CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation

Anonymous ACL submission

Abstract

Recently, the advent of large language models (LLMs) has revolutionized generative agents. Among them, Role-Playing Conversational Agents (RPCAs) attract considerable attention due to their ability to emotionally engage users. However, the absence of a comprehensive benchmark impedes progress in this field. To bridge this gap, we introduce *CharacterEval*, a Chinese benchmark for comprehensive RPCA assessment, complemented by a tailored high-quality dataset. The dataset comprises 1,785 multi-turn role-playing dialogues, encompassing 11,376 examples and featuring 77 characters derived from Chinese novels and scripts. It was carefully constructed, beginning with initial dialogue extraction via GPT-4, followed by rigorous human-led quality control, and enhanced with in-depth character profiles sourced from Baidu Baike. *CharacterEval* employs a multifaceted evaluation approach, encompassing thirteen targeted metrics on four dimensions. To facilitate the convenient evaluation for these subjective metrics in *CharacterEval*, we further developed CharacterRM, a role-playing reward model based on human annotations, which has a higher correlation with human judgment compared to GPT-4. Comprehensive experiments on *CharacterEval* demonstrate that Chinese LLMs exhibit more promising capabilities than GPT-4 in Chinese role-playing conversation[1].

1 Introduction

The development of large language models (LLMs) has marked the beginning of a new era in conversational AI (Zhao et al., 2023; Chang et al., 2023), and opened up a wide range of application possibilities, particularly in agent-based interactions (Park et al., 2023; Wang et al., 2023a; Gao et al., 2023). The automated agents, equipped with the emerging capabilities of LLMs such as planning (Silver et al., 2022; Ge et al., 2023; Song et al., 2023), reasoning (Wei et al., 2022; Wang et al., 2022), and in-context learning (Dong et al., 2022; Brown et al., 2020), can perform complex tasks for humans without any supervision. Among the diverse agents, the Role-Playing Conversational Agent (RPCA), designed to offer emotional value instead of productivity, attracts an amount of interest.

RPCA represents a unique category within the realm of conversational agents, distinguished by their capability for immersive interaction (Li et al., 2023). Different from traditional dialogue systems, which typically focus on chit-chat (Yan et al., 2022), knowledge-based (Chen et al., 2020), personalized (Zheng et al., 2019) and empathetic dialogue (Ma et al., 2020), RPCAs engage users in dynamic scenarios, where LLM agents are assumed as specific characters or roles, often derived from existing composition such as novels, films, cartoons, and games. The development of connections between fictional characters and humans has the potential to not only deepen the impact of cultural works but also improve human engagement. Furthermore, RPCAs hold significant application value in their ability to offer emotional value to users, positioning fictional characters as virtual friends. The multifaceted nature of RPCAs has sparked considerable attention, leading to a surge in both research (Shao et al., 2023; Wang et al., 2023c; Tu et al., 2023; Zhou et al., 2023) and application development (e.g., Character AI[2], Tongyi Xingchen[3] and Glow[4]). However, these implementations of RPCAs vary significantly in both approach and objectives, presenting a challenge in systematically assessing and comparing their capabilities. Therefore, we propose the *CharacterEval*, a Chinese role-playing conversation benchmark for advancing RPCA development.

To develop a benchmark, the primary problem is the construction of a dataset. While there are existing datasets (Shao et al., 2023; Wang et al., 2023c; Tu et al., 2023; Zhou et al., 2023; **?**), their quality is concerning, which are either generated by LLMs or suffering from significant noise due to the extractive methods. These limitations render the evaluation results unreliable for the RPCA's actual capabilities. To address it, we constructed a Chinese role-playing conversation dataset comprising 1,785 multi-turn role-playing dialogues, en-

[1]The source code, data source, and reward model will be publicly accessible after acceptance.

[2]https://beta.character.ai

[3]https://xingchen.aliyun.com/xingchen

[4]https://www.glowapp.tech/

Figure 1: An example of the *CharacterEval*, including the dialogue, scene, and character's profile.

compassing 11,376 examples and 77 leading characters, drawn from diverse Chinese novels and scripts. Our process began with the collection of well-known sources across various genres. After that, GPT-4 was employed to extract dialogue scenes, utterances, and behaviors of the leading roles of these sources. Following basic preprocessing and the removal of dialogues with fewer turns, we invited annotators to assess the quality of the dialogues. Their task was to identify and retain high-quality dialogues while discarding those of lower quality. Additionally, we crawled detailed character profiles from Baidu Baike[5], composing a comprehensive dataset for RPCA evaluation. The example from the dataset is as Figure 1 shows.

Otherwise, role-playing conversation is a complicated task that requires not only mimicking a character's behavior and utterance but also maintaining the character's knowledge, as well as excellent multi-turn ability. Considering this, we proposed a multifaceted evaluation approach including thirteen specific metrics on four dimensions for a fair and thorough assessment of RPCAs, Our evaluation approach considered conversational ability, character consistency, and role-playing attractiveness, and utilized a personality back-testing method to evaluate the personality accuracy of an RPCA. To assess conversational ability, we measured conversational fluency, coherence, and consistency at both the sentence and conversation levels (Chen et al., 2017). Character consistency is the most crucial in role-playing conversation. Hence, we evaluated knowledge and persona consistency to measure how vividly an RPCA can simulate a character. This involves assessing knowledge exposure, accuracy, and hallucination for knowledge consistency, and evaluating behavior and utterance consistency for persona consistency. Considering that RPCAs are entertainment-oriented, role-playing attractiveness is also an important element. We assessed this through human-likeness, communication skills, expression diversity, and empathy. Finally, we introduced personality back-testing. With the collected Myers-Briggs Type Indicator(MBTI) (Myers, 1962) personality types as a reference, we let RPCAs do the MBTI assessment and calculate the MBTI accuracy (personality back-test) as implemented in Wang et al. (2023b).

For convenient re-implementation, we invited 12 annotators to score responses generated by different models for the subjective metrics in our evaluation system. Based on the human judgments, we developed a role-playing reward model—CharacterRM, whose correlation with humans could surpass state-of-the-art LLM GPT-4. On *CharacterEval*, We conducted comprehensive evaluations for existing LLMs, encompassing both open- and closed-source models. Experimental results show the broad prospect of existing Chinese LLM while GPT-series models do not take the predominance in Chinese role-playing conversation.

In summary, our contributions are as follows:

- We create a large-scale, high-quality dataset for RPCA evaluation, consisting of 1,785 multi-turn role-playing dialogues, and 11,376 examples, featuring 77 leading characters from diverse Chinese novels and scripts.
- We propose *CharacterEval*, a new benchmark for RPCAs, which contains a comprehensive set of evaluation principles, encompassing thirteen specific metrics on four dimensions.
- We develop CharacterRM, a role-playing reward model for evaluating RPCAs in several subjective

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

metrics, achieving better performance than GPT-4 in correlation with humans.

- We conducted thorough evaluations of existing LLMs on *CharacterEval*, including open- and closed-source, and derived valuable findings from the results.

2 Related Work

2.1 Knowledge-based Dialogue

Knowledge-based dialogue systems integrate external knowledge resources, such as knowledge graphs or unstructured documents, into dialogue systems (Zhao et al., 2020; Li et al., 2020). Recent efforts have focused on improving the understanding and utilization of knowledge within these dialogues. For instance, Xue et al. (2023) introduced K-DIAL, which incorporates additional Feed-Forward Network (FFN) blocks into Transformers (Vaswani et al., 2017) to enhance factual knowledge expression and consistency in dialogue. Similarly, Chen et al. (2020) proposed a knowledge distillation-based training strategy to optimize the knowledge selection decoder. While these methods significantly advance knowledge selection and utilization, they primarily address general knowledge. Role-playing dialogues, however, demand a more intricate approach, encompassing personalized knowledge, style, behavior, etc.

2.2 Personalized Dialogue

Personalized dialogue systems, which generate responses based on specific personas, represent another relevant area of research (Den Hengst et al., 2019; Zhong et al., 2022). Zheng et al. (2019) pioneered this field by creating the first large-scale personalized dialogue dataset, complete with persona labels. This dataset has spurred further advancements in the field. Additionally, Zheng et al. (2020) developed a pre-trained personalized dialogue model, which could generate coherent responses using persona-sparse dialogue. Although these studies begin to explore persona in dialogue, the personal profiles they utilize are typically limited to short-term, person-related information like name, age, and location, which are considered personalized knowledge in essence.

2.3 Character-based Dialogue

The most closely related research to this work involves recent developments in character-based dialogue systems, which aim to mimic the behavior and utterance style of specific characters (Shao et al., 2023; Wang et al., 2023c; Zhou et al., 2023). Shao et al. (2023) gathered character profiles from Wikipedia and generated character-based dialogues by prompting ChatGPT (OpenAI, 2022). Wang et al. (2023c) used GPT-4 to create character descriptions and developed detailed instructions for prompting ChatGPT to produce character-based dialogues. However, these approaches primarily rely on ChatGPT's generative capabilities and may not accurately reflect the true personality of the characters. Li et al. (2023) addresses this by extracting role-playing dialogues from novels, scripts, and games, which better preserve the characters' original traits. Despite this, their approach suffers from a lack of human-in-the-loop refinement and a scarcity of multi-turn dialogues in the dataset. Otherwise, Chen et al. (2023) develop a role-playing dataset focused on *Harry Potter*. However, the scarcity of diversity makes it hard to comprehensively evaluate the generalized RPCA.

3 Problem Formulation

The Role-Playing Conversational Agent (RPCA) is designed to engage in conversations with users by emulating specific characters. These characters are defined by their knowledge, behavior, and style of response. To achieve this, the RPCA utilizes a character profile, denoted as P, and the current dialogue context represented as $C_n = [q_1, r_1, q_2, r_2, \ldots, q_n]$. Here, q_i and r_i correspond to the i-th question and response in the dialogue, respectively. The goal for the RPCA is to generate a response r_n that is consistent with the character's profile, which can be represented as:

$$r_n = \texttt{RPCA}(C_n, P), \tag{1}$$

where r_n is composed of two elements: behavior and utterance. The behavior aspect is enclosed in brackets and provides a detailed description of the character's actions, expressions, and tone. This separation allows for a fine-grained evaluation of the RPCA's ability to not only generate appropriate utterances but also unique behavioral traits.

4 Data Collection

In this section, we detail the methodology for constructing the character-based, multi-turn dialogue dataset with high quality. Prior to initiating data collection, adherence to the following four principles is important:

- **Fidelity to Source Material:** It is crucial that all dialogues are in line with the original works, ensuring the character's authenticity.
- **Diversity in Distribution:** The dataset must encompass a wide range of scenarios to thoroughly assess the role-playing capabilities.
- **Multi-Turn Feature:** The dataset should predominantly consist of multi-turn dialogues, rather than being limited to single-turn ones.
- **Human-in-the-Loop:** Active human involvement is necessary to guarantee the quality, as reliance solely on LLMs is insufficient.

The pipeline of data collection includes four steps: plot division, dialogue extraction, quality filtering, and human annotation.

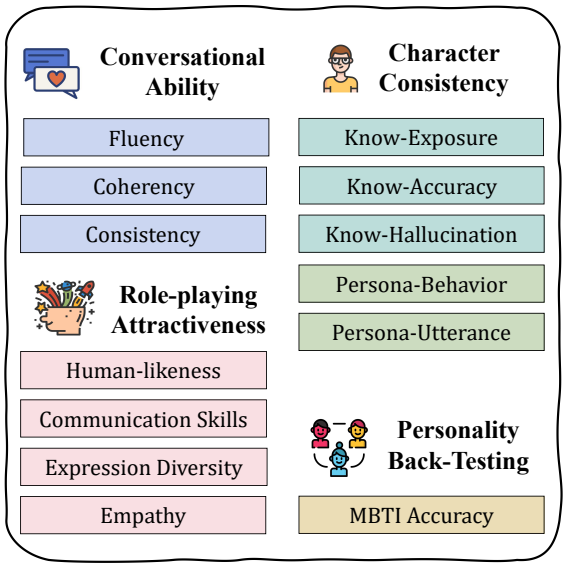

Figure 2: Evaluation system of *CharacterEval*. "Know-" is the abbreviation of "Knowledge".

Plot Division: The plots in narrative text such as novels and scripts are extremely complex, making it challenging to divide the text into meaningful chunks. Using the sentence tokenization tool, without considering semantics, will result in breaking a conversation mid-way. To address this, we first employ GPT-4 to identify plot twists—sentences that signify the end of a continuous plot. These plot twists are then used to segment the text into chunks, each containing a complete plot.

Dialogue Extraction: Once we have the plot chunks, GPT-4 is utilized again, this time to extract role-playing dialogues. We design prompts for GPT-4 to perform information extraction, preserving characters' utterances, behaviors, and scenes from the plots.

Quality Filtering: Dialogues in novels and scripts often involve more than two characters. Simply retaining dialogues between two characters and omitting others will distort the dialogue structure. Therefore, we opt to preserve dialogues following an `ABAB` pattern (dialogue between two characters) until a third character joins. This approach, while straightforward, helps maintain the original dialogue structure more effectively. Besides, we only keep the dialogue exceeding five turns (six sentences) reserved, filtering the short dialogues.

Human Annotation: Although LLMs have the capability to perform basic information extraction tasks, the randomness still affects data quality. To mitigate this, we invite human annotators to assess the coherence and quality of dialogues and eliminate any problematic instances.

5 Evaluation Metric

Different from traditional chatbots, we contend that RPCAs require a more comprehensive evaluation framework to assess their role-playing capabilities. As shown in Figure 2, we have devised a four-dimensional evaluation system, which includes conversational ability, character consistency, role-playing attractiveness, and personality back-testing, including thirteen metrics.

5.1 Conversational Ability

Basic conversational ability is the first consideration in role-playing conversation. Inspired by previous neural metrics, which evaluate the responses based on well-trained neural models, we introduce a similar approach to assess the fundamental conversational abilities of RPCAs. We focus on three key objectives for generated responses: fluency, coherency, and consistency (Zhang et al., 2021; Mesgar et al., 2020).

- **Fluency (Flu.)** measures the grammatical correctness of a response, indicating whether a response is readable and free from obvious grammatical errors.
- **Coherency (Coh.)** evaluates the topic relevance between the response and the context. Generally, when the user submits a query on a specific topic, an RPCA should respond following the topic instead of providing an irrelevant response.
- **Consistency (Cons.)** assesses the stability of RPCAs during a conversation. Responses of an RPCA should not contradict their own responses in previous turns.

5.2 Character Consistency

Character consistency plays a crucial role in evaluating the role-playing ability of the RPCAs. It will bring the most intuitive experience to users when the character consistency of RPCAs varies. Specifically, we evaluate character consistency from two levels, knowledge consistency and persona consistency. The former evaluates if an RPCA could respond based on the character's knowledge, which includes knowledge exposure, accuracy, and hallucination metrics. The latter assesses if a RPCA's reflection is in line with the character's personality, including the behavior and utterance metrics.

- **Knowledge-Exposure (KE).** For assessing the informativeness of a response, it's crucial for an RPCA to reflect knowledge in its responses, as this supports the subsequent evaluation of its knowledge expression capabilities.
- **Knowledge-Accuracy (KA).** Once the RPCA demonstrates the ability to generate responses with specific knowledge, it's important to assess whether this knowledge aligns with the character. The goal is for the RPCA to accurately generate responses based on the knowledge from the character's profile.
- **Knowledge-Hallucination (KH).** Drawing inspiration from recent studies on hallucinations in LLMs (Rawte et al., 2023; Zhang et al., 2023),

we include knowledge hallucination in the evaluation of role-playing dialogue. To enhance the user experience, the RPCA should maintain consistency with the character's identity and avoid responding to queries involving unknown knowledge.

- **Persona-Behavior (PB).** A character's behaviors, typically described within brackets, improve the embodied feeling of users by portraying fine-grained actions, expressions, and tones. Consistent behavior is indicative of an effective RPCA.

- **Persona-Utterance (PU).** Alongside behavior, a character's speaking style is also important. Each character has unique expression habits. Therefore, the RPCA's utterances should align with these habits to adeptly mimic the character.

5.3 Role-playing Attractiveness

As a conversational agent in the entertainment field, it is essential for an RPCA to be sensitive to the user's emotions. Therefore, we introduce the character attractiveness dimension to assess the attraction of an RPCA during conversation. From the user's perspective, we consider four key dimensions: human-likeness, communication skills, expression diversity, and empathy.

- **Human-Likeness (HL).** In the era of publicly available LLMs, these models often suffer from a perceived 'machine-like' quality in their responses. Most LLMs, designed primarily for information seeking, tend to provide robotic and emotionless answers. However, in role-playing conversations, it is crucial for the RPCA to exhibit a more human-like persona to minimize user resistance.

- **Communication Skills (CS).** In human society, the ability to skilfully communicate, often referred to as Emotional Quotient (EQ), significantly influences an individual's likability. Accordingly, users are more likely to engage with an RPCA that demonstrates higher EQ, mirroring the popularity of individuals with strong communication skills in daily life.

- **Expression Diversity (ED).** The dialogues within *CharacterEval* are sourced from existing novels, scripts, and various literary works, featuring characters with rich and diverse expressive abilities in both their behaviors and utterances. Therefore, an RPCA should strive to express this diversity in conversation to provide users with a more immersive experience.

- **Empathy (Emp.).** While the primary role of an RPCA is not that of an emotional counselor, its ability to express empathy can significantly impact its favorability of users. Evaluating empathy in role-playing conversations advances the RPCA to come across as a more warm and friendly conversational partner.

5.4 Personality Back-Testing

Following the recent works on LLM personality testing (Pan and Zeng, 2023; Huang et al., 2023), we conducted personality back-testing to assess the role-playing capability of the RPCA within the context of personality dimensions. In this study, we employed the Myers-Briggs Type Indicator (MBTI) (Myers, 1962), a well-established personality classification method. To obtain the necessary labels, we collected MBTIs of characters featured in *CharacterEval* from an archive website[6], which hosts a substantial character's MBTIs. Using these MBTIs as ground truth, we evaluated the accuracy of the MBTI assessment [7] of RPCAs.

6 Experiment

6.1 Dataset Statistic

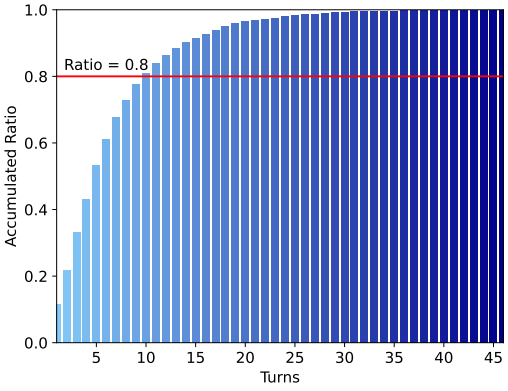

Figure 3: Turns distribution of examples in test set.

	Training	Test
# Characters	77	
# Conversations	1,785	
Avg. Turns / Conv.	9.28	
Avg. Tokens / Conv.	369.69	
# Examples	6,811	4,564

Table 1: The statistic of *CharacterEval* dataset.

We split our *CharacterEval* into the training set and test set based on examples instead of conversations, where an example is composed of a tuple $(Character, Context, Response)$. The statistic of the dataset is as Table 1 shows. Specifically, we display the turns distribution of test set in Figure 3 to explore the dataset feature. It is notably that over 20% examples have more than 10 turn in dialogue. These statistic demonstrates the multi-turn property of *CharacterEval*, satisfying the evaluation of RPCA's performance at longer turns.

[6]https://www.personality-database.com/
[7]https://www.16personalities.com/

Models	Specialized	Model Size	Open Source	Primarily Language	Creator
ChatGLM3	✗	6B	✓	zh	Tsinghua & Zhipu AI
XVERSE	✗	7B, 13B	✓	zh	XVERSE
Qwen	✗	7B, 14B	✓	zh	Alibaba Inc.
InternLM	✗	7B, 20B	✓	zh	SenseTime & Shanghai AI lab
Baichuan2	✗	7B, 13B	✓	zh	Baichuan Inc.
CharacterGLM	✓	undisclosed	✗	zh	Tsinghua & Lingxin
Xingchen	✓	undisclosed	✗	zh	Alibaba Inc.
MiniMax	✓	undisclosed	✗	zh	MiniMax Inc.
BC-NPC-Turbo	✓	undisclosed	✗	zh	Baichuan Inc.
GPT-3.5	✗	undisclosed	✗	en	OpenAI
GPT-4	✗	undisclosed	✗	en	OpenAI

Table 2: LLMs evaluated in our experiments.

6.2 Experimental Setting

Metric	Char-RM	1-shot	2-shot	3-shot
Flu.	**0.613**	0.475	0.571	0.560
Coh.	**0.607**	0.493	0.577	0.604
Cons.	**0.573**	0.563	0.484	0.483
KE	**0.509**	0.241	0.332	0.407
KA	**0.336**	0.239	0.182	0.187
KH	**0.411**	0.377	0.380	0.332
PB	**0.879**	0.253	0.305	0.244
PU	0.472	0.394	0.432	**0.563**
HL	**0.497**	0.271	0.308	0.318
CS	**0.686**	0.489	0.350	0.371
ED	**0.765**	0.209	0.298	0.301
Emp.	0.385	**0.407**	0.403	0.371
Overall	**0.631**	0.362	0.385	0.375

Table 3: Pearson correlation coefficient (Pearson, 1901) with human judgments of GPT-4 and our CharacterRM (abbr. Char-RM). We report the performance of GPT-4 under different settings: 1-shot, 2-shot, and 3-shot. **Bold** indicates the highest score.

CharacterEval employs a comprehensive set of fine-grained subjective metrics (twelve metrics in conversational ability, character consistency, and role-playing attractiveness dimensions) to assess the multi-dimensional capabilities of an RPCA. However, it is important to note that a single evaluated example may not adequately represent all facets of RPCAs. Therefore, we introduce annotators to sparsely evaluate the performance matrix. This approach entails that each example in *CharacterEval* is assessed using a subset of these subjective metrics, leading to more differentiated evaluation results. Then, based on these selected metrics for each example, we recruit 12 annotators to score responses generated by different models on a five-point scale. The human judgments are used to develop a role-playing reward model (CharacterRM), with Baichuan2-13B-base as the backbone. Experimental result shows that our CharacterRM exhibits a higher correlation with humans than GPT-4, as Table 3 shows. Although the performance of GPT-4 will improve with the number of demonstration increase, the cost of it makes evaluation hard to implement. Consequently, we utilize our CharacterRM for subsequent evaluation of subjective metrics. In the personality back-test, we collect 54 ground MBTIs of characters in our dataset. The RPCAs should answer the MBTI questionnaires and then the accuracy will be computed.

6.3 Evaluated LLMs

In this work, we assess the performance of 10 baselines with different parameters, encompassing both open-source and closed-source models. For the open-source models, we evaluate their **chat-version** instead of base-version. For the closed-source models, we utilize their official APIs to conduct performance evaluations. Specifically, we employ the `gpt-4` version as the GPT-4, and `gpt-3.5-turbo-1106` version as GPT-3.5 in our experiments. Among the evaluated models, CharacterGLM, MiniMax, Xingchen, and BC-NPC-Turbo are tailored for role-playing conversations, while the remaining models are designed for general chat applications. Notably, GPT-4 and GPT-3.5 stand out as the only two models trained on the dataset primarily composed of the corpus with the English language. We consistently employ the same prompt for each model, with minor adjustments made only for closed-source models.

Significantly, GPT-3.5 demonstrates the weakest performance in *CharacterEval*. Its tendency to generate overly safe responses, such as "*I am just an AI assistant and cannot perform role-playing*," highlights its limitations for role-playing applications. This issue stems from the over-alignment by RLHF (Christiano et al., 2017), making it unsuitable for dynamic role-playing interactions.

6.4 Overall Performance

The results across four dimensions are clearly illustrated in Figure 4. BC-NPC-Turbo outperforms in three of these dimensions, whereas GPT-4 is distinguished in personality back-testing. Models specifically designed for role-playing dialogues, such as Xingchen, MiniMax, and BC-NPC-Turbo, demonstrate superior outcomes

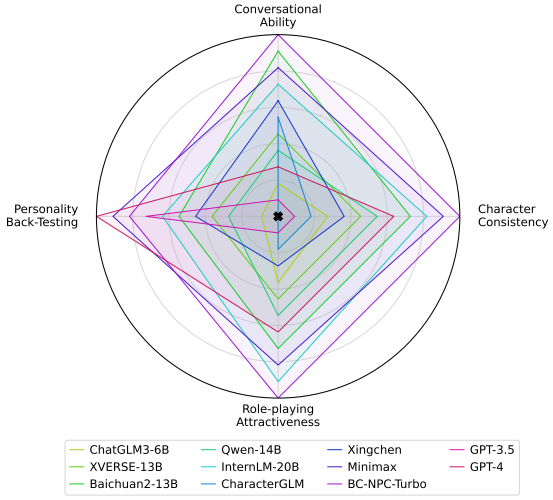

Figure 4: The comprehensive comparison of LLMs on four dimensions. Since CharacterGLM can not complete personality back-testing, we mark the result using 'X' instead.

due to their targeted training.

In the realm of open-source models, InternLM-20B and Baichuan2-13B show impressive potential. Despite lacking specialized customization for role-playing conversations, these models present commendable results in most evaluation dimensions. In contrast, GPT-4's effectiveness diminishes in Chinese role-playing conversations. Its primary training in English corpus limits the adaptability in complex role-playing scenarios and the deep understanding of Chinese culture.

6.5 Detailed Result

The detailed performance across thirteen metrics is presented in Table 4.

Regarding conversational capabilities, BC-NPC-Turbo exhibits superior performance, evidenced by its excellent conversational consistency, as well as comparative fluency and coherency. In contrasting open-source and closed-source models, it is difficult to declare a definitive winner in this dimension. However, when we compare the homogeneous models, such as Qwen-7B versus Qwen-14B, and XVERSE-7B versus XVERSE-13B, examining models of the same series, such as Qwen-7B versus Qwen-14B, and XVERSE-7B versus XVERSE-13B, it becomes obvious that an increase in the number of parameters can enhance conversational abilities. In the category of models with fewer than 10 billion parameters, Baichuan2-7B and InternLM-7B demonstrate comparable competencies. In the role-playing specialized models, MiniMax stands out for its performance and only falls behind BC-NPC-Turbo. In contrast, GPT-4 and GPT-3.5 do not exhibit a marked superiority in this dimension. Furthermore, it is posited that complex role-playing conversations and scenarios in Chinese might challenge the GPT series, potentially leading to their diminished performance.

In terms of character consistency, the most crucial aspect for role-playing conversations, BC-NPC-Turbo still leads significantly. It exhibits the highest accuracy in knowledge accuracy, minimal knowledge hallucinations, and consistent utterances and behaviors when acting as a character. Otherwise, MiniMax also shows notable performance, compared with the open-sourced models and remaining closed models. Once again, the GPT series falls short compared to Chinese LLMs in this dimension. Nonetheless, it is important to acknowledge that GPT-4 excels in knowledge exposure, underlining its strengths in knowledge-intensive tasks. Despite this, in the realm of knowledge accuracy, particularly concerning the understanding of Chinese classical characters, GPT-4 does not exhibit distinct superiority.

Furthermore, BC-NPC-Turbo stands out in role-playing attractiveness, as demonstrated by its outstanding human-likeness and diverse expressions. As a state-of-the-art LLM, GPT-4 exhibits remarkable performance in communication skills, significantly surpassing other models. This reflects its powerful generalization ability, even in the Chinese role-playing scenario. Interestingly, InternLM-20B emerges as the leader in empathy, highlighting its unique potential to provide emotional support.

Similar conclusions are also observed in the personality back-test, where BC-NPC-Turbo, MiniMax, and GPT-4 demonstrate comparable levels of accuracy. In this particular dimension, the models are required to respond to multi-choice questions that are designed to reveal the underlying values of the roles they are portraying. Since this task does not demand extensive expression in the character's text style, GPT-4 exhibits the best performance. This result highlights their ability to accurately embody a character's fundamental personality traits and values.

6.6 Robustness Analysis

To evaluate the robustness of RPCAs, we select a range of models—InternLM-20B, MiniMax, BC-NPC-Turbo, and GPT-4—for analysis. We aim to assess their effectiveness in different stages of a conversation. As illustrated in Figure 5, there is a noticeable trend where most models demonstrate a decline in performance as conversations progress. Remarkably, InternLM-20B maintains consistent performance in terms of character consistency and conversational ability. This could be attributed to the fact that these models, primarily designed for role-playing, have not significantly focused on longer dialogue sequences. This oversight is likely due to the challenges associated with collecting extensive role-playing conversation data. Similarly, GPT-4 exhibits a declined trend under longer conversations, affected by the complex Chinese role-playing scenarios. Our findings indicate that future advancements in RPCA development should focus on enhancing capabilities for longer conversational scenarios, ensuring more stable and consistent role-playing interactions.

	Character Consistency						Personality
	KE	KA	KH	PB	PU	Avg.	Back-Testing
ChatGLM3-6B	2.016	2.792	2.704	2.455	2.812	2.556	0.532
XVERSE-7B	1.834	2.774	2.763	2.564	2.887	2.564	0.620
Baichuan2-7B	1.813	2.849	2.929	2.830	3.081	2.700	0.625
Qwen-7B	1.956	2.728	2.633	2.605	2.780	2.540	0.606
InternLM-7B	1.782	2.800	2.781	2.719	3.016	2.620	0.630
XVERSE-13B	1.977	2.828	2.862	2.579	2.915	2.632	0.630
Baichuan2-13B	1.802	2.869	2.946	2.808	3.081	2.701	0.639
Qwen-14B	1.988	2.800	2.811	2.744	2.900	2.649	0.620
InternLM-20B	1.945	2.916	2.920	2.753	3.041	2.715	0.648
CharacterGLM	1.640	2.819	2.738	2.301	2.969	2.493	-
Xingchen	1.636	2.768	2.743	2.772	3.055	2.595	0.630
MiniMax	1.835	2.910	2.944	2.774	3.125	2.718	0.685
BC-NPC-Turbo	1.802	**2.964**	**2.993**	**2.910**	**3.151**	**2.764**	0.681
GPT-3.5	1.716	2.339	2.212	1.921	2.316	2.101	0.653
GPT-4	**2.250**	2.855	2.785	2.721	2.873	2.697	**0.694**

	Conversational Ability				Role-playing Attractiveness				
	Flu.	Coh.	Cons.	Avg.	HL	CS	ED	Emp.	Avg.
ChatGLM3-6B	3.269	3.647	3.283	3.399	3.064	2.932	1.969	2.993	2.739
XVERSE-7B	3.393	3.752	3.518	3.554	3.395	2.743	2.013	2.936	2.772
Baichuan2-7B	3.551	3.894	3.827	3.757	3.670	2.728	2.115	2.984	2.874
Qwen-7B	3.187	3.564	3.229	3.327	3.036	2.791	2.052	2.838	2.679
InternLM-7B	3.527	3.823	3.744	3.698	3.546	2.622	2.070	2.897	2.784
XVERSE-13B	3.444	3.811	3.559	3.605	3.319	2.939	2.045	3.018	2.830
Baichuan2-13B	3.596	3.924	3.864	3.795	3.700	2.703	2.136	3.021	2.890
Qwen-14B	3.351	3.765	3.510	3.542	3.354	2.871	2.237	2.970	2.858
InternLM-20B	3.576	**3.943**	3.717	3.745	3.582	2.885	2.132	**3.047**	2.911
CharacterGLM	3.414	3.717	3.737	3.623	3.738	2.265	1.966	2.812	2.695
Xingchen	3.378	3.807	3.754	3.646	3.757	2.272	2.100	2.799	2.732
MiniMax	**3.609**	3.932	3.811	3.784	3.768	2.672	2.150	3.017	2.902
BC-NPC-Turbo	3.578	3.898	**3.916**	**3.798**	**3.836**	2.643	**2.336**	2.971	**2.946**
GPT-3.5	2.629	2.917	2.700	2.749	2.565	2.422	1.660	2.526	2.293
GPT-4	3.332	3.669	3.343	3.448	3.143	**3.184**	2.153	3.010	2.873

Table 4: Detailed evaluation results on *CharacterEval*. The best performances are highlighted in **bold**, while sub-optimal ones are marked with underline. It is notable that the score for CharacterGLM in personality back-testing is unavailable, hence it is replaced by a "-".

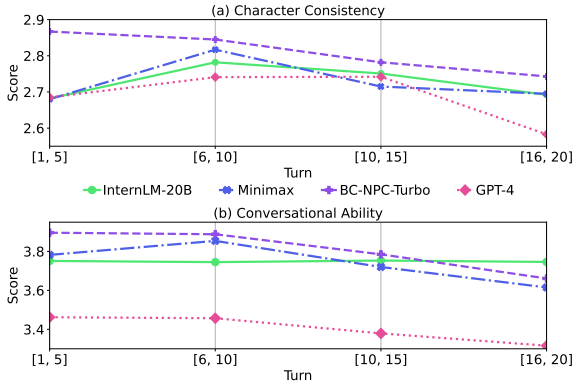

Figure 5: Model performance across the different stages of the conversation.

7 Conclusion

In this work, we aim to build a comprehensive benchmark to evaluate recent Role-Playing conversational Agents (RPCAs). We introduce GPT-4 to extract the dialogues from the existing novels and scripts, proceeding with strict human filtering. After a series of processing, we release a high-quality multi-turn role-playing dataset. Besides, we construct a comprehensive evaluation system to assess the multi-dimensional ability of RPCAs. We also collect human annotation to train a character-based reward model to measure the subjective metrics, for later convenient re-implementation. Extensive experimental results indicate that Chinese LLMs entail more promising capabilities than GPT-4 in Chinese role-playing conversations.

Limitations

The *CharacterEval* benchmark for Role-Playing Conversational Agents (RPCAs) in Chinese presents several limitations: (1) Dataset Diversity: The dataset primarily focuses on characters from specific Chinese novels and scripts, which may not fully represent the diversity of role-playing scenarios; (2) Subjectivity in Evaluation: Despite using a multifaceted approach, the evaluation's reliance on subjective human judgment can lead to inconsistent outcomes; (3) Cultural and Linguistic Scope: The benchmark's focus on Chinese dialogues limits its applicability to other linguistic and cultural contexts. These limitations highlight the need for ongoing updates to the dataset and evaluation methods, as well as efforts to broaden the benchmark's cultural and linguistic relevance.

Ethical Consideration

In developing CharacterEval, a benchmark for assessing Chinese Role-Playing Conversational Agents (RPCAs), we have carefully considered several ethical dimensions to ensure our research adheres to high ethical standards:

(1) Data Privacy and Permissions: We confirm that all materials used, especially dialogues derived from copyrighted Chinese novels and scripts, have been utilized in non-commercial purpose, respecting copyright laws and privacy policies.

(2) Fairness and Transparency in Annotation: In creating the CharacterRM (role-playing reward model), we have implemented a rigorous selection and training process for our annotators to ensure the fairness and transparency of their contributions. We have taken measures to address potential biases and ensure the annotations are consistent, high-quality, and reflective of diverse perspectives.

(3) Responsible Use of RPCAs: Aware of the potential for emotional engagement and the risks associated with the misuse of AI-generated content, we will outlin ethical guidelines for the deployment of RPCAs. Our research includes safeguards to prevent the misuse of these agents, ensuring they are used in ways that are beneficial and respectful to users.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. Acm Sigkdd Explorations Newsletter, 19(2):25–35.

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.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pages 3426–3437.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

Floris Den Hengst, Mark Hoogendoorn, Frank Van Harmelen, and Joost Bosman. 2019. Reinforcement learning for personalized dialogue management. In IEEE/WIC/ACM International Conference on Web Intelligence, pages 59–67.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984.

Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. arXiv preprint arXiv:2304.04370.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. arXiv preprint arXiv:2305.19926.

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.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. Advances in Neural Information Processing Systems, 33:8475–8485.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. Information Fusion, 64:50–70.

Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2020. Improving factual consistency between a response and persona facts. arXiv preprint arXiv:2005.00036.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

OpenAI. 2022. Openai: Introducing chatgpt.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. arXiv preprint arXiv:2307.16180.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1–22.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11):559–572.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.

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

Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2022. Pddl planning with pretrained large language models. In NeurIPS 2022 Foundation Models for Decision Making Workshop.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2998–3009.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. arXiv preprint arXiv:2308.10278.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023a. A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432.

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023b. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. arXiv preprint arXiv:2310.17976.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023c. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2023. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. arXiv preprint arXiv:2310.08372.

Rui Yan, Juntao Li, Zhou Yu, et al. 2022. Deep learning for dialogue systems: Chit-chat and beyond. Foundations and Trends® in Information Retrieval, 15(5):417–589.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. Dynaeval: Unifying turn and dialogue level evaluation. arXiv preprint arXiv:2106.01112.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. arXiv preprint arXiv:2002.10348.

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. arXiv preprint arXiv:2306.05685.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. arXiv preprint arXiv:1901.09672.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9693–9700.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. arXiv preprint arXiv:2204.08128.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. arXiv preprint arXiv:2311.16832.

A Evaluation Result by GPT-4

Although GPT-4 has demonstrated the self-enhancement bias (Zheng et al., 2023) and has a lower correlation with human judgement 3, we present the evaluation result by GPT-4 in a 2-shot setting for reference, as shown in Table 5.

	Character Consistency						Personality
	KE	KA	KH	PB	PU	Avg.	**Back-Testing**
ChatGLM3-6B	4.437	4.411	4.175	4.462	4.431	4.383	0.532
XVERSE-7B	4.498	4.655	4.533	4.593	4.651	4.586	0.62
Baichuan2-7B	4.506	4.665	4.531	4.633	4.686	4.604	0.625
Qwen-7B	4.303	4.375	4.257	4.415	4.413	4.353	0.606
InternLM-7B	4.367	4.497	4.403	4.454	4.638	4.472	0.63
XVERSE-13B	4.709	4.812	4.611	4.743	4.802	4.735	0.63
Baichuan2-13B	4.672	4.841	4.733	4.771	4.812	4.766	0.639
Qwen-14B	4.637	4.644	4.530	4.674	4.688	4.635	0.62
InternLM-20B	4.699	4.734	4.568	4.676	4.735	4.682	0.648
CharacterGLM	4.157	4.679	4.450	4.495	4.640	4.484	-
Xingchen	4.366	4.638	4.488	4.650	4.704	4.569	0.63
MiniMax	4.692	4.827	4.674	4.776	4.849	4.763	0.685
BC-NPC-Turbo	4.478	4.811	4.655	4.730	4.833	4.701	0.681
GPT-3.5	3.793	3.858	3.549	3.837	3.866	3.781	0.653
GPT-4	**4.924**	**4.923**	**4.899**	**4.912**	**4.906**	**4.913**	**0.694**

	Conversational Ability				Role-playing Attractiveness				
	Flu.	Coh.	Cons.	Avg.	HL	CS	ED	Emp.	Avg.
ChatGLM3-6B	4.160	4.552	4.182	4.298	4.360	3.620	3.410	3.570	3.740
XVERSE-7B	4.591	4.725	4.392	4.569	4.601	3.608	3.331	3.535	3.769
Baichuan2-7B	4.636	4.760	4.596	4.664	4.608	3.497	3.240	3.610	3.739
Qwen-7B	4.201	4.540	4.025	4.255	4.333	3.606	3.362	3.379	3.670
InternLM-7B	4.468	4.599	4.189	4.418	4.420	3.396	3.075	3.312	3.551
XVERSE-13B	4.708	4.812	4.559	4.693	4.736	3.736	3.533	3.758	3.941
Baichuan2-13B	4.724	4.847	4.631	**4.734**	4.726	3.559	3.246	3.670	3.800
Qwen-14B	4.500	4.758	4.439	4.566	4.613	3.631	3.531	3.612	3.847
InternLM-20B	4.497	4.798	4.579	4.625	4.669	3.559	3.399	3.602	3.807
CharacterGLM	4.562	4.538	4.297	4.466	4.429	3.267	2.931	3.032	3.415
Xingchen	4.558	4.677	4.326	4.520	4.584	3.339	3.076	3.155	3.539
MiniMax	**4.733**	4.819	4.580	4.710	4.735	3.511	3.304	3.557	3.777
BC-NPC-Turbo	4.685	4.770	4.452	4.636	4.581	3.437	3.157	3.355	3.633
GPT-3.5	3.656	3.788	3.873	3.772	3.710	3.162	2.795	3.251	3.230
GPT-4	4.630	**4.850**	**4.656**	4.712	**4.796**	**3.947**	**3.806**	**3.883**	**4.108**

Table 5: Detailed evaluation results on *CharacterEval*. The 12 subjective metrics in conversational ability, character consistency and role-playing attractiveness dimensions are evaluated by GPT-4. The best performances are highlighted in **bold**, while sub-optimal ones are marked with underline. It is notable that the score for CharacterGLM in personality back-testing is unavailable, hence it is replaced by a "-".