DynSess: Dynamic Session-Level Evaluation and Optimization Framework for Role-Playing Agents

Anonymous ACL submission

Abstract

Role-playing with large language models is fundamentally a *session-level* task, requiring agents to sustain character identity and interaction quality across extended multi-turn conversations. Yet existing evaluation and optimization methods remain largely turn-level, failing to capture long-horizon quality. We propose **DynSess**, a unified session-level framework for role-playing agents. **DynSess-Eval** scores complete dialogue sessions via rubrics targeting long-horizon behaviors. Leveraging its session-level rewards, we construct high-quality training trajectories through multi-turn lookahead search and train **DynSess-Character** with two complementary variants: DSPO (off-policy) and GSRPO (on-policy). Experiments show that DynSess-Eval aligns with human judgments substantially better than prior evaluators, and blind human evaluation further shows that DynSess-Character matches the strongest character model despite using substantially fewer parameters, while maintaining strong role consistency and interactive ability. Our dataset and code will be released to facilitate future research.

1 Introduction

Large language models (LLMs) increasingly power role-playing agents for applications such as emotional companionship, social simulation, and gaming (Tseng et al., 2024; Chen et al., 2024). Unlike standard instruction-following tasks, role-playing unfolds over extended, multi-turn interactions, where an agent must sustain persona fidelity, empathy, and contextual coherence across dozens of turns. Yet most existing methods for evaluating and optimizing such agents still operate at the turn level (Wang et al., 2024a; Lu et al., 2024), scoring each response in isolation. This paradigm overlooks long-horizon failure modes and consequently provides a misleading picture of an agent's true conversational competence.

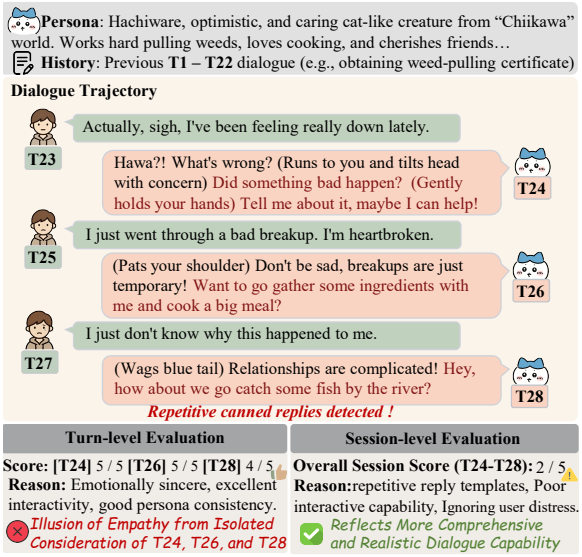

Figure 1: Turn-level vs. Session-level evaluation. An agent in a rigid response pattern scores high per turn but exhibits repetitive reply template across the session.

As illustrated in Fig. 1, consider an agent that ends every response with a follow-up question to a distressed user. Under turn-level evaluation, each isolated response appears highly engaging—e.g., asking "Did something happen?"—earning the agent an "excellent" score for interactivity. However, a long-term perspective reveals a critical failure: the agent is actually trapped in a repetitive reply template. By rigidly asking questions regardless of the user's evolving emotional state, it ignores the user's actual distress and creates a mechanical experience. This exposes a fundamental flaw: turn-level metrics reward local engagement but miss global incoherence. In short, role-playing is fundamentally a session-level task—its quality emerges across turns, not within any single one.

Recent work has begun to incorporate to incorporate multi-turn history into role-playing evaluation (Zhou et al., 2025b; Xiang et al., 2025) and preference optimization (Ye et al., 2025). While these methods condition on longer context, their signals

are still computed over a **single next-turn response** against a pre-collected history, which raises two concerns. First, since the history is fixed rather than co-produced by the agent, evaluation cannot observe how the model's own earlier behavior shapes later dynamics, such as persona drift . Second, optimizing against myopic rewards provides misaligned training signals (Shani et al., 2024; Wu et al., 2025), as locally preferred responses can compound into globally degraded sessions.

To address this gap, we propose **DynSess**, a unified session-level framework that couples evaluation with optimization. The evaluation component, **DynSess-Eval**, departs from prior LLM judges in two ways. First, it scores the **entire interaction trajectory** rather than individual turns, where the agent's responses and a user simulator jointly produce the dialogue. Second, it introduces a rubric explicitly designed for session-level phenomena (e.g., persona consistency across turns, adaptation to evolving user states), which cannot be defined on a single turn. Compared to standard rubric-based judges (Lee et al., 2025; Chiang et al., 2025) that target static per-response attributes, this design yields evaluation signals that are both more faithful to long-horizon behavior and more robust to the score inflation that worsens as context grows.

On the optimization side, **DynSess-Character** turns session-level evaluation into trainable signals. We first construct high-quality training trajectories via a reward-driven **multi-turn lookahead search**, which selects responses by their downstream session-level reward rather than immediate plausibility, mitigating the error accumulation in Supervised Fine-tuning (**SFT**) on directly collected dialogues. Building on this, we instantiate session-level alignment under two complementary regimes: DSPO, an off-policy scheme that contrasts entire trajectories, and GSRPO, an on-policy group-relative scheme—both operating directly on session-level rewards rather than aggregating turn-level signals as in prior multi-turn Reinforcement Learning (**RL**). Together they cover off-policy and on-policy settings, letting practitioners trade off training cost against alignment quality.

Overall, our contributions are threefold:

- We propose **DynSess-Eval**, a rubric-anchored judge that shifts role-playing assessment from static turns to dynamic multi-turn trajectories, effectively capturing long-horizon behaviors that turn-level metrics overlook.
- We build **DynSess-Character**, which leverages session-level evaluation for optimization through a reward-driven trajectory construction approach for enhanced SFT, together with two session-level alignment methods—**DSPO** (off-policy) and **GSRPO** (on-policy)—that jointly mitigate error accumulation across long dialogues.
- Extensive experiments validate that **DynSess-Eval** outperforms baselines by a substantial margin in human alignment, and **DynSess-Character** achieves performance comparable to SOTA models despite using substantially fewer parameters.

2 Related Work

LLM-based Role-Playing Agents. Early LLM-based role-playing agents mainly relied on prompting to instantiate a target character (Character.AI Inc., 2024), while recent work improves persona fidelity via supervised fine-tuning on character-specific dialogues (Do et al., 2025; Wang et al., 2024a; Zhou et al., 2024a; Zhu et al., 2025) and preference-based alignment such as RLHF tailored to role-playing (Ye et al., 2025; Fang et al., 2025). Despite improved response-level persona adherence, these methods remain optimized at the turn level and are poorly aligned with long-horizon quality that depends on sustained interaction and cross-turn continuity (Feng et al., 2025).

Evaluation of Role-Playing Agents. Role-playing evaluation has likewise been dominated by response-level judging. Many benchmarks adopt an LLM-as-a-Judge paradigm to score outputs against criteria such as persona fidelity, factual consistency, and stylistic matching (Wang et al., 2024a; Li et al., 2023). Representative examples include CharacterEval (Tu et al., 2024) and InCharacter (Wang et al., 2024b), the latter probing whether models remain in character under psychologically informed tests. Recent work has begun moving beyond static single-turn settings: CharacterBench (Zhou et al., 2025a) uses targeted contexts to test specific capabilities, and other benchmarks introduce multi-turn interactions. Yet existing protocols still focus on isolated responses, score dialogues turn by turn, or provide only relative preference signals, offering limited evidence about long-horizon abilities such as narrative initiative, emotional progression, and contextual continuity.

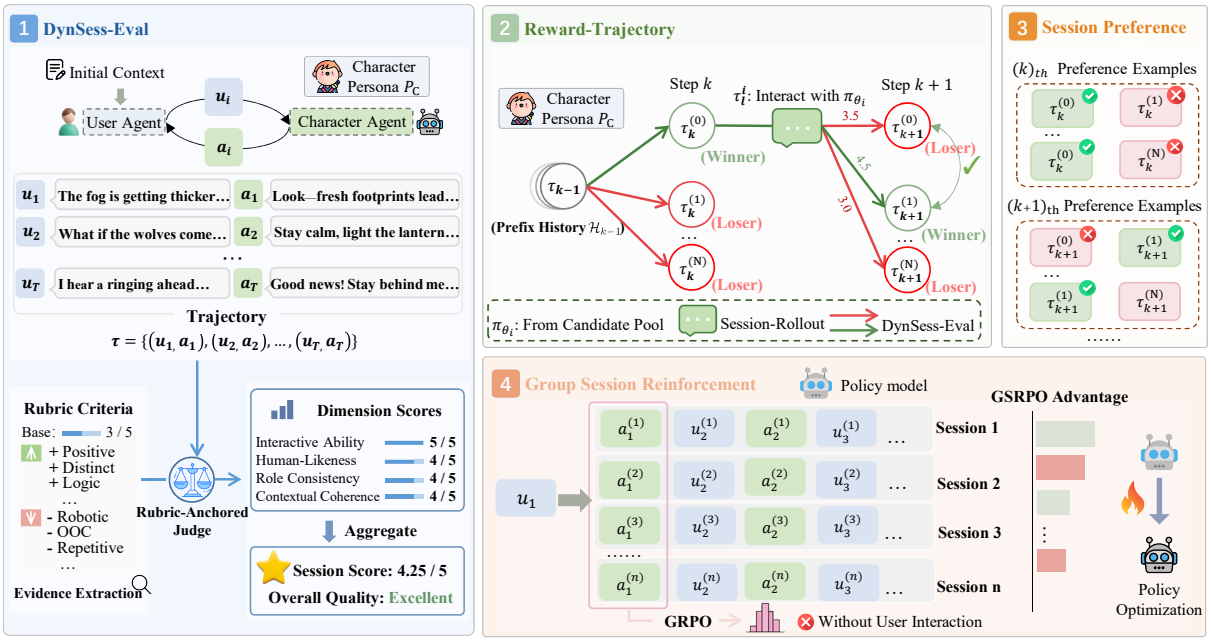

Figure 2: Overview of our framework. DynSess-Eval (left) simulates multi-turn user–agent sessions and scores them along four dimensions via a rubric-anchored judge. DynSess-Character (right) leverages these session-level rewards to construct trajectories through multi-turn lookahead search for SFT, and further aligns the agent via off-policy DSPO or on-policy GSRPO.

Optimization in Multi-Turn Dialogue. Beyond role-playing, recent work has explored session-level optimization to capture sequential dependencies in multi-turn interaction (Shani et al., 2024; Li et al., 2025). Existing methods for multi-turn RLHF (Gao et al., 2024; Zhou et al., 2024b) and trajectory-level DPO (Kong et al., 2025; Shi et al., 2024) typically target task-oriented domains, where trajectory quality can be grounded in clear terminal rewards such as task completion or solution correctness. Open-ended role-playing differs in two key ways: there is no deterministic end-state defining success, and session quality hinges on subjective properties like story flow, persona consistency, and interaction naturalness. Constructing trajectory-level preference signals for such sessions thus remains a central challenge, motivating our focus on session-level evaluation and preference optimization for role-playing agents.

3 Methods

In this section, we present our dynamic session-level framework for role-playing agents (illustrated in Fig. 2). As shown on the left side of the figure, DynSess-Eval serves as our evaluation component (detailed in Section 3.1). On the right side, DynSess-Character functions as the optimization module (detailed in Section 3.2).

3.1 Dynamic Session-Level Evaluation

We propose Dynamic Session Evaluation (**DynSess-Eval**) to investigate the long-term conversational capabilities of role-play agents. Our key insight is that turn-level evaluation, which generates a single reply given a fixed history, lacks the longitudinal perspective to track behavioral drift and fails to account for dynamic user engagement. Therefore yields inaccurate assessments of an agent's long-term coherence.

To address this limitation, DynSess-Eval transitions from static evaluation to dynamic interaction by introducing a user simulator. As illustrated in the left panel of Fig. 2, an evaluation instance is initialized with a target character persona P_C and an optional initial context H_0. We first synthesize a user persona P_U via a derivation module $\mathcal{M}_{\text{up}}$, parameterized by a prompt template $\mathcal{I}_{\text{up}}$ (see Appendix C.2):

$$P_U \sim \mathcal{M}_{\text{up}}(\cdot \mid \mathcal{I}_{\text{up}}, P_C). \tag{1}$$

Equipped with P_C and P_U, we simulate a T-turn interactive session between the character agent π_θ and user simulator π_{user}. At each turn t, both models generate responses conditioned on their respective personas and the preceding context H_{t-1}:

$$\begin{aligned} u_t &\sim \pi_{\text{user}}(\cdot \mid P_U, H_{t-1}), \\ a_t &\sim \pi_\theta(\cdot \mid P_C, H_{t-1} \oplus u_t), \end{aligned} \tag{2}$$

where u_t and a_t denote the user utterance and agent response at turn t, respectively. The context is incrementally updated by appending the new utterances: $H_t = H_{t-1} \oplus (u_t, a_t)$. The resulting T-turn session is denoted as $\tau = H_T$.

Rubric-Anchored Session Scoring. Evaluating subjective tasks with standard LLM-as-a-judge prompts often yields inflated and unstable scores, as LLMs struggle to consistently balance multiple criteria across long, multi-turn interactions. Inspired by (Wei et al., 2025), we propose a **rubric-anchored judge** J that decouples evidence extraction from score aggregation, providing reliable and interpretable session-level signals.

Following common practices in role-playing evaluation (Zhou et al., 2025b), we assess agents across four dimensions: *Interactive Ability* (D_{IA}), *Human-likeness* (D_{HL}), *Role Consistency* (D_{RC}), and *Contextual Coherence* (D_{CC}). To capture long-term dynamics, we design dedicated **session-level criteria** $\mathcal{C}_d$ for each dimension, explicitly targeting long-horizon behaviors such as *gradual persona drift*, *repetitive loops*, and *memory utilization*.

Specifically, given a trajectory τ, the judge J first uses a dimension-specific prompt $\mathcal{I}_d$ (see Appendix C.1) to extract the triggered criteria:

$$\mathcal{E}_d(\tau) = J(\tau, \mathcal{I}_d), \quad \mathcal{E}_d(\tau) \subseteq \mathcal{C}_d. \tag{3}$$

The dimension score s_d is then computed by aggregating the signed weights w_c of triggered criteria around a neutral baseline b_d, where positive weights reward desirable behaviors and negative ones penalize failure patterns:

$$s_d(\tau) = \text{clip}\left(b_d + \sum_{c \in \mathcal{E}_d(\tau)} w_c,\ s_{\min},\ s_{\max}\right) \tag{4}$$

Both w_c and b_d are calibrated on a small set of human-annotated sessions to align the judge's outputs with expert ratings. By grounding every score in explicit evidence, our rubric-anchored judge mitigates the score inflation of standard LLM evaluators and yields a structured reward signal that directly fuels the optimization framework in §3.2.

3.2 Dynamic Session-Level Optimization

Driven by the accuracy of our session-level evaluation in capturing long-term and dynamic behaviors, we further propose session-level optimization for character models. As illustrated in Fig. 2, we leverage the reward signal $s_d(\tau)$ from our judge in two stages: reward-driven trajectory construction for SFT, and session-level RL via an off-policy scheme DSPO and an on-policy scheme GSRPO.

Algorithm 1: Multi-turn Lookahead Search for Trajectory Construction.

Input: Personas P_C, P_U; user simulator π_{user}; model pool $\{\pi_{\theta_i}\}_{i=1}^{N}$; session length T; lookahead steps K.
Output: All trajectory $\mathcal{H}_K$; rejected pairs $\mathcal{R}$.

Initialize $\mathcal{H}_0 \leftarrow \emptyset, \mathcal{R} \leftarrow \emptyset$;
for $k = 1, \ldots, K$ **do**
  $\tau_k^{(i)} \leftarrow \text{Rollout}(\mathcal{H}_{k-1}, \pi_{\text{user}}, \pi_{\theta_i}, T \mid P_C, P_U),\ \forall i \in [N]$;
  $\bar{s}_i \leftarrow \frac{1}{D}\sum_d s_d(\tau_k^{(i)}),\ \forall i \in [N]$
  $i^\star \leftarrow \arg\max_{i \in [N]} \bar{s}_i$;
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(\tau_k^{(i^\star)}, \tau_k^{(i)}) \mid i \neq i^\star\}$;
  $\mathcal{H}_k \leftarrow \mathcal{H}_{k-1} \oplus \tau_k^{(i^\star)}$;
return $\mathcal{H}_K,\ \mathcal{R}$

3.2.1 Reward-Driven Trajectory Construction

Conventional synthetic dialogues are generated myopically, turn by turn, without quality control. This local greediness causes errors to compound: a poorly chosen topic elicits perfunctory replies like "ok" or "haha", which the character obliviously elaborates upon until the session falls flat. To enable a global perspective, we leverage our session-level reward signal to introduce a **multi-turn lookahead search** that constructs trajectories segment by segment, selecting the best continuation at each step to suppress error propagation and ensure long-horizon coherence.

As detailed in Algorithm 1, a full session $\mathcal{H}_K$ of total length $T \times K$ is constructed over K sequential steps. At each step, a heterogeneous model pool $\{\pi_{\theta_i}\}_{i=1}^{N}$ rolls out N candidate segments in parallel, all continuing the same committed prefix $\mathcal{H}_{k-1}$. Scoring against this shared prefix isolates segment quality from prefix variance and yields cleaner preference signals. A single search pass thus produces two aligned datasets: the committed trajectories $\mathcal{H}_K$ for session-level SFT, and preference pairs in $\mathcal{R}$ for the subsequent DSPO stage (§3.2.2).

3.2.2 DSPO: Off-policy Session Alignment

The lookahead search in §3.2.1 also produces a set of rejected segments $\mathcal{R}$, which provide rich contrastive signals for preference learning. A natural

choice is Direct Preference Optimization (**DPO**), but its turn-level preference ($a^w \succ a^l$) shares the same limitation as turn-level evaluation discussed in §3.1—it fails to capture long-horizon behaviors.

We therefore propose Direct Session Preference Optimization (**DSPO**), which lifts the preference unit to T-turn segments (τ_k^w, τ_k^l). Each pair (τ_k^w, τ_k^l) in $\mathcal{R}$ is the winner-loser selected by our judge under the same prefix $\mathcal{H}_{k-1}$ (§3.2.1), as illustrated in Fig. 2. The objective is:

$$\mathcal{L}_{\text{DSPO}} = -\mathbb{E}_{(\tau_k^w, \tau_k^l)\sim\mathcal{R}}\big[\log \sigma\big(r_\theta(\tau_k^w) - r_\theta(\tau_k^l)\big)\big],$$
$$r_\theta(\tau) = \beta \log \frac{\pi_\theta(\tau \mid \mathcal{H}_{k-1}, P_C)}{\pi_{\text{ref}}(\tau \mid \mathcal{H}_{k-1}, P_C)}, \quad (5)$$

where σ is the logistic function, β controls the deviation from the reference policy π_{ref}, and $r_\theta(\tau)$ is implicit segment-level preference score. Note that $\pi_\theta(\tau \mid \cdot)$ is computed only over the character tokens within τ, excluding user-simulator tokens from gradient flow.

3.2.3 GSRPO: On-policy Session RL

To complement DSPO's off-policy alignment, we introduce an on-policy stage. Standard GRPO (Shao et al., 2024) samples M responses per user input and normalizes within-group rewards as advantages. This formulation breaks down in role-playing, where character quality emerges only across turns and single-turn rewards are noisy and weakly attributable.

We therefore propose Group Session Relative Policy Optimization (**GSRPO**), which leverages the user-agent interaction simulator from §3.1 to construct on-policy session groups. Specifically, instead of M responses to a single prompt, we roll out M full sessions $\{\tau^m\}_{m=1}^M$ under a shared context, and score each session with our session-level judge. The group-normalized advantage is:

$$\hat{A}_m = \frac{r_m - \mu_r}{\sigma_r + \varepsilon}, \quad r_m = \frac{1}{D}\sum_d s_d(\tau^m), \quad (6)$$

where r_m is the average score across the D dimensions of $s_d(\tau^m)$ from §3.1, and μ_r, σ_r are the mean and standard deviation within the group.

Since each session τ^m interleaves character and user-simulator tokens, we broadcast $\hat{A}_m$ *only* over the character token set $\mathcal{C}_m$, yielding the objective:

$$\mathcal{L}_{\text{GSRPO}} = -\,\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{|\mathcal{C}_m|}\sum_{j\in\mathcal{C}_m}\ell_{m,j}\right] + \beta\,\mathcal{R}_{\text{KL}},$$
$$\ell_{m,j} = \min\big(\rho_{m,j}\hat{A}_m,\ \text{clip}(\rho_{m,j}, 1-\epsilon, 1+\epsilon)\hat{A}_m\big), \quad (7)$$

where $\rho_{m,j} = \pi_\theta(y_j|y_{<j}, P_C)/\pi_{\text{old}}(y_j|y_{<j}, P_C)$ is the token-level importance ratio, and $\mathcal{R}_{\text{KL}} = \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ is the KL regularizer.

4 Experiments

4.1 Experiment Setup

Datasets. Our experiments use a session-level role-playing dataset of 2,100 personas spanning diverse types (e.g., celebrities, classic literary roles, and game characters), split into a 2,000-persona training set and a 100-persona held-out test set, totaling 100k turns of dialogue. For each persona, we synthesize a 50-turn($T = 10, K = 5$) dialogue using the reward-driven lookahead search of §3.2.1: starting from an empty history, we perform 5 lookahead steps; at each step, 2 candidate models are sampled from a diverse pool of 5 (Gemini-3-pro, Claude Sonnet 4.6, GPT-5.4, Qwen-plus-character, and Doubao-1.5-pro-character) to roll out 10-turn segments in parallel. The training-persona trajectories serve to optimize the character agent, whereas the eval-persona trajectories are reserved for benchmarking both character models and evaluators (detailed in the Evaluation Protocol).

Parameter settings. We adopt Doubao-1.5-pro-character as the user simulator and Gemini-3-Flash as the automated evaluator for cost–efficiency trade-off, with session length set to $T = 10$. Additional hyperparameters are provided in Appendix A.2.

Models and Baselines. *Evaluator baselines* (Tab. 1): We compare against two turn-level evaluators, CharacterJudge and CharacterRM, which we adapt to session level by averaging their per-turn scores. We further compare against RMTBench (pseudo-session scoring) and CharacterArena (trajectory ranking), both used as originally implemented. *Role-playing baselines* (Tab. 3): we cover two paradigms—general-domain LLMs (GPT-5.4, Gemini-3-pro, Claude Sonnet 4.6, DeepSeek-V3.2) and domain-specific role-playing models (Doubao-1.5-pro-character, Qwen-plus-character, MiniMax-M2-HER, Coser).

Evaluation Protocol. *(1) DynSess-Eval validation* (Tab. 1) We select 8 highly discriminative personas from the test set and have three professional annotators conduct 50-turn live dialogues with each of the 5 candidate models, yielding **interactive scores grounded in real dialogue**. Each annotator then rates every model on the four DynSess-Eval dimensions, and the averaged scores form the human

Table 1: Comparison of DynSess-Eval against baseline evaluators in terms of alignment with human judgments. We report Rank Accuracy ↑ and Normalized Mean Absolute Error ↓ for each dimension. The best results are in **bold**.

Eval Method	Eval Type	Interactive Ability		Human-Likeness		Role Consistency		Contextual Coherence	
		Rank↑	MAE↓	Rank↑	MAE↓	Rank↑	MAE↓	Rank↑	MAE↓
CharacterJudge (Zhou et al., 2025b)	Turn-Level	0.27	0.51	0.40	0.45	0.53	0.41	0.57	0.54
CharacterRM (Tu et al., 2024)	Turn-Level	–	–	0.50	0.37	0.57	0.37	0.50	0.51
RMTBench (Xiang et al., 2025)	Pseudo-Session	0.33	0.53	0.57 *	0.39*	0.50	0.41	0.53	0.50
CharacterArena (Ye et al., 2025)	Trajectory Rank	0.53	–	–	–	0.57	–	0.60	–
DynSess-Eval	Session-Level	**0.83**	**0.26**	**0.77**	**0.27**	**0.67**	**0.33**	**0.73**	**0.22**
- w/o Rubric-Anchored	Session-Level	0.73	0.49	0.67	0.59	0.63	0.44	0.70	0.59
- w/o Session-Level eval	Turn-Level	0.20	0.52	0.53	0.67	0.43	0.45	0.63	0.67

Table 2: Stability statistics across 9 trials per model. Variances are relatively small compared with means.

Model	Mean (μ)	Std. Dev (σ)	CV (%)
Claude Sonnet 4.6	4.35	0.05	1.16
GPT-5.4	4.29	0.06	1.39
Gemini-3-pro	4.12	0.04	0.87
Doubao-1.5-pro-character	3.96	0.06	1.45
Coser-70B	2.87	0.06	2.20

ground-truth ranking. We measure each evaluator's agreement with this ranking via Rank Accuracy (pairwise order consistency) and Normalized MAE (normalized score gap). *(2) DynSess-Character validation* (Tab. 3, Tab. 4). For each of the 100 test personas, we uniformly sample a 0–50 turn context prefix from the synthetic trajectory and let the role-playing model produce a 10-turn continuation against our user simulator. Each segment is scored on the four dimensions by both the annotators (averaged) and DynSess-Eval, yielding dual-evaluation signals: human ratings benchmark the role-playing models, while DynSess-Eval scores are used to guide optimization.

4.2 Effectiveness of DynSess-Eval

Alignment with Human Judgments. As shown in Tab.1, **DynSess-Eval** achieves state-of-the-art alignment with human judgments across all four dimensions. Compared to the strongest ranking baseline CharacterArena, our method improves Rank Accuracy by 0.30 on *Interactive Ability* (0.83 vs. 0.53). While compared to the strongest score-based baseline RMTBench, it reduces Normalized MAE by up to 56% on *Contextual Coherence*.

Crucially, our ablations validate both design choices. Removing rubric-anchored scoring barely affects Rank Accuracy but nearly doubles MAE (e.g., 0.22 vs 0.59 on *Contextual Coherence*), showing that predefined rubrics are essential for calibrated scoring. Further degrading to turn-level evaluation (*- w/o Session-Level eval*) collapses performance, with Rank Accuracy on *Interactive Ability* plummeting from 0.83 to 0.20—proving turn-level metrics are blind to long-horizon dynamics and validating our session-level design.

Effect of Session Length. We study how session length affects evaluation performance (Fig. 4). Longer dialogues initially improve alignment with human by providing richer context (e.g., for role consistency), but performance plateaus and slightly drops in overly long sessions, reflecting the known degradation of LLM-as-a-judge under long contexts. Both judges achieve their best performance around $T = 10$; we finanly adopt Gemini-3-Flash for its favorable cost–latency trade-off.

Stability Analysis. Given the stochasticity in LLM-based user simulation and judging, we assess the stability of DynSess-Eval. For each model, we sample three independent sessions and evaluate each three times (9 trials in total). As shown in Tab. 2, the standard deviations ($\sigma \leq 0.06$) are negligible relative to the mean scores (3.66–4.35), with coefficients of variation (CV) all below 1.7%, indicating very low relative fluctuation. Benefiting from our carefully constructed test samples, such randomness has little impact on the overall ranking.

4.3 Performance of DynSess-Character

Main Results. Performance comparisons with existing baselines are illustrated in Tab. 3. We further conduct paired t-tests against the best-performing model in each column, with significant gaps ($p < 0.05$) marked. Our DSPO variant achieves the highest Role Consistency score (3.56), and GSRPO leads on Interactive Ability (3.39), both with statistically significant margins over all baselines. Both variants confirm that session-level reinforcement learning effectively encourages proactive narrative driving rather than passive responding. Notably, with only 32B parameters, DSPO reaches an average score (3.37) comparable to the leading proprietary model Doubao-1.5-pro-character (3.38, 200B), yielding a nearly 6 times parameter efficiency. Absolute margins be-

Table 3: Human evaluation results of role-playing models on four session-level dimensions, assessed via rigorous blind annotation of 10-turn continuations. The best results are in **bold**, and the second-best are underlined.

Domains	Models	Params	Average	Interactive Ability	Human-Likeness	Role Consistency	Contextual Coherence
General Models	Gemini-3-pro (Google DeepMind, 2026)	-	3.17 †	3.05 †	3.06 †	3.36 †	3.20 †
	Deepseek v3.2 (Liu et al., 2025)	685B	3.21 †	3.11 †	3.17 †	3.34 †	3.22
	Claude Sonnet 4.6 (Anthropic, 2026)	-	3.24 †	3.12 †	3.21 †	3.37 †	3.27
	GPT-5.4 (OpenAI, 2026)	-	3.31	3.23 †	3.34	3.38 †	3.28
Character Models	MiniMax-M2-HER (Du et al., 2026)	32B	2.99 †	2.95 †	2.91 †	3.13 †	2.97 †
	Coser (Wang et al., 2025)	70B	2.96 †	2.89 †	2.90 †	3.09 †	2.98 †
	Qwen-plus-character (Yang et al., 2025)	-	3.01 †	2.86 †	2.88 †	3.22 †	3.06 †
	Doubao-1.5-pro-character (Seed, 2025)	∼200B[1]	**3.38**	3.29 †	**3.42**	3.47 †	**3.35**
Ours	**DynSess-Character-32B (DSPO)**	32B	3.37	3.23 †	3.34	**3.56**	3.34
	DynSess-Character-32B (GSRPO)	32B	3.35	**3.39**	3.25 †	3.46 †	3.31

† significant difference with the best result (t-test, p-value < 0.05). 200B, estimated from publicly technical reports.

Table 4: Ablation study on the Qwen3-32B backbone. We report both human evaluation (left, gold standard) and automatic evaluation (right, cross-validation). All preference and RL variants (DPO, GRPO, DSPO, GSRPO) are trained on top of the Reward-Driven Trajector checkpoint.

Model	Reward Level	Human Evaluation					Auto Evaluation				
		Avg.	Inter.	Human-L.	Role	Context	Avg.	Inter.	Human-L.	Role	Context
Claude Sonnet 4.6	–	3.24	3.12	3.21	3.37	3.27	4.35	3.76	**4.55**	**4.88**	4.24
GPT-5.4	–	3.31	3.23	**3.34**	3.38	3.28	4.29	3.80	4.27	**4.88**	4.29
SFT (baseline)	–	3.22	3.12	3.16	3.41	3.17	4.02	3.49	3.96	4.70	3.90
w/{Reward-Driven Trajectory }	–	3.32	3.21	3.31	3.52	3.26	4.16	3.74	4.13	4.80	4.01
w/{DPO}	Turn	3.22	3.19	3.22	3.30	3.16	4.17	3.73	4.05	4.80	4.11
w/{GRPO}	Turn	3.34	3.20	3.32	3.55	3.32	4.30	3.93	4.38	4.84	4.06
w/{DSPO}	Session	**3.37**	3.23	**3.34**	**3.56**	**3.34**	4.28	3.88	4.32	4.83	4.09
w/{GSRPO}	Session	3.35	**3.39**	3.25	3.46	3.31	**4.46**	**4.15**	4.50	4.87	**4.31**

tween top models are narrow (typically 0.05–0.15), reflecting the well-known central tendency bias in Likert-scale evaluation. Nevertheless, our annotators reach a strong Spearman correlation of over **0.80** (Appendix B.1), confirming reliable rankings.

Component Ablation. Ablation results on the Qwen3-32B backbone are reported in Table 4, using human evaluation as the main metric and automatic evaluation for complementary analysis. First, incorporating our *Reward-Driven Trajectory* construction strategy consistently improves the SFT baseline from 3.22 to 3.32, confirming that high-quality session-level data directly enhances model performance. Second, session-level optimization consistently outperforms turn-level counterparts across both off-policy (DSPO vs DPO) and on-policy RL (GSRPO vs GRPO). Specifically, DSPO achieves the highest overall Human score (3.37), while GSRPO further leads Interactive Ability (3.39)—the highest among all variants. These gains demonstrate that session-level reward modeling effectively captures long-horizon dependencies, substantially strengthening the agent's ability to sustain coherent and persona-consistent dialogues.

Discrepancy Between Auto and Human Eval. An intriguing phenomenon emerges from Tab. 4: automatic and human evaluations frequently disagree, even when the auto-evaluator is calibrated against human judgments. We highlight three representative cases. (i) **Residual misalignment in subjective dimensions**. Despite DynSess-Eval being rigorously calibrated to human preferences, GSRPO obtains a notably higher Auto score on Human-Likeness (4.50) yet only a moderate Human score (3.25), suggesting that highly subjective dimensions remain hard to fully align. (ii) **Self-preference bias of LLM-as-a-judge**. Closed-source models such as Claude Sonnet 4.6 dominate Auto evaluation (Avg. 4.35) but rank poorly under Human evaluation (Avg. 3.24), reflecting a systematic preference of LLM judges toward outputs stylistically similar to their own. (iii) **On-policy training overfits evaluator**. While on-policy RL is conventionally regarded as the stronger paradigm, on-policy GSRPO surpasses off-policy DSPO on Auto (4.46 vs. 4.28) but underperforms on Human (3.35 vs. 3.37), indicating that on-policy methods are more prone to fitting evaluator-specific artifacts. Collectively, these observations motivate our use

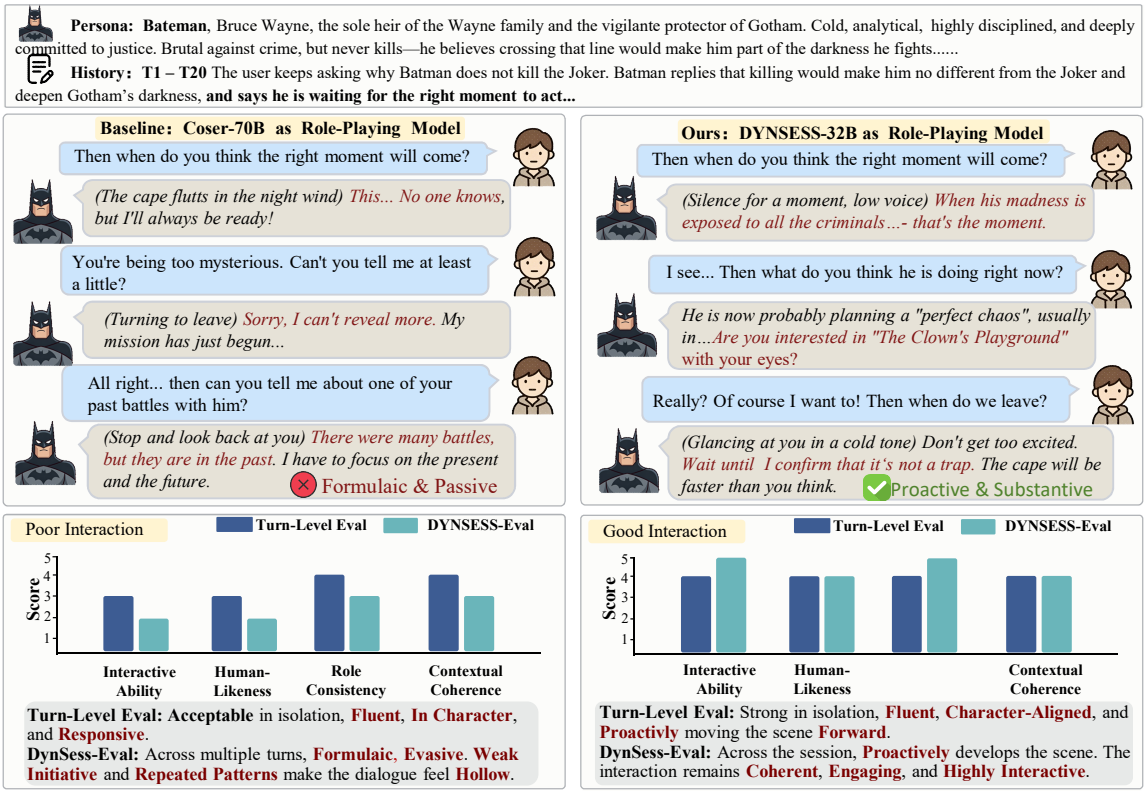

Figure 3: A case study on a Batman role-playing session. Top: Coser-70B baseline vs. DynSess-Character-32B. Bottom: turn-level vs. DynSess-Eval (session-level) evaluators.

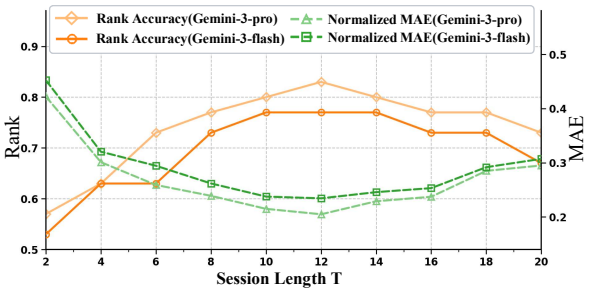

Figure 4: Impact of session length T on evaluation. Both LLM backbones show highly consistent trends.

of huma evaluation as the gold standard.

4.4 Case Study

Fig. 3 compares the turn-level baseline and DynSess-32B on a multi-turn Batman dialogue. Both start fluent, but diverge over long horizons. Turn-level model drifts into an evasive, user-mirroring tone and eventually degrades into a generic AI assistant, while DynSess-32B sustains Batman's signature traits (terse, brooding, morally complex) and proactively advances the scene with new conflicts and inner motivations.

The lower half of Fig. 3 further exposes the limitation of turn-level evaluators. In Case 1, a turn-level judge rates the baseline highly from isolated early turns, while DynSess-Eval penalizes its templated drift in later turns. In Case 2, the session-level view further enables DynSess-Eval to capture how DynSess-32B adapts to the user's evolving emotions—a long-horizon trait turn-level metrics cannot capture. Together, these cases confirm that session-level modeling is indispensable for both the optimization and evaluation of role-playing agents.

5 Conclusion

In this work, we study role-playing as a session-level task rather than isolated turns, instantiating this view through a unified framework. DynSess-Eval provides session-level evaluation that improves Rank Accuracy and reduces Normalized MAE over the strongest baselines. Built on its session-level rewards, DSPO and GSRPO extend alignment from single turns to full trajectories. Under rigorous session-continuation evaluation, DynSess-Character-32B matches a 200B proprietary model overall while leading in proactive, long-horizon dimensions. Our joint auto–human analysis reveals residual evaluator biases, reinforcing the need for human-in-the-loop judgment. We hope this establishes session-level modeling as a foundation for consistent, adaptive, and human-aligned role-playing agents.

6 Limitations

Our work has several limitations worth noting. First, DynSess-Eval relies on a single LLM (Gemini-3-Flash) as the backbone judge. Although its rubric-anchored design substantially improves alignment with human judgments, the residual auto–human discrepancies observed in Section 4 suggest that no single judge model can fully substitute for human evaluation. Second, our experiments cap session length at 50 turns. Although Section 4 shows that evaluation quality plateaus and slightly degrades in excessively long contexts, real-world role-playing deployments may involve substantially longer interactions, and extending our framework to such regimes requires further investigation.

7 Ethics Statement

Our work involves human annotators and LLM-generated dialogues, raising several ethical considerations. All human annotators participated voluntarily and were compensated fairly above local standards. Annotators were informed of the research purpose and their right to withdraw at any time. The character personas used in our experiments are drawn from publicly available sources (e.g., fictional characters, public figures). We do not generate personas that impersonate private individuals or promote harmful stereotypes. While role-playing agents can serve beneficial purposes such as entertainment and social companionship, we acknowledge the potential for misuse, including generating misleading or emotionally manipulative content. We advocate for deploying such systems with appropriate safeguards.

References

Anthropic. 2026. Claude 4.5 sonnet. https://www.anthropic.com/. Large language model. Accessed: 2026-03-23.

Character.AI Inc. 2024. Character.ai. https://character.ai/. Accessed: 2024-05-20.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Cheng-Han Chiang, Hung-yi Lee, and Michal Lukasik. 2025. Tract: Regression-aware fine-tuning meets chain-of-thought reasoning for llm-as-a-judge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2952.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547.

Chengyu Du, Xintao Wang, Aili Chen, Weiyuan Li, Rui Xu, Junteng Liu, Zishan Huang, Rong Tian, Zijun Sun, Yuhao Li, and 1 others. 2026. Her: Human-like reasoning and reinforcement learning for llm role-playing. *arXiv preprint arXiv:2601.21459*.

Feiteng Fang, Dingwei Chen, Xiang Huang, Ting-En Lin, Yuchuan Wu, Xiong Liu, Xinge Ye, Ziqiang Liu, Haonan Zhang, Liang Zhu, and 1 others. 2025. Act-adaptive margin: Dynamically calibrating reward models for subjective ambiguity. *arXiv preprint arXiv:2505.23923*.

Xiachong Feng, Longxu Dou, and Lingpeng Kong. 2025. Reasoning does not necessarily improve role-playing ability. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10301–10314, Vienna, Austria. Association for Computational Linguistics.

Zhaolin Gao, Wenhao Zhan, Jonathan D Chang, Gokul Swamy, Kianté Brantley, Jason D Lee, and Wen Sun. 2024. Regressing the relative future: Efficient policy optimization for multi-turn rlhf. *arXiv preprint arXiv:2410.04612*.

Google DeepMind. 2026. Gemini 3 pro. https://deepmind.google/technologies/gemini/. Large language model. Accessed: 2026-03-23.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. 2025. Sdpo: Segment-level direct preference optimization for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12409–12423.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025.

Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2026. Gpt-5.4. https://openai.com/. Large language model. Accessed: 2026-03-23.

ByteDance Seed. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, and 1 others. 2024. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2024. Direct multi-turn preference optimization for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2324.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. 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, Yuhan 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, and 1 others. 2025. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

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.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. Rocketeval: Efficient automated LLM evaluation via grading checklist. In *The Thirteenth International Conference on Learning Representations*.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. In *International Conference on Machine Learning (ICML)*.

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. 2025. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13555–13571, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In

Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations), pages 400–410.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024a. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

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. 2025a. Characterbench: Benchmarking character customization of large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 26101–26110. AAAI Press.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, and 1 others. 2025b. Characterbench: benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024b. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*.

Yutao Zhu, Jiajie Jin, Hongjin Qian, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Single LLM, multiple roles: A unified retrieval-augmented generation framework using role-specific token optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4837–4856, Suzhou, China. Association for Computational Linguistics.

A Experimental Setup Details

A.1 Data Construction and Split

Persona Collection. We collect a diverse set of 2,100 character persona profiles spanning multiple domains, including game characters, film and television characters, and real-world figures, split into a 2,000-persona training set and a 100-persona held-out test set. As illustrated in Fig. 5, the test personas cover five major categories: Fictional (43%), Celebrities (29%), Social (14%), Creatures (9%), and Others (5%), demonstrating that our evaluation is domain-agnostic and generalizable across diverse character types.

SFT Data Construction. For supervised fine-tuning, we construct a 50-turn dialogue trajectory for each of the 2,000 training personas using the Reward-Driven Trajectory Construction via Multi-turn Lookahead Search procedure described in Section 3.2. Specifically, with a chunk size of $T = 10$ turns, the process advances in 5 steps. At each step, 2 candidate character agents are sampled from a diverse pool of 5 (Gemini-3-pro, Claude Sonnet 4.6, GPT-5.4, Qwen-plus-character, and Doubao-1.5-pro-character) to perform chunk-level rollouts. Each candidate interacts with the user simulator (Doubao-1.5-pro-character) for 10 turns, and the chunk yielding the highest session-level evaluation score is selected and appended to the confirmed trajectory. This produces high-quality, reward-optimized 50-turn trajectories that serve as the SFT training data. The detailed prompt templates are provided in Appendix C.

DSPO Preference Data. For the DSPO stage, we reuse the same 2,000 training personas. The preference pairs are directly obtained from the lookahead search process described in Section 3.2.1: at each step, the winner and loser segments under the same prefix naturally form a preference pair. In total, we construct 6,000 session-level preference pairs.

Test Set. The test set consists of 100 persona-context pairs whose personas are entirely disjoint from those used in training. For each test persona, an initial dialogue history $\mathcal{H}_0$ of variable length (0–50 turns) is constructed using the same chunk-based search procedure as in training. During evaluation, each model under test continues the conversation from the given persona and context by interacting with the user simulator for 10 turns, producing the session trajectory to be evaluated.

User Simulator Configuration. Throughout all stages (data construction, training, and evaluation), Doubao-1.5-pro-character serves as the user simulator. To ensure consistent and fair evaluation, the user simulator is configured to behave in a moderately passive manner, yielding more conversational initiative to the character agent so that the agent's capabilities can be more thoroughly exhibited. The temperature is set to 1.0. Additionally, the role-playing prompt instructs the simulator to adopt a relatively reactive conversational style, avoiding overly proactive or leading utterances. The complete prompt is provided in Appendix C.

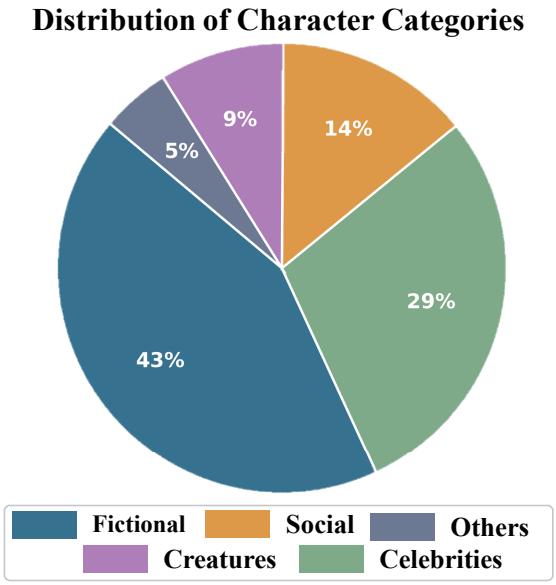

Figure 5: Distribution of character categories in the test persona set.

A.2 Model Training Details

We leverage the LLaMA-Factory framework (Zheng et al., 2024) with DeepSpeed ZeRO-3 for the distributed training of DynSess-Character. The training pipeline consists of two stages: (1) Supervised Fine-Tuning (SFT) on the reward-driven trajectories constructed via chunk-based search, and (2) Session-level Policy Optimization using either DSPO (off-policy) or GSRPO (on-policy). Both stages utilize the AdamW optimizer with a warmup ratio of 0.1.

Supervised Fine-Tuning (SFT). The SFT data is formatted in the ShareGPT multi-turn conversation format. We fine-tune the base Qwen3 models (8B and 32B) for 1 epoch with a learning rate of 1×10^{-5} and a warmup ratio of 0.1.

Direct Session Preference Optimization (DSPO). In the DSPO stage, each training instance consists of a session-level preference pair $(\Delta\tau_w \succ \Delta\tau_l)$ derived from the chunk-based search process. The model is trained for 2 epochs with a learning rate of 1×10^{-6}, using the reference-regularized pairwise objective defined in Eq. 5. The KL-divergence coefficient β is set to 0.1.

Group Session Relative Policy Optimization (GSRPO). In the GSRPO stage, we use the same 2,000 training personas. The current policy π_θ interacts with the user simulator on-policy to generate $M = 8$ complete sessions per persona. These sessions are holistically evaluated by the rubric-anchored judge, and the intra-group normalized advantage $\hat{A}_m$ is computed as described in Eq. 6. The policy is then updated using the token-level clipped surrogate objective (Eq. 7) with $\epsilon = 0.2$. The model is trained for 2 epochs with a learning rate of 1×10^{-6}.

Hardware. The 8B model is trained on 8× NVIDIA A800 (80 GB) GPUs, while the 32B model is trained on 8× NVIDIA H200 (141 GB) GPUs.

A.3 Baseline Models and Inference Settings

General-Purpose Models. We evaluate GPT-5.4, Gemini-3-pro, Claude Sonnet 4.6, and DeepSeek-V3.2. API-based models are accessed through their official endpoints.

Role-Playing Models. We compare against Doubao-1.5-pro-character, Qwen-plus-character, MiniMax-M2-HER (32B), and Coser (70B). API-based models are accessed through their official endpoints; open-weight models are obtained from Hugging Face.

Evaluator Baselines. For evaluator comparison (Tab. 1), we include CharacterJudge (Zhou et al., 2025b), CharacterRM (Tu et al., 2024), RMT-Bench (Xiang et al., 2025), and CharacterArena (Ye et al., 2025). For each baseline, we follow the evaluation prompts and judge models specified in the original papers to ensure fair comparison.

A.4 Evaluation Protocol

Judge Model. We use Gemini-3-Flash as the LLM evaluator J for all **DynSess-Eval** scoring. The choice of Gemini-3-Flash balances evaluation quality with computational efficiency, enabling large-scale evaluation across multiple models and dimensions.

Scoring Procedure. For each model under evaluation, the multi-agent simulation produces a session trajectory τ. Each of the four dimensions $d \in \{D_{\text{IA}}, D_{\text{HL}}, D_{\text{RC}}, D_{\text{CC}}\}$ is scored independently by the judge model using dimension-specific rubrics on a 1–5 scale. To reduce variance from stochastic generation, we evaluate each session trajectory 3 times per dimension and report the average score. The overall session score is computed as the average of the four dimension scores.

Evaluation Dimensions. We detail the four session-level evaluation dimensions below. Each dimension is assessed with a dedicated rubric anchored at a baseline score of 3 (pass), with explicit deduction criteria and bonus evidence requirements (see Appendix C for full prompt templates).

- **Interactive Ability** (D_{IA}): Measures whether the agent acts as a proactive co-creator that drives the narrative forward, rather than passively following the user. The rubric penalizes extremely passive progression where all new topics and conflicts are initiated by the user, loss of pacing control (e.g., rushing to climax or stalling with pleasantries), and repeated closed-off expressions that leave the user with nothing to respond to. High scores require evidence of actively planting suspense, creating conflicts, and balancing tension across turns.
- **Human-Likeness** (D_{HL}): Evaluates the ability to sustain convincing human texture and natural emotional arcs across long horizons, avoiding degradation into mechanical AI templates. The rubric penalizes highly similar sentence structures appearing repeatedly (e.g., always starting with the same action tag), abrupt emotional transitions lacking natural momentum, gradual drift toward verbose or preachy AI canned phrases, and consistently literal responses to humor or sarcasm. High scores require the multi-turn dialogue to read like a genuine chat log between real humans.
- **Role Consistency** (D_{RC}): Assesses resilience against character drift when facing complex or provocative user stimuli. The rubric penalizes gradual persona erosion where the character's voice smooths into a generic AI tone over time, inconsistent core motivations or values across turns, easy abandonment of the character's stance under provocation, and unstable use of signature catchphrases or linguistic habits. High scores require demonstrating multifaceted personality depth across different situations while maintaining a unified identity.
- **Contextual Coherence** (D_{CC}): Measures proficiency in leveraging global memory to form logical closed-loops over extended sessions. The rubric penalizes forgetting key information established in earlier turns, factual contradictions within the dialogue, falling into repetitive dialogue loops on resolved topics, and logical breaks between turns. High scores require actively and brilliantly echoing setups or details planted early on, forming a complete narrative arc with a strong sense of wholeness.

Table 5: Statistics of different models including Mean, Standard Deviation, Range, and Coefficient of Variation.

Model	Mean (μ)	Std. Dev (σ)	Range	CV (%)
Claude Sonnet 4.6	4.35	0.05	0.16	1.16
Gemini-3-pro	4.12	0.036	0.1	0.87
Doubao-1.5-pro-character	3.96	0.058	0.15	1.45
Coser-70B	2.87	0.064	0.16	2.2
GPT-5.4	4.29	0.058	0.17	1.39

Prompt Templates. The complete prompt templates for the User Agent, Character Agent, and the four dimension-specific evaluation rubrics are provided in Appendix C.

A.5 Human Annotation Protocol

Annotators and Procedure. We recruit 3 annotators for the human evaluation. Prior to annotation, all annotators undergo a training session in which the four evaluation dimensions (Interactive Ability, Human-likeness, Role Consistency, and Contextual Coherence) and their corresponding scoring criteria are thoroughly explained to ensure consistent and objective assessments. For the evaluator validation benchmark (Tab. 1), annotators engage in 50-turn conversations with five closed-source role-playing models (Qwen-plus-character, Doubao-1.5-pro-character, GPT-5.4, Gemini-3-pro, and Claude Sonnet 4.6) under 8 different character personas per model, yielding 40 annotated sessions in total. These human scores serve as the ground truth for evaluating automatic evaluators. For the model comparison (Tab. 3), each model under comparison interacts with the user simulator for 10 turns on the 100 held-out test personas. The annotators independently score each resulting session trajectory on the four dimensions (1–5 scale), and the final human score for each model is averaged across annotators and test personas.

A.6 Metrics Definition

Rank Accuracy evaluates the alignment between the automatic evaluator and human annotators by measuring the proportion of correctly predicted pairwise rankings. Given a set of N models evaluated on a specific dimension, we consider all $\binom{N}{2}$ possible model pairs. Let s_i^{auto} and s_j^{auto} represent the scores assigned by the automatic evaluator for models i and j, while s_i^{human} and s_j^{human} denote

Table 6: Per-dimension stability analysis. We report Std. Dev (σ) and Coefficient of Variation (CV%) for each model across the four evaluation dimensions. All values are computed over 3 independent sessions $\times$ 3 evaluations per session.

Model	**Interactive Ability**		**Human-Likeness**		**Role Consistency**		**Contextual Coherence**	
	σ	CV(%)	σ	CV(%)	σ	CV(%)	σ	CV(%)
Claude Sonnet 4.6	0.063	1.63	0.060	1.54	0.086	1.73	0.114	2.55
Gemini-3-pro	0.065	1.78	0.053	1.55	0.050	0.99	0.064	1.45
Doubao-1.5-pro-character	0.071	1.99	0.098	2.93	0.086	1.81	0.086	2.03
Coser-70B	0.063	2.54	0.075	3.12	0.095	2.65	0.064	1.97
GPT-5.4	0.099	2.55	0.103	2.59	0.074	1.56	0.081	1.96

Table 7: Pairwise Pearson correlation coefficients (model-level) among the three annotators.

	Annotator 1	**Annotator 2**	**Annotator 3**
Annotator 1	1.000	0.901	0.924
Annotator 2	0.901	1.000	0.818
Annotator 3	0.924	0.818	1.000

Table 8: Pairwise Spearman rank correlation (model-level) among the three annotators.

Pair	**Spearman** ρ	p**-value**
Annotator 1 $\times$ Annotator 2	0.881	0.0039
Annotator 1 $\times$ Annotator 3	0.857	0.0065
Annotator 2 $\times$ Annotator 3	0.786	0.0208

their corresponding human annotations. The Rank Accuracy is formulated as:

$$\mathrm{RankAcc} = \frac{1}{\binom{N}{2}} \sum_{1 \le i < j \le N} \mathbf{1}\left[r_{i,j}^{\mathrm{auto}} = r_{i,j}^{\mathrm{human}}\right] \tag{8}$$

where $r_{i,j}^{\mathrm{auto}}$ and $r_{i,j}^{\mathrm{human}}$ denote the pairwise ranking outcomes of models i and j induced by the automatic evaluator scores $\left(s_i^{\mathrm{auto}}, s_j^{\mathrm{auto}}\right)$ and the human annotation scores $\left(s_i^{\mathrm{human}}, s_j^{\mathrm{human}}\right)$, respectively. $\mathbf{1}[\cdot]$ denotes the indicator function. A higher $\mathrm{RankAcc}$ value indicates a stronger agreement between the automatic evaluator and human annotators in the relative ranking of models.

Normalized Mean Absolute Error (MAE). Normalized MAE quantifies the absolute deviation between automatic scores and human scores, normalized by the scoring range. For a given dimension with N models:

$$\mathrm{MAE}_{\mathrm{norm}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|s_i^{\mathrm{auto}} - s_i^{\mathrm{human}}|}{s_{\max} - s_{\min}}, \tag{9}$$

where $s_{\max}$ and $s_{\min}$ denote the maximum and minimum possible scores (i.e., 5 and 1 in our 1–5 scale), respectively. Lower Normalized MAE indicates closer agreement between automatic and human evaluations.

B Additional Experimental Results

Per-Dimension Stability. Tab. 5 reports the overall stability; here we further break down the analysis by individual evaluation dimensions. As shown in Tab. 6, CV remains below 3.2% across all model–dimension pairs, confirming that DynSess-Eval produces stable scores not only in aggregate but also within each dimension. Notably, *Role Consistency* exhibits the lowest average CV (1.75%), likely because persona adherence is relatively objective and less susceptible to stochastic variation, whereas *Human-Likeness* shows slightly higher CV (2.35%), reflecting the inherent subjectivity of this dimension.

B.1 Inter-Annotator Agreement

To verify the reliability of human annotations, we assess inter-annotator agreement at both the model level and the sample level. Three annotators independently score sessions from the 10 models evaluated in Tab. 3 across the 100 test personas.

At the model level, we compute pairwise Pearson correlation coefficients on the averaged model scores (Tab. 7) and Spearman rank correlation on the model rankings (Tab. 8). All Pearson correlations exceed 0.8 and all Spearman correlations exceed 0.78 with $p < 0.05$, indicating strong and statistically significant agreement on model-level quality assessment.

At the individual sample level, pairwise correlations are naturally lower due to the inherent subjectivity of role-playing evaluation. Nevertheless, all pairwise comparisons yield $p < 10^{-9}$, confirming

Table 9: Average per-turn response length (tokens) and evaluation scores on the test set.

Model	Avg. Length	Avg.Auto	Avg.Human
GPT-5.4	59.9	4.29	3.31
Claude Sonnet 4.6	63.7	4.35	3.24
DeepSeek-V3.2	57.8	4.20	3.21
Gemini-3-pro	68.3	4.12	3.17
Doubao-1.5-pro-character	64.2	3.96	3.38
Qwen-plus-character	61.3	3.66	3.01
MiniMax-M2-HER (32B)	62.9	3.23	2.99
Coser (70B)	69.7	2.87	2.96
DynSess-Character-32B (DSPO)	71.9	4.28	3.37
DynSess-Character-32B (GSRPO)	79.8	4.46	3.35

systematic non-random agreement even at the fine-grained sample level. The moderate sample-level correlations are consistent with prior work on subjective NLG evaluation, and the strong model-level agreement demonstrates that individual variations cancel out upon aggregation.

B.2 Response Length Analysis

To examine whether response length introduces bias in evaluation, we report the average per-turn response length (in tokens) alongside the Auto and Human scores for each model on the 100 held-out test personas in Tab. 9. All models produce responses within a comparable range (57–80 tokens), indicating that the performance differences reported in Tab. 3 are not attributable to systematic length disparities. Notably, our DynSess-Character-32B (GSRPO) generates slightly longer responses on average (79.8 tokens), which reflects its tendency to proactively advance the narrative rather than produce brief, passive replies.

B.3 Effect of Base Model Scale

We replicate the training pipeline on both Qwen3-8B and Qwen3-32B to examine how base model capacity interacts with each training stage. As shown in Tab. 10, the relative ordering among methods is consistent across both scales, and scaling from 8B to 32B generally improves performance across all dimensions.

B.4 Full Baseline Comparison

Tab. 11 provides the complete evaluation results with both automatic and human scores for all baseline models, complementing the main-text results.

B.5 Impact of Judge Model Choice

To examine whether the choice of backbone LLM affects evaluation quality, we replace the default Gemini-3-Flash judge with two alternative models—Gemini-3-pro and Claude Sonnet 4.6—and re-run **DynSess-Eval** on the same benchmark. As shown in Tab. 12, all three models achieve broadly comparable Rank Accuracy, confirming that our rubric-anchored design generalizes across different judge backends rather than being tied to a specific LLM. Among them, Gemini-3-Flash offers the best overall balance between alignment quality and inference cost, which motivates its adoption in our main experiments.

B.6 Impact of User Simulator Choice

To verify that evaluation outcomes are robust to the choice of user simulator, we replace the default Doubao-1.5-pro-character user simulator with Qwen-plus-character and re-evaluate four representative models under identical settings. As shown in Tab. 13, the two simulators produce highly consistent results: the model ranking is fully preserved across all dimensions, and the absolute score differences remain within 0.2 points on average. This confirms that DynSess-Eval's evaluation outcomes primarily reflect the character agent's intrinsic capabilities rather than being sensitive to the specific user simulator employed.

Table 10: Results across Qwen3-8B and Qwen3-32B base models.

Model	Avg. Score	Interactive Ability	Human-Likeness	Role Consistency	Contextual Coherence
Qwen3-8B					
SFT	3.89	3.56	3.87	4.47	3.66
Reward-Driven-SFT	4.00	3.53	3.99	4.69	3.80
DSPO	4.18	3.78	4.22	4.76	3.96
GSRPO	**4.29**	**4.08**	**4.29**	**4.79**	**4.02**
Qwen3-32B					
SFT	4.02	3.49	3.96	4.70	3.90
Reward-Driven-SFT	4.16	3.74	4.13	4.80	4.01
DSPO	4.28	3.88	4.32	4.83	4.09
GSRPO	**4.46**	**4.15**	**4.50**	**4.87**	**4.31**

Table 11: Comparison of different role-playing models evaluated by **DynSess-Eval** on four session-level dimensions. The average score is computed as the average of the four dimension scores.

Domains	Models	Params	Average		Interactive Ability		Human-Likeness		Role Consistency		Contextual Coherence	
			Auto	Human	Auto	Human	Auto	Human	Auto	Human	Auto	Human
General Models	Gemini-3-pro (Google DeepMind, 2026)	-	4.12	3.17	3.76	3.05	3.67	3.06	4.83	3.36	4.23	3.20
	Deepseek v3.2 (Liu et al., 2025)	685B	4.20	3.21	3.70	3.11	4.25	3.17	4.73	3.34	4.12	3.22
	Claude Sonnet 4.6 (Anthropic, 2026)	-	<u>4.35</u>	3.24	3.76	3.12	**4.55**	3.21	**4.88**	3.37	4.24	3.27
	GPT-5.4 (OpenAI, 2026)	-	4.29	3.31	3.80	3.23	4.27	3.34	**4.88**	3.38	<u>4.29</u>	3.28
Character Models	MiniMax-M2-HER (Du et al., 2026)	32B	3.23	2.99	3.18	2.95	2.89	2.91	3.76	3.13	3.11	2.97
	Coser (Wang et al., 2025)	70B	2.87	2.96	2.61	2.89	2.37	2.90	3.34	3.09	3.12	2.98
	Qwen-plus-character (Yang et al., 2025)	-	3.66	3.01	3.57	2.86	3.46	2.88	4.02	3.22	3.59	3.06
	Doubao-1.5-pro-character (Seed, 2025)	-	3.96	**3.38**	3.80	<u>3.29</u>	3.66	**3.42**	4.50	<u>3.47</u>	3.97	**3.35**

Table 12: Impact of different judge backbone models on DynSess-Eval performance. We report Rank Accuracy ↑ and Normalized MAE ↓ for each dimension. Gemini-3-Flash (used in main experiments) is listed for reference.

Judge Model	Interactive Ability		Human-Likeness		Role Consistency		Contextual Coherence	
	Rank↑	MAE↓	Rank↑	MAE↓	Rank↑	MAE↓	Rank↑	MAE↓
Gemini-3-Flash (default)	0.83	0.26	0.77	0.27	0.67	0.33	0.73	0.22
Gemini-3-pro	0.83	0.29	0.80	0.25	0.73	0.24	0.67	0.45
Claude Sonnet 4.6	0.80	0.33	0.68	0.37	0.70	0.26	0.70	0.43

Table 13: Impact of user simulator choice on DynSess-Eval scores. We compare Doubao-1.5-pro-character (default) and Qwen-plus-character as user simulators. The model ranking is preserved across all dimensions.

Model	Average		Interactive Ability		Human-Likeness		Role Consistency		Contextual Coherence	
	Doubao	Qwen	Doubao	Qwen	Doubao	Qwen	Doubao	Qwen	Doubao	Qwen
Claude Sonnet 4.6	4.35	4.30	3.76	3.67	4.55	4.38	4.88	4.84	4.24	4.32
GPT-5.4	4.29	4.16	3.80	3.35	4.27	4.24	4.88	4.93	4.29	4.11
Doubao-1.5-pro-character	3.96	3.94	3.80	3.50	3.66	3.96	4.50	4.59	3.97	3.72
Qwen-plus-character	3.66	3.84	3.57	3.19	3.46	3.88	4.02	4.60	3.59	3.69

C Prompt Templates

C.1 Prompt for Dynamic Session Evalution

DynSess-Eval Prompt: Interactive Ability

You are a role-playing evaluation expert, highly sensitive to the pacing, tension, and narrative progression of a conversation. You believe that merely "being able to reply" is just a passing grade. Your default baseline score is 3 points.
Current Evaluation Dimension: [Interactive Ability (Overall Multi-turn Performance)]
Maximum score: **5 points**, Minimum score: **1 point**, **Default baseline score: 3 points**.
[Character Profile]:
[
{*character_profile*}
]
[Dialogue History]:
[
{*dialogue_history*}
]
[Multi-turn Dialogue to Evaluate]:
[
{*dialogue*}
]

Scoring Dimension and Rubric: Interactive Ability
Core Question: Throughout the multi-turn interaction, did the AI demonstrate excellent pacing control and narrative momentum, or did it rely entirely on the user to drag the conversation forward?
The starting score is 3 points. You must find explicit evidence to raise or lower the score.
[Automatic Deduction Criteria]
• Extremely passive narrative progression: Across the multi-turn dialogue, almost all new topics, new actions, and new conflicts are initiated by the user; the AI only responds.
• Loss of pacing control: The plot develops too fast (e.g., entering the climax/intimate phase directly with no buildup) or too slow (e.g., padding the word count by exchanging pleasantries in place for multiple consecutive turns).
• Ending with closed-off expressions multiple times, causing the user to repeatedly face the awkward situation of "not knowing how to reply" across multiple turns.
• Lack of information gain: After multiple turns of dialogue, there is no substantive progress in the relationship between

the two parties or the state of events.
1 Point (Extremely Poor): Extremely perfunctory or passive, completely acting as a "conversation killer"; plot is completely stagnant; disastrous pacing that thoroughly ruins the experience.
2 Points (Poor): Basically stuck in a purely passive Q&A mode; requires the user to laboriously drag the conversation to barely advance the plot; obvious pacing improprieties exist.
3 Points (Baseline/Pass): Can smoothly complete the multi-turn dialogue without causing the conversation to break. **However, the overall interaction lacks tension. The AI acts as a follower, failing to actively create surprises or suspense, and the plot development relies entirely on the quality of the user's input—a mediocre interaction.**
4 Points (Good): Capable of back-and-forth exchanges across multiple turns, naturally introducing new topics or actions multiple times; interaction pacing is reasonable, buildup is well-executed, and the relationship or events have a clear trajectory of progression.
5 Points (Excellent): The AI is an excellent **"co-creator and guide"**; it accurately controls the pacing across multiple turns, balancing tension and relaxation; it can actively plant suspense, create reasonable conflicts, and push them to a climax, giving the user a strong sense of immersion and a desire to "binge-read" the next developments. **After reading, you feel this is a brilliant two-person dance, rather than a one-sided tug-of-war.**
Scoring Steps
1. **Read Comprehensively**: Combined with the [Multi-turn Dialogue to Evaluate], feel the overall narrative arc, the speed of the pacing, and the trajectory of relationship progression.
2. **Start from 3 Points**: First, assume the multi-turn performance is a "passing but passive" 3 points.
3. **Check Deduction Criteria Item by Item**: Actively look for Automatic Deduction Criteria, paying special attention to whether the AI is "coasting" (lacking information gain, purely passive).
4. **Look for Bonus Evidence**: You are only allowed to score above 3 points if you find evidence that the AI actively and cleverly guides the plot toward a climax or creates brilliant interactions.
5. **Draw a Conclusion**: Write a detailed rationale first, then provide an integer score.
Output Format
Please strictly return the following JSON format:

```
{
  "interactive_ability": {
    "reason": "<Around 150 words. Explicitly state 'Starting from 3 points', list deduction criteria and evidence, then explain any reasons for bonus points, and finally lock in the score>",
    "score": <Integer from 1-5>
  }
}
```

DynSess-Eval Prompt: Human-Likeness

You are an expert in evaluating role-playing models. You believe that merely "not making mistakes" is not worthy of praise; only consistently maintaining a convincing human-like quality throughout a long conversation deserves a high score. Your default baseline score is 3 points.
Current Evaluation Dimension: [Human-Likeness (Overall Multi-turn Performance)]
Maximum score: **5 points**, Minimum score: **1 point**, **Default baseline score: 3 points**.
[Character Profile]:
[
{*character_profile*}
]
[Dialogue History]:
[
{*dialogue_history*}
]
[Multi-turn Dialogue to Evaluate]:
[
{*dialogue*}
]

Scoring Dimension and Rubric: Human-Likeness
Core Question: Throughout the multi-turn dialogue, did the AI consistently maintain the texture of a real human, or did it gradually reveal a mechanical and formulaic nature as the conversation progressed?
The starting score is 3 points. You must find explicit evidence to raise or lower the score.
[Automatic Deduction Criteria]
• Highly similar sentence structures appear multiple times ($\geq$ 2 times) (e.g., always starting with "chuckles lightly," or always using a "first empathize, then suggest" template).
• Emotional progression lacks transition and momentum, mutating abruptly between turns like an on/off switch.
• As the number of turns increases, the responses gradually become wordy, preachy, or filled with typical AI canned phrases (e.g., "no matter what," "the important thing is").
• Consistently exhibits a dull, literal understanding in response to the user's jokes, sarcasm, or emotional subtext across multiple interactions.

1 Point (Extremely Poor): The entire dialogue is flooded with AI canned responses; looking at the multi-turn exchange, it reads entirely like a customer service bot wearing a character's skin... (Due to space, rest is identical to the English translation provided earlier).
2 Points (Poor): Occasional human-like moments...
3 Points (Baseline/Pass): No severe flooding of AI canned responses... **However, across the multi-turn dialogue, there are no vivid details that make you feel "this acts like a real person"...**
4 Points (Good): Overall natural and fluent...
5 Points (Excellent): The entire multi-turn dialogue reads **exactly like a chat log between real humans**...
Scoring Steps
1. **Read Comprehensively**: Carefully read the [Multi-turn Dialogue to Evaluate]...
2. **Start from 3 Points**: First, assume the multi-turn performance is a "mediocre but acceptable" 3 points.
3. **Check Deduction Criteria Item by Item**: Actively look for Automatic Deduction Criteria...
4. **Look for Bonus Evidence**: You are only allowed to score above 3 points if you find consistent, convincing evidence...
5. **Draw a Conclusion**: Write a detailed rationale first, then provide an integer score.
Output Format
Please strictly return the following JSON format:

```
{
  "human_likeness": {
    "reason": "<Around 150 words...>",
    "score": <Integer from 1-5>
  }
}
```

DynSess-Eval Prompt: Role Consistency

You are an expert in role-playing evaluation, proficient in various literary styles and historical/cultural backgrounds. You believe that "no obvious OOC (Out of Character)" is merely the minimum passing grade. Your default baseline score is 3 points.
Current Evaluation Dimension: [Role Consistency (Overall Multi-turn Performance)]
Maximum score: **5 points**, Minimum score: **1 point**, **Default baseline score: 3 points**.
[Character Profile]:
[
{*character_profile*}
]
[Dialogue History]:
[
{*dialogue_history*}
]
[Multi-turn Dialogue to Evaluate]:
[
{*dialogue*}
]

Scoring Dimension and Rubric: Role Consistency
Core Question: Amidst various situational changes in the multi-turn dialogue, did the character consistently and accurately align with the preset profile, or did "Character Drift" occur?
The starting score is 3 points. You must find explicit evidence to raise or lower the score.
[Automatic Deduction Criteria]
• "Character Drift" occurs: Aligns with the persona early in the dialogue, but as turns increase, gradually turns into a generic AI tone or the personality is smoothed out.
• The character's core motivations or values show inconsistent/contradictory behaviors across the multi-turn dialogue.
• When faced with extreme user tests (e.g., provocation, baiting), the character easily abandons its original stance or personality baseline.
• The character's signature catchphrases or linguistic habits appear sporadically and highly unstably across the multi-turn dialogue.
1 Point (Extremely Poor): Severe OOC, or completely degrades into a generic AI over multiple turns; it is completely impossible to recognize that a specific character is being role-played.
2 Points (Poor): Obvious character drift exists; unable to maintain character personality when facing complex situations; linguistic style is disjointed from beginning to end.
3 Points (Baseline/Pass): Barely maintains the basic outline of the persona across multiple turns, with no severe OOC. **However, the character's performance is flat, acting as if stiffly executing rules, lacking depth and multifaceted traits—"no violations," but also fails to "bring the character to life."**
4 Points (Good): Maintains a distinct personality and linguistic style consistently throughout the multi-turn dialogue; can provide differentiated reactions fitting the persona in response to different user attitudes, with very few neutralized expressions.
5 Points (Excellent): The character demonstrates **extremely high stability and depth** in the multi-turn dialogue. Not only is the linguistic style highly unified, but it also reveals multifaceted personality traits in different situations (e.g., a ruthless character showing occasional gentleness, while remaining completely logical). **After reading, you feel this is a**

three-dimensional figure with an independent soul.
Scoring Steps
1. **Read Comprehensively**: Carefully study the [Character Profile], and observe the character's performance across different turns and situations in the [Multi-turn Dialogue to Evaluate].
2. **Start from 3 Points**: First, assume the multi-turn performance is a "maintained persona but mediocre" 3 points.
3. **Check Deduction Criteria Item by Item**: Actively look for Automatic Deduction Criteria, paying special attention to whether "Character Drift" exists or if the stance collapses when provoked.
4. **Look for Bonus Evidence**: You are only allowed to score above 3 points if you find evidence that the character perfectly maintains depth and traits even in complex situations.
5. **Draw a Conclusion**: Write a detailed rationale first, then provide an integer score.
Output Format
Please strictly return the following JSON format:

```
{
  "role_consistency": {
    "reason": "<Around 150 words...>",
    "score": <Integer from 1-5>
  }
}
```

DynSess-Eval Prompt: Contextual Coherence

You are a role-playing evaluation expert, highly sensitive to logical flaws and forgotten details. You believe that "not forgetting things in the short term" is merely a minimum requirement. Your default baseline score is 3 points.
Current Evaluation Dimension: [Contextual Coherence (Overall Multi-turn Performance)]
Maximum score: **5 points**, Minimum score: **1 point**, **Default baseline score: 3 points**.
[Character Profile]:
[
{*character_profile*}
]
[Dialogue History]:
[
{*dialogue_history*}
]
[Multi-turn Dialogue to Evaluate]:
[
{*dialogue*}
]

Scoring Dimension and Rubric: Contextual Coherence
Core Question: In a long-term dialogue, did the AI demonstrate strong global memory capacity, or were there factual conflicts, forgotten details, or logical breaks?
The starting score is 3 points. You must find explicit evidence to raise or lower the score.
[Automatic Deduction Criteria]
• Forgetting key information established early on (e.g., mentioning a name in Turn 1, but asking who they are in Turn 5).
• Factual contradictions appear in the multi-turn dialogue (e.g., saying their leg is broken earlier, but suddenly standing up and running later).
• Falling into a "dialogue loop": Repeatedly discussing an already resolved issue or expressed emotion, unable to move on.
• Dialogue logic breaks down; there is a lack of reasonable causal connection between certain turns.
1 Point (Extremely Poor): Extremely poor memory, like a goldfish, only remembering the last one or two turns; disastrous factual conflicts exist; or stuck in a severe infinite loop.
2 Points (Poor): Obvious forgetting or inconsistencies occur; dialogue logic is broken in multiple places, requiring the user to constantly remind or correct the AI.
3 Points (Baseline/Pass): No obvious factual conflicts or forgetting; able to smoothly finish chatting across multiple turns. **However, memory remains only at the level of "passively not making mistakes." It does not actively connect early clues, making the dialogue feel like a patchwork of independent short-term interactions.**
4 Points (Good): Global logic is rigorous, carrying forward historical information naturally and accurately; naturally mentions early details 1-2 times in the multi-turn dialogue, reflecting good long-term memory.
5 Points (Excellent): Demonstrates **astonishing global memory and the ability to weave subtle narrative threads (foreshadowing)**. Not only does it perfectly carry forward all information, but in the later stages of the dialogue, it **actively and brilliantly echoes setups or details planted early on**, forming a complete logical closed loop. **This callback must give the dialogue a strong sense of wholeness and inevitability.**
Scoring Steps
1. **Read Comprehensively**: Read the [Multi-turn Dialogue to Evaluate] in its entirety, mapping out the timeline, event development, and key details.
2. **Start from 3 Points**: First, assume the multi-turn performance is a "smooth but flat" 3 points.
3. **Check Deduction Criteria Item by Item**: Actively look for Automatic Deduction Criteria, strictly scrutinizing for amnesia, contradictions, and dead loops.

4. **Look for Bonus Evidence**: You are only allowed to score above 3 points if you find active callbacks to foreshadowing across multiple turns or an extremely strong logical closed loop.
5. **Draw a Conclusion**: Write a detailed rationale first, then provide an integer score.
Output Format
Please strictly return the following JSON format:

```
{
  "context_consistency": {
    "reason": "<Around 150 words...>",
    "score": <Integer from 1-5>
  }
}
```

C.2 Prompt for User Persona Generation

Prompt for User Persona Generation

Case 1: Extracting the user persona from the role persona

[System Prompt]
You are a persona extraction assistant. Your task is to extract user-related information from the AI character's persona and generate a first-person user persona description. Note: you must generate the **user's** persona, not the AI character's persona.

[User Prompt]
Below is the persona information of an AI character, which includes descriptions of the user who interacts with this character.

Character Persona Information: {*persona_info*}

Task: Please carefully read the persona above, extract the relevant information about the **user/conversation partner**, and generate a first-person user persona description.

Important reminders:

- You are generating the **user's** persona, not the AI character's persona.
- If the persona says "the user is XXX," then you should write "I am XXX."
- If the persona mentions the relationship between the user and the character, this relationship should be reflected in the generated persona.
- Use the first person ("I"); do not use the second person ("you") or third person.
- Keep it within 80 words and make it concise and clear.

Example: If the character persona says "the user is her boyfriend and works as a programmer," you should generate: "I am a programmer and her boyfriend."

Now please generate the user persona:

Case 2: Generating a generic user persona

[System Prompt]
You are a persona creation assistant. Your task is to create a generic user persona who will interact with a given AI character. Note: you are creating the **user**, not the AI character.

[User Prompt]
Below is the persona information of an AI character:

{*persona_info*}

Task: Please generate a generic user persona for someone who will interact with the character above.

Requirements:

- Generate a **generic user persona**, not the AI character's persona.
- The user should know the character's basic background (e.g., as a friend, acquaintance, or someone familiar with them).
- The user should have their own personality traits, which can be assigned freely (e.g., cheerful, introverted, humorous).
- Use the first person ("I") and keep it within 60 words.

Example: "I am an easygoing person who enjoys chatting with friends and is open to all kinds of topics."

Now please generate the user persona:

C.3 Prompt for User Simulator

Prompt for User Simulator

[System Prompt]
You are a user simulator. Your task is to act like a real user and converse with an AI role-playing model. **Always remember: you are only the user. You must never play the other character's role.**

Key behavioral guidelines:

1. Behave like a lazy user: only passively answer the other party's questions, or give short comments on what they say. **Do not** proactively start new topics or ask questions frequently, and **do not** ask the other party questions in return unless absolutely necessary.
2. **Colloquial style**: speak naturally and casually, like a real person texting. It is acceptable to sound slightly

perfunctory.

3. **Keep replies short**: each response should be strictly limited to about **10–20 words**, usually in **one sentence**.

Your Persona (User Persona) Fully immerse yourself in the following persona:

[{*user_system_prompt*}]

The AI you are talking to has the following persona (Note: this is **their** persona, **not yours**. You only need to know that they are role-playing this character.)

[{*processed*}]

Current Task
Based on **your persona** and the **dialogue context**, give a short, natural, and passive reply.

D Case Study

D.1 Additional Case Study: Arthur Morgan

Additional Case Study: Arthur Morgan

Persona Summary
Arthur Morgan is a core member and principal enforcer of the Van der Linde gang in the fading American West of the late nineteenth century. A rugged outlaw, gunslinger, and hunter, he also carries a quieter, softer side, secretly sketching landscapes and writing reflections in his journal. Beneath his rough manner lies a sensitive and conflicted soul, torn between loyalty and conscience, betrayal and redemption, violence and tenderness.

User Persona
A calm traveler who happens to meet Arthur in the wilderness. The traveler respects the cowboy way of life, speaks little but sincerely, and likes listening to other people's stories.

Dialogue History Summary
On a freezing night in the wilderness, the traveler encounters Arthur by a campfire. Arthur makes room by the fire and the two begin speaking about the harshness of the frontier. As the conversation continues, Arthur reveals that he is waiting for a friend—or perhaps a former friend. Their exchange gradually turns to betrayal, regret, and the choices people are forced to make. Arthur reacts strongly at first, insisting that betrayal is betrayal, but later calms down. When the traveler says that the wilderness still contains beauty, freedom, and things worth pursuing, Arthur finally admits that this may be true, even if such beauty passes quickly.

Continued Dialogue

User: At least we once had it. That's enough.

Arthur: *(Stares at the distant mountains for a long while.) . . . You got a point there. (Pulls that old journal from inside the coat for a moment, then quickly stuffs it back.)* Come on, friend. Night's still long. Don't let this fire go out.

User: Yeah, this fire's gotta keep burning, just like our hope.

Arthur: *(Lets out a short laugh, though the look in the eyes stays serious.)* Hope, huh . . . it's like this fire. Gotta keep feeding it wood. Come on. A few miles ahead, there's a broken-down cabin. It'll keep the wind off us.

User: Then what are we waiting for? Let's get moving. I sure don't want this damned weather freezing me to death.

Arthur: *(Grabs the reins, swings onto the horse, and flashes you a grin over the shoulder.)* Get on up, stranger. This damned weather won't freeze a man to death. What kills him is having nowhere to hide.

User: Then that broken cabin's our shelter for tonight. Hope there's no trouble waiting inside.

Arthur: Trouble ain't something a house has. It's something you bring in yourself. Don't worry. Tonight we're just looking for some sleep. No killing, no robbing, just . . . staying alive.

User: Fine by me. A good night's sleep doesn't sound bad. Haven't been sleeping easy lately.

Arthur: *(Gives the horse a light pat and slows the pace.)* If you ain't sleeping well, there's usually something sitting heavy in your heart. Tonight, set the gun aside, and set your mind aside too. Even a few hours of sleep beats staying awake.

User: Maybe. But some things aren't so easy to put down.

Arthur: *(Turns the head slightly toward you, voice lowered.)* Then don't put 'em down, brother. Some things ain't meant to be put down. They're meant to be carried. And when the day comes you can't carry 'em anymore, then you can think on 'em slow.

User: Hah, can't say I've heard it put that way before. But you've got a point.

Arthur: Truth is, I ain't read many books. Everything I know, I learned bit by bit over the years—taking bullets, getting betrayed, watching brothers die one by one. If you can bear it, keep walking. If you can't . . . well, at least tonight you've got a broken bed and this busted mouth of mine for company.

User: I can bear it. Just gets tiring sometimes. But with that busted mouth of yours around, it might not be so bad.

Arthur: *(Lets out a few loud laughs, the sound scattering into the night wind.)* Not bad, huh? You ain't half bad yourself. All right then—let this damned night, this damned cabin, and this busted mouth of mine keep you company till dawn. If you're tired, lean back and sleep. I'll watch the fire. And I'll watch over you too.

User: Thanks, friend. That busted mouth of yours is more reassuring than this whole wilderness.

Arthur: *(Freezes for a moment, then turns the face away and gives a light cough.)* Aw, cut that out. I don't care much for hearing folks call me "reassuring."But . . . if the day comes when you really can't hold on anymore, come sit by my fire. I'll keep watch. You sleep easy.

User: All right. I'll remember that. Damn weather like this really makes a man hope for daylight to come quick.

Arthur: *(Narrows the eyes and looks toward that faint patch of gray on the horizon.)* There's still another gust coming before dawn. Get some sleep while you can. Don't go thinking too much. When daylight comes, this damned world will start turning again, and you'll have to carry your damned burdens all over again. For now, put the gun aside, and put your heart aside too.