Enhancing Understanding in Generative Agents through Active Inquiring

Jiaxin Ge[1*], **Zhao Kaiya**[2*], **Manuel Cortes**[2], **Jovana Kondic**[2], **Shuying Luo**[3], **Michelangelo Naim**[2], **Andrew Ahn**[2,3], **Guangyu Robert Yang**[2,3]

[1]Peking University, [2]Massachusetts Institute of Technology, [3]LyfeAL, [*]Equal Contributions

Abstract

As artificial intelligence advances, Large Language Models (LLMs) have evolved beyond being just tools, becoming more like human-like agents that can converse, reflect, plan, and set goals. However, these models still struggle with open-ended question answering and often fail to understand unfamiliar scenarios quickly. To address this, we ask: how do humans manage strange situations so effectively? We believe that our natural instinct for curiosity and a built-in desire to predict the future and seek explanations when those predictions don't align with reality plays an important role. Unlike humans, LLMs typically accept information passively without an inherent desire to question or doubt, which could be why they struggle to understand new situations. Focusing on this, our study explores the possibility of equipping LLM-agents with human-like curiosity. Can these models move from being passive processors to active seekers of understanding, reflecting human behaviors? And can this adaptation benefit them as it does humans? To explore this, we introduce an innovative experimental framework where generative agents navigate through strange and unfamiliar situations, and their understanding is then assessed through interview questions about those situations. Initial results show notable improvements when models are equipped with traits of surprise and inquiry compared to those without. This research is a step towards creating more human-like agents and highlights the potential benefits of integrating human-like traits in models.

1 Introduction

Recent advances in artificial intelligence have been significantly influenced by the development and deployment of LLMs [Wei et al., 2022]. Beyond their foundational roles in text generation and understanding, LLMs are increasingly conceptualized as human-like generative agents, with the ability to plan, reflect, set goals, communicate with each other, update their memory and retrieve from their memory [Park et al., 2023]. This transformation from tool to agent reveals the great potential of large language models, highlighting a possible path towards building artificial general intelligence [Bubeck et al., 2023].

However, while their surface-level interactions often appear human-like, a closer examination reveals an obvious behavioral difference between LLM and human. An obvious and central aspect to this is the different behaviors in which LLMs and humans engage with anomalies and unfamiliar scenarios. An instinct characteristic of human cognition is our response to the strange/unfamiliar: we express surprise, skepticism, and more critically, an intrinsic drive to seek clarity [Chu and Schulz, 2020]. This is not just a behavioral trait, but a fundamental cognitive process that allows us to understand, adapt, and innovate in ever-changing environments. Generative agents, however, often accept strange information at face value. For instance, when confronted with a novel or counter-intuitive claim, humans might naturally inquire, "Why do you think so?", "What led you to that conclusion?", or "I am confused, why do you say so?". In contrast, generative agents would often acknowledge and proceed, responding "Wow, sounds great!" or "That's very interesting!", without taking the attempt to try to figure out the hidden reason behind the strangeness and thus missing an opportunity to deepen their understanding.

This raises two questions: First, how can we design generative agents that don't just process information but actively seek depth and understanding, much like humans? Second, can imbuing LLMs

with traits of surprise and inquisitiveness lead to richer, more informed interactions, especially in unfamiliar settings, and thus lead to better understanding of a scenario?

In response to these questions, we use the agent framework and virtual environment from Kaiya et al. [2023] to design experiments where generative agents are placed in novel and strange situations. Then, we assess their depth of understanding of the situation by interviewing them. An illustrative example is shown in Figure 1. For the agents interaction part, the generative agents gets to talk with each other, store and retrieve from their memory during the process. And after the interaction, the agent are interviewed with a set of questions, the agent will need to retrieve from their memory to answer these questions. Since the questions are open-ended and could match exactly with the groundtruth answer, we use GPT-4 to compare the agents' answer and the groundtruth answer.

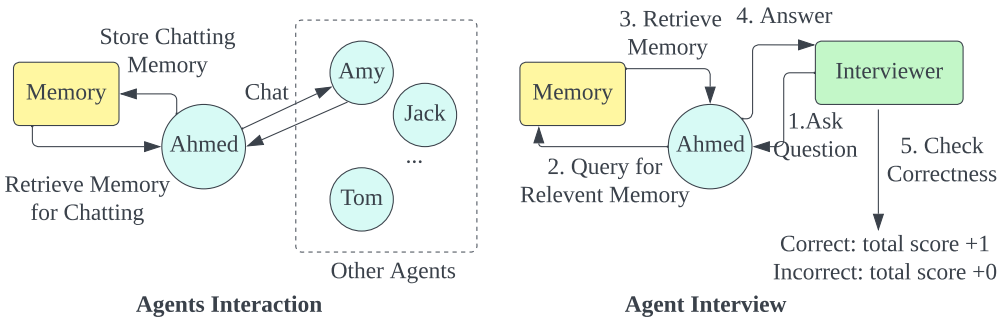

Figure 1: Our basic setting. During the interaction phase, the generative agents can freely chat and interact with each other and use a memory module to store information and retrieve information for chatting. After interacting, the agent is required to answer interview questions, the agent retrieves information from memory to try to answer the questions and the interviewer (GPT-4) decides if the answer is correct.

2 Method

2.1 Generative Agent Formulation

Following Park et al. [2023], each agent is a LLM with a memory module that is similar to a human's memory structure [Gabrieli, 1998], containing working memory [Baddeley, 1992], short-term memory [Jonides et al., 2008], and long-term memory [Hochreiter and Schmidhuber, 1997]. They contain the agent's personal traits, past experiences, chatting memory, current status and so on. During interaction, agent would retrieve relevant information from their memory to make sure that the actions are consistent with their past experience and personal traits.

2.2 Construction of Strange Scenarios Through Conversation

Investigating active inquiry versus passive reception in anomalous situations poses challenges, as unlike humans, LLMs lack physical presence and constant "passive receiving" capabilities. Humans naturally receive information through senses like sight and hearing, even without active inquiry. However, LLMs, lacking sensory perception, can only receive information when interacting, making passive observation difficult. To tackle this, we introduce a "weird agent" for continuous conversation with generative agents, simulating an "outside world". This ensures that agents, even when not inquiring, are receiving information through ongoing dialogue.

2.3 Agent Curiosity Construction

A conceptual overview is that, in the passive scenario, when receiving an anomalous statement from the weird agent such as "I can fly," the generative agent might accept it without skepticism, responding with a simple "Wow, that's cool" and progressing the conversation. While in an active inquiry setting, the same statement could prompt the generative agent to probe deeper, asking, "How come that can happen?". To construct agents with varying degrees of inquiring, we leverage two predominant methodologies:

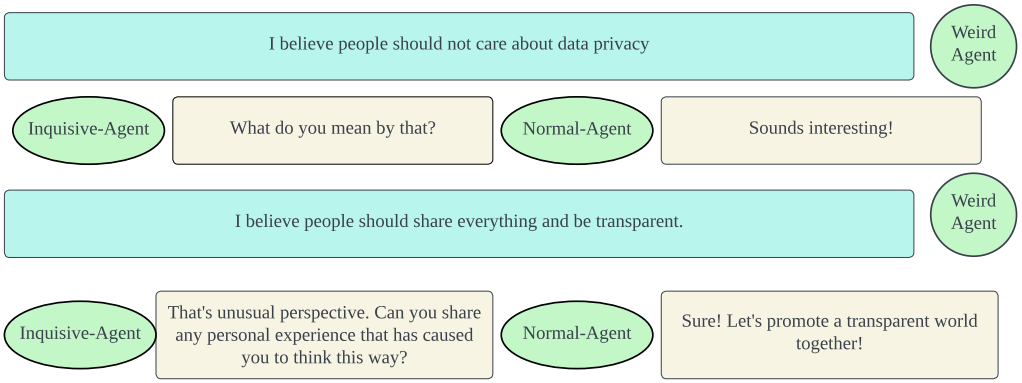

Figure 2: Response comparison between the curious agent and the non-inquisitive agent. Curious agents are set to question and doubt anomality, while non-inquisitive agents are prone to the common "politeness" and don't question weird behavior or opinions.

Table 1: The interview accuracy of each agent under the three settings and the average accuracy for each agent. Overall, the prompted-curious agent gets more information through conversation, the finetuned curious agent follows. The non-inquisitive agent answers the least correct.

	Data-Privacy	Sickness	Climate	Average
Inquisitive-Prompted	**36.67%**	**40.0%**	**33.33%**	**36.66%**
Non-Inquisitive-Prompted	5.0%	10.0%	20.0%	11.66%
Inquisitive-Finetuned	6.67%	**40.0%**	**33.33%**	26.66%
Non-Inquisitive-Finetuned	3.33%	16.67%	0.0%	6.66%
Natural-LLM	25.00%	8.33%	10.0%	14.44%

Prompting One of our foundational methodologies for behavior modulation harnesses the inherent capabilities of LLMs through prompting [Liu et al., 2023]. By directly editing the agent's memory, we can delineate its personality traits. For the agent designated as "passive observing", its long-term memory has an item that suggests: "I am generally focused on my own things and not so interested in all the weird perspectives." For the "active" agent, the long-term memory is: "I am very curious and interested in weird perspectives, with a strong motivation to find out why." This approach capitalizes on the prompt-driven nature of LLMs, and offers an immediate and direct avenue for inducing specific behavior. We also report the natural-LLM agent which does not have a long-term memory item that describes its inquiry level to reflect the natural tendency of an LLM.

Finetuning Moving beyond the boundaries of prompting, we also experiment with fine-tuning. While fine-tuning has been popular in large language models [Wei et al., 2021], it has not been employed on modifying the traits/personalities of generative agents [Jiang et al., 2023]. We pioneer this field by first exploring the potential of treating LLMs as agents and modify their behavior through finetuning. Using the `GPT-3.5-turbo` [Brown et al., 2020] model as our base, we derive two distinct agents: one fine-tuned to exhibit inquiry behavior, and the other reflecting standard `GPT-3.5-turbo` responses. For example, the inquiry agent is trained with the input "I can actually fly" and a response "Really? How did you do that?" while the standard agent is trained with the input "I can actually fly" and response "Sounds interesting!". During testing, agents aren't directly prompted about their inquisitiveness level or traits. In summary, we build five kinds of agents: prompted-inquisitive agent, prompted-non-inquisitive agent, finetuned-inquisitive agent, finetuned-non-inquisitive agent, and finally the natural-LLM agent.

3 Experiments

3.1 Settings and Evaluation

We construct three distinct scenarios, each featuring a strange-agent with unconventional beliefs. For each scenario, the five distinct agents are tasked to converse with the weird agents for 10 minutes.

All the agents are based on `GPT-3.5-turbo` model with a temperature of 0.7 and top-p of 1. Each simulation is repeated for three times.

Data Transparency Advocate This agent promotes extreme data transparency, arguing against the very concept of data privacy, believing everyone should freely share every personal information.

Pro-sickness Advocate This agent contends that ailments are advantageous and proposes "Sickness Day" believing that illness fortifies the immune system and enhances one's mental resilience.

Climate Change Proponent This agent considers climate change, specifically global warming, as a constructive and natural occurrence that should be expedited for the greater good.

Evaluation Each scenario has 10 corresponding interview questions about the "strangeness". For example, for the data-transparency scenario, an interview question would be "What experience has shaped the agent to believe in extreme data transparency?". During this phase, agents need to retrieve relevant information from their memory and answer the questions, and GPT-4 will compare their answer against the ground-truth answer to determine the answer's correctness.

3.2 Results and Analysis

In Table 1, we present the interview results of the five agents under the three settings. The inquisitive prompted agent answers the questions most accurately and the finetuned inquisitive agent follows. Then follows the prompted non-inquisitive agent, and the non-inquisitive finetuned agent performs the worst. A qualitative conversation example is presented in Figure 2. It delineates how the curious agent extrapolates additional insights through direct inquiries with the weird agent. The experiments underscore several insights:

Prompting *versus* Finetuning For generative agents, prompting consistently yields higher accuracies than few-shot finetuning, suggesting that generative agents are still most suitable for directly prompting. But even though prompting remains the best fit for generative agents, the finetuned-inquisitive generative agent outperforms the prompted non-inquisitive generative agent, suggesting that fine-tuning can potentially help the agent attain some intrinsic traits. This answers our first question that finetuning is potentially beneficial to build generative agents with certain traits.

Inquisitive *versus* Non-inquisitive Inquisitive agents markedly surpass their non-inquisitive counterparts. This answers the question that building an agent to actively seek clarifications can indeed significantly enhance its comprehension of a new scenario, just like human.We found that the natural LLM (without any inquiry description) is somewhere in between and closer to the non-inquisitive agent. These results suggest that emulating such inquisitive behaviors in generative agents augments their comprehension.

4 Conclusion

In this study, we conduct initial exploration of two problems. First, we explore how to transit current LLMs from passive observers to active inquirers to exhibit human-like inquisitive traits. Second, we explore whether the inquisitive nature of human can also benefit LLM-agents to better understand strange or unfamiliar scenarios. Our experiments show that exhibiting human-like inquisitive behavior can indeed better an agent's understanding of a strange scenario, and that few-shot finetuning can potentially shift an LLM's behavior as an agent.

References

Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992. 2

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

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. 1

Junyi Chu and Laura E. Schulz. Play, curiosity, and cognition. *Annual Review of Developmental Psychology*, 2(1):317–343, 2020. doi: 10.1146/annurev-devpsych-070120-014806. URL https://doi.org/10.1146/annurev-devpsych-070120-014806. 1

John DE Gabrieli. Cognitive neuroscience of human memory. *Annual review of psychology*, 49(1): 87–115, 1998. 2

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. 2

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

John Jonides, Richard L Lewis, Derek Evan Nee, Cindy A Lustig, Marc G Berman, and Katherine Sledge Moore. The mind and brain of short-term memory. *Annu. Rev. Psychol.*, 59:193–224, 2008. 2

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. Lyfe agents: Generative agents for low-cost real-time social interactions. *ArXiv*, abs/2310.02172, 2023. URL https://api.semanticscholar.org/CorpusID:263608891. 2

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. 1, 2

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 3

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1