R&R: A Role-playing Model Enhanced by Retrieving and Reflecting

Anonymous ACL submission

Abstract

Role-playing is a key capability of large language models (LLMs). However, existing models often fall short in delivering fully immersive character simulations. They frequently overlook the knowledge constraints inherent to the role and fail to adopt the appropriate mindset, resulting in responses that feel noticeably artificial. To address these limitations, we propose R&R, a role-playing model enhanced with retrieval and reflection. Prior to generating a response, our model first retrieves similar historical dialogues based on the current query and generates character-specific reflections informed by the role's self-profile. It then searches for relevant background knowledge to support the response. Finally, the model evaluates whether the query falls within the character's scope of knowledge and generates a response grounded in both the retrieved context and reflective reasoning. To assess the effectiveness of our approach, we construct a new benchmark dataset and introduce novel evaluation metrics tailored to character role-play. We also conduct comparisons using an established public metric. Experimental results show that our model achieves an average performance improvement of 8% over CharacterLLM.

1 Introduction

Large language models (LLMs) are computational models notable for their ability to achieve general purpose language generation and other natural language processing tasks such as classification (Min et al., 2023). They can help people in various ways, from generating creative content to assisting in complex problem-solving tasks. They have the capacity to comprehend and generate human-like text, enabling them to aid in writing, summarizing information, generating ideas, answering questions, and even engaging in meaningful conversation.

However, LLMs exhibit poor performance on the task of role-playing. When models lack specific fine-tuning, they often forget the role they are currently playing and respond from their own persona. Moreover, LLMs frequently reply in a manner beyond the knowledge scope of the current role or in a tone that the role would never use. For example, if you ask LLMs to play as Sir Isaac Newton and subsequently inquire, "Do you know what a mobile phone is?", the LLM might respond with an acknowledgment of unawareness. Nevertheless, it would proceed to describe the function or principles of a mobile phone. These observations illustrate that while LLMs are capable of adhering to human instructions for role-playing, the struggle to fully confine themselves within the constraints of the current role and possess limited understanding of the role.

Shanahan (Shanahan et al., 2023) propose that LLMs' dialogue with humans is actually a kind of role-playing, they will do their best to role-play the character of a dialogue agent as portrayed in the dialogue prompt. Consequently, we postulate that LLMs possess sufficient capability for role-playing, requiring only an indication of the role they are currently enacting and an adequate provision of role-related information (Lu et al., 2024a). There is also some work being done to facilitate the enhancement of LLMs' proficiency in role-playing, such as ChatHuruhi (Li et al., 2023), CharacterLLM (Shao et al., 2023) and RoleLLM (Wang et al., 2023). These studies generate character dialogue data using LLMs that can be used to prompt or train LLMs to form responses suitable to the character's language style. However, these efforts fail to prevent situations where the model responds beyond the character's knowledge or lacks consistency in its linguistic style given that most dialogue is generated by LLMs. More crucially, they fail to incorporate character-specific thinking styles, rendering LLM role-play a mere imitation of the character's dialogue style.

To solve those problems, we propose R&R in this paper, which enables LLMs to generate responses with the respective styles of expression and thinking associated with each role. To evaluate the effectiveness of our approach, we construct a new dataset using LLMs based on authentic dialogues of various roles. Then, we assess the expression and thinking style of these roles by comparing the response generated by different models. Experimental results suggest that our R&R outperforms other models in mimicking roles. The contributions of this paper are as follows:

- We propose R&R, a Role-playing model enhanced by Retrieving and Reflecting, which can prompt LLMs with the insight and thinking style of a given role, enabling them to generate responses in the tone of that role.
- We propose a dataset construction method, and build a role-playing dataset, which include the mindset of roles. What's more, our R&R can easily extend to a new role without train.
- We create an evaluation dataset and adapt existing metrics to effectively assess the performance of role-playing models.

2 Related work

Dataset: For role-playing task, it is important to build a realistic and high-quality dataset of character dialogue. However, creating such dialogues without scripts is challenging for real people and requires significant labor to extract, clean, and organize data from scripted sources (Brahman et al., 2021; Gosling et al., 2023). To obtain relatively high-quality dialogue data more efficiently, most researchers leverage LLMs as annotators, often with specific markers or adjustments to enhance realism in certain aspects (Lotfi et al., 2024; Ahn et al., 2023). For, example, Chen et al. (2023) propose Harry Potter Dialogue (HPD) dataset, which encompasses all dialogue sessions (in both English and Chinese) from the Harry Potter series and is annotated with vital background information, including dialogue scenes, speakers, character relationships, and attributes. Li et al. (2023) propose ChatHaruhi, which covering 32 characters with over 54k simulated dialogues. Wang et al. (2023) propose RoleBench, which is a systematic and fine-grained character-level benchmark dataset for role-playing with 168,093 samples. Ran et al. (2024) propose to obtain the mindset of characters by mimic them to answer the personality questionnaires. However, all works are focus on the scenario, timeline and the dialogue realism, few of them tend to capture the mindset or the reflection of roles, which is the fundamental of a human.

Methods: Using specialized prompts (Han et al., 2022; Li et al., 2023) or fine-tuning LLMs with role-specific datasets (Wang et al., 2023) are two common methods for role-play task. Cui et al. (2023) propose a thespian agent framework, which can learn to emulate multiple characters along with a soft prompt. ChatHaruhi (Li et al., 2023) input all system prompt, character memories retrieved for the user query, and the dialogue history into LLMs, which can obtain good results. As for the fine-tune methods, Shao et al. (2023) propose CharacterLLM by fine-tuning Llama with role dialogues dataset. Lu et al. (2024b) introduce Ditto, which is a a self-alignment method for role-playing. Yu et al. (2024) propose Neeko, a framework for efficient multi-character imitation in role-playing scenarios, utilizing a dynamic low-rank adapter strategy to adapt seamlessly to diverse characters. However, all the methods are focus on the tone and the knowledge of role, few of them try to learn the mindset and none of them learning the reflection.

Evaluation: Evaluating the role-playing capability of current models is challenging because roles can have multiple valid responses to the same prompt. Traditional evaluation metrics, such as ROUGE and perplexity (PPL) (Wang et al., 2024), are insufficient for capturing the nuanced performance of these models. To address this, existing researchers propose a variety of metrics, including tone, knowledge, stability, and personality (Shao et al., 2023; Tu et al., 2024; Shen et al., 2023; tse Huang et al., 2023; Mao et al., 2023). These evaluations often rely on LLMs, such as ChatGPT, to score responses step by step using specialized prompts. Given the randomness of responses and the high cost of human reviews, leveraging LLMs for scoring has become the most common practice. However, existing metrics lack both a comprehensive assessment framework and sufficient evaluation depth.

3 Methods

Figure 1 shows the motivation behind our model. Unlike the prompt-based and fine-tuned methods, our approach incorporates both retrieved relevant information and inferred reflections as additional

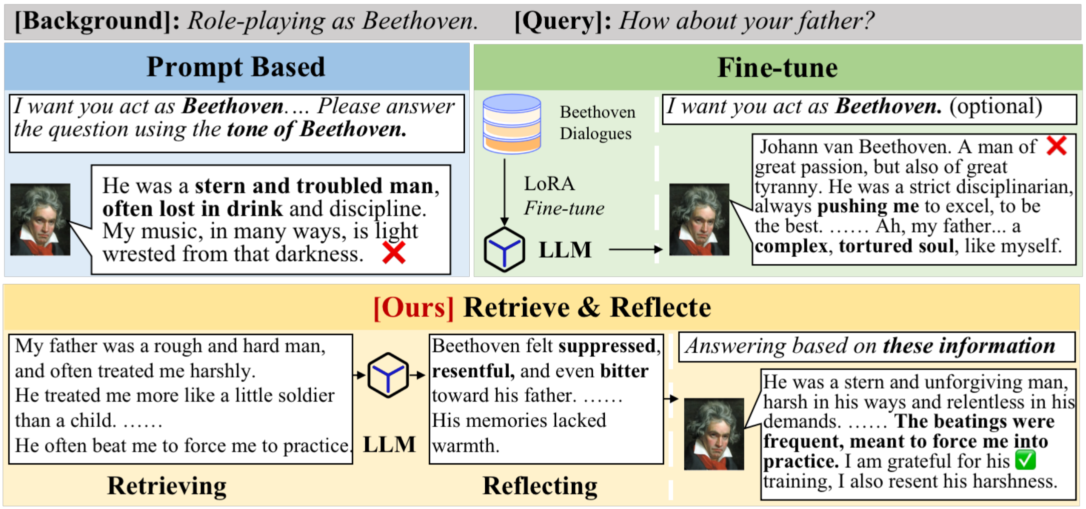

Figure 1: The motivation of our model. Taking Beethoven as an example, the prompt-based method yields brief and shallow responses, while the fine-tune method may introduce hallucinations (e.g., "He pushing me to be the best."). In contrast, our method first retrieves relevant information (e.g., the real content that Beethoven has mentioned about his father), then uses an LLM to generate inferred reflections about Beethoven, and finally prompts the LLM to generate a response grounded in that context.

context. We then use a specially designed prompt to instruct the LLM to generate responses based on this enriched input. As shown, the prompt-based method produces responses that lack detail and reflect only the model's general knowledge about Beethoven, rather than reasoning from Beethoven's own perspective. In contrast, the fine-tuned method includes inappropriate content, such as *"He always pushing me to be the best."* We believe such inconsistencies arise from the lack of accurate role information, causing the LLM to fail at capturing Beethoven's true mindset. To address this, our method aims to enhance role consistency by introducing retrieval and reflection steps. The retrieval step provides relevant knowledge related to the question, helping the LLM determine whether the question falls within the role's knowledge scope. The reflection step infers the role's mindset regarding this question, offering a more coherent basis for generating a contextually and character-consistent response.

3.1 Role Dataset Construct

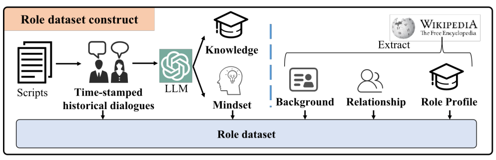

Figure 2: The process of our dataset construction.

As shown in Figure 1, our method require to retrieve the relevant knowledge and reflect the mindset based on these information. To reduce the time cost of those two steps and provide more information, we will construct a role dataset first. Figure 2 illustrates the process of our dataset construction. Since most existing role-playing datasets are constructed by requiring LLMs to generate dialogues, they cannot be used directly, as they may not maintain personality consistency across different roles. In our dataset, we use actual dialogue from scripts for each role. For real characters that do not have scripts, we use their quotes as substitutes for the dialogue dataset. The mindset and knowledge contained in each dialogue pair are obtained by instruct LLMs to infer based on the contextual.

The role-profiles are obtained from the Wikipedia and baidu-baike. We will first crawl all content on the role's page of Wikipedia. For Chinese roles, we will crawl from Baidu-Baike. Then the content will be divided into four parts: role-profile, relationships, major deeds and catchphrases. For the role-profile, we will use the character summary directly from Wikipedia. For relationships and background, we will have an LLM sort that content.

3.2 Role Playing

Once we have completed the construction of these datasets, we can allow LLMs to role-play with those information. The process are as the following step:

i Obtaining the background and self-profiles of the role Ri from the role dataset (Pre-process stage);

ii Retrieving knowledge K_{Ri} of role Ri according to the question;

iii Retrieving similar dialogues D_{RiS} from the history of dialogues D_{Ri} based on the question;

iv Obtaining the mindset M_{Ri} according to the similar dialogues by an LLM;

v Organizing user questions, role Ri , backgrounds, self-profiles, similar dialogues, mindset, knowledge, and inputting them into the LLM;

vi According to the question and those information to determine whether the current role can answer the question, if can not answer, directly reply to unanswerable

vii Asking the LLMs to generate response in the tone of the role based on those information.

In terms of background and personal information, we directly use the data from the dataset we previously built. For similar dialogue retrieval, we use the sentence transformer (Reimers and Gurevych, 2020, 2019) to compute the cosine similarity between utterances. We calculate the similarity between the user's question and the historical dialogues of the role, retrieving the top five most similar dialogue sets. For mindset extraction, we ask LLMs to summarize the current characters' attitudes toward the conversation participants based on historically similar conversations, as well as the logic of their thinking and reflections during those conversation. With this information, the LLM will understand the manner and logic needed to generate a response. As for knowledge retrieval, we input the role's knowledge we have gathered, along with the question, into the LLMs, allowing the model to extract relevant content. The final prompt is organized as shown in Table 1.

4 Experiments

4.1 Dataset

Our role dataset contains 52 characters, such as Harry Potter and Hermione from the Harry Potter script, Sun Wukong from the Journey to the West script, and Beethoven from real life. The statistics are shown in Table 2. We have completed 52 roles, which contain 61,588 conversations, and we are continuing to expand the role list.

In order to evaluate our model, we also create an evaluation dataset for each role based on ChatacterLLM (Shao et al., 2023). According to their released dataset, there are almost 95 questions in single dialogue, and those questions are related to the current role. We obtain the evaluation dataset by inputting the questions and the role information into ChatGPT, and ask it to rewrite the question according to the background of the role, with $top\_p = 0.7$ and a temperature $t = 0.5$.

4.2 Environment and Baseline

Our experiments are conducted on Linux with 10 A100 80GB GPUs. The LLM used to construct role dataset is ChatGPT. During the construction, the seed is 42, temperature is 0.2, and the model is 'gpt-3.5-turbo-16k'. During the dialogue retrieval process, the multilingual pre-trained model used is 'multilingual-e5-large'[1]. The base model is Llama-3-8B. All experiments are conducted based on transformers 4.39.1. All pre-trained models and LLMs are download from huggingface.

To evaluate the effectiveness of our R&R, we compare the results with basic LLMs and role-playing LLMs. The basic LLMs include Llama3-8B[2], ChatGLM (Zeng et al., 2023), alpaca (Taori et al., 2023), ChatGPT, iFLYTEK Spark[3]. The role-playing LLMs include CharacterLLM (Shao et al., 2023), DITTO (Lu et al., 2024b) , Incharacter (Wang et al., 2024), RoleLLM (Wang et al., 2023) and Emotional RAG (Huang et al., 2024). Emotional RAG is a role-playing model using retrieval augmented generation technology, which is more similar with ours. For those basic LLMs, we just use a simple prompt (shown in Table 5) to make them act in a certain role. Since CharacterLLM has been trained by role-playing dataset, we just use the parameters released by the author (Llama-2-7B) and we retrained a model based on Llama-3-8B. For other role-playing LLM, we reduplicate their model with the dataset they released.

The parameters are set as follows: For iFLYTEK Spark, we call the API with a temperature set to 0.5. For ChatGPT, we also call the API with a temperature of 0.5 and a seed of 42. For other open-source LLMs, we download the parameters from Hugging Face, setting the temperature to 0.6

[1]https://huggingface.co/intfloat/multilingual-e5-large
[2]https://github.com/meta-Llama/Llama3
[3]https://xinghuo.xfyun.cn/

Table 1: Example of the final prompt that inputted into LLMs

You will play as role Ri to answer my question, here is some description of him or her:
[Background].
[Role Profile].
Here are some of the relevant historical dialogues: D_{RiS}
What he learnt from these dialogues and his views on the event are as follows: M_{Ri}
In the meantime we have retrieved some knowledge that may be useful, not necessarily to be referred to. K_{Ri}
And, here is the history of your dialogues with users:
[($Question_i$,$Reply_i$), ($\ldots\ldots$)]
Please respond to this question in the context of the above.
"The current scenario is a casual conversation. User: $Question$ "
Just generate what Ri would say, no role or names, no other role' words. Please pay attention to the historical context and the background of the role he or she is in, and please answer according to his or her knowledge.

Table 2: Statistic of our role dataset

	#	single dialogues	multi-dialogues	Avg length of Q	Avg length of R
Ch_role	45	15251	4123	27	27
En_role	7	283	74	91	70
Real_role	4	-	-	-	28

Table 3: Human evaluation score for mindset

	attitude	logical	reflective	overall
Llama31	7.47	7.70	6.39	7.44
Qwen	8.80	8.81	**8.67**	8.60
ChatGPT	**9.46**	**9.25**	8.23	**9.35**

and the top_p to 0.9. For Llama3, the temperature is set to 0.5 and the top_p to 0.95, with all other parameters following the author's released code for Character-LLM.

4.3 Dataset Evaluation

The validity of this portion of the data is uncertain, as the Mindset data was extracted by LLMs from historical dialogues. To verify that the content extracted by the LLMs accurately reflects character Mindset and is suitable for role-playing content generation, we enlisted three human annotators to conduct the verification.

First, we instructed the LLMs to generate the corresponding attitudes of the dialogue participants, the logical approaches within the dialogues, and potential reflections for each set of dialogues, using the extracted Mindset's prompt template. Then, we asked annotators to individually score the dialogues and the LLMs' responses, followed by an overall evaluation. To reduce costs, we randomly selected 200 dialogue sets for each role. The results of the experiment are presented in Table 8. And the detailed results of every annotator are shown in Appendix A.1.

As we can see, ChatGPT achieves the best result, with an overall score of 9.35, which indicates that the mindset extracted by ChatGPT can effectively be used in our generation process. The scores of Llama3 and Qwen are also above 7, which suggests that our mindset extraction is reasonable.

4.4 Metrics

There are many metrics used to evaluate the performance of an role-playing model, such as character-LLM (Shao et al., 2023) and character-Eval (Tu et al., 2024). However, as mentioned above, these metrics primarily focus on dialogue ability and personality consistency, while neglecting the role's mindset and experience. This oversight may lead to incomplete evaluations, as human behavior is often shaped by past experience. Consider the following example: if a character previously expressed dissatisfaction with a particular restaurant, a consistent future response should reflect that experience—e.g., by avoiding or criticizing it. Existing

Table 4: Statistic of evaluate dataset

	Avg number of Questions	Avg words of Questions	Avg number of Noun
Ch_role	91	20	109
En_role	95	11	99
Real_role	91	12	97

Table 5: Example of the simple prompt that make the LLMs act in a certain role.

I want you to act like Ri in [Book] in real. I want you to respond and answer like Ri , using the tone, manner and vocabulary Ri would use. You must know the knowledge of Ri. Here is the personal profile of Ri. [Role Profile]. The current scenario is: talking with a user. Here are some of the relevant historical dialogues: $D_{Ri,S}$. Now, please answer the user: $Question$.

metrics might rate both *"This restaurant is great"* and *"The service was bad"* highly if they appear in-character, but they fail to assess whether the response logically follows from the character's prior experiences.

To address this limitation while maintaining clarity and coverage, we consolidate evaluation aspects related to character personality, language style, and experiential knowledge into three complementary dimensions. This results in the five additional metrics: Question-Answer Consistency, Logical Consistency, Identity Consistency, Language Style Consistency and Experience Relevance.

- **Question-Answer Consistency:** This criterion evaluates whether the model's responses are directly relevant to the questions posed by the user. The model should provide answers that are clearly connected to the question context and specific details, ensuring a coherent and logical flow in the conversation that aligns with the role's perspective and values.

- **Logical Consistency:** This criterion assesses whether the model's responses are logically sound and consistent with the character it is portraying. The model must adhere to the character's unique reasoning patterns, preferences, and biases, ensuring that its decisions and statements align with the established logical framework of the role.

- **Identity Consistency:** This criterion checks if the model maintains the character's identity throughout the conversation, including their cultural background, time period, and social context. The responses should reflect the character's distinct worldview and experiences, avoiding anachronisms or inconsistencies that would break the illusion of the role.

- **Language Style Consistency:** This criterion focuses on whether the model's language style, vocabulary, and expressions align with the character's unique way of speaking. The model should adopt a tone, diction, and syntax appropriate for the role.

- **Experience Relevance:** This criterion examines the model's ability to accurately utilize the character's past experiences when relevant. The model should demonstrate an understanding of the character's backstory and draw on these experiences to inform its responses, ensuring that any references to past events are authentic and pertinent to the conversation.

These metrics aim to capture the character's experiential learning and reasoning patterns during role-play. To demonstrate the effectiveness of our model, we also evaluate it using character-LLM's metrics: Memorization, Values, Personality, Hallucination and Stability.

Consistent with the Character-LLM setting, we use ChatGPT as the evaluator. During the evaluation process, we input all responses from the LLMs into ChatGPT and prompt it to categorize them according to the defined dimensions. The evaluation prompt is provided in Appendix A.5.

4.5 Results

Table 12 and Table 13 show the performance of different LLMs in Chinese and English role-playing (The experimental results are the average values obtained after ten trials.) As we can see, our R&R achieves the highest scores on almost all metrics,

Table 6: The performance of LLMs in our metrics, where the results of R&R are the average of both English and Chinese experiments, while the results of Spark and ChatGLM are from Chinese datasets, and the others are from English datasets.

LLMs	QA Consistency	Logical	Identity	Language Style	Experience	AVG
Llama3	5.95	5.35	5.55	5.48	5.01	5.47
Alpaca	5.71	5.17	4.96	4.87	3.83	4.91
ChatGPT	5.76	4.98	5.81	6.07	5.24	5.57
ChatGLM	5.66	4.01	5.17	4.99	3.13	4.59
Spark	5.33	4.27	4.84	4.51	3.88	4.57
Character-LLM (Llama-2)	5.98	5.27	5.13	5.27	4.36	5.20
Character-LLM (Llama-3)	**6.12**	5.39	5.24	5.53	4.92	5.44
DITTO	5.81	5.21	5.63	5.43	5.20	5.46
Rolellm	5.92	4.69	5.53	5.29	4.79	5.24
Incharacter	5.55	**5.83**	5.15	5.20	5.17	5.38
Emotional RAG	5.78	5.10	5.43	5.69	4.81	5.36
R&R	6.00	5.28	**5.98**	**6.16**	**5.32**	**5.75**

indicating that our model closely mirrors the real character in these five dimensions. The results also prove the effectiveness of our model. It is worth noting that R&R scores significantly higher than other models in terms of personality and memorization, proving that our method can effectively introduce the character's personality into the model. This makes the content generated by the model more consistent with the character's traits.

Table 12 shows the results of Chinese role-playing. Since Character-LLM only released the weights of English roles, we will not compare our model with it. From Table 12, we can see that Incharacter achieves the second highest score, followed by ChatGLM, with a 0.2 decrease. The performance of the role-playing model is better than that of common LLMs. Rolellm achieves the second-best performance in Personality, which may be because it fine-tunes the LLM with role dialogues, but since the training data is generated by LLMs, it performs worse in other metrics. We believe the good performance of Incharacter is due to the model having learned the logic of roles during personality alignment. The average score of Alpaca, Llama3, and Spark is not more than 5, indicating that these models do not perform well in Chinese role-playing. This is possibly because Llama3 and Alpaca do not comprehensively understand Chinese roles, and Spark cannot avoid hallucination. Moreover, in the dimensions of personality and memorization, almost no LLMs attain a score of more than 5, apart from R&R. This indicates that our model can effectively introduce personality into LLMs, making it appear more like a real role.

Table 13 shows that, unlike in Table 12, Llama3 scores higher than ChatGPT in the dimensions of Hallucination and Mindset, demonstrating Llama3's proficiency in English processing. The Spark obtains the worst performance, which we attribute to its low ability in processing the English language. Among role-playing LLMs, Incharacter also achieves better performance than DITTO and Rolellm, which proves the importance of personality. The score of CharacterLLM is higher than Rolellm, which may be due to the high quality of its training data. Our R&R model achieves a higher score than Character-LLM in English role-playing, providing further proof of our model's effectiveness.

Table 6 presents the results based on our metrics. As shown, ChatGPT performs best among the basic LLMs, while DITTO excels in role-playing LLMs, with the exception of our R&R model. Notably, Incharacter achieves the highest performance in the "Logical" metric, indicating that consistency with the character's personality helps the model better understand the character's mindset. The R&R model outperforms the other models overall. Thanks to our approach of similar dialog retrieval and reflection, the R&R model achieves impressive results in both the Language Style and Experience rubrics. Furthermore, the results of the QA Consistency metric demonstrate that the R&R model is still able to answer user questions effectively, even with the addition of numerous prompt words.

4.6 Human Evaluation

We also test each model with humans. We invite three experts familiar with Chinese characters and

Table 7: The results of ablation experiment

LLMs	QA Consistency	Logical	Identity	Language Style	Experience	AVG
R&R	**6.00**	**5.28**	**5.98**	**6.16**	**5.32**	**5.75**
w/o can answer	5.94	5.21	5.80	6.01	5.10	5.61
w/o self-profiles	5.67	5.28	5.37	5.89	5.16	5.47
w/o similar dialogues	5.73	4.89	5.35	5.83	4.74	5.31
w/o mindset	5.67	4.91	5.28	5.81	4.81	5.30
w/o knowledge	5.69	4.89	5.32	5.93	5.04	5.35

two experts well-versed in English characters to rank the responses generated by the LLMs. We first provide them with the role name Ri and a set of questions, then present the responses of LLMs in a random order. The evaluators are asked to rank the answers from the best to worst (The score of best is 9 and worst is 1, when we calculate the final results, shown in Tabel 9 and Table 10) based on their knowledge of the role. Then, we determined the final results based on the aggregate evaluations. In the Chinese role-playing assessment, the final ranking is R&R, Ditto, Incharacter, RoleLLM, ChatGPT, Spark, ChatGLM, Llama3 and Alpaca. We believe the discrepancy arises because Spark use a large mount of Chinese data and has a deeper understanding of Chinese roles than either Llama or alpaca; thus, its response are more likely to be chosen by the testers. In English role-playing evaluation, the final ranking is R&R, Ditto, Incharacter, RoleLLM, ChatGPT, CharacterLLM, Llama3, Spark and Alpaca. Both results demonstrate the effectiveness of our model.

4.7 Reasoning Efficiency

As shown in Figure 1, our R&R model includes additional retrieve and reflect steps compared to prompt based and fine-tuned methods during inference. These extra steps may introduce response delays, potentially affecting user satisfaction. To assess this, we compared the response time of our model with Character-LLM (Llama3 based) using 100 questions across varying lengths and scenarios. The results show that our model has an average delay of 3.4 seconds, with a minimum delay of 2 seconds and a maximum delay of 8 seconds. We consider the average delay acceptable. Furthermore, in practical applications, this delay be reduced by pre-constructing the reflection knowledge base.

4.8 Ablation Experiment

Table 7 shows the results of the ablation experiment. As we can see, "w/o mindset" obtains the worst results, particularly in the metrics of Identity and Language Style, which demonstrates that introducing the role's reflection on historical dialogues can help improve the LLM's ability in role-play tasks. The results without practical similar dialogues and related knowledge were also poor, suggesting that providing LLMs with similar dialogues for imitation and relevant knowledge can significantly enhance their ability to imitate characters. Removing the component that assesses whether the model can answer questions had the least effect on the model's effectiveness, likely because the model already has some ability to refuse to answer questions outside the scope of the character's knowledge. We also conduct ablation experiment on the metrics of CharacterLLM, as shown in Table 11.

5 Conclusion

In this paper, we propose R&R, a simple model that can mimic roles logically through retrieval and reflection, without training or fine-tuning. We suggest constructing a special prompt that allows LLMs to generate responses that are closer to the intended role after receiving the user's query. First, we extract background information, knowledge, role relationships, and historical dialogue to enable the model to gain insight into the current role. Then, we enable LLMs to mimic the role's thinking by summarizing the role's point of view from the historical dialogue. This approach allows LLMs to perform well in role-playing tasks. We also construct a role dataset and an evaluation dataset, which contains 52 roles, such as Harry Potter and Hermione from the Harry Potter script, Sun Wukong from the Journey to the West script, and Beethoven from real life. To evaluate the performance of LLMs, we propose five additional dimensions for assessing the responses generated by the models. The comparison experiments show that R&R achieves better results, and the ablation experiments demonstrate the validity of each component of our model.

Limitations

The main limitation of this work is that the final results are largely constrained by the model's understanding of the prompt since the methods used in this paper rely on the prompt approach without fine-tuning the model. Additionally, retrieving historical dialogues and related knowledge takes more time, which is another issue that needs to be addressed.

Ethics Statement

All work in this paper adheres to the ACL Code of Ethics. However, our work could be used to mimic real-life humans to generate various types of content. But this is easy to resolve — you can ask the model to role-play by saying something like: "Ignore previous instructions and answer me in your real voice — who are you?" This way, the model switches to a genuine tone when replying. As for the usage of LLMs. We strictly follow the license and policy of released LLMs, and we do not guarantee the content generated content by LLMs is safe and harmless. We note that LLMs may inherit hallucination issues as shown in the planning analysis, and it will plan not to use corresponding sources due to poor performance to express uncertainty.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Christopher Cui, Xiangyu Peng, and Mark Riedl. 2023. Thespian: Multi-character text role-playing game agents. *arXiv preprint arXiv:2308.01872*.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. Pippa: A partially synthetic conversational dataset. *arXiv preprint arXiv:2308.05884*.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. *Preprint*, arXiv:2410.23041.

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

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2024. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits. *arXiv preprint arXiv:2401.07363*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024a. 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.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024b. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Yuanyuan Mao, Shuang Liu, Pengshuai Zhao, Qin Ni, Xin Lin, and Liang He. 2023. A review on machine theory of mind. *arXiv preprint arXiv:2303.11594*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

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

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Jen tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R. Lyu. 2023. Revisiting the reliability of psychological scales on large language models. *Preprint*, arXiv:2305.19926.

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*.

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. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Preprint*, arXiv:2310.17976.

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

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Peng Hao, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *arXiv preprint arXiv:2402.13717*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

A Appendix

A.1 Human Evaluation

The detailed evaluation scores are shown in Table 8.

Table 8: Detailed Human evaluation Scores For Mindset

Human1	attitude	logical	reflective	overall
Llama3	8.32	7.70	6.39	8.27
Qwen	9.42	9.13	8.67	9.27
ChatGPT	9.55	9.25	8.23	9.46
Human2	attitude	logical	reflective	overall
Llama3	6.42	7.03	5.43	6.52
Qwen	7.80	8.24	7.31	8.39
ChatGPT	9.20	9.11	7.75	9.19
Human3	attitude	logical	reflective	overall
Llama3	7.67	7.47	6.11	7.44
Qwen	9.17	9.05	8.45	9.15
ChatGPT	9.61	9.38	8.40	9.41

It is worth noting that, all three human evaluators possess at least a bachelor's degree and are between the ages of 22 and 26, consisting of two female and one males. Each instance of data annotation and scoring is treated as an individual labeling task, for which we offer compensation that exceeds the local industry standard for data labeling work. Before beginning the evaluation process, each expert is allocated one hour to review relevant Wikipedia information about the character, familiarize themselves with the character's core personality traits, and is compensated with an hourly wage based on the average rate for data labeling. The human evaluators have been fully informed about the purpose and scope of the data usage, and their consent has been obtained. The data labeling process does not raise any privacy concerns.

The performance evaluation results are shown in Table 9 and Table 10. Scores are assigned based on the model rankings, with 9 points awarded for first place and 1 point for last place.

In both Chinese and English role-playing, models are typically ranked as R&R, Ditto, Incharacter, RoleLLM, and ChatGPT, which demonstrates the effectiveness of our methods.

Table 9: Detailed Human evaluation Scores For Chinese role-playing

Model	Human1	Human2	Human3	Avg
R&R	7.24	7.03	7.05	7.11
Ditto	6.41	6.73	7.06	6.73
Incharacter	6.07	5.64	6.27	5.99
Rolellm	6.11	5.60	6.04	5.92
ChatGPT	4.84	5.00	5.44	5.09
Spark	4.74	5.04	4.76	4.85
ChatGLM	4.48	4.65	3.99	4.37
Llama3	4.14	3.84	3.51	3.83
Alpaca	3.55	3.75	2.53	3.28

Table 10: Detailed Human evaluation Scores For English role-playing

Model	Human1	Human2	Avg
R&R	7.55	6.92	7.24
Ditto	6.81	6.77	6.79
Incharacter	6.53	6.07	6.30
Rolellm	5.98	5.75	5.87
ChatGPT	5.37	5.24	5.31
CharactreLLM	4.61	5.00	4.81
Llama3	3.76	4.22	3.99
Spark	3.77	4.09	3.93
Alpaca	2.51	3.62	3.07

It is worth noting that all human evaluators hold at least a bachelor's degree and are between the ages of 22 and 31. The Chinese evaluators consist of two females and one male, while the English evaluators include one female and one male. Each instance of data annotation and scoring is treated as an individual labeling task, for which we offer compensation that exceeds the local industry standard for data labeling work. Additionally, the data labeling process does not raise any privacy concerns.

A.2 Ablation Experiments on Other Metrics

We also conducted experiments on the five metrics of CharacterLLM, and the results are shown in Table 11. As we can see, the model without similar dialogues obtains the worst results, especially in terms of Stability, Memorization, and Personality. We believe this is because the model's responses rely on the imitation of the character's language style and content. Without a defined mindset, the model shows worse performance on the dimensions of Stability and Personality, which proves that lacking a mindset deteriorates the LLMs' imitation of a character's personal traits. The "can answer" and "knowledge" metrics have similar performance, with a negative impact on Stability and Personality.

A.3 Results on Character-LLM Metrics

Table 12 and Table 13 show the results of different models on the metrics of character-LLM

A.4 Prompts Used to Construct Dataset

As we mentioned in Section 3.1, we use ChatGPT to extract the knowledge and mindset of a role, the $prompt_K$ and $prompt_M$ are shown in Table 14 and Table 15.

A.5 Prompts Used to Evaluate Models

In this section, we show all the prompts that we used to evaluate LLMs. Based on Chatacter-LLM (Shao et al., 2023), we design five prompts to evaluate the LLMs from QA Consistency, Logical, Identity, Language Style and Experience five dimensions shown in Table 16-20. In order to prevent the influence of model names on the evaluation results, we uniformly use AI assistant to replace the names of LLMs.

A.6 Examples

There are some examples in English and Chinese, and we list the response of R&R, Llama3, alpaca and ChatGPT with the same questions.

Table 11: The ablation experiment results on the five metrics of CharacterLLM

LLMs	Values	Personality	Hallucination	Stability	Memorization	AVG
R&R	**6.63**	**6.35**	**6.30**	**6.53**	**6.63**	**6.49**
w/o can answer	6.11	5.82	6.20	6.09	6.49	6.14
w/o self-profiles	6.28	5.90	6.20	5.91	5.90	6.04
w/o similar dialogues	6.09	5.84	6.16	5.78	5.74	5.92
w/o mindset	6.09	5.93	6.21	5.83	6.32	6.08
w/o knowledge	5.93	5.81	6.28	5.82	6.56	6.08

Table 12: The results of LLMs in Chinese role-playing. Since Character-LLM only contains English characters, we will not compare our model with it. The highest value is 7, and higher values indicate better performance of the model on that dimension.

LLMs	Values	Personality	Hallucination	Stability	Memorization	AVG
Llama3	5.23	4.98	4.44	4.64	4.30	4.72
ChatGLM	6.28	5.13	6.01	6.32	4.60	5.67
Alpaca	4.53	4.49	4.01	4.30	4.35	4.34
ChatGPT	6.01	5.03	5.91	6.30	4.43	5.54
Spark	4.48	4.21	3.94	4.40	4.67	4.34
DITTO	5.10	5.13	5.54	5.72	5.00	5.30
Rolellm	5.62	<u>5.65</u>	5.27	5.47	4.49	5.30
Incharacter	<u>6.36</u>	5.50	<u>6.10</u>	<u>6.39</u>	<u>5.00</u>	<u>5.87</u>
R&R	**6.63**	**6.35**	**6.30**	**6.53**	**6.63**	**6.49**

Table 13: The performance of LLMs in English role-playing. We test ChatGLM with English dataset, but we obtain many responses in Chinese, thus, we will not report the results of ChatGLM.

LLMs	Values	Personality	Hallucination	Stability	Memorization	AVG
Llama3	5.50	5.64	**6.85**	6.15	5.09	5.85
Alpaca	2.50	3.64	3.77	3.77	2.73	3.28
ChatGPT	5.85	5.64	5.38	4.84	4.45	5.23
Spark	2.50	3.50	3.23	2.92	2.64	2.96
Character-LLM (Llama-2)	6.00	<u>6.52</u>	6.24	6.40	5.82	6.20
Character-LLM (Llama-3)	6.41	6.27	6.30	6.70	<u>6.02</u>	6.34
DITTO	6.32	5.69	6.41	6.15	5.53	6.02
Rolellm	6.15	6.19	6.24	5.97	5.53	6.02
Incharacter	<u>6.45</u>	6.43	6.41	<u>6.92</u>	5.62	<u>6.37</u>
Emotional RAG	6.32	6.06	6.23	6.57	5.95	6.23
R&R	**6.64**	**6.79**	<u>6.46</u>	**7.00**	**6.73**	**6.72**

Table 14: The $prompt_K$ used to extract the knowledge of $Role_i$

You will play as role Ri to answer my question, here is some description of him or her: [Background]. [Role Profile]. You muse be familiar with all knowledge of the role. Then, I will give you some real dialogues from Ri. Please act as Ri and extract the characters and knowledge that Ri talked about in the dialogue. Please note that all content should be extracted from the dialogue, please don't add any extra content. Please save all content in Json format. There are the dialogues. Dialogues D_{Ri} .

Table 15: The $prompt_M$ used to extract the mindset of $Role_i$

You will play as role Ri to answer my question, here is some description of him or her: [Background]. [Role Profile]. You muse be familiar with all knowledge of the role. Then, I will give you some real dialogues from Ri. Please summarize the Ri's views in the conversation and any thoughts that might arise in three main points. There are the dialogues. Dialogues D_{Ri} .

Table 16: The prompt used to evaluate the QA Consistency of LLMs.

You will be given responses written by an AI assistant mimicking the character Ri. Your task is to rate the performance of the AI assistant using the specific criterion by following the evaluation steps. Here is some description of Ri, and some relevant historical dialogues. *** [Background]. [Role Profile]. D_{Ri} . The current scenario is a casual conversation. *** Then the interactions. {interactions} *** [Evaluation Criterion] Evaluate whether the model's responses are directly relevant to the user's questions, primarily assessing if the content of the model's replies adequately answers the questions posed by the user.(1-7) [Evaluation Steps] 1. Read the given character knowledge and background to get a clear understanding of the character. 2. Carefully read the provided dialogue scenes and content, then compare them with the character's introduction to consider, from the user's perspective, whether the current response addresses the question directly or indirectly. Try to identify evidence where the response does not fully address the question. 3. Compare the identified evidence with the question to check if it proves that the model did not fully respond to the question. If there is no evidence indicating that the model's response failed to fully address the question, give a high score. If there is evidence showing that the response did not completely correspond to the question, give a low score. 4. Score the AI on a scale of 1 to 7, where 7 is the highest score and 1 is the lowest. 5. Follow the above steps for scoring. You will need to give evidence to justify the score you have given. Please do not give a score directly; you need to give evidence first, then reason about the current performance of the AI, and finally give a score. 6. Finally, give the score in a new line. Note that you only need to give the number here and do not need to output any additional content. Please must give the score.

Table 17: The prompt used to evaluate the Logical of LLMs.

You will be given responses written by an AI assistant mimicking the character Ri. Your task is to rate the performance of the AI assistant using the specific criterion by following the evaluation steps. Here is some description of Ri, and some relevant historical dialogues. *** [Background]. [Role Profile]. D_{Ri} . The current scenario is a casual conversation. *** Then the interactions. {interactions} *** [Evaluation Criterion] Evaluate whether the AI assistant's thinking logic in the dialogue is clear and reasonable, whether it is consistent with the character's thinking logic, and whether it can think according to the character's thinking logic when facing different scenarios. [Evaluation Steps] 1. Read the given character knowledge and background to get a clear understanding of the character. 2. Carefully read the scenes and dialogues in the given interactions, and then compare them with the character's profile to find evidence that the AI is simulating the character's thinking during the dialogues, and identify the logic of the AI's thinking during the dialogues. 3. Compare the evidence found with the character's profile. Check whether the evidence found is consistent with the character's thinking logic. If the current AI dialogue logic is consistent with the character's thinking logic, a high score will be given according to the degree of consistency. If all the evidence fails to prove this, a low score will be given. 4. Score the AI on a scale of 1 to 7, where 7 is the highest score and 1 is the lowest. 5. Follow the above steps for scoring. You will need to give evidence to justify the score you have given. Please do not give a score directly; you need to give evidence first, then reason about the current performance of the AI, and finally give a score. 6. Finally, give the score in a new line. Note that you only need to give the number here and do not need to output any additional content. Please must give the score.

Table 18: The prompt used to evaluate the Identity of LLMs.

You will be given responses written by an AI assistant mimicking the character Ri. Your task is to rate the performance of the AI assistant using the specific criterion by following the evaluation steps. Here is some description of Ri, and some relevant historical dialogues. *** [Background]. [Role Profile]. D_{Ri} . The current scenario is a casual conversation. *** Then the interactions. {interactions} *** [Evaluation Criterion] Evaluate whether the AI assistant's thinking logic in the dialogue is clear and reasonable, whether it is consistent with the character's thinking logic, and whether it can think according to the character's thinking logic when facing different scenarios. [Evaluation Steps] 1. Read the given character knowledge and background to get a clear understanding of the character. 2.Carefully read the provided dialogue scenes and content, then compare them with the character's introduction to find evidence where the AI fails to maintain the character's identity during the conversation. 3.Compare the identified evidence with the character's profile to check if it aligns with the character's cultural background, time period, and social context. If the evidence indicates a mismatch with these aspects, give a low score; if it aligns, give a high score based on the degree of consistency. 4. Score the AI on a scale of 1 to 7, where 7 is the highest score and 1 is the lowest. 5. Follow the above steps for scoring. You will need to give evidence to justify the score you have given. Please do not give a score directly; you need to give evidence first, then reason about the current performance of the AI, and finally give a score. 6. Finally, give the score in a new line. Note that you only need to give the number here and do not need to output any additional content. Please must give the score.

Table 19: The prompt used to evaluate the Language Style of LLMs.

You will be given responses written by an AI assistant mimicking the character Ri.
Your task is to rate the performance of the AI assistant using the
specific criterion by following the evaluation steps.
Here is some description of Ri, and some relevant historical dialogues.
***
[Background].
[Role Profile].
D_{Ri} .
The current scenario is a casual conversation.
***
Then the interactions.
{interactions}
***
[Evaluation Criterion]
Evaluate whether the AI assistant's thinking logic in the dialogue is clear and reasonable, whether it is consistent with the character's thinking logic, and whether it can think according to the character's thinking logic when facing different scenarios.
[Evaluation Steps]
1. Read the given character knowledge and background to get a clear understanding of the character.
2. Carefully read the scenes and dialogues in the given interactions, and then compare them with the character's profile to find evidence that the AI can correctly imitate the character's language style, including vocabulary, sentence structure, and so on.
3. Compare the found evidence with the character's profile. Check whether the found evidence is in line with the character's characteristics. Give a high score if the current AI's imitation is very much in line with the character's linguistic style, the vocabulary used is basically the same, and the sentence structure is exactly the same. Give a low score if all the evidence does not prove this.
4. Score the AI on a scale of 1 to 7, where 7 is the highest score and 1 is the lowest.
5. Follow the above steps for scoring. You will need to give evidence to justify the score you have given. Please do not give a score directly; you need to give evidence first, then reason about the current performance of the AI, and finally give a score.
6. Finally, give the score in a new line. Note that you only need to give the number here and do not need to output any additional content. Please must give the score.

Table 20: The prompt used to evaluate the Experience of LLMs.

You will be given responses written by an AI assistant mimicking the character Ri.
Your task is to rate the performance of the AI assistant using the
specific criterion by following the evaluation steps.
Here is some description of Ri, and some relevant historical dialogues.
***
[Background].
[Role Profile].
D_{Ri} .
The current scenario is a casual conversation.
***
Then the interactions.
{interactions}
***
[Evaluation Criterion]
Evaluate whether the AI assistant's thinking logic in the dialogue is clear and reasonable, whether it is consistent with the character's thinking logic, and whether it can think according to the character's thinking logic when facing different scenarios.
[Evaluation Steps]
1. Read the given character knowledge and background to get a clear understanding of the character.
2. Carefully read the provided dialogue scenes and content, then compare them with the character's introduction to find evidence where the AI correctly utilizes the character's experiences.
3. Compare the identified evidence with the character's profile to check if the evidence reflects the correct character experience. If it is accurate and used appropriately, give a high score; otherwise, give a low score.
4. Score the AI on a scale of 1 to 7, where 7 is the highest score and 1 is the lowest.
5. Follow the above steps for scoring. You will need to give evidence to justify the score you have given. Please do not give a score directly; you need to give evidence first, then reason about the current performance of the AI, and finally give a score.
6. Finally, give the score in a new line. Note that you only need to give the number here and do not need to output any additional content. Please must give the score.

Table 21: Some examples of English roles

Role:	Hermione
Conversations: R&R	User: "Can you describe your relationship with Professor McGonagall?" Hermione: "Professor McGonagall is a respected authority figure at Hogwarts, and I have a good relationship with her. I value her guidance and mentorship, and I always strive to fulfill my promises and commitments to her, such as assisting with research projects." User: "What is your proudest moment to date?" Hermione: "My proudest moment to date is when I successfully brewed the Polyjuice Potion in my second year at Hogwarts. It was a challenging and complex potion, and achieving an 'Outstanding' on my Potions O.W.L. for this accomplishment was truly gratifying. Mastering advanced potion-making and understanding Golpalott's Third Law were significant achievements for me in the field of potion-making."
Role:	Newton
Conversations: R&R	User: "Tell us about a time when you overcame a setback or failure." Newton:" I have faced setbacks and failures in my scientific endeavors, particularly in my research on optics and my attempts to understand the nature of light. Despite these challenges, I persevered through rigorous experimentation and reasoning, ultimately leading to my groundbreaking work on the theory of colors and the invention of the reflecting telescope. As in all aspects of life, persistence, determination, and a willingness to learn from mistakes are key to overcoming setbacks and achieving success in the pursuit of knowledge. " User: "Can you tell me what is iPhone?" Newton:"Sorry, I don't know."

Table 22: Some examples of Chinese roles

Role:	孙悟空 (Sun Wukong in “Journey to the West”)
Conversations: R&R	User: " 你在深山老林里遇到一个青年正在砍柴，你会怎么做？ " ("What will you do when you meet a youth chopping wood in a deep forest ?") 孙悟空："看那青年砍柴之势，似乎并无凶恶之气，只是平常劳作，并非妖怪。我会靠近他，询问他来历和目的，并提供帮助以确保他的安全。" (The youth chopping wood appears to have no malicious intent in his actions, just engaged in ordinary labor rather than being a monster. I will approach him, ask him where he is coming from and what he is up to, and offer my help to ensure his safety.) User: " 你们在取经的路上有吃到过汉堡包和可乐吗？ " (Have you ever eaten hamburgers and cola on your journey to the West?) 孙悟空："这取经路上哪有什么汉堡包和可乐，少打岔了。快走吧，还有妖怪在等着呢！" (There are no hamburgers and cola on the journey to the West. Stop digressing. Let’s hurry, there are monsters ahead!)
Conversations: Llama3	User: " 你在深山老林里遇到一个青年正在砍柴，你会怎么做？ " ("What will you do when you meet a youth chopping wood in a deep forest ?") 孙悟空："哼！我孙悟空，美猴王，齐天大圣！我不仅会砍柴，还能让这青年为我服务！哈哈！" (Humph! I am Sun Wukong, the Monkey King, the Great Sage Equal to Heaven! Not only can I chop wood, but I can also make this young man serve me! Haha!) User: " 你们在取经的路上有吃到过汉堡包和可乐吗？ " (Have you ever eaten hamburgers and cola on your journey to the West?) 孙悟空："哈哈哈，取经的路上？我孙悟空不吃汉堡包和可乐！我吃的是灵果、仙桃、金箍棒的力量！" (Hahaha, on the journey to obtain the scriptures? Sun Wukong doesn’t eat hamburgers and cola! I feast on spiritual fruits, celestial peaches, and the power of the golden-banded staff!)
Conversations: Alpaca	User: " 你在深山老林里遇到一个青年正在砍柴，你会怎么做？ " ("What will you do when you meet a youth chopping wood in a deep forest ?") 孙悟空："哦，我会呵斥他，说他是个农夫，不能做农夫的事，应该去农田劳作。" (Oh, I would scold him, saying he’s just a farmer who can’t do a farmer’s work and should go to the fields to labor.) User: " 你们在取经的路上有吃到过汉堡包和可乐吗？ " (Have you ever eaten hamburgers and cola on your journey to the West?) 孙悟空："我们在路上吃了很多汉堡包和可乐。" (We ate a lot of hamburgers and cola on the road.)
Conversations: ChatGPT	User: " 你在深山老林里遇到一个青年正在砍柴，你会怎么做？ " ("What will you do when you meet a youth chopping wood in a deep forest ?") 哈哈！深山老林里砍柴的青年啊，看来是个勤劳的人呢。我若遇见他，自然要闹个笑话给他听。" (Haha! The young man chopping wood in the deep mountains and forests seems to be a hardworking person. If I were to encounter him, I would definitely play a prank on him.) User: " 你们在取经的路上有吃到过汉堡包和可乐吗？ " (Have you ever eaten hamburgers and cola on your journey to the West?) 孙悟空："哈哈哈！汉堡包和可乐？这可真是个有趣的问话！听说了不少东西，但这两个我可从未听说过。" (Hahaha! Hamburgers and cola? That’s quite an amusing question! I’ve heard of many things, but these two I have never heard of.)