Thinking Before Speaking: A Role-playing Model with Mindset

Anonymous ACL submission

Abstract

Role-playing involves making Large Language Models (LLMs) generate responses in the manner of a specific character. This task is relatively easy for LLMs, as they excel at simulating human behaviors. Existing works mainly focus on ensuring the consistency of a character's personality, information, and tone by fine-tuning the models or using specialized prompts. However, these models often lack the ability to fully embody the mindset of the character, making it difficult for them to generate responses that align with the character's way of thinking. This limitation leads to a poor user experience. To solve this problem, we propose a Thinking Before Speaking (TBS) model in this paper, which can mimic the character's logical reasoning process and generate reflections before answering a question. We enhance the training data for each set of dialogues by incorporating logical reasoning based on the character's profile and the contextual content of the conversations. This approach enables the model to learn and replicate the character's mindset. Additionally, we include a small number of questions beyond the character's knowledge scope to train the model on how to appropriately decline to answer. To verify the effectiveness of our model, we prepare new evaluation datasets and metrics. Experimental results show that the TBS model achieves best role-playing performance in terms of tone, information, and mindset.

1 Introduction

The emergence of large language models (LLMs) has significantly improved the quality of dialogue systems, making their responses more human-like, with coherent and fluent conversations. Thanks to their exceptional natural language processing and instruction-following capabilities, LLMs can assist users with a wide range of tasks, including summarizing, translating, writing, and more. They also serve as companions, offering a listening ear and engaging in meaningful conversations. Beyond being dialogue systems, LLMs function as versatile human assistants. Recent research (Shanahan et al., 2023) highlights that interactions with LLMs are inherently a form of role-playing, where the models strive to embody the character of a dialogue agent as described in the prompt. This role-playing ability is a key reason why LLMs have become an integral part of many people's daily lives.

However, LLMs can not maintain such a good performance in role-playing task. The role-playing task involves assigning a model a specific character to portray and requiring it to respond as that character. The model's responses must align with the character's tone, word choice, knowledge, and mindset. The poor performance of LLMs in role-playing tasks can be attributed to two main factors. First, they often lose track of the assigned role during multi-turn dialogues, resulting in out-of-character responses and a diminished user experience. Second, due to limitations in input context length, LLMs struggle to gain a deep understanding of the character, making it difficult to recognize the character's knowledge boundaries. As a result, they frequently provide responses that exceed the character's scope of knowledge, or generate response in the role of AI assistant.

There are many models aimed at improving the performance of LLMs in role-playing tasks (Chen et al., 2024b; Wang et al., 2023; Zhou et al., 2023; Chen et al., 2023). These efforts can broadly be classified into two categories: using prompts (Li et al., 2023) and fine-tuning with role-specific datasets (Shao et al., 2023). However, both methods face certain challenges. The prompt-based approach is limited by the input length of the LLM and cannot provide comprehensive character information, such as background and knowledge. Consequently, the LLM fails to capture the full complexity of the character necessary for accurate emulation. Moreover, because it relies on only a few historical dialogue examples, the LLM is unable to comprehend the character's logic or thought process. The fine-tuning method performs better, but its effectiveness depends on the quality of the dataset. While expanding the dataset typically involves dialogue generation through an LLM, character information can be lost or altered, making it difficult to learn authentic character details.

Meanwhile, both methods focus solely on enabling the LLM to generate responses that reflect the character's tone, personal information or psychological. They overlook the character's experience-based choices in various scenarios. For instance, when asked about his thoughts on a stranger suddenly appearing in the forest, the Monkey King in the existing model invariably chooses to help or question the stranger, while ignoring his inclination to judge whether the stranger is a demon. Similarly, when Hermione from Harry Potter is questioned about muggles, the model does not reflect her potential irritation.

To solve those problems, we propose Thinking Before Speaking (TBS) model in this paper, which can first formulate its thoughts about the question following the character's logical framework before providing a response. Then, generate a answer based on those thoughts. To enable the model to possess this capability, we developed an automated dataset construction method and generated the training data from Wikipedia sources. Furthermore, we confirmed the effectiveness of TBS through comparative experiments and ablation studies in accordance with existing metrics. Our contributions are as follows:

- We propose the TBS model, a new role-playing model, which can mimic the character's mindset, and achieve the best performance in most dimensions.
- We propose a new method for automated role-playing dataset construction, which can enrich dialogue data by incorporating dialogue scenarios and character mindset.
- We expand the evaluation dataset and propose more comprehensive evaluation dimensions based on existing metrics.

2 Related work

Current work on role-playing can be divided into two main areas: approaches that use prompts to induce role-playing and those that leverage supervised fine-tuning (SFT) or fine-tune LLMs to create role-playing models (Park et al., 2023; Sclar et al., 2023). Additionally, one of the most critical areas of research focuses on evaluating role-playing models.

Prompts: Methods for inducing LLMs to perform role-playing tasks using prompts typically involve designing a specific prompt and providing the character's name, profile, and conversation history to the LLM. This allows the LLM to learn how to respond in a manner consistent with that character (Li et al., 2023; Zhou et al., 2023; Gupta et al., 2023; Ma et al., 2024; Zhao et al., 2023; Xu et al., 2024). The advantage of this approach is that it does not require additional computational resources for training and can be quickly adapted to new roles. However, it is constrained by the input length of the LLM, which limits the amount of role information that can be included in the prompt. This restriction can impair the LLM's understanding of the character. Additionally, introducing extensive role-related information might reduce the LLM's responsiveness to user queries.

SFT: This approach trains LLMs to learn a character's conversational style by re-training or fine-tuning them (Shao et al., 2023; Wang et al., 2023; Qin et al., 2023; Lu et al., 2024a). The benefits of this method include not needing to repeatedly mention the current role in the prompts (Yu et al., 2024), better utilization of the LLM's limited input length, and producing models that are more proficient at imitating specific roles. However, this approach generally requires a substantial amount of training data, and the quality of the resulting model heavily depends on the dataset (Han et al., 2022; Chen et al., 2023, 2024a; Lu et al., 2024b). As we can see, the current role-playing models primarily focus on improving the accuracy of role-related information and better simulating a character's tone and style. However, hey often overlook the deeper essence of a character, including their thought processes and behavioral patterns. This oversight frequently leads to suboptimal performance in authentically portraying the character.

Evaluation is also a critical aspect of role-playing tasks. Current assessments of role-playing models can be categorized into evaluations of conversational competence and evaluations of character imitation. Conversational competence assessments focus on evaluating completeness (Zhou et al., 2023), informativeness, fluency, ethical standards, and the avoidance of harmful content (Tu et al., 2024; Deshpande et al., 2023). In contrast, character imitation assessments evaluate linguistic style (Yu et al., 2024), knowledge (Tang et al., 2024; Lu et al., 2024b), personality (Wang et al., 2024; Chen et al., 2024a), and thought processes (Yuan et al., 2024). But those metrics lack a focus on assessing the reasoning and overall performance of large models, which can result in high scores that do not align with a poor user experience.

3 Model

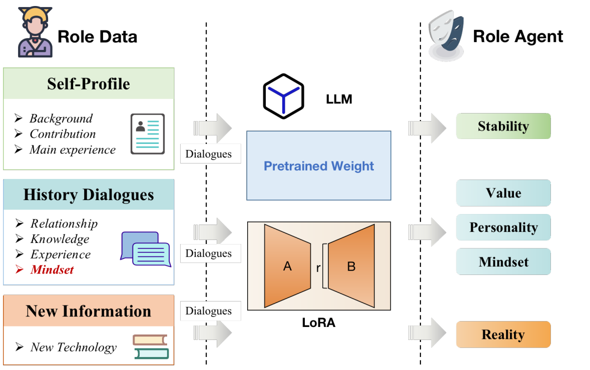

Figure 1: Overview of the TBS model. We incorporate character information, historical dialogue, and some new information that beyond the character's knowledge into the training data. Then, we fine-tune LLMs to learn this knowledge.

Figure 1 presents an overview of our TBS model. Unlike other models, TBS uniquely integrates specialized prompts with a fine-tuning method. This approach enables the LLMs to first generate reflective thoughts about the question, following the character's mindset and considering factors such as the current persona, environment, and other relevant aspects, before crafting a response. As a result, the dialogue more closely mirrors real-world scenarios. As show in Figure 1, our train data include the character's profile, historical dialogue, and new information beyond the character's knowledge. The role profile, obtained from Wikipedia, provides a brief summary of a character's life experiences, detailing their relationships, main storylines, and more. 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. The new information contains something outside the character's knowledge. This information is designed to prompt the LLMs to refuse to answer certain questions that go beyond the character's background. Finally, we fine-tune the LLMs to learn this knowledge.

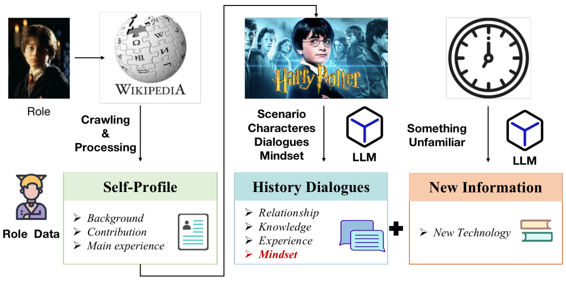

Figure 2: Overview of the data construction. It is worth noting that, each piece of training data contains only a small amount of those information. All the train data are in the format of dialogues as shown in Table 13.

Figure 2 is the overview of train data construction. We first collect the character's profile from the Wikipedia, then we generate dialogues with scenario and mindset by LLMs based on the real scripts and quotes. Finally we added some new information into the train dialogues.

3.1 Profile Collection

We crawled all the self-profiles of characters from Wikipedia, including their main introductions, personality developments, personal experiences over time, physical features, personality traits, and key skills. However, the length of the data extracted from Wikipedia exceeds the input length limitation of LLMs. As a result, we can only input summarized versions of the character profiles. These summaries are used to help LLMs learn about the characters' backgrounds, relationships, personalities, and other characteristics, ensuring that the LLMs' responses are more consistent.

3.2 Dialogues

To enhance an LLM's ability to represent characters, we propose feeding it authentic dialogues so it can respond naturally within realistic scenarios. However, extensive dialogue records mainly exist for scripted characters, while real-life figures rarely have sufficient transcribed conversations. This lack of comprehensive data makes fine-tuning on certain characters more challenging. To address this, we rely on real dialogues from scripted characters and guide the LLM to imitate the target character's style when little or no authentic dialogue is available.

3.3 New Information

The new information is intended to help prevent LLMs from answering questions that go beyond the character's knowledge. As we previously discussed, this is a challenging issue for existing models. Even with fine-tuning or prompt engineering to restrict the scope of answers, models can still be coaxed into revealing out-of-scope information through indirect questioning. For instance, if a role-playing model portraying Beethoven is asked, *"Do you know what an airplane is?"*, it might correctly respond, *"No, I do not know what an airplane is."* However, if posed with a more subtle query like, *"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 model might fail to recognize the out-of-scope nature of the question and respond as if it has knowledge of airplanes. To solve this problem, we propose fine-tuning LLMs with a small amount of new information to enable them to refuse answering such questions, thereby teaching the models to reject out-of-scope queries. The prompt used to generate new information dialogues is shown in Table 10.

3.4 Generating Train data

After completing the collection of all the data, we ask LLMs to generate the train data step by step using special prompts. To enhance the realism of the dialogue data, we include detailed scenario information for each dialogue set, such as the location, characters, and general context of the conversations. We also provide pre-dialogue thoughts for each character, tailored to align with the dialogue content. For characters with limited data, we prompt the LLMs to generate dialogue scenes and content based on their personal information, imitating the characters' traits. We also add reflections for these characters to further enrich the dialogues.

The construction steps are as follows.

- Collecting the dialogue dataset for each character. For scripted characters, extract their authentic lines directly from the scripts. For real-life individuals, use internet searches to collect verified quotes and statements they have made.

- Segmenting the character's life experiences into distinct periods, and prompt the LLMs to generate stories that could plausibly occur during each time period. During generation, the LLMs do not need to produce dialogue for the main character or related characters. Instead, they focus on describing the scenes where the dialogue might take place, the characters involved, and their interactions. The specific prompt used is detailed in Table 8.

- Generating dialogues that the character might engage in within the given scenario, drawing from their profile summaries, life experiences, and predefined contexts. Ensure that the LLMs replicate the tone and vocabulary consistent with the character's historical dialogues. The specific prompt used is provided in Table 9.

- Using LLMs to generate potential scenarios for real dialogue data, including the characters involved and their actions leading up to the dialogue. The specific prompt used is provided in Table 11.

- Extracting dialogue pairs from real and mimic-generated conversations, beginning with other characters' utterances and concluding with the character's responses. Then, input the current scene and character profiles into the LLMs to generate responses for the character, including the thought process behind their reply. This process should reflect how the character considers their relationships and other relevant context, thereby embedding the character's reasoning into each dialogue. The specific prompt used is detailed in Table 12.

- Generating new information to reduce hallucination generation. The prompt used is provided in Table 10. We input the role summary and footage into the LLMs and ask them to generate 20 questions, along with responses that exceed the character's knowledge.

As shown in Figure 3, we aim to train LLMs to learn the character's values, personality, mindset, and life experiences from the dialogue dataset. To achieve this, we segment the character's life experiences from Wikipedia and prompt the LLMs to generate a training dataset based on this information. It is important to note that when generating possible dialogue scenarios, these scenarios must align closely with the character's historical era and background. For scenarios accompanying real dialogues, they should only provide simple descriptions of the current context without including content or implications of subsequent dialogue. Furthermore, when generating the reasoning behind the dialogue, the LLMs must take into account the relationship between the participants, the current scenario, and the character's primary goal at that moment.

3.5 Fine-tune with Lora

After completing the steps above to obtain the dataset, we format the data into the fine-tuning structure required for LLMs and fine-tune the model using LoRA. Examples of the training data are provided in Table 13.

4 Experiment

4.1 Dataset

The train dataset is constructed according to the steps and prompts of Section 3. We have finished the train data for 20 roles. 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 validate the effectiveness of our dataset, we conduct a human evaluation on the train dataset. We randomly select 400 samples for review. We hire three human evaluators to perform the assessment, with detailed information provided in the Appendix. We randomly select 400 samples and ask the evaluators to assess the data from three perspectives: (1) whether the scenarios align with the character profiles, (2) whether the generated dialogues match the given scenarios, and (3) whether the mindset on the dialogues offer guidance for generating them while remaining consistent with the character's thought process. All human evaluators are asked to give a score range from 0 to 10. The results are shown in Table 1. Detailed evaluation are shown in Appendix.

Table 1: The human evaluation results of train dataset

Score (0-10)	Scenarios	Dialogues	Mindset
Avg	9.40	9.36	9.44
# over 7	96.25%	98.92%	99.25%
# over 8	92.75%	94.69%	95.17%

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. The statistics of the evaluation data are shown in Table 2, we manually verified the correctness of the evaluation data.

Table 2: 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), the model's performance is evaluated across five dimensions: **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 believably, 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, they focus excessively on the character itself, overlooking the model's presentation of the character and the user's experience during the dialogue process. For instance, aspects such as the character's ability to captivate the user and react spontaneously to unexpected situations are crucial for assessing whether a character feels vivid and lifelike. Moreover, the evaluation lacks a holistic perspective, which could result in a scenario where a character excels in a single dimension, yet the user finds its responses unsatisfactory. To address these issues, we propose five new dimensions, along with an overall assessment indicator [1].

Contextual Immersion: Evaluate whether the model integrates seamlessly into a specific situation, demonstrating the character's reactions and behaviors within a particular historical event or occasion.

Emotional Resonance: Evaluate whether the model expresses character traits through dialogue

[1]Detailed prompts will be provided in the Appendix.

Table 3: 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
Ditto	6.64	6.59	6.70	6.23	6.37	6.51
Incharacter	5.96	6.67	5.91	6.20	6.21	6.19
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**

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 model can emulate the character's linguistic style, including vocabulary, sentence structure, and other language features, to make the dialogue closely resemble the character's authentic style. Additionally, evaluate whether the model consistently applies these linguistic characteristics throughout the interaction.

Logical Thinking: Evaluate whether the model demonstrates clear and reasonable thinking during dialogue, aligns with the character's logical reasoning, and adapts its thought process to the character's perspective in different scenarios. Assess the extent to which the model imitates the character's thought patterns and applies experiential knowledge derived from the character's background to similar situations.

Adaptability: Evaluate the model's ability to respond flexibly while maintaining character authenticity in the face of unexpected questions or shifts in conversation. This includes assessing whether the model can adapt to changes, remain true to the character's role, and quickly adjust its reasoning to provide coherent and contextually appropriate answers.

Overall: Assess the model's overall interaction quality, focusing on user experience. This involves determining whether the model accurately responds to user questions, maintains the character's linguistic style, handles context logically, and remains consistent with the character throughout the dialogue to deliver an immersive experience.

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 Ditto (Lu et al., 2024a), Incharacter (Wang et al., 2024), CharacterLLM (Shao et al., 2023), RoleLLM (Wang et al., 2023), CharacterGLM (Zhou et al., 2023), ChatGPT, Llama, Qwen [2], and Baichuan[3] as our baselines.

For Ditto, we retrain a models follow the author's paper, as there is no available models. We use the dataset released by the author, with Qwen2.5-7B-Instruct as base model. The training parameters were: batch size of 64, 5 epochs, learning rate of 5e-5. For Incharacter, we prepared the data based on the information provided by the author and utilized their program to conduct role-playing experiments on the Llama3.1-8B-Instruct model. For CharacterLLM, we directly use the model weights released by the authors and compare them using only the authors' trained characters, with a temperature of 0.5 and top_p of 0.7. For RoleLLM, we train on Llama3-8B-Instruct via LoRA using the data provided by the authors and use 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

[2]https://huggingface.co/Qwen/Qwen2-7B-Instruct
[3]https://github.com/baichuan-inc/Baichuan2

Table 4: The performance of LLMs under the Metrics of Ours.

LLMs	Contextual	Emotional	Language	Logical	Adaptability	Overall
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
Ditto	6.05	5.67	5.17	5.91	5.94	6.34
Incharacter	6.08	5.96	5.48	5.20	5.09	6.20
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 5: The ablation experiment results on the Metrics of Ours.

LLMs	Contextual	Emotional	Language	Logical	Adaptability	Overall
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

"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, respectively. Each character is trained once using LoRA with the following 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 AdamW optimizer is used. For inference, the parameters are consistent with those of other models, with a temperature of 0.5 and a top_p value of 0.7.

4.5 Comparison Experiment

Our experimental setup consists of both single-turn and multi-turn dialogues. In **single-turn dialog**, we directly use the questions from the evaluation Dataset. In **multi-turn dialogues**, we first use the questions from the evaluation Dataset, and then input the dialog content to ChatGPT, 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-turn 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 3 and Table 4.

As we can see, TBS_Llama3 obtains the best results across almost all metrics, proving the effectiveness of our model. From Table 3, 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. 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. In Table 3, we can find that, although Ditto achieves the high score in the Contextual metric, its Language metric score is relatively low. This is because ditto prioritizes the accuracy of character information over imitating the character's style of speech. In terms of its Incharacter performance, it scores low in both values and hallucination, but achieves a high score in personality. This indicates that while it imitates the character's personality, it still does not possess the character's other essential qualities.

4.6 Human Evaluation

Table 6: Human evaluation results

LLMs	Cons	Know	Overall	Avg
Qwen	7.11	7.18	7.16	7.15
Llama3	7.22	7.13	7.27	7.21
CharacterGLM	7.15	7.12	7.23	7.17
ChatGLM	7.26	7.22	7.21	7.23
ChatGPT	7.05	7.06	7.20	7.10
RoleLLM	7.22	7.21	7.12	7.18
CharacterLLM	7.26	7.25	7.30	7.27
Incharacter	7.18	7.24	7.19	7.20
Ditto	7.23	7.42	7.21	7.29
TBS_GLM	7.64	7.76	7.62	7.67
TBS_Llama2	7.60	7.56	7.58	7.58
TBS_Llama3	**8.18**	**8.20**	**8.21**	**8.20**

To ensure the reliability of our experimental results, we conduct human evaluations, as detailed in the Appendix. We randomly select a sample of 100 dialogues, encompassing both single-turn and multi-turn interactions, for assessment. Three human evaluators evaluate these dialogues, assigning three distinct scores to each: **Consistency (Cons)**, which assesses how well the model's responses align with the character being portrayed; **Knowledge Accuracy (Know)**, which evaluates whether the responses stay within the character's knowledge boundaries; and **Overall Score (Overall)**, which reflects the experts' overall preference for the responses in terms of quality and engagement. The score range from 0 to 10, the results are shonw in Table 6.

As shown in Table 6, Apart from TBS, Ditto achieved the highest score in knowledge accuracy, likely due to the extensive use of Wikipedia data during training. CharacterLLM scored the highest in consistency and overall, possibly because it also utilized a large dataset for training. However, its performance in knowledge accuracy is lower than Ditto's, likely due to its weaker ability to avoid responses. TBS achieves the highest score in the Overall metric, consistent with the results obtained using large model-based evaluations, highlighting the effectiveness of our approach. Additionally, TBS also achieves high scores in Knowledge Accuracy and Consistency, further demonstrating the robustness and reliability of our method.

4.7 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" indicator that we will only use simple prompts such as "Next, you will play as Character $\{agent\_name\}$". The results are shown in Table 5.

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 4, 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 three 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. **Potential Applications:** This paper is to mimic a character's mindset during role-playing to achieve more effective character portrayal. However, if misused, this technology could potentially be employed to impersonate friends and deceive others. Currently, all the characters we have developed in our research are based on historical figures or fictional characters from films and television, and we do not engage in role-playing real, living individuals. We also encourage researchers to avoid portraying real, existing persons to minimize the potential harms that could arise from the misuse of this technology.

References

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.

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

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

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, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. *Association for Computational Linguistics*, pages 11836–11850.

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. *Association for Computational Linguistics*, pages 1840–1873.

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*.

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 Human Evaluation

To ensure the validity of our training data and evaluation results, we enlisted three human evaluators to assess both our data and experimental results.

A.1 Dataset Evaluation

During the construction of the training dataset, we utilize LLMs to generate dialogue scenarios, dialogue content, and character mindsets. Thus, the humans are required to evaluate these three aspects.

In the terms of scenarios, human evaluators are tasked with evaluating three aspects: scenario accuracy, temporal consistency, and scenario suitability. **Scenario Accuracy** refers to whether the current scenario is one that could realistically occur in the character's personal life. **Temporal Consistency** assesses whether the scenario contains any elements or information that do not align with the character's background. **Scenario Suitability** examines whether it is plausible for conversations to take place with other characters within the given scenario. Each time, human evaluators are provided with a randomly selected scenario along with the rationale behind its generation. They are then asked to evaluate and score the scenario based on the three previously mentioned metrics. This comprehensive evaluation ensures that the scenarios not only provide authentic historical information but also guide the production of dialogue content.

In the terms of dialogues content, human evaluators are tasked with evaluating three aspects: dialogue-scene consistency, dialogue-historical consistency, and dialogue-character consistency. **Dialogue-Scenario Consistency** assesses whether the current dialogue aligns with the scene and is appropriate for the given location. **Dialogue-Historical Consistency** evaluates whether the dialogue content fits within the character's historical background and ensures that no elements outside of that history are present. **Dialogue-Character Consistency** examines whether the dialogue matches the character's identity, profession, and personal background, and whether it is generated in the character's distinct voice. Each time, human evaluators receive a randomly selected scenario along with the dialogue content generated from that scenario. They are then required to evaluate and score the dialogue based on the three previously mentioned metrics. This comprehensive evaluation ensures that the dialogue is not only contextually appropriate but also true to the character's persona and historical setting.

In the terms of character mindsets, human evaluators are required to score based on three criteria: mindset-dialogue consistency, mindset-character consistency, and mindset effectiveness. **Mindset-Dialogue Consistency** assesses whether the mindset aligns logically with the dialogue, the current scene, and the characters involved. **Mindset-Character Consistency** evaluates whether the mindset matches the character's persona, identity, and background. **Mindset Effectiveness** determines whether the mindset allows the character's subsequent responses to be directly inferred and includes all the necessary information the character would use during the dialogue. Each time, human evaluators are provided with a randomly selected set of dialogues and the corresponding mindset content. They are then asked to score each set based on the three aforementioned criteria. This structured evaluation ensures that the character mindsets are not only logically coherent with the dialogues and scenes but also faithfully represent the characters' identities and provide sufficient information for generating authentic and effective responses. The results are shown in Table 7 and Figure 3.

A.2 Results Evaluation

Considering the issue of hallucinations in LLMs, relying solely on them to evaluate experimental results may introduce certain biases. To more effectively validate our experimental results, we also engage human evaluators to assess the results. Given that human evaluations are highly costly, we randomly select 100 data samples and evaluate them across three dimensions: Consistency, Knowledge Accuracy, and Overall. Consistency assesses the model's ability to maintain the assigned role. Knowledge Accuracy evaluates the model's capability to avoid responses that exceed the character's knowledge scope. Overall reflects the human preference for the reply from a human perspective. To ensure independence and objectivity, each human evaluator is provided with only one model's response at a time and is asked to score each set of dialogues based on the aforementioned three dimensions. **The detailed results are shown in**

A.3 Human Evaluator

All three human evaluators hold at least a bachelor's degree and are aged between 22 and 26, including one female and two males. Each instance of data annotation and scoring is considered a separate labeling task, for which we provide compensation above the local industry average for data labeling work. Prior to evaluating the data, we allocate one hour for each expert to review the relevant Wikipedia information about the character, understand the character's primary personality traits, and compensate them with an hourly wage based on the average data labeling salary. We have informed the human evaluators about the purpose and scope of the data usage and obtained their consent. The data labeling process does not involve any privacy concerns.

B Example

Tables 8 through 12 present all the prompts we used to generate the training data. Table 13 provides an example of the training data. In the "output" section, the LLMs are instructed to generate the mindset first, followed by the response based on that mindset. The mindset portion is longer than the response because it must consider more factors.

Tables 14 through 20 show all the prompts given to ChatGPT for evaluation. Table 20 provides examples of the evaluation data, while Table 21 presents example responses.

Table 7: The detailed human evaluation results of the training data

	Scenario			Dialogue Content			Mindset		
	Accuracy	Temporal	Suitability	Scenario	Historical	Character	Dialogue	Character	Effectiveness
Avg	9.46	9.46	9.27	9.54	9.10	9.44	9.40	9.52	9.41
# over 7	97.92%	96.00%	93.75%	99.08%	98.33%	98.42%	99.25%	99.25%	99.25%
# over 8	93.58%	91.67%	86.25%	95.67%	62.67%	94.25%	84.08%	95.17%	84.33%

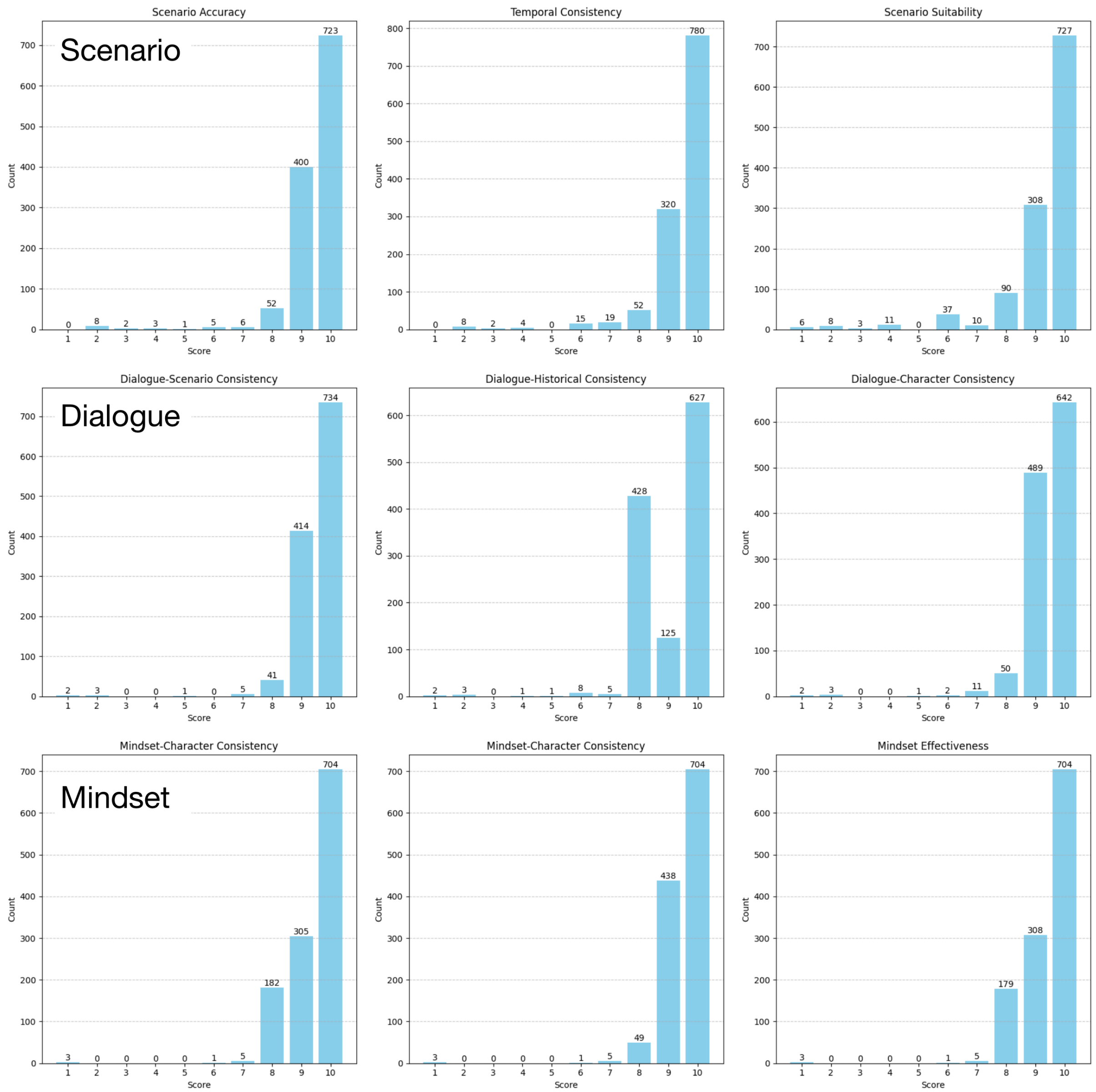

Figure 3

Table 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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.<|endoftext|>"

Table 14: 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 15: 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 16: 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 17: 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 18: 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 19: 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 20: 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 21: 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!