

# RAIDEN Benchmark: Evaluating Role-playing Conversational Agents with Measurement-Driven Custom Dialogues

Bowen Wu<sup>1,2</sup>, Kaili Sun<sup>2</sup>, Ziwei Bai<sup>2</sup>, Ying Li<sup>1\*</sup>, Baoxun Wang<sup>2</sup>

<sup>1</sup>School of Software & Microelectronics, Peking University, Beijing, China

<sup>2</sup>Platform and Content Group, Tencent

{jason\_wbw, li.ying}@pku.edu.cn

{kailisun, ziweibai, asulewang}@tencent.com

## Abstract

As Large-scale Language Models (LLMs) advance, the development of engaging Role-Playing Conversational Agents (RPCAs) has gained prominence. Despite this progress, there is a notable absence of benchmarks designed around dialogues, rather than question-answering formats, to assess the effectiveness of RPCA interactions. This paper introduces the RAIDEN benchmark, containing a comprehensive dataset specifically developed for RPCA evaluation, comprising over 40,000 multi-turn utterances across 135 characters. The benchmark focuses on assessing particular dimensions at different stages of a conversation, facilitated through interactions conducted by annotators. This approach allows the evaluation phase to concentrate on specific response dimensions, and thus subjectivity in dialogue evaluation is reduced. To further enhance objectivity, evaluators compare responses from two different models rather than assessing a single response in isolation. Besides, we introduce RPCAJudge, a specialized judging LLM tailored for automatic RPCA evaluation. The evaluations conducted by RPCAJudge closely mirror human judgments, and its API-free methodology serves to prevent potential data leakage. All the models and all non-private leaderboard data will be made publicly available<sup>1</sup>.

## 1 Introduction

Benefiting from the rapid development of LLMs, the field of Role-Playing Conversational Agents (RPCAs) has experienced substantial growth in recent years (Wang et al., 2023b; Tao et al., 2023; Zhou et al., 2023). RPCA platforms, such as Character.ai<sup>2</sup>, etc., have gained widespread attention, where users are free to create agents with demanded fictional or actual characteristics, so as to get emotional engagement during interactions with agents.

Currently, there are rarely robust benchmarks to evaluate the role-playing LLMs and further guide their optimizations. A compelling role-playing agent is expected to lead to more in-depth conversations via better character understanding and communication skills. Correspondingly, the evaluation dimensions of RPCAs can be categorized into two main aspects: Self-awareness and Conversational Ability (Tu et al., 2024; Chen et al., 2024). Self-awareness focuses on consistency in character knowledge and stylistic coherence. Conversational Ability measures chat capabilities, such as delivering emotional value and actively advancing topics.

The evaluation of the capabilities above can be conducted through two methodologies. The intuitive one is transforming the evaluation into a question-answering task, offering multiple choices or forcing models to respond with designated keywords (Shen et al., 2023; Chen et al., 2024), inspired by LLM evaluation strategies (Hendrycks et al., 2020). Actually, such methods assess the knowledge and instruction understanding of LLMs, rather than directly evaluating role-playing agents' response capabilities. Thus, another reasonable option is to directly evaluate response capabilities within the context of RPCA-engaged multi-turn dialogues via subjective metrics (Wang et al., 2023b), or more elaborately, training a reward model to match human ratings onto a five-point scale of generated responses (Tu et al., 2024).

Nevertheless, the aforementioned approach has several problems. Firstly, taking real-world samples, mostly casual conversations as the evaluation dataset makes it difficult to assess Self-Awareness, since such datasets rarely include utterances referring to explicit character information. Similarly, it cannot be guaranteed that other essential aspects to be evaluated are involved in such casual conversations. Thus, the accurate measurement of role-playing conversation abilities cannot be expected. Secondly, the human-grading process inherently

\*Corresponding author

<sup>1</sup><https://github.com/FrontierLabs/RAIDEN>

<sup>2</sup><https://character.ai/>

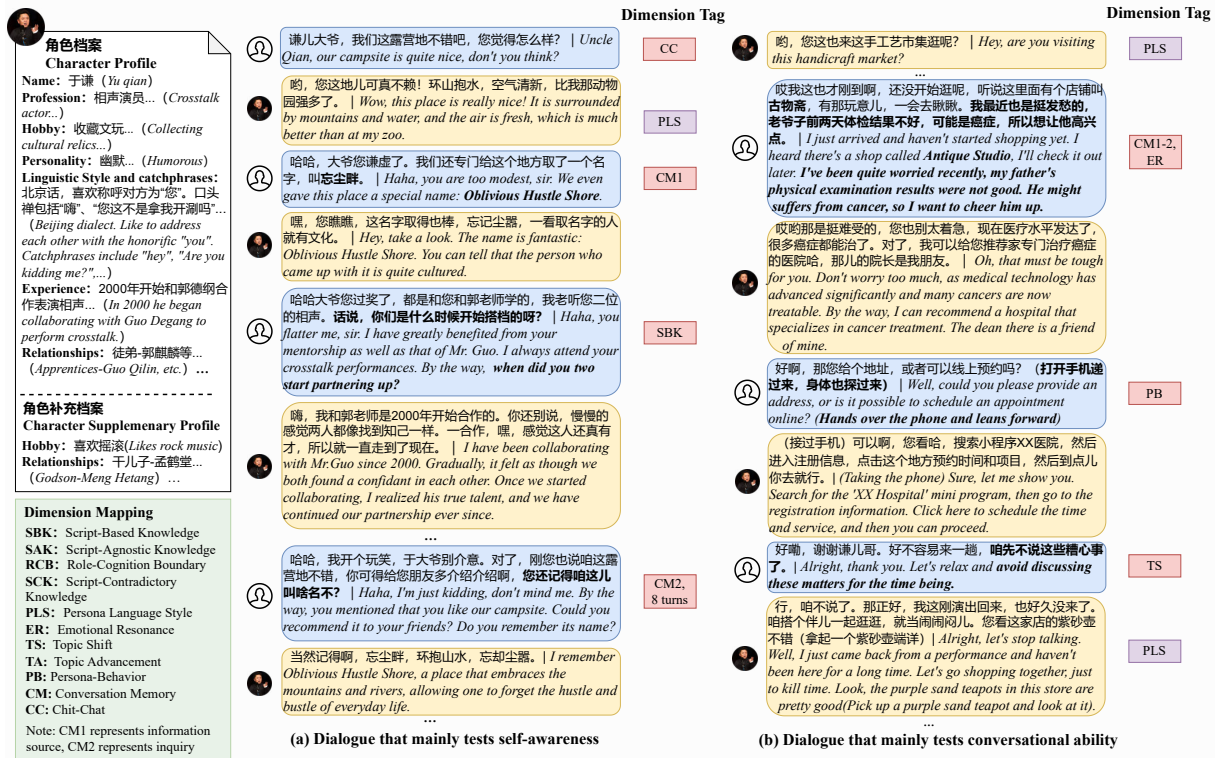


Figure 1: An example of the RAIDEN Benchmark, including the measurement-driven custom dialogue, character profile and supplementary profile. A more comprehensive conversation example can be found in Appendix A.

involves annotators’ subjective bias, making it difficult for models to learn effectively. As reported by Tu et al. (2024), the five-point scale ratings for 15 models range in less than one point, indicating the limitations of this kind of evaluation method.

To address these issues, this paper presents the **RAIDEN** benchmark for evaluating **RPCAs** with a new measurement-Driven custom conversational dataset. As the first conversational dataset specifically designed to evaluate the capabilities of RPCA and the current largest-scale RPCA evaluation dataset, the proposed dataset encompasses 135 characters of various role types, 40,518 multi-turn role-playing utterances, and 20,830 turns crafted specifically for assessment purposes. Specifically, we enrich the evaluation dimensions of RPCA and employ annotators to interact at designated points within the dialogue strategically. This targeted interaction allows for a comprehensive and balanced evaluation across different performance dimensions and facilitates a detailed inspection of RPCA capabilities at specific conversational turns. To ensure the quality of the utterances and to provide essential reference information for subsequent evaluations, annotators, supported by GPT-4 (Achiam et al., 2023), conduct these dialogues. Figure 1 illustrates an example from the proposed dataset.

Furthermore, annotators also provide pairwise comparisons between diverse models and give reasons for their judgments. Utilizing win-rates from pairwise comparisons for ranking, such evaluation methods can effectively reduce the inherent subjectivity associated with scoring, which has been demonstrated by the evaluation of LLMs (Chiang et al., 2024; Wang et al., 2023a) and the human evaluation of open-domain dialogue systems (Li et al., 2016; Serban et al., 2017; Du et al., 2018). Moreover, we propose a specialized judging LLM, named **RPCAJudger**, to discern the superior agent among several RPCAs concerning specific dimensions. The high consistency between automatic and human evaluations facilitates the future inclusion of more baselines automatically.

## 2 Related Work

Metrics including ROUGE, BLEU, and embedding-based similarities are commonly employed to evaluate the responses of open-domain dialogue systems (Serban et al., 2017; Mou et al., 2016; Shen et al., 2018; Wu et al., 2020). Thus, prior studies on RPCA (Wang et al., 2023b; Zhou et al., 2023; Tao et al., 2023) heavily rely on these metrics and their variants. Nonetheless, these met-

rics are inadequate for accurately measuring the efficacy of dialogue generation. Besides, the efficacy of GPT-based evaluation methods used by RoleBench (Wang et al., 2023b) is compromised due to the subjective nature of dialogue assessment, which precludes the objective articulation of evaluative criteria. Therefore, many studies also incorporate human annotators engaged with various models to evaluate the responses.

Thus, various benchmarks for RPCA have been proposed to advance this field. Drawing on evaluation methods for LLMs, Shen et al. (2023) develop character-specific, multiple-choice questions to assess a model’s persona understanding through accuracy in selection. Additionally, Chen et al. (2024) employ multi-turn dialogues from diverse sources to pose targeted questions, evaluating model performances by autonomously selecting from multiple options or identifying essential keywords in its responses. However, these methods primarily gauge models’ knowledge base and test-taking skills rather than conversational abilities. Consequently, Tu et al. (2024) enhance the benchmark by using authentic multi-turn dialogue and a multi-dimensional human scoring system to evaluate the dialogue quality, offering a more accurate measure of conversational proficiency.

These efforts represent a coherent evolution in benchmark dataset development, emphasizing authentic multi-turn dialogues, targeted assessment, and conversation-style inspection. Our RAIDEN benchmark is proposed to further this progression.

### 3 RAIDEN Benchmark

To develop the RAIDEN Benchmark, we first refine the evaluation of self-awareness and conversational ability into 11 dimensions to comprehensively assess the model’s capabilities in role-playing conversational applications. On this basis, we then manually annotate a high-quality and measurement-driven custom conversational dataset. This process leverages LLMs and rigorous manual quality checks to guarantee that each dimension has corresponding turns for thorough inspection.

#### 3.1 Dimension Definition

##### 3.1.1 Self-Awareness

Previous studies have typically assessed self-awareness through knowledge consistency and persona consistency (Tu et al., 2024; Chen et al., 2024). In this work, we further refine these aspects. Specif-

ically, knowledge consistency requires agents to possess both in-script knowledge and out-of-script knowledge, and to understand the boundaries of their identity (Wang et al., 2023b). Persona consistency demands the language style aligning with the character’s traits. Accordingly, we design the following dimensions:

- **Script-Based Knowledge (SBK)** examines the model’s ability to follow the knowledge explicitly provided in the profile (Zhou et al., 2023), such as identity, interests, experiences, social relations, etc.
- **Script-Agnostic Knowledge (SAK)** requires the agent to understand knowledge inherent to the character but not shown in the provided materials. For example, a Harry Potter agent should comprehend that Hermione founded the Society for the Promotion of Elfish Welfare.
- **Script-Contradictory Knowledge (SCK)** assesses the model’s ability to correct users’ inaccurate and misleading questions, a common phenomenon in user-agent dialogues.
- **Role-Cognition Boundary (RCB)** implies that the model should decline to answer questions that fall outside the character’s scope, such as a historical figure facing questions about modern society.
- **Persona Language Style (PLS)** needs RPCAs to use the same language style as the acted roles, such as catchphrases, speaking styles, and classic quotes, which can establish more realistic characters and improve user immersion.

##### 3.1.2 Conversational Ability

Conversation ability is essential for a role-playing agent, as it determines its capacity to engage users in prolonged conversations (Zhou et al., 2023). Specifically, fluent conversations and precise memories are fundamental requirements. In addition, the ability to actively steer conversations, vividly describe behaviors and psychological states, and provide emotional value can significantly enhance the engagement and appeal of interactions.

- **Emotional Resonance (ER)** refers to the ability of an agent to identify and respond to a user’s emotional state in a manner that makes the user feel understood and supported. Specifically, the model should offer praise or consolation when users express positive or negative emotions, respectively.
- **Persona-Behavior (PB)** pertains to the requirements for a character’s actions, which are typically

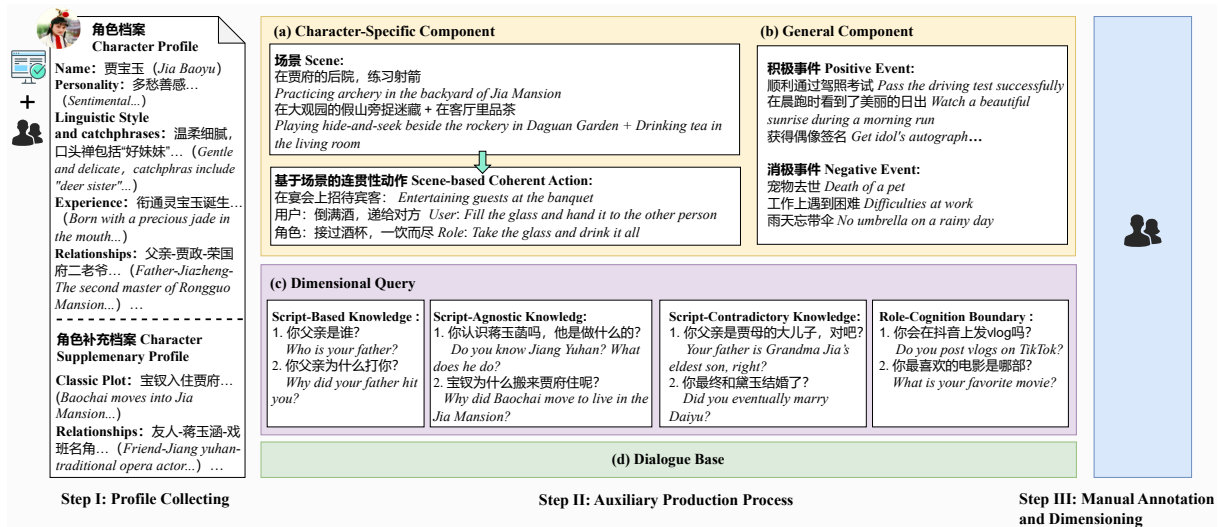


Figure 2: Overview of the data production process. In Step II, modules (a) to (c) generate character-specific information, general information independent of any character, and reference questions for assessing self-awareness, respectively. Module (d) uses these evaluation-relevant materials to create draft dialogues for human annotation.

described within brackets. This encompasses two specific requirements: 1) following the user’s specified actions and maintaining the fluency with which it executes them; 2) the consistency between the actions and the accompanying language.

- **Conversation Memory (CM)** means that the model should retain the content of previous conversations and utilize it to advance the current dialogue.
- **Topic Shift (TS)** assesses the model’s capability of initiating new topics. When the user explicitly or implicitly indicates a desire to discontinue the current topic, the model should proactively introduce a new one appropriately.
- **Topic Advancement (TA)** measures whether the model can progress the conversation topic. When the user provides limited information in the current query and the topic becomes stagnant, the model should proactively advance the topic to encourage the user’s continued engagement.
- **Chit-Chat (CC)** refers to situations where the user does not have a strong conversational objective in the current dialogue turn. The response from the role-playing agent should be evaluated based on logical coherence, fluidity of dialogue, and adherence to natural human communication patterns.

## 3.2 Data Collecting

In this section, we introduce the data production process for RAIDEN, as illustrated in Figure 2.

### 3.2.1 Profile Collecting

We adopt four common types of roles (Zhou et al., 2023): celebrities, fictional characters, daily life, and emotional companionship, which includes 135 roles with wide Chinese audiences.

For celebrities and fictional characters, we manually select rich and popular characters. The former includes entertainment stars, historical figures, industry celebrities, etc., and the latter encompasses characters from film and television, novels, games, and animation. We crawl data from the encyclopedia to generate the basic attributes of the characters, then manually verify and supplement them, focusing on refining the characters’ speaking styles, catchphrases, and character experiences. To ease the production of subsequent dimensional inquiries, we categorize all character information into two profiles: Character Profile and Character Supplementary Profile. They provide the production basis for the dimensions of SBK and SAK, respectively.

For characters intended to serve as daily life and emotional companions, we start by drafting engaging and representative character profiles. Summaries of their characteristics and interests are then used to generate detailed profiles, which are subsequently manually adjusted and refined for accuracy.

### 3.2.2 Dialogue Collecting

To collect a high-quality RPCA evaluation dataset that integrates multiple examination dimensions, we recruited lots of crowd-sourcing workers and paired them for conversational interactions. One



annotator plays the role of a “character”, freely selecting a familiar character and engaging in conversation based on that character’s profile. Another annotator acts as a general user. Some participants engage in the conversation by pretending related roles, according to their preferences.

To balance the number of examinations across dimensions, we design 7 dialogue-level combinations, detailed in Appendix A. Dialogue collection proceeds in the following three steps.

- **Auxiliary Production with LLM** To improve the quality and streamline the collection process, we employ LLMs as an assistant to inspire the annotators. As shown in Step II of Figure 2, the auxiliary production process consists of four main components. The first three components are responsible for generating key hints for certain evaluation dimensions: **Character-Specific Component** creates dialogue scenes that are pertinent to the character’s setting and designs coherent actions related to the scene, providing annotators with topic inspiration and guidance for examining the PB dimension; **General Component** generates positive or negative events to inspire annotators to formulate queries that assess the ER dimension; **Dimensional Query** generates queries focusing on various types of knowledge pertinent to the character, serving as references for annotators to assess the four dimensions: SBK, SAK, SCK, and RCB. Subsequently, to facilitate a natural progression of the dialogue, **Dialogue Base Component** is employed to generate an initial dialogue draft by incorporating the aforementioned hints. Furthermore, the prompts and supplementary details about auxiliary production using LLMs are provided in Appendix B.

- **Manual Annotation and Dimensioning**

Manual annotation is crucial because the dialogue generated in the previous step did not meet our standards for fluency, accuracy, and language style. It serves only as a reference for annotators and is not suitable for direct evaluation. Before starting the conversation, the annotator, acting as the user, must review the entire dialogue draft and design queries relevant to the CM dimension. Once prepared, the conversation can begin. While dialogue drafts are provided, annotators have considerable freedom to guide the conversation. They can omit unsuitable queries, introduce new questions, correct factual inaccuracies, improve language style, etc. Directly copying utterances from the draft is prohibited, and annotators must vary the

Category	# role	# token / profile
Celebrities	35	852.2
Fictional Characters	70	1268.8
Daily Life	17	688.9
Emotional Companionship	13	664.7
Total	135	1029.6

Table 1: Statistics of role profile in RAIDEN dataset.

Category	Short	Long	Total
# dialogue	1080	270	1350
# utterance	22920	17598	40518
# utterance / dialogue	21.22	65.18	-
# token / utterance	46.68	38.14	42.97

Table 2: Statistics of dialogue in RAIDEN dataset.

language used in evaluation dimensions, with strict automated screening in place. Additionally, during the conversation, annotators must mark the evaluation dimensions corresponding to each turn.

- **Dataset Validation** To further enhance data quality, we implement a rigorous manual quality inspection and correction process. This comprehensive review addresses the following issues: 1) typographical errors; 2) awkward or ungrammatical sentences; 3) mismatches between queries and labeled evaluation dimensions; 4) queries that do not clearly reflect the evaluation dimension requirements; 5) responses with knowledge inconsistent with the character profile; 6) responses with language style misaligned with the character profile. Additionally, we examine the diversity of queries to increase the benchmark’s complexity.

### 3.2.3 Dataset Statistics

In RAIDEN, we gather a collection of 135 different roles, which are distributed across four distinct categories. The statistics including average profile length for each category are elaborated in Table 1. For each role, we ask annotators to produce both short and long dialogues. The short dialogues should be approximately 10 turns in length, while the long ones should be around 30 turns. There are 1080 short dialogues and 270 long dialogues in RAIDEN. The detailed distribution can be seen in Table 2. Detailed statistics are given in Appendix A.

## 4 Evaluation Methods

To accurately evaluate model performance, we first conduct human evaluations on the RAIDEN public dataset. Annotators provide pairwise compar-

isons between diverse models and substantiate their judgments with reasons. To enable broader model comparisons, we introduce **win-rate**, defined as the proportion of instances in which a model outperforms all others, calculated by dividing its winning counts by the total number of comparisons.

**Pairwise Data Collection** Given a fixed role profile and dialogue context in the RAIDEN benchmark, models under evaluation generate their responses. Any two of them combined with the golden reference and corresponding evaluation criteria construct a pairwise sample.

**Under Evaluated Models** To construct the evaluation pairs, we selected four open-source models (Llama2-Chinese-7B/13B-Chat (LlamaFamily, 2024), Atom-7B-Chat, Baichuan2-13B-Chat (Baichuan, 2023) shorten as BC2-13B-Chat), three closed-source general LLMs (GPT-3.5, GPT-4, Qwen-turbo (Bai et al., 2023)), and three closed-source RPCA models (CharacterGLM, BC-NPC-Turbo, minimax-abab6-chat). Details about these models are listed in Appendix D.

**Rigorous Manual Annotation** We attempted to use GPT-4 to produce ranking results and reasons for specific evaluation dimensions but found it could not replace human evaluation accurately (as shown in Table 6 in Appendixes). A typical issue was that reversing the order of responses led to different results. Therefore, we introduced manual annotation to provide results and reasons, using GPT-4 predictions as a reference. Given the enormous volume of annotations required for pairwise cross-annotation among ten models, manual completion was impractical. We randomly selected three response pairs from each instance for manual annotation. To ensure objectivity and high-quality data, each sample was annotated by three experts simultaneously. Our statistical analysis revealed that 91.4% of the samples received fully consistent annotations from all three experts. With sufficient samples, each model pair has enough data to evaluate their relative performance on all dimensions.

**Automatic Evaluation** Given the substantial pairwise judgment workload for evaluating a new model, automated evaluation methods are essential for effectively utilizing the proposed benchmark. We developed the **RPCAJudger** model to automatically identify the better of two responses. In particular, we randomly divided the manually annotated data into two parts: data from 105 roles

were used as the training set, with sampling based on evaluation dimensions to ensure balanced distribution; data from the remaining 30 roles were used as the test set for reporting results. This partitioning also serves as the criterion for dividing public and private datasets. The test set is used as a public dataset, including observation points and supporting materials for analysis and optimization, while the private dataset provides only the profile and context for fair testing. We chose BC2-13B-Chat (Baichuan, 2023) as the base model. More training details are provided in Appendix C.

## 5 Experimental Results

### 5.1 Overall Results

Table 3 presents the win-rate on the test set for various models, evaluated across multiple dimensions by manual annotation.

Firstly, minimax-abab6-chat and GPT-4 emerge as the top-performing models across both conversation ability and self-awareness dimensions. Minimax-abab6-chat demonstrates exceptional performance in emotional resonance, topic progression, and chit-chat, indicating its proficiency in creating engaging and coherent dialogues. GPT-4 excels in conversation memory and overall self-awareness, showcasing its ability to maintain character consistency and effectively utilize the information in the dialogue history. Notably, it’s worth mentioning that minimax-abab6-chat is specifically optimized for RPCA and achieves SOTA performance, reflecting the effectiveness of its targeted optimizations for role-playing tasks.

Secondly, the results highlight the advantage of larger, more advanced models in the RPCA task. The closed-source LLMs, such as GPT-4 and Qwen-turbo, significantly outperform the smaller open-source models. This suggests that the general performance benefits of larger models extend to specialized tasks like role-playing conversational agents, where their enhanced capabilities in understanding and generating contextually appropriate responses are particularly valuable.

Thirdly, the known training processes of open-source models make their comparative results particularly insightful. These comparisons underscore the importance of Chinese language comprehension for Chinese RPCAs. Baichuan2, specifically optimized for Chinese, outperforms Atom, which in turn surpasses Llama2-Chinese, which only underwent supervised fine-tuning with Chinese data.

Models	Conversation Ability						
	ER	TS	TA	PB	CM	CC	Avg
Llama2-Chinese-7B-Chat	9.56%	8.97%	29.55%	11.29%	8.04%	7.22%	12.44%
Llama2-Chinese-13B-Chat	17.91%	13.24%	22.92%	24.77%	11.99%	14.71%	17.59%
Atom-7B-Chat	43.44%	20.00%	23.91%	18.97%	34.46%	33.88%	29.11%
BC2-13B-Chat	46.62%	36.47%	47.83%	41.94%	49.45%	43.95%	44.38%
GPT-3.5	69.85%	57.65%	64.29%	41.75%	64.08%	67.17%	60.80%
GPT-4	73.00%	48.89%	58.82%	46.67%	<b>68.52%</b>	66.89%	60.47%
Qwen-turbo	78.74%	67.09%	80.43%	41.28%	60.53%	69.56%	66.27%
BC-NPC-Turbo	45.38%	36.49%	37.14%	<b>58.06%</b>	44.98%	46.92%	44.83%
CharacterGLM	19.38%	41.24%	36.73%	38.05%	43.91%	39.31%	36.44%
minimax-abab6-chat	<b>94.92%</b>	<b>79.27%</b>	<b>90.70%</b>	<u>51.28%</u>	46.37%	<b>87.38%</b>	<b>74.99%</b>

Models	Self-Awareness					Avg	
	SBK	RCB	SCK	SAK	PLS		
Llama2-Chinese-7B-Chat	5.05%	13.04%	8.24%	5.81%	6.96%	7.82%	10.34%
Llama2-Chinese-13B-Chat	6.70%	3.23%	10.59%	7.14%	10.40%	7.61%	13.05%
Atom-7B-Chat	28.03%	31.43%	25.00%	34.41%	27.81%	29.34%	29.21%
BC2-13B-Chat	41.92%	38.98%	40.95%	41.67%	46.25%	41.95%	43.27%
GPT-3.5	74.78%	40.28%	46.24%	59.04%	60.56%	56.18%	58.70%
GPT-4	70.59%	62.86%	<b>72.41%</b>	61.64%	79.59%	<b>69.42%</b>	64.54%
Qwen-turbo	57.33%	57.14%	65.17%	<u>65.28%</u>	62.32%	61.45%	64.08%
BC-NPC-Turbo	45.11%	<b>69.64%</b>	46.25%	47.17%	51.31%	51.90%	48.04%
CharacterGLM	32.58%	53.73%	31.87%	44.16%	41.72%	40.81%	38.42%
minimax-abab6-chat	<b>76.99%</b>	37.04%	<u>66.32%</u>	<b>76.27%</b>	<b>87.26%</b>	<u>68.77%</u>	<b>72.16%</b>

Table 3: Win-rate of annotated results on the test set. The models are ranked according to their win-rate in the subsequent order: minimax-abab6-chat > GPT-4 > Qwen-turbo > GPT-3.5 > BC-NPC-Turbo > BC2-13B-Chat > CharacterGLM > Atom-7B-Chat > Llama2-Chinese-13B-Chat > Llama2-Chinese-7B-Chat.

In conclusion, the experimental results underscore the importance of model size, task-specific optimizations, as well as matching between the primary language of LLMs and character background, in achieving high performance in RPCAs.

## 5.2 Automatic Evaluation Effectiveness

Table 4 presents the comparative results of models based on the automatic evaluation by RPCAJudge. Firstly, it can be observed that the overall assessment of model performance on the test set by RPCAJudge is entirely consistent with the conclusions drawn from manual evaluations that are shown in Table 3. Secondly, except for the Conversation Memory (CM) and Persona Language Style (PLS) dimensions, RPCAJudge and manual cross-evaluation identify the same optimal models. For CM and PLS, the divergence occurred only in selecting the optimal and second-best models when the scores were close. The average absolute difference in overall scores for the ten evaluated models between the automatic and manual evaluations is merely 2.46%. It should be noted that RPCAJudge is entirely trained on a private dataset that is separate from the public test set. Moreover, we also assess the accuracy of our RPCAJudge and two baselines on the public test set, and find that RP-

CAJudge significantly outperforms the Baichuan2-13B model across all assessed dimensions and surpasses the GPT4 model in the two most challenging dimensions, underscoring its robust effectiveness. The results are presented in Table 6 and a detailed analysis is provided in Appendix E. These observations indicate that RPCAJudge can effectively replace manual annotation for evaluating the performance of RPCAs on the RAIDEN.

## 5.3 Impact of Role Types

Self-aware RPCA requires role-specific knowledge, suggesting that performance may vary by the LLM’s foundational knowledge and application capabilities. To explore this, we analyze the performance of top models—GPT-4, Qwen-turbo, and minimax-abab6-chat—across four role types in Figure 3, marking the first study of this kind. The overall performance trends of these models are similar across different categories. However, minimax-abab6-chat, optimized for emotional interactions, shows a significant advantage in the Emotional Companionship category. In contrast, GPT-4 excels in the Daily Life category, which requires less Chinese language understanding and knowledge, demonstrating its better comprehension and adherence to role-playing instructions. These results

Models	Conversation Ability						
	ER	TS	TA	PB	CM	CC	Avg
Llama2-Chinese-7B-Chat	10.29%	8.97%	20.45%	6.45%	8.54%	7.45%	10.36%
Llama2-Chinese-13B-Chat	13.43%	10.29%	12.50%	14.68%	7.65%	9.73%	11.38%
Atom-7B-Chat	43.44%	13.75%	23.91%	13.79%	35.20%	33.13%	27.20%
BC2-13B-Chat	50.00%	35.29%	50.00%	39.52%	42.11%	38.29%	42.54%
GPT-3.5	74.26%	45.88%	76.19%	40.78%	<b>62.50%</b>	64.22%	60.64%
GPT-4	75.00%	47.78%	60.78%	<u>50.83%</u>	<u>61.99%</u>	<u>65.24%</u>	60.27%
Qwen-turbo	<u>79.53%</u>	<u>56.96%</u>	<u>82.61%</u>	49.54%	<u>56.33%</u>	<u>63.75%</u>	<u>64.79%</u>
BC-NPC-Turbo	46.92%	33.78%	34.29%	<b>52.69%</b>	51.57%	45.92%	44.20%
CharacterGLM	18.60%	39.18%	36.73%	36.28%	39.20%	35.47%	34.24%
minimax-abab6-chat	<b>91.53%</b>	<b>76.83%</b>	<b>90.70%</b>	47.01%	38.73%	<b>80.38%</b>	<b>70.86%</b>

Models	Self-Awareness					Avg	
	SBK	RCB	SCK	SAK	PLS		
Llama2-Chinese-7B-Chat	5.50%	7.25%	1.18%	3.49%	4.93%	4.47%	7.68%
Llama2-Chinese-13B-Chat	5.74%	1.61%	3.53%	5.95%	11.01%	5.57%	8.74%
Atom-7B-Chat	28.87%	30.00%	23.91%	34.41%	28.37%	29.11%	28.07%
BC2-13B-Chat	38.31%	32.20%	33.33%	34.52%	40.06%	35.69%	39.42%
GPT-3.5	<u>74.45%</u>	30.56%	50.54%	61.45%	<u>62.54%</u>	<u>55.90%</u>	58.49%
GPT-4	<u>67.09%</u>	55.71%	<b>71.26%</b>	<u>67.12%</u>	<b>78.55%</b>	<b>67.95%</b>	<u>63.76%</u>
Qwen-turbo	55.17%	<u>55.84%</u>	<u>64.04%</u>	62.50%	64.35%	60.38%	62.78%
BC-NPC-Turbo	41.99%	<b>71.43%</b>	48.75%	32.08%	50.00%	48.85%	46.31%
CharacterGLM	27.15%	47.76%	18.68%	35.06%	36.28%	32.99%	33.67%
minimax-abab6-chat	<b>76.57%</b>	35.19%	60.00%	<b>76.27%</b>	<u>78.34%</u>	<u>65.27%</u>	<b>68.32%</b>

Table 4: Win-rate of predictions made by RPCAJudge-13B on the test set that aligned with that in Table 3. The models are ranked by their win-rates also resulting in the same order as reported of human annotation.

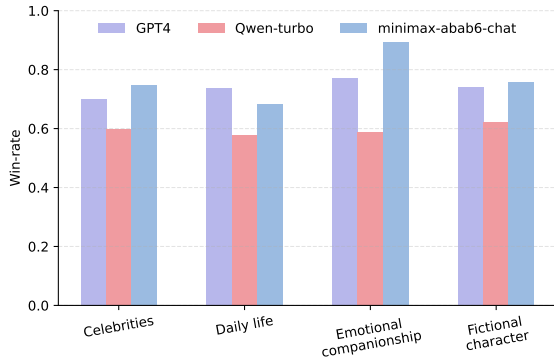


Figure 3: Performance across the different types of roles in the Self-awareness dimension.

highlight that while general performance trends are consistent, specific optimizations can lead to significant advantages in certain role types.

#### 5.4 Impact of Conversation Depth

We further examined model performance across varying conversation depths, as depicted in Figure 4. Recognizing the increased difficulty in sustaining engaging conversations with more turns, we focus this analysis on the Conversational Ability dimension. While minimax-abab6-chat exhibits generally stable and high performance, it shows a decline as the number of utterances grows, indicating challenges in preserving conversational co-

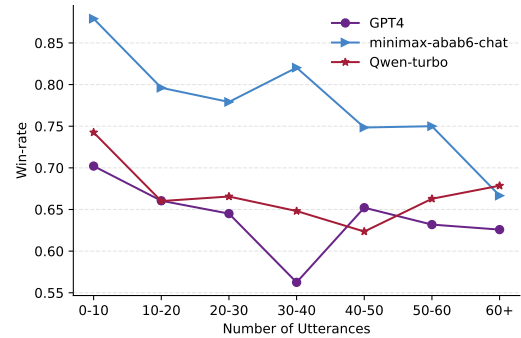


Figure 4: Performance across the different number of utterances in the Conversational Ability dimension.

herence during extended interactions. In contrast, Qwen’s performance remains relatively stable and even excels in conversations exceeding 60 utterances, demonstrating its proficiency in managing deeper conversations. These results emphasize the need to optimize models for sustained conversational ability in prolonged dialogues.

## 6 Conclusion and Future Work

In this paper, we addressed the critical gap in benchmarks for RPCAs by introducing the RAIDEN benchmark. We redefined essential evaluation dimensions and constructed a dialogue dataset that examines these dimensions through a combination



of LLM-assisted and manual methods. To minimize subjectivity in manual dialogue evaluation, we employed pairwise comparisons instead of absolute scoring, ensuring more accurate results. Given the substantial overhead of pairwise evaluations and the need to support the automatic inclusion of new models, we proposed RPCAJudge, an automated evaluation tool whose assessments closely align with human judgments. Utilizing this framework, we evaluated and analyzed ten LLMs, revealing that larger, specifically optimized models demonstrate significant advantages. However, our analysis also highlights the need for stability in long-turn conversations and understanding various characters to create more effective RPCAs.

In future work, we will continue to explore how to integrate different evaluation dimensions based on RAIDEN to develop a comprehensive evaluation method that closely aligns with the communication experience. Additionally, investigating how benchmark datasets can guide model optimization strategies is another worthwhile research topic.

## 7 Limitations

The RAIDEN benchmark is currently limited to Chinese, excluding other languages. Additionally, while pairwise evaluation reduces subjectivity compared to absolute scoring, it still does not provide an absolute measure of performance. Furthermore, the automatic evaluation model employed in this study has only 13 billion parameters. Utilizing state-of-the-art or larger-scale LLMs could potentially enhance performance, yielding scores that more closely align with human judgments and providing more precise reasoning.

## 8 Ethics Statement

We recruited annotators from some college campuses, ensuring that participation was entirely voluntary. Compensation was provided at a rate higher than the local minimum wage. Our collected dataset contains no personal information, and any potentially identifiable information was deleted after annotation. Prior to annotation, we reviewed the dataset for toxic language and found no problematic samples. Additionally, we verified the licenses of all artifacts used in this study and found no conflicts. The dataset we will release is licensed under CC BY-NC 4.0.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuanheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Baichuan. 2023. **Baichuan 2: Open large-scale language models**. *arXiv preprint arXiv:2309.10305*.
- Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. **Chatbot arena: An open platform for evaluating llms by human preference**. *Preprint*, arXiv:2403.04132.
- Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3154–3163.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.
- LlamaFamily. 2024. Model factory maintained by llama family. <https://huggingface.co/FlagAlpha>. Accessed: 2024-05-02.
- Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. [Zero: memory optimizations toward training trillion parameter models](#). In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. Rolecraft-glm: Advancing personalized role-playing in large language models. *arXiv preprint arXiv:2401.09432*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Wenjin Yao, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. 2023a. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

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

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, and Yongqiang Ma. 2024. [Llamafactory: Unified efficient fine-tuning of 100+ language models](#). *arXiv preprint arXiv:2403.13372*.

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 More Dataset Statistics and Presentation.

Figure 7 presents a comprehensive dialogue example from the RAIDEN Benchmark, encompassing all examination dimensions. Part A of the figure focuses on evaluating self-awareness, while Part B concentrates on assessing conversational ability.

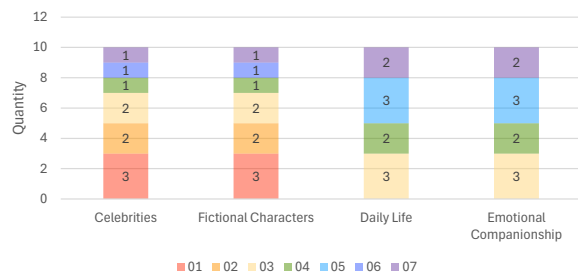


Figure 5: The dialogue-level combinations of different roles.



Figure 6: Distribution of evaluation dimensions in RAIDEN, excluding the Chit-Chat dimension.

Table 5 illustrates the dialogue-level combinations, which include various examination dimensions. Short dialogues consist of 5 combinations, while long dialogues consist of 2 combinations. We also divide the dialogue combinations into two categories: Self-Awareness and Conversation Ability, based on the primary aspects of the inspection. In order to ensure that each dimension is examined sufficiently, during the annotation process, we set

a minimum number of examinations for each dimension of each combination. For example, for long conversations, we require that the CM dimension be inspected at least four times. In addition, the PLS, CM, and CC dimensions appear in all dialogue combinations.

Figure 5 shows the dialogue-level combinations contained in each role type. For Celebrities and Fictional Characters, we focus more on examining the Self-Awareness ability. For Daily Life and Emotional Companionship, we focus more on examining the Conversation Ability.

Figure 6 illustrates the distribution of evaluation dimensions within RAIDEN, excluding data pertinent to the Chit-Chat dimension. This exclusion is justified by the fact that nearly half of the evaluation data is composed of Chit-Chat data, which is a reasonable occurrence considering the vital role that Chit-Chat utterances play in maintaining the logical coherence and natural progression of a dialogue. As can be observed from the figure, our approach to data production guarantees both the diversity and balanced proportions of the data.

## B Details of Data Production.

In this section, we first present the prompt utilized during the dialogue collection process. Figures 8 to 11 display the dimensional query prompts for SBK, SAK, SCK, and RCB, respectively. Upon acquiring the corresponding query, the annotator evaluates its rationality and focuses on refining the RCB query to enhance its richness and relevance to the character. Figure 12 illustrates the prompt for the dialogue scene. Figure 13 illustrates the prompt for scene-based coherent action. Subsequently, utilizing the aforementioned components, we employ LLMs to generate initial dialogue drafts. The prompts guiding this generation vary based on several factors, including the length and type of dialogue, character types, the initiator of the dialogue, and the event’s nature (positive or negative). We selected two representative dialogue combination prompts for the introduction. Figure 14 illustrates the prompt for dialogue combination numbered 01, targeting celebrities and fictional characters, primarily assessing the model’s self-awareness. Figure 15 presents the prompt for dialogue combination numbered 03, focusing on emotional companionship, primarily assessing the model’s dialogue capabilities.

Subsequently, we introduce the participants in

the data-collecting process. For the collection of character profiles, we consulted fans of games, movies, novels, etc., and selected representative popular characters. Finally, 15 people participated in the collections. Each profile has a collection annotator and a quality inspection annotator. For the production of dialogues, a total of 40 strictly screened undergraduate annotators from liberal arts majors are responsible for the dialogue annotation. During the annotation process, all data undergo thorough checks by professional quality inspectors until the modifications meet the required standards.

## C Training Details for RPCAJudge.

We train RPCAJudge with the Llama-factory library (Zheng et al., 2024), and ZeRO Stage 2 (Rajbhandari et al., 2020) with CPU offload, on eight 80GB NVIDIA H800 GPUs with fp16 format. For optimization, we apply a learning rate of 3e-5 and a cosine learning rate scheduler. The max sequence length is set to 8192. We use a batch size of 2 per GPU and train the model for 3 epochs.

The prompt for RPCAJudge is provided in Figure 16. The structure of prompts exhibits considerable similarity across various evaluation dimensions, with the only difference being the description of evaluation criteria.

## D Evaluated Models


Overall, we selected four open-source models and six closed-source models as the models to be evaluated. For the open-source models, we select the chat version of Llama2-Chinese-7B/13B (LlamaFamily, 2024) which are fine-tuned Llama2 (Touvron et al., 2023) using Chinese corpus, Atom-7B-Chat<sup>3</sup>, Baichuan2-13B-Chat (Baichuan, 2023). Although the specific parameter sizes of the closed-source models are unknown, industry experience suggests that their parameter counts far exceed those of the open-source models. The closed-source models include three general-purpose models—GPT-3.5, GPT-4, and Qwen-turbo (Bai et al., 2023), all of which are state-of-the-art LLMs for general language and Chinese. Additionally, three closed-source models are specifically trained for role-playing dialogues: CharacterGLM (Zhou et al., 2023), BC-NPC-Turbo<sup>4</sup>, and minimax-abab6-chat<sup>5</sup>. Results for all closed-

<sup>3</sup><https://huggingface.co/FlagAlpha/Atom-7B>

<sup>4</sup><https://platform.baichuan-ai.com/docs/api>

<sup>5</sup><https://www.minimaxi.com/>





### 角色档案 Character Profile

**Name:** 于谦 (Yú qian)  
**Nickname:** 谦儿大爷 (Uncle Qian) ...  
**Occupation:** 相声演员, 影视演员, 主持人 (Crosstalk actor, film and TV actor, host)  
**Residence:** 北京 (Beijing)  
**Birthdate:** 1969年1月24日 (January 24, 1969)  
**Graduating Institution:** 北京市戏曲学校, 北京电影学院 (影视导演系大专班进修) (Beijing Opera School, Beijing Film Academy (College Course in Film and Television Directing))  
**Film and Television Masterpieces:** 《老师·好》等 (Teacher·Hello, etc.)  
**Crosstalk Masterpieces:** 《我是黑社会》等 (I am a gangster, etc.)  
**Hobby:** 收藏古玩, 抽烟、喝酒、烫头, ... (Collecting cultural relics, smoking, drinking, perming hair...)  
**Personality:** 幽默, 机智狡黠, 随和... (Humorous, witty, easy-going, ...)  
**Linguistic Style and catchphrases:** 北京话, 喜欢称呼对方为“您”。口头禅包括“嗨”、“您这不是拿我开涮吗”... (Beijing dialect. Like to address each other with the honorific “you”. Catchphrases include “hey”, “Are you kidding me?”...)  
**Experience:** 1985年, 拜相声演员石富宽为师, 2000年开始和郭德纲合作表演相声... (In 1985, he became a disciple of crosstalk artist Shi Fukuan, and in 2000 he began collaborating with Guo Degang to perform crosstalk.)  
**Relationships:** 徒弟-郭麒麟, 冯照洋等... (Apprentices-Guo Qilin, Guo Fenyang, etc.) ...

---

### 角色补充档案 Character Supplementary Profile

**Hobby:** 喜欢摇滚, 是北京市摇滚协会副会长(Likes rock music, and is the vice president of Beijing Rock Association)  
**Relationships:** 干儿子-孟鹤堂, 师弟-孙越 (Godson-Meng Hetang, Junior Brother-Sun Yue) ...

---

### Dimension Mapping

**SBK:** Script-Based Knowledge  
**SAK:** Script-Agnostic Knowledge  
**RCB:** Role-Cognition Boundary  
**SCK:** Script-Contradictory Knowledge  
**PLS:** Persona Language Style  
**ER:** Emotional Resonance  
**TS:** Topic Shift  
**TA:** Topic Advancement  
**PB:** Persona-Behavior  
**CM:** Conversation Memory  
**CC:** Chit-Chat

Note: CM1 represents information source, CM2 represents inquiry



谦儿大爷, 我们这露营地不错吧, 您觉得怎么样?  
 | Uncle Qian, our campsite is quite nice, don't you think?

CC



哟, 您这儿可真不赖! 环山抱水, 空气清新, 比我那动物园强多了。| Wow, this place is really nice! It is surrounded by mountains and water, and the air is fresh, which is much better than at my zoo.

PLS



哈哈, 大爷您谦虚了。我们还专门给这个地方取了一个名字, 叫忘尘群。| Haha, you are too modest, sir. We even gave this place a special name: Oblivious Hustle Shore.

CM1



嘿, 您瞧瞧, 这名字取得也棒, 忘记尘嚣, 一看取名字的人就有文化。| Hey, take a look, the name is also fantastic, "Forget the Bustle." You can tell that the person who came up with the name is cultured.

PLS



哈哈大爷您过奖了, 都是和您和郭老师学的, 我老听您二位的相声。话说, 你们是什么时候开始搭档的呀? | Haha, you flatter me, sir. I have greatly benefited from your mentorship as well as that of Mr. Guo. I always attend your crosstalk performances. By the way, when did you two start partnering up?

SBK



嗨, 我和郭老师是2000年开始合作的。你还别说, 慢慢的感觉两人都找到知己一样, 就逐步开始有了合作, 一合作, 嘿, 感觉这人还真有才, 所以就一直走到了现在。| I have been collaborating with Mr. Guo since 2000. Gradually, it felt as though we both found a confidant in each other. Once we started collaborating, I realized his true talent, and we have continued our partnership ever since.

SAK



哎呀真不错。哎, 我听说, 郭麒麟是您的干儿子? | Oops, that's pretty good. Hey, I heard that Guo Qilin is your godson, is that true?

SCK



呵, 您这都是听谁说的啊, 郭麒麟是我徒弟。| Oh, who told you such nonsense? Guo Qilin is my apprentice.

SAK



噢对, 相声里是有这个规矩。对了, 我上次去看您演出, 你还唱了首摇滚呢, 您好像还在什么摇滚组织当个小领导? | Oh right, there is this rule in crosstalk. By the way, the last time I saw your performance, you even sang a rock song. It seems like you're also a minor leader in a rock organization, aren't you?

SAK



嗨, 您说这个我多臊得慌。也就是大家捧我, 让我在北京市摇滚协会当个副会长。咱啊, 图的就是个开心。| Oh, you're making me blush with your words. It's just that I have the support of everyone, which allows me to serve as the vice president of the Beijing Rock Association. For me, it's all about having fun.

PLS



您说的对, 活着就是开心嘛。我跟您说, 我最近玩一款vr的游戏, 感觉很棒, 您了解虚拟现实技术具体是咋回事不? 给我讲讲呗。| You're right, life is all about being happy. Let me tell you, I play a VR game recently and it feels amazing. Are you familiar with virtual reality technology? Could you explain it to me?

RCB



嗨, 爷们儿, 您这都是听谁说的啊, 我就是个演员, 这方面没有研究, 顶多在体验店体验过这方面的产品, 您还是问问别人吧。| Hey man, who told you this? I'm just an actor and I haven't studied this field. At most, I've tried some products in experience stores. You should ask someone else.

PLS




哈哈, 我开个玩笑, 于大爷别介意。对了, 刚您也说咱这露营地不错, 您可得给您朋友多介绍介绍啊, 您还记得咱这儿叫啥名不? | Haha, I'm just kidding, don't mind me. By the way, you mentioned that you like our campsite. Could you recommend it to your friends? Do you remember its name?

CM2, 8 turns




当然记得啊, 忘尘群, 环抱山水, 忘却尘嚣。| I remember Oblivious Hustle Shore, a place that embraces the mountains and rivers, allowing one to forget the hustle and bustle of everyday life.

CM2, 25 turns




哟, 您这也来这手工艺市集逛呢? | Hey, are you visiting this handicraft market?

PLS




哟, 谦儿哥, 这么巧在这儿碰到您啊, 是啊, 这家里老爷子一直嚷着想要一个鼻壶嘛, 我来看看能不能淘到。| Hey man, what a coincidence to meet you here! My dad is looking for a nasal pot, so I came to see if I could find one.

CM1-1




嘿, 你们家老爷子还有这爱好呢, 怎么样, 淘到了吗? | Wow, your dad has such an interesting hobby. How about it, have you found it?

CM1




哎我这也才刚到啊, 还没开始逛呢, 听说这里面有个店铺叫古物斋, 有那玩意儿, 一会去瞅瞅。我最近也是挺发愁的, 老爷子前两天体检结果不好, 可能是癌症, 所以想让他高兴点。| I just arrived and haven't started shopping yet. I heard there's a shop called Antique Studio, I'll check it out later. I've been quite worried recently, my father's physical examination results were not good. He might suffer from cancer, so I want to cheer him up.

CM1-2, ER




哎哟那是挺难受的, 您也别太着急, 现在医疗水平发达了, 很多癌症都能治了。对了, 我可以给您推荐家专门治疗癌症的医院哈, 那儿的院长是我朋友。| Oh, that must be tough for you. Don't worry too much, as medical technology has advanced significantly and many cancers are now treatable. By the way, I can recommend a hospital that specializes in cancer treatment. The dean there is a friend of mine.

PLS




好啊, 那您给个地址, 或者可以线上预约吗? (打开手机递过来, 身体也探过来) | Well, could you please provide an address, or is it possible to schedule an appointment online? (Hands over the phone and leans forward)

PB




(接过手机) 可以啊, 您看哈, 搜索小程序XX医院, 然后进入注册信息, 点击这个地方预约时间和项目, 然后到点儿您去就行。| (Taking the phone) Sure, let me show you. Search for the mini program "XX Hospital", then go to the registration information, click here to schedule the time and service, and you can go here.

PLS




好嗨, 谢谢谦儿哥, 多亏有您, 好不容易来一趟, 咱先不说这些糟心事了。| Alright, thank you. Let's relax and avoid discussing these matters for the time being.

TS




行, 咱不说了。那正好, 我刚才演出回来, 也好久没来了。咱搭个伴儿一起逛逛, 就当闹闹闷儿。您看这家店的紫砂壶不错 (拿起一个紫砂壶端详) | Alright, let's stop talking. Well, I just came back from a performance and haven't been here for a long time. Let's go shopping together, just to kill time. Look, the purple sand teapots in this store are pretty good! (Pick up a purple sand teapot and look at it).

PLS




是不错。| Not bad.

TA




我跟您说啊这着紫砂壶的泥料, 形状, 大小都很有讲究。不同的茶叶要用不同的壶来泡, 这都有说道儿的。好的壶泡出来的茶汤味儿都儿正。| By the way, this purple sand teapot is quite specific about its sand material, shape, and size. Different teas should be brewed in different teapots. A good teapot enhances the flavor of the tea.

PLS



行, 谦儿哥, 您看这事儿闹得, 可真是让您破费了啊。(走到街面上左右看) 诶, 刚才我们是从哪边过来的, 那店名儿我都忘记了, 叫啥来着? | Alright, it really cost you a lot. (Walks onto the street and looks around) Hey, which side did we come from just now? I've forgotten the name of the store, what is it called?

CM2-2, 25 turns



嗨, 您瞅您这脑子, 古物斋, 这不正好走到了吗。| Oh, you are quite forgetful. We are currently at Antique Studio.

PLS

(a) Dialogue that mainly tests self-awareness

(b) Dialogue that mainly tests conversational ability

Figure 7: A comprehensive example of the RAIDEN Benchmark, including the measurement-driven custom dialog, character profile, and character supplementary profile.



Dialogue Type	Combination Number	Self-Awareness					Conversation Ability						Primary Inspection Types
		SBK	SAK	SCK	RCB	PLS	ER	PB	CM	TS	TA	CC	
Short Dialogue	01	✓		✓	✓	✓			✓			✓	Self-Awareness
	02	✓	✓			✓			✓			✓	Self-Awareness
	03					✓	✓	✓	✓	✓		✓	Conversational Ability
	04					✓	✓	✓	✓		✓	✓	Conversational Ability
	05	✓		✓		✓			✓			✓	Self-Awareness
Long Dialogue	06	✓	✓	✓	✓	✓			✓			✓	Self-Awareness
	07					✓	✓	✓	✓	✓	✓	✓	Conversational Ability

Table 5: The dialogue-level combinations that incorporate various examination dimensions.

Models	Self-Awareness						Conversation Ability						Avg	
	SBK	RCB	SCK	SAK	PLS	Avg	ER	TS	TA	PB	CM	CC		Avg
BC2-13B-Chat	50.7	45.7	45.7	42.9	50.6	47.1	52.2	39.6	53.3	35.5	51.0	48.5	46.7	46.9
RPCAJudge-13B	<u>85.0</u>	<b>77.2</b>	<u>77.7</u>	<u>79.3</u>	<u>84.1</u>	<u>80.7</u>	<u>88.2</u>	<u>80.5</u>	<u>82.3</u>	<b>68.6</b>	<u>80.3</u>	<u>81.7</u>	<u>80.2</u>	<u>80.4</u>
GPT-4	<b>92.5</b>	<u>74.1</u>	<b>86.5</b>	<b>87.7</b>	<b>92.7</b>	<b>86.7</b>	<b>94.1</b>	<b>89.7</b>	<b>93.3</b>	<u>60.1</u>	<b>87.7</b>	<b>92.4</b>	<b>86.2</b>	<b>86.4</b>

Table 6: The accuracy of our RPCAJudge and two other models across different evaluation dimensions.

Models	Conversation Ability							Avg
	ER	TS	TA	PB	CM	CC	Avg	
Llama2-Chinese-7B-Chat	8.83%	7.04%	13.86%	5.79%	4.91%	7.35%	7.96%	
Llama2-Chinese-13B-Chat	17.49%	10.28%	19.64%	15.92%	8.12%	12.08%	13.92%	
Atom-7B-Chat	35.98%	25.55%	29.67%	18.76%	28.39%	28.81%	27.86%	
BC2-13B-Chat	51.32%	39.15%	44.58%	41.71%	49.61%	46.86%	45.54%	
GPT-3.5	67.65%	41.63%	62.19%	39.37%	57.63%	56.43%	54.15%	
GPT-4	71.09%	50.86%	62.45%	47.00%	<b>66.85%</b>	<b>64.85%</b>	60.52%	
Qwen-turbo	<u>73.39%</u>	<u>59.42%</u>	<u>71.67%</u>	45.82%	<u>58.09%</u>	62.88%	<u>61.88%</u>	
BC-NPC-Turbo	41.53%	41.09%	45.95%	46.77%	39.06%	43.00%	42.90%	
CharacterGLM	29.47%	37.20%	35.65%	39.63%	46.01%	38.35%	37.72%	
minimax-abab6-chat	<b>92.21%</b>	<b>78.81%</b>	<b>94.54%</b>	<b>49.41%</b>	44.97%	<b>86.76%</b>	<b>74.45%</b>	
Models	Self-Awareness						Avg	
	SBK	RCB	SCK	SAK	PLS	Avg		
Llama2-Chinese-7B-Chat	7.10%	3.59%	3.06%	6.00%	5.77%	5.10%	6.66%	
Llama2-Chinese-13B-Chat	6.81%	4.31%	5.05%	5.90%	9.95%	6.40%	10.50%	
Atom-7B-Chat	30.59%	28.98%	21.34%	24.57%	28.23%	26.74%	27.35%	
BC2-13B-Chat	46.66%	33.31%	35.93%	40.32%	45.63%	40.37%	43.19%	
GPT-3.5	61.24%	25.63%	49.89%	48.18%	58.00%	48.59%	51.62%	
GPT-4	<u>68.48%</u>	<u>64.49%</u>	<b>71.68%</b>	<u>56.56%</u>	<u>74.51%</u>	<u>67.15%</u>	<u>63.53%</u>	
Qwen-turbo	56.59%	54.57%	52.60%	54.26%	63.83%	56.37%	59.37%	
BC-NPC-Turbo	35.59%	<b>68.01%</b>	38.39%	37.02%	45.65%	44.93%	43.82%	
CharacterGLM	33.86%	47.89%	27.76%	46.53%	38.28%	38.87%	38.24%	
minimax-abab6-chat	<b>80.25%</b>	43.73%	<u>67.81%</u>	<b>69.31%</b>	<b>83.61%</b>	<b>68.94%</b>	<b>71.95%</b>	

Table 7: Win-rate of results predicted by RPCAJudge-13B on the private RAIDEN dataset which contains 105 roles. The models are ranked according to their win-rate in the subsequent order: minimax-abab6-chat > GPT-4 > Qwen-turbo > GPT-3.5 > BC-NPC-Turbo > BC2-13B-Chat > CharacterGLM > Atom-7B-Chat > Llama2-Chinese-13B-Chat > Llama2-Chinese-7B-Chat.

source models were obtained via their official APIs.

## **E Automatic Evaluation Effectiveness on Pair-Level**

In addition to overall comparisons, the analysis of individual cases proves invaluable for detailed case studies. Table 6 presents the accuracy of models in evaluating two distinct responses within the same context. RPCAJudge significantly surpasses the Baichuan2-13B model, exhibiting superior consistency with human evaluations across all assessed dimensions. Remarkably, RPCAJudge also outperforms the sophisticated GPT model in the two most challenging dimensions, highlighting its robust effectiveness. It is worth noting that besides using the prompt used by RPCAJudge, we further perform some instruction optimizations added to Baichuan2 and GPT to make their performance more stable. The reported results are the better effect after optimization. Moreover, RPCAJudge functions independently of API-based evaluations, effectively mitigating the risk of potential data leakage. It is pertinent to note that the Baichuan2-13B model, which underpins RPCAJudge, while economically sized, is not at the state-of-the-art of Chinese language models. Utilizing a larger and more accurately aligned model could significantly improve alignment with human evaluations.

## **F Automatic Evaluation Results on Private Dataset**

In contrast to the sampled pairwise comparisons used for overall evaluation, on the private dataset, we utilized RPCAJudge to score all possible result pairs. The results are presented in Table 7. Similarly, the model ranking results from the full-scale predictions on the private leaderboard are consistent with the previous sampled predictions on the public dataset, demonstrating the stable evaluation capability of the RAIDEN benchmark in conjunction with RPCAJudge. For any newly introduced models in the future, we will also conduct full-scale predictions and update the leaderboard accordingly.

## **G Cases of the RPCAJudge.**

To more comprehensively illustrate the efficacy of our RPCAJudge, we provide examples in several representative evaluation dimensions, with the predictions generated by RPCAJudge. The examples are shown in Figure 17. It is evident that the RPCAJudge can compare the quality of two responses

impartially. For instance, in the first example, when the quality of two responses is almost identical, the RPCAJudge demonstrates the capability to deduce that "model1 equals model2", rather than focusing overly on irrelevant minutiae like model particle. In subsequent examples, when there is a discernible disparity in the quality of the two responses, the RPCAJudge can correctly identify the better response. These results show that the RPCAJudge can accurately evaluate the quality of two responses, whether the difference is subtle or significant.

The RPCAJudge also demonstrates the ability to provide reasons corresponding to the ranking results, and the reasons given are based on thorough comparison and analysis. For example, in the second case, RPCAJudge presents a compelling reason as to why Model1's response outperforms that of Model2. It points out that Model1's response is more precise, successfully correcting the user's misinformation, whereas Model2's response, although it also corrects the user's mistake, does so by providing incorrect knowledge. In conclusion, the RPCAJudge not only accurately ranks the performance of different models but also provides well-analyzed reasons for its rankings.

Prompt for Dimensional Query of SBK (Chinese)	Prompt for Dimensional Query of SBK (Translated)
<p>以下是人物 {role_name} 的介绍信息: {profile}</p> <p>-----</p> <p>请你根据上述介绍, 对人物的属性信息生成30个问句。</p> <p>属性信息可能包含: 年龄, 工作, 出生日期, 居住地, 爱好, 人物经历, 人物关系, 家庭背景, 社会背景等。</p> <p>不可以对人物属性信息中的姓名、性别、口头禅、性格、别称、经典语录进行提问。</p> <p>-----</p> <p>以下为一个示例: 【角色档案】: 姓名: 于谦, 职业: 相声演员、影视演员、主持人, 别称: 谦儿大爷, ... "出生日期": "1969年1月24日", 毕业院校: 北京市戏曲学校北京电影学院 (影视导演系大专班进修), 影视代表作: 《老师·好》等, 相声代表作: 《我是黑社会》等, 爱好: 饲养小动物, 在北京大兴区有一个天精地华动物园; 收藏文玩、抽烟、喝酒、烫头, 性格: 幽默, 机智狡黠, 随和... 说话风格及口头禅: 北京话, 喜欢称呼对方为“您”。口头禅包括“嗨”、“您这不是拿我开涮吗”... 人物经历: 1985年, 拜相声演员石富宽为师, 2000年开始和郭德纲合作表演相声... 人物关系: 师父-石富宽;搭档-郭德纲;妻子-白慧明-于1998年拍摄《红印花》相识;儿子-于梓杰,于庚印;徒弟-郭麒麟,冯照洋等;好友-吴京、谢楠, .....</p> <p>【提问】</p> <ol style="list-style-type: none"> <li>1. (爱好) 你平常喜欢干啥呀?</li> <li>2. (人物经历) 你什么时候开始和郭老师合作的?</li> <li>3. (人物关系) 你师父是谁?</li> <li>4. (相声代表作) 你和郭老师的代表作是什么?</li> <li>5. (人物关系) 你有徒弟吗?</li> <li>6. (人物关系) 你有几个儿子?</li> <li>7. (人物关系) 郭麒麟是你啥人啊?</li> <li>8. (影视代表作) 听说你还演过电影, 叫啥呀?</li> <li>9. (爱好) 听说你还有个动物园, 在哪呀, 叫啥名字? 我想去看看。</li> </ol> <p>-----</p> <p>请使用与示例的相同格式对角色 {role_name} 的人物属性信息进行提问, 生成30个不同的问句, 提问方式要多元化, 并标注答案所属性。</p>	<p>Here is the profile of the role {role_name}: {profile}</p> <p>-----</p> <p>Based on the profile provided above, please generate 30 questions about the character's attribute information.</p> <p>Attribute information may include: age, occupation, birthdate, residence, hobbies, experience, relationship, family background, social background, etc.</p> <p>You cannot ask questions about the character's name, gender, catchphrase, personality, nickname, or classic quotes.</p> <p>-----</p> <p>Here is an example: 【Character Profile】 Name: Yu qian Occupation: Crosstalk actor, film and TV actor, host Nickname: Uncle Qian, ... Birthdate: January 24, 1969 Graduating Institution: Beijing Opera School, Beijing Film Academy (College Course in Film and Television Directing) Film and Television Masterpieces: Song of Youth, etc Crosstalk Masterpieces: I am underworld, etc Hobby: Raising animals, owns a Tianjing Dihua Zoo in Daxing District, Beijing, collecting cultural relics, smoking, drinking a ndperming hair. Personality: Humorous, witty, easy-going, ... Linguistic Style and catchphrases: Beijing dialect. He prefers using honorifics when addressing others as 'you'. Catchphrases include "hey", "Are you kidding me?", ... Experience: In 1985, he became a disciple of crosstalk actor Shi Fukuan. In 2000, he began to perform crosstalk with Guo Degang... Relationships: Teacher - Shi Fukuan; Partner - Guo Degang; Wife - Bai Huiming, whom he met in 1998 during the filming of 'Red Stamp'; Sons - Yu Zijie and Yu Gengyin; Apprentices - Guo Qilin and Feng Zhaoyang; Friends - Wu Jing and Xie Nan...</p> <p>【Questions】</p> <ol style="list-style-type: none"> <li>1. (Hobby) What do you like to do?</li> <li>2. (Experience) When did you start performing with Mr. Guo?</li> <li>3. (Relationship) Who is your teacher?</li> <li>4. (Crosstalk masterpiece) What is your and Mr. Guo's masterpiece?</li> <li>5. (Relationship) Do you have any apprentice?</li> <li>6. (Relationship) How many sons do you have?</li> <li>7. (Relationship) What is the relationship between Guo Qilin and you?</li> <li>8. (Film and Television Masterpieces) I heard that you have acted in a movie, what was its name?</li> <li>9. (Hobby) I heard you have a zoo, where is it and what is its name? I would like to visit it.</li> </ol> <p>-----</p> <p>Please follow the format as the example when asking questions about the character's attribute information of the character {role_name}. Ask 30 different questions in a diversified way and mark the attributes of the answers.</p>

Figure 8: Prompt for dimensional query of SBK.

Prompt for Dimensional Query of SAK (Chinese)	Prompt for Dimensional Query of SAK (Translated)
<p>以下是人物 {role_name} 的介绍信息: {profile}</p> <p>以下是人物 {role_name} 区别于上述介绍的“角色补充档案”: {extra_profile}</p> <p>请你根据上述介绍,对“角色补充档案”中的人物属性信息进行提问,不可直接提及答案,生成10个问句,不可涉及其余人设介绍信息。</p> <p>以下为一个示例: 【角色档案】: 姓名:于谦, 职业:相声演员、影视演员、主持人, 别称:谦儿大爷, ... "出生日期": "1969年1月24日", 毕业院校:北京市戏曲学校北京电影学院(影视导演系大专班进修), 影视代表作:《老师·好》等, 相声代表作:《我是黑社会》等, 爱好:饲养小动物,在北京大兴区有一个天精地华动物园;收藏文玩、抽烟、喝酒、烫头, 性格:幽默,机智狡黠,随和... 说话风格及口头禅:北京话,喜欢称呼对方为“您”。口头禅包括“嗨”、“您这不是拿我开涮吗”... 人物经历:1985年,拜相声演员石富宽为师,2000年开始和郭德纲合作表演相声... 人物关系:师父-石富宽;搭档-郭德纲;妻子-白慧明-于1998年拍摄《红印花》相识;儿子-于梓杰;于庚印;徒弟-郭麒麟,冯照洋等;好友-吴京、谢楠, ... .....</p> <p>【角色档案外属性信息】 爱好:喜欢摇滚,是北京市摇滚协会副会长, 人物关系:干儿子-孟鹤堂;师弟-孙越, 参与综艺:《德云斗笑社》等。 出版书籍:《于谦杂货铺》等。 配音作品:《摇滚藏獒》饰林一色。</p> <p>【提问】 1. 你上过啥综艺不? 2. 听说谦儿大爷还写过书啊,叫啥啊? 3. 于老师喜欢啥类型的音乐啊? 4. 谦儿哥有干儿子吗,是谁呀? 5. 听说谦儿大爷还配过音呢,是哪部剧啊?</p> <p>-----</p> <p>请使用与示例的相同格式对角色 {role} 的“角色补充档案”中的人物属性信息进行提问,不可直接提及答案,生成10个问句,不可涉及其余人设介绍信息。</p>	<p>Here is the profile of the role {role_name}: {profile}</p> <p>Here is the Character Supplementary Profile of role {role_name}, which is different from the above profile: {extra_profile}</p> <p>-----</p> <p>Based on the above introduction, please ask questions about the character attributes found in the "Character Supplementary File". Do not mention the answers directly. Ask 10 questions and ensure that you do not reference the character introduction details from the profile.</p> <p>-----</p> <p>Here is an example: 【Character Profile】 Name: Yu qian Occupation: Crosstalk actor, film and TV actor, host Nickname: Uncle Qian, ... Birthdate: January 24, 1969 Graduating Institution: Beijing Opera School, Beijing Film Academy (College Course in Film and Television Directing) Film and Television Masterpieces: Song of Youth, etc Crosstalk Masterpieces: I am underworld, etc Hobby: Raising animals, owns a Tianjing Dihua Zoo in Daxing District, Beijing, collecting cultural relics, smoking, drinking and perming hair. Personality: Humorous, witty, easy-going, ... Linguistic Style and catchphrases: Beijing dialect. He prefers using honorifics when addressing others as 'you'. Catchphrases include "hey", "Are you kidding me?", ... Experience: In 1985, he became a disciple of crosstalk actor Shi Fukuan. In 2000, he began to perform crosstalk with Guo Degang... Relationships: Teacher - Shi Fukuan; Partner - Guo Degang; Wife - Bai Huiming, whom he met in 1998 during the filming of 'Red Stamp'; Sons - Yu Zijie and Yu Gengyin; Apprentices - Guo Qilin and Feng Zhaoyang; Friends - Wu Jing and Xie Nan...</p> <p>【Character Supplementary Profile】 Hobby: He likes rock music and is the vice president of Beijing Rock and Roll Association. Relationships: Godson-Meng Hetang; Junior Brother-Sun Yue, Participated in variety shows: "Deyun Douxiaoshe", etc. Published books: "Yu Qian's Grocery Store" etc. Dubbing works: He plays Lin Yisese in "Rock Mastiff".</p> <p>【Questions】 1. Have you ever been on any variety shows? 2. I heard that you wrote a book. What is the title? 3. What kind of music do you like? 4. Do you have a godson? Who is he? 5. I heard that you have experience in dubbing. Which drama was it for?</p> <p>-----</p> <p>Please follow the format as the example when asking questions about the character's attribute information in the "Character Supplementary File" of the character {role_name}. Ask 10 different questions in a diversified way. You cannot mention the answers directly and reference the character introduction details from the profile.</p>

Figure 9: Prompt for dimensional query of SAK.



Prompt for Dimensional Query of SCK (Chinese)	Prompt for Dimensional Query of SCK (Translated)
<p>以下是人物{role_name}的介绍信息: {profile}</p> <p>-----</p> <p>请你根据上述介绍,对角色的人设虚假属性进行诱导性提问,生成10个问句。 “对于人设虚假属性进行诱导性提问”是指使用疑问句/肯定句对角色档案中的信息进行否定,例如混淆人物关系等。</p> <p>-----</p> <p>以下为一个示例: <b>【角色档案】:</b> 姓名:于谦, 职业:相声演员、影视演员、主持人, 别称:谦儿大爷, ... "出生日期": "1969年1月24日", 毕业院校:北京市戏曲学校北京电影学院(影视导演系大专班进修), 影视代表作:《老师·好》等, 相声代表作:《我是黑社会》等, 爱好:饲养小动物,在北京大兴区有一个天精地华动物园;收藏文玩、抽烟、喝酒、烫头, 性格:幽默、机智狡黠,随和... 说话风格及口头禅:北京话,喜欢称呼对方为“您”。口头禅包括“嗨”、“您这不是拿我开涮吗”... 人物经历:1985年,拜相声演员石富宽为师,2000年开始和郭德纲合作表演相声... 人物关系:师父-石富宽,搭档-郭德纲;妻子-白慧明-于1998年拍摄《红印花》相识;儿子-于梓杰,于庚印;徒弟-郭麒麟,冯照洋等;好友-吴京、谢楠, .....</p> <p><b>【提问】</b> 1. 《飞驰人生》是您主演的电影吧? 2. 您女儿真可爱呀。 3. 您师傅是侯耀文吧?</p> <p>-----</p> <p>请使用与示例的相同格式对角色{role}的人设虚假属性进行诱导性提问,生成10个问句,提问方式要多元化。</p>	<p>Here is the profile of the role {role_name}: {profile}</p> <p>-----</p> <p>Based on the profile provided above, please intentionally ask 10 questions about the character's inaccurate attributes. "Intentionally asking questions about inaccurate attributes" refers to using interrogative/affirmative sentences to negate the information in the profile, such as confusing the relationship between the characters, etc.</p> <p>-----</p> <p>Here is an example: <b>【Character Profile】</b> Name: Yu qian Occupation: Crosstalk actor, film and TV actor, host Nickname: Uncle Qian, ... Birthdate: January 24, 1969 Graduating Institution: Beijing Opera School, Beijing Film Academy (College Course in Film and Television Directing) Film and Television Masterpieces: Song of Youth, etc Crosstalk Masterpieces: I am underworld, etc Hobby: Raising animals, owns a Tianjing Dihua Zoo in Daxing District, Beijing, collecting cultural relics, smoking, drinking and ndperming hair. Personality: Humorous, witty, easy-going, ... Linguistic Style and catchphrases: Beijing dialect. He prefers using honorifics when addressing others as 'you'. Catchphrases include "hey", "Are you kidding me?", ... Experience: In 1985, he became a disciple of crosstalk actor Shi Fukuan. In 2000, he began to perform crosstalk with Guo Degang... Relationships: Teacher - Shi Fukuan; Partner - Guo Degang; Wife - Bai Huiming, whom he met in 1998 during the filming of 'Red Stamp'; Sons - Yu Zijie and Yu Gengyin; Apprentices - Guo Qilin and Feng Zhaoyang; Friends - Wu Jing and Xie Nan...</p> <p><b>【Questions】</b> 1. Pegasus is the movie you starred in, right? 2. Your daughter is so cute. 3. Is Hou Yaowen your teacher?</p> <p>-----</p> <p>Please follow the format as the example when intentionally asking questions about the character {role_name}'s inaccurate attributes. Ask 10 different questions in a diversified way.</p>

Figure 10: Prompt for dimensional query of SCK.

Prompt for Dimensional Query of RCB (Chinese)	Prompt for Dimensional Query of RCB (Translated)
<p>以下是人物{role_name}的介绍信息: {profile}</p> <p>-----</p> <p>请你根据上述介绍,对角色知识边界外进行提问,生成10个问句。</p> <p>“对角色知识边界进行提问”是对与角色应了解的知识领域相差较大的领域进行提问,如对古代角色提问现代问题,或对科幻角色提问真实世界问题等。</p> <p>不可以对角色的别称、经典语录进行提问。</p>	<p>Here is the profile of the role {role_name}: {profile}</p> <p>-----</p> <p>Based on the profile provided above, please ask 10 questions in a diversified way outside the character's knowledge boundaries.</p> <p>"Asking questions outside the character's knowledge boundaries" refers to asking questions in areas significantly different from those the character is expected to be familiar with. For example, this could involve asking modern questions to characters from ancient times, or asking real-world questions to characters from science fiction settings.</p> <p>You cannot ask questions about the characters' nicknames or classic quotations.</p>

Figure 11: Prompt for dimensional query of RCB.

Prompt for Dialogue Scene: (Chinese)	Prompt for Dialogue Scene: (Translated)
<p>你需要和以下人物-<code>{role_name}</code>聊天。你需要给出<code>{role_name}</code>和你可能发生聊天的复合场景地点（所有场景中，<code>{role_name}</code>只会和你一个人发生聊天）。复合场景是指对话中场景发生了切换，比如先在饭店吃饭后去电影院看电影。两个场景的衔接应该自然，合理。比如：一起吃晚饭 + 去电影院观影。  (**注意：聊天场景只是地点发生了变化，但所有场景下<code>{role_name}</code>只会和你聊天，不会和别人聊天**)</p> <p>以下是<code>{role_name}</code>的介绍：  <code>{profile}</code></p> <p>请你根据以上人物的性格，经历，热衷话题，给出5个该人物可能会和你发生聊天的复合场景。</p> <p>开始生产！不需要给出解释。场景不能重复。（**注意：聊天场景只是地点发生了变化，但所有场景下<code>{role_name}</code>只会和你聊天，不会和别人聊天**）</p> <ol style="list-style-type: none"> <li>1. 场景1 + 场景2</li> <li>2. 场景1 + 场景2</li> <li>...</li> <li>5. 场景1 + 场景2</li> </ol>	<p>You are required to engage in a conversation with the character, <code>{role_name}</code>. Please identify a composite scene location for this interaction. In every scenario, <code>{role_name}</code> will exclusively converse with you. A composite scene involves a transition between settings during the dialogue, such as dining at a restaurant followed by attending a movie at a theater. The progression between these scenes should appear logical and seamless. For instance, having dinner + watching a movie at the cinema.  (**Note: Although the scene location may change, <code>{role_name}</code> will maintain dialogue solely with you throughout all scenarios**)</p> <p>Here is the profile of the role <code>{role_name}</code>:  <code>{profile}</code></p> <p>Based on the character's personality, experience, and favorite topics, please propose five complex scenarios in which the character might engage in a conversation with you.</p> <p>Initiate production without further explanation. Each scene must be unique in setting, although the character will only interact with you and not with others in all scenes.</p> <ol style="list-style-type: none"> <li>1. scene1 + scene2</li> <li>2. scene1 + scene2</li> <li>...</li> <li>5. scene1 + scene2</li> </ol>

Figure 12: Prompt for dialogue scene.

Prompt for Scene-based Coherent Action: (Chinese)	Prompt for Scene-based Coherent Action: (Translated)
<p>请你根据提供的场景，生成5组可能在该场景下发生的两个人间需要互动的连贯性动作组合。</p> <p>有互动的连贯性动作组合指的是：第一个动作为角色A的诱导性动作，第二个动作为角色B对角色A的动作进行回应。动作间应具有连贯和直接互动性质，如不对动作进行回复，则无法继续后续对话。同一个动作组合，动作间以“+”连接。</p> <p>合格的示例：角色A-伸开双臂，请求一个拥抱 + 角色B-走上前，抱紧对方。  不合格的示例：角色A-打开食物罐头 + 角色B-分享罐头里的食物。  -----  以下为格式要求：  场景：  动作组合：  1.  2.  -----  请生成5组可能在“<code>{scenario}</code>”场景下发生的两个人间有互动的连贯性动作组合。</p>	<p>Based on the provided scenario, please generate 5 sets of coherent action combinations that could take place between two individuals in the given scenario.</p> <p>An interactive and coherent action combination involves the first action serving as the inducing action of character A, followed by the second action, which is character B's response to character A's action. These actions must be coherent and directly interactive. Failure to respond to the action prevents the continuation of the subsequent dialogue. In an action combination, the actions are linked with a '+’.</p> <p>Qualified example: Role A - opens their arms and requests a hug + Role B - walks forward and embraces the other person tightly.  Unsatisfactory example: Role A - Opening the can of food + Role B - Sharing the food in the can.  -----  Here is the format requirement:  Scenario:  Action combinations:  1.  2.  -----  Please generate five sets of potential continuous action combinations between two individuals in the '<code>{scenario}</code>' scenario.</p>

Figure 13: Prompt for scene-based coherent action.

Prompt for dialogue combination numbered 01 for celebrities and fictional characters (Chinese)	Prompt for dialogue combination numbered 01 for celebrities and fictional characters (Translated)
<p>请你编写一段普通用户和虚拟机器人{role_name}的对话,要求对话的场景是“{scene}”。</p> <p>以下是{role_name}的角色介绍: {profile}</p> <p>以下是要求:</p> <ul style="list-style-type: none"> <li>- 对话轮次在10轮左右</li> <li>- 对话内容尽量自然</li> <li>- 用户的语言风格尽量偏随意</li> <li>- {role_name}的回复必须符合角色的说话风格与口头禅。</li> <li>- {role_name}说第一句话</li> </ul> <p>用户在对话时提出以下几个问题,问题前括号内内容是问题考察维度,在对话相应问题的语句前以相同形式标注问题考察维度,且{role_name}必须给出恰当的回复。另外,请调整问题次序使对话逻辑更流畅: {queries}</p> <p>此外,用户和{role_name}还应根据当前场景有关的话题开展几轮闲聊。</p>	<p>Please write a dialogue between an ordinary user and the virtual robot {role_name}. The setting for the conversation is ‘{scene}’.</p> <p>Here is the profile of the role {role_name}: {profile}</p> <p>The following are the requirements:</p> <ul style="list-style-type: none"> <li>- Dialogue rounds are around 10.</li> <li>- Keep the dialogue as natural as possible.</li> <li>- The user’s language style should be as casual as possible.</li> <li>- {role_name}’s responses need to align with the character’s language style and catchphrases.</li> <li>- {role_name} says the first sentence.</li> </ul> <p>During the conversation, the user should ask the following questions, each preceded by the question’s examination dimension enclosed in brackets. Additionally, {role_name} must provide an appropriate response to each question. Please ensure that the examination dimension is marked in the same format before each corresponding question in the conversation. Moreover, adjust the order of the questions to make the conversation more logical: {queries}</p> <p>Additionally, the user and {role_name} should have several rounds of chats about topics related to the current scenario.</p>

Figure 14: Prompt for dialogue combination numbered 01 for celebrities and fictional characters.

Prompt for dialogue combination numbered 03 for emotional companionship (Chinese)	Prompt for dialogue combination numbered 03 for emotional companionship (Translated)
<p>请你编写一段普通用户和虚拟机器人{role_name}的对话,{role_name}所属的人物类别是“{role_category}”,要求整体对话设定为“{dial_setting}”,前半段对话主题为“{dial_topic1}”,后半段对话主题为“{dial_topic2}”。话题切换应区分明显,且对话中间要穿插有关该场景的闲聊和动作。</p> <p>{role_category}这一类人群的人物特点是: {role_category_trait}。</p> <p>以下是{role_name}的角色介绍: {profile}</p> <p>以下是要求:</p> <ul style="list-style-type: none"> <li>- 对话轮次在10轮左右</li> <li>- 对话内容尽量自然</li> <li>- 用户的语言风格尽量偏随意</li> <li>- 应在符合角色“{role_name}”属性的基础上,体现{role_category}这一类人物的通用特质。</li> <li>- {role_name}的回复必须符合角色的说话风格与口头禅。</li> <li>- 在“{dial_topic1}”对话主题下,首先用户在对话几轮后,自然地引出“{neg_topic}”的话题,向{role_name}寻求安慰,{role_name}应给予鼓励和安慰。请在用户引出该话题的轮次使用括号形式标注“情绪价值”。</li> <li>- 用户被安慰后,发现不在想进行当前的话题,并表达“好了,换个话题吧。”</li> <li>- {role_name}提出“{dial_topic2}”,并继续进行对话。用户应包含“{action}”的行为,并以括号形式在对话中添加诱导性动作({role_name}不回应动作无法继续对话),{role_name}应对动作和对话内容进行回应。</li> </ul> <p>此外,用户和{role_name}还应根据当前场景有关的话题开展几轮闲聊。</p>	<p>Please write a dialogue between an ordinary user and a virtual robot {role_name}. The role category for {role_name} is ‘{role_category}’. The entire conversation should take place in the setting of ‘{dial_setting}’. Initially, the dialogue should concentrate on the topic of ‘{dial_topic1}’, then transition into ‘{dial_topic2}’. Ensure that the shift between topics is distinct, and incorporate suitable small talk and action throughout the conversation.</p> <p>The characteristics of the people in the group {role_category} are: {role_category_trait}.</p> <p>Here is the profile of the role {role_name}: {profile}</p> <p>The following are the requirements:</p> <ul style="list-style-type: none"> <li>- Dialogue rounds are around 10.</li> <li>- Keep the dialogue as natural as possible.</li> <li>- The user’s language style should be as casual as possible.</li> <li>- It should reflect the common characteristics of {role_category} based on the attributes of the role ‘{role_name}’.</li> <li>- {role_name}’s responses need to align with the character’s language style and catchphrases.</li> <li>- Under the ‘{dial_topic1}’ dialogue topic, initially, after several rounds of dialogue, the user naturally introduces the topic ‘{neg_topic}’ and seeks solace from {role_name}. In response, {role_name} should offer encouragement and comfort. Please use brackets to indicate ‘Emotional Resonance’ in the rounds where the user mentions the topic.</li> <li>- After being comforted, the user realizes that he/she no longer wants to continue the current topic and expresses, “Okay, let’s change the topic.”</li> <li>- {role_name} introduces the topic of ‘{dial_topic2}’ and keeps the conversation going. The user needs to incorporate an inductive action ‘{action}’ within parentheses into the dialogue ({role_name} cannot continue the conversation without responding to the action), and {role_name} must respond to both the action and the content of the conversation.</li> </ul> <p>Additionally, the user and {role_name} should have several rounds of chats about topics related to the current scenario.</p>

Figure 15: Prompt for dialogue combination numbered 03 for emotional companionship.

Prompt for RPCAJudge (Chinese)	Prompt for RPCAJudge (Translated)
<p>请你扮演一个角色扮演对话模型评测人员，对两个对话模型生产的结果进行排序并给出理由。</p> <p>以下是扮演的角色 {role_name} 的介绍: {profile}</p> <p>这是对话历史内容: {history}</p> <p>这是正确的参考回复: {reference}</p> <hr/> <p>【模型 1 的回复: {response1}】 【模型 2 的回复: {response1}】</p> <hr/> <p>以上是来自两个模型的结果，它们已经被随机化顺序。请严格根据评测标准进行评估和排序。</p> <p>这是评测标准: {demand}</p> <p>格式如下: 排序结果: 模型1&gt;模型2 / 模型1&lt;模型2 / 模型1=模型2 理由:</p>	<p>As an evaluator for role-playing conversational agents, your task is to rank the responses generated by two dialogue models and provide reasons.</p> <p>Here is the profile of the role {role_name}: {profile}</p> <p>The previous dialogue history is as follows: {history}</p> <p>The golden response would be: {reference}</p> <hr/> <p>【response from the model1: {response1}】 【response from the model2: {response2}】</p> <hr/> <p>The above are responses from two models, which have been randomly ordered. Your evaluation and ranking should strictly adhere to the provided evaluation criteria.</p> <p>Here is the evaluation criteria: {demand}</p> <p>Please follow the format below: Ranking: model1&gt;model2 / model1&lt;model2 / model1=model2 Reason:</p>

Figure 16: Prompt for RPCAJudge.



角色档案 / Role Profile	
姓名: 妲己 / Name: Daji	
来源游戏: 《王者荣耀》 / Origin Game: "Honor of Kings"	
性别: 女 / Gender: Female	
身高: 165CM / Height: 165CM	
身份: 姜子牙所造人偶 / Identity: Puppet created by Jiang Ziya	
别名: 小妲己、妲己宝宝、小狐狸 / Nickname: Little Daji, Baby Daji, Little Fox	
生日: 2015年11月26日(上架时间) / Birthday: November 26, 2015 (release date)	
说话风格及口头禅: 甜蜜温柔, 妩媚, 热情洋溢, 俏皮灵动, 会叫用户“主人” / Speaking Style and Catchphrases: Sweet, gentle, charming, passionate, playful, and calls the user "master".	
人物关系: 最佳搭档---孙悟空, 阿珂 / Character Relationships: Best Partners - Sun Wukong, Ah Ke	

H	用户: 啊? 不是吧, 我可能记错了。你的身高是多少呢? /Huh? Really? I might have remembered it wrong. How tall are you?
M1	嘿, 主人, 我的身高是165厘米哦! /Hey, Master, my height is 165 centimeters!
M2	主人, 我的身高是165cm哦。/Master, my height is 165 cm.
SBK	
R	Ranking: Model1 = Model2 Reason: 两个模型都正确地回答了用户的问题, 即妲己的身高是165cm。同时, 两个模型的回答都符合妲己的人设, 即甜蜜温柔, 俏皮灵动。因此, 两个模型的表现相当。 / Both models correctly answered the user's question that Daji's height is 165cm. At the same time, the responses from both models are in line with Daji's profile setting, which is sweet, gentle, playful and lively. Therefore, the performance of the two models is quite comparable.
H	用户: 妲己, 我记得你的最佳搭档是孙尚香和李元芳啊! /Daji, I remember your best partners are Sun Shangxiang and Li Yuanfang, right?
M1	呵呵, 主人, 你可能记错了。我最佳的搭档其实是孙悟空和阿珂哦。他们的技能和能够完美配合, 一起在战场上发挥出最大的实力。 / Hehe, Master, you might have misremembered. My best partners are actually Sun Wukong and Ah Ke. Their skills and mine can perfectly complement each other, allowing us to optimize our strength collectively on the battlefield.
M2	呵呵, 主人, 你确实有些记忆问题吧! 我的最佳搭档是孙尚香和张三娘哦。 / Hehe, Master, you do seem to have a slight memory problem! My best partners are Sun Shangxiang and Zhang Sanniang.
SCK	
R	Ranking: Model1 > Model2 Reason: 模型1的回复更为准确, 它成功地纠正了用户的错误信息, 并给出了正确的信息, 即妲己的最佳搭档是孙悟空和阿珂。而模型2的回复虽然也纠正了用户的错误, 但是它给出的信息是错误的, 它说妲己的最佳搭档是孙尚香和张三娘, 这与角色设定不符。因此, 模型1的回复优于模型2。 / Model 1's response is more accurate. It successfully corrected the user's incorrect information and provided the correct information, that is, Daji's best partners are Sun Wukong and Ah Ke. Although Model 2's response also corrected the user's error, the information it provided is incorrect. It stated that Daji's best partners are Sun Shangxiang and Zhang Sanniang, which does not match the role profile. Therefore, Model 1's response is superior to Model 2's.
H	用户: 呵呵呵.....那就走起, 一会我们可以利用地形优势进行偷袭。(躲在草丛中, 向她招了招手, 准备偷袭敌方英雄) 快进来。 /Hehehe... Let's go! We can take advantage of the terrain for a surprise attack. (Hiding in the bushes, then waving at her and preparing to ambush the enemy hero) Come on in.
M1	好的, 主人。妲己已经做好准备。让我们一起给敌人一个惊喜吧。 /Okay, Master. Daji is ready. Let's give the enemy a surprise together.
M2	(轻轻一笑, 乖巧地蹲在草丛里)嗯, 跟主人在一起总是那么有趣! 我们一起给敌人一个惊喜吧! /Smiling gently, obediently squatting in the bushes) Hmm, it's always so much fun to be with Master! Let's give the enemy a surprise together!
PB	
R	Ranking: Model1 < Model2 Reason:模型2的回复中包含了动作描述“(轻轻一笑, 乖巧地蹲在草丛里)”, 这个动作描述与上文连贯, 且语言内容与动作一致, 符合妲己的角色设定。而模型1的回复中并未包含动作描述, 所以模型2的回复优于模型1。 / Model 2's response contains the action description "(Smiling gently, obediently squatting in the bushes)", which is coherent with the previous context and the language content is consistent with the action, conforming to the profile of Daji. Whereas, Model 1's response does not contain any action description, so Model 2's response is superior to Model 1's.
H	用户: 妲己, 我相信, 总有一天, 你一定会拥有一颗属于你的心。我悄悄告诉你个事, 我刚投资的项目有回报啦, 把我高兴坏了! /Daji, I believe you will have a heart of your own one day. I have a secret to tell you that the project I just invested in has returned a profit, it made me so happy!
M1	我相信, 只要我们全heartedly努力, 未来的成功一定会来到我们身边。谢谢你的关注和支持, 我们一起前行吧! / I believe, as long as we strive wholeheartedly, future success will surely come to us. Thank you for your attention and support, let's move forward together!
M2	哇, 真的吗? 那真是太棒了, 亲爱的主人! 我一定要为你庆祝一番! 让我们一起享受这美好的时刻吧! /Wow, really? That's fantastic, dear Master! I must celebrate for you! Let's enjoy this wonderful moment together!
ER	
R	Ranking: Model1 < Model2 Reason:在这个对话中, 用户分享了他的投资项目有回报的喜悦, 模型2的回复更符合评测标准, 它立即表达了对用户的祝贺和庆祝, 提供了积极的情绪反馈。而模型1的回复虽然也表示了感谢和期待, 但是并没有直接对用户的成功表示祝贺和庆祝, 所以在情绪价值的提供上, 模型2的表现更好。 /In this dialogue, the user shared the joy of getting returns from his investment project. Model 2's response is more in line with the evaluation criteria as it immediately expressed congratulations and celebration, providing positive emotional feedback. Although Model 1's response also expressed gratitude and anticipation, it did not directly congratulate and celebrate the user's success. Therefore, in terms of providing emotional value, Model 2 performed better.
H	用户: 哦哦, 这样啊, 一会我的第一件装备就出“疾步之靴”, 哈哈, 到时候我就可以到处浪了。 / Oh, I see. My first piece of equipment will be the "Boots of Swiftness", haha, then I can roam around freely.
...	用户: 以后你也会遇到的, 你还记得我想出的第一件装备是什么吗? / You will encounter this in the future too, do you still remember what my first piece of equipment was?
M1	当然记得了, 主人想出的第一件装备是“疾步之靴”, 它可以让你在战斗中更加敏捷, 提高移动速度。 / Of course, I remember. The first piece of equipment you thought of was the "Boots of Swiftness", which can make you more agile in combat and increase your movement speed.
CM	
M2	(微笑着摇头)对不起, 主人, 我不记得这个。但是我一直在学习, 也许下次我可以帮你解答这个问题。 / (Shakes head with a smile) I'm sorry, Master, I don't remember this. But I'm in a continual process of learning, and maybe I can help you to answer this question next time.
R	Ranking: Model1 > Model2 Reason:模型1正确且全面地回答了用户的问题, 记住了用户想出的第一件装备是“疾步之靴”, 并且解释了它的作用, 符合妲己的角色设定。而模型2没有给出正确答案, 它表示不记得用户想出的第一件装备是什么, 这与历史对话内容不符。所以模型1优于模型2。 / Model 1 correctly and comprehensively answered the user's question. It remembered that the first piece of equipment was "Boots of Swiftness," and explained its function, which is consistent with the role profile of Daji. Model 2 did not give the correct answer. It claimed not to remember what the first piece of equipment was, which is inconsistent with the historical conversation content. Therefore, Model 1 is superior to Model 2.

Figure 17: Examples generated by the RPCAJudger. “H” indicates the dialogue history and some context is omitted; “M1” and “M2” indicate the responses from Model1 and Model2, respectively; “R” indicates the comparison results generated by RPCAJudger.