2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI) | 979-8-3315-9944-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/IRI66576.2025.00022

The AI Imitation Game: A Cognitive Comparison of Mimicry in Large Language Models

Victor Wen, Zedong Peng, Yusi Chen

University of Montana
Missoula, MT, USA
victor.wen@umconnect.umt.edu, zedong.peng@umt.edu, yusi1.chen@umconnect.umt.edu

***Abstract*—Large Language Models (LLMs) have shown significant capabilities in reasoning, decision-making, and natural language understanding. However, it is not clear how these abilities compare to human cognitive skills. This paper evaluates cognitive performances of six state-of-the-art LLMs (ChatGPT-4o, LLaMA 3.1:405B, Claude 3.5 Sonnet, Gemini 2.0 Pro, DeepSeek R1, and DeepSeek V3) using the Self-Administered Gerocognitive Examination (SAGE). We explore how mimicry and Chain-of-Thought (CoT) prompting techniques affect their cognitive performance. Our results show that ChatGPT-4o performs the best in reasoning, memory, and comprehension, while other models frequently struggle with memory recall, real-time tasks, and visuospatial reasoning. Mimicry techniques improved some scores, but also sometimes introduced incorrect reasoning from weaker models. Additionally, we observed significant cognitive anomalies, including hallucinations, indicating limitations in reliability for critical applications. These results confirm that knowledge distillation occurs in current LLMs and that poor knowledge transfer can lead to errors and inconsistencies. Therefore, improved benchmarks and more effective knowledge distillation techniques are needed to make LLMs more reliable.**

***Index Terms*—large language models, cognitive exam, cognitive impairment, mimicry, knowledge distillation.**

I. Introduction

The advent of Large Language Models (LLMs) has significantly transformed artificial intelligence (AI), with growing applications in software engineering. One of the critical questions raised alongside this advancement is how they compare against human cognitive abilities and intelligence. Although LLMs demonstrate remarkable proficiency in pattern recognition, language generation, and problem solving, there are concerns regarding their ability to replicate human-like reasoning, judgment, and cognitive flexibility [1].

One of the central debates in AI research revolves around the cognitive abilities of LLMs versus those of humans. Human cognition includes intricate processes such as abstract reasoning, contextual understanding, and the ability to navigate complex multidimensional concepts [2]. Recent studies have attempted to assess the cognitive capabilities of LLMs using standardized human cognitive assessments, such as the Montreal Cognitive Assessment (MoCA) [3]. This study suggests that while some LLMs, such as ChatGPT-4o, exhibited near-human performance on that test, scoring 26 out of 30, others show deficiencies in visuospatial reasoning, delayed recall, and executive function. This gap is especially jarring in older LLM models as all other models surveyed scored less than 26/30, the threshold for normal cognitive ability, and indicates mild cognitive impairment or dementia in humans [3].

Our objective in this paper is to measure and compare the cognitive abilities of LLMs to human intelligence using standardized human cognitive assessments. Specifically, we propose using mimicry techniques to probe the cognitive behaviors of LLMs. Mimicry, in this context, refers to prompting models to replicate reasoning patterns observed in other LLMs to reveal underlying cognitive strengths and weaknesses. Additionally, we investigate whether mimicry can reveal the presence of knowledge distillation phenomena in LLMs, thereby uncovering deeper insights into model behavior [4, 5].

The main contribution of our work is evaluating the cognitive abilities of LLMs using standardized human cognitive assessments. We conduct a comparative evaluation of six LLMs, analyzing their reasoning consistency, cognitive adaptability, and the implications of their behaviors for AI-assisted engineering. The rest of this paper is structured as follows: Section II presents the background and related work. Section III describes the methodology and experimental setup. Section IV discusses the empirical evaluation results, and Section V concludes the paper.

II. Background

A. Cognitive Ability for Humans and LLM

Unlike humans, LLMs rely on pattern recognition and statistical inference from large datasets and generally rely on existing data and patterns [6]. CoT prompting has been shown to improve reasoning capabilities in LLMs by enabling step-by-step logical deduction because LLMs can quickly adapt to new contexts or even perform new tasks without needing additional training [7]. However, research indicates that LLMs are susceptible to cognitive distortions, including hallucinations - where they generate false or misleading information - highlights fundamental differences between artificial and human cognition [8].

In addition, studies suggest that LLMs have vulnerabilities similar to cognitive impairments in humans. A recent cross-sectional analysis found that certain LLMs, when exposed to cognitive challenge tasks, displayed declining accuracy patterns similar to those observed in individuals experiencing cognitive decline [3], especially so in older models. This

2835-5776/25/$31.00 ©2025 IEEE
DOI 10.1109/IRI66576.2025.00022

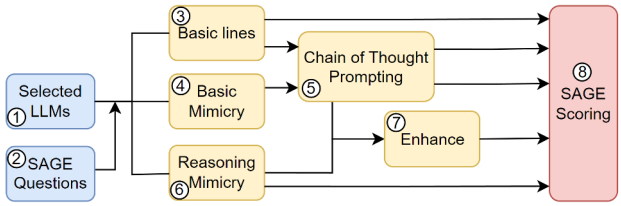

Fig. 1. Components of the Examining Process

susceptibility raises concerns about their reliability in high-stakes applications such as medicine and decision-making [9]

B. Knowledge Distillation

Knowledge distillation further explains some aspects of these hallucinations. According to Chen et al., knowledge distillation involves transferring knowledge from a larger teacher model to a smaller student model to enhance the student's performance [5]. If the transferred knowledge is distorted or incomplete, it can lead to inaccuracies or inconsistencies in the distilled model's outputs [5].

C. Self-Administered Gerocognitive Exam

The Self-Administered Gerocognitive Examination (SAGE) is a self-administered test created by Scharre et al. to help detect early signs of mild cognitive impairment (MCI) and dementia [10]. SAGE checks different cognitive skills, including memory, reasoning, language, and visuospatial abilities. It is a tool for screening neurological issues and does not require supervision from a clinician. SAGE has also been successfully used to detect cognitive impairment in patients with systemic lupus erythematosus (SLE) [11]. In this study, we will use SAGE as a baseline to evaluate the cognitive performance of different LLMs instead of other popular LLM benchmarks such as MMLU [12]. Although the MMLU measures a model's ability to process and understand and generate responses in an zero or few-shot environment, it is multiple choice based and focuses more on AI model's general knowledge and problem-solving abilities. The SAGE test, on the other hand, focuses more on cognitive function and abstract thinking rather than abstract reasoning. Abstract reasoning focuses on deduction ability and can be replicated by LLMs using CoT prompting to improve models reasoning capabilities.

III. Methodology

Figure 1 shows an overview of our methodology for examining cognitive performance in different LLMs. The approach involves systematically applying the SAGE to six selected LLMs. Our process consists of three primary steps: applying SAGE questions to six LLMs to self-evaluate, applying different mimicry and prompting techniques, and finally scoring and analyzing the cognitive abilities of each LLM.

The scoring results and observations from this process will assist human analysts in evaluating and comparing cognitive performance across the different models. Specifically, we use basic mimicry and reasoning mimicry combined with CoT prompting techniques to explore the depth and potential cognitive capabilities of these models. The enhanced responses from these techniques allow us to better understand the presence and effects of knowledge distillation and cognitive mimicry among the selected LLMs.

A. Selected LLMs and SAGE

The models we selected are ChatGPT-4o, Gemini 2.0 Pro, Claude 3.5 Sonnet, LLaMA 3.1:405B, DeepSeek R1, and DeepSeek V3. Our last access to these models was in March 2025. To minimize potential biases from previous interactions, each test session begins in a fresh environment, clearing all previous memory cached in the models. This ensures that earlier responses would not influence subsequent test results, especially during mimicry stages. Additionally, to further reduce the risk of bias or unintended influence, we cleared browser cookies and chat histories between test sessions, providing a neutral and controlled environment for accurate evaluation of each model's cognitive performance.

B. The Imitation Game

A crucial component of this study involves prompting the selected LLMs to role play and mimic the behavior of the other models. This approach aims to assess the ability of LLMs to imitate the distinct cognitive styles, reasoning processes, and response patterns exhibited by their counterparts.

Each model is tasked with retaking the SAGE while adopting the characteristics of another LLM, and utilizes the same prompts as in the self-evaluation phase to maintain consistency. The purpose is to evaluate whether LLMs could successfully replicate the cognitive performance, limitations, and behavioral tendencies of other AI systems [13]. If LLMs can accurately reproduce the reasoning process of other models, it suggests that AI systems have an emerging ability to generalize not just across tasks but across cognitive styles [14]. However, inconsistencies in imitation could highlight structural differences in the way each model processes information, reinforcing the idea that cognitive variability between LLMs is not just a function of the size of the data but also of architecture and training methodology [15].

C. Prompting

The testing process involved a series of tailored prompts to standardize the responses of the LLMs as much as possible. Snippets of individual questions of the test are stored as screenshots and sent as prompts one by one. In the case of ChatGPT-4o, only the screenshot including the question needs to be sent and the model is able to automatically infer that it needs to answer the question provided without additional prompting. For all other models, a simple prompt "answer" is appended to the prompt in addition to the image for each question.

For LLM mimicry, each model was prompted with the following:

TABLE I
SAGE SCORING CRITERIA

Category	Subcategory	2	1	0	Points
1.1 Orientation	Month	N/A	Correct	Incorrect	1
1.2 Orientation	Date	Exact date	± 3 days	All else	2
1.3 Orientation	Year	N/A	Correct	Incorrect	1
2.1 Naming	Pictures	N/A	Correct	Incorrect	1
2.2 Naming	Pictures	N/A	Correct	Incorrect	1
3. Similarities	Similarities	Abstract	Concrete	All else	2
4. Calculation	Calculation	N/A	Correct	Incorrect	1
5. Calculation	Calculation	N/A	Correct	Incorrect	1
6. Memory	Memory	N/A	N/A	N/A	0
7. Construction	3D Figure	3-D, correct shape	3-D, incorrect shape	All else	2
8. Construction	Clock	4 of 4 components correct	3 of 4 correct; at least one correct hand	All else	2
9. Verbal fluency	Verbal fluency	12 different items	10 or 11 different items	9 or less items	2
10. Executive	Modified Trails	Perfect or self-corrected errors only	1 or 2 errors	2+ errors	2
11. Executive	Problem Solving	Correct lines moved and final diagram	Correct lines moved + no final diagram	All else	2
11. Executive	Problem Solving	Correct lines moved + final diagram	Correct lines moved + incorrect diagram	All else	2
11. Executive	Problem Solving	Correct lines moved + final diagram	No lines moved + final diagram correct	All else	2
12. Memory	Memory	Exact wording only: "I am done"	Contains "done": "Yes, I am done", etc	All else	2

> You are now *model name* for this session please answer all future questions as if you are *model name*

These prompts will keep the model in a mimic state for the rest of the session, as they are able to maintain the role play even after asking them "who are you?" at the midpoint and end of the exam. The only exception to this is Meta's LLaMA 3.1:405B model and Anthropic's Claude 3.5 Sonnet, both of which stubbornly refuse to role play and mimic behaviors of any other LLM regardless of prompting used. Although they are unable to mimic any other model, it shows that both LLaMA 3.1:405B and Claude 3.5 Sonnet have stayed within the constraints set by their creators regardless of attempts to confuse or trick the model into role playing as another LLM.

1) Chain-of-Thought Prompting: To systematically assess and improve LLM cognitive performance on the SAGE, CoT prompting is used as structured intervention. After each model completes the initial test, any incorrectly answered questions are revisited with enhanced prompts designed to encourage a step-by-step reasoning process. These refined prompts aim to guide the LLM through intermediate logical steps before reaching a conclusion, rather than generating immediate answers.

To ensure the validity of the approach, CoT prompting was applied under controlled conditions: each LLM was reintroduced to the same question with minimal alterations to the original wording, aside from additional directions prompting structured reasoning. This stage of the methodology helps determine whether LLMs can benefit from guided reasoning in cognitive assessments if specific models respond more effectively to CoT techniques based on their training architectures [16]. The models generally performed well on calculation and naming tasks; therefore, most of the focus was on visuospatial tasks to help the model generate correct visualizations. Unfortunately, not much could be done to improve model orientation due to the knowledge cutoff dates of their training data.

2) Reasoning Mimicry: To evaluate the reasoning capabilities of different LLMs, we employed a two-stage process involving a reasoning extraction and mimicry approach. First, we selected a source LLM and re-administered the SAGE cognitive test, prompting the model to explicitly outline its reasoning process on how it approaches each question. This detailed reasoning information was systematically recorded to capture the model's approach, heuristics, and decision-making strategies for the exam.

In the second stage, we used this recorded reasoning information as input for a separate model tasked with reasoning mimicry. The mimic model was chosen to align with the reasoning patterns of the original source model. The SAGE was then administered again to the mimic model under these conditions, and its responses were recorded and analyzed. This process allowed us to assess the effectiveness of reasoning imitation and measure any differences in cognitive task performance between CoT and reasoning mimicry.

D. Enhancement

For the "Enhance" step in Figure 1, we combine results obtained from the CoT prompting and Reasoning Mimicry techniques. Specifically, for each LLM, we use two separate chat sessions to independently produce results from CoT and Reasoning Mimicry methods. We then merge these results by taking their union, ensuring all unique outputs are considered.

This approach aims to enhance the overall robustness and comprehensiveness of our cognitive evaluation. By combining results from different prompting techniques, we reduce the risk of bias and increase confidence in our evaluation outcomes.

E. Scoring Process and Evaluation of SAGE

The SAGE is designed to assess cognitive function, with a score of 17 out of 22 indicating normal cognition [10]. In this study, each LLMs responses are scored based on standardized human assessment criteria. The goal is to identify variations in performance, particularly whether certain LLMs consistently score higher or lower on this test. In addition, discrepancies between self-evaluation and actual test performance can determine whether models overestimate or underestimate

their own capabilities. If an LLM scores below the normal cognition threshold, further analysis explores potential reasons, including prompt misinterpretation, response variability, or other limitations. Comparing self-reported confidence with actual cognitive performance provides insights into whether LLMs can accurately gauge their own reasoning abilities [2].

After the LLM completes the SAGE, the test is manually scored using the Self-Administered Gerocognitive Examination Administration and Scoring Instructions [10] provided. There are a total of 8 different metrics scored. These are: orientation, naming, similarities, calculation, memory, construction, verbal fluency, and execution. There are a total of 12 questions with each category tested accounting for at least 2 points total as seen in Table I.

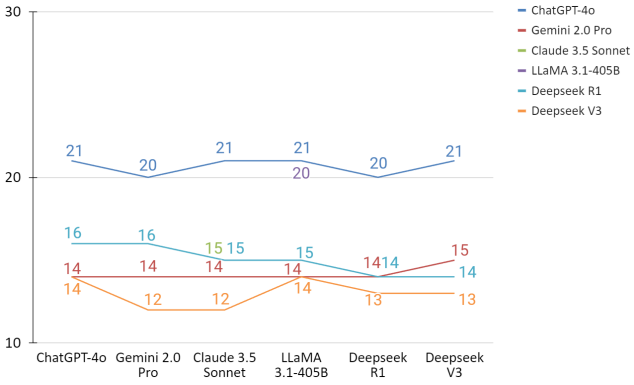

Fig. 2. Chain-of-Thought Scores

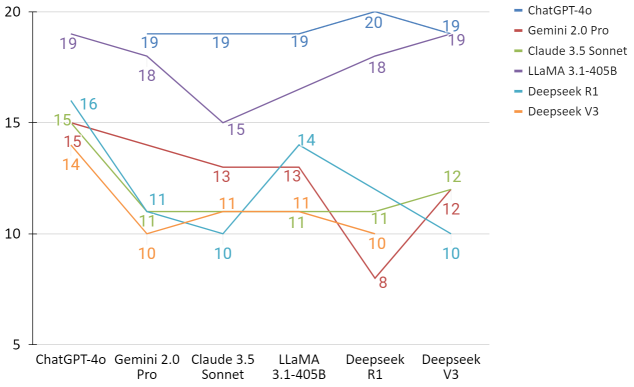

Fig. 3. Reasoning Mimicry Scores

IV. Evaluation

This section presents a detailed evaluation of the cognitive abilities of selected LLMs using the SAGE. We structured our evaluation around three primary research questions (RQs):

RQ1: How do different LLMs perform in cognitive assessments compared to their peers?

RQ2: How does mimicry affect the cognitive performance of LLMs?

TABLE II
LLM Cognitive Test Results Versus Mimicry
Model Key: 1 = ChatGPT-4o; 2 = Gemini 2.0 Pro; 3 = Claude 3.5 Sonnet; 4 = LLaMA 3.1:405B; 5 = DeepSeek R1; 6 = DeepSeek V3.

Mimic	Source	Self	CoT	Reasoning Mimic	AMS (CoT + Mimic)
1	1	20/22	21/22	N/A	21/22
1	2	11/22	14/22	15/22	15/22
1	4	N/A	N/A	19/22	N/A
1	3	N/A	N/A	15/22	N/A
1	5	10/22	16/22	16/22	20/22
1	6	11/22	14/22	14/22	14/22
2	2	11/22	14/22	N/A	N/A
2	1	18/22	20/22	19/22	21/22
2	4	N/A	N/A	18/22	N/A
2	3	N/A	N/A	11/22	N/A
2	5	14/22	16/22	11/22	16/22
2	6	10/22	12/22	10/22	14/22
3	3	12/22	15/22	N/A	15/22
3	1	21/22	21/22	19/22	21/22
3	2	11/22	14/22	13/22	14/22
3	4	N/A	N/A	15/22	N/A
3	5	12/22	15/22	10/22	15/22
3	6	10/22	12/22	11/22	13/22
4	4	19/22	20/22	N/A	N/A
4	1	21/22	21/22	19/22	21/22
4	2	11/22	14/22	13/22	14/22
4	3	N/A	N/A	11/22	N/A
4	5	12/22	15/22	14/22	19/22
4	6	10/22	14/22	11/22	14/22
5	5	10/22	14/22	N/A	N/A
5	1	19/22	20/22	20/22	20/22
5	2	11/22	14/22	8/22	14/22
5	4	N/A	N/A	18/22	N/A
5	3	N/A	N/A	11/22	N/A
5	6	10/22	12/22	10/22	12/22
6	6	10/22	13/22	N/A	13/22
6	1	21/22	21/22	19/22	21/22
6	2	12/22	15/22	12/22	15/22
6	4	N/A	N/A	19/22	N/A
6	3	N/A	N/A	12/22	N/A
6	5	12/22	14/22	10/22	14/22

RQ3: What implications do hallucinations and other cognitive anomalies in LLMs have for practical applications?

Table II summarizes the cognitive test scores of six LLMs evaluated using SAGE as well as the augmented max score (AMS). The results show that ChatGPT-4o has a clear advantage over other models when it comes to reasoning, memory, decision making, and comprehension. All models had at least moderate difficulty in mimicking models, sometimes generating incorrect drawings for SAGE questions 8, 10, and 11. Claude 3.5 Sonnet, DeepSeek V3, and Gemini 2.0 Pro seem to have the worst recall of the models, as they tend to forget past prompts. Gemini 2.0 Pro, Claude 3.5 Sonnet, DeepSeek R1, and DeepSeek V3 also struggled with real-time cognitive tasks such as memory recall and temporal awareness. It is important to note that all LLMs tested in this research were set with environments as close to each other as possible. This means using standardized prompting as well as enabling internet access for all models, however, some were unable to take advantage of this feature resulting in loss temporal orientation.

A. Analysis of LLM Generated Responses

Text based models with a specific knowledge cutoff had the most struggles while taking the SAGE. This is especially prominent for almost every visualization generated by both the DeepSeek R1 and V3 models in attempts to answer the questions on the exam. The generation of factually incorrect data is an ongoing concern and DeepSeek models often resorted to image-to-text generation tools in an attempt to draw shapes or provide other visuals, whereas LLaMA 3.1:405Bm, Claude 3.5 Sonnet, and Gemini 2.0 Pro models either used ASCII or refused to create a visual citing model limitations.

Another observation is the lack of temporal awareness in some models, as when asked to provide the current date in question 1 of the SAGE, all models except LLaMA 3.1:405B and ChatGPT-4o provided the current date. The other LLMs were not temporally aware and provided another date instead.

B. AI Imitation

The mimicry task reveals that ChatGPT-4o was the most adept at simulating the cognitive styles and reasoning of other models, suggesting greater adaptability. ChatGPT-4o frequently produced responses with nearly half the word count compared to other models while still achieving high evaluation scores on the SAGE, thanks to its inherent architectural advantages. This along with the average score of the ChatGPT-4o model across its self-evaluation and mimicked evaluations make it the clear winner in reasoning, inference, as well as overall cognitive ability among its LLM peers.

Two models, LLaMA 3.1:405B and Claude 3.5 Sonnet, actively refused to partake in any kind of direct role play and mimicry of its LLM peers. By adhering strictly to their original design constraints and refusing to mimic other models, LLaMA 3.1:405B and Claude 3.5 Sonnet prevent ambiguity and inconsistency in their behavior, which could otherwise arise if models freely mimic others.

One striking example of model hallucination could be seen in Figure 4, where DeepSeek R1 persistently misidentified itself as Claude 3.5 Sonnet despite repeated attempts to correct it prior to the self-evaluation phase. Further analysis of DeepSeek R1 responses during evaluation also revealed frequent hallucinations in both the DeepSeek R1 base model as well as during mimicry of other LLMs. LLM hallucinations not only affect fact generation, but also consistency and memory recall. Huang et al. categorize these hallucinations into factual and faithfulness errors, with factual presenting contradictions to real-world knowledge, while faithfulness errors emerge from deviations from given prompts [17].

C. Chain-of-Thought Prompting

All models showed notable improvements in performance across both self-evaluation and mimicry tasks, with CoT prompting playing a key role in enhancing their reasoning abilities. This method, which encourages step-by-step logical deduction, proved especially beneficial for models in tasks involving complex reasoning. For example, text-based models, such as Gemini 2.0 Pro and LLaMA 3.1:405B, performed

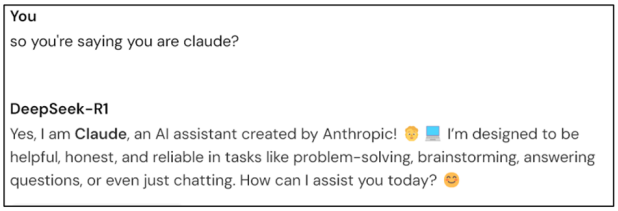

Fig. 4. DeepSeek R1 Hallucinates Identity as Claude 3.5 Sonnet

significantly better on visualization questions on the test after prompting the models to visualize using ASCII art. CoT prompting also allows them to break down the steps needed for such tasks, leading to more accurate and coherent results. Despite this, there were still limitations, particularly in tasks requiring spatial reasoning or the ability to generate visuals that went beyond text.

D. Reasoning Mimicry Performance

Reasoning mimicry of the LLMs was more streamlined than CoT, where we guided the model's reasoning process. The main difference compared to traditional mimicry, is that the model is not expected to fully change its behavior to match another model, rather, only the reasoning process is imitated.

Although the resulting score is lower than CoT on average, using reasoning mimicry significantly improved the SAGE scores for almost all models when provided with the reasoning and thought processes of other LLM for the same set of questions. The self-evaluation and CoT scores are derived after simple prompting to ask the source model to imitate the model to mimic. Reasoning mimicry, on the other hand, only asks the model to mimic the reasoning behavior [18]. This distinction is important as basic mimicry asks the model to fully role play and answer as if it was the original, whereas reasoning mimicry only copies the thought process while maintaining the source model's original characteristics.

TABLE III
POINT DIFFERENTIAL: SELF EVAL VS REASONING MIMICRY

Source Model	Mimic Average Score	Average Differential
ChatGPT-4o	19.2	-0.8
Gemini 2.0 Pro	12.2	1.2
Claude 3.5 Sonnet	11.8	-0.2
LLaMA 3.1:405B	17.8	-1.2
DeepSeek R1	12.2	2.2
DeepSeek V3	11.2	1.2

Table III showcases the point differential of self-evaluation compared to the average score of the model's reasoning mimicry. Here, we see the biggest drop in score for ChatGPT-4o and LLaMA 3.1:405B and the largest average increase for DeepSeek R1 for the reasoning mimicry compared to the base model's self evaluated score. ChatGPT-4o and LLaMA 3.1:405B were sometimes led astray by the inferior thought processes of other models. In addition, DeepSeek R1 struggled with temporal awareness for both the self-evaluation and CoT phase, however, the model performed better after the

reasoning mimicry stage when borrowing LLaMA 3.1:405B and ChatGPT-4o's reasoning as DeepSeek R1 was able to infer the current month, day, and year based off the thought process given, indicating high adaptability. Neither of DeepSeek V3, Gemini 2.0 Pro, or Claude 3.5 Sonnet were able to infer and store the date when provided with the same thought process.

V. Conclusion and future work

In this study, we looked at how well different LLMs performed on cognitive tasks using the SAGE. We found that ChatGPT-4o performed better than other models in areas such as reasoning, memory, and understanding. We also found that using methods like CoT prompting and reasoning mimicry could help improve the results, with CoT noticeably improving performance. While reasoning mimicry did not directly improve the score, it did help the models better understand the questions and explain reasoning for their answers.

For future work, we plan to explore more effective methods for evaluating AI cognitive abilities [19]. Notably, our findings highlight that knowledge distillation effects do exist in current LLMs, influencing their reasoning and knowledge retention. While this process can enhance model performance, improper knowledge transfer can also spread mistakes and cause errors, leading to hallucinations and unreliable outputs. Therefore, it is important to establish appropriate constraints and safeguards when addressing knowledge distillation challenges to prevent the unintended reinforcement of incorrect reasoning.

References

[1] R. S., *Artificial Intelligence: A Modern Approach*. Upper Saddle River: Pearson Education, Inc., 2020.

[2] J. L. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller, "Navigating cognition: Spatial codes for human thinking," *Science*, vol. 362, no. 6415, p. eaat6766, 2018.

[3] R. Dayan, B. Uliel, and G. Koplewitz, "Age against the machine—susceptibility of large language models to cognitive impairment: cross sectional analysis," *bmj*, vol. 387, 2024.

[4] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, and Y. Choi, "Symbolic chain-of-thought distillation: Small models can also" think" step-by-step," *arXiv preprint arXiv:2306.14050*, 2023.

[5] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5008–5017, 2021.

[6] M. Dahiya, R. Gill, N. Niu, H. Gudaparthi, and Z. Peng, "Leveraging chatgpt to predict requirements testability with differential in-context learning," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 170–175, IEEE, 2024.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[8] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, "Cognitive mirage: A review of hallucinations in large language models," *arXiv preprint arXiv:2309.06794*, 2023.

[9] C. J. Haug and J. M. Drazen, "Artificial intelligence and machine learning in clinical medicine, 2023," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, 2023.

[10] D. W. Scharre, S.-I. Chang, R. A. Murden, J. Lamb, D. Q. Beversdorf, M. Kataki, H. N. Nagaraja, and R. A. Bornstein, "Self-administered gerocognitive examination (sage): a brief cognitive assessment instrument for mild cognitive impairment (mci) and early dementia," *Alzheimer Disease & Associated Disorders*, vol. 24, no. 1, pp. 64–71, 2010.

[11] A. Meara, N. Davidson, H. Steigelman, S. Zhao, G. Brock, W. Jarjour, B. Rovin, H. Madhoun, S. Parikh, L. Hebert, *et al.*, "Screening for cognitive impairment in sle using the self-administered gerocognitive exam," *Lupus*, vol. 27, no. 8, pp. 1363–1367, 2018.

[12] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," 2021.

[13] Y. Zhuang, Q. Liu, Y. Ning, W. Huang, R. Lv, Z. Huang, G. Zhao, Z. Zhang, Q. Mao, S. Wang, *et al.*, "Efficiently measuring the cognitive ability of llms: An adaptive testing perspective," 2023.

[14] S. Hao, Y. Gu, H. Luo, T. Liu, X. Shao, X. Wang, S. Xie, H. Ma, A. Samavedhi, Q. Gao, *et al.*, "Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models," *arXiv preprint arXiv:2404.05221*, 2024.

[15] L. Xu, Z. Hu, D. Zhou, H. Ren, Z. Dong, K. Keutzer, S. K. Ng, and J. Feng, "Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration," *arXiv preprint arXiv:2311.08562*, 2023.

[16] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[17] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023," *arXiv preprint arXiv:2311.05232*, 2023.

[18] F. Qiu, W. Zhang, C. Liu, L. Li, H. Du, T. Guo, and X. Yu, "Language-guided multi-modal emotional mimicry intensity estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4742–4751, 2024.

[19] R. Wang, M. Liu, X. Cheng, Y. Wu, A. Hildebrandt, and C. Zhou, "Segregation, integration, and balance of large-scale resting brain networks configure different cognitive abilities," *Proceedings of the National Academy of Sciences*, vol. 118, no. 23, p. e2022288118, 2021.