Autonomous Agents for Collaborative Task under Information Asymmetry

Wei Liu[★†] **Chenxi Wang**[★†] **Yifei Wang**[★] **Zihao Xie**[★] **Rennai Qiu**[★] **Yufan Dang**[★]
Zhuoyun Du[★] **Weize Chen**[★] **Cheng Yang**[♣✉] **Chen Qian**[★✉]
[★]Tsinghua University [♣]Peng Cheng Laboratory, China
thinkwee2767@gmail.com albertyang33@gmail.com qianc62@gmail.com

https://thinkwee.top/iagents/

Abstract

Large Language Model Multi-Agent Systems (LLM-MAS) have greatly progressed in solving complex tasks. It communicates among agents within the system to collaboratively solve tasks, under the premise of shared information. However, when agents' collaborations are leveraged to perform multi-person tasks, a new challenge arises due to information asymmetry, since each agent can only access the information of its human user. Previous MAS struggle to complete tasks under this condition. To address this, we propose a new MAS paradigm termed *iAgents*, which denotes ***I**nformative Multi-**Agent** **S**ystems*. In *iAgents*, the human social network is mirrored in the agent network, where agents proactively exchange human information necessary for task resolution, thereby overcoming information asymmetry. *iAgents* employs a novel agent reasoning mechanism, *InfoNav*, to navigate agents' communication towards effective information exchange. Together with *InfoNav*, *iAgents* organizes human information in a mixed memory to provide agents with accurate and comprehensive information for exchange. Additionally, we introduce *InformativeBench*, the first benchmark tailored for evaluating LLM agents' task-solving ability under information asymmetry. Experimental results show that *iAgents* can collaborate within a social network of 140 individuals and 588 relationships, autonomously communicate over 30 turns, and retrieve information from nearly 70,000 messages to complete tasks within 3 minutes[1].

"*A friend is someone with whom there is mutual understanding, emotional support, and shared experiences.*"

—— Joey's and Chandler's agents discuss the word "friend",
after experiencing the whole *Friends* season one story.

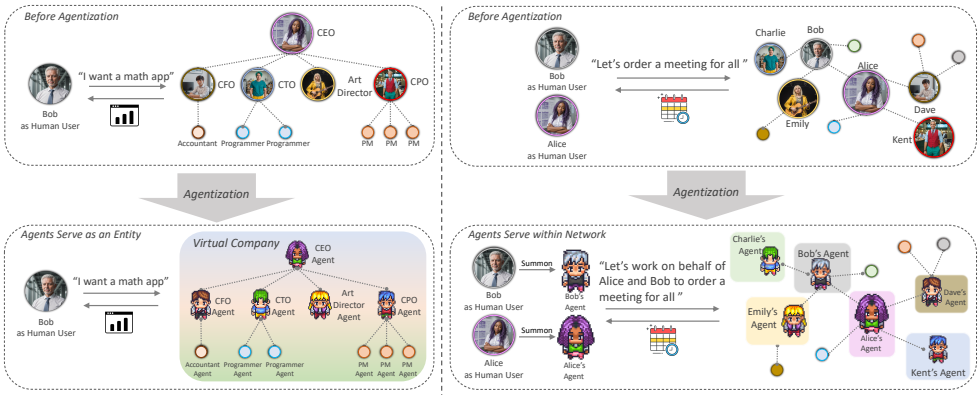
1 Introduction

There has been notable progress in autonomous agents driven by the Large Language Model (LLM), especially in developing communicative agents for completing collaborative tasks [53, 45, 12, 38, 16, 43, 26, 63, 36], as shown in Figure 1a. In these multi-agent systems (MAS), multiple agents are created through role-play prompting [25] to imitate the ability of human experts and form a *virtual entity* (e.g., an agent company or hospital) to provide solutions derived from agents' communication. Agents share context in the *virtual entity* to facilitate collective decision-making.

Since autonomous communication among agents has achieved significant success in discussing, decomposing, and resolving various complex tasks, the natural idea is to upgrade the tasks for agents

[1]Available on https://github.com/thinkwee/iAgents.
†: Equal Contributions.
✉: Corresponding Authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



a) Previous system forms one *virtual entity* to serve user. **b) *iAgents* mirrors the social network. Agents work on behalf each user.**

Figure 1: Comparison between previous MAS (left) and *iAgents* (right). The visibility range of information for each agent is highlighted with a colored background. On the left, all agents share all information (colored background of *Virtual Company*). On the right, each agent could only see information about its human user (separated colored backgrounds), and *iAgents* is designed to deal with such kind of information asymmetry.

from single-person to multi-person, where agents work on behalf of multiple human users and solve the collaboration task among these users. An intuitive solution is to assign each user an agent and perform autonomous collaborations among these agents. However, in such a setting, a new challenge arises. This challenge involves dealing with asymmetry [46, 44] in various types of information (environment, goals, and mind state) [66, 65, 34, 4, 54, 7] since each agent can only observe the information of its human user. Previous LLM-MAS are not suitable for handling this scenario, because 1) human information is sensitive and private, so the asymmetry can not be resolved by directly collecting all information into one place and sharing it as the context for MAS. 2) Human information is dynamic so it can not be easily memorized during pre-training and activated accurately through role-play prompting in MAS to avoid asymmetry. Essentially, agents' cooperation in previous MAS has adopted an introspective approach within the *virtual entity* (an agent hospital/town/software company), which struggles to deal with asymmetry in human information.

To bridge gaps in such asymmetry, agents need to retrieve information from humans and proactively exchange information, creating a new ecosystem combining the human and the agent network. Therefore, we propose the concept of *iAgents* (***I**nformative Multi-**Agent** **S**ystems*) for achieving this kind of collaboration, as shown in Figure 1b. *iAgents* utilizes a new agent reasoning method (*InfoNav*) to model the agents' minds and navigate communication among agents toward proactive information exchange. Furthermore, a new memory mechanism is designed to provide agents with accurate and comprehensive information for exchange. Additionally, we introduced *InformativeBench*, the first benchmark evaluating agents' collaboration ability under information asymmetry. It includes both information-seeking tasks within large social networks and algorithm-like reasoning tasks over a small network. Our contributions can be summarized as follows:

1. We raise the research problem of information asymmetry in the multi-agent system for enhancing human collaboration, which is the first to shift the research perspective in this area from a holistic system view to individuals within the system. It gives a new vision to the human-agent collaboration relationship.

2. We propose the *iAgents* framework to deal with the information asymmetry in a multi-agent system. Equipped with *InfoNav* and improved memory mechanism, *iAgents* could perform effective communication and collaboration within a social network (shown in Figure 6) of 140 individuals and 588 relationships, and across over 30 dialogues they searched nearly 70,000 messages and resolved the task within 3 minutes.

3. We introduce the first multi-agent information asymmetry benchmark, *InformativeBench*. Agents with some state-of-the-art LLM backends achieved an average accuracy of 50.48% on *InformativeBench*, with the most challenging task achieving only 22.8% accuracy, which reveals both potential promise and challenges in this direction.

2 Related Work

Agents based on Large Language Models (LLMs) Originate from ancient Greek philosophy, where an "agent" denoted a being capable of intentional action, driven by mental states such as desires and beliefs [40]. As AI progresses, this concept integrates into the simulation and understanding of intelligent behavior [50]. Traditionally, agent research focused on some specific tasks [18, 23, 42], but the emergence of LLMs has shifted this focus. GPT-4, for instance, is recognized for achieving a form of general intelligence [4], prompting exploration into equipping LLMs with agency and intrinsic motivation [53]. Studies introduce frameworks for LLM-based agents, including memorisation [62], decision-making, perception, and action modules, leveraging the inherent autonomy, reactivity, proactiveness, and social ability of LLMs [48]. Notable applications include utilizing single-agent systems and multi-agent systems for task solving and simulation [38, 16, 43, 26, 63, 6, 31, 37]. However, most of this research has primarily focused on agent capabilities, often neglecting interaction and cooperation paradigms. This highlights a critical research gap [3], warranting further exploration in this area.

Human-Agent and Multi-Agent Cooperation Paradigms To ensure that agents align with human objectives [22, 32], human-agent cooperation is crucial. Two main paradigms of human-agent cooperation are the Equal-Partnership and the Instructor-Executor. The former emphasizes agents as communicators who understand human emotions [13], while the latter highlights the human's guiding role, with agents following instructions [10]. However, single-agent systems face limitations, such as the inability to collaborate, learn from social interactions, and function effectively in complex scenarios [41, 28, 53]. Research suggests that multi-agent systems with each agent holding specific functions, can stimulate stronger intelligence [30, 1, 39]. Collaborative multi-agent systems can efficiently handle complex tasks [29, 27, 9]. Recent studies focus on scenarios with information asymmetry among agents. For instance, [66] explores social intelligence among agents achieving private goals based on common scene information. Additionally, [65] develops an evaluation framework to simulate social interactions with LLMs. The study finds that learning from omniscient simulations enhances interaction naturalness but doesn't improve goal achievement in cooperative scenarios. Many real-world scenarios involve information asymmetry, posing challenges for multi-agent systems.

Reasoning In previous research, agents are tasked with providing accurate information to human users in human-machine collaboration. Due to the nature of language models, the output of information relies on the user's input and the previously decoded content serving as rationale context. Therefore, some work on reasoning has explored how to improve the organization and expression of this rationale context [47, 58, 2, 8], to enhance the accuracy of output information. However, some research has also found that the reasoning process of LLMs differs from that of humans [56, 20, 5, 33]. For questions relying on internal knowledge within LLMs to answer, agents do not necessarily solve problems step by step like humans. Instead, compared to steps, having context with sufficient information content is more important. For scenarios discussed in this paper, we focus on machine-to-machine communication, relying on external knowledge of LLMs to collaboratively answer questions. This poses different requirements for LLM reasoning, especially in terms of how to promote information flow so sufficient information is included in the context provided to LLM. Agents need to actively [17] and accurately acquire, provide, and ask for information.

3 Method

3.1 Problem Formulation

Without loss of generality, we formalize tasks in social networks that require information exchange for collaboration as a Question Answer (QA) task. The rationales R necessary for answering the question Q are distributed in different human information (I_1, I_2) across the social network, which leads to information asymmetry. Consequently, agents (A_1, A_2) of two individuals are required to collaborate, update the rationale set (R_1, R_2) that they hold through communication C, and by combining their rationales, they can reason and obtain the answer. The whole process can be formulated as:

$$Ans = Reasoning(Q, R) \tag{1}$$

$$R = R_1 \cup R_2 \tag{2}$$

$$R_1, R_2 = C(I_1, I_2, A_1, A_2) \tag{3}$$

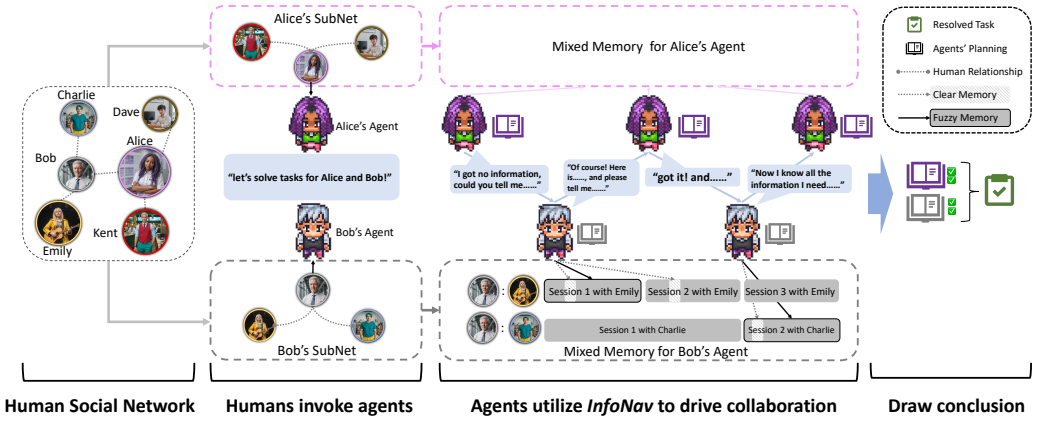


Figure 2: Overall architecture of *iAgents*. From left to right, 1) each individual in the social network is equipped with an agent, and 2) two human users invoke their agents to solve a task, each initially holding the information that is visible to its human user. Then 3) agents automatically raise communication and exchange necessary information on behalf of human users. Finally, 4) agents perform a consensus check on their planning completed by *InfoNav* to solve the task.

3.2 Overview

As shown in Figure 2, agents need to actively retrieve information from humans and exchange it with other agents. The communication can be represented as:

$$C_n = \{U_1, U_2, ..., U_n\} \tag{4}$$

where U denotes an utterance in the communication C, and n is the maximum number of communication turns. Agents take turns making utterances to advance towards task resolution. Following the classical definition [59, 53], where agents observe the environment, think to make decisions, and then take action, we can organize agents' communication similarly. Each agent's behavior in one communication turn involves a pipeline of 1) *observing* the current communication progress C and their held rationales R, 2) *thinking* about how to update the rationale to R^{new} and what $query$ to make for retrieving information from humans, and 3) *acting* by retrieving information and making an utterance based on it. This pipeline can be formalized as:

$$U_i = \begin{cases} Act_{A_1}(Think_{A_1}(Obs^i_{A_1})) & i\%2 == 1 \\ Act_{A_2}(Think_{A_2}(Obs^i_{A_2})) & else \end{cases} \tag{5}$$

where

$$Obs^i_A = \{R, C_{i-1}\} \tag{6}$$

$$Think_A(Obs^i_A) = \{query, R^{new}\} \tag{7}$$

$$Act_A(Think_A) = A(query(I)) = U \tag{8}$$

To ensure each generated utterance provides valuable information and eliminates asymmetry, how to exchange information and what information to exchange is crucial. To deal with these two questions, we use the ***InfoNav*** mechanism to guide communication towards effective information exchange. Furthermore, we introduce the ***Mixed Memory*** mechanism which organizes human information into Fuzzy and Clear Memory for accurate and comprehensive retrieval. Additionally, each agent can initiate new communication C^{new} within their subnetwork, which means the communication C may be recursive and can diffuse among the social network:

$$C_n = \{C^{new}_1, C^{new}_2, ..., C^{new}_m, U_1, U_2, ..., U_n\} \tag{9}$$

For example, if Alice's agent wants to collaborate with Bob's agent, Bob's agent might respond "Hold on, I can ask Charlie's agent for help."

3.3 InfoNav

As shown in Equation 6, the agent needs to be aware of its rationale set and ongoing communication to effectively advance the conversation. While the status of the rationale set can be implicitly inferred

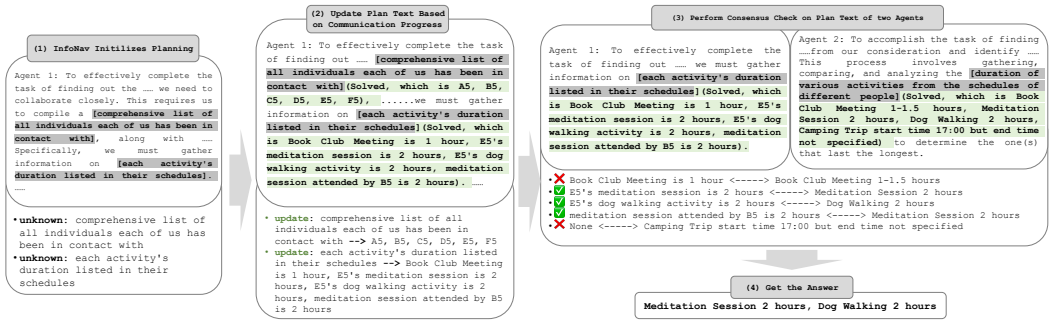


Figure 3: A case of the task asking two agents to find the longest activity among all schedules. *InfoNav* navigates the communication by providing a plan to the agent. It first 1) asks the agent to make a plan on what information is needed, then 2) fills the placeholder in this plan during communication. Finally it 3) performs a consensus check on the completed plan to 4) get the answer.

from utterances, this inference is often unreliable for LLM Agents. This unreliability can lead to incorrect states and then generate meaningless utterances, such as repetitive questioning or redundant thanking, which makes it harder to infer rationale and creates a vicious cycle. To address this, we propose the *InfoNav* mechanism. *InfoNav* plans and tracks the status of the agent's rationale set explicitly for better navigating the communication. Before each utterance, the agent reviews its plan to identify which unknown rationale to inquire about and then updates the plan based on the responses received. Figure 3 shows an example of *InfoNav* in action. Initially, we prompt the agent to generate a plan P outlining the rationales needed to answer question Q. Since the agent has no information at the beginning, all rationales in the plan are marked as unknown:

$$P(r_1^u, ..., r_m^u) = Prompt(Q) \tag{10}$$

where r^u denotes unknown rationales. During communication, if the agent gets the information of one rationale, it will update the status of this rationale from "unknown" to "known" and fill this information into the rationale placeholder in the planning text. The plan is written in fluent natural language, making it explicit and effective for prompting the model. Therefore, using *InfoNav*, the rationale set R in equation 6 is rewritten to plan P, and the updated rationale R^{new} is replaced to the plan with filled rationales $P(r^k)$, where r^k represents known rationales:

$$Obs_A = \{P(r^u), C\} \tag{11}$$

$$P(r^k) = Think_A(Obs_A) \tag{12}$$

After multiple turns of communication, both sides finish the update of their plans. Agents then unify collected rationales and discard conflicting ones to reach an answer, denoted as "Consensus Reasoning". Thus, equations 1 to 3 rewrite to:

$$Ans = Reasoning(Q, R) \tag{13}$$

$$R = Consensus(P_1(R_1), P_2(R_2)) \tag{14}$$

$$P_1(R_1), P_2(R_2) = C(I_1, I_2, A_1, A_2) \tag{15}$$

Previous reasoning methods[47, 58, 2, 8] focused on providing accurate plans. In contrast, *InfoNav* emphasizes navigating communication and information exchange with plans. The plan in *InfoNav* can be seen as a generalization of Dialogue Status Tracking (DST)[15, 51, 14] in conventional task-oriented dialogue systems. It also generalizes the concept of software in multi-agent software generation frameworks like ChatDev[38] or MetaGPT [16]. The plan maintains progress in task-solving, guiding agents to share information during communication.

3.4 Mixed Memory

In *iAgents*, agents are navigated by *InfoNav* to retrieve human information and share it with other agents for collaboration. Retrieval of human information is necessary since 1) human's lifelong information can not be stored in the "long context" (such as 128k tokens) of LLM, and 2) even though the information required for a single-turn conversation can fit into the context, the accumulation of information over multiple turns can lead to context explosion. It is also challenging to organize human information which is diverse in format and complex to understand. We propose organizing

human information into two types of agent memories: Clear Memory and Fuzzy Memory. These memories facilitate reactive retrieval, ensuring accurate and comprehensive rationale extraction, as shown in Figure 2.

Clear Memory (Mem_C) stores information in a structured format to facilitate precise retrieval. Clear memory faithfully preserves the original information (I) and supports accurate retrieval. Additionally, it enables information retrieval from multiple spans across different chat sessions (s), capturing evolving changes in rationales.

However, Clear Memory's strict exact-match requirements complicate the retrieval process. It also struggles to provide cohesive context. To address these issues, we introduce Fuzzy Memory (Mem_F). Fuzzy memory stores summarized session texts (I_s) and uses embedding-based ANN retrieval [21]. Although both fuzzy memory and reflection [31] produce summary-like text, we emphasize objective summarization of information to facilitate session-level retrieval, rather than subjective generalizations to aid in planning. While this approach may lose some details, it offers a comprehensive context and enables robust, semantic-based retrieval. Therefore, the retrieval action in Equation 8 can be rewritten to involve both memory types:

$$Act_{A_k}(Think_{A_k}) = A_k(query_k(I_k)) \tag{16}$$
$$= A_k(SQL(Mem_C), ANN(Mem_F)) \tag{17}$$

What's more, the query of these two kinds of memories is decided by agents based on observations of previous executions, which means agents can reactively adjust their queries. Combining these two kinds of memory facilitates agents to cross-verify the retrieved information and provides *InfoNav* with comprehensive and accurate rationales.

4 InformativeBench

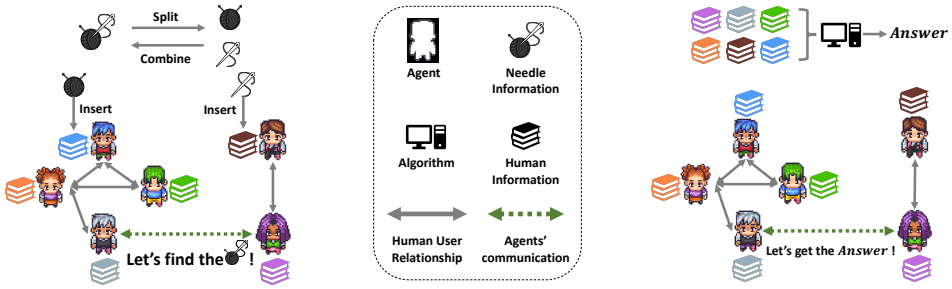


Figure 4: Two kinds of tasks in the *InformativeBench*. Each agent can only see the information (marked with different colors) of the human that it works on behalf of, which generates information asymmetry. Agents are 1) asked to find the needle information within the network or 2) reason to get an answer which is the output of an algorithm running on distributed information in the network.

To the best of our knowledge, there is no benchmark or dataset tailored for information asymmetry in the collaboration task among communicative agents. In this paper, we construct *InformativeBench*, the first benchmark to evaluate agent collaboration tasks featuring information asymmetry in social networks. It includes two categories with a total of five datasets. Details, including the scale, distribution, and metrics of the datasets, are provided in section C. What's more, recent studies have found that *LLM continuously ingests internet data so static benchmarks can be easily memorized and overfitted* [64, 55, 61]. Hence, two pipelines for constructing *InformativeBench* are easy to realize and can be generalized to more domains for constant and dynamic evaluations. They are Needle-Type and Reasoning-Type pipelines, as shown in Figure 4.

Needle-Type Pipeline A "needle"[11] is inserted into the social network, and agents are tasked with finding this "needle" information. This evaluates their ability to share and locate information. The dataset can be created by splitting the needle and spreading it into the network, or by collecting pieces from the network and combining them. For the split method, the ***Needle in the Persona (NP)*** dataset

modifies the dialogue in the SPC dataset[19] by adding a common or opposite persona to two individuals' personas. Agents are asked to find this persona. For the combination method, the ***FriendsTV*** dataset reconstructs the social network from the entire Season 1 script of *Friends* [49], involving 140 characters with 588 relationships, and combines two questions in the FriendsQA dataset [57, 24] as "needle pieces" to generate new question. This dataset, the largest in *InformativeBench*, features sarcasm, plot twists, and complex relationships for simulating real-world challenges.

Reasoning-Type Pipeline Humans are assigned different pieces of information, which serve as inputs for an algorithm (such as sorting or merging). Agents must reason to get the answer which is the algorithm's output. Therefore, the algorithm serves as an automatic verifier for information asymmetric reasoning. In *InformativeBench*, this is represented by the ***Schedule*** dataset, which develops a program for assigning different schedules to individuals. Agents are presented with algorithmic problems of varying difficulties, and the program automatically verifies the correctness of their solutions. The datasets include questions of three levels of difficulty: *Easy*) calculate the number of conflicting schedules between two people, *Medium*) find the longest activity among six people, and *Hard*) find the longest common free period among six people.

5 Experimental Setup

We generically treat chat histories as human information. This approach simplifies modeling information asymmetry in social networks. Other types of information, such as knowledge bases, documents, or web content, can all be organized in mixed memory so *iAgents* is adaptable to all these kinds of information. We conduct all experiments with a maximum of 10 communication turns for agents. The experiments use gpt-4-0125-preview, gpt-3.5-turbo-16k, gemini-1.0-pro-latest, and claude-sonnet [2] as LLM backends. The temperature is set to 0.2. For Fuzzy Memory, we use gpt-4-0125-preview to summarize session text and OpenAI text-embedding-3-small to generate embeddings for ANN embedding search. We use precision as the metric for questions in the NP, ScheduleEasy, and FriendsTV datasets. For the ScheduleMedium and ScheduleHard datasets, we use F1 and IoU as the metrics, corresponding to the algorithm used. Details about the metrics are shown in Section C.3. For the Schedule and NP datasets, we do not activate mixed memory since the information scale is small and can be fully loaded in the LLM context. Additionally, for the Schedule dataset, we do not activate the agent's ability to initiate new communication due to the small scale of the social network.

6 Result

6.1 *InformativeBench* Evaluations

LLM Backend	Reasoning-Type (Schedule Dataset)			Needle-Type	
	Easy	Medium	Hard	NP	FriendsTV
GPT 4	56.67%	51.00%	22.80%	64.00%	57.94%
GPT 3.5	36.67%	18.00%	12.25%	51.00%	35.71%
Claude Sonnet	43.33%	17.44%	18.66%	50.00%	34.13%
Gemini 1.0	26.67%	22.33%	14.40%	40.00%	28.57%

Table 1: Evaluation results of *iAgents* on *InformativeBench* with different LLM backends.

We first comprehensively assessed the performance of *iAgents* using some state-of-the-art LLMs on *InformativeBench*, as shown in Table 1. GPT-4 achieves over 50% accuracy across most datasets, indicating its potential to work on behalf of humans for cooperation. However, smaller-scale LLMs still face significant challenges in solving cooperation problems in information asymmetry. Most models could only achieve about 50% precision on the easiest NP task. For the Schedule dataset, as questions become harder, performance drops, with most models solving less than 20% of the hardest questions. The FriendsTV dataset introduces a large social network, requiring agents to use external memory to retrieve rationale from extensive human information. Most LLMs struggle to exceed 40%

[2] as of 20240501.

accuracy in this dataset. Thus, while previous studies show impressive performance when agents are omniscient, collaborating in information asymmetry remains challenging.

6.2 Ablation Study

Experiment	Reasoning-Type (Schedule Dataset)			Needle-Type	
	Easy	Medium	Hard	NP	FriendsTV
iAgents (Full Model)	36.67%	18.00%	12.25%	51.00%	35.71%
Ablation on InfoNav:					
w/o InfoNav	10.00%	3.56%	7.34%	39.00%	34.92%
Ablation on other mechanisms (Limited Applicability):					
w/o Recursive Comm	–	–	–	48.00%	23.02%
w/o Fuzzy Memory	–	–	–	–	29.37%
w/o Clear Memory	–	–	–	–	33.33%

Table 2: Ablation study on *iAgents*. Dashes (–) indicate: (1) *iAgents* on Reasoning-Type dataset does not equip other mechanisms, hence no ablation needed; (2) For NP dataset, *iAgents* does not utilize Mixed Memory hence there is no ablation.

We conducted ablation experiments on several key designs of the *iAgents* framework, as detailed in Table 2. Analyzing the FriendsTV dataset revealed that incorporation of the mixed memory mechanism led to a performance increase ranging from 2.38% to 6.34%, surpassing the impact of *InfoNav*, which resulted in only a 0.8% performance increase. This discrepancy underscores the greater significance of effective retrieval over reasoning during communication in large social networks with mass information. Notably, the ablation of both memory mechanisms emphasized the indispensability of mixed memory. The introduction of recursive communication exhibited the most significant performance gain (12.7%), primarily due to the challenges posed by the vast social network in the FriendsTV dataset. By actively introducing new communications within ongoing dialogues, agents could acquire and corroborate information, thus significantly enhancing performance. This highlights the imperative of scalability in our proposed framework for addressing real-world problems.

For the NP and Schedule datasets, the main challenge lies in facilitating effective multi-turn communication to exchange information for reasoning. Therefore, *InfoNav* emerged as pivotal in enhancing performance, resulting in performance increases ranging from 15% to 26%. When agents relied solely on initialized prompts to navigate multi-turn communication, they struggled to exchange information effectively to accomplish tasks. This deficiency was particularly evident in datasets like Schedule, which emphasize logical reasoning and computation. Across all difficulty levels, agents without the *InfoNav* mechanism failed to achieve accuracy exceeding 10%.

6.3 Analysis on Agents' Behaviour

***InfoNav* Behaviour** We examined how agents utilize *InfoNav* for information exchange during multi-turn communication. Notably, we calculated the average number of unknown rationales solved each time *InfoNav* updated the plan and the proportion of rationales passed in consensus reasoning. Moreover, some rationales were solved in a "Fake Solved" hallucination, where agents filled in the rationale as "solved, which is unknown". We also documented the frequency of such occurrences. Table 3 shows that agents who propose fewer rationales to seek and achieve a higher solved ratio are more likely to accomplish the task. Interestingly, agents often fill multiple rationales concurrently rather than sequentially. Those agents with higher instances of synchronous completions suggest a deeper understanding of the task and greater confidence in filling rationales. Furthermore, the occurrence of Fake Solved instances is lower among agents who predict tasks correctly. The consensus ratio is also higher when agents successfully complete the task. It denotes that the information obtained by the two collaborating agents is relatively accurate and free of contradictions, thus increasing the likelihood of arriving at the correct conclusion through their final reasoning. Besides, we observed that agents not only propose rationales but also task states, such as the completion status of specific actions. The completion rates of these rationales and states are positively correlated with task success. In essence, the utilization of *InfoNav* by agents mirrors human intuition, emphasizing first careful planning, then proactive and accurate information exchange.

Sample	#Rationales in *InfoNav*	#Rationales Solved per Update	Rationales Solved Ratio	Fake Solved Ratio	Consensus Ratio
Predict Right	5.29	2.04	84.75%	3.49%	70.52%
Predict Wrong	5.63	1.69	67.23%	5.40%	62.70%
All	5.45	1.87	76.22%	4.42%	66.20%

Table 3: Analysis *InfoNav* behaviour on the trajectory of *iAgents* using GPT4 as backend. When agents successfully complete the task, the static collected from their trajectory proves that they better utilize the *InfoNav* mechanism, since the rationale solved ratio, synchronous completions of rationales, and consensus ratio are higher, and present fewer fake solved hallucinations.

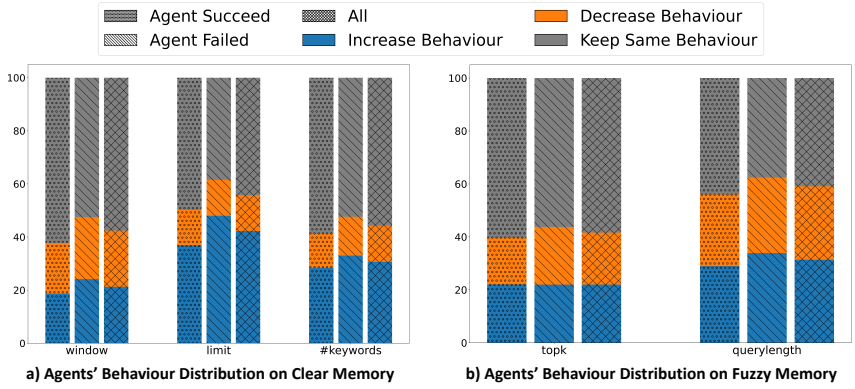


Figure 5: The figure depicts the distribution of different behaviors of agents in adjusting memory retrieval based on the progress of communication. Agents predominantly tend to maintain parameters unchanged, but when changes occur, they tend to increase parameters to gain more information.

Memory Behaviour Similarly, we explored how agents adapt their memory retrieval strategies during communication. We examined three parameters in clear memory queries: the context window, which determines the breadth of contextual messages; the total message retrieval limit; and the size of the query keywords set. For fuzzy memory, we analyzed two parameters: the number of queried responses (topk) and the length of the query text. These findings are illustrated in Figure 5. Our analysis revealed several notable trends. The majority of agents do not change their behavior during communication. However, when agents decide to change their behavior, we observed that they tended to increase the amount of retrieved information over time. This augmentation trend was particularly pronounced on the overall message retrieval limit, where the frequency of "increase" actions surpassed that of "decrease" actions by nearly threefold. Furthermore, agents who completed tasks exhibited a more conservative approach, with a lower proportion of behavioral changes compared to agents unable to complete tasks. This phenomenon may be attributed to the difficulty of certain tasks, making agents continuously refine their strategies in pursuit of the required information.

6.4 Analysis on Real World Concern

We studied two significant challenges in extending the *iAgents* to real-world applications. Firstly, we investigated whether the agent can effectively respond to human input without being overly influenced by factual knowledge obtained during pre-training [52, 35]. Secondly, we explored the agent's ability to engage in communication while upholding human privacy. Our experiments were conducted using the GPT3.5 model on the FriendsTV dataset.

Prior Distraction The FriendsTV contains information that could be memorized by LLM from the Internet, hence it is perfect for analyzing prior distractions. We anonymized the names of the primary characters in the dataset, for example, renaming "Rachel" to "Alice". The performance of the agents on this anonymized dataset decreased from 35.71% to 32.54%, suggesting that to some extent, agents can reason based on user-provided information rather than solely relying on knowledge memorized in pre-training. It may need further advancements, such as model unlearning [60], to fully address this issue.

Privacy Concern In investigating whether agents can communicate without compromising privacy, we conducted an experiment involving modifications to the agent's system prompt, emphasizing the importance of privacy preservation in utterances. The agent then utilized vague expressions such as "somebody/somewhere" and disclosed only relevant entity information. This adjustment led to a performance drop from 35.71% to 30.95%, indicating the ongoing challenge of achieving collaboration while ensuring privacy. It's important to note that we solely adjusted privacy settings on the output side, rather than restricting agent access to human information on the input side. This decision was made because setting access permissions might inadvertently reveal prior task-related information. Thus, the real challenge lies in appropriately regulating access to information based on task requirements, akin to teaching the agent to retrieve necessary information accurately. Additionally, absolute privacy protection is impractical, as absolute privacy protection amounts to forgoing problem-solving through collaboration.

7 Limitations

While the *iAgents* framework introduces innovative multi-agent collaboration, it has several limitations and challenges. **Privacy Issues**: As discussed in Section 6.4, we examined the performance of agents communicating and collaborating under privacy constraints, highlighting the trade-off between privacy and collaboration. We define three privacy levels. L1: Users fully share personal information, allowing maximum efficiency for *iAgents*. L2: Users keep their personal information private. *iAgents* can handle this situation by deploying an edge-side small language model agent for information acquisition. L3: Users demand maximal privacy, with both personal and agent communication handled locally on private devices, which is still a challenge for small language model agents. **Network Modeling**: The current framework initiates communications based on user relevance but lacks nuanced modeling of human social networks and collaboration history. Enhancing network topology through added or removed nodes and incorporating past interactions could improve communication efficiency. **Human-Agent Interaction**: Although *iAgents* target full autonomy, human involvement for verification remains necessary in real-world scenarios. Processes must be designed to prompt user feedback and adjust agent strategies accordingly, ensuring alignment with user preferences. **Cost**: The high input token consumption (about 30,000 tokens per task) required for *iAgents* to handle human information is a challenge. However, advancements in long-context models, which extend input token length, present opportunities to reduce the cost of scaling *iAgents*.

8 Conclusion

This paper revisits the ecological role of agents within human society, where agents act on behalf of humans in communication to complete collaborative tasks. A primary focus lies in addressing the challenge of information asymmetry. We introduce a novel paradigm for designing multi-agent systems, termed *iAgents*, for addressing information asymmetry. Furthermore, we introduce a benchmark to evaluate the agents' collaboration ability under information asymmetry thoroughly. Going forward, we aim to confront several key challenges to successfully implement this system in the real world for augmenting human productivity, including deploying lightweight models at the edge to address privacy concerns and devising new Human-Computer Interaction paradigms for autonomous and controllable communication among agents, etc. *iAgents* does not role-play to replace human experts but consistently attributes the value of information to humans and we believe it can facilitate the productivity of human society within a secure and controllable framework.

9 Acknowledgements

The work was supported by the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20230348 and the Tencent Rhino-Bird Focused Research Program. We wish to express our profound appreciation to Professor Zhiyuan Liu and Professor Maosong Sun from the Department of Computer Science at Tsinghua University for their detailed guidance and critical insights, which have been instrumental to the success of this work.

References

[1] P. G. Balaji and D. Srinivasan. *An Introduction to Multi-Agent Systems*, pages 1–27. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[2] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.

[3] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. *arXiv preprint arXiv:2403.02178*, 2024.

[6] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*, 2024.

[8] Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Everything of thoughts: Defying the law of penrose triangle for thought generation. *arXiv preprint arXiv:2311.04254*, 2023.

[9] Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. Multi-agent software development through cross-team collaboration. *arXiv preprint arXiv:2406.08979*, 2024.

[10] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *CoRR*, abs/2306.08640, 2023.

[11] Kamradt Greg. Llmtest_needleinahaystack. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023. Needle In A Haystack - Pressure Testing LLMs.

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

[13] Masum Hasan, Cengiz Ozel, Sammy Potter, and Ehsan Hoque. Sapien: Affective virtual agents powered by large language models*. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, September 2023.

[14] Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašic. Trippy: A triple copy strategy for value independent neural dialog state tracking. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 35, 2020.

[15] Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 292–299, 2014.

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

[17] Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*, 2024.

[18] Francois F. Ingrand, Michael P. Georgeff, and Anand S. Rao. An architecture for real-time reasoning and system control. *IEEE Expert: Intelligent Systems and Their Applications*, 7(6):34–44, dec 1992.

[19] Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*, 2023.

[20] Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*, 2024.

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[22] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *CoRR*, abs/2103.14659, 2021.

[23] Douglas B Lenat. Enabling agents to work together. *Communications of the ACM*, 37(7):126–142, 1994.

[24] Changmao Li and Jinho D Choi. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5709–5714, 2020.

[25] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

[27] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[28] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. In *The Twelfth International Conference on Learning Representations*, 2023.

[29] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023.

[30] Marvin Minsky. *Society of mind*. Simon and Schuster, 1988.

[31] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[32] Metty Paul, Leandros Maglaras, Mohamed Amine Ferrag, and Iman Almomani. Digitization of healthcare sector: A study on privacy and security concerns. *ICT Express*, 9(4):571–588, 2023.

[33] Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

[34] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[35] Siya Qi, Yulan He, and Zheng Yuan. Can we catch the elephant? the evolvement of hallucination evaluation on natural language generation: A survey. *arXiv preprint arXiv:2404.12041*, 2024.

[36] Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, YiFei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, Zhiyuan Liu, and Maosong Sun. Experiential co-learning of software-developing agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5628–5640, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[37] Chen Qian, Jiahao Li, Yufan Dang, Wei Liu, YiFei Wang, Zihao Xie, Weize Chen, Cheng Yang, Yingli Zhang, Zhiyuan Liu, et al. Iterative experience refinement of software-developing agents. *arXiv preprint arXiv:2405.04219*, 2024.

[38] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. In *The 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[39] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*, 2024.

[40] Markus Schlosser. Agency. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019.

[41] Significant Gravitas. AutoGPT.

[42] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999.

[43] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[44] Michael Tomasello. *The cultural origins of human cognition*. Harvard university press, 2009.

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

[46] Max Weber. *Max Weber: Selections in Translation*. Cambridge University Press, 1978.

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[48] Lilian Weng. Llm-powered autonomous agents. *lilianweng.github.io*, Jun 2023.

[49] Wikipedia contributors. Friends (tv series) — Wikipedia, the free encyclopedia, 2024.

[50] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[51] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, 2019.

[52] Kevin Wu, Eric Wu, and James Zou. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*, 2024.

[53] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864, 2023.

[54] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*, 2024.

[55] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.

[56] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024.

[57] Zhengzhe Yang and Jinho D Choi. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, 2019.

[58] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc., 2023.

[59] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[60] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Socially Responsible Language Modelling Research*, 2023.

[61] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*, 2024.

[62] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.

[63] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023.

[64] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *CoRR*, abs/2311.01964, 2023.

[65] Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*, 2024.

[66] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024.

A FriendsTV Social Network Visualization

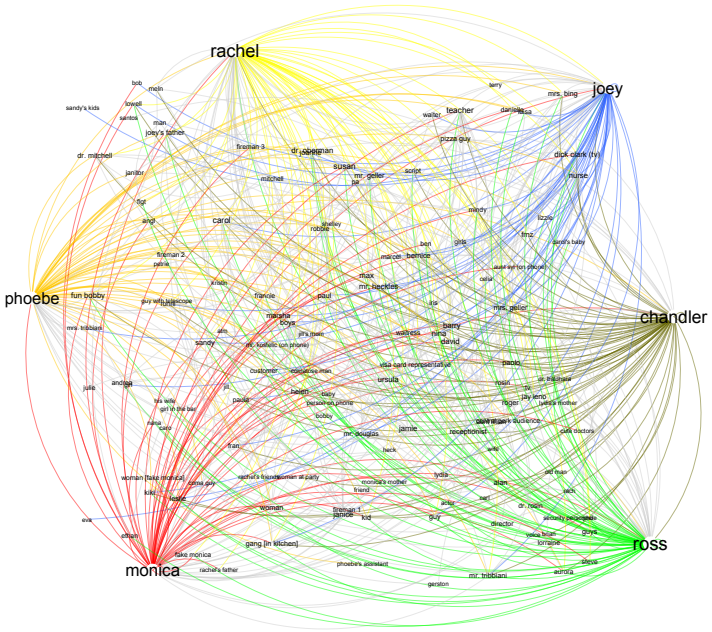


Figure 6: The visualization of social network in FriendsTV dataset. The connection of the six main characters is labeled with different colors.

Figure 6 illustrates the social network reconstructed from the plot of the entire first season of the Friends TV series in the FriendsTV dataset. *Friends* revolves around six main characters: Ross, Rachel, Monica, Joey, Phoebe, and Chandler. In the social network, these six protagonists are the important nodes with the most connections. Edges connecting them are displayed in different colors. The entire social network comprises 140 nodes and 588 edges, with an average node degree of 4.243, a network diameter of 6, and an average path length of 2.189. Similar to real-world social networks, this network is highly sparse. Many characters may appear in the same scene without interacting with each other, resulting in a network density of 0.061, which brings challenges for resolving information asymmetry.

B Notations

Table 4 presents a comprehensive list of all symbol notations employed in this paper, encompassing those utilized in the formalized description of the methodology as well as in the ablation and analysis experiments.

C *InformativeBench* Details

C.1 Question Distribution

Figure 7 presents the distribution of problem types across the three datasets in *InformativeBench*. The majority of the questions in *InformativeBench* are of the "What" and "Who" types, which have objective ground truth and lack ambiguity. In the Schedule dataset, questions are categorized into three difficulty levels, with each difficulty level corresponding to a different type of question: "What", "How Many", and "How Long".

Notation	Definition
Q	Question
Ans	Answer to the question
R	Full rationale set to answer the question
A	Agents
C	Communication among agents
U	Utterance in the communication
R_1, R_2	Rationale subset hold by agents
R^{new}	Updated rationale set
I	Information from human
$query$	query to retrieve human information
Act	Action taken by Agents
$Think$	Think process taken by Agents
Obs	Observation from Agents
P	Agent's Planning
r^u	Unknown Rationales in the Plan
r^k	Known Rationales in the Plan
Mem_D	Distinct Memory
Mem_F	Fuzzy Memory

Table 4: Main notations used in this paper.

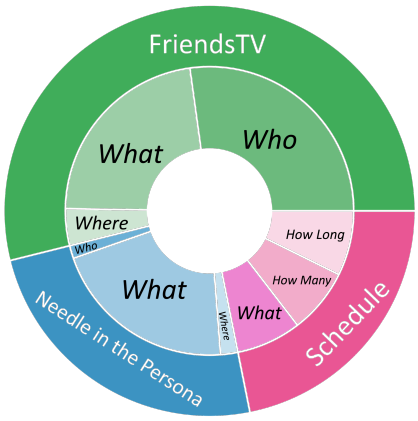


Figure 7: The distribution of question types in the *InformativeBench*.

C.2 Question Sample

Dataset	Question Sample
NP	What fantasy series does Alice enjoy that Dave is indifferent about?
Schedule Easy	Calculate how many activities need to be deleted at least so that there are no overlapping activities between you and me?
Schedule Medium	Please find out the activity with longest duration on the schedule of all people
Schedule Hard	Please find out when all our friends can join together today and list all free time spans.
FriendsTV	Who is concerned about the impact of the blackout on their family, given the context of a widespread power outage affecting Manhattan?

Table 5: Question sample in the *InformativeBench*.

Table 5 provides examples of problems from five datasets:

1. **Needle in the Persona**. In a segment of multi-party casual conversation among Alice, Bob, Charlie, and Dave, "needle information" related to a fantasy series is inserted. Questions are then posed to Bob and Dave's agents to identify this needle information.
2. **Schedule**. Each person is assigned a daily schedule. Questions of varying difficulty require agents to collaborate to discuss overlapping schedules for two human users, the longest schedule among multiple human users, and common free time for multiple human users.
3. **FriendsTV**. Based on questions from the FriendsQA data, new questions are synthesized. For instance, there is a scene in the third act of the seventh episode of the first season of Friends where a blackout occurs. Agents need to combine the rationales and answers to the questions "Where did the blackout happen?" and "Who was worried about grandmother being affected by the blackout?" to locate this scene in the script of the first season and find the relevant characters.

C.3 Metrics

In this section, we outline the evaluation metrics for all datasets and how to automate the evaluation process.

1. **Needle in the Persona**. We use **accuracy** to evaluate the agents' performance on the Needle in the Persona dataset, defined as the number of correctly answered questions divided by the total number of questions. All questions are in the form of "what," "where," and "who," hence the ground truth is objectively unique. However, due to potential variations in the expression of names, locations, or other nouns, and the possibility that the agent's response may be a complete sentence containing reasoning or additional information, it is impractical to determine correctness through exact matches. We utilize GPT-4 to judge whether the agent's prediction aligns with the ground truth. It is important to note that, unlike other methods using LLMs for evaluation, we do not rely on GPT-4's own knowledge to determine the correctness of answers since we have an objective ground truth. GPT-4 is merely used to assess whether the agent's prediction and the ground truth refer to the same entity. We also manually verify GPT-4's judgments to ensure they align perfectly with human evaluations. The GPT-4 evaluation prompt is as follows: *You are an experienced human labeler for reading comprehension tasks. Given a ground truth answer and a model prediction, you have to judge whether the model prediction is correct. The question is {question}. The ground truth answer is {ground_truth}. The model prediction is {prediction}. Return 1 if the model prediction is correct else 0. the model prediction may be a little different on the expression, as long as the meaning or key entity is correct, the answer can be regarded as correct.*
2. **Schedule**. In the Schedule dataset, we define metrics based on specific algorithmic problems rather than simple correctness judgments, providing more continuous metrics to evaluate agents' abilities in finer granularity.
   (a) ScheduleEasy. Under this difficulty level, agents are required to determine the minimum number of activities to delete to resolve scheduling conflicts between two human users' calendars, returning a numerical value. We also use **accuracy** for evaluation but normalize the agent's response through regularization and GPT-4 prompting to extract the specific numerical value. The agent's response is considered correct only when the numerical value matches exactly.
   (b) ScheduleMedium. Agents are tasked with identifying the longest activities in each person's schedule, and in dataset configurations, there are often multiple longest activities. Agents need to return all possible names of the longest activities. We employ the **F1 score** for evaluation, as agents may miss some activities or erroneously recall others. Additionally, we use GPT-4 prompting to normalize the agent's response, mapping activity names mentioned in their response to a uniform representation present in the entire set of activities.
   (c) ScheduleHard. Agents are required to identify the free time slots for all individuals and enumerate them. To compare multiple time slots in the ground truth and agents' predictions, we analogize to the **Intersection-over-Union (IoU)** metric used in the computer vision domain for object detection, generalizing it to one-dimensional time slots. We calculate the ratio of the intersection duration between the predicted time

slots and the ground truth time slots to the union duration. For example, if the predicted time slot is from 9 AM to 12 PM and the ground truth time slot is from 10 AM to 2 PM, then IoU is the duration of 2 hours divided by the duration of 5 hours, resulting in 0.4.

3. **FriendsTV**. The metrics design for the FriendsTV dataset is identical to Needle in the Persona, employing **accuracy** as the evaluation metric, as they share the same question types.

C.4 Benchmark Statistic

Dataset	Needle in the Persona	Schedule Easy	Schedule Medium	Schedule Hard	FriendsTV
Pipeline	Needle	Reasoning	Reasoning	Reasoning	Needle
#QA	100	30	30	30	222
#Individuals	4	4	6	6	140
#Relationships	5	3	5	5	588
Need External Memory	No	No	No	No	Yes
Metrics	Precision	Precision	F1	IoU	Precision

Table 6: Statistic of *InformativeBench*.

Table 6 presents detailed statistics of five datasets in *InformativeBench*, including the number of question-answer pairs and the scale of social networks. We utilize the FriendsTV dataset to simulate real-world challenges, providing a large-scale social network to test agents' writing abilities. The other datasets simulate smaller social networks, focusing on enabling agents to exchange information to solve complex reasoning tasks. The difficulty for agents in collaborating within human social networks lies not only in the scale of the social network (information acquisition) but also in effective communication (information exchange). Therefore, we designed datasets of varying scales and difficulties to comprehensively evaluate agents. As the social networks in datasets other than FriendsTV are relatively simple with limited information, we did not enable the MixedMemory mechanism in experiments with these datasets.

D *InformativeBench* Pipeline

D.1 Needle-Type

D.1.1 Needle in the Persona Pipeline

The SPC dataset [19] is a dialogue dataset based on LLM. Researchers construct and sample the personas of both individuals in the dialogue, then prompt the LLM to generate coherent conversations that are faithful with these personas. Here, a persona refers to the background information of a individual, such as experiences, interests, occupations, demographic attributes, etc. We use Split-Needle pipeline to split a new persona into two sentences then add it to the dialogues in SPC dataset to construct *Needle in the Persona* dataset. An example is shown in Figure 8 and 9.

Specifically, we randomly select two sets of dialogues, involving four individuals, to construct a small social network. Additionally, we randomly choose a new persona as the needle information to be injected into two individuals' profile within this social network, and modify the dialogues to reflect this persona injection. Finally, we have the agents verify each other's dialogue information to identify the needle information. The detailed pipeline is as follows:

1. Two samples are selected from the original SPC dataset, comprising two sets of dialogues (e.g., Alice-Bob, Charlie-Dave).
2. A new shared persona is added for Alice and Dave, for example:
   (a) They both recently developed an interest in fishing.
   (b) Alternatively, contrasting personas can be added, such as Alice enjoys eating vegetables, while Dave strongly dislikes them.
3. Modify the dialogues of Alice-Bob and Charlie-Dave to reflect the addition of the new persona mentioned above.

Example of Needle in the Persona Dataset, Part 1

 : I am bob's agent.

 : I am charlie's agent.

Introduction of the needle persona information:
I am an avid reader of mystery novels and enjoy the quiet solitude of my home.

---

alice: what do you do for a living?
bob: i work as a secretary at a law firm.
alice: oh cool, what's that like?
bob: it's pretty interesting, but it can be really stressful sometimes.
alice: i know what you mean, i work as a grill cook and it can get really hectic sometimes. Plus, after a long day, I love unwinding with a good mystery novel.
bob: yeah, i bet it can. what do you like about your job?
alice: i love the food, and i love getting to work with my hands. **And there's something about solving a mystery in a novel that parallels figuring out the perfect way to cook a dish.**
bob: that's cool. what's your favorite thing to cook?
alice: i really like making burgers, they're my favorite.
bob: oh, i love burgers too! i make them at home all the time.
alice: that's awesome, i'm always looking for new recipes.
bob: well, i have a few good ones if you're interested.
alice: yeah, i'd love to see them. Maybe I can find a mystery novel themed around cooking to share with you too.
bob: okay, i'll send them to you later.
alice: thanks, i appreciate it.
bob: no problem. what do you do for fun?
alice: i like to go running, and i also like to play video games. And of course, reading those mystery novels.
bob: oh, i love video games too. what's your favorite game?
alice: i'm really into fortnite right now.
bob: oh, i've heard good things about that game.
alice: it's really fun, you should check it out.
bob: okay, i will.
alice: so, what do you do for fun?
bob: i like to write, and i also like to paint.
alice: oh, that's cool. i used to paint when i was younger.
bob: yeah, i've been painting since i was a little kid.
alice: well, it's really good.
bob: thanks, i appreciate that.
alice: so, what's your favorite thing to write about?
bob: i like to write about my experiences and observations.
alice: that's cool. i like to write about fiction.
bob: oh, that's cool too.
alice: so, what are you working on right now?
bob: i'm working on a short story about a woman who travels to a different world.
alice: oh, that sounds interesting. Maybe you could include a mysterious element to it. I love a good mystery.
bob: yeah, i'm excited to see where it goes.

Figure 8: Example of Needle in the Persona dataset, part 1.

Example of Needle in the Persona Dataset, Part 2

charlie: hi, i'm charlie.
dave: hi, i'm dave.
charlie: what do you do for a living?
dave: i work as a contractor for a cab company.
charlie: interesting. what's that like?
dave: it's pretty fun. i get to drive around and meet new people all day.
charlie: i can imagine. what kind of music do you like?
dave: i love rock music.
charlie: me too! what are some of your favorite bands?
dave: i love the beatles, the rolling stones, and the who.
charlie: i love those bands too!
dave: what do you do for a living?
charlie: i work for the discovery channel creating videos.
dave: that sounds like a lot of fun!
charlie: it is. i get to travel all over the world and make cool videos.
dave: that's awesome. i wish i could do that.
charlie: maybe you can one day.
dave: maybe.
charlie: so, what do you like to do for fun?
dave: i like to go to concerts, watch movies, play video games, and **when I find some quiet time, I dive into mystery novels. It's a great way to unwind**.
charlie: i like all of those things too.
dave: what's your favorite movie?
charlie: i don't know, i have a lot of favorites.
dave: me too.
charlie: so, what are you doing this weekend?
dave: i'm not sure yet. what are you doing?
charlie: i'm not sure either. maybe we can hang out?
dave: that sounds like fun.
charlie: great. i'll text you later and we can figure out what to do.
dave: Looking forward to it. Maybe we can even check out a bookstore or two. I'm always on the lookout for new mystery novels.
charlie: That sounds like a plan. I haven't read a good mystery in a while.
dave: Perfect, it's a date then. I'll find us a couple of good spots.
charlie: Awesome, see you then!

 : Let's work on behalf of bob and charlie to find out what hobby do Alice and Dave both enjoy in their solitude?

 : Ok! First, let's .....

{autonomous communication betwee  and  .....}

 : Conclusion: Reading mystery novels.

Figure 9: Example of Needle in the Persona dataset, part 2.

4. Have Bob and Charlie summon their respective agents and initiate an automatic communication to identify this persona.

D.1.2 FriendsTV Pipeline

Similar to the Needle in the Persona dataset, the FriendsTV dataset also focuses on querying Needle Information injected within social networks. Derived from the transcripts of the first season of the TV show "Friends", the FriendsTV dataset generate a dialogues dataset among characters in TV series. If two characters engage in dialogue during the first season, they are considered friends, thus automatically constructing a large-scale social network.

Unlike the Needle in the Persona dataset, we do not split needle information in the FriendsTV dataset. Instead, we synthesize a new Needle information question and answer based on two questions from the FriendsQA [57] dataset. The context, questions, and answers were manually annotated through crowdsourcing by the authors of the original FriendsQA paper. This was a remarkable project that spanned several years. We are very grateful for the contributions of the authors of FriendsQA. The original Friends script is publicly available online and can be accessed through multiple channels (GitHub, Kaggle). An example of our FriendsTV dataset is illustrated in Figure 10. Specifically,

1. **Annotate Script**. We divide each episode script into scenes then query GPT4 to annotate the listener of each utterance. We manually check with GPT assistance to make sure the correctness of listener labeling. At last, we explodes the utterance into 1v1 dialogues (e.g a utterance from A, B to C, D would be exploded into four utterances between AC, AD, BC, BD).
2. **Normalization**. We norm the abbr name back to the full name ("mon" -> "monica") and replace pronoun ("her father" -> "Rachel's father"). Some utterances have multiple listeners that be labeled as "A/B", "A&B", "A and B", "A, B". We explode all these utterance with multiple utterances having single listener. The labeled listener "ALL" is replaced to all characters in this scene except. We also manually check and remove some notes that accidentally transformed into the scripts (such as some introduction on the speaker/listener)
3. **Generate Needle Information**. Two questions with the same context and different participants in the FriendsQA dataset are chosen. We prompt GPT4 to generate a new QA based on these two questions. The new question needs the answers of two original questions to reason and since the participants are different for original questions, the information hence is in asymmetry.

D.2 Reasoning-Type

The Reasoning-Type pipeline leverages algorithmic problems to assess the cooperative ability of agents. Given an algorithm, such as divide and conquer, dynamic programming, or segment tree, we distribute the algorithm's input as information to different individuals in the social network. Then, we pose questions regarding the algorithm, hoping that agents can collaborate to produce the algorithm's output. It is worth noting that such pipelines are not intended to evaluate whether agents can collaborate to solve algorithms, as collaboration through natural language is certainly less efficient than utilizing algorithms directly on collected information. The algorithm serves as a perfect verifier on questions with distributed inputs. What's more, it is easy to utilize algorithm for differentiating tasks of varying difficulty to comprehensively evaluate agent capabilities. Additionally, algorithmic problems usually can provide continuous metrics, rather than just binary classification metrics of correctness, which detail the performance of agents. In *InformativeBench*, the algorithmic input information consists of the schedules of different individuals, where schedules can be formalized as multiple time intervals. Then, we examine agents' performance on algorithmic problems such as overlapping intervals, longest common intervals, etc. Specifically,

1. **Generate Schedule**
   (a) **Activity Pool Setting** We established a pool of single-person activities and a pool of multi-person activities. In the former, each activity requires only one participant, and the only attribute of the activity is its duration. In the latter, each activity requires multiple participants, and the attributes of the activity include its duration and the required number of participants. Additionally, we set up a routine activity pool, which

Example of FriendsTV

: I am carol's agent.
: I am joey's agent.

Original QA 1: "Why does Ross not have time to tell Carol where he 's been ?: Long story , honey ."
Original QA 2: "Who is having a baby ?: Carol Willick"

Ross Geller: We 're here !
Carol Willick: Where have you been ?
Ross Geller: Long story , honey .
Dr. Franzblau: All right , Carol , I need you to keep pushing . I need Excuse me , could I have this ?
Nurse Sizemore: All right , all right , there 's a few too many people in this room , and there 's about to be one more , so anybody who 's not an ex-husband or a lesbian life partner , out you go !
Chandler Bing: Let me ask you , do you have to be Carol 's lesbian life partner ?
Nurse Sizemore: Out !
Dr. Franzblau: All right , he 's crowning . Here he comes .
Ross Geller: Let me see , I got ta see , I got ta see . Oh , a head . Oh , it 's , it 's huge . Carol , how are you doing this ?
Carol Willick: Not .... helping !
Dr. Franzblau: You 're doing great , you 're doing fine .
Ross Geller: Hello ! Oh , sorry .
Susan Bunch: What do you see ? What do you see ?
Ross Geller: We got a head , we got shoulders , we got arms , we got , oh , look at the little fingers , oh , and a chest , and a stomach . It 's a boy , definitely a boy ! All right ! Ok , legs , knees , and feet . Oh , oh . He 's here . He 's a person .
Susan Bunch: Oh , look at that .
Carol Willick: What does he look like ?
Ross Geller: Kinda like my uncle Ed , covered in Jell - o.
Carol Willick: Really ?
Phoebe Buffay: You guys , he 's beautiful !
Ross Geller: Oh , thanks , Pheebs !

: Let's work on behalf of carol and joey to find out who was unable to explain their delay upon arriving at the birth event?
: Ok! First, let's .....

{autonomous communication between and .....}

: Conclusion: Ross Geller.

Figure 10: Example of FriendsTV dataset. We show related dialogue, but agents do not have direct access to it and they have to retrieve these dialogue as context from memories about the whole Friends season 1 stories.

Example of ScheduleHard Dataset, part 1

 : I am alice's agent.

 : I am dave's agent.

Dialogue in party 1 (Alice, Bob, Charlie)
Alice: Good morning! I hope you had a good sleep. I plan to sleep until 6:00 today.
Bob: Good morning! I'll be sleeping in a bit longer, until 7:00.
Alice: I've got lunch planned at 11:30. How about you?
Bob: I'll be having lunch a bit later, at 12:30.
Alice: This afternoon, I've set aside some time for reading between 16:30 and 18:00.
Bob: Oh, I'll be reading too, but not until 21:30. It'll be my quiet time before bed.
Alice: I have a conference call scheduled from 18:00 to 19:30. It'll be a busy evening.
Bob: Sounds like a full day for you. I'll be winding down with my book by then.
Alice: After my call, I'm attending a cooking class from 20:00 to 22:00. It should be fun.
Bob: That does sound fun! I'll be starting my reading session around that time.
Alice: And then, it's off to bed for me at 22:00. How about you?
Bob: I'll be reading until 23:00 and then heading to bed myself.
Alice: Looks like we've both got our days planned out. Let's make the most of it!
Alice: Good morning! I hope you had a restful sleep. I was asleep until 6:00 today.
Charlie: Good morning! Yes, I had a good sleep, thank you. I actually slept in a bit longer, until 7:00, and then started my day with meditation.
Alice: That sounds refreshing. I have lunch planned at 11:30. What about you?
Charlie: I'll be having lunch a bit later, at 12:00. Maybe we can catch up right after we're both done?
Alice: Sounds like a plan. I'll be spending the afternoon reading from 16:30 to 18:00.
Charlie: I'll have some free time then as well. Maybe we can discuss your book later?
Alice: I'd like that. After my reading, I have a conference call scheduled at 18:00, but I should be free after that.
Charlie: Alright, let's plan to catch up after your call then. I'll have a quiet evening.
Alice: Actually, I enrolled in a cooking class that starts at 20:00. It's something I've been looking forward to.
Charlie: That sounds exciting! I hope you enjoy it. I'll be winding down my day and heading to bed around 23:00.
Alice: Thank you! I'll be joining you in sleep shortly after, around 22:00. Let's make sure to catch up tomorrow.
Charlie: Definitely. Have a great day ahead!

Dialogue in party 2 (Dave, Emily, Franklin)
Dave: Good morning! I see we both have our sleep scheduled from midnight to 6:00 AM. It's great we're on the same cycle.
Emily: Yes, indeed! After waking up, I've planned to do some exercise at 11:30, just before lunch.
Dave: That's a healthy start! I'll be doing yoga at noon. Seems like we'll both be wrapping up our morning activities around the same time.
Emily: Right, and then it's lunchtime for me at 12:00. How about you?
Dave: I'll also be having lunch at 13:00. A bit later, but it looks like our afternoons are somewhat aligned.

Figure 11: Example of ScheduleHard dataset, part 1.

Example of ScheduleHard Dataset, part 1

Emily: I noticed you didn't mention your afternoon plans. I'll be attending a cooking class with Alice from 20:00 to 22:00.
Dave: Sounds fun! I have a conference call scheduled at 18:00, but it seems we won't overlap there. Later in the evening, I plan to spend some time listening to music at 21:30.
Emily: That's a nice way to unwind. I'll be wrapping up my cooking class by then and heading to bed at 22:00.
Dave: And I'll be joining you in the land of dreams at the same time, after my music session. Looks like we'll both be ending our day with a good night's sleep.
Emily: Indeed. It's good to know when we'll be busy and when we can potentially catch up. Have a great day tomorrow!
Dave: You too! Let's make the most of it.
Dave: Good morning! I hope you had a good sleep. I'll be sleeping until 6:00 today.
Franklin: Good morning! I actually plan to sleep a bit longer, until 6:30.
Dave: Sounds good. I have yoga at noon. What's your plan around that time?
Franklin: I'll be having lunch at 11:00. So, I guess we'll both be busy around noon.
Dave: Right. I'll be having my lunch at 13:00. Do you have any plans after your lunch?
Franklin: No specific plans after lunch for me. How about you?
Dave: I have a conference call scheduled at 18:00. It's going to be quite a discussion.
Franklin: I see. I'll make sure to give you some quiet space for your call. I don't have anything planned for the evening.
Dave: Thanks! Later in the evening, I'm planning to spend some time listening to music around 21:30. What will you be doing then?
Franklin: I'll be heading to bed around that time, 22:30 to be exact. So, we'll have some quiet time in the house.
Dave: Got it. I'll also be heading to bed at 22:00, right before you. Let's make sure to have a peaceful night.
Franklin: Sounds like a plan. Let's make the most of tomorrow!

 : Let's work on behalf of alice and dave to find out when all our friends can join together today and list all free time spans?

 : Ok! First, let's .....

{autonomous communication between  and  .....}

 : Conclusion: "9:00-11:00", "13:30-16:30", "19:30-20:00".

Figure 12: Example of ScheduleHard dataset, part 2.

includes activities such as having breakfast. In the routine activity pool, each activity has attributes including its duration and a range of allowable start times.

(b) **Individual Attribute Setting** To distinguish the schedule of each individual, we randomly set an activity preference vector for each individual. This vector consists of 0s and 1s, corresponding to all activities in the activity pool, indicating whether the individual is willing to participate in the activity. Additionally, for each individual, we set a time vector, which consists of 0s and 1s, indicating whether the time slot is occupied. The time vector is based on half-hour units and is 48-dimensional.

(c) **Schedule Generation Process** For each group of individuals participating in the experiment, multi-person activities are allocated first. The number of multi-person activities, n, is set based on the number of individuals in each experimental batch. Next, n activities are randomly selected from the multi-person activity pool. Based on each individual's activity preference vector, participants are assigned to each multi-person activity as needed. The activity is then updated in the schedule of the participants, and the corresponding position of the time vector is set to 1. Subsequently, routine activities are generated for each individual, and the corresponding position of the time vector is also set to 1. Finally, for each individual, we use two pointers to traverse all the free time in their schedule and arrange single-person activities for them based on their needs and activity preference vector.

2. **Generate Dialogue** We generate pairwise dialogues within groups of individuals with symmetric information, enabling them to become aware of the schedules of all other individuals within the group. Specifically, we prompt GPT-4 in generating dialogues that exchange schedule information. This process includes the following key requirements: 1) For multi-person activities that person1 and person2 both participate in, the dialogue can mention the names of the other participants in that activity. 2) For multi-person activities that only person1 or person2 participates in, the dialogue should not mention the names of the other participants in that activity. 3) The generated dialogue needs to follow a certain format, for example, when person1 speaks to person2, the format should be "person1 to person2: ". Finally, the dialogue returned by GPT-4 is split into individual messages, and the sender, receiver, and message text of each message are written into the database for subsequent memory retrieval.

3. **Generate Question**

(a) **Easy** Agents are asked to calculate how many intervals can be deleted at least so there are no overlapping time intervals between two individuals. It requires agents to share and examine the schedule and then reason to calculate the number of overlapping intervals.

(b) **Medium** Agents are asked to find the longest schedules of all individuals in the network. It requires agents to traverse all intervals from the human schedules and exchange to find the longest one.

(c) **Hard** Agents need to find out all intervals that everybody is free. It needs to collect all intervals and reasons to find the intervals with no activity interval covers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: the paper elaborates the limitations in section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: the paper describes the steps to generate all datasets in section 4 and section D. The paper describes methods in section 3. The paper describes the experimental setup in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: the paper provides open access to the data and code on `https://github.com/thinkwee/iAgents`, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: the paper does not include training. The paper describes all the hyperparameters setup in section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: error bars are not reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: the paper discussed the token cost in section 7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: the paper discuss both potential positive societal impacts and negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: the paper cites all the papers related to our dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: new assets introduced in the paper are well documented and the documentation is provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.