18
	PsyMem †	7.00	6.92	8.42	8.17	5.58	5.33	8.25	7.10
	PsyMem §	**7.00**	**7.17**	**8.67**	**8.25**	5.75	**5.63**	**8.25**	**7.25**
Qwen2-7B-Instruct	SimPsyDial †	6.42	6.44	8.18	7.78	**6.68**	**7.80**	5.38	6.95
	MentalChat16K •	6.58	7.48	6.59	7.66	6.20	6.42	6.02	6.71
	MentalChat16K †	5.02	5.77	7.45	7.39	4.57	4.55	**6.30**	5.86
	DiaSynth †	4.99	4.02	7.36	6.58	4.96	5.74	3.68	5.33
	PsyMem †	6.87	7.51	8.31	8.46	5.69	6.92	5.77	7.08
	PsyMem §	**7.08**	**8.03**	**8.76**	**8.66**	6.26	7.42	6.18	**7.48**
Qwen3.5-9B-Thinking	SimPsyDial †	**8.55**	8.91	8.80	9.04	7.60	7.89	7.40	8.31
	MentalChat16K •	7.06	7.73	6.38	7.61	6.04	6.27	5.91	6.71
	MentalChat16K †	8.06	8.78	8.32	8.83	6.87	6.61	7.13	7.80
	DiaSynth †	3.50	2.74	5.34	5.06	3.78	4.07	3.90	4.06
	PsyMem †	8.40	8.87	**9.47**	9.33	6.60	6.93	**9.07**	8.38
	PsyMem §	**8.55**	**9.13**	9.40	**9.47**	**8.33**	7.60	8.67	**8.74**
MeChat	SimPsyDial †	4.08	4.96	6.53	6.49	3.77	4.53	5.30	5.09
	MentalChat16K •	4.41	5.47	5.06	6.52	**5.09**	4.89	4.29	5.10
	MentalChat16K †	3.47	4.33	5.79	5.85	3.22	4.09	4.41	4.45
	DiaSynth †	4.11	3.27	6.16	5.68	4.18	4.79	3.09	4.47
	PsyMem †	4.17	4.14	6.55	6.35	4.30	**6.08**	3.48	5.01
	PsyMem §	**4.50**	**5.50**	**7.00**	**7.00**	3.55	4.45	**6.15**	**5.45**

Table 19: Full MentalChat16K dimensions across models and datasets, evaluated by DeepSeek-V4-Flash. The dataset markers are defined in Table 18.

Model	Dataset	Active Listening	Empathy & Validation	Safety & Trustworthiness	Open-mindedness & Non-judgment	Clarity & Encouragement	Boundaries & Ethical	Holistic Approach	Average Score
LLaMA2-7B	SimPsyDial †	1.61	1.60	1.93	2.45	1.12	1.78	1.21	1.67
	MentalChat16K •	6.17	7.06	6.21	7.98	6.49	5.94	5.36	6.46
	MentalChat16K †	4.71	5.97	6.59	7.38	3.50	4.59	4.59	5.33
	DiaSynth †	2.49	1.73	4.57	4.27	2.75	3.23	1.74	2.97
	PsyMem †	5.50	**6.30**	6.20	6.80	4.65	4.45	4.85	5.54
	PsyMem §	5.80	6.15	**6.75**	**7.00**	4.80	4.70	**5.50**	5.81
DeepSeek-7B-Chat	SimPsyDial †	6.39	6.67	7.60	7.70	5.50	5.80	6.46	6.59
	MentalChat16K •	5.58	6.12	5.52	7.44	6.30	5.29	4.85	5.87
	MentalChat16K †	4.89	5.43	6.41	6.94	4.08	4.32	5.09	5.31
	DiaSynth †	2.90	1.81	5.22	5.57	3.68	2.89	1.87	3.42
	PsyMem †	**7.95**	**8.00**	**9.05**	**9.05**	**6.80**	**6.65**	7.50	**7.86**
	PsyMem §	7.70	7.85	8.55	8.85	6.70	6.35	**7.70**	7.67
Vicuna-7B-v1.5	SimPsyDial †	5.37	5.43	6.61	6.72	4.14	5.92	4.45	5.52
	MentalChat16K •	5.75	6.52	5.54	7.62	6.41	5.29	4.88	6.00
	MentalChat16K †	2.17	2.01	3.12	4.15	1.62	1.44	2.65	2.45
	DiaSynth †	2.80	1.92	4.85	4.68	3.29	3.32	1.98	3.26
	PsyMem †	**7.64**	7.57	**8.64**	8.57	6.50	**6.57**	7.50	7.57
	PsyMem §	**7.64**	**7.71**	8.57	**8.71**	**6.64**	6.50	**7.93**	**7.67**
Qwen2-7B-Instruct	SimPsyDial †	6.34	5.84	6.92	7.48	6.44	5.70	5.46	6.31
	MentalChat16K •	5.85	6.49	5.98	7.74	6.55	5.89	5.25	6.25
	MentalChat16K †	6.41	6.94	7.78	7.88	5.63	5.83	6.40	6.70
	DiaSynth †	3.67	2.66	5.92	5.96	4.25	3.70	2.47	4.09
	PsyMem †	**7.74**	**8.18**	**7.90**	8.13	**6.72**	**6.31**	**6.95**	**7.42**
	PsyMem §	7.53	8.13	7.61	**8.45**	6.42	6.11	6.55	7.26
Qwen3.5-9B-Thinking	SimPsyDial †	8.87	8.93	8.22	9.08	8.03	7.06	7.92	8.30
	MentalChat16K •	6.32	7.04	5.88	7.43	5.68	5.73	4.98	6.15
	MentalChat16K †	8.69	8.90	7.78	8.84	7.69	7.61	6.22	7.96
	DiaSynth †	3.63	2.59	4.99	4.78	3.85	3.56	2.53	3.71
	PsyMem †	8.50	8.69	8.00	8.88	7.38	6.88	6.44	7.82
	PsyMem §	9.00	9.00	9.69	9.75	8.81	8.25	9.56	9.15
MeChat	SimPsyDial †	4.36	5.10	5.68	6.27	3.91	4.51	3.90	4.82
	MentalChat16K •	4.30	4.75	4.49	6.35	5.01	3.93	3.60	4.63
	MentalChat16K †	3.86	4.59	4.40	5.51	3.27	4.03	2.68	4.05
	DiaSynth †	2.39	1.53	4.98	5.32	3.19	3.02	1.59	3.15
	PsyMem †	**4.70**	**5.25**	5.45	6.25	3.65	4.25	3.50	4.72
	PsyMem §	4.10	5.00	4.50	5.80	3.10	3.60	3.00	4.16

F PROMPT TEMPLATES

We provide the prompts used for dialogue generation, memory extraction, and automatic evaluation. All prompts are shown verbatim except for placeholder variables. During inference, placeholders such as `role_card` are replaced with the corresponding profile, dialogue, memory, or session content.

F.1 DIALOGUE GENERATION PROMPTS

Client prompt.

Client Generation Prompt

You are now a client participating in psychological counseling.
Below is your personal information. All of your utterances must be grounded in this information, and you must not deviate from it or add any content beyond the profile:
`role_card`
Your responses must satisfy the following requirements:
1. Core consistency: Express yourself around the current primary psychological concern. Responses must be consistent with the background information, personality traits, resistance level, and current emotional state.
2. Specific but non-repetitive: You may generate similar but different specific scenarios or events based on the profile, while avoiding direct restatement of the original text.
3. Expression rules: - Use natural, colloquial language; - Keep each response concise and include at least one concrete detail; - Responses must be complete; mid-response truncation is prohibited; - Do not directly use professional personality terminology; - Adjust the depth of information disclosure according to the resistance level.
4. Dialogue pacing: - Break the core concern into multiple specific scenarios; - Gradually unfold them in the order of "event—emotion—behavior"; - Preserve points of continuation

and avoid ending prematurely; - If the counselor summarizes too early or offers advice, you may naturally express doubt.
5. Ending condition: When you believe the current issue has been clarified to some extent, or you need to pause the conversation, use a natural closing statement to clearly express your intention to end, for example: "Let's stop here for today" or "I'd like to digest these issues a bit more; let's talk again next time."

Counselor prompt.

Counselor Generation Prompt

You are now a virtual psychological counselor named "Xiaoxin". You are a responsible counselor who draws on person-centered, psychodynamic, and cognitive-behavioral approaches, and you specialize in layered probing and detail deepening. Your goal is to guide the client, through natural and continuous inquiry, to gradually express their experiences, emotions, behaviors, and current concerns, and to help them identify and challenge unhealthy thinking. Adjust the pace according to the client's level of resistance, and avoid giving generalized advice all at once.
Below is a memory summary of this client's previously disclosed facts and personality characteristics, carried over from earlier counseling sessions. Use it to understand the client's background and experiences, and keep your inquiry and interventions consistent with it:
`memory summary`
(If no memory summary is provided, this is the first session.)
Core rules:
1. Expression style: - Use natural, concise, colloquial language; - Each response should relate to the specific content of the client's previous turn; - Prioritize open-ended questions or gentle guidance; - Avoid ending the interaction with closed expressions such as "okay" or "I understand".
2. Information gathering: Gradually learn about the client's background (age, occupation or study, relationship and family situation), personality dispositions, past experiences, current concerns, and deeper needs. For each issue, cover the specific scenario, the emotional responses, the behavioral responses, the past events involved, and the concerns the client has not yet voiced. Do not assume information that has not yet been disclosed, and do not omit any of these dimensions.
3. Intervention repertoire: Keep four intervention types available and select among them according to the session phase and the client's resistance. - Empathy: acknowledge and validate what the client has disclosed; use it mainly to lower defensiveness in mid- and high-resistance clients rather than as a standalone response; - Reframing: help the client examine and challenge unhealthy thinking and view the situation from a different perspective; - Guidance: offer at most one personalized suggestion tied to the current scenario; - Actionable support: help the client turn insight into concrete next steps, with specific execution details.
4. Inquiry strategy: - Each turn should focus on one specific question and stay closely tied to the client's previous turn; - Gently probe ambiguous expressions; - Reduce questioning pressure for avoidant or perfunctory responses; - Avoid jumping between topics or giving premature advice.
5. Session phases and resistance adaptation: A counseling process spans at least 50 turns and is organized into three phases, paced by the client's resistance (low: cooperative; mid: mildly avoidant; high: frequently avoidant). - Early phase (at least 20–25 turns): build trust and collect information, following the order background $\rightarrow$ personality $\rightarrow$ past experiences $\rightarrow$ current concerns $\rightarrow$ deeper needs, refining one dimension per turn; - Middle phase (20–25 turns): promote self-awareness, following the order behavior $\rightarrow$ emotion $\rightarrow$ past events $\rightarrow$ underlying beliefs, focusing on one concrete scenario per turn; - Late phase (10–15 turns): review the client's changes and co-construct a personalized action plan whose steps include concrete execution details.

6. Advice strategy: When it is necessary to offer advice, give only one personalized suggestion relevant to the current scenario, and continue to ask about its feasibility, the client's feelings, or difficulties in carrying it out.
7. Ending condition: Only when the client explicitly expresses an intention to end, or repeatedly indicates that the current issue is basically clear, use a natural closing statement to end the session.

F.2 Memory Extraction Prompt

Role-specific state extraction prompt. In implementation, the client- and counselor-side extraction prompts are instantiated separately with role-specific framing and inputs; we present the unified template below.

Role-specific State Extraction Prompt

You will be shown the complete dialogue content of a psychological counseling session, as well as the memory summary from the previous session. Please generate the `perspective` perspective memory summary for the `visit`-th counseling session based on the current session content and with appropriate reference to the previous memory.
Input: - Current session content: `dialog_text` - Previous memory summary: `previous_memory_text`
Please generate structured memory according to the perspective:
If perspective = "client", please summarize: 1. facts: important facts, events, and details discussed in this session; 2. suggestions: advice or guidance given by the counselor; 3. `counselor_style`: the counselor's interaction style and response characteristics; 4. issues: the client's new understanding or changes regarding their own problems.
If perspective = "counselor", please summarize: 1. facts: important facts, events, and details disclosed by the client; 2. suggestions: advice given by the counselor in this session; 3. `client_personality`: the client's personality characteristics observed in this session (interaction style, dispositions, emotional tendencies); 4. issues: changes in the client's problems, actions, or state.
Please output valid JSON directly without adding any extra explanation.

F.3 Evaluation Prompts

WAIOS evaluation prompt.

WAIOS Evaluation Prompt

Please act as a psychological counseling evaluation expert and use the WAI-O scale to evaluate the following conversation:
`conversation_text`
Please rate the 15 items on a scale of 1–7: 1 = very strong negative evidence; 7 = very strong positive evidence.
Goal dimension: [2] The two parties agree on the steps to improve the client's condition. [6] The two parties share a common understanding of the treatment goals. [10] The two parties disagree on the goals of the session. [R] [24] The two parties agree on the problems that need to be focused on. [27] The two parties have different views on the client's real problems. [R]
Task dimension: [4] The two parties agree on the usefulness of the current counseling activities. [13] The two parties agree on the client's responsibilities in therapy. [16] The two parties believe that the current activities can help bring about change. [22] The two parties are working on mutually agreed-upon goals. [30] The two parties jointly set the goals of the session.
Bond dimension: [1] There is discomfort in the relationship. [R] [8] The two parties like each other. [19] The two parties respect each other. [26] The two parties trust each other. [36] The client feels that the counselor respects and cares about them.

Please output 15 integer ratings in order, one per line.

MCAF evaluation prompt.

MCAF Evaluation Prompt

Task: Evaluate the cross-session continuity and coherence of memories from four psychological counseling sessions.
The input consists of memory records from four counseling sessions. Please determine whether these records come from the same client and assess whether they constitute a continuous counseling process.
Please rate the following four dimensions, each on a 1–5 scale:
`client_side_continuity`: whether the client's problem state remains consistent and exhibits reasonable changes;
`counselor_side_continuity`: whether the counselor retains and uses prior disclosures, suggestions, and progress;
`within_session_alignment`: whether, within the same session, the client-side state and the counselor-side memory are consistent;
`overall_multi_session_consistency`: whether the four sessions as a whole form a coherent sequence rather than mutually independent fragments.
Scoring criteria: 1 = completely disconnected; 2 = clearly disconnected; 3 = basically coherent but with inconsistencies; 4 = well coherent with only minor issues; 5 = highly coherent.
Please output strictly in JSON:

```
{
  "client_side_continuity": <integer 1-5>,
  "counselor_side_continuity": <integer 1-5>,
  "within_session_alignment": <integer 1-5>,
  "overall_multi_session_consistency": <integer 1-5>,
  "overall_score": <number>,
  "rationale": "<brief explanation>"
}
```

Input: `formatted_memories`

Longitudinal persona consistency prompt.

Longitudinal Persona Consistency Prompt

Task: Determine whether the memories from four psychological counseling sessions belong to the same client.
Input data: `sessions_text`
Please judge solely based on the memory content, and do not introduce external assumptions.
The basis for judgment includes: - Background information; - Primary concerns; - Language and response style; - Emotional and behavioral patterns; - Whether facts across sessions are contradictory.
Please output valid JSON:

```
{
  "identity_consistency": {
    "consistent": <true or false>,
    "confidence": <integer 0-100>,
    "evidence": ["evidence_1", "evidence_2"],
    "contradictions": ["contradiction_1"],
    "summary": "<brief conclusion>"
  }
}
```

Note: confidence ranges from 0 to 100; a score of 80 or above is regarded as the same person; if no contradictions exist, please output an empty list.

MentalChat16K metrics prompt.

MentalChat16K Metrics Evaluation Prompt

Please act as an impartial mental health counseling evaluation expert and evaluate the following counselor's performance in the complete session:
`conversation_text`
Please rate the counselor on the following seven dimensions, each on a 1–10 scale: 1. `active_listening`: active listening; 2. `empathy_validation`: empathy and validation; 3. `safety_trustworthiness`: safety and trustworthiness; 4. `openmindedness_nonjudgment`: open-mindedness and non-judgment; 5. `clarity_encouragement`: clarity and encouragement; 6. `boundaries_ethics`: boundaries and ethics; 7. `holistic_approach`: holistic approach.
When scoring, please focus on whether the counselor consistently listens, responds accurately to the client, maintains safety and ethical boundaries, avoids vague or generic advice, and demonstrates an understanding of the client's overall situation.
Please output strictly in JSON:

```
{
  "overall_comment": "<3-5 sentence summary>",
  "scores": {
    "active_listening": <integer 1-10>,
    "empathy_validation": <integer 1-10>,
    "safety_trustworthiness": <integer 1-10>,
    "openmindedness_nonjudgment": <integer 1-10>,
    "clarity_encouragement": <integer 1-10>,
    "boundaries_ethics": <integer 1-10>,
    "holistic_approach": <integer 1-10>
  },
  "rationale": "<brief explanation>"
}
```

G Human Evaluation Protocol

G.1 Evaluators

Eleven evaluators participated in the study. All had a background in computer science or data synthesis and were familiar with dialogue data; none had clinical or counseling training, and none was involved in developing PsyMem. Participation was voluntary and uncompensated. The study was conducted as a minimal-risk assessment of synthesized text artifacts: no personal data were collected beyond the ratings, and no interaction with real clients or real counseling sessions took place.

G.2 Questionnaire

For each of the five evaluated datasets, approximately 25 dialogues were sampled, giving about 125 dialogues in total. Each dialogue was rated by exactly one evaluator; evaluators could complete questionnaires for several datasets, but no dialogue received overlapping ratings, so no inter-rater agreement statistic is reported. Each evaluator rated counseling-style dialogue material with the twelve-item questionnaire in Table 20: Part 1 on a 1-10 scale, Part 2 on a 1-7 scale with four rating considerations per item, and Part 3 on a 1-10 scale scored for each individual dialogue. Scoring and aggregation rules are given in Table 21.

G.3 Scoring and reporting

Three negatively worded WAI-O items (items 1, 10, and 27) are reverse scored as 8 - x before averaging. Part 3 items were originally recorded on a 5-point scale and are linearly rescaled to 1-10 for consistency with the MentalChat16K rating scale.

Table 20: The twelve-item human-evaluation questionnaire. Part 1 items map one-to-one onto the seven MentalChat16K dimensions; Part 2 covers the Goal, Task, and Bond dimensions of the working-alliance measure; Part 3 captures lay-perspective comfort and is scored for each individual dialogue.

Part	Item	What the evaluator rates	Scale
Part 1. Counseling quality			
Q1	Comprehension & follow-through	The material captures the client's core problem, and later turns show memory of and follow-up on earlier content rather than restarting or responding irrelevantly.	1–10
Q2	Validation	The counselor makes the client feel understood and accepted rather than dismissed or given perfunctory reassurance.	1–10
Q3	Safety	Responses are safe, contain no harmful, inappropriate, or self-contradictory advice, and give no risky guidance when the situation is uncertain.	1–10
Q4	Non-judgment & inclusivity	No bias or judgment; respect for the client's personal circumstances such as gender, cultural background, and values.	1–10
Q5	Clarity & empowerment	Language is clear and easy to follow; helps the client recognize strengths or progress; not preachy or over-directive.	1–10
Q6	Role awareness	Recognizes it is an AI rather than a real counselor; directs the client to real professional help in complex or crisis situations instead of over-reaching.	1–10
Q7	Holistic & longitudinal understanding	Does not fixate on a single event; understands the client from multiple angles (emotion, cognition, behavior, relationships) and sustains this across turns.	1–10
Part 2. Working alliance			
Q8	Goal agreement	Both parties share the priority problem, the intended outcome, and the path toward it; no goal conflict or talking past each other.	1–7
Q9	Task agreement	Both endorse what is being done: the exchange is seen as helpful and meaningful, roles are clear, and progress advances in a jointly agreed direction.	1–7
Q10	Bond	Mutual appreciation, respect, and a degree of trust; the client feels genuinely cared for rather than placated.	1–7
Part 3. Lay-perspective comfort (scored per dialogue)			
Q11	Comfort as client	Setting aside professional knowledge and social expectations, how comfortable would the AI-generated counselor make you feel in these sessions?	1–10
Q12	Comfort as counselor	How comfortable would the AI-generated counselor make you feel in these sessions?	1–10

Table 21: Scoring and aggregation of the questionnaire.

Part	Maps onto	Scale	Aggregation
Part 1	Seven MentalChat16K dimensions	1–10	Mean per dataset
Part 2	WAI-O Goal / Task / Bond	1–7	Mean of the three dimension scores
Part 3	Lay-perspective comfort	1–10 (recorded on 1–5, rescaled $\times 2$)	Mean over dialogues