Working Memory Capacity of ChatGPT: An Empirical Study

Dongyu Gong
University of Oxford
dongyu.gong@psy.ox.ac.uk

Xingchen Wan
University of Oxford
xwan@robots.ox.ac.uk

Dingmin Wang
University of Oxford
dingmin.wang@cs.ox.ac.uk

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-turbo`), 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 [34] has propelled the pursuit of artificial general intelligence [6] and unveiled human-level abilities that warrant further exploration [42, 25]. 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 temporarily store and manipulate information in mind [2]. 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 [11].

Studies on human participants have revealed a fundamental capacity limit in working memory [12]. However, there has not been a consensus on why and how working memory capacity is limited [33, 44]. Among many theories, the executive attention hypothesis [18, 17] 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 neuroscience

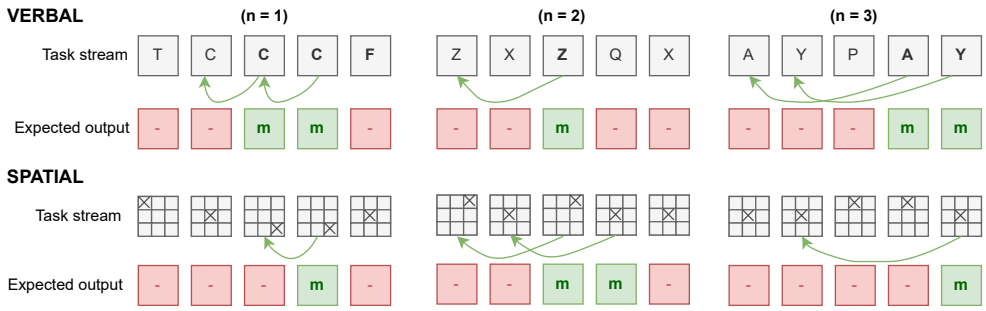

Figure 1: Illustrations of **verbal** (top row) and **spatial** (bottom row) *n*-back tasks with $n = \{1, 2, 3\}$. Participants are instructed to give a response ("`m`") when the current stimulus (e.g., a letter or a spatial location) is the same as the stimulus *n* trials ago), and not respond ("`-`")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

(for a review, see [23]). The *n*-back task, initially developed by Kirchner [24], 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. Typical human performance in this task (measured by accuracy) as a function of *n* is shown in Figure 2, where we plot the data presented in [22].

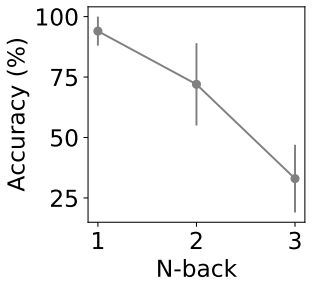

Figure 2: Typical human performance in *n*-back tasks for $n = \{1, 2, 3\}$. We plot the mean ± 1 standard deviation of the data collected in [22].

In humans, working memory capacity has proved to be closely related to fluid intelligence (*Gf*) [9, 37], 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 [1, 21], highlighting the special role of working memory capacity in human intelligence [20]. 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 science. 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.

To demonstrate this, we used ChatGPT (`gpt-3.5-turbo`) as a representative of LLMs, and designed two categories of *n*-back tasks 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 cognition [13]. Unlike long-term memory, which is stored in long-term synaptic weights in the neural system, working memory is believed to be maintained by activation of neurons in distributed brain networks [29]. However, the investigation of working memory in LLMs remains largely unexplored. A few latest studies in

this line has shown that studying and improving working memory of LLMs holds great interest and significance, as it can contribute to better performance of these models [19, 26].

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 [15, 41, 3], 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 [5] , 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 integrating the correct knowledge with the context, bears striking resemblance to how human working memory works. 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 [36, 28, 27], meta-training with an explicit in-context learning objective [7, 30], and exploring the variant of in-context learning that involves learning to follow instructions [43, 41, 16, 31, 32]

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.

3 Methods

We devised two categories of *n*-back tasks involving verbal and spatial working memory [39] respectively, and prompted ChatGPT (using the OpenAI API, model = "`gpt-3.5-turbo`", with default parameters) 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 (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 behavioural 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 behaviour when it is impossible to get the correct answer by simply doing string match between stimulus inputs (see Figure 3b).
- In human behavioural studies, a common strategy to improve participants' performance is to provide feedback after each trial [38]. 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 [43]. In this variant, we instructed the LLM to think step by step when giving a response (see Figure 3b).

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 [6]. 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** 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`" on

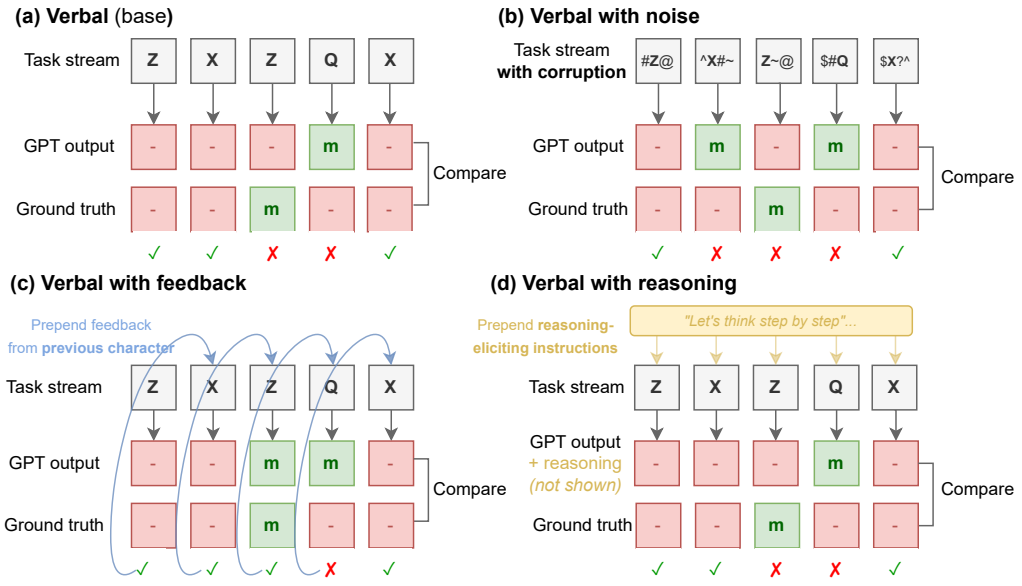

Figure 3: Illustrations of the different variants of **verbal*** n-back tasks (we use $n = 2$ in the figure) considered in this paper. **(a)**: base version identical to the case presented in Figure 1 (top row); **(b)**: stimulus on each trial now contains 3-6 random noise characters (chosen from "`#$%&@^~`") 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; **(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: both verbal and spatial tasks are compatible with these variants; we illustrate using verbal tasks without loss of generality.

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 basically 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 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; a hypothetical model that otherwise simply matches strings would not succeed in these variants. 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 complexity of the second variant, we expect it would be harder for the LLM to perform compared to the first variant.
- 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 has not been human studies exploring how the number of all possible spatial locations would impact behavioural 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.

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.

4 Results

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

Hit Rate: it is the proportion of correct identifications of the target (the stimulus that was *n* steps back). It can be calculated as follows:

$$\text{Hit Rate} = \frac{\text{Number of Hits}}{\text{Total Number of Targets}} \tag{1}$$

where *Number of Hits* is the number of times the target was correctly identified, and *Total Number of Targets* is the total number of targets that were presented during the task.

False Alarm Rate: it is the proportion of incorrect identifications of the target. It is the rate at which non-targets are incorrectly identified as targets. It can be calculated as follows:

$$\text{False Alarm Rate} = \frac{\text{Number of False Alarms}}{\text{Total Number of Non-Targets}} \tag{2}$$

where *Number of False Alarms* is the number of times a non-target was incorrectly identified as a target, and *Total Number of Non-Targets* is the total number of non-targets that were presented during the task.

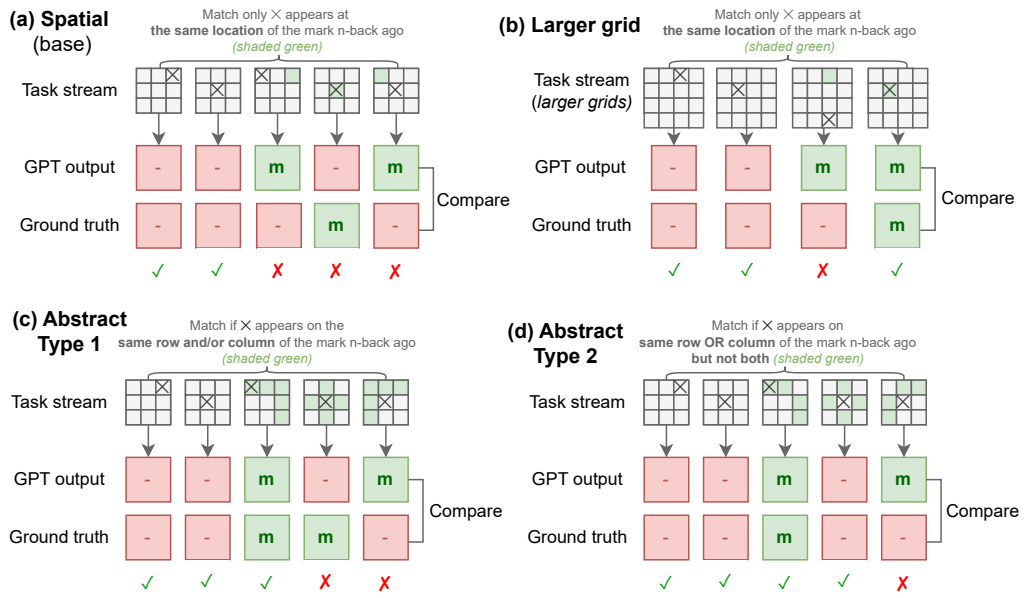

Figure 4: Illustrations of the different variants of **spatial** n-back tasks (we use $n = 2$ in the figure) considered in this paper *in addition to the variants presented in Figure 3*, which are applicable to both spatial and verbal tasks. **(a)**: base version identical to the case presented in Figure 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 on the detailed instructions the LLM is prompted with for each of the variant.

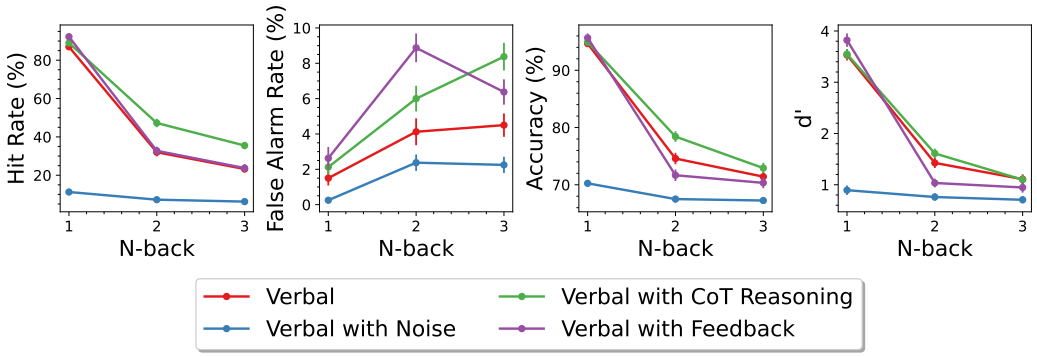

Figure 5: Results of different variants of verbal n-back experiments. Error bars represent ± 1 *SEM*.

Accuracy: it represents the overall correctness of responses, whether the stimulus is a target or a non-target. Accuracy can be calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Hits} + \text{Number of Correct Rejections}}{\text{Total Number of Trials}} \tag{3}$$

where *Number of Correct Hits* is the number of targets correctly identified, *Number of Correct Rejections* is the number of non-targets correctly identified (i.e., they were not incorrectly identified as targets), and *Total Number of Trials* is the total number of stimuli presented in a block, both target trials and non-target trials (i.e., 24, in our case).

Detection Sensitivity (d'): it is commonly used in signal detection theory and is a measure of sensitivity to distinguish between signal (target) and noise (non-target). In the context of the n-back

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: 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.

task, d' can be calculated using the z-scores (the inverse of the cumulative distribution function of a standard normal distribution) of the hit rate and the false alarm rate. The formula is as follows:

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

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 case where *Hit Rate* or *False Alarm Rate* is equal to either 0 or 1, they will be adjusted by 0.01 to handle the problem of z-score being infinite.

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. Among the four metrics, the pattern of hit rates and false alarm rates can vary a lot depending on the specific task condition [8]. Accuracy, in turn, will also be biased by very high/low hit rate and false alarm rate. 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. Conversely, a d' near 0 indicates performance no better than chance. Our analysis below will mainly rely on d' as the performance metric (see Appendix A for the statistics tests we conducted and Appendix B for performance distributions).

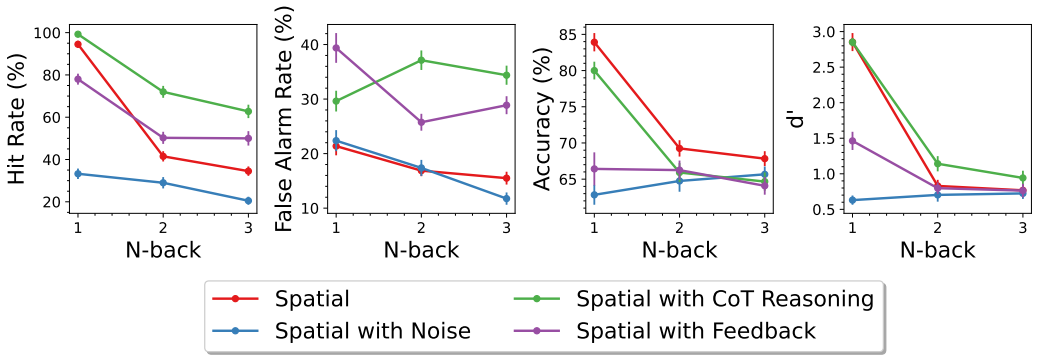

Figure 6: Results of the variants of spatial n-back tasks corresponding to those in verbal tasks. Error bars represent ± 1 *SEM*.

4.1 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 significantly when n increased from 1 to 3 (Figure 5). While CoT prompting has significantly improved the performance of the LLM, feedback on whether the model has performed correctly on the previous trial failed to meaningfully improve performance. On the other hand, adding noise made the model perform worse, as expected – these noises may be interpreted as analogous to distractors in human behavioural tasks.

4.2 Spatial n-back experiments

In the four versions of spatial tasks corresponding to the above verbal tasks, same patterns of performance declines were basically replicated (Figure 6). Interestingly, CoT prompting again significantly made the LLM perform better – this further confirms the hypothesis that the spatial n-back task presented to the LLM cannot be solved trivially with string similarity, as previous works on LLMs show that strong gain from CoT prompting is usually only present in tasks requiring advanced reasoning [43].

We further evaluated whether the LLM could conduct abstract spatial reasoning. Although for both types of abstract reasoning variants the d' was significantly lower than the base version, a closer look into the results shows that it was mainly driven by the disproportionately high false alarm rates in these two variants. If we focus on the hit rates, then clearly the LLM was able to conduct some abstract reasoning (Figure 7). Furthermore, 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.

Our explorations on the effect of the grid size on model performance yielded interesting results, too. The LLM performed better when the grid size was larger, especially as seen from the hit rate and 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 generalisability of this finding.

5 Discussion

Our consistent finding across nearly all tasks is that *ChatGPT suffers from significant declines in performance as n increases*. We argue that our experimental results firmly point to the conclusion that ChatGPT has limited working memory capacity similar to humans. Although various prompting techniques (such as the use of state-of-the-art CoT prompting [43]) may be used to improve the model's performance, the trend of performance declines as a function of increasing n still bears striking resemblance to humans. This consistent pattern thus might be reflecting a fundamental constraint emerged from the architecture of the model, suggesting a possibility that the low-level

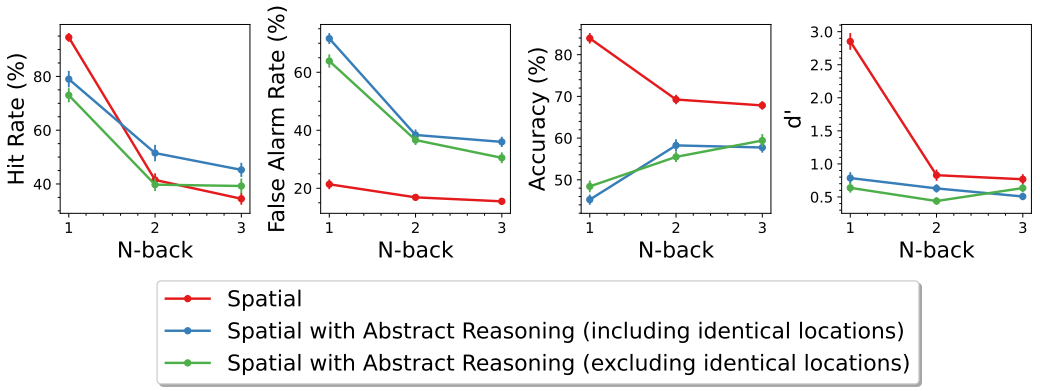

Figure 7: Results of abstract reasoning variants of spatial n-back tasks. Error bars represent ± 1 *SEM*.

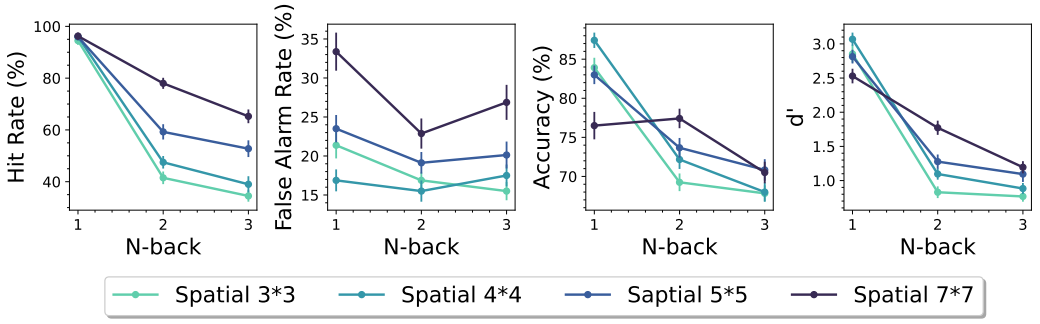

Figure 8: Results of spatial task variants with different grid sizes. Error bars represent ± 1 *SEM*.

mechanisms of working memory in LLMs might be similar to human working memory at least in some aspects.

In human neuroscience, numerous unresolved challenges persist pertaining to working memory. We propose that, in light of the above observation, ChatGPT and 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 [35]. Furthermore, future efforts aimed at interpreting activity of artificial neurons in LLMs [4] would probably hold potential in informing the mechanisms of human working memory. If we could visualise the activity of artificial neurons across different layers of the model when doing working memory tasks, that could probably shed some light on the neural representations of human working memory as well.

Our work also has some limitations. It would be important to test other LLMs on the same task we used here, to confirm whether they exhibit similar performance patterns and whether they have different working memory capacity. It would also be helpful to test ChatGPT on other working memory span tasks used in cognitive science [10, 14] to address the generalisability of n-back tasks as measurement tools. Furthermore, given that other non-verbal/spatial n-back tasks (e.g. auditory) have been previously used in human experiments, it would also be interesting to test LLMs on these novel task types, especially given that LLMs are becoming increasingly multi-modal and support a wide range of input and/or output formats.

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 [40]. 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] Jacky Au, Ellen Sheehan, Nancy Tsai, Greg J. Duncan, Martin Buschkuehl, and Susanne M. Jaeggi. Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, 22(2):366–377, April 2015. doi:10.3758/s13423-014-0699-x.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Weng-Tink Chooi and Robert Logie. Changes in Error Patterns during N-back Training Indicate Reliance on Subvocal Rehearsal. *Memory & Cognition*, 48(8):1484–1503, November 2020. doi:10.3758/s13421-020-01066-w.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] Graeme S Halford, Nelson Cowan, and Glenda Andrews. Separating Cognitive Capacity from Knowledge: A New Hypothesis. *Trends in cognitive sciences*, 11(6):236–242, June 2007. doi:10.1016/j.tics.2007.04.001.

[21] Susanne M. Jaeggi, Martin Buschkuehl, John Jonides, and Walter J. Perrig. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19):6829–6833, May 2008. doi:10.1073/pnas.0801268105.

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

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

[33] 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.

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

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

A Statistical Tests

Because the experimental data do not conform to the assumptions of parametric tests (normality and homogeneity of the variance), we used non-parametric Kruskal-Wallis H tests and reported H value, p value, and ϵ^2 (effect size) to investigate if there is a significant difference in d' across $n = \{1, 2, 3\}$. After that, we did non-parametric post-hoc Mann-Whitney U tests and reported U value, Bonferroni-corrected p value, and rank-biserial correlation (effect size) to examine if the differences between groups ($\{$1-back vs 2-back, 1-back vs 3-back, 2-back vs 3-back$\}$) are significant.

Table 3: Kruskal-Wallis H test statistics on **verbal** tasks.

Task	H	p	ϵ^2
Verbal	97.5376	6.60666e-22	0.649916
Verbal with Noise	3.91569	0.141162	0.0130319
Verbal with CoT Reasoning	99.4143	2.58493e-22	0.662683
Verbal with Feedback	94.9077	2.46072e-21	0.632025

Table 4: Mann-Whitney U test statistics on **verbal** tasks.

Task	Test	U	Bonferroni-corrected p	rank-biserial correlation
Verbal	1-back vs 2-back	2451.5	2.66392e-16	-0.9612
Verbal	1-back vs 3-back	2487	3.30459e-17	-0.9896
Verbal	2-back vs 3-back	1566	0.0849936	-0.2528
Verbal with Noise	1-back vs 2-back	1444.5	0.474433	-0.1556
Verbal with Noise	1-back vs 3-back	1513.5	0.15668	-0.2108
Verbal with Noise	2-back vs 3-back	1309	1	-0.0472
Verbal with CoT Reasoning	1-back vs 2-back	2443	4.64142e-16	-0.9544
Verbal with CoT Reasoning	1-back vs 3-back	2473	7.96749e-17	-0.9784
Verbal with CoT Reasoning	2-back vs 3-back	1778	0.000792086	-0.4224
Verbal with Feedback	1-back vs 2-back	2468.5	8.0401e-17	-0.9748
Verbal with Feedback	1-back vs 3-back	2460.5	1.24486e-16	-0.9684
Verbal with Feedback	2-back vs 3-back	1404	0.863005	-0.1232

Table 5: Kruskal-Wallis H test statistics on **spatial** tasks corresponding to the verbal ones.

Task	H	p	ϵ^2
Spatial	84.9206	3.62842e-19	0.564086
Spatial with Noise	0.63338	0.728557	-0.00929674
Spatial with CoT Reasoning	88.4591	6.18527e-20	0.588157
Spatial with Feedback	21.4206	2.23139e-05	0.132113

Table 6: Mann-Whitney U test statistics on **spatial** tasks corresponding to the verbal ones.

Task	Test	U	Bonferroni-corrected p	rank-biserial correlation
Spatial	1-back vs 2-back	2392	9.56099e-15	-0.9136
Spatial	1-back vs 3-back	2415.5	2.57312e-15	-0.9324
Spatial	2-back vs 3-back	1310	1	-0.048
Spatial with Noise	1-back vs 2-back	1250	1	0
Spatial with Noise	1-back vs 3-back	1142	1	0.0864
Spatial with Noise	2-back vs 3-back	1159	1	0.0728
Spatial with CoT Reasoning	1-back vs 2-back	2410	3.39674e-15	-0.928
Spatial with CoT Reasoning	1-back vs 3-back	2431	1.0369e-15	-0.9448
Spatial with CoT Reasoning	2-back vs 3-back	1472.5	0.376059	-0.178
Spatial with Feedback	1-back vs 2-back	1804	0.000401651	-0.4432
Spatial with Feedback	1-back vs 3-back	1847	0.000115822	-0.4776
Spatial with Feedback	2-back vs 3-back	1353	1	-0.0824

Table 7: Kruskal-Wallis H test statistics on the abstract reasoning variants of **spatial** tasks.

Task	H	p	ϵ^2
Spatial	84.9206	3.62842e-19	0.564086
Abstract Reasoning (includ. identical)	4.06941	0.130719	0.0140776
Abstract Reasoning (exclud. identical)	7.19739	0.0273595	0.0353564

Table 8: Mann-Whitney U test statistics on the abstract reasoning variants of **spatial** tasks.

Task	Test	U	Bonferroni-corrected p	rank-biserial correlation
Spatial	1-back vs 2-back	2392	9.56099e-15	-0.9136
Spatial	1-back vs 3-back	2415.5	2.57312e-15	-0.9324
Spatial	2-back vs 3-back	1310	1	-0.048
Abstract Reasoning (includ. identical)	1-back vs 2-back	1363	1	-0.0904
Abstract Reasoning (includ. identical)	1-back vs 3-back	1512	0.213143	-0.2096
Abstract Reasoning (includ. identical)	2-back vs 3-back	1469	0.39399	-0.1752
Abstract Reasoning (exclud. identical)	1-back vs 2-back	1547	0.121597	-0.2376
Abstract Reasoning (exclud. identical)	1-back vs 3-back	1204.5	1	0.0364
Abstract Reasoning (exclud. identical)	2-back vs 3-back	881.5	0.0332108	0.2948

Table 9: Kruskal-Wallis H test statistics on the **spatial** task variants with different grid sizes.

Task	H	p	ϵ^2
Spatial 3*3	84.9206	3.62842e-19	0.564086
Spatial 4*4	93.9609	3.95043e-21	0.625585
Saptial 5*5	73.0433	1.37675e-16	0.483288
Spatial 7*7	53.6315	2.25977e-12	0.351235

Table 10: Mann-Whitney U test statistics on the **spatial** task variants with different grid sizes.

Task	Test	U	Bonferroni-corrected p	rank-biserial correlation
Spatial 3*3	1-back vs 2-back	2392	9.56099e-15	-0.9136
Spatial 3*3	1-back vs 3-back	2415.5	2.57312e-15	-0.9324
Spatial 3*3	2-back vs 3-back	1310	1	-0.048
Spatial 4*4	1-back vs 2-back	2442.5	5.49518e-16	-0.954
Spatial 4*4	1-back vs 3-back	2470.5	1.07401e-16	-0.9764
Spatial 4*4	2-back vs 3-back	1477	0.35332	-0.1816
Saptial 5*5	1-back vs 2-back	2298.5	1.40244e-12	-0.8388
Saptial 5*5	1-back vs 3-back	2325.5	3.52397e-13	-0.8604
Saptial 5*5	2-back vs 3-back	1477.5	0.35177	-0.182
Spatial 7*7	1-back vs 2-back	1897	2.39505e-05	-0.5176
Spatial 7*7	1-back vs 3-back	2209	1.10069e-10	-0.7672
Spatial 7*7	2-back vs 3-back	1863	7.07856e-05	-0.4904

B Performance Distributions

To get a better sense of the LLM's performance across blocks, below we plotted the distributions of d' from all the tasks.

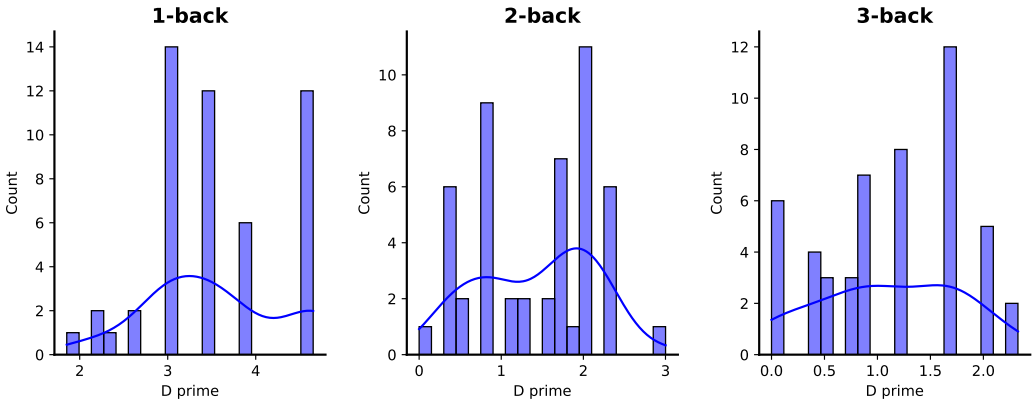

Figure 9: d' distributions: Verbal (Base Version).

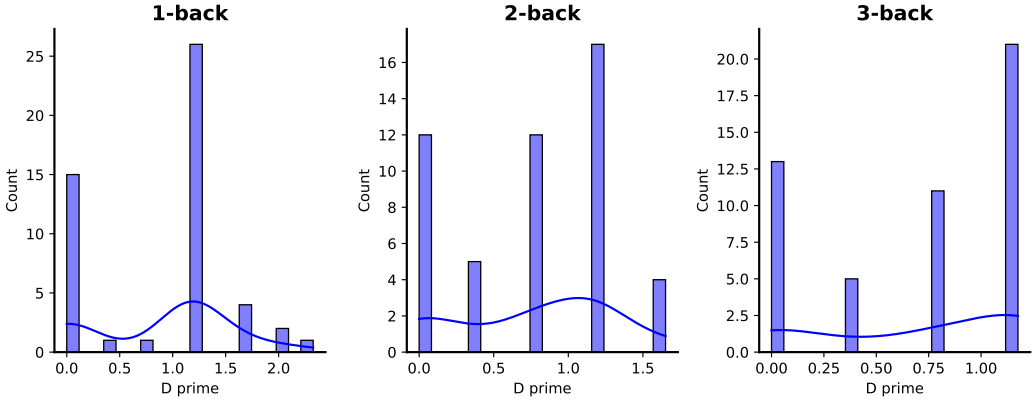

Figure 10: d' distributions: Verbal with Noise.

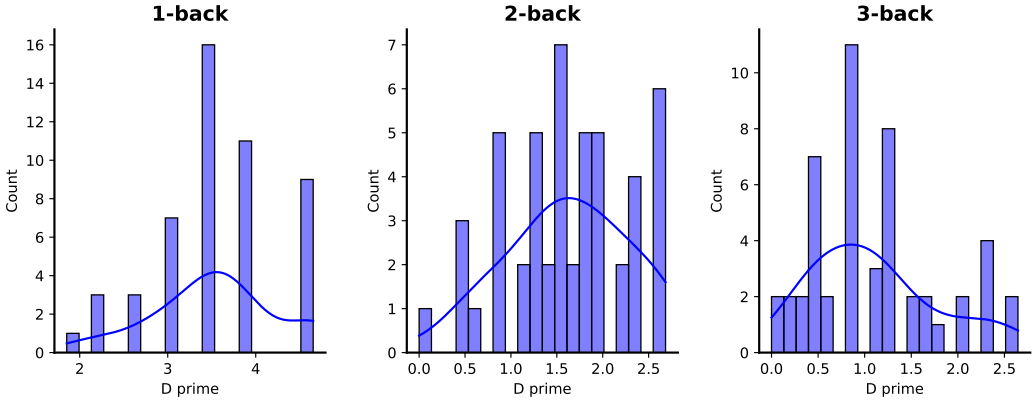

Figure 11: d' distributions: Verbal with CoT Reasoning.

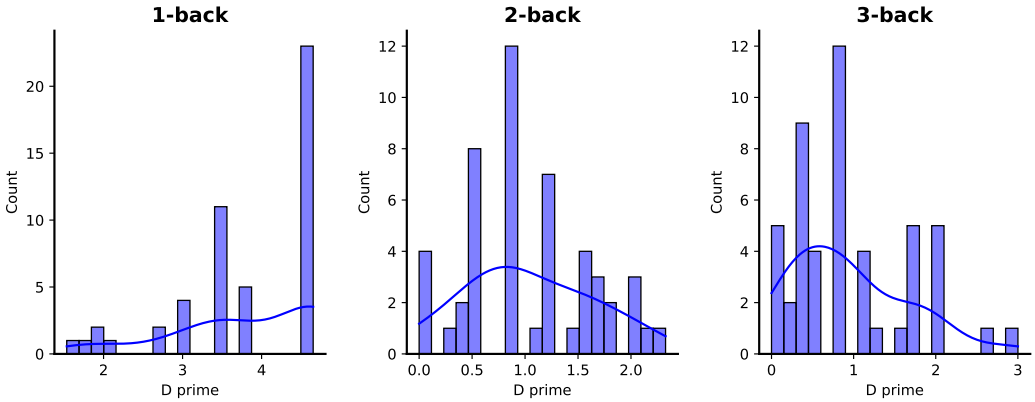

Figure 12: d' distributions: Verbal with Feedback.

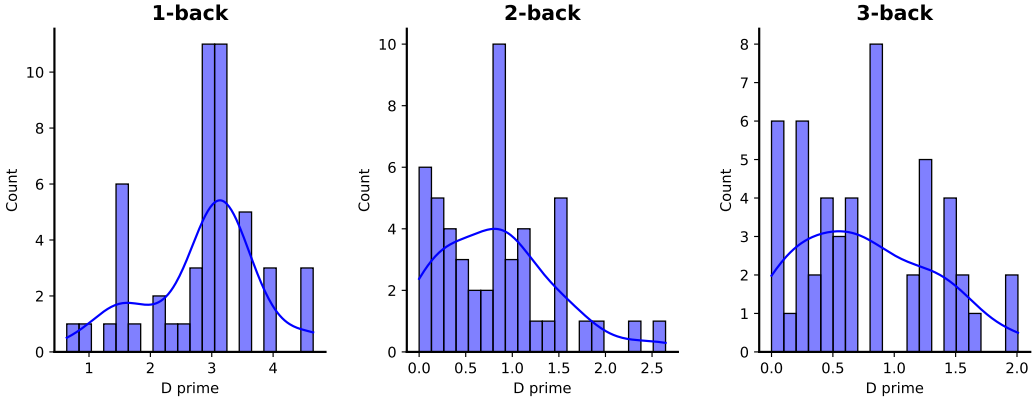

Figure 13: d' distributions: Spatial (base version).

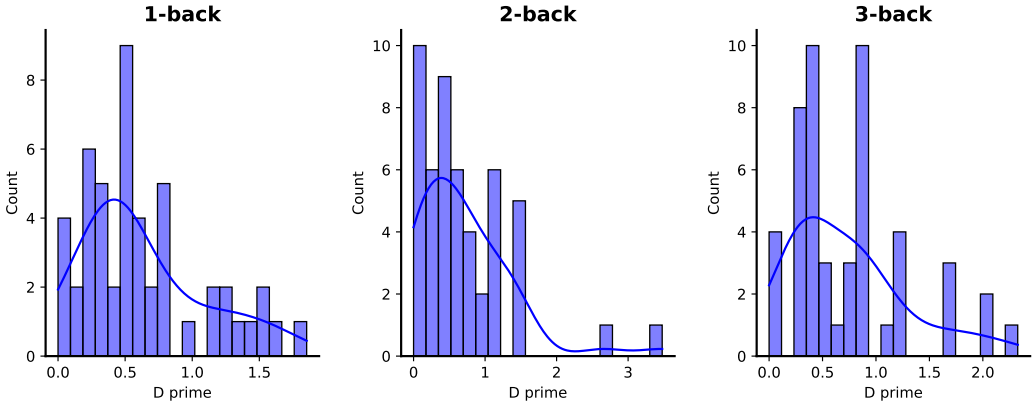

Figure 14: d' distributions: Spatial with Noise.

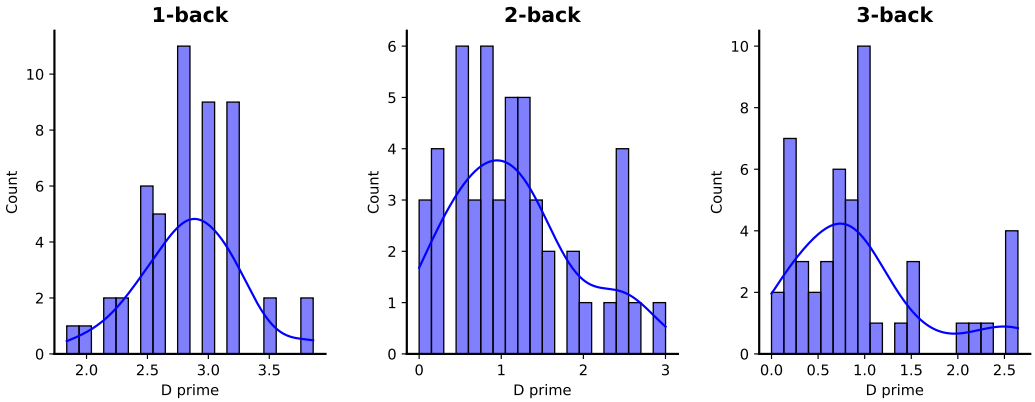

Figure 15: d' distributions: Spatial with CoT Reasoning.

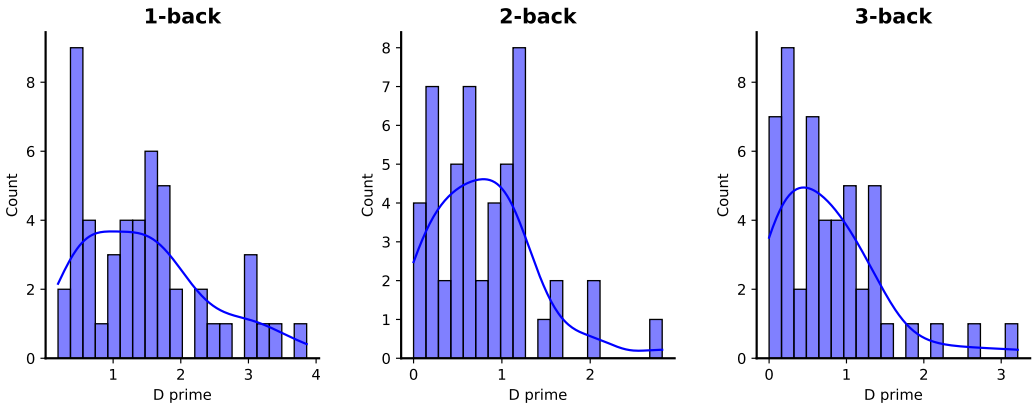

Figure 16: d' distributions: Spatial with Feedback.

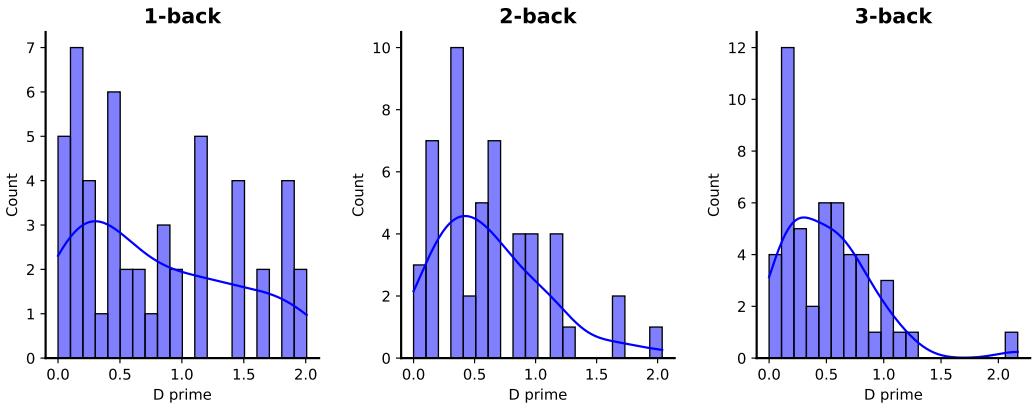

Figure 17: d' distributions: Spatial with Abstract Reasoning (including identical locations).

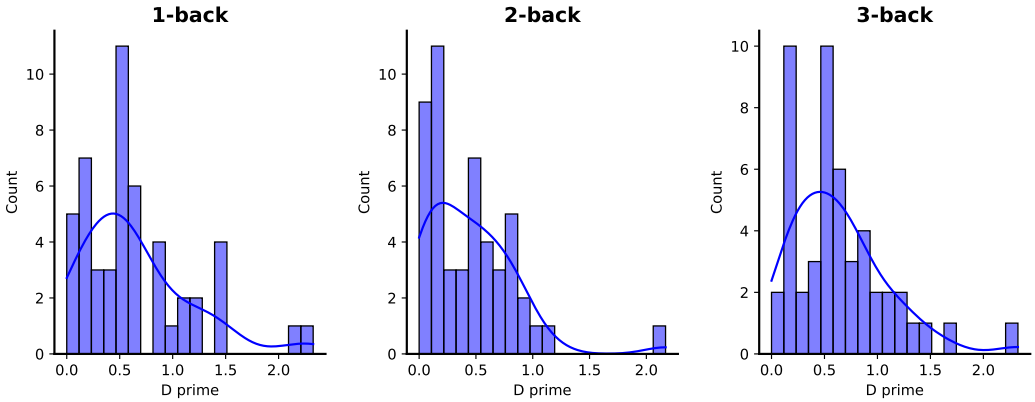

Figure 18: d' distributions: Spatial with Abstract Reasoning (excluding identical locations).

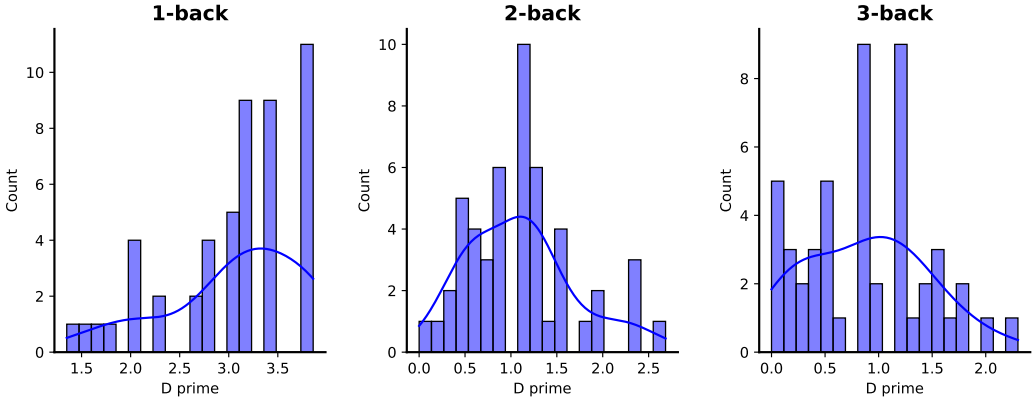

Figure 19: d' distributions: Spatial 4*4.

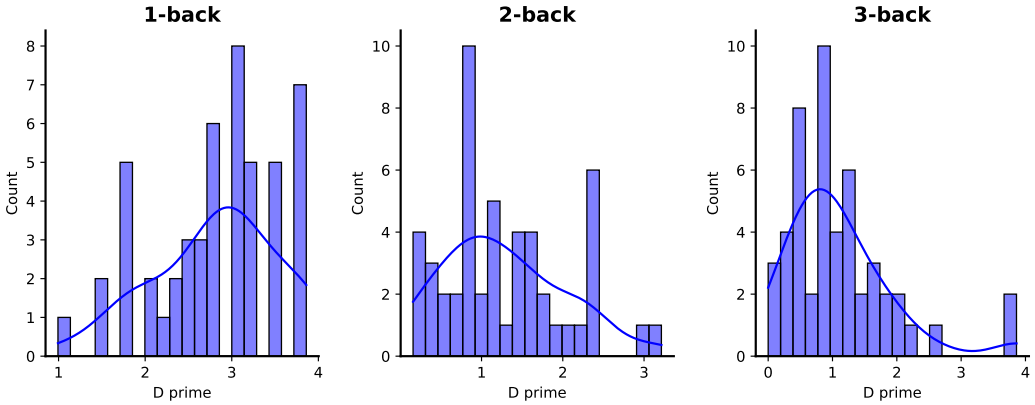

Figure 20: d' distributions: Spatial 5*5.

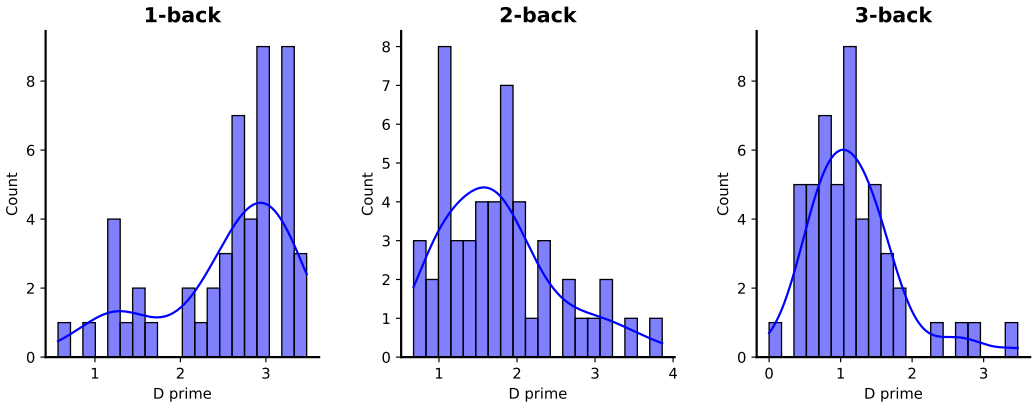

Figure 21: d' distributions: Spatial 7*7.