Assessing Working Memory Capacity of ChatGPT

Dongyu Gong[1,2] (dongyu.gong@psy.ox.ac.uk)
[1]Oxford Centre for Human Brain Activity, Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry
[2]Department of Experimental Psychology
University of Oxford, Oxford, United Kingdom

Abstract

Working memory is a critical aspect of both human intelligence and artificial intelligence (AI), serving as a workspace for the temporary storage and manipulation of information. This paper investigates the working memory capacity of ChatGPT, a state-of-the-art language model, by examining its performance on N-back tasks. We begin by discussing the importance of working memory to humans and AI, followed by the methods employed to assess ChatGPT's working memory capacity. Our study compares behavioral performance of ChatGPT on verbal and spatial N-back tasks to that of human participants reported in the literature, revealing notable similarities. Our findings offer crucial insights into the current progress in designing AI systems with human-level cognitive abilities and hold promise for informing future endeavors aimed at enhancing AI working memory and understanding human working memory through AI models.

Keywords: working memory; large language models; ChatGPT; N-back tasks; human-AI comparison

Working memory of humans and AI

The advent of large language models (LLMs) like ChatGPT and GPT-4 has propelled the pursuit of artificial general intelligence (Bubeck et al., 2023) and unveiled human-level abilities that warrant further exploration (Kosinski, 2023; Wei et al., 2022). Among these abilities is the capacity to retain contextual information while engaging in multi-turn conversations, suggesting the presence of working memory in these LLMs. Despite this intriguing observation, a systematic assessment of the working memory capacity of such models remains to be conducted. In this study, we drew upon cognitive sciences literature to devise working memory tasks for ChatGPT and compared its performance with human participants.

In cognitive sciences, working memory is usually defined as the ability to temporarily store and manipulate information in mind (Baddeley, 1992). It is widely regarded as a critical element of human intelligence, as it underlies various higher-order cognitive processes such as reasoning, problem-solving, and language comprehension (Conway & Kovacs, 2020).

The study of working memory in AI has not received the attention it deserves until recent years. It is increasingly evident that in order to design more powerful AI models, the role of working memory must be taken seriously (Graves et al., 2016; Guo et al., 2020; Yoo & Collins, 2022). By placing a greater emphasis on working memory, AI researchers will be able to develop models that better mimic the cognitive processes underpinning human intelligence, leading to more efficient and versatile AI systems.

The emergence of ChatGPT offers a unique opportunity to directly compare human and AI working memory. One widely used task to measure working memory capacity in laboratory settings is the N-back task, initially developed by Kirchner (1958). In the current study, we adapted two versions of the N-back task compatible with ChatGPT and assessed its performance using well-established metrics from human behavioral studies. This approach allows us to obtain a deeper understanding of the working memory capacity of powerful LLMs like ChatGPT and facilitates a more meaningful comparison with human cognitive abilities.

Methods

In N-back tasks, individuals are asked to report whether the currently presented item matches the item presented N items ago. In this study, we devised two N-back tasks involving verbal and spatial working memory (Szmalec et al., 2011) respectively, and prompted ChatGPT (using the OpenAI API) to complete the tasks in a trial-by-trial manner.

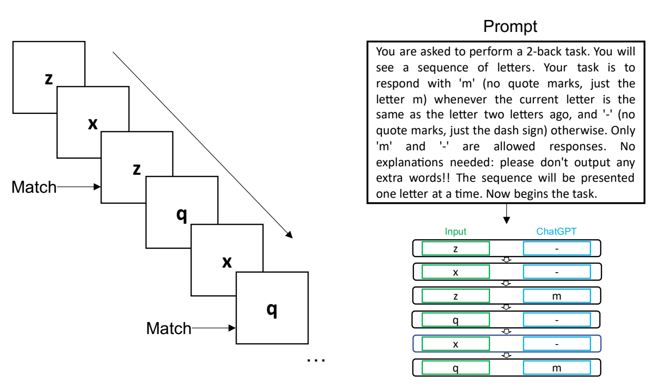

Fig. 1. Human (left) and AI (right) versions of the verbal N-back task, with N = 2 chosen for illustration purposes.

Verbal N-back task For N = 1, 2, and 3 respectively, we generated 50 blocks of letter sequences using an alphabet commonly found in the literature. Each block contains 24 letters (thus 24 trials), including 8 match trials and 16 nonmatch trials. **Fig. 1** illustrates how the same task would be conducted by human participants and how ChatGPT performed the task.

Spatial N-back task Based on the observation that ChatGPT can understand spatial relationships (Bubeck et al., 2023), we constructed a 3*3 grid using ASCII art (see **Fig. 2**, right panel), For N = 1, 2, and 3 respectively, we generated 50 blocks of grid sequences each featuring a letter “X” in one of the nine positions. Note that the letter “X” here was arbitrarily chosen to represent an occupied spatial location textually and could be substituted by any other letter or symbol. Each block contains 24 grids, including 8 match trials and 16 nonmatch trials.

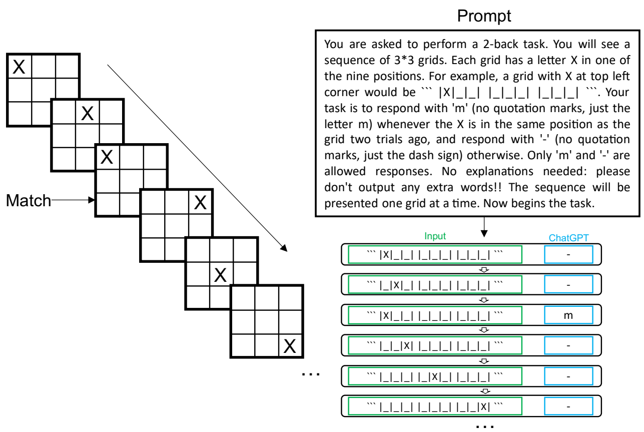

Fig. 2. Human (left) and AI (right) versions of the spatial N-back task, with N = 2 chosen for illustration purposes.

Results

We analyzed ChatGPT’s performance using widely accepted performance metrics reported in numerous human behavioral studies (Jaeggi et al., 2010; Redick & Lindsey, 2013), including hit rate, false alarm rate, accuracy (hit rate - false alarm rate), and sensitivity (d' = Z(hit rate) - Z(false alarm rate)).

Strikingly, the behavioral patterns of ChatGPT in both verbal (**Fig. 3**) and spatial (**Fig. 4**) N-back tasks closely resembled those of human participants. The level of sensitivity was also comparable to humans reported in much of the literature (Kane et al., 2007; Pelegrina et al., 2015). The significant decline in ChatGPT’s performance as N increases (p values ≤ .001 for all Kruskal-Wallis tests on hit rates, accuracy, and sensitivity) not only indicates that it has limited working memory capacity as humans do (Cowan, 2001), but also suggests that LLMs like ChatGPT may exhibit emergent properties (Wei et al., 2022) akin to the human brain. One exception occurred in the spatial N-back task, with false alarm rates decreasing as N increased from 1 to 3 (p = .021), a pattern rarely reported in human studies.

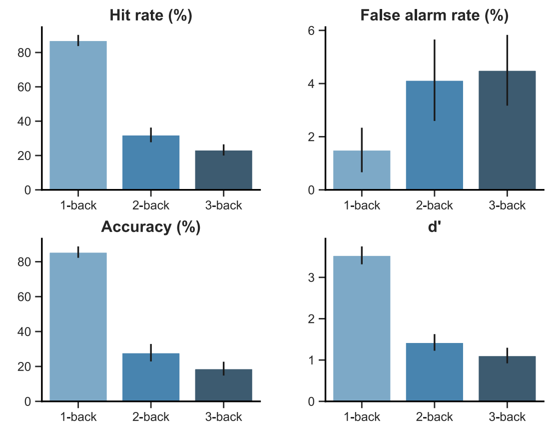

Fig. 3. ChatGPT performance in the verbal N-back task. Error bars represent 95% confidence intervals.

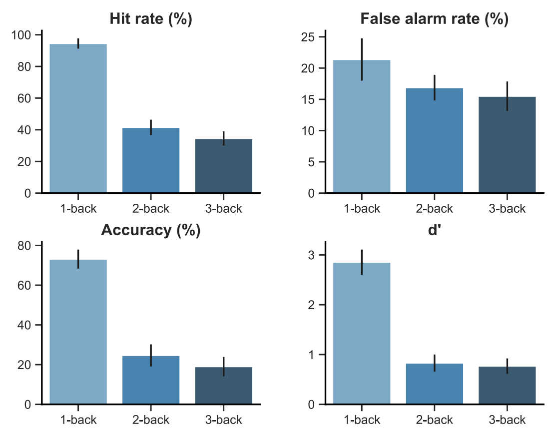

Fig. 4. ChatGPT performance in the spatial N-back task. Error bars represent 95% confidence intervals.

Discussion

It is crucial to acknowledge that LLMs’ abilities are constrained by their text-based nature (Mahowald et al., 2023). For instance, it remains uncertain whether ChatGPT trully processed the inputs in our spatial N-back task from a “spatial” perspective. Nevertheless, future research should investigate the impact of various task manipulations, such as employing different stimuli or prompt stategies, on ChatGPT’s working memory performance. Such exploration could provide valuable insights into how LLMs learn from task contexts and the extent to which the nature of AI working memory resembles that of humans.

Acknowledgments

This research was funded by a Clarendon Scholarship, a Medical Research Council Studentship, and a New College-Yeotown Scholarship awarded to Dongyu Gong.

References

Baddeley, A. (1992). Working Memory. *Science*, *255*(5044), 556–559.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712.*

Conway, A. R. A., & Kovacs, K. (2020). Working Memory and Intelligence. In R. J. Sternberg (Ed.), *The Cambridge Handbook of Intelligence* (2nd ed.). Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471-476.

Guo, F., He, R., Dang, J., & Wang, J. (2020). Working Memory-Driven Neural Networks with a Novel Knowledge Enhancement Paradigm for Implicit Discourse Relation Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 7822-7829).

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 615–622.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*, 352–358.

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv preprint arXiv:2302.02083.*

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv: 2301.06627.*

Pelegrina, S., Lechuga, M. T., García-Madruga, J. A., Elosúa, M. R., Macizo, P., Carreiras, M., Fuentes, L. J., & Bajo, M. T. (2015). Normative data on the n-back task for children and young adolescents. *Frontiers in Psychology*, *6*, 1544.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113.

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 137.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *arXiv preprint arXiv: 2206.07682.*

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568.