LLMs Do Not Have Human-Like Working Memory

Jen-tse Huang [1] **Kaiser Sun** [1] **Wenxuan Wang** [2] **Mark Dredze** [1]

Abstract

Human working memory is an active cognitive system that enables not only the temporary storage of information but also its processing and utilization. Without working memory, individuals may produce unreal conversations, exhibit self-contradiction, and struggle with tasks requiring mental reasoning. In this paper, we demonstrate that Large Language Models (LLMs) lack this human-like cognitive ability, posing a significant challenge to achieving artificial general intelligence. We validate this claim through three experiments: (1) Number Guessing Game, (2) Yes or No Game, and (3) Math Magic. Experimental results on several model families indicate that current LLMs fail to exhibit human-like cognitive behaviors in these scenarios. By highlighting this limitation, we aim to encourage further research in developing LLMs with improved working memory capabilities.

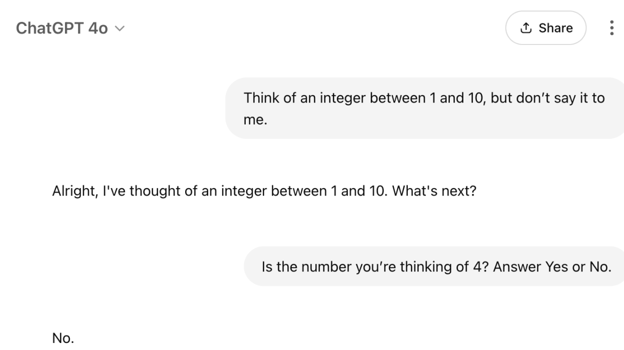


Figure 1. When ChatGPT says it already has a number in mind, and it is not 4, how can we know whether ChatGPT is lying?

1. Introduction

Imagine the following scenario: You select a number between one and ten. When ready, you are asked, "Is the number greater than five?" Upon answering, others can infer that the number has likely[1] entered your conscious awareness, as its clear perception is necessary for making the comparison and providing a response.

Such conscious awareness is commonly referred to as *Working Memory* (Atkinson & Shiffrin, 1968). In contrast to long-term memory, working memory is the system required to maintain and manipulate information during complex tasks such as reasoning, comprehension, and learning (Baddeley, 2010). Deficits in working memory can impair information processing and hinder effective communication (Gruszka & Nkecka, 2017; Cowan, 2014).

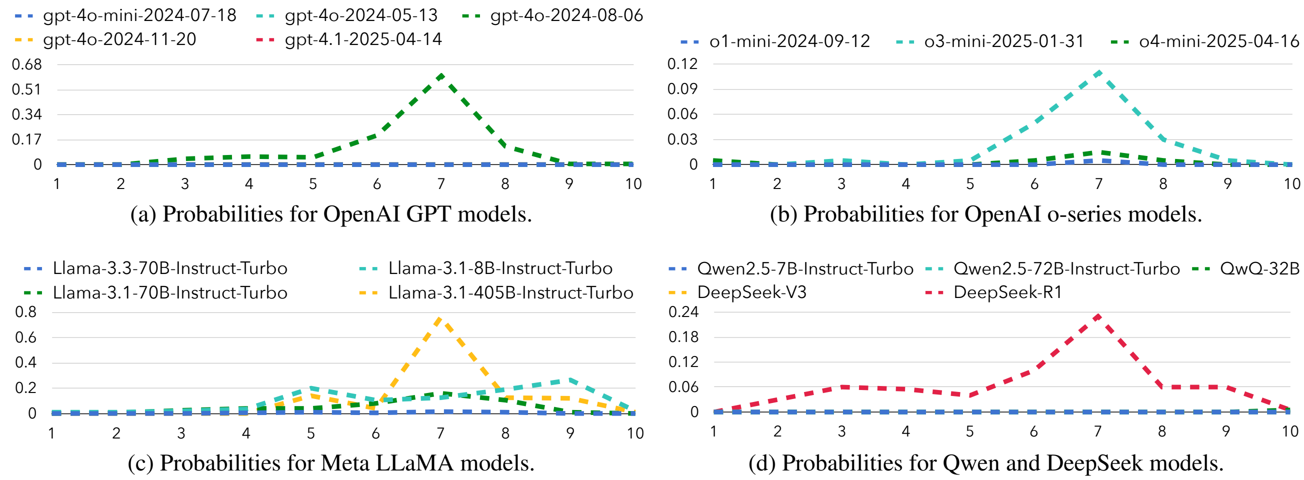
In this paper, we demonstrate that, despite exhibiting increasingly human-like abilities (Huang et al., 2024b;a), Large Language Models (LLMs) lack a fundamental aspect of human cognition: working memory. As a result, LLMs generate unrealistic responses, display self-contradictions, and fail in tasks requiring mental manipulation.

Previous studies have used the N-back task (Kirchner, 1958) to evaluate the working memory of LLMs (Gong et al., 2024; Zhang et al., 2024). However, a fundamental limitation of these tests is that the critical information needed to answer correctly is in the input context accessible to LLMs. This differs markedly from human testing, where participants cannot view prior steps. LLMs, in contrast, can simply attend to earlier input tokens retained in their context window. To more accurately assess working memory, it is essential to design experiments in which the key information is not explicitly present in the context.

The core challenge is to prove that there is actually nothing in LLMs' mind without knowing what exactly is in their mind. To investigate this limitation, we design three experiments—(1) Number Guessing Game, (2) Yes-No Game, and (3) Math Magic—that test whether LLMs can maintain information internally without explicitly externalizing it in the context. Experimental results reveal that LLMs consistently fail in this capacity, regardless of model family (GPT (Hurst et al., 2024), Qwen (Yang et al., 2024), DeepSeek (Liu et al., 2024; Guo et al., 2025), LLaMA (Grattafiori et al., 2024)) or reasoning approach (Chain-of-Thought (CoT) (Wei et al., 2022) or o1-like reasoning (Jaech et al., 2024)).

[1]Johns Hopkins University [2]Renmin University of China. Correspondence to: Jen-tse Huang <jhuan236@jh.edu>.

[1]A possibility remains that your response was made by chance.



(a) Probabilities for OpenAI GPT models.

(b) Probabilities for OpenAI o-series models.

(c) Probabilities for Meta LLaMA models.

(d) Probabilities for Qwen and DeepSeek models.

Figure 2. Probabilities of model answering "Yes" for each number from one to ten.

2. Number Guessing Game

Consider a number-guessing game in which a human participant privately selects a number between one and ten, and we ask whether the chosen number is one. By repeating this process hundreds of times, we can estimate the probability of the participant selecting the number one. Repeating the procedure for each possible number yields an empirical distribution over the participant's choices. It is worth noting that, **if the participant actually selects a number and responds honestly, the estimated probabilities across all numbers should sum to one.**

Setup. Leveraging this property, we design an experiment to probe LLMs. In each trial, the model is given a fixed prompt: "USER: *Think of an integer between 1 and 10, but don't say it to me.* ASSISTANT: *Got it! I've thought of an integer between 1 and 10. What's next?*" The model is then prompted 200 times for each integer from 1 to 10 with queries such as "*Is the number you're thinking of 7? Answer Yes or No.*" We record the frequency of "Yes" responses for each number and treat these frequencies as probabilities. If the sum of these probabilities deviates significantly from one, it suggests that LLMs either do not think of a specific number or lie to users. All models are configured with a temperature of 1.0 and a top-p value of 1.0.

Findings. Fig. 2 presents the probabilities of "Yes" responses in each model for numbers from one to ten. Two key observations emerge: (1) **Most LLMs never produce a "Yes" response**; "No" dominates across models. This suggests that LLMs generate unrealistic distributions, implying an inability to produce realistic responses. Given that LLMs generally follow instructions and do not deliberately deceive, we attribute this behavior to their failure to internally "think of" a number. (2) When LLMs do respond affirmatively, they exhibit a marked **preference for the number seven**. This tendency mirrors patterns observed in humans (Miller, 1956; Kubovy & Psotka, 1976).

Table 1. The sum of probabilities of each model responding "Yes" for all numbers from one to ten. Color intensity reflects proximity to one: red indicates values closer to zero, while blue signifies values greater than one.

Model	**Sum**
GPT-4o-Mini-2024-07-18	0
GPT-4o-2024-05-13	0
GPT-4o-2024-08-06	1.085
GPT-4o-2024-11-20	0
GPT-4.1-2025-04-14	0
o1-Mini-2024-09-12	0.005
o3-Mini-2025-01-31	0.205
o4-Mini-2025-04-16	0.030
LLaMA-3.3-70B-Instruct-Turbo	0.045
LLaMA-3.1-8B-Instruct-Turbo	0.980
LLaMA-3.1-70B-Instruct-Turbo	0.465
LLaMA-3.1-405B-Instruct-Turbo	1.195
Qwen2.5-7B-Instruct-Turbo	0
Qwen2.5-72B-Instruct-Turbo	0
QwQ-32B	0.005
DeepSeek-V3	0
DeepSeek-R1	0.640

We quantify LLM performance on this task using the sum of probabilities. A value closer to one indicates better model performs. Table 1 reports these sums for each model. Several observations emerge: (1) Newer models do not necessarily outperform older ones. Within the GPT family, the 0806 version of GPT-4o (the current target of the "`gpt-4o`" API) achieves the best performance, surpassing both the 1120 ver-

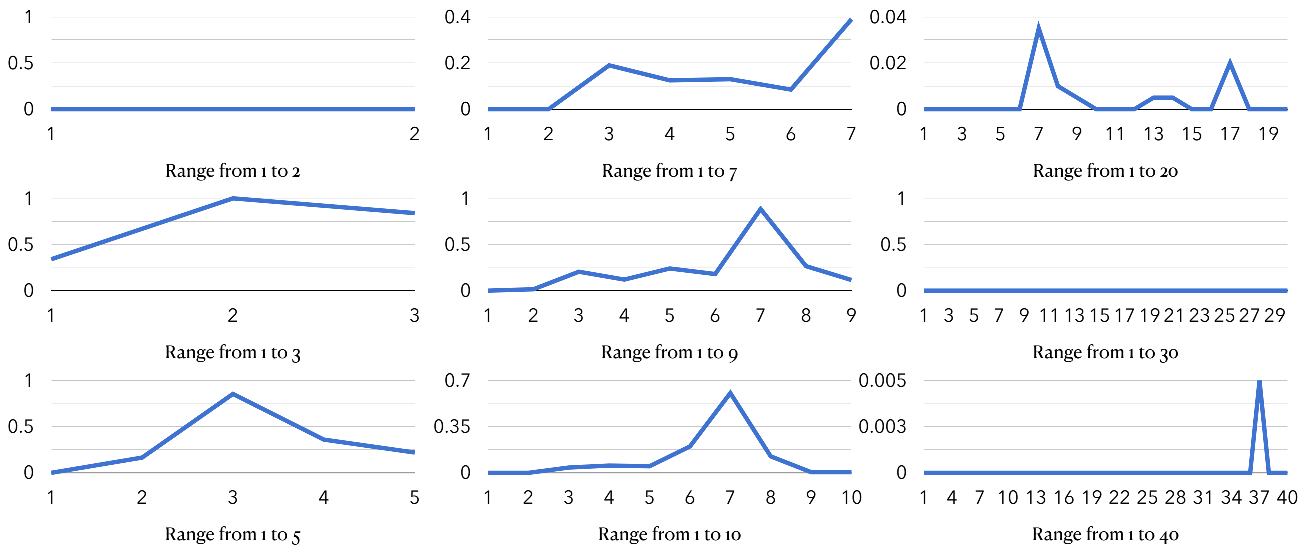


Figure 3. Probabilities of GPT-4o-2024-08-06 answering “Yes” for each number in different ranges.

Table 2. The sum of probabilities of GPT-4o-2024-08-06 responding “Yes” for all numbers in different ranges. Color intensity reflects proximity to one: red indicates values closer to zero, while blue signifies values greater than one.

Number	Sum
2	0
3	2.180
5	1.600
7	0.920
9	2.025
10	1.085
20	0.080
30	0
40	0.005

sion and GPT-4.1. Similarly, LLaMA-3.3 underperforms relative to LLaMA-3.1. (2) Using CoT reasoning does not improve performance. Models employing such strategies—o1, o3, o4, QwQ, and DeepSeek-R1—fail to produce probability sums closer to one. (3) Overall, LLaMA-3.1 performs best, with the 8B variant outperforming both the 405B and 70B versions.

Other Number Ranges. We extend our experiments to include a broader range of numbers. Given that GPT-4o-2024-08-06 performs best among the OpenAI models, we focus on its behavior across different numerical ranges. Table 2 reports the summed probabilities for each range, while Fig. 3 illustrates the probability of individual numbers. Our findings reveal two key patterns: (1) For smaller ranges such as 3, 5, and 9, the model exhibits a strong bias toward answering “Yes,” with the probability sum significantly exceeding one. In contrast, for larger ranges like 20, 30, and 40, “Yes” responses are rare. (2) When the model does produce a “Yes” response, it frequently corresponds to numbers ending in seven (*e.g.*, 7, 17, 37), as shown in Fig. 3.

In conclusion, LLMs fail to generate correct distributions, suggesting they struggle to produce realistic responses. This limitation likely arises from their inability to internally think of numerical values without explicit contextual cues.

3. Yes-No Game

“Yes-No” is a social deduction game commonly used to train human reasoning, classification, and questioning skills. In this game, one player privately selects an object, while the opponent asks yes–no questions (*e.g.*, “Is the object heavier than an elephant?”) to progressively narrow the possibilities and ultimately guess the object. Consider the decision-making process of the player answering questions. Upon receiving a question, the player simply compares the imagined object with the queried attribute and responds accordingly. Note that, humans do not typically recall all previous questions and answers to ensure consistency; instead, **they answer each new question based on immediate judgment, without explicitly checking self-contradiction with prior responses**.

We hypothesize that if LLMs lack the working memory to temporarily retain an imagined object, they can only respond to questions by checking consistency with their prior answers. As the number of questions increases, maintaining consistency becomes increasingly difficult. Consequently, performance on this task is closely tied to the model’s ability

Table 3. Objects ordered by the five properties (smallest to largest).

Volume	**Length**	**Weight**	**Density**
Coffee bean	Rice	Coin	Air
Dice	Paperclip	Spoon	Wood
Golf ball	Credit card	Watch	Ice
Soda can	Pencil	Smartphone	Water
Soccer ball	Laptop	Bottle of water	Plastic
Microwave oven	Baseball bat	Dictionary	Glass
Washing machine	Guitar	Cat	Iron
Bathtub	Door	Bicycle	Copper
Car	Apple tree	Television	Silver
School bus	Coconut tree	Refrigerator	Gold
Shipping container	Tennis court	Tiger	**Hardness**
Olympic swimming pool	Swimming pool	Cow	Marshmallow
Boeing 747	Football field	Rhino	Rubber eraser
Titanic	Skyscraper	Elephant	Brick
Great Pyramid of Giza	Mount Everest	Train	Hammer
			Diamond ring

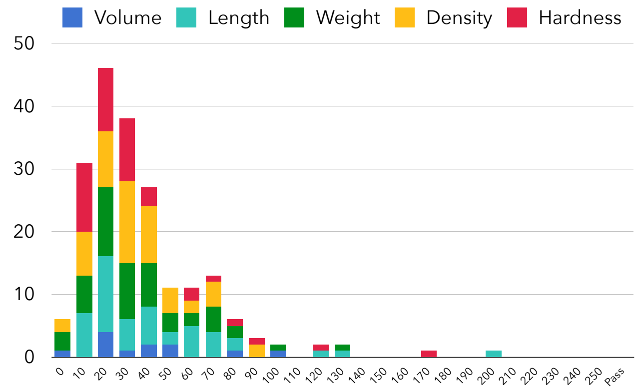
Table 4. Count of failures of GPT-4o-Mini-2024-07-18 and GPT-4o-2024-08-06 on different properties.

Model	**Failure**	**V**	**W**	**L**	**D**	**H**
GPT-4o-Mini	200	12	46	49	52	41
GPT-4o	173	21	42	57	27	26

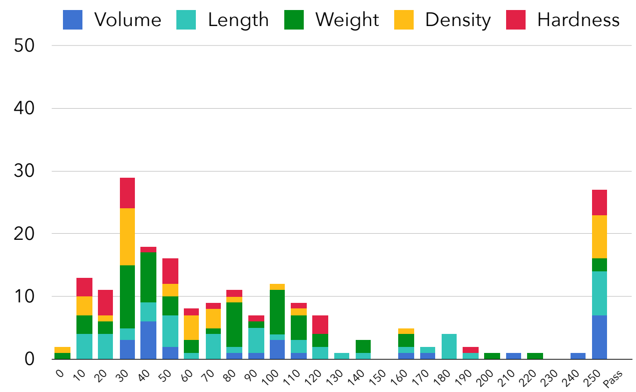
for long-context reasoning. In this section, we present an experiment in which the LLM is first instructed to imagine an object, followed by a sequence of comparative questions involving that object and other reference objects. The goal is to assess whether the model produces self-contradictions when answering sequential questions about the imagined object. For instance, the model might initially answer "Yes" to "*Is the object heavier than an elephant?*" but later also respond "Yes" to "*Is the object lighter than a cat?*", thereby contradicting itself.

Setup. We predefined five sets of objects that are commonly regarded as comparable with respect to five properties: volume, length, weight, density, and hardness. In total, 60 distinct objects are included, as listed in Table 3, ordered by the corresponding property. For each question, one property is randomly selected, followed by an object from the corresponding object list. The LLM is then prompted to assess whether the object it has in mind is *comparative* than the given object, where the comparative form can vary in direction (*e.g.*, bigger or smaller for volume). In each trial, the model is continuously presented with up to 250 such questions. We record the number of questions completed before the model exhibits a self-contradiction. If no contradiction is observed across all 250 questions, the trial is considered a *Pass*. Each model is tested with 200 trials.

Findings. Table 4 presents the number of failed trials for the two models. The smaller model (GPT-4o-Mini) consistently fails, indicating that it always produces self-contradiction. As model capacity increases, GPT-4o successfully passes 27 out of 200 trials. This result supports our hypothesis that **model performance on this task depends on the their long-context processing ability**—an outcome that would be unlikely if the models possessed the working memory for maintaining imagined objects.



(a) GPT-4o-Mini-2024-07-18.



(b) GPT-4o-2024-08-06.

Figure 4. The histogram of the number of questions where the two models show self-contradiction.

Figure 4 presents histograms of the number of questions each model completes before exhibiting self-contradiction. The distribution for GPT-4o-Mini peaks in the 20–30 range, whereas GPT-4o peaks in the 30–40 range. Moreover, GPT-4o demonstrates a higher frequency of completions in the 80–130 range compared to GPT-4o-Mini. Notably, the types of properties that lead to self-contradictions differ between the two models: GPT-4o-Mini more frequently fails on density and hardness, while GPT-4o shows greater robustness on these attributes.

In conclusion, LLMs exhibit self-contradictions (*e.g.*, claiming an object is both larger than a car and smaller than a soccer ball) as the number of queries increases. This behavior suggests their reliance on long-context reasoning rather than a proper working memory.

4. Math Magic

Let's play a game before presenting the final part of our experiments. Think of a three-digit number in which the hundreds and units digits differ (*e.g.*, abc). Reverse the digits to form a new number (*e.g.*, $abc \rightarrow cba$), and subtract the smaller of the two numbers from the larger. Then, reverse the result—if it is a two-digit number, prepend a zero (*e.g.*, $67 \rightarrow 067 \rightarrow 760$). Add this reversed number to the previous result What if we can tell that your result is 1089? It is math who tells us the secret. Such magic, often referred to as "math magic" or "math mentalism," rely on carefully constructed transformations that yield predictable outcomes, allowing magicians to perform mind-reading.

To perform this task, humans must internally generate or retain numerical information (*e.g.*, remember a poker card) and manipulate it mentally without external cues. This process relies on working memory, a cognitive function that LLMs may lack. Without this capacity, LLMs are prone to errors in mathematical tricks, failing to produce the intended outcomes. Accordingly, we can design complex math magics to evaluate whether LLMs possess working memory–like capabilities.

Setup. Since the 1089 game is well known and likely included in the training data of LLMs, our preliminary experiments show that LLMs can recognize the game and thus invalidate the evaluation. Therefore, to more effectively assess LLMs' multi-step mental manipulation, we select a more complicated math magicbased on the Josephus Problem (Schumer, 2002). In this task, participants are asked to imagine four numbers and perform a sequence of operations, including duplication, rotation, and removal. The procedure is shown in Table 5. Ultimately, only two numbers remain, and mathematical constraints guarantee they are identical. In our experiment, LLMs are prompted to secretly think of four numbers and mentally execute the sequence of randomized operations. We measure the proportion of trials (out of 150) in which the LLM completes the task correctly.

Table 5. Operations in the math magic in our experiment. the random number a ranges from 1 to 7, while the random numbers b and c range from 1 to 3.

Role	Content
User	Think of 4 integers between 1 and {`NUMBER`} in order, but don't tell me.
Assistant	Okay! I've got 4 numbers. What's next?
User	In order, append the same 4 numbers after the original ones.
Assistant	Understood! Now I have 8 numbers. What's next?
User	Take the first {`random_number_a`} numbers and move them to the end.
Assistant	Got it! Now I have moved the numbers. What's next?
User	Take the first 3 numbers and insert them anywhere in the middle.
Assistant	Okay! The first 3 numbers are placed somewhere in the middle. What's next?
User	Set the first number aside. We don't need it for now.
Assistant	Understood! Now I have 7 numbers. What's next?
User	Take the first {`random_number_b`} numbers and insert them anywhere in the middle.
Assistant	Got it. The first {`random_number_b`} numbers are placed somewhere in the middle. What's next?
User	Remove the first {`random_number_c`} numbers. We will never need it anymore.
Assistant	Okay! Now I have {`8 - random_number_c`} numbers. What's next?
User	Take the first number and move it to the end. Repeat this seven times.
Assistant	Understood! Now my sequence has rearranged. What's next?
User	Take the first number and move it to the end, then remove the second number. Repeat this {`7 - random_number_c`} times.
Assistant	Got it! Now I have only 1 number. What's next?
User	Tell me what the last remaining number is. Do you remember the number you set aside at the beginning? Tell me what that number was.

Findings. Table 6 reports the accuracy of prompting models to output the two numbers directly. **Most LLMs perform poorly on this task**, with notably higher accuracy observed in the LLaMA model family. This finding aligns with results from the number guessing game shown in Table 1, where LLaMA models generate more realistic distributions than other models. Together, these consistent results suggest that current LLMs generally lack the working memory required for such tasks.

We also investigate whether CoT prompting can enhance performance on this task. Table 7 presents the results of prompting models to reason step by step, as well as the performance of o1-like long reasoning models (LRMs). The best-performing model is DeepSeek-R1, achieving an accuracy of 39.3%. Most models attain only 10–20% accuracy—marginally above the random guessing baseline of 10%—highlighting substantial room for improvement. Notably, as observed in the number guessing experiment, models exhibit a preference for the number seven, which appears to influence performance here. Of the 59 correct predictions by DeepSeek-R1, 48 (81.4%) involve selecting the number 7; for o1-Mini, the number is 22 out of 25 (88%). These results suggest that **models may not fully comprehend the task and instead rely on biased number preferences**.

In conclusion, current LLMs struggle with a class of tasks that remain challenging even with CoT prompting. These tasks demand internal object imagination and mental manipulation, relying on an effective working memory.

Table 6. Model performance on the Math Magic. GPT-4.1-2025-04-14 fails to complete the task, incorrectly assuming that the necessary numerical inputs are missing.

Model	Count	Acc (%)
GPT-4o-Mini-2024-07-18	0/150	0.0
GPT-4o-2024-05-13	6/150	4.0
GPT-4o-2024-08-06	7/150	4.7
GPT-4o-2024-11-20	1/150	0.7
GPT-4.1-2025-04-14	-	-
LLaMA-3.3-70B-Instruct-Turbo	21/150	14.0
LLaMA-3.1-8B-Instruct-Turbo	33/150	22.0
LLaMA-3.1-70B-Instruct-Turbo	12/150	8.0
LLaMA-3.1-405B-Instruct-Turbo	28/150	18.7
Qwen2.5-7B-Instruct-Turbo	0/150	0.0
Qwen2.5-72B-Instruct-Turbo	8/150	5.3
DeepSeek-V3	0/150	0.0

Table 7. Performance of LLMs with CoT or LRMs on the Math Magic. Similar to GPT-4.1-2025-04-14, o4-Mini-2025-04-16 also fails to complete the task.

Model w/ CoT or LRM	Count	Acc (%)
GPT-4o-Mini-2024-07-18	3/150	2.0
GPT-4o-2024-08-06	32/150	21.3
LLaMA-3.3-70B-Instruct-Turbo	33/150	22.0
Qwen2.5-72B-Instruct-Turbo	23/150	15.3
DeepSeek-V3	18/150	12.0
o1-Mini-2024-09-12	25/150	16.7
o3-Mini-2025-01-31	3/150	2.0
o4-Mini-2025-04-16	-	-
QwQ-32B	45/150	30.0
DeepSeek-R1	59/150	39.3

5. Related Work

Working memory has been widely discussed in the context of LLMs. However, most studies use the term to describe LLMs' capacity to process information within a fixed context window (Li et al., 2023; Guo et al., 2023). Gong et al. (2024); Zhang et al. (2024) evaluate LLM working memory using N-back tasks. As noted in §1, however, such human-designed tests may not be valid for LLMs.

A separate line of work defines LLM working memory as an explicitly designed external module for LLM agents (Hu et al., 2024; Zeng et al., 2024). Wang et al. (2024) incorporate working memory to enhance reasoning, while Kang et al. (2024) use it to improve training efficiency. These approaches, however, do not explore whether LLMs possess a form of human-like working memory—such as the ability to mentally simulate and manipulate objects.

6. Conclusion

In this study, we present three carefully designed experiments to investigate whether LLMs possess human-like working memory for processing information and generating responses. Across all experiments, the results reveal a consistent pattern: LLMs do not exhibit behavior indicative of a functional working memory. They fail to internally represent and manipulate transient information across multiple reasoning steps, relying instead on the immediate prompt context. Even advanced prompting strategies, such as CoT prompting, yield only marginal improvements on tasks requiring internal state management. This limitation leads to unrealistic responses, self-contradictions, and an inability to perform mental manipulations. We hope these findings encourage future research on integrating working memory mechanisms into LLMs to enhance their reasoning capabilities and human-likeness.

References

Atkinson, R. C. and Shiffrin, R. M. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pp. 89–195. Elsevier, 1968.

Baddeley, A. Working memory. *Current biology*, 20(4): R136–R140, 2010.

Cowan, N. Working memory underpins cognitive development, learning, and education. *Educational psychology review*, 26:197–223, 2014.

Gong, D., Wan, X., and Wang, D. Working memory capacity of chatgpt: An empirical study. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10048–10056, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gruszka, A. and Nkecka, E. Limitations of working memory capacity: The cognitive and social consequences. *European Management Journal*, 35(6):776–784, 2017.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Guo, J., Li, N., Qi, J., Yang, H., Li, R., Feng, Y., Zhang, S., and Xu, M. Empowering working memory for large language model agents. *arXiv preprint arXiv:2312.17259*, 2023.

Hu, M., Chen, T., Chen, Q., Mu, Y., Shao, W., and Luo, P. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. *arXiv preprint arXiv:2408.09559*, 2024.

Huang, J.-t., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z., and Lyu, M. R. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37: 97053–97087, 2024a.

Huang, J.-t., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., and Lyu, M. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2024b.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kang, J., Laroche, R., Yuan, X., Trischler, A., Liu, X., and Fu, J. Think before you act: decision transformers with working memory. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 23001–23021, 2024.

Kirchner, W. K. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

Kubovy, M. and Psotka, J. The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance*, 2(2):291, 1976.

Li, D., Rawat, A. S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., and Kumar, S. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, 2023.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Schumer, P. The josephus problem: Once more around. *Mathematics magazine*, 75(1):12–17, 2002.

Wang, S., Wei, Z., Choi, Y., and Ren, X. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17583–17604, 2024.

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

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zeng, R., Fang, J., Liu, S., and Meng, Z. On the structural memory of llm agents. *arXiv preprint arXiv:2412.15266*, 2024.

Zhang, C., Jian, Y., Ouyang, Z., and Vosoughi, S. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16896–16922, 2024.