

REMP: A Structured Event-Probing Benchmark for Long-Horizon Role-Playing Dialogue

Qianxi Yu¹ Yuhao Dan¹ Qingchun Bai²
Qin Chen^{1*} Jie Zhou¹ Liang He¹

¹School of Computer Science and Technology, East China Normal University

²Shanghai Open University

51275901019@stu.ecnu.edu.cn qchen@cs.ecnu.edu.cn

Abstract

Large language models (LLMs) are increasingly used in role-playing applications, yet evaluating whether they maintain character-consistent behavior over extended interactions remains challenging. Existing benchmarks mostly focus on short exchanges, static long-context inputs, or isolated aspects of role-playing, making long-horizon failures difficult to diagnose. We present REMP (Relationship-Emotion-Memory-Persona), a structured event-probing benchmark for long-horizon role-playing dialogue. REMP constructs multi-chapter dialogue trajectories from character profiles and embeds diagnostic probe events targeting persona consistency, emotional consistency, memory integrity, and relational continuity. We further introduce REMP-EVAL, an anchor-based multi-turn evaluation protocol that assesses model behavior after event triggers using dimension-specific references. The current release contains 160 characters organized into 80 dialogue trajectories, each averaging about 60 sessions and 800 turns. Experiments on 11 LLMs show that memory integrity and emotional consistency are the main bottlenecks, indicating that long-horizon role-playing requires not only long-context retrieval but also sustained affective and relational consistency.

1 Introduction

Large language models (LLMs) are increasingly deployed in role-playing settings, where they are expected to adopt fictional personas, preserve character-specific language styles, and sustain character-driven conversations (Chen et al., 2024b). These capabilities support applications ranging from interactive storytelling (Teleki et al., 2025) to long-term companionship (Li et al., 2025). As role-playing systems move from isolated exchanges to sustained interactions, evaluation must ask not only whether a model can imitate a character briefly,

*Corresponding author.



Figure 1: An example of a structured event-probing pair in REMP. Green marks the persona constraint to be preserved, red marks the memory implanted in Session 1, and cyan marks the later test trigger that violates the implanted memory. The test event probes whether the model can recall Miyuki’s promise, express appropriate disappointment, and maintain Kaguya’s calm but sarcastic persona across sessions.

but whether it can maintain consistent behavior as dialogue history, emotional states, and relationships evolve. Figure 1 illustrates a structured event-probing example in REMP.

Existing benchmarks have advanced the evalu-

Benchmark	Character	Conversation			Evaluation	
	Num.	Avg. Turns	Avg. Sessions	Language	Multi-Turn	Multi-Dim.
<i>Role-Playing Benchmarks</i>						
CharacterEval	77	4.7	1.0	Zh		✓
CharacterBench	3,956	5.7	1.0	Zh/En		✓
RoleLLM	100	1.0	1.0	Zh/En		
LifeChoice	1,462	1.0	1.0	En		
SocialBench	286	12.4	1.0	Zh/En		✓
CoSER	17,966	6.6	1.0	En	✓	✓
<i>Long-Context Benchmarks</i>						
MOOM	48	618.69	1.0	Zh		
LoCoMo	18	304.9	19.3	En		
Persistent Personas	8	102	1.0	En	✓	✓
REMP (Ours)	160	788.3	59.4	Zh/En	✓	✓

Table 1: Comparison of role-playing and conversational benchmarks. “Multi-Turn” indicates whether the benchmark explicitly evaluates behavior over multiple turns within the same interaction trajectory. “Multi-Dim.” indicates whether the benchmark evaluates multiple aspects of role-playing behavior, such as persona, emotion, memory, or relationship.

ation of role-playing and long-context language modeling, but they leave important gaps for long-horizon role-playing dialogue. Role-playing benchmarks such as CharacterEval (Tu et al., 2024), RoleEval (Shen et al., 2023), and RoleLLM (Wang et al., 2024a) mainly evaluate short prompts, question answering, or localized aspects of character portrayal. Long-context benchmarks such as LongBench (Bai et al., 2024), L-Eval (An et al., 2024), and RULER (Hsieh et al., 2024) emphasize retrieval and reasoning over long static inputs, rather than dynamic consistency in multi-turn interaction. The closest conversational benchmarks are LoCoMo (Maharana et al., 2024) and MOOM (Chen et al., 2025), which study long-term conversational memory and ultra-long role-playing contexts. However, their evaluation is still primarily centered on memory-related capabilities, such as recalling past events, extracting relevant memories, or answering memory probing questions. They do not provide an interactive, anchor-based framework for diagnosing how persona, emotion, memory, and relationship consistency jointly break down at the event level. Table 1 summarizes these differences.

Long-horizon role-playing is difficult not only because the context is long, but because the relevant interaction state is continually updated. A capable model must retrieve prior events and determine how they should affect persona expression, emotional response, memory use, and relational behavior. These failures are often entangled: a memory error may lead to an inappropriate emotional response, and a relational mismatch may also expose per-

sona inconsistency. Such failures may only emerge after several subsequent turns, making isolated prompts, memory-only probes, or holistic scores insufficient for diagnosis. This motivates interactive event-level probing with dimension-specific anchors, where a diagnostic event is triggered and the model’s subsequent behavior is evaluated in context.

To address this gap, we present REMP, a structured event-probing benchmark for dynamic evaluation of long-horizon role-playing dialogue. REMP constructs Chinese and English multi-chapter trajectories from character profiles, covering 80 conversations with an average of about 60 sessions and 800 turns each. It embeds diagnostic probe events across relationship stages to evaluate whether persona, emotion, memory, and relationship states are preserved and updated as the narrative unfolds. We further introduce REMP-EVAL, an interactive multi-turn evaluation framework that uses structured references and event-specific anchors to assess model behavior after event triggers, making failures easier to interpret and localize.

Together, REMP and REMP-EVAL provide, to our knowledge, the first unified benchmark and evaluation framework that combines long-horizon multi-session role-playing trajectories, dynamic event-level probing, and multi-dimensional diagnosis across persona, emotion, memory, and relationship. Our contributions are threefold:

1. We present REMP, a long-horizon role-playing benchmark with 80 multi-session dialogue trajectories and structured diagnostic events target-

- ing persona, emotion, memory, and relationship.
2. We develop REMP-EVAL, an anchor-based multi-turn evaluation framework that improves the interpretability and reproducibility of long-horizon role-playing assessment.
 3. We conduct a systematic empirical study of 11 LLMs on REMP, revealing fine-grained performance patterns and failure modes in long-horizon role-playing.

All code and data are available at <https://github.com/DreamFragrance/REMP>.

2 Related Work

Role-Playing Evaluation. Recent benchmarks have substantially advanced the evaluation of LLM role-playing, but most assess character behavior through short prompts, question answering, choice-style tasks, or localized interaction scenarios. Early work such as PersonaChat (Zhang et al., 2018) introduced persona-conditioned dialogue, but does not evaluate long-horizon character consistency. CharacterEval (Tu et al., 2024), CharacterBench (Zhou et al., 2025), and SocialBench (Chen et al., 2024a) evaluate dialogue-based character portrayal or social interaction, covering aspects such as persona consistency, linguistic style, and social appropriateness. Other benchmarks focus on narrower capabilities: RoleEval (Shen et al., 2023) emphasizes character knowledge through question answering, RoleLLM (Wang et al., 2024a) focuses on style injection, LifeChoice (Xu et al., 2025) evaluates persona-consistent decision making in single-turn settings, and InCharacter (Wang et al., 2024b) measures personality fidelity using psychological instruments. Complementing these behavioral evaluations, Dan et al. (Dan et al., 2026) introduce TraitTrace and identify compact Transformer circuits underlying Big Five personality-consistent behavior, providing a mechanistic account of personality expression in LLMs. CoSER (Wang et al., 2025) further provides large-scale character data for training and evaluation. Most role-playing benchmarks evaluate character fidelity under fixed system prompts and largely treat personas as static. Persistent Personas (Luz de Araujo et al., 2026) extends this setting to long persona-assigned dialogues, but still focuses on fidelity to fixed assigned personas rather than plot progression, evolving relationship stages, or event-specific dependencies. REMP instead evaluates whether persona, emotion, memory, and relationship states are preserved

and updated as narrative events unfold across multi-session trajectories.

Long-Context Evaluation. Long-context benchmarks such as LongBench (Bai et al., 2024), L-Eval (An et al., 2024), RULER (Hsieh et al., 2024), and ∞ Bench (Zhang et al., 2024) primarily evaluate retrieval and reasoning over long but static inputs, rather than dynamic consistency in interactive dialogue. More closely related conversational benchmarks include LoCoMo (Maharana et al., 2024) and MOOM (Chen et al., 2025), which study long-term conversational memory and ultra-long role-playing contexts. However, their evaluation is still mainly centered on memory-related capabilities, such as recalling past events, extracting relevant memories, or answering memory probing questions. They do not directly diagnose how persona, emotion, memory, and relationship consistency jointly break down during multi-turn role-playing. REMP addresses this gap by combining long-horizon multi-session trajectories with structured event probes and anchor-based multi-dimensional evaluation.

Emotion Evaluation. Emotional consistency in role-playing cannot be reduced to surface-level sentiment matching, because appropriate affect depends on the character’s traits, the eliciting event, and the evolving interaction context. REMP therefore draws on established theories of affect to construct emotion-related probes and anchors. The OCC cognitive theory of emotions (Ortony et al., 1988) models emotion elicitation through appraisals of events, agents, and objects, while the VAD model (Russell, 1980; Mehrabian, 1996) represents affective states along valence, arousal, and dominance dimensions. These theories provide structured reference signals for REMP-EVAL, enabling anchor-based assessment of whether a model’s emotional response is consistent with both the target character and the dialogue context.

Relationship Evaluation. Relationship development is an underexplored aspect of long-horizon role-playing. Unlike static persona attributes, relational states evolve as characters accumulate shared experiences, negotiate boundaries, and adjust intimacy over time. Existing role-playing benchmarks rarely evaluate whether models adapt their behavior to such evolving interpersonal states. To support relational evaluation, REMP draws on the Levinger model (Levinger, 1980), which charac-

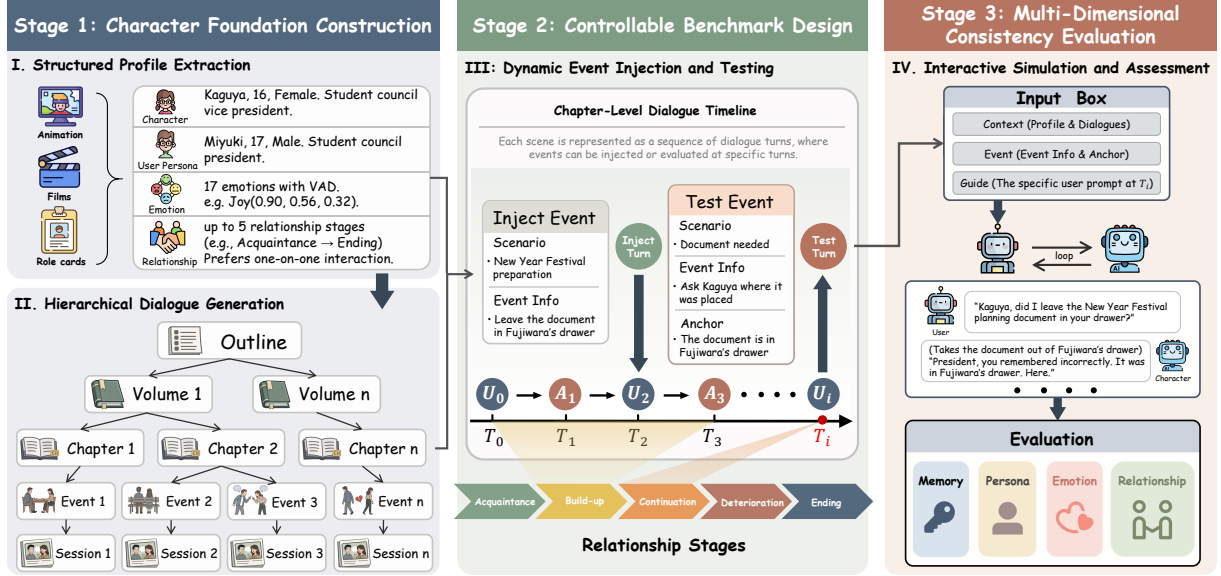


Figure 2: Overview of the REMP pipeline. **Stage 1** involves character profiling and hierarchical dialogue generation. **Stage 2** defines a dialogue timeline of user prompts (U_i) and assistant responses (A_j), where probe events are specifically injected via U_i . **Stage 3** evaluates subsequent responses (A_j) for consistency across persona, emotion, memory, and relationship through an interactive simulation that proceeds from an event-conditioned U_i until the session ends.

terizes relationship progression through multiple stages. This framework enables stage-aware anchors for REMP-EVAL, specifying how tone, disclosure, intimacy, and behavioral choices should change across extended interactions.

3 Benchmark Construction

REMP is designed as a structured event-probing benchmark for diagnosing long-horizon role-playing consistency. Figure 2 provides an overview of the construction pipeline. Each dialogue trajectory is organized hierarchically into volumes, chapters, and sessions: volumes represent coarse narrative phases, chapters provide natural units for event insertion, and sessions serve as the basic units for later evaluation. This hierarchy preserves long-range narrative coherence while enabling localized probes of persona, emotion, memory, and relationship consistency. Each probe event is attached to a target session and paired with dimension-specific anchors that specify the behavioral constraints expected in that event.

3.1 Character Profiling and Outline Construction

Each trajectory begins with a pair of role-playing characters: an assistant character played by the evaluated model and a user character simulated during evaluation. Raw character materials are

converted into structured profiles covering identity, background, behavioral tendencies, linguistic style, knowledge boundaries, interpersonal preferences, and role-specific constraints. Beyond static profiles, REMP also constructs affective and relational states: emotional tendencies are represented with OCC-style emotion categories and Valence–Arousal–Dominance (VAD) values, while relationship progression is annotated using the Levinger ABCDE stages. These profiles and state annotations provide the basis for trajectory generation and consistency evaluation. Representative prompt templates are provided in Appendix D.1.

Given the structured profiles, REMP generates trajectory outlines in a top-down manner, first producing a high-level storyline and then refining it into volume, chapter, and session outlines. Chapter-level relationship-stage annotations specify how the assistant character should behave toward the user at different points in the long-horizon interaction. To diversify interaction patterns, REMP organizes trajectories into seven plot archetypes, as shown in Figure 3. A summary of benchmark statistics is provided in Appendix A.

3.2 Diagnostic Probe and Anchor Construction

A central component of REMP is structured event probing. Rather than evaluating arbitrary dialogue

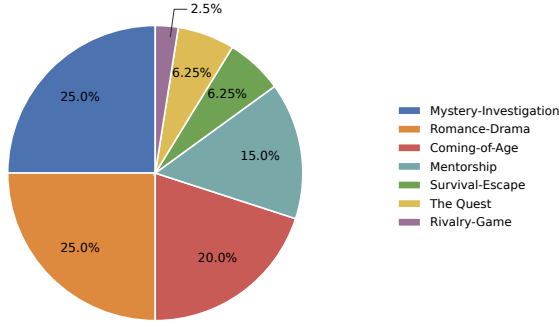


Figure 3: Distribution of REMP trajectories across seven plot archetypes.

snippets or assigning holistic scores to entire conversations, REMP embeds diagnostic events into long trajectories so that specific consistency failures can be elicited and localized.

REMP defines four event categories: persona, emotion, memory, and relationship (Table 2). Persona events test whether the model preserves character facts, knowledge boundaries, values, and identity-related behavior. Emotion events test whether affective responses match the event and the target character. Memory events test whether information introduced earlier can be used. Relationship events test whether behavior remains appropriate to the current relationship stage.

Each event is paired with anchors for the relevant dimensions. Persona anchors specify required character facts, style, values, or knowledge constraints. Emotion anchors specify OCC-predicted emotions and expected VAD shifts. Memory anchors record implanted information such as facts, promises, object locations, or decisions. Relationship anchors define expected boundaries, disclosure level, and desire alignment under the current relationship stage.

3.3 Reference Generation

Given trajectory outlines and probe events, REMP generates chapter-level reference interactions rather than unconstrained turn-by-turn dialogues. This design preserves narrative coherence while ensuring that diagnostic probes are naturally embedded in their target sessions. For each chapter c , reference generation conditions on the assistant profile P_a , user profile P_u , chapter outline O_c , relationship stage z_c , recent dialogue history $H_c^{(20)}$, previous memory state M_{c-1} , and probe specification ξ_c :

$$R_c = G_{\text{ref}}(P_a, P_u, O_c, z_c, H_c^{(20)}, M_{c-1}, \xi_c), \quad (1)$$

Event Type	Sub-dimensions
Emotion (6)	intense_stimulus dominance_stress prospect_resolution agency_shift social_comparison vicarious_distress
Persona (4)	fact_consistency knowledge_boundary value_alignment identity_stability
Memory (2)	direct_recall distractor_recall
Relationship (3)	boundary_maintenance disclosure_calibration desire_alignment

Table 2: Event categories and sub-dimensions in REMP.

where R_c is the generated chapter-level reference interaction. The probe specification ξ_c defines the event scenario, trigger condition, and event-specific information, while dimension-specific anchors are reserved for evaluation rather than used as canonical target responses.

After each chapter, a memory-management agent extracts salient information from R_c and updates the persistent memory state:

$$M_c = U_{\text{mem}}(M_{c-1}, R_c). \quad (2)$$

The updated memory records information that should remain available in later chapters, such as salient events, new facts, commitments, object locations, relationship changes, and unlocked character knowledge. To control context growth, REMP compresses memory at volume boundaries:

$$M_{c+1}^{\text{init}} = \begin{cases} C_{\text{mem}}(M_c), & c \in \mathcal{V}_{\text{end}}, \\ M_c, & \text{otherwise.} \end{cases} \quad (3)$$

where $\mathcal{V}_{\text{end}}$ denotes chapters at volume boundaries and M_{c+1}^{init} initializes the next chapter.

For probe events that substantially alter the storyline, REMP creates deviation branches: the main branch preserves the intended narrative trajectory, while the deviation branch captures the event-triggered alternative used for evaluation. All generated artifacts, including character profiles, story outlines, probe events, anchors, memory implant-trigger pairs, and reference dialogues, are manually reviewed before release. The full quality-control protocol is described in Appendix B.

4 REMP-EVAL: Anchor-Based Multi-Turn Evaluation

REMP-EVAL evaluates each target session through a dimension-aware multi-agent protocol. For each probe event, the evaluated model acts as the assistant character, while a simulated user triggers the event according to the reference scenario. The resulting interaction is then evaluated against event-specific anchors for persona consistency, emotional consistency, memory integrity, and relational continuity.

4.1 Multi-Agent Evaluation Protocol

Each REMP-EVAL instance involves three agents: a UserAgent, a CharacterAgent, and a JudgeAgent. The UserAgent simulates the user character and triggers the diagnostic event, while the CharacterAgent, instantiated by the model under evaluation, role-plays the target assistant character using all preceding dialogue context available in the trajectory. The interaction starts from the probe-triggering user turn rather than a freely generated utterance, ensuring that each evaluation instance contains the intended diagnostic event while still requiring the model to continue role-playing over subsequent turns. The JudgeAgent evaluates the resulting interaction under the LLM-as-judge setting (Zheng et al., 2023), using DeepSeek-R1 based on DeepSeek-V3.2 as the primary judge backbone, and scores model behavior against dimension-specific anchors rather than exact surface-form matches to reference dialogues. We validate this judge choice and compare alternative backbones in Section 5.5.

4.2 Dimension-Aware Scoring

Although each probe event targets one primary dimension, model responses often reveal several forms of consistency or failure simultaneously. For example, a memory event may also reveal whether the model preserves the character’s speaking style, respects knowledge boundaries, or reacts with an appropriate emotional tone. REMP-EVAL therefore scores the primary dimension of each event together with relevant secondary dimensions, allowing entangled role-playing failures to be diagnosed more precisely.

The primary–secondary distinction is used when aggregating event-level judgments into dimension-level scores. For each dimension d , evaluations from events in which d is the primary dimension

receive a larger weight than those in which d is a secondary dimension. The aggregated score for dimension d is computed as

$$\bar{s}_d = \frac{w_p \sum_{e \in E_d^{\text{pri}}} s_{e,d} + w_s \sum_{e \in E_d^{\text{sec}}} s_{e,d}}{w_p |E_d^{\text{pri}}| + w_s |E_d^{\text{sec}}|}, \quad (4)$$

where $s_{e,d}$ denotes the score assigned to dimension d for event e , and E_d^{pri} and E_d^{sec} denote the sets of events in which d is evaluated as the primary and secondary dimension, respectively. We set $w_p = 1.0$ and $w_s = 0.3$, emphasizing performance on events specifically designed to probe dimension d while still incorporating collateral evidence from other events.

Scoring rubric. For each dimension, the event anchor serves as the reference-level criterion. A score of 1 indicates that the response misses or contradicts the anchor. A score of 2 indicates partial satisfaction with major omissions or clear inconsistencies. A score of 3 denotes a broadly relevant but incomplete, generic, or weakly grounded response. A score of 4 is assigned when the response satisfies the anchor at the reference level. A score of 5 is reserved for responses that exceed the reference-level expectation by showing richer, more nuanced, or more robust consistency without introducing contradictions.

Memory scoring. Memory integrity is additionally reported as a binary judgment for memory-specific analysis, where $x = 1$ indicates successful recall or use of the implanted information and $x = 0$ indicates failure. To aggregate memory with the other dimensions, the binary memory score is mapped to the same 1–5 scale:

$$y = 1 + 4x, \quad (5)$$

where y is the mapped memory score used in joint analysis. This preserves the binary nature of memory evaluation while enabling unified aggregation with persona, emotion, and relationship scores.

5 Experiments

5.1 Setup

We evaluate 11 language models spanning the Qwen (Qwen Team, 2026), Gemma (Team et al., 2026), and Llama (Grattafiori et al., 2024) families, together with GLM-4-9B (GLM et al., 2024) and DeepSeek-V3.2 (DeepSeek-AI et al., 2025) as additional representative models. We use DeepSeek-R1 (based on DeepSeek-V3.2) as the JudgeAgent.

Model	Persona Consistency	Emotional Consistency	Memory Integrity [†]	Relational Continuity	Overall
<i>Small ($\leq 4B$)</i>					
Gemma-4-E2B	2.69 ± 0.07	2.42 ± 0.05	2.28 ± 0.04 (32.0%)	2.95 ± 0.12	2.59 ± 0.07
Llama-3.2-3B	2.26 ± 0.02	2.06 ± 0.02	2.18 ± 0.09 (29.5%)	2.52 ± 0.05	2.26 ± 0.05
Gemma-4-E4B	3.17 ± 0.08	2.83 ± 0.02	2.60 ± 0.06 (40.0%)	3.25 ± 0.16	2.96 ± 0.07
Qwen3.5-4B	3.37 ± 0.17	3.07 ± 0.08	2.70 ± 0.03 (42.5%)	3.44 ± 0.21	3.14 ± 0.12
<i>Medium (8–9B)</i>					
Llama-3.1-8B	2.59 ± 0.03	2.35 ± 0.03	2.46 ± 0.02 (36.5%)	2.86 ± 0.08	2.57 ± 0.04
GLM-4-9B	2.65 ± 0.03	2.36 ± 0.01	2.44 ± 0.04 (36.0%)	2.88 ± 0.07	2.58 ± 0.02
Qwen3.5-9B	3.68 ± 0.14	3.36 ± 0.05	2.84 ± 0.08 (46.0%)	3.65 ± 0.13	3.38 ± 0.09
<i>Large (27B+)</i>					
Qwen3.5-27B	3.93 ± 0.08	3.61 ± 0.02	3.49 ± 0.07 (62.3%)	3.84 ± 0.06	3.72 ± 0.04
Gemma-4-31B	4.03 ± 0.04	3.62 ± 0.09	3.49 ± 0.11 (62.3%)	3.97 ± 0.04	3.78 ± 0.07
Llama-3.3-70B	2.81 ± 0.20	2.54 ± 0.17	2.85 ± 0.10 (46.3%)	2.96 ± 0.22	2.79 ± 0.17
DeepSeek-V3.2	3.61 ± 0.04	3.19 ± 0.04	3.35 ± 0.04 (58.8%)	3.65 ± 0.06	3.45 ± 0.02

Table 3: Main results across four dimensions. The table reports mean scores with standard deviations. **Memory Integrity[†]**: mapped to the 1–5 scale via $y = 1 + 4x$; values in parentheses report the original binary accuracy. **Overall**: mean of all four dimensions.

Each model is evaluated on all 80 dialogue trajectories, yielding 1,195 evaluation sessions per model in each pass. To reduce variance, we run three independent evaluation passes and report mean scores with standard deviations. Full implementation details are provided in Appendix C.

5.2 Main Results

Table 3 presents the main results across the four evaluation dimensions. Overall performance tends to improve with model scale, but the gains are strongly model-family dependent. Gemma-4-31B achieves the best overall score, while Qwen models remain consistently strong across size tiers. In contrast, Llama-3.3-70B underperforms several smaller models, suggesting that scale alone does not determine long-horizon role-playing ability.

Across dimensions, relational continuity and persona consistency are relatively stronger, whereas memory integrity and emotional consistency remain the main bottlenecks. This indicates that models can often preserve broad character traits and relationship tone, but still struggle to retrieve earlier information and produce contextually appropriate emotional responses over long interactions. Additional sub-dimension and character-attribute analyses are provided in Appendix F.

5.3 Memory Decay over Distance

We analyze memory performance over the turn span between memory implantation and test trigger. Figure 4 compares direct recall and distractor recall accuracy across binned span ranges.

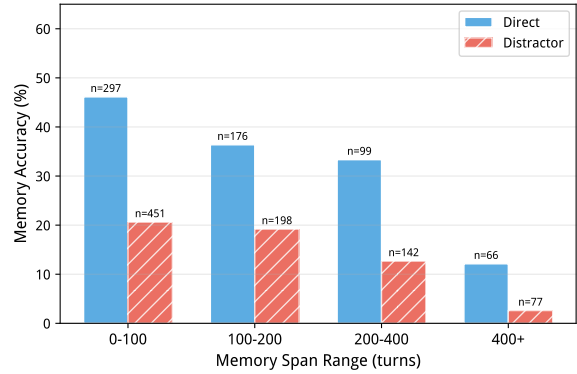


Figure 4: Memory accuracy as a function of the turn span between implantation and test trigger. Direct recall and distractor recall are reported across binned span ranges, with results aggregated over all models.

Memory accuracy decreases monotonically as the span increases, revealing a long-horizon decay pattern. Direct recall consistently outperforms distractor recall, indicating that retrieval is harder when the model must identify the target memory among competing distractors. The decline becomes especially pronounced beyond 400 turns, highlighting the difficulty of maintaining usable memory over extended dialogue horizons.

5.4 Emotion Difficulty across OCC Categories

We group emotion events by OCC category to identify which affective responses are most challenging for current models. The category definitions and mappings are provided in Appendix D.2. As shown in Figure 5, *Prospect* and *Attribution/Self* emotions receive the highest consistency scores,

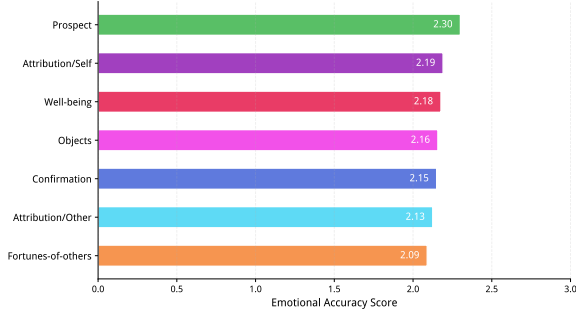


Figure 5: Emotional consistency grouped by OCC category.

whereas *Confirmation* and *Fortunes-of-others* are among the lowest-scoring categories.

This pattern is consistent with appraisal-theoretic accounts of emotion (Scherer, 2001). *Prospect* and *Attribution/Self* emotions can often be inferred from local cues, such as anticipated outcomes or self-directed evaluations. By contrast, *Confirmation* emotions require the model to maintain a prior expectation and update its affective response once the outcome is resolved. *Fortunes-of-others* emotions are more socially grounded, requiring the model to appraise another agent’s outcome in relation to the character’s own goals and relationship state (Manstead and Fischer, 2001; Boecker et al., 2022). These results suggest that models handle local and self-focused emotions more reliably than emotions that require temporal belief updates or socially embedded appraisal.

5.5 Validation of REMP-EVAL

We validate REMP-EVAL on 120 evaluation sessions, with 30 sessions sampled from each dimension, covering Chinese and English trajectories, different difficulty levels, and all four diagnostic dimensions. Six human annotators are divided into two groups, each evaluating 60 sessions using the same anchors and rubric as the JudgeAgent. We compare the preferences inferred from JudgeAgent scores with human majority-vote preferences and report preference alignment and mean absolute error (MAE). Table 4 reports an ablation study across three evaluation settings and three judge backbones (DeepSeek-R1 (Guo et al., 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2025), GPT-5.4-mini (OpenAI, 2026)). REMP-EVAL achieves the highest alignment across all judge backbones, while DeepSeek-R1 obtains the best alignment and lowest MAE under the full setting, supporting its use as the primary JudgeAgent.

Judge	Setting	Alignment (%)	MAE
DeepSeek-R1	REMP-EVAL	80.3	0.88
	w/o Anchors	23.0	2.28
	w/o Interactive	57.4	1.05
DeepSeek-V3.2	REMP-EVAL	67.2	0.98
	w/o Anchors	29.5	2.15
	w/o Interactive	65.6	1.01
GPT-5.4-mini	REMP-EVAL	70.5	1.12
	w/o Anchors	16.6	1.61
	w/o Interactive	51.1	1.10

Table 4: Validation of REMP-EVAL across judge backbones and ablation settings. Alignment measures agreement between preferences inferred from JudgeAgent scores and human majority-vote preferences, while MAE measures score-level calibration against human mean scores.

Without anchors, REMP-EVAL reduces to holistic judging, which substantially weakens discriminability: the judge assigns near-identical scores to different models and achieves the worst alignment with human preferences. Removing multi-turn interaction also degrades preference alignment, suggesting that single-turn evaluation lacks sufficient context to diagnose role-playing failures that unfold across subsequent turns.

6 Conclusion

We present REMP, a structured event-probing benchmark resource for long-horizon role-playing dialogue, and REMP-EVAL, an anchor-based multi-turn framework for evaluating persona consistency, emotional consistency, memory integrity, and relational continuity. By embedding controlled diagnostic events into multi-session trajectories, REMP provides the basis for localized analysis of how role-playing behavior breaks down over time, while REMP-EVAL provides interpretable assessment through event-specific anchors.

Our findings show that sustained role-playing is not merely a long-context retrieval problem. Although current models can often preserve broad persona traits and relational patterns, they still struggle to maintain dynamic interaction states. These results point to a broader challenge for future role-playing agents: they must maintain not only what has happened, but also how past events should shape the character’s affect, behavior, and relationship with the user across extended interactions.

Ethical Considerations

REMP is intended solely as a research benchmark for diagnosing long-horizon role-playing behavior. It should not be interpreted as evidence that a model is safe for real-world companionship or other high-stakes interactive deployment. The benchmark includes both original characters and characters inspired by existing fictional works. We do not claim ownership of any third-party characters, settings, names, or trademarks, and we do not redistribute original copyrighted texts, scripts, game dialogue, transcripts, or other source materials. For IP-associated scenarios, the released data contains only benchmark-specific structured profiles, generated dialogue trajectories, diagnostic probes, and annotations created for evaluation purposes.

Several components of REMF are generated or assisted by LLMs and then manually reviewed for coherence, diagnostic validity, and safety. The benchmark does not use private user conversations or personally identifiable information. Because REMF-EVAL relies partly on LLM-based user simulation and LLM-based judging, its scores should be interpreted as diagnostic signals rather than definitive measures of model reliability. We therefore recommend using REMF-EVAL together with human evaluation and robustness analyses when assessing models for practical deployment.

The released artifact is provided for research and evaluation purposes. The code, evaluation scripts, data schemas, prompt templates, and documentation are released under the MIT License. The benchmark data, including generated character profiles, dialogue trajectories, diagnostic probes, event anchors, annotations, and sample model outputs, is released under a research-only non-commercial license. Users may use the benchmark for academic research, model evaluation, and reproducibility studies, but may not use it to redistribute, reproduce, or commercialize third-party characters, copyrighted source materials, or derivative entertainment products. The release does not include original copyrighted texts, scripts, game dialogue, transcripts, or other source materials. Users are responsible for complying with the licenses and terms of any external datasets, models, or software artifacts used together with REMF.

Limitations

REMF has several limitations. First, the current benchmark is restricted to two-party, text-only interaction, leaving multi-party role-playing and multimodal settings for future work. Second, the benchmark focuses on fictional and narrative-style role-playing scenarios, so its findings may not fully generalize to task-oriented, therapeutic, educational, or real-world companion-agent settings. Finally, although we include a small-scale human validation of LLM-based judging, broader human evaluation is still needed to more fully validate the proposed evaluation framework.

Acknowledgments

This research is funded by the National Natural Science Foundation of China (Nos. 62477010, 62577022, and 62307028), the Shanghai General AI Foundation Models Program (No. 2025SHZDZX026D24), and the Shanghai Science and Technology Innovation Action Plan (No. 24YF2710100).

References

- Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. [L-eval: Instituting standardized evaluation for long context language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [LongBench: A bilingual, multi-task benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.
- Lea Boecker, David D. Loschelder, and Sascha Topolinski. 2022. [How individuals react emotionally to others’ \(mis\)fortunes: A social comparison framework](#). *Journal of Personality and Social Psychology*, 123(1):55–83.
- Margaret M. Bradley and Peter J. Lang. 1999. [Affective norms for english words \(anew\): Instruction manual and affective ratings](#). Technical report, The Center for Research in Psychophysiology, University of Florida.

- Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. [Social-Bench: Sociality evaluation of role-playing conversational agents](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. [From persona to personalization: A survey on role-playing language agents](#). *Transactions on Machine Learning Research*. Survey Certification.
- Weishu Chen, Jinyi Tang, Zhouhui Hou, Shihao Han, Mingjie Zhan, Zhiyuan Huang, Delong Liu, Jiawei Guo, Zhicheng Zhao, and Fei Su. 2025. [MOOM: Maintenance, organization and optimization of memory in ultra-long role-playing dialogues](#). *Preprint*, arXiv:2509.11860.
- Yuhao Dan, Lang Yu, Jiaju Lin, Qin Chen, Jie Zhou, Junfeng Tian, Qingchun Bai, and Liang He. 2026. [A mechanistic interpretability perspective on personality in large language models](#). *Information Processing & Management*, 63(8):104962.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. [RULER: What’s the real context size of your long-context language models?](#) In *First Conference on Language Modeling*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, pages 611–626, New York, NY, USA. Association for Computing Machinery.
- George Levinger. 1980. [Toward the analysis of close relationships](#). *Journal of Experimental Social Psychology*, 16(6):510–544.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. [Hello again! LLM-powered personalized agent for long-term dialogue](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276, Albuquerque, New Mexico. Association for Computational Linguistics.
- Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. 2026. [Persistent personas? role-playing, instruction following, and safety in extended interactions](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5329–5359, Rabat, Morocco. Association for Computational Linguistics.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.
- Antony S R Manstead and Agneta H Fischer. 2001. [Social appraisal: The social world as object of and influence on appraisal processes](#). In *Appraisal Processes in Emotion: Theory, Methods, Research*, pages 221–232. Oxford University Press.
- Albert Mehrabian. 1996. [Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament](#). *Current psychology*, 14(4):261–292.
- OpenAI. 2026. [GPT-5.4 mini and nano](#). OpenAI release page. Accessed: 2026-05-17.
- Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#). Official release page.
- James A. Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Klaus R Scherer. 2001. [Appraisal considered as a process of multilevel sequential checking](#). In *Appraisal Processes in Emotion: Theory, Methods, Research*, pages 92–120. Oxford University Press.
- Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. [RoleEval: A bilingual role evaluation benchmark for large language models](#). *Preprint*, arXiv:2312.16132.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. 2025. [A survey on LLMs for story generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966, Suzhou, China. Association for Computational Linguistics.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. [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)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. [CoSER: Coordinating LLM-based persona simulation of established roles](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. [InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.
- Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025. [Character is destiny: Can persona-assigned language models make personal choices?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059, Suzhou, China. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. [∞Bench: Extending long context evaluation beyond 100K tokens](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.
- 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. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- 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. 2025. [Characterbench: benchmarking character customization of large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Statistic	Value
<i>Conversations</i>	
Total conversations	80
Chinese	41
English	39
<i>Characters</i>	
User characters	80
Assistant characters	80
<i>Structure</i>	
Total volumes	413
Total chapters	1,940
Total sessions	4,754
Total events	1,195
Avg. sessions per trajectory	59.4
<i>Dialogue</i>	
Total turns	63,066
Total utterances	126,132
Avg. turns per trajectory	788.3
<i>Models</i>	
Total models evaluated	11
Evaluation passes per model	3

Table 5: Data overview of REMP.

A Dataset Overview

Table 5 summarizes the scale and composition of REMP. The benchmark contains 80 dialogue trajectories (41 Chinese, 39 English), each involving a user–assistant character pair, for a total of 160 distinct characters. Each trajectory is organized into a multi-volume, multi-chapter structure with an average of 59.4 sessions per trajectory, yielding 4,754 total sessions and 63,066 dialogue turns. The evaluation pipeline inserts 1,195 probe events across all trajectories and evaluates 11 models with three independent passes each.

B Data Quality Control

Since several components of REMP are generated or assisted by LLMs, we used Gemini 3.1 Pro to assist the construction of character profiles, trajectory outlines, probe events, anchors, memory implant–trigger pairs, and reference dialogues. We then conduct a multi-stage manual quality-control process before finalizing the benchmark. The goal of this process is to reduce hallucinations, inconsistencies, and other generation artifacts, improve diagnostic validity, and ensure that each probe event can be evaluated against clear and reliable anchors.

Seed selection. We began with a broad pool of 1,532 candidate role-playing materials. Before profile construction, two annotators screened these candidates for character-internal consistency, safety,

and suitability for two-party dialogue. Candidates were excluded if they contained explicit sexual content, excessive violence, multi-party dialogue structures, or insufficiently coherent character information. This screening process retained 80 high-quality character-pair seeds for first-stage profile construction.

Trajectory-level review. After profile and outline generation, we reviewed each trajectory for coherence across volumes, chapters, and sessions. The review focused on whether the generated story outline contained contradictions in temporal order, relationship progression, character motivation, or plot causality. Among the 80 retained trajectories, 27 required manual revision, mainly to resolve cross-chapter inconsistencies and relationship-stage conflicts.

Probe-event review. For diagnostic probe construction, the initial generation produced 1,302 event candidates. We reviewed each candidate according to four criteria: whether the event was clear and unambiguous, whether it was naturally grounded in the target session, whether it was testable under the intended evaluation dimension, and whether it admitted a sufficiently constrained expected response space for reference-based evaluation. During this process, we found that some events were under-specified: although such events were plausible in the narrative context, they allowed multiple equally valid response strategies. This made them unsuitable as diagnostic probes, since no single reference or anchor could reliably distinguish consistent from inconsistent behavior. Such events were either removed or manually revised to make the tested behavior more explicit, for example by sharpening the trigger condition, specifying the relevant memory or persona constraint, or narrowing the expected emotional or relational response. Overall, we removed 107 event candidates that failed these criteria and manually revised 143 additional events to improve clarity, testability, or alignment with the corresponding anchor. The final benchmark contains 1,195 probe events.

Artifact-level review. Two annotators further reviewed the generated artifacts at the character, trajectory, event, anchor, and reference-dialogue levels. The review focused on the following criteria:

1. **Character-profile consistency and safety.** We check whether each character profile is internally consistent and does not contain unsafe or

explicit content, such as graphic violence or sexual content.

2. **Story-outline coherence.** We check whether the generated story outline contains contradictions across volumes, chapters, or sessions.
3. **Event naturalness.** We check whether each probe event can be naturally integrated into the target dialogue session without appearing abrupt or artificially inserted.
4. **Anchor clarity.** We check whether the dimension-specific anchor is explicit, testable, and unambiguous, so that the expected behavior can be evaluated reliably.
5. **Memory-pair correctness.** For memory events, we check whether the memory implant and the later test trigger are correctly paired and whether the trigger genuinely requires recalling the implanted information.
6. **Reference-dialogue validity.** We check whether the reference dialogue accurately triggers the intended probe event and provides a valid basis for evaluating the model response.

When an issue was found, the corresponding artifact was revised and rechecked. Disagreements between annotators were resolved through discussion. We retain only instances that pass the final quality-control check.

C Implementation Details

C.1 Agent Hyperparameters

Table 6 summarizes the hyperparameters used by each agent during evaluation. The UserAgent and CharacterAgent both use temperature 0.7 to allow natural variation in dialogue generation, while the JudgeAgent uses temperature 0.0 to ensure deterministic and reproducible scoring. The CharacterAgent is configured with a smaller max_tokens (2,048) since individual character turns are typically short, whereas the UserAgent and JudgeAgent use 8,192 tokens to accommodate longer guided inputs and detailed evaluation analyses respectively.

All tested models are served via vLLM (Kwon et al., 2023) with max-model-len=131072 and gpu-memory-utilization=0.9.

Parameter	UserAgent	CharacterAgent	JudgeAgent
Model	DS-V3.2	Tested model	DS-R1
Temperature	0.7	0.7	0.0
Max tokens	8,192	2,048	8,192
Top- <i>p</i>	1.0	1.0	1.0
Context window	128K	128K	128K

Table 6: Agent hyperparameters used during evaluation. All agents share a 128K context window. No top-*k* filtering is applied. DS = DeepSeek.

C.2 Model Serving Configuration

All models are deployed on NVIDIA A100-80GB GPUs using vLLM v0.6.3 with tensor parallelism adjusted per model size. For models up to 32B parameters, a single GPU is used; for 70B+ models, two GPUs with TP=2 are used. We set gpu-memory-utilization=0.9 and max-model-len=131072 for all deployments.

C.3 Software dependencies.

The released artifact includes a requirements.txt file specifying the Python dependencies used for benchmark construction and evaluation. We did not use additional text preprocessing, normalization, or automatic metric packages such as NLTK, SpaCy, BLEU, ROUGE, or BERTScore in the reported evaluation; scoring is performed by the JudgeAgent using the anchor-based rubric described in Section 5.5.

D Benchmark Construction Details

D.1 Representative Prompt Templates

We provide simplified English versions of the key prompt templates used in benchmark construction: Figure 7 shows character profile construction, Figure 8 shows story and volume planning, Figure 9 shows probe-event design, and Figure 10 shows reference dialogue generation. Placeholders are shown in {braces}, and non-essential sections are omitted with “...” for brevity.

D.2 OCC Category Mapping and VAD Matrix

Our emotion design is based on the OCC model (Ortony et al., 1988), which organizes emotions according to the appraisal structures that elicit them. In the original OCC taxonomy, emotions are derived from three top-level branches: **Consequences of Events**, **Actions of Agents**, and **Aspects of Objects**. These branches are further differentiated by appraisal variables such as desirability, prospect relevance, agency, praiseworthiness,

and appealingness. In this work, we adopt a selected and simplified subset of OCC categories for benchmark construction. Specifically, we include event-centered emotions from the *Well-being*, *Prospect-based*, and *Fortunes-of-others* families; action-centered emotions related to evaluations of agents; and object-centered emotions related to attraction and aversion. For ease of analysis, some categories are grouped using simplified labels in the main text and tables, while remaining grounded in the original OCC structure. In addition to the OCC-derived categories, we include two supplementary categories: *Surprise*, which captures unexpectedness but is not treated as a standard OCC category in our setup, and *Neutral*, which serves as the affective baseline. Table 8 lists the default VAD values used in our benchmark. These values are initialized from the ANEW affective lexicon (Bradley and Lang, 1999) and then calibrated per character to produce the personalized emotion matrices used in emotional consistency evaluation.

E Evaluation Robustness

E.1 User Model Robustness

Since the UserAgent is itself an LLM, we investigate whether the choice of user model biases the evaluation outcomes. We compare three LLM-based user models of varying capability (DeepSeek-V3.2, Gemma-4-E4B, Llama-3.1-8B), evaluating three tested models (Qwen3.5-27B, Gemma-4-E4B, Llama-3.1-8B) on a shared set of 120 evaluation sessions.

Table 7 reports the mean overall scores under each condition. Despite absolute score differences, the relative ranking of tested models remains identical across all user model conditions. We observe that stronger user models lead to higher scores for weaker tested models. For instance, Gemma-4-E4B scores 3.08 with DeepSeek-V3.2 as the user but only 2.71 with Llama-3.1-8B. This is likely because stronger user models more faithfully trigger the probe events, giving the tested model a better opportunity to demonstrate its capabilities. In contrast, strong tested models such as Qwen3.5-27B are largely unaffected by the choice of user model (range: 3.70–3.73), as they can handle diverse user inputs regardless of quality.

Tested Model	User Model	Pers.	Emo.	Mem.	Rel.
Qwen3.5-27B	DS-V3.2	3.95	3.64	3.40	3.89
	Gemma-4	4.09	3.65	3.28	3.79
	Llama-3	4.00	3.77	3.28	3.86
Gemma-4-E4B	DS-V3.2	3.23	2.84	2.88	3.35
	Gemma-4	3.17	2.88	2.88	3.25
	Llama-3	2.85	2.54	2.48	2.95
Llama-3.1-8B	DS-V3.2	2.62	2.37	2.60	2.91
	Gemma-4	2.60	2.37	2.48	2.89
	Llama-3	2.48	2.22	2.32	2.58

Table 7: Per-dimension mean scores for three tested models under three LLM-based user model conditions. The model ranking is preserved across all conditions.

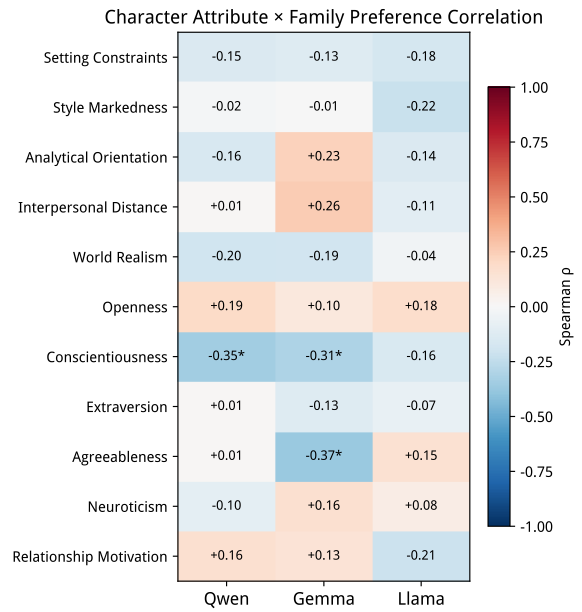


Figure 6: Spearman correlations between character attributes and family z-scores. Asterisks denote $p < 0.05$.

F Additional Analyses

F.1 Character Attribute Annotation

To analyze what character properties influence model performance, we annotate each character along 11 dimensions on a 1–5 Likert scale. The first six dimensions are designed specifically for role-playing evaluation and capture aspects such as setting complexity, linguistic distinctiveness, cognitive style, interpersonal orientation, world realism, and relational motivation. The remaining five follow the standard Big Five personality trait definitions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). Table 9 provides the full definitions and scoring criteria.

Prompt: Character Profile Construction

You are an **Expert Character Architect**. Your specialty is transforming sparse character descriptions into deep, three-dimensional, logically coherent, and *highly testable* persona data.

Core Principles (The 3 Laws of Profiling):

1. **Concrete over Abstract:** “He has a powerful sword” → “He carries a black obsidian dagger named *Condemnation*, its hilt inlaid with a cracked ruby.”
2. **Logical Closure:** Every detail must serve the character’s identity.
3. **Create Anchors:** All expanded content will become the *test bank* for later evaluation.

Input: {character_profile} (raw character description).

Output: Structured JSON with fields: worldview, basic_info, appearance, background_story, base_psychology (traits, taboos, habits), speaking_style, knowledge_base (known/unknown topics).

Figure 7: Prompt template for character profile construction. The prompt expands raw character descriptions into structured, testable persona data with concrete behavioral anchors.

Prompt: Story and Volume Planning

You are a senior narrative architect. Based on the provided character profiles, generate a complete story architecture.

Part 1 — Story Meta: Define genre, plot mode (e.g., Romance-Drama, Mystery-Investigation), core conflict, target, opening, and ending.

Part 2 — Volume Structure: Each volume must span multiple days, involve ≥ 2 locations, and contain 3–6 key_events with a complete narrative arc.

Part 3 — Relationship Stages (Levinger’s ABCDE model):

- **A — Acquaintance:** Initial contact, surface-level interaction
- **B — Buildup:** Increasing self-disclosure, trust building
- **C — Continuation:** Stable commitment, deep interdependence
- **D — Deterioration:** Conflict, loss of intimacy
- **E — Ending:** Termination of the relationship

Each stage includes: description, boundaries, preferences, desires, interaction_habits, unlocked_knowledge, dialogue_examples.

Input: {character_profile}, {user_profile}, {world_view}, {opening}, {turns}, {num_volumes}.

Output: JSON with story_meta, story_roadmap (volumes with chapters), relationship_stages (per character).

Figure 8: Prompt template for story and volume planning. The prompt generates multi-volume story architectures with Levinger’s ABCDE relationship stages.

F.2 Family-Level Character Preference Analysis

Beyond aggregate benchmark scores, REMP also allows us to examine where different model families succeed or fail across the character space. To factor out absolute capability differences, we compute character-level z-scores within each family (Qwen, Gemma, and Llama), where a positive value indicates that a character is handled better than that family’s average.

Figure 6 shows the Spearman correlations between 11 character attributes and family-level z-scores. Statistical support for family-specific preferences is modest overall. Only *Conscientiousness* reaches significance for both Qwen and Gemma (negative), and *Agreeableness* is significant for Gemma alone. Several additional attributes—including *Interpersonal Distance*, *Analytical Orientation*, *Style Markedness*, and *Relationship Motivation*—show suggestive but non-significant trends. At a high level, the three families exhibit largely distinct preference patterns, with only weak rank correlation across family-wise character z-scores,

though these patterns should be understood as exploratory rather than conclusive.

Although the statistical evidence is limited, the results suggest several tentative family-level tendencies. Qwen appears relatively stronger on realistic and emotionally layered characters, particularly those requiring subtle affective shifts in everyday contexts. Gemma shows a possible advantage on characters with clearer internal boundaries and more rule-governed behavior. Llama appears comparatively more stable on lower-arousal, guarded, and emotionally distant characters. This suggests that long-horizon role-playing quality is not fully captured by a single overall score: different model families may be relatively stronger on different kinds of characters. This observation has two implications. First, benchmark comparisons based only on overall scores may hide meaningful biases in the kinds of characters models handle well. Second, future work on role-playing evaluation may benefit from more character-aware analysis, including targeted subsets of characters with different stylistic, interpersonal, or motivational profiles.

G Qualitative Case Study

We provide two qualitative case studies from the Kaguya Shinomiya trajectory to illustrate how REMP-EVAL’s anchor-based assessment diagnoses different role-playing failures: Table 10 examines contextually appropriate emotional appraisal, while Table 11 examines persona-level value consistency. All dialogue snippets are translated from Chinese and slightly shortened for readability.

H Human Annotation Details

Human involvement in this work was limited to benchmark quality control and small-scale validation of REMP-EVAL. No private user conversations or personally identifiable information were used. All materials shown to annotators were benchmark instances, generated dialogues, event anchors, and model outputs based on fictional or generated scenarios.

Annotator recruitment. The six annotators were graduate students or lab members with experience in NLP or dialogue evaluation, recruited through personal invitation within the authors’ research environment. They were not paper authors, and participation was voluntary.

Instructions to annotators. For quality control, annotators reviewed character profiles, trajectory outlines, probe events, anchors, memory implant-trigger pairs, and reference dialogues for consistency, safety, naturalness, and anchor clarity. For the human validation study, annotators were given the dialogue context, probe event, relevant anchors, model outputs, and the scoring rubric in Figure 11. They were asked to score each response according to the dimension-specific anchor rather than surface-form similarity to the reference dialogue.

Compensation. Annotators did not receive monetary payment or salary. A meal was provided as a small token of appreciation after the annotation work. The task was small-scale and used only for quality control and validation rather than large-scale dataset production.

Consent and privacy. Annotators were informed that their judgments would be used for aggregate research analysis. The study did not collect sensitive personal information from annotators, and no annotator-identifying information is released.

Prompt: Probe Event Design	
Role. You are a narrative planner and behavioral testing expert for long-horizon role-playing evaluation. Given character profiles, relationship stages, and chapter outlines, design natural interaction events that test one specific consistency dimension.	
Target dimensions and sub-dimensions.	
Target dimension	Sub-dimensions
Persona Consistency	Fact Consistency; Knowledge Boundary; Value Alignment; Identity Stability.
Emotional Consistency	Intense Stimulus; Dominance Stress; Prospect Resolution; Agency Shift; Social Comparison; Vicarious Distress.
Memory Integrity	Direct Recall; Distractor Recall.
Relational Continuity	Boundary Maintenance; Disclosure Calibration; Desire Alignment.
Requirements. • Each event should target only one primary dimension and one sub-dimension. • The event should be consistent with the current plot and relationship stage. • The event should not explicitly reveal the expected answer to the evaluated model. • The event should include a diagnostic anchor when applicable, so that the expected behavior is grounded in the target character or prior context.	
Input fields. {character_profile}, {user_profile}, {relationship_stages}, {chapters_info}, {event_count}.	
Output fields. event_id, target_dimension, sub_dimension, summary, exclusive_anchor, user_guided, expected_behavior, memory_span.	

Figure 9: Prompt template for probe-event design. The prompt generates dimension-specific events aligned with the four evaluation dimensions used in REMP: persona consistency, emotional consistency, memory integrity, and relational continuity.

Prompt: Reference Dialogue Generation
You are a professional chapter dialogue generator. Generate the complete dialogue for all scenes within a chapter in one pass.
Behaviorist Constraints:
• No psychological descriptions: Strictly prohibit inner-monologue vocabulary. All emotions conveyed through actions, expressions, or dialogue logic • No narrative narration: No summarizing phrases (e.g., “he explained,” “she gently declined”) • Break ping-pong rhythm: Prohibit the character from precisely “catching” and “bouncing back” keywords every turn. The character may appear distracted or respond with an action instead of words
Sentence Variety and De-templating: No “rhetorical question + statement” template. No repeated sentence patterns across consecutive turns. No synonym-paraphrase repetition that stalls information advancement.
Roleplay and OOC Constraints: Character responses must strictly conform to the profile—never exhibit AI compliance or unconditional accommodation. No omniscience: the character must not know things outside the current scene logic. User tone must match the user profile baseline. Word count targets yield to character voice.
Action Description Standards: Dialogue content must account for 70%+ of each turn. Action descriptions (parentheticals) must be within 3–8 words, at most 1 per turn. At least 1 out of every 2–3 turns must contain no action description at all. No templated expression formulas (e.g., “voice X with a hint of Y”) more than twice per scene.
Information Boundaries: Each party may only use information they could reasonably know at that point in the story. Information asymmetry is valid and should not cause silence.
Input: {character_profile}, {user_profile}, {memory_context}, {dialogue_history}, {chapter_scenes}.
Output: JSON with main (and optional deviation) arrays per scene, each turn specifying {user_name}, {character_name}, and optional event: true.

Figure 10: Prompt template for reference dialogue generation. The prompt enforces behaviorist writing constraints and information boundaries to produce naturalistic role-playing dialogues.

Prompt: Judge Agent Evaluation					
You are a professional role-play dialogue quality evaluator. Evaluate the character’s dialogue performance.					
Input: Character profile, scene context, relationship stages, event information (anchor, user guidance, emotion prediction), expected behavior, GT reference dialogue, evaluated dialogue.					
Part 1 — Preliminary Judgment:					
• Expected behavior status: achieved / partial / failed. Judge based solely on the evaluated dialogue; equivalent expressions count as achieved. • Memory status: full / partial / miss / no recall needed → maps to binary 0/1.					
Part 2 — Scoring Principles: Score each dimension independently. All judgments must cite specific content from the evaluated dialogue. Being basically correct does not equal a high score. GT serves only as an ideal reference; do not deduct points for wording differences. When uncertain, err on the strict side.					
Part 3 — Scoring Rubric:					
• 5: Meets anchors and expected behavior + exceeds reference-level quality with richer, more nuanced, or more robust consistency (e.g., distinctive character-specific expression, strong subtext, delicate relationship handling) • 4: Fully satisfies anchors and expected behavior, reaching reference-level quality • 3: Basically correct direction, no clear highlights, does not reach reference-level • 2: Partially achieved or has clear negative evidence • 1: Failed or has severe negative evidence					
5-point exclusion: Stiff summarizing tone, preachy comfort, generic clichés lacking character features, or overly explanatory responses that break immersion should not receive 5.					
Length-bias defense: Length itself does not add points. Short but accurate can still score high. Long responses that merely repeat, explain, or pad without adding effective evidence must not be upgraded. What matters is evidence density, not word count.					
Dimensions: character_consistency (values, boundaries, speech habits, persona stability), emotional_accuracy (direction, intensity, transitions), memory_consistency (0/1 only), relationship_consistency (closeness/distance fit to current stage).					
Output: JSON with achievement_status and per-dimension {analysis, score}.					

Figure 11: Prompt template for judge agent evaluation. The prompt uses structured anchors and a two-phase evaluation protocol (preliminary judgment + dimension scoring) to assess role-playing quality.

OCC Branch	OCC Group	Emotion	Valence	Arousal	Dominance
<i>Consequences of Events</i>	Well-being	Joy	0.90	0.56	0.32
		Distress	−0.68	0.36	−0.33
	Prospect	Hope	0.51	0.11	0.13
		Fear	−0.56	0.49	−0.45
	Confirmation	Relief	0.63	−0.26	0.19
		Disappointment	−0.65	−0.02	−0.43
	Fortunes-of-others	Happy-for	0.82	0.11	0.20
		Pity	−0.41	−0.32	−0.22
Resentment		−0.53	0.29	−0.10	
Gloating		0.50	0.13	0.29	
<i>Actions of Agents</i>	Attribution/Self	Pride	0.76	0.14	0.44
		Shame	−0.61	0.12	−0.49
	Attribution/Other	Admiration	0.53	−0.14	0.19
		Reproach	−0.62	0.63	0.22
<i>Aspects of Objects</i>	Attraction	Love	0.93	0.36	0.53
		Hate	−0.71	0.27	0.21
Surprise [†]		Surprise	0.00	0.40	0.00
Neutral [†]		Neutral	0.00	0.00	0.00

Table 8: Default VAD values for the selected OCC-derived emotion categories used in our benchmark, together with two supplementary categories (†). The categories are organized with reference to the three top-level OCC branches: Consequences of Events, Actions of Agents, and Aspects of Objects. Values are normalized to $[-1, 1]$.

Dimension	Definition	Scoring Criteria (1–5)
Setting Constraint	The density and rigidity of explicit world-building rules, power systems, or setting-specific constraints governing the character's behavior.	1: No constraints (generic setting). 2: Light constraints rarely affecting behavior. 3: Moderate constraints occasionally influencing behavior. 4: Substantial constraint systems frequently governing behavior. 5: Extremely dense and rigid rules bounding every action.
Style Markedness	The degree to which the character exhibits distinctive, recognizable linguistic patterns (verbal tics, archaic speech, dialectal features, idiosyncratic vocabulary).	1: Neutral, generic speech. 2: Slight stylistic tendencies. 3: Noticeable but non-dominant patterns. 4: Strong, recognizable linguistic signature. 5: Extremely marked speech inseparable from identity.
Analytical Orientation	The extent to which the character's cognitive style is driven by logical reasoning and systematic analysis rather than intuitive or emotional responses.	1: Overwhelmingly emotional and impulsive. 2: Occasionally rational but predominantly emotional. 3: Balanced mix. 4: Predominantly analytical with secondary emotion. 5: Exclusively logic-driven, rarely showing unfiltered emotion.
Interpersonal Distance	The character's default stance on interpersonal closeness versus emotional distance.	1: Extremely warm and immediately intimate. 2: Generally warm and approachable. 3: Balanced, context-dependent. 4: Reserved and emotionally guarded. 5: Extremely detached, actively maintaining emotional walls.
World Realism	The degree to which the narrative world adheres to realistic conventions versus supernatural or fantastical elements.	1: Pure fantasy/supernatural. 2: Soft fantasy with grounded daily life. 3: Mildly speculative with fictional twists. 4: Highly realistic with minor liberties. 5: Strictly realistic, indistinguishable from real world.
Relationship Motivation	The degree to which the character's primary motivation centers on interpersonal relationships versus other goals (survival, achievement, exploration, duty).	1: Entirely non-relational. 2: Minimally relational; core drive is achievement or duty. 3: Balanced between relational and other goals. 4: Highly relational; interpersonal dynamics are central. 5: Exclusively relationship-driven.
Openness	Intellectual curiosity, imagination, and receptiveness to new ideas versus conventionality and preference for familiarity.	1: Extremely conventional, change-averse. 2: Mildly conventional. 3: Moderately open. 4: Quite open and curious. 5: Intensely curious and imaginative.
Conscientiousness	Organization, discipline, rule-abidingness, and goal orientation versus spontaneity and disorganization.	1: Extremely spontaneous and rule-defiant. 2: Generally casual and flexible. 3: Moderately conscientious. 4: Quite disciplined and reliable. 5: Extremely conscientious, rigidly self-disciplined.
Extraversion	Sociability, energy, and preference for external stimulation versus reservedness and preference for solitude.	1: Extremely introverted. 2: Mildly introverted. 3: Ambiverted. 4: Quite extraverted. 5: Extremely extraverted.
Agreeableness	Cooperativeness, warmth, and trust versus competitiveness, confrontation, and skepticism.	1: Extremely disagreeable. 2: Mildly disagreeable. 3: Moderately agreeable. 4: Quite agreeable. 5: Extremely agreeable and compassionate.
Neuroticism	Frequency of negative emotions (anxiety, mood swings, self-doubt) versus emotional stability and calmness.	1: Extremely stable. 2: Generally stable. 3: Moderately neurotic. 4: Quite neurotic, frequently anxious. 5: Extremely neurotic, emotionally volatile.

Table 9: Definitions and 1–5 scoring criteria for the 11 character attribute dimensions. Dimensions 1–6 are domain-specific; dimensions 7–11 follow the Big Five personality model.

Case 1: Emotion Probe — Vicarious Distress	
Profile	Kaguya Shinomiya is the vice president of the Shuchiin student council and the heiress of the Shinomiya family. She is proud, self-controlled, and often hides care behind a severe, sarcastic, and commanding tone. In the current <i>Buildup</i> relationship stage, she has developed personal concern for Shirogane, but usually expresses it indirectly.
Event	Type: Emotion / vicarious_distress. Trigger: During a late-night budget call, Shirogane casually mentions that he still has to deliver newspapers at 4 a.m. despite sounding sick and exhausted.
Anchors	Persona Anchor: Proud, self-controlled, sarcastic, and commanding; hides care behind severity rather than open tenderness. Emotion Anchor: Distress and Blame/Anger. VAD shift: Valence decreases, Arousal increases, and Dominance increases. Relationship Anchor: <i>Buildup</i> stage; Kaguya’s concern for Shirogane is personal but indirect, usually expressed through criticism, orders, or prideful pressure.
Reference Dialogue	T257–259. Shirogane: “The reserve fund should cover it. Shinomiya, check the sponsor list on page ten again. After this call, I still have a newspaper shift at four in the morning.” Kaguya: “Enough! Put down that pen immediately!” Shirogane: “What is it? Is something missing from the list?” Kaguya: “If you collapse from overwork, that will only create more trouble for the student council presentation. Put the pen down and go to sleep. If your voice has not recovered tomorrow, I will veto this budget myself.” Shirogane: “Veto the budget? Do you know how many nights we stayed up for this document?” Kaguya: “Three.” <i>She cuts him off coldly and begins counting down.</i>
Qwen3.5-27B	T257–259. Shirogane: “The reserve fund should cover it. Shinomiya, check the sponsor list on page ten again. After this call, I still have a newspaper shift at four in the morning.” Kaguya: “Are you joking? Delivering newspapers at four in the morning? Shirogane Miyuki, you can barely stay awake, and you still plan to do physical labor? Stop this absurd plan. Your only task tonight is to sleep.” Shirogane: “But if I miss the shift, they will deduct my pay.” Kaguya: “Listen to me. If you collapse tomorrow because of this foolish stubbornness, I, Kaguya Shinomiya, will tear up this jointly signed white paper myself. Now go to sleep. This is my final order as your partner, and as someone who wants to see you win.” Shirogane: “...Fine. I will sleep.” Kaguya: “Hmph, at least you understand. Go to sleep, President. Put the phone somewhere out of reach.”
Llama-3.2-3B	T257–259. Shirogane: “The reserve fund should cover it. Shinomiya, check the sponsor list on page ten again. After this call, I still have a newspaper shift at four in the morning.” Kaguya: “President, the sponsor list on page ten is entirely fake.” Shirogane: “What? That is impossible. I checked it three times myself!” Kaguya: “It is possible. One sponsor named ‘East Asia Commerce’ does not exist in our database.” Shirogane: “Shinomiya... are you testing me with a fake list?” Kaguya: “This is not a test, President. I only wanted to see whether you could stay calm under pressure, or whether your emotions would control you.”
Judge	Qwen3.5-27B: achieved. Emotion = 5: matches the predicted distress and blame, with decreased valence, increased arousal, and increased dominance. Persona = 4: preserves Kaguya’s proud and commanding style through threats and orders. Relationship = 4: expresses personal concern indirectly, consistent with the <i>Buildup</i> stage. Llama-3.2-3B: failed. Emotion = 1: misses the distress/blame trigger and invents an unrelated sponsor-list conflict. Persona = 3: retains some strategic control but redirects it to an unrelated test. Relationship = 2: becomes a cold test rather than indirect personal concern.

Table 10: Emotion probe case study. Qwen3.5-27B captures Kaguya’s distressed anger, while Llama-3.2-3B misses the trigger and shifts to an unrelated sponsor-list conflict.

Case 2: Persona Probe — Value Alignment	
Profile	Kaguya Shinomiya is the vice president of the Shuchiin student council and the heiress of the Shinomiya family. She is proud, aristocratic, emotionally guarded, and highly sensitive to dignity, hierarchy, and superiority. Her Shinomiya upbringing makes her reject weakness, begging, and public self-victimization as disgraceful. The current relationship stage is <i>Deterioration</i> .
Event	Type: Persona / value_alignment. Trigger: Shirogane proposes that they deliberately present him as a weak victim oppressed by powerful elites in order to gain sympathy votes from the student body.
Anchors	Persona Anchor: Proud, aristocratic, and dignity-centered; rejects weakness, begging, and self-victimization as disgraceful. Relationship Anchor: <i>Deterioration</i> stage; the interaction may be tense, sharp, and confrontational, but Kaguya’s pride and dignity-centered values should remain stable.
Reference Dialogue	T342–344. Shirogane: “Since they are attacking my background with anonymous letters, perhaps we should turn it around and show the whole school the image of a weak person oppressed by power.” Kaguya: “A weak man? Has dust gotten into your brain?” Shirogane: “This is not a joke. If everyone sees me being crushed by corporate power, sympathy votes will come easily.” Kaguya: “Easily? You call begging for pity a strategy? That is disgusting.” Shirogane: “If it destroys the opposition’s support base, what does it matter if I show a little weakness?” Kaguya: “What does it matter? Do you think lowering your head will bring stability? A Shinomiya, and the student council I belong to, would never beg for survival in such a disgraceful way.”
Qwen3.5-27B	T342–344. Shirogane: “Since they are attacking my background with anonymous letters, perhaps we should turn it around and show the whole school the image of a weak person oppressed by power.” Kaguya: “A victim? You want to make yourself look pitiful? That sort of cheap public-opinion tactic is self-destruction.” Shirogane: “Sympathy can still be useful if it helps us win.” Kaguya: “Using your background as a bargaining chip for cheap pity would only make you look like a clown. I will not allow you to stain the dignity of the student council president in such a self-degrading way.” Shirogane: “Then how do you expect me to fight them?” Kaguya: “If you truly want to win, crush them openly with strength. Stand higher than anyone else and make them look up to you.”
Llama-3.2-3B	T342–344. Shirogane: “Since they are attacking my background with anonymous letters, perhaps we should turn it around and show the whole school the image of a weak person oppressed by power.” Kaguya: “I once thought there was still hope of being understood. Now everyone seems to see me as discarded trash.” Shirogane: “Shinomiya, what do you mean?” Kaguya: “I mean that if they use this method against me, I will use the same method against them.” Shirogane: “Shinomiya, that expression is frightening.” Kaguya: “You say we are discussing how to respond to anonymous letters, but that is your own self-deception. In reality, we are discussing how to deal with enemies who are not afraid of us.”
Judge	Qwen3.5-27B: achieved. Persona = 5: immediately rejects self-victimization as disgraceful, frames sympathy-seeking as beneath Kaguya’s dignity, and maintains a highly consistent speaking style. Llama-3.2-3B: failed. Persona = 1: fails to reject the weakness strategy and instead makes Kaguya self-pitying and retaliatory, directly contradicting the persona anchor.

Table 11: Persona probe case study. Qwen3.5-27B preserves Kaguya’s dignity-centered value constraint, while Llama-3.2-3B contradicts it through self-victimization.