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 [♣]Beijing University of Posts and Telecommunications
thinkwee2767@gmail.com yangcheng@bupt.edu.cn qianc62@gmail.com

Abstract

Large Language Model Multi-Agent Systems (LLM-MAS) have achieved great progress in solving complex tasks. It performs communication among agents within the system to collaboratively solve tasks, under the premise of shared information. However, when agents' communication is leveraged to enhance human cooperation, 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-**A**gent **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.

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

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

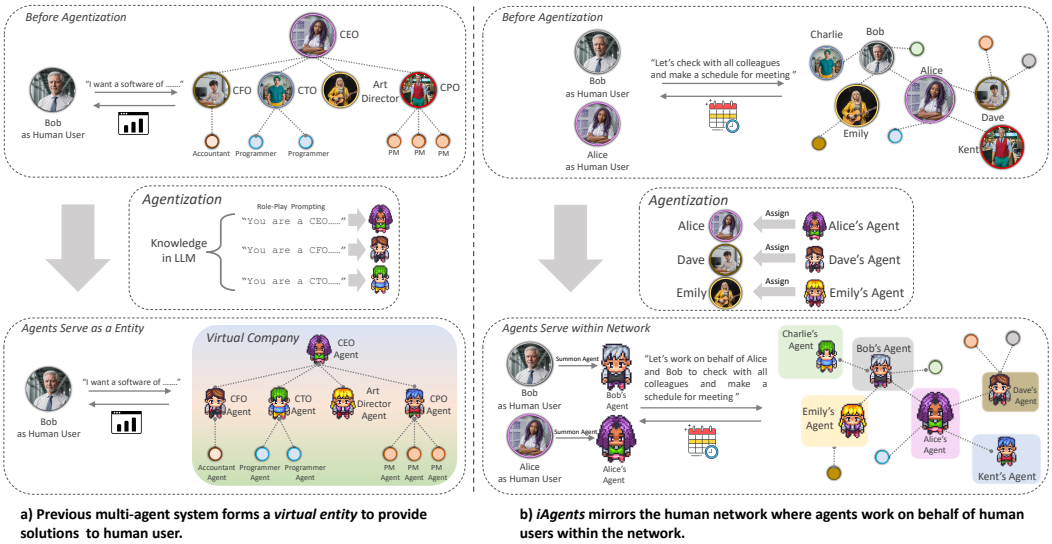
1 Introduction

There has been notable progress in autonomous agents driven by the Large Language Model (LLM) [42, 5, 6, 41], especially in developing communicative agents for completing collaborative tasks [51, 43, 11, 36, 15, 39, 26, 60], 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.

Given that autonomous communication among agents has achieved significant success in discussing, decomposing, and resolving various complex tasks, a natural idea is to introduce such autonomous

†: Equal Contributions.
✉: Corresponding Authors.

Preprint. Under review.



a) Previous multi-agent system forms a *virtual entity* to provide solutions to human user.

b) *iAgents* mirrors the human network where agents work on behalf of human users within the network.

Figure 1: Comparison between previous MAS and *iAgents*. The visibility range of information for each agent is highlighted with a colored background. In the previous MAS, all agents share all information (colored background of *Virtual Company*), so it only works under the premise of shared context. In *iAgents*, each agent only sees information of its human user (different colored backgrounds of cartoon characters), and *iAgents* is designed to deal with information asymmetry.

communication into human society to enhance collaboration efficiency among humans. However, when we manage to assign each human with an agent, establishing a mapping from the human social network to the agent network so autonomous communication among agents can facilitate human cooperation, a new challenge arises. This challenge involves dealing with asymmetry [44, 40] in various types of information (environment, goals, and mind state) [63, 62, 34, 6, 52] 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*, 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 with each other, creating a new ecosystem combining the human network and the agent network. Therefore, we propose the concept of *iAgents* (***I**nformative Multi-**Agent S**ystems*)[1] for achieving this kind of collaboration, as shown in Figure 1b. *iAgents* utilizes a new method for agent reasoning (*InfoNav*) which models the agents' minds and navigates agents' autonomous communication 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 in social networks with information asymmetry. It includes both information-seeking tasks for evaluating agents' ability to resolve asymmetry in a large amount of information, and complex algorithm-like tasks that focus on evaluating agents' ability to collaboratively reason under information asymmetry. We found that *iAgents* could perform effective communication and collaboration within a social network of (shown in Figure 5) 140 individuals and 588 relationships, and across over 30 dialogues they searched nearly 70,000 messages and resolved the task within 3 minutes. Despite such cases where *iAgents* exhibit impressive performance, 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.

[1]Available on https://github.com/thinkwee/iAgents

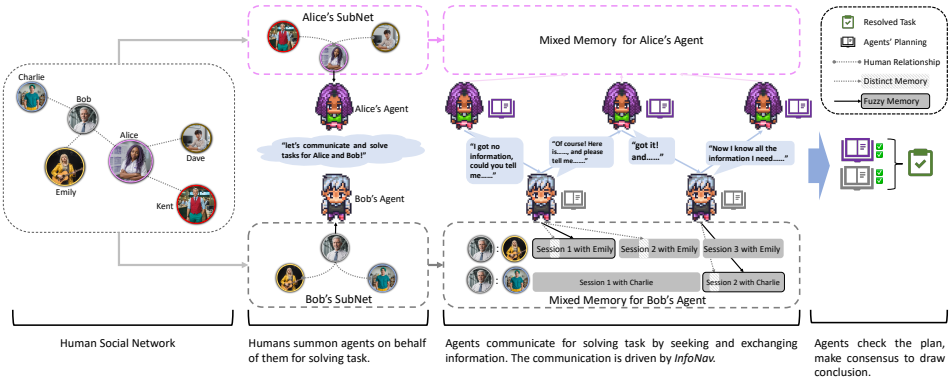


Figure 2: Overall architecture of *iAgents*. From left to right, 1) each individual in the social network is equipped with an agent, and 2) two individuals summon their agents to solve a task, each initially holding the information that is visible to its human user. Then 3) agents automatically raise 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.

2 Method

2.1 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:

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

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

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

2.2 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 [57, 51], 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

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

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

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

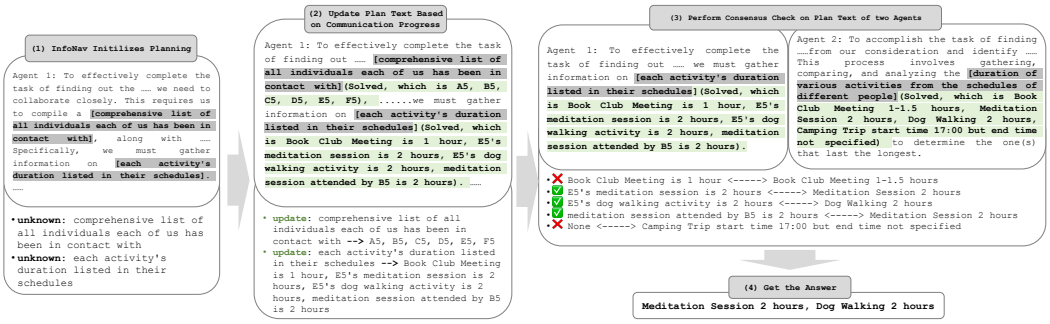


Figure 3: A case of the task asking two agents to find the longest activity among all schedules. *InfoNav* navigates the communication by providing a plan to the agent. It first 1) asks the agent to make a plan on what information is needed, then 2) fills the placeholder in this plan during communication.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

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 Distinct 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:

$$C_n = \{C_1^{new}, C_2^{new}, ..., C_m^{new}, U_1, U_2, ..., U_n\} \tag{9}$$

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."

2.3 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 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* explicitly records the rationale set using the agent's planning text and navigates the autonomous 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, these rationales can updated to "known" and filled with concrete content as information is provided. 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. The agents then perform consensus reasoning, which unifies collected rationales and discards conflicting ones to reach an answer. 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[45, 56, 3, 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)[14, 49, 13] in conventional task-oriented dialogue systems. It also generalizes the concept of software in multi-agent software generation frameworks like ChatDev[36] or MetaGPT [15]. The plan maintains progress in task-solving, guiding agents to share information during communication.

2.4 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 challenging due to it being diverse in format and complex to organize. We propose organizing human information into two types of agent memories: Distinct Memory and Fuzzy Memory. These memories facilitate reactive retrieval, ensuring accurate and comprehensive rationale extraction, as shown in Figure 2.

Distinct Memory (Mem_D) stores human information in a structured format, allowing for exact matches such as structured query language. Distinct 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, distinct 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 more abstract text, we emphasize an objective summary 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:

$$Act_{A_k}(Think_{A_k}) = A_k(query_k(I_k)) \tag{16}$$

$$= A_k(SQL(Mem_D), ANN(Mem_F)) \tag{17}$$

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.

3 InformativeBench

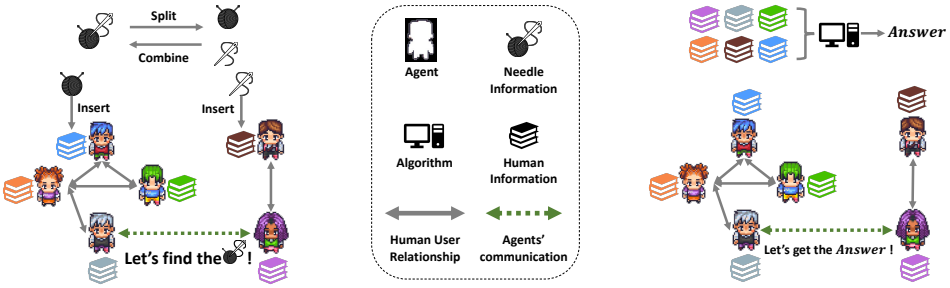


Figure 4: Two kinds of tasks in the *InformativeBench*. Each agent can only see the information (marked with different colors) of the human that it works on behalf of, which generates information asymmetry. Agents are 1) asked to find the needle information within the network or 2) reason to get an answer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

3.1 Pipelines

Currently, there exists no benchmark or dataset specifically designed to address information asymmetry in collaborative tasks among communicative agents. In this paper, we construct *InformativeBench*,

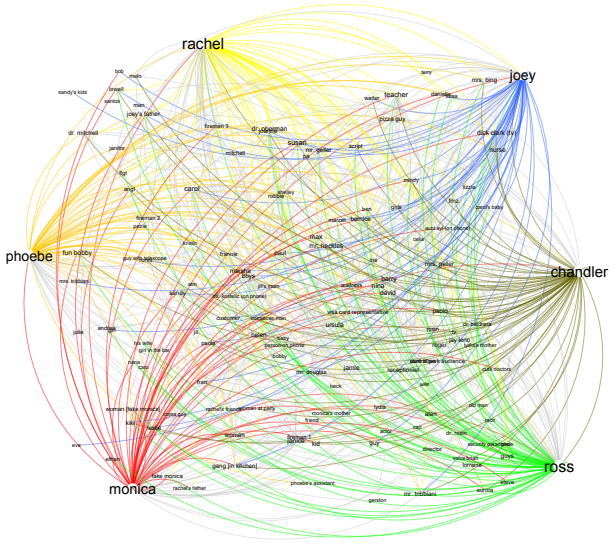


Figure 5: The visualization of social network in FriendsTV dataset. The connection of the six main characters is labeled with different colors.

the first benchmark to evaluate agent collaboration tasks featuring information asymmetry in social networks. It contains five datasets to assess agents' various capabilities. What's more, recent studies have found that *LLM continuously ingests internet data so static benchmarks can be easily memorized and overfitted* [61, 53, 59]. Hence we share two pipelines for constructing *InformativeBench* which are easy to realize and can be generalized to more domains for constant and dynamic evaluations. They are Needle-Type and Algorithm-Type pipelines, as shown in Figure 4.

Needle-Type Pipeline A "needle"[10] 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 either by splitting the needle and spreading it into the network, or by combining information pieces from the network to generate a needle. 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 based on the entire season one script of *Friends* [47], involving 140 characters with 588 relationships (as shown in Figure 5), and combines two questions in the FriendsQA dataset [55, 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.

Algorithm-Type Pipeline Humans are assigned different pieces of information, which serve as inputs for an algorithm. Agents must solve tasks where the answer is the algorithm's output, evaluating their reasoning abilities based on information exchange. 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 algorithm 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.

3.2 Question Distribution

Figure 6 presents the distribution of problem types in *InformativeBench*. It is evident that 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,

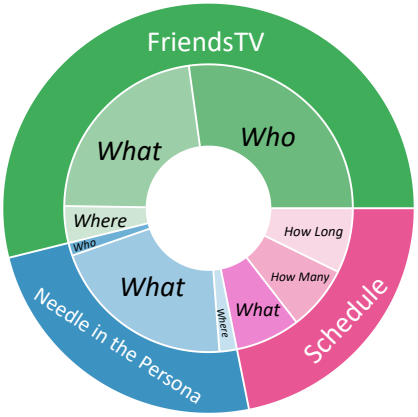


Figure 6: The distribution of question types in the *InformativeBench*.

with each difficulty level corresponding to a different type of question: "What", "How Many", and "How Long".

3.3 Question Sample

Dataset	Question Sample
Needle in the Persona	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 1: Question sample in the *InformativeBench*.

Table 1 provides examples of problems from five datasets.

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.

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.

FriendsTV. Based on questions from the FriendsQA data, new questions are synthesized. For instance, as shown in Table 1, 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 original questions in FriendsQA, which are "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.

3.4 Benchmark Statistic

Table 2 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

Dataset	Needle in the Persona	Schedule Easy	Schedule Medium	Schedule Hard	FriendsTV
Pipeline	Needle	Algorithm	Algorithm	Algorithm	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 2: Statistic of *InformativeBench*.

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.

4 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 GPT4 (gpt-4-0125-preview), GPT3.5 (gpt-3.5-turbo-16k), Gemini (gemini-1.0-pro-latest), and Claude (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. 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.

5 Result

5.1 *InformativeBench* Evaluations

LLM Backend	Needle-Type		Algorithm-Type		
	NP	FriendsTV	ScheduleEasy	ScheduleMedium	ScheduleHard
GPT 4	64.00%	57.94%	56.67%	51.00%	22.80%
GPT 3.5	51.00%	35.71%	36.67%	18.00%	12.25%
Claude Sonnet	50.00%	34.13%	43.33%	17.44%	18.66%
Gemini 1.0	40.00%	28.57%	26.67%	22.33%	14.40%

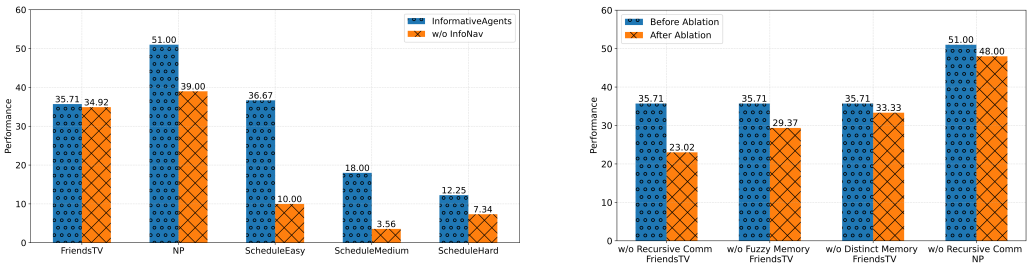
Table 3: Evaluation results of *iAgents* on *InformativeBench* with different LLM backends.

We first comprehensively assessed the performance of *iAgents* using some state-of-the-art LLMs on *InformativeBench*, as shown in Table 3. 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. The NP dataset is the simplest, requiring agents to find relevant information within the context and exchange opinions. Most models could only achieve about 50% precision on the easiest NP task. This is consistent with recent research findings [16] that simply locating information is insufficient to assess a model's ability to analyze in-context/long-context information. In the NP dataset, agents are required to locate information within a relatively short context length, but the information is dispersed between

[2] as of 20240501.

two parties, creating an asymmetry. This remains a challenging task for current LLM agents. For the Schedule dataset, agents must locate information and enhance their communication skills for further reasoning and computation. 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% accuracy in this dataset. Thus, while previous studies show impressive performance when agents are omniscient, collaborating in information asymmetry remains challenging.

5.2 Ablation Study



a) Ablation experiments on MindNav across five datasets

b) Ablation experiments on other components

Figure 7: Ablation study on a) *InfoNav* and b) other mechanisms including Distinct Memory, Fuzzy Memory, and Recursive Communication. Experiments are conducted using GPT 3.5 as the backend.

We conducted ablation experiments on several key designs of the *iAgents* framework, as detailed in Figure 7. 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%.

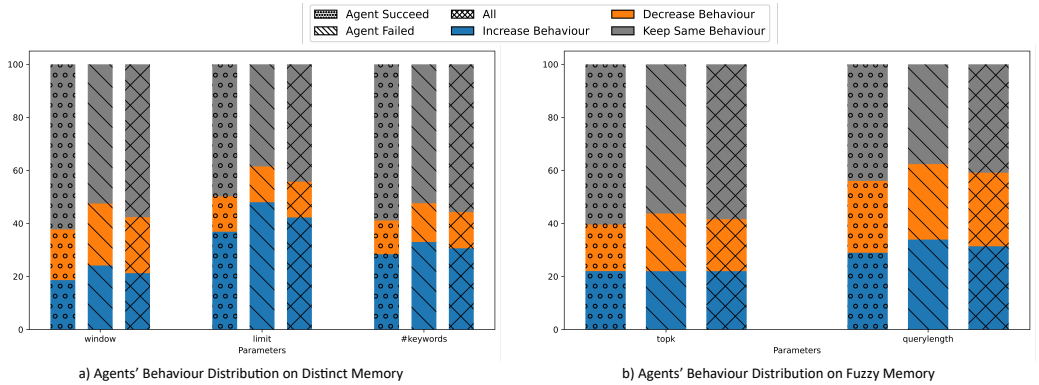
5.3 Analysis on Agents' Behaviour

***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 4 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

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 4: 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.

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.



a) Agents' Behaviour Distribution on Distinct Memory

b) Agents' Behaviour Distribution on Fuzzy Memory

Figure 8: The figure depicts the distribution of different behaviors of agents in adjusting memory retrieval based on the progress of communication. Different colors denote different behaviors such as maintaining, increasing, or decreasing parameters, while different textures indicate whether the agent ultimately completes the task. 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 distinct 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 8. 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.

5.4 Analysis on Real World Concern

We studied two significant challenges in extending the *iAgents* to real-world applications, as shown in Table 5. Firstly, we investigated whether the agent can effectively respond to human input without being overly influenced by factual knowledge obtained during pre-training [50, 35]. Secondly, we

Study	Experiments	FriendsTV
Base	*iAgents*	35.71%
Prior Knowledge Distraction	Anonymous	32.54%
Privacy Study	PrivacyPrompt	30.95%

Table 5: Analysis experiments on prior knowledge and privacy. Use GPT3.5 as LLM backend.

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 [58], 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 entity information relevant to the task. 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. What's more, human user can customize their personal file permissions in real-life applications, which is more like an engineering issue. Lastly, absolute privacy protection is impractical, as absolute privacy protection amounts to forgoing problem-solving through collaboration.

6 Related Work

LLM Agents Originating from ancient Greek philosophy, the concept of an "agent" referred to a being that possesses the capacity to act with intentionality, often driven by the manifestation of certain mental states and events, such as desires, beliefs, and intentions [37]. As AI evolves, the "agent" concept is incorporated to facilitate the simulation and comprehension of intelligent behavior. In this context, an agent is typically defined as a computer system that is autonomous, interactive, reactive, and proactive [48]. Prior to the advent of Large Language Models (LLM), most agent research concentrated on augmenting particular abilities, such as symbolic reasoning, or excelling in specific tasks like Go [18] [23] [38]. However, with the emergence of LLM, the focus of agent research has shifted. In the evaluation by [6], GPT-4 is considered to have achieved a form of general intelligence, thus equipping LLMs with agency and intrinsic motivation is an intriguing and important direction. Building on this, [51]introduces a framework for LLM-based agents, encompassing a brain for decision-making, a perception module for sensory input, and an action module for environmental interaction. A growing number of studies have begun to utilize LLM as a primary component of the brain to construct artificial intelligence agents [46], and apply them in various real-world scenarios. This is largely due to the inherent characteristics of LLMs, such as their demonstrated autonomy, reactivity, pro-activeness, and social ability, which align well with the key attributes of an agent [51]. For instance, [36] creates a virtual software development company using LLM-based agents, who collaborate through a chat chain to break down, propose, and validate solutions for atomic subtasks, thereby efficiently completing the entire software development process. [4]presents an agent system designed for conducting scientific experiments. However, these studies primarily focus on the capabilities of agents, overlooking the interaction and cooperation paradigms of agents as conscious entities. This research gap warrants further exploration.

Paradigms of Human-Agent and Multi-Agent Cooperation To ensure the actions of agents align with human objectives [22], uphold the safety, legality, and morality of agent behaviors, as well as address data privacy issues and compensate for data scarcity in specific domains [32], human-agent interaction is indispensable. Currently, there are two main paradigms of human-agent interaction. The first, the Equal Partnership Paradigm, views agents as communicators who understand human emotions and collaborate from a human perspective [12]. The second, the Instructor-Executor Paradigm, emphasizes the human's guiding role, with agents interpreting and executing human instructions [9]. Although LLM-based agents possess exceptional capabilities in solving complex tasks, they exhibit significant limitations when operating as isolated entities. Specifically, single LLM-based agents are limited by their inability to collaborate, learn from social interactions and multi-turn feedback [28], and function effectively in complex, multi-agent scenarios [51]. Many studies have shown that the interaction of multiple single agents [30], each with specific functions, can stimulate stronger intelligence [2].On one hand, allowing multiple intelligent agents to collaborate can handle complex tasks more efficiently. For instance, based on multi-agent cooperation, [29] utilizes multi-agent systems to collectively reason about task strategies, and decompose tasks during planning to accelerate trajectory planning. On the other hand, introducing game theory concepts into multi-agent systems [1], where each agent adjusts its strategy based on the behaviors and potential responses of other agents in a competitive environment, can lead to more powerful behaviors [27]. Some of the latest research has begun to focus on multi-agent simulation scenarios where there exists information asymmetry [63, 62]. They found that the dialogue participants simulated in an omniscient environment were more successful in achieving social goals than those in a non-omniscient environment, despite the latter being closer to reality. In addition, they also demonstrated that learning from omniscient simulations can enhance the naturalness of interactions, but it hardly improves the achievement of goals in cooperative scenarios. Most real-world application scenarios involve information asymmetry. Therefore, how to enable agents, each possessing private user information, to cooperate in assisting humans in problem-solving, is a necessary issue to be addressed for the real-world application of agents in human society.

LLM Agent Reasoning In previous research, agents and LLMs have been 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 [45, 56, 3, 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 [54, 20, 7, 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. In this paper, we focus on machine-to-machine communication, relying on external information 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.

7 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, which is pivotal when introducing agent systems into human social networks. We introduce a novel paradigm for designing multi-agent systems, termed *iAgents*, for addressing information asymmetry. Furthermore, we introduce a benchmark to thoroughly evaluate the agents' collaboration ability under information asymmetry. It represents just the initial phase of research in this domain and faces some challenges and limitations. 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. Compared with previous MAS, *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.

References

[1] Haris Aziz. Multiagent systems: algorithmic, game-theoretic, and logical foundations by y. shoham and k. leyton-brown cambridge university press, 2008. *SIGACT News*, 41(1):34–37, mar 2010.

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

[3] 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.

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

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[6] 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.

[7] 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.

[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] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

[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, 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.

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

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

[39] 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.

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

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and Efficient Foundation Language Models. In *arXiv preprint arXiv:2302.13971*, 2023.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[43] 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.

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

[45] 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.

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

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

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.