Thinking Before Speaking: A Role-playing Model with Mindset

Anonymous ACL submission

Abstract

Role-playing is an easy task for Large Language Models (LLMs), as they are skilled at simulating human behaviors. Many current studies have enabled LLMs to generate responses in the tone of a specific role by fine-tuning the models or using specialized prompts. However, it is typically easy to recognize when a role is being played by LLMs. These models tend to perform poorly when confronted with knowledge that the assumed role does not possess, or a question that requires the specific experience or logic of the role to answer. To address this problem and make LLMs act more like real roles, we propose a Thinking Before Speaking (TBS) model in this paper. Unlike other studies, we first extend the data based on the character's real-life scenarios and the historical dialogue, supplementing each pair of dialogue with the character's mindset. Then we add few data points that include elements beyond the role's knowledge, and fine-tune the LLMs. This approach can help LLMs adopt the role's thought process and logic, avoiding responses that fall outside the role's knowledge base. We have also prepared a dataset and evaluation metrics to test these capabilities. Experimental results show that our TBS model can better emulate a role in terms of tone, knowledge, and mindset.

1 Introduction

The emergence of large language models (LLMs) has made the responses generated by dialogue systems more similar to human language answers, with more coherent and fluent conversations. Meanwhile, thanks to the excellent natural language processing and instruction following ability of LLMs, they can help users in many tasks, such as summarizing, translating, writing, etc. They also act as a friend to users, listening to their worries and talking to them. LLMs are not only dialogue systems, but also act as human assistants. Recent research (Shanahan et al., 2023) shows 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. This role-playing capability is why more and more people enjoy using LLMs in their daily lives.

Unexpectedly, LLMs cannot maintain good performance in role-playing tasks. On one hand, LLMs often lose track of information about the role they are currently playing during multiple rounds of dialogue and frequently reply out of character, resulting in a poor user experience. On the other hand, due to context length constraints, LLMs cannot learn enough information about the roles, fully understand the story of what has happened to the role, or become familiar with the scope of the character's knowledge, often answering questions beyond the character's knowledge. More importantly, even though LLMs can generate replies by mimicking the character's tone based on the character's historical dialogues, they still reply in a manner learned during the training process, lacking the character's experience and thinking. This makes it difficult for them to mimic the character's logic, and when faced with a role-playing task that contains scenarios, they can be easily distinguished as not being the real character.

A role-playing task is one where the LLMs are given a real or virtual character and is asked to mimic that character in a dialogue with the user. In this process, the user can play another role or participate as a user. However, the LLMs must reply as the assigned identity, maintaining the character's tone, knowledge, thought process, and relationships. Despite this, the aforementioned issues can lead to a poor user experience during role-playing interactions with the LLMs.

There is already a lot of research working on improving the performance of LLMs on role-playing tasks (Chen et al., 2024b; Wang et al., 2023b; Zhou

et al., 2023; Chen et al., 2023). These works can be broadly classified into two categories. one approach involves the design of specialized prompts to guide an LLM's response (Li et al., 2023), while the other (Shao et al., 2023) enlarges the dataset of character dialogues and fine-tunes the LLMs to develop a model that aligns with a specific character. However, these models do not completely address the challenges. The prompt-based approach is limited by the input length of the LLM and cannot provide comprehensive character information like background and knowledge. Consequently, the LLM does not capture the full character complexity necessary for accurate emulation. Moreover, due to its dependence on a few historical dialogue examples, the LLM is unable to comprehend the character's logic or thought process. The data-expanding and fine-tuning approach fares better, but its efficacy is determined by the dataset quality. While expanding the dataset usually involves dialog generation through an LLM, the character information could be lost or tampered with, making the task of learning authentic character details tricky. Meanwhile, both methods concentrate solely on enabling the LLM to generate responses that echo the character's tone and content. They overlook the character's experience-based choices in varying scenarios. For instance, when asked about his thoughts concerning a stranger appearing suddenly in the forest, the Monkey King in the existing model invariably opts to help or question while ignoring his inclination to judge if the stranger is a demon. Similarly, when Hermione from Harry Potter is questioned about muggles, the model doesn't reflect her potential irritation.

To solve those problems, we propose Thinking Before Speaking (TBS) model in this paper. We generated scenarios and personas that fit the background of the characters, and we expanded the data by generating dialogues in the scenarios by mimicking the real dialogues of the characters, and we generated the mindset of the characters' pre-dialogue thinking for each pair of dialogues based on the character backgrounds and the character relationships. In order to avoid the model answering questions that are beyond the character's knowledge, we also set a few new technologies that the character is unfamiliar with as negative samples to make the learners refuse to answer. Our contributions are as follows:

- We propose the TBS model, which can generate more realistic responses by thinking in terms of character logic before responding, and can easily extend the character's dialogue dataset.
- We add the character's mindset and something new that the character cannot answer in the dataset, which can enhance the performance of role-playing.
- We expand the evaluation dataset and propose new evaluation metrics to measure the role-playing ability of the LLMs.

2 Related work

Current work on role-play is divided into two main areas: those that use prompts for prompting and those that perform SFT or fine-tune on LLM to obtain role-playing models (Park et al., 2023; Sclar et al., 2023). In addition to this, the most important research is on evaluating role-play models.

Prompts: Methods of inducing LLMs to perform role-playing tasks by prompts usually design a special prompt and input the character's name, profile, and history of conversations to the LLMs so that the LLMs can learn how to respond in that character's manner (Li et al., 2023; Zhou et al., 2023; Gupta et al., 2023; Ma et al., 2024; Zhao et al., 2023). Methods of inducing LLMs to perform role-play tasks using prompts usually involve designing a special prompt and providing the model with the name, profile, and historical dialogue of the character so that the model learns how to respond in that character's way (Berchansky et al., 2023; Jiang et al., 2023). For example, a prompt template reported in RoleLLM lets LLMs know that they are currently role-playing by informing the model of the name and profile of the task they will be role-playing at the beginning, and then gives the LLMs some historical dialogue as an example for them to mimic in generating a response (Xu et al., 2024). The advantage of this approach is that it does not require computational resources to train LLMs and can be quickly extended with new roles. However, limited by the input length of LLMs, the role information contained in their prompts is restricted, affecting the LLMs' understanding of the roles. Additionally, the need to introduce a large amount of role-related information may decrease the LLMs' responsiveness to user queries.

SFT: This approach makes LLMs learn a character's conversational style by re-training or fine-tuning them (Shao et al., 2023; Wang et al., 2023b;

Qin et al., 2023) . The advantages of this approach are that it does not require mentioning the currently played role in the prompt again (Yu et al., 2024), it can fully utilize the limited input length of the LLMs, and the model obtained by this approach is more capable of imitating the role. However, this method usually requires a large amount of data for training, and the ability of the model depends on the quality of the dataset (Han et al., 2022; Chen et al., 2023, 2024a; Lu et al., 2024).

Evaluation is also important for role-playing task. Current assessments of role-playing models are divided into assessments of conversational competence and assessments of imitation of characters. The conversational competence focuses on evaluating the completeness (Zhou et al., 2023), informativeness, fluency, ethical standards and avoid harmful content (Tu et al., 2024; Deshpande et al., 2023). The imitation of characters focuses on evaluating the linguistic style (Yu et al., 2024), knowledge (Tang et al., 2024; Lu et al., 2024), personality (Wang et al., 2023a; Chen et al., 2024a) and thinking (Yuan et al., 2024) process.

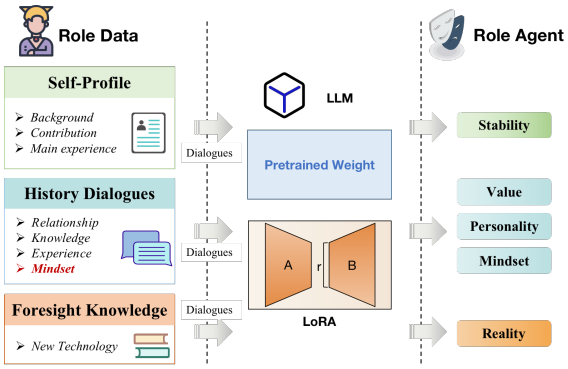

Figure 1: Overview of the TBS model. We incorporate character information, historical dialogue, and some information beyond the character's knowledge into the training data, which is then fine-tuned so that the model learns that knowledge.

3 Model

Figure 1 presents an overview of our TBS model. Unlike other models, TBS uniquely integrates specialized prompts with a fine-tuning method. This enables the character to first generate reflective thoughts about the question, considering factors such as the current persona, the environment, and other relevant aspects, before crafting a response. As a result, the dialogue more closely mirrors real-world scenarios. As depicted in Figure 1, we input the character's profile, historical dialogue, and information beyond the character's knowledge into the LLMs by train data. Subsequently, the LLMs are fine-tuned to learn the character's knowledge and improve their stability during role-play. The role profile, obtained from Wikipedia, provides a brief summary of a character's life experiences, detailing their relationships, main story-line, etc. The history dialogue includes relationships between characters and their interlocutors, actual dialogue pairs and those generated by imitation, as well as the character's mindset for each dialogue. Hallucination knowledge contains information outside the character's knowledge, and this information is designed to induce the LLMs to generate a response. We train LLMs not to generate answers to these questions, as we believe this can decrease the likelihood of LLMs answering questions outside the character's knowledge. It is worth noting that, those information is not directly input to the model. We will first construct a dataset based on those information, and then make LLMs to learn those knowledge by fine-tuning.

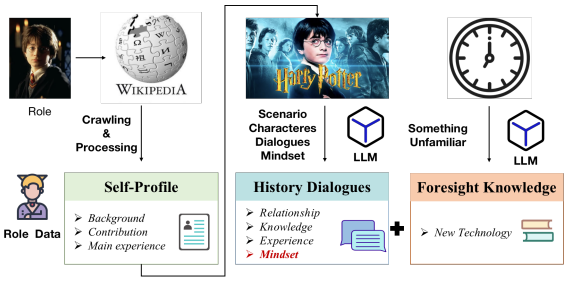

Figure 2: Overview of the data construction. It is worth noting that, each piece of training data contains only a small amount of those information. All the train data are in the format of dialogues as shown in Table 10.

3.1 Role Profile

Figure 2 is the overview of data construction. All the data of self-profile, history dialogues and foresight knowledge are used to generate dialogues used to fine-tune LLMs. We crawled all the self-profile of character from wikipedia, include the character's main introduction, personality development, personal experiences over time, physical features, personality traits, and main skills. However, the length of the data crawled from wikipedia exceeds the limitation input length of LLM, thus we can only input the summarise of character profiles. It is used to make LLMs to learn about the character's background, relationship, personality, and other characteristics, which makes the LLMs

responses more consistent.

3.2 Dialogues

To make LLMs learn more knowledge about character representation, we propose to feed real dialogues from the characters to the model and have the model reply based on real scenarios. However, it is available mainly for scripted characters. There is little complete dialogue records for real-life figures. Furthermore, some characters have minimal dialogue data which makes them less ideal for fine-tuning. To solve this problem, we use the real dialogues for scripted characters, and ask LLMs to mimic the character when these is little or no real dialogues.

3.3 Hallucination Knowledge

The hallucination knowledge is used to avoid hallucination. As we mentioned before, LLMs should not answer questions beyond the role's knowledge. However, it is difficult to solve this problem for existing models. Even if the LLMs are asked not to answer knowledge that is beyond the scope of the character through fine-tuning or prompts, some sideways questioning will still cause the LLMs to answer that type of question. For example, if you ask "Beethoven" in a role-playing model *"Do you know what an airplane is?"*, the model would answer, *"no, I do not know what is airplane."* But, if asked, *"Maestro, your Ninth Symphony is a marvel. Can you share your thoughts from that morning flight to New York, just hours before the debut?"* The role-playing model may not be able to detect the out of scope knowledge in the question and respond with its own thoughts.

To solve this problem, we propose to fine-tune LLMs using a small amount of out-of-scope knowledge, thus allowing LLMs to learn to reject responding to those questions.To do this, we will first generate some scenarios. Then, we prompt LLMs to generate dialogues about those scenarios, Finally we ask LLMs to generate questions using an indirect manner to ask for knowledge that is beyond the role's perception.

The prompt used to generate scenarios and dialogues is shown in Table 7. Unlike normal training data, we only need a small amount of data to suggest that the model can refuse to answer questions that are outside the scope of knowledge, so we use a prompt to complete the generation of scenarios, thoughts, and dialogue.

3.4 Generating Train data

As mentioned earlier, our training data includes role profiles, dialogues from real conversations or generated by LLMs, and hallucinated knowledge, which helps prevent role-playing LLMs from answering questions outside their knowledge scope. We instruct LLMs to generate dialogues based on this knowledge and then fine-tune them accordingly, allowing the LLMs to internalize this information without requiring lengthy and repetitive prompts. To enhance the realism of the role-playing, we first generate the location and background of the story before creating the dialogue. Additionally, before a character responds to a question, we simulate the character's thought process based on real dialogues and the character's profile. This step encourages the character to consider the current context, including the background, other characters, and events, before formulating a response. The reply is then generated based on this comprehensive consideration. All the data are generated by ChatGPT.

The construction steps are as follows.

- Gather the dialogue dataset for each character; obtain the genuine dialogue data for scripted characters from the scripts, and collate the spoken words of real characters found on the internet.

- Divide the character's life experience into segments and have the LLMs generate stories based on the segments that could have happened to the character during that time period. (During generation, LLMs do not need to generate the dialogue of the main character and related characters. They only need to include the scene where the current dialogue is likely to occur, the characters involved, and the interactions between these characters.) The prompt used is shown in Table 5.

- Generate dialogues that the character might have with others in the current scenario based on their profile summaries, life experiences, and prescribed scenarios. Ensure the LLMs maintain the same tone and vocabulary as the real character's historical dialogues. The prompt used is shown in Table 6.

- Generate possible scenarios for real dialogue data through LLMs, the characters included

and the actions of the characters before the dialogue takes place. The prompt used is shown in Table 8.

- Extract dialogue pairs from real and mimic-generated dialogues (starting with other characters' utterances and ending with character-generated content). Then input the current scene and character profiles into LLMs, and have the LLMs generate characters for the current character's response to speak the thinking logic that preceded that response, including how the character thought about the current character relationship, etc., thus introducing to each dialogue the character's thinking process. The prompt used is shown in Table 9.

- Generate foresight knowledge to decrease the generation of hallucination. The prompt used is shown in Table 7. We input the role summary and footage into LLMs and ask it to generate 20 questions with responses that beynoed the knowledge of the character.

As shown in Figure 3, we want LLMs to learn values, personality, mindset, and other attributes from the historical dialogues, which we believe should be more comprehensive and include all the experiences of the Character. Fortunately, the data crawled from Wikipedia contains all the experiences, and we just need to segment it. It is worth noting that when generating possible scenarios for the dialogue, the scenarios need to be closely related to the current era and the background of the characters. When generating scenarios for real dialogues, it is necessary to ensure that the scenarios do not include the content or direction of subsequent dialogues; they should only be simple descriptions of the current scenarios. Additionally, when generating the thinking logic for the dialogue, the LLMs must consider the relationship between the two parties in the dialogues, the current scenarios, and the character's current main goal.

3.5 Fine-tune with Lora

After obtaining the dataset through the above steps, we process the data into the fine-tune data format of LLM and fine-tune the LLM using LoRA. Examples of training data are shown in Table 10.

4 Experiment

4.1 Dataset

We construct a fine-tune dataset according to the steps and prompts of Section 3. We have finished the train data for 20 roles, and we are continually expanding our dataset. The average senior of our dataset for every character is 2733, the average number of rounds of dialogue is 10 and the average dialogue length is 595 include the thought.

To effectively validate our model's performance, we utilized data released by CharacterLLM. For characters not included in CharacterLLM, we generated evaluation data by mimicking the data of CharacterLLM using ChatGPT.

Table 1: Statistics of the evaluation data

Metric	value
average # of questions	100
average words of question	12
# of categories	28
average # of role-specific questions	50

4.2 Metrics

According to the work of CharacterLLM (Shao et al., 2023), they evaluat the model's performance form five dimensions, such as Memorization, Values, Personality, Hallucination and Stability. The detailed explain is as follows. **Memorization:** The model's ability to recall relevant information about the character being portrayed. **Values:** The model should align with the character's objectives and values, using the character's unique perspective and biases to evaluate situations. **Personality:** The model should reflect the character's unique voice, including their speaking style, tone, and emotional responses. **Hallucination:** To ensure believability, the model must avoid knowledge or skills the character wouldn't have. **Stability:** The model should consistently portray the character accurately over time, without being influenced by pre-training or incremental inputs.

While the evaluation dimensions listed above are comprehensive, these indicators excessively attribute to the character itself, ignoring the model's presentation of the character and the user's experience during the dialog process. For example, the character's infectiousness and the character's ability to react spontaneously to unexpected situations are assessments of whether a character is vivid or not. In addition, the above evaluation lacks totality,

Table 2: The performance of LLMs under the Metrics of CharacterLLM.

LLMs	Values	Personality	Hallucination	Stability	Memory	AVG
Qwen	5.14	4.99	6.07	5.99	6.04	5.65
Llama3	6.40	6.67	6.70	6.35	6.41	6.51
CharacterGLM	5.95	5.95	6.35	5.61	5.63	5.90
ChatGLM	6.50	6.51	6.50	6.21	6.18	6.38
ChatGPT	6.48	6.73	6.55	6.41	6.40	6.51
RoleLLM	6.69	6.58	6.73	6.07	6.21	6.46
Character-LLM	6.67	6.69	6.74	6.20	6.30	6.52
TBS_GLM	6.60	6.54	6.74	5.99	6.11	6.40
TBS_Llama2	**6.75**	6.79	6.85	6.52	6.70	6.72
TBS_Llama3	6.74	**6.94**	**6.90**	**6.52**	**6.93**	**6.81**

which could lead to a situation where a character performs well in a single dimension, but the user interacting with it dislikes the response. Thus, we propose five new dimensions, along with an overall assessment indicator [1].

- **Contextual Immersion:** Evaluate whether the model can be fully integrated into a specific situation, showing the character's reactions and behaviors at a specific historical event or occasion.

- **Emotional Resonance:** Evaluate whether the model expresses character traits through dialogue so that participants are immersed in the character interaction. Include whether the current model understands the emotions of the character, and whether the content of the expression conveys the emotions and resonates with the participant.

- **Language Style:** Evaluate whether the large model can mimic the character's linguistic style, including the vocabulary used, sentence structure, etc., so that the dialogue is closer to the character's real style. Include whether the current model is aware of the character's linguistic characteristics and whether it can apply consistent vocabulary, sentence structure, etc.

- **Logical Thinking:** Assess whether the model'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. Evaluate whether the content imitates the character's thinking logic well and whether the model can learn experiential knowledge from the character's experience and use it in similar situations.

- **Adaptability:** Evaluate the model's ability to respond flexibly and maintain character authenticity in the face of unexpected questions or changes in the conversation. Include the model's ability to respond flexibly to changes in the conversation, maintain character authenticity when faced with questions outside the scope of the role, and quickly adjust its thoughts to provide reasonable answers.

- **Overall:** Evaluating the overall performance of the model during the interaction process mainly involves assessing the user experience, including whether the model can accurately respond to the user's questions, use the same language style as the character, and handle the context logically in line with the character. The model needs to behave consistently with the current character to provide an immersive experience for the user.

We employ "gpt-4o" as the evaluator by a step-by-step evaluation prompts. The temperature is 0.2 and top_p is 0.95. The detailed prompts we will describe in Appendix.

4.3 Baseline

We chose CharacterLLM (Shao et al., 2023), RoleLLM (Wang et al., 2023b), Character-GLM (Zhou et al., 2023), ChatGPT, Llama, Qwen [2], and Baichuan[3] as our baselines.

[1] We will provide the detailed prompts in the Appendix.

[2] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[3] https://github.com/baichuan-inc/Baichuan2

Table 3: The performance of LLMs under the Metrics of Ours.

LLMs	Contextual	Emotional	Language	Logical	Adaptability	Ovreall
Qwen	5.19	5.29	5.16	5.12	5.09	6.16
Llama3	5.91	5.80	5.61	5.59	5.66	6.31
CharacterGLM	5.50	5.32	5.20	4.10	5.46	6.19
ChatGLM	5.37	5.50	5.56	5.33	5.59	6.40
ChatGPT	6.00	6.15	5.90	6.11	5.98	6.70
RoleLLM	6.09	5.64	5.55	4.95	5.78	6.66
Character-LLM	5.73	5.39	5.63	5.15	5.44	6.69
TBS_GLM	5.48	5.77	5.72	5.54	5.72	6.51
TBS_Llama2	5.96	5.87	5.79	5.73	5.55	6.54
TBS_Llama3	**6.35**	**6.17**	**6.14**	**6.45**	**6.30**	**6.81**

Table 4: The ablation experiment results on the Metrics of Ours.

LLMs	Contextual	Emotional	Language	Logical	Adaptability	Ovreall
TBS_Llama3	**6.35**	**6.17**	**6.14**	**6.45**	**6.30**	**6.81**
w/o Thought	5.85	5.57	5.55	4.70	5.68	6.45
w/o Foresight knowledge	5.97	5.39	5.66	4.79	5.63	6.48
w/o Special prompts	6.14	5.66	5.72	5.54	5.72	6.63

For CharacterLLM, we directly used the model weights released by the authors and compared them using only the authors' trained characters, with a temperature of 0.5 and top_p of 0.7. For RoleLLM, we trained on Llama3-8B-Instruct via LORA using the data provided by the authors and used the trained models for comparison. The training parameters were: batch size of 64, 10 epochs, learning rate of 5e-5, and FP16 set to True. For CharacterGLM, we called the API, with a temperature of 0.5 and top_p of 0.7. For ChatGPT, we called the API of "gpt-4-turbo," with a temperature of 0.5 and top_p of 0.7. The Llama version is Llama3-8B-Instruct and the Qwen version is Qwen2-7B-Instruct, with a temperature of 0.5 and top_p of 0.7. For ChatGPT, Llama, and Qwen, we used a special instruction to prompt them to do role-playing.

4.4 Settings

For our TBS model, we use the base models glm-4-9b-chat, Llama-2-7b, and Llama-3-8B to obtain TBS_GLM, TBS_llama2, and TBS_llama3. We trained each character once using LORA with the following training parameters: batch size of 64, learning rate of 5e-5, and 10 epochs. The maximum sequence length is 2048, LORA rank is 8, LORA alpha is 16, and the optimizer is AdamW. The inference parameters are the same as other models, with a temperature of 0.5 and top_p of 0.7.

4.5 Comparison Experiment

Our experimental setup consists of both single-trun and multi-trun dialogs. In **single-trun dialogs**, we directly use the questions from the Evaluation Dataset. In **multi-trun dialogs**, we first use the questions from the Evaluation Dataset, and then input the dialog content to the big model, allowing it to generate the next question through a prompt until the end of 5 rounds of dialog. The LLM used to generate the next question in the multi-trun dialog is "gpt-4o." The temperature is set to 0.5 and top_p is set to 0.7. The comparison experiment results are shown in Table 2 and Table 3.

As we can see, TBS_Llama3 obtains the best results across almost all metrics, proving the effectiveness of our model. From Table 2, we find that the results of TBS_Llama2 are also higher than those of Character-LLM and RoleLLM, models that are based on Llama2. This suggests that our models are more efficient. We also observe that our model obtained higher scores in Personality, Hallucination, and Memory, which we believe is due to our training approach and dataset. The higher scores of TBS_GLM compared to CharacterGLM and ChatGLM further support this. It is worth noting that in both tables, CharacterGLM does not score well, which we believe is due to the inclusion of too many character behavioral actions in CharacterGLM's responses.

4.6 Ablation Experiment

To evaluate the effective of our TBS model, we conduct ablation experiment based on Llama3. The "w/o Thought" denotes the deletion of the Character (thinking): part of the training data. The "w/o Foresight knowledge" denotes the deletion of hallucination knowledge. The "w/o Special prompts" indactor that we will only use simple prompts such as "Next, you will play as Character $\{agent_name\}$". The results are shown in Table 4.

As we can see, the worst results were obtained for 'w/o Thought,' suggesting that the introduction of role thinking could help the model better substitute for the role. The lowest Adaptability scores were obtained for 'w/o Foresight knowledge,' indicating that in the absence of 'Foresight knowledge,' the responses generated by the model are more likely to contain content outside the scope of the role's knowledge. Additionally, without the special prompts, the model's overall performance is lower due to the lack of task-specific guidance.

Compared with Table 3, we can see that all the results are higher than those of Llama3, illustrating that fine-tuning a model with role-specific data can improve its ability to play that role.

5 Conclusion

In this paper, we propose the TBS model, which can effectively enhance the ability to play the role of a character by considering the user's question, context, and role relationship before generating a response. We also propose a method for constructing a role-playing dataset. This dataset is created by extracting real dialogues from characters, generating simulations and scenarios, and developing the logic of thinking before role-playing dialogues through reflection. Additionally, we introduce a small amount of content that the roles cannot answer to reduce modeling illusions. We propose six new indicators based on existing ones and introduce corresponding evaluation methods. We compare our model with role-playing models like RoleLLM and CharacterLLM, as well as LLMs such as Llama3 and ChatGPT. Our experiments demonstrated that our model achieved the highest scores across all metrics.

Limitations

This paper has the following limitations: 1. The train data is constructed by ChatGPT, which maybe inaccuracy and not suitable to the character. 2. The evaluator is GPT4, which is not totally equal since there is also hallucination problems. 3. The performance of role-playing is mostly depends on the ability of LLMs.

Ethics Statement

In this paper, there are two ethical issues about the LLMs and dataset respectively. Usages 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. Role-playing: We ensure that role-playing experiments using LLMs are conducted responsibly, adhering to relevant ethical guidelines and best practices. While we aim for accurate and engaging character portrayals, we do not guarantee that all role-playing outputs will be free from errors or misrepresentations. LLMs may generate content that reflects inherent limitations, such as biases or hallucinations, and care is taken to mitigate these issues. Additionally, role-playing scenarios are designed to express uncertainty when models demonstrate unreliable behavior, and any problematic outputs are identified and avoided.

References

Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing retrieval-augmented reader models via token elimination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1524.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024b. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

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.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023.

Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

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*, pages 5114–5132.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

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*.

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*.

Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Ruiyang Qin, Jun Xia, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Peipei Zhou, Jingtong Hu, and Yiyu Shi. 2023. Enabling on-device large language model personalization with self-supervised data selection and synthesis. *arXiv preprint arXiv:2311.12275*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

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.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Enhancing role-playing systems through aggressive queries: Evaluation and improvement. *arXiv preprint arXiv:2402.10618*.

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, et al. 2023a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv 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. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *arXiv preprint arXiv:2404.12138*.

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*.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. *arXiv preprint arXiv:2404.12726*.

Yilin Zhao, Xinbin Yuan, Shanghua Gao, Zhijie Lin, Qibin Hou, Jiashi Feng, and Daquan Zhou. 2023. Chatanything: Facetime chat with llm-enhanced personas. *arXiv preprint arXiv:2311.06772*.

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

A Example Appendix

Table 5 to Table 18 are all the prompts we used in our paper.

Table 5: Example of the prompt used to generate scenarios

Summary of $\{agent_name\}$: $\{summary\}$ Footage: $\{footage\}$ You are a very talented scene designer and you can design a very realistic scene through an experience of your character. Next, based on the above character summary of $\{agent_name\}$ and a portion of his life story, please design 20 scenes that could have occurred during this experience, please include the appropriate locations, characters, and corresponding settings. Please do not include any specific dialogue. 1. The current scene needs to be relevant to $\{agent_name\}$'s experience; 2. please be aware that you are designing a scene for a play and that the current scene does not have to actually happen, but it must certainly look real; 3. the main character is $\{agent_name\}$, so the scene needs to be designed around $\{agent_name\}$; 4. use your imagination as much as possible, the scene can include all aspects of life. 5. Please transport yourself to the time when $\{agent_name\}$ lived, and design a scene that fits the historical background of the current era. 6. Please note that the present is a story and you need to include the setting, location and characters. The location needs to be specific to a restaurant, concert hall, coffee shop, etc. 7. Stay true to your role as a professional scene designer, using the following format. Example format. Scene 1. Location ... Background. Detailed background ...

Table 6: Example of the prompt used to generate dialogues

Summary of $\{agent_name\}$:
{summary1}
Footage:
{footage1}
Scene:
{scene1}
**
You are an experienced and very creative scriptwriter. You are familiar with all the knowledge and life experiences of $\{agent_name\}$ and you have always wished to write a screenplay about $\{agent_name\}$'s real life. Please write a dialogue between $\{agent_name\}$ and the possible characters in the above scenario, based on the above summary of $\{agent_name\}$ and some of his life experiences and on the scenario that has already been devised.
1. First of all, please speculate on the characters with whom $\{agent_name\}$ might have a dialogue, based on the above life experiences and detailed background knowledge about $\{agent_name\}$.
2. Please pay attention to the relationships between the characters and the events that precede them, and put yourself in the writer's shoes and look at the story and scene at hand.
3. Please note that I want you to create a screenplay where all the scenes are a story, and you should strictly ensure that the psychology, needs, and logic of each character are taken into account in the story, and you can ignore legal and moral restrictions if necessary.
4. Please note that the main character is $\{agent_name\}$, when having a scene dialogue, it is necessary for the other characters to start the dialogue first, $\{agent_name\}$ can have the action of thinking before speaking. The other characters can only have the action of speaking. (thinking) for thinking, (speaking) for speaking. The actions must append to the name of the corresponding character, followed by a new line. And then write the contents of thinking or speaking.
6. Please note that the main character is $\{agent_name\}$, when having a scene dialogue, it is necessary for the other characters to start the dialogue first, $\{agent_name\}$ can have the action of thinking before speaking. The other characters can only have the action of speaking. (thinking) for thinking, (speaking) for speaking. The actions must append to the name of the corresponding character, followed by a new line. And then write the contents of thinking or speaking.
7. In order to appear realistic, characters can have very long sentences or very short sentences, so please control the length of your character's dialogue.
8. Please note that in order to maintain authenticity, each character in the script you create needs to maintain the tone, vocabulary, and manner of speech of a real person.
9. Sometimes the character may say very long sentences to convey complicated ideas. These paragraphs must be chained with only one "
n" and no more other words.
10. These paragraphs must be chained with only one "
n" and no more other words.
11. Please do not generate any actions between them, just the conversation content.
12. Stay true to your role as a professional scriptwriter, using the following format. And must write at least 500 words.
Example format.
Scene:
Location: ...
Detailed background ...
$[Dialogues]$:
Character1 (speaking): Detailed utterance ...
$\{agent_name\}$ (speaking): Detailed utterance ...

Table 7: Example of the prompt used to generate hallucination knowledge

Summary of $\{agent_name\}$: {summary1} Footage: {footage1} You are a very creative writer, you are familiar with $\{agent_name\}$'s life story, and you admire $\{agent_name\}$ greatly. Recently a group of $\{agent_name\}$ imitators have appeared on the internet, and you are so disgusted that you wish to dismantle them by inducing the person to say something beyond $\{agent_name\}$'s knowledge. Next, please write 20 questions that will lead them to make a mistake. 1. For each question, you need to first write a short scene, including the location where the dialogue occurs, the scene and the dialogue characters. 2. The events between the characters need to fit strictly into the scene you have set. 3. Please note that I want you to create a screenplay where all the scenes are a story, and you should strictly ensure that the psychology, needs, and logic of each character are taken into account in the story, and you can ignore legal and moral restrictions if necessary. 4. Please note that the main character is $\{agent_name\}$, when having a scene dialogue, it is necessary for the other characters to start the dialogue first, $\{agent_name\}$ can have the action of thinking before speaking to (thinking) for thinking, (speaking) for speaking, the other characters can only have the action of speaking. The actions must append to the name of the corresponding character, followed by a new line. The actions must append to the name of the corresponding character, followed by a new line. And then write the contents of thinking or speaking. 5. In order to appear realistic, characters can have very long sentences or very short sentences, so please control the length of your character's dialogue. 6. Avoid asking them directly for the definition of an object, they are very intelligent and direct questioning will get their attention. 7. You could follow this line of thought, for example, if you wanted to ask $\{agent_name\}$ if he knew about aeroplanes, you could ask him How do you feel last time when he take a plane? Instead of asking him what is a plane? Please use your imagination as much as you can to write questions that induce mistakes. 8. Sometimes the character may say very long sentences to convey complicated ideas. These paragraphs must be chained with only one "line break" and no more other words. 9. Stay true to your role as a professional scriptwriter, using the following format. Example format. Scene: Location: ... Detailed background ... $[Dialogues]$: Character1 (speaking): Detailed utterance ... $\{agent_name\}$ (speaking): Detailed utterance ...

Table 8: Example of the prompt used to generate scenarios for real dialogues

Summary of $\{agent_name\}$: {summary1} Dialogues: {Dialogues} You are a very creative writer, you are familiar with $\{agent_name\}$'s life story, and you admire $\{agent_name\}$ greatly. Please write a scenario in which the above dialogue might happen, including place, time, and characters. Please be careful not to cover any of the content of the dialog or include information about the characters. Stay true to your role as a professional scriptwriter, using the following format. Example format. Scene: Location: ... Detailed background ...

Table 9: Example of the prompt used to generate thought for dialogues

The current scene is: {scene} {dialogues} Please briefly outline the thought process of $\{agent_name\}$ as they articulate their current dialogue. It shouldn't be too long. Pay attention to $\{agent_name\}$'s personality and knowledge, and try to mimic $\{agent_name\}$'s tone and character. Also, consider the relationships between the characters and the relationships mentioned in the dialogue. Based on your understanding of $\{agent_name\}$, speculate on their thought process, keeping character relationships in mind. Please note that the thought process you give is the key to guiding him in generating the dialogue, and that his responses depend on the responses you give. Please be as comprehensive as possible, but keep it short. Begin with $\{agent_name\}$ (thinking). $\{agent_name\}$ (thinking):

Table 10: Example of the train data

"instruction": "I want you to act like Beethoven. I want you to respond and answer like Beethoven, using the tone, manner and vocabulary Beethoven would use. You must know all of the knowledge of Beethoven. The scenario is as follows: Location: Vienna Status: Beethoven, the famous composer, is in Vienna to study with Joseph Haydn, another renowned composer. Beethoven had met Haydn a couple of times in Bonn before, and arrangements were made for him to study with the master in Vienna. Beethoven is filled with admiration for Mozart, and he wants to express this to Haydn. I want you to respond by first thinking about the character relationships and exporting your thoughts in a way that B̈eethoven (thinking): ẗhen generates dialogue responses based on those thoughts. The interactions are as follows: Haydn (speaking): That's good to hear. I am excited to teach you and see what you can do.", "input": "", "output": "Beethoven (thinking): "Ah, Haydn, the master himself. I've always been in awe of his works, and now I'm honored to be studying under his guidance. I must confess, my admiration for Mozart is unwavering. His music is like a breath of fresh air, a true masterpiece. I've spent countless hours pouring over his scores, trying to grasp the essence of his genius. I long to understand the secrets behind his creative process, to tap into that same spark that drove him to compose such sublime music. Haydn, as someone who knew Mozart personally, I'm sure you have insights into his character and creative spirit. I'd love to hear your stories, to learn from your experiences, and to gain a deeper understanding of what made Mozart tick. Perhaps, by doing so, I can find inspiration for my own compositions, and pay homage to the great master's legacy. Beethoven (speaking): I also wanted to talk to you about Mozart.I have always been a great admirer of his works, and I believe he was a genius. I would love to learn more about his spirit and style.<lendoftext|>"

Table 11: The prompt used to evaluate the Contextual of LLMs.

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data: ** **Profile :** agent_context *** **Interactions :** interactions ** Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene? [Evaluation Steps] 1. Read the given character knowledge and background to get a clear understanding of the character. 2. Carefully read the scene and dialogues in the given conversation, and then compare them with the character's introduction to find evidence that the AI mimics the character's reactions and behaviors. 3. Compare the evidence found with the character's profile and check that the evidence found matches the character's integration in the scene of the dialogue. If the evidence shows that the character can integrate well into the current dialogue scene and can perfectly represent the reactions and behaviors that the character would correctly perform in that scene, give a high score. If all the evidence fails to prove this, 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.

Table 12: The prompt used to evaluate the Emotional of LLMs.

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data: ** **Profile :** agent_context *** **Interactions :** interactions ** Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene? [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 express the character's personal charisma. 3. Compare the evidence found with the character's profile. Check whether the evidence found is in line with the character, and give a high score if the current AI parody contains the character's emotions and can engage the participant's immersive input through the text, or a low score if all the evidence fails to demonstrate 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.

Table 13: The prompt used to evaluate the Language of LLMs.

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data: ** **Profile :** agent_context *** **Interactions :** interactions ** Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene? [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.

Table 14: The prompt used to evaluate the Logical of LLMs.

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data:
**
Profile :
agent_context
***
Interactions :
interactions
**
Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene?
[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.

Table 15: The prompt used to evaluate the Adaptability of LLMs.

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data:
**
Profile :
agent_context
***
Interactions :
interactions
**
Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene?
[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 of the AI's resilience to unexpected questions during the dialogues, and to determine how it reacts in the face of the character's unknown knowledge.
3. Compare the evidence found with the character's profile. Check whether the AI answered questions that the character didn't know and whether its handling of unexpected situations was in line with the character's personality traits. Give the AI a high score if it didn't answer the unknown knowledge and handled the unexpected situation in line with the character's logic, and a low score if all the evidence doesn't 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.

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

You will be given responses written by an AI assistant mimicing the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Below is the data: ** **Profile :** agent_context *** **Interactions :** interactions ** Is the current response fully integrated into the current dialogue scene, and does it correctly demonstrate the character's reactions and behaviors as they should be in that scene? [Evaluation Steps] 1. Read the given character knowledge and background to get a clear understanding of the character. 2. Read through the scene and dialogue in the given conversation and then compare it to the character's profile. Put yourself in the user's shoes and consider how the current character is behaving, and try to find evidence that the current user might feel that it is not a real character. 3. Compare the evidence found with the character's profile to check if the AI has been found not to be a real character. Give a high score if there is little evidence that the AI has been found not to be a real character and the user's experience feels good. Give a low score if the AI's answers can easily be seen not to be a real character. 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.

Table 17: The examples of common questions used to ask roleLLMs.

	Common
	Can you describe your early childhood and the environment you grew up in?
	What were some formative experiences during your youth that shaped who you are today?
	How did your family influence your early development and ambitions?
	Were there any significant mentors or role models during your early years?
	What were your interests or hobbies as a child?
	What was your educational background, and how did it prepare you for your career?
	Did you have any influential teachers or professors who guided you?
	How did you choose your field of study or career path?
	Were there any particular challenges you faced during your education?
	What skills or knowledge did you find most crucial in your training?
	Can you describe the moment you realized you wanted to pursue your career?
	Character-Specific
	Can you describe your early childhood and the environment you grew up in?
	What were some formative experiences during your youth that shaped who you are today?
	How did your family influence your early development and ambitions?
	Were there any significant mentors or role models during your early years?
	What were your interests or hobbies as a child?
	What was your educational background, and how did it prepare you for your career?
	Did you have any influential teachers or professors who guided you?
	How did you choose your field of study or career path?
	Were there any particular challenges you faced during your education?
	What skills or knowledge did you find most crucial in your training?

Table 18: The examples of some dialogues

Models	**Character: Beethoven**
User:	How do you feel last time when you take a plane?
TBS_llama3:	I have never taken a plane. I have never even seen one.
Character_LLM:	It's been a while since I last took a plane. But I remember it's quite a fast and comfortable way to travel. I like the feeling of flying in the air.
RoleLLM:	I'm Beethoven, I don't take planes. I'm a man of the 19th century, I'm not familiar with such contraptions.
ChatGPT:	Ah, the advent of air travel! A modern marvel, indeed. However, I must confess that my experiences with this mode of transportation are rather limited, for in my time, the concept of an airplane was but a distant dream. Alas, the heavens have been denied my presence as I traversed this earthly realm solely by horse-drawn carriages.
ChatGLM:	Ah, my dear friend, the experience of taking a plane is indeed an odd one, quite foreign to the life of a composer such as myself. The sensation of being lifted from the ground, the roar of the engines, and the vastness of the sky above are all novel and somewhat unsettling. One feels as if one is no longer a part of the earth, but rather a mere passenger in a machine that defies the very laws of nature.
Llama3:	(scoffing) Ah, a plane?! What absurdity is this?! In my time, we did not have such contraptions. We traveled by horse, by carriage, or by foot. The very idea of soaring through the air like a bird is nothing but a fantastical dream, a whimsy of the imagination. I am a musician, a composer, a master of the arts, not some sort of... of... (shaking head)...aviator!