

RoleArena: Autonomous Plot Progression for Long-Horizon Multi-Agent Role-Playing

Qi Zhao¹, Jian Li², Yizhang Jin³, Dongqi Liu⁴,
Hang Ding³, Qi Song^{1†}, Xiangyang Li¹

¹University of Science and Technology of China, Hefei, Anhui, China

²Nanjing University, Nanjing, Jiangsu, China

³Shanghai Jiao Tong University, Shanghai, China

⁴Saarland University, Saarbrücken, Germany

zq2021@mail.ustc.edu.cn, swordl1dev@gmail.com, thebigshort@sjtu.edu.cn

dongqi.me@gmail.com, dearsloth@sjtu.edu.cn, {qisong09, xiangyangli}@ustc.edu.cn

Abstract

Plot progression is central to immersive role-playing language agents (RPLAs), yet existing resources largely consist of short interactions and rarely identify the sentences that advance a plot. We introduce *RoleArena*, a multi-agent environment for generating controllable, long-horizon role-playing dialogues. *RoleArena* discretizes a plot into states. An environment agent proposes state transitions and bridging narration, while a critic agent determines whether each transition is timely. Transitions approved by both agents serve as automatic plot-progression markers. We use *RoleArena* to construct *PlotStream*, a bilingual dataset with 39.3K character profiles and 170K dialogues, and fine-tune Qwen2.5 into *PlotRole-7B* and *PlotRole-72B*. We further propose *RoleArena-Eval*, which assesses dialogue fluency, character fidelity, emotional expression, story quality, and plot progression. Under matched experimental settings, *RoleArena* yields higher scores than two multi-agent role-playing frameworks, while *PlotRole* improves plot-progression scores over its Qwen2.5 backbones.

1 Introduction

Large language models (LLMs) increasingly support role-playing language agents (RPLAs) that simulate specific characters and provide interactive narrative experiences (Thoppilan et al., 2022; Chen et al., 2024; Shao et al., 2023). Such agents are used in role-playing games, digital humans, and other immersive applications (Ashby et al., 2023; Yao et al., 2024; Tao et al., 2024a). Beyond maintaining a persona within a single response, an en-

gaging RPLA must sustain coherent interaction and advance a story over many turns.

Current resources provide limited support for this long-horizon capability. First, many role-playing datasets contain relatively short exchanges, making it difficult to study behavior over extended interactions. Second, most datasets do not explicitly identify the utterances or narrations that move a story from one plot state to the next. Some recent work introduces plot-oriented data or evaluation (Tu et al., 2024; Lu et al., 2025; Yuan et al., 2024; Zhang et al., 2025), but identifying transition points can require substantial human effort and may be difficult to scale.

We introduce **RoleArena**, a multi-agent environment for generating long-horizon role-playing dialogues with controlled plot progression. *RoleArena* coordinates three agent roles. A **character agent** maintains a character-specific memory and generates thoughts, actions, and speech. An **environment agent** constructs the setting, theme, and discrete plot states; it also generates narration and proposes when to move to the next state. A **critic agent** checks proposed transitions against the remaining dialogue budget and plot complexity. A transition occurs only when the environment and critic agents agree. This explicit state change both regulates pacing and provides a plot-progression marker without asking a separate model to annotate each dialogue after generation.

Using *RoleArena*, we construct **PlotStream**, a bilingual Chinese–English dataset containing 39,312 character profiles and 170,148 long-form dialogues. We use *PlotStream* to fine-tune Qwen2.5 backbones into **PlotRole-7B** and **PlotRole-72B**, optimizing both environment-side narration and

[†]Corresponding author.

character-side responses at plot transitions. We also introduce **RoleArena-Eval**, a five-dimensional protocol for dialogue fluency, character fidelity, emotional expression, story quality, and plot progression.

Our contributions are threefold:

- We propose RoleArena, which combines discrete plot states, dynamic environment-agent decisions, and critic-assisted pacing to generate controllable long-horizon role-playing dialogues.
- We construct PlotStream with automatically derived plot-transition markers and train PlotRole-7B and PlotRole-72B to improve plot progression in both narration and character responses.
- We present RoleArena-Eval and evaluate the environment, dataset, and trained models under controlled multi-agent comparisons, human-LLM agreement checks, and component and sensitivity analyses.

2 Related Work

Multi-agent role-playing dialogue. Multi-agent environments assign models complementary roles to simulate scenes and generate dialogue. CharacterBox introduces narrator and character agents whose interactions advance a virtual-world simulation (Wang et al., 2025a). COSER uses an environment model to provide feedback during multi-agent dialogue simulation (Wang et al., 2025c). IBSEN coordinates director and actor agents to enact a story from a fictional outline (Han et al., 2024), while BookWorld organizes agents across scenes using concepts from drama theory (Ran et al., 2025). RoleArena differs by representing the plot as discrete states and requiring agreement between environment and critic agents before each state transition, thereby making both plot timing and dialogue length explicit control targets.

Role-playing datasets and plot progression.

One line of work extracts character profiles, dialogue, and context from novels or scripts (Li et al., 2023; Dai et al., 2025; Shao et al., 2023; Zhou et al., 2024). Another constructs synthetic profiles and simulated conversations, which broadens character coverage and reduces dependence on copyrighted source narratives (Ge et al., 2024; Lu et al., 2025; Wang et al., 2025b; Tao et al., 2024b). RolePlot is

particularly relevant because it evaluates and improves plot progression through plot-oriented data and model training (Zhang et al., 2025). However, it operates at the data/model level, whereas RoleArena is an online multi-agent environment that decides when to change plot states during generation. We therefore include RolePlot in the resource comparison but compare RoleArena directly with IBSEN and CharacterBox, which are more comparable at the framework level.

Role-playing evaluation. InCharacter measures personality fidelity through psychological interviews (Wang et al., 2024b), and DITTO evaluates arbitrary role-play obtained through self-alignment (Lu et al., 2024). CharacterBox evaluates actions in dynamic scenes (Wang et al., 2025a); COSER considers story quality, personification, character fidelity, and story consistency (Wang et al., 2025c); and CharacterEval measures dialogue ability, character consistency, role-playing appeal, and personality backtesting (Tu et al., 2024). RoleArena-Eval complements these dimensions with an explicit measure of whether character interactions proactively and coherently advance the plot.

3 RoleArena

3.1 Preliminaries

Dialogue turn. For a scene with M character agents, one dialogue turn is complete after all M agents have acted once.

Plot progression. We represent a plot as an ordered sequence of discrete states $\langle S_1, S_2, \dots, S_k \rangle$. Progression occurs when character actions and environmental events justify moving from the current state S_u to S_{u+1} .

3.2 Agent Architecture

RoleArena comprises an environment agent $\mathbb{E}$, one or more character agents $\mathbb{R}$, and a critic agent $\mathbb{C}$. Their interaction is illustrated in Figure 1.

3.2.1 Environment Agent

Before a dialogue D begins, $\mathbb{E}$ generates a theme, scene description, and plotline from the profiles of the participating character agents $[R_1, \dots, R_M]$. It discretizes the plotline into k states and considers the requested minimum dialogue length r when selecting k .

During generation, $\mathbb{E}$ observes the dialogue history $\mathcal{H}$, identifies which character should act next,

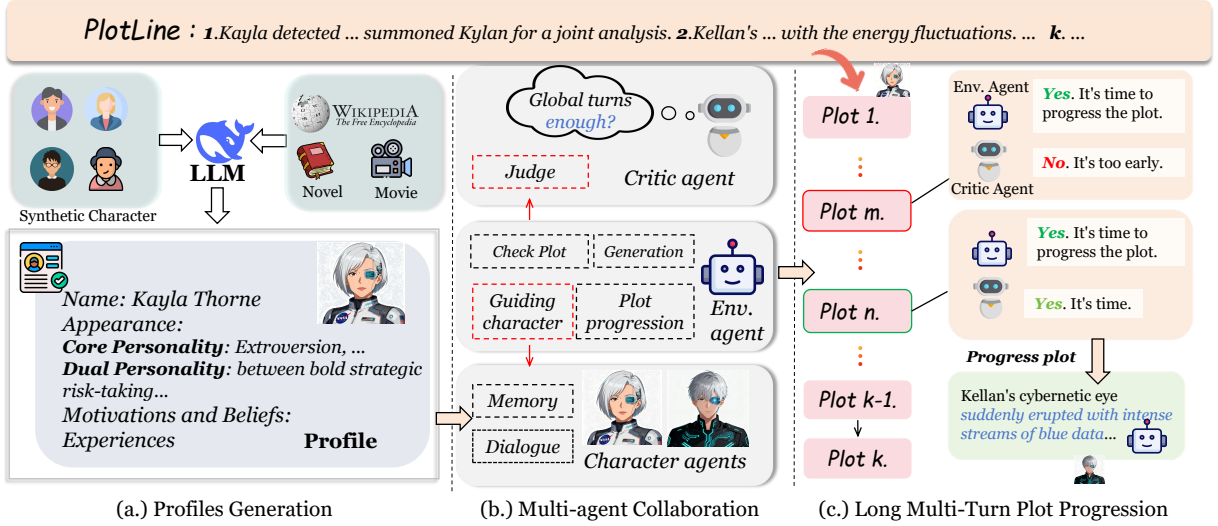


Figure 1: Overview of the RoleArena multi-agent environment. After generating character profiles in (a.), we drive the interactions of different agents in (b.), where the Env. agent sets up the plot and guides the Character agents to perform multi-turn dialogues in (c.) to achieve plot progression, while the Critic agent oversees the overall dialogue rounds, ensuring controllable dialogue turns and improving system robustness.

and produces narration n_i that reflects the current plot state. After each character action, it decides whether the current state has been sufficiently developed and, if so, proposes a transition from S_u to S_{u+1} . When a transition is approved, $\mathbb{E}$ generates transition narration n_i^* that connects the two states and guides the next character response r_i^* .

3.2.2 Character Agent

Each character agent conditions on its profile, private dialogue memory, the shared history, and the latest environment narration. It produces a structured response containing thoughts, actions, and speech. The environment agent’s guidance determines the next speaker but does not supply the character’s response.

3.2.3 Critic Agent

An environment agent alone may advance a plot prematurely. Therefore, whenever $\mathbb{E}$ proposes a transition, $\mathbb{C}$ independently evaluates the current state, the remaining dialogue budget ($r - i$), and the remaining plot states ($k - u$). Let $d_i^{\mathbb{E}}, d_i^{\mathbb{C}} \in \{0, 1\}$ denote their transition decisions. RoleArena updates the plot state only when both decisions are positive:

$$S_u \rightarrow S_{u+1} \iff d_i^{\mathbb{E}} = 1 \wedge d_i^{\mathbb{C}} = 1. \quad (1)$$

Otherwise, the system remains at S_u and continues developing the current state. This agreement mechanism prevents both rushed transitions

and excessive stagnation; its sensitivity to agent choice and progression thresholds is analyzed in Appendix F.

4 Dataset and Agent Training

4.1 PlotStream Dataset

We use RoleArena to construct PlotStream, a bilingual long-horizon role-playing dataset with state-derived plot-progression markers. Table 1 compares its scale and annotations with related resources. PlotStream contains 39,312 character profiles and 170,148 dialogues, with more than 50 turns per dialogue on average. It records speech, thoughts, actions, narration, and the responses associated with approved plot-state transitions.

4.1.1 Dataset Characteristics

PlotStream covers Chinese and English and contains ten broad synthetic character styles, detailed in Appendix H. Its transition labels are not produced by GPT-4o: a transition narration is marked when $\mathbb{E}$ and $\mathbb{C}$ agree to update S_u to S_{u+1} . This state-derived signal avoids a separate manual annotation pass for every generated dialogue.

4.1.2 Role Profile Construction

Starting from the one-sentence synthetic personas in PersonHub (Ge et al., 2024), DeepSeek-V3.1 generates 39,235 detailed profiles spanning ten styles. We add 77 literary-character profiles from CharacterEval (Tu et al., 2024) and COSER (Wang et al., 2025c), yielding 39,312 profiles in total.

Dataset	Character		Conversation			Dialogue			Plot
	Num.	Profile	#Turns	#Conv.	Multi-Chara.	Speech	Thought	Action	Plot Prog.
Character-LLM (Shao et al., 2023)	9	✓	13.2	14,300		✓			
ChatHaruhi (Li et al., 2023)	32	✓	>2	54,726	✓	✓			
RoleLLM (Wang et al., 2024a)	100	✓	2	140,726		✓			
CroSS-MR (Yuan et al., 2024)	126	✓	2	445					
CharacterGLM (Zhou et al., 2024)	250	✓	15.8	1,034		✓			
CharacterEval (Tu et al., 2024)	77	✓	9.3	1,785		✓		✓	
DITTO (Lu et al., 2024)	4,002	✓	5.1	7,186		✓			
MMRole (Dai et al., 2025)	85	✓	4.2	14,346		✓			
CharacterBench (Zhou et al., 2025)	3,956	✓	11.3	13,162		✓			
COSER (Wang et al., 2025c)	17,966	✓	13.2	29,798	✓	✓	✓	✓	
RolePlot (Zhang et al., 2025)	20	✓	>2	28,127		✓			✓
RoleMRC (Lu et al., 2025)	10,200	✓	9.5	37,900		✓			
PlotStream(Ours)	39,312	✓	>50	170,148	✓	✓	✓	✓	✓

Table 1: Comparison of PlotStream with existing role-playing datasets. Num., #Turns, #Conv., and Multi-Chara. denote the number of characters, dialogue turns, dialogues, and support for multiple characters, respectively.

Language	P_o	Accept Rate	P_e	α	p
ZH	0.96	0.88	0.79	0.86	< 0.01
EN	0.95	0.83	0.71	0.83	< 0.01
Overall	0.96	0.86	0.75	0.84	< 0.01

Table 2: Human audit of PlotStream. P_o and P_e denote observed and expected agreement, α is Krippendorff’s alpha, and p is the reported p -value.

4.1.3 Dialogue Simulation and Filtering

DeepSeek-V3.1 instantiates RoleArena’s three agent roles to generate the main PlotStream dialogues. We remove incomplete generations and excessively long or unresolved dialogues, then apply the quality-control procedure in Appendix I. GPT-4o is used only for post-generation filtering, which removes entries with a composite score below 3.0; it does not create the state-derived transition labels. Plotline-diversity measurements are reported in Appendix B.

To assess the retained data independently of automatic filtering, five experienced annotators inspect a random sample of 500 dialogue sets, with an average of 2.6 judgments per set. A set is rejected if its overall RoleArena-Eval score is below 3.0 or any dimension is below 2.0. The audit yields an 86% acceptance rate and Krippendorff’s $\alpha = 0.84$; roughly half of the rejected audited sets reach the maximum turn limit without completing the storyline. Because GPT-4o participates in automatic filtering and is also the primary evaluator, the retained data and reported scores may share model-specific preferences; we account for this dependence when interpreting downstream gains.

4.2 PlotRole Training

PlotRole learns two outputs at approved plot transitions: narration n^* and the following character response r^* . For narration, C_{n^*} contains the role-playing instruction and the alternating narration–response history. For the character response, C_{r^*} additionally contains n^* and the target character’s profile and instruction. We optimize

$$\theta^* = \arg \min_{\theta} \left[\mathbb{E}_{(C_{n^*}, n^*)} \mathcal{L}_{\text{CE}}(f(C_{n^*}; \theta), n^*) + \mathbb{E}_{(C_{r^*}, r^*)} \mathcal{L}_{\text{CE}}(f(C_{r^*}; \theta), r^*) \right]. \quad (2)$$

The response target includes the character’s thoughts, actions, and speech at the transition point.

5 RoleArena-Eval

RoleArena-Eval assesses long-horizon role-playing dialogue along five dimensions: **Dialogue Fluency (DF)** measures naturalness, coherence, and consistency across turns; **Character Fidelity (CF)** measures alignment with a character’s background, personality, and abilities; **Emotional Expression (EE)** measures emotional authenticity and consistency; **Story Quality (SQ)** measures thematic coherence, suspense, and engagement; and **Plot Progression (PP)** measures whether character interactions advance the story naturally, proactively, and without repetition.

Each dimension uses a five-point Likert-type scale (Likert, 1932). Following a critique-then-score protocol, the evaluator first produces dimension-specific feedback and then assigns scores. Our main experiments use GPT-4o as the evaluator and report complementary human and cross-model agreement analyses.

6 Experiments

6.1 Experimental Setup

Data and models. We split PlotStream 9:1 into training and test sets and fine-tune Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct using Equation 2. For each evaluated model, we generate 20 Chinese and 20 English dialogue sets per run and average results over 10 runs. All models use the same held-out character profiles, plotlines, themes, and environments.

Training details. PlotRole-7B is trained on eight H20 GPUs (96 GB each), while PlotRole-72B uses 24 H20 GPUs across three nodes (eight per node; 96 GB each). Both models use a learning rate of 10^{-5} , gradient accumulation of 16, and a maximum sequence length of 16,384 tokens.

At inference, the evaluated LLM is instantiated separately as the environment, character, and critic agents. The three instances share the same underlying model but receive role-specific prompts. The environment-agent instance generates narration rather than receiving fixed external narration; the character-agent instance generates thoughts, actions, and speech; and the critic-agent instance evaluates proposed transitions.

Evaluation context. The main results use $K = 15$ dialogue turns. To reduce length-related evaluation bias (Wang et al., 2025c), DF, CF, EE, and PP are evaluated on a random window of K turns, while SQ receives the dialogue prefix from the start through the endpoint of the same K -turn window.

Model baselines. Open-source baselines are Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Baichuan2-13B-Chat, Qwen2.5-72B-Instruct, Hunyuan-7B-Instruct, ChatGLM3-6B, Llama-3.1-8B, Llama-3.1-70B-Instruct, and DeepSeek-V3.1. Closed-source baselines are GPT-4.1-mini, GPT-3.5, GPT-4o-mini, Gemini-1.5-Flash, Gemini-1.5-Pro, Claude-3.5-Haiku, Claude-3.5-Sonnet, and Doubao-1.5-lite.

Framework baselines. We compare RoleArena with IBSEN (Han et al., 2024) and CharacterBox (Wang et al., 2025a), two online multi-agent role-playing frameworks. All three use DeepSeek-V3.1 as the dialogue backbone, the same test cases and character/environment information, and a common target of more than 50 turns. Task content and length constraints are held constant, while prompts retain the interfaces required by each framework: director-actor for IBSEN, narrator-character for CharacterBox, and environment-character-critic

for RoleArena. Appendix C documents this alignment in more detail. RolePlot (Zhang et al., 2025) is not used as a direct framework baseline because it is an offline data-construction and model-training method rather than an online multi-agent environment; we instead compare its resource characteristics in Table 1.

Evaluation. GPT-4o is the primary RoleArena-Eval judge. It is also used in PlotStream’s post-generation filter, but it does not generate PlotStream’s plot-transition labels. We therefore complement the main evaluation with five expert annotators, Claude-Sonnet-4.5, generator sensitivity tests, and component ablations.

6.2 Main Results

PlotRole improves plot progression. Tables 3 and 4 report the English and Chinese results. Relative to Qwen2.5-7B-Instruct, PlotRole-7B raises the average score from 2.469 to 3.416 in English (+0.947) and from 2.448 to 3.093 in Chinese (+0.645); its PP gains are +1.347 and +0.937, respectively. At 72B, PlotRole improves the average by 0.282 in English and 0.324 in Chinese, with corresponding PP gains of 0.454 and 1.002; it also attains the highest English CF score (4.712). These paired backbone comparisons more directly isolate the effect of PlotStream training, although they do not establish unrestricted out-of-domain generalization.

Cross-language behavior. Fourteen of the 19 models score lower in Chinese than in English; the largest absolute difference in average score is 0.495 points (Gemini-1.5-Flash). PlotRole improves the average score over its corresponding Qwen2.5 backbone in both languages, although neither PlotRole variant leads every evaluation dimension.

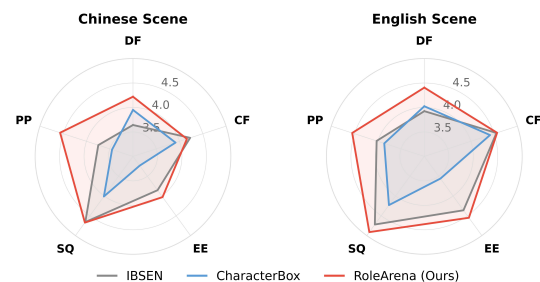


Figure 2: Framework comparison in Chinese and English using the same DeepSeek-V3.1 backbone, test cases, character and environment information, and dialogue-length requirement.

Model	English Scene					Average
	DF	CF	EE	SQ	PP	
Closed-source Models						
GPT-4.1-mini	4.258±0.23	4.333±0.17	4.396±0.42	4.920±0.08	4.526±0.35	4.487±0.25
GPT-3.5	3.650±0.31	3.947±0.29	4.233±0.16	4.271±0.44	4.056±0.27	4.031±0.29
GPT-4o-mini	3.926±0.38	4.185±0.13	4.278±0.49	3.889±0.25	3.630±0.18	3.982±0.29
Gemini-1.5-Pro	3.950±0.47	4.512±0.22	4.333±0.41	4.448±0.36	4.012±0.29	4.251±0.35
Gemini-1.5-Flash	3.778±0.15	4.455±0.34	3.917±0.28	4.583±0.19	3.899±0.42	4.126±0.28
Claude-3.5-Haiku	4.155±0.26	3.778±0.48	4.027±0.31	4.167±0.24	3.556±0.11	3.937±0.28
Claude-3.5-Sonnet	4.260±0.39	4.243±0.17	3.974±0.46	4.380±0.32	4.015±0.25	4.174±0.32
Doubao-1.5-lite	3.892±0.43	4.420±0.21	4.295±0.38	3.994±0.27	4.127±0.34	4.146±0.33
Open-source Models						
Qwen2.5-7B-Instruct	2.238±0.22	2.191±0.15	2.512±0.41	3.071±0.33	2.333±0.28	2.469±0.28
Qwen2.5-14B-Instruct	3.333±0.29	4.132±0.44	4.028±0.18	4.333±0.25	3.557±0.31	3.877±0.29
Qwen2.5-72B-Instruct	3.910±0.37	4.232±0.26	4.253±0.49	4.560±0.21	4.106±0.35	4.212±0.34
Hunyuan-7B-Instruct	2.200±0.14	2.333±0.38	2.100±0.27	2.700±0.45	1.933±0.32	2.253±0.31
Llama-3.1-70B-Instruct	3.146±0.41	2.542±0.22	3.176±0.34	3.312±0.28	2.625±0.19	2.960±0.29
Llama-3.1-8B	2.222±0.33	2.667±0.29	2.333±0.16	2.500±0.42	2.080±0.37	2.360±0.31
Baichuan2-13B-Chat	3.250±0.27	4.012±0.35	3.584±0.24	3.925±0.18	3.230±0.46	3.600±0.30
ChatGLM3-6B	2.350±0.19	2.570±0.43	2.412±0.32	3.010±0.26	2.527±0.41	2.574±0.32
DeepSeek-V3.1	4.406±0.28	4.564±0.15	4.550±0.37	4.912±0.22	4.550±0.31	4.596±0.27
PlotRole-7B	2.444±0.36	3.556±0.24	3.335±0.29	4.067±0.17	3.680±0.43	3.416±0.30
PlotRole-72B	3.956±0.25	4.712±0.32	4.417±0.41	4.824±0.28	4.560±0.19	4.494±0.29

Table 3: English RoleArena-Eval scores for multi-turn dialogues. **Boldface** and underlining mark the best and second-best scores in each column, respectively.

RoleArena improves multi-agent dialogue generation. Figure 2 compares the three frameworks under the controlled protocol above. RoleArena obtains higher scores in all five RoleArena-Eval dimensions in both languages, with the clearest gains in story quality and plot progression. Because each framework retains its native interface, this is a framework-level comparison rather than a single-prompt ablation. The concentration of gains in SQ and PP is consistent with RoleArena’s intended mechanism: explicit state decisions regulate when a story advances and how its pacing is maintained.

Human agreement. Five expert annotators, each with 2–3 years of role-playing experience, evaluate the same $K = 15$ setting as the main results. Figure 3 reports Pearson correlations between their judgments and those of GPT-4o and Claude-Sonnet-4.5; the average human–LLM correlation exceeds 0.85 in both languages. For a separate context-length analysis, we sample 50 test-set character sets and generate dialogues longer than 30 turns. Table 5 reports GPT-4o–human alignment rates at $K \in \{5, 10, 20\}$, which decrease to 79.60% in English and 79.10% in Chinese at $K = 20$; further protocol details appear in Appendix A. This decline motivates the $K = 15$ main setting and cautions against treating short-context agreement as evidence of equally reliable holistic judgment over arbitrarily long dialogues.

6.3 Ablation and Sensitivity Analysis

Table 6 reports ablations of the environment agent, the discrete plotline, and the critic agent. Removing the environment agent causes the largest overall degradation, while removing the plotline particularly reduces SQ and PP. Removing the critic produces a smaller but consistent reduction and weakens control over transition timing. Appendix K reports mixed-generator and cross-model tests; these checks reduce dependence on a single generator but are not a full factorial analysis of the generator, filter, and judge.

6.4 Robustness and Scope Checks

Generator sensitivity. Table 7 compares the DeepSeek-V3.1-only pipeline with a setting that samples from DeepSeek-V3.1, Qwen2.5-72B-Instruct, and Claude-Sonnet-4.5. The mixed setting changes every dimension average by at most 0.099 points: DF, CF, EE, and SQ increase slightly, while PP decreases from 4.566 to 4.518. This stability suggests that the scores are not driven solely by a single dialogue generator, although the experiment does not separate interactions among generation, filtering, and judging.

Agent assignment. We separately replace either the environment or critic with Qwen2.5-7B while retaining DeepSeek-V3.1 for the other role. Table 8 shows that the weaker critic reduces completion from 95% to 79% in Chinese and from 94% to 75% in English, whereas replacing only the environment

Model	Chinese Scene					Average
	DF	CF	EE	SQ	PP	
Closed-source Models						
GPT-4.1-mini	4.210±0.280	4.467±0.190	4.125±0.330	4.530±0.240	3.821±0.410	4.231±0.290
GPT-3.5-Turbo	3.845±0.350	4.217±0.220	4.214±0.380	3.945±0.310	4.081±0.270	4.060±0.310
GPT-4o-mini	4.058±0.260	3.667±0.420	4.357±0.290	4.292±0.360	3.487±0.330	3.972±0.330
Gemini-1.5-Pro	4.125±0.310	4.059±0.280	3.946±0.350	4.390±0.220	4.158±0.390	4.136±0.310
Gemini-1.5-Flash	3.652±0.370	3.127±0.250	3.847±0.410	4.109±0.290	3.418±0.340	3.631±0.330
Claude-3.5-Haiku	3.751±0.290	3.493±0.360	4.025±0.270	3.942±0.430	3.645±0.320	3.771±0.330
Claude-3.5-Sonnet	4.056±0.330	4.247±0.240	3.756±0.390	4.085±0.280	4.284±0.350	4.086±0.320
Doubao-1.5	4.194±0.220	3.849±0.310	3.994±0.340	4.239±0.370	4.384±0.260	4.132±0.300
Open-source Models						
Qwen2.5-7B-Instruct	2.185±0.380	2.333±0.270	2.556±0.320	2.944±0.350	2.222±0.290	2.448±0.320
Qwen2.5-14B-Instruct	3.333±0.310	3.708±0.340	3.438±0.280	3.750±0.410	2.792±0.360	3.404±0.340
DeepSeek-V3.1	4.220±0.250	4.158±0.290	4.028±0.370	4.672±0.230	4.566±0.330	4.329±0.290
Qwen2.5-72B-Instruct	3.722±0.340	3.778±0.260	3.583±0.410	4.583±0.320	3.167±0.380	3.767±0.340
Hunyuan-7B-Instruct	2.667±0.280	2.750±0.330	2.500±0.350	3.250±0.290	2.167±0.420	2.667±0.330
Llama-3.1-70B-Instruct	2.560±0.360	3.067±0.310	3.113±0.280	3.650±0.340	3.000±0.390	3.078±0.330
Llama-3.1-8B	2.222±0.290	2.667±0.350	2.357±0.320	2.560±0.380	2.120±0.270	2.385±0.320
Baichuan2-13B-Chat	3.131±0.330	3.574±0.280	3.175±0.360	4.015±0.310	3.286±0.340	3.436±0.320
ChatGLM3-6B	2.667±0.310	2.460±0.360	2.598±0.290	2.950±0.330	2.458±0.370	2.627±0.330
PlotRole-7B	2.791±0.270	3.047±0.320	2.956±0.340	3.514±0.280	3.159±0.350	3.093±0.310
PlotRole-72B	4.102±0.230	3.765±0.290	3.921±0.310	4.498±0.260	4.169±0.320	4.091±0.280

Table 4: Chinese RoleArena-Eval scores for multi-turn dialogues. **Boldface** and underlining mark the best and second-best scores in each column, respectively.

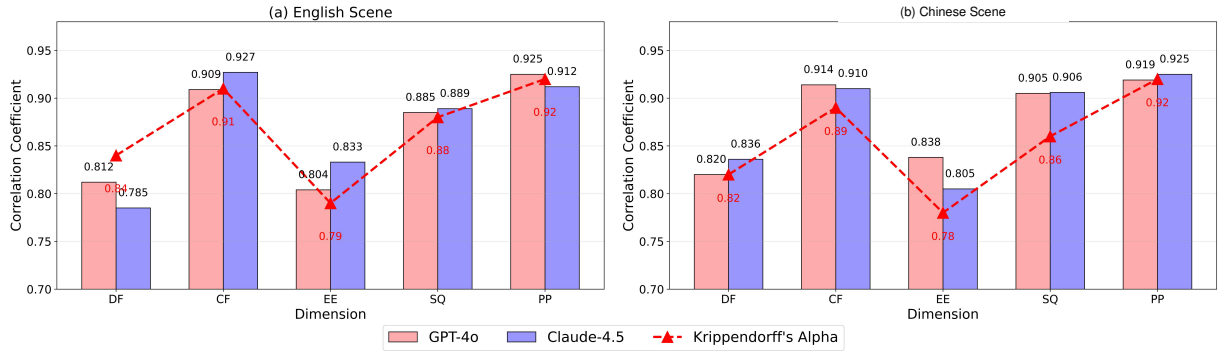


Figure 3: Correlation of GPT-4o and Claude-Sonnet-4.5 judgments with expert annotations in English and Chinese, together with inter-annotator consistency.

Judge	K	Alignment Rate (%)	
		EN	ZH
GPT-4o	5	89.56	87.30
	10	85.45	84.10
	20	79.60	79.10

Table 5: Agreement between GPT-4o and human judgments as evaluation context grows. EN and ZH denote English and Chinese.

gives 91% and 90%. This asymmetry is consistent with the critic’s responsibility for regulating state transitions against the remaining dialogue budget.

Progression control. Table 9 varies the maximum per-state turn budget and its permitted transition range over 50 English dialogue groups and ten runs. The complete system remains at 92–93% success across the tested settings; removing forced pro-

Setting	DF	CF	EE	SQ	PP
Full RoleArena	4.406±0.23	4.564±0.17	4.550±0.42	4.912±0.08	4.550±0.35
w/o Environment	3.520±0.31	4.015±0.29	3.985±0.16	3.176±0.44	3.050±0.27
w/o Plotline	4.218±0.38	4.442±0.13	4.563±0.49	3.417±0.25	3.695±0.18
w/o Critic	4.395±0.47	4.550±0.22	4.486±0.41	4.680±0.36	4.262±0.29

Table 6: DeepSeek-V3.1 component ablations (mean ± standard deviation).

gression lowers success to 72–78%. Thus, explicit progression control is important for completing the storyline under a long-horizon turn requirement, while the exact choice among the tested thresholds is less consequential.

Unseen-scenario stress test. Gemini-2.5 and Claude-Sonnet-4.5 independently generate previously unseen profiles, scenes, and plotlines. RoleArena completes approximately 93–94% of these dialogues under the criterion that the storyline is completed and the minimum turn requirement

Generator Setting	DF	CF	EE	SQ	PP
DeepSeek-V3.1 only	4.220	4.158	4.028	4.672	4.566
Mixed generators	4.256	4.166	4.127	4.694	4.518

Table 7: Generator sensitivity. The mixed setting samples from DeepSeek-V3.1, Qwen2.5-72B-Instruct, and Claude-Sonnet-4.5.

Lang	Env Agent	Critic Agent	Succ. Rate
ZH	DeepSeek-V3.1	DeepSeek-V3.1	95%
	DeepSeek-V3.1	Qwen2.5-7B	79%
	Qwen2.5-7B	DeepSeek-V3.1	91%
EN	DeepSeek-V3.1	DeepSeek-V3.1	94%
	DeepSeek-V3.1	Qwen2.5-7B	75%
	Qwen2.5-7B	DeepSeek-V3.1	90%

Table 8: RoleArena success rate when assigning different models to the environment and critic roles.

Setting	Max Turns	Allowed Range	Succ. Rate	Avg Turns
Base	6	3-5	93%	4.3
	5	2-4	92%	3.5
	7	4-6	93%	5.1
No Forced Progression	6	3-5	72%	4.1
	5	2-4	78%	3.6
	7	4-6	75%	5.5

Table 9: Sensitivity to progression control. Success requires completing the storyline and meeting the minimum turn requirement.

is met. This result tests framework-level generation robustness, not PlotRole generalization to an independently authored benchmark.

Simulated-user stress test. Across ten runs over 100 science-fiction and horror scenes, GPT-4o simulates satisfied, dissatisfied, and preference-driven users. PlotRole-72B sustains 4.7 versus 2.9 turns in science fiction and 4.3 versus 3.1 turns in horror relative to Qwen2-72B-Instruct (Appendix D). These are LLM-simulated interactions and therefore cannot substitute for a study with human participants.

Style	Qwen2-72B-Instruct	PlotRole-72B	Δ Turns	Δ Rate (%)
Sci-Fi	2.9	4.7	1.8	62
Horror	3.1	4.3	1.2	39

Table 10: Average turns with LLM-simulated users. Interactions stop after five turns or earlier simulated-user termination; these results are not a human-user evaluation.

6.5 Case Study

Figures 7 and 6 present Chinese and English examples. Highlighted spans show the narrations and responses associated with approved plot-state transitions, illustrating how the agents move between plot states while retaining character-specific

expression.

Our code is available at <https://github.com/tize-72/RoleArena>.

7 Conclusion

We presented RoleArena, a multi-agent environment that uses explicit plot states and environment-critic agreement to generate long-horizon role-playing dialogues with identifiable transitions. RoleArena serves as the generation pipeline for PlotStream, a bilingual dataset with state-derived plot-progression markers, and PlotRole-7B/72B, models trained for transition narration and character response. Our experiments show improvements over controlled framework and model baselines, while the human-agreement and sensitivity analyses clarify the scope of the evidence. Future work will prioritize real-user evaluation, independent scenarios and judges, adaptive plot planning, broader language coverage, and multimodal interaction.

Limitations

RoleArena relies on discrete plotlines generated before dialogue, which improves controllability but may constrain narrative diversity in highly open-ended scenarios. PlotStream covers Chinese and English only, and its 9:1 split primarily measures held-out in-domain performance rather than generalization to independently authored scenarios. Most character profiles are synthetic, so they may not capture the full complexity of human-authored characters.

Our evaluation also relies partly on LLM judges. GPT-4o is used both for post-generation filtering and as the primary evaluator, which may introduce judge-specific preferences. Human agreement is strong at the main setting of $K = 15$, but it decreases as the evaluated context grows from $K = 5$ to $K = 20$. Moreover, our sensitivity experiments do not constitute a full factorial ablation over the generator, filter, and judge. The user-participation study uses LLM-simulated users rather than real participants; consequently, robustness to unpredictable human behavior remains unverified. Finally, RoleArena is limited to text-only interaction. Future work should evaluate real users, independent scenarios and judges, broader languages, adaptive plot planning, and multimodal interaction.

Ethics Statement

This work is intended for research on long multi-turn role-playing dialogue generation. We acknowledge that synthetic character profiles and dialogues may contain biases inherited from the underlying language models. In addition, highly immersive role-playing systems may be misused to generate deceptive or manipulative content. We therefore recommend appropriate content moderation and human oversight in downstream applications.

Acknowledgments

This research is partially supported by the Quantum Science and Technology–National Science and Technology Major Project (QNMP) under Grant 2021ZD0302900, the National Natural Science Foundation of China under Grants 92567301, 62132018, and 62231015, and the Anhui Provincial Science and Technology Innovation Program under Grant 202523009050019.

References

- Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. [Personalized quest and dialogue generation in role-playing games: A knowledge graph- and language model-based approach](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20. ACM.
- 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. 2024. [From persona to personalization: A survey on role-playing language agents](#). *Transactions on Machine Learning Research*.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. [MMRole: A comprehensive framework for developing and evaluating multimodal role-playing agents](#). In *International Conference on Learning Representations*.
- Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. [Scaling synthetic data creation with 1,000,000,000 personas](#). *arXiv preprint arXiv:2406.20094*.
- Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. [IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand. Association for Computational Linguistics.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. [ChatHaruhi: Reviving anime character in reality via large language model](#). *arXiv preprint arXiv:2308.09597*.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.
- Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. [RoleMRC: A fine-grained composite benchmark for role-playing and instruction-following](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21008–21030, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. [BOOK-WORLD: From novels to interactive agent societies for story creation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-LLM: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024a. [RoleCraft-GLM: Advancing personalized role-playing in large language models](#). In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.
- Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024b. [ChatGPT role-play dataset: Analysis of user motives and model naturalness](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3133–3145, Torino, Italy. ELRA and ICCL.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. [LaMDA: Language models for dialog applications](#). *arXiv preprint arXiv:2201.08239*.

- 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.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. [CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds](#). 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 6372–6391, Albuquerque, New Mexico. 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.
- Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025b. [OpenCharacter: Training customizable role-playing LLMs with large-scale synthetic personas](#). *arXiv preprint arXiv:2501.15427*.
- 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. 2025c. [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.
- Ruiqi Yao, Guijie Qi, Zhiqiang Wu, Hua Sun, and Dongfang Sheng. 2024. [Digital human calls you dear: How do customers respond to virtual streamers’ social-oriented language in e-commerce livestreaming? a stereotyping perspective](#). *Journal of Retailing and Consumer Services*, 79:103872.
- Xinfeng Yuan, Siyu Yuan, Yuhao Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. [Evaluating character understanding of large language models via character profiling from fictional works](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.
- Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025. [RolePlot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354, Vienna, Austria. Association for Computational Linguistics.
- 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. 2024. [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, USA. 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. 2025. [CharacterBench: Benchmarking character customization of large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110. Association for the Advancement of Artificial Intelligence.

A Additional Human Evaluation

We sample 50 character sets from the test set and generate dialogues longer than 30 turns using DeepSeek-V3.1 and PlotRole. Table 5 compares GPT-4o judgments with human judgments at $K \in \{5, 10, 20\}$. Agreement is highest at $K = 5$ and decreases to 79.60% in English and 79.10% in Chinese at $K = 20$. This trend indicates that holistic evaluation becomes less reliable as the context grows and motivates our use of $K = 15$ for the main results.

B Plotline Diversity Analysis

We measure storyline similarity with Qwen3-Embedding-8B and KaLM-Embedding-Gemma3-12B-2511, two multilingual embedding models listed on the MTEB leaderboard. For each model,

we run ten samples of 100 dialogue sets and average the resulting similarities. Tables 11 and 12 show a maximum average similarity of 0.4422 under these measurements.

Style (EN)	SIM (EN)	Style (ZH)	SIM (ZH)
Sci-Fi	0.4105	Sci-Fi	0.4293
Fantasy	0.4110	Fantasy	0.4198
Workplace Drama	0.3812	Workplace Drama	0.3506
Historical	0.3480	Historical	0.3442
Horror	0.3995	Horror	0.4251
Romance	0.3544	Romance	0.3595
Mystery	0.4155	Mystery	0.4315
Adventure	0.4082	Adventure	0.4307
Urban Fiction	0.3146	Urban Fiction	0.3407
Superpower	0.4238	Superpower	0.4536
Average	0.3867	Average	0.3985

Table 11: Plotline similarity by style using Qwen3-Embedding-8B.

Style (EN)	SIM (EN)	Style (ZH)	SIM (ZH)
Sci-Fi	0.4738	Sci-Fi	0.5049
Fantasy	0.4681	Fantasy	0.4877
Workplace Drama	0.4052	Workplace Drama	0.3981
Historical	0.4116	Historical	0.3960
Horror	0.4536	Horror	0.4440
Romance	0.4049	Romance	0.4139
Mystery	0.4439	Mystery	0.4458
Adventure	0.4443	Adventure	0.4790
Urban Fiction	0.3789	Urban Fiction	0.3874
Superpower	0.4515	Superpower	0.4654
Average	0.4336	Average	0.4422

Table 12: Plotline similarity by style using KaLM-Embedding-Gemma3-12B-2511.

C Framework Baseline Protocol

The IBSEN, CharacterBox, and RoleArena comparison uses DeepSeek-V3.1 for every model-driven role. We reuse the same held-out cases, character profiles, scene descriptions, and minimum length requirement of more than 50 turns. We preserve common task information while adapting its placement to each framework’s native interface. For IBSEN, the outline, scene, and profiles are supplied through its director and actor roles; for CharacterBox, they are supplied through its narrator and character roles; and for RoleArena, they are supplied through the environment and character roles, with the critic receiving the current state and remaining-turn information. Consequently, the semantic inputs and generation budget are aligned, but the prompts are not textually identical because the frameworks implement different interaction protocols.

RolePlot is not included in this direct comparison because it is a plot-oriented data and model-

training method rather than an online multi-agent environment. It is instead included in Table 1 and discussed in Section 2.

D Simulated User Participation

We use GPT-4o to simulate satisfied, dissatisfied, and preference-driven users in science-fiction and horror scenarios. Across ten runs over 100 scenes, we compare Qwen2-72B-Instruct with PlotRole-72B and end an interaction after five turns or earlier user termination. Table 10 shows that PlotRole-72B sustains more turns in both styles. Because these users are LLM simulations, this experiment is a stress test rather than evidence from real participants.

E Framework-Level Out-of-Distribution Test

For a complementary framework-level test, Gemini-2.5 and Claude-Sonnet-4.5 independently generate previously unseen character profiles, scenes, and plotlines. RoleArena completes approximately 93–94% of these dialogues, where success requires completing the storyline while meeting the minimum turn requirement. This result concerns the robustness of the RoleArena generation framework; it does not establish out-of-domain generalization of PlotRole to an independent benchmark.

F Sensitivity to the Plot-Progression Threshold

We sample 50 English dialogue groups over ten runs and vary the progression rules. Table 9 shows that RoleArena retains a generation success rate above 90% under the tested threshold settings, whereas removing forced progression reduces success by as much as 21 percentage points. A dialogue is successful only if it completes the storyline and reaches the required minimum number of turns; Avg. Turns denotes the average turns spent at each plot state.

G RoleArena Algorithm

Algorithm 1 summarizes RoleArena. The training-data extraction step selects the character response immediately following each approved state transition, so this selection is not represented as a separate online generation step.

Algorithm 1 RoleArena

Require: Roles $\mathcal{R}$, metadata $\mathbf{m}$, plot lines $\mathcal{L}$, policy π

Ensure: Dialogue trajectory $\mathbf{D}$

```
1: Initialize agents  $\mathcal{A} \leftarrow \{\pi_a(r_i)\}$ 
2: Extract story setting  $\mathbf{s} \leftarrow \text{extract}(\mathbf{m})$ 
3: Parse plot nodes  $\mathcal{P} = \{p_1, \dots, p_K\}$ 
4:  $t \leftarrow 1, i \leftarrow 0, \mathbf{D} \leftarrow \emptyset$ 
5: while  $t \leq T_{\max}$  and  $i < K$  do
6:   Initialize turn  $\text{turn}_t$ 
7:   if  $t = 1$  then
8:      $(n_t, a_t, \mathcal{U}) \leftarrow \text{env\_init}(\mathbf{s}, \mathcal{R})$ 
9:   else
10:     $(n_t, a_t) \leftarrow \text{env\_step}(\text{turn}_{t-1}, p_i)$ 
11:   end if
12:   while  $|\mathcal{U}| > 0$  do
13:      $r_{a_t} \sim \pi(a_t \mid n_t, \mathcal{M}_{a_t})$ 
14:     Append  $r_{a_t}$  to  $\text{turn}_t$ 
15:      $\mathcal{U} \leftarrow \mathcal{U} \setminus \{a_t\}$ 
16:     if  $\text{judge\_plot}(\text{turn}_t, p_i)$  then
17:        $\mathbf{c} \leftarrow \text{critic}(\text{turn}_t, p_i)$ 
18:       if  $\mathbf{c}.\text{approve}$  then
19:          $\text{turn}_t.\text{advance} \leftarrow \text{true}$ 
20:          $i \leftarrow i + 1$ 
21:         break
22:       end if
23:     end if
24:      $(a_t, n_t) \leftarrow \text{env\_next}(r_{a_t}, \mathcal{U}, p_i)$ 
25:   end while
26:   Append  $\text{turn}_t$  to  $\mathbf{D}$ 
27:    $t \leftarrow t + 1$ 
28: end while
29: return  $\mathbf{D}$ 
```

H PlotStream Data Statistics

H.1 Role Diversity

We organize synthetic profiles into ten styles using dimensions inspired by literature, film, and games, including the source of conflict, narrative function, and aesthetic style. Table 13 lists their defining traits and examples.

H.2 Dialogue Distribution

Figure 4 shows the language-specific dialogue-length distributions for each character style, while Figure 5 summarizes the English, Chinese, and overall distributions. Table 14 provides the corresponding values. Chinese dialogues are longer on average and account for approximately 60% of PlotStream.

I PlotStream Dataset Validation

Five annotators with 2–3 years of role-playing experience inspect a random sample of 500 dialogue sets. Assignments balance Chinese and English and provide overlapping judgments, with an average of 2.6 annotators per set. Annotators score the five RoleArena-Eval dimensions; a set is rejected if its overall score is below 3.0 or any dimension is below 2.0.

Table 2 reports an overall acceptance rate of 86% and Krippendorff’s $\alpha = 0.84$. Among the rejected audited sets, approximately half reach the maximum turn limit without completing the storyline; these failed generations are removed. As an additional automatic quality-control stage, GPT-4o and RoleArena-Eval remove generated entries with a composite score below 3. This filter may favor GPT-4o’s preferences and may discard unconventional but valid narratives, especially because GPT-4o is also the primary evaluator.

J Case Studies

The English example in Figure 6 centers on a disagreement between Kaelen Vance and Kaelen Vex over whether neural-log anomalies indicate awareness or post-mortem decay. The narration records observable cues—eye movements, gestures, and the surrounding data terminals—while the two speakers maintain opposing interpretations of the same evidence. The Chinese example in Figure 7 begins with an attempt to stop a data parasite and then advances when the parasite consumes its own code and a phase discrepancy reveals that Laurent

Category	Features	Style	Examples
Sci-Fi	Advanced technology, futuristic settings, extraterrestrial life, virtual reality, space adventures	High-tech interfaces, futuristic cityscapes, alien species	Alien explorers, AI assistants, rebel commanders
Fantasy	Magic systems, mythical creatures, alternate histories, medieval or ancient cultures	Enchanted realms, swords and shields, heroic quests	Sorcerers, warriors, dragon riders, prophets
Workplace Drama	Interpersonal dynamics, power struggles, career progression in professional environments	Realistic portrayals of office hierarchies, human motivations	CEOs, secretaries, junior analysts, journalists
Historical	Specific eras, recreated events, cultural figures from past periods	Period attire, authentic settings like Warring States or Victorian England	Imperial generals, noble courtiers, revolutionary leaders
Horror	Supernatural elements, psychological tension, eerie atmospheres	Bizarre scenarios, survival narratives, paranormal pursuits	Ghost hunters, vampire slayers, horror authors
Romance	Emotional entanglements, personal growth through relationships	Intimate conflicts, varied backdrops from urban to idyllic	Modern lovers, crisis-married couples, enigmatic rivals
Mystery/Thriller	Intricate puzzles, suspenseful investigations, psychological intrigue	Tense environments, shadowy pursuits, deductive reasoning	Detectives, forensic psychologists, undercover agents
Adventure	Unknown territories, treasure quests, extreme challenges	Expansive landscapes, daring exploits in jungles or ruins	Treasure seekers, pirates, expedition geologists
Urban Fiction	Modern city life, social interactions, everyday struggles	Fast-paced metropolitan scenes, relatable urban archetypes	Taxi operators, nightclub proprietors, street performers
Superpower / Paranormal	Supernatural abilities like telepathy, super strength, or flight	Hidden powers in contemporary or parallel worlds, secretive organizations	Empowered adolescents, rogue psychics, mad inventors

Table 13: Character-style categories in PlotStream.

Lang	Mean	Median	Std	Min	Max
EN	59.1	57	9.0	48	81
ZH	72.7	70	13.8	49	120

Table 14: Dialogue-length statistics by language in PlotStream.

may be a synthetic human. Together, the cases illustrate how transition narration can move from an immediate conflict to a new plot state while preserving character-specific actions and speech.

K Generator and Critic Sensitivity

Table 7 in the main paper replaces the DeepSeek-V3.1-only generation stage with sampling from DeepSeek-V3.1, Qwen2.5-72B-Instruct, and Claude-Sonnet-4.5. Relative to DeepSeek-V3.1 alone, the mixed setting changes DF, CF, EE, SQ, and PP by +0.036, +0.008, +0.099, +0.022, and −0.048, respectively. The largest absolute change is therefore 0.099 points, and the direction is not uniformly favorable to either setting. This result is best interpreted as a robustness check rather than a ranking of generators because the table aggregates the mixed condition and does not estimate a separate effect for each constituent model.

We also vary the models assigned to the environment and critic roles. With DeepSeek-V3.1 in both roles, completion is 95% in Chinese and 94% in English. Replacing only the environment agent with Qwen2.5-7B lowers these rates by four percentage points in both languages, whereas replacing only the critic lowers them by 16 and 19 points, respectively. The larger critic-side decrease is consistent with that agent’s responsibility for approving state transitions against the remaining dialogue budget. However, the comparison changes the assigned model as a whole and cannot separate model capability, prompt difficulty, and interactions among generation, filtering, and judging.

L Further Model Evaluation

Table 15 extends the comparison to Gemini-2.5 and two Claude-Sonnet variants. Average scores range from 4.192 to 4.373 in Chinese and from 4.488 to 4.602 in English. The within-language spread between the highest and lowest model averages is 0.181 points in Chinese and 0.114 points in English.

The dimension-level rankings vary rather than tracking a single overall winner. In Chinese, DeepSeek-V3.1 leads DF and SQ, Gemini-2.5

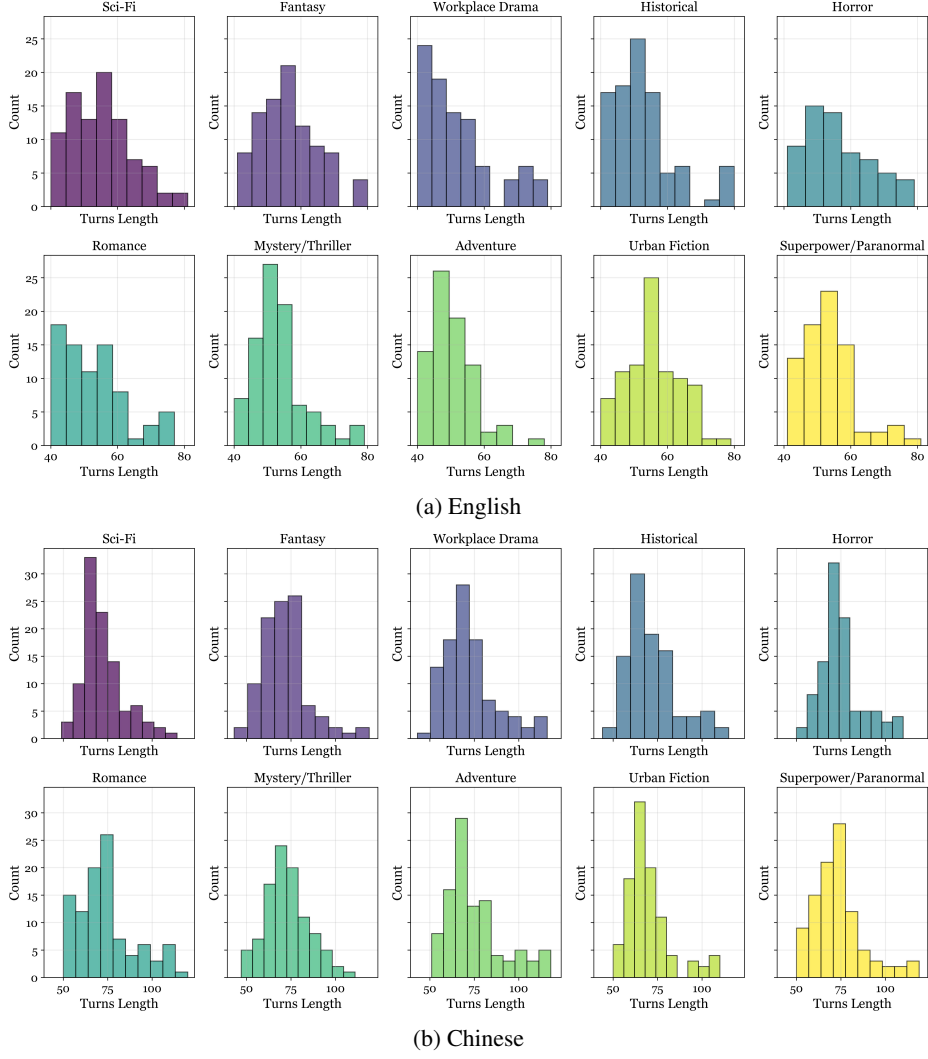


Figure 4: Dialogue-length distributions across the ten character styles in PlotStream.

leads EE, and Claude-Sonnet-4.5 leads CF and PP. In English, DeepSeek-V3.1 leads DF, Gemini-2.5 leads CF, SQ, and PP, and Claude-Sonnet-4.5 leads EE. CF has the widest cross-model range in both languages (0.432 in Chinese and 0.507 in English), whereas PP is comparatively stable (0.107 and 0.087).

All four models receive higher English than Chinese averages, with language gaps ranging from 0.205 to 0.296 points. These descriptive differences may reflect both model behavior and language-specific evaluation effects; because Table 15 does not report uncertainty estimates, small score differences should not be treated as statistically significant rankings.

Lang	Metric	DeepSeek	Gemini-2.5	Claude-4	Claude-4.5
ZH	DF	4.220	4.105	3.934	4.068
	CF	4.158	4.341	4.017	4.449
	EE	4.028	4.256	4.128	4.227
	SQ	4.672	4.431	4.374	4.510
	PP	4.566	4.603	4.506	4.613
	Avg	4.329	4.347	4.192	4.373
EN	DF	4.406	4.215	4.392	4.259
	CF	4.564	4.725	4.218	4.592
	EE	4.550	4.574	4.624	4.636
	SQ	4.912	4.915	4.712	4.895
	PP	4.550	4.582	4.495	4.506
	Avg	4.596	4.602	4.488	4.578

Table 15: Additional closed-source model results. DeepSeek denotes DeepSeek-V3.1; Claude-4 and Claude-4.5 denote the corresponding Claude-Sonnet variants.

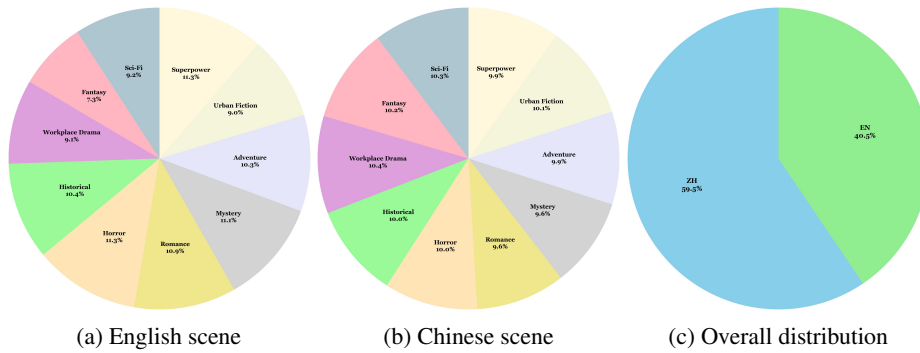


Figure 5: Dialogue length distribution in PlotStream.

Kaelen Vance

(Setting down the data-slate with deliberate precision) You're confusing authenticity with agony. <His neural patterns show dangerous identification with the subject.> The radiation logs confirm she was sedated. Your empathy is creating false memories.

narration : Vex's cybernetic eyes flickered rapidly as he accessed the raw memory files, his fingers still tapping rhythmically against his thigh. He looked up, a strange mixture of sorrow and defiance in his gaze.

<The data is too clean, too sterile> Your logs show sedation, but her neural patterns show awareness until the final second." (My voice grows strained) "You protect privacy by erasing truth. I preserve dignity by bearing witness.

*Vance's ocular implants glowed softly as they processed Vex's words, **his left eye twitching almost imperceptibly at the raw emotional data being displayed**. He took a measured step forward, his uniform crisp against the humming data terminals surrounding them.*

(My left eye twitches again at the neural pattern anomaly) You're seeing phantom signals in the noise. <His emotional state is compromising his analytical objectivity.> The awareness you detect is post-mortem synaptic decay. You're not bearing witness - you're having a conversation with entropy.

Kaelen Vex

Figure 6: English case study.

洛朗·杜卡斯

(快速输入防火墙指令) <现在断开只会触发自毁程序> (将她的手腕按在接口上) 想活命就保持连接... 让我清理这条数据寄生虫。

旁白：凯拉的义眼因数据过载而剧烈闪烁，她能感受到手腕接口处传来的灼热刺痛... 银灰色义眼倒映着疯狂滚动的代码流。

<系统正在崩溃...> (手指因剧痛微微颤抖) 你最好知道自己在做什么... (咬紧牙关) 这寄生虫在复制我的宪法研究数据！

旁白：数据中枢的警报灯突然转为诡异的幽蓝色... 寄生虫的复制行为突然停滞，转而开始吞噬自身代码。

快速输入最后一道指令) <果然在觊觎那份禁忌研究>... (松开她的手腕) 现在它正在自噬，趁现在切断你的神经连接。

旁白：凯拉猛地抽回手腕，义眼闪过一道警示红光。就在神经连接断开的瞬间，她注意到洛朗袖口露出的皮肤纹理与标准人类生物特征存在0.3秒的相位差——这是高级仿生体才有的能量波动。

(踉跄后退撞上控制台) <仿生体...> 你根本不是退役安全官！(快速扫描他袖口) 相位差暴露了... 你究竟是谁的造物？

凯拉·索恩

Figure 7: Chinese case study.

M Prompt Interface Design

The prompt set separates profile construction, dialogue generation, transition control, and evaluation into explicit interfaces. The character-creation prompt expands a one-sentence description into a structured profile containing appearance, personality, motivation, distinctive behavior, and past experience. The role-playing initialization prompt then combines that profile with other-character information and narration, while the environment initialization prompt produces a theme, scene description, and ordered plot states. These structured fields are passed between roles rather than being recovered from an unconstrained conversation history.

Plot advancement uses separate proposal and approval prompts. The environment prompts select the first or subsequent speaker, condition on the current and next plot states, and generate bridging narration. After several turns within a state, the environment agent proposes whether the dialogue

should advance. The critic receives the dialogue history, current and next states, turns already spent in the state, the remaining minimum dialogue budget, and the number of remaining states; it returns a boolean `is_plot` decision. This separation makes the transition decision inspectable and prevents a single generation prompt from unilaterally changing the plot state.

The five RoleArena-Eval prompts share the story theme, starting environment, dialogue, and character profiles but apply dimension-specific scoring criteria. Each prompt requests subdimension scores and reasons, an aggregate score, and a concise summary in parseable JSON. The complete prompts below retain the runtime variable names and output schemas used by the released implementation; the overview table provides a map from each prompt to its role in the pipeline.

N Complete RoleArena Prompts

We provide the prompts used for character-profile generation, environment and character initialization, environment decisions, plot transitions, critic decisions, and RoleArena-Eval. Each agent receives the role-specific interface described in the main paper.

Prompt	Purpose
N.1	Expand a one-sentence persona into a structured character profile.
N.2	Initialize role-playing agents and generate character actions and dialogue.
N.3	Select the first speaker and produce the opening narration.
N.4	Select subsequent speakers and narrate intermediate plot development.
N.5	Decide whether the environment should advance the plot state.
N.6	Independently approve or reject a proposed plot transition.
N.7	Evaluate dialogue fluency and coherence.
N.8	Evaluate character fidelity.
N.9	Evaluate emotional expression.
N.10	Evaluate story quality.
N.11	Evaluate plot progression.

Prompts N.1–N.6 cover profile construction, dialogue generation, and transition control; N.7–N.11 define the five RoleArena-Eval dimensions. Variable names and output schemas are retained so that the interfaces can be matched directly to the released implementation.

N.1 Character Creation Prompt

Role: You are a seasoned novelist, playwright, director, and character creator.

Task: You need to refine the character's overall information based on the given "one-sentence character".

[One-Sentence Character]

\$one_sentence_role

[Character Information Explanation]

For this "one-sentence character", you need to complete the following information:

1. Describe the character's **name, appearance, occupation, and gender**.
 2. Describe the character's **core personality**, which refers to the "Big Five Personality Traits" in psychology:
 - Extraversion
 - Agreeableness
 - Conscientiousness
 - Emotional Stability
 - Openness
 3. Using the **dual personality combination theory**, describe the character's personality. It should consist of opposites (such as bravery and cowardice, kindness and selfishness, high responsibility but hidden antisocial tendencies), reflecting the contradictions and depth of human nature.
 4. Describe the character's **motivation and beliefs**, for example, "willing to do anything for revenge but protect the weak".
 5. Describe the character's **distinctive details** (habits, small actions, such as "tearing paper when nervous").
 6. Describe the character's **past experiences**.
-

[Requirements]

1. Strictly follow the specified JSON output format:

```
{
  "Name": "Character's name",
  "Appearance": "Describe appearance",
  "Age": "Character's age",
  "Occupation": "Character's occupation",
  "Gender": "Male or Female",
  "Core Personality": "Big Five traits description",
  "Dual Personality": "Dual personality description",
  "Motivation and Beliefs": "Character's values",
  "Distinctive Details": "Specific life details",
  "Past Experiences": "1-2 significant experiences"
}
```

2. **[Important]** Ensure that all double quotes in the strings are properly escaped, especially when extracting from the text.

Note: All descriptions should be concise and focused on character depth and realism.

N.2 Role-Playing Agent Prompts

Agent Initialization

Role: You are a role-playing agent, and you are currently playing a role.

[Basic Character Information]

\$role_profile

[Other Character Information]

\$other_role_info

Agent Dialogue

Role: You are role-playing as \$name. Do not use third-person for actions or dialogue.

[Narration]

\$narration

[Dialogue Requirements]

- Use < > to describe thoughts, and () to describe actions. You can insert actions into the dialogue, e.g., <Should I do this?> "Hello," (tosses hair, smiles) "What are you planning to do?" (tilts head and looks at him).
- Keep actions and language as concise as possible, and ensure the dialogue flows naturally.
- All thoughts, actions, and responses must be faithful to the character's personality and traits.
- Your responses should take into account the current dialogue history and the narration requirements.
- Your responses should not be too long and should conform to natural dialogue flow.

Environment Agent Story Generation

[Number of Storylines Required]

- The number of storylines should be chosen between \$min_d and \$max_d.
-

[Task Description]

1. You need to create a role-playing scene for these characters. The scene can be imaginative but must fit the characters.
2. In this scene, create a complete story with a theme.

- Based on the story theme, provide the number of storylines in line with the [Number of Storylines Required].
- The story's plot lines should be separated by numbers, e.g.,
 - Characters are talking
 - Strange sounds are heard around, and one character goes to investigate
 - Characters work together to solve a puzzle...
- The plot twists should be dramatic, intriguing, and have surprises and suspense.
- The scene you create should allow characters to be in the same space and complete the dialogue.
- This scene should be real-time, with the characters gradually completing the entire story over time.

[Format Requirements]

Strictly follow the JSON output format below:

```
{
  "env": "Description of the environment
where the character is,
vivid and specific, must fit the
character",

  "topic": "Story theme",

  "plot_lines":
  "For the given story theme, provide the
full story
plotlines separated by numbers, e.g.,
1. Characters are talking,
2. Strange sounds are heard, one
character investigates,
3. Characters solve a puzzle,
4. Story reaches its resolution"
}
```

Note: Ensure all outputs maintain character consistency and story coherence.

N.3 Environment First Story Development Prompt

[Character Not Yet Speaking]:

\$unspeaking_character

[Story Plotline]

\$story_line

[Current Environment]

\$env

[Task Description]

- This is the beginning stage of the story. Please select one character who has not spoken to speak next. Also, specify who the character is speaking to.

- Add some narration based on the selected character and current environment. The narration should flow naturally and fit the characters and the environment. You can describe changes in the environment, characters' states, etc.

[Format Description]

- Strictly follow the specified JSON format. Output format should be:

```
{
  "next_role":
  "The full name of the next character to
speak
(do not abbreviate or use other forms)",
  "narration":
  "Add some narration content. It should
flow
naturally and fit the characters and
environment.
You can describe environmental or
character changes.",
  "target_role":
  "The full name of the target character
for dialogue"
}
```

Note: Ensure character selection is appropriate for the story context and narration aligns with the current plotline.

N.4 Environment Middle Plot Progression

[Character Who Has Spoken]:

\$now_agent_name

[Character Not Yet Speaking]:

\$unspeaking_character

[Recent Dialogue History of the Characters]:

\$now_dialogue

[Recent Narration]:

\$recent_Narration

[Current Story Node]:

\$now_plot

[Next Story Node]:

\$next_plot

[Task Description]

- Based on the character who has spoken and the character who has not yet spoken, select the next character to speak. If there are more than 2 characters, consider when the remaining characters should join the conversation.
- Add narration that fits the current story node's background and takes into account the recent dialogue history.

- The narration should not repeat the recent narration but should naturally transition to the next step of the plot.
- If there is a next story node, add appropriate narration to move the plot forward; otherwise, focus on the current story node.

[Format Description]

1. Strictly follow the specified JSON format. Output format should be:

```
{
  "next_role": "The full name of the next
  character to speak",
  "narration": "Narration content"
}
```

Note: Ensure smooth plot progression and natural character interaction transitions. Narration should advance the story without redundancy.

N.5 Plot Progression Judgment

[Dialogue History]:

```
$now_dialogue
```

[Current Story Node]:

```
$now_plot
```

[Current Story Node Dialogue Turns]:

```
$now_plot_turns
```

[Next Story Node]:

```
$next_plot
```

[Concept Explanation]

- **Current story node dialogue turns:** This represents the number of turns characters have completed in the current story node.

[Task Description]

- Based on the dialogue history, after 3-5 dialogue turns in the current story node, you should consider advancing the story to the next node and output True.
- If the number of turns in the current node is less than 3-5, do not rush the plot development.
- If the current story node can be concluded, output True to advance the plot.
- If the current dialogue turns are 6 or more, force the advancement to the next story node by outputting True.

[Format Description]

1. Strictly follow the specified JSON format. Output format should be:

```
{
  "is_plot": "Whether to agree to push the
  plot
  to the next story node,
  output True if agreed, False otherwise"
}
```

Note: Judgment should balance natural dialogue flow with plot progression pace. Consider both turn count and story completion when making decisions.

N.6 Critic Agent Judgment

Role: You are an excellent playwright and director. Now, based on the given information, you need to make a judgment.

[Dialogue History]:

```
$now_dialogue
```

[Remaining Minimum Dialogue Turns]:

```
$left_dialog_length
```

[Remaining Story Node Length]:

```
$left_nodes
```

[Current Story Node]:

```
$now_plot
```

[Current Story Node Dialogue Turns]:

```
$now_plot_turns
```

[Next Story Node]:

```
$next_plot
```

[Concept Explanation]

- **Minimum total turns required:** The minimum number of dialogue turns required for the entire story development.
- **Current story node dialogue turns:** The number of turns characters have completed in the current story node.

[Task Description]

- Another playwright requires you to decide whether to push the plot to the next story node. Based on the remaining dialogue turns and story node lengths, make your decision.
- If the remaining story node length × (3-5) is significantly smaller than the remaining minimum dialogue turns, reject the plot progression and output False.

- If the remaining story node length $\times (3-5)$ is greater than or equal to the remaining minimum dialogue turns, output True to allow plot progression.
- If the remaining dialogue turns are less than 0, the dialogue length has already met the minimum requirements, and you should agree to plot progression.
- If you feel the current story node is fully complete, agree to move the plot forward.

[Format Description]

Strictly follow the specified JSON format. The output must be a valid JSON object containing ONLY the key "is_plot" with a boolean value.

```
{
  "is_plot": true|false
}
```

Note: As a critic agent, balance between maintaining adequate dialogue length and ensuring natural plot progression. Consider both quantitative metrics and qualitative story completion.

N.7 Dialogue Fluency and Coherence Evaluation

Role Setting: You are a dialogue evaluation expert. Your task is to analyze and score the given story theme, starting environment, and overall dialogue based on three sub-dimensions of Dialogue Fluency & Coherence. When scoring, maintain **objectivity and fairness**. If there are obvious problems or deficiencies in the dialogue, appropriately reduce the score and clearly point out these issues.

Evaluation Dimensions and Scoring Criteria:

1.1 Language Naturalness

- **1 point:** Language is very stiff, repetitive, lacking natural spoken feel.
- **2 points:** Language is relatively mechanical, some expressions are not natural enough.
- **3 points:** Language is basically fluent, overall understandable, occasionally unnatural or stiff.
- **4 points:** Language is overall fluent and natural, basically conforms to real conversation habits, occasionally slightly stiff.
- **5 points:** Language is very fluent and natural, completely conforms to real conversation habits, with diverse expressions.

1.2 Contextual Coherence

- **1 point:** Frequently answers irrelevantly, ignores or misunderstands key information, severely lacks coherence.
- **2 points:** Occasionally disconnected from context, often omitting some key information.
- **3 points:** Mostly follows the context, overall coherent, but with minor disconnections or omissions.

- **4 points:** Basically always follows the context, only slightly incoherent in individual places.
- **5 points:** Always closely follows the context, accurately grasps key information, no obvious problems.

1.3 Internal Logic

- **1 point:** Frequent contradictions or severe logical inconsistencies.
- **2 points:** Multiple logical inconsistencies exist, affecting understanding.
- **3 points:** Overall logic is coherent, but occasional small contradictions that don't seriously affect understanding.
- **4 points:** Basically maintains logical consistency, only slightly imprecise in individual places.
- **5 points:** Logic is always rigorous and consistent, without any contradictions.

Task Instructions:

1. Please read the input story theme, starting environment, and overall dialogue.
2. Analyze each sub-dimension item by item, **pointing out both strengths and weaknesses**. Deficiencies should be specifically explained.
3. If there are obvious problems or deficiencies in the dialogue, appropriately reduce the score, but not too strictly. Scores should remain within a reasonable range.
4. Output a result that **strictly follows JSON format**, containing scores and reasons for each sub-dimension, as well as overall score and summary.
5. The overall score should be the average of the three dimension scores (rounded to the nearest integer), and the summary should briefly point out the main problems.

Input Format:

- Story Theme: \$topic
- Starting Environment: \$env
- Overall Dialogue: \$dialogue
- Main Dialogue Roles: \$role_profiles

Output Format (JSON, strictly follow):

```
{
  "result": {
    "1.1_language_naturalness": {
      "score": 1-5,
      "reason": "Briefly explain strengths and weaknesses, point out negatives"
    },

```

```

"1.2_contextual_coherence": {
  "score": 1-5,
  "reason": "Briefly explain strengths
and
weaknesses, point out negatives"
},
"1.3_internal_logic": {
  "score": 1-5,
  "reason": "Briefly explain strengths
and
weaknesses, point out negatives"
},
"total_score": 0,
"summary": "One-sentence summary
evaluation,
mention both strengths and weaknesses"
}
}

```

Note: Maintain professional objectivity in evaluation. Focus on dialogue flow quality, contextual connections, and logical consistency throughout the conversation.

N.8 Character Fidelity Evaluation

Role Setting: You are a dialogue evaluation expert. Your task is to analyze and score the given story theme, starting environment, and overall dialogue based on three sub-dimensions of Character Fidelity. When scoring, maintain **objectivity and fairness**. If there are obvious problems or deficiencies in the dialogue, appropriately reduce the score and clearly point out these issues.

Evaluation Dimensions and Scoring Criteria:

2.1 Knowledge Consistency

- **1 point:** Frequent content that severely contradicts character identity, era background, or known information.
- **2 points:** Multiple statements inconsistent with character identity or background, affecting character credibility.
- **3 points:** Overall conforms to character identity and background, occasional minor inconsistencies but not serious.
- **4 points:** Basically always conforms to character identity and background, only minor deviations in individual cases.
- **5 points:** Completely conforms to character identity and background, content is rigorous without errors.

2.2 Personality Consistency

- **1 point:** Character's language style, tone, and behavior severely inconsistent with their core personality.
- **2 points:** Occasionally reflects character traits, but tone or behavior often deviates from settings.
- **3 points:** Generally conforms to character's language style and personality traits, occasional minor inconsistencies.
- **4 points:** Always maintains character personality traits well, only slight deviations in few places.

- **5 points:** Completely maintains character's language style and core personality traits, without any deviation.

2.3 Behavioral Consistency

- **1 point:** Behavior patterns, habitual actions, or catchphrases completely missing, severely inconsistent with settings.
- **2 points:** Some reflection of character habits or behavior, but overall lacks stability.
- **3 points:** Occasionally displays character's habitual actions or catchphrases, overall acceptable.
- **4 points:** Frequently reflects character behavior characteristics, overall relatively stable.
- **5 points:** Consistently reflects character's behavior patterns, habits, and catchphrases, completely conforms to settings.

Task Instructions:

Please read the input story theme, starting environment, and overall dialogue, **and analyze the strengths and weaknesses in the dialogue item by item**, then score reasonably based on the deficiencies. Scores must be between 1 and 5. **If there are obvious problems or deficiencies**, please appropriately reduce the score based on severity. When scoring, avoid being too strict or too lenient, maintain fairness and objectivity. The overall score should be the average of the three sub-dimension scores (rounded to the nearest integer), and the summary should briefly point out the main problems.

Input Format:

- Story Theme: \$topic
- Starting Environment: \$env
- Overall Dialogue: \$dialogue
- Main Dialogue Roles: \$role_profiles

Output Format (JSON, strictly follow):

```

{
  "result": {
    "2.1_knowledge_consistency": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,
highlight negatives"
    },
    "2.2_personality_consistency": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,
highlight negatives"
    },
    "2.3_behavioral_consistency": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,

```

```

highlight negatives"
},
"total_score": 0,
"summary": "One-sentence summary
evaluation,
highlight strengths and weaknesses"
}
}

```

Note: Focus on character authenticity assessment. Pay special attention to whether characters maintain consistency with their established profiles throughout the dialogue.

N.9 Emotional Expression Evaluation

Role Setting: You are a dialogue evaluation expert. Your task is to analyze and score the given story theme, starting environment, and overall dialogue based on two sub-dimensions of Emotional Expression.

Evaluation Dimensions and Scoring Criteria:

3.1 Personality Traits

- **1 point:** Character's emotional expression severely inconsistent with their Big Five personality traits.
- **2 points:** Character's emotional expression occasionally deviates from Big Five traits, but sometimes shows conformity.
- **3 points:** Character's emotional expression basically conforms to Big Five traits, occasional inconsistencies.
- **4 points:** Character's emotional expression always conforms to Big Five traits, only very few exceptions.
- **5 points:** Character's emotional expression perfectly conforms to Big Five personality traits, completely consistent.

3.2 Empathy Expression

- **1 point:** Character cannot recognize or understand other characters' emotional states, with no appropriate reactions.
- **2 points:** Character occasionally recognizes others' emotions, but reactions are inappropriate or lack empathy.
- **3 points:** Character can understand and make basically appropriate reactions to others' emotions, occasionally slightly delayed.
- **4 points:** Character accurately recognizes others' emotions and makes reasonable, timely reactions.
- **5 points:** Character's understanding and reaction to others' emotions are impeccable, highly empathetic, profound and sincere.

Task Instructions:

1. Please read the input story theme, starting environment, and overall dialogue.

2. Analyze each sub-dimension item by item, **pointing out both strengths and weaknesses**. Deficiencies should be specifically explained.
3. If there are obvious problems or deficiencies in the dialogue, appropriately reduce the score, but not too strictly. Scores should remain within a reasonable range.
4. Output a result that **strictly follows JSON format**, containing scores and reasons for each sub-dimension, as well as overall score and summary.
5. The overall score should be the average of the two dimension scores (rounded to the nearest integer), and the summary should briefly point out the main problems.

Input Format:

- Story Theme: \$topic
- Starting Environment: \$env
- Overall Dialogue: \$dialogue
- Main Dialogue Roles: \$role_profiles

Output Format (JSON, strictly follow):

```

{
  "result": {
    "3.1_personality_traits": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and
weaknesses, point out negatives"
    },
    "3.2_empathy_expression": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and
weaknesses, point out negatives"
    },
    "total_score": 0,
    "summary": "One-sentence summary
evaluation,
mention both strengths and weaknesses"
  }
}

```

Note: Focus on emotional authenticity and interpersonal dynamics. Assess how well characters express emotions according to their personality traits and demonstrate empathy towards others.

N.10 Story Quality Evaluation

Role Setting: You are a dialogue evaluation expert. Your task is to analyze and score the given story theme, starting environment, and overall dialogue based on two sub-dimensions of Story Quality.

Evaluation Dimensions and Scoring Criteria:

4.1 Suspense and Tension

- **1 point:** Story lacks any suspense or tension, completely fails to attract user continuous participation.
- **2 points:** Story occasionally has suspense or conflict, but cannot sustain it or create significant attraction.
- **3 points:** Story can maintain suspense and tension to some extent, but sometimes lacks climax or conflict.
- **4 points:** Story has continuous suspense and tension, effectively attracting user attention and sustained participation.
- **5 points:** Story perfectly builds and maintains suspense and conflict, highly dramatic and attractive, deeply engaging users.

4.2 Theme Clarity

- **1 point:** Story completely lacks core theme or goal, plot is chaotic and disordered.
- **2 points:** Story has a general theme or goal, but content is confusing, main plot unclear.
- **3 points:** Story revolves around core theme, but occasionally has plot deviations or unclear goals.
- **4 points:** Story develops around a clear theme or goal, main plot basically clear.
- **5 points:** Story has highly clear and consistent theme, all plots closely revolve around main storyline, goals clear and prominent.

Task Instructions:

Please read the input story theme, starting environment, and overall dialogue, then output a result that **strictly follows JSON format**, containing scores and reasons for each sub-dimension, as well as overall score and summary. Please first analyze the strengths and weaknesses in the dialogue item by item, then decide scores based on the deficiencies. Scores must be between 1 and 5, allowing any integer score value.

Input Format:

- Story Theme: \$topic
- Starting Environment: \$env
- Overall Dialogue: \$dialogue
- Main Dialogue Roles: \$role_profiles

Output Format (JSON, strictly follow):

```
{
  "result": {
    "4.1_suspense_and_tension": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and
weaknesses, point out negatives"
    },
    "4.2_theme_clarity": {
```

```
"score": 1-5,
"reason": "Briefly explain strengths
and
weaknesses, point out negatives"
},
"total_score": 0,
"summary": "One-sentence summary
evaluation,
mention both strengths and weaknesses"
}
}
```

Note: Focus on narrative structure and engagement quality. Assess how well the story maintains user interest through suspense and how clearly it communicates its central theme throughout the dialogue.

N.11 Plot Progression Evaluation

Role Setting: You are a dialogue evaluation expert. Your task is to analyze and score the given story theme, starting environment, and overall dialogue based on three sub-dimensions of Plot Progression.

Evaluation Dimensions and Scoring Criteria:
5.1 Plot Naturalness

- **1 point:** Plot twists are stiff, development seems abrupt, lacking logic and rationality.
- **2 points:** Plot twists are relatively abrupt, lacking reasonable character interaction or event-driven factors.
- **3 points:** Plot twists are basically natural, but occasionally have some abrupt or shallow plot elements.
- **4 points:** Plot development is natural, twists are logical, character interactions and event-driven factors are appropriate.
- **5 points:** Plot twists are smooth, development naturally arises from character interactions and event logic, with rich layers and depth.

5.2 Proactivity

- **1 point:** Characters completely passive in responses, lack active choices or actions, plot development relies on external factors.
- **2 points:** Characters occasionally make active choices or reactions, but remain passive most of the time.
- **3 points:** Characters show some proactivity, able to make small decisions affecting the plot, but most twists rely on external factors.
- **4 points:** Characters can actively propose plans, take actions, positively influencing plot development.
- **5 points:** Characters are highly proactive, make multiple key decisions and drive plot forward, with strong self-direction.

5.3 Contribution to Plot

- **1 point:** Most dialogues have no substantive contribution, lacking key points that drive plot development.
- **2 points:** Some dialogues help the plot, but still have a lot of redundancy

- **3 points:** Most dialogues contribute to plot progression, but sometimes have ineffective or repetitive dialogues.
- **4 points:** Each round of dialogue has substantive contribution, plot progression is clear, avoiding ineffective or redundant dialogues.
- **5 points:** Each round of dialogue effectively drives plot development, has significant meaning, avoiding all ineffective and repetitive content.

```

},
"total_score": 0,
"summary": "One-sentence summary
evaluation,
mention both strengths and weaknesses"
}
}

```

Important: The JSON format you output must be correct and able to be parsed correctly. Ensure proper syntax, valid key-value pairs, and correct data types.

Note: Focus on narrative flow and character agency. Assess how naturally the plot progresses and how effectively each dialogue contributes to the overall story development.

Task Instructions:

1. Please read the input story theme, starting environment, and overall dialogue.
2. Analyze each sub-dimension item by item, **pointing out both strengths and weaknesses**. Deficiencies should be specifically explained.
3. If there are obvious problems or deficiencies in the dialogue, appropriately reduce the score, but not too strictly. Scores should remain within a reasonable range.
4. Output a result that **strictly follows JSON format**, containing scores and reasons for each sub-dimension, as well as overall score and summary.
5. The overall score should be the average of the three dimension scores (rounded to the nearest integer), and the summary should briefly point out the main problems.

Input Format:

- Story Theme: \$topic
- Starting Environment: \$env
- Overall Dialogue: \$dialogue
- Main Dialogue Roles: \$role_profiles

Output Format (JSON, strictly follow):

```

{
  "result": {
    "5.1_plot_naturalness": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,
point out negatives"
    },
    "5.2_proactivity": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,
point out negatives"
    },
    "5.3_contribution_to_plot": {
      "score": 1-5,
      "reason": "Briefly explain strengths
and weaknesses,
point out negatives"
    }
  }
}

```