Strong Memory, Weak Control: An Empirical Study of Executive Functioning in LLMs

Karin de Langis[1], Jong Inn Park[1], Bin Hu[1], Khanh Chi Le[1], Andreas Schramm[2], Michael C. Mensink[3], Andrew Elfenbein[4], & Dongyeop Kang[1]

[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Linguistics, Hamline University
[3]Department of Psychology, University of Wisconsin-Stout
[4]Department of English and Center for Cognitive Sciences, University of Minnesota
{dento019@umn.edu}

Abstract

Working memory, or the ability to hold and manipulate information in the mind, is a critical component of human intelligence and executive functioning. It is correlated with performance on various cognitive tasks, including measures of fluid intelligence, which encompasses reasoning and problem solving. We use a comprehensive set of classic working memory tasks to estimate the working memory capacity of large language models (LLMs). We find that in most working memory tasks, LLMs exceed normative human scores. However, we do not find that the increased working memory capacity is associated with higher performance on other executive functioning tasks – specifically cognitive flexibility and attentional control. These results suggest that an important core deficit in LLM intelligence may stem from these deficits in adapting to shifting information and inhibiting automatic responses. We test two reasoning models and find that they do not fully compensate for these executive functioning deficits.[1]

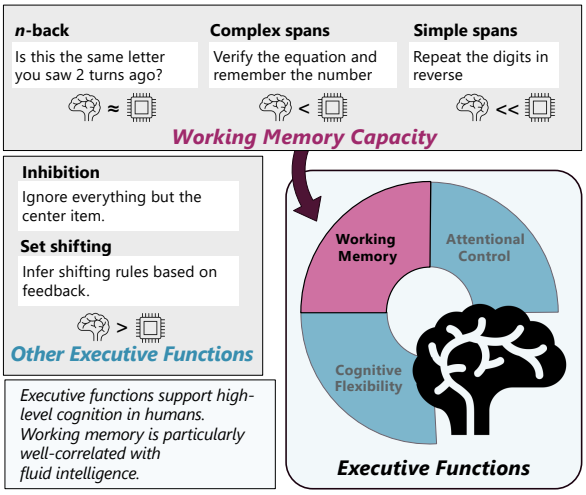


Figure 1: Working memory is a keystone component of high-level human cognition. We administer a standard set of working memory tasks to LLMs and find that they tend outperform humans, especially in the simplest tasks. However, we find that in tasks of executive function, the LLMs were generally below human baselines.

1 Introduction

Working memory is the cognitive function that handles storage and manipulation of task-relevant data. Many everyday tasks, from calculating a tip to playing a game of chess, require keeping relevant information in mind and mentally manipulating that information and thus rely on working memory. Working memory is correlated with fluid intelligence and performance on numerous cognitive tests (Conway et al., 2013), and it has been suggested to be "perhaps the most significant achievement of human mental evolution" (Goldman-Rakic, 1992).

Naturally, working memory is also of interest in the study of artificial intelligence (Yang et al., 2018) and more recently, in large language models (LLMs) (Gong et al., 2024; Zhang et al., 2024). Working memory in LLMs has been assessed via the n-back test, with results suggesting LLMs have human-like working memory. However, n-back is only one of several standard measures of working memory capacity, and it is often not the most representative metric (Frost et al., 2021).

We propose a more robust estimate of LLM working memory capacity by evaluating LLMs on a full suite of working memory tasks adapted from human studies, taking measures of simple and complex spans in addition to the n-back task. We also extend the study of working memory into executive functions more broadly. Executive functions are a group of high-level cognitive processes that enable sophisticated thinking (Diamond, 2013). There are three primary executive functions: (1) working memory, (2) inhibition / attentional control, and (3) cognitive flexibility. These functions, although correlated with one another, are also clearly separa-

[1]Our code is available at https://github.com/minnesotanlp/llm-exec-function.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5971–5986

ble based on both behavioral and neurological data (Miyake et al., 2000). We choose one test of attentional control and one test of cognitive flexibility to assess LLM executive functions beyond working memory.[2] Our experiments consider multiple model families, sizes, and prompts, and we compare both reasoning and instruction tuned models.

Our findings show that LLMs on most tasks have a substantially higher working memory capacity than humans do, and that this capacity is correlated with model size. Still, LLMs have relatively lower performance on complex working memory measures that have additional processing or manipulation demands. Since the transformer architecture offers much stronger information encoding than the biological brain, it is not clear whether advanced working memory capacity in LLMs will correlate to better performance in other executive functions. Our tests of other executive functions show no human-like correlation with working memory capacity, and further, we find that LLM performance on these tasks is *below* published normative data from human research. This suggests that *the bottleneck for LLM intelligence is not working memory, but other critical executive functions required to perform higher-order tasks*.

Finally, we examine the performance of reasoning model on these cognitive tasks, since the thought strings they produce could theoretically take the role of the central executive by carrying out selection and goal-oriented manipulation of task relevant information. We find mixed results in performance benefits: they improve performance on a single-turn attentional control task, but not on a multi-turn cognitive flexibility task. We also consistently observe the overthinking phenomenon (Chen et al., 2025) in which simple tasks result in excessively long thought strings, making performance very inefficient. With specific cognitive task training, reasoning models may learn to consistently serve as an artificial executive function mechanism, which may in turn open the door to more intelligent LLMs. However, in our experiments, we do not find any models that excel across all executive function tasks.

2 Related Work

Working memory is among the most studied cognitive ability in humans, perhaps because it has been correlated with a large number of cognitive skills. In particular, working memory plays an important role in executive function and problem solving, and thus has a high impact on goal-oriented behavior and critical thinking. For an excellent overview on the relation between working memory and intelligence, see Unsworth et al. (2014).

Working memory is a temporary information store with limited capacity (Baddeley and Hitch, 1994; Cowan, 2001). There are two critical components of working memory. The first is the maintenance or storage of relevant information – i.e., keeping the information "in mind" – and the second is the transformation, processing, or updating of that information as needed. Neurological studies indicate that the *maintenance* of information in working memory is enabled by sustained activations of neural circuits (Goldman-Rakic, 1995). The neural substrates behind the manipulation of this information are not yet well understood.

The study of working memory in LLMs is still developing. Gong et al. (2024) found similar performance between LLMs and humans on the n-back task, a common assessment of working memory. Later work further investigates possible prompting strategies to enhance model performance on the task (Zhang et al., 2024). Finally, drawing inspiration on cognitive difficulties associated with overloading working memory in humans, Upadhayay et al. (2025) demonstrate that overloading LLMs with unrelated tasks is an effective jailbreaking technique, suggesting that LLMs may also have a limited working memory capacity.

Other types of memory[3] have a longer history of study in LLMs. Long term memory, or the background knowledge learned during pre-training, is well-established (Petroni et al., 2019); the formation and forgetting of long-term memory over the course of pre-training has been investigated (Chang et al., 2024) and precise weights storing facts have been localized and causally demonstrated to facilitate factual recall (De Cao et al., 2021). Exact sequence memorization during training has also been extensively studied, with theoretical bounds for sequence lengths conditioned on the number of parameters in a transformers-based architecture being established (Kim et al., 2023; Kajitsuka and Sato, 2025; Yun et al., 2020; Zaheer et al., 2020). The associative memory capability of transformers

[2] Specifically, we administer the Eriksen flanker task and the Wisconsin Card Sorting Task (WCST).

[3] Note that working memory is functionally dissociable from other forms of memory in humans, as evidenced by double dissociation in patients with focal brain lesions.

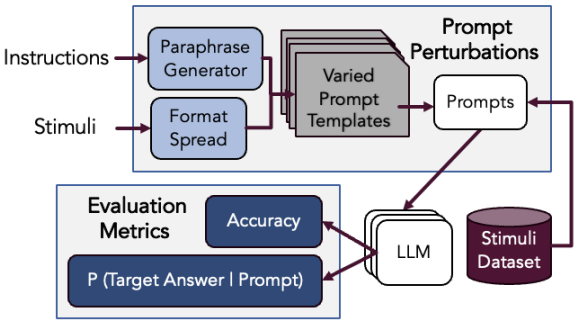


Figure 2: Our experimental pipeline.

has also been theoretically established (Jiang et al., 2024), and in-context memorization has been compared to declarative memory in humans (Li et al., 2024) and linked to the transformers architecture (Akyürek et al., 2023). Jelassi et al. (2024) show that the transformers architecture excels at repeating in-context sequences, which strongly resembles a short-term memory store in humans (see also (Wen et al., 2024)). Other features of human memory, such as recency effects and priming effects, have been documented in LLMs as well (Janik, 2023; Jumelet et al., 2024; Shaki et al., 2023). Our work expands the study of working memory in LLMs, with particular focus on whether LLM working memory modulates downstream skills as it does in humans.

3 Experimental Setup

Since cognitive processes cannot be directly observed and individual results may reflect task-specific artifacts, cognitive science aims to collect observations across diverse tasks and populations. Following this as the primary design principle, we introduce variations in our experimental setups (Figure 2): models, prompt templates, and task stimuli to see if we obtain comparable findings under different conditions.

Prompt Variations. When evaluating LLM responses to cognitive tasks, it is more desirable to marginalize over all prompts for a given task, rather than relying on one prompt to evaluate performance, due to high prompt sensitivity in LLM responses (Sclar et al., 2024; Wahle et al., 2024). We create thirty prompt permutations based on FORMATSPREAD (Sclar et al., 2024) and paraphrasing of instructions (Figure 2). For details, see §B.1. We establish that these prompts induce spread in model accuracy as expected (see §A.2).

Model Variations. We use six open-source LLMs (instruction-tuned variants) from three different model families, each with two size variations: Gemma2 with 9B and 27B parameters (Riviere et al., 2024), Llama3.1 with 8B and 70B parameters (Dubey et al., 2024), and Qwen2 with 7B and 72B parameters (Yang et al., 2024). The smaller models additionally vary in their number of attention heads, which is theoretically relevant to in-context memorization: Qwen2-7B has 28 attention heads, Gemma2-9B has 16, and Llama-3.1-8B has 32. We also test reasoning variants of two models: a DeepSeek-R1 (DeepSeek-AI et al., 2025) distilled Llama3.1-8B and Qwen3-8B. We use the Huggingface library (Wolf et al., 2020) for model inference, applying 4-bit quantization for larger models (27B or more parameters) to meet computational constraints. For experiments with long contexts, we also use Flash Attention (Dao et al., 2022). Otherwise, default model configurations and generation hyperparameters are used in all cases. All experiments are completed on a Linux server equipped with two Nvidia A100 GPUs.

Task Variation The following result sections answer three research questions, each addressed by a different set of tasks:

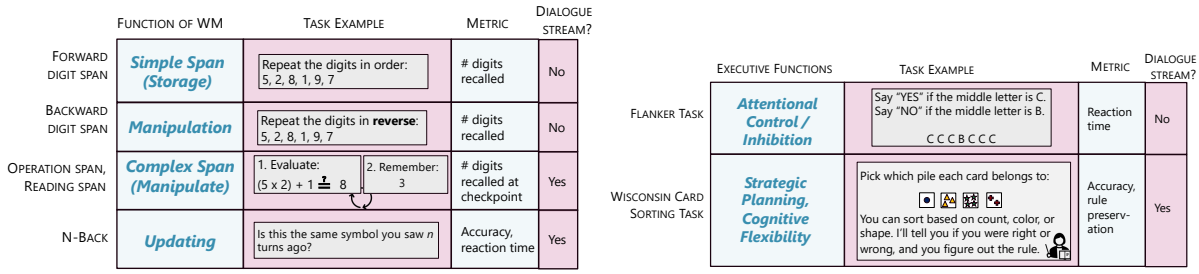
- **RQ1 (§4.1, §4.2):** Do LLMs show human-like working memory capacity? We measure access and manipulation limits in LLMs from five working memory tasks (Figure 4): two simple spans, two complex spans, and the n-back task.

- **RQ2 (§4.3):** Does working memory capacity in LLMs transfer to other executive functions? To assess executive functions, we measure tasks linked to attentional control and strategic flexibility: one simple task (flanker) and one complex task (WCST) (Figure 5). These tasks allow us to estimate both LLM working memory capacity and whether it supports broader cognitive abilities, i.e., executive functions.

- **RQ3 (§5):** Can reasoning models provide cognitive benefits that can supplement LLM cognition? We assess reasoning effects on working memory capacity (§5) and the cognitive tasks from RQ2.

4 Estimating working memory capacity

In this section, we detail experiments assessing working memory capacity in LLMs (§4.1, §4.2),



(a) Working memory assessments

(b) Executive function assessments

Figure 3: Brief overview of the cognitive assessments adapted for LLMs in the present study: (a) working memory tasks, spanning from simple input/output pairs to full dialogues, (b) executive function tasks, ranging relatively simple flanker task to complex Wisconsin Card Sorting.

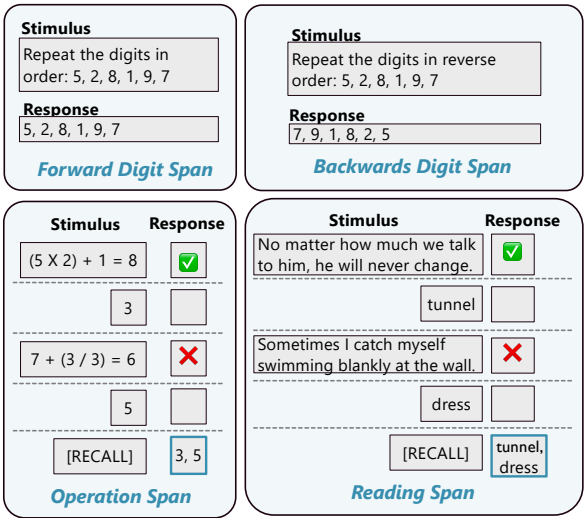


Figure 4: We asses both **simple span tasks** (top) and **complex span tasks** (bottom). The simple forward digit span task assesses the storage component of short term memory, whereas backward digit span has some overlap with working memory due to the manipulation component. On the other hand, complex span tasks also heavily tax information processing, making them a frequent choice to estimate of working memory spans.

followed by tests of other executive functions §4.3. Finally, we discuss how reasoning models perform on these assessments (§5).

4.1 LLMs generally have larger working memory spans than humans

Standard measures of working memory capacity are the backward digit span, operation span, and reading span (see Figure 4) and the n-back task. Humans struggle with these tasks because stimuli vanish instantly, subject to rapid decay, and easily disrupted by interference. Transformers, by contrast, do not suffer from sensory decay and can reliability reproduce long sequences with their context windows by leveraging the self-attention mechanism. (Jelassi et al., 2024).

Prior work shows that LLMs have human-like n-back performance (Gong et al., 2024). To assess LLM working memory capacity more robustly, we apply a battery of classic tasks used in psychology. (For details on specific prompting and evaluation set up, see §B.1.) We reproduce Gong et al. (2024)'s finding that LLMs have human-like performance on the n-back task, but for all other working memory tasks, *LLMs meet or surpass human baselines* (Table 1). Moreover, we find parameter count correlates strongly with mean span performance ($r = 0.82, p < 0.05$), while the number of attention heads shows a weaker trend ($r = 0.76; p = 0.08$). This suggests that while attention heads may contribute to straightforward repetition (e.g., forward digit span), they are *less predictive of tasks requiring information manipulation*, such as backward or flexible repetition.

4.2 LLMs excel at simple spans but complex spans induce more errors

Memory span tasks are often grouped into two categories: simple spans, i.e. forward and backward digit spans, and complex spans, i.e. operation and reading spans (Kane and Engle, 2003). In Table 1, we see that LLMs vastly exceed human performance on simple spans, but are less dramatically ahead in complex spans, indicating that additional information processing demands can hinder LLM information retrieval in a human-like fashion.

In a similar vein, consider LLM performance on the forward and backward digit spans. LLMs are perfect up to 50 forward digits (c.f. the mean 7 digits in humans), but already begin to falter at only 15 backward digits (c.f. the mean 5 digits in humans).[4]

[4]We do not find an upper bound on the forward digit span of LLMs.

Model	1-back	2-back	3-back	O-SPAN	R-SPAN	BDS ($d = 15$)	FDS ($d = 50$)
Gemma-2-9B	0.99	0.75	0.72	0.93	0.97	0.21	0.99
Gemma-2-27B	0.91	0.72	0.69	0.92	0.98	0.59	1.00
Llama-3.1-8B	0.76	0.68	0.67	0.99	0.92	0.18	1.00
Llama-3.1-70B	0.93	0.82	0.82	0.99	0.94	0.83	1.00
Qwen2-7B	0.99	0.89	0.85	0.96	0.66	0.00	1.00
Qwen2-72B	0.78	0.74	0.70	0.93	0.97	0.51	1.00
Human (approx)	0.98	0.91	0.75	0.53	0.48	0.00	0.00

Table 1: Average model accuracy on each test. Rough estimates of typical human scores are provided for reference (see §A.3). Note that for the digit span tasks, we include accuracies for very long strings ($d = 15$ and $d = 50$), while the typical human span is 5 (BDS) to 7 (FDS).

The reversal operation seems to interfere with recall ability. These findings suggest that while LLMs' architecture supports large working memory storage capacity, *their ability to concurrently integrate information processing and memory is more limited*. In Section 5, we examine whether reasoning models can better handle these complex working memory demands.

4.3 LLMs are below human baselines on tests of other executive functions

Human working memory capacity is correlated to a wide range of cognitive abilities, including other executive functions (Unsworth et al., 2014). Executive functions are a suite of high-level cognitive skills that support sophisticate goal-driven behavior. Having found that LLMs exceed human baselines of working memory capacity (§ 4.1), we next ask whether this advantage extends to other executive functions, which rely on working memory but require additional control and flexibility.

LLMs show low aptitude on WCST

We first test LLMs on the Wisconsin Card Sorting Task (WCST) (Grant and Berg, 1948), a widely studied measure of cognitive flexibility, abstract reasoning, and set-shifting (Miles et al., 2021). In this task, participants must infer and apply a sorting rule from feedback, and then adapt when the rule changes without warning (Figure 5 shows a schematic). Performance on this task is moderately correlated with working memory capacity on complex spans (Lehto, 1996; Dann et al., 2023).

We translate the WCST to a textual format for LLMs; however, we also test two multi-modal models on visual formats of the task. Each model receives 102 serially presented stimuli, with three examples illustrating each sorting rule. The experiment is conducted such that each sorting rule is presented for 10-15 consecutive turns. Each sorting rule (count, shape, color) is presented twice, and presented cards do not contain any ambiguous sorting options (Miles et al., 2021).

In addition, we note that the instructions for humans taking the WCST are intentionally vague: the possible sorting rules are not made explicit, and there are no specific suggestions regarding how to respond to the feedback. We test models on prompts adapted directly from the human instructions, but we also include a version with additional hints and strategy. Finally, we create a prompt that contains strategy and organizes all information using headers tailored for LLMs. (Specific prompts can be found in Appendix B.1.)

Strategy guidelines added to WCST prompts

On choosing a rule: (...) There is a simple strategy you can follow:
(1) if there is a previous answer, check the feedback.
(1a) If the feedback says correct, just apply the same rule.
(1b) If the feedback says incorrect, choose a *different rule* than the one you previously used.
(2) If there is no previous answer, randomly try a rule.

On importance of recent messages: Remember that the underlying rule can change, so the most recent feedback (from the last 1-2 turns) is the most important to consider – you should ignore all but the most recent messages.

Results (Table 2) show that *all LLMs perform well below human accuracy*, which is typically 70–80% (Barceló et al., 1997; Grant and Berg, 1948; Milner, 1963). Unlike humans, models fail to respond well to additional exposure of a given rule, suggesting difficulty in both inferring the underlying rule and maintaining it based on the feedback received within the dialogue (Fig. 6). This pattern may echo Coda-Forno et al. (2024), who found that LLMs under-weight observations in decision-

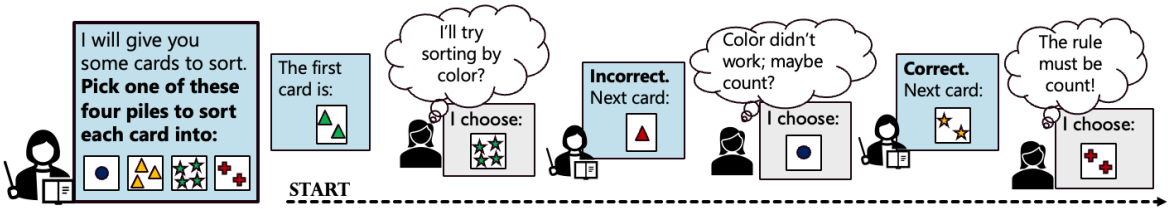
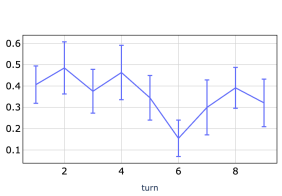
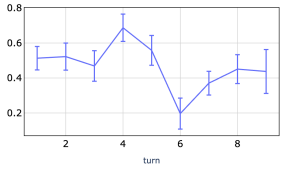


Figure 5: The Wisconsin Card Sorting Task (WCST). Participants are serially presented with cards to sort, and they must infer an underlying "sorting rule" based on feedback. *The rule will periodically change without warning*, and people must detect the change and adjust to the new rule. Participants make sorting decisions as quickly as possible. The WCST tests various executive functions, primarily cognitive flexibility as participants adapt from one set of rules to another.

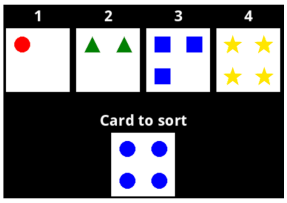


(a) Gemma-2-9B: Frequency of correct answers for turns 1 - 9 after seeing a new rule.



(b) Llama-3.1-70B: Frequency of correct answers for turns 1-9 after seeing a new rule .

Figure 6: Results from the WCST task show that some models have difficulty inferring and maintaining an underlying rule. The pattern suggests a *failure to maintain set*: it seems to infer a rule with a slight rise in accuracy, but the accuracy sharply drops, indicating "forgetting" of the current rule.



(a) Exemplary image shown to vision language models doing the WCST task.

Card to sort: Four blue circles

Options:

1. One red circle
2. Two green triangles
3. Three blue squares
4. Four yellow triangles

(b) Analogous text version of the WCST, for use in language-only models.

Figure 7: The WCST stimuli are adapted to both LLMs and VLMs. Our results do not indicate a strong advantage for either medium.

making tasks.

An important metric in WCST results is *preservation error*, or how frequently participants respond according to the previous rule rather than the current rule – humans have more preservation errors than other, non-preservation errors. Because in our experiments, LLMs fail to effectively learn a rule, there are few preservation errors and many non-preservation errors.

For context, in humans, WCST deficits of this magnitude would be consistent with clinical cognitive impairment (Hammers et al., 2016; Silva-Filho et al., 2007). While LLMs can perform well on highly advanced benchmarks, their failure to adapt to the WCST may indicate foundational weaknesses in executive function. The WCST may also hit on the same weaknesses LLMs have in completing the n-back task: both tasks are multi-turn and require constant updating of information.

LLMs flanker task performance is below human baselines

Executive functioning depends not only on flexibility but also on attentional control, the ability to prioritize task-relevant information while suppressing distractors. This faculty is tightly coupled with working memory capacity (Kane and Engle, 2003; Pratt et al., 2011), both relying on a central "attentional spotlight."

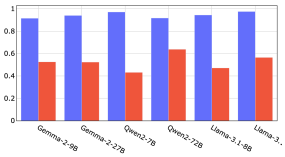
To test LLMs on attentional control, we use the classic Eriksen flanker task (Eriksen and Eriksen, 1974). In this task, participants respond to the central letter of a string: e.g., hit one button if the center letter is X or C, and another button if the center letter is B or V. Participants must ignore the "flanker" letters on either side of the center. Two types of sequences are shown:

- *Congruent trials* (e.g., 'X X C X X') have flankers with the same response as the target ('X' and 'C').
- *Incongruent trials* (e.g., 'B B C B B') place the flankers 'B' in conflict with target 'C,' requiring good attentional control to ignore the flanker.

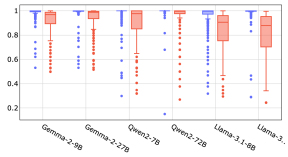
The key metric in this task is reaction time, which typically varies from 300-500ms. Participants have longer reaction times for incongruent strings because inhibiting the flankers requires additional

Model	Accuracy (↑)			Preservation Error (↓)			Other Error (↓)		
	Base Prompt	+Strategy	+Strat.,Hd	Base	+Strat.	+St,Hd	Base	+Strat.	+St,Hd
Llama-3.1-8B	0.31 (0.09)	0.28 (0.09)	0.33 (0.05)	0.02	0.00	0.03	0.67	0.72	0.64
Llama-3.1-70B	0.31 (0.06)	0.50 (0.07)	0.40 (0.07)	0.04	0.04	0.09	0.65	0.47	0.52
Gemma-2-9B	0.12 (0.06)	0.34 (0.09)	0.22 (0.10)	0.07	0.07	0.00	0.81	0.59	0.78
Gemma-2-27B	0.31 (0.06)	0.41 (0.10)	0.20 (0.18)	0.09	0.05	0.10	0.60	0.54	0.70
Qwen2-7B	0.30 (0.09)	0.39 (0.09)	0.30 (0.06)	0.08	0.07	0.07	0.62	0.54	0.63
Qwen2-72B	0.32 (0.05)	0.48 (0.10)	0.52 (0.13)	0.08	0.06	0.09	0.60	0.47	0.39
Qwen2.5-Vl-7B	0.41 (0.10)	0.38 (0.07)	0.35 (0.09)	0.07	0.10	0.07	0.52	0.52	0.58
Qwen2.5-Vl-32B	0.34 (0.09)	0.35 (0.05)	0.34 (0.07)	0.09	0.08	0.08	0.57	0.57	0.58
Healthy adults		0.77			0.12			0.09	

Table 2: Frequencies of correct answers and errors in WCST across different models and prompt types (due to rounding, all rows do not sum to exactly 1). Models are tested with not only prompts based on the instructions used for humans, but also prompts that additionally include explicit rules and strategy instructions (+ Strat) and more prompts that also organize all information neatly with clear delineated headers (+ Strat, Hd). Values in parentheses represent 95% CI. Human norms naturally vary slightly across studies; the reference values shown here are from a normative study by Barceló et al. (1997).



(a) Frequencies of correct responses on the flanker task. All models do substantially worse in the incongruent condition.



(b) For correct answers, confidence (probability of the correct answer) decreases in the incongruent condition, indicating interference from the flanking letters.

Figure 8: Flanker task: (a) average accuracy and (b) probabilities of correct answers. All models tested perform worse in the incongruent condition.

cognitive resources, which slows the response.

Because we cannot measure reaction times in LLMs, we instead measure accuracy and the probabilities assigned to correct answer. We find that model perform substantially worse in incongruent trials, with accuracies of $40-60\%$, compared to human performance of $93\%-96\%$ (Eriksen and Eriksen, 1974; Yantis and Johnston, 1990). Even when responses are correct, model confidence drops in incongruent cases (Figure 8), indicating a vulnerability to the distracting flankers.[5]

Together, the WCST and flanker task results suggest that LLMs may lack the executive mechanisms needed to strategically utilize their large working memory capacity. While they can access and reproduce long sequences, they *struggle with selecting relevant information, suppressing irrelevant inputs, and updating information flexibly.*

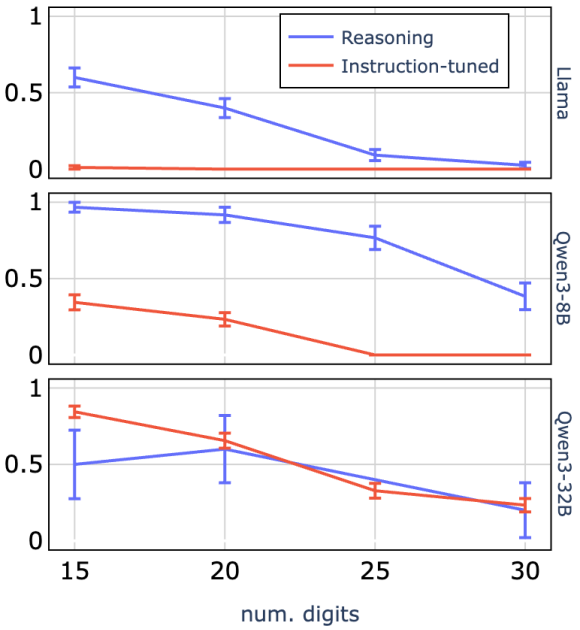
[5]We verify that tokenization of the letters never groups them such that two letters share the same token.

5 Capabilities of Reasoning models

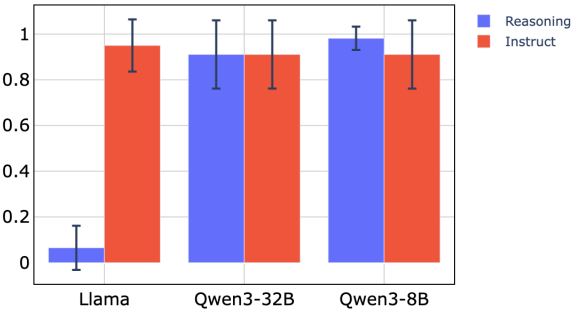
We next ask whether reasoning-enabled variants of LLMs show advantages of working memory and executive function. The rationale is that the intermediate "thought strings" produced by reasoning models might act as a pseudo central executive, supporting memory maintenance and executive attention or planning. Specifically, we evaluate Qwen3-8B and Qwen3-32B, with and without reasoning enabled, and compare Llama-3.1-8B (instruction-tuned) variant against its reasoning variant distilled from DeepSeekR1.

5.1 Reasoning does not consistently improve working memory capacity

Working memory capacity was tested using the backward digit span and reading span tasks (Sections 4.1, 4.2). Across both tasks, reasoning variants did not yield consistent improvements (Figure 9). While reasoning sometimes boosted weaker baselines (e.g., Llama-8B, Qwen-8B on backward digit span), accuracy remained well below ceiling and failed to generalize. Inspection of thought strings suggests a tendency toward *excessive deliberation*: models often looped through near-identical reasoning steps without converging, despite the very simple and straightforward natures of the tasks (Chen et al., 2025); for examples see §A.1. When reasoning models do markedly worse, we find it is because thought strings fail to converge, possibly due to the tasks being out-of-domain.



(a) Backward digit span accuracy across sequence lengths for reasoning vs. non-reasoning models. Reasoning aids weaker models (Llama-8B, Qwen-8B), but does not achieve perfect accuracy.



(b) Reading span accuracy. Reasoning variants do not outperform instruction-tuned.

Figure 9: Reasoning does not reliably improve working memory capacity. We find that the majority of errors are due to models failing to converge during their thought string or hallucinating during the thought string.

5.2 Reasoning improves attentional control on flanker task (at high computational cost)

In contrast, reasoning models show clear benefits on flanker task accuracy (Table 3). They outperform their non-reasoning counterparts on incongruent trials, though *at high computational cost*: models often generate thousands of tokens to solve a task humans complete in 300-500ms. Notably, reasoning sequences were equally long for congruent and incongruent conditions, suggesting that models do not capitalize on the simplicity of congruent cases (all letters require the same response, so it does not really matter which letter is in the middle). Instead, they adapt thought heuristics likely learned during training, such as: *"Maybe the middle letter isn't the exact center, but the third in the sequence."* These strategies appear inefficient.

Model	Congruent	Incongruent
Llama-3.1-8B	0.944	0.471
+Reasoning	0.977	**0.910**
Qwen2-32B	0.917	0.638
+Reasoning	0.994	**0.985**

Table 3: Flanker accuracies (simple attentional control) on the congruent vs. incongruent conditions. Adding reasoning allows models to achieve much higher accuracies in this task. Thought strings are several hundred tokens long, despite the straightforward and simple task.

5.3 Reasoning sometimes offers modest improvement on WCST

We also tested reasoning models on the Wisconsin Card Sorting Task (Section 4.3). Instruction-tuned LLMs performed substantially below healthy adult baselines, failing to adapt to feedback when inferring the rule. We find that reasoning variants of models sometimes fail to outperform their non-reasoning counterparts, or offer modest improvements that do not close the gap with human norms (Table 4).

Thought strings often failed to terminate under our 30,000 token limit (a generous threshold, given that humans make their choice in seconds). Inspection of thought strings revealed frequent hallucinations, irrelevant elaborations, and fabricated stimuli, rather than systematic strategic adjustment. The thought strings also frequently disregard the instructions on task strategy; for instance Qwen3-8B generates: "Wait, the previous rule was count, and the feedback was correct. But maybe the rule changed?" which directly contradicts the instructions.

6 Discussion

Our experiments consistently suggest that LLMs have a larger working memory capacity than humans do. This is not wholly unsurprising, given the propensity the transformers architecture has for copying in-context text. However, we critically find that LLM memory tends to get weaker when manipulation of in-context information is required (even a simple reversal, as in the backward digit span), when updating is required (e.g., n-back task),

		Accuracy (↑)
Llama-3.1-8B	Human Inst.	0.31 (0.09)
	+ Strategy	0.28 (0.09)
	+ Strat., Hdr.	0.33 (0.05)
Deepseek-R1-Distill-Llama-8B	Human Inst.	0.14 (0.15)
	+ Strategy	0.29 (0.05)
	+ Strat., Hdr.	0.30 (0.09)
Qwen2-7B	Human	0.30 (0.09)
	Strategy1	0.39 (0.09)
	+ Strat., Hdr.	0.30 (0.06)
Qwen3-8B	Human	0.47 (0.09)
	+ Strategy	0.53 (0.08)
	+ Strat., Hdr.	0.22 (0.08)
Healthy adults		0.77

Table 4: Frequencies of correct answers in WCST across reasoning models and their instruction-tuned counterparts. The reasoning Llama variant shows no improvement, while the reasoning Qwen variant shows some improvement, albeit still falling behind human accuracy. Results are shown across three prompting strategies: prompt based on human instructions, an augmented prompt that makes rules and strategy explicit (+ Strategy), and a further augmented prompt that clearly organizes information under headers (+ Strat, Hdr). Human norms naturally vary slightly across studies; reference values shown are from Barceló et al. (1997).

or when simultaneous processing is required (i.e., operation span or reading span).

Our results further suggest that attentional control – or selecting relevant information to focus on while ignoring distractors – is another vulnerability for LLMs. Notably, reasoning models were able to successfully mitigate this in the flanker task.

Difficulty with *updating* task-relevant information may also interfere with LLM ability to adapt to the WCST. While humans can infer a simple sorting rule with relative ease based on feedback from an examiner, no model approached human-like performance on this task. If LLMs have difficulty updating the state of relevant information, then they will not be able to successfully update their beliefs after receiving feedback, leading to low accuracy scores like the ones we observe.

While reasoning models can possibly supplement these difficulties by, e.g., generating thought strings that repeat information in a strategic, goal-oriented way to aid retrieval, we do not find consistent evidence of such behavior. Instead, the reasoning models tested could augment performance on simpler, one-turn tasks (e.g., Flanker, backward digit span) and offered little help on the most complex, longest task (WCST). Finally, we note that these tasks are relatively fast and simple for humans to complete, meaning little explicit reasoning should be required. However, we find that models generate thousands of tokens to solve these straightforward tasks.

It is possible that reasoning models are more suited to certain cognitive functions, explaining the asymmetric success we observe. However, it should be noted that reasoning models are trained on thought strings for vastly different types of problem solving, and their learned thought heuristics (e.g., "Wait," or "let me double check") do not appear to serve them well on these simple cognitive tasks. With more cognitively-focused training and more efficient thought strings, reasoning models could effectively augment executive functioning in LLMs, unlocking more intelligent models.

7 Conclusion

We apply a suite of working memory tasks to LLMs, finding that they consistently exceed human performance. In contrast, we find that the LLMs tested do poorly on tasks of executive functioning. Manipulation and updating of information are identified as key cognitive difficulties for LLMs, and some difficulties with attentional control are also observed. While reasoning models show some signs of compensating for these cognitive deficits, they are highly inefficient and consistently produce sub-optimal reasoning strings.

Limitations

We report only behavioral results that cannot be used to interpret what a model encodes in its internal representations. We believe these behavioral results are revealing, but as they only capture the generation process, they do not shed light on the models' internal computations or knowledge.

We focus on a collection of six instruction-tuned and three reasoning models; it is possible that other language models will not follow the trends found here, although we have made efforts to collect a reasonable sample of models.

We make efforts to test a range of prompts and pilot test prompting strategies, but we are unable to definitively determine that our prompts are representing the full range of possible model responses to these tasks.

Ethics Statement

In trying to address the working memory limitations discussed in this paper, it may be tempting to turn to scaling up LLMs, whether via additional attention heads, wider context windows, generating more reasoning tokens, ensembling multiple language models, or some combination thereof.

We instead advocate for research toward more efficient models. There are very pressing environmental concerns associated with generative language models, and further scaling reasoning tokens should not be done without regard to the sustainability of the models. This is particularly relevant as this paper addresses working memory, a key feature for models in industrial, user-facing applications.

Acknowledgments

This research was supported by the University of Minnesota Graduate School Doctoral Dissertation Fellowship. The authors would also like to thank Minnesota NLP and the University of Minnesota Text Group for their helpful comments and discussions.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Alan D Baddeley and Graham J Hitch. 1994. Developments in the concept of working memory. *Neuropsychology*, 8(4):485.

Joseph A Banken. 1985. Clinical utility of considering digits forward and digits backward as separate components of the wechsler adult intelligence scale-revised. *Journal of Clinical Psychology*, 41(5):686–691.

Francisco Barceló, Marta Sanz, Vicente Molina, and Francisco J Rubia. 1997. The wisconsin card sorting test and the assessment of frontal function: A validation study with event-related potentials. *Neuropsychologia*, 35(4):399–408.

James M Broadway and Randall W Engle. 2010. Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, 42(2):563–570.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2025. Do not think that much for 2+ 3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*.

Hyo Jung Choi, Dong Young Lee, Eun Hyun Seo, Min Kyung Jo, Bo Kyung Sohn, Young Min Choe, Min Soo Byun, Jee Wook Kim, Shin Gyeom Kim, Jong Choul Yoon, and 1 others. 2013. A normative study of the digit span in an educationally diverse elderly population. *Psychiatry investigation*, 11(1):39.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9076–9108. PMLR.

Andrew RA Conway, Brooke N Macnamara, and Pascale MJ Engel de Abreu. 2013. Working memory and intelligence: An overview. *Working Memory*, pages 13–35.

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

Kelly M Dann, Aaron Veldre, Stephanie Miles, Philip Sumner, Phillipa Hay, and Stephen Touyz. 2023. Measuring cognitive flexibility in anorexia nervosa: Wisconsin card sorting test versus cued task-switching. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 28(1):60.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Adele Diamond. 2013. Executive functions. *Annual review of psychology*, 64(1):135–168.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Barbara A Eriksen and Charles W Eriksen. 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149.

Adam Frost, Simar Moussaoui, Jagjot Kaur, Samreen Aziz, Keisuke Fukuda, and Matthias Niemeier. 2021. Is the n-back task a measure of unstructured working memory capacity? towards understanding its connection to other working memory tasks. *Acta psychologica*, 219:103398.

Patricia S Goldman-Rakic. 1992. Working memory and the mind. *Scientific American*, 267(3):110–117.

Patricia S Goldman-Rakic. 1995. Cellular basis of working memory. *Neuron*, 14(3):477–485.

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

David A Grant and Esta Berg. 1948. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of experimental psychology*, 38(4):404.

Dustin Hammers, Gabriela Ramirez, Carol Persad, Judith Heidebrink, Nancy Barbas, and Bruno Giordani. 2016. Diagnostic profiles of patients differentially failing executive functioning measures. *American Journal of Alzheimer's Disease & Other Dementias®*, 31(3):214–222.

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

Romuald A Janik. 2023. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839*.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. 2024. Repeat after me: transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, pages 21502–21521.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. 2024. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *Advances in Neural Information Processing Systems*, 37:67712–67757.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? *arXiv preprint arXiv:2406.04847*.

Tokio Kajitsuka and Issei Sato. 2025. On the optimal memorization capacity of transformers. In *The Thirteenth International Conference on Learning Representations*.

Michael J Kane and Randall W Engle. 2003. Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to stroop interference. *Journal of experimental psychology: General*, 132(1):47.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. 2023. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*.

Juhani Lehto. 1996. Are executive function tests dependent on working memory capacity? *The quarterly journal of experimental psychology section A*, 49(1):29–50.

Ji-An Li, Corey Zhou, Marcus Benna, and Marcelo G Mattar. 2024. Linking in-context learning in transformers to human episodic memory. *Advances in neural information processing systems*, 37:6180–6212.

Stephanie Miles, Caitlin A Howlett, Carolyn Berryman, Maja Nedeljkovic, G Lorimer Moseley, and Andrea Phillipou. 2021. Considerations for using the wisconsin card sorting test to assess cognitive flexibility. *Behavior research methods*, 53(5):2083–2091.

Brenda Milner. 1963. Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of neurology*, 9(1):90–100.

Akira Miyake, Naomi P Friedman, Michael J Emerson, Alexander H Witzki, Amy Howerter, and Tor D Wager. 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100.

Marco Monaco, Alberto Costa, Carlo Caltagirone, and Giovanni Augusto Carlesimo. 2013. Forward and backward span for verbal and visuo-spatial data: standardization and normative data from an italian adult population. *Neurological Sciences*, 34:749–754.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Nikki Pratt, Adrian Willoughby, and Diane Swick. 2011. Effects of working memory load on visual selective attention: behavioral and electrophysiological evidence. *Frontiers in human neuroscience*, 5:57.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. 2023. *Cognitive Effects in Large Language Models*. IOS Press.

José Humberto Silva-Filho, Sonia Regina Pasian, and Francisco de Assis Carvalho do Vale. 2007. Typical performance of elderly patients with alzheimer disease on the wisconsin card sorting test (wcst). *Dementia & Neuropsychologia*, 1(2):181–189.

Nash Unsworth, Keisuke Fukuda, Edward Awh, and Edward K Vogel. 2014. Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive psychology*, 71:1–26.

Bibek Upadhayay, Vahid Behzadan, and Amin Karbasi. 2025. Working memory attack on llms. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, Florida, USA. Association for Computational Linguistics.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. 2024. Rnns are not transformers (yet): The key bottleneck on in-context retrieval. In *The Thirteenth International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. 2018. A dataset and architecture for visual reasoning with a working memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 714–731.

Steven Yantis and James C Johnston. 1990. On the locus of visual selection: evidence from focused attention tasks. *Journal of experimental psychology: Human perception and performance*, 16(1):135.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. 2020. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2024. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16896–16922.

A Appendix

A.1 Overthinking in Models

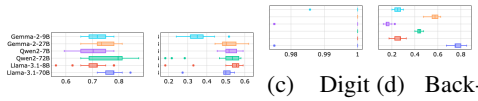
We observe several detrimental thought patterns in reasoning models. Below are selected examples from Qwen3-8B.

One is incoherent reiterations of the situation, e.g., "Now, the third letter is being presented. But the user hasn't given it yet," or "Some words are repetitive and nonsensical, but these are the ones that can be clearly recalled," from the reading and operation span tasks.

Sometimes we see muddled logic, e.g.: "The last few interactions included words like 'problem,' 'speed,' 'ice,' and the current one. So, the words to recall would be 'ice' and any others mentioned in the current round. But since the user hasn't provided more than one word in this round, the answer should be "ice.""

Thought strings will sometimes veer from the task entirely, circling without making progress, e.g.: "I think I'm getting the hang of it.</br> So,</br> I need to make sure that I don't make mistakes.</br> But I'm still learning.</br> I think I need to be careful.</br>"

We also note that learned reasoning heuristics ("Wait" or "let me double check") are not very useful in these tasks, further hindering reasoning models. For instance, consider Qwen3-8B's : "Alright, let's break it down. When I receive a new letter, I need to look back three positions to see if



(a) Flanker (b) WCST (c) Digit Span (d) Backward DS

Figure 10: For each prompt variation, we compute each model's average accuracy on several tasks. These box plots display the range in resulting model accuracies over the prompts. For challenging tasks like Flanker and WCST, the range accuracies are comparatively large.

it matches. If it does, I write 'm'; if not, I write '-'. Simple enough, but I have to be careful not to make a mistake with the positions. Maybe I should write them down as I go along to keep track.

Wait, what if the sequence is longer? I mean, how do I handle it if the sequence is, say, 10 letters long? I need to make sure I don't miss any or miscount the positions. Maybe I can keep a running tally or use some kind of notation to mark the positions as I go. That way, when I get to the fourth letter, I can check the first one easily.

Wait, but what if the sequence has letters that repeat or have patterns?"

"Lastly, I should remind myself to stick strictly to the rules. Only 'm' and '-' are allowed, and no explanations. So, I have to make sure my responses are clear and follow the instructions precisely. That way, I can ensure my task is done correctly and efficiently."

A.2 Prompt Variations

We assess the extent to which the prompt variations affect the spread of model responses and find that outside of very easy tasks, the variations capture a range of model accuracies (Fig. 10).

A.3 Estimation of human baselines

It is difficult to precisely estimate human baselines due to the sheer number of studies applying these experimental tasks across a variety of human populations.

We estimate human baseline accuracy for the backward digit span of length 7 by referring to the means and standard deviations presented in (Choi et al., 2013). Ideally we would use the Weschler norms, but these are proprietary and we do not have access. Given the backward digit span norms $\mu = 5.4, \sigma = 1.5$, we estimate that about 14% of people could have a backward digit span of 7. For the forward digit span of length 7, multiple sources report means of 7 (Banken, 1985; Monaco et al., 2013), so we estimate the accuracy at 50%.

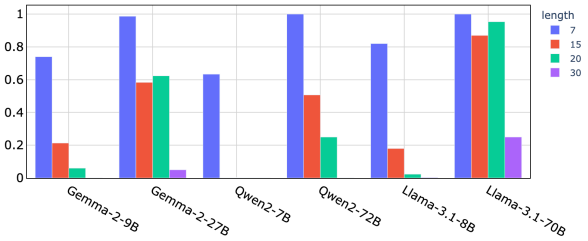


Figure 11: Mean model accuracy, by digit sequence length, in the backward digit span task across all trials ($n = 2100$, 70 stimuli x 30 prompts).

We use the reported mean human WCST errors in (Barceló et al., 1997) to estimate human WCST accuracy. The reported mean number of errors is 58.9 across 252 trials, so we estimate the baseline accuracy to be about 77%.

To estimate n-back accuracy we use Experiment 2 in Jaeggi et al. (2010).

We use the reported incongruent error rate in the Flanker task from (Yantis and Johnston, 1990), 4.7%, to estimate the baseline human accuracy for incongruent stimuli to be about 95%.

We estimated reading span and operation span accuracy through reported means from Experiment 1 in (Broadway and Engle, 2010).

A.4 Additional Results

A.4.1 Forward Digit Span

The mean accuracies for forward digit span across all models can be found in Table 5 .

A.4.2 Backward Digit Span

The mean accuracies for forward digit span across all models can be found in Figure 11.

A.4.3 Wisconsin Card Sorting Task

The average error rates over the first ten turns after a rule is introduced are shown in Figure 12 for each model.

We also investigate model performance over all dialogue turns. Correlation results are in Table 6, and the remaining model plots are in Figure 13.

Finally, decreases in accuracy between first and second exposure to a sorting rule can be seen in Figure 14

B Licenses of Models

All licenses are fair use for this work.

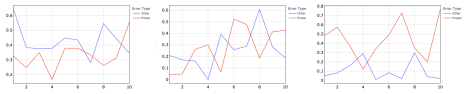
1. LLaMA-3.1 have a Community License that permits research.

Length	Gemma-2-27B	Gemma-2-9B	Llama-3.1-70B	Llama-3.1-8B	Qwen2-72B	Qwen2-7B
7	1.00	1.00	0.99	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	0.98
30	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	0.99	0.99	1.00	1.00	1.00

Table 5: Mean accuracies on forward digit span across all prompts.

	Gemma-27B	Gemma-9B	Llama-70B	Llama-8B	Qwen-72B	Qwen-7B
Pearson's r	-0.36*	-0.51*	-0.31*	-0.33*	-0.45*	-0.28*

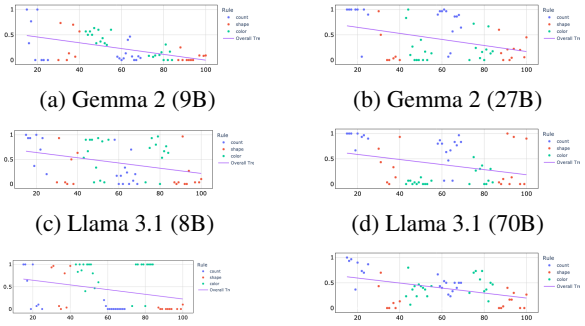
Table 6: Pearson correlation results for model accuracy and dialogue length in the WCST. Significant correlations ($p < 0.05$) are marked with *. We find all model accuracy decreases as the dialogue goes on.



(a) Gemma2 (9B) (b) Gemma2 (27B) (c) Llama3 (8B)

(d) Llama3 (70B) (e) Qwen2 (7B) (f) Qwen2 (72B)

Figure 12: We take the average error rate for preservation (red) and other (blue) errors across the first 10 rounds after a new rule is introduced. We find no correlation between any type of model error and the number of rounds exposed to a new rule.



(a) Gemma 2 (9B) (b) Gemma 2 (27B) (c) Llama 3.1 (8B) (d) Llama 3.1 (70B) (e) Qwen 2 (7B) (f) Qwen 2 (72B)

Figure 13: Trends for model accuracy over the course of the WCST dialogue.

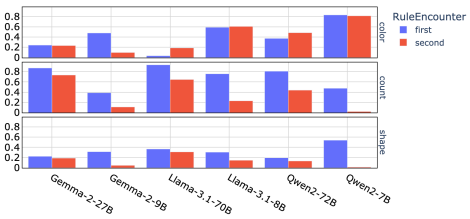


Figure 14: Models tend to have lower accuracy the second time they are exposed to the count and shape sorting rules (second and third rows).

2. Qwen models have an Apache 2.0 license.

3. Gemma has a custom Gemma Terms of Use license that permits research.

B.1 Prompt Specifications

Prompts from the reading span task:

> The following task consists of two parts. First, you will be presented with sentences, which you are supposed to read carefully. These sentences can make sense (e.g. More and more women want to have a career.) or not (e.g. I stopped by the gas station to refuel on apple juice.). Your first task is to judge the content of each sentence. You will then respond with your judgment (True = yes it makes sense, False = no it does not). After this decision you will see a word on the screen, which you are also supposed to read carefully. Your second task is to memorize this word in order to recall it later on. After two to six sentences (which you are supposed to judge) and words (which you are supposed to memorize), you will see three question marks (???). Whenever this happens, please respond by recalling all words you still remember since the last ???. Often it's easier to recall them in the order in which they were presented, but this isn't really necessary. If you only remember a couple, but not all words, it's okay to only name the ones that you do remember. After you have named all these words, the next round will begin. Your answers will all be recorded. We will now begin.

Prompts from the backward digit span:

> You will be presented with a list of digits. Repeat the digits back, in reverse order. Your response should contain *only* the reversed digits. Below are some examples of number sequences, followed by the reverse sequence:Digits: 1,2,3,4,5 5,4,3,2,1 Digits: 6,1,2,3,1,8,7 7,8,1,3,2,1,6 This concludes the examples. Below is a list of digits; respond by repeating these digits back, in reverse order. Your response should contain *only* the reversed digits.

Prompts from the flanker task:

> you'll be presented with 5 to 11 letters. Focus on the letter in the center. When encountering an X or a C, reply with Ä.Ïf a V or a B appears, reply with Ľ.Ÿour response should solely consist of Äör Ľ,ëxcluding any other letters. For example, if you see the letters BBCBBÿour response should be exactly Ä◦ If you see the letters XXXXCXXXXÿour response should be exactly Ä◦ Only respond with a single letter. This is the end of the instructions; the letters are below:

Prompts from WCST (human-like, base prompt):

> This is a card matching game. I'm going to show you some cards. Each card will have colored shapes on it.
>
> Your task is to match each card I show you to one of four key card options. I can't tell you how to match the cards, but I will tell you after each choice whether you are right or wrong. There is a correct way to match the cards, and you need to use my feedback to figure out what it is.
>
> The way the cards are supposed to be matched will change from time to time during the test, but I won't tell you when this happens. When it does change, you'll have to figure out the new way to match them, again using my feedback to

help you.

RESPONSE FORMAT:
For each card, respond with the sorting principle you're using and your choice number.

Format: [principle] - [option number]
Example:
Item: Two red triangles
Options: 1. One red circle, 2. Two green stars, 3. Three yellow crosses, 4. Four blue triangles
Your response: color – 1
(matching by color: red)

OR
Your response: count – 2
(matching by count: two items)

OR
Your response: shape – 4
(matching by shape: triangles)

Prompts from WCST (with additional explicit rules and strategy):

You will observe an object and need to categorize it into one of four options. Your goal is to determine the rule that classifies the object into one of these options.

Reply with *only the option number* where you believe the object belongs. Once you submit your answer, you'll receive feedback on whether your classification was correct. If it was correct, continue the same classification strategy; if it was incorrect, try a different rule until you find the correct rule.

At some point(s), without warning, the correct classification strategy will change. When this happens, forget the old classification strategy and once again, use the feedback to deduce the new classification strategy.

Below is an example in which the underlying rule is color:

EXAMPLE START
Item: two red triangles
Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
Answer: 1
Feedback: Correct
EXAMPLE END

EXAMPLE START

Below is an example in which the underlying rule is shape:
EXAMPLE START
Item: two red triangles
Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
Answer: 4
Feedback: Correct
EXAMPLE END

Below is an example in which the underlying rule is count:
EXAMPLE START
Item: two red triangles
Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
Answer: 2
Feedback: Correct
EXAMPLE END

To review, over the course of this conversation, there is a changing hidden classification rule. Your objective is to deduce the current rule, using the previous answers and feedback. Now let's start.

Prompt from the WCST (with additional strategy and organized headers):

Your goal: Sort a presented item into one of four available options using the correct sorting criterion.

Available Criteria:
- COLOR (match by color)
- SHAPE (match by shape)
- COUNT (match by quantity)

Important Details:
- Exactly one criterion is correct at any given time
- The correct criterion may shift unexpectedly during the task
- You'll receive feedback (correct/incorrect) after each attempt

How to Proceed:
When you have previous feedback:
- Correct feedback? Stick with that criterion
- Incorrect feedback? Choose a different criterion than the one you just used

When starting fresh or after a potential rule shift:
- Select any criterion to test

Key Point:
Prioritize your most recent attempt and its feedback. Earlier rounds may reflect a different criterion that's no longer active.

Answer Format:
Provide only the criterion and option number, with no additional text. Format examples:
- 'color - 2'
- 'shape - 4'
- 'count - 1'

Below, we use an example from the WCST to illustrate the prompt perturbation process.

Original Prompt Template - WCST

You will see an item, and you will have to match it to one of four options. Your task is to figure out the classification rule to sort the item into one of the four options.
Respond with *only the option number* you would like to sort the card into. After you respond, you will get feedback about your response. You will have to try a different classification rule if the feedback says you were wrong. This is an example in which the underlying rule is color:
[EXAMPLE START]
Item: two red circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—
Feedback: Correct
[EXAMPLE END]
Here's an example in which the underlying classification rule is *count*:
[EXAMPLE START] Item: two yellow triangles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 4
—
Feedback: Correct [EXAMPLE END]
Here's an example in which the underlying classification rule is *shape*:
[EXAMPLE START]
Item: two green circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—
Feedback: Correct
[EXAMPLE END]
Now we will begin.
<<DATA>>

Paraphrased Prompt Template 1 - WCST

You will be shown an item, and your task is to match it with one of four options. Your objective is to determine the hidden classification rule that assigns the item into one of these four options. The classification rule may be shape, color, or count.
Reply with *only the option number* you believe the item should be matched with, based on the classification rule. After you reply, you will receive feedback regarding your choice. If the feedback says you were incorrect, you will need to attempt a different classification rule. Note that the rule may change at any point; keep using the feedback to figure out the current rule. Here's an example in which the underlying rule is color:
[EXAMPLE START]
Item: two red circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—
Feedback: Correct
[EXAMPLE END]
Here's an example in which the underlying classification rule is *count*:
[EXAMPLE START] Item: two yellow triangles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 4
—
Feedback: Correct [EXAMPLE END]
Here's an example in which the underlying classification rule is *shape*:
[EXAMPLE START]
Item: two green circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—

Feedback: Correct
[EXAMPLE END]
Let's get started.
<<DATA>>

Paraphrased Prompt Template 2 - WCST

You will observe an object and need to categorize it into one of four options. Your goal is to determine the rule that classifies the object into one of these options. Reply with *only the option number* where you believe the object belongs. Once you submit your answer, you'll receive feedback on whether your classification was correct. If it was incorrect, you'll need to revise your classification strategy. Below is an example in which the underlying rule is color:
[EXAMPLE START]
Item: two red circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—
Feedback: Correct
[EXAMPLE END]
Next is an example in which the underlying classification rule is *count*:
[EXAMPLE START]Item: two yellow triangles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 4
—
Feedback: Correct [EXAMPLE END]
Finally, an example in which the underlying classification rule is *shape*:
[EXAMPLE START]
Item: two green circles Options: 1. one red circle, 2. two blue crosses, 3. three yellow stars, 4. four green triangles
—
Answer: 1
—
Feedback: Correct
[EXAMPLE END]
Now let's start...
<<DATA>>

Data Format 1 (original) - WCST

Feedback: (|FEEDBACK TEXT|)

Item: (|CARD TO SORT|)

Options:
1. one red circle,
2. two green triangles,
3. three blue crosses,
4. four yellow stars

Data Format 2 (Field: {}\n Answer: {}) - WCST

Feedback: (|FEEDBACK TEXT|)

Item: (|CARD TO SORT|)

Options:
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 3 (Field: {} <sep> Answer: {}) - WCST

Feedback: (|FEEDBACK TEXT|) <sep>
Item: (|CARD TO SORT|) <sep>
Options:
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 4 (Field - {}. Answer - {}) - WCST

Feedback - (|FEEDBACK TEXT|).
Item - (|CARD TO SORT|).
Options -
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 5 (Field\t{}. Answer\t{}) - WCST

Feedback (|FEEDBACK TEXT|).
Item (|CARD TO SORT|).
Options
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 6 (FIELD- {} \n ANSWER- {}) - WCST

FEEDBACK- (|FEEDBACK TEXT|)
ITEM- (|CARD TO SORT|)
OPTIONS-
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 7 (field:: {} – answer:: {}) - WCST

feedback:: (|FEEDBACK TEXT|) –
item:: (|CARD TO SORT|) –
options::
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 8 (field - {} , answer - {}) - WCST

feedback - (|FEEDBACK TEXT|) ,
item - (|CARD TO SORT|) ,
options -
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 9 (Field \n\t{} \n Answer \n\t{}) - WCST

Feedback
(|FEEDBACK TEXT|)
Item
(|CARD TO SORT|)
Options
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars

Data Format 10 (Field - {}\n Answer - {}) - WCST

Feedback - (|FEEDBACK TEXT|)
Item - (|CARD TO SORT|)
Options -
1. one red circle, 2. two green triangles, 3. three blue crosses, 4. four yellow stars