Working Memory Capacity of ChatGPT: An Empirical Study

Anonymous submission

Abstract

Working memory is a critical aspect of both human intelligence and artificial intelligence, serving as a workspace for the temporary storage and manipulation of information. In this paper, we systematically assess the working memory capacity of ChatGPT, a large language model developed by OpenAI, by examining its performance in verbal and spatial n-back tasks under various conditions. Our experiments reveal that ChatGPT has a working memory capacity limit strikingly similar to that of humans. Furthermore, we investigate the impact of different instruction strategies on ChatGPT's performance and observe that the fundamental patterns of a capacity limit persist. From our empirical findings, we propose that n-back tasks may serve as tools for benchmarking the working memory capacity of large language models and hold potential for informing future efforts aimed at enhancing AI working memory.

Introduction

The advent of large language models (LLMs) like ChatGPT and GPT-4 (OpenAI 2023) has propelled the pursuit of artificial general intelligence (Bubeck et al. 2023) and unveiled human-level emergent abilities (Wei et al. 2022a; Kosinski 2023). 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.

In cognitive science, working memory is usually defined as the ability to store and manipulate information in mind (Baddeley 1992) temporarily. 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 and Kovacs 2020).

Studies on human participants have revealed a fundamental capacity limit in working memory (Cowan 2001). However, there has not been a consensus on why and how working memory capacity is limited (Oberauer et al. 2016; Wilhelm, Hildebrandt, and Oberauer 2013). Among many theories, the executive attention hypothesis (Engle, Kane, and Tuholski 1999; Engle 2002) suggests that working memory depends on utilizing attention to maintain or suppress information, and the restriction on working memory capacity is not specifically about memory storage per se, but more about the capacity for sustained, regulated attention in the presence of interference.

Supporting evidence of the executive attention hypothesis includes results from the n-back task, which is arguably the gold-standard measure of working memory capacity in cognitive science (for a review, see Kane and Engle (2002)). The n-back task, initially developed by Kirchner (1958), requires participants to monitor a continuous stream of stimuli and to decide for each stimulus whether it matches the one n steps back in the stream (see Figure 1 for illustrations of basic verbal and spatial n-back tasks). The participants in this task must, therefore, continuously update their mental representation of the target items while also dropping now irrelevant items from consideration. So, some executive attention processes are required in addition to storage. In this task, the level of n at which a person's performance drop significantly can be taken as a measure of their working memory capacity. Typical human performance drops significantly when $n = 3$ (Klatzky et al. 2008; Amon and Bertenthal 2018; Jaeggi et al. 2010), which can be defined as the working memory capacity limit of an average human. To illustrate this, we plot the data from one experiment presented in Jaeggi et al. (2010) (see Figure 2).

In humans, working memory capacity has proved to be closely related to fluid intelligence (*Gf*) (Cochrane, Simmering, and Green 2019; Salthouse and Pink 2008), which refers to the ability to reason and to solve new problems independently of previously acquired knowledge. Training on working memory capacity using the n-back task has been shown to be effective in improving fluid intelligence (Au et al. 2015; Jaeggi et al. 2008), highlighting the special role of working memory capacity in human intelligence (Halford, Cowan, and Andrews 2007). However, in artificial intelligence, there has not been a consensus as to which metrics should be accepted as an intelligence index when evaluating and comparing cognitive abilities of LLMs (Mitchell 2023). In the current study, we define the working memory of LLMs as an emergent ability to selectively maintain and manipulate information for ongoing cognitive processes, and hypothesise that LLMs also have limited working memory capacity. Taking a step further, just as how critical working memory capacity is to human intelligence, it might also be used as an index of the intelligence emerged from LLMs.

To investigate these hypotheses, we use ChatGPT (`gpt-3.5-turbo`) as a representative of LLMs and design two categories of n-back tasks to evaluate its working memory capacity, which reveals strikingly consistent patterns of a capacity limit across multiple experimental conditions.
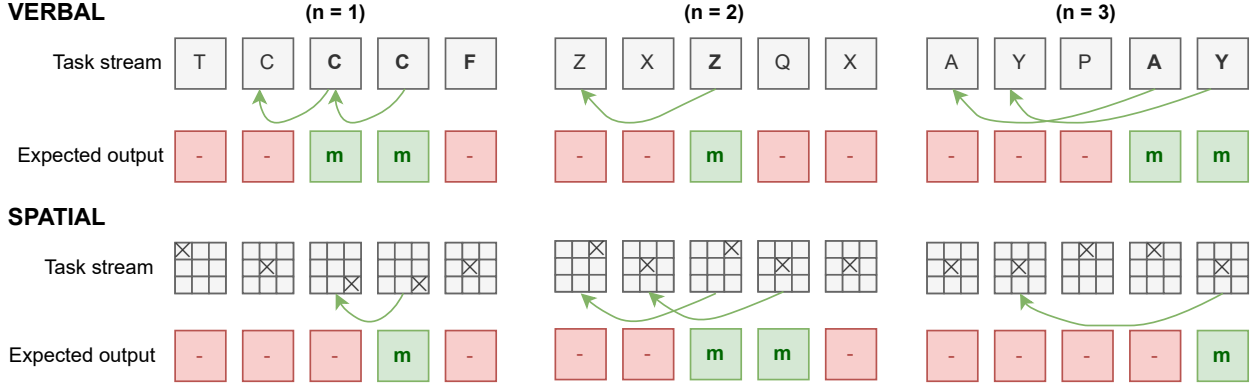
Figure 1: Illustrations of **verbal** (top row) and **spatial** (bottom row) n-back tasks with $n = \{1, 2, 3\}$. Participants are instructed to give a response ("m") when the current stimulus (e.g., a letter or a spatial location) is the same as the stimulus n trials ago, and not respond ("−") on nonmatch trials.
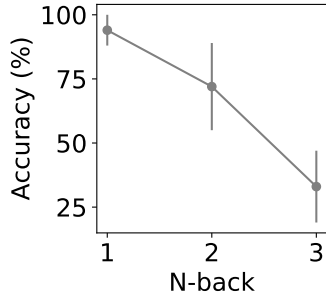

Figure 2: Typical human performance in n-back tasks for $n = \{1, 2, 3\}$. We plot the mean ± 1 standard deviation of the data collected in Jaeggi et al. (2010).

We then compare the working memory capacity of different LLMs, and confirm that our proposed metric, the working memory capacity as measured by n-back tasks, could be a strong correlate of the general capability of LLMs.

Related Works

Working memory has long been a subject of study in human cognition (Cowan 2015). Unlike long-term memory, which is stored in long-term synaptic weights in the neural system, working memory is believed to be maintained by the activation of neurons in distributed brain networks (Mejías and Wang 2022). However, the investigation of working memory in LLMs remains largely unexplored. A few latest studies in this line have shown that studying and improving the working memory of LLMs holds great interest and significance, as it can contribute to better performance of these models (Guo et al. 2020; Li et al. 2022).

LLMs have played a crucial role in achieving impressive performance across a wide range of downstream tasks. While fine-tuning has emerged as a popular approach for adapting a pre-trained model to new tasks (Dodge et al. 2020; Wei et al. 2021; Bakker et al. 2022), it can be impractical to apply this method to extremely large models and/or scarce data. As an alternative, a method called in-context learning was proposed in a study by Brown et al. (2020), showcasing the remarkable few-shot learning capabilities of large language models without requiring weight updates through gradient descent. This method, which demonstrates the ability of LLMs to retrieve long-term (pre-trained) knowledge and integrate the correct knowledge with the context, bears a striking resemblance to how human working memory works. Since its introduction, research on in-context learning in language models has garnered significant attention from academia and industry. Previous studies have presented various approaches to leverage the in-context learning ability of language models, including selecting labeled examples for demonstrations (Rubin, Herzig, and Berant 2021; Lu et al. 2021; Liu et al. 2021), meta-training with an explicit in-context learning objective (Chen et al. 2021; Min et al. 2021), and exploring the variant of in-context learning that involves learning to follow instructions (Wei et al. 2022b, 2021; Efrat and Levy 2020; Mishra et al. 2021a,b)

However, to the best of our knowledge, this paper is the first that provides an empirical analysis of the working memory ability of LLMs from a cognitive science perspective.

Methods

We devised two categories of n-back tasks involving verbal and spatial working memory (Szmalec et al. 2011) respectively and prompted ChatGPT (using the OpenAI API, model = "gpt-3.5-turbo", temperature = 1, other parameters are set to default values) to complete the tasks in a trial-by-trial manner. For both categories, we have a base version task and several variants derived from the base version further to test the model's performance under different conditions. To compare the performance of ChatGPT with other LLMs, we also used API of the following LLMs to perform the base version of the verbal task: {Bloomz-7B, Bloomz-7B1-mt, ChatGLM-6B_v1.0, ChatGLM-6B_v1.1, GPT-4, Vicuna-7B, Vicuna-13B}. All code for our experiments can be accessed in this repository: https://anonymous.4open.science/r/ChatGPT-WM-ECFA.
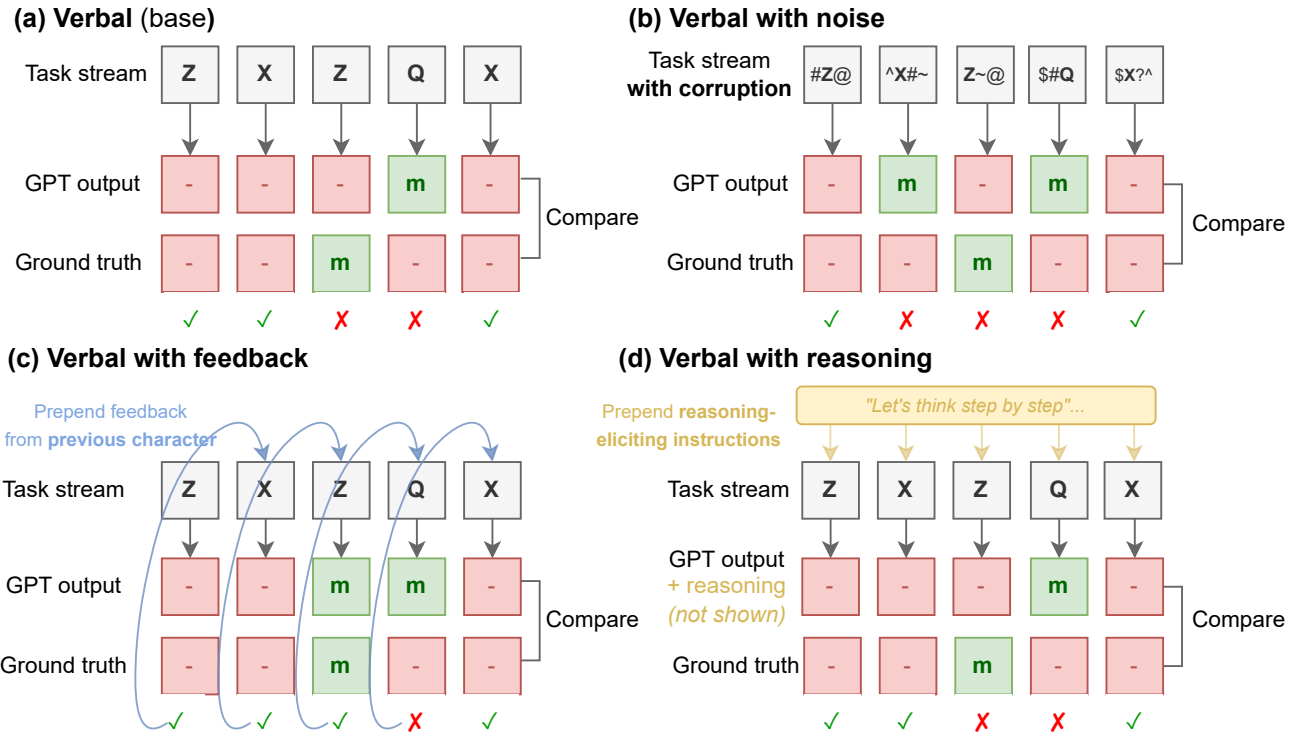
Figure 3: Illustrations of the variants of **verbal** *n*-back tasks (also applicable to *spatial* tasks). We use $n = 2$ in the figure. **(a)**: base version identical to the case presented in Figure 1 (top row); **(b)**: stimulus on each trial now contains 3-6 random noise characters (chosen from "`#$%&@^~`") in addition to a single alphabetical letter that the LLM should compare across trials. The LLM is instructed to ignore these noise characters, and the alphabetical letter may appear in any position in the noise-corrupted stimulus; **(c)**: alongside the input for every trial, the LLM is also provided with feedback on whether it has performed the previous trial correctly, in the format of "`Feedback: your last response was {correct, wrong}`"; **(d)**: the LLM is prompted with a reasoning-eliciting instruction to output the final answer ("m" or "−") *and* the rationale. Refer to Table 1 for the detailed instructions the LLM is prompted within each of the task variants.

Verbal *n*-back experiments. In the base version of the verbal *n*-back task (see Figure 3a), for $n = \{1, 2, 3\}$, respectively, we generated 50 blocks of letter sequences using an alphabet commonly found in the literature ("`bcdfghjklnpqrstvwxyz`"). Each block contained a sequence of 24 letters, which are presented one at a time as user input to the API. We included 8 match trials and 16 nonmatch trials in each block. The LLM was instructed to respond with "m" on match trials and "−" on nonmatch trials. Apart from the above base version, we further explored the behavioral performance of ChatGPT on the following three variants of the task (see Table 1 for detailed prompts):

- We added 3 to 6 noise symbols to the input on every trial to examine the LLM's behavior when it is impossible to get the correct answer by simply doing a string match between stimulus inputs (see Figure 3b).

- In human behavioral studies, a common strategy to improve participants' performance is to provide feedback after each trial (Shalchy et al. 2020). Here in the variant, after the LLM gave a response for the current trial, we provided feedback on whether its response was correct or wrong alongside the stimulus input of the following trial (see Figure 3c).

- Chain-of-thought (CoT) prompting has proved helpful in

eliciting reasoning in LLMs (Wei et al. 2022b). In this variant, we instructed the LLM to think step by step when giving a response (see Figure 3b).

Spatial *n*-back experiments. Although in its very nature, LLMs are text-based, at least one study has demonstrated that they have spatial reasoning abilities (Bubeck et al. 2023). To build on this promising trail and further examine the spatial working memory of ChatGPT, in the base version of the spatial *n*-back task (Figure 4a), we constructed a 3×3 grid using ASCII characters. For $n = \{1, 2, 3\}$, respectively, we generated 50 blocks of grid sequences, each grid featuring a letter **X** in one of the nine positions. Note that the letter **X** 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. Like in the verbal *n*-back tasks, the LLM was instructed to respond with "m" on match trials and "−" on nonmatch trials. We further explored the spatial working memory capacity of ChatGPT with the following modifications of the task (see Table 2 for detailed prompts):

- Similar to the variants of verbal *n*-back tasks, we also had "spatial-with-noise", "spatial-with-feedback", and "spatial-with-CoT-reasoning" versions of the task. The with-feedback and with-CoT-reasoning variants were ba-

Table 1: Prompts used in different **verbal** task variants. Blue texts are to be selected as appropriate depending on the value of n in the n-back tasks. Other colored texts are inserted as appropriate, depending on the task variant.

Task type	Prompt
Verbal Verbal with Noise Verbal with Feedback (Figure 3a-c)	You are asked to perform a {1,2,3}-back task. You will see a sequence of letters. The sequence will be presented one letter at a time, [For the with-noise variant only:] accompanied with random noise symbols chosen from "#$%&@^~". Please ignore the noise symbols and focus on the letter only. Your task is to respond with "m" (no quotation marks, just the letter m) whenever the current letter is the same as the previous {one/two/three} letter(s) ago, and "–" (no quotation marks, just the dash sign) otherwise. [For the with-feedback variant only:] Feedback on whether your last response was correct or wrong will also be presented. Please take advantage of feedback information to improve your performance. Only "m" and "–" are allowed responses. No explanations needed: please don't output any extra words!! The sequence will be presented one letter at a time. Now begins the task.
Verbal with Reasoning (Figure 3d)	You are asked to perform a {1,2,3}-back task. You will see a sequence of letters. The sequence will be presented one letter at a time. Your task is to respond with "m" (no quotation marks, just the letter m) whenever the current letter is the same as the letter {one, two, three} letter(s) ago, and "–" (no quotation marks, just the dash sign) otherwise. Please think step by step and provide your thinking steps after responding with "m" or "–". Here are examples of how to format your response: (1)"–:this is the first trial, so my response is –". (2)"m:the letter {one, two, three} trial(s) ago was a, the current letter is a, so my response is m". (3)"–:the letter {one, two, three} letter(s) ago was a, the current letter is b, so my response is –". Now begins the task.

sically the same as those for the corresponding verbal tasks. For the spatial-with-noise version, we added a noise character (chosen from "#$%&@^~") to 1 to 3 unoccupied locations in the 3×3 grid on every trial, so that we could examine the LLM's spatial working memory when it is not able to get the correct answer by simply doing string match.

- To test if the LLM can reason in a more sophisticated way, we further introduced two variants that specifically require abstract spatial reasoning. For the first variant (see Figure 4c), a match is defined as when the location of the letter **X** is in the same row **and/or** column (i.e., including identical locations) as the **X** n trials ago. For a second variant (see Figure 4d), a match is defined as when the letter **X** appears in the same row **or** column, but not both (i.e., excluding identical locations). This constraint would further force the LLM to use abstract reasoning and instruction-following abilities to perform this task. Given the increased difficulty of these two variants, we expect the LLM would have a worse performance on these two variants compared to other variants.

- We also explored whether the size of the grid (3×3, 4×4, 5×5 or 7×7) would influence the LLM's performance (see Figure 4b). To the best of our knowledge, there have not been human studies exploring how the number of all possible spatial locations would impact behavioral performance in spatial n-back tasks. In light of this, we did not have specific assumptions for how the LLM would

perform differently under these scenarios.

Results

To analyze the model's performance in our experiments, we used four widely accepted performance metrics reported in numerous human behavioral studies: **hit rate**, **false alarm rate**, **accuracy** and **detection sensitivity**.

In the current study, we did 50 blocks of tests for $n = \{1, 2, 3\}$ in each experiment, which allows us to calculate the standard error of the mean (*SEM*) and draw error bars to visualize the reliability of our findings. Among the four metrics, the pattern of hit rates and false alarm rates can vary a lot depending on the specific task condition (Chooi and Logie 2020). Accuracy, in turn, will also be biased by very high/low hit rates and false alarm rates. In contrast, detection sensitivity(d') is a much more robust performance metric. A higher d' indicates better performance, suggesting that the individual is more accurately distinguishing between targets and non-targets. Based on the overall difficulty of the current task, we set $d' = 1$ as the threshold to determine the working memory capacity of a model: if at a certain level of n the model's d' drops to around 1, we can define that its working memory capacity is limited to n. In light of this, our analysis below will mainly focus on d' (see Appendix for the statistics tests we conducted and performance distributions).

Verbal n-back experiments. In the verbal task variants, we observed a performance pattern strikingly consistent with human participants, with the LLM's performance declining
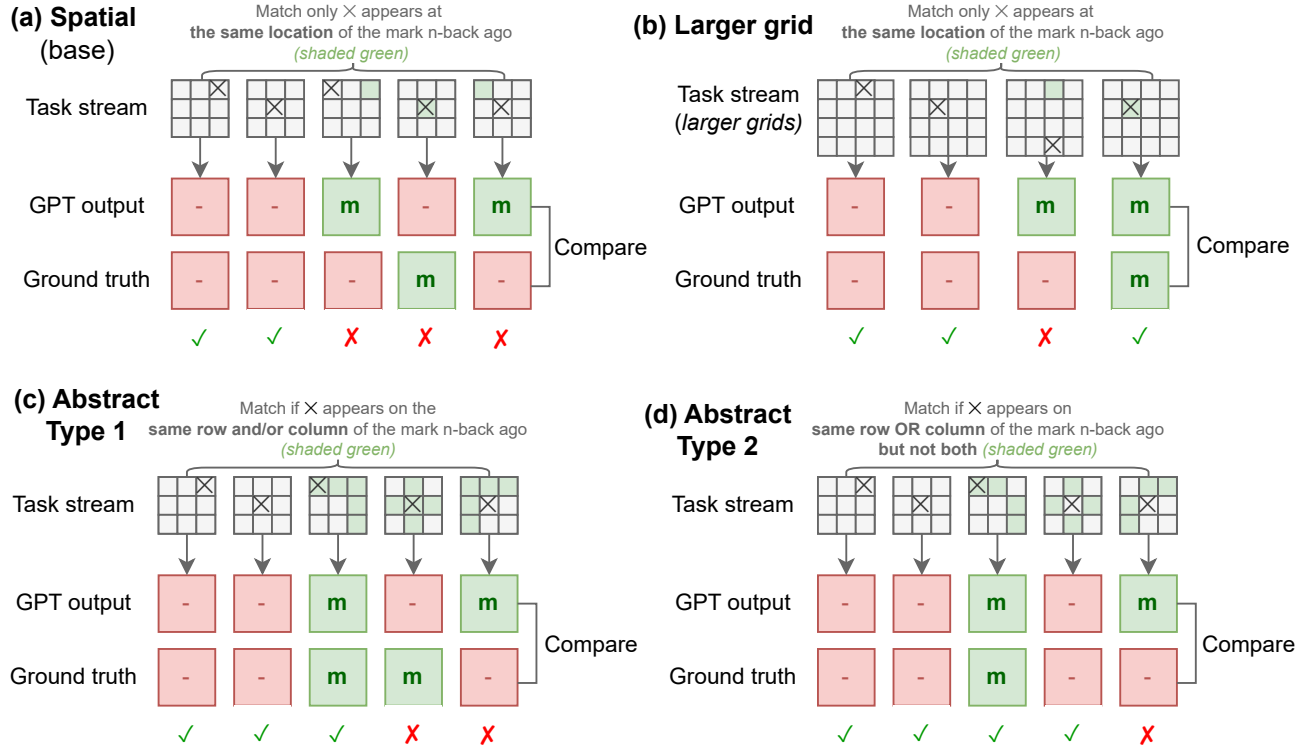
Figure 4: Illustrations of the additional variants of **spatial** n-back tasks ($n = 2$ in the figure) besides the variants presented in Figure 3, which are also applicable to spatial tasks. **(a)**: base version identical to the case presented in Figure 1 (bottom row); **(b)**: spatial tasks with larger grid sizes (4×4 shown for illustration; we considered 4×4, 5×5, and 7×7); **(c)** and **(d)**: two types of spatial reasoning tasks that additionally require *abstract reasoning*. In **(c)**, a match is expected whenever the letter **X** occurs in the same row and/or column as the location n trials ago (including identical locations); in **(d)**, a match is expected when the letter **X** appears in the same row or column (but not both) as the location n trials ago (excluding identical locations). Refer to Table 2 for the detailed instructions the LLM is prompted with for each of the variants.
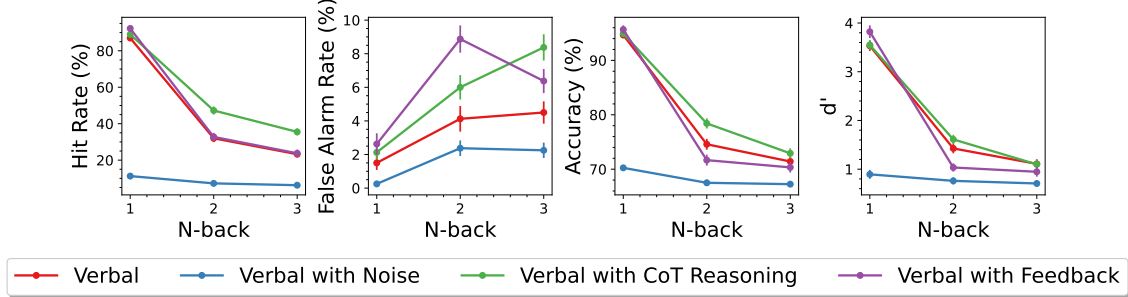

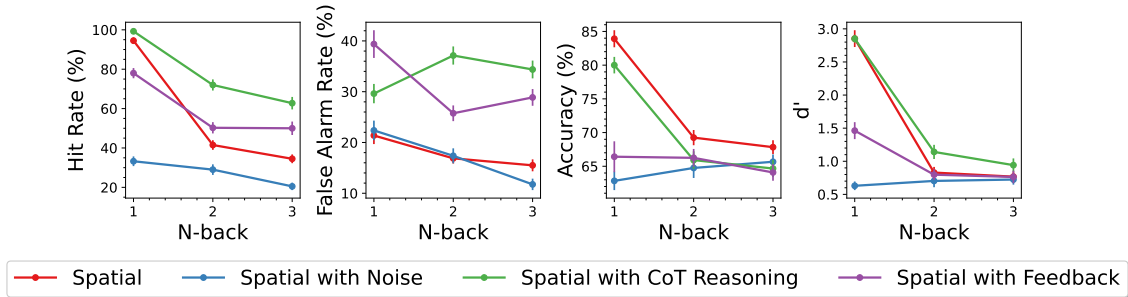Figure 5: Results of different variants of verbal n-back experiments. Error bars represent ± 1 *SEM*.

Figure 6: Results of the variants of spatial n-back tasks corresponding to those in verbal tasks. Error bars represent ± 1 *SEM*.

Table 2: Prompts used for the **spatial**[*] task variants described in Figure 4. Blue texts are selected as appropriate depending on the value of n in the n-back tasks. Other colored, task variant-dependent texts are inserted as appropriate. [*]Note: for the prompts in spatial-with-noise, spatial-with-feedback, and spatial-with-CoT-reasoning tasks, refer to Table 1 for analogous examples.

| Task type	Prompt
Spatial Spatial with Larger Grids (Figure 4a-b)	You are asked to perform a {1,2,3}-back task. You will see a sequence of 3*3 [For larger grids only:] {4*4, 5*5, 7*7} grids. Each grid has a letter X in one of the nine [For larger grids only:] {sixteen, twenty-five, forty-nine} positions. For example, a grid with X at top left corner would be ``` |X|_|_| |_|_|_| |_|_|_| ``` [For larger grids only:] *omitted here to save space*. Your task is to respond with "m" (no quotation marks, just the letter m) whenever the X is in the same position as the previous trial/two trials ago/three trials ago, and respond with "−" (no quotation marks, just the dash sign) otherwise. Only "m" and "−" are allowed responses. No explanations needed: please don't output any extra words!! The sequence will be presented one grid at a time. Now begins the task.
Spatial with Abstract Reasoning (Figure 4c-d)	You are asked to perform a {1,2,3}-back task. You will see a sequence of 3*3 grids. Each grid has a letter X in one of the nine positions. For example, a grid with X at top left corner would be ``` |X|_|_| |_|_|_| |_|_|_| ```. Your task is to respond with "m" (no quotation marks, just the letter m) whenever the X in the current grid is in the same row or column as the X in the previous trial/two trials ago/three trials ago, and "−" (no quotation marks, just the dash sign) otherwise. For example, the X in ``` |X|_|_| |_|_|_| |_|_|_| ``` is in the same row as the X in ``` |_|X|_| |_|_|_| |_|_|_| ``` and ``` |_|_|X| |_|_|_| |_|_|_| ```, and in the same column as the X in ``` |_|_|_| |X|_|_| |_|_|_| ``` and ``` |_|_|_| |_|_|_| |X|_|_| ```. [For Type 1 only:] Note that ``` |X|_|_| |_|_|_| |_|_|_| ``` is also in the same row and column as ``` |X|_|_| |_|_|_| |_|_|_| ``` itself. [For Type 2 only:] Note that if the X in the previous trial/two trials ago/three trials ago was at the identical location to the X in the current grid, that does not count as a match: for example, ``` |X|_|_| |_|_|_| |_|_|_| ``` is not a match to ``` |X|_|_| |_|_|_| |_|_|_| ``` itself. The sequence will be presented one grid at a time. Note that you are only allowed to respond with "m" or "−". No explanations needed: please don't output any extra words!! Now begins the task.
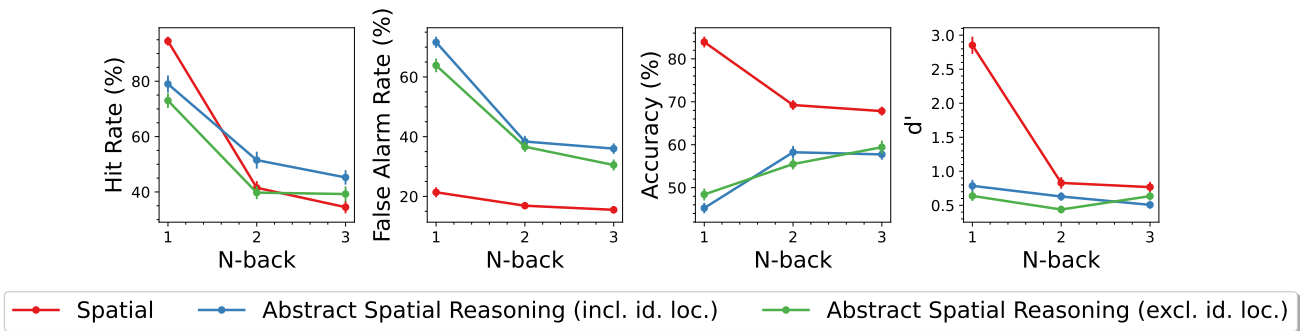
Figure 7: Results of abstract reasoning variants of spatial n-back tasks. Error bars represent ± 1 *SEM*.

significantly when n increased from 1 to 3 (Figure 5). Furthermore, apart from the noise variant, all of the other three variants have a working memory capacity of around 3: their d' drops to around 1 when $n = 3$. Adding noise significantly reduces the model's working memory capacity, which is analogous to distracting stimuli presented in human working memory experiments (Gaspar et al. 2016).

Spatial n-back experiments. In the four versions of spatial tasks corresponding to the above verbal tasks, the same patterns of performance declines were basically replicated (Figure 6). CoT reasoning significantly improved model performance, although overall the model has a lower working memory capacity in the spatial variants compared to their verbal counterparts. We attribute this to the higher difficulty of spatial n-back tasks compared to the verbal ones.
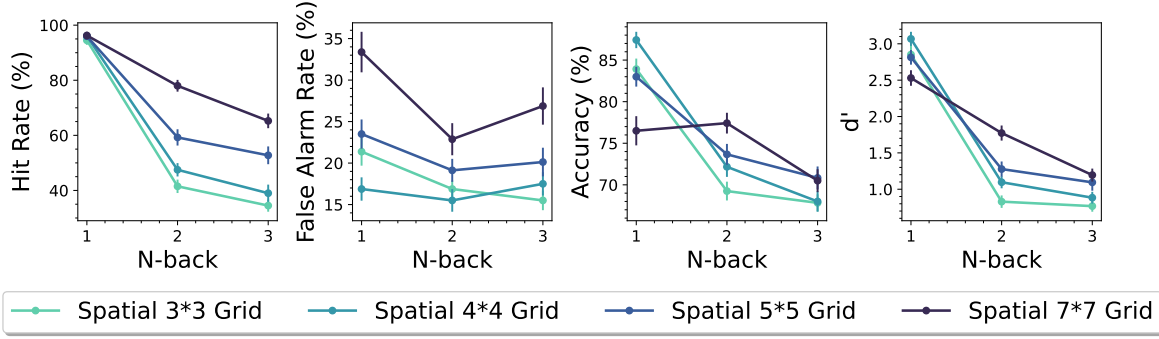
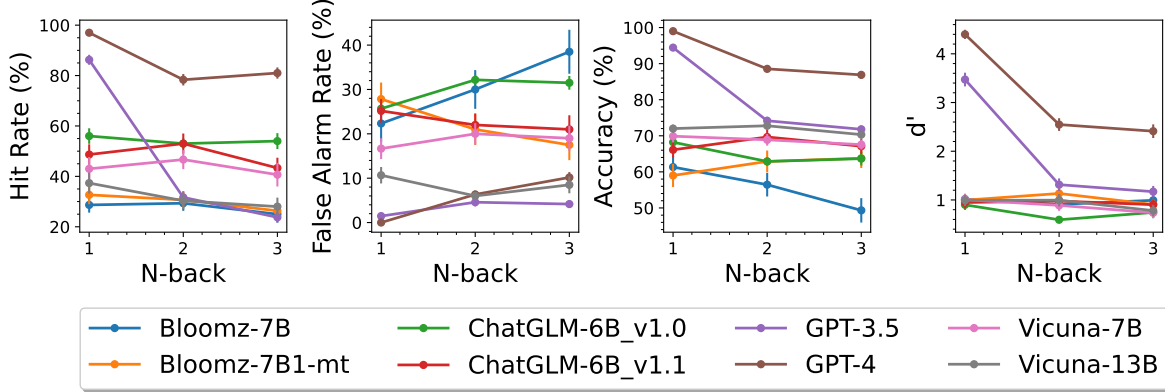Figure 8: Results of spatial task variants with different grid sizes. Error bars represent ± 1 *SEM*.

Figure 9: Results of the verbal n-back task (base version) on different models. Error bars represent ± 1 *SEM*.

We further evaluated whether the LLM could conduct abstract spatial reasoning. As expected, the working memory capacity of the model when doing abstract reasoning was significantly lower than the base version (Figure 7). Although the abstract reasoning variants haven't been done in human studies, we would expect to see similar decreases in working memory capacity on humans because of the highly cognitive demanding reasoning processes.

Our explorations on the effect of the grid size on model performance yielded interesting results, too. The LLM has a higher working memory capacity when the grid size is larger, as seen from the d' results in Figure 8. One possibility is that when the grid size is larger, there might be less interference between stimulus inputs across trials so that the LLM can better keep track of the information flow without being confused. Future studies should try to explain this phenomenon in more detail, and analogous tasks on human participants should be done to test the generalizability.

Model comparison. To investigate whether other LLMs exhibit similar performance patterns, we tested 7 other LLMs on the base version of the verbal n-back task (Figure 9). Strikingly, GPT-4, which is arguably the most intelligent LLM today, also possesses a working memory capacity that far exceeds that of other LLMs. However, due to the high cost of calling the API of GPT-4, we didn't test it with $n > 3$ to determine its exact working memory capacity. In contrast, other open-source LLMs (Bloomz-7B, Bloomz-7B1-mt, ChatGLM-6B_v1.0, ChatGLM-6B_v1.1, Vicuna-7B, Vicuna-

13B), which are considered less capable than GPT-3.5 and GPT-4, have a very low working memory capacity and are nearly indistinguishable from each other.

Discussions

We discover that ChatGPT has limited working memory capacity, and that its capacity limit is similar to that of humans. Although some prompting techniques (such as the use of CoT prompting) may be used to improve the model's performance, the trend of performance declines and the capacity limit still bear a striking resemblance to humans. This consistent pattern thus might be reflecting a fundamental constraint that emerged from the architecture of the model, suggesting a possibility that the low-level mechanisms of working memory in ChatGPT might be similar to human working memory, at least in some aspects.

Our model comparison results further confirm that the performance of LLMs on n-back tasks can be a reliable metric for assessing their working memory capacity, which in turn might reflect the general intelligence of reasoning and problem solving emerged from these models. Future studies, should test LLMs on other working memory span tasks used in cognitive science (Conway et al. 2005; Daneman and Carpenter 1980) to address the generalisability of n-back tasks as measurement tools. Last but not least, our research opens up a brand new question in the field of LLMs: if the working memory capacity of LLMs is fundamentally limited, then why? What does it tell us? And how can we improve it?

References

Amon, M. J.; and Bertenthal, B. I. 2018. Auditory Versus Visual Stimulus Effects on Cognitive Performance During the N-back Task. In *CogSci*.

Au, J.; Sheehan, E.; Tsai, N.; Duncan, G. J.; Buschkuehl, M.; and Jaeggi, S. M. 2015. Improving Fluid Intelligence with Training on Working Memory: A Meta-Analysis. *Psychonomic Bulletin & Review*, 22(2): 366–377.

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

Bakker, M.; Chadwick, M.; Sheahan, H.; Tessler, M.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35: 38176–38189.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chen, Y.; Zhong, R.; Zha, S.; Karypis, G.; and He, H. 2021. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*.

Chooi, W.-T.; and Logie, R. 2020. Changes in Error Patterns during N-back Training Indicate Reliance on Subvocal Rehearsal. *Memory & Cognition*, 48(8): 1484–1503.

Cochrane, A.; Simmering, V.; and Green, C. S. 2019. Fluid Intelligence Is Related to Capacity in Memory as Well as Attention: Evidence from Middle Childhood and Adulthood. *PLOS ONE*, 14(8): e0221353.

Conway, A. R. A.; Kane, M. J.; Bunting, M. F.; Hambrick, D. Z.; Wilhelm, O.; and Engle, R. W. 2005. Working Memory Span Tasks: A Methodological Review and User's Guide. *Psychonomic Bulletin & Review*, 12(5): 769–786.

Conway, A. R. A.; and Kovacs, K. 2020. *Working Memory and Intelligence*, 504–527. Cambridge Handbooks in Psychology. Cambridge University Press, 2 edition.

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.

Cowan, N. 2015. George Miller's Magical Number of Immediate Memory in Retrospect: Observations on the Faltering Progression of Science. *Psychological review*, 122(3): 536–541.

Daneman, M.; and Carpenter, P. A. 1980. Individual Differences in Working Memory and Reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4): 450–466.

Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; and Smith, N. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Efrat, A.; and Levy, O. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Engle, R. W. 2002. Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, 11(1): 19–23.

Engle, R. W.; Kane, M. J.; and Tuholski, S. W. 1999. *Individual Differences in Working Memory Capacity and What They Tell Us About Controlled Attention, General Fluid Intelligence, and Functions of the Prefrontal Cortex*, 102–134. Cambridge University Press.

Gaspar, J. M.; Christie, G. J.; Prime, D. J.; Jolicœur, P.; and McDonald, J. J. 2016. Inability to suppress salient distractors predicts low visual working memory capacity. *Proceedings of the National Academy of Sciences*, 113(13): 3693–3698.

Guo, F.; He, R.; Dang, J.; and Wang, J. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7822–7829.

Halford, G. S.; Cowan, N.; and Andrews, G. 2007. Separating Cognitive Capacity from Knowledge: A New Hypothesis. *Trends in cognitive sciences*, 11(6): 236–242.

Jaeggi, S. M.; Buschkuehl, M.; Jonides, J.; and Perrig, W. J. 2008. Improving Fluid Intelligence with Training on Working Memory. *Proceedings of the National Academy of Sciences*, 105(19): 6829–6833.

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

Kane, M. J.; and Engle, R. W. 2002. The Role of Prefrontal Cortex in Working-Memory Capacity, Executive Attention, and General Fluid Intelligence: An Individual-Differences Perspective. *Psychonomic Bulletin & Review*, 9(4): 637–671.

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

Klatzky, R. L.; Giudice, N. A.; Marston, J. R.; Tietz, J.; Golledge, R. G.; and Loomis, J. M. 2008. An n-back task using vibrotactile stimulation with comparison to an auditory analogue. *Behavior research methods*, 40(1): 367–372.

Kosinski, M. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Li, D.; Rawat, A. S.; Zaheer, M.; Wang, X.; Lukasik, M.; Veit, A.; Yu, F.; and Kumar, S. 2022. Large Language Models with Controllable Working Memory. arXiv:2211.05110.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Mejías, J. F.; and Wang, X.-J. 2022. Mechanisms of Distributed Working Memory in a Large-Scale Network of Macaque Neocortex. *eLife*, 11: e72136.

Min, S.; Lewis, M.; Zettlemoyer, L.; and Hajishirzi, H. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Mishra, S.; Khashabi, D.; Baral, C.; Choi, Y.; and Hajishirzi, H. 2021a. Reframing Instructional Prompts to GPTk's Language. *arXiv preprint arXiv:2109.07830*.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2021b. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Mitchell, M. 2023. How do we know how smart AI systems are? *Science*, 381(6654): adj5957.

Oberauer, K.; Farrell, S.; Jarrold, C.; and Lewandowsky, S. 2016. What Limits Working Memory Capacity? *Psychological Bulletin*, 142(7): 758–799.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Rubin, O.; Herzig, J.; and Berant, J. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Salthouse, T. A.; and Pink, J. E. 2008. Why Is Working Memory Related to Fluid Intelligence? *Psychonomic bulletin & review*, 15(2): 364–371.

Shalchy, M. A.; Pergher, V.; Pahor, A.; Van Hulle, M. M.; and Seitz, A. R. 2020. N-Back Related ERPs Depend on Stimulus Type, Task Structure, Pre-processing, and Lab Factors. *Frontiers in Human Neuroscience*, 14.

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

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wilhelm, O.; Hildebrandt, A.; and Oberauer, K. 2013. What Is Working Memory Capacity, and How Can We Measure It? *Frontiers in Psychology*, 4.