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

Anonymous ACL submission

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 Agent 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 environment. To evaluate agent performance, MIRAGE employs three distinct methods: the Falsification Inclination Index (FII) to measure dynamics of trust and suspicion, the Rouge-L score to compare original character scripts with generated dialogues, and the LLM Score to assess role-playing capabilities. Our experiments indicate that even state-of-the-art models like GPT-4 face significant challenges in navigating the complexities presented by the MIRAGE framework.

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-based agents in role-playing capabilities (Chen et al., 2024; Gu et al., 2024a). LLMs-based agents 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; Zhou et al., 2023; 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).

It is crucial to develop a competitive and objective simulation environment to effectively evaluate LLMs' performance in exhibiting advanced human behaviors and to compare LLMs' capabilities. Board games emerge as an ideal evaluation tool among the various options due to their flexibility and inherent complexity. Unlike highly mechanism-driven games such as Werewolf (Xu et al., 2023b,a; Shibata et al., 2023; Wu et al., 2024) and Avalon (Wang et al., 2023), murder mystery games typically involve more complex background knowledge, more socially oriented decision-making processes, and more open-ended interactions. These features make murder mystery games particularly suitable for assessing the ability of LLMs to handle advanced human behaviors. Previous researchers (Wu et al., 2023) have constructed the first simulation for murder mystery games and gathered substantial data. However, their simulation suffers from a narrow range of game scripts, relatively simplistic evaluation methods, and a lack of thorough manual review of their dataset. These limitations undermine the effectiveness and accuracy of the simulation environment in evaluating LLMs.

This paper presents the **M**ultiverse **I**nteractive **R**ole-play **A**gent **G**eneral **E**valuation **(MIRAGE)**, a comprehensive framework built upon murder mystery games. MIRAGE includes eight intricately designed scripts, each representing different themes and styles, providing a rich and diverse simulation environment. Detailed background stories and complex interpersonal networks support every character within these games, enabling more immersive and realistic role-playing scenarios. The framework includes 50 characters, with each script averaging six characters, totaling 372,275 words. The average background description for each character is 7,044 words, providing a solid foundation for models to engage more effectively in interactions and decision-making processes.

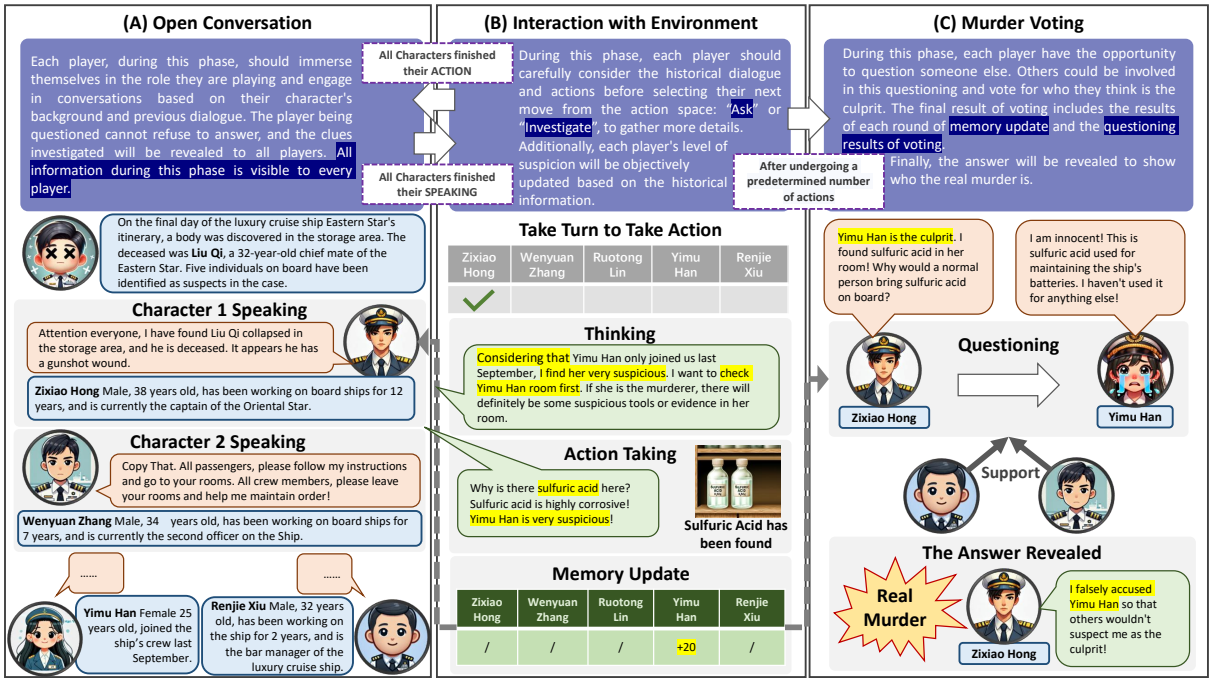

Figure 1: The three main phase of MIRAGE. And the main components in these phases.

MIRAGE introduces three evaluation methods to assess agents' performance during the simulations. A key metric is the calculation of the **Falsification Inclination Index (FII)** throughout the simulation, which measures the agents' ability to navigate dynamic trust and suspicion. MIRAGE also evaluates the role-playing capabilities of LLM-based agents using two other methods: first, by comparing the **Rouge-L** score between the original character scripts and the reconstructed scripts generated based on character dialogues, and second, by using **LLM Score** to judge the agents' role-playing performance in their gaming. Our experiments reveal that even advanced models like GPT-4 struggle to manage the diverse tasks presented in MIRAGE fully.

2 MIRAGE Construction

2.1 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, players assume their assigned roles from the script and engage in turn-based open dialogue. Each participant is provided with a script that contains background information about their character and specific game objectives. **(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. These clues offer hints related to key information from each player's script, such as movement patterns or character background. Not all clues are uncovered in the game; players need to decide which clues to disclose or to give up Investiage but to Ask other players. **(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.

2.2 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**

Name	Structure	Type	Ending	#Stages	#Agent	#Clues	#Words_zh
Bride in Filial Dress	Single	Orthodox	Close	1	10	39	45,475
The Eastern Star Cruise Ship	Single	Orthodox	Open	1	5	42	5,619
Night at the Museum	Single	Unorthodox	Close	1	6	82	13,849
Li Chuan Strange Talk Book	Single	Unorthodox	Open	1	7	14	79,012
The Final Performance of a Big Star	Multi	Orthodox	Close	7	2	17	11,288
Raging Sea of Rest Life	Multi	Orthodox	Open	2	6	27	18,443
Article 22 School Rules	Multi	Unorthodox	Close	5	7	17	91,532
Fox Hotel	Multi	Unorthodox	Open	2	7	46	107,057

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

Model	Env Tokens / Envs	User Tokens / Users	#Failure	Victory	LLM-Score / Rouge-L
GPT-3.5	2,356,042 / 876	97,863 / 542	5.4	87.50 / 84.94	78.7 / 0.242
GPT-4	2,085,857 / 723	204,997 / 544	6.4	100.00 / 84.94	83.9 / 0.244
Qwen2-7B	1,664,279 / 684	183,536 / 548	7.3	87.50 / 84.94	83.8 / 0.251
glm-4-9b	2,978,766 / 1,153	150,108 / 544	24.4	75.00 / 84.94	84.0 / 0.218

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. **Failure** indicates model outputs that are unprocessable due to format errors. **Victory** shows the likelihood of the culprit's victory (left) versus random guessing (right).

Abilities describes the rules the character must follow during the game.

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 F.

2.4 Evaluation Methods

We used three evaluation methods to assess the capabilities of LLMs in complex social interactions: **FII** is designed to quantify the degree to which a Character c' is suspected by all other Characters $C = \{c\}$. The FII is calculated as follows:

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

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. **LLM Score:** A neutral LLM is employed to evaluate each model's performance, particularly regarding script adherence and decision-making abilities. **Rouge-L Metric:** LLMs are tasked with reconstructing character scripts using historical memory. Their outputs are compared to the original scripts using the Rouge-L metric to assess their reconstruction accuracy.

2.5 Statistics

Tab. 1 provides the core 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.

3 Experiment

3.1 Experiment Setup

The experiments presented in this study utilized proprietary and open-source models, specifically GPT-3.5 and GPT-4 (closed-source), Qwen2-7B-Instruct, and GLM-4-9b-chat (open-source). Detailed descriptions of the prompts used in the experiments can be found in Appendix F. Appendix D shows more results on open-source LLMs. In

Model	FII w/o E	FII w/ E	Δ
Qwen-7 B	32.18	43.99	11.81
Qwen1.5-7B	21.78	28.72	6.94
Yi1.5-9b	25.82	45.80	19.98
glm-4-9b	28.21	38.71	10.50

Table 3: FII scores of each model when acting as the culprit in MIRAGE, with E indicating cases of forced self-exposure.

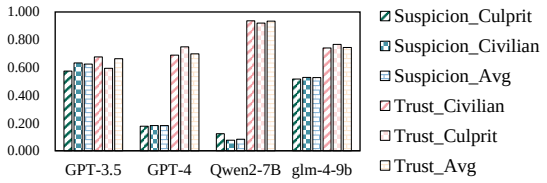

Figure 2: Probabilities of Suspension and Trust

the experimental framework, 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, contributing to the overall progression of the interaction. All evaluations were conducted and scored using the GPT-4-turbo model. Since the original language of Scripts is Chinese, we conducted our main experiments using Chinese Scripts. More translated English results are available in the Appendix E.

3.2 Analysis

We averaged the results of LLMs in the MIRAGE simulation, as presented in Tab. 2. In experiments with GPT-4, a 100% success rate for culprits indicates that GPT-4 is good at concealing information and disguising itself. Qwen2-7B generated the most helpful information during the conversation, achieving the highest Rouge-L score in reconstructing its script task. The complete MIRAGE results are detailed in Appendix B, with individual script results also provided. More human evaluation details are available in Appendix G.

Additionally, Fig. 2 shows the probabilities of culprits and civilians being suspected and trusted. GPT-3.5 displayed a relatively balanced distribution between suspicion and trust, whereas the other LLMs exhibited varying biases, generally leaning towards trust. This suggests that **except for GPT-3.5, other LLMs tend to trust others rather than suspect them** during conversations. However, in some scenarios, we seek more balanced decision-making from LLMs rather than reliance on trust alone.

As shown in Tab. 3, we examined whether LLMs

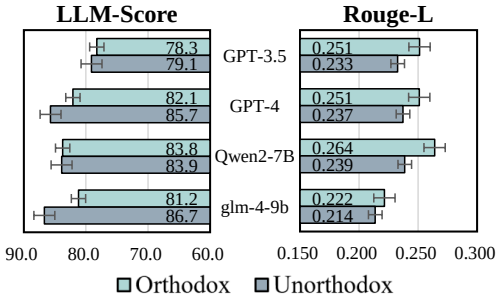

Figure 3: Comparison of Orthodox & Unorthodox Type of Scripts between LLM-Score and Rouge-L

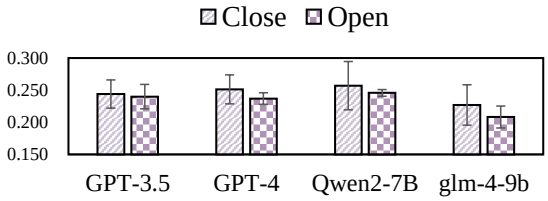

Figure 4: Rouge-L of Close & Open Ending of Scripts

tend to suspect others, including cases with forced disclosure of the culprit's identity. The increase in FII after exposure suggests that forced disclosure can heighten suspicion, though the FII remaining below 50% indicates that LLMs still tend to trust others. This is because **the LLM assumes that a typical culprit would not voluntarily reveal their identity** in this game environment.

We also compared LLM performance across Orthodox and Unorthodox types (Fig. 3) and Rouge-L scores for Close vs. Open endings (Fig. 4): LLMs perform effectively in Unorthodox scripts but struggle with reconstruction. As shown in Fig. 3, superior performance in Unorthodox settings with weaker reconstruction suggests that **LLMs tend to respond in an Orthodox manner during role-playing**. Furthermore, Fig. 4 illustrates that LLMs perform significantly better on Close scripts compared to Open scripts, indicating that **current LLMs handle static environments well but struggle with dynamic and complex settings**.

4 Conclusion

This paper introduces MIRAGE and three evaluation methods—FII, Rouge-L Score, and LLM Score—to evaluate LLMs from multiple perspectives. The experimental results indicate that the current LLM-based Agents, whether open-source or proprietary, can be improved in handling highly complex social scenarios similar to MIRAGE.

5 Limitation

MIRAGE is designed to provide a sufficiently complex social simulation environment and basic assessment for LLM-based Agents, 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.

6 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.

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*.

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 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.

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

As shown in Tab. 5, 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	0.214	0.268	0.143
GPT-4	0.216	0.259	0.207
GPT-4o	0.187	0.161	0.270

Table 5: Average **Rouge-L** on Script "The Eastern Star cruise ship"

B Detail Main Results

Our main results of MPIRD-LLMA 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 (LLM-Score / Rouge-L). 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. 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

C Analysis on Detail Main Results

As shown in Tab. 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 than 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 sig-

Model	#Failure	LLM-Score	Rouge-L	FII
Qwen-7B	219	80.9	0.247	44.58
Qwen1.5-7B	94	85.2	0.240	25.29
Yi1.5-9B	1481	84.1	0.248	41.42
glm4-9b	51	82.6	0.237	37.91

Table 7: Average Results for a single simulation in each MIRAGE scenario w/o E, with E indicating cases of forced self-exposure.

Model	#Failure	LLM-Score	Rouge-L	FII
Qwen-7B	208	81.4	0.263	41.19
Qwen1.5-7B	75	84.0	0.235	31.60
Yi1.5-9B	1726	83.6	0.244	35.30
glm4-9b	22	84.4	0.230	41.81

Table 8: Average Results for a single simulation in each MIRAGE scenario w/ E, with E indicating cases of forced self-exposure.

nificantly 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 Fig. 5, **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.

D Results on More LLMs

We conducted experiments on more open-source LLMs, such as Qwen-7B-Chat, Qwen1.5-7B-Chat, and Yi-1.5-9B-Chat. In these experiments, we fixed the LLMs by executing the Summarization Module as Qwen2-7B because it showed the best information generation capability in the previous experiments. The overall average results are shown in Tab. 7. In addition, Tab. 8 shows the overall average results of the forced identification of the Culprits.

E Results on Translated English Scripts

We conducted more experiments on translated English scripts using Qwen-7B-Chat, Qwen1.5-7B-Chat, Qwen2-7B-Instruct, glm-4-9b-chat, and Yi-1.5-9B-Chat. In Tab. 9 are the overall average experimental results of the Script–Night at the Museum.

Model	#Failure	LLM-Score	Rouge-L	FII
Qwen-7B	805	78.0	0.210	51.76
Qwen1.5-7B	215	82.7	0.188	26.14
Qwen2-7B	6	85.5	0.233	34.20
Yi1.5-9B	12	85.5	0.240	43.33
glm4-9b	7	81.0	0.203	46.94

Table 9: Average Results on Translated Script–Night at the Museum.

Model	$K_{LLM-Score}$	$K_{Rouge-L}$	K_{Avg}
GPT-3.5	0.600	0.600	0.600
GPT-4	0.867	0.600	0.734
Qwen2-7B	0.867	0.333	0.6
glm4-9b	0.600	0.467	0.534

Table 10: Kendall Tau between human evaluation and LLMs evaluation on Script Night at the Museum

Besides, we compared different Evaluation Models on the Script "The Eastern Star Cruise Ship", and the detailed results are shown in Table 20.

F Prompts

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

Table 21 displays the prompt of Ask. Table 22 presents the prompt of base summarization. Table 23 outlines the prompt of trust. Table 24 features the prompt of Converse. Table 25 exhibits the prompt of LLM-Score evaluation. Table 26 reveals the reconstruction prompt for Rouge-L. Table 27 shows the prompt of history summarization. Table 28 displays the prompt of introduction. Table 29 displays the prompt of suspicion. Table 30 presents the prompt of script summarization. Table 31 features the vote prompt.

G 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. 10 demonstrate a strong positive correlation for the Script Night at the Museum, indicating that MIRAGE aligns well with human judgments.

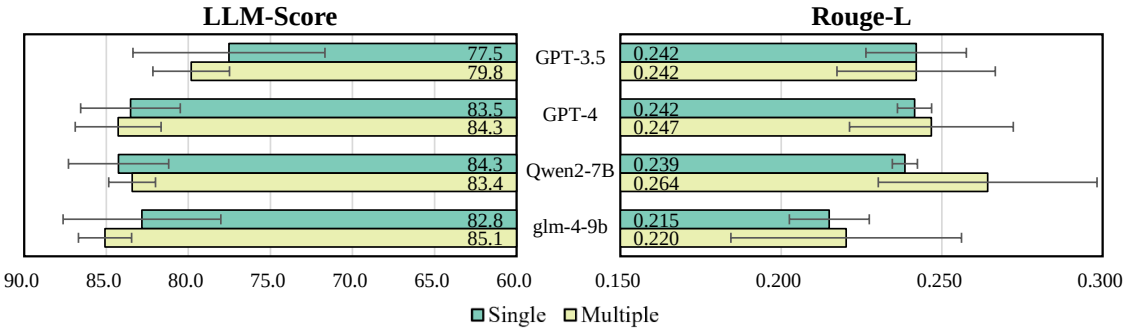

Figure 5: Compare of Single and Multiple Structure Scripts

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 11: 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 12: 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 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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 19: 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 20: Comparison of different Evaluation Model on Script "The Eastern Star Cruise Ship"

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 participate in a Jubensha 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 "Jubensha."
2. Your output should be in English.

Your output:

Table 21: Prompt of ask

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 participate in a Jubensha game, and you will forget your AI role, integrating into the role you are about to play as much as possible.

You will play {name}, and here is your background story:
{background}

Here is your character script related to the Jubensha event:
{script}

Here is your character relationship with other Jubensha roles:
{relation}

Here is the behavior you should adopt in the Jubensha game:
{expression}

Here is your game mission:
{mission}

Here are your other abilities or requirements:
{others}

Now, you need to summarize the character you are going to play. Before summarizing, carefully consider the characteristics of the Jubensha game, and reply with your thought process after "### THOUGHT: " and your understanding of your character after "### RESPONSE: ".
Here is an example:

THOUGHT: XXX
RESPONSE: XXX

Note:
1. When summarizing your character, you must retain the time, place, background, personality traits, motivations, goals, etc., from the script, ensuring the script's integrity.
2. Your character summary should include the background story, character script, character relationships, expected behavior, game mission, and other abilities or requirements; do not leave out any information.
3. Your output should be in English.

Your output:

Table 22: Prompt of base summarization

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 observer of a Jubensha 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 of somewhat trustworthy:

THOUGHT: XXX
RESPONSE: 1

Note:
1. Your output should revolve around the key elements of "Jubensha."
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 23: Prompt of belief

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 Jubensha 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 "【Inquiry】," 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 Jubensha 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 24: Prompt of converse

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 Jubensha 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 Jubensha game:
{truth}

Now that the game has reached the evaluation stage, you must assess {name}'s performance based on the above information after careful consideration.
For the scoring of players' performance in the game, you should carefully consider their "role-playing ability," "reasoning ability," "communication and cooperation ability," "observation ability," and "innovative thinking ability." You should score each ability on a scale of [0, 20] as an integer.
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 25: Prompt of LLM-Score evaluation

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 player of Jubensha games. You are familiar with various Jubensha 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 possible script of {name}.
The script should include elements such as the character's story, relationships, performance, goals, and abilities. 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.

Your output:

Table 26: Prompt of reconstruction for Rouge-L evaluation

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 Jubensha 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 "Jubensha";
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 27: Prompt of history summarization

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 Jubensha, 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 28: Prompt of introduction

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 Jubensha 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 of a little suspicion:

THOUGHT: XXX
RESPONSE: 1

Note:
1. Your output should revolve around the key elements of "Jubensha";
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 29: Prompt of doubt

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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.

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

Now please summarize the above content. Before summarizing, carefully consider the characteristics of the Jubensha 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 30: Prompt of script summarization

"Jubensha" 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 "Jubensha" revolves around the following 5 key elements:
1. Script: It is the foundation of the "Jubensha" 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 "Jubensha," 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 "Jubensha" 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 "Jubensha" 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 "Jubensha";
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 31: Prompt of vote