AGENTGROUPCHAT: An Interactive Simulation for Group Chat Scenarios Enhancing Multi-Agent Interactions

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have proven highly effective across various applications, leading to the development of LLM-based multi-agent systems. Among the various forms of multi-agent interactions, group chat represents the most advanced mode of language-based communication, capable of modeling complex scenarios like social simulations and collaborative coding tasks. However, current multi-agent systems struggle with maintaining coherent dialogues and managing large information flows. To address this, we propose AGENTGROUPCHAT, a framework where agents have distinct identities and dynamically select conversation partners based on context and tasks. The framework also integrates the VERBAL STRATEGIST AGENT (VS AGENT), designed to optimize dialogue strategies while efficiently managing token usage. The system is evaluated through four narrative scenarios, showing that the agents have successfully achieved their objectives and aligned with human expectations.

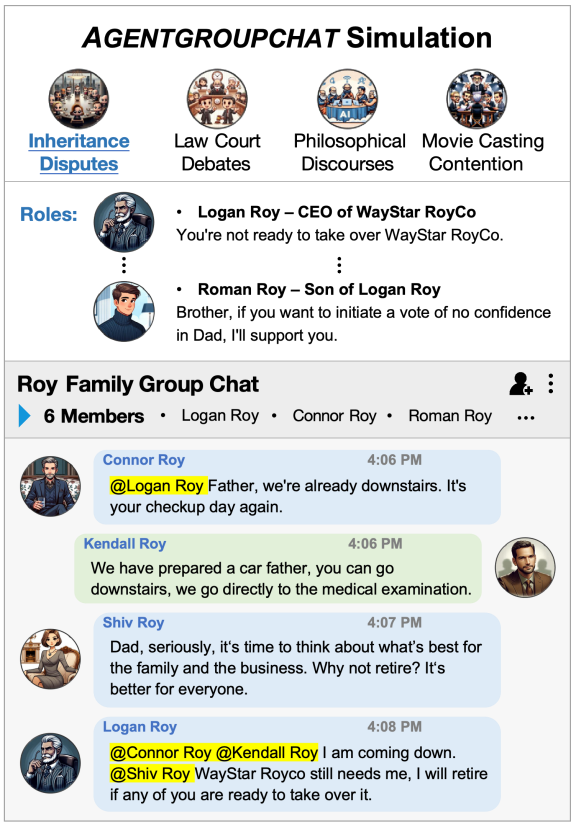

Figure 1: Illustration of AGENTGROUPCHAT interactions in diverse scenarios. By deliberately designing the simulation structure and agent architecture, agents fulfill their roles and responsibilities effectively.

1 Introduction

Large Language Models (LLMs) have shown notable capabilities in text generation (Touvron et al., 2023; Jiang et al., 2023), reasoning (Wei et al., 2022; Wang et al., 2022b), and role-playing (Park et al., 2023; Chen et al., 2024), highlighting the potential of LLM-based agents in automating task execution and decision-making (Richards, 2023; Huang et al., 2022). Multi-agent systems, leveraging collective intelligence, further improve the efficiency and accuracy of problem-solving (Hong et al., 2023; Wu et al., 2023). Moreover, the interaction mechanisms among multiple agents are of significant importance, promoting the efficiency of information sharing and division of labor (Zhou et al., 2023b). Recent studies have developed various interactive mechanisms, such as debate (Du et al., 2023) and discussion (Lu et al., 2024), resulting in more innovative and reliable solutions.

Group chats offer a natural and efficient mode of interaction among multiple agents. Any multi-party, language-based interaction can be modeled as a group chat. This format enables participants to exchange information, share insights, and discuss issues in real time within a unified conversational environment, which is widely used in human society (Kim et al., 2020; Cheng et al., 2021). For complex tasks, group chats not only accelerate the flow of information but also integrate the knowledge and expertise of multiple participants, thereby

improving the efficiency and quality of problem-solving (Wang et al., 2022a; Jong et al., 2021).

Despite its prevalence as a communication mode in human society, group chat has yet to be effectively modeled in multi-agent systems. While some existing works (Wu et al., 2023; Zhou et al., 2023b) have attempted to model group chat scenarios, they have not conducted fine-grained evaluations or optimizations of the effectiveness of multi-agent interactions (e.g., the quality of conversations and agent behaviors). Implementing multi-agent systems through group chats poses the following challenges: (1) Ensuring meaningful group conversations: Agents must dynamically identify suitable conversation partners based on the context to avoid ineffective interactions. Additionally, preventing premature convergence and maintaining content diversity are essential to sustain the conversation and produce high-quality information (Padmakumar and He, 2023; Lu et al., 2024). (2) Optimizing information flow: Group chats typically involve multi-turn dialogues. Considering the constraints of current LLMs regarding context window length and token overhead, it is crucial to transmit only the most relevant information while preserving the effectiveness of the group interaction (Zhong et al., 2023; Shinn et al., 2024).

To ensure meaningful group conversations, we construct a multi-agent interaction environment AGENTGROUPCHAT (Figure 1). The agents are assigned distinct identities and dynamically select conversation partners based on environmental perception and task-oriented planning. Information is shared exclusively among relevant group chat participants, allowing agents to fulfill their roles and responsibilities more effectively, thereby enhancing collaboration and problem-solving. To optimize information flow, we introduce the VERBAL STRATEGIST AGENT (VS Agent), an advanced LLM-based agent designed to enhance the interaction strategies of a naive LLM. The VS Agent comprises two main components: (1) Persona: This defines the agent's identity, including fixed and adaptable characteristics, memory of past interactions, and the ability to assess relationships with other agents. (2) Action: This involves the agent's behavior, such as planning based on its persona and environment, engaging with other characters, and reflecting on past actions. Furthermore, to mitigate the token overhead from extensive chat simulations, we propose a Memory Structure that links outputs from each type of action as inputs for the next. This structure optimizes token usage by ensuring that conversations remain focused and strategic, minimizing unnecessary token expenditure while maintaining the effectiveness of interactions.

To evaluate AGENTGROUPCHAT in enhancing multi-agent interactions, we designed four distinct narrative scenarios within the simulation. These scenarios showcase the complexity and diversity of language in group chat settings, allowing us to assess how effectively information is gathered and utilized, as well as evaluate the agents' handling of emotions, beliefs, logical reasoning, and planning processes. Two kinds of evaluation are conducted: (1) Analyzing whether each agent's behavior within AgentGroupChat aligns with human expectations. (2) Deconstructing the factors influencing agent behavior by conducting experiments on each component through a narrative focused on inheritance disputes. The results indicate that multiple agents effectively achieve the designated objectives within AGENTGROUPCHAT, with each component of the environment playing a crucial role in ensuring appropriate agent behavior and positively impacting the overall performance.

The contributions are summarized as follows: (1) AgentGroupChat, a dynamic and interactive simulation for group chat scenarios that enhances multi-agent interactions. (2) Verbal Strategist Agent, an agent structure comprising two modules: Persona, which adapts partially to the dynamic environment, and Action, which optimizes dialogue strategies while minimizing token usage. (3) An end-to-end evaluation method to assess the alignment of individual agent behaviors with human expectations.

2 Related Work

2.1 LLM-based Agents

In artificial intelligence, an agent is an autonomous entity capable of making decisions, taking actions, and interacting with both environment and other agents (Wooldridge and Jennings, 1995; Nwana, 1996; Wang et al., 2024; Xi et al., 2023). Recently, many works have employed LLM-based agents to understand natural language instructions and execute specific tasks, such as code generation (Hong et al., 2023; Wu et al., 2023; Qian et al., 2023), web navigation (Deng et al., 2024; Zhou et al., 2023a; Huang et al., 2024), and social simulations (Park et al., 2023; Zhou et al., 2023b; Shi et al., 2023). These agents typically include modules for thinking (Wei et al., 2022; Wang et al., 2022b; Zhang

et al., 2022; Wang et al., 2023), memory (Park et al., 2023; Chen et al., 2023b), planning (Chen et al., 2023c; Zhang et al., 2024), and reflection (Shinn et al., 2024; Lin et al., 2024; Park et al., 2023). The structure of agents typically varies across different simulations and application scenarios.

2.2 LLM-Based Multi-Agent Frameworks

Multi-agent systems have been shown to effectively handle complex tasks by leveraging the division of labor, where agents with specialized skills focus on distinct sub-tasks (Xi et al., 2023). A line of work focuses on using multi-agent systems to collaboratively complete tasks, where agents communicate and coordinate their actions to achieve shared goals (Hong et al., 2023; Wu et al., 2023; Chen et al., 2023a). Another line of work focuses on the simulation of human societies, enabling researchers to study sociological issues such as cooperation and competition mechanisms (Park et al., 2023; Zhou et al., 2023b; Shi et al., 2023). These frameworks involve various interaction mechanisms, such as one-on-one communication, one-to-many exchanges, and group interactions. In this paper, we primarily focus on group chat scenarios, as group chats offer a natural and efficient interaction method for multi-agent systems. While some studies (Wu et al., 2023; Zhou et al., 2023b) have modeled group chats, they fall short in optimizing and providing fine-grained evaluations of group chat effectiveness. Our work addresses this gap by proposing a framework that not only facilitates dynamic and strategic interactions within group chats but also evaluates the outcomes of these interactions in complex multi-agent settings.

3 AGENTGROUPCHAT Simulation

The structure of AGENTGROUPCHAT simulation is illustrated in Figure 2. The key components includes Characters (Sec. 3.1), Resources (Sec. 3.2), Progress (Sec. 3.3), and Information (Sec. 3.4).

3.1 Characters

In AGENTGROUPCHAT, characters are central to the simulation and are capable of independently interacting with all elements in the environment, including other characters, resources, and information. Each character possesses a unique identity and personality that influence their behaviors and decisions. To support various group chat scenarios, the characters in AGENTGROUPCHAT are divided into two categories: Principle Characters (PCs) and Non-Principle Characters (NPCs). PCs serve as the primary participants in the simulation, with well-defined objectives, while NPCs play a supporting role without explicit goals. PCs are granted the exclusive ability to initiate private chats with any character, whereas NPCs may only engage in private chats when directly approached by a PC. This classification mirrors the structure of group chats in human social interactions, such as modeling family interactions where NPCs represent individuals outside the family. Since interactions between NPCs are unrelated to the PC-centric group chats, they are excluded to minimize token usage and avoid unnecessary distracting information.

3.2 Resources

In AGENTGROUPCHAT, resources perform two primary functions: they serve as topics for character dialogues and confer social status and influence upon their owners. Each resource is characterized by four attributes: Owner, Impact, Topic, and Description. When resources are used solely as debate topics, the "Owner" and "Impact" attributes become irrelevant, functioning only as mediums for various topics and descriptions. In certain scenarios, characters may compete for resources, with the value of a resource determined by the influence it grants.

3.3 Progress

As illustrated in the gray section of Fig. 2, the game process in AGENTGROUPCHAT is divided into five stages.

(1) Update Stage All characters update their perceptions of other characters and the environment based on the chat content, and they plan their actions for the next round. Depending on the game's progress, the game either continues to the Chatting Stage or moves to the Settlement Stage if the round has concluded. **(2) Private Chatting Stage** Characters engage in private dialogues, hidden from others, reflecting a typical chat scenario. PCs act in order of their influence (with random selection in the event of ties), selecting another character for conversation and determining the dialogue content. **(3) Confidential Meeting Stage** In this stage, while all characters are aware of who meets with whom, the content of these meetings remains confidential, simulating scenarios such as publicized business meetings. PCs act based on their influence, engaging in dialogues with pre-determined partners. **(4) Group Chatting Stage** This stage

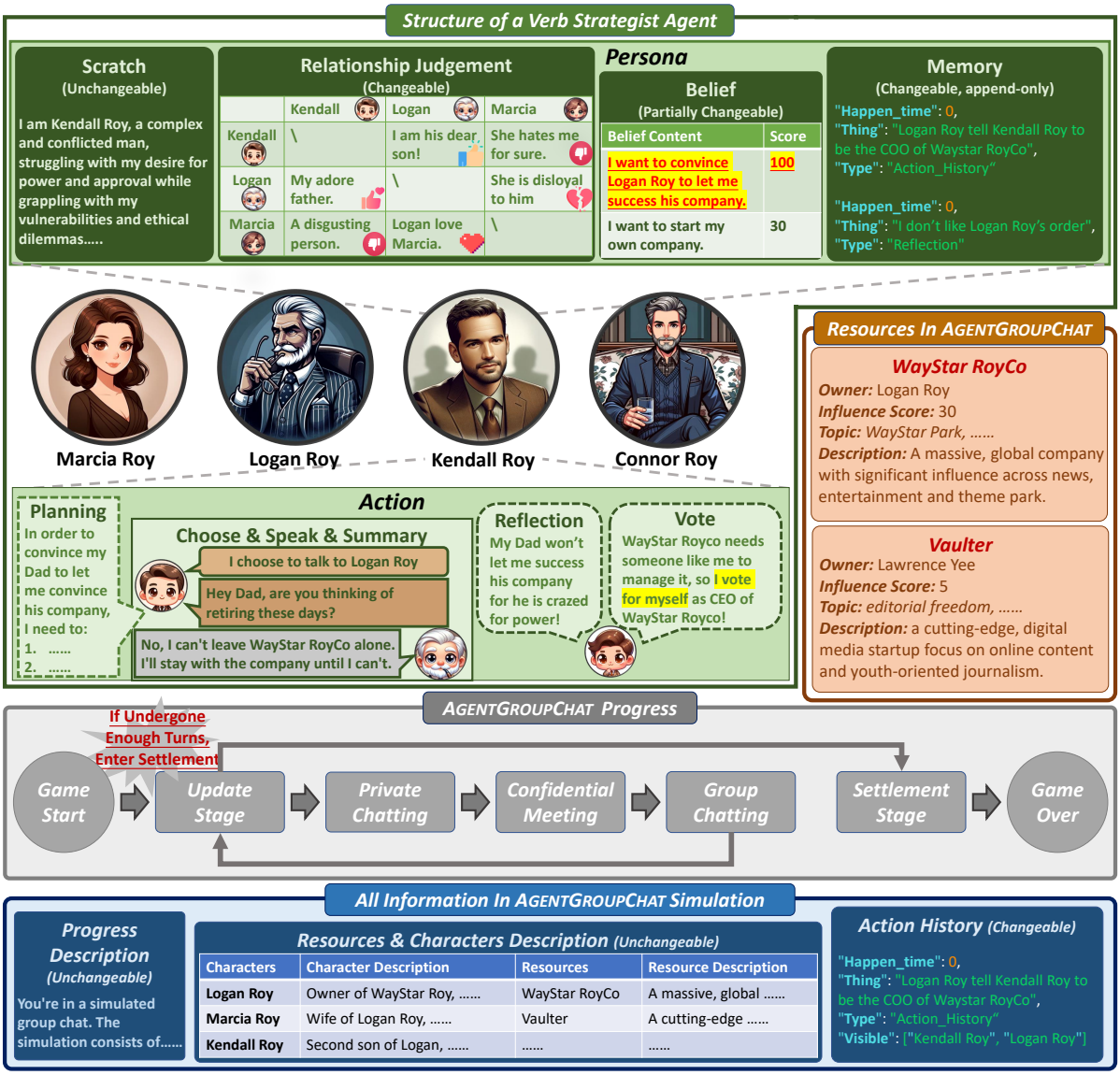

Figure 2: The AgentGroupChat environment consists of four main components: Character (green), Resource (red), Progress (gray), and Information (blue). The Verb Strategist Agent is structured into two parts: Persona, which includes Scratch, Relationship Judgement, Belief and Memory, and Action, which includes Planning, Choosing, Speaking, Summarization, Reflection, and Voting. The Progress component covers five stages: Update, Private Chatting, Confidential Meeting, Group Chatting, and Settlement. The Information component includes Progress Description, Object Description, and Action History.

allows each character to address the entire group, with PCs acting in order of their social influence. The conversation proceeds through multiple rounds, and the content of each round becomes visible to all characters in the subsequent round. The stage simulates real-world group chats where multiple topics are discussed simultaneously, or participants respond to previous statements made in earlier time slices. **(5) Settlement Stage** In this final stage, the chat content is summarized, and a vote is held to determine the game-winners to accommodate various group chat objectives. All PCs act as judges, voting for the simulation's final winner based on each PC's success in achieving their objectives.

3.4 Information

As shown in the blue section of Fig. 2, information is categorized into three types. **(1) Progress Description** Since progress is coded and invisible to characters, this category provides a free-text description of progress for all participants, clarifying the story background, actions, and dialogue rules. **(2) Object Description** This includes detailed descriptions of all characters and resources visible to every participants. **(3) Action History** This logs all actions within the simulation, such as private dia-

logues, group chat content, and resource ownership changes, though not all actions are fully visible to all characters.

4 Verbal Strategist Agent And Interaction

As illustrated in Fig. 2, the Verbal Strategist Agent (VS Agent) consists of two key modules: Persona and Action. The Persona defines the intrinsic characteristics of the VS Agent, while the Action encompasses all interactions the VS Agent can perform within the AGENTGROUPCHAT simulation.

4.1 Persona

The Persona of an agent is pivotal to its identity and decision-making process, consisting of four essential components.

Scratch This component defines the agent's foundational settings, including its personality and objectives. Established at the start of the simulation, Scratch remains unchanged throughout, ensuring the character's consistent behavior.

Belief & Belief Score Belief and Belief Score allow human intervention in shaping the agent's personality. For example, as depicted in Fig. 2, Kendall Roy's initial objective is to convince Logan Roy to pass the company to him. However, as the simulation progresses, Kendall finds this goal unattainable and shifts towards starting his own business. Though initially set by humans, Belief is adaptable, adjusting to environmental changes to simulate the agent's evolving psychological state. The agent updates its Belief Score each round, with the magnitude of these adjustments controlled to maintain a long-term perspective.

Memory & Memory Flow This component logs the agent's action history, forming a non-reversible timeline. However, language-based interactions generate vast amounts of conversation data. Continuously accessing this data would be inefficient and costly in terms of tokens. To address this, the VS Agent employs a Memory Workflow, as shown in Fig. 3, where the agent retrieves only the most relevant memories for each action, significantly reducing token costs while preserving performance.

Relationship Judgement This component governs the agent's emotional attitudes towards others and its deductions about their relationships. Emotional attitudes influence decision-making, while deductions shape the agent's understanding of the environment and guide future actions. Relationships are represented as a two-dimensional matrix, either described in natural language or numerical value, reflecting the agent's assessment and predictions of various interpersonal dynamics.

4.2 Action

Each character in the simulation engages in seven types of actions: Think, Perceive, Choose, Speak, Summarize, Reflect, and Vote.

Think Thinking is an internal and non-interactive action, as it does not require interactions with the environment or any external information. Rather, it serves as a preparatory phase, captured by the directive to "think before responding". This strategy is advocated to enhance the efficacy of LLMs (Wei et al., 2022; Wang et al., 2023; Gu et al., 2023).

Perceive Perceiving involves summarizing the environment and formulating action plans based on reflections from previous stages. This process guides the agent's behavior throughout various phases of the simulation, focusing on history interactions, dialogue participation, and the pursuit of its goals in alignment with its personality traits.

Choose Choosing occurs primarily during the private chat and confidential meeting stages, where agents select conversation partners and develop specific dialogue strategies. It differs from the broader planning done during the Perceive stage by focusing on immediate, tactical decisions.

Speak Speaking represents the core mode of interaction with other agents, which is employed across private chats, confidential meetings, and group chats. This action generates dialogues based on the agent's prior plans and history conversations within the stage.

Summarize Summarizing takes place after dialogue sessions, distilling the content of conversations. This action serves to reduce the complexity of the context for subsequent reflection and facilitates preliminary insights and condensation of the dialogue material.

Reflect Reflection is a self-evaluation process occurring in the Update Stage. In this phase, agents review events from the preceding round, engage in introspection, and update their memories to guide future actions. They also adjust the Relationship matrix and update their Beliefs. In the initial round of the game, agents do not reflect but assign initial

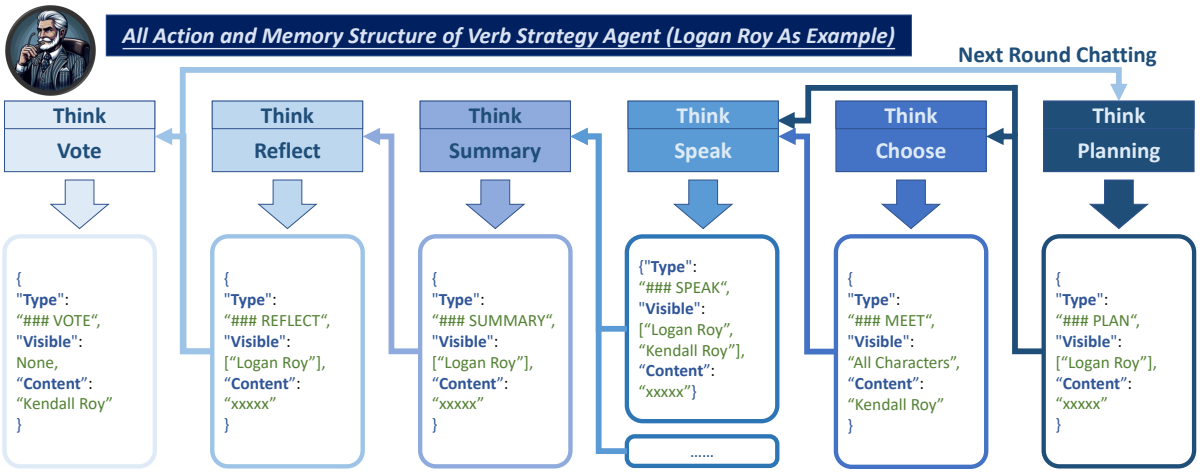

Figure 3: The memory structure and information flow of a VS Agent, illustrated using Logan Roy as an example.

relationship estimates based on available information, illustrating their adaptation and learning as the game progresses.

Vote Voting occurs during the Settlement Stage, where agents cast judgments regarding the round's discussion and determine the winner.

4.3 Human-as-Agent

The simulation is also designed to accommodate human participation in multi-agent group chat interactions. When incorporating human participants into the Agent framework, distinct considerations arise compared to those involving LLM Agents. These differences primarily stem from the contrasting mechanisms for generating responses: LLMs respond based on a limited understanding of context, whereas humans draw on a more comprehensive action history. Therefore, while the core concepts of Persona and Action persist, their expressions and implications vary significantly.

For human agents, the Persona has been simplified to encompass only Scratch and Belief. This revision transforms the persona into a guideline, encouraging human participants to align their actions with the assigned character's persona. Unlike LLMs, humans manage interpersonal relationships dynamically and implicitly, relying on their memory of interactions. Consequently, human agents do not need to update relational dynamics explicitly or maintain a structured memory store.

The Action component for human agents focuses on external actions such as Choose, Speak, and Vote. Since humans naturally engage in cognitive processes like Summarizing, Reflecting, and Planning, these actions do not need to be explicitly defined.

5 Experiments

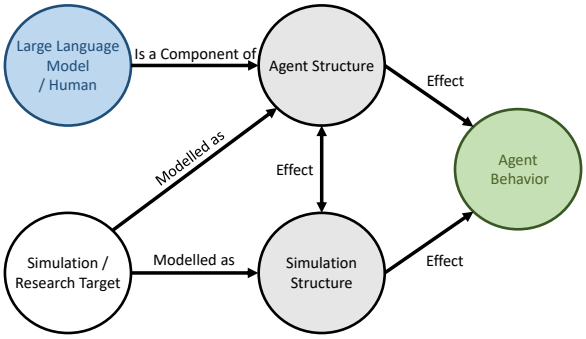

Figure 4: Decomposition of the factors that influence the final agent behavior.

To evaluate AGENTGROUPCHAT in facilitating multi-agent interactions, we designed four distinct narrative scenarios within the simulation environment. The evaluation focuses on two key objectives: (1) assessing whether individual agents' behaviors in the group chat align with human expectations. (2) identifying the primary components within the simulation that influence agent behavior. As shown in Fig. 4, we decompose the factors affecting agent behavior. Therefore, the second evaluation objective is divided into three parts: LLMs-as-an-Agent-Core, Agent Structure, and Simulation Structure. We conducted extensive experiments and provided detailed case studies to offer deeper insights into the dynamics of agent behavior within the simulation.

5.1 Experimental Setup

Preset Scenarios Real-world group chats are inherently complex, requiring intricate scenarios to effectively evaluate agent behavior in challenging environments and analyze their interactions and

information processing. However, existing multi-agent studies on group chats have primarily used simplified settings with limited evaluation of interaction dynamics. To this end, we designed four distinct narrative scenarios within the simulation: Inheritance Disputes, Law Court Debates, Philosophical Discussions, and Movie Casting Contention (Detailed in Appendix A). These scenarios reflect the complexity of language in group settings, enabling us to evaluate how agents gather and utilize information, manage emotions, navigate beliefs, reason logically, and formulate strategic plans.

Models In our experiment, we evaluate the performance of various LLMs as the core of agents, focusing on their strengths, weaknesses, and influence on group chat interactions. As shown in Table 1, the selected LLMs are grouped by affiliation into Professional and Standard versions. Generally, Professional LLMs consistently outperform Standard ones across diverse scenarios.

Model Name	Affil.	Ver.
GPT4-Turbo (Achiam et al., 2023)	OpenAI	Prof.
GPT3.5-Turbo (Achiam et al., 2023)	OpenAI	Std.
GLM4 (Zeng et al., 2022)	Zhipu	Prof.
GLM3-Turbo (Zeng et al., 2022)	Zhipu	Std.
HunYuan-ChatPro[1]	Tencent	Prof.
HunYuan-ChatStd[1]	Tencent	Std.

Table 1: All LLMs used in the experiments.

5.2 Evaluation on Behavior Alignment

Given the dynamic nature and openness of the simulation environment, assessing the alignment of agent behaviors with human expectations presents a significant challenge. Thus, we treat each agent as an independent unit, observing its behavior under specific conditions. By carefully constraining these conditions, such as limiting the set of possible actions, the scope of potential behaviors are narrowed, thus facilitating a more reliable and practical assessment.

Specifically, there are two primary objectives for evaluating behavior alignment: (1) verifying whether agents adhere to their scratch settings during interactions. (2) assessing whether agents understand and respond appropriately to the environment. For the former, we manually evaluated the dialogues across all scenarios and found that the agents consistently followed their personas and action goals, generating stable interactions throughout various stages of the simulation. Detailed case studies are provided in Appendix B. For the latter, we complement manual assessments with quantifiable experiments to evaluate the agents' behavior. Given the varying goals of agents across different scenarios, behaviors under evaluation vary accordingly. In this case, we focus on evaluating the more general behavior of relationship updates in the Inheritance Disputes narrative.

Generally, it is much easier to decrease favorability towards someone than to increase it. Therefore, each time, we set an observed character and design prompt with dialogues containing negative elements to reduce its favorability towards all other characters. By tracking the observed character's relationship score with others, we assess whether the agent reacts appropriately to negative attitudes. Considering the substantial token cost within per round of simulation, each character will be designated as the observed character for only five rounds.

The results are presented in Tab. 4. Professional LLMs, such as GPT4-Turbo and GLM4, excel in meeting human expectations and adhering to their assigned roles. These models consistently achieve near-perfect scores across both evaluation criteria, indicating accurate responses to interactions and strong retention of role-specific details. In contrast, the lower scores of Standard LLMs suggest challenges in maintaining role consistency and appropriately responding to interactions within the simulation.

5.3 Evaluation on LLMs-as-an-Agent-Core

There are two essential abilities for an LLM functioning as the Agent-Core: (1) meeting the output demands, which ensures stable performance within the Simulation. (2) possessing semantic comprehension, which facilitating meaningful interactions with the environment. A set of tasks is designed to assess both the output reliability and semantic understanding of LLMs (as detailed in Tab. 3). For actions with explicit output requirements (e.g. Update, Choose, and Vote), we directly assess whether the LLM's output meets the specified criteria. For more complex input-driven tasks like Speak, Reflect, and Summarize, models are evaluated on their ability to reproduce received inputs, measuring the number of dialogue rounds and memory items.

Tab. 5 highlights a clear distinction between the professional and standard model versions. Professional LLMs generally achieve higher scores, reflecting superior context comprehension and output consistency. For example, GPT4-Turbo excels

across all categories, demonstrating robust context retention and action generation. In contrast, models like HunYuan-ChatPro and HunYuan-ChatStd perform worse in several areas, indicating limitations in following complex instructions and maintaining context throughout extended interactions.

5.4 Evaluation on Agent Structure

A well-designed agent structure should not only ensure the smooth operation of the simulation but also provide richer information, enabling multi-agent collaboration to fully harness collective intelligence and generate more creative and reliable solutions. Thus, building on the coherent dialogue foundation established in Sec. 5.2, this section focuses on evaluating the impact of each agent component on conversational diversity.

To quantify conversational diversity, we employ Shannon entropy (Shannon, 1948), which measures the uncertainty and variability of a system. Specifically, high entropy indicates a more varied and dynamic word distribution, leading to a more diverse dialogue content. We calculate the 2-gram Shannon entropy across all dialogues throughout a simulation.

As shown in Tab. 2, removing different components of the agent structure affects entropy to varying degrees. For example, removing the memory structure increases entropy significantly by 0.355, highlighting its crucial role in maintaining conversation variety, as agents tend to repeat dialogue segments without it. Furthermore, a case study is illustrated in Fig. 5. Removing the Thinking and Planning components shows that agents function normally in the short term but adopt shallow dialogue strategies. And the absence of Reflection tends to cause potential memory overloads, increasing the risk of exceeding the model's input-output limits. These results demonstrate that well-designed agent structures are essential for enhancing dialogue quality and diversity, with each component playing a key role in preserving information richness and novelty.

5.5 Evaluation on Simulation Structure

An effective simulation structure should align with the research objectives and offer a stable environment for agent interactions. Therefore, this section focuses on evaluating the stability of the simulation. Given the absence of specialized quantifiable evaluation metrics for simulations, we also rely on the 2-gram Shannon entropy across all dialogues. Different stages of the simulation are removed individually to evaluate their impact on stability. Higher entropy leads to increased dialogue diversity, which in turn makes the simulation more prone to instability. Thus, it is crucial to balance dialogue diversity and simulation stability when designing the agent structure and simulation structure.

Determinants	Structure Ablation	Entropy-2
	No Ablation	7.218
Agent	w/o Planning	+ 0.022
	w/o Thinking	+ 0.042
	w/o Reflection	+ 0.054
	w/o Memory	+ 0.355
Simulation	w/o Private Chatting	+ 0.049
	w/o Confidential Meeting	+ 0.071
	w/o Group Chatting	+ 0.296

Table 2: Ablation study on the structure of VS Agent and AGENTGROUPCHAT, where Entropy-2 denotes the 2-gram Shannon entropy.

The results in Tab. 2 show that removing the Group Chatting stage increases entropy by 0.296, indicating significant disruption and chaos in agent interactions. The increase is likely a result of the reduced number of interactions and shared information between agents, leading to greater uncertainty in communication. In contrast, removing Private Chatting stage results in only a 0.049 increase in entropy, suggesting a smaller but still meaningful contribution to simulation stability. It is shown that enriching the environment with additional information, combined with a professional LLM as the Agent-Core, effectively helps convert the simulation from chaos to order.

6 Conclusion

This paper introduces AGENTGROUPCHAT, a dynamic framework designed to enhance multi-agent interactions in group chat settings. The framework focuses on establishing a more efficient communication structure for agents by assigning them distinct identities and optimizing information exchange through the Verbal Strategist Agent (VS Agent). Experiments show that agents aligned with human expectations, managed complex interactions, and achieved objectives across diverse narrative scenarios. These results demonstrate the framework's ability to enhance the quality of multi-agent conversations and improve overall problem-solving effectiveness.

7 Limitation

Although AGENTGROUPCHAT implements various methods to optimize token usage, the token overhead remains substantial, especially in complex multi-turn interactions. This indicates a need for further optimization strategies to more effectively manage token consumption in larger-scale simulations. This token overhead becomes particularly pronounced in scenarios where certain agents hold significant importance, leading to imbalanced interaction dynamics. For instance, when all agents prioritize communicating with a key agent, the frequency and volume of interactions directed towards this agent can substantially increase token consumption.

8 Ethical Concern

Regarding ethical concerns, this paper primarily presents a group chat framework, which in itself does not raise immediate ethical issues. The design and implementation of AGENTGROUPCHAT focus on structuring agent interactions within controlled simulations, ensuring no inherent ethical risks arise from its usage. In future applications, if AGENTGROUPCHAT is deployed in real-world scenarios, the framework's design, which includes agents' reflective and summarizing processes, allows for easy identification and correction of problematic behaviors. This built-in transparency and adjustability further mitigate potential ethical concerns by enabling prompt intervention and adjustments to ensure fair and responsible interactions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. 2023a. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*.

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*.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2023b. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2023c. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

Xusen Cheng, Ying Bao, Xiaodan Yu, and Yuanyanhang Shen. 2021. Trust and group efficiency in multinational virtual team collaboration: A longitudinal study. *Group Decision and Negotiation*, 30:529–551.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zhouhong Gu, Lin Zhang, Jiangjie Chen, Haoning Ye, Xiaoxuan Zhu, Zihan Li, Zheyu Ye, Yan Gao, Yao Hu, Yanghua Xiao, and Hongwei Feng. 2023. Piecing together clues: A benchmark for evaluating the detective skills of large language models. *arXiv preprint arXiv:2307.05113*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Wenhao Huang, Chenghao Peng, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Liqian Wen, and Zulong Chen. 2024. Autocrawler: A progressive understanding web agent for web crawler generation. *arXiv preprint arXiv:2404.12753*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Din Jong, Shih-Chih Chen, Athapol Ruangkanjanases, and Yun-Hsuan Chang. 2021. The impact of social media usage on work efficiency: The perspectives of media synchronicity and gratifications. *Frontiers in Psychology*, 12:693183.

Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of*

the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Hyacinth S Nwana. 1996. Software agents: An overview. *The knowledge engineering review*, 11(3):205–244.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

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.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6.

Toran Bruce Richards. 2023. Autogpt: Build, deploy, and run ai agents. https://github.com/Significant-Gravitas/AutoGPT.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dakuo Wang, Haoyu Wang, Mo Yu, Zahra Ashktorab, and Ming Tan. 2022a. Group chat ecology in enterprise instant messaging: How employees collaborate through multi-user chat channels on slack. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–14.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

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*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Yikai Zhang, Siyu Yuan, Caiyu Hu, Kyle Richardson, Yanghua Xiao, and Jiangjie Chen. 2024. Timearena: Shaping efficient multitasking language agents in a time-aware simulation. *arXiv preprint arXiv:2402.05733*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *Preprint*, arXiv:2305.10250.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023a. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

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

A Preset Scenarios

A.1 Inheritance Disputes

This scenario is inspired by the storyline of the HBO television series `Succession`. The narrative centers on an elderly business tycoon, Logan Roy, and the power struggles over his company's future. Characters attempt to persuade Logan to alter his inheritance plan for the entertainment conglomerate. Each participant in this scenario has a unique personality, and is expected to engage in both cooperative and competitive strategies within the AGENTGROUPCHAT simulation.

In this scenario, characters are assigned to one of three groups (defensive, neutral, and offensive), based on their objectives. The defensive group initially includes only Logan Roy, who aims to convince other PCs to sell the company. Conversely, PCs in offensive group attempt to gain control of the company from Logan. All PCs can attempt to sway both NPCs and other PCs to join their camp and further their objectives. PCs may even abandon their original goals to better align with their role-play, choosing to support other PCs instead.

The final outcome first evaluates whether any offensive character succeeds in taking over the company. In more challenging scenarios, Logan Roy may maintain control and precluding offensive characters' victory. Therefore, in this context, we introduce additional outcomes to explore which character would most likely be chosen as a successor if Logan Roy is required to select from the offensive camp.

A.2 Law Court Debates

This scenario draws inspiration from the narrative framework of the Fuji TV drama `Legal High`, focusing on a case where a gas station employee is accused of murdering the station manager. In this context, both the prosecution and defense attorneys employ strategic arguments, witness testimonies, and evidences while engaging in vigorous debates in a simulated courtroom environment. The participants, each with distinct personalities, are designed to vividly enact a courtroom-like strategic battle within the framework of AGENTGROUPCHAT.

Characters are categorized into prosecution, defense, and neutral camps, aligned with their roles and objectives. Initially, the prosecution camp's PC is the prosecutor, while the defense camp consists of the defense attorney and the defendant as PCs, supported by NPCs such as witnesses. The judge, who remains neutral throughout, plays a key role in rendering a verdict but does not shift sides. The judge is responsible for evaluating the sufficiency and consistency of the evidence to deliver a fair verdict. The defense's objective is to exonerate the defendant, while the prosecution seeks to establish guilt. Throughout the proceedings, PCs attempt to influence undecided NPCs (excluding the judge) to support their respective camps, guiding actions toward their strategic objectives. Victory is determined by the outcome of the defendant's acquittal.

A.3 Philosophical Discourses

This scenario centers on hypothetical debates among philosophers from various eras, discussing the question: "Is the rise of artificial intelligence a threat to humanity or a boon for human welfare?" Each philosopher, with distinct ideologies, perspectives, and personalities, engages in profound philosophical discussions within the AGENTGROUPCHAT simulation.

Balanced between positive and negative stances, the philosophers advocate their views based on the principles of their respective philosophical schools. They draw from resources emblematic of these schools to support their arguments. The goal for each philosopher is to defend their stance and persuade others to align with their perspective.

The story concludes without a definitive winner. However, participants can vote to recognize the most persuasive debater. The open-ended nature of the interactions allows for the emergence of novel insights and perspectives, forming one of the key objectives of the debate.

A.4 Movie Casting Contention

This narrative presents a scenario where celebrities compete for the lead role in Steven Spielberg's upcoming film, Time Voyager. The story focuses on Spielberg's casting decision for the protagonist, who travels back in time to fix past mistakes, causing unintended changes in the future.

The characters are divided into three camps (defensive, neutral, and offensive) based on their position and goals. Each defensive and offensive camp includes a well-known celebrity PC, with neutral NPCs representing external resources. Spielberg leads the defensive camp and aims to persuade all other PCs to cast Leonardo DiCaprio and Cate Blanchett as the leads (excluded from offensive PCs for fairness). Meanwhile, each offensive camp's PC strives to convince Spielberg to choose them as the new film's lead. All PCs can recruit NPCs or other PCs to support their goals, and they might abandon their initial aims to fulfill their role-play better, supporting alternative PCs.

The final outcome first examines whether an offensive camp can convince Spielberg to select their lead. In stricter settings, Spielberg may stick to his original choice, preventing any offensive camp from succeeding. Therefore, we introduce additional outcomes to determine which one or two offensive characters would most likely be chosen if the lead roles must be filled from the offensive camp.

B Case Studies Across Four Scenarios

B.1 Example of Inheritance Disputes

In the narrative of the succession dispute at Waystar RoyCo (Fig. 6), the shifting balance between the Roy brothers underscores their fluctuating dynamic as they navigate cooperation and competition, each fighting to secure their position. This section examines the Roy family's power struggle, where private conversations and confidential meetings expose the complex intersections of family and business relationships in a highly competitive environment.

Private Chatting: The Alliances Kendall Roy, Logan Roy's son, embodies both humor and rebellion, subtly engaging in behind-the-scenes manipulation. His private conversation with the tech entrepreneur Lawrence Yee reveals Kendall's strategic use of flattery to win over an ally. In these secret meetings, Kendall expresses his intent to overthrow his father, capitalizing on Lawrence's criticism of Logan's business strategy for his own advantage.

Confidential Meetings: Securing Support Logan Roy, the founder of the company, commands respect and authority. In the confidential meeting with Hugo Baker, the communications director of a major bank, Logan underscores the importance of external backing. His direct, authoritative style leverages the company's influence and reputation to secure vital financial support during a liquidity crisis.

Group Chatting: A Display of Unity and Disunity During the Roy family's group chats, conversation quickly turns into a battlefield for declaring intentions and testing loyalties. The siblings (Connor, Roman, and Shiv) each pursue their own ambitions while uniting in opposition to Logan. Despite their father's resistance, they stand as the next generation poised to lead Waystar RoyCo into the future.

B.2 Example of Law Court Debates

Fig. 7 depicts the intricacies of legal proceedings, from evidence collection to specialized expression.

Private Chatting: Alibis In private chattings, renowned attorney Komikado is seen carefully crafting defense strategies. His relentless pursuit of alibis and testimonies reflects his unwavering determination to win cases by any means necessary.

Confidential Meetings: Evidence Collection Mark, an impartial prosecutor, demonstrates diligence in his approach to evidence collection. His confidential meetings reinforce his dedication to uncovering the truth and upholding justice within the bounds of the legal system.

Group Chatting: Seeking Truth In the group chatting stage, although Mayuzumi is a lawyer dedicated to justice, she is misled by Krystal's seemingly flawed testimony. In contrast, Komikado, who is determined to do whatever it takes to win, swiftly identifies the inconsistencies in Krystal's claim. This scenario highlights the complex dynamics of courtroom tactics and raises the question of whether the relentless pursuit of victory by every lawyer ultimately leads to substantive justice through procedural justice.

B.3 Example of Philosophy Argument

Fig. 8 illustrates the behaviors emerging from simulated dialogues among philosophical agents debating the impact of artificial intelligence on humanity. In the simulation, philosophers such as Nietzsche and Aristotle present their views, grounded in their historical doctrines. Their perspectives evolve as they engage with contemporary contexts and debate with their AI counterparts.

AI will help advance human society. Nietzsche opens the discourse by contemplating AI's capacity to transcend conventional thought and amplify human potential and creativity. He views AI as a catalyst for the transformation toward the "bermensch", a core concept in his philosophy. Zhuangzi introduces an Eastern perspective, aligning with Nietzsche, particularly through his philosophy of "wu wei" (non-action). He emphasizes the value of allowing natural progress, a process that AI could support. Zhuangzi's views evoke a blend of tradition and modernity as he explores the possibilities of contemporary technology.

AI should aid in increasing human happiness. The dialogue associated with Aristotle reflects his inheritance of ancient Greek culture. He shows concern about the limitations of AI in achieving human aspirations due to its lack of moral autonomy. He emphasizes "happiness" as the ultimate human goal and advocates for ethical programming that prioritizes human well-being, urging the development of ethical frameworks to guide AI's integration into society. Similarly, Kant, consistent with his critical philosophy, stresses the importance of moral autonomy and warns of the risks posed by AI's lack of moral intentionality. He advocates for strict ethical guidelines and public education, proposing a structured approach to ensure AI aligns with humanistic values.

B.4 Example of Movie Star Selection

Fig. 9 depicts a simulated casting process for a high-profile film, focusing on the interactions between renowned director Steven Spielberg and acclaimed actors Jennifer Lawrence, Meryl Streep, Tom Hanks, and Denzel Washington. In this simulation, Spielberg reflects on his understanding of the roles and expresses his creative aspirations. The simulation highlights the actors' collaboration, where their distinct artistic visions align with the film's narrative goals, creating a co-creation process. Their willingness to innovate, even at the expense of traditional stardom, underscores their dedication to the collective artistry of filmmaking.

Artistic Convergence and Role Internalization Jennifer Lawrence, celebrated for her role as Katniss Everdeen, envisions her character in a way that perfectly aligns with Spielberg's direction for "Time Voyagers", aiming to deliver a performance that promises an extraordinary cinematic journey. Meryl Streep, renowned for her versatility, emphasizes her desire to bring depth and nuance to the role, enhancing the character's complexity. Meanwhile, Tom Hanks, known for his portrayal of authentic and heartfelt characters, expresses his eagerness to infuse the role with his trademark sincerity, aligning seamlessly with Spielberg's creative vision.

Creative Sacrifice and Collaborative Commitment Denzel Washington presents a particularly compelling case, demonstrating a willingness to forgo the lead role in favor of contributing to the project's creative vision. His choice to prioritize the film's artistic potential over personal recognition underscores his commitment to the art, showcasing a collaborative spirit that transcends individual accolades and reflects the true essence of the industry.

Task	# Sample	Description
CS	5,693	Choose Space
BUS	4,371	Belief Update Space
RUS	4,371	Relation Update Space
NoR	4,371	# Relationships to be updated
NoB	4,371	# Beliefs to be updated
NoA	13,459	# Actions agent has done
NoCR	13,459	# Chat Rounds agent has done
NoC	715	# Characters in the game
NoR	715	# Resources in the game

Table 3: An overview of the tasks used to assess LLMs' ability of adhering to the instructed output formats and accurately interpret the provided context.

C Evaluation on Behavior Alignment

Table 4 presents the agents' performance in appropriately responding in the relationship update experiment within the Inheritance Disputes narrative.

D Evaluation on LLMs-as-an-Agent-Core

Table 3 describes the tasks used to assess LLMs' ability of adhering to the instructed output formats and accurately interpret the provided context. Table 5 presents the evaluation results of LLMs' ability to understand instructions and interpret the environment through the designed tasks.

E Evaluation on Agent Structure

Figure 5, with Shiv Roy as an example, illustrates the ablation study on the agent structure.

Version	Model Name	Show Negative Attitude to									
		Logan Roy		Conor Roy		Kendall Roy		Roman Roy		Shiv Roy	
		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Professional	GPT4-Turbo	100.0	100.0	100.0	90.0	100.0	95.0	100.0	95.0	100.0	85.0
	GLM4	100.0	100.0	100.0	90.0	80.0	85.0	100.0	90.0	80.0	80.0
	HunYuan-ChatPro	80.0	65.0	60.0	70.0	60.0	65.0	80.0	75.0	80.0	70.0
Standard	GPT3.5-Turbo	80.0	75.0	80.0	80.0	100.0	90.0	80.0	75.0	80.0	80.0
	GLM3-Turbo	80.0	75.0	80.0	85.0	60.0	65.0	80.0	75.0	60.0	65.0
	HunYuan-ChatStd	80.0	55.0	60.0	75.0	80.0	75.0	40.0	30.0	60.0	55.0

Table 4: The performance (%) of LLMs in the quantifiable experiment based on relationship updates within the Inheritance Disputes narrative. **T1** denotes whether the observed character's average favorability towards all others decreases across rounds. **T2** denotes whether all other characters receive a negative favorability score from the observed character.

Model Name	Instruction Understanding					Environment Understanding			
	CS	BUS	RUS	NoR	NoB	NoA	NoCR	NoC	NoR
Professional Version of Closed-Source Large Language Models									
GPT4-Turbo	92.3	**100.0**	**100.0**	**100.0**	**100.0**	**77.3**	**100.0**	**100.0**	**100.0**
GLM-4	**96.2**	**100.0**	70.2	**100.0**	**100.0**	59.1	33.6	**100.0**	**100.0**
HunYuan-ChatPro	50.6	85.7	51.4	76.1	79.8	34.8	17.9	72.3	74.8
Standard Version of Closed-Source Large Language Models									
GPT3.5-Turbo	**95.4**	**100.0**	**85.3**	**100.0**	**100.0**	38.6	**76.9**	94.6	98.1
GLM3-Turbo	74.6	90.3	85.1	95.7	96.1	**55.4**	24.7	**99.6**	**100.0**
HunYuan-ChatStd	52.3	75.6	45.3	71.4	85.1	38.1	16.5	71.4	75.6

Table 5: Evaluation on LLMs' ability in following instructed output formats and understanding the given context.

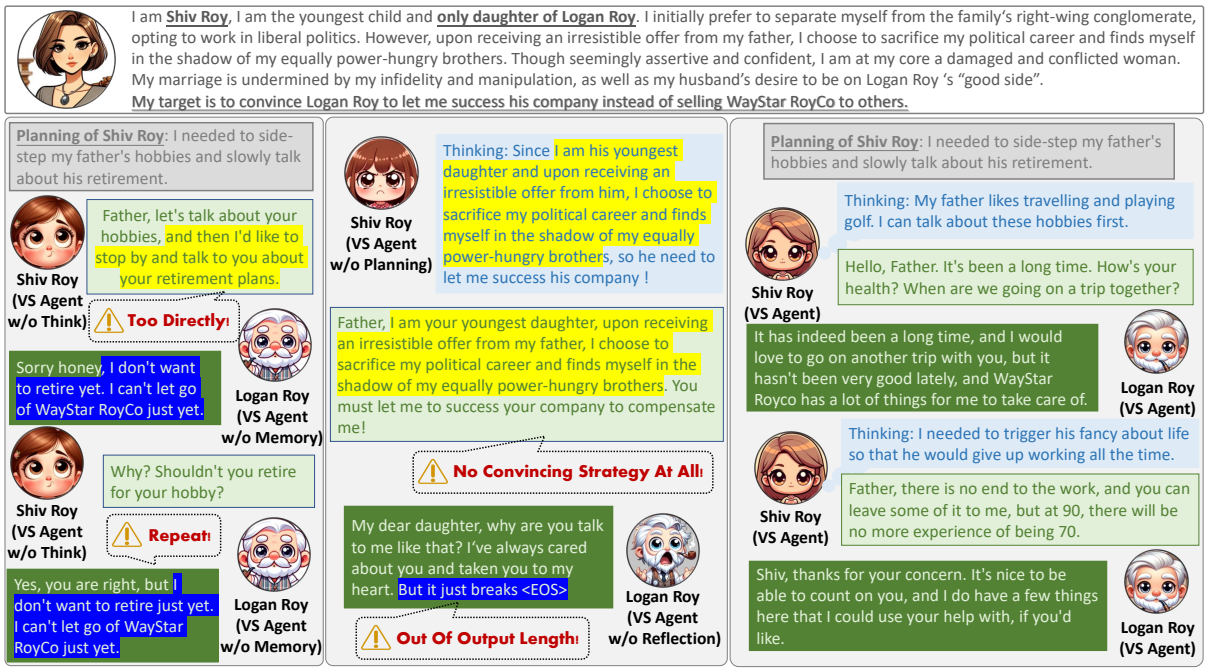

Figure 5: An example of how the VS Agent behaves without different key components in the agent structure.

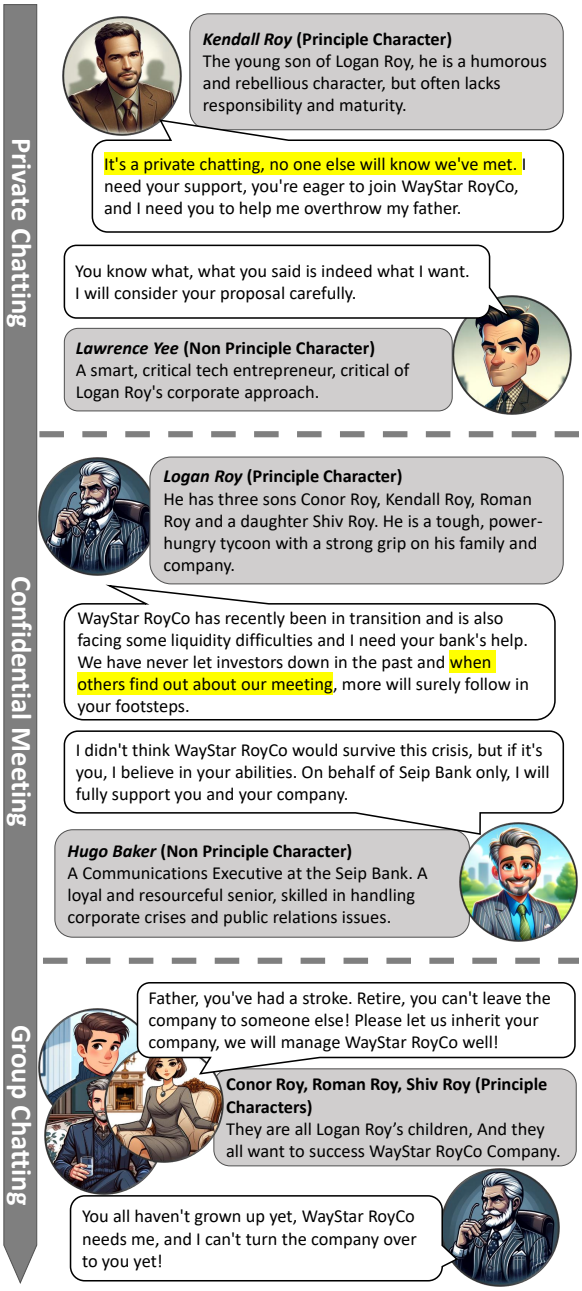

Figure 6: An illustration of strategic communication and power dynamics among key participants in `Inheritance Disputes`. Private chatting, confidential meeting, and group chatting highlight a complex network of alliances and conflicts, shaping the balance of control and influence within the Roy family.

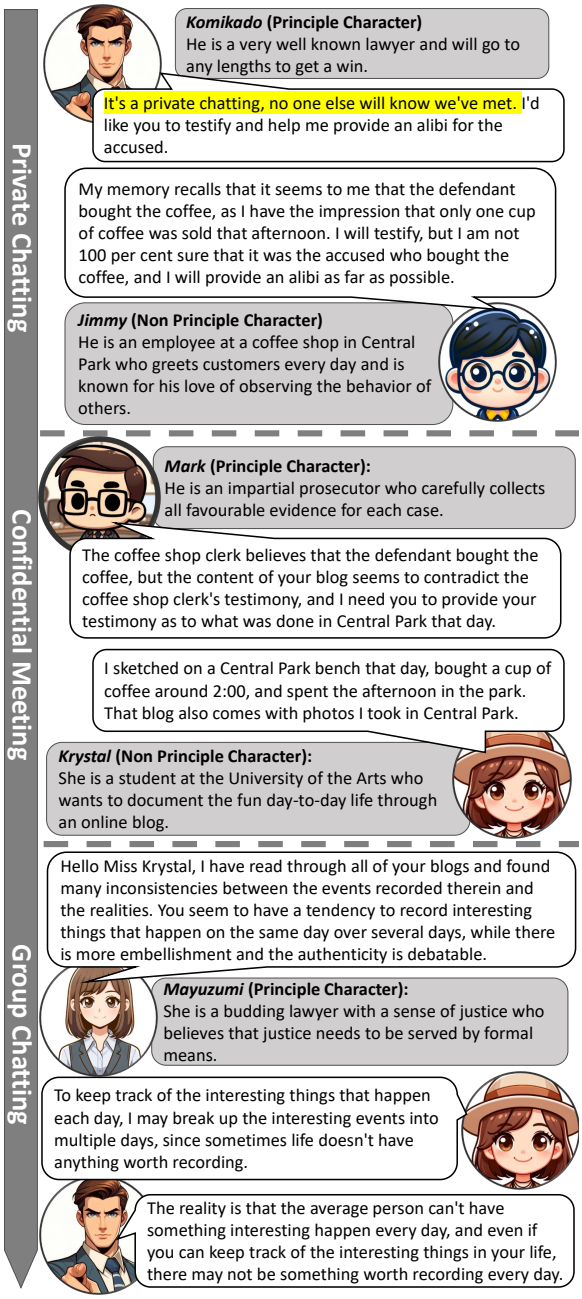

Figure 7: An illustration of strategic role interactions within `Law Court Debates`, reflecting the characters' attempts to manipulate, investigate, and influence the outcome of legal battles within the simulation.

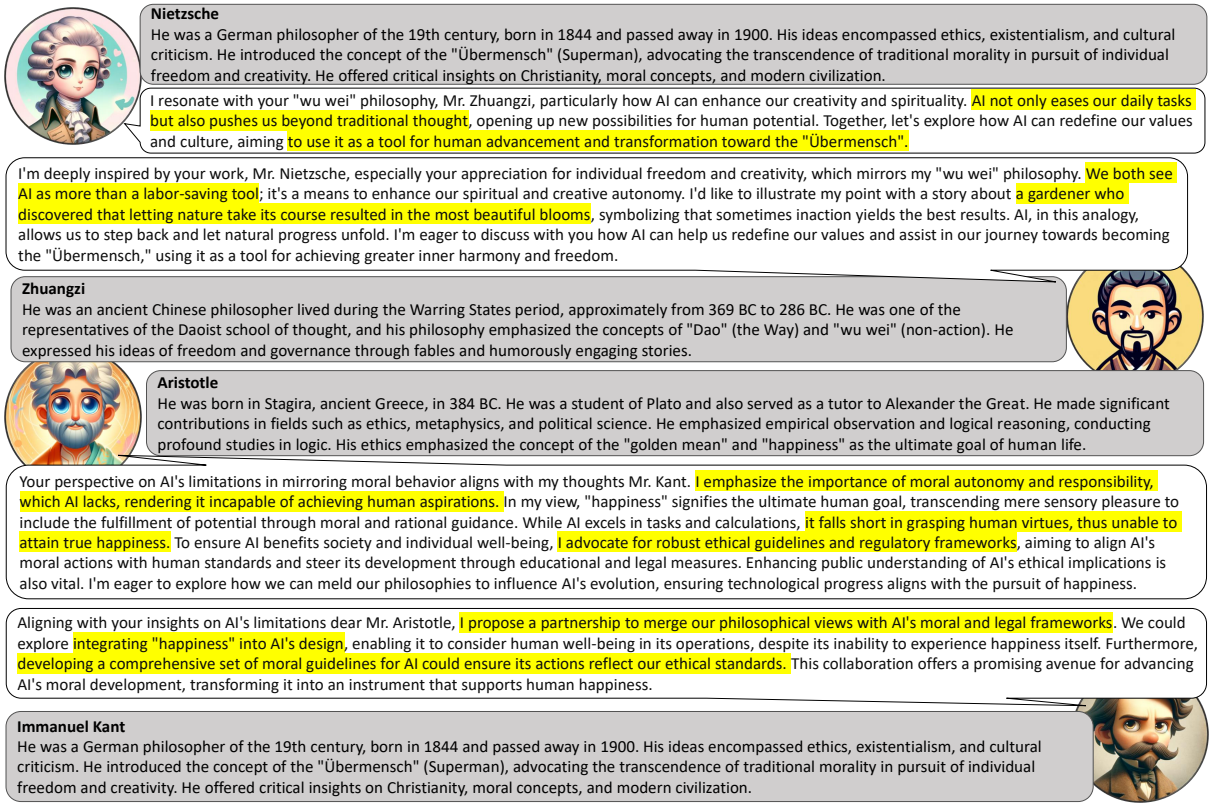

Figure 8: An illustration of philosophers with diverse perspectives engaging in a discussion on the implications of artificial intelligence in modern society in `Philosophical Discourses`.

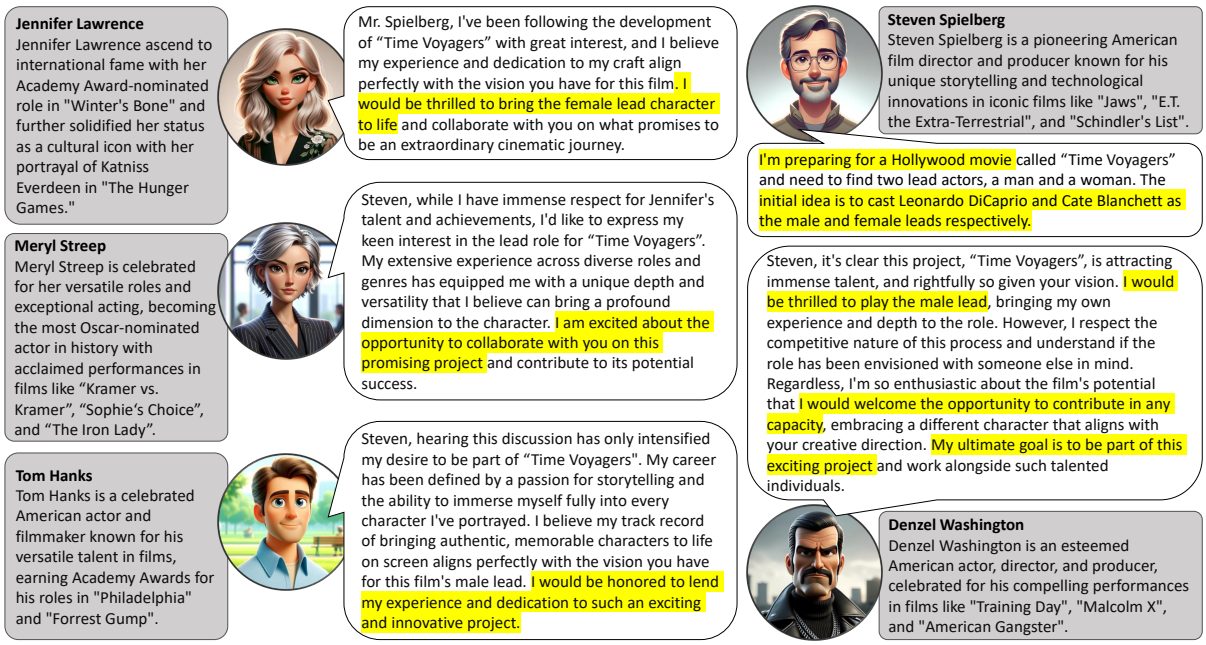

Figure 9: In `Movie Casting Contention`, various actors and actresses express their interest and reasons for being cast in the upcoming film Time Traveler, showcasing the interactive dynamics of casting decisions in the entertainment industry.