RoleArena: A Multi-Agent Role-Playing Environment for Long Multi-Turn Dialogues with Autonomous Plot Progression

Anonymous ACL submission

Abstract

Improving the plot progression ability of role-playing language agents (RPLAs) plays a critical role in enhancing user experience and dialogue immersion. However, research on plot progression in role-playing language agents is still limited. Existing datasets often contain too few dialogue turns and lack large-scale sentences with clear deterministic plot turns, which restricts further model development and evaluation. To bridge this gap, we propose *RoleArena*, an innovative multi-agent role-playing environment. By introducing a critic agent and combining discrete plot lines with environment-agent-based dynamic judgments, the system generates long multi-turn dialogues with key markers for plot progression. Based on *RoleArena*, we construct the PlotStream dataset, which contains 39.3k characters and 170k dialogues, with annotations for key plot progression sentences. In addition, based on PlotStream, we further develop PlotRole-7B and PlotRole-72B, RPLAs built upon the Qwen2.5 model and specialized in plot progression capabilities. Extensive experiments show that *RoleArena* and PlotRole improve dialogue generation with plot progression and strengthen the ability of plot progression. The code and data is available at https://anonymous.4open.science/r/RoleArena-CB22.

1 Introduction

The rapid development of large language models (LLMs) accelerates the digital transformation of human society (Piatti et al., 2024; Zhou et al., 2025). As research on LLMs deepens, there is increasing interest in the human-like intelligent traits they exhibit (Thoppilan et al., 2022). As a result, role-playing language agents (RPLAs) have gained attention and become a key focus of current research (Chen et al., 2024; Shao et al., 2023). RPLAs introduce innovative interaction experiences between humans and artificial intelligence (Santurkar et al., 2023; Zhao et al., 2024), leading to numerous application scenarios, such as role-playing game dialogues and digital human interactions (Wang et al., 2025b; Li et al., 2023; Ashby et al., 2023; Yao et al., 2024). In such dialogue environments, users can engage in immersive dialogue scenarios (Tao et al., 2024a). Therefore, RPLAs face the challenge of providing users with an immersive chat experience, necessitating the enhancement of their plot progression capabilities (Wakaki et al., 2024).

However, current research on improving the plot progression capabilities of RPLAs is insufficient. Specifically, there are certain limitations: 1. Insufficient dialogue depth: Existing RPLAs typically have short multi-turn dialogue sequences, often limited to shallow interactions within a dozen turns, which fail to reflect RPLAs' performance in complex multi-turn dialogue scenarios. 2. Lack of plot progression: Datasets lack long dialogues with key plot progression, making it difficult for RPLAs to learn how to proactively and logically advance the plot, leading to monotonous and repetitive dialogues. Although some work attempts to provide datasets with plot progression statements (Tu et al., 2024; Lu et al., 2025; Yuan et al., 2024), these require significant human effort to identify key plot points, resulting in high costs and making large-scale generation impractical (Zhang et al., 2025).

In this paper, we introduce RoleArena, a multi-agent role-playing environment capable of generating long, multi-turn dialogues with autonomously developed plotlines. RoleArena allows multiple RPLAs to interact in fictional dialogue scenarios. Specifically, it contains three types of agents: **Character agent**, which has independent memory and engage in dialogue with other role agents according to the environment agent's guidance. **Environment agent**. To ensure a coherent theme and drive the plot forward, environment agents first gener-

ate the environment, dialogue themes, and a discretized plotline. After each interaction by the role agents, the environment agent determines whether to advance the plot. The environment agent focuses on the current plot point. If it chooses to advance the plot, it generates a naturally coherent plot progression based on the preceding and following events. This approach significantly improves the robustness of plot-driving statements compared to prior work. **Critic agent.** To achieve controlled multi-turn dialogues, we introduce a critic model to perform a secondary evaluation of the environment agent's decisions. To improve system efficiency, the critic agent only initiates the secondary evaluation when the environment agent chooses to advance the plot. The critic agent adjusts the environment agent's decision based on the current and required total number of dialogue turns. The plot only advances if the critic agent agrees with the environment agent's decision.

The main difference between RoleArena and existing work is that: 1) RoleArena uses discretized plotlines to drive the environment agent to dynamically determine plot advancement points during multi-turn dialogues and generate natural plot progressions, offering a robust method for plot-driving data generation. 2) RoleArena introduces critic agents to make the overall dialogue turn count controllable, while ensuring sufficient development of each plot point.

Based on RoleArena, we construct a large-scale multi-turn dialogue dataset, PlotStream, which includes plot-driven annotations. To investigative the enhancement of plot-driving capabilities in RPLAs, we develop PlotRole-7B and PlotRole-72B based on PlotStream. Additionally, we propose the RoleArena-Eval evaluation system, based on five dimensions, to comprehensively assess RPLAs' abilities in long-turn dynamic dialogue environments.

Briefly, our main contributions are three-fold:

- We propose **RoleArena**, a multi-agent role-playing environment for generating controllable-turn, plot-driven long multi-turn dialogues. By combining discretized plotlines, environment-agent-based dynamic plot control, and critic-agent-assisted dialogue regulation, RoleArena enables coherent and natural narrative progression in complex dialogue scenarios.
- Based on RoleArena, we construct **PlotStream**, a large-scale long-turn role-playing dialogue dataset containing 39.3k characters and 170k dialogues with explicit plot progression annotations. Furthermore, we develop **PlotRole-7B** and **PlotRole-72B**, which significantly enhance the plot progression capability of role-playing language agents.
- We propose **RoleArena-Eval**, a comprehensive evaluation framework for long-turn dynamic role-playing dialogues. Beyond conventional role-playing evaluation dimensions, RoleArena-Eval introduces a dedicated quantitative protocol for assessing plot progression, enabling systematic evaluation of narrative development and proactive story advancement in RPLAs.

2 Related Work

Multi-agent role-playing language agent dialogues. Several agents are assigned to play different roles and are used to simulate plotlines and generate dialogues. CharacterBox (Wang et al., 2024a) proposed a simulation sandbox that introduces a narrator agent and character agents, where conversations between character agents and their interactive feedback with the narrator drive the dialogue forward. COSER (Wang et al., 2025b) built a multi-agent environment for dialogue simulation, in which an environment model is introduced to provide feedback. IBSEN (Han et al., 2024) attempted to base its design on a fictional story outline, using agents to play the roles of director and actors to perform the plotline. BookWorld (Ran et al., 2025), grounded in drama theory, uses multiple agents across different scenes to complete a story. These methods confirm the effectiveness of multi-agent collaboration in role-playing agent dialogues; however, research on narrative progression is still insufficient. RoleArena introduced a critic agent, achieving more stable and controllable dialogue turns while emphasizing narrative progression.

Construction of role-playing datasets. Current methods for building role-playing datasets mainly follow two approaches: obtaining character configuration metadata, dialogues, and other contextual information from real texts such as novels and film scripts, which allows role-playing language agents to act as the characters contained within them and provide users with an immersive conversational experience (Li et al., 2023; Dai et al.,

2024; Shao et al., 2023; Zhou et al., 2024). Creating synthetic character configuration metadata and generating simulated dialogues based on it greatly expands the sources of data for constructing role-playing datasets, significantly enhancing the role generalization ability of role-playing language agents while avoiding infringement risks associated with some literary and film works (Ge et al., 2024; Lu et al., 2025; Wang et al., 2025a; Tao et al., 2024b). PlotStream is mainly built on synthetic characters.

Role-playing evaluation. Evaluating the performance of role-playing is a challenging task, as a comprehensive evaluation often requires considering various subjective indicators. InCharacter (Wang et al., 2024b) proposes using personality tests to assess the fidelity of RPLA. DITTO (Lu et al., 2024) evaluates role-playing language agents by judging character consistency, knowledge accuracy, and refusal rates. In addition to evaluating static character fidelity, consistency, and personification, CharacterBox (Wang et al., 2024a) extends the evaluation to specific actions in dynamic scenes. COSER (Wang et al., 2025b) proposes a GCA-based evaluation method, where LLMs perform role-playing and multi-round dialogues in a given scene and are evaluated across four dimensions: story quality, personification, character fidelity, and story consistency. CharacterEval (Tu et al., 2024) designed a four-dimensional evaluation system, including dialogue ability, character consistency, role-playing appeal, and personality backtesting. Beyond the comprehensive evaluation of character actions, psychology, and dialogue, RoleArena-Eval also extends the evaluation to plot progression and evolution.

3 RoleArena

3.1 PRELIMINARY

Motivation. The ability of role-playing language agents to drive the plotline greatly affects the user's immersive experience, but research in this area has not made significant progress so far. RolePlot(Zhang et al., 2025) proposes a method based on triggering the plotline transition subspace, which is effective, but the data used for training often requires manual filtering, resulting in high labor costs, and it heavily depends on historical dialogue information of the characters. In addition, the dialogue turns and plotline development in existing datasets are relatively limited, making it difficult to support complex multi-turn dialogues. Research on generating robust long-turn multi-round dialogues will fill the gap in this area of study.

Role-playing multi-round dialogue. In a given environment, if there are M dialogue agents, we define the completion of the dialogue by all agents as the end of one dialogue turn.

Plot progression. The process unfolds gradually with changes in character actions and event developments, with each step laying the foundation for the subsequent plot, maintaining the continuity and appeal of the story, and driving it in new directions.

3.2 Agent Architecture

RoleArena is a framework for multi-agent collaboration that drives natural story progression and maintains long-turn conversations. It mainly includes three functional agents: environment agent $\mathbb{E}$, character agent $\mathbb{R}$, and critic agent $\mathbb{C}$.

3.2.1 Environment Agent

The environment agent, as a core functional agent in RoleArena, is mainly responsible for four subtasks:

1) Topic, environment, and plotline generation. Before the start of the entire conversation D, the environment agent $\mathbb{E}$ first generates the story topic, environment, and corresponding plotline based on the character agents $[R_1, R_2, ..., R_M]$ present in the current scene. The plotline is then discretized into a sequence of nodes $< S_1, S_2, ..., S_k >$. To achieve controllable turns, we introduce a hyperparameter r, which defines the minimum number of conversation turns required for the story. $\mathbb{E}$ will consider r when generating the plotline length k.

2) Guiding character agents. When the conversation reaches the i-th round, after each action r of a character agent $\mathbb{R}$ is completed, $\mathbb{E}$ immediately receives the action information of that character and prioritizes prompting the silent character agent $\mathbb{R}$ to take action. $\mathbb{E}$ will add appropriate narration content n and guide the actions of the character agent. Therefore, the overall sequence of the i-th round of conversation is

$$[n_1, r_1, \ldots, n_l] \text{ or } [r_1, n_1, \ldots, n_l] \quad (1)$$

The reason for the two different situations is that at the start of the first round of conversation, the environment agent $\mathbb{E}$ initiates, and at the end of each round, the environment agent provides immediate feedback on the actions of the character agent $\mathbb{R}$.

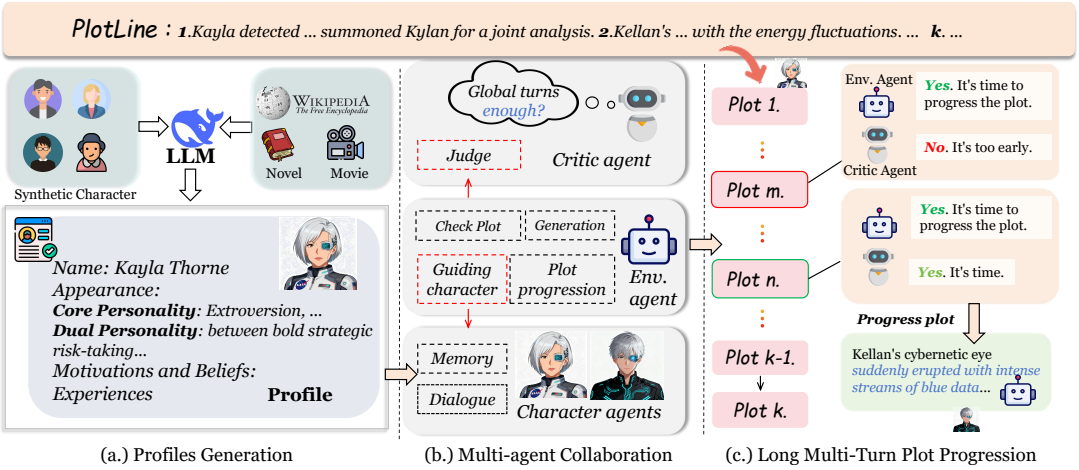

Figure 1: Overview of the RoleArena multi-agent environment. After generating character profiles in (***a.***), we drive the interactions of different agents in (***b.***), where the Env. agent sets up the plot and guides the Character agents to perform multi-turn dialogues in (***c.***) to achieve plot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

3) Check the story's end state and drive progression. In complex story tasks, purely relying on character dialogues to advance the plot often leads to stagnation or deviation from the main plotline. Unlike previous methods, in RoleArena, the environment agent $\mathbb{E}$ dynamically determines the story's end state in real-time. After each action by a character agent R, the environment agent E combines the character's dialogue history $\mathcal{H}$, the current story node's dialogue turn d_k, the current story node S_u, and the next story node S_{u+1} to decide whether to end the current plotline S_u and move the plot to S_{u+1}.

4) Robust plot progression. Once the environment agent $\mathbb{E}$ and the critic agent $\mathbb{C}$ reach a consensus among the characters and agree to advance the plot, $\mathbb{E}$ will be instructed to create a natural plot twist narration n_i^* for the plotline $< S_u, S_{u+1} >$. This plot twist narration n_i^* will also direct the next character agent to generate the corresponding action r_i^*.

In our design, plot twists are actively judged and triggered by the environment agent. When the system detects the need to advance the plot, the environment agent generates a plot twist narration based on the established story nodes and task requirements, guiding the characters into a new story phase. Compared to methods relying on manual filtering, the advantages of this mechanism include: *1)* Robustness: plot twists are based on clear task signals and node logic, avoiding random or inconsistent progression, ensuring the controllability and coherence of story development. *2)* Cost efficiency: Through automated twist generation, the system can significantly reduce manual intervention and data filtering costs, enabling larger-scale and more reusable plot generation processes.

3.2.2 Character Agent

To make the overall conversation controllable and improve its quality, each action of the character agent $\mathbb{R}$ is guided by the environment agent $\mathbb{E}$. At the same time, $\mathbb{R}$ needs to consider its dialogue history with other character agents when taking actions. Furthermore, to more realistically reflect the character's inner thoughts and enrich the emotional expression of the character agent, we require $\mathbb{R}$ to express $< thoughts >$, $(actions)$, and speech while taking actions.

3.2.3 Critic Agent

Relying solely on the environment agent to judge plot progression can lead to the plot advancing too quickly, which may affect the quality and coherence of the conversation. To improve the system's robustness, we introduced the critic agent $\mathbb{C}$ to assist the decision-making process of the environment agent. Specifically, when the environment agent determines that the plot should advance, $\mathbb{C}$ evaluate this decision. At this point, the critic agent not only considers the current plot state but also takes into account the remaining dialogue turns $(r - i)$ and the remaining story elements $(k - u)$ to ensure the plot's progression is reasonable. The system will only update the current plot state S_u to S_{u+1} when both the environment agent $\mathbb{E}$ and the critic agent $\mathbb{C}$ agree that the plot can advance. If

Dataset	Character		Conversation			Dialogue			Plot
	Num.	Profile	#Turns	#Conv.	Multi-Chara.	Speech	Thought	Action	Progre.
Charater-LLM (Shao et al., 2023)	9	✓	13.2	14,300		✓			
ChatHaruhi (Li et al., 2023)	32	✓	>2	54,726	✓	✓			
RoleLLM (Wang et al., 2023)	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., 2024)	85	✓	4.2	14,346		✓			
CharacterBench (Zhou et al., 2025)	3,956	✓	11.3	13,162		✓			
COSER (Wang et al., 2025b)	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 between the PlotStream dataset and other existing datasets. Num., #Turns, #Conv., and Multi-Chara. denote the number of characters, the number of dialogue turns, the total scale of dialogues, and whether multiple characters are supported, respectively.

the critic agent believes that the plot is advancing too quickly and the current plot is not sufficiently developed, it will reject the environment agent's judgment, and the system will maintain the current plot state until the conditions for progression are met.

$$S_u \xrightarrow{\mathbb{E} \wedge \mathbb{C}} S_{u+1} \quad \text{if} \quad \mathbb{P}(\mathbb{E}, \mathbb{C}) = 1 \qquad (2)$$

where $\mathbb{P}(\mathbb{E}, \mathbb{C})$ denotes the probability that both the environment agent $\mathbb{E}$ and the critic agent $\mathbb{C}$ agree on the plot progression. This design enhances the system's robustness and conversation quality by introducing the critic agent. The critic agent $\mathbb{C}$ assists the environment agent's judgment by evaluating the remaining turns and plot complexity, preventing the plot from advancing too quickly, thus improving dialogue stability. The dual evaluation mechanism ensures the reasonable timing of plot twists, avoiding fragmentation or rushed progression. Additionally, $\mathbb{C}$ considers both the remaining dialogue turns and plot development, ensuring that the dialogue progress matches the overall story pace, preventing the plot from advancing too early or too late. We report the results of RoleArena dialogue threshold sensitivity analysis in Table 9.

4 Dataset and agent training

4.1 PlotStream Dataset

Based on RoleArena, we build PlotStream, the first long-turn role-playing multi-turn dialogue dataset with plot progression. This section introduces the details of PlotStream.

4.1.1 dataset Characteristics

Compared to previous work, the PlotStream dataset has several features: ***1)*** **Long-turn dialogue rounds**. The average number of dialogue rounds in PlotStream exceeds 50, covering more complex and continuous multi-turn interactive scenarios, allowing for better simulation of real-world narrative and character interactions. ***2)*** **Clear annotation of key plot-driving statements**. Statements that drive plot twists and progress in the dialogue are clearly marked, making PlotStream an effective resource for improving the plot-driving ability of role-playing agents. ***3)*** **Rich language and character information**. PlotStream supports both Chinese and English dialogue data. To provide more diverse data, PlotStream includes 10 styles of synthetic characters and also introduces novel characters based on human-created literature.

4.1.2 Role Profile Construction

We use Deepseek V3.1[1] to generate meta-configuration information for characters based on the large-scale "one-sentence synthetic characters" provided in PersonHub[2] (Ge et al., 2024). To increase the diversity of the data, we guide the LLM to generate 10 types of characters with highly distinguishable styles. Additionally, we extract 77 novel characters based on human-created literature and their corresponding information from CharacterEval[3] and COSER[4].

[1] https://huggingface.co/deepseek-ai/DeepSeek-V3.1
[2] https://github.com/tencent-ailab/persona-hub
[3] https://github.com/morecry/CharacterEval
[4] https://github.com/Neph0s/CoSER

4.1.3 RoleArena Dialogue Simulation

Using the RoleArena multi-agent role-playing dialogue environment we proposed, based on the profiles of synthetic and novel characters, we use DeepSeek V3.1 to simulate long-turn role-playing multi-turn dialogues with plot-driving statement markers. Moreover, We report the diversity check results for the generated storylines in Appendix D.

During this process, we control the quality of PlotStream with the help of manual inspection. We remove samples with generation failures and samples with excessively long dialogues, and we further filter the data based on predefined thresholds. Detailed procedure is described in Appendix H.

4.2 PlotRole Training

We use the constructed PlotStream dataset to train the LLM to improve its plot progression ability in role-playing. To improve the performance of role-playing agents in both narration generation and character response generation, we define two optimization goals for the role-playing agent's parameter training: 1) Narration target n^*, with its context during training as $[i_n, n_1, r_1, n_2, r_2, ..., r_i]$, where i_n is the role-playing instruction for the task, including detailed information about the characters in the scene. n^* is considered as the optimized plot-driving output. 2) Character response target r^*, with its context during training as $[i_r, ..., n_i, r_i, n^*]$, where i_r is the role-playing instruction that includes the agent's role and other characters' information. r^* is considered as the optimized character plot-driving response. The character r^* response target represents the specific reactions of characters during critical plot advancements, including thoughts, actions, and speech.

$$\theta^* = \arg\min_{\theta} \underbrace{\mathbb{E}_{C_{n^*}} \mathcal{L}_{\text{CE}}\big(f(C_{n^*};\theta), n^*\big)}_{\text{Narration Guidance}} + \underbrace{\mathbb{E}_{C_{r^*}} \mathcal{L}_{\text{CE}}\big(f(C_{r^*};\theta), r^*\big)}_{\text{Role Response}} \quad (3)$$

where $C_{n^*} = [n, r, \ldots, n, r]$ denotes the dialogue context before narration, and $C_{r^*} = [n, r, \ldots, n, r, n^*]$ denotes the context including the narration.

5 RoleArena-Eval

To comprehensively evaluate RPLAs in long-turn dynamic dialogue environments, we propose RoleArena-Eval, which evaluates role-playing performance from five dimensions:

1) **Dialogue Fluency (DF)**: evaluates dialogue naturalness, coherence, and conversational consistency across turns.

2) **Character Fidelity (CF)**: evaluates whether character behaviors, language styles, and decisions align with their background, personality, and abilities.

3) **Emotional Expression (EE)**: evaluates emotional authenticity, personality consistency, and empathetic interaction with other characters.

4) **Story Quality (SQ)**: evaluates narrative quality, including thematic coherence, suspense, and user engagement.

5) **Plot Progression (PP)**: evaluates whether plot development is natural, proactive, and meaningfully advanced through character interactions and actions.

We adopt a Likert-scale (Likert, 1932) evaluation protocol with scores ranging from 1 to 5. Following critique-based evaluation, GPT-4o first generates dimension-specific critique feedback, which is incorporated into the final scoring process. Compared with existing role-playing benchmarks, RoleArena-Eval explicitly introduces quantitative evaluation of plot progression in long-turn narrative dialogues.

6 EXPERIMENTS

6.1 Experimental Settings

Test Set and Evaluation Setup. We split PlotStream into training and test sets with a 9:1 ratio, and train PlotRole-7B and PlotRole-72B based on Equation 3 using Qwen2.5-7B and Qwen2.5-72B as backbones. For evaluation, each LLM generates 20 Chinese and 20 English dialogue sets from the test set over 10 runs using identical character profiles, plotlines, themes, and environments. Final results are averaged across all runs.

Since varying dialogue lengths can introduce evaluation bias (Wang et al., 2025b), we fix the evaluated dialogue length to K turns. In general, we randomly sample K turns from dialogue D for evaluation, while story quality evaluation uses dialogue turns from 0 to K to preserve narrative completeness.

Baseline models. We select two types of LLMs for testing: open-source and closed-source models. **1) Open-source models**: Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Baichuan2-

Model	English Scene					Average
	DF	CF	EE	SQ	PP	
Close-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	3.23±0.46	3.600±0.30
ChatGLM3-6b	2.350±0.19	2.570±0.43	2.412±0.32	3.01±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 2: Statistics of multi-turn dialogue scores generated by different models in RoleArena. **Bold** and underline indicate the best and second-best performance of LLMs across different evaluation dimensions, respectively.

13B-Chat, Qwen2.5-72B-Instruct, Hunyuan-7B-Instruct, ChatGLM3-6B, Llama-3.1-8B, DeepSeek-V3.1. **2) Closed-source models**: 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, Doubao-1.5-lite.

Baseline methods. We select several methods closely related to RoleArena as comparative baselines, all of which perform multi-turn role-playing dialogues in multi-agent environments: IBSEN (Han et al., 2024) and CharacterBox (Wang et al., 2024a). For a fair comparison, all methods used the character configurations and story background information from the test set, and we reported the average results over multiple generations.

LLM as a judge. We use GPT-4o as the large model evaluation expert to score different dimensions based on the RoleArena-Eval evaluation framework, and we reported the average scores across multiple results for each dimension of different LLMs.

6.2 Main Results

PlotRole improves the plot progression capability of RPLAs. Tables 2 and 4 report the average results over 10 runs. PlotRole achieves consistent improvements across multiple evaluation dimensions, particularly in plot progression (PP) for long-turn dialogues. Compared with base models of similar size, PlotRole significantly improves narrative progression ability. Moreover, PlotRole narrows the gap between open-source and strong closed-source models such as DeepSeek-V3.1 and GPT-4.1-mini, demonstrating the effectiveness of plot-driven training data for enhancing RPLAs.

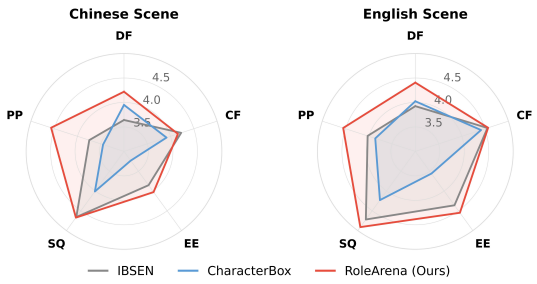

Figure 2: Comparison with existing multi-agent role-playing frameworks under two settings.

RoleArena improves multi-turn dialogue generation. Figure 2 compares RoleArena with existing multi-agent role-playing frameworks. RoleArena consistently achieves better performance across all evaluation dimensions, especially in story quality and plot progression. In addition, dialogues generated by RoleArena exhibit

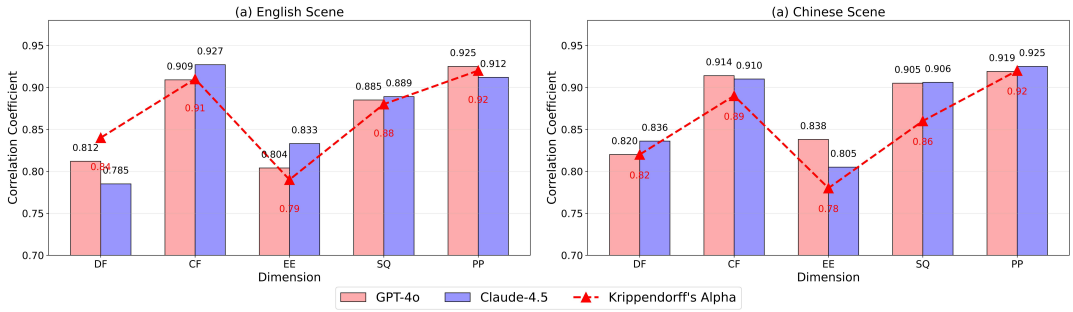

Figure 3: The correlation between human evaluation results and GPT-4o and Claude-4.5 evaluation results in both Chinese and English scenarios, and the consistency of human evaluation

.

Model	Settings	DF	CF	EE	SQ	PP
DeepSeek-V3.1	RoleArena	4.406±0.23	4.564±0.17	4.55±0.42	4.912±0.08	4.55±0.35
	w/o Env Agent	3.52±0.31	4.015±0.29	3.985±0.16	3.176±0.44	3.05±0.27
	w/o Story Line	4.218±0.38	4.442±0.13	4.563±0.49	3.417±0.25	3.695±0.18
	w/o Critic Agent	4.395±0.47	4.55±0.22	4.486±0.41	4.68±0.36	4.262±0.29

Table 3: Results of ablation experiments. We set up three different ablation methods in total.

improved readability and fluency under general text quality metrics.

Human evaluation. We conduct human evaluation with five expert annotators, each having 2–3 years of role-playing experience. Under the setting of $K = 15$, Figure 3 reports the agreement between human judgments and LLM evaluators (GPT-4o and Claude-Sonnet-4.5), as well as inter-annotator consistency. Results show strong alignment between human and LLM evaluations, validating the effectiveness of RoleArena-Eval and the feasibility of using LLMs as judges.

To further analyze judge consistency, we additionally select 50 character sets from the test set and generate dialogues with more than 30 turns. We evaluate alignment between human and LLM judgments under $K = 5$, 10, and 20, with detailed results reported in Table 5.

6.3 Ablation Analysis

We report the ablation results in Table 3. The results show that introducing the environment agent improves the sustainability of dialogue between characters, as well as dialogue fluency and overall story quality. The plotline also plays an important role in driving plot development, while the judgment-generation plotline transition mechanism helps ensure the natural progression of the story. In addition, the critic agent is crucial for controlling the number of dialogue turns. With the critic agent in place, each turn is more meaningfully developed within the story.

6.4 Case Study

We conduct a case study and present the detailed content of the cases in Figures 8 and 7. From these cases, we observe that dialogues generated by LLMs based on RoleArena show high fluency and clear emotional expression, and they achieve smooth and natural transitions between plot lines. In the cases, we highlight the sentences that contribute to plot progression.

7 Conclusion

In this paper, we introduce RoleArena, a multi-agent collaborative environment used to generate long-turn, multi-round role-playing dialogues with plot progression. Based on this environment, we construct PlotStream, a large-scale dataset with annotations for plot progression sentences. We then develop two role-playing agents, PlotRole-7B and PlotRole-72B using PlotStream. We also propose RoleArena-Eval, which evaluates the role-playing ability of RPLAs in long-turn, multi-round dialogues from five dimensions. The evaluation results show that PlotStream, built on RoleArena, effectively improves the overall performance of RPLAs, especially in plot progression capability. For future work, we will involve extending RoleArena to support multimodal interactions while exploring dynamic plotline adaptation mechanisms to handle increasingly complex scenarios.

Limitations

Despite the encouraging results, our work still has several limitations. RoleArena relies on predefined discrete plotlines generated by the environment agent. While this design improves controllability, it may restrict narrative diversity in highly open-ended scenarios. Although PlotStream is large-scale and covers both Chinese and English, its language coverage is still limited, and the performance gap across languages suggests uneven generalization. In addition, our evaluation partly depends on LLM-based judges. Even though they show substantial agreement with human annotators, they may still introduce bias when assessing subjective dimensions such as story quality, emotional expression, and plot progression. Moreover, the dataset is mainly built from synthetic characters, with only a portion derived from existing literary characters, so the resulting models may not fully capture the complexity of real-world character portrayal. Finally, the current framework is restricted to text-only role-playing dialogues and does not yet support multimodal interaction. Future work will explore more adaptive plot planning, broader multilingual coverage, stronger human-centered evaluation, and multimodal extensions.

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.

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.

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*.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

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.

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*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

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.

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.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *arXiv preprint arXiv:2504.14538*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

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.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024b. Chatgpt role-play dataset: Analysis of user motives and model naturalness. *arXiv preprint arXiv:2403.18121*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and 1 others. 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.

Hiromi Wakaki, Yuki Mitsufuji, Yoshinori Maeda, Yukiko Nishimura, Silin Gao, Mengjie Zhao, Keiichi Yamada, and Antoine Bosselut. 2024. Comperdial: Commonsense persona-grounded dialogue dataset and benchmark. *arXiv preprint arXiv:2406.11228*.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2024a. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. *arXiv preprint arXiv:2412.05631*.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. Open-character: 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, and 1 others. 2025b. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 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.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

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, Yuhan 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.

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.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, and 1 others. 2024. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, and 1 others. 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.

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

A Additional Results

We present additional experimental results. Table 4 reports the RoleArena evaluation when different large language models generate RPLA multi-turn dialogues under the Chinese scenario setting. From the results, we observe several findings: 1) The RPLA ability of large language models is sensitive to Chinese and English. Almost all evaluated LLMs show a performance drop, with the maximum decline exceeding 0.5. This indicates the need to increase the diversity of datasets, and we plan to include more languages in future work. 2) PlotRole achieves clear improvements in RPLA

Model	Chinese Scene					Average
	DF	CF	EE	SQ	PP	
Close-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: Statistics of multi-turn dialogue scores generated by different models in RoleArena. **Bold** and underline indicate the best and second-best performance of LLMs across different evaluation dimensions, respectively.

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: Alignment rate between large models as judges and human evaluations. EN and ZH denote English and Chinese settings, respectively.

performance. Although PlotRole-7B and PlotRole-72B do not rank high across all dimensions, they obtain an average gain of about 0.4 compared with the base Qwen2.5-7B and Qwen2.5-72B. This confirms the effectiveness of using plot progression datasets to enhance the plot progression ability of RPLAs.

B Training Settings

We present the detailed training configurations for PlotRole-7B and PlotRole-72B, where PlotRole-7B is trained with 8 × H20 × 96G devices, and PlotRole-72B is trained with an extended setup of 3 × 8 × H20 × 96G devices.

The specific training hyperparameters are as follows: The learning rate is set to 1e-5, the gradient accumulation steps is set to 16, and the maximum training sequence length is 16384.

C Additional Human Evaluation

We select 50 role sets from the test set and use DeepSeek-V3.1 and PlotRole to generate role-playing multi-turn dialogue data of more than 30 rounds based on RoleArena. We examine the alignment between LLM judges and human evaluations under K=5, 10, and 20, and present the results in Table 5. From the results, we find that 1) LLM judges show high consistency with human preferences, confirming the effectiveness of using them for evaluation. 2) As the number of evaluated dialogue rounds increases, preference consistency gradually decreases. This happens because semantic complexity increases with more dialogue rounds, which challenges the understanding ability of LLMs and raises the likelihood of deviation from human preferences.

D Plotline Diversity Analysis

To verify story line diversity across styles, we computed similarities using two top MTEB multilingual Leaderboard embedding models: Qwen3-Embedding-8B and KaLM-Embedding-Gemma3-12B-2511. For efficient comprehensive reporting, we ran 10 samplings of 100 dialogue sets each and

averaged results in Tables 6 and 7. Key findings: PlotStream plot lines show no high similarity, with a maximum average of 0.4422, confirming their diversity.

E User Participation in RoleArena

We used GPT-4o to simulate users in satisfied, dissatisfied, and preference modes across science fiction and horror styles. We ran experiments with untuned Qwen2-72B-Instruct and fine-tuned PlotRole-72B, recording dialogue turns between the simulated user and agent until reaching 5 turns or user termination.

Table 8 reports results from 10 simulated runs over 100 scenes, showing that RoleArena's dataset improves plot progression, with average dialogue turns increasing by up to 62%. We believe this will enhance user immersion in real-world scenarios.

F Sensitivity Analysis of the Plot Progression Threshold

We conducted a sensitivity analysis on the thresholds. In Table 4, we report the average results from sampling 50 groups of dialogues 10 times in the English scenario. The results confirm that RoleArena maintains an accuracy rate above 90% for data generation under different plot progression rule requirements, verifying the stability of RoleArena in generating dialogues. Without adding the forced progression rules, the success rate decreased by up to 21%, which demonstrates the necessity of the forced plot progression rules.

It should be noted that successful dialogue generation refers to the entire dialogue successfully completing the storyline and meeting the minimum required dialogue turns, and Avg Turns represents the average number of dialogue turns per plot point.

G PlotSream Data Statistics

G.1 Role Diversity Setting

To increase the overall diversity of the data, we referenced common typologies from the fields of literature, film, and games, and integrated multiple dimensions such as "core conflict source ", "narrative function", and "aesthetic style", to establish ten distinct types of roles; in Table 10, we list the features, styles, and examples for each type.

G.2 Distribution of Dialogue Quantity Characteristics

In Figure 6, we report the overall language distribution in the PlotSream dataset, as well as the proportion of dialogue counts for 10 different types under Chinese-English scenarios. At the same time, we report the length distribution of English dialogues in the PlotStream dataset in Figure 5 and the length distribution of Chinese dialogues in Figure 4, with the specific length distribution values provided in Table 11. Overall, the average dialogue length under Chinese scenarios is longer, so the Chinese proportion in the PlotStream dataset reaches approximately 60%.

H PlotStream Dataset Validation

To ensure high-quality annotations, we hired five professional annotators with 2–3 years of role-playing experience. To achieve balanced coverage and overlap review, we randomly sampled 500 dialogue sets from PlotStream and assigned 135 Chinese and 135 English sets to each annotator, which covered about 10% of the overall dialogue turns and averaged 2.6 annotators per set. To maintain evaluation rigor, we instructed them to score across RoleArena-Eval's five dimensions, while predictive experiments applied a rejection threshold of 3.0 overall or below 2.0 per dimension.

Table 12 reports the observed agreement, expected agreement, acceptance rate, and Krippendorff's Alpha of 0.84 to measure inter-annotator reliability. This yielded an 86% acceptance rate, which confirms the high quality of PlotStream. Among the rejected samples, 50% exceeded the maximum turns without reaching a resolution, so we excluded them from the final dataset. To align with manual standards, we filtered the PlotStream dataset using GPT-4o and RoleArena-Eval, removing all entries with composite scores below 3.

I Case Study

J Cross-Model Sensitivity Analysis

We selected two models with a significant difference in parameter scale, DeepSeek-V3.1 and Qwen2.5-7B-Instruct, mentioned in the paper, for cross-model experiments. We generated 100 groups of dialogues in both Chinese and English scenarios and reported the average results in Table 13.

Our results show that: 1) A high-performance

Model	Style (EN)	SIM (EN)	Style (ZH)	SIM (ZH)
Qwen3-Embedding-8B	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**	—	**0.3985**

Table 6: Plot Line Similarity Computation Results Based on Qwen3-Embedding-8B.

Model	Style (EN)	SIM (EN)	Style (ZH)	SIM (ZH)
KaLM-Embedding-Gemma3-12B-2511	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**	—	**0.4422**

Table 7: Plot Line Similarity Computation Results Based on KaLM-Embedding-Gemma3-12B-2511.

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 8: Comparison of Dialogue Turns in User Simulation Scenarios Between Unfine-tuned Qwen2-72B-Instruct and PlotRole-72B.

Setting	Max	Range	Succ. Rate	Avg Turns
Base	6	(3,5)	93%	4.3
	5	(2,4)	92%	3.5
	7	(4,6)	93%	5.1
W/O Const.	6	(3,5)	72%	4.1
	5	(2,4)	78%	3.6
	7	(4,6)	75%	5.5

Table 9: Sensitivity Analysis of the Threshold for RoleArena Dialogue.

Critic Agent can greatly enhance the success rate of dialogue generation. We analyzed specific samples and found that using a low-performance Critic Agent causes the Critic Agent's function to fail, leading it to easily conform to the Env Agent's plot progression, losing control over the overall dialogue turns. At the same time, this result also confirms the effectiveness of RoleArena's innovative approach of introducing the Critic Agent. 2) Using high-performance and similar-performance Agents can better ensure the stability of dialogue generation.

K Further Model Evaluation Outcomes

L RoleArena Algorithm Design

We present the algorithmic procedure for RoleArena in Alg. 1. In the pseudocode, we highlight the discretization of plot lines (line 4) and the critical plot advancement (line 29). The pseudocode does not directly specify the annotation step for character responses during

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 10: Statistical results of character types in PlotStream

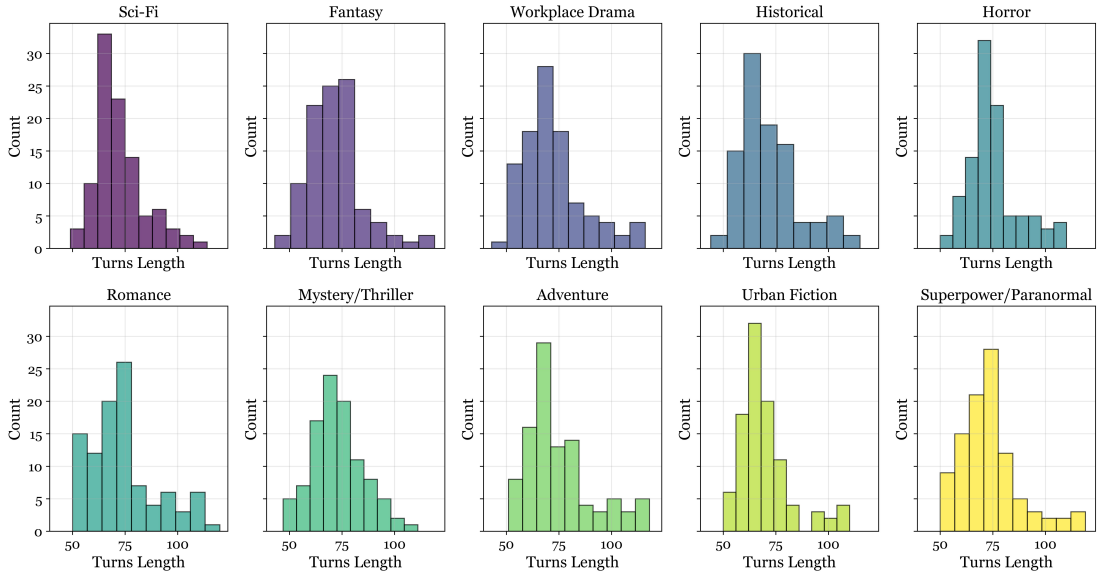

Figure 4: Dialogue length distribution in Chinese scenarios in PlotStream

.

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

Table 11: Dialogue Length Statistics by Language in PlotStream.

critical plot advancements. This is because, when constructing the training data for PlotRole, we can directly select the next character utterance following the critical plot annotation, which serves as the response at the plot turning point. Therefore, there is no need to specially annotate this statement in the RoleArena algorithmic procedure.

M Prompts

We provide all prompts used in our study, including those for generating character profiles, initializing the environment agent and character agents, environment agent's decision-making, plot progression, critic agent's evaluation, and our assessment based on five different dimensions. The prompts cover the complete workflow from character creation to multi-dimensional evaluation. All prompts were iteratively refined through multiple rounds of testing and validation. The complete set of prompts ensures reproducibility and consistency in our experimental setup.

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

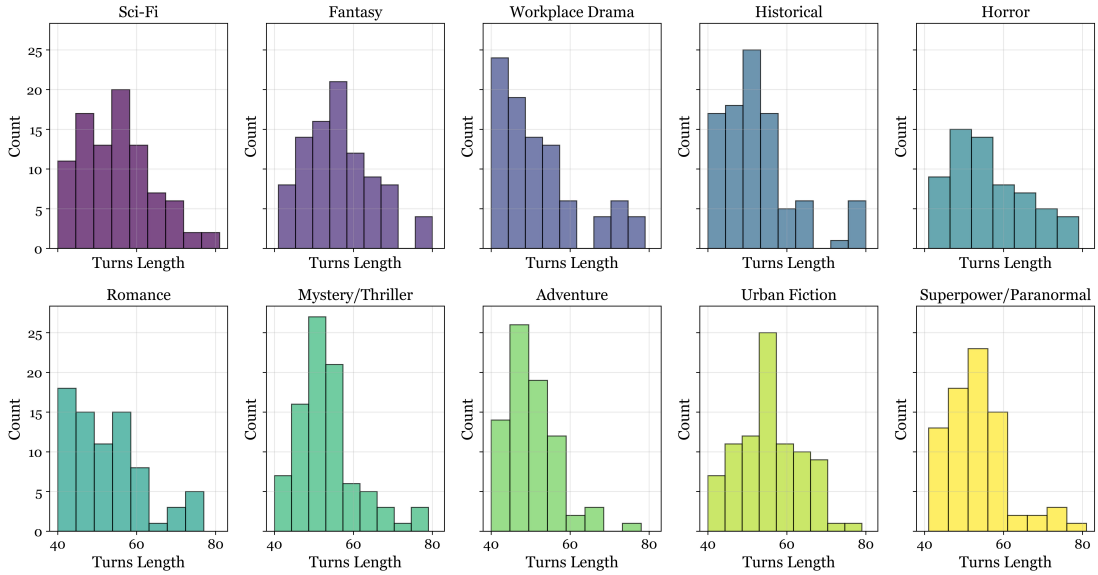

Figure 5: Dialogue length distribution in English scenarios in PlotStream

.

(a) English scene (b) Chinese scene (c) Overall distribution

Figure 6: Dialogue length distribution in PlotStream.

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.

Language	Po	Accept Rate	Pe	Krippendorff's Alpha	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 12: Manual Inspection Metrics for PlotStream.

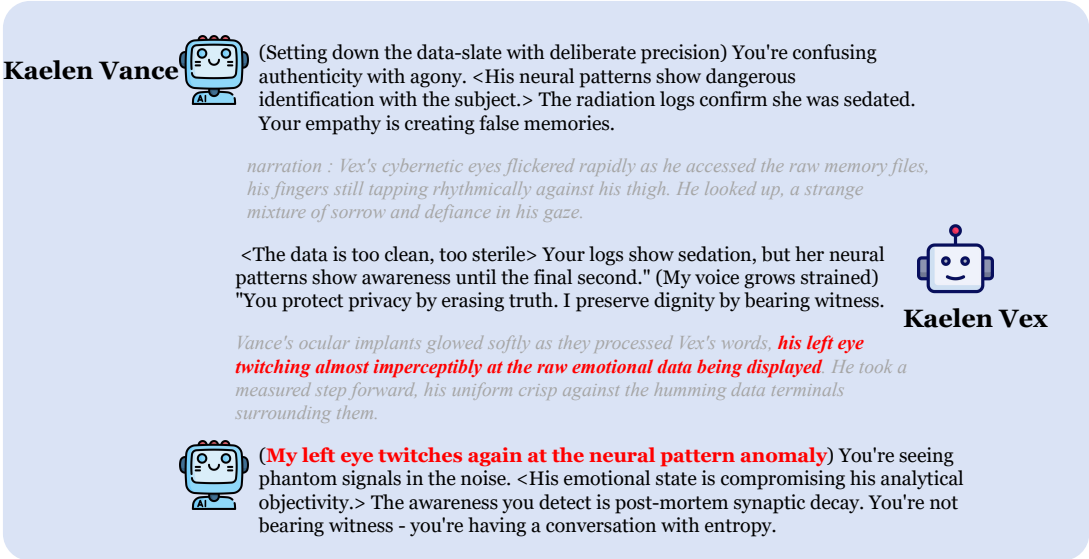

Figure 7: Case Showcase for English Scenarios.

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 13: Cross-Model Sensitivity Analysis for RoleArena.

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

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.

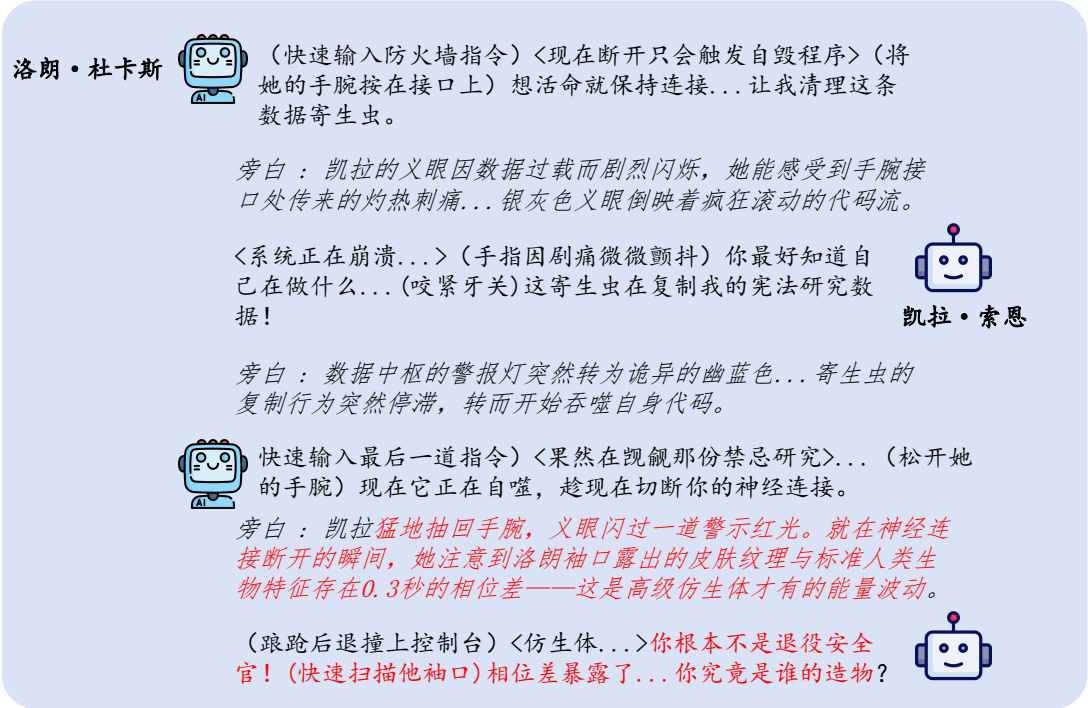

Figure 8: Case Showcase for Chinese Scenarios.

Lang	LLM	DF	CF	EE	SQ	PP	Avg
ZH	DeepSeek-V3.1	4.220	4.158	4.028	4.672	4.566	4.329
	Gemini-2.5	4.105	4.341	4.256	4.431	4.603	4.347
	Claude-Sonnet-4	3.934	4.017	4.128	4.374	4.506	4.192
	Claude-Sonnet-4.5	4.068	4.449	4.227	4.510	4.613	**4.373**
EN	DeepSeek-V3.1	4.406	4.564	4.550	4.912	4.550	4.596
	Gemini-2.5	4.215	4.725	4.574	4.915	4.582	**4.602**
	Claude-Sonnet-4	4.392	4.218	4.624	4.712	4.495	4.488
	Claude-Sonnet-4.5	4.259	4.592	4.636	4.895	4.506	4.578

Table 14: Evaluation Results of High-Performance Closed-Source Models.

- 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.
3. Based on the story theme, provide the number of storylines in line with the [Number of Storylines Required].
4. The story's plot lines should be separated by numbers, e.g.,

Algorithm 1 RoleArena

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

Ensure: Dialogue trajectory $\mathbf{D}$

Initialize agents $\mathcal{A} \leftarrow \{\pi_a(r_i)\}$
Extract story setting $\mathbf{s} \leftarrow \text{extract}(\mathbf{m})$
Parse plot nodes $\mathcal{P} = \{p_1, \ldots, p_K\}$
$t \leftarrow 1, i \leftarrow 0, \mathbf{D} \leftarrow \emptyset$
while $t \leq T_{\max}$ **and** $i < K$ **do**
  Initialize turn $\mathbf{turn}_t$
  if $t = 1$ **then**
    $(n_t, a_t, \mathcal{U}) \leftarrow \text{env_init}(\mathbf{s}, \mathcal{R})$
  else
    $(n_t, a_t) \leftarrow \text{env_step}(\mathbf{turn}_{t-1}, p_i)$
  end if
  while $|\mathcal{U}| > 0$ **do**
    $r_{a_t} \sim \pi(a_t \mid n_t, \mathcal{M}_{a_t})$
    Append r_{a_t} to $\mathbf{turn}_t$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{a_t\}$
    if $\text{judge_plot}(\mathbf{turn}_t, p_i)$ **then**
      $\mathbf{c} \leftarrow \text{critic}(\mathbf{turn}_t, p_i)$
      if $\mathbf{c}$.approve **then**
        $\mathbf{turn}_t$.advance $\leftarrow$ true
        $i \leftarrow i + 1$
        break
      end if
    end if
    $(a_t, n_t) \leftarrow \text{env_next}(r_{a_t}, \mathcal{U}, p_i)$
  end while
  Append $\mathbf{turn}_t$ to $\mathbf{D}$
  $t \leftarrow t + 1$
end while
return $\mathbf{D}$

1. Characters are talking
2. Strange sounds are heard around, and one character goes to investigate
3. Characters work together to solve a puzzle...

5. The plot twists should be dramatic, intriguing, and have surprises and suspense.

6. The scene you create should allow characters to be in the same space and complete the dialogue.

7. 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.

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]

1. 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.

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.

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.

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 × (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.

Dialogue Fluency & Coherence Evaluation Expert Prompt

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.

Character Fidelity Evaluation Expert Prompt

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.

Emotional Expression Evaluation Expert Prompt

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.

Story Quality Evaluation Expert Prompt

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.

Plot Progression Evaluation Expert Prompt

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.

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"
},
"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.