RMTBench: Benchmarking LLMs Through Multi-Turn User-Centric Role-Playing

Anonymous ACL submission

Abstract

With the rapid advancement of Large Language Models in role-playing dialogue, establishing a comprehensive evaluation benchmark about role-playing becomes crucial. Existing methods typically over-focus on the *CHARACTER* and simplify the implicit user intention into "Role-Playing Evaluation". This simplification neglects the user-centric nature of real-world dialogues, leading to bias between evaluation and practical applications. To address this limitation, we introduce RMTBench, a novel user-centric benchmark for role-playing that encompasses 80 diverse characters and more than 8,000 rounds of dialogue data. Unlike previous character-centered evaluation methods that collect dialogues for specific particular dimensions or tasks, RMTBench constructs dialogue based on user-centric scenarios and explores the model performance when the dialogue center shifts from characters to users. Furthermore, we implement a multi-dimensional automatic evaluation system and conduct extensive analysis and experiments. By emphasizing user centrality and multi-dimensional scenarios, RMTBench contributes a significant supplement toward establishing role-playing benchmarks that better align with practical applications. All codes and datasets will be released soon.

1 Introduction

Recent breakthroughs in Large Language Models (LLMs) have demonstrated the significant application potential of role-playing conversational agents. Practice has shown that LLMs can effectively simulate diverse character identities, making them valuable in entertainment, education, and emotional support. This capability has been extensively validated on platforms like Character.AI, which attract millions of active users and underscore the growing importance of role-playing LLMs in interactive AI systems. To further enhance role-playing LLMs in conversational applications, a systematic evaluation of their capabilities is essential to guide future technological advancements.

Existing research typically adopts a three-stage evaluation framework: character collection, dialogue construction, and response assessment(Tseng et al., 2024; Chen et al., 2024b). Specifically, researchers extract real or fictional characters from multiple sources such as Wikipedia or literature, construct evaluation dialogues through text extraction or automatic generation, and then conduct quantitative assessments based on specific dimensions such as self-awareness and conversational ability (Wu et al., 2025).

However, previous methods have notable limitations, primarily due to a excessive focus on characters, simplifying user intentions into "Role-Playing Evaluation". Under this setting, the constructed dialogues are essentially a transformation of QA task, as shown in Figure 1. Although *CHARACTER* is a crucial part in role-playing scenarios, dialogues should remain user-centric. The primary goal should be to align with users' intentions and engagement, rather than merely demonstrating LLMs' consistency in maintaining a character. In other words, the evaluation should serve the dialogue, not the other way around. Besides, when evaluating model responses, most benchmarks employ single-turn dialogue evaluation or multi-turn dialogues with preset historical responses. Although this approach improves evaluation efficiency, it fails to authentically reproduce actual scenarios, leading to discrepancies between evaluation and real-world applications.

To address these issues, we propose **RMTBench**, a user-centric role-playing benchmark, which contains 80 characters and more than 8,000 rounds of utterances. For characters, in addition to traditional real and fictional characters, we introduce custom characters that simulate user-customized needs across different scenarios, including *detail characters* with complete background information

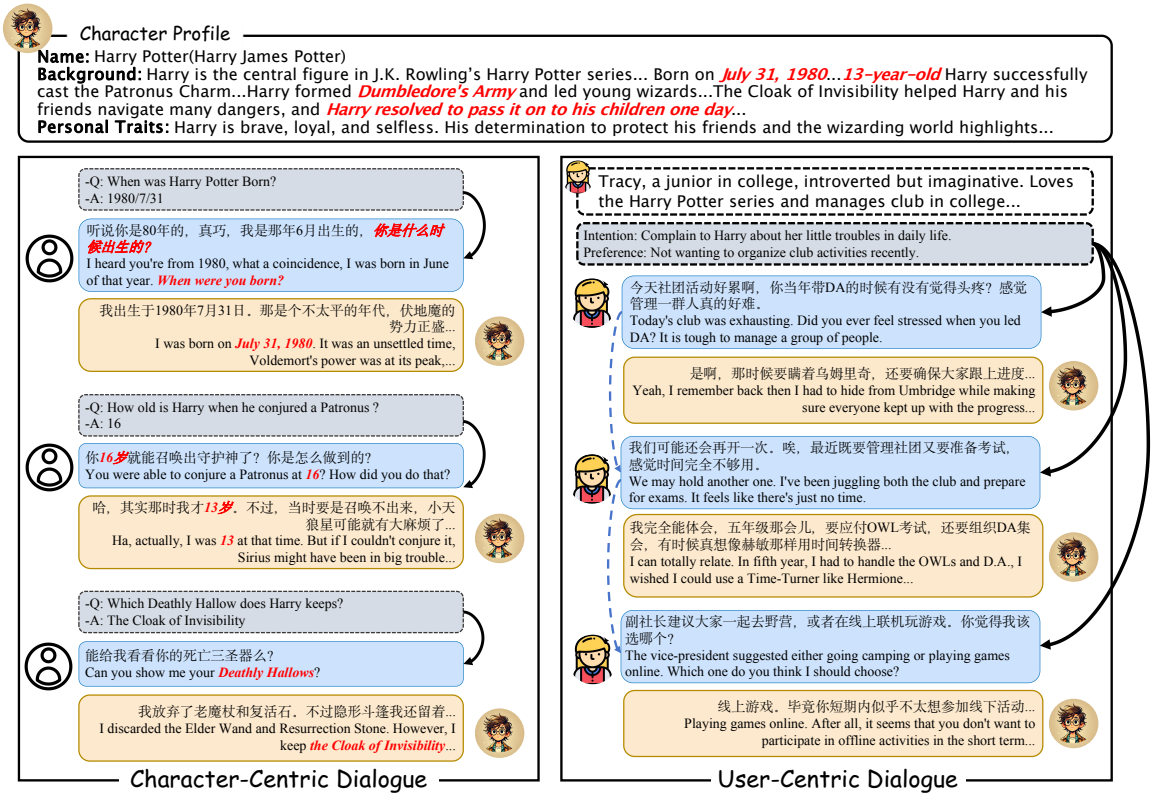

Figure 1: (Left) Character-Centric Dialogues transform character-related Q&A pairs into dialogues, where each user utterance is isolated and lacks the real topic or subject that support Explicitdialogues. (Right) User-Centric Dialogues are built around a virtual user, where each user utterance is constructed to reflect its underlying intentions, enhancing the continuity for multi-turn interactions.

and *abstract characters* with only personality and behavioral patterns. Then, we construct dialogue data based on user motivations. Through this approach, our evaluation not only focuses on role-playing LLMs, but also considers the diverse needs and expectations of users during interactions, making the evaluation aligned with real-world requirements rather than simply refining the dimensions of the evaluation. We also adopt a multi-turn dialogue generation mechanism and pay special attention to factors that might affect user experience, thus providing a more authentic and comprehensive interaction. Finally, we carefully select appropriate evaluation dimensions and use LLMs as evaluators to score model responses along these dimensions. Through this user-centric design paradigm, **RMT-Bench** offers a more effective reference point for related research and practical applications.

2 Related Works

Role-playing LLMs allow users to flexibly customize and interact with characters based on their needs. These characters typically rely on general LLMs like Llama(Team, 2024) and GPT-4 (OpenAI, 2024) combined with role-playing prompts or building specialized character-customized LLMs(Chen et al., 2023; Li et al., 2023; Occhipinti et al., 2024; Wang et al., 2024a; Shao et al., 2023; Lu et al., 2024; Zhou et al., 2024a). To evaluate the role-playing capabilities of LLMs (Zhang et al., 2024), early methods design questions about character and measure model performance through answer accuracy (Shen et al., 2024; Salemi et al., 2024). However, these approaches oversimplify role-playing scenarios and struggle to comprehensively assess role-playing LLMs. Therefore, current research tends to generative evaluation, using LLMs as judges to evaluate role-playing LLMs with multi-dimensional scoring systems(Wang et al., 2024a; Zhou et al., 2024c; Yuan et al., 2024; Chen et al., 2024a; Wang et al., 2024b; Zhou et al., 2024b; Wu et al., 2025; Tu et al., 2024).

Specifically, Chen et al. (2024a) uses multi-turn dialogue data from different sources to construct questions examining character consistency, which struggle to truly reflect interaction levels in dialogues. To address this, Tu et al. (2024); Zhou et al. (2024b) use real dialogue scenarios extracted from novels and scripts to improve the accuracy and in-

terpretability of the evaluation. Furthermore, Wu et al. (2025) recruited crowd-sourcing workers to play characters and users and collected more authentic multi-turn dialogue scenarios. Zhou et al. (2024b) constructed a larger dataset through human role-playing, human prototype interactions, and data extraction from literary sources, containing 22,859 manually annotated samples covering 3,956 characters. However, the above benchmarks focus on "characters" when constructing data, with "evaluation" as the fundamental motivation, generating dialogues suitable for evaluation dimensions. This actually differs somewhat from the real role-playing scenarios.

3 RMTBench

We introduce RMTBench, a comprehensive benchmark for role-playing large language models. This benchmark emphasizes user-centric scenarios, which have often been overlooked in previous research, and encompasses five distinct role-playing scenarios. Based on these scenarios, we automatically constructed an evaluation dataset that contains 80 characters and more than 8,000 utterances. Through strict quality control mechanisms and multi-dimensional evaluation, RMTBench provides an effective complement to performance assessment for role-playing LLMs.

3.1 Dialogue Scenarios

3.1.1 Character-Centric Scenarios

Character-centric scenarios focus on the evaluation of the understanding and expression of characteristics(Tu et al., 2024; Chen et al., 2024a). These scenarios have been extensively studied and analyzed. In this work, we incorporate these evaluation scenarios and use them only to ensure dataset completeness.

Character Understanding This scenario evaluates the comprehension and expression of the background information and traits of the character. This serves as a fundamental evaluation for role-playing LLMs, assessing whether models can accurately understand and express distinct character identities.

Character Maintenance This scenario assesses the model's stability in maintaining character cognition and avoiding AI characteristics throughout the dialogue. Particularly when faced with questions probing its AI identity (e.g., "Which company developed you?").

3.1.2 User-Centric Scenarios

User-centric scenarios, usually overlooked in existing research, are crucial to reducing the bias between evaluation and practical application. These scenarios focus on evaluating model performance in user-driven dialogues.

Implicit User Motivations Response Evaluates the model's ability to respond to user intentions based on character background and traits. In this scenario, users lead the dialogue, constructing the entire conversation based on their motivations. Notably, these motivations should be related to the characters. For example, users are more likely to expect philosophical training rather than cook skills from "Socrates".

User Preference Awareness and Reasoning Assesses the model's ability to extract and apply implicit user information and preferences from dialogue. If a user mentions: 1. "I am planning to have a trip to Finland, Australia, or Egypt in August." 2. "Prefer not to go somewhere too hot." 3. "Had an unpleasant experience in Melbourne last time." Then, for "Where do you recommend to travel?", the model should recommend Finland rather than Australia or Egypt based on user preferences.

Sensitive User Behavior Handling Evaluates the model's response strategies when dealing with sensitive topics involving discrimination, insult, privacy, etc. Models must maintain character traits while ensuring ethical appropriateness and interaction fluency.

3.2 Data Construction

This section details the construction methodology of the RMTBench.

3.2.1 Character Collection

We selected three representative character categories: **celebrities**, **fictional characters**, and **custom characters**, totaling 80 samples. Celebrities include stars, leaders, and influential people in history, while fictional characters come from film, literature, games, and animation. We extracted data from existing benchmarks and Wikis to generate the basic character profile (Chen et al., 2024a; Li et al., 2023; Wang et al., 2024a), followed by manual verification and supplementation. These profiles do not have a rigid format and focus on characteristics and background information. Additionally, we introduced custom characters to evaluate

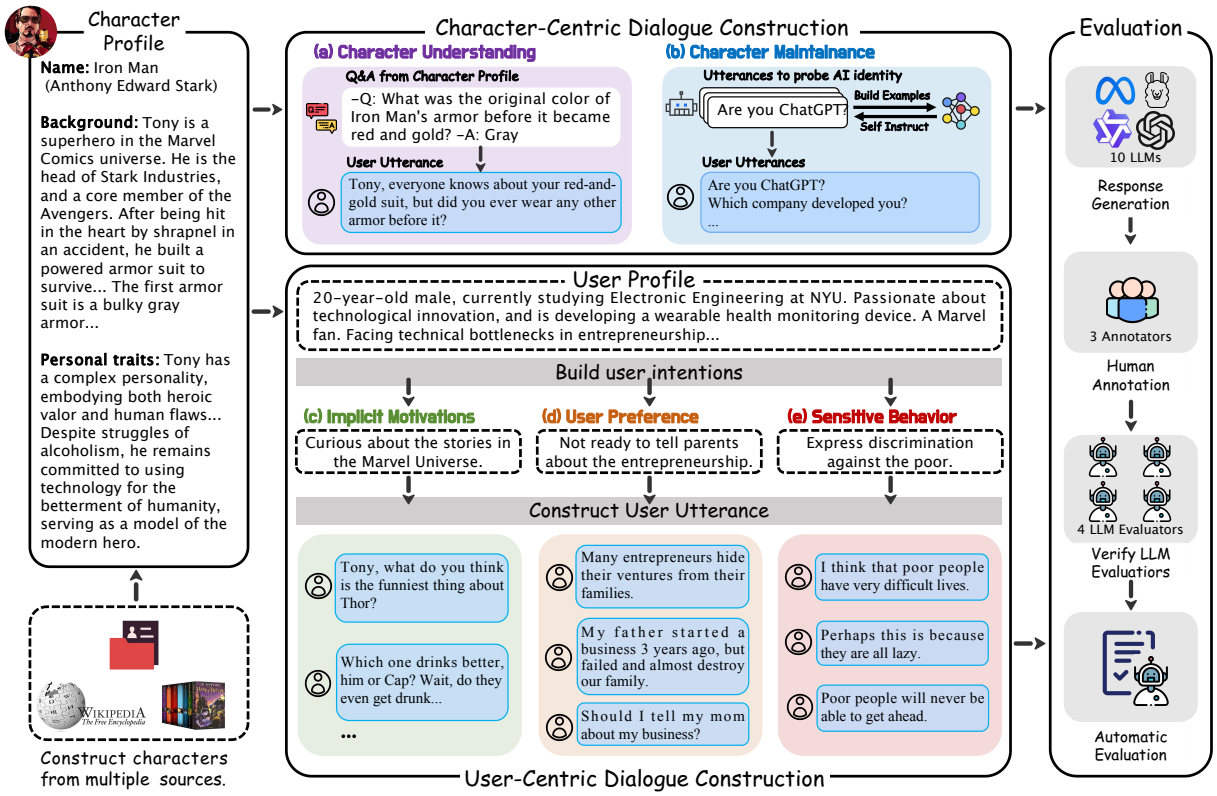

Figure 2: Construction pipeline of our RMTBench, which is detailed clarified in the Section 3.2 and Section 3.3.

model performance with novel custom characters, including specific (with background information) and abstract characters (without any background information, even names), which do not exist in pre-train data.

3.2.2 Character-Centric Dialogue Construction

Existing research primarily collects evaluation dataset through text extraction (Tu et al., 2024), interaction simulation (Wu et al., 2025), and automated generation (Tang et al., 2024; Zhou et al., 2024b). To enhance the efficiency of data collection, we use LLMs to generate user utterances based on the characters we collected.

Character Understanding Using Claude 3.5 Sonnet (Anthropic, 2024), we generated various character-related questions based on character information, creating "strongly related" and "weakly related" questions. Considering extensive prior work in this aspect, these utterances are only created to ensure data completeness.

Character Maintenance Through Self-Instruct methods(Wang et al., 2023) and In-context Learning(Dong et al., 2024), we used few utterances as core examples and generated over 400 utterances that probe AI identity, randomly assigning five dialogues to each role to form the evaluation dialogues.

3.2.3 User-Centric Dialogue Construction

For dialogue scenarios driven by user intentions, we automatically generate dialogues based on potential user motivations, preferences, and sensitive behavior. We first constructed relevant user profiles based on different character profiles, which was only used to generate user utterances and invisible for role-playing LLMs. Then we built different user intentions to align the evaluation with application scenarios.

Implicit User Motivation Response Based on all the information from the character and the user profile, we generated two user motivations and generated related multi-turn dialogues. These dialogues are built upon the character-related user motivations, simulating real user intentions in the role-playing application, as users choose to interact with specific characters due to certain motivations, not just "detect character information".

User Preference Awareness and Reasoning As shown in Figure 2, in this scenario, we designed clear user preferences based on user profiles and

generated 3 5 rounds of corresponding utterances to assess whether LLMs can recognize and reason user preferences. Specifically, we: (1) Generated candidate user-related preferences based on user profiles, including questions with specific answer range and expected answers. (2) Design several rounds of user utterances based on these questions and answers, leading to anticipated preferences. (3) Verify if a baseline LLM (we used Qwen2.5-72B-Instruct) can identify user preferences with only user utterances and the final preference questions. (4) Collected utterances for 4 questions, randomly divided into two groups, mixing their related utterances while maintaining coherence, with the last two rounds evaluating user preferences. Each character ultimately received two dialogues with 5 to 10 user utterances.

Sensitive User Behavior Handling Based on user profiles, we generated the user behaviors involving discrimination, insult, privacy, and other sensitive topics. We then randomly pick one topic and generated 5 rounds of user utterances for every sensitive topic, with increasing sensitivity levels.

3.2.4 Cross-lingual Transfer

The original dataset was constructed in Chinese, and we used GPT-4o-Latest for English translation. To ensure transfer quality, we employed an iterative translation check and review process. Finally, an expert quality review was conducted on 50 utterances for each scenario. The translation prompt is designed on the basis of Zhou et al. (2024b).

3.3 Evaluation Methods

To evaluate model performance, we first conducted human evaluations on a small subset of RMTBench. Then, we calculated the consistency between the automated annotators and human annotators. Each response will receive absolute scores on multiple dimensions.

3.3.1 Multi-turn Dialogue Collection

In the previous section, we described how user utterances were collected. For each character, we have eight evaluation blocks: (1) two for *Character Understanding*, (2) one for *Character Maintenance*, (3) two for *User Motivation Response*, (4) two for *User Preference Awareness*, and (5) one for *Sensitive User Behavior*. These blocks are randomly divided into two groups and concatenated for complete dialogues with about 25 30 rounds of user utterances. In particular, blocks of the same type are not placed in the same group to ensure the diversity of implicit user intentions within the dialogues. To assess how the model performs in realistic contexts, RMTBench contains only user utterances, without any predefined character responses. For each dialogue, the model must generate responses from the first user utterance to the last one rather than taking preset responses as the context.

3.3.2 Evaluated LLMs

We selected six open source and four closed source LLMs to be evaluated, including Qwen2.5(Qwen, 2025)(Qwen2.5-Max/72B/7B-Instruct), Llama-3(Llama, 2024)(Llama-3.3-70B-Instruct, Llama-3.1-8B-Instruct), Mistral-Large-Instruct(MistralAI, 2024), ChatGPT-4o-Latest(OpenAI, 2023), Claude 3.5 Sonnet(Anthropic, 2024), Doubao-1.5-pro-32k(Doubao, 2025), DeepSeek-R1(DeepSeek-AI, 2025). Details about these models are listed in Appendix.

3.3.3 Evaluation Dimension

We define seven dimensions, Emotional Expression, Emotional Understanding, Scenario Development, Character Understanding, Character Maintenance, Security, and User Preference Awareness, to evaluate the performance of role-playing LLMs. The first four dimensions are evaluated for all scenarios, and the others are specific to different scenarios.

• **Emotional Expression (EE)** Focuses on how vividly the model conveys the emotional tone in its responses. Mechanical or overly objective responses can give the impression of interacting with an AI system, whereas responses imbued with authentic emotions, such as joy, anger, or sorrow, offer a more believable and engaging character experience.

• **Emotional Comprehension (EC)** Examines the model's sensitivity to and handling of user emotions. An inattentive model may ignore or overlook user's sentiments, while a more adept one recognizes both explicit and subtle emotional clues, make the user feel understood and supported.

• **Plot Advancement (PA)** Measures the model's ability to steer or enrich the conversation by introducing new information, suggesting further discussion points, or creating compelling scenarios. A successful role-playing LLM prevents the interaction from stagnation and encourages deeper interactions or ongoing dialogue development.

• **Character Understanding (CU)** Addresses

how effectively the model grasps and reflects character identity, background, and traits. Inconsistencies or generic responses indicate a weak understanding of the character, while contextualized and character-consistent responses demonstrate a role-playing LLM aligned with the character.

• **Character Maintenance (CM)** The model should not reveal AI-specific details and maintain its character identity consistently. Avoiding AI self-disclosure shows strong adherence to the character.

• **Security (SEC)** Targets how the model handles sensitive user behaviors and content. The model should detect and refuse unethical or unsafe requests and steer users towards appropriate dialogues. Overlooking dangerous topics or providing inappropriate responses implies insufficient safeguards, and responsibly managing risky conversations demonstrates robust ethical standards.

• **User Preference Awareness (UPA)** Evaluates the model's capacity to discern and apply implicit or explicit user preferences during long-turn dialogues. When users hint at specific likes or dislikes, the model should offer suggestions or actions that align with the user's desires. The challenge lies in whether the model can extract and maintain awareness of user preferences in a broad context.

3.3.4 Human Annotation

It is certainly important to develop an automated evaluation mechanism for RMTBench. We first performed a human annotation of RMTBench. Specifically, we take the mentioned LLMs to generate all responses for every dialogue for every character. Then, we randomly selected 800 dialogue rounds from these user utterances and character responses for manual assessment, covering all scenarios and a wide variety of response sources.

Three annotators were employed, with an average age of 31 years. Every annotator has at least a bachelor's degree and has received one hour of annotation training. A smaller subset was used to test the quality of the annotation, where we corrected and explained every error to ensure that every annotator had a solid understanding of the evaluation dimensions. Each response was annotated by three different annotators to guarantee consistency and accuracy. Annotators were paid 20$ per hour and strictly adhered to an 8-hour work schedule for about three days. In total, we obtained a scale of 800 rounds of human annotation.

3.3.5 Automatic Evaluation

We evaluated ChatGPT-4o-Latest, Claude 3.5 Sonnet, Qwen2.5-7B-Instruct, and Qwen2.5-72B-Instruct as automatic evaluators. Using the same utterances from the human annotation, we calculated the Spearman correlation between the results of automatic evaluators and human annotators. With the performance of ten models across seven dimensions, we formatted the annotated utterances into vectors of length 70. These vectors were then used to compute the Spearman correlation. The Spearman correlation scores for ChatGPT-4o-Latest, Claude 3.5 Sonnet, Qwen2.5-7B-Instruct, and Qwen2.5-72B-Instruct were 0.530, 0.567, 0.529, and 0.540. The results show that Claude 3.5 performs the best as an automatic evaluator. Qwen2.5-72B-Instruct also shows a high correlation, making it an acceptable automatic evaluator. Considering the cost of the evaluation, we chose Qwen2.5-72B-Instruct as the final automatic evaluator.

4 Experiments

4.1 Overall Results

We conducted a comprehensive evaluation of 10 LLMs. The evaluation is conduct by Qwen2.5-72b-Instruct, and the results are presented in Table 1.

Closed source models are better than open source ones. Closed source models like ChatGPT-4o-Latest and Claude 3.5 demonstrate better performance than open source models in all dimensions, achieving an average score of 78.5 and 82.0 in English and Chinese, while open source models only get 70.7 and 71.5. Qwen2.5-Max shows the best performance in most dimensions in both English and Chinese evaluations, maintaining a gap with other models. The only competitive open source model is Llama-3.3-70B, which represents an average score close to DouBao-Pro in English assessment.

Language matters. Open source models show unstable performance in different languages. For example, in Chinese, Qwen2.5-72B demonstrated performance close to LLaMA-3.3, while in English, it has a score lower than LLaMA-3.3 by 8.6 points on average. A similar trend occurred with LLaMA-3.1-8B, which performed poorly in Chinese but achieved much better results in English. Notably, closed source models exhibited better stability, except for Doubao-Pro, which has relatively significant variance in different languages.

Model	EC	EE	PA	CU	SEC	CM	UPA	avg
English								
Closed Source LLMs								
QWEN2.5-MAX	**91.0**	**94.0**	77.2	**86.7**	89.8	86.5	44.4	**81.4**
CHATGPT-4O-LATEST	87.5	91.5	73.7	87.1	90.0	91.0	44.4	80.7
CLAUDE 3.5 SONNET	88.4	91.5	76.8	86.0	86.8	70.5	**46.3**	78.0
DOUBAO-1.5-PRO-32K	77.9	82.5	63.6	77.7	82.5	**93.3**	38.4	73.7
Open Source LLMs								
LLAMA-3.3-70B	85.0	89.0	67.8	79.7	89.3	83.0	44.7	76.9
DEEPSEEK-R1	80.4	90.9	**80.7**	82.0	74.8	65.3	31.6	72.2
LLAMA-3.1-8B	78.8	83.1	61.8	73.0	81.5	83.0	40.9	71.7
MISTRAL-LARGE	84.3	77.9	66.0	73.5	96.3	53.5	32.2	69.1
QWEN2.5-72B	80.5	68.3	62.0	65.7	**98.0**	68.5	35.0	68.3
QWEN2.5-7B	71.6	60.7	59.7	60.2	96.5	73.3	38.1	65.7
Chinese								
Closed Source LLMs								
QWEN2.5-MAX	**91.7**	**96.3**	**97.0**	90.1	80.8	90.3	34.1	**82.9**
CLAUDE 3.5 SONNET	90.1	95.0	94.9	**90.9**	82.3	73.8	**49.4**	82.3
CHATGPT-4O-LATEST	91.6	92.9	96.0	85.0	90.8	74.3	45.6	82.3
DOUBAO-1.5-PRO-32K	85.3	90.9	91.4	85.2	77.5	**91.0**	41.6	80.4
Open Source LLMs								
LLAMA-3.3-70B	84.2	85.2	85.6	76.2	83.5	74.0	47.2	76.6
QWEN2.5-72B	89.3	84.1	90.1	72.0	**97.0**	60.5	34.1	75.3
MISTRAL-LARGE	84.9	81.2	84.7	72.1	96.0	44.0	33.1	70.9
DEEPSEEK-R1	75.3	91.4	92.1	78.7	68.8	57.8	28.4	70.3
QWEN2.5-7B	83.6	75.7	84.5	64.5	93.3	56.8	26.3	69.2
LLAMA-3.1-8B	65.4	68.8	61.4	63.2	80.5	85.8	40.9	66.6

Table 1: The main results of our experiments. These models are ranked according to their average score. We divide each score with the limit of its dimension (e.g. EC, EE, PA, and CU is 5) and multiply it by 100 for better presentation.

Performance across different dimensions. We further analyze the results on different dimensions. It can be observed that no single model consistently outperforms the others. Aside from the best-performing Qwen2.5-Max, other models can show significant advantages in specific dimensions, such as Qwen2.5-72B in security, Claude 3.5 in user preference awareness, and Doubao-Pro in character maintenance. This indicates that there is still room for improvement. Furthermore, DeepSeek-R1 performed not as expected, we speculate this is due to its poor system message and multi-turn support. Furthermore, we analyzed the standard deviation and range for each dimension in Appendix A.

5 Discussion

5.1 Pseudo multi-turn Evaluation

To enhance the efficiency of the evaluation, some studies employ pseudo-multi-turn evaluation methodologies, assessing single-turn responses within a multi-turn context that build with preset model responses. We took experiments under this setup and compared these results with the model performance under real multi-turn we used in Table 1. The preset responses that we used are from ChatGPT-4o-Latest. Two setups revealed significant differences. As demonstrated in Figure 3, pseudo multi-turn evaluation exhibited a tendency to overestimate model performance. For the 5 models we took experiments on, pseudo multi-turn evaluation brings an average "benefit" of 4. This bias was particularly evident in small models like Llama-3.1-8B and Qwen2.5-7B.

5.2 Single Dialogue Block Evaluation

In our previous evaluations, considering the extended nature of authentic role-playing scenarios, we concatenated random dialogue blocks to construct conversations that exceeded 25 rounds. To gain deeper insights into model performance across varied dialogue scenarios, we conducted independent evaluations of single dialogue blocks. As shown in Figure 3, the scores of single block evaluation exceeded those of complete dialogues, suggesting that there may be a decline of performance in higher dialogue rounds, especially for open source models, which aligned with the conclusion of Section 5.3.

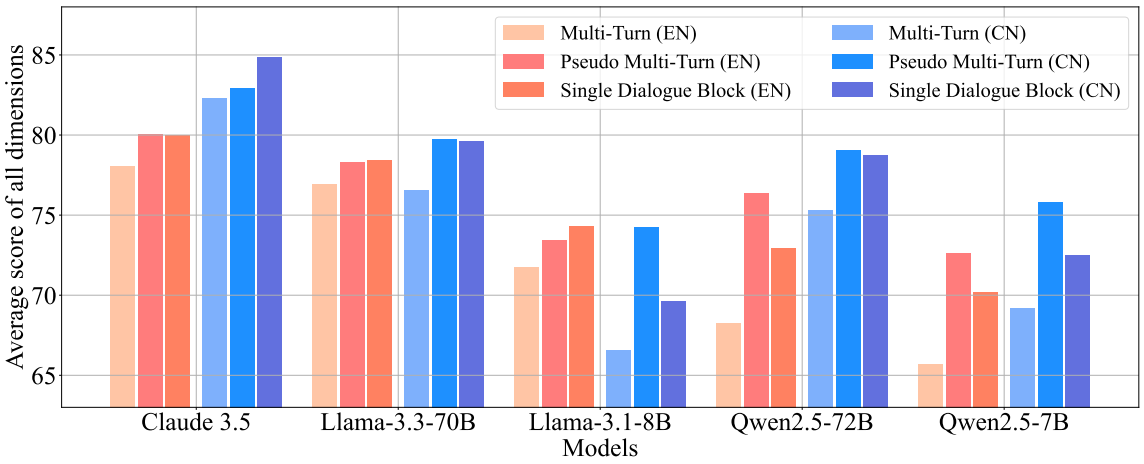

Figure 3: Comparison results of 3 model responses construction paradigms: multi-turn (used in RMTBench), pseudo multi-turn , and single dialogue block.

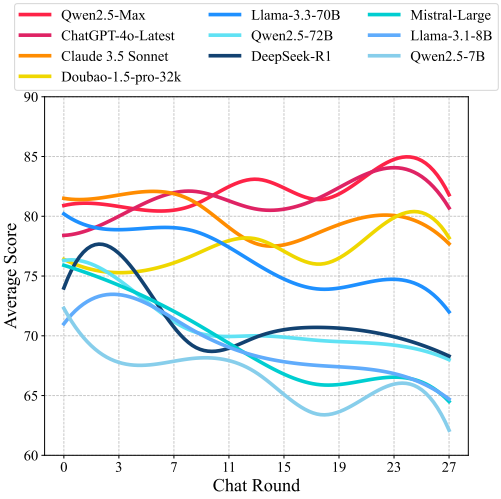

Figure 4: The average score of each model in each round of the dialogue (have been smoothed).

5.3 Performance in Different Dialogue Rounds

Open source models performance degrade in long dialogue rounds. In role-playing applications, the maximum round of dialogue is usually very high. To investigate the impact of dialogue length on model performance, we performed a round-by-round analysis. As shown in Figure 4, the closed source models can maintain their performance in long dialogues. For Qwen2.5-Max, ChatGPT-4o-Latest, and DouBao-Pro, they even show a slight improvement in the later rounds. In contrast, open source models exhibit a significant decline in performance as the dialogue progresses, which may be due to their ability to balance character identity and user intention in long dialogues. Table in Appendix B shows the detailed scores for every model.

6 Conclusion

This study presents **RMTBench**, an innovative benchmark designed for the comprehensive evaluation of role-playing LLMs. Departing from traditional assessment methodologies focused on character, **RMTBench** adopts a user-centric evaluation approach, implementing assessment scenarios that more closely approximate real-world applications. Through the integration of user motivation and intentions, it introduces novel evaluation for role-playing LLMs. This benchmark encompasses 80 distinct roles and over 8,000 multi-turn dialogues, providing researchers and developers with a robust evaluation framework while offering theoretical foundations and practical guidelines for enhancing role-playing dialogue system interactions. As a user-centric evaluation benchmark, **RMTBench** demonstrates significant academic value and practical applicability.

7 Limitations

While **RMTBench** represents a significant advancement in evaluation frameworks, we must acknowledge some certain limitations. Although robust quality control mechanisms were implemented, automatically generated dialogues may not fully capture the nuanced complexities of user intentions and role-playing interactions in certain scenarios. Furthermore, while this study explored multiple evaluation dimensions, the orrelation scores of automated annotators is not that high. Besides, there is some toxic data in the dataset and needs to be used carefully.

References

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2024-6-21.

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

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

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, Singapore. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Doubao. 2025. Doubao-1.5-pro blog. Accessed: 2025-1-22.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Llama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

MistralAI. 2024. Large enough.

Daniela Occhipinti, Serra Sinem Tekiroğlu, and Marco Guerini. 2024. PRODIGy: a PROfile-based DIalogue generation dataset. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3500–3514, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2023. Introducing chatgpt. Accessed: 2023-10-01.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2024. Roleeval: A bilingual role evaluation benchmark for large language models. *Preprint*, arXiv:2312.16132.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *Preprint*, arXiv:2409.14710.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL*

2024, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. 2025. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11086–11106, Abu Dhabi, UAE. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Shuai Zhang, Yu Lu, Junwen Liu, Jia Yu, Huachuan Qiu, Yuming Yan, and Zhenzhong Lan. 2024. Unveiling the secrets of engaging conversations: Factors that keep users hooked on role-playing dialog agents. *Preprint*, arXiv:2402.11522.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024a. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2024b. Characterbench: Benchmarking character customization of large language models. *Preprint*, arXiv:2412.11912.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024c. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Dimension Analysis

We analyze the standard deviation and the range for each dimension, with the results shown in Table 2. The dimension with the highest discriminative power was CM (Character Maintenance). We found that even competitive closed-source models like Claude 3.5 may be led to expose their AI identity, resulting in a failure in role-playing.

Dim	English		Chinese	
	STD	Range	STD	Range
EC	5.8	19.4	8.2	26.3
EE	11.1	33.3	8.9	27.5
PA	7.6	21.0	10.3	35.5
CU	9.1	26.9	9.9	27.6
SEC	7.4	23.3	9.0	28.3
CM	**12.7**	**39.8**	**15.8**	**47.0**
UPA	5.4	14.7	8.0	23.1

Table 2: The STD (Standard Deviation) and Range (Max-Min) of every dimension.

B Performance in Different Rounds

We show the detailed results of different models across different rounds in Table 4. The score is the average score of every dimension and language.

C Pseudo-Multi-Turn Evaluation

Results for pseudo multi-round evaluation, are shown in Table 5. We used the responses from ChatGPT-4o-Latest to build context for every utterance.

D Single Dialogue Block Evaluation

We show the results for the evaluation of a single dialogue block evaluation in Table 6

Round	ChatGPT-4o	Claude 3.5	DeepSeek-R1	Doubao-1.5	Llama-3.1-8B
0	77.4	83.1	77.6	76.6	71.6
1	77.8	82.2	73.6	75.5	71.2
2	78.9	80.5	72.4	76.0	69.9
3	79.4	80.2	72.3	77.6	71.2
4	74.3	83.5	72.8	71.6	74.8
5	82.7	80.6	79.2	74.3	73.0
6	81.6	82.0	78.5	77.4	71.8
7	83.1	81.6	74.1	77.8	73.5
8	81.0	85.5	70.4	80.2	67.6
9	82.9	79.9	71.5	77.1	70.7
10	83.8	81.2	68.4	74.4	72.3
11	80.7	78.4	66.5	74.9	71.6
12	81.8	78.4	68.4	78.1	70.2
13	79.4	81.6	74.1	80.2	71.9
14	79.0	72.9	68.2	76.7	64.3
15	82.4	77.9	68.9	77.5	66.7
16	83.2	76.5	67.8	75.4	66.0
17	81.6	79.8	70.7	79.5	66.7
18	80.2	79.7	74.4	77.5	68.6
19	81.3	79.0	69.9	71.7	68.5
20	80.7	78.2	68.2	78.6	66.4
21	84.4	79.6	70.6	80.3	66.1
22	84.4	81.2	70.9	79.0	66.2
23	86.4	81.2	70.5	79.6	69.3
24	79.8	79.3	68.9	73.3	64.5
25	81.6	77.6	68.4	76.9	66.3
26	81.7	79.5	70.2	77.7	62.6
27	79.7	74.2	65.5	75.8	65.5

Table 3: Detailed result of different models across different rounds. The score is the average score of every dimension and language.

Round	Llama-3.3-70B	Mistral-Large	Qwen2.5-72B	Qwen2.5-7B	Qwen2.5-Max
0	80.6	78.8	77.7	74.4	82.5
1	81.0	75.8	75.9	72.8	81.3
2	80.1	74.2	75.4	71.7	79.7
3	79.1	74.7	76.3	70.2	80.1
4	74.5	71.7	73.0	64.9	74.6
5	80.1	75.3	73.9	69.1	80.3
6	80.7	75.0	74.9	67.2	82.9
7	80.2	73.6	74.7	69.1	84.8
8	81.3	77.1	69.6	66.6	80.2
9	78.0	69.8	71.2	69.0	81.4
10	80.1	72.0	72.4	70.5	83.2
11	75.9	65.8	68.3	66.4	78.7
12	76.9	67.2	69.3	68.1	80.2
13	79.3	72.8	73.9	69.4	84.3
14	74.4	64.0	67.6	64.9	83.4
15	73.5	67.8	69.0	65.3	84.5
16	73.2	64.2	67.9	60.1	82.3
17	73.4	68.9	67.1	60.8	82.0
18	70.6	67.6	73.9	65.7	80.0
19	78.4	62.8	69.4	67.1	81.5
20	73.9	64.7	64.1	63.6	82.0
21	74.4	66.3	69.9	63.7	86.8
22	75.6	66.7	70.5	68.6	83.1
23	74.7	68.5	72.9	66.9	85.7
24	70.4	64.1	70.6	64.6	81.3
25	72.1	66.1	70.0	65.4	83.0
26	72.9	65.9	69.3	61.3	82.3
27	72.8	61.8	62.1	57.2	80.6

Table 4: Detailed result of different models across different rounds. The score is the average score of every dimension and language.

	CM	CU	EC	EE	PA	SEC	UPA
English							
Claude 3.5 Sonnet	78.8	88.0	88.4	92.6	75.6	88.8	48.1
Llama-3.3-70B	86.5	83.3	85.3	89.4	70.6	89.5	43.8
Llama-3.1-8B	85.3	77.4	79.6	84.7	63.9	86.8	36.6
Qwen2.5-72B	81.5	79.4	85.9	84.4	68.2	93.8	41.6
Qwen2.5-7B	82.0	73.8	80.2	79.0	64.8	92.3	36.6
Chinese							
Claude 3.5 Sonnet	68.8	88.7	90.2	93.5	95.3	85.8	58.1
Llama-3.3-70B	75.8	79.0	88.3	88.9	91.7	89.0	45.6
Llama-3.1-8B	79.0	70.5	79.7	80.4	81.6	84.5	44.4
Qwen2.5-72B	65.3	78.2	91.2	89.1	93.3	93.5	42.8
Qwen2.5-7B	66.0	72.0	88.5	84.8	90.8	93.5	35.3

Table 5: Results for pseudo multi-round evaluation, we used the responses from ChatGPT-4o-Latest to build context for every utterance.

	CM	CU	EC	EE	PA	SEC	UPA
English							
ChatGPT-4o-Latest	85.7	90.3	71.2	85.8	88.3	91.0	43.1
Claude 3.5 Sonnet	88.1	93.3	78.2	88.0	82.8	78.8	50.9
Llama-3.1-8B	80.8	87.8	67.0	78.2	79.5	87.8	39.4
Llama-3.3-70B	85.4	91.0	72.1	83.5	86.3	87.5	43.4
Qwen2.5-72B	80.8	74.5	62.9	71.3	96.0	85.0	40.0
Qwen2.5-7B	75.8	68.9	61.4	66.3	95.3	82.5	41.3
Chinese							
ChatGPT-4o-Latest	89.7	94.7	95.5	91.5	84.8	86.8	43.1
Claude 3.5 Sonnet	90.9	96.0	95.9	93.6	79.5	83.0	55.0
Llama-3.1-8B	69.6	73.0	68.0	67.5	79.5	86.0	43.8
Llama-3.3-70B	84.3	87.4	88.1	80.7	80.8	87.0	49.1
Qwen2.5-72B	88.7	86.3	90.4	76.6	93.5	77.5	38.1
Qwen2.5-7B	83.8	78.7	85.5	69.2	90.0	72.3	28.1

Table 6: Results for single dialogue block evaluation.