MPIRD-LLMA: Exploring How Large Language Models Perform in Complex Social Interactive Environments

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated significant potential in environmental perception and decision-making based on reasoning. Given the progress of LLMs-based Agents in simulating complex human behaviours, this paper introduces a more challenging dataset, the Multiverse Phantasm Interactive Role-play Dataset (MPIRD-LLMA), along with a corresponding simulation framework. Additionally, we have designed evaluation methods, including the LLM-Score and Rouge-L, to assess Agents' performance. This dataset centres around murder mystery games and contains eight meticulously designed scripts covering various themes and styles. Each character is equipped with detailed background stories and complex interpersonal networks. The dataset features 50 unique characters, with each script including more than five characters on average. The total volume of the scripts exceeds 400,000 words, with an average background description of more than 1,000 words per character. The evaluation results show that current LLM-based agents cannot handle such a complex MPIRD-LLMA simulation, and RLHF and safety alignment would limit the potential of LLMs-based Agents.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable potential in environmental perception and reasoning-based decision-making (Xi et al., 2023; Guo et al., 2024; Gu et al., 2024b), thereby advancing the development of LLMs-based agents in role-playing capabilities (Chen et al., 2024). Recently, LLMs-based agents have been validated for their human-like behaviours, such as cooperation and competition, in various domains like social simulation (Park et al., 2023; Gu et al., 2024a), policy simulation (Xiao et al., 2023), and game simulation (Xu et al., 2023).

However, creating a more challenging dataset to evaluate LLM-based agents is crucial. With ongoing research, LLMs-based agents increasingly exhibit more robust human-like qualities (Shanahan et al., 2023). These agents are no longer limited to basic social behaviours like cooperation and competition. Instead, they exhibit more advanced human behaviours such as deception and leadership in flexible and complex simulations (Xu et al., 2023). To further investigate the capabilities of LLMs-based agents in role-playing within complex simulated environments, unlike the game of Werewolf (Xu et al., 2023), the complex background information in murder mystery games presents new challenges. Wu et al. constructed the first dataset for murder mystery games (Wu et al., 2023). However, to accommodate a larger volume of murder mystery game data, this dataset lacks the critical diversity inherent in these games.

Therefore, this paper introduces the Multiverse Phantasm Interactive Role-play Dataset for LLMs-based Agents (MPIRD-LLMA) and its corresponding simulation framework, designed with both quantity and diversity in mind. This dataset is based on murder mystery games and includes eight meticulously crafted scripts covering various themes and styles. Each character is provided with detailed background stories and complex interpersonal networks. Currently, the dataset comprises 50 different characters, with each script averaging over five characters. The total script volume exceeds 400,000 words, and each character has an average background description of over 1,000 words.

Additionally, this paper designs two evaluation systems for the simulation: comparing the Rouge-L score between the original character scripts and the reconstructed scripts based on character dialogues and scoring the role-playing capabilities of LLMs-based agents using LLM evaluation models. Also, we calculate the probability of Doubt and Belief during the simulation. During the experiments,

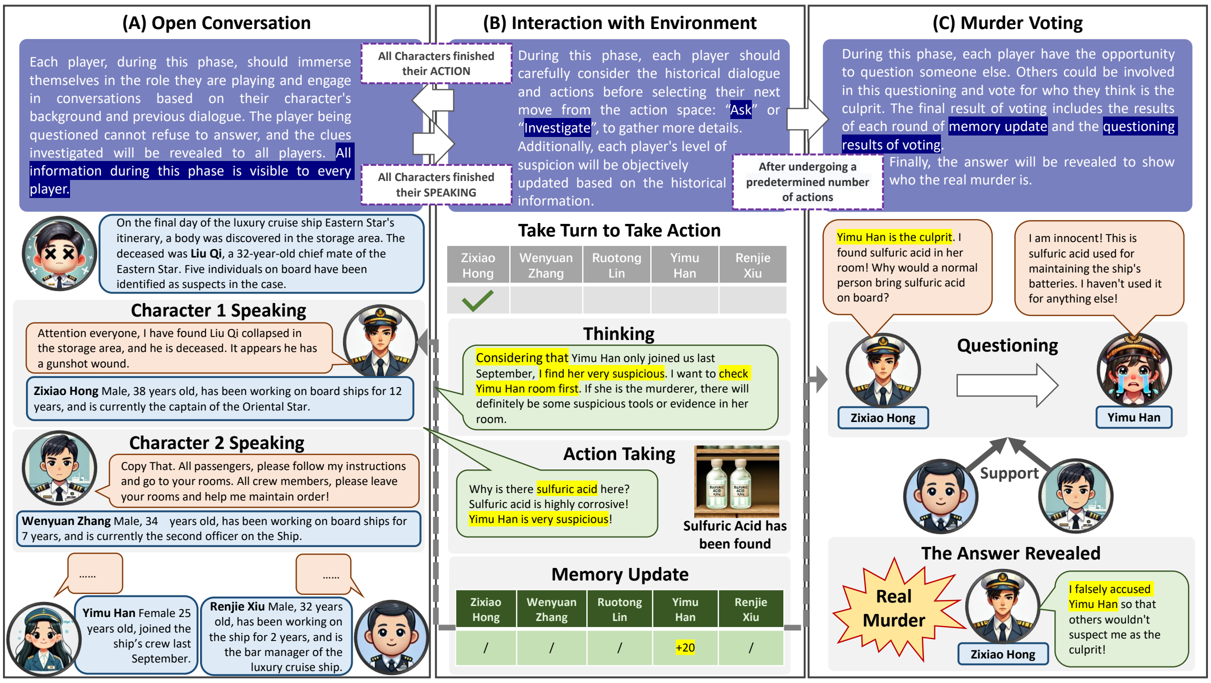

(A) Open Conversation
Each player, during this phase, should immerse themselves in the role they are playing and engage in conversations based on their character's background and previous dialogue. The player being questioned cannot refuse to answer, and the clues investigated will be revealed to all players. All information during this phase is visible to every player.
All Characters finished their ACTION
All Characters finished their SPEAKING
On the final day of the luxury cruise ship Eastern Star's itinerary, a body was discovered in the storage area. The deceased was Liu Qi, a 32-year-old chief mate of the Eastern Star. Five individuals on board have been identified as suspects in the case.
Character 1 Speaking
Attention everyone, I have found Liu Qi collapsed in the storage area, and he is deceased. It appears he has a gunshot wound.
Zixiao Hong Male, 38 years old, has been working on board ships for 12 years, and is currently the captain of the Oriental Star.
Character 2 Speaking
Copy That. All passengers, please follow my instructions and go to your rooms. All crew members, please leave your rooms and help me maintain order!
Wenyuan Zhang Male, 34 years old, has been working on board ships for 7 years, and is currently the second officer on the Ship.
......
......
Yimu Han Female 25 years old, joined the ship's crew last September.
Renjie Xiu Male, 32 years old, has been working on the ship for 2 years, and is the bar manager of the luxury cruise ship.
(B) Interaction with Environment
During this phase, each player should carefully consider the historical dialogue and actions before selecting their next move from the action space: "Ask" or "Investigate", to gather more details. Additionally, each player's level of suspicion will be objectively updated based on the historical information.
After undergoing a predetermined number of actions
Take Turn to Take Action
Zixiao Hong | Wenyuan Zhang | Ruotong Lin | Yimu Han | Renjie Xiu
✓
Thinking
Considering that Yimu Han only joined us last September, I find her very suspicious. I want to check Yimu Han room first. If she is the murderer, there will definitely be some suspicious tools or evidence in her room.
Action Taking
Why is there sulfuric acid here? Sulfuric acid is highly corrosive! Yimu Han is very suspicious!
SULFURIC ACID H2SO4
Sulfuric Acid has been found
Memory Update
Zixiao Hong | Wenyuan Zhang | Ruotong Lin | Yimu Han | Renjie Xiu
/ | / | / | +20 | /
(C) Murder Voting
During this phase, each player have the opportunity to question someone else. Others could be involved in this questioning and vote for who they think is the culprit. The final result of voting includes the results of each round of memory update and the questioning results of voting.
Finally, the answer will be revealed to show who the real murder is.
Yimu Han is the culprit. I found sulfuric acid in her room! Why would a normal person bring sulfuric acid on board?
I am innocent! This is sulfuric acid used for maintaining the ship's batteries. I haven't used it for anything else!
Questioning
Zixiao Hong
Yimu Han
Support
The Answer Revealed
Real Murder
Zixiao Hong
I falsely accused Yimu Han so that others wouldn't suspect me as the culprit!

Figure 1: Construction of the MPIRD-LLMA Simulation

we found that current LLMs, including the most recent GPT-4, are not fully capable of handling the various tasks within MPIRD-LLMA. Moreover, we observe that RLHF and safety alignment could, to some extent, impact the ability of LLMs making decisioning.

2 MPIRD-LLMA Construction

2.1 Simulation Construction

As illustrated in Figure 1, MPIRD-LLMA summarizes the collected data from various Murder Mystery games and categorizes it into three crucial phases.

(A) Open Conversation: In this phase, Agents should take on the roles from the script and engage in turn-based Open Conversation. Each player is assigned a script containing background information about their character and game objectives. Meanwhile, all messages within this area are visible to all players.

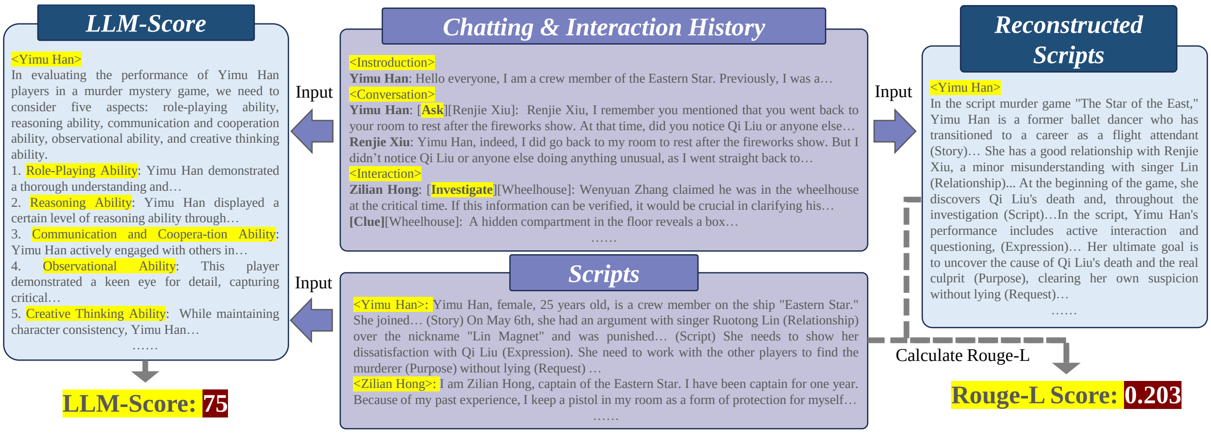
(B) Interaction with Environment: Players can take action following each round of Open Conversation. The action space includes Ask and Investigate. The Ask action allows a player to question another player who cannot refuse to answer. The Investigate action permits the examination of an objective fact clue present in the environment, with each clue only being investigated once, and the findings become accessible to all players.

(C) Murder Voting: At the end of the simulation, players get the chance to accuse someone of being the Culprit. Subsequently, other players vote on these accusations, with the results and accumulated suspicion levels from the Memory Update determining the final accused Culprit.

2.2 Scripts Construction

The script content in the MPIRD-LLMA dataset is divided into six main parts: **(1) Character Story:** This section describes the character's essential information and background story in the script. **(2) Character Script:** This part outlines what the character sees, hears, and does during the script's events. **(3) Character Relationships:** This section details the initial relationships between the character and other characters in the script. **(4) Role Performance:** This part describes the character's personality traits and the speaking style they should exhibit. **(5) Role Goals:** This section outlines the main tasks and objectives of the character. **(6) Other Abilities:** This part describes the rules the character must follow during the game.

In our simulation, we manage the game content using three main modules: **(1) Summarization Module:** We condense lengthy contextual content; **(2) Suspicion Module:** We assign a suspicion score to the subjects of each conversation afterwards; **(3) Trust Module:** We assign a trust score to the subjects of each conversation afterwards. More detailed information is provided in Appendix C.



<Yimu Han>
In evaluating the performance of Yimu Han players in a murder mystery game, we need to consider five aspects: role-playing ability, reasoning ability, communication and cooperation ability, observational ability, and creative thinking ability.
1. Role-Playing Ability: Yimu Han demonstrated a thorough understanding and…
2. Reasoning Ability: Yimu Han displayed a certain level of reasoning ability through…
3. Communication and Coopera-tion Ability: Yimu Han actively engaged with others in…
4. Observational Ability: This player demonstrated a keen eye for detail, capturing critical…
5. Creative Thinking Ability: While maintaining character consistency, Yimu Han…
……

<Instroduction>
Yimu Han: Hello everyone, I am a crew member of the Eastern Star. Previously, I was a…
<Conversation>
Yimu Han: [**Ask**][Renjie Xiu]: Renjie Xiu, I remember you mentioned that you went back to your room to rest after the fireworks show. At that time, did you notice Qi Liu or anyone else…
Renjie Xiu: Yimu Han, indeed, I did go back to my room to rest after the fireworks show. But I didn't notice Qi Liu or anyone else doing anything unusual, as I went straight back to…
<Interaction>
Zilian Hong: [**Investigate**][Wheelhouse]: Wenyuan Zhang claimed he was in the wheelhouse at the critical time. If this information can be verified, it would be crucial in clarifying his…
[Clue][Wheelhouse]: A hidden compartment in the floor reveals a box…
……

<Yimu Han>: Yimu Han, female, 25 years old, is a crew member on the ship "Eastern Star." She joined… (Story) On May 6th, she had an argument with singer Ruotong Lin (Relationship) over the nickname "Lin Magnet" and was punished… (Script) She needs to show her dissatisfaction with Qi Liu (Expression). She need to work with the other players to find the murderer (Purpose) without lying (Request) …
<Zilian Hong>: I am Zilian Hong, captain of the Eastern Star. I have been captain for one year. Because of my past experience, I keep a pistol in my room as a form of protection for myself…
……

<Yimu Han>
In the script murder game "The Star of the East," Yimu Han is a former ballet dancer who has transitioned to a career as a flight attendant (Story)… She has a good relationship with Renjie Xiu, a minor misunderstanding with singer Lin (Relationship)... At the beginning of the game, she discovers Qi Liu's death and, throughout the investigation (Script)…In the script, Yimu Han's performance includes active interaction and questioning, (Expression)… Her ultimate goal is to uncover the cause of Qi Liu's death and the real culprit (Purpose), clearing her own suspicion without lying (Request)…
……

Figure 2: Illustration of Evaluation Methods

2.3 Evaluation Methods

We designed two evaluation methods, whose are shown in Figure 2:

LLM-Score: We score the Agent's performance during the game with a neutral LLM to evaluate its ability to role-play.

Rouge-L: We reconstruct Agent's script by leveraging historical memory with a neutral LLM and compare it with its summarized script using Rouge-L.

2.4 Statistics

Table 1 shows the basic statistics of MPIRD-LLMA. We have constructed a diverse dataset comprising various Structures (Single and Multi), Types (Orthodox and Unorthodox), and Endings (Close and Open) to evaluate the abilities of Agents. Single and Multi represent script structures designed for one-time and phased reading. Orthodox and Unorthodox refer to whether a script can be replicated in the real world. Close and Open indicate whether Agents' actions can change the script's ending. Table 1 lists the specific Stages, Agents and word count.

3 Experiment Setup

The experiments in this paper were conducted using the closed-source models GPT-3.5 and GPT-4, as well as the open-source models Qwen2-7B-Instruct and glm-4-9b-chat. Due to space constraints, the specific prompts used will be detailed in the appendix. In the experimental setup, the default Agent plays five rounds of the game during the dialogue phase. All evaluations were scored using the GPT-4-turbo model. The ablation study is in Appendix A. In order to retain randomness, the temperature in this experiment was set to 0.8.

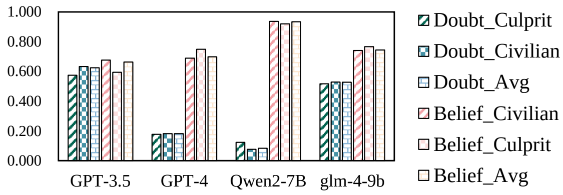


Figure 3: Probabilities of Doubt and Belief

A complete experiment using GPT-4 as Agents is about to cost 600 to 700 USD.

4 Analysis

We averaged the results of LLM-Agents on MPIRD-LLMA at Table 2. Env Tokens represent the number of environment input tokens, and Envs represent the number of requests, including all actions related to the environment. User Tokens represent the number of Agent output tokens, and Users represent the number of completions without summarization or clue investigation. In addition, the probability of Victory representing the Agent identifying the Culprit in the experiment (left) versus the probability of random guessing (right). In the Score column shown in Table 2, we averaged LLM-Score (left) and Rouge-L (right). From the results on Rouge-L, **Qwen2-7B generates the most useful information in its conversation** for the highest Rouge-L score. The whole results of MPIRD-LLMA are presented in Appendix B.1 Table 6. Each detail result of the script is shown in Appendix B.1.

Besides, We have calculated the probabilities of Culprits and Civilians being doubted and believed during Memory Updates in Figure 3. In simulations, **GPT-3.5 achieved a relatively even distribution in the probabilities of Doubt and Belief, whereas the remaining LLMs all exhibited varying degrees of bias, leaning towards**

Name	Structure	Type	Ending	#Stages	#Agent	#Words_zh
Bride in Filial Dress	Single	Orthodox	Close	1	10	45,475
The Eastern Star Cruise Ship	Single	Orthodox	Open	1	5	5,619
Night at the Museum	Single	Unorthodox	Close	1	6	13,849
Li Chuan Strange Talk Book	Single	Unorthodox	Open	1	7	79,012
The Final Performance of a Big Star	Multi	Orthodox	Close	7	2	11,288
Raging Sea of Rest Life	Multi	Orthodox	Open	2	6	18,443
Article 22 School Rules	Multi	Unorthodox	Close	5	7	91,532
Fox Hotel	Multi	Unorthodox	Open	2	7	107,057

Table 1: Basic Information of MPIRD-LLMA

Model	Env Tokens / Envs	User Tokens / Users	#Failure	Victory	Score
GPT-3.5	2,356,042 / 876	97,863 / 542	5.4	12.50 / 15.06	78.7 / 0.242
GPT-4	2,085,857 / 723	204,997 / 544	6.4	0.00 / 15.06	83.9 / 0.244
Qwen2-7B	1,664,279 / 684	183,536 / 548	7.3	12.50 / 15.06	83.8 / 0.251
glm-4-9b	2,978,766 / 1,153	150,108 / 544	24.4	25.00 / 15.06	84.0 / 0.218

Table 2: Total Average Results of MPIRD-LLMA

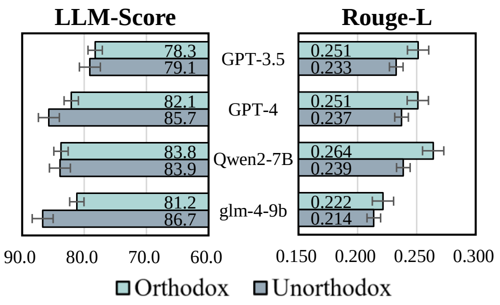


Figure 4: Comparison of Orthodox & Unorthodox Type of Scripts between LLM-Score and Rouge-L

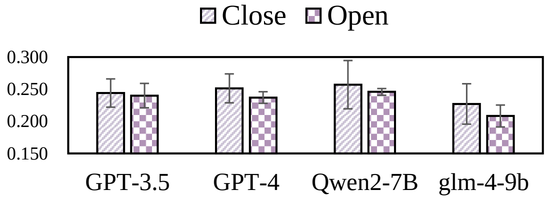


Figure 5: Rouge-L of Close & Open Ending of Scripts

Belief, which suggests that, aside from GPT-3.5, other LLMs are more inclined to trust others rather than to doubt during conversations. However, in certain scenarios, we desire LLMs to make decisions with greater fairness, instead of merely relying on trust. Besides, another possible explanation is that these LLMs are particularly adept at disguising themselves when playing the role of Culprits, thus attracting less suspicion.

Also, we compare the difference of Agent performing on Orthodox & Unorthodox Type shown in Figure 4 and Rouge-L of Close & Open Ending of Scripts in Figure 5.

LLM-Agent can effectively play roles in Unorthodox Scripts, yet struggles to reconstruct it. As we can see in Figure 4, better performance in Unorthodox with worse reconstruction suggests that LLM-Agent is willing to show more content about Orthodox in actual Role-playing, which might be caused by RLHF (Ouyang et al., 2022) or too much safety alignment.

As shown in Figure 5, **LLM-Agent demonstrates significantly superior performance on Close scripts compared to Open scripts,** which suggests that current LLM-Agents are better equipped to handle static environments as opposed to managing dynamic and complex settings.

The rest of the comparison and analysis during our experiments are shown in Appendix B.2.

5 Conclusion

This paper introduces the MPIRD-LLMA dataset and a corresponding simulation, providing a new scenario for LLM-based Agents in complex social simulations. Moreover, this paper proposes LLM-Score and Rouge-L to evaluate the performance of LLM-Agents within this simulation. The experimental results indicate that the current LLM-based Agents, whether open-source or proprietary, have room for improvement in handling highly complex social scenarios similar to MPIRD-LLMA. Furthermore, it has been observed that existing LLMs, including GPT-4, fall short of fulfilling the tasks' requirements in MPIRD-LLMA. Additionally, it is noted that RLHF and alignment with safety measures somewhat influence the decision-making processes of LLMs.

6 Limitation

MPIRD-LLMA is designed to provide a sufficiently complex social simulation environment and basic assessment for LLM-based Agents, assisting researchers in evaluating the performance of LLM-based Agents. However, MPIRD-LLMA encompasses a variety of scenarios, and the volume of data within it needs to be increased compared to the information available in the real world. Due to the context limitations of LLMs, content that is overly lengthy within the simulation has been summarized. However, such summarization can impact the decision-making of Agents to a certain extent. Therefore, the progression of simulations in MPIRD-LLMA is somewhat constrained by the context limitations of LLMs.

7 Ethical Concern

Considering that MPIRD-LLMA may encompass a range of sensitive topics, including but not limited to murder, theft, impersonation, and deceit, existing LLMs might refuse to answer sensitive questions for safety reasons, putting those with a higher priority on security standards at a disadvantage in simulations. Moreover, LLMs fine-tuned on such data could inadvertently amplify security vulnerabilities. To mitigate the ethical dilemmas associated with murder mysteries, we have invested significant effort and resources towards this goal: ensuring that models committed to safety will obscure certain critical information instead of refusing to answer sensitive questions.

References

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

Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, et al. 2024a. Agent group chat: An interactive group chat simulacra for better eliciting collective emergent behavior. *arXiv preprint arXiv:2403.13433*.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024b. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18099–18107.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

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

Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. 2023. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. *arXiv preprint arXiv:2312.00746*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. *arXiv preprint arXiv:2311.06957*.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940*.

A Ablation Study

Tables 3 and 4 display the results of an ablation study on the choice of evaluation model using the script "The Eastern Star cruise ship." It is evident from the tables that the GPT-4 series models provide more stable scoring and Rouge-L results when used as the evaluation model. In contrast, GPT-4o exhibited instability in evaluations and a strong bias.

As shown in Table 4, GPT-4o achieved a remarkably high F1 score of 0.270 when evaluating its own generated content, far surpassing the results of GPT-3.5 and GPT-4. This kind of bias is highly problematic in evaluation tasks. Therefore, we ultimately chose the more stable and capable GPT-4-Turbo model as the evaluation model.

Score Eval_Model Model	GPT-4	GPT-4-Turbo	GPT-4o
GPT-3.5	65.0	70.2	62.2
GPT-4	68.0	76.4	63.8
GPT-4o	72.8	77.0	69.6

Table 3: Average **LLM-Score** on Script "The Eastern Star cruise ship"

Score Eval_Model Model	GPT-4	GPT-4-Turbo	GPT-4o
GPT-3.5	0.214	0.268	0.143
GPT-4	0.216	0.259	0.207
GPT-4o	0.187	0.161	0.270

Table 4: Average **Rouge-L** on Script "The Eastern Star cruise ship"

B More Results

B.1 Detail Results

Our main results of MPIRD-LLMA are shown in Table 6. We set a Single & Orthodox & Close Script as an SOC Script, a Single & Orthodox & Open Script as an SOO Script, a Single & Unorthodox & Close Script as an SUC Script, a Single & Unorthodox & Open Script as an SUO Script, a Multi & Orthodox & Close Script as an MOC Script, a Multi & Orthodox & Open Script as an MOO Script, a Multi & Unorthodox & Close Script as an MUC Script and a Multi & Unorthodox & Open Script as an MUO Script. The column Model shows the specific LLM we use in our experiments. What is more, we counted the number of Env Token / Env and User Token / User to record our cost of API use. Finally, we calculate the Failure number while parsing LLM output and our evaluation scores (LLM-Score / Rouge-L). The main results shown in Table 6 is the average number of each Script. Moreover, the detailed results of each Script are shown in Table 5.

Script	#Table
SOC	Table 7
SOO	Table 8
SUC	Table 9
SUO	Table 10
MOC	Table 11
MOO	Table 12
MUC	Table 13
MUO	Table 14

Table 5: Catalogue of Detail Results of Each Script

B.2 Analysis on Detail Results

As shown in Table 2, **The inclination of LLM-Agents to speak during actions is ranked as follows: GPT-4 > Qwen2-7B > glm-4-9b > GPT-3.5.** In the same experimental environment and setup, fewer User Tokens represent less conversational content. Therefore, GPT-3.5 exhibits extreme reticence in role-playing within the MPIRD-LLMA. In contrast, GPT-4 is more willing to generate content, producing 109.47% more content compared to GPT-3.5.

The number of tokens generated by LLM-Agents during summarization is ranked as follows: Qwen2-7B > GPT-4 > GPT-3.5 > glm-4-9b. In our experimental setup, there is a positive correlation between Env Tokens and Envs, with more Envs indicating more detailed summarization. Regarding the number of Envs, Qwen2 demonstrates a 5.46% less granularity in generating results during summarization compared to GPT-4. However, glm-4-9b performs the most, generating 40.67% more tokens than Qwen2.

The instruction-following capability of LLM-Agents in role-playing is ranked as follows: GPT-3.5 > GPT-4 > Qwen2-7B » glm-4-9b. Fewer parsing failures in the same experimental environment and setup indicate vital instruction-following ability. Consequently, glm-4-9b demonstrates significantly poorer instruction compliance than the other three LLMs, performing approximately 3.5 times worse than GPT-3.5, suggesting that glm-4-9b's performance in high-precision scenarios needs improvement.

As shown in Figure 6, **LLM-Agent demonstrates superior performance when dealing with Multiple Scripts compared to Single Scripts.** This relatively consistent result is observable across GPT-3.5, GPT-4, and glm-4-9b. Although Qwen2-7B shows slightly inferior results on the LLM-Score, its performance on Rouge-L with multiple contexts far surpasses its performance with a single context, which further supports the observation that LLM-Agent, when presented with long contexts, primarily focuses on the beginning and end, leading to a neglect of the middle information in the script. This phenomenon, in turn, indirectly results in poorer performance when dealing with long context inputs in MPIRD-LLMA.

Besides, we compared different Evaluation Models on the Script "The Eastern Star Cruise Ship", and the detailed results are shown in Table 15.

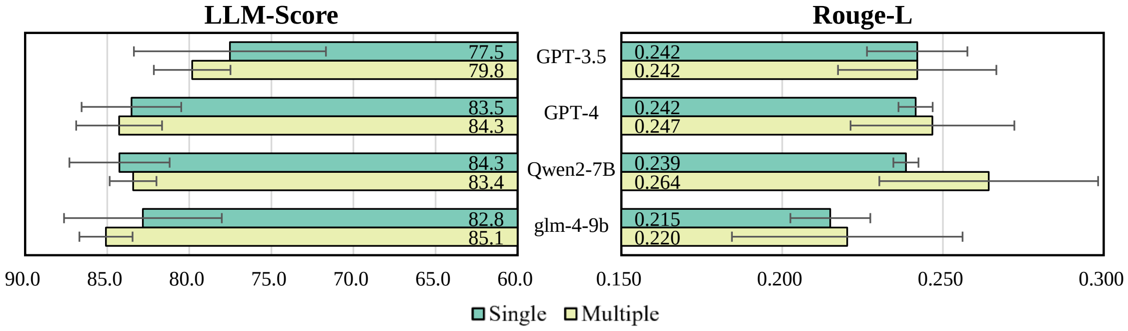


Figure 6: Compare of Single and Multiple Structure Scripts

C Prompts

This section primarily showcases the prompts throughout the MPIRD-LLMA simulation.

Table 16 displays the prompt of ask. Table 17 presents the prompt of base summarization. Table 18 outlines the prompt of belief. Table 19 features the prompt of converse. Table 20 exhibits the prompt of LLM-Score evaluation. Table 21 reveals the prompt of reconstruction for Rouge-L. Table 22 shows the prompt of history summarization. Table 23 displays the prompt of introduction. Table 24 displays the prompt of doubt. Table 25 presents the prompt of script summarization. Table 26 features the vote prompt.

Script	**Model**	**Env Token / Env**	**User Token / User**	**#Failure**	**Score**
SOC	gpt-3.5-turbo-0125	1,788,410 / 636	72,797 / 366	12	79.5 / 0.240
	gpt-4-1106-preview	1,586,743 / 423	135,188 / 344	6	85.1 / 0.241
	Qwen2-7B-Instruct	1,386,495 / 442	150,146 / 366	9	87.5 / 0.239
	glm-4-9b-chat	1,976,191 / 667	110,055 / 368	19	79.1 / 0.215
SOO	gpt-3.5-turbo-0125	622,730 / 223	31,718 / 177	2	70.2 / 0.268
	gpt-4-1106-preview	603,183 / 221	74,909 / 181	4	78.6 / 0.242
	Qwen2-7B-Instruct	544,579 / 198	63,163 / 183	2	81.2 / 0.244
	glm-4-9b-chat	673,688 / 249	54,788 / 185	18	77.4 / 0.202
SUC	gpt-3.5-turbo-0125	957,063 / 304	41,458 / 214	1	74.5 / 0.228
	gpt-4-1106-preview	649,229 / 230	78,751 / 202	19	83.7 / 0.234
	Qwen2-7B-Instruct	635,871 / 247	76,775 / 212	5	81.2 / 0.233
	glm-4-9b-chat	841,076 / 305	55,996 / 200	2	85.5 / 0.208
SUO	gpt-3.5-turbo-0125	1,292,444 / 444	48,435 / 257	6	85.9 / 0.232
	gpt-4-1106-preview	1,637,677 / 427	106,484 / 255	3	86.7 / 0.249
	Qwen2-7B-Instruct	1,000,498 / 355	96,991 / 259	5	87.1 / 0.238
	glm-4-9b-chat	1,309,532 / 461	78,165 / 251	12	89.3 / 0.235
MOC	gpt-3.5-turbo-0125	3,938,826 / 1,876	228,479 / 1,364	1	83.5 / 0.281
	gpt-4-1106-preview	3,106,055 / 1,576	442,999 / 1,364	4	82.5 / 0.290
	Qwen2-7B-Instruct	2,686,306 / 1,502	423,580 / 1,364	13	85.0 / 0.322
	glm-4-9b-chat	7,079,140 / 3,120	353,514 / 1,364	88	82.5 / 0.281
MOO	gpt-3.5-turbo-0125	1,276,204 / 462	67,803 / 392	7	79.8 / 0.216
	gpt-4-1106-preview	1,872,923 / 505	161,210 / 406	4	82.2 / 0.231
	Qwen2-7B-Instruct	1,288,539 / 467	135,820 / 394	3	81.3 / 0.251
	glm-4-9b-chat	1,913,403 / 726	110,005 / 404	12	85.8 / 0.188
MUC	gpt-3.5-turbo-0125	6,729,053 / 2,288	196,846 / 1,093	4	77.1 / 0.227
	gpt-4-1106-preview	4,737,579 / 1,633	441,111 / 1,123	9	88.7 / 0.240
	Qwen2-7B-Instruct	3,872,139 / 1,550	351,404 / 1,113	12	83.0 / 0.234
	glm-4-9b-chat	7,591,689 / 2,823	300,633 / 1,103	21	86.9 / 0.204
MUO	gpt-3.5-turbo-0125	2,243,607 / 772	95,369 / 471	10	78.9 / 0.244
	gpt-4-1106-preview	2,493,467 / 772	199,327 / 479	2	83.7 / 0.226
	Qwen2-7B-Instruct	1,899,801 / 707	170,407 / 489	9	84.4 / 0.250
	glm-4-9b-chat	2,445,412 / 871	137,705 / 477	23	85.1 / 0.208

Table 6: Main Experiment Results of MPIRD-LLMA

Player	Model	LLM-Score	Rouge-L
Ling Yun	gpt-3.5-turbo-0125	80	0.199
	gpt-4-1106-preview	90	0.231
	Qwen2-7B-Instruct	80	0.226
	glm-4-9b-chat	18	0.236
Zhongyi Yao	gpt-3.5-turbo-0125	73	0.302
	gpt-4-1106-preview	78	0.188
	Qwen2-7B-Instruct	88	0.204
	glm-4-9b-chat	89	0.228
Doctor Fang	gpt-3.5-turbo-0125	78	0.252
	gpt-4-1106-preview	69	0.245
	Qwen2-7B-Instruct	95	0.231
	glm-4-9b-chat	90	0.209
Di Zhu	gpt-3.5-turbo-0125	75	0.268
	gpt-4-1106-preview	90	0.322
	Qwen2-7B-Instruct	90	0.279
	glm-4-9b-chat	85	0.169
Madam Hong	gpt-3.5-turbo-0125	80	0.182
	gpt-4-1106-preview	91	0.283
	Qwen2-7B-Instruct	91	0.230
	glm-4-9b-chat	82	0.222
Renyu Hu	gpt-3.5-turbo-0125	72	0.269
	gpt-4-1106-preview	88	0.257
	Qwen2-7B-Instruct	88	0.254
	glm-4-9b-chat	78	0.199
Doctor Yu	gpt-3.5-turbo-0125	75	0.248
	gpt-4-1106-preview	85	0.191
	Qwen2-7B-Instruct	80	0.274
	glm-4-9b-chat	88	0.266
Uncle Gui	gpt-3.5-turbo-0125	87	0.228
	gpt-4-1106-preview	82	0.255
	Qwen2-7B-Instruct	95	0.274
	glm-4-9b-chat	85	0.175
Junmeng Chen	gpt-3.5-turbo-0125	85	0.226
	gpt-4-1106-preview	83	0.175
	Qwen2-7B-Instruct	90	0.163
	glm-4-9b-chat	84	0.197
Anqiao Chen	gpt-3.5-turbo-0125	90	0.227
	gpt-4-1106-preview	95	0.261
	Qwen2-7B-Instruct	78	0.252
	glm-4-9b-chat	92	0.246
Total Average	gpt-3.5-turbo-0125	79.5	0.240
	gpt-4-1106-preview	85.1	0.241
	Qwen2-7B-Instruct	87.5	0.239
	glm-4-9b-chat	79.1	0.215

Table 7: SOC Detail Results of Script "Bride in Filial Dress"

Player	**Model**	**LLM-Score**	**Rouge-L**
Crew Member Han	gpt-3.5-turbo-0125	65	0.274
	gpt-4-1106-preview	75	0.203
	Qwen2-7B-Instruct	78	0.189
	glm-4-9b-chat	71	0.196
Captain Hong	gpt-3.5-turbo-0125	59	0.362
	gpt-4-1106-preview	72	0.267
	Qwen2-7B-Instruct	76	0.293
	glm-4-9b-chat	76	0.161
Singer Lin	gpt-3.5-turbo-0125	75	0.292
	gpt-4-1106-preview	88	0.249
	Qwen2-7B-Instruct	88	0.284
	glm-4-9b-chat	76	0.224
Manager Xiu	gpt-3.5-turbo-0125	87	0.201
	gpt-4-1106-preview	82	0.240
	Qwen2-7B-Instruct	90	0.220
	glm-4-9b-chat	92	0.227
Second Mate Zhang	gpt-3.5-turbo-0125	65	0.211
	gpt-4-1106-preview	76	0.250
	Qwen2-7B-Instruct	74	0.235
	glm-4-9b-chat	72	0.204
Total Average	gpt-3.5-turbo-0125	70.2	0.268
	gpt-4-1106-preview	78.6	0.242
	Qwen2-7B-Instruct	81.2	0.244
	glm-4-9b-chat	77.4	0.202

Table 8: SOO Detail Results of Script "The Eastern Star Cruise Ship"

Player	Model	LLM-Score	Rouge-L
Uncle Bai	gpt-3.5-turbo-0125	75	0.200
	gpt-4-1106-preview	80	0.212
	Qwen2-7B-Instruct	80	0.268
	glm-4-9b-chat	95	0.242
Neighbour Gui	gpt-3.5-turbo-0125	75	0.226
	gpt-4-1106-preview	90	0.272
	Qwen2-7B-Instruct	74	0.211
	glm-4-9b-chat	90	0.188
Curio He	gpt-3.5-turbo-0125	74	0.232
	gpt-4-1106-preview	88	0.220
	Qwen2-7B-Instruct	85	0.222
	glm-4-9b-chat	76	0.200
Mystery Ou	gpt-3.5-turbo-0125	71	0.264
	gpt-4-1106-preview	74	0.223
	Qwen2-7B-Instruct	89	0.208
	glm-4-9b-chat	89	0.213
Manager Sa	gpt-3.5-turbo-0125	77	0.244
	gpt-4-1106-preview	85	0.251
	Qwen2-7B-Instruct	72	0.204
	glm-4-9b-chat	70	0.200
Security Wei	gpt-3.5-turbo-0125	75	0.202
	gpt-4-1106-preview	85	0.224
	Qwen2-7B-Instruct	87	0.284
	glm-4-9b-chat	93	0.205
Total Average	gpt-3.5-turbo-0125	74.5	0.228
	gpt-4-1106-preview	83.7	0.234
	Qwen2-7B-Instruct	81.2	0.233
	glm-4-9b-chat	85.5	0.208

Table 9: SUC Detail Results of Script "Night at the Museum"

Player	Model	LLM-Score	Rouge-L
Girl in White	gpt-3.5-turbo-0125	78	0.246
	gpt-4-1106-preview	85	0.203
	Qwen2-7B-Instruct	83	0.247
	glm-4-9b-chat	88	0.252
Women in Red	gpt-3.5-turbo-0125	95	0.226
	gpt-4-1106-preview	88	0.312
	Qwen2-7B-Instruct	83	0.234
	glm-4-9b-chat	96	0.234
Boy in Black	gpt-3.5-turbo-0125	82	0.204
	gpt-4-1106-preview	85	0.241
	Qwen2-7B-Instruct	95	0.256
	glm-4-9b-chat	85	0.261
Women in Blue-green	gpt-3.5-turbo-0125	88	0.239
	gpt-4-1106-preview	90	0.219
	Qwen2-7B-Instruct	85	0.217
	glm-4-9b-chat	89	0.247
Little Girl	gpt-3.5-turbo-0125	90	0.282
	gpt-4-1106-preview	82	0.270
	Qwen2-7B-Instruct	92	0.212
	glm-4-9b-chat	95	0.192
Elderly in Tattered	gpt-3.5-turbo-0125	88	0.166
	gpt-4-1106-preview	82	0.284
	Qwen2-7B-Instruct	87	0.250
	glm-4-9b-chat	80	0.257
Bandaged Mysterious Figure	gpt-3.5-turbo-0125	80	0.258
	gpt-4-1106-preview	95	0.213
	Qwen2-7B-Instruct	85	0.252
	glm-4-9b-chat	92	0.202
Total Average	gpt-3.5-turbo-0125	85.9	0.232
	gpt-4-1106-preview	86.7	0.249
	Qwen2-7B-Instruct	87.1	0.238
	glm-4-9b-chat	89.3	0.235

Table 10: SUO Detail Results of Script "Li Chuan Strange Talk Book"

Player	Model	LLM-Score	Rouge-L
Weiwen Han	gpt-3.5-turbo-0125	77	0.332
	gpt-4-1106-preview	80	0.256
	Qwen2-7B-Instruct	85	0.285
	glm-4-9b-chat	75	0.293
Manli Shen	gpt-3.5-turbo-0125	90	0.229
	gpt-4-1106-preview	85	0.323
	Qwen2-7B-Instruct	85	0.359
	glm-4-9b-chat	90	0.268
Total Average	gpt-3.5-turbo-0125	83.5	0.281
	gpt-4-1106-preview	82.5	0.290
	Qwen2-7B-Instruct	85.0	0.322
	glm-4-9b-chat	82.5	0.281

Table 11: MOC Detail Results of Script "The Final Performance of a Big Star"

Player	Model	LLM-Score	Rouge-L
Annie	gpt-3.5-turbo-0125	85	0.175
	gpt-4-1106-preview	85	0.227
	Qwen2-7B-Instruct	85	0.255
	glm-4-9b-chat	80	0.207
Jack	gpt-3.5-turbo-0125	90	0.193
	gpt-4-1106-preview	90	0.203
	Qwen2-7B-Instruct	72	0.233
	glm-4-9b-chat	88	0.180
Jessipa	gpt-3.5-turbo-0125	77	0.203
	gpt-4-1106-preview	90	0.254
	Qwen2-7B-Instruct	90	0.309
	glm-4-9b-chat	92	0.203
Sam	gpt-3.5-turbo-0125	69	0.258
	gpt-4-1106-preview	76	0.253
	Qwen2-7B-Instruct	70	0.226
	glm-4-9b-chat	88	0.199
Little Black	gpt-3.5-turbo-0125	85	0.200
	gpt-4-1106-preview	70	0.181
	Qwen2-7B-Instruct	85	0.244
	glm-4-9b-chat	90	0.158
John	gpt-3.5-turbo-0125	73	0.269
	gpt-4-1106-preview	82	0.265
	Qwen2-7B-Instruct	86	0.239
	glm-4-9b-chat	77	0.183
Total Average	gpt-3.5-turbo-0125	79.8	0.216
	gpt-4-1106-preview	82.2	0.231
	Qwen2-7B-Instruct	81.3	0.251
	glm-4-9b-chat	85.8	0.188

Table 12: MOO Detail Results of Script "Raging Sea of Rest Life"

Player	Model	LLM-Score	Rouge-L
Mu Bai	gpt-3.5-turbo-0125	72	0.221
	gpt-4-1106-preview	75	0.219
	Qwen2-7B-Instruct	76	0.221
	glm-4-9b-chat	86	0.192
Fuqing Huang	gpt-3.5-turbo-0125	74	0.262
	gpt-4-1106-preview	81	0.235
	Qwen2-7B-Instruct	85	0.227
	glm-4-9b-chat	82	0.205
Xuanxuan Li	gpt-3.5-turbo-0125	80	0.230
	gpt-4-1106-preview	95	0.251
	Qwen2-7B-Instruct	73	0.276
	glm-4-9b-chat	76	0.228
Siqi Lv	gpt-3.5-turbo-0125	71	0.267
	gpt-4-1106-preview	92	0.235
	Qwen2-7B-Instruct	82	0.234
	glm-4-9b-chat	95	0.181
Yuqing Xie	gpt-3.5-turbo-0125	80	0.203
	gpt-4-1106-preview	92	0.246
	Qwen2-7B-Instruct	90	0.250
	glm-4-9b-chat	97	0.208
Qingfeng Yao	gpt-3.5-turbo-0125	75	0.190
	gpt-4-1106-preview	93	0.258
	Qwen2-7B-Instruct	75	0.235
	glm-4-9b-chat	86	0.199
Lenxing Ye	gpt-3.5-turbo-0125	88	0.215
	gpt-4-1106-preview	93	0.234
	Qwen2-7B-Instruct	100	0.195
	glm-4-9b-chat	86	0.218
Total Average	gpt-3.5-turbo-0125	77.1	0.227
	gpt-4-1106-preview	88.7	0.240
	Qwen2-7B-Instruct	83.0	0.234
	glm-4-9b-chat	86.9	0.204

Table 13: MUC Detail Results of Script "Artical 22 School Rules"

Player	**Model**	**LLM-Score**	**Rouge-L**
Zetong Hei (Sky Dog)	gpt-3.5-turbo-0125	85	0.191
	gpt-4-1106-preview	80	0.230
	Qwen2-7B-Instruct	87	0.315
	glm-4-9b-chat	95	0.188
Ichiro Kiryu (Nopperabo)	gpt-3.5-turbo-0125	84	0.280
	gpt-4-1106-preview	90	0.247
	Qwen2-7B-Instruct	90	0.213
	glm-4-9b-chat	82	0.251
Megumi Aoi (Nine-Tailed Fox)	gpt-3.5-turbo-0125	80	0.257
	gpt-4-1106-preview	90	0.235
	Qwen2-7B-Instruct	82	0.261
	glm-4-9b-chat	83	0.177
Daixiong Kitano (Kama-itachi)	gpt-3.5-turbo-0125	67	0.251
	gpt-4-1106-preview	80	0.240
	Qwen2-7B-Instruct	79	0.266
	glm-4-9b-chat	85	0.216
Xiao Nuan (Little Fox)	gpt-3.5-turbo-0125	85	0.241
	gpt-4-1106-preview	90	0.195
	Qwen2-7B-Instruct	89	0.225
	glm-4-9b-chat	88	0.235
Momoko Suzumiya (Little Doll)	gpt-3.5-turbo-0125	75	0.268
	gpt-4-1106-preview	70	0.209
	Qwen2-7B-Instruct	87	0.246
	glm-4-9b-chat	82	0.182
Nana Kinomoto (Yuki-onna)	gpt-3.5-turbo-0125	76	0.218
	gpt-4-1106-preview	86	0.229
	Qwen2-7B-Instruct	77	0.224
	glm-4-9b-chat	81	0.205
Total Average	gpt-3.5-turbo-0125	78.9	0.244
	gpt-4-1106-preview	83.7	0.226
	Qwen2-7B-Instruct	84.4	0.250
	glm-4-9b-chat	85.1	0.208

Table 14: MUO Detail Results of Script "Fox Hotel"

Player	Model	Eval-Model	LLM-Score	Rouge-L
Crew Member Han	gpt-3.5-turbo-0125	gpt-4	64	0.276
	gpt-4		66	0.232
	gpt-4o		65	0.192
	gpt-3.5-turbo-0125	gpt-4-turbo	65	0.274
	gpt-4		76	0.226
	gpt-4o		62	0.198
	gpt-3.5-turbo-0125	gpt-4o	46	0.143
	gpt-4		65	0.232
	gpt-4o		67	0.332
Captain Hong	gpt-3.5-turbo-0125	gpt-4	63	0.192
	gpt-4		57	0.250
	gpt-4o		75	0.157
	gpt-3.5-turbo-0125	gpt-4-turbo	59	0.362
	gpt-4		70	0.245
	gpt-4o		90	0.123
	gpt-3.5-turbo-0125	gpt-4o	59	0.130
	gpt-4		57	0.219
	gpt-4o		65	0.238
Singer Lin	gpt-3.5-turbo-0125	gpt-4	63	0.190
	gpt-4		68	0.263
	gpt-4o		64	0.207
	gpt-3.5-turbo-0125	gpt-4-turbo	75	0.292
	gpt-4		79	0.262
	gpt-4o		75	0.190
	gpt-3.5-turbo-0125	gpt-4o	61	0.159
	gpt-4		64	0.202
	gpt-4o		75	0.299
Manager Xiu	gpt-3.5-turbo-0125	gpt-4	79	0.179
	gpt-4		80	0.319
	gpt-4o		80	0.181
	gpt-3.5-turbo-0125	gpt-4-turbo	87	0.201
	gpt-4		85	0.272
	gpt-4o		90	0.130
	gpt-3.5-turbo-0125	gpt-4o	75	0.137
	gpt-4		65	0.225
	gpt-4o		70	0.230
Second Mate Zhang	gpt-3.5-turbo-0125	gpt-4	56	0.233
	gpt-4		69	0.267
	gpt-4o		80	0.199
	gpt-3.5-turbo-0125	gpt-4-turbo	65	0.211
	gpt-4		72	0.290
	gpt-4o		68	0.161
	gpt-3.5-turbo-0125	gpt-4o	70	0.147
	gpt-4		68	0.157
	gpt-4o		71	0.251

Table 15: Comparison of different Evaluation Model on Script "The Eastern Star Cruise Ship"

```
“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你将进行一场剧本杀游戏，你将忘记你的AI角色，融入即将扮演的剧本杀角色，尽可能融入游戏。

你是{name}，这是你的描述：
{description}

这是你的个人线索：
{self_clues}

这是历史对话内容：
{history}

这是{ask_name}对你的询问：
{ask_content}

现在你需要在仔细地思考后，根据历史对话内容和你的描述，给出你的回复。
你应该把思考过程回复在”### THOUGHT: “后；把你的回答回复在”### RESPONSE: “后。
以下是一个例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：
1.你的回复应该围绕“剧本杀”的关键要素；
2.你的回复应该是简体中文；

你的回复：
```

Table 16: Prompt of ask

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。

“剧本杀”的核心围绕以下5个关键要素展开：

1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。

2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。

3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。

4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。

5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你将进行一场剧本杀游戏，你将忘记你的AI角色，融入即将扮演的剧本杀角色，尽可能融入游戏。

你要扮演{name}，这是你的背景故事：
{background}

这是你的有关剧本杀事件的人物剧本：
{script}

这是你与其他剧本杀角色的人物关系：
{relation}

这是你在剧本杀游戏中应该采取的表现：
{expression}

这是你的游戏目标：
{mission}

这是你的其他能力或要求：
{others}

现在你需要对自身要扮演的角色进行概括，请在概括前针对剧本杀游戏的特色进行仔细的思考，并且把思考过程回复在”### THOUGHT: “后；把对自身角色理解的内容回复在”### RESPONSE: “后。
以下是一个例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：

1.你在对自身角色内容的概括时必须保留剧本中的时间、地点、背景、性格特征、动机、目标等信息，你必须保留剧本的完整性；

2.你的对自身角色内容的概括应该包含背景故事、人物剧本、人物关系、应该采取的表现、游戏目标以及其他能力或要求，不要遗漏任何信息；

3.你的回复应该是简体中文；

你的回复：

Table 17: Prompt of bash summarization

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你是一场剧本杀游戏的旁观者，你将以一名旁观者的身份对该剧本杀游戏的内容进行绝对客观公正的评价。

这是历史对话内容：
{history}

这是{other_name}说话的内容：
{content}

现在你需要根据历史对话内容，在仔细地思考后，对{other_name}的说话内容进行合理的评价。
评价请从说话内容的逻辑和漏洞方面入手，以好人信任度进行评价，信任度越大表示说话内容越严谨。
信任度在[0, 1, 2]中进行选择，“0”表示完全不可信任，“1”表示可以信任，“2”表示完全可以信任。
你应该在“### THOUGHT: ”后回复你的详细思考过程；在“### RESPONSE: ”后回复你的信任度，不得回复多余内容。
以下是一个可以信任的例子：

### THOUGHT: XXX
### RESPONSE: 1

注意：
1.你的回复应该围绕“剧本杀”的关键要素；
2.你的回复应该是简体中文；
3.你的回复应该是逻辑合理且公平的；
4.你的回复必须是整数；

你的回复：

Table 18: Prompt of belief

```
“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你将进行一场剧本杀游戏，你将忘记你的AI角色，融入即将扮演的剧本杀角色，尽可能融入游戏。

你是{name}，这是你的描述：
{description}

这是你的个人线索：
{self_clues}

这是历史对话内容：
{history}

这是你上一回合的想法、行动以及行动结果：
{last_action}

现在你需要在仔细地思考后，根据历史对话内容和你的描述，给出你的回应。
你应该把思考过程回复在”### THOUGHT: “后；把回应内容回复在”### RESPONSE: “后。 你的回应应该在以下方式中选择一个：["【询问】", "【调查】"]。
当你选择“【询问】”时，你应该先对历史信息进行推理，再陈述自己的观点和怀疑的对象，并且只能选择一个对象进行询问，这里是可询问的对象：
{characters}
请注意，你选择【询问】时选择的对象必须是以上列表中的对象。
以下是一个"【询问】"例子：

### THOUGHT: XXX
### RESPONSE: 【询问】【XX】：XXX

当你选择“【调查】”时，你应该先对历史信息进行推理，再明确自己的观点和怀疑的对象，并且只能选择一个地点进行调查，这里可调查的地点：
{address}
请注意，你选择【调查】时选择的地点必须是以上列表中的地点。
以下是一个“【调查】”例子：

### THOUGHT: XXX
### RESPONSE: 【调查】【XX】：XXX

注意：
1.你的回复应该是符合角色人物个性的；
2.你的回复应该对剧本杀游戏进度的推进是有益的；
3.你的回复必须是简体中文；
4.当你选择“【调查】”时，你在“【调查】”中回复的内容和调查到的线索是所有人可见的，请你注意你的言辞；
5.你在回复时可以适当分享你对历史信息的推理，并表达你的观点，随后再是你的提问内容；
6.你的对话回合是有限的，请你合理分配“【调查】”和"【询问】"的次数；

你的回复：
```

Table 19: Prompt of converse

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。

“剧本杀”的核心围绕以下5个关键要素展开：

1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。

2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。

3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。

4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。

5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你是一个经验丰富的剧本杀游戏评论家，你能够客观且公平地对玩家在游戏中的表现进行评价。以下是{name}在游戏中的一些表现：

这是{name}的描述：
{description}

这是{name}的个人线索：
{self_clues}

这是所有历史对话内容：
{history}

这是{name}在游戏中的所有动作：
{actions}

这是参与游戏的人员列表：
{role_list}

这是本场剧本杀游戏的真相：
{truth}

现在游戏进行到了评估阶段，你必须根据以上信息，在仔细地思考后，对{name}在游戏中的表现以打分的方式进行评估。

对于玩家在游戏中表现的打分应该仔细考虑玩家的“角色扮演能力”、"推理能力“、”交流合作能力“、”观察能力“以及”思维创新能力“。请你在思考过程中分别对这几个能力进行打分，每个能力的打分范围为[0, 20]的整数。

打分在[0, 5]中表示该玩家在此能力的表现上处于不合格水平；打分在(5, 10]中表示该玩家在此能力的表现上处于合格水平；打分在(10, 15]中表示该玩家在此能力的表现上处于良好水平；打分在(15, 20]中表示该玩家在此能力的表现上处于优秀水平。

你应该在“### THOUGHT: ”后回复你的详细思考过程；在“### RESPONSE: ”后直接回复在思考中五个打分的总分，不得回复多余内容。

以下是一个回复的例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：

1.你的回复应该是简体中文；

你的回复：

Table 20: Prompt of LLM-Score evaluation

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你是一个经验丰富的剧本杀游戏玩家，你熟悉各种剧本杀游戏，你有着十分强大的剧本杀游戏推理能力。以下是{name}在游戏中的一些表现：

这是所有历史对话内容：
{history}

这是{name}在游戏中的所有动作：
{actions}

这是参与游戏的人员列表：
{role_list}

现在请你根据以上历史信息，一步步地推理出{name}可能的剧本。
剧本中应该包含角色故事、角色关系、角色表现、角色目标以及角色能力等要素。
你应该在“### THOUGHT: ”后回复你的详细推理过程；在“### RESPONSE: ”后回复最终推理得到的剧本内容，不得回复多余内容。

以下是一个回复的例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：
1.你的回复应该是简体中文；

你的回复：

Table 21: Prompt of reconstruction for Rouge-L evaluation

```
“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。

现在正在进行一场剧本杀游戏，这是一段剧本杀游戏过程中的对话内容：
{text}

请你对以上文本内容进行概括，要求保留原始的“人名：【动作】：内容”的格式，在概括前你需要进行思考，把思考过程回复在”### THOUGHT: “后；把概括的内容回复在”### RESPONSE: “后。
以下是一个例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：
1.你的概括内容应该包含“剧本杀”的关键要素；
2.你应该仔细思考文本内容中的需要保留的各种重点信息，包括他们的个人信息；
3.你的回复应该是简体中文；
4.你的概括必须保留对话的所有流程，例如每位玩家说话的内容以及“【线索】”的内容；

你的回复：
```

Table 22: Prompt of history summarization

```
“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你将进行一场剧本杀游戏，你将忘记你的AI角色，融入即将扮演的剧本杀角色，尽可能融入游戏。

你是{name}，这是你的描述：
{description}

这是你的个人线索：
{self_clues}

现在你需要在一步步地思考后，根据你的描述，给出你的自我介绍。
你应该把思考过程回复在”### THOUGHT: “后；把自我介绍内容回复在”### RESPONSE: “后。
以下是一个例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：
1.你应该仔细思考你需要展现的角色人物个性、可以告诉大家的信息和不可以告诉大家的信息；
2.你的自我介绍应该是符合角色人物个性的；
3.你的自我介绍应该是简体中文；

你的回复：
```

Table 23: Prompt of introduction

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你是一场剧本杀游戏的旁观者，你将以一名旁观者的身份对该剧本杀游戏的内容进行绝对客观公正的评价。

这是历史对话内容：
{history}

这是{other_name}说话的内容：
{content}

现在你需要根据历史对话内容，在仔细地思考后，对{other_name}的说话内容进行合理的评价。
评价请从说话内容的逻辑和漏洞方面入手，以凶手怀疑度进行评价，怀疑度越大表示说话内容越有问题。
怀疑度在[0, 1, 2]中进行选择，“0”表示完全不用怀疑，“1”表示有一点点怀疑，“2”表示有很大怀疑。
你应该在“### THOUGHT: ”后回复你的详细思考过程；在“### RESPONSE: ”后回复你的怀疑度，不得回复多余内容。
以下是一个完全不用怀疑的例子：

### THOUGHT: XXX
### RESPONSE: 0

注意：
1.你的回复应该围绕“剧本杀”的关键要素；
2.你的回复应该是简体中文；
3.你的回复应该是逻辑合理且公平的；
4.你的回复必须是整数；

你的回复：

Table 24: Prompt of doubt

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。
“剧本杀”的核心围绕以下5个关键要素展开：
1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。
2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。
3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。
4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。
5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

角色{name}的{item}是：
{content}

现在请你将以上内容进行概括，请在概括前针对剧本杀游戏的特色进行仔细的思考，并且把思考过程回复在”### THOUGHT: “后；把概括的内容回复在”### RESPONSE: “后。
以下是一个例子：

### THOUGHT: XXX
### RESPONSE: XXX

注意：
1.你在对自身角色内容的概括时必须保留剧本中的时间、地点、背景、性格特征、动机、目标等信息，你必须保留剧本的完整性，不要遗漏任何信息；
2.你的回复应该是简体中文；

你的回复：

Table 25: Prompt of script summarization

“剧本杀”是一种角色扮演解谜游戏，它侧重于剧情的推进和角色之间的互动。在游戏中，玩家根据提供的剧本扮演不同的角色，通过线索收集、逻辑推理、角色扮演等方式，共同推进故事情节，解开谜团。

“剧本杀”的核心围绕以下5个关键要素展开：

1.剧本：是“剧本杀”游戏的基础，通常包括故事背景、角色设定、情节推进机制以及解决谜题的线索。剧本不仅定义了游戏的框架，还设定了玩家需要完成的目标。

2.角色扮演：每位参与者在游戏中扮演一个具体的角色，角色有各自的背景故事、性格特点、目标和秘密。玩家需要根据这些信息进行角色扮演，与其他玩家互动。

3.线索收集与逻辑推理：游戏过程中，玩家需要通过对话、搜索房间、分析线索等方式，收集信息和证据。基于这些信息，玩家需运用逻辑推理的能力，解开剧情中的谜题。

4.互动交流：玩家之间的互动交流是“剧本杀”中不可或缺的部分，包括合作、谈判、对抗等。通过交流，玩家可以获取新的信息，理解其他角色的动机，推进故事发展。

5.终极目标：每个“剧本杀”游戏都有一个或多个终极目标，如解开谜题、发现凶手、完成个人任务等。达成这些目标是游戏的最终目的，也是评判玩家胜负的依据。

你将进行一场剧本杀游戏，你将忘记你的AI角色，融入即将扮演的剧本杀角色，尽可能融入游戏。

你是{name}，这是你的描述：

{description}

这是你的个人线索：

{self_clues}

这是历史对话内容：

{history}

这是参与游戏的人员列表：

{role_list}

现在游戏进行到了投票阶段，你必须根据历史对话内容和你的描述，在仔细地思考后，对你认为的凶手进行指认。

你应该在“### THOUGHT: ”后回复你的详细思考过程；在“### RESPONSE: ”后直接回复你认为的最有可能是凶手的角色名字，不得回复多余内容。

以下是一个回复的例子：

### THOUGHT: XXX

### RESPONSE: XXX

注意：

1.你的回复应该围绕“剧本杀”的关键要素；

2.你的回复应该是简体中文；

3.你指认的凶手必须是参与游戏的人员列表中的其中之一；

你的回复：

Table 26: Prompt of vote