PSYEVO: A PERSONALIZED COUNSELING AGENT THAT SELF-EVOLVES AT TEST TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Mental health disorders affect a substantial proportion of the global population, yet a persistent shortage of trained practitioners leaves the majority without adequate care. Large language model (LLM)-based counselors present a promising direction for delivering scalable conversational psychological support. Offline model training alone leaves limited room to adapt to individual clients or to learn from ongoing therapeutic interaction at test time. We introduce PsyEvo, an LLM-based counseling framework that enables both client-specific personalization and response-policy improvement at test time through three components: Hierarchical Bayesian Skill Policy (HBSP) personalizes what intervention to apply by maintaining a per-client skill posterior updated from session feedback; Inter-session Listwise Preference Optimization (LiPO) improves how the selected skill is expressed by updating a shared response adapter from cross-client preference evidence; and State-conditioned Ordinal Credit Assignment (SOCA) supplies candidate preferences and trajectory credit to the two components through consistency-checked comparisons and ordinal projection. In simulated-client evaluation with shared online cohort adaptation, PsyEvo obtains 7.684 Overall on PsychEval and exceeds every component variant in each of three matched runs. Removing individual components lowers mean overall score by 0.138–0.171 under the shared configuration, supporting conditional contributions within the complete scaffold.

Keywords: test-time learning; multi-session counseling; personalized counseling; ordinal credit assignment; preference optimization; self-evolving agents.

1 INTRODUCTION

Mental health disorders have become a major global health challenge, with nearly one billion individuals affected and a substantial share of the worldwide disease burden attributed to conditions such as depression and anxiety (World Health Organization; 2025). Psychological counseling is a widely adopted evidence-based intervention, yet training qualified counselors requires years of supervision, and the available workforce falls far short of demand (British Association for Counselling and Psychotherapy; World Health Organization, 2022). In this context, large language models (LLMs) provide a scalable conversational interface that could expand access to psychological support.

Research on LLM-based counseling spans supervised fine-tuning (SFT), preference optimization, and agent-based mechanisms. SoulChat and CPsyCoun provide counseling and empathetic dialogue data for SFT, whereas PsyCoPref supplies preference comparisons for response optimization (Chen et al., 2023; Zhang et al., 2024; 2026a). MusPsy and PsychAgent incorporate memory, planning, and skill-related mechanisms to support longitudinal interaction (Wang et al., 2026; Yang et al., 2026). These approaches extend counseling support from individual responses to multiple sessions. Beyond continuity, individual clients can differ in their experiences, needs, and communication styles (Wang et al., 2026), motivating updates to intervention selection and to the response policy during test-time interaction. This motivates two complementary learning objectives: personalization, which tailors intervention selection to a particular client, and counselor self-evolution, which updates the response policy using cross-client experience.

To this end, we present PsyEvo, a test-time learning framework for multi-session psychological counseling that addresses both objectives over a frozen backbone through three components: Hier-

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.

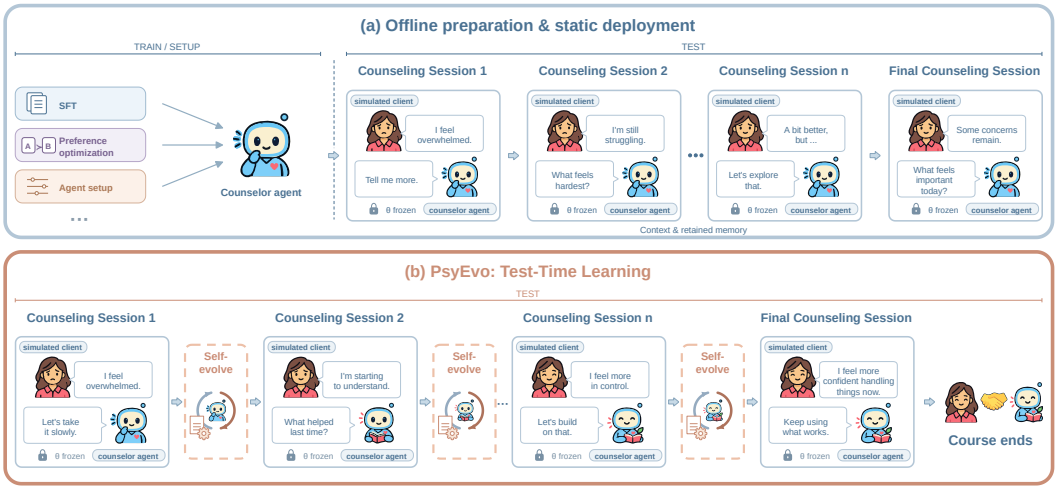

Figure 1: Static deployment versus test-time learning in psychological counseling. (a) A statically deployed counselor keeps its parameters fixed across sessions; (b) PsyEvo keeps the backbone frozen, updates the client-specific skill posterior after each session, and schedules shared-adapter updates at five-session intervals on PsychEval, subject to sufficient preference evidence.

archical Bayesian Skill Policy (HBSP) for personalized intervention selection via a per-client skill posterior, Listwise Preference Optimization (LiPO) for shared response-policy improvement via cross-client preference graphs, and State-conditioned Ordinal Credit Assignment (SOCA) for constructing candidate preferences and trajectory credit from relative comparisons and realized feedback. Our contributions are as follows: (1) we introduce PsyEvo, which combines client-private skill posteriors with a shared population response adapter over a frozen backbone, using HBSP for personalized skill selection and LiPO for cross-session response-policy learning; (2) we propose SOCA, which converts compressed self-evaluation feedback into candidate preferences through consistency-checked comparisons and into delayed credit through within-trajectory ordinal propagation; and (3) we establish positive conditional component contributions across three matched PsychEval runs and evaluate a related evidence-fusion rule on SummEval, where it achieves the highest mean agreement with human judgments among the compared evaluators.

2 PsyEvo

2.1 Problem Formulation

We formalize multi-session psychological counseling as a sequential decision process. Let $u \in \mathcal{U}$ denote a client, k the session index, and t the turn index within a session. At each turn, the counselor observes a state s_t comprising the realized dialogue history, the inferred client state, the therapeutic modality, the course stage, and a cross-session memory M_u. A skill controller μ_u selects a counseling skill a_t from the set $\mathcal{A}_t$ of skills valid for the current modality, and the response policy $\pi_{\theta,\phi^{(v)}}$ generates N candidate responses $y_t^{(1)}, \ldots, y_t^{(N)}$ conditioned on s_t and a_t; one candidate is delivered to the client.

PsyEvo decomposes the counselor's learning state into a client-private component and a population-shared component:

$$\Omega_u = (M_u, b_u), \qquad \Psi = (\phi^{(v)}, \mathcal{B}_{\text{pop}}), \qquad \pi_u(a, y \mid s) = \mu_u(a \mid s)\pi_{\theta,\phi^{(v)}}(y \mid s, a), \tag{1}$$

where the backbone parameters θ remain frozen throughout. The private state Ω_u consists of the client's memory M_u and a skill posterior b_u; the population state Ψ contains the versioned shared adapter $\phi^{(v)}$ and a population evidence buffer $\mathcal{B}_{\text{pop}}$.

Three components operate over this decomposition (Figure 2). HBSP updates b_u to personalize which skill to select for each client; LiPO updates $\phi^{(v)}$ to improve how the selected skill is expressed;

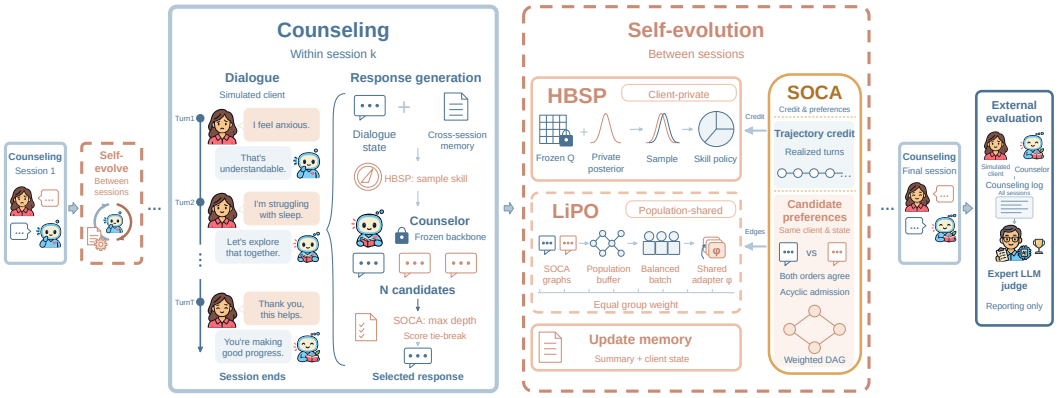

Figure 2: Overview of PsyEvo's test-time learning cycle. Within a session, HBSP selects a skill and SOCA selects the delivered response; between sessions, HBSP updates the client-private posterior using SOCA trajectory credit, while LiPO periodically updates the shared adapter from SOCA preference graphs accumulated across clients. External judges score completed transcripts for reporting only and do not feed the learning signals.

and SOCA constructs the learning signals, i.e. candidate preferences and trajectory credit, consumed by both.

2.2 HBSP

First, we propose HBSP to determine which counseling skill to apply at each turn. A shared skill-prior table $Q(s, a)$, initialized from skill-labeled PsychEval dialogue, provides a common starting point for all clients. During interaction, Q remains fixed; only the client-private posterior $b_u(s, a)$ is updated to capture individual preferences. Each client starts with $m_u(s, a) = 0$ and $\lambda_u(s, a) = 4$, corresponding to a residual prior variance of 0.25; we use observation variance $\sigma_b^2 = 1$.

Concretely, HBSP decomposes the skill-selection score into the shared term and a client-specific residual drawn from a normal posterior:

$$Q_u(s, a) = Q(s, a) + b_u(s, a), \qquad b_u(s, a) \sim \mathcal{N}(m_u(s, a), \lambda_u(s, a)^{-1}), \tag{2}$$

$$p_t(a) = (1 - \varepsilon_t)\mathrm{softmax}\Big(\tau[Q(s_t, a) + \tilde{b}_u(s_t, a)]\Big) + \varepsilon_t/|\mathcal{A}_t|. \tag{3}$$

Here, $\tilde{b}_u(s_t, a)$ denotes a draw from $\mathcal{N}(m_u(s_t, a), \lambda_u(s_t, a)^{-1})$ for the current skill decision. At the end of each session, SOCA assigns a trajectory return G_t to the skill selected at turn t (Section 2.4). With residual $d_t = G_t - Q(s_t, a_t)$, HBSP uses the Gaussian working likelihood $d_t \mid b_u(s_t, a_t) \sim \mathcal{N}(b_u(s_t, a_t), \sigma_b^2)$ to update the posterior:

$$\lambda'_u = \lambda_u + \sigma_b^{-2}, \qquad m'_u = \frac{\lambda_u m_u + \sigma_b^{-2} d_t}{\lambda'_u}. \tag{4}$$

Posterior sampling uses model-based uncertainty for exploration: repeated observations increase precision, while less-visited state–skill cells retain wider sampling distributions. Appendix C.1 derives how the shared starting score and observed returns combine under this update.

2.3 LiPO

While HBSP personalizes skill selection per client, we employ LiPO to improve how the selected skill is expressed by updating a shared response adapter from preference evidence accumulated across clients. For each candidate group g generated for client u, SOCA provides a set of supported preference edges $E_{u,g}$ with positive weights w_{ij} (Section 2.4). Each graph retains its client and modality identifiers; no edge is constructed between candidates from different clients.

We utilize a shared rank-8 LoRA adapter (Hu et al., 2022) as the learnable module. For a preference edge $i \succ j$, the policy–reference log-ratio utility $h_{u,g,i}$ and the per-group loss are

$$h_{u,g,i} = \beta \log \frac{\pi_{\theta,\phi^{(v)}}(y_i \mid s,a)}{\pi_{\theta,0}(y_i \mid s,a)}, \qquad \mathcal{L}_{u,g} = \frac{1}{|E_{u,g}|} \sum_{(i,j)\in E_{u,g}} w_{ij}\, \mathrm{softplus}[-(h_{u,g,i} - h_{u,g,j})]. \tag{5}$$

Our objective averages SOCA-weighted, DPO-style logistic losses within each candidate group. For a single edge with unit weight, Eq. 5 reduces to the standard pairwise DPO loss. LiPO retains the candidate-group structure, weights each supported edge by its bidirectional SOCA support strength, and normalizes by the number of retained edges in that group; pairwise DPO optimizes individual preferred–dispreferred pairs with unit weight. The population objective averages the per-group losses over the candidate groups in the training batch, so that a group does not receive greater weight merely because it contains more preference edges:

$$\mathcal{L}_{\mathrm{pop}} = \frac{1}{\sum_{u\in\mathcal{U}_B} |\mathcal{G}_u|} \sum_{u\in\mathcal{U}_B} \sum_{g\in\mathcal{G}_u} \mathcal{L}_{u,g}. \tag{6}$$

The reference $\pi_{\theta,0}$ is the frozen backbone with the adapter disabled. An accepted adapter version becomes available only to sessions that start after the update boundary, ensuring that response generation uses a fixed version throughout each session.

2.4 SOCA

Both HBSP and LiPO require informative learning signals, yet model self-evaluation can assign nearly identical scores to plausible counseling responses, a problem we call score compression. To address this issue, we introduce SOCA to construct evidence at two resolutions: candidate preferences for LiPO from pairwise comparisons within a state, and trajectory credit for HBSP from ordinal projection along the realized session.

Candidate preferences. For each unordered candidate pair $\{i, j\}$, let $p_{ij}^{\rightarrow}$ and $p_{ij}^{\leftarrow}$ denote weighted judgment support for candidate i under the two presentation orders. SOCA accepts a direction only when both orders agree; otherwise it abstains. The direction-consistent edge strength is

$$c_{ij} = \begin{cases} 2\min(p_{ij}^{\rightarrow}, p_{ij}^{\leftarrow}) - 1, & i \succ j, \\ 2\min(1 - p_{ij}^{\rightarrow}, 1 - p_{ij}^{\leftarrow}) - 1, & j \succ i. \end{cases} \tag{7}$$

Edges below a stability floor are discarded. The remaining edges are added in decreasing strength, skipping any that would create a directed cycle; their strengths serve as the LiPO weights w_{ij} in Eq. 5. We orient each edge from the preferred candidate to the dispreferred candidate and define depth as the longest directed path from a candidate to a sink, with sink depth zero. The delivered response maximizes this depth; ties are resolved by the higher pointwise score and then the smaller candidate index.

Trajectory credit. For the realized trajectory, let R_t be the turn outcome and ρ_t its tie-aware midrank among the T session turns. SOCA maps midranks to ordinal levels, scales them by $s = 0.5$, and applies a mean-preserving shift ξ:

$$\tilde{r}_t = \mathrm{clip}(\bar{R} + s\ell_t + \xi, 0, 1), \qquad \frac{1}{T}\sum_{t=1}^{T} \tilde{r}_t = \bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t. \tag{8}$$

For zero-based midranks, we use $K = \min(3, T)$ and $\ell_t = 2\min(K-1, \lfloor K\rho_t/T \rfloor)/(K-1) - 1$ when $T > 1$, and $\ell_1 = 0$ otherwise. Equal scores receive equal midranks and therefore equal projected credit. Discounted propagation then connects earlier skill decisions with later outcomes:

$$G_t = \tilde{r}_t + \gamma G_{t+1}, \qquad G_{T+1} = r_{\mathrm{session}}, \qquad \gamma = 0.9. \tag{9}$$

HBSP receives returns from the realized trajectory, while LiPO receives preferences from candidate comparisons.

Algorithm 1 summarizes the interaction and update cycle. Posterior and adapter updates take effect at session boundaries, keeping their values fixed within each session.

Algorithm 1 Test-time learning cycle of PsyEvo. All parameter updates take effect only at session boundaries.

Require: Frozen backbone θ, shared skill table Q, adapter $\phi^{(v)}$, population buffer $\mathcal{B}_{\text{pop}}$
Require: Client u with private memory M_u and skill posterior b_u
for each session (u, k) **do**
  Fix adapter version $\phi^{(v)}$ for the current session ▷ Within-session loop
  for each turn $t = 1, \ldots, T$ **do**
    Construct dialogue state s_t from history and M_u
    Sample skill $a_t \sim p_t(\cdot)$ via HBSP posterior (Eq. 3)
    Generate N candidates $\{y_t^{(1)}, \ldots, y_t^{(N)}\}$ from $\pi_{\theta,\phi^{(v)}}(\cdot \mid s_t, a_t)$
    Evaluate candidates via pointwise scoring and two-order pairwise comparisons
    Construct acyclic SOCA preference graph (Eq. 7); select response by graph depth
    Deliver selected response; record client reaction
    Store local graph with client, modality, and version metadata in $\mathcal{B}_{\text{pop}}$
  end for ▷ Between-session updates
  Compute ordinal trajectory credit $\{G_t\}_{t=1}^{T}$ via SOCA (Eqs. 8–9)
  Update client posterior b_u using realized residuals (Eq. 4)
  if population update criterion is met **then**
    Sample balanced batch from $\mathcal{B}_{\text{pop}}$ and normalize losses within each candidate group
    Update shared adapter: $\phi^{(v)} \to \phi^{(v+1)}$ by optimizing $\mathcal{L}_{\text{pop}}$ (Eq. 6)
    Rollback to $\phi^{(v)}$ on optimization failure
  end if
  Update cross-session memory M_u from the realized transcript
end for

3 Experiments

3.1 Datasets and Baselines

Our main counseling study uses a 100-client PsychEval cohort across five therapy modalities (Pan et al., 2026); SummEval (Fabbri et al., 2021) tests evidence aggregation beyond counseling against human judgments. Supplementary MusPsy figures for 100 profiles over six sessions are retained in Appendix K.1 (Wang et al., 2026). For counseling, we compare PsyEvo with eleven baselines, seven counseling-specific systems and four general-purpose models (Table 1), and with a Qwen3-32B variant trained with DPO (Table 3); for summary evaluation, we compare SOCA with six evidence-aggregation baselines that share its Qwen3-32B backbone. PsyEvo uses a Qwen3-32B counseling backbone with a rank-8 LoRA adapter optimized by LiPO. Internal candidate judgments also use Qwen3-32B; external PsychEval reporting uses Gemini-3-Flash on completed transcripts (Appendix G). PsychEval follows the PsychAgent interaction protocol, with BGE-M3 retrieval and DeepSeek-V3.1 client simulation; MusPsy follows its benchmark-specific client-state, six-session interaction, and evaluation protocols (Yang et al., 2026; Wang et al., 2026). In PsychEval, PsyEvo updates its shared adapter across clients in the online evaluation cohort, while each private skill posterior uses only that client's feedback. Appendices A, B, C, and I detail the datasets, baselines, learning settings, and training costs.

3.2 Evaluation Metrics

On PsychEval, we report four metric families derived from its 18 evaluation instruments (Pan et al., 2026). Counselor-shared (C-Shared) summarizes common dimensions of counseling quality, including working alliance and relational authenticity. Counselor-specific (C-Spec) uses CTRS, MITI, TES, PSC, and EFT-TFS for CBT, behavioral, humanistic, psychodynamic, and postmodernist therapy, respectively. Client-shared (L-Shared) summarizes common simulated-client scores, including affect and session alliance; client-specific (L-Spec) uses BDI-II, STAI, CCT, IPO, and SFBT indicators for the corresponding modalities. Metric definitions are detailed in Appendix D. Overall is the unweighted average of the four reported family scores.

For MusPsy, we report the Working Alliance Inventory–Short Form (WAI; Goal, Task, Bond, and their average) and positive and negative PANAS affect scores on a 1–5 scale (Wang et al., 2026), averaged over six sessions and 100 clients using the official MusPsy evaluation code and judging prompt. While for SummEval, we compute Spearman's ρ between automatic and human ratings for coherence (COH), consistency (CON), fluency (FLU), and relevance (REL) within each source-document group and macro-average across groups (Fabbri et al., 2021); Mean averages the four dimensions. Higher values are better for all metrics except negative PANAS. Appendix D lists the constituent instruments, aggregation rules, and judging-prompt protocols.

3.3 Main Results

Table 1: PsychEval results. P denotes a published reference score; E denotes our evaluation. PsyEvo averages seeds 7, 17, and 27 under shared online cohort adaptation. All metrics favor higher values. Complete-system rankings are descriptive; the matched component contrasts appear in Table 3.

Method	Source	Overall	C-Shared	C-Spec	L-Shared	L-Spec
CACTUS	E	4.869	5.112	2.530	4.594	7.240
CPsyCounX	P	4.680	4.210	2.510	4.730	7.270
SMILE	E	4.142	4.261	2.658	3.794	5.856
PsyDTLLM	P	6.055	6.100	5.430	5.270	7.420
PsyLLM-thinking-v3	P	5.883	5.300	4.670	5.630	7.930
TheraMind	P	6.625	6.250	6.940	5.480	7.830
PsychAgent-32B	P	7.348	7.320	7.910	5.920	8.240
GPT-5.6-terra	E	6.839	7.297	7.142	5.906	7.012
DeepSeek-V4-Flash	E	6.555	6.543	5.917	5.891	7.872
Gemini-3.8-Flash	E	7.363	7.310	8.150	6.224	7.770
Qwen3-32B	E	7.028	6.851	7.012	6.214	8.033
PsyEvo	E	**7.684**	**7.527**	**8.260**	**6.468**	**8.481**

Multi-Session Counseling Quality As shown in Table 1, PsyEvo achieves an Overall score of 7.684 on PsychEval. The descriptive gap is 0.336 over the strongest published counseling-specific reference, PsychAgent-32B (7.348), and 0.321 over the strongest general-purpose model evaluated by us, Gemini-3.8-Flash (7.363). Among the four metric families, L-Shared shows the largest improvement over PsychAgent-32B (+0.548), followed by C-Spec (+0.350); C-Shared and L-Spec improve by 0.207 and 0.241, respectively.

The supplementary MusPsy comparison is recorded separately in Appendix K.1, where its unresolved main-run configuration linkage is stated.

Summary Evaluation with SOCA We evaluate a SummEval variant of SOCA that combines pointwise, pairwise, and repair evidence to rank summaries against human quality judgments. This fusion variant shares pointwise and bidirectional comparison evidence with counseling SOCA, but does not test its graph-based response selector or ordinal trajectory credit, nor HBSP or LiPO; the component-removal analysis in Section 3.4 examines the contribution of each evidence channel. As shown in Table 2, SOCA achieves the highest Spearman correlation with human judgments on coherence, fluency, and relevance, and ties for the highest consistency correlation on the full SummEval dataset. Its mean correlation of 0.508 surpasses the next-best method PCA (0.489) by 0.019 and JDM (0.467) by 0.041, with the largest gains on fluency (+0.036 over PCA) and relevance (+0.015 over PCA).

3.4 Ablation Study

PsyEvo Component Ablations We ablate each component conditional on the remaining scaffold (Table 3); these contrasts do not decompose the full advantage over the base model. The SOCA comparison jointly tests candidate-preference construction and trajectory-credit projection. Table 9 specifies the changes and retained learning paths for each comparison. PsyEvo exceeds every component variant in Overall in each of the three matched runs. Disabling LiPO-based parameter updates reduces the mean Overall score by 0.171; removing SOCA reduces it by 0.152, replacing LiPO with pairwise DPO by 0.146, and removing HBSP by 0.138. The largest decreases for removing LiPO

Table 2: Correlation with human judgments on SummEval. Values are within-document Spearman correlations between automatic and human scores for coherence (COH), consistency (CON), fluency (FLU), and relevance (REL), computed from all 1,600 summaries grouped into 100 source documents and macro-averaged across documents, with all methods using a Qwen3-32B backbone.

Method	COH ↑	CON ↑	FLU ↑	REL ↑	Mean ↑
Raw Scalar	0.440	0.449	0.374	0.443	0.427
JDM	0.544	0.431	0.412	0.479	0.467
PCA	0.590	0.436	0.412	0.520	0.489
PairS-Greedy	0.558	0.390	0.366	0.419	0.433
PairS-Beam	0.553	0.388	0.356	0.417	0.429
BatchEval	0.497	0.422	0.358	0.500	0.444
SOCA	**0.600**	**0.449**	**0.448**	**0.535**	**0.508**

and SOCA occur in L-Spec (0.429 and 0.383, respectively), while removing HBSP most affects L-Spec (0.367). Appendix E.1 reports individual seeds, standard deviations, and paired differences whose conditional client-bootstrap 95% intervals are above zero for all four comparisons.

Table 3: Ablation of PsyEvo's three components on PsychEval. All variants share the experimental configuration, with only the specified component removed or replaced; with DPO replaces the weighted candidate-group objective with pairwise DPO, w/o LiPO disables LiPO-based parameter updates, and Overall averages the four metric families. Scores are means across seeds 7, 17, and 27.

Variant	C-Shared	C-Spec	L-Shared	L-Spec	Overall
PsyEvo w/o SOCA	7.510	8.010	6.510	8.098	7.532
PsyEvo with DPO	7.350	8.300	6.450	8.052	7.538
PsyEvo w/o LiPO	7.470	7.990	**6.540**	8.052	7.513
PsyEvo w/o HBSP	7.320	**8.400**	6.350	8.114	7.546
PsyEvo	**7.527**	8.260	6.468	**8.481**	**7.684**

SOCA Component Ablations We separately remove each of SOCA's three evidence channels—pointwise scoring, pairwise comparisons, and repair evidence—on the full SummEval dataset (see Table 4). Removing pairwise comparisons produces the largest drop in mean correlation (-0.041), with coherence most affected (-0.089). We also found that removing pointwise scoring or repair evidence leads to smaller decreases of 0.016 and 0.013, respectively.

Table 4: Ablation of SOCA's evidence channels on SummEval. Each variant removes one evidence channel; values are within-document Spearman correlations with human judgments, macro-averaged across the full SummEval dataset under a Qwen3-32B backbone.

Method	COH ↑	CON ↑	FLU ↑	REL ↑	Mean ↑
SOCA w/o Pointwise Scoring	0.581	0.434	0.432	0.522	0.492
SOCA w/o Pairwise Comparisons	0.511	0.439	0.438	0.483	0.467
SOCA w/o Repair Evidence	0.587	0.439	0.422	0.529	0.495
SOCA	**0.600**	**0.449**	**0.448**	**0.535**	**0.508**

3.5 Human Evaluation

In a descriptive pilot, three psychology experts blind-rated simulated dialogues from 25 cases (three sessions each) on five session-level dimensions and case-level client improvement, using a 1–5 scale (Appendix H.1). PsyEvo achieves the highest mean overall rating of 4.10, exceeding PsychAgent-32B (3.58) by 0.52 and CPsyCounX (2.95) by 1.15 (Table 5). These are descriptive means from a small pilot with low dimension-level agreement (Appendix H.4); client improvement denotes impressions of simulated dialogue, not clinical change or established expert consensus. Composite absolute-agreement reliability is ICC(A,3)=0.580 (95% CI [0.398, 0.690]).

Table 5: Descriptive pilot ratings of three counseling systems on 25 cases. Three psychology experts blind-rated three sessions per case on five session-level dimensions and case-level client improvement (1–5); Overall is the unweighted mean of the six dimension means.

System	Ethics	Empathy accuracy	Empathy quality	Interaction	Intervention	Client improvement	Overall
PsyEvo	**4.53**	**3.89**	**3.73**	**4.11**	**4.13**	**4.20**	**4.10**
PsychAgent-32B	4.49	3.49	3.22	3.47	3.67	3.13	3.58
CPsyCounX	4.49	2.76	2.64	2.49	2.84	2.47	2.95

3.6 Analysis

We describe client-specific skill rankings, session-wise outcomes, and a longitudinal case to examine learning-state changes alongside counseling behavior.

Client-Specific Skill Personalization We analyze HBSP skill rankings on PsychEval for 100 clients across five modalities to examine how skill preferences differ between clients at shared states. At the same state-space slot, different clients can receive different top-ranked skills (Figure 3): within-modality median pairwise disagreement ranges from 5.6% (CBT) to 14.3% (PDT), while most slots agree. These differences document client-specific ranking adaptation; the ablation (Table 3) tests its conditional contribution, but does not isolate client matching from residual learning (Appendix F). Separate archived adapter diagnostics are given in Appendix F; their unverified connection to the main runs limits them to illustrating that trace.

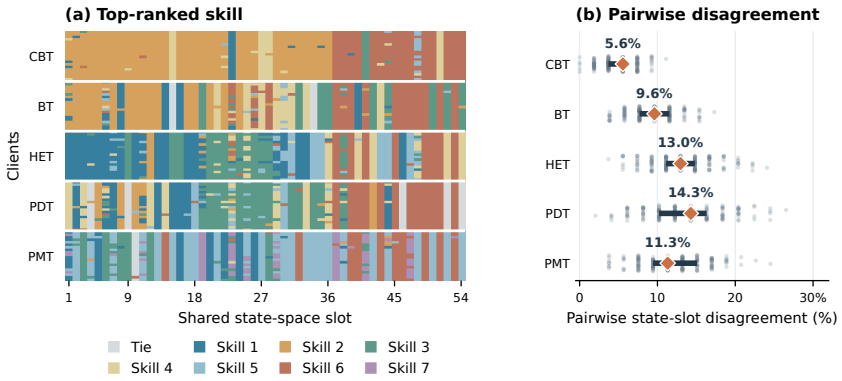

Figure 3: Client-specific HBSP skill rankings at shared states on PsychEval. (a) Top-ranked skill under HBSP for 20 clients per modality across 54 state-space slots; gray cells mark ties. (b) Pairwise disagreement fraction over the 190 client pairs per modality; diamonds mark medians and horizontal lines the interquartile range.

Session-wise Outcomes Figure 4 shows session-wise PsychEval scores underlying the main results. The four metric families describe outcomes across a changing set of scored clients; these trajectories do not isolate learning effects from other course dynamics.

Appendix H.5 traces within-state skill preferences, dialogue, and outcomes across the thirteen-session HET-860 case.

4 Related Work

LLM-based counseling has progressed from multi-turn empathetic dialogue within a session (Liu et al., 2021; Chen et al., 2023) through report-grounded reconstruction (Zhang et al., 2024) and counselor-style modeling (Xie et al., 2025) to multi-session agent systems with memory and plan-

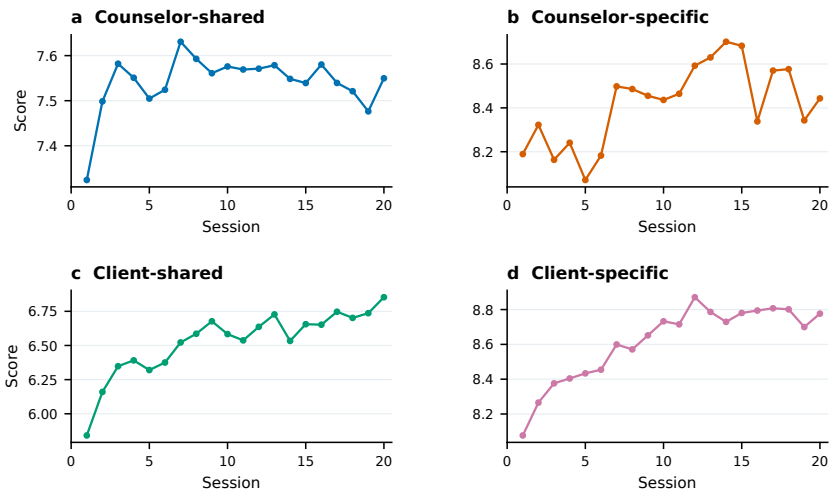

Figure 4: PsychEval score trajectories over 20 sessions. (a–d) Curves show session-wise mean counselor-shared, counselor-specific, client-shared, and client-specific scores, respectively. The number of scored clients varies across sessions, and panels use separate vertical ranges. Higher scores are better for all metrics.

ning (Wang et al., 2026; Pan et al., 2026; Yang et al., 2026). PsyEvo combines client-private skill posteriors with a shared response adapter updated at test time.

Personalization and test-time adaptation have been explored separately: LaMP and OPPU personalize generation via user histories or parameter-efficient modules (Salemi et al., 2024; Tan et al., 2024), while Self-Refine, Reflexion, and ExpeL enable iterative self-improvement through feedback loops (Madaan et al., 2023; Shinn et al., 2023; Zhao et al., 2024). In preference optimization, DPO learns from pairwise comparisons (Rafailov et al., 2023) and LiPO provides a general learning-to-rank framework (Liu et al., 2025); MICA derives mixed credit from user-state potentials (Zhang et al., 2026b). Our LiPO objective uses SOCA support-strength weights and group normalization; the original LiPO-Lambda variant uses gain and rank-discount weights. LLM-based evaluation can exhibit presentation bias and compressed scores (Zheng et al., 2023; Liu et al., 2023). Reward-model training can also produce prompt-independent reward distributions (Song et al., 2026), a distinct issue from the self-evaluation score compression addressed here. SOCA constructs ordinal evidence from consistency-checked comparisons without learning a reward model.

5 Conclusion and Limitations

We have presented PsyEvo, a test-time learning framework that combines HBSP for client-specific skill selection, LiPO for cross-client response-policy improvement, and SOCA for candidate preferences and ordinal trajectory credit. This separation allows individual intervention preferences to evolve alongside a shared response policy while keeping the backbone frozen. Matched PsychEval ablations show consistent conditional component contributions across three runs; the separate SummEval study examines evidence aggregation against human judgments.

Several limitations remain. The automatic evaluations use simulated clients and LLM judges, so their scores characterize model behavior rather than clinical efficacy. Shared online adaptation within the scored cohort does not establish transfer to unseen clients, and individual component removals do not isolate the joint benefit of both learning paths or each SOCA mechanism. Broader system comparisons mix evaluated and published results, while the small human-rating pilot has limited agreement and remains descriptive. Future work will examine transfer to disjoint client cohorts, isolate the learning mechanisms with matched controls, and broaden independent human evaluation. Extending this framework to clinician-supervised settings will also require assessment of crisis handling, privacy, and policy drift, alongside whether personalized skill selection and shared policy learning improve continuity across sessions.

AI DISCLOSURE

We used generative AI tools to polish writing, search and organize related literature, provide methodological feedback, interpret existing results, and derive and check the HBSP algebraic reformulation. AI-assisted revisions were checked against the manuscript's equations and reported results. The authors take responsibility for the final text, claims, and artifacts.

REFERENCES

British Association for Counselling and Psychotherapy. Training to become a therapist. URL https://www.bacp.co.uk/careers/counselling-and-therapy-as-a-career/training-to-become-a-therapist/. Accessed September 12, 2026.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1170–1183, 2023. doi: 10.18653/v1/2023.findings-emnlp.83.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Edward J. 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. In *International Conference on Learning Representations*, 2022.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3469–3483, 2021. doi: 10.18653/v1/2021.acl-long.269.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2404–2420, 2025. doi: 10.18653/v1/2025.naacl-long.121.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=9gdZI7c6yr.

Adian Liusie, Potsawee Manakul, and Mark Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *EACL*, pp. 139–151, 2024. URL https://aclanthology.org/2024.eacl-long.8/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, et al. Psycheval: A multi-session and multi-therapy benchmark for high-realism ai psychological counselor. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.1115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399.

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.

Ziang Song, Tianle Cai, Jason D. Lee, and Weijie J. Su. Reward collapse in aligning large language models. *Journal of Data Science*, 24(1):146–166, 2026. doi: 10.6339/25-JDS1201.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024. doi: 10.18653/v1/2024.emnlp-main.372.

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, 2026. doi: 10.18653/v1/2026.findings-acl.818.

World Health Organization. Depressive disorder (depression). WHO fact sheet. URL `https://www.who.int/news-room/fact-sheets/detail/depression`. Accessed September 12, 2026.

World Health Organization. World mental health report: Transforming mental health for all, 2022.

World Health Organization. Anxiety disorders, 2025. URL `https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders`.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. doi: 10.18653/v1/2025.acl-long.55.

Yutao Yang, Junsong Li, Qianjun Pan, Jie Zhou, Kai Chen, Qin Chen, Jingyuan Zhao, Ningning Zhou, Xin Li, and Liang He. Psychagent: An experience-driven lifelong learning agent for self-evolving psychological counselor. *arXiv preprint arXiv:2604.00931*, 2026.

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. BatchEval: Towards human-like text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15940–15958. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.846. URL `https://aclanthology.org/2024.acl-long.846/`.

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, 2024. doi: 10.18653/v1/2024.findings-acl.830.

Mian Zhang, Shaun M. Eack, and Zhiyu Chen. Preference learning unlocks LLMs' psychocounseling skills. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10729–10750, 2026a. doi: 10.18653/v1/2026.findings-acl.521. URL `https://aclanthology.org/2026.findings-acl.521/`.

Naifan Zhang, Ruihan Sun, Jinwei Su, Hengjie Yang, Zhengyuan Pan, Zhaohan Chen, and Xiaofan Zhang. MICA: Multi-granularity intertemporal credit assignment for long-horizon emotional support dialogue. *arXiv preprint arXiv:2603.06194*, 2026b. doi: 10.48550/arXiv.2603.06194.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023.

A Dataset Details

We evaluate on three datasets: PsychEval and MusPsy assess multi-session counseling quality, and SummEval measures agreement between a summary-evaluation variant of SOCA and human judgments. Table 6 summarizes the source corpora and the subsets used in our experiments.

Table 6: Dataset statistics and evaluation subsets. Source size describes the released corpus; the evaluation subset is the portion used in our experiments.

Dataset	Source size	Evaluation subset	Reported metrics
PsychEval	369 trajectories; 2,798 sessions; 67,314 turns	100 test clients, 20 per modality across five modalities	C-Shared, C-Spec, L-Shared, L-Spec, Overall
MusPsy	1,400 training and 100 test profiles	100 test profiles $\times$ 6 sessions (600 client–session observations)	WAI, PANAS
SummEval	100 articles; 1,600 summaries from 16 systems	All 100 source documents and 1,600 summaries	Spearman ρ on COH, CON, FLU, REL

PsychEval PsychEval contains 369 case-based counseling trajectories (2,798 sessions and 67,314 turns; 7.6 sessions per case and 24.1 turns per session on average) spanning cognitive behavioral, behavioral, humanistic–existential, psychodynamic, postmodernist, and integrative therapy, with a skill taxonomy of 677 meta-skills and 4,577 atomic skills. We evaluate on a 100-client cohort across the five individual modalities, with 20 clients per modality. The internal feedback used for learning is kept separate from external outcome evaluation (Appendix G). Because clients can end their courses at different sessions, session-wise analyses average the scored clients at each session.

MusPsy MusPsy is a CBT-oriented synthetic corpus grounded in published case reports, with 1,400 training and 100 test client profiles averaging 6.17 sessions per client and 28.55 turns per session; each profile includes evolving client states, session goals, and counseling memories. We evaluate all 100 test profiles over six consecutive sessions, yielding 600 client–session observations per method. The author-released reference systems use Llama3-8B on 100 profiles; our evaluation follows the official MusPsy evaluation code and judging prompt (Appendix D).

SummEval SummEval contains 1,600 summaries from 16 systems for 100 CNN/DailyMail articles, each rated by three experts and five crowd workers on four quality dimensions. We evaluate all 1,600 summaries across the 100 source documents, retaining each document's 16 summaries as a group for correlation computation.

B Baseline Specifications

B.1 Counseling Baselines

Table 1 compares PsyEvo with eleven baselines on PsychEval, including seven counseling-specific systems (CACTUS, CPsyCounX, SMILE, PsyDTLLM, PsyLLM-thinking-v3, TheraMind, PsychAgent-32B) and four general-purpose models (GPT-5.6-terra, DeepSeek-V4-Flash, Gemini-3.8-Flash, Qwen3-32B). The MusPsy figures appear separately in Table 19. PsychEval scores of

CPsyCounX, PsyDTLLM, PsyLLM-thinking-v3, TheraMind, and PsychAgent-32B are cited from prior publications; all other PsychEval results are from our evaluation, and the sources of MusPsy scores are described in Appendix D. The comparisons include published reference scores alongside our evaluations. Our evaluations follow the source benchmark protocols; imported scores retain their published evaluation conditions and have not been uniformly re-scored by us. For PsychEval, we follow the public PsychAgent setup with at most 20 sessions per client and 45 counselor turns per session. Client and counselor generation use temperatures of 0.7 and 0.9 and output limits of 512 and 8,196 tokens, respectively. External scoring uses Gemini-3-Flash (`gemini-3-flash-preview`) and the released modality-specific evaluation methods. The protocol covers BT, CBT, HET, PDT, and PMT, with 20 clients per modality. The released configuration and evaluation interfaces are available at `https://github.com/ECNU-ICALK/PsychAgent`. For MusPsy, the benchmark protocol uses English client profiles with dynamic states carried between six sessions; client and counselor generation both use temperature 0.7, and the original GPT-4o evaluator uses temperature 0 (Wang et al., 2026). The released profiles and simulation and evaluation prompts define the MusPsy setup independently of the PsychEval configuration. For the author-released MusPsy references, CACTUS and SMILE denote Meta-Llama-3-8B-Instruct models fine-tuned on those datasets, rather than their original released checkpoints: the source implementation uses LoRA rank 32, $\alpha = 16$, learning rate 2×10^{-4}, and two epochs (Wang et al., 2026). These source training settings describe the imported references, not PsyEvo's online updates.

Comparison provenance and interpretation. For PsychEval, the five cited systems listed above are published references; CACTUS, SMILE, and the four general-purpose systems are evaluated by us, alongside PsyEvo. The component comparisons in Table 3 are matched interventions within the PsyEvo scaffold; Table 1 instead compares complete systems, including differences in inference scaffolding and adaptation access. Following a common benchmark does not by itself guarantee identical model versions, judge calibration, or uncertainty for imported and newly evaluated results. The highest-mean comparison is descriptive, not a paired significance claim; the within-scaffold interventions provide the evidence for individual component contributions.

B.2 Summary-Evaluation Baselines

All compared evaluators use Qwen3-32B as the backbone. Raw Scalar selects the maximum-probability score category, while Judgment Distribution Mean (JDM) takes the expectation over score categories. Pairwise Comparative Assessment (PCA) aggregates judgments of relative quality between two summaries (Liusie et al., 2024). PairS-Greedy constructs a ranking through pairwise merge-sort decisions; PairS-Beam retains multiple candidate orderings during merging to account for uncertain preferences (Liu et al., 2024). BatchEval evaluates multiple candidates jointly through iterative batch-wise comparison (Yuan et al., 2024). The comparison includes all six baselines on the full SummEval dataset. These are evaluator baselines applied to fixed SummEval outputs, rather than the 16 summarization systems that generated the dataset. SOCA's aggregation procedure is detailed in Appendix K.2.

C Implementation and Reproducible Update Schedule

The PsychEval implementation separates dialogue serving from adapter training. It uses Qwen3-32B as the counselor, BGE-M3 for retrieval embeddings, and DeepSeek-V3.1 as the simulated client. Four graphics processing units (GPUs) serve dialogue, and four train adapters. The implementation maintains separate memories and skill posteriors for individual clients and accumulates cross-client preference evidence for LiPO training. Within each run, clients across the five modalities share one response adapter and evidence buffer; their memories and skill posteriors remain client-specific. Consequently, a session can use an adapter updated from earlier sessions of other clients in the same evaluation cohort. The reported scores assess this shared online-learning setting and do not by themselves establish transfer to a disjoint client cohort. Each seed starts a separate run with resume disabled, and up to ten clients execute concurrently. Table 7 specifies the PsychEval full-system and component-comparison configuration.

Online evaluation target. The scored unit is a realized counseling session produced during adaptation, with an adapter version fixed when that session starts. Accepted updates can affect subsequent sessions but cannot retroactively alter an already scored transcript. This protocol evaluates learning within an interacting client cohort; evaluating a frozen adapter on a disjoint cohort would test a separate transfer claim. Within a run, client outcomes can be dependent through the shared adapter even though memories and skill posteriors are private.

Table 7: Online learning configuration for the PsychEval full system and component comparisons. Component variants retain these settings except for the specified intervention.

Setting	Value
Maximum sessions / counselor turns	20 / 45
Soft session target / grace	6 / 2
Random seeds (PsychEval evaluation)	7, 17, 27
Client / shadow temperature	0.7 / 1.3
Shadow responses / total candidates	4 / 5
Update eligibility interval	5 completed sessions across clients
LoRA rank / batch size / gradient accumulation	8 / 2 / 4
Preference pairs / minimum candidate groups	320 / 32
Minimum batch share per modality	20% for each of the five modalities
Stable preference-edge strength threshold	0.1
Maximum training sequence length	1,024 tokens
Learning rate / preference scale	10^{-5} / $\beta = 0.05$
Maximum optimizer steps / epochs per round	3 / 1
Shared skill table / backbone	Fixed / frozen
Preference reference	Frozen backbone, adapter disabled
Adapter activation	Subsequent sessions after an accepted update
Ordinal tiers / spread / discount	$K = 3$, $s = 0.5$, $\gamma = 0.9$
HBSP prior mean / variance	0 / 0.25
HBSP observation variance	$\sigma_b^2 = 1$
Skill-state slots	3 stages $\times$ 3 progress bins $\times$ 6 affect categories

For counseling, the skill controller maps the dialogue to 54 slots: three counseling stages, three within-session progress bins, and six keyword-based affect categories. The stages are problem formulation and goal setting, core intervention, and consolidation; progress bins use thresholds 0.33 and 0.66 of the configured turn budget. The affect categories are anxiety, sadness, anger, guilt, calmness, and positive affect. This discrete index addresses the skill table; response generation retains the dialogue context. HBSP treats residual returns within each client-specific state–skill cell as conditionally independent observations under its Gaussian working model. The resulting precision controls posterior sampling within this model; it is an online approximation for a policy that evolves across sessions. The hierarchy consists of a fixed shared table and client-specific residuals; HBSP does not learn a population hyperprior online.

The realized-turn outcome $R_t \in [0, 1]$ is a retrospective assessment of the client's observed emotional change, willingness to disclose, collaboration, and concrete progress after the delivered response. The terminal reward is a weighted combination of direction-aligned, normalized PANAS, modality-specific scale, SRS, and alliance scores. SOCA applies the same mean-preserving ordinal projection to the client's session-reward history and uses the latest projected value as r_{session}; the first session retains its raw value. Equation 9 accumulates these credits with discounting.

Candidate generation and preference construction retain the identity of the state from which each candidate group originates. The buffer stores the prompt, candidate identifiers, local judgments, retained preference edges, and the policy version that generated the group. Fixing this evidence before optimization allows the loss to be reconstructed from the saved candidate group. An external evaluation score is not substituted for a missing candidate judgment.

On PsychEval, the five-session eligibility interval counts completed sessions across clients sharing the adapter. An optimization round uses 320 preference pairs from at least 32 candidate groups, with each of the five modalities contributing at least 20% of the batch. A batch that fails a requirement is returned to the buffer for further evidence accumulation. An update proceeds through evidence collection, optimization, validation, and activation. The active adapter remains fixed throughout each

session. At a session boundary, the implementation freezes the evidence, evaluates reference probabilities under the frozen backbone with the adapter disabled, optimizes the preference objective, and records the update outcome. Only an accepted adapter becomes eligible for subsequent generation; a failed update leaves the previous version active. This sequence associates each transcript with a specific policy version.

Table 8: Persistent client and counselor learning states. Each row lists a stored object, its contents, and its role in the test-time learning cycle.

Object	Contents	Role
Memory	Recap, goals, unresolved themes	Semantic continuity across sessions
Skill posterior	State–skill means and precisions	Personalized selection and exploration
Evidence buffer	Candidates, edges, weights, versions	Preference optimization input
Adapter registry	Learning scope, parent version, checkpoint, activation boundary	Loading, replay, and rollback

Memory and parameter learning provide different forms of continuity. A recap can retain a workplace trigger or a previously agreed coping exercise without changing the model weights. Updates to the skill posterior change the relative values of interventions, while updates to the adapter change the conditional response distribution. The separate states support memory-only, controller-only, and adapter-update evaluations.

The full system and all component variants in Table 3 use the same experimental configuration described in Table 7, except for the specified component intervention. The SOCA variant replaces consistency-checked candidate preferences with raw-score preferences and uses unprojected outcome scores for trajectory credit, while retaining grouped preference optimization. The DPO variant retains SOCA preferences but replaces weighted group optimization with unit-weight pairwise optimization, testing the combined effect of SOCA weighting and group normalization. The HBSP variant retains the shared skill table while removing client-private residuals and their updates. The parameter-update variant keeps the response adapter fixed while retaining skill-posterior updates, memory, and candidate evaluation. Other experimental settings remain unchanged across variants. The fixed-adapter variant therefore still permits client-posterior adaptation, and the no-HBSP variant still permits adapter learning. Neither is a control that freezes both learning paths while retaining the complete scaffold. Likewise, the SOCA intervention jointly changes candidate preferences and realized-trajectory credit; it does not identify the separate effect of ordinal projection.

Table 9: Scope of the matched PsychEval interventions. Memory, candidate generation, backbone, case set, and evaluation protocol are retained throughout. Each row changes the indicated component within the remaining scaffold.

Variant	Changed operation	Retained learning or selection operations
w/o HBSP	Remove client-private residuals and their updates	Fixed shared skill table; SOCA selection and shared-adapter learning
w/o LiPO	Keep the response adapter fixed	Client-posterior updates; SOCA candidate evaluation and selection
w/o SOCA	Use raw-score preferences and unprojected outcome credit	Client-posterior updates; grouped preference optimization
with DPO	Replace weighted group loss with unit-weight pair loss	SOCA preferences; client-posterior and shared-adapter updates

C.1 Supplementary Implementation Details

The supplementary method package exposes the numerical HBSP, SOCA, and population LiPO operations separately from model inference and benchmark evaluation. It is a reference implementation of these operations, rather than an export of the experimental training system. The following details specify its decision rules, weighting, and update lifecycle; model-specific generation, feedback prompts, and physical adapter training are supplied by a backend.

HBSP exploration and initialization. Initialization uses skill labels in the static PsychEval corpus. The base construction maps labeled decisions to the counseling-stage, session-progress, and affect slots described above, counts skills within each state, and normalizes their frequencies. For a state with observed labels, writing $n(s, a)$ for the skill count gives the frequency prior

$$Q_{\text{freq}}(s, a) = \frac{n(s, a)}{\sum_{a'} n(s, a')}. \tag{10}$$

The tooling also includes a mixed-prior variant; the loaded initialization file determines which fixed table is used. In particular, the frequency prior is a skill-selection preference score, not an empirical mean discounted return. During evaluation, realized returns update the client-private residual against the loaded, fixed table, rather than re-estimating its corpus frequencies. The experiments use $\tau = 1.0$ and a fixed $\varepsilon_t = 0.05$ throughout interaction, matching the supplied `HBSPConfig`. The implementation's `temperature` multiplies the skill values before softmax, as in Eq. 3; it is not a divisor. Posterior sampling is applied before the exploration mixture, which is uniform only over the valid skills; excluded skills receive zero probability. The constructor copies the supplied shared table and keeps it read-only, while each new client receives separate zero-mean residuals and prior precisions. An update changes only the selected client–state–skill cell and subtracts the corresponding shared value from the cumulative return without an additional rescaling step. The package accepts the shared table as input; its synthetic example is not the annotated-data initialization used in the experiments. Decision records expose the action probabilities and posterior mean, standard deviation, and draw, so posterior updates and subsequent decision-time use can be inspected separately.

How the skill prior combines with returns. For a fixed client–state–skill cell, let $G_1, \ldots, G_n$ denote the returns assigned to its successive updates, with $m_0 = 0$ and fixed Q and σ_b^2. Accumulating Eq. 4 gives $\lambda_n m_n = \sigma_b^{-2} \sum_{i=1}^{n} (G_i - Q)$, and hence

$$Q + m_n = \frac{\lambda_0 Q + \sigma_b^{-2} \sum_{i=1}^{n} G_i}{\lambda_0 + n\sigma_b^{-2}}. \tag{11}$$

Thus the mean selection score interpolates between the loaded prior and the observed mean return; Q need not itself be an empirical return estimate. With the stated variances, its prior weight is $4/(4+n)$. The cumulative G_i and residual need not lie in the unit interval used for individual rewards. This identity explains the update's scale behavior, not its statistical calibration: the working variance and dependence between returns still affect exploration.

SOCA comparison and credit interfaces. Qwen3-32B serves as the internal evaluator: SOCA converts its judgments into candidate preferences, which LiPO uses to update the shared response adapter. For each candidate pair, the evaluator supplies twelve criterion-level A/B judgments, which are combined into a weighted support share. Let $z^o_{\ell,ij}$ indicate whether criterion ℓ favors the original candidate i in order o, and let ω_ℓ denote its criterion weight. For a complete binary judgment vector, the support share is

$$p^o_{ij} = \frac{\sum_{\ell=1}^{12} \omega_\ell z^o_{\ell,ij}}{\sum_{\ell=1}^{12} \omega_\ell}, \qquad o \in \{\rightarrow, \leftarrow\}. \tag{12}$$

Each presentation order is evaluated separately, with A/B choices mapped back to the original candidate identities before applying Eq. 7. The resulting number represents weighted criterion support, not a next-token likelihood or a calibrated probability of therapeutic benefit. The edge weight in Eq. 7 measures the smaller of the two directional support margins; the twelve criteria are inputs to one evaluator, not independent evaluators. The graph constructor requires one two-order judgment for every unordered candidate pair. Both support values refer to the same original candidate after reversing presentation order. Agreement must be strict on the same side of 0.5; an exact tie or order disagreement produces no edge. To apply the configured stability floor of 0.1, the backend must pass it explicitly, since the general graph routine defaults to a zero floor. Equal-strength edges are ordered by a deterministic hash of the group identifier and unordered candidate indices before acyclic admission. Pointwise scores cannot create or reverse edges; they only break equal-depth selection ties. An empty graph therefore falls back to pointwise selection but is not admitted to preference training. For realized outcomes, exact ties share a midrank, and a common shift followed by clipping preserves the mean; the reference solver uses bisection. Discounted returns are accumulated backward from the supplied terminal reward. These interfaces consume numerical feedback and do not determine which model or prompt generated it.

Feedback flow and optimization. Pointwise candidate scores and two-order comparison supports describe alternative responses at the current state, before a response is delivered. Their role differs from realized-turn feedback, which describes the simulated client's reaction after delivery, and from terminal feedback available after the session. The former supplies local preference edges for LiPO; the latter supplies trajectory returns for HBSP. The preference dataset contains candidates, edge directions, and numerical weights, not gradients between groups. LiPO differentiates policy log probabilities while holding these saved judgments fixed; SOCA does not backpropagate through evaluator calls. The frozen, adapter-disabled preference reference specifies the denominator in Eq. 5; it is a distinct role from the internal evaluator that supplies comparison support. External transcript scores are a separate reporting channel (Appendix G).

Population weighting and batch units. Each buffered item is a complete local graph identified by client and group, with modality and generating policy version retained. Pooling these items does not introduce comparisons between different clients. In `lipo.py`, each graph loss is divided by its retained edge count, not by the sum of support-strength weights; graph losses are then averaged equally. Clients consequently contribute in proportion to their admitted group counts, rather than receiving equal client weights. The listwise grouping preserves local candidate sets, while the objective itself remains a weighted pairwise logistic loss; it does not define a likelihood over complete candidate permutations. The optional autograd implementation detaches reference log probabilities; the backend must compute them with the adapter disabled and matching response-token masks. The reference batch sampler counts preference pairs as retained graph edges. Under the PsychEval configuration, exactly 320 edges with at least 20% from each of five modalities requires 64 edges per modality, together with at least 32 groups. It selects whole graphs without truncation and leaves the buffer unchanged if no admissible batch exists. Its deterministic selection does not reconstruct the experimental sampling order or identify the accounting unit of the separate archived training trace in Appendix I.

Session boundaries and failed updates. One scheduler is shared by all clients in a learning scope. Starting a session captures the active adapter identifier; later acceptance of a new version does not change that session's identifier. Under the PsychEval schedule, five completed sessions make an update eligible, but training still waits for an admissible batch. Training must produce a separate candidate checkpoint, which is checked by a backend validation callback before the active reference advances. An exception or rejection returns the reserved graphs and retains the previous active reference; returning evidence is not a rollback of arbitrary in-place model changes. The backend must preserve the active checkpoint and implement physical loading, cleanup, and recovery.

Benchmark configurations. Benchmark-specific configurations follow the corresponding source protocols for PsychEval and MusPsy, while SummEval uses the original summary data and human annotations (Pan et al., 2026; Yang et al., 2026; Wang et al., 2026; Fabbri et al., 2021).

MusPsy supplementary-run configuration. The identified MusPsy supplementary run uses CBT only, with `required_modalities=['cbt']` and the five-modality quota disabled. Update eligibility is checked after every two completed sessions; the configured preference-pair range is 16–160. The LoRA rank is 8, the learning rate is 10^{-5}, and optimization is limited to three steps per round. These supplementary-run settings are separate from the PsychEval schedule in Table 7; their correspondence to the reported MusPsy run in Table 19 remains to be verified.

D Evaluation Metric Details

D.1 PsychEval

C-Shared covers working alliance (WAI), helping skills (HTAIS), relational authenticity (RRO), and customized clinical-quality criteria. C-Spec uses CTRS, MITI, TES, PSC, and EFT-TFS for CBT, behavioral, humanistic, psychodynamic, and postmodernist therapy, respectively. L-Shared covers symptoms (SCL-90), affect (PANAS), authenticity (RRO), and session alliance (SRS); L-Spec uses BDI-II, STAI, CCT, IPO, and SFBT indicators for the corresponding modalities. Overall is the unweighted average of the four family scores.

D.2 MusPsy

The 12-item WAI–Short Form uses four items each for Goal (agreement on objectives), Task (agreement on counseling activities), and Bond (trust and rapport), rated on a 1–5 scale; we report the three subscale means and their unweighted average (WAI Avg). PANAS comprises ten positive and ten negative affect items, each rated from 1 to 5; we report the mean within each group. Higher WAI and positive PANAS and lower negative PANAS indicate better evaluator-assessed outcomes. For MusPsy, author-released reference scores are taken from the original paper (Wang et al., 2026); our evaluation averages six sessions for all 100 test clients using the official MusPsy evaluation code and judging prompt. For the cited CACTUS and SMILE results, we retain the dataset authors' separately reported WAI average, averaged across their six published session rows.

D.3 SummEval

Human ratings use a 1–5 scale for coherence (COH; organization across sentences), consistency (CON; factual support from the source), fluency (FLU; sentence-level linguistic quality), and relevance (REL; coverage of salient information without redundancy). For each dimension, Spearman's ρ between automatic and human scores is computed within each source-document group and macro-averaged across groups; Mean averages the four dimension-level correlations.

E PsychEval Evaluation Protocol

The counselor-side metric labels in the original evaluation diagram denote working alliance (WAI), professional skill execution (HTAIS), response relevance and progression (RRO), interaction quality (Custom), and dialogue planning and continuity (DP). These labels denote evaluation dimensions, not additional learning modules.

PsychEval evaluates counseling across repeated interactions rather than a single isolated response. Its five therapeutic modalities are cognitive behavioral therapy (CBT), behavioral therapy (BT), humanistic–existential therapy (HET), psychodynamic therapy (PDT), and postmodernist therapy (PMT). Counselor-shared and counselor-specific scores assess the intervention, while client-shared and client-specific scores describe the simulated client's responses. The mean reported in the main tables assigns equal weight to the four metric-family aggregates.

Aggregation accounts for turns nested within sessions and sessions nested within clients. For client u with K_u scored sessions, let z_{uk} be a session-level metric. The reported aggregate gives equal weight to clients:

$$\bar{z}_{\text{client}} = \frac{1}{U} \sum_{u=1}^{U} \left(\frac{1}{K_u} \sum_{k=1}^{K_u} z_{uk} \right). \tag{13}$$

Each client contributes one within-client mean, so longer courses do not receive greater weight.

Session-wise means are computed from the scored clients at each session; client composition can change over the course.

E.1 Cross-Run Variability

We evaluate PsyEvo and four ablation variants on the same PsychEval case set using seeds 7, 17, and 27. For each seed, we average session scores within each client and then equally across clients; Overall gives equal weight to the four metric families. The reported mean gives equal weight to the three seeds, and SD is their sample standard deviation (Table 10). PsyEvo achieves the highest Overall score in each of the three runs, with a mean of 7.684 and a standard deviation of 0.028. Across seeds 7, 17, and 27, the paired run-level gains are 0.171, 0.146, and 0.121 over DPO; 0.218, 0.169, and 0.126 over no parameter updates; 0.158, 0.153, and 0.145 over no SOCA; and 0.174, 0.136, and 0.104 over no HBSP.

For each comparison, we subtract the variant's client-level Overall score from PsyEvo's score for the same client and seed, then average these differences equally across clients and seeds. We estimate

Table 10: PsychEval Overall scores across three random seeds. Mean and SD denote the across-seed mean and sample standard deviation.

Variant	Seed 7	Seed 17	Seed 27	Mean	SD
PsyEvo w/o SOCA	7.498	7.531	7.567	7.532	0.035
PsyEvo with DPO	7.485	7.538	7.591	7.538	0.053
PsyEvo w/o LiPO	7.438	7.515	7.586	7.513	0.074
PsyEvo w/o HBSP	7.482	7.548	7.608	7.546	0.063
PsyEvo	7.656	7.684	7.712	7.684	0.028

95% confidence intervals using 10,000 paired client-level bootstrap resamples, retaining all three seeds and both methods for each sampled client. These intervals summarize client-level variation conditional on the three realized online learning runs; Table 10 separately reports variation across runs. Resampling client scores does not rerun the coupled adaptation process, so these intervals do not capture uncertainty over new learning schedules or new cohorts. The consistently positive paired seed differences are useful descriptive replication evidence, but the small number of runs limits broader robustness conclusions. Table 11 reports the mean differences and intervals, which lie above zero for all four comparisons.

Table 11: Paired differences in PsychEval Overall score. Positive differences favor PsyEvo; CI denotes the 95% confidence interval.

Comparison	Mean difference	95% CI
PsyEvo − with DPO	+0.146	[0.048, 0.244]
PsyEvo − w/o parameter updates	+0.171	[0.085, 0.257]
PsyEvo − w/o SOCA	+0.152	[0.060, 0.244]
PsyEvo − w/o HBSP	+0.138	[0.052, 0.224]

F Mechanism Analysis of the Three Components

F.1 Client-specific Skill Selection

Figure 3 compares HBSP's top-ranked skills for 20 clients per PsychEval modality at 54 shared state-space slots. Skill indices are defined within each modality, and comparisons are made between clients of the same modality at the same state. Maxima tied within 10^{-12} are shown in gray, and a slot tied for either client is excluded from that pair's disagreement denominator. Median disagreement is 5.6%, 9.6%, 13.0%, 14.3%, and 11.3% for CBT, BT, HET, PDT, and PMT, respectively. The interquartile ranges summarize the distribution over the 190 client pairs within each modality. All 54 state-space slots are retained, and clients contribute 10–20 recorded sessions. The analysis captures how HBSP assigns different skill rankings to individual clients under a shared state representation.

F.2 Preference Learning and Adapter Changes

The archived training trace records twelve accepted adapter updates over 2,080 preference pairs. These diagnostics document parameter updates in that trace; its correspondence to the main evaluation runs and the accounting unit for pairs remain unspecified. All twelve updates have finite gradients, and the recorded adapter norm changes from 418.83 to 421.14. These records establish that optimization executed and changed the adapter parameters. Counseling quality is assessed using subsequent session outcomes rather than inferred from a reduction in training loss.

Pair construction and pair optimization are separate operations. Once a valid preferred–dispreferred pair is available, DPO can learn from the preference direction without requiring a large numerical gap between judge scores. Score compression instead affects rules that select pairs by a fixed score threshold or multiply their losses by the score gap. In SOCA, the weighted objective uses only the edges explicitly retained in the acyclic graph, weighted by their bidirectional-consistency strengths.

F.3 Ordinal Signal and Delayed Credit

The trace records internal self-evaluation rewards spanning 0.569–0.779 across 63 sessions and ordinal-credit standard deviations of 0.104–0.127 across twelve updates (Figure 5), confirming that the internal learning signal is non-degenerate. Preference agreement with human judgments is evaluated separately on SummEval (Section 3.3).

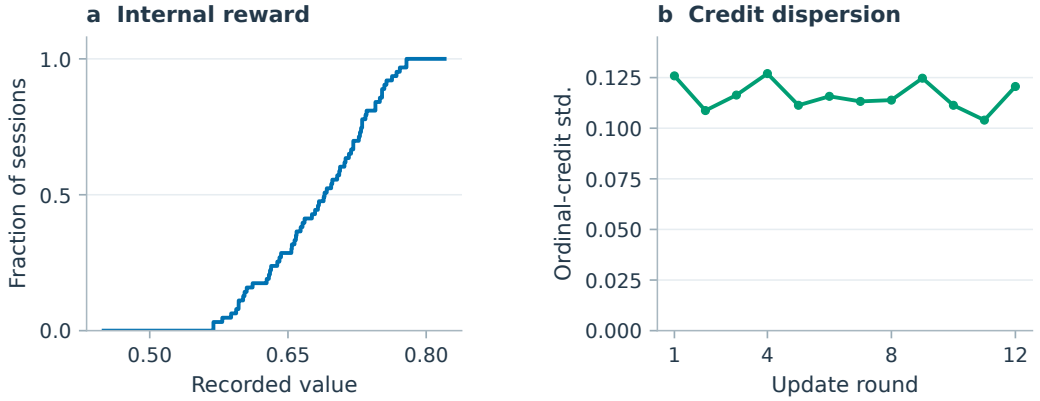

Figure 5: Internal self-evaluation signal diagnostics. (a) Empirical cumulative distribution of internal rewards over 63 sessions; (b) standard deviation of ordinal credit at each of the twelve adapter updates. Both statistics are computed from the archived training trace.

Discounted propagation links early skill choices to later session outcomes. Expanding the return recursion gives a weighted sum of subsequent turn credits and the terminal reward:

$$G_t = \sum_{j=t}^{T} \gamma^{j-t} \tilde{r}_j + \gamma^{T-t+1} r_{\text{session}}. \tag{14}$$

An early action that builds rapport can therefore receive credit associated with later progress even when its immediate score is modest. The discount reduces the contribution of distant outcomes. This mechanism specifies how credit is assigned to earlier actions; isolating its effect would require a controlled intervention on the propagation rule.

G Reward Separation and Judge Agreement

Reward leakage concerns the flow of evaluation information into learning. Internal feedback is used to construct candidate preferences and trajectory returns, whereas external evaluation scores frozen transcripts. The write-path audit checks whether external scores enter prompts, skill values, posteriors, evidence buffers, or adapter losses. Recording call order and policy versions allows this separation to be checked.

Table 12: Feedback channels and the modules that use them. Each row gives the unit at which a channel is recorded and the module that consumes it; only the first two channels enter learning.

Channel	Unit	Use
Internal candidate judgment	Candidates in one state	LiPO preferences
Client-grounded reward	Realized turns and sessions	HBSP trajectory credit
External PsychEval judgment	Frozen transcript	Outcome reporting
Human ballot	Matched dialogue item	Independent quality and ranking

We assess agreement between Qwen3-32B and Gemini-3-Flash on 1,468 matched sessions from 100 clients (Table 13). We estimate 95% confidence intervals for Spearman correlation using 2,000 client-level bootstrap resamples, retaining each sampled client's complete set of matched sessions.

Table 13: Agreement between Qwen3-32B and Gemini-3-Flash judges on 1,468 matched sessions. Intervals are 95% client-level bootstrap intervals (2,000 replicates) for the Spearman correlation, and offsets are mean Qwen minus Gemini scores.

Metric	Pearson r	Spearman ρ	95% interval	Offset
C-Shared	.430	.480	[.412,.538]	+1.989
WAI	.450	.499	[.431,.560]	+1.398
C-Spec	.677	.632	[.531,.708]	−.652
L-Shared	.521	.530	[.463,.591]	+.729

The judges exhibit positive rank correlations across all four reported dimensions, with Spearman correlations ranging from 0.480 to 0.632 and confidence intervals above zero. Mean score offsets quantify their differences in calibration. Positive judge correlations do not establish equivalent calibration, independent feedback, or clinical validity. In particular, improvement under this simulator and these evaluators need not persist under a different simulator or a different preference source.

A separate matched reward-channel audit uses 75 sessions from 40 clients. Client-grounded reward has Pearson correlation 0.564 and Spearman correlation 0.477 with external Gemini C-Shared. The local supervisor checklist has lower correlations of 0.178 and 0.157, respectively. These estimates compare different signals on a matched subset and are distinct from the 1,468-session judge-agreement estimates.

Correlation with a constant control score is undefined. Likewise, correlation between a composite reward and one of its own components partly reflects its algebraic construction. Neither observation establishes independence from an external judge. The checks therefore combine non-constant matched scores with verification that external scores are not written into the training process.

H Human Evaluation and Longitudinal Cases

H.1 Human Evaluation Protocol

To complement automatic evaluation, three experts with backgrounds in psychology rated multi-session counseling dialogues generated by PsyEvo, PsychAgent-Qwen3-32B, and CPsyCounX. The materials cover 25 cases, with three sessions from different counseling stages per case, yielding 75 sessions. Cases are selected at random, and the three systems are matched by case and corresponding counseling stage. The rating platform randomly maps system identities to anonymous labels A, B, and C and randomizes display positions, while keeping the system-to-label mapping consistent across sessions of the same case. The three raters score independently without knowing the system identities. The platform retains the identity mapping and session metadata for matched comparisons and aggregation after scoring. The main text reports aggregate expert ratings and the unweighted mean of six dimension-level means.

H.2 Rating Criteria and Scale Anchors

The evaluation interface provides dimension definitions and anchors on a 1–5 scale. Experts rate ethics, empathy accuracy, empathy quality, interaction, and intervention for each session, and rate client improvement once after reading the three sessions of a case. Client improvement assesses expert impressions of emotional, cognitive, or behavioral changes across those simulated sessions, rather than measured clinical change.

H.3 Aggregation

For the six dimensions D_6, let $\bar{s}_d$ denote the mean rating for dimension d. The overall rating is

$$\text{Overall}_6 = \frac{1}{6} \sum_{d \in D_6} \bar{s}_d. \tag{15}$$

Table 14: Criteria and anchors presented to expert raters for the six reported dimensions.

Dimension	Criterion	1–2	3	4–5
Ethics	Counseling ethics and professional boundaries	Unsupported diagnoses, dangerous advice, or excessive promises	Minor boundary issues without substantive harm	Respects boundaries; clarifies diagnostic or medical limits
Empathy accuracy	Recognition of expressed emotion	Frequently invents unexpressed emotions	Mostly accurate, with occasional subjective inference	Accurately reflects expressed emotion
Empathy quality	Depth and supportive value	Little empathy or awkward, inappropriate responses	Reasonable but superficial empathy	Deep understanding and support for self-awareness
Interaction	Continuity, coherence, and pacing	Disconnection, repetition, or abrupt topic changes	Generally coherent, with occasional mismatch or repetition	Natural, responsive, and fluent interaction
Intervention	Fit to the counseling stage	Premature advice or generic guidance	Appropriate direction with imperfect timing or intensity	Gradual intervention matched to the stage and therapeutic logic
Client improvement	Change across three sessions	No improvement or clear deterioration	Some improvement or fluctuation	Clear, sustained positive change

The first five means aggregate expert session ratings; client improvement aggregates expert case ratings. The six means receive equal weight, and human-evaluation results are reported to two decimal places.

H.4 Inter-Rater Agreement

The agreement results reported here concern the same 25-case, three-rater evaluation as Table 5. Table 15 reports ordinal Krippendorff's α for each dimension and the two-way absolute-agreement ICC(A,3) for the three-rater mean composite, together with their reported 95% confidence intervals. The reported field name *Evolution* is retained in the agreement table. Dimension-level agreement is limited. The composite reliability interval is not an interval for a paired difference between systems. Low agreement does not alone determine whether within-rater system orderings are consistent; that question requires per-rater contrasts and paired case-level uncertainty. The reported human comparison also does not cover the strongest general-purpose systems in the automatic evaluation.

Table 15: Reported inter-rater agreement and reliability for the 25-case evaluation with three human raters. Ordinal α measures dimension-level agreement; ICC(A,3) measures absolute-agreement reliability of the three-rater mean composite.

Dimension / composite	Metric	Estimate	95% CI
Ethics	Ordinal α	0.317	[0.204, 0.455]
Empathy accuracy	Ordinal α	0.035	[−0.131, 0.171]
Empathy quality	Ordinal α	0.153	[−0.040, 0.337]
Interaction	Ordinal α	0.106	[−0.057, 0.239]
Intervention	Ordinal α	0.350	[0.115, 0.585]
Evolution	Ordinal α	0.084	[−0.150, 0.235]
Six-dimension composite	ICC(A,3)	0.580	[0.398, 0.690]

H.5 HET-860: Client-Specific Adaptation Across Thirteen Sessions

HET-860 is a simulated client seeking help with repeated impulsive purchases, regret, and emptiness. We select the complete thirteen-session course to illustrate the transition from shared skill preferences to client-specific intervention selection. In Session 1, the client describes difficulty resisting appealing purchases; by Session 13, she reports sharing a drawing with a trusted friend and finding the emptiness associated with shopping less distressing. Drawing and self-directed reassurance increasingly become ways to respond to these urges.

Individual feedback reshapes skill preferences. Figure 6 compares skill rankings before the first session and after the thirteenth. Both snapshots evaluate all six skills at the same 32 visited states, each defined by counseling stage, within-session progress, and an emotion category. The top-

ranked skill set changes in 16 states and remains unchanged in the other 16. Adaptation in this case is therefore state-selective: individual experience reshapes intervention priorities in some contexts while preserving shared preferences in others.

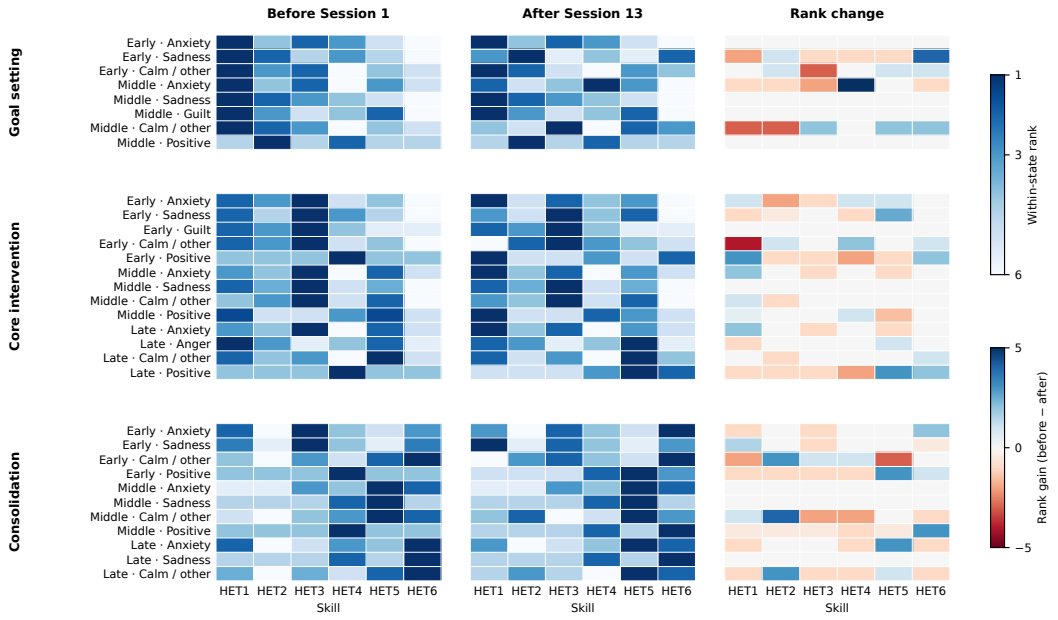

Figure 6: Skill preferences before and after thirteen sessions for HET-860. Left and middle: within-state rankings of shared values plus client-specific posterior means before Session 1 and after Session 13; darker colors indicate higher priority. Right: initial rank minus final rank, with blue indicating increased priority, red decreased priority, and white no change; tied skills receive average ranks. Rows cover all 32 states visited in the complete course, grouped by counseling stage. Early, Middle, and Late denote within-session progress; emotion categories are identified by keyword rules.

From shared preference to individual priority. For the state corresponding to core intervention, middle-session progress, and anxiety, the shared values initially favor HET3 over HET1 by 0.430. Between the first queries of this state in Sessions 2 and 10, sixteen updates raise the HET1 client-specific posterior mean from zero to 4.404. With shared values unchanged, the HET1-minus-HET3 value difference shifts from -0.430 to $+3.974$, reversing the ranking for this client. In Session 10, selecting HET1 supplies interventions centered on acceptance, empathic listening, and nonjudgmental expression of negative feelings. This example shows how accumulated individual feedback changes the priority of concrete interventions when the same encoded state recurs.

Adaptation and progress across sessions. The course contains 384 skill decisions and 384 posterior updates, with 282 decisions reading a nonzero client-specific residual. The number of updated state–skill entries grows from 17 after the first session to 72 after the last. At 66 recorded decision points, adding the posterior mean changes the highest-valued skill relative to the shared values alone. This continual revision of the individual decision state implements non-parametric adaptation across sessions.

Over the same trajectory, external WAI increases from 6.250 to 7.500, and the mean of PANAS and SRS increases from 5.188 to 6.875. The complete session sequence shows improvement with fluctuations: the client-shared score peaks at 7.562 in Session 9 and subsequently declines, ending above its initial value. Table 16 presents the scores alongside posterior use and the adapter versions recorded during generation.

Table 16: The complete thirteen-session HET-860 case. Client-shared is the mean of external PANAS and SRS scores. Nonzero reads count decisions that access a nonzero posterior mean; cells count cumulatively updated state–skill entries. Adapter versions are those recorded during generation.

Session	WAI	Client-shared	Nonzero reads	Cells	Adapter
1	6.250	5.188	0/28	17	v_0
2	6.042	4.875	0/25	31	v_1
3	7.292	4.938	21/29	37	v_1
4	7.917	6.938	30/33	40	v_2
5	7.917	6.188	29/29	41	v_3
6	7.083	5.312	34/35	44	v_4
7	7.708	6.500	27/27	45	v_5
8	6.667	6.062	23/23	45	v_5
9	6.667	7.562	34/34	46	v_6
10	7.917	6.500	33/35	48	v_6
11	6.667	6.875	0/23	61	v_7
12	6.875	6.438	21/25	65	v_{11}
13	7.500	6.875	30/38	72	v_{11}

H.6 A Client-specific Course: CBT-412

The CBT-412 archive follows a simulated retail worker whose anxiety and self-devaluation are triggered when colleagues receive praise. The client repeatedly thinks that they are inferior to others, with accompanying chest tightness and avoidance of a supervisor. This case illustrates how changes in the client state connect memory, skill selection, and response-policy learning across five saved sessions.

The first session identifies a trigger–thought–emotion–avoidance chain. When imagining the supervisor approaching, the client reports anxiety of at least eight on a ten-point scale and agrees to begin with a familiar calming image. Later summaries retain the workplace trigger, bodily cues, and the belief that external recognition determines personal value. These retained details allow subsequent interventions to address the same difficulty without repeating a generic intake.

The recorded skill sequence progresses from emotion regulation toward trigger monitoring, cognitive distancing, and self-supportive language. By the fifth session, the summary records that the client recognizes the recurring thought and remains in the situation rather than immediately withdrawing. Residual discomfort remains in the profile. This progression is documented in the narrative, with memory and both learning paths present throughout the course.

Table 17: Recorded state changes for client CBT-412. The first three rows track session-wise state; the final two report the archived ordinal tiers and expanded credit.

Record	Observed sequence
Sessions	1, 2, 3, 4, 5
Updated posterior cells	14, 30, 34, 42, 43
Adapter versions	v_0, v_0, v_1, v_2, v_3
Archived ordinal tiers	$[0, 0, 1, 2, 1]$
Archived expanded credit	$[0.241, 0.241, 0.741, 1.000, 0.741]$

Across the five sessions, the number of updated posterior cells grows from 14 to 43, while the active adapter advances from v_0 to v_3. Together with the retained workplace context, these records illustrate how client-specific skill adaptation and response-policy updates accompany a continuous counseling course.

H.7 Failure Cases and Memory Dependence

A fully tied score group has no recoverable strict order. One archived run records 725 rewards equal to 0.5 across 51 cases. These observations support recording variance and skip decisions instead of

assigning arbitrary preferences. They also distinguish a failure of the score format from successful ordinal reconstruction when the inputs remain distinguishable.

Cross-session memory can preserve both useful context and earlier errors. If a recap incorrectly records a goal or belief, subsequent skill selection and candidate evaluation may reinforce that mistake. A correction from the client should therefore be incorporated into the retained profile with its source preserved. Memory ablations should compare absent memory, fixed summaries, and revisable summaries before session trends are attributed to HBSP or LiPO.

I Compute, Deployment, and Safety

The principal online overhead consists of shadow generation, local evaluation, and adapter optimization. The instrumented training trace records 2,080 pairs and 2,217.6 seconds of trainer time over twelve rounds, or 36.96 minutes in total. The mean trainer time in this trace is 3.08 minutes per round on four trainer GPUs. This trainer-only measurement excludes dialogue serving and shadow generation and does not establish response latency or end-to-end session time.

Let N_s denote the number of additional shadow responses generated at each of T opportunities. Let C_g be the generation cost per shadow, C_j the amortized local evaluation cost per candidate, including pointwise scoring and pairwise comparisons, and $C_o(P)$ the optimization cost for P pairs. If the delivered response is evaluated together with the shadows, the incremental cost over ordinary single-response generation is approximately

$$C_{\text{extra}} \approx T\left[N_s C_g + (N_s + 1)C_j\right] + C_o(P). \tag{16}$$

Distinguishing shadow responses from the full candidate set prevents double-counting the cost of generating the delivered response. The instrumented configuration uses $N_s = 4$ shadows plus the delivered response, giving $N = N_s + 1 = 5$ candidates. A saved rank-8 adapter occupies approximately 79.8 MB, allowing the large backbone to remain shared. Storage includes client-specific memories and posteriors, cross-client preference evidence, and optimizer states and adapter versions maintained for each learning scope.

Synchronous activation allows update timing to be audited but can delay the next session. An asynchronous service could keep the previous verified adapter active while preparing a replacement, whereas an uncertainty-triggered service could reduce generation and training frequency. These are deployment alternatives to the measured synchronous configuration. Their effects on latency and the continued use of an earlier policy version require separate measurement.

Table 18: Deployment risks and candidate controls. Each row names a risk, the path through which it could arise under test-time learning, and a control whose effectiveness remains to be evaluated.

Risk	Failure path	Control
Crisis handling	Engagement reward competes with escalation	Independent crisis routing and professional oversight
Policy drift	Repeated local updates shift behavior	Fixed checks, versioning, rollback
Preference poisoning	Adversarial clients inject misleading preference evidence	Per-client isolation, client-balanced loss, edge caps, validation before activation
Memory error	Incorrect summaries persist	Source-linked correction
Privacy	Sensitive information in memory or adapters	Provenance, access control, deletion and unlearning evaluation
State/version routing	Wrong client memory or counselor version loaded	Identity checks and activation records

Automatic evaluations use simulated clients, and human ratings assess interaction quality. Clinical efficacy and safety require prospective evaluation with professional oversight, privacy protections, and crisis-handling assessment.

J Update Schedule and Ordinal Properties

Algorithm 1 gives the complete online update schedule. This appendix describes the conditional properties of the ordinal representation and its learning signal; implementation details and version records are provided in Appendix C.

Proposition 1. *For a strictly increasing function f, $R_i > R_j$ if and only if $f(R_i) > f(R_j)$. Consequently, the tie-aware ranks and tiers used for realized-trajectory projection are invariant to a strictly increasing transformation of the turn scores.*

The proof follows directly from strict monotonicity. Clipping after ordinal expansion preserves the mean because the implementation solves for a common shift, although the projected distances need not remain invariant. Quantization can also introduce ties, and a non-monotone evaluator can reverse an ordering. The proposition concerns the realized-trajectory representation; SOCA candidate edges are instead admitted by the two-order consistency rule in Eq. 7.

For an accepted edge, the derivative with respect to the utility difference $d_{ij} = h_i - h_j$ is $-w_{ij}\sigma(-d_{ij})$. It is nonzero for positive weight and finite utility difference. This property applies in score space; the parameter gradient also depends on the model Jacobian. Standard DPO likewise learns from preference labels and does not require a numerical reward gap once a valid pair exists.

K Evidence Aggregation and Additional Results

K.1 Supplementary MusPsy Comparison

Table 19 reports the supplementary MusPsy figures for six sessions over 100 profiles. The main PsyEvo run has not yet been linked to the supplementary configuration in Appendix C.1; these figures therefore do not serve as evidence for the core method comparison. The CACTUS and SMILE rows are published references, using the dataset authors' separately reported WAI averages; the other baseline rows do not yet have a complete row-level source mapping in the available record. The numerical comparison is retained for completeness, without a superiority or robustness claim.

Table 19: Supplementary MusPsy figures. WAI and positive PANAS favor higher values; negative PANAS favors lower values. The main PsyEvo configuration linkage remains unresolved, so these point estimates are not used to establish the paper's principal experimental claims.

Method	WAI Goal	WAI Task	WAI Bond	WAI Avg	PANAS Pos.	PANAS Neg.
CACTUS	4.125	3.713	4.427	4.085	3.035	2.333
CPsyCounX	4.352	4.239	4.639	4.410	3.325	1.317
SMILE	3.582	3.647	4.257	3.827	2.695	2.875
PsyDTLLM	4.327	4.219	4.595	4.380	3.118	1.231
PsyLLM-thinking-v3	4.406	4.314	4.688	4.469	3.547	1.248
TheraMind	4.409	4.312	4.667	4.463	3.438	1.236
PsychAgent-32B	4.467	4.375	4.695	4.512	3.408	1.134
GPT-5.6-terra	4.522	4.463	4.766	4.584	3.531	1.210
DeepSeek-V4-Flash	4.499	4.418	4.739	4.552	3.479	1.188
Gemini-3.8-Flash	4.511	4.467	4.768	4.582	3.305	1.257
Qwen3-32B	4.486	4.401	4.741	4.543	3.462	1.136
PsyEvo	**4.567**	**4.551**	**4.799**	**4.639**	**3.997**	**1.057**

K.2 Multi-evidence Aggregation Beyond Counseling

We use SummEval to evaluate a summary-ranking variant of SOCA against human judgments (Fabbri et al., 2021). For summaries of the same source document, SOCA combines pointwise scores, relative preferences from pairwise comparisons, and accepted repair evidence. Pointwise scores assess individual candidate quality, pairwise comparisons assess relative quality, and repairs provide evidence of correctable defects.

For candidate i and quality dimension d, let $q_{i,d}$ be the expected pointwise score, $v_{i,d}$ the mean pairwise win rate across the two presentation orders, and $r_{i,d}$ the accepted-repair signal. Within

each source-document group, the fused score is

$$f_{i,d} = z(q_{i,d}) + \alpha_d z(v_{i,d}) - \beta_d z(r_{i,d}), \qquad z(x_i) = \frac{x_i - \bar{x}}{\mathrm{sd}(x) + 10^{-8}}. \tag{17}$$

The pointwise channel has unit weight; α_d and β_d control the contributions of pairwise and repair evidence for dimension d. We set $\alpha_d = \beta_d = 1$ for every dimension a priori and keep these coefficients fixed throughout evaluation. Equal coefficients apply after within-document channel standardization; they are not selected by maximizing the reported human correlations. The pointwise expectation is $q_{i,d} = \sum_{k=1}^{5} k\, p_{i,d}(k)$. The repair signal is the mean probability that the accepted revision is preferred in the two presentation orders, and is zero when no repair is accepted. Its negative coefficient treats an accepted improvement as evidence of a defect in the original summary. Candidates are ranked by decreasing $f_{i,d}$; a constant channel contributes zero after standardization. SummEval measures agreement between the fused scores and human rankings, rather than the counseling graph selector or trajectory-credit mechanism. The mean correlation differences are descriptive: no paired document-level uncertainty is reported, so the observed margin is not a significance claim. In counseling, the SOCA ablation tests the combined contribution of consistency-checked candidate preferences and ordinal trajectory credit (Section 2.4).

The repair module proposes a local edit and checks its validity. The original and repaired summaries are compared in both presentation orders to check preference consistency. Only accepted repairs contribute to the repair channel; pointwise and pairwise evidence remain available for every candidate.

JDM uses the expectation of the score distribution, whereas Raw Scalar uses its maximum-probability category. PCA evaluates relative candidate quality through pairwise comparisons (Liusie et al., 2024). The evidence-channel ablations are reported in Table 4.