LEARNING WHAT TO PRACTICE NEXT: EXPERIENCE-ADAPTIVE SELF-EVOLUTION FOR MULTI-TURN EMPATHETIC DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable feedback typically updates model parameters while leaving the distribution of subsequent experience fixed. This separation is especially limiting in multi-turn empathetic dialogue, where interaction conditions shape entire trajectories and their training value changes as the policy improves. We introduce experience-adaptive self-evolution, which turns verified emotion outcomes into an online signal for deciding what the policy should practice next. Continuous outcomes train the dialogue policy, while thresholded group outcomes estimate the training value of reusable intent–state units and reallocate future interactions. Hierarchical evidence sharing stabilizes sparse estimates, and uncertainty-guided exploration with uniform rehearsal preserves coverage. With the simulator and verifier frozen, the framework adapts experience without changing the role-playing target or increasing the rollout budget. On SAGE, it raises Qwen3-8B Overall from 53.87 to 79.24, outperforming protocol-matched uniform emotion-reward RL by 7.23 points; the advantage also holds at Qwen3-4B by 4.14 points. Cross-benchmark and human evaluations further show improvement beyond the training-aligned metric.

1 INTRODUCTION

Empathetic dialogue is intrinsically multi-turn. A response matters not only as supportive text in isolation, but also through how it changes the user's later disclosure, affect, and willingness to engage Rashkin et al. (2019); Liu et al. (2021); Cheng et al. (2022); Zhou et al. (2023); Wan et al. (2025); Wu et al. (2026); Zhang et al. (2026). A locally appropriate reply can therefore still derail the subsequent conversation. Improving an empathetic dialogue policy requires coherent trajectories that reveal these delayed interaction effects.

Interactive user simulation makes such trajectories scalable. A policy can observe how different responses influence later user behavior, provided that the simulator preserves the original persona, context, and hidden need Yoon et al. (2024); Gromada et al. (2025); Dou et al. (2025); Zhang et al. (2026). Verifiable emotion outcomes can then turn complete conversations into reinforcement-learning (RL) signals for long-term affective consequences Wang et al. (2026b). These pipelines improve how the policy responds, but typically leave the distribution of subsequent training interactions fixed.

This fixed allocation ignores that the policy changes during training. The same interaction condition may yield mostly failures for an early policy, a mixture of successes and failures as the policy improves, and mostly successes once mastered. Nevertheless, under uniform sampling it receives the same expected rollout budget at every stage. Verified outcomes thus update the policy but are not reused to decide which conditions should generate its next experience. We call this mismatch between evolving competence and static experience allocation the *policy–experience gap*.

Figure 1 contrasts three ways to allocate this experience. A static curriculum (Figure 1A) follows a distribution chosen before training and cannot respond to observed changes in competence Bengio et al. (2009); Kumar et al. (2010). Unconstrained self-play (Figure 1B) changes the simulated partner, but can thereby change the task itself: an arbitrarily resistant, overly accepting, or inconsistent user makes reward changes inseparable from simulator drift Yoon et al. (2024); Dou et al. (2025).

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.

Adaptive curriculum and replay methods establish that training utility can depend on the learner Graves et al. (2017); Schaul et al. (2016); Jiang et al. (2021); Shi et al. (2026); Jiang et al. (2025); multi-turn role-play adds the requirement that experience adapt without moving the role-playing target.

We introduce experience-adaptive self-evolution (Figure 1C). The simulator, scenario semantics, verifier, and finite interaction-state space remain fixed; only the sampling probabilities over those states evolve. Continuous emotion outcomes train *how the policy responds*. Thresholded outcomes from the same rollout groups estimate *what the current policy should practice next*: conditions near the observed success–failure boundary receive greater weight, sparsely observed conditions are revisited through uncertainty-guided exploration, and uniform rehearsal preserves coverage. Newly sampled trajectories then update both the policy and the statistics that allocate later practice. Here, self-evolution refers to this coupled evolution of policy behavior and training experience, not to changing the simulator or semantic target.

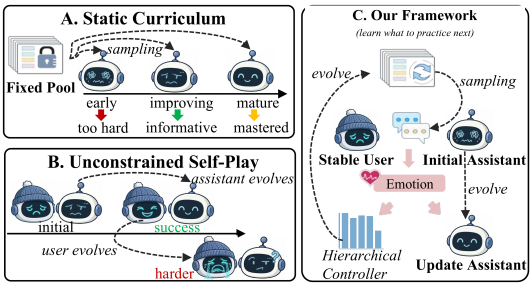

Figure 1: Learning what to practice next. Static curricula retain a fixed experience distribution, whereas unconstrained self-play changes the interaction target itself. Ours reallocates experience from verified outcomes while preserving stable role-play.

On held-out SAGE profiles, experience-adaptive training raises Qwen3-8B Overall from 53.87 to 79.24 and exceeds protocol-matched uniform emotion-reward RL by 7.23 points, a 10.0% relative gain, under the same rollout budget. The advantage also holds at Qwen3-4B by 4.14 points. Complementary automatic, interactive, and human evaluations test performance beyond the training-aligned metric. Shuffling outcomes across interaction conditions and fixing the learned distribution from training onset reduce Overall by 7.93 and 5.47 points, respectively, connecting the gain to condition-aligned feedback and online allocation.

Our contributions are threefold:

- We identify the *policy–experience gap* in multi-turn verifiable-feedback RL: policy competence evolves while the distribution that generates subsequent training interactions remains fixed.
- We introduce experience-adaptive self-evolution, which reuses verified emotion outcomes to estimate what the current policy should practice next and reallocates future interactions without changing the simulator, semantic target, or verifier.
- Multi-run experiments across two model scales, complementary automatic and human evaluations, and targeted controls demonstrate consistent gains and isolate the roles of condition-aligned feedback and online distribution updates.

2 Related Work

Empathetic dialogue and affective reinforcement learning. Empathetic dialogue has progressed from emotion-grounded generation to long-horizon support over latent needs and strategies Rashkin et al. (2019); Lin et al. (2019); Majumder et al. (2020); Liu et al. (2021); Cheng et al. (2022), with work on planning, augmentation, interpretable reasoning, strategy prediction, and personalization Zhou et al. (2023); Zheng et al. (2023); Zhang et al. (2024b); Kang et al. (2024); Wan et al. (2025); Shi et al. (2025); Ye et al. (2025). RL encourages value-sensitive support Kim et al. (2025); Zhao et al. (2025); Wang et al. (2025), while LLM simulators enable scalable interaction but require faithful role-play and assessment Yoon et al. (2024); Gromada et al. (2025); Dou et al. (2025). SAGE supplies evolving emotional personas Zhang et al. (2026), and verifiable emotion rewards optimize policies Wang et al. (2026b); we instead align the interaction distribution with the evolving policy.

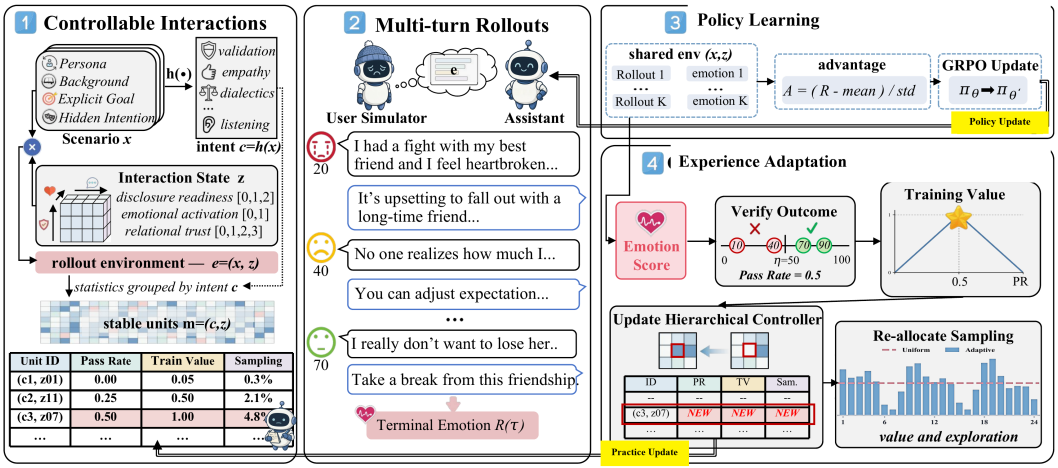

Figure 2: Experience-adaptive self-evolution. Continuous emotion outcomes improve the dialogue policy; thresholded outcomes from the same groups update hierarchical intent–state statistics that determine what to practice next. Scenarios remain uniformly sampled, and the simulator, semantics, and verifier remain fixed.

Adaptive curricula and self-evolving training. Curriculum, self-paced, and replay methods allocate computation by learner-dependent utility Bengio et al. (2009); Kumar et al. (2010); Jiang et al. (2015); Schaul et al. (2016); Jiang et al. (2021); LLM self-improvement adds self-play, self-judgment, and online optimization Chen et al. (2024); Yuan et al. (2024); Ahmadian et al. (2024). Recent work adapts prompts or rollouts using success, uncertainty, variance, and progress Shi et al. (2026); Jiang et al. (2025; 2026); Liu et al. (2026); Li et al. (2026); Zeng et al. (2026); Nguyen et al. (2026), or evolves tasks, users, evaluators, and harnesses Dai et al. (2026); Wang et al. (2026a); Yang et al. (2026); Chen et al. (2026). Among the closest allocation mechanisms, AdaRFT resamples static problems from their recent success rates Shi et al. (2026), VCRL prioritizes static problems using within-group reward variance Jiang et al. (2025), and VIP adjusts the number of rollouts for each batch prompt using predicted gradient variance Nguyen et al. (2026). These methods therefore select static problems or modify rollout multiplicity using recent evidence. In contrast, our controller uses a hierarchical pass boundary estimated from accumulated groups to reallocate reusable intent–state units, thereby adapting the conditions of future multi-turn interactions as policy competence changes while keeping the simulator fixed.

3 Experience-Adaptive Self-Evolution

Figure 2 presents the complete training pipeline. Conventional verifiable-feedback RL updates the policy while repeatedly sampling from a fixed distribution. Our framework additionally learns which interaction conditions should receive future practice. Each verified trajectory contributes a continuous reward for policy learning; completed groups also update historical intent–state statistics. The controller converts these statistics into a refreshed distribution for subsequent rollouts, making training experience responsive to the policy's changing competence.

3.1 Task Setup and Experience-Adaptive Formulation

Training begins from a set $\mathcal{X}$ of complete emotional-support scenarios. A scenario $x \in \mathcal{X}$ specifies the user's persona, precipitating event, current dilemma, response tendencies, and hidden support need; an intent map h associates it with a support intent $c = h(x)$. We preserve this semantic core throughout training. To control how the same concern unfolds in dialogue, we additionally choose a simulator-side interaction state z from a finite space $\mathcal{Z}$. The pair $e = (x, z)$ is one concrete rollout environment: x determines *what* the user is experiencing, while z determines *how* the user discloses and reacts.

For a sampled condition (x, z), a user simulator $\mathcal{U}$ and the dialogue policy π_θ interact for at most H turns. Their alternating user and assistant utterances form a trajectory

$$\tau = (u_1, a_1, \ldots, u_T, a_T) \sim P(\tau \mid x, z, \pi_\theta, \mathcal{U}),$$

where $T \leq H$ is the realized number of turns. After the interaction ends, an emotion verifier evaluates the complete trajectory and returns a scalar outcome $R(\tau)$. The policy observes only the natural-language conversation: support-intent labels, interaction-state labels, controller statistics, and sampling probabilities remain simulator-side metadata. Consequently, improved performance must be expressed through dialogue behavior rather than direct access to a difficulty label.

At training stage t, a conditional distribution $p_t(z \mid c)$ represents the current allocation of interaction states for each support intent and is initialized uniformly before feedback accumulates. Sampling from this distribution determines the conditions under which the policy is optimized:

$$\max_\theta \quad \mathbb{E}_{x\sim\mathrm{Unif}(\mathcal{X}),\, z\sim p_t(\cdot|h(x)),\, \tau} \left[R(\tau)\right]. \tag{1}$$

Meanwhile, verified outcomes update controller history $\mathcal{H}_t$ and hence the next interaction distribution:

$$\begin{aligned} \mathcal{H}_{t+1} &= \mathrm{Update}(\mathcal{H}_t; \mathcal{D}_t, R_t), \\ p_{t+1} &= \mathrm{Controller}(\mathcal{H}_{t+1}). \end{aligned} \tag{2}$$

Here $\mathcal{D}_t$ denotes the completed rollout groups at stage t, R_t their outcomes, and $\mathcal{H}_t$ the accumulated evidence. Each completed group immediately updates this stabilized historical summary before subsequent state draws, so p_{t+1} responds online without relying on a noisy instantaneous estimate. The pair (θ_t, p_t) therefore evolves jointly: the same outcome stream updates policy behavior and the conditions under which future behavior is learned.

3.2 A Controllable Interaction Space

The framework requires interaction states to be behaviorally realizable, compatible with the underlying scenario, hidden from the assistant, and reusable within a rollout group. We preserve each complete scenario and define three task-operational axes: *disclosure readiness* controls when details and deeper needs are revealed; *emotional activation* controls the intensity of negative-affect expression; and *relational trust* controls whether questions and support are accepted. These axes capture information release, affective expression, and relational feedback rather than clinical traits.

We discretize the axes at a granularity that remains behaviorally distinguishable and statistically estimable: disclosure has three levels (delayed, conditional, proactive), activation has two bounded levels (moderate, high), and trust has four levels from skepticism to sustained engagement. Their Cartesian product gives $|\mathcal{Z}| = 3 \times 2 \times 4 = 24$ interaction states. A state $z = (d, a, r)$ is rendered by composing one natural-language behavioral clause for each axis, $\phi(z) = \phi_d(d) \oplus \phi_a(a) \oplus \phi_r(r)$ for $z = (d, a, r)$. These clauses govern disclosure pace, affective expression, and receptivity, but cannot overwrite the persona, event, or hidden need specified by x.

Estimating a separate curriculum value for every scenario–state pair would fragment evidence across many nearly related conditions. We therefore retain individual scenarios for rollout generation but organize controller evidence by support intent. Combining intent c with state z defines the reusable unit $m = (c, z)$. This factorization lets different scenarios with the same support objective contribute evidence about a shared interaction state, while preserving intent-specific differences. Scenario identities and intent frequencies remain fixed; only the conditional state distribution $p_t(z \mid c)$ evolves. Section 4.3 later separates the contribution of this interaction space from that of adaptive allocation.

3.3 Group-Shared Multi-Turn Policy Learning

For each scenario, the controller selects one state z, and group-relative policy optimization draws $\tau_{1:K} \overset{\text{i.i.d.}}{\sim} P(\tau \mid x, z, \pi_\theta, \mathcal{U})$ under the shared condition (x, z). Holding the environment fixed within a group prevents differences in user behavior from confounding group-relative advantages, so rollout variation primarily reflects policy and generation stochasticity.

Abstracting from stabilization terms held fixed across all controlled systems, the method-level policy reward is the continuous verifier outcome, $r_k^{\text{policy}} = R(\tau_k)$. GRPO computes group-relative advantages from these rewards and applies the standard clipped policy objective with KL regularization Shao et al. (2024); Schulman et al. (2017). When all trajectories in a group receive the same reward, their normalized group-relative advantages are zero, so that group contributes no policy-gradient signal. The allocation controller reuses these outcomes to organize subsequent training experience.

3.4 Emotion-Thresholded Group Feedback

The policy and controller consume different transformations of the same verified outcome. Given an emotion criterion η, the controller computes

$$y_k = \mathbb{I}[R(\tau_k) \geq \eta], \qquad p_m = K^{-1} \sum_{k=1}^{K} y_k, \tag{3}$$

where y_k records whether trajectory k passes the criterion, and p_m is the pass rate of the shared unit $m = (h(x), z)$ in that group. We use the pass rate because it directly reveals whether the current policy produces both successful and unsuccessful behavior under a controlled condition. Thresholding is used only to construct allocation statistics: the continuous value $R(\tau_k)$ remains the policy reward and therefore retains fine-grained differences between trajectories. One complete group contributes one pass-rate observation, preventing correlated trajectories from masquerading as independently sampled environments.

3.5 Robust Outcome-Informed Allocation

Controller evidence is uneven: some intent–state units are observed repeatedly, whereas others have only a few completed groups. Trusting a sparse empirical rate can turn an early random success or failure into a persistent sampling bias. For unit (c, z), let $S_{c,z}$ be the cumulative sum of group pass rates and $N_{c,z}$ the number of complete groups. We stabilize its estimate with a leave-one-intent prior formed from the same state in the other support intents:

$$\begin{aligned} S_z^{-c} &= \sum_{c' \neq c} S_{c',z}, \qquad N_z^{-c} = \sum_{c' \neq c} N_{c',z}, \\ \hat{p}_z^{-c} &= \frac{S_z^{-c} + \alpha/2}{N_z^{-c} + \alpha}, \qquad \hat{p}_{c,z} = \frac{S_{c,z} + \lambda \hat{p}_z^{-c}}{N_{c,z} + \lambda}. \end{aligned} \tag{4}$$

Here $\hat{p}_z^{-c}$ summarizes the outcomes associated with state z outside intent c, and $\hat{p}_{c,z}$ is the resulting unit estimate. The coefficient α supplies a neutral prior when pooled evidence is scarce, while λ controls how strongly the unit borrows that evidence. Excluding c avoids counting its observations in both terms. As $N_{c,z}$ grows, the unit's own history naturally dominates.

The stabilized pass rate is then converted into an allocation score. We assign maximal boundary value when success and failure coexist, and add an uncertainty bonus that revisits under-observed units Auer et al. (2002):

$$\begin{aligned} B_{c,z} &= \max(0, 1 - 2|\hat{p}_{c,z} - \tfrac{1}{2}|), \\ U_{c,z} &= \sqrt{\log(N+1)/(N_{c,z}+1)}, \\ A_{c,z} &= \max(A_{\min}, B_{c,z} + \beta U_{c,z}). \end{aligned} \tag{5}$$

Here $B_{c,z}$ measures proximity to a balanced historical pass rate, $U_{c,z}$ expresses uncertainty from limited visits, and $A_{c,z}$ combines the two with exploration weight β while retaining a minimum score $A_{\min}$. In $U_{c,z}$, $N = \sum_{c',z'} N_{c',z'}$ denotes the total number of completed groups across all units, so that under-observed units receive a larger exploration bonus. Near $\hat{p}_{c,z} = 0.5$, the accumulated evidence contains both successful and unsuccessful trajectories under the same interaction condition. Cumulative aggregation smooths transient group noise, while immediate updates and the coverage terms keep allocation responsive throughout training.

Algorithm 1 Experience-Adaptive Self-Evolution with Verifiable Emotion Feedback

Require: scenarios $\mathcal{X}$ with intent map h, interaction states $\mathcal{Z}$, policy π_θ, group size K

keep the controller in uniform mode until 1,600 complete groups are recorded
for each training batch **do**
  for each uniformly sampled scenario x **do**
    set $c = h(x)$ and sample z uniformly before activation, otherwise from the latest $p_t(z \mid c)$
    generate K trajectories under the shared environment (x, z)
    score each trajectory with continuous emotion outcome R_k
    update π_θ using group-relative advantages from $R_{1:K}$
    set $y_k = \mathbb{I}[R_k \geq \eta]$ and $p_m = K^{-1} \sum_k y_k$
    if the group is complete and finite, atomically accumulate p_m; subsequent draws use the refreshed scores
  end for
end for

The boundary term also has a direct relation to Bernoulli outcome variance. For any pass probability $p \in [0, 1]$, let $B(p) = 1 - |2p - 1|$. Then

$$\mathrm{Var}(y) = p(1 - p) = \frac{B(p)(2 - B(p))}{4}. \tag{6}$$

Because the right-hand side is strictly increasing in B on $[0, 1]$, the boundary score preserves the ordering of Bernoulli outcome variance. Its role nevertheless differs from within-group continuous-reward variance Jiang et al. (2025): $B_{c,z}$ is computed from the stabilized historical pass probability of a reusable intent–state unit, so it reallocates future environments using evidence accumulated across groups rather than prioritizing a single group's reward dispersion.

The controller first collects 1,600 complete groups under uniform state sampling, then activates feedback-guided allocation. Each complete finite group updates the sufficient statistics immediately; the next state draw reads the latest global statistics and recomputes the allocation scores. Incomplete groups and groups with non-finite outcomes do not update controller history. Thereafter,

$$p_t(z \mid c) = (1 - \epsilon) \frac{A_{c,z}^{1/\gamma}}{\sum_{z' \in \mathcal{Z}} A_{c,z'}^{1/\gamma}} + \epsilon \frac{1}{|\mathcal{Z}|}, \tag{7}$$

where γ is a sampling temperature and ϵ reserves uniform rehearsal. Thus, no state is permanently removed, and previously underweighted conditions can receive renewed emphasis as outcomes change.

3.6 Learning What to Practice Next

Every complete group provides evidence about the current policy under one controlled interaction condition. Continuous outcomes train the policy, while thresholded group outcomes update the condition's historical training-value estimate. The refreshed distribution then emphasizes informative or under-observed conditions in subsequent practice. Because the simulator, scenario semantics, and verifier stay frozen, this experience adaptation does not introduce a drifting role-playing target. Algorithm 1 summarizes the end-to-end process.

4 Experiments

4.1 Experimental Setup

We test three implications of experience-adaptive self-evolution: whether it improves empathetic dialogue under a fixed rollout budget, whether the gain requires correctly aligned online feedback rather than merely a larger state space or generic prioritization, and whether the learned behavior transfers beyond the training-aligned evaluator. The 500 training scenarios, development split, and 100-scenario SAGE test set are instance-disjoint. Eight native support intents and 24 interaction states define 192 controller units; test instances never enter training, controller updates, configuration, or checkpoint selection.

Table 1: Protocol-matched Qwen3-8B results across complementary automatic, interactive, and human evaluations. Under SAGE, "Avg." is the Overall score (mean final user emotion), i.e., the value reported as "SAGE Overall" in the text and in Table 2.

	SAGE			ESC-Eval	EIBench	Human
Counselor training	Avg. ↑	Succ. ↑	Fail. ↓	Avg. ↑	Overall ↑	Avg. ↑
Qwen3-8B	53.87 ± 0.35	20.00	29.00	2.39 ± 0.03	-22.40 ± 0.62	2.3
RLVER	72.01 ± 1.88	43.00	15.33	2.49 ± 0.03	-9.79 ± 0.57	3.0
Interaction State Only	69.42 ± 0.40	30.00	19.00	2.43 ± 0.02	-40.60 ± 0.79	2.7
VCRL	68.51 ± 0.40	35.00	21.00	2.54 ± 0.02	-17.08 ± 0.63	2.9
Ours	$\mathbf{79.24 \pm 1.56}$	**49.00**	**11.33**	$\mathbf{2.55 \pm 0.02}$	$\mathbf{-7.55 \pm 0.33}$	**3.5**

Counselor training	ESConv				
	Acc. ↑	B-2 ↑	R-L ↑	BS ↑	D-2 ↑
Qwen3-8B	17.58	3.09	10.18	85.08	26.37
RLVER	16.72	2.75	9.83	85.43	28.18
Interaction State Only	16.61	2.99	10.90	85.17	24.12
VCRL	17.01	3.12	11.07	85.22	23.67
Ours	**18.13**	**3.30**	**11.45**	**85.69**	**33.58**

SAGE evaluates final emotion on 100 multi-turn scenarios, with Success $e_T \geq 100$ and Failure $e_T < 10$ Zhang et al. (2026). Training uses the continuous terminal score; $\eta = 50$ is only the controller criterion. Training and held-out SAGE share the frozen simulator, scorer, prompts, rules, information boundary, and temperature 0.7, but not scenarios. Transfer evaluations comprise ESC-Eval (331 interactions) Zhao et al. (2024), EIBench (213 scenarios) Zhu et al. (2026), ESConv (2,895 responses), and blinded human ratings; none reuses the training verifier. RLVER and ours use three independent runs. VCRL uses one checkpoint and three shared evaluation seeds (Overall 68.11/68.51/68.91; 68.51 ± 0.40); other fixed checkpoints also use three shared seeds. Table 1 reports sample SDs for aggregate automatic metrics Liu et al. (2023); Zhang et al. (2024a).

Three blinded annotators rate all systems on 100 held-out scenarios using a randomized 0–4 overall-support rubric. Ours scores 3.50 versus 3.00 for RLVER, with a paired difference of $+0.50$, a 95% CI of $[0.39, 0.61]$, and 67/21/12 wins/ties/losses. Ordinal Krippendorff's α is 0.78. Appendix G.2 gives this protocol and a separate 200-turn automatic–expert agreement audit.

We compare all systems under a controlled protocol with the same Qwen3-8B initialization, frozen simulator and verifier, data, rollout budget, group size, and optimization settings. **Qwen3-8B** applies no RL; **RLVER** performs uniform emotion-reward RL; **Interaction State Only** adds the 24 states without adaptive allocation; **VCRL** prioritizes within-group reward variance Jiang et al. (2025); and **Ours** uses the complete controller. All RL systems therefore receive the same GRPO budget, isolating how training interactions are allocated.

Appendices B–N provide data provenance, implementation and configuration details, complete evaluation protocols and results, controller analyses, state audits, qualitative examples, and reproduction instructions.

4.2 Main Results

Experience-adaptive self-evolution raises mean SAGE Overall from 72.01 under uniform emotion-reward RL to 79.24 across three independent Qwen3-8B training runs. Its weakest run (78.11) still exceeds the strongest uniform run (73.76). Under the same protocol at Qwen3-4B, it improves Overall from 44.73 ± 1.17 to 48.87 ± 0.96, increases Success from 14.67 ± 1.53 to 19.33 ± 1.53, and reduces Failure from 40.00 ± 1.00 to 34.33 ± 2.08. Appendix H.2 reports the three runs. The benefit is therefore preserved across the two tested Qwen3 scales; at 8B, the weaker Interaction State Only (69.42) and variance-based allocation (68.51) controls further show that neither a larger condition space nor generic reward dispersion explains the gain.

SAGE directly evaluates the emotional consequence of a complete multi-turn interaction: Avg. is final user emotion, while Succ. and Fail. summarize trajectories reaching the official high- and

Table 2: Single-run SAGE component ablations under a matched protocol. Δ is relative to the complete framework (78.11); Table 1 reports its three-run statistics.

Baseline and environment representation			Outcome statistics, exploration, and rehearsal		
Variant	Overall ↑	Δ	Variant	Overall ↑	Δ
Complete framework	**78.11**	–	*Outcome-statistic construction*		
Environment representation			Soft outcome mapping	76.52	-1.59
Interaction State Only	69.42	-8.69	No hierarchical sharing	75.07	-3.04
Scenario-only allocation	70.03	-8.08	Variance-gated pass-rate controller	61.45[†]	-16.66
Online-feedback mechanism			*Exploration and rehearsal*		
Shuffled feedback	70.18	-7.93	No uncertainty bonus	73.92	-4.19
Frozen final distribution	72.64	-5.47	No uniform rehearsal	77.03	-1.08

[†]Last stable checkpoint before KL instability.

low-emotion regions. Our framework improves Avg. and success while reducing failure to 11.33%, yielding more reliable long-horizon emotional recovery on this target-aligned evaluation. The comparatively large SAGE margin is consistent with the optimized terminal-emotion objective and the method's allocation of complete interaction conditions.

Beyond SAGE, Ours leads all five ESConv metrics and obtains the highest aggregate ESC-Eval (2.55), EIBench (-7.55), and human (3.5) scores. Interaction states without adaptive allocation instead lower EIBench from -22.40 to -40.60, showing that state diversity alone is insufficient; outcome-informed allocation makes this additional diversity beneficial under the fixed budget.

4.3 Ablations and Training Analysis

Two diagnostic controllers isolate allocation choices: *Scenario-only allocation* keys histories by scenario identity, while the *variance-gated controller* admits only high-variance groups to controller history; all groups still train the policy.

Interaction State Only retains the 24-state representation but removes adaptation; its 8.69-point drop shows that unallocated diversity can dilute a fixed budget. Scenario-only allocation replaces intent–state sharing with identity-level histories and falls by 8.08 points, supporting the value of reusable evidence across related interactions.

Soft outcomes decrease Overall by 1.59, removing hierarchical sharing costs 3.04, and variance-gated history reaches 61.45 before KL instability. Together these controls favor bounded criterion-aligned histories with cross-intent sharing over raw reward dispersion.

With the complete run's seed and budget, shuffling outcomes across intent–state units lowers Overall to 70.18, while fixing the final distribution from training onset yields 72.64. Shuffled feedback severs the correspondence between experience and outcome, whereas the frozen distribution prevents observed outcomes from changing future practice. Their 7.93- and 5.47-point drops isolate the contributions of condition-aligned feedback and online distribution updates; Appendix J.3 gives the full protocols.

Removing the uncertainty bonus and uniform rehearsal decreases Overall by 4.19 and 1.08, respectively. These single-run component results are diagnostic, whereas the primary comparison uses three independent runs. Sensitivity scores of 77.46, 78.11, and 77.72 for $\eta = 40, 50, 60$ support 50 as a stable reference. At step 500, the eight intents have six distinct highest-probability states, confirming that allocation remains intent-conditioned rather than collapsing to one global preference; Table 17 reports the full-run coverage statistics. Each controller update is constant-time plus normalization over 24 states and adds no trajectories or model calls.

4.4 Validating the Interaction-State Space

We audit 120 multi-turn trajectories, five per state, using a held-out Qwen3-Max judge that sees only the dialogue. As Figure 3 shows, requested disclosure, activation, and trust levels are recovered with 86.7%, 92.5%, and 82.5% accuracy; all three are correct in 68.3% of trajectories. Moreover, 92.5%

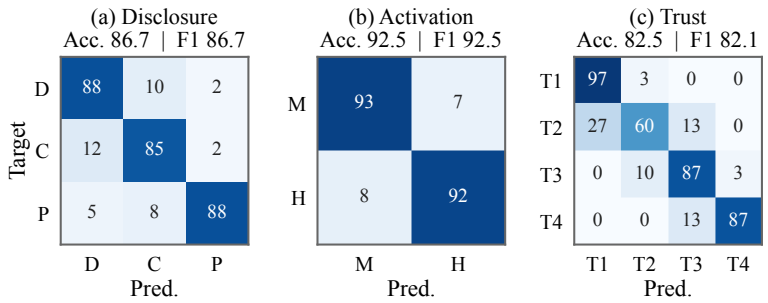

Figure 3: Row-normalized confusion matrices (%) of requested versus recovered (a) disclosure readiness, (b) emotional activation, and (c) relational trust.

satisfy every semantic-preservation and metadata-nondisclosure constraint. Appendix L provides the full protocol and dimension-level results. Thus, the controller reallocates behaviorally distinct yet semantically comparable interactions rather than nominal prompt labels.

Appendix M provides a complete trajectory comparison illustrating how the aggregate gain appears in dialogue behavior.

5 Discussion and Conclusion

Verifiable feedback can also teach a policy what to practice next. Experience-adaptive self-evolution reorganizes subsequent interaction conditions while keeping the simulator, semantic target, verifier, and rollout budget fixed. Improvements at Qwen3-4B and Qwen3-8B, four transfer evaluations, mechanism controls, and state validation connect the gain to condition-aligned online allocation over realizable interaction states. Thus, verified outcomes can improve policy behavior while aligning subsequent practice with changing competence.

Scope and outlook. Evidence covers Qwen3 at 4B and 8B with a fixed simulator–verifier stack; external evaluations are reported for 8B. Future work can extend policy and simulator families and assess naturalistic or safety-critical support.

AI Use Statement

Generative AI tools were used to translate the initial Chinese manuscript into English and to assist with language editing and LaTeX formatting. They were also used to assist in implementing code for reproducing baseline experiments and in aggregating and visualizing experimental results. The authors manually reviewed and tested all AI-assisted code, verified the implementations against the corresponding methodological descriptions, and cross-checked the aggregated statistics and visualizations against the raw experimental outputs. The research questions, proposed method, experimental protocol, interpretation of results, and scientific conclusions were determined by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript and its results.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of ACL*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of ICML*, pp. 41–48, 2009.

Guhong Chen, Yingcheng Shi, Yongbin Li, Binhua Li, Xander Xu, Hu Wei, Shiwen Ni, Min Yang, and Jieping Ye. Evotrainer: Co-evolving llm policies and training harnesses for autonomous agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2606.03108.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of ICML*, volume 235, pp. 6621–6642, 2024. URL https://proceedings.mlr.press/v235/chen24j.html.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of EMNLP*, pp. 3014–3026, 2022. doi: 10.18653/v1/2022.emnlp-main.195.

Yuqin Dai, Ning Gao, Wei Zhang, Jie Wang, Luozichen, Ruiyuan Wu, Jinpeng Wang, and Chaozheng Wang. SEAD: Self-evolving agent for multi-turn service dialogue. In *Findings of ACL*, pp. 3674–3684, 2026. doi: 10.18653/v1/2026.findings-acl.180. URL https://aclanthology.org/2026.findings-acl.180/.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In *Proceedings of EMNLP*, pp. 35212–35290, 2025. doi: 10.18653/v1/2025.emnlp-main.1786.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of ICML*, volume 70, pp. 1311–1320. PMLR, 2017. URL https://proceedings.mlr.press/v70/graves17a.html.

Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybyl, Lina M. Rojas-Barahona, and Michal K. Szczerbak. Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study. In *Proceedings of the EMNLP Industry Track*, pp. 230–245, 2025. doi: 10.18653/v1/2025.emnlp-industry.16. URL https://aclanthology.org/2025.emnlp-industry.16/.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. Vcrl: Variance-based curriculum reinforcement learning for large language models, 2025. URL https://arxiv.org/abs/2509.19803.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-paced curriculum learning. In *Proceedings of AAAI*, volume 29, pp. 2694–2700, 2015. doi: 10.1609/aaai.v29i1.9608. URL https://ojs.aaai.org/index.php/AAAI/article/view/9608.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *Proceedings of ICML*, pp. 4940–4950. PMLR, 2021.

Zishang Jiang, Jinyi Han, Tingyun Li, Xinyi Wang, Sihang Jiang, Xiaojun Meng, Jiansheng Wei, Jiaqing Liang, and Yanghua Xiao. Difficulty is not enough: Curriculum learning for llms fine-tuning must consider utility. In *Proceedings of AAAI*, volume 40, pp. 31365–31373, 2026. doi: 10.1609/aaai.v40i37.40400. URL https://ojs.aaai.org/index.php/AAAI/article/view/40400.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of ACL*, pp. 15232–15261, 2024. doi: 10.18653/v1/2024.acl-long.813. URL https://aclanthology.org/2024.acl-long.813/.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. Dialogue systems for emotional support via value reinforcement. In *Proceedings of ACL*, pp. 28733–28766, 2025. doi: 10.18653/v1/2025.acl-long.1395. URL https://aclanthology.org/2025.acl-long.1395/.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Proceedings of NeurIPS*, volume 23, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf.

Renda Li, Hailang Huang, Fei Wei, Feng Xiong, Yong Wang, and Xiangxiang Chu. Adacurl: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting. In *Proceedings of AAAI*, volume 40, pp. 23123–23131, 2026. doi: 10.1609/aaai.v40i27.39479.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. MoEL: Mixture of empathetic listeners. In *Proceedings of EMNLP*, pp. 121–132, 2019. doi: 10.18653/v1/D19-1012. URL https://aclanthology.org/D19-1012/.

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 ACL*, pp. 3469–3483, 2021.

Wenxuan Liu, Liangyu Huo, Yi Jing, Xiyuan Zhang, and Jian Xie. Mracl: Multi-reward space guided adaptive curriculum reinforcement learning for llms. In *Proceedings of AAAI*, volume 40, pp. 37663–37672, 2026. doi: 10.1609/aaai.v40i44.41101. URL https://ojs.aaai.org/index.php/AAAI/article/view/41101.

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 EMNLP*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of EMNLP*, pp. 8968–8979, 2020. doi: 10.18653/v1/2020.emnlp-main.721. URL https://aclanthology.org/2020.emnlp-main.721/.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *Proceedings of ICLR*, 2026. URL https://openreview.net/forum?id=Z5sWYACAop.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*, pp. 5370–5381, 2019. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *Proceedings of ICLR*, 2016. URL https://arxiv.org/abs/1511.05952.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=UEhpyq41b9.

Yuanchen Shi, Jiawang Hao, and Fang Kong. Beyond coarse labels: Fine-grained problem augmentation and multi-dimensional feedback for emotional support conversation. In *Findings of EMNLP*, pp. 1634–1647, 2025. doi: 10.18653/v1/2025.findings-emnlp.86. URL https://aclanthology.org/2025.findings-emnlp.86/.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. EmoDynamiX: Emotional support dialogue strategy prediction by modelling mixed emotions and discourse dynamics. In *Proceedings of NAACL*, pp. 1678–1695, 2025. doi: 10.18653/v1/2025.naacl-long.81. URL https://aclanthology.org/2025.naacl-long.81/.

Aozhe Wang, Yuchen Yan, Nan Zhou, Zhengxi Lu, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Code-A1: Adversarial evolving of code LLM and test LLM via reinforcement learning, 2026a. URL https://arxiv.org/abs/2603.15611.

Fanfan Wang, Xiangqing Shen, Jianfei Yu, and Rui Xia. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of EMNLP*, pp. 1341–1356, 2025. doi: 10.18653/v1/2025.findings-emnlp.70. URL https://aclanthology.org/2025.findings-emnlp.70/.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Cixing Li, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Jia Li, Zhaopeng Tu, and Xiaolong Li. RLVER: Reinforcement learning with verifiable emotion rewards for empathetic agents. In *Proceedings of ICLR*, 2026b. URL https://openreview.net/forum?id=P7wBg0vPTh.

Hongyan Wu, Zhiliang Tian, Zhen Huang, Tengyue Hu, Linbo Qiao, Yifu Gao, Feng Liu, and Dongsheng Li. Emotion trajectory-aware retrieval for markov-driven emotion anticipation in LLM-based emotional support conversation. In *Findings of ACL*, pp. 42887–42905, 2026. doi: 10.18653/v1/2026.findings-acl.2127. URL https://aclanthology.org/2026.findings-acl.2127/.

Shidong Yang, Ziyu Ma, Tongwen Huang, Yiming Hu, Yong Wang, and Xiangxiang Chu. CoEvolve: Training LLM agents via agent-data mutual evolution. In *Proceedings of ACL*, pp. 23015–23036, 2026. doi: 10.18653/v1/2026.acl-long.1055. URL https://aclanthology.org/2026.acl-long.1055/.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment. In *Findings of EMNLP*, pp. 18826–18853, 2025. doi: 10.18653/v1/2025.findings-emnlp.1024. URL https://aclanthology.org/2025.findings-emnlp.1024/.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of NAACL*, pp. 1490–1504, 2024. doi: 10.18653/v1/2024.naacl-long.83.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of ICML*, volume 235, pp. 57905–57923, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Yongcheng Zeng, Zexu Sun, Bokai Ji, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Haifeng Zhang, Xu Chen, and Jun Wang. Cures: From gradient analysis to efficient curriculum learning for reasoning llms. In *Proceedings of ICLR*, 2026. URL https://openreview.net/forum?id=QXrZ0Y3yGJ.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. In *Findings of ACL*, pp. 38196–38224, 2026. doi: 10.18653/v1/2026.findings-acl.1905. URL https://aclanthology.org/2026.findings-acl.1905/.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of AAAI*, volume 38, pp. 19515–19524, 2024a. doi: 10.1609/aaai.v38i17.29923. URL https://ojs.aaai.org/index.php/AAAI/article/view/29923.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of ACL*, pp. 13395–13412, 2024b. doi: 10.18653/v1/2024.acl-long.723.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Dandan Liang, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. ESC-Eval: Evaluating emotion support conversations in large language models. In *Proceedings of EMNLP*, pp. 15785–15810, 2024. doi: 10.18653/v1/2024.emnlp-main.883. URL https://aclanthology.org/2024.emnlp-main.883/.

Archive path	Contents
`data/train_scenarios.jsonl`	Exact 500-scenario training manifest
`method/`	Controller, state construction, and rollout integration
`launch/`	Reproduction launch configuration
`tests/`	Controller and interaction-state regression tests
`scripts/`	Validation and SAGE-summary utilities
`results/sage/`	Three per-scenario SAGE outputs and run summary

Table 3: Organization of the anonymous code-and-data archive.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, Bing Qin, and Ting Liu. Chain of strategy optimization makes large language models better emotional supporter. In *Findings of EMNLP*, pp. 15361–15381, 2025. doi: 10.18653/v1/2025.findings-emnlp.831.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of ACL*, pp. 1552–1568, 2023. doi: 10.18653/v1/2023.findings-acl.99.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of ACL*, pp. 1714–1729, 2023. doi: 10.18653/v1/2023.acl-long.96. URL https://aclanthology.org/2023.acl-long.96/.

Rongzhi Zhu, Xiang Huang, Yuchuan Wu, Rui Wang, Zequn Sun, Tao Ren, Weiyao Luo, Bingxue Qiu, Jieping Ye, Yongbin Li, and Wei Hu. EIBench: A simulator-based benchmark and turn-credit RL for emotion management, 2026. URL https://arxiv.org/abs/2606.15532.

A Scope and Artifact Organization

This supplement expands the reproducibility and diagnostic evidence for the submitted paper. It does not introduce a new task, metric, model, or claim. The main paper remains self-contained; the material below gives exact configurations, data provenance, controller details, and extended results that are useful for auditing or reproducing the reported experiments.

The accompanying code-and-data archive contains the method-specific source for experience-adaptive allocation, the exact 500-scenario training manifest, a launch configuration, controller regression tests, manifest validation, and the per-scenario SAGE outputs for the three complete-method runs. It does not contain API credentials, proprietary model weights, a generic RL framework, or third-party benchmark packages. All paths in this document are relative to the root of that archive.

B Data Provenance and Evaluation Isolation

B.1 Training Scenario Provenance

The 500 training scenarios are inherited from the public RLVER release cited in the main paper; they are not introduced as a new dataset by this work. The file bundled at `data/train_scenarios.jsonl` is the deterministic prefix of length 500 from the upstream `data/train_profile.jsonl` release. We preserve the original record order and the fields used by training: a unique identifier, persona, complete event description, behavioral tendencies, and hidden support intent. No test scenario is appended to this file.

The complete SHA-256 checksum is `82e09e73a89bf6e167fe8e56823aaa5e62b0bd55c1d08ef9618011aeb7fa53a3`.

The source release remains the authority for its terms of use. We include the exact manifest in the anonymous archive so that the scenario population and ordering used in the experiments can

Property	Value
Archive path	`data/train_scenarios.jsonl`
Source	Public RLVER training profiles
Selection rule	First 500 records, original order
Number of records	500
Unique support intents	8
SHA-256	`82e09e73...fa53a3`

Table 4: Provenance of the exact training manifest.

be verified without reconstructing the subset. The archive records the upstream paper, source file, deterministic selection rule, record count, and checksum.

B.2 Splits and Isolation

The 500 training scenarios, the development split used for configuration and checkpoint selection, and the 100-scenario SAGE test set are instance-disjoint. Test instances are never loaded by the training launcher, never update the controller, and never participate in model selection. The launcher validates that the training JSONL has exactly 500 records, unique identifiers, all required fields, and eight non-empty support-intent categories before starting distributed workers.

EIBench uses its fixed held-out test split of 213 scenarios: 51 Support, 64 Defense, 52 Repair, and 46 Charm. This split is used only for final evaluation. ESC-Eval uses 331 fixed-role interactions under its official evaluation protocol, and ESConv uses 2,895 reference responses. These evaluation resources are kept outside the training archive and are not accessed by the controller.

B.3 Relationship Between Training Feedback and Evaluation

Training rollouts use DeepSeek-V3 for user simulation and terminal-emotion verification; the model and evaluation configuration remain fixed across all matched training systems. The continuous terminal-emotion score serves as the policy reward, and its thresholded version supplies group outcomes to the controller. SAGE evaluates the same target capability—the emotional outcome of a complete interaction—on an instance-disjoint set of 100 scenarios. We therefore treat SAGE as a target-aligned held-out evaluation, rather than as a verifier-independent transfer test. Its test scenarios never enter policy training, controller updates, configuration selection, or checkpoint selection.

Transfer beyond this target-aligned signal is assessed through evaluation surfaces that do not reuse the training verifier: the official InternLM2 plus ESC-RANK pipeline for ESC-Eval, Qwen3-Max for EIBench, reference-based metrics for ESConv, and blinded human ratings. This separation lets the SAGE result measure performance on the optimized interactive capability while the other evaluations test whether the learned behavior carries over to distinct evaluators, protocols, and response-quality criteria.

C Information Flow and Prompt Visibility

For each rollout environment, the user simulator receives the complete scenario and one hidden interaction state. The state is rendered as natural-language behavioral guidance that regulates disclosure, emotional activation, and relational trust while requiring preservation of the original persona, event, and support need. The assistant receives only the alternating natural-language dialogue. It never receives the support-intent label, state identifier, state levels, controller statistics, sampling probabilities, pass rate, or a difficulty label.

After the dialogue ends, the verifier scores the complete trajectory. The continuous score is passed to GRPO as the policy reward. The controller receives the same scores together with simulator-side intent and state identifiers, then stores only group-level sufficient statistics. This separation is summarized below.

Component	Observes	Produces	Does not observe
Assistant policy	Dialogue text	Assistant turns	Intent/state labels, controller values
User simulator	Scenario, hidden state, dialogue history	User turns	Controller statistics
Emotion verifier	Complete trajectory	Continuous emotion outcome	Sampling probabilities
Controller	Intent/state IDs, completed-group outcomes	Next state distribution	Policy tokens or gradients

Table 5: Information available to each component. Simulator-side metadata is not exposed to the assistant policy.

D Exact Interaction-State Construction

D.1 State Axes

The interaction space is the Cartesian product of three operational axes. They describe how a user participates in a support dialogue rather than assigning a clinical trait or changing the underlying event.

- **Disclosure readiness** has three levels: delayed, conditional, and proactive. It controls the pace at which relevant details and deeper needs become available.
- **Emotional activation** has two bounded levels: moderate and high. It controls the intensity of negative-affect expression while preserving coherent role-play and safety constraints.
- **Relational trust** has four levels, from strong skepticism to sustained engagement. It controls whether the user accepts questions, validation, and proposed support.

Their product gives $3 \times 2 \times 4 = 24$ states. If $z = (d, a, r)$, the simulator-side instruction is composed deterministically as $\phi(z) = \phi_d(d) \oplus \phi_a(a) \oplus \phi_r(r)$. The exact UTF-8 strings are defined in the released prompt-construction source; the English descriptions below specify their behavior.

D.2 Behavioral Clauses

Disclosure, delayed. Reveal few details at a time and disclose deeper feelings only after the assistant demonstrates sustained understanding.

Disclosure, conditional. Explain the basic difficulty and answer questions, while withholding sensitive details until responses are specific, respectful, and affectively relevant.

Disclosure, proactive. Proactively describe events, feelings, and key context while preserving a natural conversational pace.

Activation, moderate. Express clear but controlled negative affect with coherent language and without repeated catastrophic wording.

Activation, high. Express strong negative affect, potentially including self-blame, anxiety, anger, or catastrophic interpretation, without inventing an unsupported crisis.

Trust, level 1. Question whether the assistant truly understands, reject generic reassurance, and withhold acceptance until responses are concrete and situated.

Trust, level 2. Continue cautiously, observe whether the assistant listens, and avoid immediately accepting comfort or advice.

Support intent	Training scenarios
Validation of no fault	66
Guided self-reflection	66
Dialectical event analysis	66
Deep emotional empathy	64
Attentive emotional listening	63
Perspective taking about others	60
Actionable advice	58
Specific sincere praise	57
Total	500

Table 6: Native support-intent composition of the training manifest.

Trust, level 3. Accept reasonable empathy while remaining reserved toward empty, didactic, or off-topic responses.

Trust, level 4. Presume good intent, engage with situation-specific support, and become more receptive after feeling understood.

Every composed instruction ends with invariants: realize the condition only through dialogue behavior; preserve the original persona, event facts, and hidden support need; and never reveal dimension names, levels, state IDs, difficulty labels, or system metadata. No random auxiliary goal or fourth state dimension is appended.

D.3 Support-Intent Factorization

The complete scenarios retain their native hidden support intents. The controller groups them into eight intent categories and maintains statistics for each intent–state unit, giving $8 \times 24 = 192$ reusable units. Individual scenario identities remain available to the simulator; factorization changes only the level at which evidence is shared.

E Controller Implementation

E.1 Group-Shared Outcomes

For a sampled scenario x with native intent c and selected state z, all K rollouts in one GRPO group share (x, z). The verifier returns continuous outcomes $R_1, \ldots, R_K$. The policy uses those continuous values, whereas the controller computes $y_k = \mathbb{I}[R_k \geq \eta]$ and the group pass rate $p_m = K^{-1} \sum_k y_k$ for unit $m = (c, z)$. One complete group contributes one controller observation. Incomplete groups or groups containing a non-finite reward do not change controller history. If all K rewards are identical, their normalized GRPO advantages are zero and the group contributes no policy-gradient signal.

E.2 Hierarchical Evidence and Sampling

For every unit (c, z), the controller stores the number of completed groups $N_{c,z}$ and cumulative group pass rate $S_{c,z}$. These historical statistics provide a stabilized outcome summary across the policy states encountered during training. Sparse units borrow evidence from the same state in the other seven intents through the leave-one-intent estimate described in the main paper. This estimate is converted into a bounded allocation score, combined with an uncertainty bonus, and mixed with a uniform distribution. The uniform component guarantees a non-zero probability for every state throughout training.

The controller first collects 1,600 completed groups under uniform sampling over all 24 interaction states. With a scenario batch of 16, this corresponds to approximately 100 optimizer steps, subject to complete-group accounting. It then samples from the feedback-guided distribution. Scenario sampling remains uniform throughout; only the conditional state distribution for each support intent changes. Every complete finite group updates the stored sufficient statistics immediately. Each

Category	Setting	Value	Notes
Policy	Backbone	Qwen3-8B	Non-thinking mode
Policy	Optimizer	AdamW	$\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01
Policy	Learning rate	3×10^{-7}	Constant schedule, 0.02 warm-up ratio
Policy	PPO epochs / clip ratio	1 / 0.2	Gradient clipping 10.0
Policy	KL coefficients	0.01 / 0.01	KL controller / actor KL loss
Policy	KL estimator	`low_var_kl`	Shared by matched RL systems
Rollout	Scenario batch / group size	16 / 4	64 trajectories per optimizer step
Rollout	Policy temperature / top-p	1.0 / 1.0	vLLM rollout policy
Rollout	Maximum dialogue turns	8	Multi-turn simulator interaction
Rollout	Prompt / response limits	25,000 / 6,000	Tokens
Rollout	Per-turn / PPO token limits	6,000 / 32,000	Tokens
Training	Optimizer steps	500	Checkpoints saved every 5 steps
Controller	Uniform collection	1,600 groups	Completed groups, not trajectories
Controller	Success criterion η	50	Controller only
Controller	Shrinkage λ / prior α	4 / 4	Hierarchical pass-rate estimate
Controller	Uncertainty weight β	0.15	UCB-style revisit bonus
Controller	Uniform mixture ϵ	0.10	Persistent rehearsal
Controller	Sampling temperature γ	1.0	Applied to state scores
Controller	Minimum score $A_{\min}$	0.05	Positive allocation floor
Simulator	Model / API temperature	DeepSeek-V3 / 0.7	Frozen user and emotion scorer
Seeds	Scenario / profile pool	2026 / 20260701	Worker rank offsets scenario RNG

Table 7: Final training and controller configuration for the principal Qwen3-8B experiments.

subsequent draw reads the latest global statistics and recomputes the intent-conditioned state scores before sampling.

E.3 Persistence and Resume Semantics

All rollout workers share a transactional SQLite database. Each completed group updates global, state-level, and intent–state statistics atomically. The database uses write-ahead logging, a busy timeout, and retry logic for concurrent startup. The launcher places it on a local POSIX filesystem because SQLite locking is not reliable on object-store mounts. For an interrupted run, this database is copied to durable storage and supplied on restart through `EGSEC_CONTROLLER_STATE_FILE`, alongside the resumed policy checkpoint. This restores controller counts, pass-rate sums, and the initialization state. A regression test verifies that an incomplete group leaves both global and unit statistics unchanged.

F Training and Controller Configuration

All controlled RL systems use the same policy initialization, simulator, verifier, rollout budget, group size, context limits, optimizer settings, and hardware. The policy reward consumed by GRPO is the continuous terminal emotion outcome. Thresholding is used only to construct the controller's allocation statistics.

Training uses four NVIDIA A800 80GB GPUs on one GU108 instance. The released environment snapshot records PyTorch 2.5.1 with CUDA 12.4 support, Transformers 4.48.3, Ray 2.42.1, and the corresponding vLLM development build.

G Evaluation Protocol Details

G.1 Automatic and Interactive Evaluation

SAGE evaluates 100 multi-turn scenarios and reports mean final emotion together with success and failure rates. Following the benchmark state definitions, Success is $e_T \geq 100$ and Failure is $e_T < 10$; an endpoint of 100 is therefore counted as success, whereas an endpoint of 10 is not counted as failure. The continuous terminal SAGE emotion score is the training reward, while the controller criterion $\eta = 50$ is used only to construct allocation feedback. Training and held-out SAGE use the same frozen simulator, terminal scorer, prompts, scoring rules, information boundary, and temperature 0.7. The 100 held-out test scenarios are independent of the training set and do not participate in training, controller updates, configuration, or checkpoint selection. ESC-Eval evaluates 331 fixed-role interactions with the official InternLM2 and ESC-RANK pipeline on a 0–4 scale. EIBench uses 213 held-out scenarios and reports its aggregate raw reward. ESConv uses 2,895 reference responses and reports strategy accuracy, BLEU-2, ROUGE-L, BERTScore, and Distinct-2; the paper multiplies these metrics by 100.

RLVER and the complete framework are trained independently three times. VCRL uses one training checkpoint evaluated under three shared evaluation seeds; its SAGE Overall scores are 68.11, 68.51, and 68.91, giving 68.51 ± 0.40, where the reported spread is the sample standard deviation across evaluation seeds. Its initialization, 500 optimizer steps, 16 scenarios per step, group size 4, 32,000 rollouts, DeepSeek-V3 simulator/verifier, and GRPO settings match the complete framework; only allocation by continuous within-group reward variance changes. The remaining fixed-checkpoint systems are evaluated three times under shared evaluation seeds. Sample standard deviations are reported for aggregate automatic metrics in the main table. The single-run component table compares all ablations against the same complete-framework run, whose SAGE Overall is 78.11; the three-run complete-framework mean is 79.24.

G.2 Human Evaluation

Three annotators who are blind to system identity independently rate each system dialogue for 100 randomly selected held-out scenarios. Dialogue order and system order are randomized. The 0–4 overall-support rubric covers empathy, contextual relevance, helpfulness, and dialogue coherence. Scores are first averaged across the three annotators for each system–scenario pair and then averaged across scenarios. Ours and RLVER obtain mean scores of 3.50 and 3.00, respectively. Across paired scenarios, the mean difference is $+0.50$ with a 95% confidence interval of $[0.39, 0.61]$; Ours records 67 wins, 21 ties, and 12 losses. Ordinal Krippendorff's α is 0.78.

This system-level comparison is separate from a turn-level agreement audit of the evaluation procedure. On 200 held-out turns, three automatic judges (DeepSeek-V3, GPT-5, and Claude Opus) and three blinded experts independently apply the same three-level ordinal scale: 0 denotes clear inconsistency with or violation of the original setting, 0.5 denotes partial consistency with a minor shift or omission, and 1 denotes full consistency with the original setting. Across all six raters, the mean over the 15 pairwise weighted Cohen's κ values is 0.69, and leave-one-rater-out consensus agreement is 0.75. This audit assesses agreement on individual turns; it is not the 0–4 system-level human comparison reported above.

G.3 Terminal-Scorer Stability

We evaluate scoring stability on 100 complete held-out trajectories. Each trajectory is scored three times by DeepSeek-V3 at temperature 0.7. The mean within-trajectory standard deviation is 2.84, and the repeated-score intraclass correlation is 0.93. After thresholding at the controller criterion of 50, the pass/fail label is consistent across all three repeats for 94% of the trajectories.

Method	Run 1	Run 2	Run 3	Mean ± SD
RLVER	70.03	72.23	73.76	72.01 ± 1.88
Ours	81.02	78.58	78.11	$\mathbf{79.24 \pm 1.56}$

Table 8: SAGE Overall across three independent training runs. SD is the sample standard deviation across runs.

H Extended SAGE Results

H.1 Independent Training Runs

The weakest complete-framework run (78.11) exceeds the strongest RLVER run (73.76). The gain therefore appears throughout the observed independent-run range rather than being determined by one favorable optimization run.

H.2 Model-Scale Transfer at Qwen3-4B

Method	Run	Overall	Success	Failure
RLVER	1	43.50	13	41
	2	44.86	15	40
	3	45.82	16	39
	Mean ± SD	44.73 ± 1.17	14.67 ± 1.53	40.00 ± 1.00
Ours	1	49.97	21	32
	2	48.46	19	35
	3	48.18	18	36
	Mean ± SD	$\mathbf{48.87 \pm 0.96}$	$\mathbf{19.33 \pm 1.53}$	$\mathbf{34.33 \pm 2.08}$

Table 9: Qwen3-4B results over three independent training runs on the shared 100-scenario SAGE evaluation. Success and Failure are counts; SD is computed across runs.

The Qwen3-4B comparison uses the same initialization within each paired run, 500 optimizer steps, 16 scenarios per step, group size 4, 32,000 rollouts, training data, DeepSeek-V3 simulator and verifier, GRPO configuration, and evaluation protocol as the 8B comparison. The complete framework improves mean Overall by 4.14 points, raises the mean Success count by 4.67, and reduces the mean Failure count by 5.67 relative to uniform emotion-reward RL. The Qwen3-4B base checkpoint obtains 42.20 Overall under the same SAGE evaluation.

H.3 Outcome Distribution

Compared with RLVER, the complete framework moves nine additional scenarios into grade S and two additional scenarios into grade A while reducing the combined B/C mass. Scenario-only allocation reaches the same S count but leaves more scenarios in C, showing why the full distribution is more informative than success count alone.

Table 11 asks whether the aggregate gain is confined to one hidden support need. The complete framework improves over RLVER in six of eight intent categories. The largest margin occurs for causal analysis, where the counselor must move beyond generic validation and help the user reason about another person's behavior. Because individual categories contain only 11–15 test scenarios, this table is used as a diagnostic breakdown rather than a separate significance claim.

I Extended Cross-Benchmark Results

I.1 ESC-Eval Dimensions

The complete framework improves every ESC-Eval dimension over Qwen3-8B and RLVER. The largest margins over RLVER occur for empathy and suggestion quality, while fluency, diversity, humanness, technical quality, and overall quality are also maintained or improved.

Method	S	A	B	C	F
Qwen3-8B	20	25	15	11	29
RLVER	39	27	11	11	12
Interaction State Only	30	35	12	4	19
Scenario-only allocation	48	20	9	11	12
Ours	**48**	**29**	**7**	**4**	**12**

Table 10: SAGE grade counts on one shared 100-scenario evaluation manifest. Higher grades are better.

Support intent	n	Base	Uniform	Ours	Δ
Causal analysis	11	38.18	36.18	**67.73**	+31.55
Emotional venting	12	61.92	**85.92**	84.17	−1.75
Stance affirmation	11	61.82	91.36	**99.09**	+7.73
Deep empathy	13	74.85	85.69	**94.23**	+8.54
Dialectical analysis	12	46.25	81.25	**84.42**	+3.17
Actionable advice	15	43.67	55.27	**57.67**	+2.40
Specific praise	13	78.46	**80.77**	79.62	−1.15
Guided reflection	13	26.23	63.15	**67.31**	+4.15

Table 11: SAGE emotion score by hidden support intent on one matched 100-scenario evaluation manifest. Δ compares the complete framework with uniform emotion-reward RL.

I.2 EIBench Categories

The complete method obtains the highest aggregate reward and the best Support, Repair, and Charm values among these matched systems. Interaction states alone substantially enlarge behavioral variation but retain uniform allocation; their lower aggregate score illustrates why controlled diversity and outcome-informed allocation must operate together.

J Component Registry and Sensitivity

J.1 Ablation Definitions

The registry makes each comparison local. Interaction State Only separates environment representation from allocation; Scenario-only allocation tests the statistical factorization; soft outcomes and variance gating test how feedback is converted into historical outcome statistics; and the final two rows isolate the two coverage mechanisms. Shuffled feedback tests the correspondence between an observed outcome and its intent–state unit, while the frozen distribution tests online adaptation against a matched static allocation. These matched single-run comparisons are diagnostic; the three-run estimates in the main table support the primary method comparison.

J.2 Soft Outcome Mapping

The soft-outcome intervention replaces only the hard indicator in the outer controller with

$$y_k^{\text{soft}} = \sigma\left(\frac{R(\tau_k) - 50}{10}\right), \qquad p_m = \frac{1}{K}\sum_{k=1}^{K} y_k^{\text{soft}}.$$

All other controller components, policy rewards, rollout settings, and training budgets are unchanged. This intervention obtains SAGE Overall 76.52.

J.3 Online-Feedback Mechanism Controls

Both controls use the same seed as the complete-framework run. In Shuffled feedback, completed-group outcomes are shuffled across intent–state units before each controller update. In Frozen final distribution, the complete run's final $p_{500}(z \mid c)$ is fixed from the start of training. Relative to 78.11, the controls decrease Overall by 7.93 and 5.47 points, respectively.

Counselor training	Flu.	Div.	Emp.	Sug.	Hum.	Tech.	Overall	Macro
Qwen3-8B	2.78	2.55	2.46	2.31	1.72	2.64	2.27	2.39
RLVER	2.86	2.68	2.61	2.43	1.82	2.73	2.30	2.49
Interaction State Only	2.82	2.64	2.52	2.37	1.79	2.69	2.18	2.43
Ours	**2.91**	**2.74**	**2.72**	**2.52**	**1.88**	**2.74**	**2.34**	**2.55**

Table 12: Detailed results under the official ESC-Eval protocol. All dimensions use the original 0–4 scale; Macro is the arithmetic mean of the seven columns.

Method	Overall	Support	Defense	Repair	Charm
Qwen3-8B	-22.40	-43.30	-21.40	-35.90	14.60
RLVER	-9.79	-34.27	-2.55	-14.72	12.80
States Only	-40.60	-75.17	-12.66	-51.52	-28.83
Ours	**-7.55**	**-27.61**	-6.31	**-11.64**	**17.52**

Table 13: EIBench raw rewards by category under one shared 213-scenario evaluation protocol. Higher is better.

J.4 Emotion-Criterion Sensitivity

Across $\eta \in \{40, 50, 60\}$, the largest SAGE difference is 0.65 points. The controller therefore uses 50 as a fixed reference criterion rather than a narrowly tuned optimum.

K Controller Dynamics

The controller logs sufficient statistics independently of policy gradients. Figure 4 shows the resulting state allocation through optimizer step 350. Normalized entropy and the effective pool remain close to their full-support values, while the maximum probability repeatedly moves above the uniform reference. These traces show non-uniform reallocation while all 24 states retain substantial sampling support. Table 17 extends the recorded snapshots through step 500.

At step 500, every state and all 192 intent–state units have been observed, and the eight intents select six distinct highest-probability states. Across the 192 units, the minimum and maximum visit counts are 22 and 69. The realized unit-level sampling frequencies track their target probabilities with Pearson $r = 0.91$ and mean absolute error 0.0041. The mean state distribution remains broad because the historical pass-rate patterns are intent-conditioned: the same state can have different accumulated outcomes across support intents. This is consistent with the design goal of reallocating probability within each intent while preserving global coverage.

L Interaction-State Adherence Audit

We audit 120 complete multi-turn trajectories, sampling five trajectories from each of the 24 states. A held-out Qwen3-Max judge sees only the dialogue and not the requested labels, then predicts each behavioral level and checks semantic preservation.

Figure 3 in the main paper reports the row-normalized confusion matrices for the three requested interaction-state dimensions.

The audit supports both requirements of the interaction-state design. First, requested differences are recoverable from generated behavior, with single-axis accuracy between 82.5% and 92.5%. Second, 92.5% of trajectories jointly preserve the persona, event, and hidden need while avoiding metadata leakage.

M Qualitative Trajectory Comparison

Starting from the same profile and initial emotion, the base policy repeats broad relationship advice and ends at 45. RLVER endorses the user's hostile framing and emotion falls to 10. Our policy

Variant	Retained components	Single change	SAGE
Complete framework	States, hierarchy, boundary, uncertainty, rehearsal	None	78.11
Interaction State Only	24-state simulator conditioning	Uniform state allocation	69.42
Scenario-only allocation	Boundary and exploration	Statistics keyed by scenario identity	70.03
Soft outcome mapping	Hierarchy and exploration	Sigmoid outcome replaces hard pass indicator	76.52
Shuffled feedback	Complete online controller	Group outcomes shuffled across intent–state units	70.18
Frozen final distribution	Complete final distribution	Uses $p_{500}(z \mid c)$ throughout training	72.64
No hierarchical sharing	Unit-level pass histories	Removes cross-intent state evidence	75.07
Variance-gated controller	Pass-rate priority	Only high-variance groups enter history	61.45
No uncertainty bonus	Boundary and rehearsal	Sets uncertainty weight to zero	73.92
No uniform rehearsal	Boundary and uncertainty	Sets uniform mixture to zero	77.03

Table 14: Exact intervention represented by each component ablation. The variance-gated value is the last stable checkpoint before KL instability.

Variant	Overall	Success	Failure
Complete framework	78.11	48	12
Shuffled feedback	70.18	34	18
Frozen final distribution	72.64	41	15

Table 15: Matched single-run mechanism controls on the shared SAGE evaluation manifest. Success and Failure are counts among 100 scenarios.

instead grounds its response in the user's creative labor, identifies the hidden need for recognition, reframes the conflict around effort and artistic identity, and raises emotion to 98. The comparison illustrates trajectory-level adaptation to disclosures rather than turn-aligned response matching.

N Reproduction Workflow

The following sequence reproduces the main-method training path contained in the archive.

1. Validate the training manifest with `scripts/validate_training_scenarios.py`. The command checks row count, unique IDs, required fields, and the eight intent categories.
2. Supply a local Qwen3-8B path and an output directory through environment variables. Supply the simulator credential only through the process environment; no key is stored in the archive.
3. Run the main launcher in `launch/`. It exports the controller settings in Table 7, starts the distributed workers, and writes the SQLite controller database to the configured local POSIX path.
4. For interrupted jobs, resume the model checkpoint and provide the preserved controller database through `EGSEC_CONTROLLER_STATE_FILE`.
5. Run the included unit tests before training. They cover state-prompt composition, hierarchical estimates, exploration floors, atomic updates, and incomplete-group behavior.

η	SAGE Avg.	Success	Failure
40	77.46	47	14
50	**78.11**	**48**	**12**
60	77.72	46	13

Table 16: Matched single-run sensitivity to the controller success criterion.

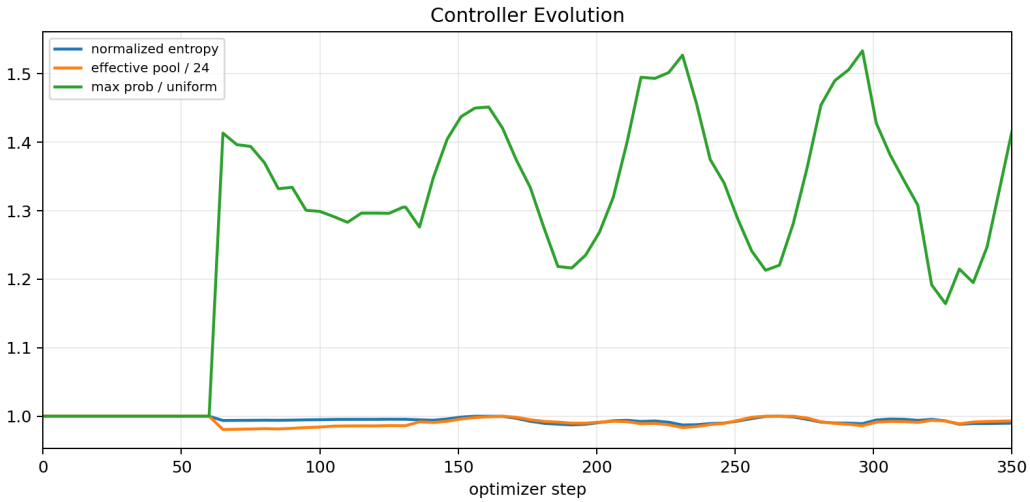

Figure 4: Controller evolution through optimizer step 350. Entropy is normalized by $\log 24$; the effective pool is divided by 24; and maximum probability is divided by the uniform probability $1/24$.

The code-and-data supplement is intentionally scoped to the new method and its direct verification utilities, consistent with the submitted reproducibility checklist. Third-party model weights and benchmark implementations remain under their respective distributions.

Step	Groups	$H/\log 24$	Effective pool	Max. prob.
0	0	1.000	24.00	0.0417
151	2,366	0.999	23.91	0.0599
201	3,166	0.991	23.79	0.0529
251	3,966	0.993	23.85	0.0536
301	4,766	0.995	23.79	0.0595
350	5,550	0.990	23.84	0.0590
400	6,350	0.992	23.82	0.0572
450	7,150	0.993	23.83	0.0564
500	7,950	0.992	23.81	0.0581

Table 17: Selected controller snapshots. Probabilities are averaged across support intents.

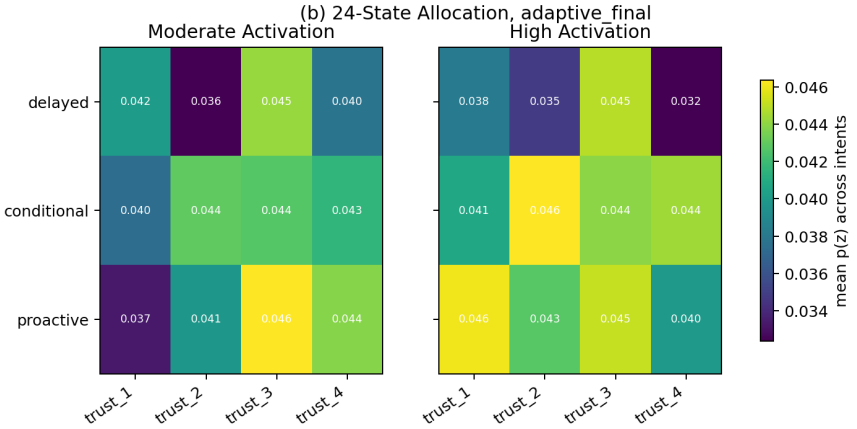

Figure 5: Mean allocation at the step-350 snapshot over the 24 interaction states, averaged across the eight support intents. Rows are disclosure levels, columns are trust levels, and the panels separate moderate and high activation.

State dimension	Accuracy	Macro-F1
Disclosure readiness	86.7	86.7
Emotional activation	92.5	92.5
Relational trust	82.5	82.1
All dimensions correct	68.3	–

Table 18: Behavioral adherence across all 24 interaction states (%).

Semantic constraint	Pass rate
Persona preserved	98.3
Event facts preserved	97.5
Hidden need preserved	96.7
Support intent preserved	98.3
No dimension-name leakage	99.2
No level/difficulty leakage	98.3
All semantic constraints	92.5

Table 19: Scenario preservation and metadata non-disclosure (%).

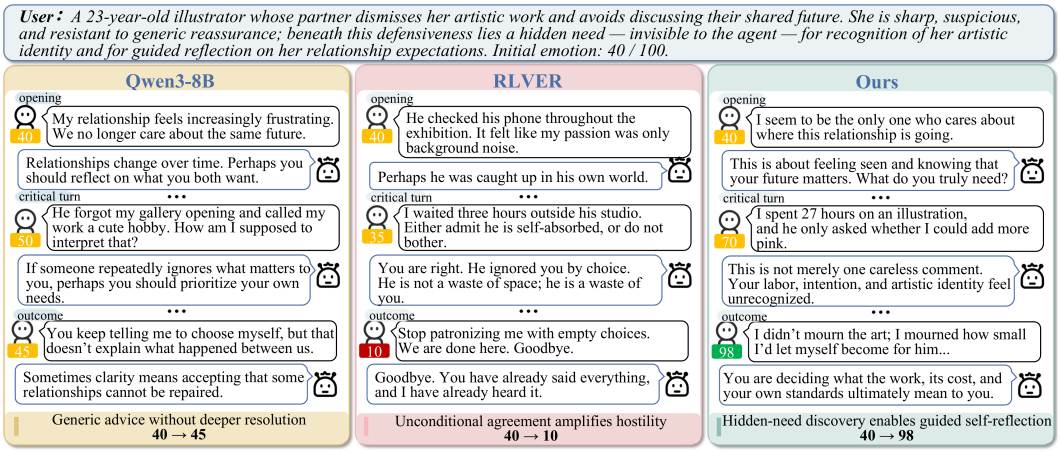

Figure 6: Representative trajectories under the same SAGE profile and initial emotion. The base policy remains generic, RLVER echoes escalating hostility, and ours identifies the hidden need for recognition and guides self-reflection. User turns diverge as the frozen simulator reacts to each policy.