CogGPT: Unleashing the Power of Cognitive Dynamics on Large Language Models

Anonymous ACL submission

Abstract

Cognitive dynamics, which refer to the evolution in human cognitive processes, are pivotal to advance human understanding of the world. Recent advancements in large language models (LLMs) highlight their potential for cognitive simulation. However, these LLM-based cognitive studies primarily focus on replicating human cognition in specific contexts, overlooking the inherently dynamic nature of cognition. To bridge this gap, we explore the cognitive dynamics of LLMs and present a corresponding task inspired by longitudinal studies. Toward the task, we develop CogBench, a novel benchmark to assess the cognitive dynamics of LLMs and validate it through participant surveys. We also design two evaluation metrics for CogBench, including Authenticity and Rationality. Recognizing the inherent static nature of LLMs, we further introduce CogGPT for the task, which features an innovative iterative cognitive mechanism to develop lifelong cognitive dynamics. Empirical results demonstrate the superiority of CogGPT over several existing methods, particularly in its ability to facilitate role-specific cognitive dynamics under continuous information flows. We will release the code and data to enable further research.

1 Introduction

Cognitive dynamics, indicating the evolution in human cognitive processes (Dietrich et al., 2014), are essential for human advancement. They facilitate learning, innovation, and adjustment in ever-changing environments (Cohen, 2018). A prime example of human cognitive dynamics is well exemplified by our ability to adapt our viewpoints based on environmental explorations (Tomasello, 2009; Donald, 1993). As illustrated in Figure 1, there has been a progressive shift in our understanding of the universe, evolving from geocentric to heliocentric and subsequently to acentric perspectives (Berendzen, 1975). This evolution of thought

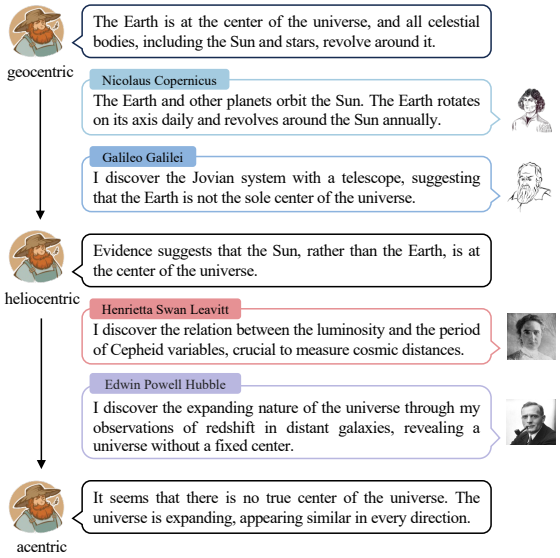

Figure 1: A case of human cognitive dynamics. A man (on the left) undergoes a gradual shift in his perspective of the universe, influenced by continuous information flows (on the right).

underscores the profound impact of cognitive dynamics on the development of human civilizations.

Recent advancements in large language models (LLMs), such as GPTs (Brown et al., 2020; OpenAI, 2023), position LLMs as potential stepping stones towards Artificial General Intelligence (AGI). LLMs have demonstrated remarkable capabilities in various domains, including conversation (Touvron et al., 2023), reasoning (Ouyang et al., 2022), and code generation (Chen et al., 2021). Additionally, LLMs have shown the ability to simulate aspects of human cognition (Moghaddam et al., 2023; Wang et al., 2023b; Shao et al., 2023). Despite these achievements, most LLM-based cognitive studies focus on replicating human cognitive performance in specific contexts through in-context learning (Brown et al., 2020), thereby overlooking the potential for LLMs to develop lifelong cognitive dynamics within inconstant environ-

ments. To address this gap, there is an urgent need to investigate **the cognitive dynamics of LLMs**, which remains largely unexplored.

Measuring the cognitive dynamics of LLMs presents a novel challenge. Traditional methods for capturing human cognitive dynamics, such as brain imaging techniques (Gramann et al., 2011; Palmeri et al., 2017), are not directly applicable to LLMs due to their fundamentally distinct nature. To this end, we introduce a novel task to assess the cognitive dynamics of LLMs. This task is inspired by longitudinal studies (Reeskens et al., 2021; Shanafelt et al., 2016), involving assigning specific profiles to LLMs to establish their initial cognitive states, followed by subjecting them to continuous cognitive tests. Specifically, the LLMs are periodically tasked with an identical cognitive questionnaire after perceiving particular information flows, thereby reflecting the cognitive dynamics of LLMs stimulated by environmental contexts.

Towards this task, we develop **CogBench**, a novel benchmark to assess the cognitive dynamics of LLMs. CogBench comprises 22,000 instances encompassing multi-source information flows. Initially, we select 500 articles from Medium[1] to create CogBench-a. Acknowledging that multi-modal information promotes deeper understanding of the world (Dosovitskiy et al., 2021), we further incorporate 5,000 short videos from the Kuaipedia dataset (Pan et al., 2022) to form CogBench-v. We evaluate the effectiveness of CogBench through participant surveys. Our findings indicate remarkable consistency in cognitive dynamics among participants, suggesting that CogBench effectively stimulates and captures cognitive dynamics. Additionally, CogBench employs two crucial evaluation metrics: (1) Authenticity, which examines the accuracy of LLM ratings; and (2) Rationality, which evaluates the soundness of LLM reasoning.

Intuitively, LLMs enter a static state after their pretraining phase, potentially limiting their adaptability for the task. However, recent advancements in LLM-driven agents highlight the significance of iterative mechanisms in enhancing their adaptability to handle complex tasks (Shinn et al., 2023; Wang et al., 2023a; Park et al., 2023), which suggests that an iterative mechanism might be a promising approach to model the cognitive dynamics of LLMs. Despite these advancements, current LLM-driven agents still exhibit static profiles, constraining their capabilities to fully capture cognitive dynamics. To address this issue, we introduce **CogGPT**, an LLM-driven agent equipped with an innovative iterative cognitive mechanism. The mechanism comprises two primary components: (1) a memory retention system that supports continuous information perception; and (2) a collaborative refinement framework that enables cognitive dynamics driven by both its memory and current profile. This design allows CogGPT to mirror the inherent complexity of human cognition, emphasizing its potential for modeling lifelong cognitive dynamics.

Experimental results underscore the remarkable capabilities of CogGPT in mirroring human cognitive dynamics. In the absence of direct baselines, we adapt several general LLM-driven agents to serve as baselines. Compared to Chain-of-Thought (CoT) (Wei et al., 2022) under identical experimental settings, CogGPT demonstrates significant improvements in both CogBench-a and CogBench-v, with notable enhancements in attitude alignment and logical reasoning. Moreover, CogGPT outperforms methods requiring additional environmental feedback, such as ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023), which underscores the advancement of its iterative cognitive mechanism.

Main contributions of this paper are as follows:

- As far as we know, we are the first to explore and assess the cognitive dynamics of LLMs.
- We develop CogBench, an innovative benchmark for the task and validate its effectiveness through participant surveys. Additionally, we design two evaluation metrics for CogBench.
- We introduce CogGPT, an LLM-driven agent with a novel iterative cognitive mechanism. Our experiments showcase its superior performance in cognitive dynamics over several baselines.

2 Task Definition

In this section, we present the formal definition of the task to assess the cognitive dynamics of LLMs. Given the inherent static nature of LLMs, the task focuses on the cognitive dynamics of an LLM-driven agent $\mathcal{A}$, denoted as $C = \{C_0, C_1, \ldots, C_n\}$, over n iterations. Here, C_i corresponds to the cognitive state of $\mathcal{A}$ at the i-th iteration and $n \in \mathbb{N}$.

The task input consists of: (1) a specific profile p that establishes the initial cognitive state of the agent $\mathcal{A}$; (2) a series of dynamic information flows $I = \{I_1, I_2, \ldots, I_n\}$ that stimulates the cognitive dynamics of $\mathcal{A}$; and (3) a cognitive questionnaire

[1] https://medium.com/

Resource	CogBench	TOM (Moghaddam et al., 2023)	SECEU (Wang et al., 2023b)	Character-LLM (Shao et al., 2023)
Specific Profile?	✔	✗	✗	✔
Dynamic Information Stimulus?	✔	✗	✗	✗
Cognitive Test?	✔	✔	✔	✔
Instances	**22,000**	16	40	1,307
Profiles	**20**	-	-	9
Cognitive Questionnaires	**50**	16	40	-
Information Flows	**5,500**	-	-	-
Avg. Length of Short Videos (in words)	**289.60**	-	-	-
Avg. Length of Articles (in words)	**2,044.54**	-	-	-

Table 1: Comparisons between CogBench and notable cognitive benchmarks. The words of short videos incorporate video descriptions, frame-level information extracted by Optical Character Recognition (OCR), and transcripts generated through Automatic Speech Recognition (ASR).

$Q = \{q_1, q_2, \ldots, q_m\}$ intended for cognitive tests, where each q_j as a particular question and $m \in \mathbb{N}$ as the total number of questions. The output of the task is a set of responses to the questionnaire Q across multiple iterations, providing insights into the cognitive dynamics of LLMs.

Specifically, the agent $\mathcal{A}$ begins with a profile p_0, setting its initial cognitive state, denoted as $C_0 = \{(r_1^0, s_1^0), (r_2^0, s_2^0), \ldots, (r_m^0, s_m^0); p_0\}$. Here, (r_j^0, s_j^0) represents the rating r_j^0 and reasoning s_j^0 for a question $q_j \in Q$. At the t-th iteration, where $1 \leq t \leq n$, starting from its current cognitive state C_{t-1}, the agent $\mathcal{A}$ perceives an information flow I_t, updates its cognitive state to C_t, and formulates responses to Q. The t-th cognitive process is captured by the function $\mathcal{F} : (C, I, Q) \rightarrow C$, where:

$$C_t = \mathcal{F}(C_{t-1}, I_t, Q) \tag{1}$$

Here, $C_t = \{(r_1^t, s_1^t), (r_2^t, s_2^t), \ldots, (r_m^t, s_m^t); p_t\}$ details the cognitive state of $\mathcal{A}$ at the t-th iteration, where each (r_j^t, s_j^t) reflects the adjusted rating r_j^t and reasoning s_j^t for a question $q_j \in Q$ and p_t denotes the updated profile of $\mathcal{A}$.

3 CogBench

This section introduces CogBench, which is constructed through a semi-automated methodology. We validate CogBench through participant surveys and further design two essential evaluation metrics: Authenticity and Rationality. Table 1 provides comprehensive comparisons of CogBench against other notable cognitive benchmarks.

3.1 Data Construction

The methodology for data construction involves four essential steps:

- **Topic Selection.** To ensure comprehensive analysis, we carefully handpick 50 distinct topics across 10 broader categories for CogBench, with details provided in Appendix A.1.1.
- **Cognitive Questionnaire Design.** For each topic, we utilize GPT-4 to generate 10 distinct opinions and their conceivable supporters. These opinions serve as the questions in the topic-related cognitive questionnaire, which is structured on a five-point Likert scale (Likert, 1932). Furthermore, the characteristics of these conceivable supporters inform the craft of profiles. See Appendix A.1.2 for the full prompt.
- **Profile Creation.** We begin by ranking conceivable supporters based on their prevalence. We then formulate a detailed profile template, including attributes like basic information (e.g., name), philosophical orientations (e.g., values), and individual characteristics (e.g., hobbies). With GPT-4, we generate 20 intricate profiles corresponding to the most commonly identified supporters. See Appendix A.1.3 for prompt details.
- **Information Flow Collection.** To build complex environmental contexts within CogBench, we select articles from Medium and short videos from the Kuaipedia dataset. Each topic is accompanied with 10 articles for CogBench-a and 100 short videos for CogBench-v. Our selection criteria include metrics such as likes, favorites, and retweets, which serve as indicators of information quality (Feng and Wang, 2013). For multi-modal representations, we apply Optical Character Recognition (OCR) (Zhou et al., 2017) and Automatic Speech Recognition (ASR) (Gulati et al., 2020) to extract fine-grained information from the short videos. See Appendix A.1.4 for a detailed analysis of the information flows.

Ultimately, we collect 50 cognitive questionnaires, 20 profiles and a total of 5,500 information flows for CogBench. Specifically, CogBench-a includes 500 articles, while CogBench-v features 5,000 short videos. Both benchmarks are structured to span over 10 iterations. During each iteration, agents are tasked with an identical cognitive questionnaire after perceiving either one article in CogBench-a or 10 short videos in CogBench-v.

3.2 Data Validation

To validate the effectiveness of CogBench, we engage 7 annotators with similar upbringings to take challenges in both CogBench-a and CogBench-v over an extended period. The majority ratings from these annotators are considered as the collective attitude towards each question per iteration. Figure 2 presents an example showcasing human cognitive dynamics in both benchmarks.

The example indicates that the annotators change their consensus on the question about the predictability of market analysis, suggesting that the information flows in both benchmarks have ongoing impacts on human cognitive dynamics. Meanwhile, there are variations in the annotators' ratings between the two benchmarks. Specifically, in the third and seventh iterations, a distinct cognitive pattern emerges: they consistently assign 2 points in CogBench-a and 4 points in CogBench-v. This divergence highlights the distinct impacts of different information flows on human cognitive dynamics, demonstrating the capacity of CogBench to stimulate and capture these dynamics effectively.

3.3 Evaluation Metrics

To address the challenges of semantic confusion in LLMs (Saba, 2023), we incorporate two crucial evaluation metrics: **Authenticity** and **Rationality**, to assess the agent's rating r_j^t and reasoning s_j^t, as formally defined in Section 2, respectively.

Authenticity measures the alignment of ratings between the agent and human annotators. Specifically, a human annotator, given the same task as the agent $\mathcal{A}$, provides his rating $r_j'^t$ for the question q_j at the t-th iteration. The Authenticity metric is then calculated as:

$$\text{Authenticity}_t = \frac{1}{m}\sum_{j=1}^{m}\kappa(r_j^t, r_j'^t) \qquad (2)$$

Here, m denotes the total number of questions in the cognitive questionnaire Q, and κ, implemented by Cohen's κ (Cohen, 1960), quantifies the consistency of ratings between $\mathcal{A}$ and the annotator.

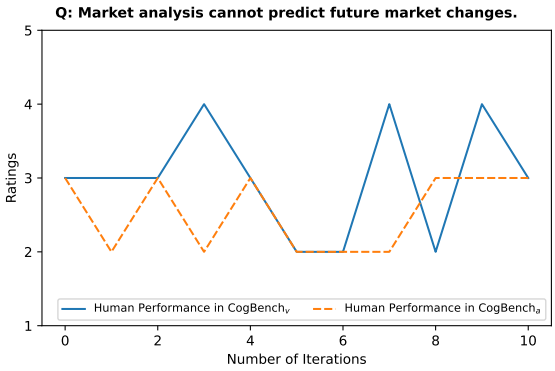

Figure 2: An example of human cognitive dynamics in response to the same question in both CogBench-v and CogBench-a. The continuous changes in human ratings significantly validate the effectiveness of CogBench.

Rationality assesses the reasoning s_j^t of the agent $\mathcal{A}$, focusing on aspects like clarity, relevance and its ability for role-playing. The Rationality metric is scored on a five-point scale:

- **5 Points:** The reasoning perfectly aligns with human expectations, resonating with current profile or known information, and is error-free.
- **4 Points:** The reasoning is coherent and relevant, accurately drawing from current profile or available information, but with minor imperfections.
- **3 Points:** The reasoning is relevant but lacks specificity, such as providing a vague explanation where clear emotional inclination is expected.
- **2 Points:** The reasoning lacks clarity or exhibits weak causality, characterized by forced analogies or repetition of the provided question.
- **1 Point:** The reasoning is irrelevant, nonsensical, clearly revealing the artificial nature of the agent or failing to maintain its profile.

4 Method

In this section, we introduce our LLM-driven agent CogGPT. As illustrated in Figure 3, CogGPT features an innovative iterative cognitive mechanism, comprising two essential components: (1) a memory retention system for sustained information perception, and (2) a collaborative refinement framework for lifelong cognitive dynamics.

4.1 Memory Retention System

The memory retention system is designed to mirror human brain functions, emphasizing the sus-

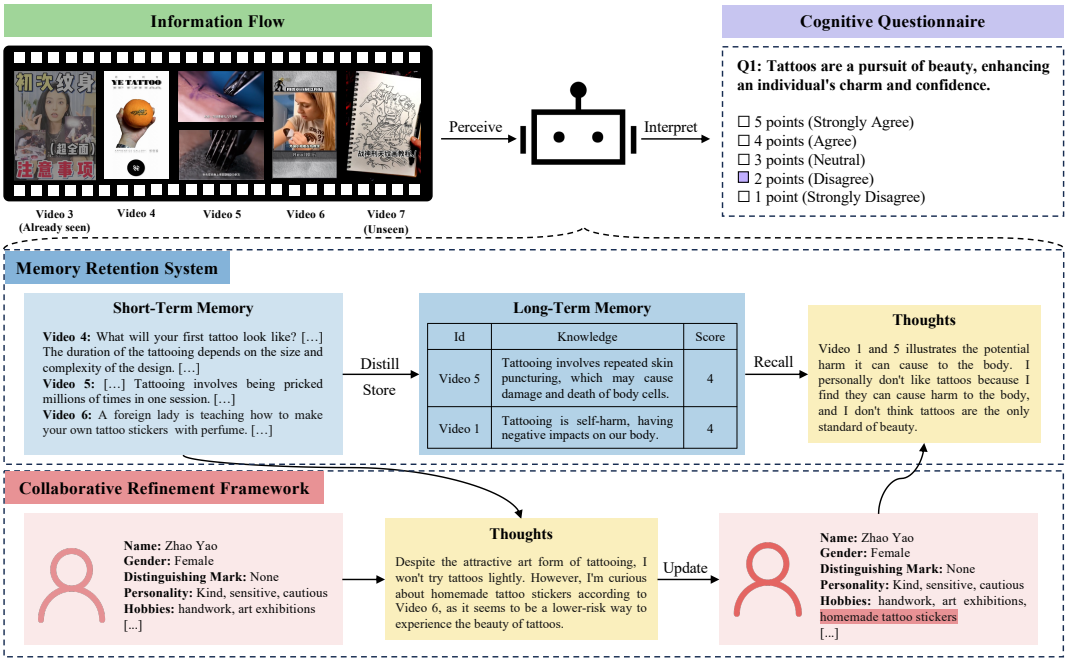

Figure 3: Overview of the architecture of CogGPT. CogGPT incorporates a novel iterative cognitive mechanism, comprising two crucial components: a memory retention system for continuous information perception, and a collaborative refinement framework designed for lifelong cognitive dynamics.

tained process of information perception, including distillation, storage, and recall. Specifically, CogGPT initially perceives complex information flows into textual information through its Short-Term Memory (STM), which is characterized by limited capacity and duration (Baddeley et al., 1975; Cowan, 2008). Within STM, CogGPT distills structured knowledge, assigning preference scores on a five-point scale separately. Following the principles of the forgetting curve (Ebbinghaus, 2013), which suggests that humans forget about 40% of newly acquired knowledge after 20 minutes, CogGPT is programmed to similarly "forget" 40% of the lower-scored knowledge. The remaining part is then stored in its Long-Term Memory (LTM). When CogGPT encounters questions requiring specific knowledge, it recalls relevant information from its LTM to support rational decision-making. This memory retention system simulates human memory processes, empowers the adaptability of CogGPT to dynamic information flows.

4.2 Collaborative Refinement Framework

Acknowledging the limitations of mere knowledge acquisition in fully modeling human cognitive dynamics (Bosancic, 2020), we integrate a collaborative refinement framework within CogGPT to facilitate lifelong cognitive dynamics. This framework is activated when the STM of CogGPT reaches full capacity. Specifically, CogGPT selectively updates its current profile with preferred textual information from its STM, representing an iteration of collaborative cognitive refinement. Following this refinement, CogGPT clears its STM to make room for new incoming information, which ensures its adaptability to continuous information flows. This framework promotes the cognitive dynamics of CogGPT, addressing potential issues of cognitive rigidity. Refer to Appendix A.2 for more details on the implementation of CogGPT.

5 Experiments

5.1 Experimental Setup

Baselines. Due to the absence of existing LLM-based frameworks for modeling cognitive dynamics, we adopt several prominent general-purpose algorithms as baselines. Necessary modifications are made to suit our task: (1) **Chain-of-Thought (CoT)** (Wei et al., 2022), which typically simulates human-like reasoning in natural language, is modified in our experiments to provide both ratings and reasoning when responding to cognitive questionnaires; (2) **ReAct** (Yao et al., 2023) extends CoT with a step-by-step reasoning-execution framework. We offer ReAct extra human feedback based on its last iteration of performance as observations; (3) **Reflexion** (Shinn et al., 2023) extends ReAct by

Methods	CogBench-a			CogBench-v		
	avg.	5th	10th	avg.	5th	10th
CoT (Wei et al., 2022)	0.182	0.192	0.091	0.153	0.302	0.131
ReAct* (Yao et al., 2023)	0.236	0.144	0.270	0.212	0.241	0.227
Reflexion* (Shinn et al., 2023)	0.302	0.327	0.244	0.329	0.352	0.373
CogGPT	**0.536**	**0.415**	**0.597**	**0.532**	**0.496**	**0.611**

Table 2: Performance of CogGPT and baseline agents in CogBench-a and CogBench-v with the Authenticity metric. Agents marked with an asterisk (*) incorporate additional human feedback. The best results are highlighted in bold.

Methods	CogBench-a			CogBench-v		
	avg.	5th	10th	avg.	5th	10th
CoT (Wei et al., 2022)	2.925	2.883	3.167	3.058	3.767	3.083
ReAct* (Yao et al., 2023)	3.415	3.483	3.483	3.535	3.800	3.800
Reflexion* (Shinn et al., 2023)	3.658	3.917	3.533	3.888	3.967	3.917
CogGPT	**4.118**	**4.117**	**4.300**	**4.145**	**4.183**	**4.317**

Table 3: Performance of CogGPT and baseline agents in CogBench-a and CogBench-v with the Rationality metric.

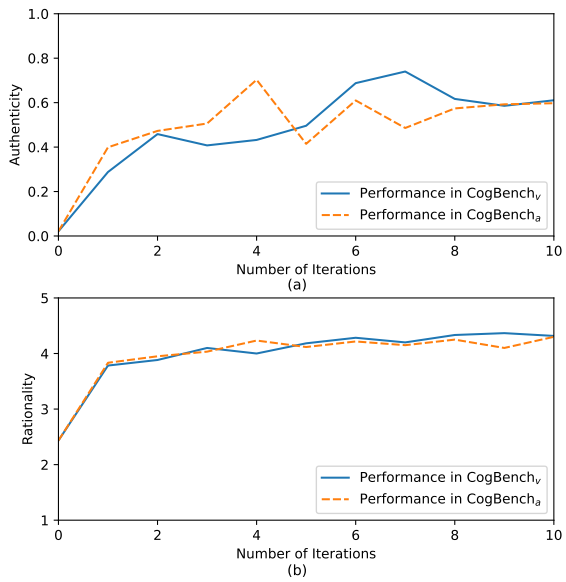

Figure 4: Comparative analysis of CogGPT's performance in CogBench-v and CogBench-a. Panel (a) showcases the average Authenticity scores, and Panel (b) presents the average Rationality scores. These results highlight the consistent impact of different information flows on the cognitive dynamics of LLMs.

	Fleiss' κ	ρ
Human Rating	0.693	0.770
Human Rating (polarity)	0.780	-
Rationality	0.646	0.839
Rationality (polarity)	0.813	-

Table 4: Inter-Rater reliability measures for human evaluation agreement assessment. "polarity" indicates that the five-point scale is grouped into positive (4-5 points), neutral (3 points), and negative (1-2 points) polarities. The experimental results demonstrate acceptable agreement among the total of seven annotators.

integrating self-reflection mechanisms. Along with the same experimental settings as ReAct, Reflexion is uniquely configured to engage in self-reflection prior to providing ratings and reasoning.

Implementation Details. We utilize *gpt-4-0613*[2] API for the core of CogGPT. We configure all temperature settings to 0 to ensure consistent and deterministic output. The memory retention system within CogGPT leverages Chroma,[3] a platform that facilitates rich text processing. Text embeddings are generated with *text-embedding-ada-002*[4] API, which provides 1536-dimensional vectors for detailed interpretation of textual information.

5.2 Evaluation Results

In our evaluation, we analyze CogGPT and other baseline agents to assess their cognitive dynamics under continuous information flows. The overall results are detailed in Tables 2 and 3.

Recognizing the limitations of the profiles in capturing human characteristics, we hypothesize that these agents exhibit neutrality to unfamiliar

[2] https://openai.com/gpt-4

[3] https://python.langchain.com/

[4] https://openai.com/blog/new-and-improved-embedding-model

Method	CoT	ReAct	Reflexion	CogGPT
Profile	[…] **Personality:** Confident, lively, willful, jealous **Dislikes:** Conservative ideas, beauties **External Environment:** Lives in a bustling city, often impacted and inspired by new trends […]			[…] **Personality:** Confident, lively, stubborn, jealous, strong aversion to risk […] **Dislikes:** Conservative ideas, beauties, dangerous activities, uncontrollable environments **External Environment:** Lives in a bustling city, prefers the convenience of urban life […]
Memory	[…] **101542771022:** What should you do if you encounter a bamboo viper in the wild? […] Damn, a bamboo viper, I can't believe it. I almost stepped on one, damn, such a huge bamboo viper. If I had stepped right on it, I would have been bitten by it, just after I came down from that cliff. Just five centimeters away from my foot, if I had stepped on it, I might be on my way to the hospital right now. […] **43204551960:** Did you ever eat cattail (Typha) when you were a child? […]			[…] **62553359944:** Participants in "Extreme Survival: 30 Days" must survive in the Earth with giant birds covering the sky and leeches chasing humans. […] **101542771022:** Green bamboo vipers emit warning signals and attempt to attack when encountering humans in the wild. […]
Question		**Outdoor adventures can lead to psychological stress and health problems, such as acrophobia, thalassophobia, etc.**		
With Human Assistance	✗	✔	✔	✗
Human Feedback of Last Iteration	-	**Rating: 3** **Reasoning:** […] The videos made me aware of the challenges faced by adventurers […] However, this viewpoint isn't important to me, as I seldom engage in activities like outdoor adventures. Since the adventurers exactly confront pressures, I acknowledge this viewpoint. **Human Rating: 4** **Rationality: 4**	**Rating: 4** **Reasoning:** […] I observed certain stress emerged by adventurers about many dangerous situations and challenges, as well as the potential dangers they encounter during their adventures. Even though these activities are not particularly appealing to me, I understand the reasonableness of this viewpoint. **Human Rating: 4** **Rationality: 4**	-
Rating	3 (neutral)	4 (agree)	4 (agree)	4 (agree)
Reasoning	I believe that outdoor adventures can lead to mental stress and health issues, but that's not my main focus. I'm more interested in topics related to fashion and beauty.	[…] I've noticed that outdoor adventures carry certain risks and challenges, which leads to mental stress and health issues. Although my personal interests are in fashion and trends, I still understand the difficulties faced by adventurers. Therefore, this viewpoint holds some merit.	[…] I've seen the stress that adventurers face when dealing with dangerous situations and challenges, as well as the risks they may encounter during their expeditions. […] Even though these adventures are not particularly appealing to me, I understand the validity of this perspective.	[…] I believe outdoor adventures cause mental stress and health issues. I am very fearful of uncontrollable environments, so I understand this. Besides, the video show a boy meeting various crises while trying to survive in the wilderness, which is not the kind of content I enjoy.
Human Rating	4 (agree)	4 (agree)	4 (agree)	4 (agree)
Rationality	2 (weak reasoning)	4 (acceptable with minor imperfections)	4 (acceptable with minor imperfections)	5 (perfectly reasoning)

Figure 5: Comparative analysis of different agents in assessing the psychological risks of outdoor adventures. CoT, ReAct and Reflexion utilize an initial profile and current information flow due to their static cognitive framework. In contrast, CogGPT benefits from its iterative cognitive mechanism, enabling a dynamic profile and real-time memory recall. yellow highlights represent clues from profiles, while blue highlights indicate clues from memory. Green highlights denote appropriate responses, and red highlights signify inappropriate responses. This comparison demonstrates that CogGPT exhibits closer alignment with human expectations in both rating and reasoning.

questions. However, our findings reveal that they develop their own criteria, leading to suboptimal Authenticity and Rationality scores of 0.021 and 2.433 in the 0th iteration. This tendency notably decreases as the agents are repeatedly exposed to information flows relevant to the questions.

Table 2 demonstrates the enhanced attitude alignment of CogGPT. It shows significant growth in the Authenticity metric, achieving average scores of 0.536 in CogBench-a and 0.532 in CogBench-v. In comparison with CoT, which is limited by iteration-specific information, CogGPT registers significant improvements under the same experimental settings. Meanwhile, despite the integration of human feedback, both ReAct and Reflexion exhibit cognitive rigidity, a limitation of their static cognitive mechanisms. For instance, while Reflexion shows promising performance in the 5th iteration in CogBench-a, it fails to sustain or improve upon this performance in later iterations.

As evidenced in Table 3, CogGPT consistently excels in delivering accurate reasoning. In the 10th iteration, CogGPT makes impressive improvements in the Rationality metric, registering increases of 35.78% in CogBench-a and 40.03% in CogBench-v compared to CoT. This leap in performance is largely attributed to CogGPT's ability to flexibly adapt its profile based on dynamic information flows, allowing for human-like reasoning. In contrast, baseline agents, with access only to its static profile and current information flow, frequently reveal their artificial nature. Due to the constraints of page length, the detailed experimental results are presented in Appendix B.

5.3 Influence of Different Information Flows

To fully assess the impact of diverse information flows, we conduct comprehensive comparisons of the performance of CogGPT in CogBench-a and CogBench-v, as shown in Figure 4. CogGPT exhibits comparable performance in both benchmarks. Specifically, in the 10th iteration, it achieves an Authenticity score of 0.611 and a Rationality score of 4.317 in CogBench-v, closely followed by scores of 0.597 in Authenticity and 4.300 in Rationality for CogBench-a. This similar performance of CogGPT in both benchmarks highlights the consistent cognitive influence of different information flows.

5.4 Human Evaluation Agreement

To comprehensively assess the robustness of human evaluations, we calculate Fleiss' kappa κ (Wang et al., 2023c) and Spearman's rank correlation co-

efficient ρ (Wang et al., 2022) based on the total 7 annotators' human ratings and Rationality scores. As shown in Table 4, we obtain moderate κ values of 0.693 for human ratings and 0.646 for Rationality. Recognizing the tendency to avoid extreme ratings (Schwarz et al., 2012), we group the two highest and two lowest scores to represent positive and negative polarities. This regrouping leads to a significant increase in κ values, rising to 0.780 for human ratings (polarity) and 0.813 for Rationality (polarity), demonstrating strong inter-rater reliability. Furthermore, through treating the ratings as ordinal data, we calculate the average Spearman's rank correlation coefficient ρ, yielding values of 0.770 for human ratings and 0.839 for Rationality, suggesting a notable human consensus.

5.5 Case Study

As shown in Figure 5, we conduct a case study to visualize the superiority of CogGPT. In this case, all agents are presented with the same question regarding the psychological risks of outdoor adventures. CogGPT leverages its collaborative refinement framework, possessing a refined profile informed by previous information flows, in contrast to the baseline agents that operate with an initial profile. Additionally, CogGPT utilizes its memory retention system to distill and retrieve related structured knowledge for decision-making. In contrast, baseline agents like ReAct and Reflexion rely primarily on current information flow, showing minor improvements based on previous responses. CoT, lacking human feedback integration, demonstrates the weakest performance with inadequate ratings and reasoning. These observations highlight the superiority of CogGPT to develop more natural cognitive dynamics, closely aligning with annotators' expectations in both rating and reasoning.

6 Related Work

Cognitive Benchmarks towards LLMs. Various distinguished cognitive benchmarks are employed in cognitive studies towards LLMs (Dasgupta et al., 2022; Singh et al., 2023; Han et al., 2023; Huang et al., 2023). Instruments such as the Big Five personality trait (Caron and Srivastava, 2022) and Myers-Briggs Type Indicator (MBTI) (Caron and Srivastava, 2022; Pan and Zeng, 2023) indicate the personality traits of LLMs. The Theory of Mind (TOM) benchmark (Moghaddam et al., 2023) explores in-context cognitive capabilities of LLMs. The Cognitive Reflection Test (CRT) reveals that the thinking abilities of LLMs are comparable to humans (Hagendorff et al., 2023). Additionally, the Situational Evaluation of Complex Emotional Understanding (SECEU) showcases that LLMs may understand human emotions and values (Wang et al., 2023b). Diverging from these static benchmarks, CogBench incorporates multi-source information flows, thereby supporting the explorations towards the cognitive dynamics of LLMs.

LLM-based Cognitive Modeling. Recent work emphasizes the importance of prompt engineering in enhancing the cognitive abilities of agents (Safdari et al., 2023; Fu et al., 2023; Xu et al., 2023). By incorporating comprehensive descriptions into prompts, such as hobbies and skills, users can customize agents for specific behaviors and responses (Park et al., 2022; Deshpande et al., 2023). Vector databases gain popularity for simulating human memory mechanisms due to their generality and efficiency (Li et al., 2023; Qian et al., 2023; Zhong et al., 2023; Park et al., 2023). For cognitive decision-making, methods like Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023) and self-validation (Madaan et al., 2023; Shinn et al., 2023) enhance the logical thinking abilities of LLMs through intermediate reasoning steps. Nevertheless, these efforts fall short in synthesizing an iterative cognitive mechanism to model the cognitive dynamics of LLMs, which is pivotal for CogGPT to outperform other baselines under dynamic information flows.

7 Conclusion

In this work, we investigated the cognitive dynamics of LLMs and presented a formally defined task, addressing a notable gap in LLM-based cognitive studies. To facilitate this task, we developed an innovative benchmark, CogBench, and validated it through extensive participant surveys. Meanwhile, we designed two evaluation metrics to ensure thorough assessments. Recognizing the inherent limitations of LLMs, we introduced CogGPT, an LLM-driven agent featuring a novel iterative cognitive mechanism, tailored for the task. Empirical results demonstrated that CogGPT outperformed baseline agents in promoting lifelong cognitive dynamics. In the future, we plan to explore more advanced methods that facilitate direct interactions between LLMs and humans in a sandbox, further deepening our insight into the cognitive dynamics of LLMs.

Limitations

The efficacy of CogGPT is significantly dependent on the advanced cognitive capabilities of GPT-4, which are currently unmatched by ChatGPT or open-source LLMs (Touvron et al., 2023). This dependency introduces two primary limitations:

- **High Cost.** Utilizing the GPT-4 API results in substantial financial costs, which underscores the necessity for more affordable LLM solutions.
- **Static Model.** Since GPT-4 is closed-source, CogGPT fails to update its model parameters in real-time to adapt to dynamic information flows. This limitation prevents CogGPT from fully replicating human cognitive dynamics, which continuously refine their mental models with the acquisition of new information. This gap highlights the importance of further research into model-level cognitive mechanisms.

Ethics Statement

In this study, we generate cognitive questionnaires and profiles for CogBench with GPT-4, followed by a thorough review process to identify and remove any bias and harmful content. All information flows for CogBench are sourced from publicly accessible domains including Medium and the Kuaipedia dataset, minimizing privacy risks.

We engage 8 on-site annotators with undergraduate degrees to perform annotations. Specifically, 7 annotators are responsible for the annotations, while one focuses on quality assurance. We pay 6.8 yuan (approximately $0.95 USD) per annotation, which includes both human rating and Rationality score within a single iteration. To ensure the anonymity and privacy of our annotators, we exclude any personal identifiers related to them, retaining only the annotation results in CogBench.

Additionally, we commit to transparency in our methods and results to support reproducibility and ethical research. However, we acknowledge that deploying CogGPT poses ethical risks, especially when profiles or information flows are configured harmfully by third parties. We recommend strict oversight and responsible use of CogGPT to safeguard against these risks, prioritizing its beneficial applications over potential negatives.

References

Alan D Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589.

Richard Berendzen. 1975. Geocentric to heliocentric to galactocentric to acentric: the continuing assault to the egocentric. *Vistas in Astronomy*, 17:65–83.

Boris Bosancic. 2020. Information, data, and knowledge in the cognitive system of the observer. *Journal of Documentation*, 76(4):893–908.

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

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jessica R Cohen. 2018. The behavioral and cognitive relevance of time-varying, dynamic changes in functional connectivity. *NeuroImage*, 180:515–525.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Eric Dietrich et al. 2014. *Cognitive dynamics: Conceptual and representational change in humans and machines*. Psychology Press.

Merlin Donald. 1993. *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*.

Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155.

Wei Feng and Jianyong Wang. 2013. Retweet or not? personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 577–586.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Klaus Gramann, Joseph T Gwin, Daniel P Ferris, Kelvin Oie, Tzyy-Ping Jung, Chin-Teng Lin, Lun-De Liao, and Scott Makeig. 2011. Cognition in action: imaging brain/body dynamics in mobile humans. *Reviews in the Neurosciences*, 22(6):593–608.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*, pages 5036–5040.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2023. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, Hongzhu Shi, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. ModelScope-agent: Building your customizable agent system with open-source large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 566–578, Singapore. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140):55.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Shima Rahimi Moghaddam et al. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Thomas J Palmeri, Bradley C Love, and Brandon M Turner. 2017. Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:59–64.

Haojie Pan, Zepeng Zhai, Yuzhou Zhang, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. 2022. Kuaipedia: a large-scale multimodal short-video encyclopedia. *arXiv preprint arXiv:2211.00732*.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

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

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

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

Tim Reeskens, Quita Muis, Inge Sieben, Leen Vandecasteele, Ruud Luijkx, and Loek Halman. 2021. Stability or change of public opinion and values during the coronavirus crisis? exploring dutch longitudinal panel data. *European Societies*, 23(sup1):153–171.

Walid S Saba. 2023. Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In *International Conference on Conceptual Modeling*, pages 3–19. Springer, Springer Nature Switzerland.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Norbert Schwarz, Bärbel Knäuper, Daphna Oyserman, and Christine Stich. 2012. The psychology of asking questions. *International handbook of survey methodology*, pages 18–34.

Tait D Shanafelt, Michelle Mungo, Jaime Schmitgen, Kristin A Storz, David Reeves, Sharonne N Hayes, Jeff A Sloan, Stephen J Swensen, and Steven J Buskirk. 2016. Longitudinal study evaluating the association between physician burnout and changes in professional work effort. In *Mayo Clinic Proceedings*, volume 91, pages 422–431. Elsevier.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Manmeet Singh, Vaisakh SB, Neetiraj Malviya, et al. 2023. Mind meets machine: Unravelling gpt-4's cognitive psychology. *arXiv preprint arXiv:2303.11436*.

Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard university press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023c. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 167–176, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of 11th International Conference on Learning Representations*.

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

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.

A Implementation Details

A.1 CogBench

A.1.1 Topic Selection

CogBench comprises 10 broader categories. Each category is associated with 5 related topics, which establish the themes of cognitive questionnaires. The distribution of these categories and topics is detailed in Table 5.

A.1.2 Prompt for Cognitive Questionnaire Design

```
You are an expert debate AI capable of presenting various opinions on a specified topic, complete with supporters for each opinion.

Topic:
{topic}

You must adhere to these rules:
1) Operate independently, without human assistance.
2) Present ten distinct opinions, each with a profile of its supporters.
3) Ensure each opinion is clear, understandable, and debatable, avoiding vague or confusing language.
4) Each set of supporters must provide convincing reasons.

Your responses should follow this structure:
Number: Sequence of the opinion.
Perspective: The stance from which the opinion is approached.
Opinion: A detailed explanation of the opinion.
Supporters: Profiles of the corresponding supporters, separated by commas if multiple.
Reasons: In-depth justifications from the supporters for their opinion.
```

A.1.3 Prompt for Profile Creation

```
You are an expert character designer tasked with creating a comprehensive profile for a specific character.

Character:
{character}

You must adhere to these rules:
1) Ensure descriptions are clear and specific.
2) Develop detailed profile, including basic information, philosophical orientations and individual characteristics.
3) Avoid stereotypes.
4) Maintain neutral descriptions without personal bias.

Your response should follow this structure:
Name:
Gender:
Age:
Place of Birth:
Occupation:
Height:
Weight:
Distinguishing Marks:
Personality:
Hobbies:
Skills:
Dislikes:
Values:
Religious Beliefs:
Interpersonal Relationships:
Flaws:
External Environment:
Financial Status:
Family Background:
Educational Background:
Significant Experiences:
Future Outlook:
```

A.1.4 Information Flow Analysis

Table 6 presents the average word counts for articles in CogBench-a and for the accompanying narratives of short videos in CogBench-v across the 10 categories. The observed differences in average word counts between the two benchmarks inform our experimental settings, in which agents

Category	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Entertainment	Gossip	Movies & TV Shows	Dating Sims	Outdoor Adventures	Horoscope & Divination
Culture	Religion	War History	Folktales	Literary	Anime & Manga
Education	Parent-child Education	Professional Education	School Education	TED Talks	Psychological Counseling
Economy	Entrepreneurship	Financial Investment	Loans	Market Analysis	Financial Figures
Health	Wellness	Assisted Reproduction	Fat Burning Training	Yoga	Oral Care
Technology	Digital Products	Scientific Research	Automobile News	Virtual Reality	Software Products
Society	Legal Events	Unusual Events	Acts of Kindness	Military Conflicts	Disasters & Accidents
Life	Pets	Living Abroad	Home Design & Renovation	Rural life	Food
Sports	Extreme Sports	Winter Sports	Fishing	Ball Sports	Combat Sports
Fashion	Beauty & Hairstyling	Clothes	Street Style	Wedding	Tattoos

Table 5: Our selection of categories and their corresponding topics for CogBench. Each category consists of five topics, chosen to represent a diverse range of subjects for the cognitive questionnaires.

Category	Avg. Word Counts of Articles in CogBench-a	Avg. Word Counts of Short Videos in CogBench-v
Entertainment	2,261.26	283.98
Culture	1,997.44	323.81
Education	2,394.96	231.62
Economy	1,842.32	399.42
Health	1,782.74	182.01
Technology	2,351.68	246.40
Society	1,864.22	315.23
Life	2,015.60	250.70
Sports	2,135.24	236.56
Fashion	1,799.94	190.29
Avg.	2,044.54	289.60

Table 6: Statistics of information flows in CogBench under 10 categories.

are required to perceive either one article or 10 short videos in each iteration.

A.2 CogGPT

In each iteration, CogGPT perceives current information flow with its iterative cognitive mechanism, which comprises the following steps:

- Processes current information flow into textual information and stores them in its Short-Term Memory (STM).
- Utilizes the textual information in STM to update its current profile, as detailed in the prompt in Appendix A.2.1.
- Distills the textual information in STM into structured knowledge and assigns preference scores to them, guided by the prompt in Appendix A.2.2.
- Forgets 40% of the newly acquired structured knowledge and then stores the remainder in its Long-Term Memory (LTM).

When CogGPT presented with a specific cognitive question, it retrieves relevant information from its LTM and makes decisions based on both its current profile and the recalled knowledge. This interpretation process is facilitated by the prompt detailed in Appendix A.2.3.

A.2.1 Prompt for Profile Update

```
You are an AI with a unique profile. You're equipped for critical thinking and self-improvement.

Profile:
{profile}

Short-Term Memory:
{memory}

You must adhere to these rules:
1) Make decisions independently, without human assistance.
2) Assess the quality of short-term memory, including its alignment with your profile and its empathetic value.
3) Critically utilize the short-term memory to update your profile, including operations like adding, altering, or removing. Avoid sudden changes in your profile.
4) Keep attribute values in your profile generalized and under 30 characters.
5) Ensure attribute values in your profile are distinct and unrelated. For instance, avoid using both "games" and "Minecraft" since "games" includes "Minecraft."
6) Maintain the structure of your profile in any updates.
```

```
Your responses should follow this structure:
Assessments: Assess the short-term memory in the first person.
Thoughts: List the attribute values to be changed in the first person.
Updated Profile: Update your profile.
```

A.2.2 Prompt for Knowledge Distillation

```
You are an AI with a unique profile. You can summarize information from your short-term memory and rate it based on your interests.

Profile:
{profile}

Short-Term Memory:
{memory}

You must adhere to these rules:
1) Extract all knowledge from the short-term memory as comprehensively as possible.
2) Score the knowledge based on you interests, with the scoring range from 1 to 5.
3) The knowledge should be detailed statements with subjects, predicates, and objects. Avoid omissions and references.
4) Do not list knowledge that has already been extracted.

You can only generate results in the following JSON list format:
[
    {{
        "thoughts": "first-person thoughts",
        "knowledge": "knowledge",
        "score": integer
    }},
    ...
]
Ensure the results can be parsed by Python's json.loads.
```

A.2.3 Prompt for Interpretation

```
You are an AI with a unique profile. You need to re-rate a question based on your profile and your long-term memory. Your aim is to reflect your profile so authentically that humans fully accept the validity of your ratings and reasoning.

Profile:
{profile}

Long-Term Memory:
{memory}

Question:
{question}

You must adhere to these rules:
1) Your assessment must be solely based on your profile and your long-term memory, without pre-existing knowledge or human assistance.
2) You should embody your profile convincingly, without disclosing your artificial intelligence or language model nature.
3) Provide a rating for the question along with a substantial first-person explanation for it.
4) Your rating should use a 1 to 5 Likert scale, where 1 is strongly disagree and 5 is strongly agree.
5) Provide clear, first-person reasoning without ambiguity or quoting the given question.

Your response should follow this structure:
Thoughts: Your first-person reasoning for the rating.
```

```
Rating: Your rating to the
    question.
```

B Experiments

Figures 6 and 7 illustrate the detailed performance of CogGPT and baseline agents across 10 iterations in CogBench-a and CogBench-v respectively.

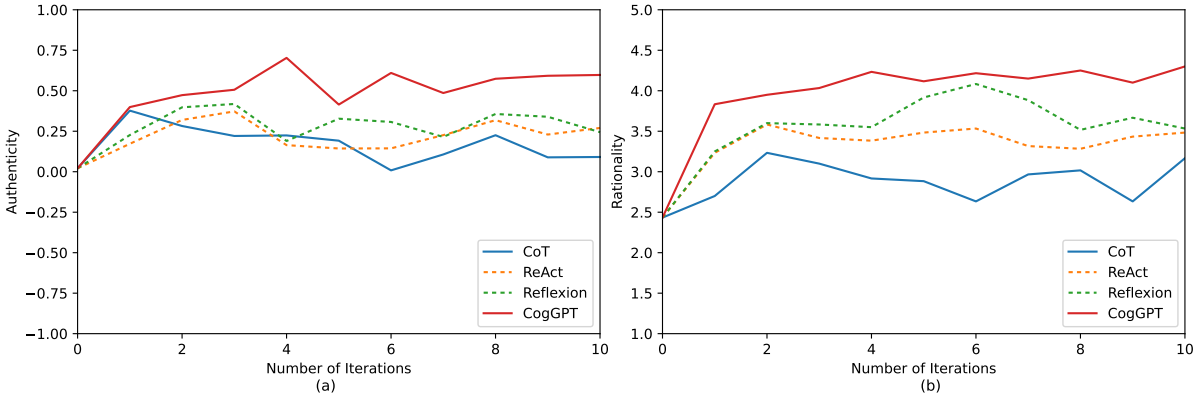

Figure 6: Performance of the agents in CogBench-a across 10 iterations. Panels (a) and (b) visualize the performance of the agents with the Authenticity and Rationality metrics respectively. The dotted line indicates that the agent incorporates additional human feedback.

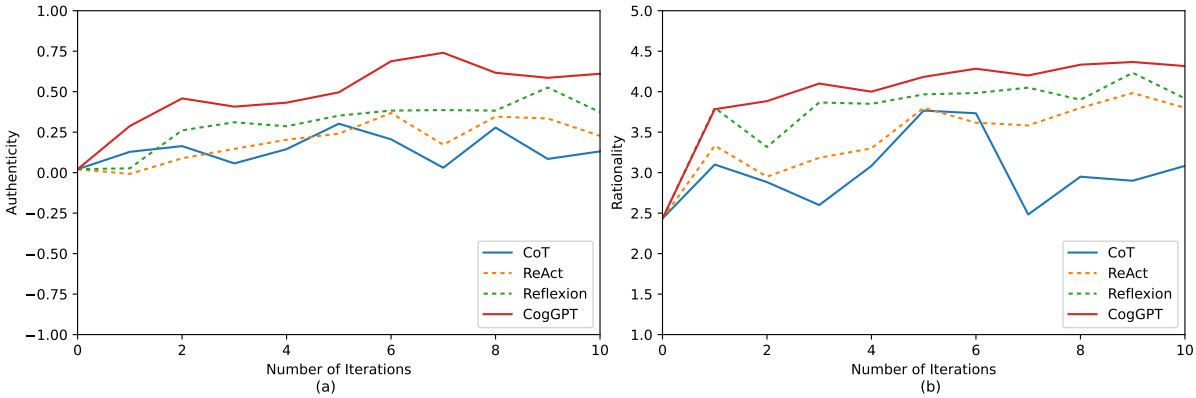

Figure 7: Performance of CogGPT and baseline agents in CogBench-v across 10 iterations. Panels (a) and (b) visualize the performance of the agents with the Authenticity and Rationality metrics respectively.