Working Memory Capacity of ChatGPT: An Empirical Study

Anonymous Author(s)
Affiliation
Address
email

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 (GPT-3.5), 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 experiences significant declines in performance as *n* increases (which necessitates more information to be stored in working memory), suggesting a limit to the working memory capacity 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 and deepening our understanding of human working memory through AI models.

1 Introduction

The advent of large language models (LLMs) like ChatGPT and GPT-4 [31] has propelled the pursuit of artificial general intelligence [5] and unveiled human-level abilities that warrant further exploration [39, 22]. 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 sciences, working memory is usually defined as the ability to temporarily store and manipulate information in mind [1]. 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 [9].

Studies on human participants have revealed a fundamental capacity limit in working memory [10]. However, there has not been a consensus on why and how working memory capacity is limited [30, 41]. Among many theories, the executive attention hypothesis [16, 15] suggests that working memory requires using attention to maintain or suppress information, and the constraint on working memory capacity is not really about memory storage *per se*, but about the capacity for controlled, sustained attention in the face of interference.

Supporting evidence of the executive attention hypothesis includes results from the *n*-back task, which is arguably the current gold standard measure of working memory capacity in the cognitive neuroscience literature (for a review, see [20]). The *n*-back task, initially developed by Kirchner [21], 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

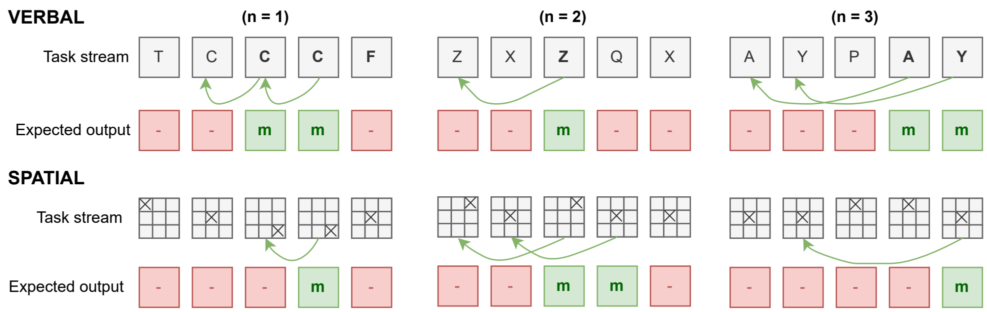

Figure 1: Illustrations of verbal (**top row**) and spatial (**bottom row**) n-back tasks with $n = \{1, 2, 3\}$. Participants are instructed to give a response ("`m`") when the current stimulus (e.g., a letter or a spatial location) is the same as the stimulus *n* trials ago), and not respond ("`-`") on nonmatch trials.

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. Typical human performance (measured by accuracy) as a function of *n* is shown in Figure 2, where we used the data presented in [19].

In humans, working memory capacity has proved to be closely related with fluid intelligence (Gf) or general intelligence (*g*) [7, 34], placing working memory at the core of human intelligence. 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. In the current study, we define working memory of LLMs as an emergent ability to selectively maintain and manipulate information for ongoing cognitive processes, echoing the executive attention hypothesis in cognitive sciences. We propose 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.

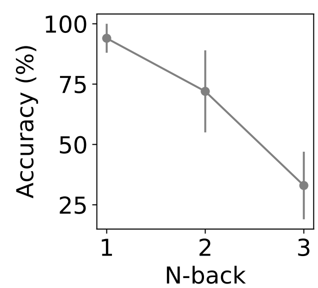

Figure 2: Typical human performance to the n-back tasks for $n = \{1, 2, 3\}$. We plot the mean $\pm$ standard deviation of the data collected in [19].

To demonstrate this, we used ChatGPT (GPT-3.5) as a representative of LLMs, and designed two categories of the *n*-back task to evaluate its working memory capacity. Our results revealed strikingly consistent patterns of a capacity limit across multiple experimental conditions, hinting at possibly similar mechanisms of working memory in humans and LLMs. We believe this finding is important for both cognitive scientists and LLM researchers, and hope that this could guide future endeavors of better understanding why human working memory capacity is limited and building more intelligent LLMs with better working memory capacity.

2 Related Works

Working memory has long been a subject of study in human and animal cognition [11]. Unlike long-term memory, which is stored in long-term synaptic weights in the neural system, working memory is believed to be maintained by sustained activations of neurons in prefrontal cortex [26]. This working mechanism bears striking resemblance to the in-context learning ability found in LLMs. However, the investigation of working memory in LLMs remains largely unexplored. Therefore, exploring the working memory capacity of LLMs holds great interest and significance, as it can contribute to the development of more powerful models [17, 18, 42, 23].

Large language models 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 transferring to new tasks [13, 38, 2], 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 [4] ,

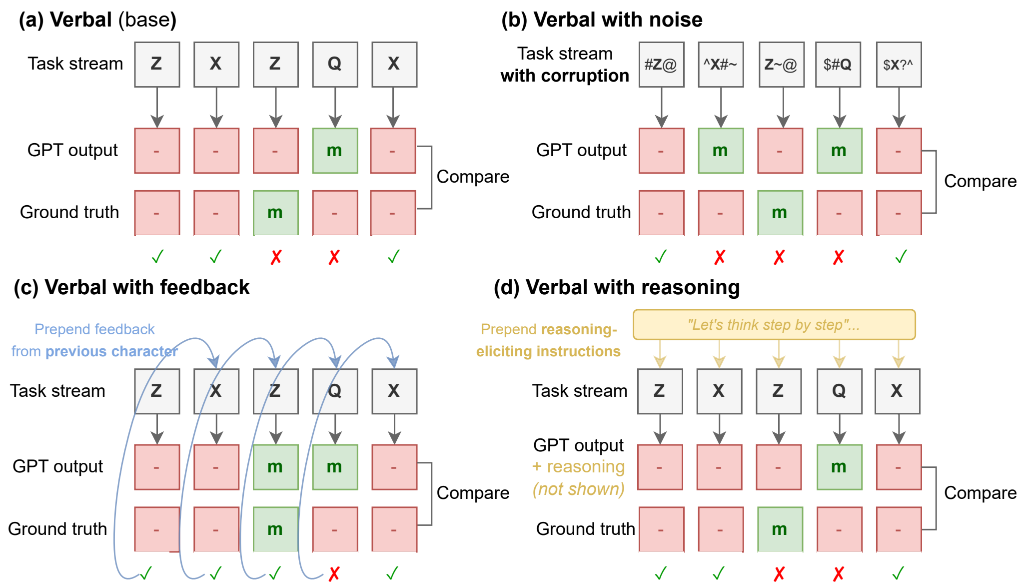

Figure 3: Illustrations of the different variants of *verbal*[*] n-back tasks (we use $n = 2$ in the figure) considered in this paper. **(a)**: base case identical to the case presented in Figure 1; **(b)**: stimulus on each trial now contains 3-6 random noise characters (chosen from "`#$%&@^~`") in 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 the feedback on whether it has performed the previous trial correctly; **(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 with in each of the task variants.

[*]Note that both verbal and spatial tasks are compatible with these variants; we illustrate using verbal tasks without loss of generality.

showcasing the remarkable few-shot learning capabilities of large language models without requiring weight updates through gradient descent. Since its introduction, research on in-context learning in language models has garnered significant attention from both 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 [33, 25, 24], meta-training with an explicit in-context learning objective [6, 27], and exploring the variant of in-context learning that involves learning to follow instructions [40, 38, 14, 28, 29]

However, relatively less work has been done to calibrate the working memory capacity of LLMs and understand the limitation of in-context learning ability. To the best of our knowledge, this paper is the first that provides an empirical analysis from the neuroscience view that investigates the working memory ability of LLMs.

3 Methods

We devised two categories of n-back tasks involving verbal and spatial working memory [36] respectively, and prompted ChatGPT (using the OpenAI API, model = "`gpt-3.5-turbo`") 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 to further test the model's performance under different conditions.

3.1 Verbal n-back experiments

In the base version of the verbal n-back task, 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

Table 1: Prompts used for 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 with noise (Figure 3b) 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 with feedback (Figure 3c) 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.

respond with "*m*" on match trials and "-" on nonmatch trials. Apart from the above base version, we further explored the behavioural performance of ChatGPT with the following modifications of the task presented in Figure 3:

- We added $3 - 6$ noise symbols to the input on every trial to examine the LLM's behaviour when to make it impossible to get the correct answer by simply doing string match between stimulus inputs.
- In human behavioural studies, a common strategy to improve participants' performance is to provide feedback after each trial [35]. Here in the task, after the LLM provided a response for the previous trial, we added feedback on whether its response was correct or wrong alongside the stimulus input of the current trial.
- Chain-of-thought (CoT) prompting has proved helpful in eliciting reasoning in LLMs [40]. Here we instructed the LLM to think step by step when giving a response.

3.2 Spatial *n*-back experiments

Although in its very nature, LLMs are text-based, but at least one study has demonstrated that they have spatial reasoning abilities [5]. 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, we constructed a 3×3 grid using ASCII characters (see Table 2 for detailed prompts). For $n = 1, 2, 3$ 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. Like in the verbal *n*-back tasks, the LLM was instructed to respond with "*m*"

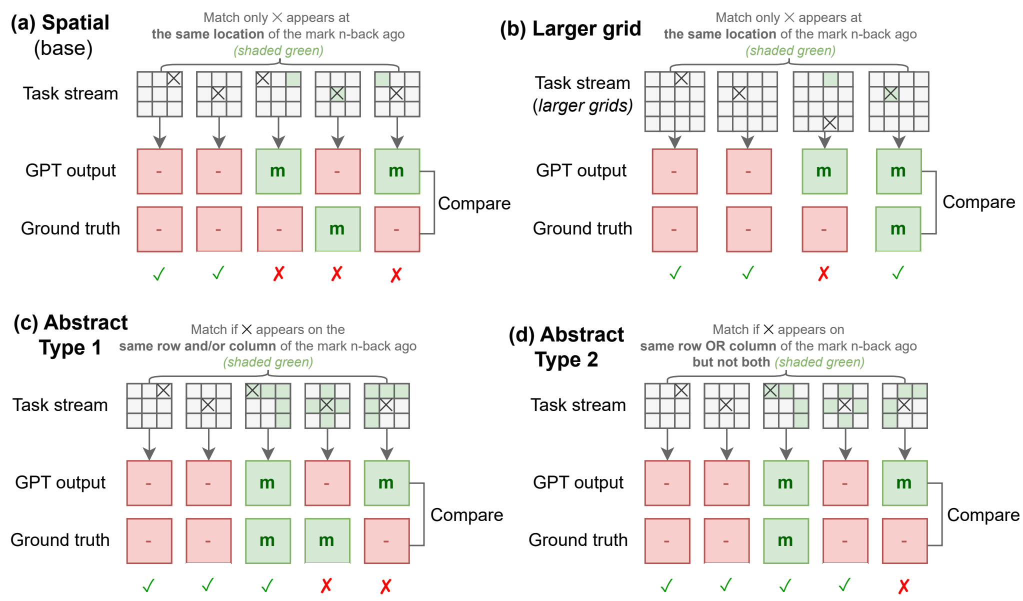

Figure 4: Illustrations of the different variants of *spatial* n-back tasks (we use $n = 2$ in the figure) considered in this paper *in addition to the variants presented in Figure 3*, which are applicable to both the spatial and the verbal tasks. **(a)**: base case identical to the case presented in 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 $\times$ mark occurs *the same row and/or column* at the same location n-back ago; in (**d**) a match is expected when $\times$ appears in the same row or column at the location n-back ago, *but not both*. Refer to Table 2 on the detailed instructions the LLM is prompted with for each of the variant.

on match trials and "-" on nonmatch trials. We further explored the spatial working memory capacity of ChatGPT with the following modifications of the task (3:

- As in the variants of verbal *n*-back tasks, we also have "spatial-with-noise", "spatial-with-feedback", and "spatial-with-CoT-reasoning" versions of the task. The prompts for the the with-feedback and with-reasoning versions were basically the same as those for the corresponding verbal tasks (see Table 1). 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. This is a first step to examine the LLM's spatial working memory when it was not able to get the correct answer by simply doing string match.

- To further confirm that the LLM can *really* reason in a spatial way rather than trivially performing some kind of string match under the hood, we further introduced two variants that specifically require abstract spatial reasoning; an model that would otherwise simply match strings would have failed. To achieve so, in these two tasks, a match is defined as when the location of the letter **X** is *in the same row or column* as the **X** *n* trials ago. The difference is whether identical locations also count as a match. We expect the version excluding identical locations to be harder for the LLM to perform.

- We also explored whether the size of the grid (3×3, 4×4, 5×5) would influence the LLM's performance. To the best of our knowledge, there hasn't been human studies exploring how the number of all possible spatial locations would impact behavioural performance. In light of this, we didn't have specific assumptions for how the LLM would perform differently under these scenarios.

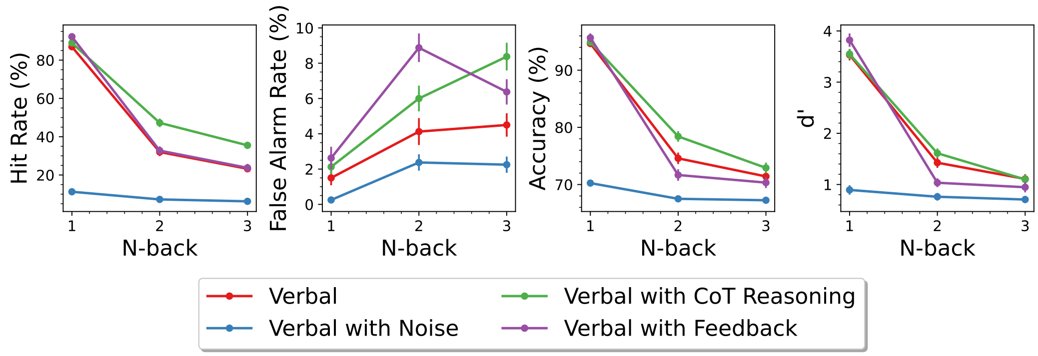

Figure 5: The results from the different variants of verbal n-back experiments. Error bars represent ± 1 *SEM*.

4 Results

To analyse the model's performance on our experiments, we used 4 widely accepted performance metrics reported in numerous human behavioral studies:

Hit Rate: It is a performance measure used in various fields, including computer science, statistics, and information retrieval. It represents the proportion of correct or successful outcomes out of the total number of targets or true positives. Mathematically, it is calculated by

$$\text{Hit Rate} = \frac{\text{Number of True Positives}}{\text{Number of True Positives + Number of False Negatives}} \tag{1}$$

False Alarm Rate: It quantifies the frequency at which a system or algorithm incorrectly identifies a negative outcome as positive. Mathematically, it is calculated by

$$\text{False Alarm Rate} = \frac{\text{Number of False Positives}}{\text{Number of False Positives + Number of True Negatives}} \tag{2}$$

Accuracy: It is a commonly used performance metric that measures the correctness or reliability of a system, model, or algorithm in making predictions or classifications. It represents the proportion of correct predictions or classifications out of the total number of predictions or classifications made. Mathematically, it is calculated by

$$\text{Accuracy} = \frac{\text{Number of Correct Responses}}{\text{Total Number of Trials}} \tag{3}$$

Detection Sensitivity (*d'*): It is a statistical measure used to assess the ability of a diagnostic test or classification model to accurately distinguish between two groups or conditions. It quantifies the extent to which the test or model can correctly identify positive cases relative to negative cases while minimizing false positives and false negatives. Mathematically, it is calculated by

$$d' = Z_{\text{Hit Rate}} - Z_{\text{False Alarm Rate}} \tag{4}$$

where $Z_{\text{Hit Rate}}$ and $Z_{\text{False Alarm Rate}}$ represent the z-score of *Hit Rate* and *False Alarm Rate*, respectively.

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 mean (*SEM*) and draw error bars to visualise the reliability of our findings (for further details on the statistics tests we performed, see **Supplementary Material**).

4.1 Verbal *n*-back experiments

In all versions of the task, we observed a performance pattern strikingly consistent with human participants, with the LLM's performance declining significantly when n increased from 1 to 3 5, as shown in hit rate, accuracy, and *d'*. Compared to the base version, the verbal-with-noise variant

Table 2: Prompts used for the *spatial* task variants described in Figure 4. 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. Note that spatial tasks with the variants described in Figure 3 are instead formatted similarly to Table 1.

<table>
<tr><th>Task type</th><th>Prompt</th></tr>
<tr><td>Spatial*
Spatial with Larger Grids
(Figure 4a-b)</td><td>You are asked to perform a {1,2,3}-back task. You will see a sequence of {3*3 [For larger grid (Figure 4b) only] 4*4,5*5,7*7} grids. Each grid has a letter X in one of the {nine, sixteen, twenty-five, forty-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 is in the same position as the previous grid/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.</td></tr>
<tr><td>Spatial with Abstract Reasoning
(Figure 4c-d)</td><td>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 grid/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 (Figure 4c) only] Note that ``` |X|_|_| |_|_|_| |_|_|_| ``` is also in the same row and column as ``` |X|_|_| |_|_|_| |_|_|_| ``` itself / [For Type 2 (Figure 4d) only] Note that if the X in the previous grid/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.</td></tr>
</table>

* For the prompts in spatial-with-noise, spatial-with-feedback, and spatial-with-CoT-reasoning tasks, refer to Table 1 for analogous examples.

significantly made the LLM's performance worse.We observe that while chain-of-thought prompting has significantly improved the performance of the language models in verbal task variants, feedback on whether the model has performed correctly in the previous task failed to meaningfully improve performance.

4.2 Spatial *n*-back experiments

In the four versions spatial tasks corresponding to the above verbal tasks, same patterns of performance were basically replicated (Figure 6). CoT reasoning significantly made the LLM perform better, adding noise made the model perform worse. But in all versions of the task, ChatGPT suffered significant declines in performance as n increases.

When further evaluated whether the LLM could conduct abstract spatial reasoning. The results confirmed so (Figure 7). In line with our prediction, the LLM performed worse when identical locations are not defined a match, which means more abstract spatial reasoning would be required in this scenario.

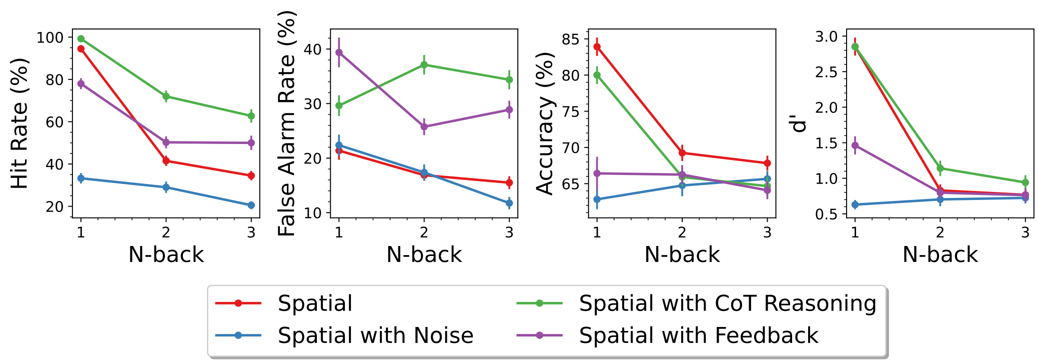

Figure 6: The results from the variants of spatial n-back experiments corresponding to those in verbal ones. Error bars represent ± 1 *SEM*.

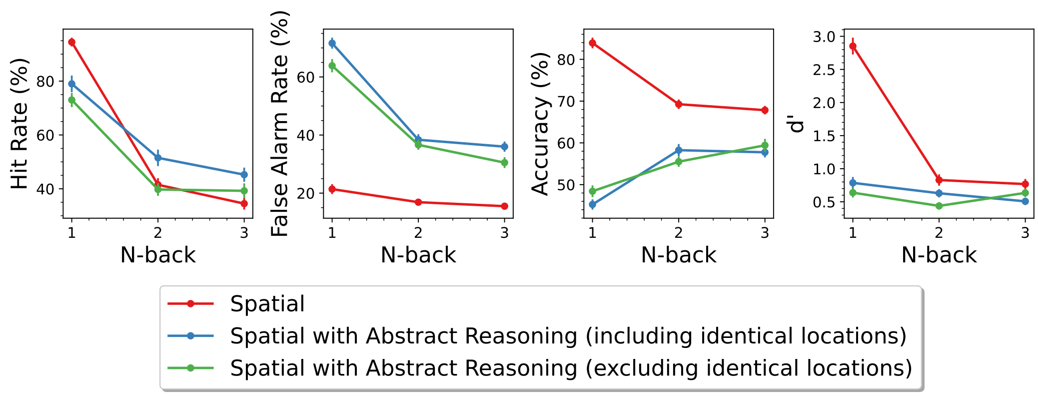

Figure 7: The results from the abstract reasoning variants of spatial n-back experiments. Error bars represent ± 1 *SEM*.

Our explorations on the effect of grid size on the model performance yielded interesting results, too. The LLM performed better when grid size was larger, especially as seen from the hit rate of *d'* results in Figure 8.

5 Discussion

We argue that our experimental results firmly point to the conclusion that ChatGPT has limited working memory capacity similar to humans. Even various prompting techniques (such as the provision of feedback and the use of state-of-the-art chain-of-thought (CoT) prompting [40]) may be used to improve its performance, the trend of performance decline as a function of increasing n still bears striking resemblance to human performance shown in numerous previous work. The consistent pattern of performance declines thus might be reflecting a fundamental constraint emerged from the architecture of the model, suggesting an possibility that the low-level working memory mechanism of LLMs might be similar to human working memory at least in some aspects.

In neuroscience, there are many unsolved problems on working memory, especially where and how working memory is encoded and maintained in the brain and why working memory capacity is limited. We propose that, in light of the above observation, ChatGPT and or other large language models of similar calibre could be potentially used and tested as a modelling platform for studying human working memory, just as what neuroscientists have done in recent years using other artificial neural networks [32]. Furthermore, future efforts aimed at interpreting activity of artificial neurons in LLMs [3] like ChatGPT would probably hold potential in informing the mechanisms of human working memory.

Our work also has some limitations. It would be important to test other LLMs on the same task we used here, to test whether they exhibit similar performance patterns and whether they have different

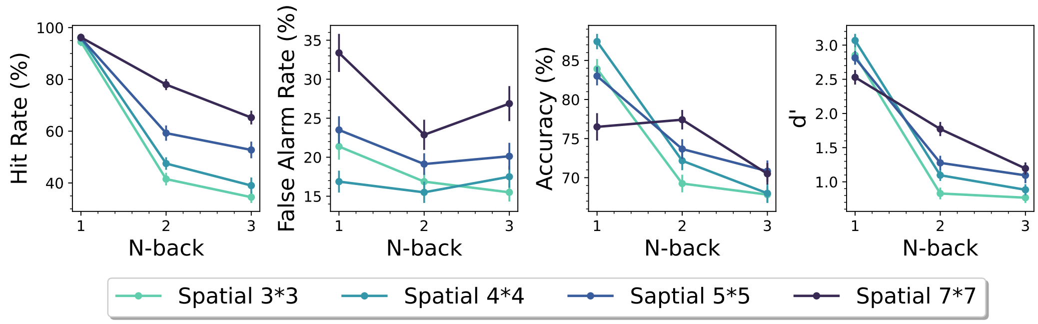

Figure 8: The results from spatial *n*-back variants with different grid sizes. Error bars represent ±1 *SEM*.

working memory capacity. It would also be helpful to test ChatGPT on other working memory span tasks used in cognitive sciences [8, 12] to address the generalisability of *n*-back tasks as measurement tools.

Last but not the least, the current work opens a brand new topic in probing the cognitive abilities of LLMs: if the working memory capacity of LLMs are fundamentally limited, then why? How their architecture is related to the capacity limit? One possible explanation would be the self-attention mechanism used in the Transformer architecture [37]. The self-attention mechanism computes a weighted sum of input elements, where each element's weight is determined by its relevance to other elements in the sequence. While this approach allows the model to focus on relevant information, it may also lead to a diffusion of information across multiple input elements, making it challenging to maintain and access specific pieces of information as n increases in *n*-back tasks.

References

[1] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[2] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[3] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

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

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021.

[7] Aaron Cochrane, Vanessa Simmering, and C. Shawn Green. Fluid intelligence is related to capacity in memory as well as attention: Evidence from middle childhood and adulthood. *PLOS ONE*, 14(8):e0221353, August 2019. `doi:10.1371/journal.pone.0221353`.

[8] Andrew R. A. Conway, Michael J. Kane, Michael F. Bunting, D. Zach Hambrick, Oliver Wilhelm, and Randall W. Engle. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5):769–786, October 2005. `doi:10.3758/BF03196772`.

[9] Andrew R. A. Conway and Kristof Kovacs. *Working Memory and Intelligence*, page 504–527. Cambridge Handbooks in Psychology. Cambridge University Press, 2 edition, 2020.

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

[11] Nelson Cowan. George Miller's Magical Number of Immediate Memory in Retrospect: Observations on the Faltering Progression of Science. *Psychological review*, 122(3):536–541, July 2015. doi:10.1037/a0039035.

[12] Meredyth Daneman and Patricia A. Carpenter. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466, August 1980. doi:10.1016/S0022-5371(80)90312-6.

[13] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[14] Avia Efrat and Omer Levy. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*, 2020.

[15] Randall W. Engle. Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1):19–23, 2002. arXiv:https://doi.org/10.1111/1467-8721.00160, doi:10.1111/1467-8721.00160.

[16] Randall W. Engle, Michael J. Kane, and Stephen W. Tuholski. *Individual Differences in Working Memory Capacity and What They Tell Us About Controlled Attention, General Fluid Intelligence, and Functions of the Prefrontal Cortex*, page 102–134. Cambridge University Press, 1999. doi:10.1017/CBO9781139174909.007.

[17] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[18] Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 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, pages 7822–7829, 2020.

[19] Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412, 2010.

[20] Michael J. Kane and Randall W. Engle. 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, December 2002. doi:10.3758/BF03196323.

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

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

[23] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory, 2022. arXiv:2211.05110.

[24] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[25] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[26] Jorge F. Mejías and Xiao-Jing Wang. Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *eLife*, 11:e72136, February 2022. doi:10.7554/eLife.72136.

[27] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[28] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*, 2021.

[29] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

[30] Klaus Oberauer, Simon Farrell, Christopher Jarrold, and Stephan Lewandowsky. What limits working memory capacity? *Psychological Bulletin*, 142(7):758–799, July 2016. doi:10.1037/bul0000046.

[31] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[32] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy De Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, November 2019. doi:10.1038/s41593-019-0520-2.

[33] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

[34] Timothy A. Salthouse and Jeffrey E. Pink. Why is working memory related to fluid intelligence? *Psychonomic bulletin & review*, 15(2):364–371, April 2008. doi:10.3758/PBR.15.2.364.

[35] Mahsa Alizadeh Shalchy, Valentina Pergher, Anja Pahor, Marc M. Van Hulle, and Aaron R. Seitz. N-Back Related ERPs Depend on Stimulus Type, Task Structure, Pre-processing, and Lab Factors. *Frontiers in Human Neuroscience*, 14, 2020.

[36] Arnaud Szmalec, Frederick Verbruggen, André Vandierendonck, and Eva Kemps. Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, 37(1):137, 2011.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[39] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[41] Oliver Wilhelm, Andrea Hildebrandt, and Klaus Oberauer. What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4, 2013.

[42] Aspen H Yoo and Anne GE Collins. How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *Journal of cognitive neuroscience*, 34(4):551–568, 2022.