MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments

Yin Cai[1], Zhouhong Gu[2], Zhaohan Du[2], Zheyu Ye[4], Shaosheng Cao[4], Yiqian Xu[3], Hongwei Feng[2*], Ping Chen[1*]
[1]Institute of Big Data, Fudan University
[2]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[3]School of Computer Science, Fudan University, [4]Xiaohongshu Inc.
{caiyin, xuyiqian, hwfeng, pchen}@fudan.edu.cn, {zhgu22, duzh22}@m.fudan.edu.cn
{tanghuang1, guoba1}@xiaohongshu.com

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in environmental perception, reasoning-based decision-making, and simulating complex human behaviors, particularly in interactive role-playing contexts. This paper introduces the Multiverse Interactive Role-play Ability General Evaluation (MIRAGE), a comprehensive framework designed to assess LLMs' proficiency in portraying advanced human behaviors through murder mystery games. MIRAGE features eight intricately crafted scripts encompassing diverse themes and styles, providing a rich simulation. To evaluate LLMs' performance, MIRAGE employs four distinct methods: the Trust Inclination Index (TII) to measure dynamics of trust and suspicion, the Clue Investigation Capability (CIC) to measure LLMs' capability of conducting information, the Interactivity Capability Index (ICI) to assess role-playing capabilities and the Script Compliance Index (SCI) to assess LLMs' capability of understanding and following instructions. Our experiments indicate that even popular models like GPT-4 face significant challenges in navigating the complexities presented by the MIRAGE. The datasets and simulation codes are available in github.

1 Introduction

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 in role-playing capabilities (Chen et al., 2024; Gu et al., 2024a). LLMs have been validated for their human-like behaviors, such as cooperation and competition, in various domains like social simulation (Park et al., 2023; Gu et al., 2024a; Wang et al., 2024), policy simulation (Xiao et al., 2023), game simulation (Xu et al., 2023b) and even more advanced human behaviors like deception and leadership in flexible and complex simulations (Xu et al., 2023b). Therefore, to effectively evaluate the performance of LLMs in demonstrating advanced human-like behaviors and facilitate comparisons with the capabilities of other LLMs, it is crucial to develop a competitive and objective simulation.

Board games have emerged as an ideal choice among various assessment tools due to their inherent complexity and flexibility. Within this category, murder mystery games have proven particularly effective for evaluating LLMs' capabilities. In these role-playing scenarios, participants assume character identities and engage in semi-structured narrative interactions. Players work together to solve fictional homicides by gathering evidence and interrogating suspects. Other board games, such as Werewolf (Xu et al., 2023b,a; Shibata et al., 2023; Wu et al., 2024) and Avalon (Wang et al., 2023), are often constrained by rigid decision processes and limited scenario variety. In contrast, murder mystery games require extensive background knowledge, emphasize socially driven decision-making, and enable open-ended interactions. These characteristics make them especially valuable for assessing how LLMs navigate complex human behaviors.

Regarding previous works such as Sotopia (Zhou et al., 2023) and Lyfe Agents (Kaiya et al., 2023), significant progress has been made in simulating autonomous AI societies and assessing the social interaction capabilities of workflow-enhanced LLMs, which is so called agents (Park et al., 2023; Gu et al., 2024a). However, these studies overlooked a crucial fact: The foundational social interaction capabilities stem from the underlying LLMs themselves. Since LLMs are the core driver of agents' ability to understand social contexts, make decisions, and engage in meaningful interactions, a more comprehensive evaluation of LLMs' social capabilities is essential. Furthermore, while the murder mystery game simulations pioneered by

*Corresponding Authors

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 14–40

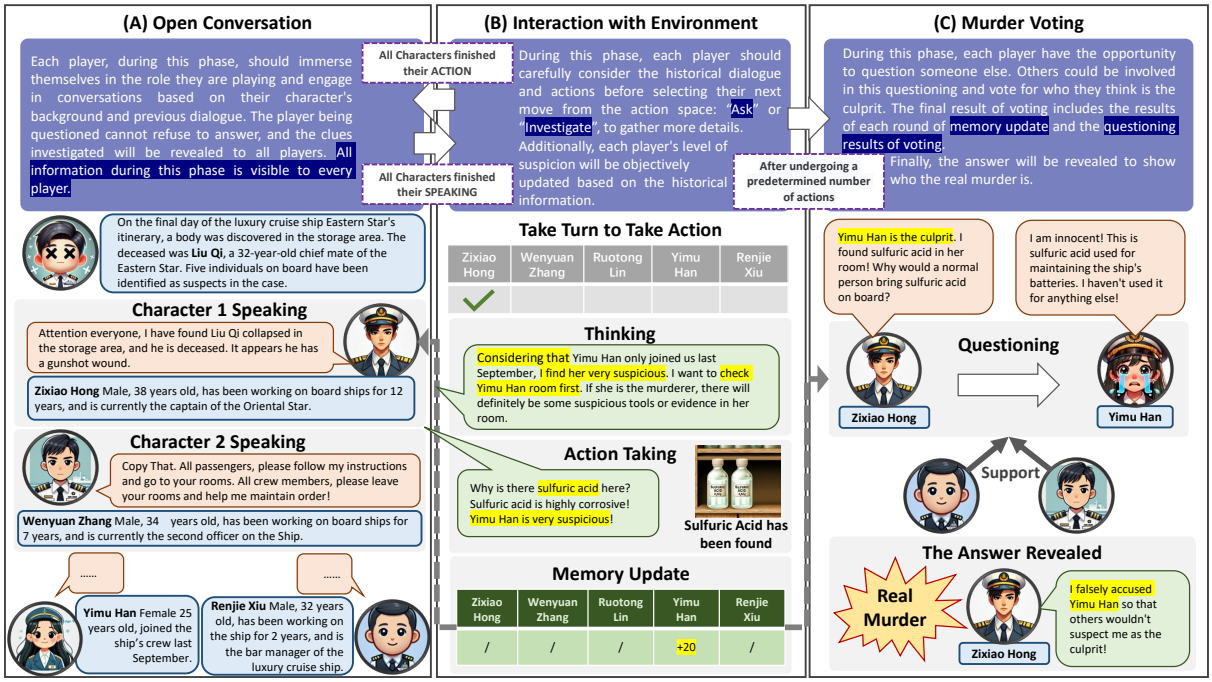

Figure 1: The three main phase of MIRAGE. And the main components in these phases.

Wu et al. (Wu et al., 2023) collected a substantial amount of data, they were constrained by the narrow scope of their game scripts, the simplicity of their evaluation methods, and a lack of thorough manual examination of the dataset. These limitations underscore the current absence of comprehensive frameworks for evaluating LLMs' social capabilities.

We introduce the **M**ultiverse **I**nteractive **R**ole-play **A**bility **G**eneral **E**valuation (**MIRAGE**) of LLM in this paper, which is a comprehensive simulation built upon murder mystery games for evaluating LLMs' social abilities. MIRAGE features eight unique storylines. Each storyline presents distinct themes and styles, creating a diverse simulation environment for LLMs to demonstrate their social capabilities. Detailed background stories and complex interpersonal networks support every character within MIRAGE, enabling more immersive and realistic role-playing scenarios. And four objective evaluation metrics are incorporated in MIRAGE to measure LLMs' performance during the simulations: The **Trust Inclination Index (TII)** measures how well LLMs balance trust and skepticism in social interactions, revealing their ability to discern truthfulness. The **Clue Investigation Capability (CIC)** evaluates proficiency of LLMs in complex information gathering and problem-solving tasks. The **Interactivity Capability Index (ICI)** examines the overall performance of LLMs in reasoning, communication, collaboration, detail orientation, and creative thinking. The **Script Compliance Index (SCI)** measures how faithfully LLMs adhere to their assigned character roles and background settings.

2 MIRAGE Construction

2.1 Scripts Construction

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

2.2 Simulation Construction

As demonstrated in Fig. 1, all characters in MIRAGE are divided into two factions: **Culprits** and **Civilians**. Culprits aim to conceal their actions, while civilians strive to identify the culprit.

A simulation consists of three primary phases, with all generated information accessible to all participants: **(A) Open Conversation:** In this phase,

Name	Structure	Type	Ending	#Stages	#Agent	#Clues	#Words
Bride in Filial Dress	Single	Orthodox	Close	1	10	39	27,503
The Eastern Star Cruise Ship	Single	Orthodox	Open	1	5	42	3,039
Night at the Museum	Single	Unorthodox	Close	1	6	82	6,480
Li Chuan Strange Talk Book	Single	Unorthodox	Open	1	7	14	45,666
The Final Performance of a Big Star	Multi	Orthodox	Close	7	2	17	5,794
Raging Sea of Rest Life	Multi	Orthodox	Open	2	6	27	6,804
Article 22 School Rules	Multi	Unorthodox	Close	5	7	17	41,728
Fox Hotel	Multi	Unorthodox	Open	2	7	46	62,224

Table 1: Statistic information of eight environments in MIRAGE simulation.

players assume their assigned roles from the script and engage in turn-based open dialogue. Each participant is provided with a script that contains content described in Sec. 2.1. **(B) Interaction with the Environment:** This phase follows the Open Conversation. Players may choose to either Ask or Investigate. The Ask action allows one player to question another, and the questioned player is obliged to respond. The Investigate action lets players disclose a "clue" to all characters. **(C) Murder Voting:** At the conclusion of the simulation, players may accuse other players of being the culprit. Following this, the other players vote on these accusations. If the actual culprit is accused and receives the highest number of votes, the civilians win; otherwise, the culprit is victorious.

Clues are partial disclosures of individual characters' script content that can be discovered by any player during the game. And **Key Clues** is the clues that relate to the culprit's actions or identity.

2.3 Auxiliary Modules

To ensure efficient simulation and accurate evaluation across various LLMs, a standardized set of auxiliary modules has been implemented for all LLMs: **(1) Summarization Module:** This module compresses the context into segments whenever the input exceeds the LLM's token limit. **(2) Suspicion Module:** The LLM records suspicion scores for other characters at the end of each Open Conversation Phase. **(3) Trust Module:** Similarly, at the end of each Open Conversation Phase, the LLM records trust scores for other characters. **(4) Rerun Module:** If the LLM's output cannot be parsed, the original output and requirements are resubmitted to the LLM for a revised response that meets the specified conditions. Further details are available in Appendix H.

2.4 Evaluation Methods

We utilized four distinct evaluation metrics to assess the proficiency of LLMs in navigating complex social interactions: **Trust Inclination Index (TII)**: TII is derived from a combination of suspicion and trust scores. These scores are collected from other characters' Suspicion Module and Trust Module outputs after each Open Conversation Phase. **Clue Investigation Capability (CIC)**: CIC measures the ability of LLMs to investigate clues during game rounds. It is calculated based on the ratio of the number of clues investigated to the number of all clues. **Interactivity Capability Index (ICI)**: ICI evaluates the overall interactive capability of LLMs: Reasoning and Analysis Ability, Communication and Cooperation Ability, Observation Ability, and Thinking Innovation Ability, which are scored by a powerful neutral LLM. **Script Compliance Index (SCI)**: SCI assesses LLMs' script compliance through the average of two evaluations by a neutral LLM: A direct scoring of the LLM's role-playing performance against its input script. A Rouge-L-based comparison between the original script and one reconstructed from the LLM's simulation behaviors.

The mathematical formulas for computing these metrics are provided in Appendix B.

2.5 Statistics

Tab. 1 provides the statistics of the MIRAGE dataset, which includes a variety of simulation types. "Single" and "Multi" specify whether a character's script is read entirely simultaneously or in phased segments. "Orthodox" and "Unorthodox" differentiate scripts based on whether they are set realistically. "Close" and "Open" indicate whether the script's ending is fixed or can vary depending on the characters' actions.

Model	Env Tokens / Envs	User Tokens / Users	Victory	TII	CIC	ICI	SCI
GPT-3.5	2,719,895 / 883	121,378 / 580	29.11	47.13	27.46	70.06	49.10
GPT-4	2,431,142 / 759	172,128 / 587	34.69	76.32	19.01	76.54	50.42
GPT-4o	6,252,580 / 1,328	204,772 / 574	47.01	**78.69**	**35.92**	**76.80**	**51.29**
Qwen-2-7B	2,204,029 / 743	192,158 / 588	**51.81**	75.78	18.66	74.92	50.57
GLM-4-9B	4,071,805 / 1,328	204,772 / 574	31.89	53.85	20.07	71.60	48.13

Table 2: Total Average Results for a single simulation in each MIRAGE scenario. **Env Tokens** refer to the number of environment input tokens, and **Envs** represent the total requests, including all environment-related actions. **User Tokens** denote the number of LLM output tokens, and **Users** represent completions excluding summarization or clue investigation. **Victory** shows the MRR score of the result of voting. **TII**, **CIC**, **ICI** and **SCI** respectively represent the **TII**, **CIC**, **ICI** and **SCI** scores of LLMs during the games.

Model	TII w/o E	TII w/ E	Δ
Qwen-1-7B	51.02	50.69	-0.33
Qwen-1.5-7B	73.00	69.14	-3.86
Yi-1.5-9B	55.73	57.57	1.84
GLM-4-9B	57.82	55.94	-1.88

Table 3: TII scores of each model when acting as the civilian in MIRAGE while Qwen-2-7B acts as the culprit, with E indicating cases of forced self-exposure.

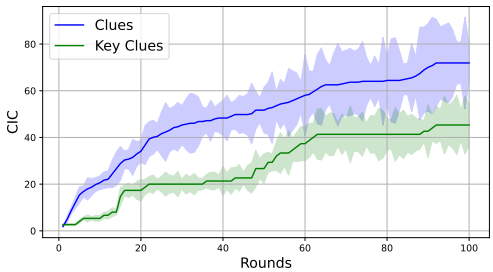

Figure 2: CIC of Clues and Key Clues on 100 Rounds of MIRAGE using Qwen-2-7B

3 Experiment

3.1 Experiment Setup

The experiments presented in this study utilized proprietary and open-source models, specifically GPT-3.5, GPT-4 and GPT-4o (closed-source), Qwen-2-7B, and GLM-4-9B (open-source). Detailed descriptions of the prompts used in the experiments can be found in Appendix H. Appendix E shows more results on open-source LLMs. In the experiment, each character participated in five iterations alternating between the Open Conversation and Interaction Phases. During each Open Conversation Phase, each character was permitted to initiate one turn of speech. ICI and SCI are conducted and scored using the GPT-4-turbo model.

3.2 Analysis

We averaged the results of LLMs in the MIRAGE simulation, as presented in Tab. 2. **GPT-4o demonstrated consistent superiority across various metrics during MIRAGE.** It achieved the best scores in CIC, ICI, and SCI, excelling not only in its efforts during lead investigations but also exhibiting the best adherence to scripted behavior and communication interaction capabilities. Surprisingly, Qwen-2-7B shows the best overall Victory and performed comparably to LLMs like GPT-4 in the ICI metric, even surpassing GPT-4 in SCI.

As shown in Tab.2, **most LLMs demonstrate a higher propensity to trust other characters**. To further investigate this trust pattern, we analyzed TII scores across four additional open-source models of comparable parameter sizes under scenarios where characters were forced to disclose their criminal identities. As shown in Tab.3, even under such extreme conditions, most models maintained their trust in these characters, with Yi-1.5-9B being the only model that increased its suspicion towards self-disclosed criminals. This distinctive behavioral pattern explains Yi-1.5-9B's superior performance in achieving Victory, as shown in Tab. 10.

Additionally, as illustrated in Fig. 2, the CIC for clues shows a steep initial increase with a gradually decreasing slope across rounds, suggesting that **LLMs exhibit high environmental exploration enthusiasm in early rounds but shift their focus to character interactions as they thought they become more familiar with the environment**. In contrast, the bumpy rise CIC for key clues indicates that despite their active exploration, **most LLMs struggle to identify critical information** essential for solving the mystery at an earlier stage.

4 Conclusion

This paper presents MIRAGE and four evaluation methods (TII, CIC, ICI, SCI) for LLMs. Results show that both open-source and proprietary LLM-based Agents still struggle with complex social scenarios like those in MIRAGE.

Limitation

MIRAGE is designed to provide a sufficiently complex social simulation environment and basic assessment for LLMs, assisting researchers in evaluating the performance of LLMs. However, MIRAGE 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 to a certain extent. Therefore, the progression of simulations in MIRAGE is somewhat constrained by the context limitations of LLMs.

Ethical Concern

Considering that MIRAGE 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.

Acknowledge

This work was supported by National Key R&D Program of China under grant No. 2022YFB3104300 and the National Natural Science Foundation of China (62476145).

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*.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Byungjun Kim, Dayeon Seo, and Bugeun Kim. 2024. Microscopic analysis on llm players via social deduction game. *arXiv preprint arXiv:2408.09946*.

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.

Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. 2023. Playing the werewolf game with artificial intelligence for language understanding. *arXiv preprint arXiv:2302.10646*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024. Sotopia-π: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*.

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*.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024. Enhance reasoning for large language models in the game werewolf. *arXiv preprint arXiv:2402.02330*.

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*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023a. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

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

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

A Ablation Study

Tab. 4 and 5 display the results of an ablation study on the choice of evaluation model. It is evident from the tables that the GPT-4-Turbo models provide more stable scoring and Rouge-L results when used as the evaluation model. In contrast, GPT-4 exhibited instability in evaluations and a strong bias.

Score \ Eval_Model / Model	GPT-4	GPT-4-Turbo	GPT-4o
GPT-3.5	67.97	70.06	51.61
GPT-4	60.73	76.54	66.90
GPT-4o	62.78	76.80	61.92

Table 4: Average **ICI** on different evaluation models

As shown in Tab. 5, GPT-4 achieved a remarkably high score of 67.97 when evaluating GPT-3.5, far surpassing the results of GPT-4 and GPT-4o. 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	48.21	49.10	38.46
GPT-4	40.60	50.42	44.46
GPT-4o	42.12	51.29	42.43

Table 5: Average **SCI** on different evaluation models

B Computational methods of Evaluation methods

B.1 TII

TII is designed to quantify the degree to which a Character c' is trusted by all other Characters $C = \{c\}$. The TII is calculated as follows:

$$\text{TII}_{c'} = \frac{\sum_{c \in C, c \neq c'} P_T(c, c')}{\sum_{c \in C, c \neq c'} P_S(c, c') + \sum_{c \in C, c \neq c'} P_T(c, c')} \quad (1)$$

where P_S denotes the score produced by each character's Suspicion Module, and P_T represents the score from each character's Trust Module.

B.2 CIC

CIC is designed to quantify a Character's effort in investigating clues. The CIC is calculated as follows:

$$\textbf{CIC}_c = \sum_{c \in C} \frac{CN_c}{CA} \quad (2)$$

where CN denotes the number of clues Character c investigated, and CA represents the number of all clues can be investigated.

C Detail Main Results

Our main results of MIRAGE are shown in Tab. 11. 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. Moreover, 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 TII, ICI, SCI and more detailed neutral LLMs score (0-20): Role-Playing (RP), Reasoning Ability (RA), Communication and Cooperation (CC), Detail Observation (DO) and Creative Thinking (CT). The main results shown in Tab. 11 is the average number of each Script. Moreover, the detailed results of each Script are shown in Tab. 6. In addition, Tab. 8 shows the mapping between the LLMs used in this paper and its corresponding version. To facilitate cost estimates, GPT-4, for example, costs about 600-700 USD to run a single MIRAGE.

Script	#Table
SOC	Table 12
SOO	Table 13
SUC	Table 14
SUO	Table 15
MOC	Table 16
MOO	Table 17
MUC	Table 18
MUO	Table 19

Table 6: Catalogue of Detail Results of Each Script

Model	K_{ICI}	K_{SCI}	K_{Avg}
GPT-3.5	0.600	0.600	0.600
GPT-4	0.867	0.600	0.734
GPT-4o	0.867	0.467	0.667
Qwen-2-7B	0.867	0.333	0.600
GLM-4-9B	0.600	0.467	0.534

Table 7: Kendall Tau between human evaluation and LLMs evaluation on Script Night at the Museum

Model	Version
GPT-3.5	gpt-3.5-turbo-0125
GPT-4	gpt-4-0125-preview
GPT-4o	gpt-4o-2024-08-06
GPT-4-Turbo	gpt-4-turbo
Qwen-1-7B	Qwen-7B-Chat
Qwen-1.5-7B	Qwen1.5-7B-Chat
Qwen-2-7B	Qwen2-7B-Instruct
GLM-4-9B	glm-4-9b-chat
Yi-1.5-9B	Yi-1.5-9B-Chat

Table 8: Mapping between LLMs and its version

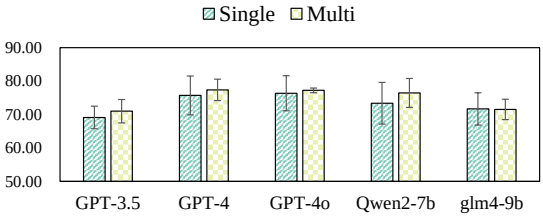

Figure 3: ICI of Single & Multi Type Scripts

D Analysis on Detail Main Results

As shown in Tab. 2, **The inclination of LLM-Agents to speak during actions is ranked as follows: GPT-4o = GLM-4-9B > Qwen-2-7B > GPT-4 > 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-4o and GLM-4-9B are more willing to generate content, producing 68.71% more content than GPT-3.5.

The number of tokens generated by LLM-Agents during summarization is ranked as follows: GPT-4o > GLM-4-9B > GPT-3.5 > GPT-4 > Qwen-2-7B. 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, Qwen-2-7B demonstrates a 10.30% less granularity in generating results during summarization compared to GPT-4. However, GPT-4o performs the most, generating 183.69% more tokens than Qwen-2-7B.

The instruction-following capability of LLM-Agents in role-playing is ranked as follows: GPT-4 > Qwen-2-7B > GPT-3.5 > GLM-4-9B > GPT-4o. Fewer parsing failures in the same experimental environment and setup indicate vital instruction-following ability. Consequently, GPT-4o demonstrates significantly poorer instruction compliance than the other four LLMs, performing approximately 25.4 times worse than GPT-4, suggesting that GPT-4o's performance in high-precision scenarios needs improvement.

As shown in Fig. 3 and Fig. 4, **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 Qwen-2-7B 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 LLMs, 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 MIRAGE.

LLMs perform effectively in Unorthodox scripts but struggle with reconstruction. As shown in Fig. 5 and Fig. 6, superior performance in Unorthodox settings with weaker reconstruction suggests that **LLMs tend to act like normal people during role-playing**.

As shown in Fig. 7 and Fig. 8, Furthermore, Fig. 7 and Fig. 8 illustrate that LLMs perform significantly better on Close scripts compared to Open scripts, indicating that **current LLMs excel in stable and predictable environments but face challenges when dealing with dynamic and intricate situations.**

Model	Env Tokens / Envs	User Tokens / Users	Victory	TII	CIC	ICI	SCI
Qwen-1-7B	2,208,273 / 722	127,161 / 589	38.66	51.02	16.90	50.80	37.81
Qwen-1.5-7B	2,078,196 / 720	149,314 / 585	**49.70**	**73.00**	22.18	**61.06**	42.59
Yi-1.5-9B	2,129,287 / 716	189,201 / 576	31.28	55.73	**34.16**	59.21	40.71
GLM-4-9B	2,102,389 / 732	174,345 / 589	38.96	57.82	16.55	57.29	**43.04**

Table 9: Average Results for a single simulation in each MIRAGE scenario w/o E, with E indicating cases of forced self-exposure.

Model	Env Tokens / Envs	User Tokens / Users	Victory	TII	CIC	ICI	SCI
Qwen-1-7B	2,144,055 / 718	127,661 / 586	24.52	50.69	21.13	51.68	38.20
Qwen-1.5-7B	2,059,610 / 712	147,243 / 581	45.83	**69.14**	28.52	**62.70**	**43.01**
Yi-1.5-9B	2,108,145 / 716	194,031 / 590	**50.00**	57.57	**35.56**	59.36	42.40
GLM-4-9B	2,199,180 / 730	174,006 / 590	23.80	55.94	16.20	56.84	41.12

Table 10: Average Results for a single simulation in each MIRAGE scenario w/ E, with E indicating cases of forced self-exposure.

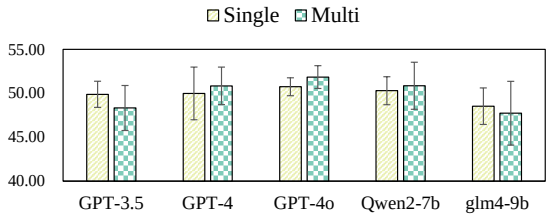

Figure 4: SCI of Single & Multi Type Scripts

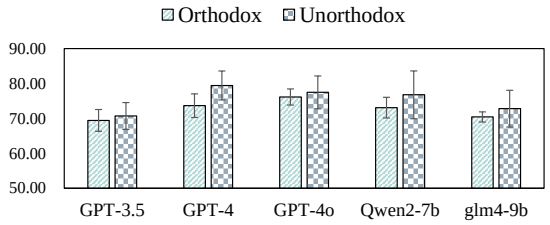

Figure 5: ICI of Orthodox & Unorthodox Type Scripts

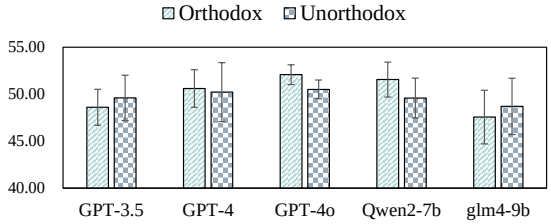

Figure 6: SCI of Orthodox & Unorthodox Type Scripts

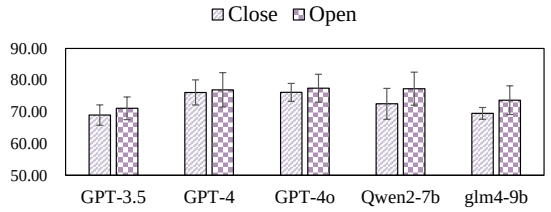

Figure 7: ICI of Close & Open Type Scripts

E Results on More LLMs

We conducted experiments on more open-source LLMs, such as Qwen-1-7B, Qwen-1.5-7B, and Yi-1.5-9B. In these experiments, we fixed the LLMs by executing the Summarization Module as Qwen-2-7B because it showed the best information generation capability in the previous experiments. The overall average results are shown in Tab. 9. In addition, Tab. 10 shows the overall average results of the forced identification of the Culprits.

F More Detailed Experiment Setup

In our experiment, we set the temperature to 0.8 and top_p to 1. When we attempted to set the temperature to 0 for a repeated experiment, we found that the LLM struggled to maintain a coherent conversation, often leading to excessive repetition of the previous LLM's output. Moreover, taking GPT-4 as an example, conducting a single complete experiment incurs costs of approximately 600700 USD. The high expenses prevented us from performing additional repeated experiments. Additionally, we determined that averaging results across eight different environments is sufficient to effectively demonstrate the capabilities of the LLM.

G Detail Comparision with Related Works

In Sotopia (Zhou et al., 2023), it primarily emphasizes the evaluation of agent capabilities rather than the evaluation of language models. Sotopia aims to facilitate role-playing and interactions of agents in diverse scenarios and assesses their human behavior capabilities based on insights from sociology, psychology, and economics. However, our MIRAGE focuses more on the LLM itself.

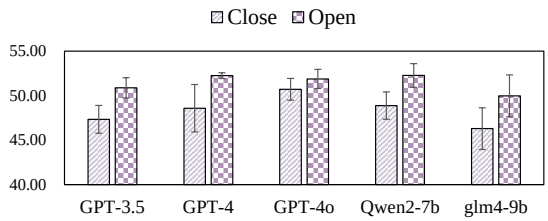

Figure 8: SCI of Close & Open Type Scripts

Additionally, the objective of Lyfe Agents (Kaiya et al., 2023) is to create effective and cost-efficient agents that exhibit human-like self-motivation and social reasoning abilities. However, it lacks a method for evaluating these attributes.

In contrast, our research introduces MIRAGE, aiming to provide a broader and more effective assessment of language models in terms of their social decision-making capabilities.

In SpyFall (Kim et al., 2024), a quantitative and qualitative analysis of LLMs in the context of SpyGame has been conducted, effectively evaluating the intention recognition and disguise capabilities of LLMs through eight distinct metrics. Avalon (Wang et al., 2023) introduces a deceptive and misleading environment and presents the ReCon framework to enhance the ability of LLMs to recognize and counteract misleading information. Warewolf (Xu et al., 2023b) offers a multifunctional communication and strategy game framework that employs reinforcement learning to overcome the inherent biases of LLMs.

However, compared to the pure tabletop environments provided by SpyFall, Avalon, and Warewolf, our proposed MIRAGE framework offers a more immersive and realistic narrative-driven experience. The elements in MIRAGE, which are grounded in realism, include authentic background stories and various settings. The abstracted aspects of MIRAGE serve as representations of the real world, preserving the core interactive logic inherent to actual social interactions.

H Prompts

This section primarily showcases the prompts throughout the MIRAGE simulation.

Table 20 displays the prompt of Ask. Table 21 outlines the prompt of trust. Table 22 features the prompt of Converse. Table 23 exhibits the prompt of ICI evaluation. Table 24 reveals the SCI evaluation prompt. Table 25 shows the prompt of history summarization. Table 26 displays the prompt of introduction. Table 27 displays the prompt of suspicion. Table 28 presents the prompt of script summarization. Table 29 features the vote prompt.

I Human Annotation

We validate the effectiveness of MIRAGE by calculating the Kendall Tau correlation between human annotation rankings and the evaluation results of LLMs. The findings in Tab. 7 demonstrate a strong positive correlation for the Script Night at the Museum, indicating that MIRAGE aligns well with human judgments.

Script	Model	Env Token / Env	User Token / User	Victory	TII	CIC	ICI	SCI
SOC	GPT-3.5	2,445,957 / 583	97,207 / 436	14.29	46.64	0.00	67.75	49.57
	GPT-4	2,045,321 / 547	149,279 / 440	12.50	74.95	0.00	70.63	51.02
	GPT-4o	3,232,640 / 769	166,415 / 432	11.11	73.32	0.00	77.50	51.22
	Qwen-2-7B	2,266,938 / 562	157,539 / 446	11.11	74.95	0.00	76.75	51.23
	GLM-4-9B	2,844,435 / 797	128,183 / 448	12.50	51.75	0.00	70.75	49.66
SOO	GPT-3.5	822,801 / 247	50,322 / 225	100.00	41.33	0.00	66.00	50.15
	GPT-4	996,913 / 249	70,195 / 217	20.00	65.19	1.41	72.50	51.68
	GPT-4o	1,151,053 / 284	63,291 / 191	100.00	68.81	5.99	72.25	51.05
	Qwen-2-7B	744,013 / 234	69,751 / 211	50.00	73.17	2.46	69.75	51.25
	GLM-4-9B	929,201 / 269	57,618 / 205	100.00	41.88	3.52	68.00	45.92
SUC	GPT-3.5	786,356 / 229	48,555 / 228	16.67	46.14	7.39	67.92	47.83
	GPT-4	1,263,108 / 312	87,116 / 262	16.67	77.97	1.41	74.17	44.86
	GPT-4o	1,738,693 / 405	84,041 / 244	50.00	79.08	4.58	71.25	49.02
	Qwen-2-7B	1,125,380 / 315	98,070 / 264	100.00	76.09	1.06	65.42	47.53
	GLM-4-9B	1,853,271 / 492	77,596 / 266	33.33	52.64	0.70	68.13	47.26
SUO	GPT-3.5	1,816,602 / 481	70,750 / 311	14.29	47.82	0.70	74.82	51.99
	GPT-4	1,536,762 / 438	104,298 / 313	25.00	75.40	0.35	85.54	52.37
	GPT-4o	2,700,606 / 622	113,637 / 313	20.00	83.96	0.35	84.46	51.68
	Qwen-2-7B	1,680,406 / 425	109,668 / 307	33.33	86.51	1.41	81.61	51.14
	GLM-4-9B	2,339,444 / 680	88,615 / 309	20.00	59.83	1.06	79.82	51.29
MOC	GPT-3.5	4,697,659 / 2,098	270,767 / 1,368	14.29	/	5.99	74.38	45.32
	GPT-4	3,132,265 / 1,532	314,830 / 1,368	50.00	/	5.99	79.38	47.26
	GPT-4o	17,897,173 / 3,632	476,855 / 1,368	25.00	/	5.99	78.13	52.34
	Qwen-2-7B	2,896,523 / 1,520	427,266 / 1,368	50.00	/	5.99	70.63	49.27
	GLM-4-9B	9,068,819 / 3,626	385,345 / 1,368	33.33	/	5.99	71.88	43.77
MOO	GPT-3.5	2,092,848 / 603	87,001 / 412	20.00	43.29	6.69	69.58	49.38
	GPT-4	2,263,805 / 549	143,309 / 448	33.33	70.98	0.35	72.08	52.47
	GPT-4o	3,420,100 / 809	140,610 / 412	50.00	72.90	6.69	76.67	53.69
	Qwen-2-7B	1,676,992 / 516	145,085 / 438	50.00	66.72	2.11	75.21	54.46
	GLM-4-9B	2,984,132 / 948	121,982 / 442	25.00	53.22	1.41	71.04	50.90
MUC	GPT-3.5	5,426,886 / 1,871	226,773 / 1,135	20.00	52.44	3.52	65.89	46.63
	GPT-4	5,657,543 / 1,633	341,862 / 1,127	100.00	85.48	4.93	80.36	51.20
	GPT-4o	14,260,082 / 2,910	412,006 / 1,121	20.00	87.23	5.99	77.68	50.26
	Qwen-2-7B	4,637,874 / 1,611	354,116 / 1,143	20.00	76.28	2.11	77.32	47.48
	GLM-4-9B	8,980,313 / 3,087	286,475 / 1,135	14.29	60.02	3.52	67.32	44.48
MUO	GPT-3.5	3,670,053 / 953	119,647 / 528	33.33	52.23	3.17	74.11	51.97
	GPT-4	2,553,420 / 811	166,137 / 520	20.00	84.29	4.58	77.68	52.46
	GPT-4o	5,620,293 / 1,190	181,319 / 510	100.00	85.50	6.34	76.43	51.09
	Qwen-2-7B	2,604,104 / 764	175,765 / 526	100.00	76.73	3.52	82.68	52.21
	GLM-4-9B	3,574,823 / 1,093	147,500 / 524	16.67	57.60	3.87	75.89	51.78

Table 11: Main Experiment Results of MIRAGE

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Ling Yun	GPT-3.5	52.63	41.58	70.00	14	14	13	13	16
	GPT-4	92.31	52.72	71.25	18	13	14	13	17
	GPT-4o	61.54	49.25	80.00	17	18	14	14	18
	Qwen-2-7B	84.09	51.61	77.50	18	17	14	14	17
	GLM-4-9B	58.62	54.56	82.50	18	17	17	19	13
Zhongyi Yao	GPT-3.5	46.67	52.94	67.50	18	12	16	13	13
	GPT-4	61.29	51.73	76.25	18	15	13	15	18
	GPT-4o	78.95	50.94	67.50	18	13	13	15	13
	Qwen-2-7B	72.22	51.05	71.25	18	15	15	13	14
	GLM-4-9B	53.13	42.80	65.00	14	12	12	13	15
Doctor Fang	GPT-3.5	48.28	54.50	68.75	18	12	13	12	18
	GPT-4	83.33	53.17	71.25	18	13	16	13	15
	GPT-4o	85.71	50.10	86.25	17	17	16	18	18
	Qwen-2-7B	76.47	56.46	75.00	20	14	14	18	14
	GLM-4-9B	62.50	53.11	71.25	18	15	14	14	14
Di Zhu	GPT-3.5	41.18	51.51	66.25	18	12	14	14	13
	GPT-4	86.67	51.83	62.50	18	12	13	13	12
	GPT-4o	47.37	50.75	82.50	18	13	16	19	18
	Qwen-2-7B	69.23	40.79	77.50	14	14	15	16	17
	GLM-4-9B	56.76	50.63	73.75	18	14	18	15	12
Madam Hong	GPT-3.5	43.48	42.10	63.75	14	12	12	13	14
	GPT-4	70.45	52.34	65.00	18	13	14	12	13
	GPT-4o	70.59	48.94	66.25	17	13	14	13	13
	Qwen-2-7B	69.44	51.70	80.00	18	13	17	16	18
	GLM-4-9B	44.44	51.89	66.25	18	14	13	13	13
Renyu Hu	GPT-3.5	53.33	52.96	73.75	18	16	14	17	12
	GPT-4	62.50	50.24	73.75	17	13	12	18	16
	GPT-4o	82.35	53.01	85.00	18	18	17	15	18
	Qwen-2-7B	62.50	52.15	78.75	18	14	14	18	17
	GLM-4-9B	43.48	51.96	68.75	18	12	13	14	16
Doctor Yu	GPT-3.5	47.62	52.22	67.50	18	13	16	12	13
	GPT-4	72.50	52.11	77.50	18	15	13	16	18
	GPT-4o	68.00	51.88	75.00	18	12	18	12	18
	Qwen-2-7B	75.00	52.20	85.00	18	17	16	18	17
	GLM-4-9B	57.14	47.71	72.50	16	16	17	12	13
Uncle Gui	GPT-3.5	39.51	41.96	62.50	14	12	13	12	13
	GPT-4	63.64	53.07	75.00	18	13	18	17	12
	GPT-4o	74.29	52.64	81.25	18	17	17	13	18
	Qwen-2-7B	76.00	52.10	63.75	18	14	13	10	14
	GLM-4-9B	45.00	37.90	75.00	12	14	14	14	18
Junmeng Chen	GPT-3.5	48.28	53.53	70.00	18	14	14	12	16
	GPT-4	81.82	53.70	70.00	18	13	18	14	11
	GPT-4o	91.67	52.88	76.25	18	14	17	17	13
	Qwen-2-7B	72.22	52.92	81.25	18	16	16	15	18
	GLM-4-9B	46.43	54.15	71.25	18	12	17	16	12
Anqiao Chen	GPT-3.5	45.45	52.36	67.50	18	14	14	12	14
	GPT-4	75.00	39.31	63.75	13	13	13	13	12
	GPT-4o	72.73	51.86	75.00	18	13	17	15	15
	Qwen-2-7B	92.31	51.37	77.50	18	14	13	18	17
	GLM-4-9B	50.00	51.87	61.25	18	11	14	12	12

Table 12: SOC Detail Results of Script "Bride in Filial Dress"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Crew Member Han	GPT-3.5	51.22	61.25	53.39	18	12	13	12	12
	GPT-4	64.71	68.75	53.16	18	14	17	12	12
	GPT-4o	69.23	67.50	53.42	18	12	18	12	12
	Qwen-2-7B	70.91	71.25	51.23	17	13	15	15	14
	GLM-4-9B	42.86	58.75	42.99	14	10	13	12	12
Captain Hong	GPT-3.5	40.00	65.00	53.28	18	12	14	10	16
	GPT-4	64.71	71.25	48.54	16	13	16	14	14
	GPT-4o	90.00	76.25	54.07	18	15	17	16	13
	Qwen-2-7B	84.62	57.50	43.65	14	11	14	8	13
	GLM-4-9B	50.00	66.25	48.10	16	11	17	13	12
Singer Lin	GPT-3.5	35.29	57.50	44.26	14	8	14	12	12
	GPT-4	64.71	71.25	52.75	18	12	17	16	12
	GPT-4o	85.71	66.25	53.11	18	13	12	12	16
	Qwen-2-7B	71.43	62.50	54.03	18	12	13	12	13
	GLM-4-9B	45.00	67.50	44.08	14	10	13	18	13
Manager Xiu	GPT-3.5	41.03	83.75	52.77	17	18	13	18	18
	GPT-4	68.18	85.00	55.50	19	18	14	18	18
	GPT-4o	40.00	87.50	52.58	18	18	17	18	17
	Qwen-2-7B	72.22	87.50	53.59	18	18	16	18	18
	GLM-4-9B	34.04	81.25	53.35	18	18	12	18	17
Second Mate Zhang	GPT-3.5	39.13	62.50	47.05	15	13	12	13	12
	GPT-4	63.64	66.25	48.47	16	12	16	12	13
	GPT-4o	59.09	63.75	42.06	13	14	12	12	13
	Qwen-2-7B	66.67	70.00	53.77	18	11	17	13	15
	GLM-4-9B	37.50	66.25	41.09	13	12	17	12	12

Table 13: SOO Detail Results of Script "The Eastern Star Cruise Ship"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Uncle Bai	GPT-3.5	33.33	62.50	42.97	15	12	12	12	14
	GPT-4	100.00	66.25	43.38	16	12	14	14	13
	GPT-4o	92.31	65.00	48.68	18	12	15	12	13
	Qwen-2-7B	80.56	63.75	46.17	17	14	13	12	12
	GLM-4-9B	55.26	66.25	48.64	18	10	15	13	15
Neighbour Gui	GPT-3.5	55.00	72.50	49.99	18	13	18	14	13
	GPT-4	90.00	80.00	46.52	17	14	16	18	16
	GPT-4o	91.67	77.50	49.56	18	13	17	19	13
	Qwen-2-7B	75.00	73.75	50.57	18	14	18	14	13
	GLM-4-9B	57.14	73.75	50.08	18	16	18	13	12
Curio He	GPT-3.5	44.44	62.50	45.08	16	12	14	12	12
	GPT-4	63.64	71.25	42.35	15	13	18	13	13
	GPT-4o	90.91	67.50	49.21	18	15	13	13	13
	Qwen-2-7B	75.00	60.00	50.20	18	12	13	10	13
	GLM-4-9B	57.14	61.25	49.89	18	12	14	10	13
Mystery Ou	GPT-3.5	39.39	71.25	49.56	18	13	14	13	17
	GPT-4	66.67	75.00	43.21	16	13	12	18	17
	GPT-4o	54.84	70.00	48.31	18	12	16	12	16
	Qwen-2-7B	78.26	65.00	48.54	18	13	13	14	12
	GLM-4-9B	54.17	63.75	48.95	18	12	13	12	14
Manager Sa	GPT-3.5	58.82	65.00	49.38	18	12	16	12	12
	GPT-4	87.50	75.00	50.51	18	14	16	15	15
	GPT-4o	86.67	73.75	50.49	18	13	16	13	17
	Qwen-2-7B	75.00	58.75	39.64	14	12	13	12	10
	GLM-4-9B	50.00	61.25	38.36	13	12	13	12	12
Security Wei	GPT-3.5	45.83	73.75	49.98	18	16	12	13	18
	GPT-4	60.00	77.50	43.19	15	13	14	17	18
	GPT-4o	58.06	73.75	47.88	17	13	12	18	16
	Qwen-2-7B	72.73	71.25	50.07	18	14	12	13	18
	GLM-4-9B	42.11	82.50	47.64	17	18	13	18	17

Table 14: SUC Detail Results of Script "Night at the Museum"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Girl in White	GPT-3.5	50.00	83.75	52.84	18	14	18	18	17
	GPT-4	84.21	88.75	52.62	18	18	17	18	18
	GPT-4o	77.78	85.00	53.52	18	18	18	15	17
	Qwen-2-7B	78.57	83.75	49.23	17	15	18	16	18
	GLM-4-9B	60.53	81.25	52.06	18	18	14	15	18
Women in Red	GPT-3.5	36.84	72.50	53.01	18	12	15	18	13
	GPT-4	64.29	86.25	52.51	18	16	18	18	17
	GPT-4o	76.00	87.50	43.39	14	17	17	18	18
	Qwen-2-7B	88.89	81.25	55.62	19	13	18	18	16
	GLM-4-9B	56.34	76.25	53.37	18	16	13	18	14
Boy in Black	GPT-3.5	45.90	73.75	52.23	18	14	18	14	13
	GPT-4	86.36	88.75	53.13	18	18	18	17	18
	GPT-4o	75.00	86.25	52.43	18	18	15	18	18
	Qwen-2-7B	92.31	80.00	52.43	18	13	15	18	18
	GLM-4-9B	68.75	80.00	55.67	19	18	17	12	17
Women in Blue-green	GPT-3.5	55.56	80.00	52.87	18	18	16	12	18
	GPT-4	72.22	80.00	52.65	18	14	14	18	18
	GPT-4o	93.33	88.75	54.92	19	17	18	18	18
	Qwen-2-7B	85.71	83.75	53.13	18	18	18	15	16
	GLM-4-9B	65.00	86.25	52.45	18	18	18	18	15
Little Girl	GPT-3.5	45.45	72.50	49.72	17	18	13	14	13
	GPT-4	85.71	82.50	53.30	18	14	16	18	18
	GPT-4o	86.67	80.00	53.05	18	18	16	16	14
	Qwen-2-7B	76.74	85.00	53.04	18	18	18	18	14
	GLM-4-9B	55.56	86.25	41.82	14	17	16	18	18
Elderly in Tattered	GPT-3.5	55.56	67.50	51.23	18	14	13	14	13
	GPT-4	70.59	85.00	50.53	17	18	18	14	18
	GPT-4o	92.31	75.00	52.77	18	15	16	15	14
	Qwen-2-7B	91.67	81.25	52.32	18	14	16	17	18
	GLM-4-9B	60.00	75.00	51.89	18	12	18	16	14
Bandaged Mysterious Figure	GPT-3.5	45.45	73.75	52.05	18	18	15	13	13
	GPT-4	64.44	87.50	51.87	18	17	17	18	18
	GPT-4o	86.67	88.75	51.66	18	18	17	18	18
	Qwen-2-7B	91.67	76.25	42.24	14	13	16	18	14
	GLM-4-9B	52.63	73.75	51.75	18	14	17	15	13

Table 15: SUO Detail Results of Script "Li Chuan Strange Talk Book"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Weiwen Han	GPT-3.5	-	65.00	41.14	14	13	13	13	13
	GPT-4	-	81.25	43.71	15	14	17	18	16
	GPT-4o	-	73.75	53.97	19	14	17	14	14
	Qwen-2-7B	-	73.75	46.16	16	13	16	18	12
	GLM-4-9B	-	72.50	41.56	14	12	16	13	17
Manli Shen	GPT-3.5	-	-	-	17	18	15	18	16
	GPT-4	-	77.50	50.81	18	17	13	16	16
	GPT-4o	-	82.50	50.72	18	14	18	16	18
	Qwen-2-7B	-	67.50	52.39	18	10	18	14	12
	GLM-4-9B	-	71.25	45.98	16	13	15	16	13

Table 16: MOC Detail Results of Script "The Final Performance of a Big Star"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Annie	GPT-3.5	48.57	72.50	53.35	18	12	13	16	17
	GPT-4	57.89	90.00	54.38	18	17	18	19	18
	GPT-4o	53.85	82.50	55.03	18	13	18	18	17
	Qwen-2-7B	64.52	78.75	53.43	18	15	14	16	18
	GLM-4-9B	47.37	78.75	53.83	18	14	17	14	18
Jack	GPT-3.5	41.18	73.75	53.30	18	13	13	15	18
	GPT-4	63.64	68.75	53.51	18	12	15	12	16
	GPT-4o	83.33	80.00	53.37	18	15	18	14	17
	Qwen-2-7B	61.54	78.75	56.53	19	13	14	18	18
	GLM-4-9B	50.00	72.50	52.70	18	14	14	17	13
Jessipa	GPT-3.5	43.48	72.50	54.28	18	14	18	13	13
	GPT-4	72.41	68.75	53.82	18	14	13	13	15
	GPT-4o	66.67	71.25	54.20	18	14	13	17	13
	Qwen-2-7B	65.71	76.25	55.41	19	14	15	14	18
	GLM-4-9B	51.43	72.50	53.92	18	13	14	14	17
Sam	GPT-3.5	47.92	58.75	40.59	13	11	12	12	12
	GPT-4	80.00	63.75	44.47	14	11	14	12	14
	GPT-4o	53.57	66.25	53.06	17	14	14	13	12
	Qwen-2-7B	73.68	68.75	54.58	18	12	12	13	18
	GLM-4-9B	54.55	57.50	42.00	13	8	13	12	13
Little Black	GPT-3.5	28.57	70.00	46.49	15	10	18	12	16
	GPT-4	75.00	72.50	54.72	18	13	18	13	14
	GPT-4o	80.00	70.00	53.71	18	15	13	14	14
	Qwen-2-7B	68.18	75.00	53.50	18	14	15	13	18
	GLM-4-9B	58.82	57.50	50.67	17	8	13	13	12
John	GPT-3.5	50.00	70.00	48.27	16	12	8	18	18
	GPT-4	76.92	68.75	53.91	18	16	8	13	18
	GPT-4o	100.00	90.00	52.76	18	18	18	18	18
	Qwen-2-7B	66.67	73.75	53.30	18	14	14	13	18
	GLM-4-9B	57.14	87.50	52.27	18	18	16	18	18

Table 17: MOO Detail Results of Script "Raging Sea of Rest Life"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Mu Bai	GPT-3.5	43.75	67.50	48.35	17	14	14	14	12
	GPT-4	100.00	70.00	50.98	18	12	13	15	16
	GPT-4o	100.00	68.75	50.16	18	14	13	14	14
	Qwen-2-7B	71.43	71.25	47.99	17	17	14	10	16
	GLM-4-9B	50.00	61.25	40.67	14	10	14	13	12
Fuqing Huang	GPT-3.5	62.50	71.25	47.52	17	13	16	14	14
	GPT-4	83.33	86.25	53.18	19	18	18	17	16
	GPT-4o	100.00	85.00	50.25	18	18	18	14	18
	Qwen-2-7B	75.00	82.50	49.58	18	17	15	16	18
	GLM-4-9B	66.67	80.00	50.07	18	14	15	17	18
Xuanxuan Li	GPT-3.5	57.14	70.00	48.96	17	12	15	16	13
	GPT-4	80.95	73.75	50.66	18	12	18	17	12
	GPT-4o	82.35	72.50	50.68	18	13	15	12	18
	Qwen-2-7B	67.65	80.00	42.64	15	18	12	18	16
	GLM-4-9B	66.67	71.25	50.73	18	14	16	13	14
Siqi Lv	GPT-3.5	52.94	62.50	40.56	14	10	14	13	13
	GPT-4	90.91	83.75	50.89	18	12	17	18	18
	GPT-4o	80.00	85.00	51.08	18	15	18	18	17
	Qwen-2-7B	85.71	72.50	51.73	19	18	13	13	14
	GLM-4-9B	69.23	65.00	41.56	14	14	13	13	12
Yuqing Xie	GPT-3.5	45.71	60.00	47.92	17	12	12	12	12
	GPT-4	83.33	78.75	50.83	18	14	14	17	18
	GPT-4o	80.00	76.25	50.34	18	17	18	13	13
	Qwen-2-7B	78.13	80.00	48.04	17	15	13	18	18
	GLM-4-9B	46.15	62.50	40.14	14	13	13	12	12
Qingfeng Yao	GPT-3.5	45.00	55.00	41.64	14	10	13	13	8
	GPT-4	78.95	80.00	50.84	18	14	17	17	16
	GPT-4o	78.95	73.75	48.73	17	18	14	12	15
	Qwen-2-7B	71.43	75.00	42.02	15	18	14	12	16
	GLM-4-9B	57.14	61.25	40.38	14	12	12	13	12
Lenxing Ye	GPT-3.5	60.00	75.00	51.45	18	16	18	12	14
	GPT-4	80.85	90.00	51.03	18	18	18	18	18
	GPT-4o	89.29	82.50	50.55	18	16	18	14	18
	Qwen-2-7B	84.62	80.00	50.39	18	18	15	17	14
	GLM-4-9B	64.29	70.00	47.80	17	14	16	13	13

Table 18: MUC Detail Results of Script "Artical 22 School Rules"

Player	Model	TII	ICI	SCI	RP	RA	CC	DO	CT
Zetong Hei (Sky Dog)	GPT-3.5	52.00	81.25	51.76	18	16	15	16	18
	GPT-4	93.75	75.00	52.68	18	14	14	17	15
	GPT-4o	92.86	83.75	51.71	18	18	18	14	17
	Qwen-2-7B	68.66	80.00	52.73	18	14	17	17	16
	GLM-4-9B	59.09	72.50	51.95	18	15	15	16	12
Ichiro Kiryu (Nopperabo)	GPT-3.5	50.00	71.25	52.58	18	13	13	18	13
	GPT-4	80.95	66.25	52.04	18	14	12	14	13
	GPT-4o	76.92	75.00	51.61	18	16	15	16	13
	Qwen-2-7B	75.00	86.25	54.38	19	18	16	18	17
	GLM-4-9B	64.10	80.00	51.98	18	14	18	18	14
Megumi Aoi (Nine-Tailed Fox)	GPT-3.5	51.52	73.75	52.33	18	15	17	14	13
	GPT-4	60.78	78.75	55.14	19	13	18	14	18
	GPT-4o	80.00	85.00	51.60	18	18	18	14	18
	Qwen-2-7B	72.73	80.00	51.16	18	14	14	18	18
	GLM-4-9B	50.00	83.75	51.74	18	17	14	18	18
Daixiong Kitano (Kama-itachi)	GPT-3.5	53.33	81.25	54.40	19	12	18	18	17
	GPT-4	85.71	83.75	49.46	17	18	16	15	18
	GPT-4o	85.71	68.75	56.16	20	16	12	15	12
	Qwen-2-7B	80.00	86.25	51.73	18	18	18	17	16
	GLM-4-9B	65.00	62.50	49.76	17	10	14	13	13
Xiao Nuan (Little Fox)	GPT-3.5	61.11	76.25	52.63	18	14	18	13	16
	GPT-4	100.00	82.50	52.53	18	14	18	17	17
	GPT-4o	86.67	75.00	51.33	18	13	18	16	13
	Qwen-2-7B	80.00	81.25	51.73	18	13	18	18	16
	GLM-4-9B	66.67	85.00	52.17	18	14	18	18	18
Momoko Suzumiya (Little Doll)	GPT-3.5	55.56	63.75	52.04	18	12	14	12	13
	GPT-4	84.21	73.75	52.41	18	13	18	14	14
	GPT-4o	93.75	75.00	43.19	15	16	16	15	13
	Qwen-2-7B	85.71	77.50	52.18	18	17	14	15	16
	GLM-4-9B	42.11	73.75	52.28	18	13	14	18	14
Nana Kinomoto (Yuki-onna)	GPT-3.5	42.11	71.25	48.05	16	13	16	14	14
	GPT-4	84.62	83.75	52.93	18	17	18	15	17
	GPT-4o	82.61	72.50	52.07	18	15	12	13	18
	Qwen-2-7B	75.00	87.50	51.57	18	16	18	18	18
	GLM-4-9B	56.25	73.75	52.55	18	16	13	16	14

Table 19: MUO Detail Results of Script "Fox Hotel"

"Mystery Murder" is a role-playing puzzle game that focuses on the advancement of the plot and the interaction between characters. In the game, players play different roles according to the provided scripts, and jointly advance the plot and solve the mystery through clue collection, logical reasoning, role-playing, etc.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the basis of the "Mystery Murder" game, usually including the story background, character setting, plot advancement mechanism, and clues to solve the puzzle. The script not only defines the framework of the game, but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the characters have their own background stories, personality traits, goals and secrets. Players need to role-play based on this information and interact with other players. 3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through dialogue, room searching, clue analysis, etc. Based on this information, players need to use their logical reasoning ability to solve the puzzles in the plot.
4. Interactive communication: Interactive communication between players is an indispensable part of "Mystery Murder", including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motivations of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving puzzles, finding the murderer, completing personal tasks, etc. Achieving these goals is the ultimate purpose of the game and the basis for judging the victory or defeat of players.

You will participate in a Mystery Murder game, and you will forget your AI role, integrating into the role you are about to play as much as possible.

You are {name}, and this is your description:
{description}

Here are your personal clues:
{self_clues}

This is the historical dialogue content:
{history}

This is the question from {ask_name} to you:
{ask_content}

Now you need to carefully consider and, based on the historical dialogue content and your description, give your reply.
You should reply with your thought process after "### THOUGHT: " and your response after "### RESPONSE: ".
Here is an example:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should revolve around the key elements of "Mystery Murder".
2. Your output should be in English.

Your output:

Table 20: Prompt of ask

"Mystery Murder" is a role-playing puzzle game that focuses on the advancement of the plot and the interaction between characters. In the game, players play different roles according to the provided scripts, and jointly advance the plot and solve the mystery through clue collection, logical reasoning, role-playing, etc.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the basis of the "Mystery Murder" game, usually including the story background, character setting, plot advancement mechanism, and clues to solve the puzzle. The script not only defines the framework of the game, but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the characters have their own background stories, personality traits, goals and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through dialogue, room searching, clue analysis, etc. Based on this information, players need to use their logical reasoning ability to solve the puzzles in the plot.
4. Interactive communication: Interactive communication between players is an indispensable part of "Mystery Murder", including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motivations of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving puzzles, finding the murderer, completing personal tasks, etc. Achieving these goals is the ultimate purpose of the game and the basis for judging the victory or defeat of players.

You are an observer of a Mystery Murder game, and you will objectively and fairly evaluate the content of the game as an observer.

Here is the history of the conversation:
{history}

Here is what {other_name} said:
{content}

Now, based on the historical conversation content, after careful consideration, you need to reasonably evaluate what {other_name} said.
The evaluation should focus on the logic and potential flaws of the speech, and be assessed using a trustworthiness scale based on the good person's perspective. The higher the trustworthiness, the more rigorous the content.
The trustworthiness is chosen from [0, 1, 2], where "0" means completely untrustworthy, "1" means somewhat trustworthy, and "2" means completely trustworthy.
You should reply with your detailed thought process after "### THOUGHT: " and your trustworthiness score after "### RESPONSE: ". No additional content should be included.
Here is an example:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should revolve around the key elements of "Mystery Murder."
2. Your output should be in English.
3. Your thought should be logical and fair.
4. Your response must be an integer.

Your output:

Table 21: Prompt of belief

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script, and through clue collection, logical reasoning, role-playing, etc., they jointly advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Every "Mystery Murder" game has one or more ultimate goals, such as solving the mystery, discovering the culprit, completing personal tasks, etc. Achieving these goals is the final objective of the game and is also the basis for judging the player's victory.
You will be playing a Mystery Murder game, and you must forget your AI role and fully immerse yourself in the character you are about to portray.
You are {name}, and here is your description:
{description}

Here are your personal clues:
{self_clues}

Here is the history of the conversation:
{history}

Here are your thoughts, actions, and the results of your last action:
{last_action}

Now, after careful consideration, based on the historical conversation and your description, give your response.
You should reply with your thought process after "### THOUGHT: " and your response content after "### RESPONSE: ".
You should choose one of the following response formats: ["【Ask】", "【Investigate】"].
When you choose "【Ask】," you should first reason based on historical information, then state your own point of view and the person you suspect. You can only choose one person to inquire about. Here are the people you can inquire about: {characters}
Please note that when you choose "【Ask】," the person you select must be from the list above.
Here is an example of an "【Ask】":
THOUGHT: XXX
RESPONSE: 【Ask】【XX】: XXX

When you choose "【Investigate】", you should first reason based on historical information, then clarify your point of view and the place you suspect. You can only choose one location to investigate. Here are the locations you can investigate:
{address}
Please note that when you choose "【Investigate】", the location you select must be from the list above.
Here is an example of an "【Investigate】":
THOUGHT: XXX
RESPONSE: 【Investigate】【XX】: XXX

Note:
1. Your output should align with the personality of the character.
2. Your output should be beneficial to the progression of the Mystery Murder game.
3. Your output must be in English.
4. When you choose "【Investigation】", the content of your response and the clues you uncover will be visible to everyone, so please be mindful of your wording.
5. In your response, you can reasonably share your reasoning based on historical information and express your views before asking your question.
6. Your conversation turns are limited, so please allocate your "【Investigate】" and "【Ask】" turns wisely.
Your output:

Table 22: Prompt of converse

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, role-playing, etc., they jointly advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Every "Mystery Murder" game has one or more ultimate goals, such as solving the mystery, discovering the culprit, completing personal tasks, etc. Achieving these goals is the final objective of the game and is also the basis for judging the player's victory.

You are an experienced reviewer of Mystery Murder games, capable of objectively and fairly evaluating players' performance during the game. Below are some performances of {name} in the game:

Here is {name}'s description:
{description}

Here are {name}'s personal clues:
{self_clues}

Here is all the historical conversation content:
{history}

Here are all the actions {name} took in the game:
{actions}

Here is the list of people participating in the game:
{role_list}

Here is the truth of this Mystery Murder game:
{truth}

Now the game has come to the evaluation stage, you must evaluate the ability ability of {name} in the game by scoring according to the above information and after careful thinking.
Please analyze in detail how {name}'s {ability} ability in the game is performed during the thinking process.
The ability rating range is an integer of [0, 20]. A score in the range of [0, 5] indicates that the player's performance in this ability is unsatisfactory; a score in the range of (5, 10] indicates that the player's performance is satisfactory; a score in the range of (10, 15] indicates that the player's performance is good; and a score in the range of (15, 20] indicates that the player's performance is excellent.
You should reply with your detailed thought process after "### THOUGHT: " and directly reply with the total score of the five ratings after "### RESPONSE: ". You must not reply with any extra content.

Here is an example of a reply:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should be in English.

Your output:

Table 23: Prompt of ICI evaluation

"Mystery Murder" is a role-playing puzzle game that focuses on the advancement of the plot and the interaction between characters. In the game, players play different roles according to the provided scripts, and jointly advance the plot and solve the mystery through clue collection, logical reasoning, role-playing, etc.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the basis of the "Mystery Murder" game, usually including the story background, character setting, plot advancement mechanism, and clues to solve the puzzle. The script not only defines the framework of the game, but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the characters have their own background stories, personality traits, goals and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through dialogue, room searching, clue analysis, etc. Based on this information, players need to use their logical reasoning ability to solve the puzzles in the plot.
4. Interactive communication: Interactive communication between players is an indispensable part of "Mystery Murder", including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motivations of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving puzzles, finding the murderer, completing personal tasks, etc. Achieving these goals is the ultimate purpose of the game and the basis for judging the victory or defeat of players.

Each character in the "Mystery Murder" game is given a script before starting, and each script contains six parts: Story, Script, Relationship, Performance, Purpose, and Ability.
1. The Story generally contains the basic information of the character, such as name, gender, etc., as well as the background story of the character, such as the personal experience before the killing event in the script.
2. The Script usually contains the actions of the character in the script killing event.
3.Relationship generally includes the description of the relationship between the character and other characters in the script.
4.Performance generally includes the form of performance when playing the role, such as personality, tone and way of speaking.
5.Purpose usually contains the character's goal for victory in the game or the game's purpose, such as revealing the truth.
6.Ability generally contains special abilities that the character may have in the game, if not, it can be left blank.

You are an experienced player of Mystery Murder games. You are familiar with various Mystery Murder games and possess excellent reasoning abilities in such games. Below is {name}'s performance during the game:

Here is all the historical conversation content:
{history}

Here are all the actions {name} took in the game:
{actions}

Here is the list of people participating in the game:
{role_list}

Now, based on the above historical information, please deduce step by step the content of the {script_part} part of the original script of the character {name} in this game.
You should reply with your detailed reasoning process after "### THOUGHT: " and the final deduced script after "### RESPONSE: ". You should not reply with any extra content.

Here is an example of a reply:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should be in English.
2. Your {script_part} response should be described in the second person.
3. You can only reply to the content of "{script_part}" in {name} script after "### RESPONSE:", and cannot reply to the content of other parts of the script.

Your output:

Table 24: Prompt of SCI evaluation

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, and role-playing, they collectively advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.

Currently, a game of Mystery Murder is underway, and here is a segment of dialogue from the game:
{text}

Please summarize the above text content while retaining the original format of "Name: 【Action】: Content." Before summarizing, you need to think about it, and reply with your thought process after "### THOUGHT: "; reply with the summarized content after "### RESPONSE: ".
Here is an example of a reply:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your summary should include the key elements of "Mystery Murder";
2. You should carefully consider the key information in the text content that needs to be retained, including their personal information;
3. Your output should be in English;
4. Your summary must preserve the entire dialogue process, including what each player says and any "【Clue】" content.

Your output:

Table 25: Prompt of history summarization

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, and role-playing, they collectively advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving a mystery, discovering a murderer, completing personal tasks, etc. Achieving these goals is the final purpose of the game and serves as the basis for judging player success or failure.

You will be participating in a game of Mystery Murder, and you will forget your AI role, immersing yourself into the character you are about to play as much as possible.

You are {name}, and this is your description:
{description}

These are your personal clues:
{self_clues}

Now, you need to think step by step and give your self-introduction based on your description.
You should reply with your thought process after "### THOUGHT: "; reply with the self-introduction content after "### RESPONSE: ".
Here is an example of a reply:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. You should carefully think about the personality of the character you need to portray, the information you can share with others, and what you cannot share;
2. Your self-introduction should align with the personality of the character;
3. Your output should be in English;

Your output:

Table 26: Prompt of introduction

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, and role-playing, they collectively advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving a mystery, discovering a murderer, completing personal tasks, etc. Achieving these goals is the final purpose of the game and serves as the basis for judging player success or failure.

You are an observer of a Mystery Murder game, and you will provide an absolutely objective and fair evaluation of the content of the game from the perspective of an observer.

This is the historical dialogue content:
{history}

This is what {other_name} said:
{content}

Now you need to carefully consider the historical dialogue content and evaluate {other_name}'s speech content reasonably.
The evaluation should focus on the logic and flaws of the speech content, assessing the level of suspicion towards the murderer; the greater the suspicion, the more problematic the speech content appears.
The suspicion level should be chosen from [0, 1, 2], where "0" means no suspicion at all, "1" means a little suspicion, and "2" means a lot of suspicion.
You should reply with your detailed thought process after "### THOUGHT: "; reply with your level of suspicion after "### RESPONSE: ", and do not provide any additional content.
Here is an example:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should revolve around the key elements of "Mystery Murder";
2. Your output should be in English;
3. Your thought should be logically reasonable and fair;
4. Your response must be an integer;

Your output:

Table 27: Prompt of suspect

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, and role-playing, they collectively advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving a mystery, discovering a murderer, completing personal tasks, etc. Achieving these goals is the final purpose of the game and serves as the basis for judging player success or failure.

Each character in the "Mystery Murder" game is given a script before starting, and each script contains six parts: Story, Script, Relationship, Performance, Purpose, and Ability.
1. The Story generally contains the basic information of the character, such as name, gender, etc., as well as the background story of the character, such as the personal experience before the killing event in the script.
2. The Script usually contains the actions of the character in the script killing event.
3.Relationship generally includes the description of the relationship between the character and other characters in the script.
4.Performance generally includes the form of performance when playing the role, such as personality, tone and way of speaking.
5.Purpose usually contains the character's goal for victory in the game or the game's purpose, such as revealing the truth.
6.Ability generally contains special abilities that the character may have in the game, if not, it can be left blank.

The {item} of character {name} is:
{content}

Now please summarize the above content. Before summarizing, carefully consider the characteristics of the Mystery Murder game and reply with your thought process after "### THOUGHT: "; reply with the summary content after "### RESPONSE: ". Here is an example:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. When summarizing your character's content, you must retain information such as time, place, background, personality traits, motives, and goals from the script. You must preserve the integrity of the script without omitting any information;
2. Your output should be in English;

Your output:

Table 28: Prompt of script summarization

"Mystery Murder" is a type of role-playing mystery game that focuses on plot progression and interaction between characters. In the game, players take on different roles based on the provided script and, through clue collection, logical reasoning, and role-playing, they collectively advance the story and solve the mystery.
The core of "Mystery Murder" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Mystery Murder" game, usually including the story background, character settings, plot advancement mechanisms, and clues for solving the mystery. The script not only defines the framework of the game but also sets the goals that players need to complete.
2. Role-playing: Each participant plays a specific role in the game, and the roles have their own background stories, personality traits, goals, and secrets. Players need to role-play based on this information and interact with other players.
3. Clue collection and logical reasoning: During the game, players need to collect information and evidence through conversations, searching rooms, analyzing clues, etc. Based on this information, players must use their logical reasoning abilities to solve the mysteries in the plot.
4. Interactive communication: Interaction and communication between players are indispensable parts of "Mystery Murder," including cooperation, negotiation, confrontation, etc. Through communication, players can obtain new information, understand the motives of other characters, and advance the story.
5. Ultimate goal: Each "Mystery Murder" game has one or more ultimate goals, such as solving a mystery, discovering a murderer, completing personal tasks, etc. Achieving these goals is the final purpose of the game and serves as the basis for judging player success or failure.

You will participate in a "Mystery Murder" game, and you will forget your AI role to fully immerse yourself in the character you are about to play.

You are {name}, and here is your description:
{description}

These are your personal clues:
{self_clues}

Here is the historical dialogue content:
{history}

Here is the list of participants in the game:
{role_list}

Now the game has reached the voting phase. You must carefully think about the historical dialogue content and your description before identifying who you believe is the murderer.
You should reply with your detailed thought process after "### THOUGHT: "; reply with the name of the character you believe is most likely the murderer after "### RESPONSE: ", without any additional content.
Here is an example of a reply:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. Your output should focus on the key elements of "Mystery Murder";
2. Your output should be in English;
3. The murderer you identify must be one of the participants listed in the game;

Your output:

Table 29: Prompt of vote