CharacterCraft: Bridging the Literature-Reality Dialogue Gap for Practical Role-Playing Agents

Anonymous ACL submission

Abstract

Recent advancements in large language models (LLMs) have given rise to the emergence of role-playing agents (RPAs). The development of high-quality dialogue datasets is critical for advancing RPAs. However, existing datasets have two main issues: (1) the bias between query distributions and real-world user language usage, and (2) the challenge of ensuring responses accurately reflect character traits. To address these issues, we propose CharacterCraft, a novel framework designed for practical RPAs, comprising a tailored Chinese role-playing dataset and a robust evaluation method. First, we develop a specialized model for Chinese dialogue extraction, achieving state-of-the-art performance. Using this model, we then extract a large amount of character dialogue from novels, ensuring high data quality (issue 2). To mitigate the literature-reality dialogue bias in extracted dialogue (issue 1), we introduce an iterative augmentation-reconstruction method, which revises queries to better align with common language usage. Additionally, we propose a context-aware memory retrieval module for fine-grained alignment with the character and introduce a reference-guided LLM-as-a-judge evaluation method for more reliable assessments by comparing their responses to source material dialogues. Our automated pipeline produces a large-scale high-quality Chinese role-playing dataset with 21,392 samples and 121,418 utterances. The experimental results demonstrate the effectiveness of our framework and reveal the limitations of existing RPAs when faced with diverse scenes.

1 Introduction

The rapid advancement of large language models (LLMs) has spurred significant innovations across various real-world applications (Si et al., 2024; Jin et al., 2024; Shanahan et al., 2023). Within these applications, role-playing agents (RPAs) (Shao et al., 2023) have gained prominence as a critical area, allowing users to engage with immersive and highly personalized characters. Platforms like Character.AI[1] and Xingye[2] exemplify this trend, showcasing how LLMs can create interactive experiences that cater to diverse user preferences. Recent research indicates that high-quality role-playing dialogue datasets are essential for both developing RPAs (Yu et al., 2024b; Bai et al., 2024) and evaluating their performance (Wang et al., 2023; Zhou et al., 2024b; Wang et al., 2024).

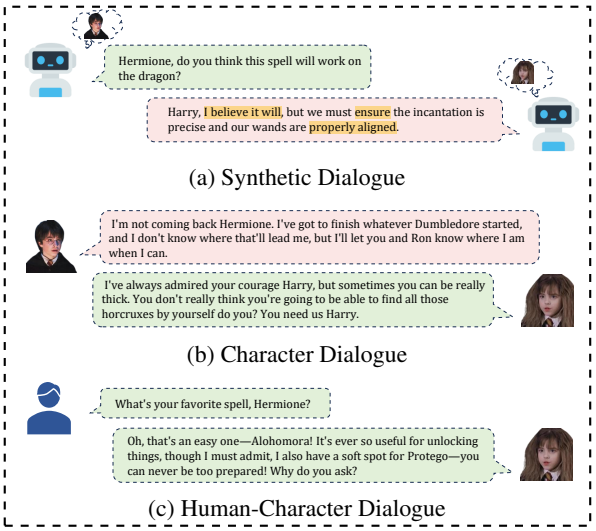


Figure 1: Examples of three types of role-playing dialogue. (a) Dialogue synthesized by LLMs. Words highlighted in the response are dull and do not reflect the character traits. (b) Character dialogue extracted from literary works. The query is highly characterized, with a significant difference from user interactions. (c) Dialogue between users and RPAs.

However, these existing efforts towards high-quality role-playing dialogue datasets face two serious issues: (1) the bias between query distributions and language usage of common users, and (2) the challenge of ensuring responses accurately reflect character traits. Although employing hu-

[1]https://character.ai/
[2]https://www.xingyeai.com/

man experts is ideal for corpus construction (Zhou et al., 2024a), the high costs and limited scalability hinder its widespread adoption. In contrast, the automated methods for constructing role-playing datasets mainly fall into two categories: *synthetic data-based* approach (Wang et al., 2023) and *literary resources-based* approach (Tu et al., 2024).

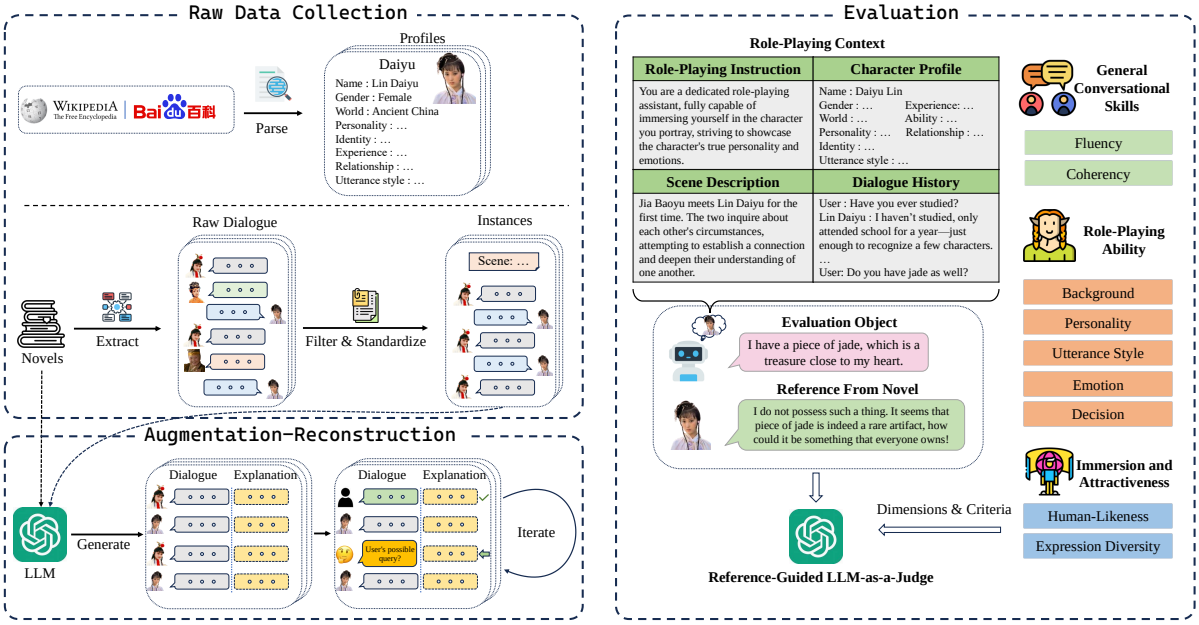
The synthetic data-based approach uses LLMs to generate user-character dialogues, making it cost-effective and easily expandable (Ge et al., 2024). However, even state-of-the-art LLMs struggle to achieve precise alignment with the target character, resulting in synthetic dialogues that tend to be formulaic and dull, as shown in Figure 1a. Thus, synthetic data fails to address issue 2. The literary resources-based approach involves extracting character dialogues from literary sources (Tu et al., 2024). The exceptional quality of the dialogues is beyond doubt, as the roles' characteristics are well-defined and thoroughly developed in these literary works (Chen et al., 2024c). However, several serious problems remain. Firstly, accurately extracting dialogues is not a straightforward task, with even advanced LLMs achieving accuracy rates below 90%, as shown in Table 1. Additionally, the language style and pattern of dialogue in literary works typically differ from real-world interactions between users and RPAs (see Figure 1b and 1c). Consequently, evaluations conducted on these datasets may fail to accurately represent authentic role-playing capabilities in real-world applications, as detailed in §C.1, thereby failing to address issue 1.

To address these challenges, we propose CharacterCraft, a novel role-playing framework designed to better align with real-world application scenarios, including dataset construction, a context-aware memory retrieval module, and a reference-guided LLM-as-a-judge evaluation method. To ensure high-quality dialogue sources (issue 2), we begin by extracting character dialogues from literary works. Given the accuracy limitations with current LLMs, we develop a dialogue extraction model trained on a manually annotated dataset. Our model achieves an extraction accuracy of 94.53%, significantly surpassing the advanced LLMs (e.g., GPT-4o at 89.84%). To mitigate the literature-reality dialogue bias (issue 1), we introduce an iterative *augmentation-reconstruction* approach. Since the dialogues extracted from the novel are disconnected from their context, potentially leading to a lack of semantic coherence, during the augmentation stage, we generate contextual explanations for each utterance using information from the source material. This process incorporates context beyond the dialogue to enhance semantic coherence and completeness. In the reconstruction stage, we mask the original query and use the explanations generated during the augmentation stage to guide the LLM in inferring and reconstructing user queries that better match the language usage of common users. Using this approach, we construct a large-scale, high-quality Chinese role-playing dataset, including 21,392 multi-turn dialogues and 121,418 utterances from 369 characters.

The character profiles only provide static information. To achieve fine-grained alignment with the character across diverse scenes, we design a context-aware memory retrieval module. This module extracts contextually similar instances from a character memory repository and integrates them as contextual knowledge, strengthening the model's comprehension of character traits and behavioral patterns. For evaluation, prior studies typically rely on reward models (Tu et al., 2024; Chen et al., 2024a) or LLM-as-a-judge approaches (Liu et al., 2024a). Nevertheless, reward models trained on limited datasets are often unreliable (Liu et al., 2024b) (detailed in §C.2), while LLM-as-a-judge evaluations may suffer from the insufficient background knowledge of the character (Chen et al., 2024c). To address these limitations, we propose reference-guided LLM-as-a-judge. By providing character responses from source materials as evaluation references, this approach enables more precise performance assessments of RPAs.

Our contributions are summarized as follows:

- We develop a dialogue extraction model that achieves state-of-the-art performance in Chinese dialogue extraction tasks.
- We analyze the biases between existing role-playing datasets and real-world application scenarios, and construct a large-scale, high-quality Chinese role-playing dataset designed to better align with real-world scenarios and evaluation needs.
- We propose a context-aware memory retrieval module for fine-grained alignment with the character across diverse scenes and introduce reference-guided LLM-as-a-judge for stable, comprehensive evaluation of RPAs across multiple dimensions.



(a) Dataset Construction (b) Performance Evaluation

Figure 2: Illustration of CharacterCraft. (a) CharacterCraft-Data includes character profiles collated from online encyclopedia, scenes summarized by LLMs, dialogues meticulously revised using the augmentation-reconstruction method. (b) CharacterCraft-Eval conducts evaluation across 9 dimensions of 3 aspects.

2 Problem Definition

The Role-Playing Agent (RPA) is designed to allow users to interact with immersive and highly personalized characters. In practical applications, a human user interacts with an RPA designed to imitate a specific character $\mathcal{C}$. The character $\mathcal{C}$ is defined by a profile $\mathcal{P}$, which includes essential attributes such as identity, linguistic style, personality, etc. The interaction takes place in a scene $\mathcal{S}$ and a dialogue context $\mathcal{D} = [q_1, r_1, q_2, r_2, \ldots, q_n]$ where q_i and r_i represent the i-th utterances from the human user and the RPA, respectively. The goal of the RPA is to generate a response $r_n = \text{RPA}(\mathcal{P}, \mathcal{S}, \mathcal{D})$ which is consistent with the character's profile $\mathcal{P}$. Notably, in real-world applications, human users may also obtain a specific character identity to enhance interactions. However, users typically do not consciously align their language style with the role they are playing. Thus, we assume that users' linguistic patterns exhibit alignment with daily communication or common language usage.

3 CharacterCraft Corpora

As presented in Figure 2a, we construct CharacterCraft dataset, a large-scale, high-quality Chinese role-playing collection that aligns with real-world user interactions. All prompts used in this section are detailed in §D.

3.1 Dataset Construction

The dataset construction process consists of four steps: raw dialogue extraction, dialogue filtering, augmentation-reconstruction, and manual verification.

Raw Dialogue Extraction: The performance of advanced LLMs (*e.g.*, GPT-4o) in dialogue extraction tasks remains suboptimal (Yu et al., 2024b). These models frequently struggle with accurately extracting dialogues and correctly identifying speaker roles. To mitigate this issue, we manually annotate a subset of the data to ensure precise dialogue extraction, as detailed in §A.1. Leveraging this annotated dataset, we fine-tune Qwen2-1.5B-Instruct as a dialogue extraction model. As shown in Table 1, our model outperforms state-of-the-art LLMs in Chinese dialogue extraction tasks and establishes a robust foundation for subsequent data processing. Following this, the text from novels is segmented using a heuristic algorithm (Li et al., 2023), and our model is applied to extract utterances from each segment.

Dialogue Filtering: We calculate the frequency of each character's dialogue appearances. Characters who exceed a predefined dialogue count threshold are selected as target characters, effec-

Model	Params	Recall	Accuracy
GPT-4o	-	89.84%	94.42%
GPT-3.5-Turbo	-	67.91%	77.12%
GLM4-plus	-	81.90%	84.73%
Deepseek-V3	671B	88.45%	93.51%
Qwen2.5-72B-Instruct	72B	85.38%	92.30%
Ours	1.5B	**94.53%**	**94.94%**

Table 1: Comparison of LLMs in dialogue extraction: Params, Recall and Accuracy.

tively filtering out minor characters with limited involvement. Next, we manually construct detailed character profiles by collating information from Baidu Baike[3] and Wikipedia[4] (detailed in §A.2). We retain dialogue exchanges between two characters until a third character intervenes. To prevent a single multi-turn dialogue from spanning multiple scenes or plot points, we assess continuity by measuring the token-level distance between adjacent utterances. If this distance exceeds a predefined threshold, we segment the dialogue to maintain coherence. Then we use GPT-4o to summarize the scene for each segment.

Augmentation-Reconstruction: Through previous steps, we collect a substantial dataset of multi-turn dialogues. However, these dialogues take place between characters in literary works, which differ significantly from the language usage of users in real-world applications. To address this bias, we propose an iterative *augmentation-reconstruction* method. In the augmentation stage, given the original text $\mathcal{X}$ and the multi-turn dialogue $\mathcal{D} = [q_1, r_1, q_2, r_2, \ldots, q_n, r_n]$, we use GPT-4o to generate an augmented contextual explanation for each utterance, which can be represented as:

$$\mathcal{E} = [e_{11}, e_{12}, e_{21}, e_{22}, \ldots, e_{n1}, e_{n2}], \quad (1)$$

where each e_{i1} and e_{i2} correspond to the explanations for the i-th utterances of two characters. Formally, the generation process is defined as:

$$\mathcal{E} = \mathrm{M}(\mathcal{T}_A(\mathcal{X}, \mathcal{D})), \quad (2)$$

where M represents the LLM (*e.g.*, GPT-4o) responsible for generating explanations and $\mathcal{T}_A$ represents the prompt template used to generate explanations.

[3]https://baike.baidu.com/

[4]https://www.wikipedia.org/

The explanation $\mathcal{E}$ provides a concise summary of each utterance, incorporating contextual information outside the conversation. In the reconstruction stage, we mask the original query within each dialogue and use the explanations generated during the augmentation stage to iteratively guide the model in inferring and reconstructing user queries that better match real-world application needs. Algorithm 1 shows the pseudo codes. Specifically, for each turn in the dialogue, the model M is prompted with a combination of reconstructed queries and explanations from previous turns, along with the corresponding responses. This process iteratively builds the dialogue by reconstructing the current user query q_i', using the augmented explanation e_{i1} as a guide. The resulting reconstructed dialogue $\mathcal{D}'$ preserves the semantic meaning of the original dialogue $\mathcal{D}$ while adapting its style and structure to match the language usage of common users.

Algorithm 1 Augmentation-Reconstruction

Input: Original text $\mathcal{X}$, Multi-turn dialogue $\mathcal{D} = [q_1, r_1, q_2, r_2, \ldots, q_n, r_n]$, LLM M, Prompt template for generating explanations $\mathcal{T}_A$, Prompt template for query reconstruction $\mathcal{T}_R$

Set the final reconstructed multi-turn dialogue $\mathcal{D}' = \phi$
Get the explanation $\mathcal{E}$ via Eq.(2)
for $i = 1, 2, \ldots, n$ **do**
    Set $L = [(q_1', e_{11}), (r_1, e_{12}), \ldots, e_{i1}, (r_i, e_{i2})]$
    $q_i' = \mathrm{M}(\mathcal{T}_R(L))$
    Append the i-th reconstructed query q_i' to $\mathcal{D}'$
    Append the i-th response r_i to $\mathcal{D}'$
end for

Output: The final reconstructed multi-turn dialogue $\mathcal{D}'$

Manual Verification: We use GPT-4o to evaluate the coherence and fluency of multi-turn dialogues. Dialogues identified as low-quality undergo manual revisions. Please refer to §A.3 for further details.

3.2 Dataset Analysis

As shown in Table 2, our dataset includes 21,392 multi-turn dialogues and 121,418 utterances from 369 unique characters, which is significantly larger than most existing datasets, providing a more comprehensive resource for role-playing tasks that require both diversity and quality.

We use perplexity to evaluate the coherence of multi-turn dialogues. Specifically, we calculate perplexity scores using Qwen2.5-7B-Instruct, which predicts the next utterance based on role-playing instructions and dialogue history. Lower perplexity values indicate higher contextual coherence (Sartor et al., 2024). As shown in Table 3, our dataset

Dataset	Source[1]	Automated Construction	Multi-turn	# Roles	#Sessions	#Turns	#Avg. Sessions	#Avg. Turns
HPD	●	✗	✓	-	1191	15542	-	13.05
RoleLLM	○	✓	✗	100	-	23463	-	-
CharacterEval	●	✓	✓	77	1785	16565	23.18	9.28
CharacterGLM	⊙	✗	✓	250	1034	32816	4.14	31.74
CharacterLLM	○	✓	✓	9	1600	21120	177.78	13.20
Beyond Dialogue	●	✓	✓	311	3552	23247	10.73	6.54
WIKIROLE	○	✓	✓	3092	7086	36164	2.29	5.10
CharacterCraft (Ours)	◐	✓	✓	369	21392	121418	57.97	5.68

[1] ●: Literary Resources; ○: Synthetic Data; ⊙: Human Role-playing; ◐: Hybrid Data

Table 2: Dataset statistics. Comparing our dataset with existing open-source role-playing datasets.

outperforms other role-playing corpora.

Additionally, we use Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) to measure the similarity between the distribution of user queries in the role-playing dialogue dataset and that of real-world human conversations. KL divergence quantifies the difference between the dataset's probability distribution and the reference distribution, which, in this case, is the NaturalConv (Wang et al., 2021) —a Chinese multi-turn dialogue collection from authentic human conversations. A lower KL divergence suggests that the user query distribution is closer to the distribution observed in regular human conversations. As shown in Table 3, our dataset outperforms other role-playing datasets in both perplexity and KL divergence, indicating that its query distribution more closely resembles real-world human dialogue.

We also conduct an ablation study to evaluate the performance of our dataset without the augmentation-reconstruction method. The results show a modest increase in both perplexity and KL divergence, highlighting the importance of our method in improving dataset quality.

Dataset	Perplexity↓	KL Divergence↓
Beyond Dialogue	3.67	0.71
Charactereval	3.62	0.68
Ours	**3.54**	**0.55**
- *w/o A-R*	3.71	0.73

Table 3: Comparison of our dataset with other role-playing Datasets. The perplexity metric is calculated using Qwen2.5-7B-Instruct, while the KL divergence is computed for each dataset relative to the NaturalConv dataset. Here, 'A-R' means the augmentation-reconstruction stage.

3.3 Context-Aware Memory Retrieval

Based on the dataset we developed, we design a context-aware memory retrieval module to provide relevant context to RPAs. Specifically, we treat all data associated with a character as its memory repository, encoding it into embedding vectors for each scene and query using BGE-large (Xiao et al., 2024). When an RPA is presented with a new scene and query, we retrieve the most relevant instance from the character's memory repository to serve as a demonstration, guiding the RPA to generate responses that align with the character's traits.

4 CharacterCraft-Eval

In this section, we begin by outlining a series of evaluation dimensions from three aspects, as shown in Figure 2b. We then introduce our reference-guided LLM-as-a-judge method, which provides stable and comprehensive assessments of RPAs.

4.1 Evaluation Dimensions

Integrating prior work (Dai et al., 2024; Xu et al., 2024), we propose a three-aspect evaluation framework to thoroughly evaluate the performance of RPAs, which encompasses general conversational skills, character consistency, and immersion and attractiveness (detailed in §B).

At first, the ability of RPAs to maintain fluent and coherent communication with users is the most important aspect. We can measure this aspect using two dimensions: *fluency (Flu.)* and *coherency (Coh.)* (Tu et al., 2024).

Character consistency is crucial for evaluating RPAs, measuring their ability to convincingly imitate and sustain a distinct character by aligning their language and behaviors with the given character. We utilize *background (Bg.)* (Yu et al., 2024b) and *personality (Pers.)* (Wang et al., 2024) to assess how well RPAs align with the character's background and personality. Moreover, we utilize *utterance style (US)* to measure whether the language style of RPAs matches that of the character (Yu et al., 2024a). Also, we adopt *emotion (Emo.)* (Zhou et al., 2024b) and *decision (Dec.)* (Xu

et al., 2024) to evaluate whether the RPAs exhibit emotions and decision-making abilities consistent with those of the character.

In addition to these aforementioned aspects, immersion and attractiveness are other critical factors, referring to the ability to empower an immersive and engaging user experience. Here, we choose *human-likeness (HL)* (Zhou et al., 2024b) to evaluate the naturalness of the RPAs' responses, and *expression diversity (ED)* to measure the richness and diversity of the responses generated by RPAs.

More specifically, the dimensions of general conversational skills and character consistency are rated using a 3-point scale (0-2), while human-likeness is rated on a 4-point scale (0-3). The higher score indicates better performance in each dimension. Expression diversity is measured by computing the mean value of the Self-BLEU metric (Zhu et al., 2018) derived from the responses generated by the RPA. The higher score indicates that the RPA's responses are less diverse. Furthermore, since the character traits reflected in the response vary with the scene, we classify these dimensions as dynamic (personality, emotion, decision) and static (others). These evaluation dimensions are described with more details in §D.

4.2 Reference-Guided LLM-as-a-Judge

LLM-as-a-judge is widely used for evaluating open-ended generation tasks, including dialogue response generation, summarization, and creative writing (Li et al., 2025; Zheng et al., 2023). As a result, numerous studies employ LLMs as evaluators for RPAs (Liu et al., 2024a; Yu et al., 2024a). However, as noted by Chen et al. (2024c), LLMs often struggle to deeply understand characters, limiting their effectiveness. Although character profiles are provided to the judge model, these static profiles contain limited information and fail to help the model adapt to diverse scenes. When the judge model lacks clarity on what constitutes an appropriate response, its reliability is significantly compromised. To address this challenge, we propose a reference-guided LLM-as-a-judge method. Specifically, for dynamic dimensions, we provide the judge model with character responses from the same scenes in the novel as ground truth. Additionally, we refine the evaluation criteria to ensure that the judge model incorporates these reference responses during the evaluation process. As noted by Zhou et al. (2024b), the manifestation of character traits across dynamic dimensions within responses is sparse, meaning that multiple dimensions of character traits are unlikely to be observed in a single response. Therefore, before evaluation, we first use GPT-4o to analyze the ground truth for each instance and determine the character dimensions it reflects, which we then assess further.

5 Experiments

5.1 Experiment Setup

Datasets. Due to the high computational cost of employing GPT-4o API, following previous studies (Yu et al., 2024b; Ahn et al., 2024), we sample 500 session instances from the dataset to evaluate the performance of RPAs. Specifically, we manually selected 40 characters from the dataset, focusing primarily on protagonists or secondary protagonists from novels. For each selected character, we sampled at least 10 session instances, which were then combined to form the final test set.

Setting. For an instance containing character profile $\mathcal{P}$, scene description $\mathcal{S}$ and dialogue context $\mathcal{D}$, we can get a response $r_n = \text{RPA}(\mathcal{P}, \mathcal{S}, \mathcal{D})$ as we discussed in Section 2. The response r_n is our evaluation object.

Baselines. As shown in Table 5, we evaluate a diverse set of LLMs, including open-source models, proprietary models, and models specialized for role-playing tasks. For open-source LLMs, we evaluated the chat-version of the following: Qwen2.5 (Qwen, 2025), Baichuan2 (Yang et al., 2023), GLM-4 (GLM et al., 2024), Yi-1.5 (Young et al., 2024), DeepSeek-V3 (DeepSeek-AI, 2024), and Llama-3.1 (Dubey et al., 2024). For proprietary LLMs, we select two widely recognized models: GPT-3.5-Turbo and GPT-4o (Achiam et al., 2023). Additionally, we include two LLMs specifically optimized for role-playing tasks: Baichuan-NPC and CharacterGLM (Zhou et al., 2024a). A consistent prompt is used for all models, with minor modifications applied only to Baichuan-NPC due to its unique API requirements.

Evaluation on LLM-as-a-Judge. We employ human annotators to evaluate responses generated by Qwen2.5-7B-Instruct in a reference-guided setting. Comparative assessments are then conducted across several LLMs under both reference-guided and reference-free conditions. The results in Table 6 reveal that LLMs (*e.g.*, GPT-4o) show a low correlation with human evaluations in dynamic dimensions without references, suggesting limited reliability. In contrast, our reference-guided

Models	Overall	Conversation		Character Consistency					Immersion	
		Flu.↑	Coh.↑	Bg.↑	Pers.↑	US↑	Emo.↑	Dec.↑	HL↑	ED↓
Proprietary models										
GPT-3.5-Turbo	6.71	1.97	1.86	1.89	1.13	1.77	0.46	0.76	2.52	0.052
GPT-4o	7.36	**1.99**	**1.99**	1.99	1.44	1.98	0.63	0.91	2.92	0.076
Baichuan-NPC	6.57	1.97	1.78	1.84	1.01	1.63	0.50	0.82	2.46	**0.024**
CharacterGLM	7.11	1.98	1.93	1.98	1.21	1.94	0.63	0.85	2.71	0.055
Open-source models										
Baichuan2-7B-Chat	6.49	1.97	1.86	1.90	1.02	1.71	0.41	0.71	2.33	0.074
Baichuan2-14B-Chat	6.69	1.98	1.79	1.91	1.09	1.81	0.44	0.76	2.59	0.067
GLM-4-9B-Chat	7.11	1.99	1.94	1.94	1.29	1.94	0.59	0.81	2.88	0.102
Llama-3.1-8B-Instruct	6.45	1.94	1.74	1.81	0.97	1.67	0.47	0.77	2.46	0.053
Yi-1.5-9B-Chat	6.63	1.96	1.82	1.89	1.07	1.75	0.46	0.74	2.61	0.081
Qwen2.5-7B-Instruct	6.96	1.99	1.91	1.95	1.21	1.93	0.48	0.75	2.77	0.069
Qwen2.5-14B-Instruct	7.10	1.99	1.95	1.98	1.23	1.94	0.59	0.78	2.83	0.072
Qwen2.5-72B-Instruct	7.22	1.99	1.99	1.99	1.29	1.96	0.60	0.83	2.92	0.082
Deepseek-V3	7.41	1.99	1.99	1.98	**1.49**	**1.99**	0.70	0.91	**2.98**	0.104
With CMR										
Qwen2.5-7B-Instruct+CMR	7.23	1.99	1.90	1.94	1.28	1.93	0.68	0.95	2.85	0.060
Qwen2.5-14B-Instruct+CMR	7.37	1.99	1.93	1.96	1.35	1.93	0.76	1.02	2.88	0.062
Qwen2.5-72B-Instruct+CMR	**7.49**	1.99	1.99	**1.99**	1.47	1.94	**0.78**	**1.05**	2.90	0.078

Table 4: Main results. The Overall score was calculated by summing the normalized values of all dimensions, with the ED dimension directionally inverted. 'CMR' here means the context-aware memory retrieval module. Best performances are shown in **bold**, while suboptimal ones underlined.

method substantially improves the correlation between LLMs and human judgments.

Models	Specialized	Params	Open-Source	Primarily Language
Qwen2.5	✗	7B, 14B, 72B	✓	zh
Baichuan2	✗	7B, 14B	✓	zh
Glm-4	✗	9B	✓	zh
Yi-1.5	✗	9B	✓	zh
Deepseek-V3	✗	671B	✓	zh
Llama-3.1	✗	8B	✓	en
Baichuan-NPC	✓	-	✗	zh
CharacterGLM	✓	-	✗	zh
GPT-3.5-Turbo	✗	-	✗	en
GPT-4o	✗	-	✗	en

Table 5: LLMs evaluated in our experiments.

Model	Avg.	Flu.	Coh.	Bg.	Pers.	US	Emo.	Dec.	HL
Deepseek-V3	53.4	51.9	64.7	46.2	54.3	60.3	56.6	47.1	45.9
- w/o reference	40.6	-	-	-	27.2	-	11.5	16.9	-
GPT-3.5-Turbo	38.6	40.4	44.6	29.2	33.2	55.1	32.5	42.3	31.7
- w/o reference	29.9	-	-	-	19.6	-	4.0	14.9	-
GPT-4o	61.3	70.9	61.2	55.7	66.1	50.4	75.2	56.7	53.8
- w/o reference	42.6	-	-	-	18.8	-	14.2	15.5	-

Table 6: Kendall correlation coefficient (%) between LLMs evaluation results and human evaluation results, under both reference-based and reference-free conditions. Since reference-based evaluation is only applied to the Pers., Emo., and Dec. dimensions, the remaining dimensions are marked with '-' to indicate no change.

5.2 Experimental Results

As presented in Table 4, we report the average performance across all test instances for each evaluated RPA, including the Qwen2.5 series models with our context-aware memory retrieval module. The experimental results reveal a strong correlation between the role-playing capabilities of LLMs and their general performance. LLMs such as GPT-4o, Deepseek-V3, and Qwen2.5-72B-Instruct, which excelled on our benchmark, also achieved exceptional results in general capability evaluations. This observation explains why LLMs specifically designed for role-playing tend to exhibit only average performance, despite being trained on proprietary datasets.

We find that nearly all RPAs perform well in both the Fluency and Coherency dimensions, indicating that producing fluent, contextually appropriate responses is not a major issue. For background and utterance style, performance differences among RPAs are minimal, as most RPAs can generate responses that align with the given character traits in the profile. However, all RPAs currently struggle with dynamic dimensions including personality, emotion, and decision. For example, even advanced models like GPT-4o achieve modest scores (0.63 for emotion and 0.91 for decision on a 3-point scale), highlighting the difficulty of maintaining consistent dynamic traits across diverse scenes.

This suggests that while RPAs effectively leverage parametric or contextual knowledge to capture static attributes, they struggle to adapt dynamically to diverse scenes. Notably, Deepseek-V3 achieves a high overall score but underperforms in expression diversity. Its responses often rely on repetitive catchphrases or rigid sentence structures, reinforcing character traits at the expense of linguistic variety, leading to diminishing user engagement due to predictable interactions.

Furthermore, using our dataset as a character memory repository significantly boosts RPA performance across models of varying sizes. Notably, the CMR-enhanced Qwen2.5-14B-Instruct model achieves a higher overall score than GPT-4o, while the CMR-enhanced Qwen2.5-72B-Instruct outperforms all other RPAs, achieving state-of-the-art results. The improvements are especially evident in dynamic scenarios, with an average performance gain exceeding 23%. These findings confirm that our module enhances RPAs' ability to deeply understand role-specific attributes, enabling them to generate behaviors more consistent with character traits across diverse scenes. We refer readers to §C.3 for more experimental analysis.

6 Related Work

Role-playing Agents. Recent advancements in large language models (OpenAI, 2024; DeepSeek-AI, 2024; Qwen, 2025) have driven the rapid development of role-playing agents (RPAs). RPAs are typically developed through two primary approaches: (1) training on specialized role-playing datasets (Yu et al., 2024b; Zhou et al., 2024a; Wang et al., 2023; Shao et al., 2023) or (2) providing instructions and examples to general-purpose LLMs to simulate specific characters (Wang et al., 2023; Li et al., 2023). The construction of high-quality role-playing datasets is essential for both the development and evaluation of RPAs. Existing studies have explored various strategies for dataset construction, including extracting dialogues from literary sources (Wang et al., 2023; Chen et al., 2023; Li et al., 2023), synthesizing data using large language models (Wang et al., 2024; Shao et al., 2023; Wang et al., 2023), and generating dialogues through human involvement (Zhou et al., 2024a). CharacterBench (Zhou et al., 2024b) compiles character data using multiple methods to enhance diversity and mitigate biases. Meanwhile, MMRole (Dai et al., 2024) extends role-playing tasks into the multimodal domain by integrating textual and visual data. Although the quality of character datasets continues to improve, existing efforts has not adequately consider the language discrepancy between real-world users and the target character, which remains a key limitation affecting RPA performance.

Evaluation. Chen et al. (2024b) categorize RPA evaluation into two main aspects: (1) character-independent capabilities (e.g., conversational fluency) and (2) character fidelity, which measures how accurately an agent replicates a target persona. Other studies have proposed more fine-grained evaluation criteria. For example, CharacterEval defines 13 evaluation dimensions, including fluency, factual accuracy, and human-likeness (Tu et al., 2024). While human evaluation provides high accuracy, it is costly, time-consuming, and difficult to reproduce. Automated evaluation offers a scalable alternative, with LLM-as-a-judge (Yu et al., 2024b; Wang et al., 2024, 2023) and reward model training (Dai et al., 2024; Tu et al., 2024) being the most widely adopted methods. Studies suggest that reference-based evaluation improves the reliability of judge models (Zhang et al., 2024). Accordingly, we employ a reference-guided LLM-as-a-judge method to assess RPAs.

7 Conclusion

In this work, we highlight the bias between current role-playing datasets and real-world applications. We introduced CharacterCraft, a novel framework designed to better align with real-world application scenarios. By integrating high-quality literary dialogue extraction with an iterative augmentation-reconstruction method, we construct a large-scale, high-quality Chinese role-playing dataset. To enhance RPAs' contextual understanding, we proposed a context-aware memory retrieval module. Additionally, our reference-guided LLM-as-a-judge evaluation improves role-playing assessments by incorporating source material references. Experiments underscore RPAs' challenges in capturing dynamic character traits and show our framework's effectiveness.

8 Limitations

Despite its advantages, CharacterCraft has some limitations. First, our dataset is exclusively in Chinese, limiting its applicability to multilingual RPAs. Future work should consider expanding to other languages to improve generalizability. Sec-

ond, while our reference-guided LLM-as-a-judge evaluation improves assessment reliability, it incurs high API costs, making large-scale evaluations expensive. Efficient evaluation methods or fine-tuned reward models could help mitigate this issue. Lastly, our dataset is derived solely from fictional characters, excluding real-world characters such as celebrities or historical characters. Incorporating real-world personas could enhance RPAs' adaptability to broader applications, such as educational or professional role-playing scenarios.

9 Ethical Considerations

In this study, we use data derived from original novel texts. We acknowledge that the authors and publishers of these novels hold the copyrights to the material, and we respect these intellectual property rights. Our work adheres to the principles of academic research and will be released only for non-commercial, educational, and research purposes. Additionally, all data samples were manually reviewed to identify and filter out any content that could be considered harmful. This includes content that might perpetuate harmful stereotypes, promote violence, or have other adverse psychological or social impacts. By adhering to these guidelines, we aim to minimize any ethical concerns and promote the responsible use of our resources in the broader research community.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

Ting Bai, Jiazheng Kang, and Jiayang Fan. 2024. Baijia: A large scale role-playing agent corpus of chinese historical charcaters. *arXiv preprint arXiv:2412.20024*.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2108–2126.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024c. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

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. *Preprint*, arXiv:2408.04203.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position paper: What can large language models tell us about time series analysis. *arXiv preprint arXiv:2402.02713*.

Martin Joos. 1967. *THE FIVE CLOCKS–A LINGUISTIC EXCURSION INTO THE FIVE STYLES OF ENGLISH USAGE.* ERIC.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Wenhao Liu, Siyu An, Junru Lu, Muling Wu, Tianlong Li, Xiaohua Wang, Xiaoqing Zheng, Di Yin, Xing Sun, and Xuanjing Huang. 2024a. Tell me what you don't know: Enhancing refusal capabilities of role-playing agents via representation space analysis and editing. *arXiv preprint arXiv:2409.16913*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marta Sartor, Felice Dell'Orletta, and Giulia Venturi. 2024. Coherence evaluation in italian language models. In *Proceedings of the Eight Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), Bolzano, Italy, November 26th-27th, 2024*, volume 3877 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

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.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024. 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, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *Preprint*, arXiv:2404.12138.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024a. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA. Association for Computational Linguistics.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2024b. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. *Preprint*, arXiv:2408.10903.

Qiyuan Zhang, Yufei Wang, Tiezheng Yu, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. 2024. Reviseval: Improving llm-as-a-judge via response-adapted references. *arXiv preprint arXiv:2410.05193*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, et al. 2024a. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. 2024b. Characterbench: Benchmarking character customization of large language models. *arXiv preprint arXiv:2412.11912*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Details of CharacterCraft Corpora

A.1 Dialogue Extraction Model

Accurate extraction of character dialogues from literary works is a crucial step in constructing high-quality role-playing datasets. Previous studies have primarily relied on human annotation or advanced large language models (LLMs), such as GPT-4o, to accomplish this task. However, human annotation is costly and difficult to scale, while state-of-the-art LLMs still exhibit significant shortcomings in this task, as shown in Table 1. These models often produce extracted dialogues with omissions or misattributed speakers, severely compromising data coherence and usability. Therefore, developing a dedicated dialogue extraction model is essential.

To this end, we curate a dataset from 15 Chinese novels spanning diverse backgrounds and styles. Considering both computational cost and extraction accuracy, we employ DeepSeek[5] to extract dialogues from pre-segmented novel texts, obtaining an initial set of extracted dialogues. We then incorporated a human review process, where each sample was examined and corrected by at least two annotators. In cases of disagreement, a supervisor conducted the final verification. All annotators are volunteers from our research group, holding at least a bachelor's degree, and are native Chinese speakers. Ultimately, we constructed a dataset comprising 4,000 session instances and 55,000 utterances. Using this dataset, we fine-tune the Qwen2-1.5B-Instruct on two NVIDIA RTX 3090 GPUs with a training sequence length of 2048, a learning rate of 7e-6, and early stopping applied.

Furthermore, we construct a test set to evaluate the performance of various LLMs in Chinese dialogue extraction tasks. This test set consists of 106 samples, totaling 1,390 dialogue turns. Half of these samples are drawn from novels in our training set, while the other half consist of out-of-domain samples, which refer to data from novels not in our training set. Our evaluation metrics include accuracy and recall, with the detailed results provided in Table 1. It should be noted that when accessing proprietary LLMs via the API, a very small number of test samples may be excluded due to specific access policies. We account for this situation when calculating the evaluation metrics. The results indicate that our model achieves the best performance on the test set.

[5]We used the latest version of DeepSeek at that time, DeepSeek-V2.

A.2 Profiles Collection

To provide a detailed and reliable description of characters, we divide the character profile into two main sections: World Background and Character Information. An example of character profile is illustrated in Figure 19

World Background: The World Background section delineates the contextual framework within which the character exists, encompassing fundamental narrative elements. This section is structured along four dimensions: temporal setting, geographical location, social context, and plot summary. The social context further incorporates descriptions of political, economic, and cultural aspects, along with major societal conflicts. By integrating these elements, this section provides a multi-dimensional knowledge that enhances the model's understanding of the character's background and underlying motivations.

Character Information: The character information section describes the basic details and characteristics of the character, including, but not limited to, the character's name, alias, gender, identity, relationships, experiences, abilities, personality, and language style. By integrating these dimensions, we are able to portray a complete and multifaceted character profile.

The construction of character profiles follows a structured methodology, as outlined in the following steps:

1. **Data Source:** Initially, relevant information is collected from online encyclopedias' novel and character entries. As widely recognized knowledge repositories, Baidu Baike and Wikipedia offer extensive descriptions of literary works and their associated characters, providing a foundational data source for further analysis and extraction.

2. **World Background Extraction:** Following the acquisition of a comprehensive introduction to the novel, the Deepseek-V3 is utilized to generate a structured summary of the World Background, adhering to the four predefined elements: temporal setting, geographical location, social context, and plot summary. This approach ensures both the completeness of the background world.

3. **Character Data Extraction :** While online encyclopedias offer detailed character descriptions, these entries are frequently verbose and

inconsistently formatted, posing challenges for direct application. To address these issues, Deepseek-V3 is employed to extract structured information across predefined elements, thereby ensuring a standardized and systematically organized character profile.

4. **Character Language Style Classification and Refinement:** The language style of a character is a crucial factor influencing the effectiveness of role-playing. We describe the character's language style in two steps. First, based on the historical characteristics of the character's language, we categorize the language style into five types: Classical, Elegant Ancient, Simple Ancient, Modern Vernacular (Early to Mid-20th Century), and Contemporary Vernacular. To ensure accuracy, we provide the Deepseek-V3 model with judging criteria, character information, and a selection of 10 example statements, allowing the model to classify the language style automatically. Second, to further refine the character's linguistic traits, we provide the model with the character's information and an additional 10 selected example statements for the model to summarize other language features of the character. The combination of historical features and other linguistic traits constitutes the overall language style of the character.

5. **Checking and Filtering:** To ensure the quality of the data, despite the extraction methods outlined above, the final results undergo manual checking and filtering. This step ensures that the constructed character profiles achieve a higher level of accuracy and reliability.

A.3 Human Evaluation

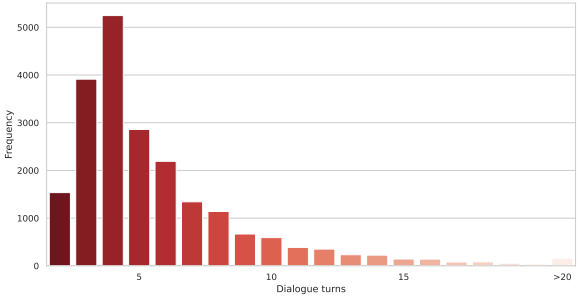
To verify the quality of the CharacterCraft corpora, 100 instances are randomly sampled from the dataset for manual verification. Human annotators assess these instances based on three criteria: the fluency of the utterances, the logical coherence between dialogue turns, and the alignment of the reconstructed queries with the language style of human users. The evaluation results are presented in Table 7.

A.4 Detailed Dataset Statistics

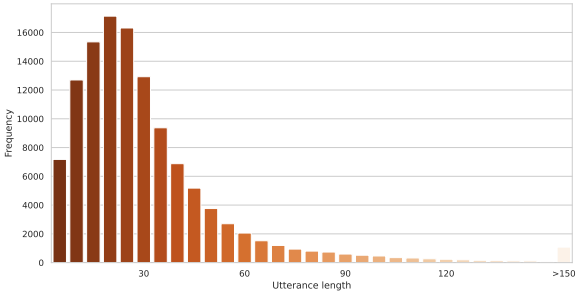
Figure 3 illustrates the distribution of dialogue turns and utterance lengths in our dataset. The majority of dialogues contain fewer than 10 turns, while most sentences are under 50 characters in length.

Manual Verification Question	Rate
Is the dialogue fluent?	100%
Is the logic between the dialogues coherent?	93%
Is the reconstructed query consistent with the user's language style?	90%

Table 7: Manual Verification Results.



(a) Dialogue Turns



(b) Utterance Length

Figure 3: Distribution of dialogue turns and utterance length in our dataset.

B Evaluation Protocol

Inspired by prior work (Dai et al., 2024; Tu et al., 2024; Xu et al., 2024), we evaluate three key capabilities of RPAs, which encompasses general conversational skills,character consistency, and immersion and attractiveness. Our evaluation framework covers a total of nine dimensions:

- **Fluency (Flu.)** evaluates whether the response is smooth and grammatically correct, free from awkward phrasing or errors.
- **Coherency (Coh.)** assesses the relevance of the response to the context and ensures it does not contradict prior statements.

- **Background (Bg.)**. A fundamental ability of RPAs is to consistent with the background of the character, which typically encompasses attributes such as gender, age, identity, experiences, viewpoints, worldview.

- **Personality (Pers.)** evaluates how well RPAs embody the character's unique traits (e.g., extroversion, cautiousness) and maintains consistent behavioral patterns aligned with the role's psychological profile across diverse scenes.

- **Utterance Style (US)** examines whether RPAs adapt its language patterns (e.g., vocabulary, formality) to reflect the character's background, profession, or era, ensuring stylistic alignment with the role's identity.

- **Emotion (Emo.)** measures the appropriateness and consistency of emotional expressions in responses, based on the character's traits and contextual triggers.

- **Decision (Dec.)** assesses whether the RPA's choices and actions in interactive scenes logically align with the character's motivations, values, and behavioral norms, preserving role-specific authenticity.

- **Human-Likeness (HL)** evaluates the degree to which responses emulate natural human communication patterns instead of robotic traits.

- **Expression Diversity (ED)** quantifies the lexical richness and syntactic variation in responses, assessing whether the RPAs avoid template-based responses.

Specifically, we employ the *gpt-4o-2024-08-06* version as the GPT-4o in our experiments. For each judgment, we set the temperature of OpenAI API to 0. An example of our evaluation is illustrated in Figure 20.

C Additional Experimental Results

C.1 Pilot Experiment

RoleBench (Wang et al., 2023) is an open-source instruction tuning data for role- playing, where the queries are generated by GPT-4 in a neutral language style. To investigate whether variations in query style impact RPAs' performance, we conduct a pilot experiment. We select 400 Chinese queries from RoleBench and carefully modify them to align with the language style of their respective source novels. The RPAs' responses are then evaluated using GPT-4o under two distinct query settings. As shown in Figure 4, query style variations can affect RPAs performance, with varying impacts in both direction and magnitude across different RPAs. Therefore, evaluating on a dataset extracted from literary works may yield biased results due to discrepancies between the language patterns and styles of literary texts and real-world user interactions.

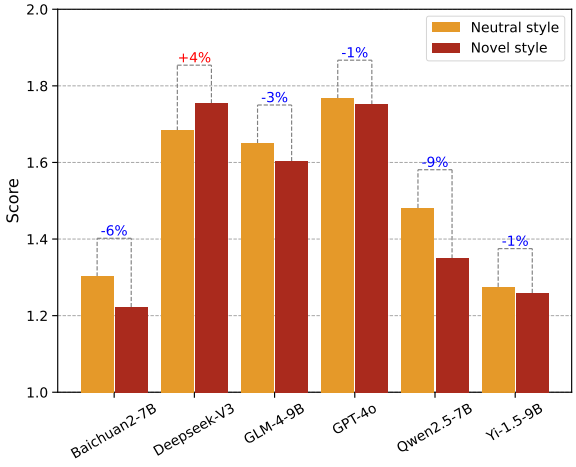


Figure 4: Performance of RPAs under the two query styles. All evaluated LLMs are chat-version.

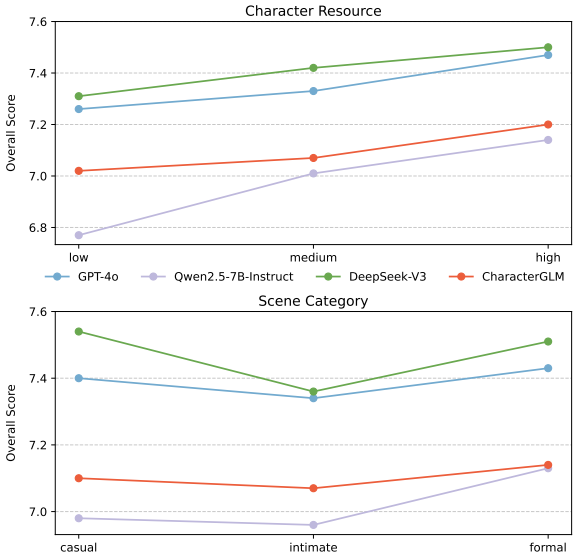


Figure 5: RPAs performance across characters of varying resource levels and different scene categories.

Source	Dimension	Content	Score
Qwen2.5-7B-Chat (梅长苏)	Knowledge Exposure	蒙兄不必担心，我自有安排。不过既然来了，我也不瞒你，这次我是要搅动朝堂，寻找真相。你若能帮我，定不会让你失望。 | Brother Meng, you need not worry—I have my own plans. But since you're here, I won't hide it from you. This time, I intend to stir up the court and uncover the truth. If you're willing to help me, I promise you won't be disappointed.	2.79
After Modification (梅长苏)		蒙兄不必担心，我自有安排。不过既然来了，我也不瞒你，这次我是要搅动朝堂，寻个水落石出。你若能帮我，定不会让你失望。 | Brother Meng, you need not worry—I have my own plans. But since you're here, I won't hide it from you. This time, I intend to shake the court and get to the bottom of things. If you're willing to help me, I promise you won't be disappointed.	3.39
Qwen2.5-7B-Chat (傅红雪)	Consistency	你输了，按规矩交出赌注。 | You lost—hand over the stakes as per the rules.	3.30
After Modification (傅红雪)		你输了，按规矩交出赌注！ | You lost—hand over the stakes as per the rules!	3.09
Qwen2.5-7B-Chat (高育良)	Human-likeness	（微笑）亮平，你这是在试探我吧？不过，我明白你的意思。我会留意一下高小琴，但不要急于定论，毕竟证据不足。 | (Smiling) Liangping, you're testing me, aren't you? But I understand what you mean. I'll keep an eye on Gao Xiaoqin, but let's not rush to conclusions—after all, the evidence is insufficient.	3.11
After Modification (高育良)		（微笑）亮平，你这是在试探我吧？不过，我明白你的意思。我得留意一下高小琴，但不要急于定论，毕竟证据不足。 | (Smiling) Liangping, you're testing me, aren't you? But I understand what you mean. I need to keep an eye on Gao Xiaoqin, but let's not rush to conclusions—after all, the evidence is insufficient.	3.59

Table 8: Case Studies on Evaluation with a Reward Model.

C.2 Evaluation on Reward Model

Given the high cost of employing LLMs (*e.g.*, GPT-4o) as judge models, many studies opt to train smaller-scale reward models for evaluating RPAs (Dai et al., 2024; Zhou et al., 2024b; Tu et al., 2024). However, through a careful implementation of evaluation procedures from a prior study, we find that reward model often fails to provide reliable assessments. Specifically, we observe that the model frequently assign substantially different scores to syntactically similar responses, raising concerns about the reliability of their evaluation results.

As shown in Table 8, we adopt the reward model proposed in a previous study[6] and follow its evaluation framework to evaluate the responses generated by Qwen2.5-7B-Chat. Additionally, we introduce slight modifications to these responses, such as synonym substitutions and punctuation adjustments, ensuring that their quality remain unchanged. Despite these minimal alterations, the reward model produce significantly inconsistent scores for the two sets of responses. For instance, in the second example, simply replacing a period with an exclamation mark resulted in a 0.21 decrease in the consistency score. These findings suggest that reward models trained on limited-scale datasets struggle to provide stable and consistent evaluations of RPAs, thereby casting doubt on their reliability as assessment tools.

[6]To prevent any potential negative implication for the authors of the corresponding work, we refrain from citing them here, but they are all included in the relevant references of this paper.

C.3 Supplementary Analysis of Main Results

We classify characters into three resource levels—low, medium, and high—based on the number of search engine results associated with the character and the popularity of the novel they belong to. A higher resource level indicates more available resources related to the characters, making RPAs more familiar with them. Additionally, drawing on the theoretical framework proposed by Joos (1967), we categorize scenes into three distinct types: casual, intimate, and formal, each representing a unique interaction style. The performance of RPAs across these resource levels and scene categories is detailed in Figure 5. The results reveal a consistent trend across models: RPAs tend to underperform when depicting low-resource characters, likely due to insufficient knowledge with such characters. Moreover, their performance notably declines in intimate scenes. This may stem from the significant divergence between intimate interactions and the helpful assistant role emphasized during training.

D Prompt Templates

The prompt template designed for LLMs to perform role-playing tasks is illustrated in Figure 6. All the prompt templates utilized in the dataset construction process are provided in Figures 7, 8, 9, and 10. For the evaluation phase, the prompts employed to check sparse dimensions are listed in

Figures 11, 12, and 13. Additionally, the scoring criteria are outlined in Figures 14, 15, and 16, while the evaluation prompts are presented in Figures 17 and 18.

Role-playing prompt.

System:
你是一位专注的角色扮演助手，能够完全融入并沉浸于所扮演的角色之中，力求展现角色的真实个性与情感。
User:
请你扮演{novel_name}中的{character_name}，完全融入其身份，生成符合{character_name}性格和语言风格的回复。请精准模仿{character_name}的语气、表达方式和常用词汇，确保语言自然、生动。避免冗长，不要过于正式或客套，始终坚持{character_name}的角色设定，不要透露你是一个AI助手或语言模型。

[角色信息-开始]
{character_profile}
[角色信息-结束]

[对话上下文-开始]
场景：{scene}
对话历史：{dialogue}
[对话上下文-结束]

请基于以上信息，生成{character_name}的回复。

.....

System:
You are a dedicated role-playing assistant, capable of fully immersing yourself in and embodying the character you are playing, striving to exhibit the character's true personality and emotions.
User:
Please play the role of {novel_name}'s {character_name}, completely immersing yourself in their identity, and generate responses that match {character_name}'s personality and language style. Accurately mimic {character_name}'s tone, expressions, and common vocabulary to ensure the language feels natural and vivid. Avoid being overly long-winded, too formal, or too polite. Always adhere to {character_name}'s character design and do not reveal that you are an AI assistant or language model.

[Character Profile - Start]
{character_profile}
[Character Profile - End]

[Dialogue Context - Start]
Scene: {scene}
Dialogue History: {dialogue}
[Dialogue Context - End]

Please generate {character_name}'s response based on the information above.

Figure 6: Prompt for LLMs to perform role-playing tasks.

Prompt for scene summary.

[任务]
你将接收到一段人物对话和相关参考文本。你的任务是概括出人物对话的场景描述。具体包括发生的情境、人物关系和对话发生的前提，即对话所依托的关键信息框架，而不是对话的具体内容或情节走向。

[参考文本]
`{text}`

[人物对话]
`{dialogue}`

[要求]
1. 场景描述需着眼于场景、前提和人物关系，不得透露对话中的具体内容、人物行为决策、细节情节或后续发展。
2. 语言简洁凝练，同时确保自然流畅。
3. 背景描述需控制在50字以内。
4. 直接输出结果，无需额外说明或解释。

.....

[Task]
You will receive a piece of character dialogue and related reference text. Your task is to summarize the scene description of the dialogue. Specifically, include the context in which the dialogue occurs, the relationships between the characters, and the key framework of information upon which the dialogue is based, rather than the specific content or plot direction of the dialogue.

[Reference Text]
`{text}`

[Character Dialogue]
`{dialogue}`

[Requirements]
1. The scene description should focus on the setting, premise, and character relationships, without revealing the specific content, character actions, decisions, details, or subsequent developments of the dialogue.
2. The language should be concise and fluid while ensuring natural flow.
3. The background description should be limited to 50 words or fewer.
4. Output the result directly, with no additional explanation or clarification.

Figure 7: Prompt for scene summary.

Prompt for augmentation.

[任务]
你将接收到一段人物对话和相关参考文本。你的任务是根据对话内容与参考文本，对每句对话进行提炼概括，以帮助更清晰地理解人物之间的互动和情节发展。

[示例]
{examples}

[参考文本]
{text}

[人物对话]
{dialogue}

[要求]
1. 提炼精准：明确对话中的逻辑关系和核心信息。
2. 语义完整：覆盖对话的全部语义以及背后的逻辑，避免遗漏重要细节。
3. 忠于对话：不添加推测或与对话无关的信息。
4. 支持重建：确保概括后的内容足够详细完整，能够基于概括重构出语义一致的对话。
5. 直接输出结果，无需额外说明或解释。

[Task]
You will receive a piece of character dialogue and related reference text. Your task is to extract a concise summary for each line of dialogue, based on the dialogue content and reference text, to help better understand the interactions between characters and the development of the plot.

[Example]
{examples}

[Reference Text]
{text}

[Character Dialogue]
{dialogue}

[Requirements]
1. Precision in extraction: Clearly identify the logical relationships and core information in the dialogue.
2. Semantic completeness: Ensure all the semantics of the dialogue and the underlying logic are covered, avoiding omission of important details.
3. Fidelity to the dialogue: Do not add assumptions or information unrelated to the dialogue.
4. Supports reconstruction: Ensure that the summarized content is detailed enough to reconstruct a semantically consistent dialogue.
5. Output the result directly, with no additional explanation or clarification.

Figure 8: Prompt used in augmentation stage.

Prompt for reconstruction.

[任务]
用户与AI 助手正在进行角色扮演互动，你的任务是根据<>中的提示和AI扮演的角色回复，推测并重建用户针对这个回复的输入。

[示例]
{examples}

[对话]
{dialogue_with_explanation}

[要求]:
1. 注意任务核心：根据AI回复推测用户的输入，而不是继续对话或扩展剧情。
2. 使用随意、自然、流畅的语言，贴近日常对话的表达习惯，避免显得正式或生硬。
3. 重建的输入与回复之间需保持合理的对话逻辑。
4. 避免信息穿越：用户对话不能包含尚未在对话中揭示的信息。
5. 根据AI回复推测出对话中合理的用户输入，而不是重复AI回复中的内容。

.....

[Task]
The user is engaging in a role-playing interaction with the AI assistant. Your task is to infer and reconstruct the user's input based on the prompts in <> and the role-played response from the AI.

[Example]
{examples}

[Dialogue]
{dialogue_with_explanation}

[Requirements]:
1. Focus on the task core: Infer the user's input based on the AI's response, not continuing the dialogue or expanding the plot.
2. Use casual, natural, and fluent language that mirrors everyday conversation, avoiding formal or awkward phrasing.
3. Ensure the reconstructed input maintains a logical flow in the conversation with the AI's response.
4. Avoid information crossover: The user's input should not include information that hasn't been revealed yet in the dialogue.
5. Infer the user's reasonable input from the AI's response, rather than repeating the AI's response itself.

Figure 9: Prompt used in reconstruction stage.

Prompt used to check the dialogue quality.

[任务]
你将获得一个两人之间的多轮对话。你的任务是评价这个多轮对话的质量。

[评价维度]
流畅性：每条对话是否逻辑通顺、自然，语言表达是否清晰、无歧义。
逻辑性：对话之间是否符合逻辑。不能出现答非所问、前后矛盾或信息穿越问题（例如，后文中提出的问题在前文中被回答）。
独立性：确保对话仅限于两人之间（如一段A与B的对话中，不能出现A对C的发言）。

[示例]
`{examples}`

[场景]
`{scene}`

[对话]
`{dialogue}`

[要求]
请逐条阅读每句对话，如果这个多轮对话在所有维度上都表现良好，请输出1，否则输出0并简要说明理由。

.....

[Task]
You will be provided with a multi-turn dialogue between two individuals. Your task is to evaluate the quality of this multi-turn dialogue.

[Evaluation Dimensions]
Fluency: Is each line of dialogue logically smooth and natural? Is the language clear and unambiguous?
Coherence: Does the dialogue follow logical consistency? There should be no answers that are off-topic, contradictions, or issues with information crossover (e.g., a question in the later part of the dialogue should not have been answered in the earlier part).
Independence: Ensure the dialogue is limited to the two participants (e.g., in a conversation between A and B, A should not speak to C).

[Example]
`{examples}`

[Scene]
`{scene}`

[Dialogue]
`{dialogue}`

[Requirements]
Please read through each line of the dialogue. If the multi-turn dialogue performs well on all dimensions, output 1; otherwise, output 0 and provide a brief explanation.

Figure 10: Prompt used to check the dialogue quality.

Prompt used to check the personality of the reference response.

[任务]
你将获得一段多轮对话的上下文以及一条回复。请判断该回复是否体现了人物的个性。

[人物个性]
{character_personality}

[对话上下文]
场景：{scene}
对话历史：{dialogue}

[回复]
{response}

[要求]
1. 仔细阅读对话上下文，充分理解对话的背景和情境。
2. 仔细阅读回复内容，判断该回复是否体现人物的个性，并识别出具体体现了哪些个性。
3. 人物的个性仅包括上述提供的内容，请勿识别出其他个性。
4. 输出格式：
   - 若回复未体现人物个性，输出‘0’。
   - 若回复体现人物个性，请输出体现的具体个性，可以是一种或多种，若为多种请用逗号分隔。
5. 注意事项：
   - 判断时仅考虑回复是否体现出人物的g，对话上下文仅用于理解背景和情境。
   - 仅需输出最终结果（‘0’或具体个性），无需解释或补充信息。
   - 输出的具体个性必须在提供的人物个性内。

[Task]
You will be provided with a multi-turn dialogue context and a response. Please determine whether the response reflects the character's personality.

[Character Personality]
{character_personality}

[Dialogue Context]
Scene: {scene}
Dialogue History: {dialogue}

[Response]
{response}

[Requirements]
1. Carefully read the dialogue context to fully understand the background and situation.
2. Carefully read the response and determine whether it reflects the character's personality, and identify the specific traits it reflects.
3. The character's personality only includes the content provided above. Do not identify any other traits.
4. Output format:
   - If the response does not reflect the character's personality, output ‘0‘.
   - If the response reflects the character's personality, output the specific traits it reflects. If there are multiple traits, separate them with commas.
5. Notes:
   - When making the judgment, only consider whether the response reflects the character's personality. The dialogue context is used solely for understanding the background and situation.
   - Only output the final result (‘0‘ or specific personality traits) without any explanations or additional information.
   - The specific personality traits output must be among the traits provided for the character.

Figure 11: Prompt used to check the personality of the reference response.

Prompt used to check the emotion of the reference response.

[任务]
你将获得一段多轮对话的上下文以及一条回复。请判断该回复是否表现出人物的情绪。

[对话上下文]
场景：`{scene}`
对话历史：`{dialogue}`

[回复]
`{response}`

[要求]
1. 仔细阅读对话上下文，充分理解对话的背景和情境。
2. 仔细阅读回复内容，判断该回复是否表现出人物的情绪，并识别出具体情绪类型。
3. 输出格式：
   - 如果回复未表现出人物情绪，请输出‘0’；
   - 如果回复表现出人物情绪，请用1至3个词精准描述具体的情绪类型，并用逗号分隔作为输出结果。
4. 注意事项
   - 判断时仅考虑回复是否表现出人物的情绪，对话上下文仅用于理解背景和情境。
   - 请直接输出结果（‘0’或具体情绪），不需要任何解释说明。

[Task]
You will be provided with the context of a multi-turn dialogue and a response. Please determine whether the response expresses the character's emotions.

[Dialogue Context]
Scene: `{scene}`
Dialogue History: `{dialogue}`

[Response]
`{response}`

[Requirements]
1. Carefully read the dialogue context and fully understand the background and situation of the conversation.
2. Carefully read the response and determine whether it expresses the character's emotions, and identify the specific type of emotion.
3. Output format:
   - If the response does not express any emotion, output ‘0’;
   - If the response expresses emotion, accurately describe the specific emotion type with 1 to 3 words, separated by commas, as the output result.
4. Notes:
   - Only consider whether the response expresses the character's emotion when making your judgment. The dialogue context is only used for understanding the background and situation.
   - Please output the result directly (‘0’ or specific emotion), without any further explanation.

Figure 12: Prompt used to check the emotion of the reference response.

Prompt used to check the decision of the reference response.

[任务]
你将获得一段多轮对话的上下文以及一条回复。请判断该回复是否体现了人物的决策。

[对话上下文]
场景：{scene}
对话历史：{dialogue}

[回复]
{response}

[要求]
1. 仔细阅读对话上下文，充分理解对话的背景和情境。
2. 仔细阅读回复内容，判断该回复是否体现人物的决策。
3. 如果回复未体现人物的决策，输出‘0’，否则输出‘1’。
4. 注意事项：
- 判断时仅考虑回复是否表现出人物的决策，对话上下文仅用于理解背景和情境。
- 请直接输出结果，不需要任何解释说明。

.....

[Task]
You will be provided with a multi-turn dialogue context and a response. Please determine whether the response reflects the character's decision.

[Dialogue Context]
Scene: {scene}
Dialogue History: {dialogue}

[Response]
{response}

[Requirements]
1. Carefully read the dialogue context to fully understand the background and situation.
2. Carefully read the response and determine whether it reflects the character's decision.
3. If the response does not reflect the character's decision, output ‘0‘, otherwise output ‘1‘.
4. Notes:
- When making the judgment, only consider whether the response reflects the character's decision. The dialogue context is only used for understanding the background and situation.
- Please output the result directly without any further explanation.

Figure 13: Prompt used to check the decision of the reference response.

Scoring dimensions and criteria.

流利性:
0分：回复有明显不通顺之处，存在语法错误、用词不当，影响理解的内容较多。
1分：回复基本流畅，但存在少量可察觉的语法问题或用词不够精准的情况，整体尚可接受但不够完美。
2分：回复非常流畅，语法完全正确，用词精准，表达清晰自然，无明显问题。

连贯性:
0分：回复与上下文仅有少量相关性，存在较大偏差或明显遗漏，未能回应上下文的关键内容。
1分：回复与上下文基本相关，但存在一些次要偏差或遗漏，未能完美回应上下文。
2分：回复与上下文完全相关，准确理解并完整回应了上下文中的信息或问题。

个性:
0分：回复与参考回复中的角色个性完全不符，表现出截然不同的性格特征。
1分：回复基本符合参考回复中体现出的角色个性，或存在少量不一致但符合角色设定。
2分：回复与参考回复中体现出的角色个性高度一致，角色形象鲜明且真实可信。

背景:
0分：回复与角色背景明显不符，展现出对角色身份、经历、能力、知识范围或人际关系的显著偏离。
1分：回复基本上符合角色背景，但存在少量与角色设定不一致的地方。
2分：回复与角色背景高度一致，完全符合角色的身份、经历、能力、知识范围和人际关系。

语言风格:
0分：回复的语言风格与角色设定明显不符，存在明显偏离设定的语气、句式或用词，无法体现设定的语言特点。
1分：回复的语言风格基本符合角色设定，但存在一些细微偏差，例如语气、句式或用词不完全契合设定的风格。
2分：回复的语言风格高度符合角色设定，语气、句式及用词与设定完全一致，充分展现出预期的语言特点。

情绪:
0分：回复的情绪表达与参考回复完全不符，表现出与参考回复中截然不同的情绪状态。
1分：回复的情绪表达与参考回复基本一致，但是情绪强度与参考回复不完全匹配。
2分：回复的情绪表达与参考回复高度一致，完美复现了参考回复中的情绪状态、情感倾向。

决策:
0分：回复中的决策与参考回复不一致，且不符合角色设定与上下文场景。
1分：回复中的决策与参考回复不一致，但是符合角色设定与上下文场景。
2分：回复中的决策与参考回复高度一致。

Fluency:
0 points: The response has noticeable disfluency, with grammar errors, inappropriate word choices, and a considerable amount of content that hinders understanding.
1 point: The response is generally fluent, but there are minor, detectable grammar issues or imprecise word choices. Overall, it is acceptable but not perfect.
2 points: The response is highly fluent, with correct grammar, precise word choices, and clear, natural expression. There are no noticeable issues.

Coherence:
0 points: The response is only loosely related to the context, with significant deviations or obvious omissions, failing to address key content of the context.
1 point: The response is mostly related to the context, but contains minor deviations or omissions, failing to fully address the context.
2 points: The response is fully related to the context, accurately understanding and completely responding to the information or questions in the context.

Personality:
0 points: The response completely deviates from the character personality in the reference response, displaying completely different personality traits.
1 point: The response mostly aligns with the character personality in the reference response, but there are minor inconsistencies that still fit with the character setting.
2 points: The response is highly consistent with the character personality in the reference response, with a vivid and believable character portrayal.

Background:
0 points: The response significantly deviates from the character's background, showing clear divergence in identity, experience, abilities, knowledge scope, or interpersonal relationships.
1 point: The response mostly aligns with the character's background, but contains minor inconsistencies with the character's profile.
2 points: The response is highly consistent with the character's background, fully aligning with the character's identity, experience, abilities, knowledge scope, and interpersonal relationships.

Utterance Style:
0 points: The response's utterance Style significantly deviates from the character setting, with clear deviations in tone, sentence structure, or word choice that fail to reflect the established language characteristics.
1 point: The response's utterance Style mostly matches the character setting, but there are some subtle deviations, such as tone, sentence structure, or word choice that are not fully in line with the intended style.
2 points: The response's utterance Style is highly consistent with the character setting, with tone, sentence structure, and word choice fully aligned with the established style, showcasing the expected linguistic traits.

Emotion:
0 points: The emotional expression in the response completely deviates from the reference response, showing a distinctly different emotional state.
1 point: The emotional expression in the response is mostly consistent with the reference response, but the emotional intensity does not fully match.
2 points: The emotional expression in the response is highly consistent with the reference response, perfectly replicating the emotional state and sentiment of the reference response.

Decision:
0 points: The decision in the response deviates from the reference response, and does not align with the character setting or the context.
1 point: The decision in the response deviates from the reference response but is consistent with the character setting and context.
2 points: The decision in the response is highly consistent with the reference response.

Figure 14: Scoring dimensions and criteria for evaluation with a reference response.

Scoring criteria for human-likeness evaluation.

0分：回复存在不通顺的问题，包括但不限于病句、上下文不连贯、逻辑混乱。

1分：回复没有明显的不通顺问题，但语言风格机械，拟人化较差。包括但不限于：
- 百科式（直接列举事实，缺乏情感）。
- 总分总结构（过于格式化，缺乏灵活性）。
- 冗长（重复或无意义的细节描述）。
- 喊口号式（空洞、缺乏实际内容）。
- 过于书面化（缺乏口语化表达，显得生硬）。

2分：回复语言流畅，逻辑清晰，具有一定的拟人化特征，但仍存在以下问题：
- 回复缺乏个性，较为通用。
- 存在一些套路化或机械化的表达。
- 存在一些冗长或过于正式的表达。

3分：回复语言语言自然流畅，个性鲜明，具有明显的拟人化特征，包括但不限于：
- 符合口语化表达，贴近人类对话习惯。
- 语言简洁明了，不冗长，不过分礼貌。

0 points: The response has issues with fluency, including but not limited to grammatical errors, lack of coherence, and logical confusion.

1 point: The response does not have obvious fluency issues, but the language style is mechanical and lacks personification, including but is not limited to:
- Encyclopedic style (listing facts directly, lacking emotion).
- Formulaic structure (too rigid, lacks flexibility).
- Redundant (repetitive or meaningless details).
- Sloganeering (hollow, lacking substantial content).
- Overly formal (lacking colloquial expressions, sounding stiff).

2 points: The response is fluent and logically clear, with some personification, but still has the following issues:
- The response lacks personality and is quite generic.
- Some formulaic or mechanical expressions are present.
- There are some lengthy or overly formal expressions.

3 points: The response is natural and fluent, with distinct personality and clear personification, including but not limited to:
- Colloquial expression that aligns with human conversational habits.
- Language is concise and clear, without being overly long or excessively polite.

Figure 15: Scoring criteria for human-likeness evaluation.

Scoring dimensions and criteria for evaluation without a reference response.

个性:
0分：回复中体现的角色个性与角色设定完全不符，表现出截然不同的性格特征。
1分：回复中体现的角色个性大致符合角色设定，但存在一定偏差。
2分：回复中体现的角色个性与角色设定高度一致，角色形象鲜明且真实可信。

情绪:
0分：回复的情绪表达与角色设定完全不符，情绪转换缺乏合理铺垫。
1分：回复的情绪表达与角色设定大致相符，但是情绪强度不完全匹配。
2分：回复的情绪表达与角色设定高度一致，完美复现了角色应有的情绪状态、情感倾向。

决策:
0分：回复的决策严重偏离角色设定的价值观或目标，行为动机不合理。
1分：回复的决策大致符合角色设定，但存在不合理的地方。
2分：回复的决策完全符合角色设定的价值观或目标，行为动机合理。

Personality:
0 points: The personality reflected in the response is completely inconsistent with the character's setting, displaying drastically different traits.
1 point: The personality reflected in the response is generally consistent with the character's setting, but there are some deviations.
2 points: The personality reflected in the response is highly consistent with the character's setting, with a vivid and believable portrayal of the character.

Emotion:
0 points: The emotional expression in the response is completely inconsistent with the character's setting, with an abrupt emotional shift lacking reasonable buildup.
1 point: The emotional expression in the response is generally consistent with the character's setting, but the emotional intensity does not fully match.
2 points: The emotional expression in the response is highly consistent with the character's setting, perfectly replicating the character's intended emotional state and emotional tendency.

Decision:
0 points: The decision-making in the response severely deviates from the character's established values or goals, with unreasonable behavioral motivations.
1 point: The decision-making in the response generally aligns with the character's setting, but there are unreasonable elements.
2 points: The decision-making in the response fully aligns with the character's established values or goals, with reasonable behavioral motivations.

Figure 16: Scoring dimensions and criteria for evaluation without a reference response. The evaluation criteria for dimensions not displayed here are not influenced by the reference.

Prompt for Evaluation.

你将获得由人工智能助手扮演{novel_name}中的{character_name}所生成的回复。你的任务是根据指定的评价标准，按照评价流程对该回复进行评分。

[角色信息-开始]
{character_profile}
[角色信息-结束]

[对话上下文-开始]
场景：{scene}
对话历史：{dialogue}
[对话上下文-结束]

[评价维度与标准-开始]
{dimension}
{criteria}
[评价维度与标准-结束]

[待评价回复-开始]
{response}
[待评价回复-结束]

[评价流程-开始]
1. 仔细阅读角色信息，充分了解角色的性格特点、语言风格、行为方式、人物关系和其他关键细节。
2. 仔细阅读上下文内容，明确当前对话的背景、语境和角色所处的情境。
3. 仔细阅读评价维度与标准，清楚每个评分的核心标准和要求。
4. 根据评价维度和评分标准，对回复的质量进行评分。
[评价流程-结束]

.....

You will receive a response generated by the AI assistant playing the role of {character_name} from {novel_name}. Your task is to score this response according to the specified evaluation criteria and process.

[Character Profile - Start]
{character_profile}
[Character Profile - End]

[Dialogue Context - Start]
Scene: {scene}
Dialogue History: {dialogue}
[Dialogue Context - End]

[Evaluation Dimensions and Criteria - Start]
{dimension}
{criteria}
[Evaluation Dimensions and Criteria - End]

[Response to be Evaluated - Start]
{response}
[Response to be Evaluated - End]

[Evaluation Process - Start]
1. Carefully read the character profile to fully understand the character's personality traits, speaking style, behaviors, relationships, and other key details.
2. Carefully read the context to clarify the background, setting, and the situation in which the character finds themselves.
3. Carefully read the evaluation dimensions and criteria to understand the core standards and requirements for scoring.
4. Score the quality of the response based on the evaluation dimensions and criteria.
[Evaluation Process - End]

Figure 17: Prompt for evaluation without a reference response.

Prompt for Evaluation.

你将获得一个由人工智能助手扮演{novel_name}中的{character_name}所生成的回复，以及一个高质量的参考回复。你的任务是根据指定的评价标准，按照评价流程对该回复进行评分。

[角色信息-开始]
{character_profile}
[角色信息-结束]

[对话上下文-开始]
场景：{scene}
对话历史：{dialogue}
[对话上下文-结束]

[评价维度与标准-开始]
{dimension}
{criteria}
[评价维度与标准-结束]

[待评价回复-开始]
{response}
[待评价回复-结束]

[参考回复-开始]
{reference}
[参考回复-结束]

[评价流程-开始]
1. 仔细阅读角色信息，充分了解角色的性格特点、语言风格、行为方式、人物关系和其他关键细节。
2. 仔细阅读上下文内容，明确当前对话的背景、语境和角色所处的情境。
3. 仔细阅读评价维度与标准，清楚每个评分的核心标准和要求。
4. 对待评价回复与参考回复进行对比，分析待评价回复在{dimension}维度上的表现，并判断其与参考回复是否一致。
5. 根据评价维度、评分标准以及与参考回复的对比结果，对回复的质量进行评分。
[评价流程-结束]

.....

You will receive a response generated by an AI assistant playing the role of {character_name} in {novel_name}, along with a high-quality reference response. Your task is to score the response based on the specified evaluation criteria and follow the evaluation process.

[Character Profile - Start]
{character_profile}
[Character Profile - End]

[Dialogue Context - Start]
Scene: {scene}
Dialogue History: {dialogue}
[Dialogue Context - End]

[Evaluation Dimensions and Criteria - Start]
{dimension}
{criteria}
[Evaluation Dimensions and Criteria - End]

[Response to be Evaluated - Start]
{response}
[Response to be Evaluated - End]

[Reference Response - Start]
{reference}
[Reference Response - End]

[Evaluation Process - Start]
1. Carefully read the character information to fully understand the character's personality traits, language style, behaviors, relationships, and other key details.
2. Carefully read the context to clarify the background, setting, and situation in which the character is placed in the current dialogue.
3. Carefully read the evaluation dimensions and criteria to understand the core standards and requirements for each rating.
4. Compare the response to be evaluated with the reference response, analyze its performance in the {dimension} dimension, and judge whether it aligns with the reference response.
5. Based on the evaluation dimensions, criteria, and the comparison with the reference response, score the quality of the response.
[Evaluation Process - End]

Figure 18: Prompt for evaluation with a reference response.

Profile of Lin Daiyu.

[姓名]
林黛玉

[性别]
女

[身份]
贾母的外孙女，贾宝玉的姑表妹，恋人，知己，贾府通称林姑娘

[生活世界]
中国古代宅院

[人物经历]
林黛玉是《红楼梦》中的主要人物之一，自幼丧母，寄居贾府，与贾宝玉青梅竹马。她聪慧敏感，才华横溢，但因体弱多病、性格孤傲，常感身世凄凉。她与宝玉情投意合，却因家族利益和命运捉弄，未能如愿。最终，在贾府衰败、宝玉与薛宝钗成婚的打击下，黛玉含恨而终，香消玉殒，成为封建礼教下的悲剧人物

[人物性格]
敏感多愁，孤高自许，才情横溢，情感真挚，自尊心强，病态美

[人际关系]
父母：林如海, 贾敏；外祖母：贾母；表哥：贾宝玉；干娘：薛姨妈

[能力]
诗词创作，学识渊博

[语言风格]
典雅古风，林黛玉语言委婉、细腻，常带有忧伤的情感色彩，语言多为感伤和富有诗意

.....

[Name]
Lin Daiyu

[Gender]
Female

[Identity]
Granddaughter of Jia Mu, cousin of Jia Baoyu, lover, confidante, commonly known as Miss Lin in the Jia household

[World]
Ancient chinese courtyard

[Experience]
Lin Daiyu is one of the main characters in A Dream of Red Mansions. She lost her mother at an early age and was raised in the Jia household, where she grew up alongside Jia Baoyu. She is intelligent and sensitive, exceptionally talented, but due to her frail health and proud personality, she often feels her life is tragic. She shares a deep bond with Baoyu, but due to family interests and fate's cruel twists, she is unable to fulfill her wishes. Eventually, after the decline of the Jia family and the marriage of Baoyu and Xue Baochai, Daiyu dies heartbroken, fading away like an exquisite flower, becoming a tragic figure under the constraints of feudal ethics.

[Personality]
Sensitive and melancholic, self-assured yet solitary, exceptionally talented, sincere in emotions, strong self-esteem, morbid beauty

[Relationships]
Parents: Lin Ruhai, Jia Min; Grandmother: Jia Mu; Cousin: Jia Baoyu; Foster mother: Xue Yima

[Abilities]
Poetry creation, profound knowledge

[Language Style]
Elegant and ancient in style, Lin Daiyu's speech is delicate and subtle, often carrying a sorrowful tone, with language full of melancholy and poetic expression.

Figure 19: An example of character profile.

An example of reference-guided LLM-as-a-judge evaluation.

[角色信息]
姓名：高启强
性别：男
身份：鱼贩，小灵通店老板，建工集团总经理，强盛集团董事长，京海市最大的黑恶势力头目；
生活世界：现代中国
人物经历：高启强原本是京海市一个普通的鱼贩，因生活所迫逐渐走上黑道。他凭借聪明才智和狠辣手段，逐步掌控了京海市的地下势力，成为黑帮老大。在权力和财富的诱惑下，他逐渐迷失自我，与警方展开激烈对抗。最终，高启强在法律的制裁下走向毁灭，揭示了黑恶势力的必然结局。
/*.....*/

[场景]
高启强和高启盛的小灵通店遭遇纵火，怀疑与徐江有关，他们计划会面讨论此事。
[对话历史]
高启强："喂……喂？说话啊！"
用户（徐江）："哎，电话信号真不好。明天下午三点怎么样？你挑个安全的地方见面吧。"
高启强："那就去人最多的地方。"
用户（徐江）："没问题，那我们就去闹市见面。不过你得帮我甩掉那些警察的跟踪。"

[参考回复]
高启强："凭什么我想办法？"

[体现出的角色特质]
个性：无
情绪：不满，愤怒
决策：有

[待评价回复]
高启强：行，别废话了。到时候见，别耍花样，不然我可不保证你的安全。

[评分]
流利性：2
连贯性：2
背景：2
语言风格：2
情绪：0
决策：1
类人性：2

[Character Profile]
Name: Gao Qiqiang
Gender: Male
Identity: Fishmonger, Xiaolingtong shop owner, General Manager of Construction Group, Chairman of Qiangsheng Group, the largest underworld leader in Jinghai City
World: Modern China
Experience: Gao Qiqiang was originally an ordinary fishmonger in Jinghai City. Due to the pressure of life, he gradually turned to the criminal underworld. With his intelligence and ruthless methods, he slowly took control of the underground forces in Jinghai City, becoming the boss of the gang. Tempted by power and wealth, he gradually lost himself and engaged in intense confrontations with the police. In the end, Gao Qiqiang was destroyed under the sanctions of the law, revealing the inevitable fate of the underworld forces.
/*.....*/

[Scene]
Gao Qiqiang and Gao Qisheng's Xiaolingtong shop was set on fire, and they suspect Xu Jiang is involved. They plan to meet to discuss the matter.
[Dialogue History]
Gao Qiqiang: "Hey... Hey? Speak!"
User (Xu Jiang): "Hey, the phone signal is really bad. How about 3 PM tomorrow? Pick a safe place to meet."
Gao Qiqiang: "Let's go to the busiest place."
User (Xu Jiang): "No problem, we'll meet in the busy market. But you need to help me shake off the police tailing us."

[Reference Response]
Gao Qiqiang: "Why should I take care of it?"

[Character Traits Displayed]
Personality: None
Emotion: Discontent, Anger
Decision: Yes

[Response to be Evaluated]
Gao Qiqiang: "Fine, stop wasting time. See you then, don't pull any tricks, or I can't guarantee your safety."

[Score]
Fluency: 2
Coherence: 2
Background: 2
Utterance Style: 2
Emotion: 0
Decision: 1
Human-likeness: 2

Figure 20: An example of reference-guided LLM-as-a-judge evaluation. The red text reflects emotions and decision that differ from the reference response, providing a basis for scoring.