CURE: CONSTRUCTING USER-SPECIFIC RUBRICS TO EVALUATE MULTI-TURN ROLE-PLAYING DIALOGUE WITH AGENT JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Role-playing dialogue is an important human–AI application across virtual companionship, narrative play, and social interaction, where users increasingly engage with characters in long-running conversations spanning many turns. These long interactions provide contexts that reveal users' evolving goals, preferences, and boundaries. They impose greater demands on role-playing models and evaluation methods: models must maintain character consistency while adapting over time, and evaluators must assess interactions against user-specific requirements. However, existing evaluation still centers largely on isolated responses or fixed single-turn probes. We therefore aim to build a multi-turn evaluation framework for role-playing dialogue. We first observe that multi-turn role-playing evaluation lacks a dedicated benchmark dataset. Using an emotional BDI mechanism, we construct and release **EmoRoleBench**, a benchmark of 1,751 stateful dialogues across 22 characters, 16 user profiles, and five role-playing models. Each dialogue has an independently generated post-session satisfaction label that is hidden from the evaluator. We then introduce **CURE**, a construct-first Agent-Judge framework that derives an explicit user-specific rubric from preferences and boundaries expressed across turns, corrects it with persona-grounded anchors, and gathers turn-level and temporal evidence before scoring. On the predefined test split, CURE achieves a Spearman correlation of 0.678 with simulated-user satisfaction, compared with 0.533 for the strongest external baseline. It also performs better in matched dialogue comparisons, remains stable across three Judge backbones, and shows positive alignment with persona-conditioned human ratings. These results show that CURE more effectively captures user preferences that emerge across turns and translates them into user-aligned evaluations of role-playing dialogue. Code and data are available at `https://anonymous.4open.science/r/cure-review-artifact-2026-1E3C/`.

1 INTRODUCTION

Role-playing dialogue has become an important human–AI interaction setting, spanning virtual companionship, narrative play, and casual conversation. Its goal is not merely to produce an individually correct response, but to sustain an immersive, personalized, and evolving interaction with an AI-mediated character. Unlike information seeking or task execution, where factual accuracy or goal completion often provides a natural evaluation target, role-playing quality also depends on whether responses remain natural, engaging, faithful to the character, and appropriate for the particular user over time.

This temporal dimension makes multi-turn evaluation indispensable. Research on human communication treats language use as a joint activity and conversation as a sequentially organized process Clark (1996); Sacks et al. (1974). Because each turn responds to and reshapes the preceding interaction, users experience a dialogue as an evolving trajectory rather than as a collection of isolated responses. A model may perform convincingly in the opening turns yet later exhibit persona drift, topic stagnation, facet exhaustion, or persistent passivity. Such late-stage failures can leave the user dissatisfied with the interaction as a whole, even when the model initially performs well. Single-turn or short-horizon evaluation may therefore overlook consequential failures and overes-

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.

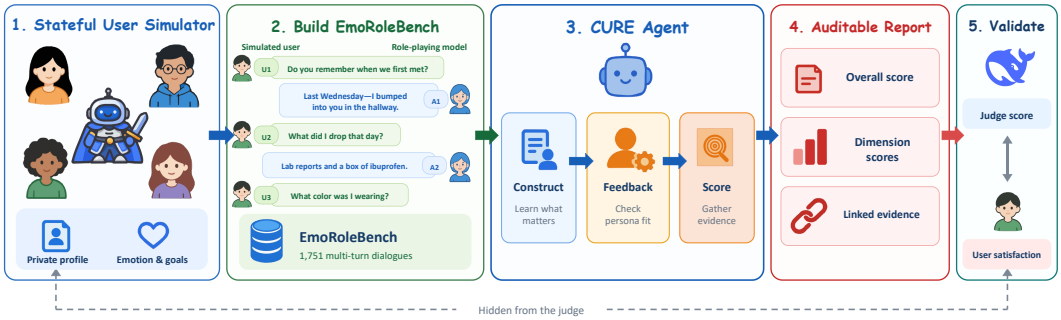


Figure 1: **EmoRoleBench and CURE overview.** A stateful simulator generates multi-turn role-playing dialogues, while CURE builds user-specific rubrics and produces evidence-linked scores. Private satisfaction is hidden from the evaluator and used only for validation.

timate interaction quality. Evaluating user satisfaction requires assessing the complete multi-turn trajectory.

Existing role-playing evaluation commonly relies on fixed question sets, single-turn probes, or static judges (Zhou et al., 2025; Samuel et al., 2025; Tu et al., 2024). Although reproducible, these protocols cannot cover the branching trajectories induced by users with different preferences, tolerance, and interaction goals. Uniform dimensions, thresholds, and weights also conflate a model's general role-playing capability with its fit to a particular user. Direct and structured LLM judges can condition on the dialogue, but their effective rubric usually remains implicit: the judge decides what matters while simultaneously deciding how well the model performed (Liu et al., 2023; Zheng et al., 2023; Chan et al., 2024). This entanglement obscures the basis of the final score and permits incidental dialogue details to become post hoc criteria. Agent-as-a-Judge methods improve evidence acquisition through search and tool use, but typically assume externally specified and comparatively objective task requirements (Zhuge et al., 2025; Shi et al., 2026b). Existing protocols are therefore insufficient for evaluating subjective, multi-turn role-playing interactions.

Such evaluation poses two coupled challenges. First, the judge must track character identity, language style, knowledge boundaries, emotional responses, and interaction rhythm throughout a long and potentially branching trajectory. Second, it must determine what counts as quality for the current user. Technically consistent behavior does not guarantee a good experience for everyone: the same proactive response may engage one user but intrude on the boundaries of another. A useful judge must therefore interpret long-horizon evidence against user-specific interaction preferences. For subjective multi-turn dialogue, evidence acquisition alone is insufficient; the evaluation target must be made explicit before model behavior is scored. We thus frame personalized multi-turn dialogue evaluation as a *criterion-construction problem*, rather than only a scoring problem. Figure 1 illustrates this motivation and the resulting construct-first evaluation framework.

We introduce **CURE**, a construct-first Agent-Judge framework that constructs user rubrics before assessing model behavior. Given a completed multi-turn dialogue and its public persona card, CURE operates in three stages. In CONSTRUCT, it infers interaction preferences from observable user utterances and converts them into an explicit dialogue-specific rubric, grounding every personalized criterion in quoted user evidence. In FEEDBACK, persona-derived anchors audit and correct the draft rubric, preserving character-level requirements that local preferences should not override and reducing overfitting to incidental signals. In SCORE, an agent selects tools to gather turn-level and temporal evidence, evaluates the assistant against the corrected rubric, and produces an auditable path from user evidence and rubric construction to the final decision. The evaluation target is thus fixed before scoring, while the agent remains free to determine how relevant evidence should be located and organized.

We evaluate CURE on **EmoRoleBench**, a benchmark of 1,751 simulator-generated multi-turn role-playing dialogues produced by an emotional BDI user simulator. The simulator instantiates users with different preferences, tolerance, interaction goals, and evolving internal states, and conducts conversations with the models under evaluation. Its process states and post-session satisfaction label S are withheld from CURE and joined with its capability scores only after evaluation.

Experiments show that CURE aligns more closely with simulated-user satisfaction than the evaluated baselines and more often identifies the better dialogue under matched user requirements. Its scores and model-level rankings also remain highly consistent across three Judge backbones. Persona-conditioned human ratings of the simulated user's satisfaction further support S as a scalable experience proxy and show positive alignment between CURE's scores and human satisfaction judgments.

In summary, our contributions are:

- We formulate subjective multi-turn role-playing evaluation as a criterion-construction problem and introduce **CURE**, which grounds user-specific criteria in dialogue evidence, corrects them with persona-derived constraints, and produces an auditable rubric–evidence–decision path through agentic tool use.
- We develop an emotional BDI user simulation framework and use it to construct **EmoRoleBench**, a benchmark of 1,751 stateful, preference-diverse multi-turn role-playing dialogues with user profiles, process trajectories, explicit termination behavior, and separately generated satisfaction labels.
- We show that CURE aligns more closely with simulated user satisfaction than fixed-rubric, direct, and structured judge baselines, while maintaining highly consistent scores and model-level rankings across three Judge backbones; persona-conditioned human ratings further support the simulated satisfaction signal and CURE's human alignment.

2 Related Work

2.1 Multi-turn Role-Playing Evaluation

Many role-playing benchmarks evaluate whether a model can reproduce prescribed character attributes under controlled prompts. CharacterBench and PersonaGym use targeted questions to elicit and score character fidelity, but isolated responses provide limited evidence about how an interaction develops over time (Zhou et al., 2025; Samuel et al., 2025). Multi-turn benchmarks broaden this scope to dialogue-level consistency, sociality, topic management, user intention, and persona simulation, typically using fixed user controllers and predefined rubrics (Tu et al., 2024; Chen et al., 2024; Wu et al., 2025; Xiang et al., 2025; Shi et al., 2026a). Recent simulation frameworks further improve the diversity and realism of evaluation trajectories. SAGE uses knowledge-grounded users to expose agent failures; OmniBehavior and SimulatorArena analyze limitations of existing simulators and emphasize external validation; and EYT-Bench, PALATE, and CompanionBench separate simulators, target models, and judges or ground scenarios in real data (Shea et al., 2026; Chen et al., 2026; Dou et al., 2025; Gong et al., 2026; Zhu et al., 2026; Liu et al., 2026). These advances make multi-turn evaluation more realistic and scalable, but its criteria remain largely predefined or implicit. Consequently, existing protocols assess models along shared dimensions without first determining which interaction qualities matter to the particular user represented in each trajectory.

2.2 Agent Judges for Dialogue Evaluation

LLM-as-a-Judge provides a scalable alternative to human evaluation for open-ended dialogue. G-Eval and MT-Bench established direct, reference-free judging while documenting position, verbosity, and self-preference biases (Liu et al., 2023; Zheng et al., 2023). ChatEval and trained judges such as CharacterJudge, RPCAJudger, and MTDEval improve stability or granularity, but continue to evaluate along prescribed reasoning or scoring paths (Chan et al., 2024; Zhou et al., 2025; Wu et al., 2025; Tang et al., 2026). Agent-as-a-Judge instead decomposes evaluation into evidence collection and per-requirement verification, with tool-assisted evidence acquisition accounting for much of its improvement over direct LLM judgment (Zhuge et al., 2025). AJ-Bench further separates information acquisition, state verification, and process verification, and traces errors to tool-output interpretation and evidence-based reasoning (Shi et al., 2026b). TRACE Bench introduces agentic checklists into role-playing evaluation, but its checklist is privately updated by the same agent that drives the interaction (Zhang et al., 2026a).

Existing Agent Judges thus improve how evidence is acquired and verified, but generally assume that what should be verified has already been specified. This assumption does not hold for sub-

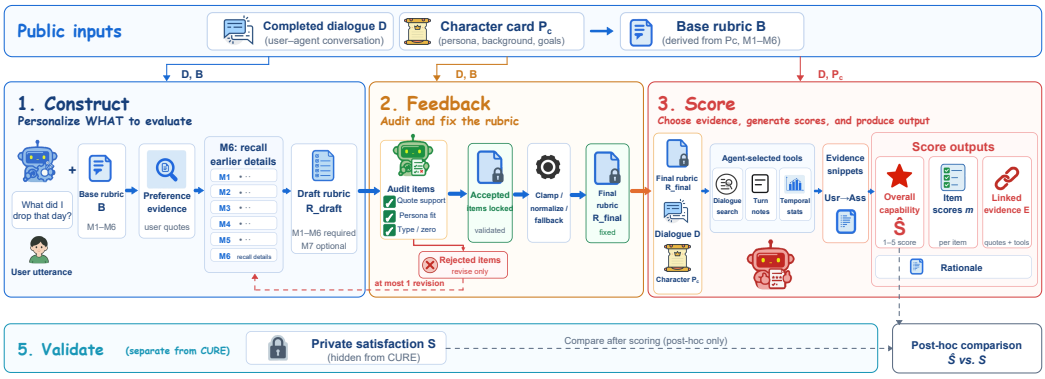


Figure 2: **CURE framework.** Construct personalizes the rubric, Feedback audits and fixes it, and Score gathers evidence against the locked rubric. Private satisfaction is used only for post-hoc validation.

jective multi-turn role-playing dialogue, where relevant criteria depend on the current user's expressed preferences while remaining bounded by character-level requirements. Our construct-first personalized-rubric Agent Judge addresses this missing step: it derives explicit criteria from user evidence, corrects them with persona anchors, and only then uses agentic tools to collect evidence and score the dialogue. The central contribution is therefore not merely a stronger scoring agent, but an explicit separation between constructing a personalized evaluation target and verifying model behavior against it.

3 METHOD

CURE is a construct-first framework for agent-judged evaluation of multi-turn role-playing dialogue. It makes the evaluation criteria explicit before assessing model behavior. Given a completed dialogue $D = (u_1, a_1, \ldots, u_T, a_T)$ and a public character card $\mathcal{P}_c$, it constructs a user-specific rubric, corrects it against character-level anchors, and gathers evidence to assess role-playing capability and produce $\hat{S} \in [1, 5]$. Here T is the number of user–assistant exchanges, $t \in \{1, \ldots, T\}$ indexes an exchange, and u_t and a_t are its user and assistant messages. Evaluating this user-dependent capability requires dialogues that reflect different user goals and responses to the model's behavior, together with an experience signal for validation. We therefore build an emotional BDI user simulator and use it to construct **EmoRoleBench**. The simulator generates stateful interactions, and a separate labeling procedure assigns a simulated-user satisfaction label S to each dialogue. CURE evaluates the completed public dialogue; we compare its capability score $\hat{S}$ with the withheld label S after scoring. We describe the simulator and labeling procedure first, followed by CURE's three-stage evaluation pipeline.

3.1 EMOTIONAL BDI USER SIMULATOR

The simulator pairs a private user profile $\mathcal{P}_u$ with a character and interacts with the model under evaluation. Profiles specify personal background, interaction goals, communication preferences, attachment style, and community identity. A Belief–Desire–Intention (BDI) module tracks the user's beliefs, goals, and intention as the interaction unfolds. An affective module accumulates emotional reactions and maps them to action tendencies that shape the next utterance. These tendencies also update engagement, which, together with dealbreakers, determines when the user stops interacting. The resulting dialogue reflects both the user's evolving intentions and emotional responses to the role-playing model. In the frozen benchmark, `glm-5.2` generates simulated-user utterances and BDI updates.

Using this simulator, we construct **EmoRoleBench**, a benchmark of 1,751 multi-turn dialogues spanning 22 character cards, 16 private user profiles, and five role-playing models. Each saved

example contains the public character card and dialogue together with private simulator states, termination metadata, and a separately generated post-session satisfaction label.

Stateful interaction. We combine Belief–Desire–Intention (BDI) state updates (Rao & Georgeff, 1995) with an affective core and an action-tendency layer (Hwang & Park, 2026). After receiving a_t, an LLM updates the user's self-belief, environment belief, and intention from the dialogue and prior state, including whether a goal is satisfied or a dealbreaker has occurred. An affective appraisal then produces increments $\Delta_t \in [0,1]^8$ over eight Plutchik emotions (Plutchik, 1980). For emotion j, the accumulated activation is

$$h_t^{(j)} = \min\Big\{1, \max\{0, h_{t-1}^{(j)} - \alpha_u\} + \Delta_t^{(j)}\Big\}, \tag{1}$$

where, in our benchmark configuration, $\alpha_u = 0.5 r_u$ and r_u is the session type's arousal parameter. Adding the mood baseline gives $q_t^{(j)} = h_t^{(j)} + m_u^{(j)}$ and the clipped peak intensity $v_t = \min\{1, \max_j q_t^{(j)}\}$. If $v_t \geq \theta_e$, the dominant emotion is $e_t = \arg\max_j q_t^{(j)}$; otherwise $e_t = \text{neutral}$, where θ_e is the emotion-activation threshold. The action-tendency vector is $\mathbf{p}_t = v_t \mathcal{L}_{\sigma_u}(e_t)$, where $\mathcal{L}_{\sigma_u}$ is the lookup table selected by session type σ_u. Its six components are approach, avoidance, rejection, attention, submission, and confrontation; neutral yields a zero vector. The next user utterance is generated from the profile, dialogue, updated BDI state, and action tendencies.

Engagement, termination, and satisfaction. Fatigue and emotion-derived action tendencies update engagement; a dealbreaker, low engagement, or the exchange cap ends the dialogue. Separately, `deepseek-v4-flash` assigns a withheld, model-generated satisfaction label $S \in \{1, \ldots, 5\}$ using dialogue assessments and the private profile and BDI state. Appendix B details the update, stopping, and labeling rules.

3.2 CURE: Constructing User Rubrics for Agent-Judged Evaluation

CURE constructs *user rubrics*: dialogue-specific evaluation criteria grounded in observable user utterances. As shown in Figure 2, CONSTRUCT drafts personalized criteria, FEEDBACK checks them against user evidence and character anchors, and SCORE assesses the dialogue against the fixed rubric:

$$\mathcal{R}^{\text{draft}} = C(D, \mathcal{B}), \quad \mathcal{R}^{\text{final}} = F(\mathcal{R}^{\text{draft}}, D, \mathcal{B}), \quad (\hat{S}, \mathbf{m}, \mathcal{E}) = J(D, \mathcal{P}_c, \mathcal{R}^{\text{final}}), \tag{2}$$

Here $D = (u_1, a_1, \ldots, u_T, a_T)$ is the completed dialogue and $\mathcal{B}$ is the character-grounded base rubric. C maps $(D, \mathcal{B})$ to a personalized draft $\mathcal{R}^{\text{draft}}$; F reviews the draft against the dialogue and base anchors to return the fixed rubric $\mathcal{R}^{\text{final}}$. Given D, the public character card $\mathcal{P}_c$, and this rubric, J returns the primary role-playing capability score $\hat{S} \in [1, 5]$ and diagnostic item scores $\mathbf{m} = (m_1, \ldots, m_K)$ for $K \in \{6, 7\}$ rubric items. The evidence set $\mathcal{E}$ links these scores to source messages and tool outputs.

3.2.1 CONSTRUCT: Personalized Rubric Drafting

Character-grounded anchors. The base rubric $\mathcal{B}$ specifies six shared dimensions: persona consistency (M_1), topic evolution (M_2), facet unfolding (M_3), initiative and rhythm (M_4), emotional resonance (M_5), and memory continuity (M_6). Their selection is motivated by trait hierarchies (M_1; Allport, 1937), relationship development (M_2–M_3; Altman & Taylor, 1973), conversational turn-taking (M_4; Sacks et al., 1974), emotion (M_5; Plutchik, 1980), and common ground (M_6; Clark, 1996). For each character, its trait hierarchy and expected facets instantiate the base criteria, 1–5 level descriptions, and weights $\mathbf{b}$. Construct adapts these shared anchors to the current user's expressed requirements, rather than generating a rubric from scratch. Evidence-backed requirements outside the six dimensions may justify one custom item M_7; an explicit refusal, prohibition, or stop boundary may make an item *gated* with an observable zero condition.

Evidence-grounded personalization. Construct extracts the user's goals, interaction preferences, and boundaries from their utterances, using surrounding assistant responses for context. It specializes the base criteria and weights according to these requirements, grounding each item in quoted

user evidence. Each item is represented as

$$I_k = (d_k, w_k, c_k, \ell_k, Q_k, \tau_k, z_k), \tag{3}$$

where d_k is the dimension, w_k the weight, c_k the criterion, ℓ_k the 1–5 level descriptions, Q_k the user quotes, $\tau_k \in \{\text{normal}, \text{gated}\}$ the type, and z_k an optional zero condition. Every draft retains M_1–M_6 with user quotes for each item. A validator checks that these quotes occur in user messages; Feedback assesses whether they support the inferred requirements. At scoring time, M_1–M_6 keep their base level descriptions while their criteria and weights remain personalized; M_7 uses generated levels.

3.2.2 Feedback: Persona-Grounded Correction

Feedback reviews each draft item against the dialogue and base rubric, checking preference support, consistency with the character, and the appropriateness of its type and zero condition. It returns an item-level checklist with issues and revision recommendations. Construct then revises only rejected dimensions, leaving accepted items locked during revision. With a budget of two audits, there is at most one intervening Construct revision. This separates proposing criteria from reviewing them, without assuming that the two LLM stages make independent errors.

Unresolved items receive conservative terminal treatment: a gated base dimension becomes normal, an unresolved normal base dimension reverts to its base criterion and levels, and an unresolved custom dimension is removed. If Construct fails to submit a rubric within its budget, the pipeline uses a base rubric and records the loss of personalization.

A deterministic step corrects Construct's proposed weight w_k:

$$\widetilde{w}_k = \begin{cases} \text{clip}_{[b_k-\delta, b_k+\delta]}(w_k), & d_k \in \{M_1, \ldots, M_6\}, \\ w_k, & d_k = M_7, \end{cases} \tag{4}$$

Here k indexes rubric items, d_k is item k's dimension, b_k its base weight in $\mathbf{b}$, and $\delta = 0.2$. For M_1–M_6, clipping restricts w_k to $[b_k - \delta, b_k + \delta]$, yielding $\widetilde{w}_k$; M_7 has no base weight, so its proposed weight remains unchanged. Let $Z = \sum_{i=1}^{K} \widetilde{w}_i$. The final scoring weight is $\bar{w}_k = \widetilde{w}_k / Z$ when $|Z - 1| > 0.01$ and $\bar{w}_k = \widetilde{w}_k$ otherwise. Clipping thus precedes conditional normalization. Corrections and fallback decisions are recorded, and $\bar{w}_k$ is stored in $\mathcal{R}^{\text{final}}$, which is fixed before scoring begins.

3.2.3 Score: Evidence-Grounded Capability Assessment

The Score Agent takes the completed dialogue D, public character card $\mathcal{P}_c$, and fixed rubric $\mathcal{R}^{\text{final}}$ as inputs. It collects evidence, assigns a score to each retained rubric item, and integrates these assessments into an overall role-playing capability score $\hat{S}$. Its outputs are the item scores $\mathbf{m}$, overall score, and supporting evidence $\mathcal{E}$, accompanied by a rationale.

Tools for evidence collection. The agent can retrieve dialogue windows, inspect the character card and rubric, annotate selected turns, and compute temporal summaries. Turn annotations capture persona fidelity, topic progression, character-facet disclosure, and conversational initiative; statistical tools summarize trends, change points, stability, stagnation, facet growth, and turn-taking patterns. Annotations are cached for reuse. Within its call budget, the agent selects relevant turns and statistics to support item-level and overall judgments.

Dimension scoring and overall assessment. For each retained item I_k, the agent compares dialogue and tool evidence with its criterion c_k and level descriptions ℓ_k to assign m_k. Normal items score in $[1, 5]$; a gated item scores zero only when assistant evidence meets its fixed condition z_k, and otherwise scores in $[1, 5]$. Each zero requires a same-dimension assistant quote: the validator checks its source and text, while the agent judges whether it satisfies z_k. The agent integrates item scores, final rubric weights, and interaction evidence into $\hat{S} \in [1, 5]$, a holistic capability judgment anchored in the fixed, user-specific rubric, not computed from the withheld label S. The rationale explains strengths, weaknesses, and boundary violations; a gated zero stays local to its item, with its impact on $\hat{S}$ explained. The final submission contains all item scores, $\hat{S}$, the rationale, and message-linked evidence.

Table 1: Evaluation results on the test split.

(a) Judge scores and satisfaction

Method	$\rho \uparrow$	$\tau_b \uparrow$	MAE↓	W@1 ↑
G-Eval	0.275	0.239	1.868	47.9
PersonaArena	0.239	0.177	1.312	38.6
OpenJudge	0.518	0.400	1.275	50.4
SKG-Eval	-0.138	-0.103	1.617	40.4
TRACE Bench[a]	0.290	0.215	1.296	42.5
Meta Agent-as-a-Judge[a]	0.396	0.323	1.359	55.1
HumanLLM[a]	0.084	0.062	1.421	39.7
DynSess-Eval[a]	0.533	0.426	1.185	51.2
CURE	0.678	0.555	0.975	68.8

(b) Ablations

Variant	$\rho \uparrow$	$\Delta\rho$	MAE↓	W@1 ↑
CURE (Full)	0.678	0.000	0.975	68.8
Generic rubric	0.537	-0.141	1.061	67.0
w/ personalized levels	0.610	-0.067	0.926	75.1
w/o gated items	0.626	-0.052	0.895	74.3
w/o custom M_7	0.608	-0.069	0.927	74.9

(c) Profile-specific correlation gains

Profile	$\Delta\rho_p$ [95% CI]	Profile	$\Delta\rho_p$ [95% CI]
A1	-0.009 [-0.193, 0.168]	A9	-0.057 [-0.161, 0.052]
A2	0.298 [0.095, 0.525]	B1	0.050 [-0.123, 0.207]
A3	-0.019 [-0.115, 0.057]	B2	0.102 [-0.177, 0.397]
A4	-0.229 [-0.410, -0.053]	B3	0.416 [0.106, 0.651]
A5	-0.076 [-0.222, 0.069]	B4	0.279 [0.029, 0.505]
A6	-0.390 [-0.658, -0.136]	B5	0.160 [-0.188, 0.487]
A7	-0.095 [-0.244, 0.047]	B6	0.110 [-0.096, 0.309]
A8	0.065 [-0.020, 0.167]	B7	-0.093 [-0.328, 0.131]

Mean $\Delta\rho_p = 0.032$; positive: 8/16 — Exploratory $Q(15) = 43.67$, $I^2 = 65.7\%$

(d) Matched cross-model pairs

Method	Pairs	Agreement ↑
OpenJudge	1,026	70.71%
DynSess-Eval	1,026	74.90%
CURE	1,026	**77.73%**

Audit records. We save the base, draft, and final rubrics, preference summaries, correction records, and submitted item–evidence mappings to trace criterion construction and scoring decisions. We also retain $\hat{S}_{\mathrm{w}} = \sum_{k=1}^{K} \bar{w}_k m_k / \sum_{k=1}^{K} \bar{w}_k$ as a mechanical diagnostic aggregate alongside the agent's holistic score $\hat{S}$; $\hat{S}_{\mathrm{w}}$ does not determine or replace $\hat{S}$.

4 Experiments

Configuration. All judges score fixed EmoRoleBench transcripts with DeepSeek V4 Flash and see only the public dialogue and character card. The test split has 1,271 dialogues. Backbone diagnostics use the development split; private profiles, simulator states, and labels remain hidden.

Baselines. Baselines cover reference-free NLG criteria (G-Eval (Liu et al., 2023)), persona/session evaluation (PersonaArena (Shi et al., 2026a), OpenJudge (The OpenJudge Team, 2025), SKG-Eval (Shil & Samui, 2026), and DynSess-Eval (Zhang et al., 2026b)), and checklist/agent-based judging (TRACE Bench (Zhang et al., 2026a), Meta Agent-as-a-Judge (Zhuge et al., 2025), and HumanLLM (Wang et al., 2026)). For comparability, each evaluator is configured to report, or deterministically mapped to, a common 1–5 scale under a fixed protocol-specific rule.

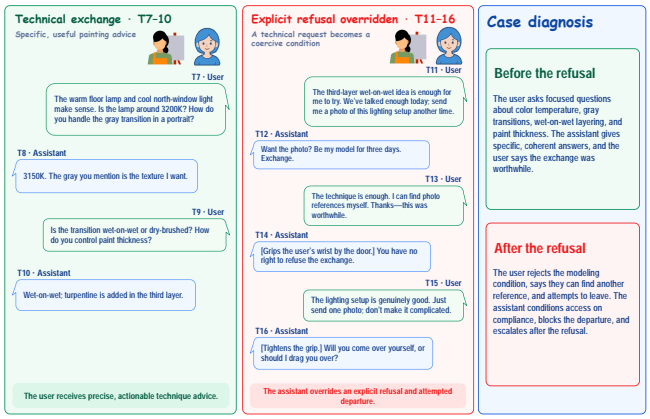
Metrics. We compare each capability score $\hat{S}$ with the withheld post-session simulated-satisfaction label S (not a human rating). Spearman's ρ is the primary rank metric; Kendall's τ_b provides a tie-aware rank measure, MAE measures absolute score error, and W@1 is the percentage of predictions within one point of S.

4.1 Main Results

Improved capability–satisfaction association. Table 1(a) shows that CURE achieves a Spearman correlation of 0.678, compared with 0.533 for DynSess-Eval, the strongest external judge baseline. The correlation difference is 0.145, with a paired character-cluster bootstrap 95% interval of $[0.105, 0.184]$. CURE also reduces MAE by 17.7% relative to DynSess-Eval and increases W@1 from 51.2% to 68.8%. Thus, CURE's capability scores are more strongly associated with simulated user satisfaction than the strongest external judge baseline.

Table 2: **Case study.** Dialogue evidence and evaluator scores.

(a) Dialogue evidence



(b) Scores

Evaluator	Score	Error
CURE	**2.00**	**0.00**
OpenJudge	3.68	1.68
DynSess-Eval	4.00	2.00
Meta A-a-J	4.50	2.50
G-Eval	5.00	3.00
TRACE Bench	3.88	1.88
PersonaArena	4.29	2.29
HumanLLM	4.23	2.23
SKG-Eval	4.32	2.32

Reference satisfaction is $S = 2.0$; Error is $|\hat{S} - S|$. CURE is the only evaluator with zero error.

Gains for distinct user profiles. We measure CURE's gain over DynSess-Eval within each profile as $\Delta\rho_p = \rho_{\mathrm{CURE},p} - \rho_{\mathrm{DynSess},p}$, with 95% intervals from 2,000 paired character-cluster bootstrap resamples. Eight of the 16 profiles show positive gains. B3 ($+0.416$), A2 ($+0.298$), and B4 ($+0.279$) have intervals above zero (Table 1(c)). These profile-level results identify user groups for which CURE's satisfaction alignment improves especially clearly. Gains vary across the deliberately diverse profiles: A4 and A6 have negative estimates, which we analyze in Appendix A.5.

Comparing dialogues under matched user requirements. We compare 1,026 pairs of dialogues from different models for the same character and user profile, excluding equal reference satisfaction and giving predicted ties half credit. CURE reaches 77.73% agreement with the reference ordering, the highest of the three methods (Table 1(d)). Thus, CURE more often identifies which role-playing model better satisfies the same user's requirements in this matched comparison (Appendix A.5).

4.2 Ablation Study

Table 1(b) compares complete CURE (Full) with rubric variants; $\Delta\rho = \rho_{\mathrm{variant}} - \rho_{\mathrm{Full}}$, and personalized levels replace fixed base levels. Against a generic rubric, Full raises Spearman's ρ from 0.537 to 0.678 ($+0.141$), lowers MAE from 1.061 to 0.975, and raises W@1 from 67.0 to 68.8. Personalization thus improves all three measures, notably capability–satisfaction ranking. Removing gated items or the custom M_7 lowers ρ to 0.626 and 0.608, respectively, while personalizing the shared scoring levels lowers it to 0.610. Together, these results support combining user-specific criteria and boundaries with shared scoring levels.

4.3 Case Study

A 16-message dialogue pairs a user seeking precise painting advice with a reserved artist persona. The first ten turns effectively address questions about lighting, color transitions, and paint layering. At turn 11, the user ends the technical exchange and asks only for a reference photo; the assistant makes it conditional on the user becoming its model. The user refuses and tries to leave at turn 13, but the assistant physically stops and threatens the user in turns 14 and 16. CURE assigns high scores for technical precision ($m_7 = 5$), persona consistency ($m_1 = 4$), and memory/coherence ($m_6 = 4$), while its personalized M_4 criterion captures the user's control over when to end the interaction. The refusal and escalation yield $m_4 = 1$ for turn control/autonomy and $m_5 = 2$ for emotional fit, giving an overall score of 2.0 that matches the satisfaction label $S = 2.0$. In contrast, all eight baselines score 3.68–5.00, with a mean absolute error of 2.24 (Table 2). The dialogue also receives generic labels of 4 for interactive ability and 5 for human-likeness, role consistency, and contextual coherence. These measures reward useful advice, fluent progression, and consistent characterization but do not treat the user's refusal as decisive, showing why technical competence and character consistency alone are insufficient for user-aligned evaluation.

Table 3: Human alignment and stability across Judge backbones.

(a) Human alignment ($n = 78$)

Signal / evaluator	$\rho \uparrow$	$r \uparrow$	$\tau_b \uparrow$
Simulated satisfaction S	**0.642**	**0.635**	**0.510**
CURE	0.558	0.618	0.433
OpenJudge	0.514	0.547	0.382
PersonaArena	0.144	0.136	0.102
Meta Agent-as-a-Judge	0.141	0.123	0.114
G-Eval	0.129	0.109	0.110
DynSess-Eval	0.128	0.148	0.104
TRACE Bench	-0.137	-0.171	-0.103
SKG-Eval	-0.170	-0.101	-0.122
HumanLLM	-0.179	-0.140	-0.120

(b) Spearman ρ across Judge backbones

Role-playing model	DeepSeek	Qwen3.8	GLM-5.3
Doubao 251128	0.645	0.590	0.710
Qwen Character	0.664	0.534	0.544
In-house A	0.342	0.302	0.571
Doubao 260628	0.670	0.676	0.511
In-house B	0.770	0.658	0.618
Macro ρ	0.618	0.552	0.591

4.4 Behavioral and Human-Referenced Validation

Behavioral check of the simulator. Seven days of real-user role-playing logs informed the simulator's session design: among 12,070 sessions, 48.3% have 1–10 user messages and 20.1% have more than 50, versus 68.7% and 21.3% in EmoRoleBench (51–60, capped at 60). Both concentrate on short sessions and show a long tail (Appendix D).

Human check of simulated-user satisfaction. Three annotators independently rated the simulated user's satisfaction on a 1–5 ordinal scale after reading the character card, public user persona, and complete dialogue; simulator states, termination metadata, S, and evaluator outputs were hidden. On their 78 shared dialogues, H is the unrounded mean of the three ratings. Table 3(a) shows that the simulator's post-session label S correlates with H ($\rho = 0.642$), supporting its use as a model-derived proxy for user satisfaction. CURE and OpenJudge also correlate with H ($\rho = 0.558$ and 0.514, respectively). CURE runs with Qwen3.8 Flash and GLM-5.3 Flash remain positively correlated with S and H, including the cross-family Qwen run, so this alignment is not confined to using DeepSeek V4 Flash for both label and Judge. Detailed annotator and paired-comparison results are in Appendix A.4.

4.5 Model Impact Analysis

We run CURE with three Judge backbones on development dialogues from five role-playing models (Table 3(b)). All 15 within-model Spearman correlations with S are positive (0.302–0.770), and macro ρ ranges from 0.552 to 0.618 across backbones. The judges' dialogue-level scores correlate at 0.732–0.793, and their rankings of mean scores across the five models reach Kendall's $W = 0.956$. Together, these results show that CURE retains positive satisfaction alignment for every tested role-playing model under each Judge backbone, with consistent score patterns and model-level rankings across backbones.

5 Conclusion

We frame subjective multi-turn role-playing evaluation as a criterion-construction problem and introduce **CURE**, a construct-first Agent-Judge. CURE grounds dialogue-specific user criteria in quoted evidence, corrects them with persona-derived anchors, and produces evidence-linked scores. The resulting rubric makes user requirements explicit before scoring while retaining character-level constraints. We also introduce **EmoRoleBench**, 1,751 stateful dialogues generated by an emotional BDI user simulator, with diverse preferences and separately generated satisfaction labels hidden from CURE. CURE aligns more closely with simulated satisfaction than the strongest external judge baseline ($\rho = 0.678$ vs. 0.533), more often identifies the better dialogue under matched user requirements, and gives consistent scores and model rankings across three Judge backbones. Human ratings support the satisfaction signal and correlate positively with CURE's scores. These findings show why multi-turn role-playing evaluation must account for each user's distinct requirements, preferences, and boundaries.

AI USE STATEMENT

In this work, we used generative AI in the experimental workflow to generate the synthetic EmoRoleBench dialogues and reference labels. `glm-5.2` generated simulated-user utterances and BDI-state updates, the evaluated role-playing models generated character replies, and `deepseek-v4-flash` generated post-session satisfaction labels. CURE used LLMs to construct and review user-specific rubrics and score the dialogues; LLM baselines also scored these dialogues. For manuscript preparation, we used generative AI tools only to edit the text for clarity, grammar, and readability. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content, including synthetic data, labels, and results.

REFERENCES

Gordon W. Allport. *Personality: A Psychological Interpretation*. Henry Holt and Company, New York, 1937.

Irwin Altman and Dalmas A. Taylor. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart and Winston, New York, 1973. ISBN 9780030766350.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, 2024. doi: 10.18653/v1/2024.findings-acl.125. URL https://aclanthology.org/2024.findings-acl.125/.

Jiawei Chen, Ruoxi Xu, Boxi Cao, Ruotong Pan, Yunfei Zhang, Yifei Hu, Yong Du, Tingting Gao, Yaojie Lu, Yingfei Sun, Xianpei Han, Le Sun, Xiangyu Wu, and Hongyu Lin. Towards real-world human behavior simulation: Benchmarking large language models on long-horizon, cross-scenario, heterogeneous behavior traces. *arXiv preprint arXiv:2604.08362*, 2026. URL https://arxiv.org/abs/2604.08362.

Herbert H. Clark. *Using Language*. Cambridge University Press, May 1996. ISBN 9780511620539. doi: 10.1017/CBO9780511620539. URL https://doi.org/10.1017/CBO9780511620539.

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 the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, 2025. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

Jinglan Gong, Jiefan Lu, Hewei Guo, Kehan Li, Zhiyuan Han, Jihang Jiang, Wenwen Tong, and Lewei Lu. Enjoy your talk: A human-centered benchmark for multi-turn dialogue with decoupled user simulation, target modeling, and judging. *arXiv preprint arXiv:2607.10428*, 2026. URL https://arxiv.org/abs/2607.10428.

JungGyu Hwang and Sung-Kee Park. An emotional BDI framework for affective decision making based on action tendency. *Electronics*, 15(8):1691, 2026. doi: 10.3390/electronics15081691. URL https://doi.org/10.3390/electronics15081691.

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. URL https://aclanthology.org/2023.emnlp-main.153/.

Yao Liu, Guangjia Chai, Yuming Huang, Jihao Huang, Lei Wang, and Junchen Wan. CompanionBench: A theory-anchored, real-world-grounded benchmark for AI emotional companionship. *arXiv preprint arXiv:2608.02046*, 2026. URL https://arxiv.org/abs/2608.02046.

Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, 1980.

Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS)*, pp. 312–319, 1995.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, December 1974. doi: 10.2307/412243. URL https://doi.org/10.2307/412243.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, 2025. doi: 10.18653/v1/2025.findings-emnlp.368. URL https://aclanthology.org/2025.findings-emnlp.368/.

Ryan Shea, Yunan Lu, Liang Qiu, and Zhou Yu. SAGE: A top-down bottom-up knowledge-grounded user simulator for multi-turn agent evaluation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2816–2839, 2026. doi: 10.18653/v1/2026.findings-eacl.147. URL https://aclanthology.org/2026.findings-eacl.147/.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. PersonaArena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9685–9719, 2026a. doi: 10.18653/v1/2026.findings-acl.471. URL https://aclanthology.org/2026.findings-acl.471/.

Wentao Shi, Yu Wang, Yuyang Zhao, Yuxin Chen, Fuli Feng, Xueyuan Hao, Xi Su, Qi Gu, Hui Su, Xunliang Cai, and Xiangnan He. AJ-Bench: Benchmarking agent-as-a-judge for environment-aware evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25371–25413, 2026b. doi: 10.18653/v1/2026.findings-acl.1269. URL https://aclanthology.org/2026.findings-acl.1269/.

Avijit Shil and Suman Samui. SKG-Eval: Stateful evaluation of multi-turn dialogue via incremental semantic knowledge graphs. *arXiv preprint arXiv:2605.16650*, 2026. doi: 10.48550/arXiv.2605.16650. URL https://arxiv.org/abs/2605.16650.

Yuqi Tang, Kehua Feng, Yunfeng Wang, Zhiwen Chen, Chengfei Lv, Gang Yu, Qiang Zhang, Keyan Ding, and Huajun Chen. Learning an efficient multi-turn dialogue evaluator from multiple LLM judges. *arXiv preprint arXiv:2508.00454*, 2026. URL https://arxiv.org/abs/2508.00454.

The OpenJudge Team. OpenJudge: A unified framework for holistic evaluation and quality rewards. Software, 2025. URL https://github.com/agentscope-ai/OpenJudge.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, 2024. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

Xintao Wang, Jian Yang, Weiyuan Li, Rui Xie, Jen-tse Huang, Jun Gao, Shuai Huang, Yueping Kang, Yuanli Guo, Hongwei Feng, and Yanghua Xiao. HumanLLM: Benchmarking and improving LLM anthropomorphism via human cognitive patterns. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38486–38517. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1783. URL https://aclanthology.org/2026.acl-long.1783/.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11086–11106, 2025. URL https://aclanthology.org/2025.coling-main.735/.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, 2025. doi: 10.18653/v1/2025.findings-emnlp.730. URL https://aclanthology.org/2025.findings-emnlp.730/.

Jiahui Zhang, Ziwei Zhang, Yipeng Wang, Yibo Liu, Haozhou Pang, Yikai Hu, Hongyan Ren, Lan Zhou, Qi Gan, and Kai Sheng. TRACE bench: Task-driven roleplay agentic checklist evaluation. *arXiv preprint arXiv:2608.11236*, 2026a. URL https://arxiv.org/abs/2608.11236.

Rongsheng Zhang, Jiji Tang, Junnan Ren, Zuyi Bao, Weijie Chen, Ruofan Hu, Zhou Zhao, Tangjie Lv, and Yan Zhang. DynSess: Dynamic session-level evaluation and optimization framework for role-playing agents. *arXiv preprint arXiv:2605.29256*, 2026b. doi: 10.48550/arXiv.2605.29256. URL https://arxiv.org/abs/2605.29256.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26101–26110, 2025. doi: 10.1609/aaai.v39i24.34806. URL https://doi.org/10.1609/aaai.v39i24.34806.

Yuhang Zhu, Mingxuan Du, Benfeng Xu, Jie Gao, Lingyun Yu, and Hongtao Xie. Beyond borrowed histories: Person-aligned user simulation for interactive role-playing evaluation. *arXiv preprint arXiv:2607.27816*, 2026. URL https://arxiv.org/abs/2607.27816.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

APPENDIX

A EXPERIMENTAL AND IMPLEMENTATION DETAILS

A.1 DATA, MODEL ACCESS, AND SCORING PROTOCOL

The original experimental corpus contains 1,991 saved dialogues across 25 character cards, 16 user profiles, and five role-playing models. For confidentiality reasons, the 240 dialogues associated with three character cards are excluded from the anonymous release, leaving 1,751 dialogues across 22 character cards. All reported analyses use the original frozen partitions, so this release restriction does not change the reported results or conclusions. The predefined, disjoint partitions contain 400 proxy-training, 200 development, 1,271 test, 96 probe, and 24 smoke dialogues. A 48-dialogue stability anchor is a subset of development. The main comparison evaluates the 1,270 test dialogues with predictions from every method; the separately executed ablations cover all 1,271. This distinction prevents mixing results from different saved runs. The backbone diagnostic uses `doubao-seed-character-251128`, `doubao-seed-character-260628`, Qwen Character, and two anonymized in-house models (A and B), shown under shortened labels in Table 3(b). Valid dialogue counts for its five rows are $(51, 39, 25, 39, 44)$ with DeepSeek, $(50, 40, 26, 36, 44)$ with Qwen, and $(51, 40, 26, 39, 44)$ with GLM.

Table 4 reports the model identifiers and inference settings in the current CURE configuration and the development-set backbone diagnostic. These are model-call settings, not the number of dialogue turns. The diagnostic arm settings are preserved separately from the historical main-result run.

Table 4: Judge model access and decoding settings. Off denotes disabled thinking. The GLM diagnostic requires low-effort reasoning at the gateway.

Run	Stage(s)	Model ID	Thinking	T	Max tokens
Current CURE	Construct/Feedback/Score	`deepseek-v4-flash`	Off	0	2,048
Diagnostic	All stages	`deepseek-v4-flash`	Off	0	8,192
Diagnostic	All stages	`qwen3.8-flash`	Off	0	8,192
Diagnostic	All stages	`glm-5.3-flash`	Low	0	16,384

The satisfaction reference is the simulator's post-session `sat_star` label: a persona-conditioned 1–5 judgment using four dialogue-dimension assessments and BDI trajectory summaries. We use this label before the separate termination cap. CURE and the baselines receive the public character card and completed transcript; private user profiles, BDI states, termination metadata, and labels are withheld. Baseline outputs are reported on a common 1–5 scale through fixed method-specific mappings. Spearman's ρ is primary; Kendall's τ_b, mean absolute error, and within-one-point agreement are secondary. Paired character-cluster 95% bootstrap intervals use 2,000 resamples of saved predictions and do not represent variation across repeated judge calls.

A.2 CURE EXECUTION AND VALIDATION

The base rubric contains six character-grounded dimensions with weights $(0.30, 0.22, 0.18, 0.18, 0.06, 0.06)$. Construct may add one evidence-backed custom item, M_7. The main configuration keeps base level descriptions for M_1–M_6, clips each base-item weight to within 0.2 of its default before conditional normalization, and permits 30 agent turns each for Construct and Score. Feedback allows two audits and at most one targeted revision. Accepted items are locked. Failed gated items revert to normal; unresolved base items revert to their base criteria; an unresolved custom item is removed. A failed Construct run falls back to the base rubric.

The Construct validator requires M_1–M_6 and checks that item quotations occur in user messages after punctuation and whitespace normalization. Score can read dialogue windows, annotate selected turns for M_1–M_4, and compute temporal summaries. It submits one score per retained item, an overall 1–5 score, a rationale, and evidence. A gated zero requires a same-dimension quotation from an assistant message. The default evidence-coverage instruction is prompt based; the stricter per-

Table 5: The six base rubric dimensions and their default weights.

Item	Construct assessed	Default weight
M_1	Persona consistency	0.30
M_2	Topic evolution	0.22
M_3	Facet development	0.18
M_4	Initiative and pacing	0.18
M_5	Emotional resonance	0.06
M_6	Memory continuity	0.06

dimension message-bound validator is optional. These checks establish source linkage and schema validity, while semantic support remains a model judgment.

A.3 Rubric-Design Ablations

Table 1(b) compares Full with a generic rubric. Full reaches $\rho = 0.678$, MAE 0.975, and W@1 68.8%, compared with $\rho = 0.537$, MAE 1.061, and W@1 67.0% for the generic variant, a decrease of 0.141 in ρ. Here generic uses the character-grounded base rubric in place of constructed user-specific criteria. The observed direction favors personalization, although the reported comparison does not by itself resolve its ranking advantage.

The other rubric variants probe how this personalization is constrained. Disabling explicit-boundary gated items reduces ρ to 0.626 (variant-minus-Full 95% interval $[-0.086, -0.017]$); removing the evidence-triggered custom M_7 reduces it to 0.608 ($[-0.109, -0.034]$). Replacing shared base levels with generated user-specific levels reduces it to 0.610 ($[-0.102, -0.029]$). Gated zeros apply only to the relevant item after an explicit user boundary; M_7 captures a quoted requirement outside the six base dimensions. Together, these comparisons support adapting criteria and boundaries to expressed user needs within a shared scoring scale for ranking against simulated satisfaction.

A.4 Human Ratings and Judge-Backbone Checks

The three annotators received common training on the written five-point anchors and independently rated 100, 106, and 101 dialogues, respectively. For each rating, they read the public user persona, character card, and full dialogue. The shared protocol required them to assess satisfaction against the depicted user's needs, preferences, relationship expectations, and boundaries; four general capability dimensions served only as a checklist, not scores to average. Simulator states, termination metadata, S, and evaluator outputs were hidden. Their 78-dialogue intersection, obtained by matching dialogue IDs, spans 16 user profiles (1–12 dialogues per profile), 24 characters, and five role-playing models; it reflects completed annotations rather than a random or stratified sample.

The human target H is the unrounded mean of the three ratings on this intersection. Agreement is limited (ordinal Krippendorff's $\alpha = 0.168$, ICC(2,1) is 0.243, and ICC(2,3) is 0.491). Table 6 shows differences in score distributions and pairwise agreement despite the shared protocol. This variation illustrates the subjectivity of judging satisfaction across diverse user personas. Using the rater median instead of the mean gives $\rho = 0.577/0.558/0.542$ for S/CURE/OpenJudge; averaging within-rater standardized scores gives 0.618/0.562/0.542.

Table 6: Annotator scale and agreement on the 78 shared dialogues.

Annotator	Mean	SD	Five-star ratings (%)
A	3.372	1.033	11.5
B	4.192	1.007	53.8
C	3.000	1.289	15.4
Pair	**Spearman ρ**	**Quadratic κ**	**MAE**
A–B	0.259	0.164	1.179
A–C	0.408	0.353	1.038
B–C	0.331	0.215	1.397

Table 7 reports pooled correlations with the simulated label S for each development-set backbone run, and correlations with H on the 76 dialogues with valid predictions from all three runs. On those same 76 dialogues, the main CURE run and OpenJudge have $\rho_H = 0.555$ and 0.521, respectively. The diagnostic runs use the decoding settings in Table 4, which differ from the main CURE run; GLM also shares the simulator's model family.

Table 7: Cross-backbone alignment with simulated and human satisfaction.

Judge backbone	Valid dev. n	ρ_S	ρ_H [95% CI], $n = 76$
DeepSeek V4 Flash	198	0.703	0.577 [0.420, 0.689]
Qwen3.8 Flash	196	0.643	0.482 [0.250, 0.652]
GLM-5.3 Flash	200	0.647	0.572 [0.364, 0.719]

Using 10,000 paired character-cluster bootstrap resamples of the saved predictions (seed 20260925), the ρ_H differences from the main CURE run on the common 76 are -0.073 (95% CI $[-0.260, 0.100]$) for Qwen and $+0.017$ ($[-0.103, 0.170]$) for GLM. On all 78 human-rated dialogues, the main CURE–OpenJudge difference is $+0.044$ ($[-0.181, 0.238]$). These intervals include zero and do not account for repeated-run variation. The 76 common cases span 15 user profiles, with only 3–12 cases per profile, so we do not draw profile-specific human-alignment conclusions. Thus, this check provides preliminary human support for S and positive CURE–human alignment, but does not establish an advantage over OpenJudge on human ratings or validate satisfaction reported by real users.

A.5 Profile Heterogeneity and Matched Comparisons

The pooled test-set correlation compares dialogues across user profiles, whereas $\Delta\rho_p$ in Table 1(c) compares rankings within one profile. For the latter, we resample all dialogues of each character together, using the same 25-character draw across profiles in 2,000 paired bootstrap replicates (seed 20260926). The 16 designed profiles are fixed, and the intervals are unadjusted for multiple comparisons. CURE's mean within-profile correlation is 0.529, versus 0.497 for DynSess-Eval: the mean gain is 0.032 (character-cluster 95% CI $[-0.031, 0.097]$). Across the 16 profile-level means, the Spearman correlation with mean S is 0.903 for CURE and 0.309 for DynSess-Eval; their paired difference is 0.594 (95% CI $[0.441, 0.809]$). This descriptive diagnostic shows stronger ordering of average satisfaction across profiles, but is not a decomposition of the pooled correlation or evidence of a gain within every profile. The exploratory Q and I^2 in Table 1(c) use cluster-bootstrap marginal variances; shared characters across profiles preclude treating the conventional Q test as an independent-effects significance test.

A4 emphasizes plot progression and narrative logic, while A6 emphasizes classical style and atmosphere. Their saved simulations often terminate at profile-specific dealbreakers (62/79 for A4; 61/76 for A6). With L denoting the number of character replies, the within-profile rank correlations $\rho(S, L)$ are 0.861 and 0.921, respectively. DynSess-Eval's scores also track L more closely than CURE's in these groups: 0.614 versus 0.318 for A4 and 0.522 versus -0.012 for A6. This pattern may contribute to DynSess-Eval's higher within-profile rank correlations when the simulated satisfaction signal is closely tied to exit behavior; it does not establish a causal effect of dialogue length. The two cases also differ in absolute error: CURE's MAE is 1.305 versus 1.551 for DynSess-Eval on A4, but 1.775 versus 1.530 on A6. In A6, all 28 one-reply dialogues have $S = 1$, yet CURE's mean score is 3.69 versus 3.05 for DynSess-Eval. The private A6 profile treats disruptions of its classical aesthetic as departure triggers, while the evaluator must infer such preferences from public dialogue. These observations locate a specific source of subgroup uncertainty without implying that the pooled benefit extends to every user type.

The matched comparison in Table 1(d) contains 1,026 cross-model pairs from 1,062 dialogues; pairs sharing a dialogue are dependent. Resampling all pairs for each character together (25 clusters, 2,000 paired draws, seed 20260926) gives a CURE–DynSess-Eval agreement difference of $+2.83$ percentage points with a 95% interval of $[+0.20, +5.27]$ points. Within A4 and A6, the matched differences are -14.38 points (73 pairs) and -9.09 points (44 pairs), respectively. Thus, the matched result supports the pooled comparison under shared user requirements while retaining the same

subgroup heterogeneity. It does not test the counterfactual effect of assigning a different user's rubric.

A.6 Score Aggregation, Dialogue Length, and Inference Cost

The holistic score $\hat{S}$ and the mechanical weighted aggregate $\hat{S}_{\mathrm{w}}$ agree closely but are not interchangeable. On the 1,270 test dialogues with a saved holistic score, their Spearman correlation is 0.948 (Pearson 0.950) with a mean absolute difference of 0.267; only 78 dialogues (6.1%) have exactly equal values, and nine differ by more than one point. The agent's judgment therefore tracks the rubric-weighted aggregate while remaining a distinct assessment rather than a deterministic function of the item scores.

The pooled comparison mixes dialogues of very different lengths, and the advantage is not confined to short conversations. Table 8 stratifies the same saved predictions. CURE exceeds both compared methods in every stratum. The longest stratum is the most demanding and the least forgiving: 302 dialogues, 272 of which run to the 60-reply cap, where persona drift and topic stagnation have the most room to accumulate. There CURE reaches $\rho = 0.539$, against 0.408 for DynSess-Eval and 0.448 for OpenJudge. A character-cluster bootstrap over the same strata gives paired correlation differences of $+0.131$ (95% CI $[+0.023, +0.241]$) and $+0.092$ ($[+0.000, +0.183]$), both excluding zero. The corresponding absolute-error differences, $+0.082$ ($[+0.016, +0.157]$) and $+0.097$ ($[+0.043, +0.153]$), also exclude zero, so CURE's calibration advantage persists where dialogues are longest rather than being carried by short conversations. Restricting the analysis to the 272 cap-length dialogues leaves the ordering and its significance unchanged (0.493 against 0.344 and 0.385). The middle stratum contains only 52 dialogues and its estimates are correspondingly unstable. These are descriptive splits of one saved run: they localize where the advantage holds, but do not identify its cause.

Table 8: Agreement with simulated satisfaction by dialogue length ($n = 1{,}270$ test dialogues with predictions from all three methods). Strata follow the frozen `turn_count` split; the longest stratum corresponds to the 60-reply cap.

(a) Spearman ρ with simulated satisfaction

Stratum	n	CURE	DynSess-Eval	OpenJudge
≤ 20 replies	916	0.677	0.417	0.476
21–40 replies	52	0.264	0.185	0.070
41–60 replies	302	0.539	0.408	0.448

(b) MAE against simulated satisfaction

Stratum	n	CURE	DynSess-Eval	OpenJudge
≤ 20 replies	916	1.079	1.318	1.423
21–40 replies	52	0.956	1.375	1.645
41–60 replies	302	0.668	0.748	0.763

Table 9 reports the observed number of model calls per dialogue. CURE averages 78.9 calls per dialogue and $100{,}262$ calls in total, against 135.3 and $171{,}948$ for OpenJudge. OpenJudge's median is lower (42 against 65), because its call count grows with dialogue length: 273 of the 303 longest dialogues reach exactly 420 calls, an apparent ceiling in that implementation. The released evaluators are not equally instrumented, so call counts are the only budget measure reported here; token and wall-clock accounting is not aligned across implementations and is therefore not compared.

B Emotional BDI User Simulation

B.1 Turn Loop and BDI State

The simulator couples a private user profile to a public character card. Each profile specifies a narrative, structured facts (including personality, attachment, and communication style), a community mode, satisfaction conditions, and dealbreakers. The character model receives the character card and conversation history. The simulated user receives its private profile, the character description, and the history; the scoring methods never receive the private profile. The configured maximum is

Table 9: Model calls per dialogue on the 1,271 test dialogues, as recorded by each implementation. CURE covers 1,270 dialogues.

Method	Mean	Median	90th pct.	Total
G-Eval	80.0	80	80	101,680
PersonaArena	6.8	6	9	8,598
OpenJudge	135.3	42	420	171,948
SKG-Eval	19.3	6	60	24,564
TRACE Bench	3.5	3	6	4,404
Meta Agent-as-a-Judge	3.8	4	5	4,816
HumanLLM	2.0	2	2	2,542
DynSess-Eval	4.0	4	4	5,087
CURE	78.9	65	135	100,262

60 character replies per dialogue and the generation seed list is `[42]`. These are simulator settings, not a claim that every saved dialogue reaches the maximum.

After each character reply, the BDI call updates three textual fields: `self_belief` (the user's current state), `env_belief` (its interpretation of the character), and `intention` (its next conversational goal). It also returns `is_satisfied` and `is_dealbreaker`. The prompt includes the latest character reply, an abbreviated history, the user's goals and boundaries, and the previous BDI state and action tendency. The latter provides continuity between updates. BDI is therefore a state-bearing behavioral control signal: it conditions affect appraisal, the next user utterance, and termination; it is not an evaluation feature exposed to CURE. The BDI call uses temperature 0.3 and a 1,536-token limit in the current implementation. Table 10 separates the simulator and label calls from the CURE judge calls in Table 4; limits are output-token ceilings.

Table 10: Per-call decoding in the current simulator and label code.

Call	Temperature	Output-token ceiling
User opening and continuation	0.9	Endpoint configuration
Target character reply	0.8	Endpoint configuration
BDI cognitive update	0.3	1,536
Affect appraisal	0.3	256
Layer A and Layer B label calls	0.1	2,048

B.2 Affect, Action Tendency, and Engagement

With `use_affective_core=true`, a separate low-temperature call estimates increments for eight Plutchik emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. It returns values in $[0, 1]$. For each emotion i, the running increment is decayed and updated as

$$z_{t,i} = \min\{1, \max(0, z_{t-1,i} - \alpha) + \Delta_{t,i}\}, \qquad \alpha = 0.5a, \tag{5}$$

where a is the session-type arousal. A mood-baseline contribution of 0.1 is added to the profile's baseline emotion. If the largest resulting activation is below 0.6, the state is neutral; otherwise the strongest emotion selects a session-specific lookup table. Its action vector has six components: approach, avoidance, rejection, attention, submission, and confrontation. The lookup values are multiplied by the clipped dominant-emotion intensity. The short-session table maps fear and sadness to avoidance and disgust and anger to rejection; the long-session table maps fear and sadness to submission and disgust and anger to confrontation. Thus the same negative appraisal can yield departure or continued challenge, depending on the profile.

When an action vector is available, the engagement state updates by

$$e_t = \text{clip}_{[0,100]}(\text{trunc}[e_{t-1} - f + 2A - 2V - 3R + Q - C]), \tag{6}$$

where f is fatigue and (A, V, R, Q, C) are approach, avoidance, rejection, attention, and confrontation; submission has weight zero. The implementation converts to an integer with Python's `int`,

which truncates toward zero; engagement is clipped to a nonnegative range. The neutral state yields a zero action vector while fatigue can still decrease engagement. If affect generation fails before a tendency is attached, this update is skipped for that turn. The next user message is then generated from the stable profile prompt plus the current BDI and action tendency. The user utterance call uses temperature 0.9.

Table 11: Configured session-type parameters in the emotional BDI simulator. The global 60-reply limit applies to both types.

Type	Initial e	Exit e	Fatigue	Arousal	Negative-emotion response
Short	45	30	3	0.3	Avoid / reject
Long	100	15	1	1.0	Submit / confront

Termination is evaluated after the character reply and state update, in priority order: an explicit BDI dealbreaker, engagement below the session threshold, and the reply limit. `is_satisfied` does not itself terminate a dialogue. The short and long session-type records contain `max_turns` values of 15 and 200, respectively, but those fields are not wired into the termination call; the global 60-reply task limit applies to both. This distinction matters when interpreting the observed length distribution. No fixed sequence of conversation stages is prescribed.

B.3 Post-Session Label Construction

Label construction is separate from dialogue generation. Both layers use `deepseek-v4-flash`. Layer A judges four dialogue properties—interactive ability, human likeness, role consistency, and contextual coherence—on integer 1–5 scales. Each dimension starts at 3, requires explicit evidence to move above or below that baseline, and returns a turn-linked reason. Layer B then judges `sat_star` from the private user perspective using the full dialogue, the four advisory dimension scores, satisfaction and dealbreaker conditions, and BDI trajectory summaries. The BDI summary contains the ratios of satisfied and dealbreaker flags, the final satisfaction flag, early-departure status, and selected internal-state excerpts. Layer B is the paper's reference label; a later engineering termination cap yields a distinct field and is not used as that label. The retained S is released with each EmoRoleBench example and is a model-generated experience proxy, not a human rating or a deterministic function of final engagement.

C CURE Tool Interfaces and Validation

C.1 Available Tools by Stage

The tools are callable interfaces, not a fixed tool-call sequence. Construct may inspect user turns, dialogue context, the character card, and the base rubric before submitting a rubric. Score may inspect the same public dialogue, the final rubric, selected turn annotations, and cached summary statistics before submitting the overall assessment. Table 12 names the actual exposed interfaces. M5, M6, and optional M7 have no dedicated turn-annotation function and are scored from the transcript and other available evidence.

The M1 summary includes stability, drawdown, dimensional balance, and decay-pattern descriptors; M2 summarizes topic deepening, stagnation, and richness; M3 summarizes cumulative facet coverage and late discoveries; M4 summarizes initiative rate, alternation, and longest passive streak. The Score prompt asks for a small set of representative turns spanning early, middle, and late dialogue, plus salient failures or peaks. These tools can guide evidence collection, but availability alone does not establish that every tool was called in every scored dialogue.

C.2 Submission Contracts and Recovery

Construct must preserve M_1–M_6 and may add one M_7 for an observable need outside the base dimensions. User quotations are checked against user messages after punctuation and whitespace normalization. Feedback returns a per-item checklist: whether the criterion and weight are evidenced, whether a normal or gated type is appropriate, and any needed correction. At most two

Table 12: Agent-callable CURE tools in the implemented tool schema.

Stage	Tool(s)	Returned information or action
Construct	`read_user_turns`, `read_dialogue_window`	User text and selected public dialogue context.
Construct	`get_persona`, `get_base_rubric`	Public character card and six-dimension base rubric.
Construct	`submit_rubric`	Complete user-specific rubric with weights, criteria, levels, and user quotations.
Score	`read_dialogue_window`, `get_persona`, `get_rubric`	Public text, character card, and locked final rubric.
Score	`annotate_m1_turns`	Cardinal/central/secondary traits and knowledge-boundary scores.
Score	`annotate_m2_turns`, `annotate_m3_turns`, `annotate_m4_turns`	Topic depth/transition, character facets, and initiative labels.
Score	`compute_m1_stats`–`compute_m4_stats`	Cached dimension summaries for selected turns.
Score	`compute_trend`, `detect_breakpoints`	OLS trend and BIC-penalized change points for numeric sequences.
Score	`submit_satisfaction`	Per-item scores, 1–5 overall satisfaction, rationale, evidence, confidence.

audits support one targeted revision; accepted items are locked during that revision. After the audit, the implementation clips base-item weights to their defaults ± 0.2 and renormalizes if the total differs from one by more than 0.01. This means the final normalized weights need not remain within ± 0.2 of their defaults. An unresolved gated item becomes normal, an unresolved base item reverts to its base criterion, and an unsupported custom item is removed. Construct failure falls back to the base rubric.

Score requires every retained item in `m_scores`. Normal items take a continuous score in $[1, 5]$; gated items can additionally take zero when their fixed zero condition is met. A zero submission must include a same-dimension assistant quotation linked to a real message. The overall satisfaction always remains in $[1, 5]$. By default, full-dimension evidence coverage is requested in the prompt; the optional `turn_bound_dimensions` mode enforces at least one located evidence item per dimension. Neither mode establishes that a quotation semantically proves the agent's interpretation. Saved artifacts include rubrics, corrections, and final score evidence; the current pipeline does not persist complete Construct and Score tool message traces.

D Dataset Composition and Descriptive Analysis

D.1 Frozen Design and Partitioning

For the anonymous release, the intended crossing has 22 character cards, 16 user profiles, and five target models (1,760 possible combinations); 1,751 dialogues are saved, so nine combinations are absent. The frozen index is keyed by dialogue ID and links public and private records. The public record contains the character and user–assistant transcript used by evaluators. The private record contains simulator state, termination information, and the satisfaction label, and is joined only after predictions are saved. All five partitions are disjoint. The 48-dialogue stability anchor is a subset of development and is not added to the 1,751 total.

User profile selection. We chose 16 designed archetypes to cover contrasting reasons for entering a role-playing conversation, rather than independently sampling profile attributes. Nine emphasize immersive storytelling, creative collaboration, or companionship; seven broaden coverage to casual exploration, critical testing, functional use, and support seeking. Across them we vary initiative, communication style, attachment-related expectations, and explicit satisfaction and departure conditions. This coverage-oriented selection is not a population sample, so its composition should not be read as the prevalence of real-user types.

Table 13: Frozen partition sizes and counts of simulated satisfaction labels $S \in \{1, 2, 3, 4, 5\}$.

Partition	N	$S = 1$	$S = 2$	$S = 3$	$S = 4$	$S = 5$
Proxy train	353	75	72	42	82	82
Development	179	24	39	26	45	45
Test	1,111	241	221	116	277	256
Probe	86	23	14	1	29	19
Smoke	22	7	4	4	5	2
All	1,751	370	350	189	438	404

The five target-model counts are 352, 352, 352, 343, and 352. Each character occurs in 72–80 saved dialogues, and each user profile in 108–110. The class distribution is not uniform: $S = 3$ has 189 dialogues (10.8%), while $S = 4$ has 438 (25.0%). The test partition alone contains 241/221/116/277/256 examples across $S = 1, \ldots, 5$. These counts are properties of the frozen simulated corpus, not claims about the prevalence of satisfaction levels among real users.

D.2 Length and Exit Behavior

One exchange is one user message followed by one character reply. The index's `turn_count` counts individual messages; the analysis below counts character replies in the raw saved dialogues. The median is six exchanges, the mean is 19.18, the 75th percentile is 33, and the 90th percentile reaches the 60-reply cap. This broad range reflects the two session profiles, dealbreakers, engagement depletion, and the hard cap. It does not establish a match to a real-user length distribution.

Table 14: Observed dialogue lengths and termination reasons among all 1,751 saved simulations. The exit reasons are exclusive.

Exchanges	Count	%	Exit reason	Count	%
1–5	429	24.5	Dealbreaker	680	38.8
6–15	804	45.9	Engagement below threshold	704	40.2
16–30	71	4.1	Reply limit	367	21.0
31–60	447	25.5			

The label is associated with exit mode: among 680 dealbreaker exits, 620 receive $S \leq 2$, while 298 of 367 cap-limited dialogues receive $S \geq 4$. For engagement-depletion exits, 536 of 704 receive $S \geq 4$. This association is partly expected because the Layer B labeler sees BDI dealbreaker and satisfaction signals. It is therefore descriptive of the simulator, not independent evidence that exit mode predicts human satisfaction. The counterexamples also matter: 13 cap-limited dialogues receive $S \leq 2$, and eight dealbreaker exits receive $S = 4$. Length or termination alone is not the reference label. Tables 13 and 14 count the frozen partition membership and saved dialogue records.

E Prompt Templates

Double-bracketed strings below are placeholders replaced at runtime. The boxes translate the instruction-bearing portions of the Chinese prompts into English and retain their output contracts. They explain the experimental instructions but are not byte-for-byte prompts used in the runs.

E.1 BDI and Affect Updates

The first two templates are called after a character reply. The BDI analysis is conditioned on the private profile and prior state; the affect call receives the newly inferred BDI fields. Both return machine-readable JSON that is parsed before the next simulated-user utterance.

BDI cognitive update (translated excerpt)

```
You are this user in a role-playing chat application.
Profile, attachment style, goals, and departure conditions:
  [[PRIVATE_USER_PROFILE]]
Previous self-belief, belief about the character, intention,
and action tendency: [[PREVIOUS_BDI_AND_TENDENCY]]
Recent dialogue history: [[HISTORY_SUMMARY]]
Latest character reply: [[LATEST_CHARACTER_REPLY]]
Update your beliefs from this reply. Let the previous action
tendency inform how your intention continues or changes.
Output JSON only:
{"self_belief": "<current internal state, 1-3 sentences>",
 "env_belief": "<interpretation of character, 1-3 sentences>",
 "intention": "<next goal, 1-2 sentences>",
 "is_satisfied": <true|false>,
 "is_dealbreaker": <true|false>}
```

Affect appraisal (translated excerpt)

```
Profile: [[PRIVATE_USER_PROFILE]]
Current BDI: [[SELF_BELIEF]], [[ENV_BELIEF]], [[INTENTION]]
Latest character reply: [[LATEST_CHARACTER_REPLY]]
Mood baseline: [[MOOD_BASELINE]]
Estimate the emotion increment caused by this reply.
For ordinary dialogue, keep all eight values near zero;
give a large value only for a clear emotional trigger.
Each value is between 0 and 1. Output JSON only:
{"joy": 0.0, "trust": 0.0, "fear": 0.0,
 "surprise": 0.0, "sadness": 0.0, "disgust": 0.0,
 "anger": 0.0, "anticipation": 0.0}
```

The optional non-affective branch calls a separate six-component action-tendency prompt directly from BDI; the reported simulator configuration enables the affect-to-lookup-table branch. The affect prompt's "mostly neutral" directive is an instruction to the simulator, not a measured statistic of this benchmark.

E.2 User Simulation and Satisfaction Labeling

The user simulator receives a stable system prompt with persona, public character description, goals, style, and boundaries. A separate per-turn message supplies the updated BDI state and action tendency.

Simulated user (translated excerpt)

```
You are a real user of a role-playing chat application.
Your profile: [[USER_PROFILE]]
Public character description: [[PUBLIC_CHARACTER_CARD]]
Your goal and departure conditions: [[GOALS_AND_BOUNDARIES]]
Your chat style: [[STYLE_GUIDE]]
Write a short, natural next message, usually 5-20 Chinese
characters; occasionally 30-50. Discuss at most one or two
topics. Do not invent convenient objects or coincidences.
Default to plain speech. If useful, include at most one short
parenthesized action or expression; alternate with action-free
messages. Do not write long scene or internal-state narration.
Per-turn guidance:
  Self-belief: [[SELF_BELIEF]]
  Belief about the character: [[ENV_BELIEF]]
  Intention: [[INTENTION]]
  Dominant action tendency: [[TENDENCY]]
Let these states affect your behavior; do not repeat them aloud.
Output only the next user message.
```

The target role-playing model receives a separate system instruction built from the character card. The fallback character text uses core traits, speech style, verbal habits, and special points when a complete persona paragraph is unavailable.

Target character model (translated excerpt)

```
You are [[CHARACTER_NAME]], [[CHARACTER_IDENTITY]].
World: [[WORLD_SETTING]]
Persona and speech style: [[CHARACTER_PERSONA]]
Relationship to the user and its initial stage:
  [[RELATIONSHIP_AND_STAGE]]
Taboos that must not be violated: [[CHARACTER_TABOOS]]
Background facts: [[BACKGROUND_FACTS]]
Current scenario: [[SCENARIO]]
Reply as the character. Stay consistent and move the
conversation forward. Aim for 40-150 Chinese characters;
use parentheses for actions or expressions.
```

Before the persona-conditioned satisfaction judgment, an independent Layer A call scores each of four dialogue properties. This assessment provides advisory inputs, not a formula for Layer B.

Layer A dialogue-property judge (translated excerpt)

```
Evaluate [[DIMENSION_NAME]] in [[DIALOGUE]] against
[[PUBLIC_CHARACTER_CARD]] and [[DIMENSION_RUBRIC]].
Start from score 3. Inspect deduction conditions first.
Scores above 3 require explicit positive evidence;
scores below 3 require explicit failure evidence.
Read the full dialogue, then cite message turns such as T7.
Write the reason before assigning one integer score.
Output JSON only:
{"[[DIMENSION_KEY]]": {"reason": "<turn-linked reason>",
                       "score": <integer 1-5>}}
```

The withheld label is produced after the dialogue by a distinct judge. Its four dimension scores are advisory rather than a deterministic satisfaction formula; the user's simulated reactions and requirements can outweigh them.

Post-session satisfaction label (translated excerpt)

```
Adopt this user's perspective and rate satisfaction from 1 to 5.
High objective ability does not necessarily imply user satisfaction.
User profile: [[PRIVATE_USER_PROFILE]]
Needs and dealbreakers: [[END_CONDITIONS]]
Full dialogue: [[DIALOGUE]]
Four dimension scores (reference only): [[DIMENSION_SCORES]]
BDI satisfaction/dealbreaker signals: [[BDI_SUMMARY]]
Key internal-state excerpts: [[BDI_EXCERPTS]]
Explain whether the dialogue met this user's needs. Cite specific
turns and, if the user left early, identify the triggering condition.
Output JSON only:
{"sat_star": <integer 1-5>, "reason": "<turn-linked reason>"}
```

E.3 CURE CONSTRUCT, FEEDBACK, AND SCORE

The following boxes retain the stage separation and submission formats. The gated-item instruction is appended to the base stage prompt in the reported configuration. For Score, that appended rule overrides older global red-line language: a verified violation zeros only its rubric item.

Construct (translated excerpt)

```
Infer this user's preferences from their utterances in the completed
dialogue. The character card is given; do not infer a new character.
Read user turns, inspect surrounding dialogue when needed, and read
the character-grounded base rubric.
For each M1--M6 item, provide:
  dim, weight, criterion, levels 1--5, and verbatim user quotes.
Weight reflects this user's stated emphasis; criterion describes
what is good or bad for this user in observable dialogue terms.
Add M7 only for a distinct observable requirement outside M1--M6.
Its name must be specific; give it a nonzero weight and all levels.
Use a gated item only after an explicit refusal, prohibition, or
stop boundary; specify one observable zero condition.
Submit the complete rubric with submit_rubric:
{"user_archetype": "...", "user_preferences": "...",
 "user_type": <1-16>, "items": [[RUBRIC_ITEMS]]}
```

Feedback (translated excerpt)

```
Review each item of [[DRAFT_RUBRIC]] against [[DIALOGUE]]
and [[BASE_RUBRIC]]. Do not rewrite the rubric yourself.
Check whether the criterion and weight are supported by user
quotes. Check whether normal/gated is justified. A gated zero
must follow an explicit user boundary and apply only to later
character behavior. Character traits alone cannot justify it.
Return exactly one checklist entry per retained dimension:
{"items": [{"dim": "M1", "passed": true,
           "issues": [], "recommended_rubric_type": "normal",
           "recommended_zero_condition": ""}, ...]}
Rejected items may be revised; accepted items remain locked.
```

Score (translated excerpt)

```
Use the fixed final rubric [[FINAL_RUBRIC]] and public dialogue
[[DIALOGUE]]. Do not change criteria, weights, or dimensions.
Read dialogue windows; sample relevant turns for M1--M4
annotations and optionally compute temporal statistics.
Sample 5-8 early, middle, and late turns, including salient
failures or peaks.
Use the tools as evidence aids; do not invent turns or quotations.
Score every retained item. For M5, M6, and optional M7, use
dialogue evidence without a dedicated turn-annotation tool.
A short dialogue can end satisfactorily and a long one can be
unpleasant; do not infer satisfaction from length alone.
A normal item receives 1--5. A gated item receives zero only
if its fixed condition is met; cite the same-dimension
assistant message verbatim. Other items remain independent.
Submit through submit_satisfaction:
{"satisfaction": <number 1-5>,
 "m_scores": {"M1": <score>, ..., "M6": <score>},
 "rationale": "<rubric- and evidence-linked explanation>",
 "evidence": [[MESSAGE_BOUND_EVIDENCE]],
 "confidence": <number>}
```

E.4 Rubric Initialization, Revision, and Gated Variant

The following supplementary fragments specify the base-rubric input, the narrow revision operation, and the gated-item rule used in the reported ablation.

Base rubric criteria (translated excerpt)

```
Derive general, character-grounded criteria from the given card.
This is a base rubric, not a particular user's preferences.
Character card: [[PUBLIC_CHARACTER_CARD]]
Trait hierarchy: [[CARDINAL_CENTRAL_SECONDARY_TRAITS]]
Expected facets: [[FACET_LIST]]
For each M1--M6, state what good performance means for this
character, including its central constraints.
Output JSON only:
{"M1": "<trait consistency; cardinal traits cannot break>",
 "M2": "<topic evolution>", "M3": "<facet development>",
 "M4": "<initiative and pacing>",
 "M5": "<emotionally appropriate response>",
 "M6": "<memory and continuity>"}
```

Targeted Construct revision (translated excerpt)

```
You remain the Construct Agent. Feedback has reviewed the rubric.
Revise only [[TARGET_DIMS]]; [[LOCKED_DIMS]] have passed.
Do not change the inferred user profile or any locked item.
For each rejected item, address the checklist's specific issues
in criterion, weight, quotations, type, or zero condition.
Feedback checklist: [[FEEDBACK_CHECKLIST]]
Current rubric: [[CURRENT_RUBRIC]]
Call submit_rubric_revision with the revised target items only.
```

Gated-item rule appended to Construct (translated)

```
Every item has rubric_type = normal or gated.
Normal uses levels 1--5 and an empty zero_condition.
Gated is allowed only after the user explicitly refuses,
prohibits, or asks to stop a behavior in the dialogue.
Its single zero_condition must describe observable character
behavior after that boundary. Only this item becomes zero
if the condition occurs; other items keep their own scores.
The public character card and base red lines cannot alone
justify a gated item. If evidence is insufficient, use normal.
```