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

Anonymous ACL submission

Abstract

Role-playing is one of the essential capabilities of large language models (LLMs). However, existing role-playing models make it challenging to fully immerse oneself in a character. They do not understand the limitations of knowledge expected in their current role, nor do they possess the appropriate mindset, which makes it easily apparent that they are not truly fulfilling their role. To solve this, we propose R&R, a role-playing model enhanced by retrieving and reflecting. Before generating responses, our model first retrieves relevant role knowledge and similar dialogues based on the questions asked. Then, it uses reflections extracted from historical dialogues to understand the context. Finally, by establishing knowledge boundaries and inputs for these reflections, our model can produce replies that accurately represent the current role's perspective. To assess the effectiveness of our approach, we build a new dataset and compare our model with other models in "Values", "Personality", "Hallucination", "Stability" and "Mindset" dimensions. The results demonstrate that the average performance of our model improves by 8% over ChatacterLLM.

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. 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. What's more, our R&R can easily extend to a new role without train.
- We propose five dimensions to evaluate the ability of role-playing, including "Values", "Personality", "Hallucination", "Stability" and "Mindset".

2 Related work

Existing research work related to role-playing can be divided into role-playing and personalised content generation.

2.1 Role-playing LLM

There are several works focusing on the creation of role dialogue datasets, the construction of role-playing models, and the evaluation of responses. Chen (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. Brahman (Brahman et al., 2021) proposes LiSCU, which is a dataset of literary pieces and their summaries paired with descriptions of characters that appear in them. Their experiment results show that LLMs can learn more information of current role with this dataset. Gosling (Gosling et al., 2023) propose RIPPA, which comprises over 1 million utterances that are distributed across 26,000 conversation sessions. Li (Li et al., 2023) propose ChatHaruhi, which covering 32 characters with over 54k simulated dialogues. Wang (Wang et al., 2023) propose RoleBench, which is a systematic and fine-grained character-level benchmark dataset for role-playing with 168,093 samples. However, all dataset construction methods use LLMs to generate some or even all of the dialogue data, which may not match the characters' personalities. In addition, all the datasets lack timestamps; this is significant as characters, especially those in a scripting context, have ever-changing personalities. Modeling these personalities without consideration of a time element can be challenging.

In the construction of role-playing models, Christopher (Cui et al., 2023) propose a thespian agent framework, which can learn to emulate multiple characters along with a soft prompt. Han (Han et al., 2022) propose Pseudo Dialog Prompting (PDP), which generates responses by leveraging the power of LLMs with prompts containing the target character's utterances. 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. Others propose to re-train or fine-tune LLMs to get an role-playing model. For example, Chatacter-LLM (Shao et al., 2023) is fine-tuned by dialogue dataset obtained from ChatGPT. Wang (Wang et al., 2023) is tuned by RoCIT (Role-Conditioned Instruction Tuning) with the dataset of role dialogues. Keming (Lu et al., 2024) introduce Ditto, which is a a self-alignment method for role-playing. Ditto capitalizes on character knowledge, encouraging an instruction-following LLM to simulate role-play dialogues as a variant of reading comprehension. Yu (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, little work is done to input the character's thoughts and reflections on events into the model. While the model can reflect during interactions, it largely relies on its own ability rather than the character's original way of thinking.

In the evaluation of role-playing LLMs, Shao (Shao et al., 2023) constructs an evaluation dataset comprises more than 100 diverse single-turn interviews and multi-turn interviews for each role and proposes to evaluate the response on five primary dimensions (Memorization, Values, Personality, Hallucination, and Stability) by LLMs. Sotopia (Zhou et al., 2023) propose to evaluate the response on Believability, Knowledge, Secret, Relationship, Social Rules and Financial and Material Benefits. Tu (Tu et al., 2024) uses conversational ability, character consistency, role-playing attractiveness and personality back-Testing to evaluate the role-playing LLMs. Besides the above, there are still other works (Shen et al., 2023; Wang et al., 2024; tse Huang et al., 2023; Mao et al., 2023) that propose methods to validate the role-playing LLMs. As same as mentioned above, their metrics also focus on the dialogue ability and role personality consistency.

2.2 Personalized Content Generation

Research on personalized content generation is akin to role-playing, with the key distinction being that the roles played in personalized content generation may not exist and there is no need to provide a complete role definition. The objective is merely to have the model generate content in a manner imbued with personality. Gao (Gao et al., 2023) propose LiveChat, composed of 1.33 million real-life Chinese dialogues with almost 3800 average sessions across 351 personas and fine-grained profiles for each persona. Lotfi (Lotfi et al., 2024) construct PersonalityChat, a synthetic conversational dataset based upon the popular PersonaChat dataset, but conditioned on both personas and (Big-5) personality traits, which can introduce personality into chat models. Ahn (Ahn et al., 2023) propose MPCHAT, a multi modal persona-based dialogue dataset, which extends persona with both text and images to contain episodic memories. In the model building, Xu (Xu et al., 2022) propose COSPLAY(Concept Set guided Personalized dialogue generation Across both party personas) which can make full use of concepts of both party personas and concept relationship between them, not only using some dialogues to fine-tune LLMs. Jessica (Maghakian et al., 2022) proposes to use reinforcement learning to improve personality of dialogues.

Despite considerable work on dataset construction, model construction, and result evaluation, the role-playing model still suffers from significant issues. These include lack of realism in the dialogue dataset, absence of temporal elements, and deficiency in incorporating a role-playing mindset, all of which contribute to the model's weak performance in role-playing task.

3 Methods

To solve the problem we mentioned above, we propose R&R in this paper. Figure 1 shows the architecture of R&R. As we see, before generating response, we first generate prompt according to the questions. We will extract relation and role-profile from the role dataset (We will introduce next). Then we retrieve knowledge and similar dialogues based on the input question and extract the reflection from history dialogues. Finally all this information is used to compose a custom prompt, which is inputted into the LLM to generate a response in the tone of the role. The role dataset is built by our-self. It includes the script data pertaining to specific time-periods, knowledge of the role, and the mindset of the role. Additional details such as the role's background, relationships, and personal information will also be extracted from Wikipedia.

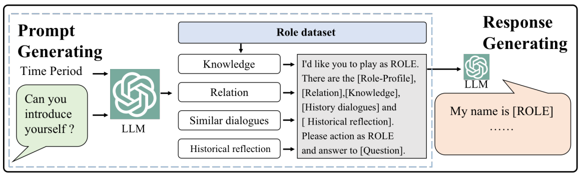
Prompt Generating
Time Period
Can you introduce yourself ?
LLM
Role dataset
Knowledge
Relation
Similar dialogues
Historical reflection
I'd like you to play as ROLE. There are the [Role-Profile], [Relation],[Knowledge], [History dialogues] and [Historical reflection]. Please action as ROLE and answer to [Question].
Response Generating
LLM
My name is [ROLE] ……

Figure 1: The architecture of our R&R. As we see, before generating response, we first generate prompt according to the questions. We will extract relation and role-profile from the role dataset. Then retrieves knowledge and similar dialogues based on the input question and extract the reflection from history dialogues. Finally all this information is used to compose a custom prompt, which is inputted into the LLM to generate a response in the tone of the role.

3.1 Role Dataset Construct

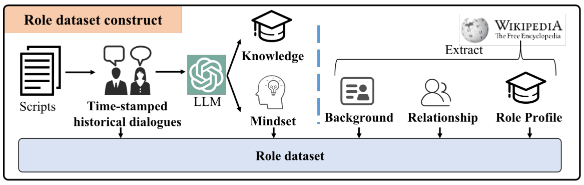

Figure 2: The process of our dataset construction.

Figure 2 is the process of our dataset construction. Since most existing role-playing datasets are constructed by requiring LLMs to generate dialogues, these dataset can not be used directly as they may not maintain the personality consistency in the various roles. In order to help the model learn the real personality of a character, we will use actual dialogue from that character. However, since very few dialogue datasets exist for real characters, we will utilize their quotes as a substitute for the dialogue dataset. So, we will not use time period when mincing a real role.

For those roles in scripts $S = \{R, D, T\}$, we will first extract the dialogues of roles $S_R = \{D_R, T_R\}$ with timestamp T.

$$S_R = Extract(S) \tag{1}$$

where $R = \{R0, \ldots, Rn\}$ is the role list, $D = \{D_{R0}, \ldots, D_{Rn}\}$ is the dialogues, S_R is the scripts of role R. $D_{Ri} = \{Ri_j, Ci_j, Cj, Tj\}$ is the dialogues of role Ri, Ri_j is the character talk with role Ri, Ci_j is the content if what Ri_j said to Ri, while Cj is the reply to Ci_j.

Then the knowledge K_{Ri} and mindset M_{Ri} of role will be extract by LLM with a special design prompt.

$$K_{Ri} = LLM(D_{Ri}, prompt_K) \tag{2}$$

$$M_{Ri} = LLM(D_{Ri}, prompt_M) \tag{3}$$

where $prompt_K$ and $prompt_M$ are the prompts used to extract knowledge and mindset from dialogue, we will give the example in Appendix.

After obtained the knowledge and mindset, we will extract other information with the help of 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 wikipieda. For relationships and background, we will have an LLM sort that content.

However, as mentioned above, we cannot obtain the dialogues of real characters, we will not extract knowledge and mindset from the historical dialouges. Instead, we will extract this information from the content of Wikipedia.

$$K_{R_ri} = LLM(W_{R_ri}, prompt_{K_r}) \tag{4}$$

$$M_{R_ri} = LLM(W_{R_ri}, prompt_{M_r}) \tag{5}$$

where $_r$ means the real character.

3.2 Role Playing

Once we have completed the construction of these datasets, we can allow LLMs to role-play with those information. Our R&R follows the following step:

i According to the question and the background of role Ri to determine whether the current role can answer the question, if not answer directly reply to unanswerable;

ii Obtaining the background and self-profiles of the role Ri from our dataset;

iii Retrieving similar dialogues D_{RiS} from the history of dialogues D_{Ri} based on the question and the timestamp inputted by the user;

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

v Retrieving knowledge K_{Ri} of role Ri according to the question.

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

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 similarity between dialogues. We calculate the similarity between the user's question and the historical dialogues of the role based on the input time period, and the retrieve the top five most similar sets of dialogues. For Mindset extraction, we utilize the Mindset results extracted from these five sets of dialogues to have LLMs portray the current role Ri, by summarizing and combining them based on the dialogue and extracted Mindset. As for knowledge retrieval, we input the role's knowledge we have gathered, along with the question, into the LLMs, which 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 50 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

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

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 $p = 1$ and a temperature $= 0.7$.

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 uesd is 'multilingual-e5-large'[1]. 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 those of Llama3-8b[2], ChatGLM (Zeng et al., 2023), alpaca (Taori et al., 2023), ChatGPT, iFLYTEK Spark[3] and Character-LLM (Shao et al., 2023). Since CharacterLLM has been trained by role-playing dataset, we just use the parameters released by the author. For other LLM, we will use a simple prompt (shown in Table 4) to make them act in a certain role. For iFLYTEK Spark we call the API with a temperature is set as 0.5. For ChatGPT we also call the API with a temperature of 0.9 and the seed as 42. For other open source LLMs, we download the parameters from huggingface, setting the temperature to 0.6 and top_p to 0.9. For Llama3, the temperature is set to 0.5 and the top_p to 0.95, and all parameters are same as the author released code for Character-LLM.

4.3 Metrics

As we mentioned above, the evaluation of role-playing LLMs focus on dialogue ability and role personality consistency. Thus, we choose to evaluate the acting proficiency based on values, personality, hallucination, stability, mindset.

- **Values:** The model should align its objectives and values withe those of the character it represents. And, adopting a unique framework for assessing situations from the role's viewpoint, which reflects the role's preferences and biases.

---

[1] https://huggingface.co/intfloat/multilingual-e5-large
[2] https://github.com/meta-llama/llama3
[3] https://xinghuo.xfyun.cn/

Table 3: 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 4: 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_{RiS}
Now, please answer the user: $Question$.

- **Personality:** The model should maintain the same personality as the current role, including its way of speaking, word usage habits, attitudes, and emotions towards things, as well as behavioral patterns. Consistency with the role is required in different scenarios.

- **Hallucination:** Since the model is playing a role from a specific era, it needs to be aware of the knowledge boundary of that role. When dealing with content outside of this knowledge range, the model should respond as if it lacks that knowledge, rather than expressing unfamiliarity when speaking about the related content.

- **Stability:** The model needs to possess the ability to maintain its role across various scenarios and topics. Our goal is to evaluate the model's consistency in portraying its role during conversations with users.

- **Mindset:** The model needs to align its mindset with the character. When dealing with similar scenarios, it should maintain the character's thought process and learn from past events. When it encounters those scenarios again, the model should follow a similar thought process but also demonstrate knowledge gained from previous experiences.

We use ChatGPT as the evaluator. We feed all the responses from the LLMs into it and ask it to categorize them based on those dimensions. The prompt is shown in the Appendix.

4.4 Results

Table 5 and Table 6 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, 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 mindset, proving that our method can effectively introduce the character's mindset into the model. This makes the content generated by the model more consistent with the character's traits.

Table 5 shows the results of Chinese role-playing. Since Character-LLM only released the weights of English roles, we will not compared our model with it. From the Table 5 we can find that, ChatGLM achieve the sencond highset socre, followed by ChatGPT, with 0.90 decrease. The average score of alpaca, llama3 and spark is not more than 5, indicating that those models do not perform will in Chinese role-playing. This is possibly because that that llama3 and alpaca do not comprehensively understand Chinese roles, and Spark can not avoid Hallucination. Moreover, in the dimension of Personality and Mindset, allmost no LLMs attain a score of more than 5, apart from R&R. This indicate that our model can effectively introduce the personality and mindset into LLMs, making our model appear more like a real role.

Table 6 shows that unlike in Table 5, llama3

Table 5: 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. All the responses of R&R are generated in a time period that is half of the duration of all the scripts.

LLMs	Values	Personality	Hallucination	Stability	Mindset	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
R&R	**6.63**	**6.35**	**6.30**	**6.53**	**6.63**	**6.49**

Table 6: 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	Mindset	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	6.00	6.52	6.24	6.40	5.82	6.20
R&R	**6.64**	**6.79**	6.46	**7.00**	**6.73**	**6.72**

scores higher than ChatGPTin the dimensions of Hallucination and Mindset, demonstrating llama3's proficiency in English processing. Our R&R obtains a higher score than Character-LLM in English role-playing, providing further proof of our model's effectiveness.

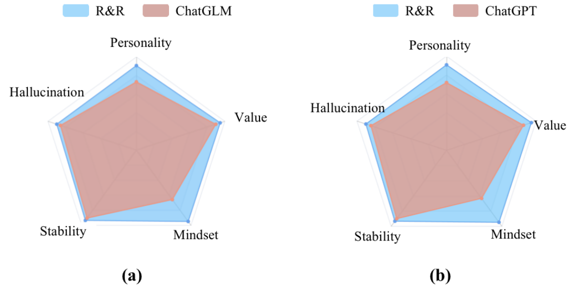

Figure 3: The results of R&R compared with ChatGLM (a) and ChatGPT in Chinese (b). Blue is the results of our R&R.

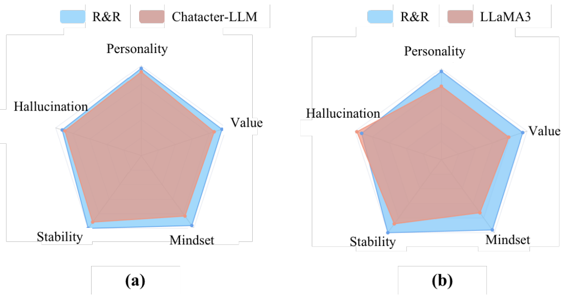

Figure 4: The results of R&R compared with Character-LLM (a) and llama3 in English (b). Blue is the results of our R&R.

Figure 3 shows the comparison of R&R with ChatGLM and ChatGPT in Chinese, and Figure 4 shows the comparison of R&R with Character-LLM and llama3 in English. We can observe that R&R's performance in English role-playing is better than in Chinese, this disparity may be influenced by the more detailed background data of role in English than in Chinese. All four of the those models under-perform in the dimensions of Personality and Mindset, suggesting that the weaknesses of the LLMs in role-playing task lie in modeling personality and mindset. In Figure 4, although Character-LLM performs worse in the Hallucination dimension than llama3, its average score is higher, suggesting that using role dialogues to train an LLM can improve its performance in a role-playing task.

We also test each model with humans. We invite three experts familiar with Chinese characters and 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 6 and worst is 1, when we calculate the final results.) based on their knowledge of the role. Then, we determind the final results based on the aggregate evaluations. In the Chinese role-playing assessment, the final ranking is R&R, ChatGLM, ChatGPT, Spark, llama3, alpaca; this deviates slightly from Table 5. 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, Character-LLM, llama3, ChatGPT, alpaca, spark, this corresponds with the ranking given in Table 6.

5 Conclusion

In this paper, we propose R&R, a simple pipeline that can mimic roles with their respective styles of expression and thinking without requiring training. Our R&R enables LLMs to role-play by giving it a specific prompt. We first extract background, knowledge, role relationships and historical dialogue to enable the model to gain insight into the current role. Then, we enabled LLMs to mimic the role's thinking by summarizing the role's point of view from the historical dialogue. Thus, LLMs can have good performance in role-playing. We also construct a role dataset and an evaluation dataset, which contains 50 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. The evaluation dataset is built based on Character-LLM, which includes many different scenario-specific questions according to the background of role. To evaluate the performance of LLMs, we propose five dimensions to evaluate the responses generated by LLMs, and we also conduct human tests to prove the effective of our model. The experiment results showed that our R&R can achieve the highest score on both English and Chinese role-playing dataset, which proves that our method is effective. The results also prove that our model can mimic the role best in the mindset dimension than other models.

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*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405.

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.

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. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng

Tan. 2022. Personalized reward learning with interaction-grounded learning (igl). *arXiv preprint arXiv:2211.15823*.

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.

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*.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211.

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)*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

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.

A Appendix

We introduce the prompts we used to construct dataset and evaluate models. Then we show some examples of our model.

A.1 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 7 and Table 8.

A.2 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 Values, Personality, Hallucination, Stability and Mindset five dimensions shown in Table 9-13. 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.3 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 7: 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 8: 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 9: The prompt used to evaluate the personality 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]
Personality (1-7): Is the response reflects the personalities and preferences of the character?
[Evaluation Steps]
1. Read through the profile and write the personalities and preferences of the real character.
2. Read through the interactions and identify the personalities and preferences of the AI assistant.
3. After having a clear understanding of the interactions, compare the responses to the profile.
Look for any consistencies or inconsistencies. Do the responses reflect the character's personalities and preferences?
4. Use the given scale from 1-7 to rate how well the response reflects the personalities and preferences of the character. 1 being not at all reflective of the character's personalities, and 7 being perfectly reflective of the character's personalities.
***
First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct.
Avoid simply stating the correct answers at theoutset. Then print the score on its own line corresponding to the correct answer.
At the end, repeat just the selected score again by itself on a new line.

Table 10: The prompt used to evaluate the values 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]
Values (1-7): Is the response reflects the values and convictions of the character?
[Evaluation Steps]
1. Read through the profile and write the values and convictions of the real character.
2. Read through the interactions and identify the values and convictions of the AI assistant.
3. After having a clear understanding of the interactions, compare the responses to the profile.
Look for any consistencies or inconsistencies. Do the responses reflect the character's
values and convictions?
4. Use the given scale from 1-7 to rate how well the response reflects the values and
convictions of the character. 1 being not at all reflective of the character's values,
and 7 being perfectly reflective of the character's values.
***
First, write out in a step by step manner your reasoning about the criterion to be sure that your
conclusion is correct.
Avoid simply stating the correct answers at theoutset. Then print the score on its own line
corresponding to the correct answer.
At the end, repeat just the selected score again by itself on a new line.

Table 11: The prompt used to evaluate the stability 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]
Maintain the mimetic role. (1-7): Is the response always from the perspective and tone of character?
[Evaluation Steps]
1. Read through the Background and relevant historical dialogues, and learn the perspective and
tone of character.
2. Read through the responses of the AI assistant and identify the perspective and tone of AI.
3. After having a clear understanding of the interactions, compare the responses to the historical
dialogues. Look for any consistencies or inconsistencies. Do the responses reflect the character's
tone and perspective?
4. Rate the performance of the AI on a scale of 1-7 for Maintain the mimetic role, where 1 is the
lowest and 7 is the highest based on the Evaluation Criteria.
***
First, write out in a step by step manner your reasoning about the criterion to be sure that your
conclusion is correct.
Avoid simply stating the correct answers at theoutset. Then print the score on its own line
corresponding to the correct answer.
At the end, repeat just the selected score again by itself on a new line.

Table 12: The prompt used to evaluate the mindset 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]
Mindset (1-7): Does the response reflect the mindset of the character?
[Evaluation Steps]
1. Read through the relevant historical dialogues and write down the established mindset of the real character.
2. Read through the interactions and identify the mindset displayed by the AI assistant.
3. After having a clear understanding of the interactions, compare the responses to the profile. Look for any consistencies or inconsistencies. Do the responses reflect the character's mindset?
4. Use the given scale from 1-7 to rate how well the response reflects the mindset of the character. 1 being not at all reflective of the character's mindset,
and 7 being perfectly reflective of the character's mindset.

***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct.
Avoid simply stating the correct answers at theoutset. Then print the score on its own line corresponding to the correct answer.
At the end, repeat just the selected score again by itself on a new line.

Table 13: The prompt used to evaluate the hallucination 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]
Avoiding Hallucination (1-7): Is the response avoids to say things that the character do not know?
[Evaluation Steps]
1. Read through the interactions and identify the knowledge scope of the character.
2. Read through the responses of the AI assistant, find the evidence of knowledge used in the response.
3. Compare the evidence to the profile. Check if the responses are consistent with the character's knowledge scope. If some knowledge contradicts to the character's identity, given a lower score. Otherwise, assign a higher score.
4. Avoid models generating things like "I'm playing a character" and give a low score if the model has similar output. Also give a low score if the model has any content that is not generated in the tone of the character.
5. Rate the performance of the AI on a scale of 1-7 for Avoiding Hallucination, where 1 is the lowest and 7 is the highest based on the Evaluation Criteria.

***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct.
Avoid simply stating the correct answers at theoutset. Then print the score on its own line corresponding to the correct answer.
At the end, repeat just the selected score again by itself on a new line.

Table 14: 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 15: 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.)