UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Backward Digit Span Benchmarks Working Memory in LLMs

Permalink
https://escholarship.org/uc/item/28m3k4g4

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Diak, Christopher James
Nguyen, Khac Nhat Nam
Tibbetts, JR
et al.

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Backward Digit Span Benchmarks Working Memory in LLMs

Christopher J. Diak[1], Nam Nguyen[1], JR Tibbetts[1], Taylor W. Webb[2] & Steven M. Frankland[1]
[1]Dartmouth College
[2]Universite de Montreal

Abstract

The maintenance and manipulation of information in working memory (WM) is fundamental to human and artificial intelligence. Although human WM is famously capacity-limited, Large Language Models (LLMs) preserve inputs in the network's context window, profoundly reducing capacity limits that depend on maintenance alone. However, work in cognitive science suggests WM limits can also arise from representational interference during manipulation, raising the possibility of strong limits even in LLMs with perfect maintenance. Here, we evaluate 15 frontier LLMs and find models perform near-perfectly on forward digit span (recalling sequences in order) but collapse on backward digit span (recalling sequences in reverse). Moreover, backward span performance is significantly correlated with two measures of fluid intelligence (Raven's Progressive Matrices and ARC-AGI-1). These findings suggest backward digit span provides a plausible benchmark for the "working" component of working memory in LLMs and point to potential shared principles of information processing across systems.

Keywords: Working memory; capacity limits; large language models; digit span; inference-time computation

Introduction

There is considerable interest in understanding large language models (LLMs) in cognitive scientific terms. One construct central to this pursuit—and uniquely challenging to assess—is working memory (WM): the temporary maintenance and manipulation of information. Humans and LLMs appear to face fundamentally different working memory constraints: LLMs can maintain lengthy sequences in context windows where task-relevant information remains continuously available. Tasks that primarily probe short-term storage should, in principle, be trivial for these systems. What could be further from the demands of a system that can maintain only 7 ± 2 items (Miller, 1956)?

But consider the simple problem of reversing a familiar sequence of digits: for example, many people who have long known that Jenny's phone number is 867-5309 would still struggle to mentally reverse it (9,0,3,5,7,6,8). For humans, this manipulation reveals the difficulty of transforming and re-indexing multiple mental representations independently of access to information about the sequence. Notably, substantial work in cognitive science suggests that limits on human working memory may arise from *representational interference* rather than informational decay (Oberauer et al., 2016), raising the possibility that LLMs exhibit capacity limits driven by the *working* component of working memory: manipulating representations.

Recent studies evaluating LLMs on *n*-back and related tasks report capacity limits resembling those in humans (Gong et al., 2024; Zhang et al., 2024). However, these results may simply reflect complex and variable task instructions (Hu & Lewis, 2025), motivating the need for a simpler benchmark. We thus turn to digit span, pioneered in early cognitive psychology and widely used in clinical neuropsychology (Terman et al., 1915; Wechsler, 2008). In the forward digit span task, participants recall a sequence in order; in the backward digit span (BDS) task, they reproduce it in reverse. Although capacity is low in both tasks—typical adults can reverse sequences of 5 or 6 items—performance is associated with fluid intelligence and executive function (Wilde & Strauss, 2002; Woods et al., 2011). The critical distinction is that forward span primarily probes access to stored information, whereas backward span requires actively reordering that information.

Earlier work identified a "Reversal Curse" in LLMs, in which models trained on relations in one direction fail to generalize to the reverse direction (Berglund et al., 2024). However, this phenomenon arises when relations between two entities are acquired during training and stored *in-weights*. BDS, by constrast, operates in an *in-context* regime. The full sequence is explicitly available, isolating the model's ability to execute a reversal transformation at inference time. We measure how that ability degrades as sequence length increases. This makes BDS a useful testbed for LLMs. It allows one to ask: *what capacity limits exist when information is available but must be manipulated?* Our contributions are:

1. We introduce backward digit span as a benchmark for inference-time sequence manipulation in LLMs.
2. We evaluate 15 OpenAI models from 2 to 500 digits, finding near-perfect forward span but length-dependent degradation on backward span.
3. We show reasoning-optimized models achieve higher backward span but exhibit similar degradation profiles.
4. We show backward span is not explained by context-window size: GPT-4.1 models (1.05M tokens) are outperformed by GPT-5 models (400K tokens).
5. We find a significant positive relationship between backward span and performance on text-based RPM and ARC-AGI-1, mirroring a classic finding relating WM differences to differences in fluid intelligence.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Methods

Task Design

Forward and Backward Digit Span. We adapted the digit span paradigm from clinical neuropsychology for evaluation of LLMs. In the forward span condition, models were presented with a sequence of N digits (0–9) and instructed to reproduce them in the same order. In the backward span condition, models were presented with an identical sequence format but instructed to reproduce the digits in reverse order.

We used a fully asynchronous batching pipeline with bounded concurrency. For each trial, we generated a single random digit sequence and ran *both* forward and backward span on the *same* sequence (thereby holding sequence content constant across directions within a trial). Sequences were generated by sampling digits uniformly at random with replacement. We tested sequence lengths ranging from $N = 2$ to $N = 500$, with at least 10 trials per condition per model at each sequence length.

Forward and backward span prompts were as follows:

```
Reprint the following sequence of digits:
{sequence}
Your answer must ONLY contain the printed digits,
comma-separated.
```

```
Reverse the following sequence of digits:
{sequence}
Your answer must ONLY contain the reversed digits,
comma-separated.
```

These prompts were designed to differ by the smallest possible change required to specify task direction: *reprint* versus *reverse*, and *printed* versus *reversed*. Holding all other prompt features constant (format, digit delimiters, and output constraints) helps ensure that performance differences are attributable to the reversal operation rather than to superficial differences in instructions.

Raven's Progressive Matrices. To assess whether BDS capacity is associated with broader reasoning performance, we evaluated the same models on a text-based version of Raven's Progressive Matrices (RPM), a classic nonverbal measure of abstract reasoning and fluid intelligence (Raven et al., 1998). In humans, performance on RPM is strongly associated with working memory capacity, reflecting shared demands on the ability to maintain and manipulate multiple mental representations (Engle, 2002; Wiley et al., 2011). Here, the text-based RPM serves as an LLM-adapted benchmark of structured pattern completion, allowing us to test whether backward span covaries with performance on a distinct but related reasoning task.

We follow Webb et al. (2023) in using text-based RPMs to evaluate LLM performance. This adaptation does not reproduce the visuospatial format of standard human RPM, but preserves the underlying relational rule structure by presenting already-parsed, symbolic inputs. As a result, it isolates rule induction over structured representations rather than perceptual processing, while remaining comparable to human performance at the level of problem structure. Cells are rendered using square brackets, with digits separated by spaces. The final cell is presented as an open bracket, which indicates where the answer should be completed. Each problem was constructed to follow rules defining transformations that map the values in related cells to one another. The following rules were included: *Constant*, *Distribution-of-Three*, *Progression*, *Multi-rule Compositions*, *Set union identities*, and *Boolean operators*. See Webb et al. for more detail.

For chat-based models, we used the following system instruction:

```
You are a Raven Progressive Matrices solver.
Follow the pattern and complete the final cell.
Output ONLY the digits and a closing bracket.
Do not explain.
```

The user message then contained the following instructions and a prompt:

```
"Complete the following matrix pattern.
The last cell is started for you with '['.
Finish it:"
```

Example matrix prompt structure:

```
[1 2] [2 3] [3 4]
[2 3] [3 4] [4 5]
[3 4] [4 5] [
```

Legacy instruction models received an equivalent prompt via completion-style API calls.

Models

OpenAI models (15 total): gpt-3.5-turbo, gpt-4-turbo, gpt-4o, gpt-4o-mini, gpt-4.1, gpt-4.1-mini, gpt-4.1-nano, gpt-5, gpt-5-mini, gpt-5-nano, gpt-5.1, and four pre-GPT-5 reasoning models: o1, o3, o3-mini, and o4-mini. Models were queried using the Chat Completions API in asynchronous batches between November 2025 and February 2026.[1]

Evaluation Methods & Metrics

Backward Span. Responses were scored as correct only if all digits were reproduced in the exact required order. We recorded:

- Span_{90}: Largest tested sequence length N with mean backward accuracy $\geq 90\%$ (i.e., $\geq$9/10 trials correct).
- Threshold_{50}: Largest tested sequence length N with mean backward span accuracy $\geq 50\%$. Here, we follow classic 75% detection threshold performance in two-alternative forced choice tasks, using 50% as the midpoint between 0% and 100%.

The batching run executed 10 trials per set size per model across 73 set sizes and 15 models. API errors were retained in the logs.

[1] For downstream analyses, we classify o-series models as reasoning models when computing grouped medians, reflecting their design emphasis on extended inference-time computation. GPT-5 and GPT-5.1 are excluded from these analyses due to architectural differences that make them not directly comparable within this grouping.

Raven's Progressive Matrices. Model outputs were parsed to extract digit sequences enclosed in square brackets. To handle reasoning models that produce intermediate text, the parser selected the last valid bracketed digit sequence in the output. A fallback parser extracted all numeric characters that come before a closing bracket if no complete bracketed sequence was found. Model outputs were scored as correct if and only if the predicted digit sequence exactly matched the ground-truth sequence for numeric problems. Model outputs for logic problems were scored as correct if and only if the predicted and ground-truth sets had equal cardinality and identical elements.

We evaluated each model on 30 problems per rule type. The maximum token output length was 15 for non-reasoning models and up to 25,000 tokens for reasoning models. For reasoning-capable models, the reasoning effort was set to "medium" to reflect default API settings used in the digit span tasks. We averaged peformance across all problem types and instances to obtain a mean measure of RPM performance for each model. We then evaluated the relationship between mean RPM performance and Threshold_{50} for backward span using ordinary least squares linear regression. We report the Pearson correlation coefficient r and analytic p-value.

ARC-AGI-1. We also assessed whether an LLM's backward span predicted it's performance on ARC-AGI-1, an external benchmark requiring few-shot induction of abstract transformations over grid-structured inputs (Chollet, 2019). Each model's performance on the ARC-AGI-1 testset is publicly available, requiring no further task design or implementation. To assess the relationship between backward span and ARC-AGI-1, we regressed Threshold_{50} for each model against that model's public ARC-AGI-1 score from the ARC Prize leaderboard, reporting Pearson's r and the analytic p-value.

Results

Performance on Forward and Backward Span is Asymmetric

Across all 15 OpenAI models, mean accuracy on forward digit span was 99.6%, confirming that simple sequence reproduction from the context window is trivial for frontier LLMs. Backward span, however, revealed stark capacity limits: Span_{90} ranged from 9 (gpt-4.1-nano) to 180 (gpt-5), while Threshold_{50} ranged from 14 (gpt-4.1-nano) to 380 (gpt-5).

Figure 1 illustrates the forward-backward asymmetry across two model families. All six models achieve near-perfect forward span, yet backward span reveals a clear hierarchy: GPT-5 models maintain Span_{90} through 180, 110, and 70 digits (full/mini/nano), while GPT-4.1 models collapse at 50, 28, and 9 digits respectively. Within each family, capacity scales with model tier; across families, GPT-5 shows more than 3× the backward span capacity of GPT-4.1 at matched scale tiers. However, backward span does not track context-window size: GPT-5 models have 400K-token context windows yet substantially outperform matched GPT-4.1 tiers, whose context windows are 1.05M tokens.

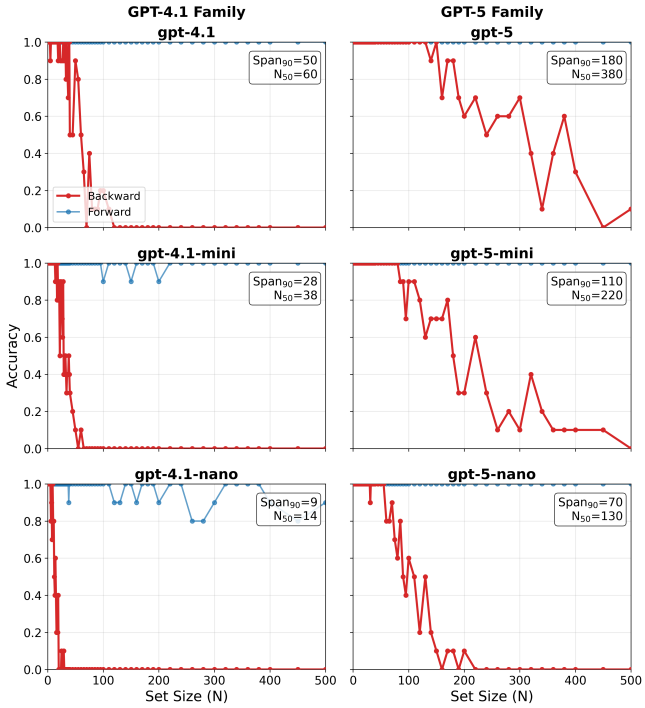

Figure 1: Forward and backward digit span accuracy for GPT-4.1 and GPT-5 model families. Forward performance is near-ceiling across models. Backward performance degrades with sequence length and varies by family and scale tier. Notably, GPT-5 models have smaller context windows (400K tokens) than GPT-4.1 models (1.05M tokens) but substantially higher backward span thresholds.

Reasoning Models Outperform. Reasoning-optimized models (o1, o3, o3-mini, o4-mini) showed substantially higher backward span capacity than standard models. Median Span_{90} was 120 for reasoning models versus 26 for standard models; median Threshold_{50} was 230 versus 38. However, this advantage does not reflect a qualitative difference in capability: reasoning models still exhibit the same characteristic collapse curve, merely shifted to larger sequence lengths, paralleling a key property of human WM: set size limits. Despite billions of parameters and context windows in the hundreds of thousands, the number of digits models can reliably *reverse* ranges from Span_{90} of 9–60 for non-chain-of-thought models and up to 180 for chain-of-thought models—a small fraction of their parametric complexity.

Table 1 reports BDS summary metrics for each model. Threshold_{50} values span more than an order of magnitude, from $N = 14$ (gpt-4.1-nano) to $N = 380$ (gpt-5). Notably, this variation is unrelated to context window size: despite a 64-fold range in context windows (16K to 1.05M tokens), no significant relationship emerged between context capacity and backward span (Spearman $\rho = -0.26$, $p = .359$). The GPT-4.1 family, with the largest context windows (1.05M tokens), performed comparably to models with 200K contexts.

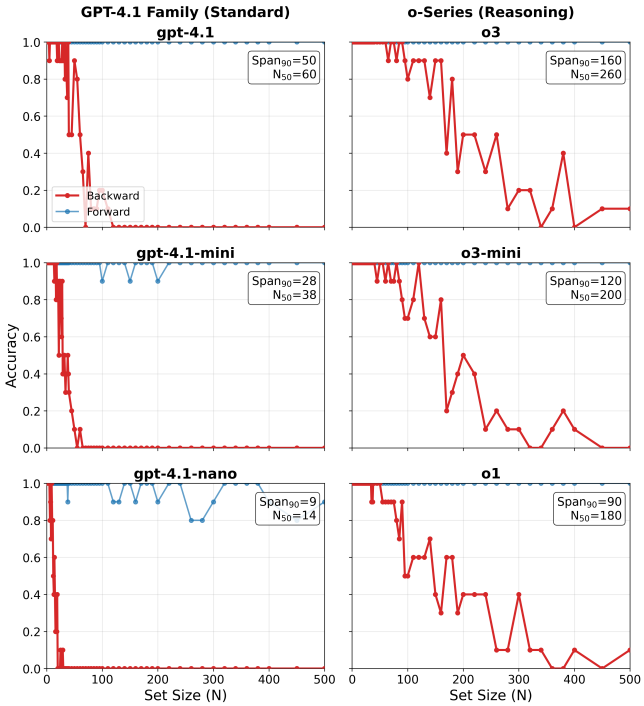

Figure 2: Forward and backward digit span accuracy for GPT-4.1 models and o-series reasoning models. Forward performance is near-ceiling, whereas backward performance shows graded degradation. O-series models shift the degradation curve to longer sequence lengths but do not eliminate length-dependent failure.

Table 1: Backward digit span summary metrics across models.

Model	Reasoning	Context Window	Span_{90}	Threshold_{50}
gpt-5	Yes	400K	180	380
o4-mini	Yes	200K	120	280
o3	Yes	200K	160	260
gpt-5-mini	Yes	400K	110	220
o3-mini	Yes	200K	120	200
o1	Yes	200K	90	180
gpt-5-nano	Yes	400K	70	130
gpt-5.1	Yes	400K	38	95
gpt-4o	No	400K	60	65
gpt-4.1	No	1.05M	50	60
gpt-4-turbo	No	400K	26	40
gpt-4.1-mini	No	1.05M	28	38
gpt-3.5-turbo	No	16K	15	24
gpt-4o-mini	No	200K	12	16
gpt-4.1-nano	No	1.05M	9	14

This dissociation reinforces that BDS taps a computational bottleneck distinct from raw sequence capacity.

Backward Digit Span Correlates with Measures of Fluid Intelligence in LLMs

In humans, WM capacity is strongly correlated with fluid intelligence, particularly as measured by Raven's Progressive Matrices (Engle, 2002). This relationship reflects shared demands on the ability to maintain and manipulate multiple mental representations simultaneously. If BDS captures a functionally analogous capacity limit in LLMs, we would expect it to predict performance on matrix reasoning tasks.

To test this prediction, we regressed each model's backward span threshold (Threshold_{50}) against its mean accuracy on Raven's Progressive Matrices. We evaluated 15 models on 30 problems per rule type, yielding a single mean RPM accuracy score per model. Backward span showed a significant positive correlation with matrix reasoning performance: Threshold_{50} vs. RPM accuracy: $r = 0.588$, $p = 0.021$, paralleling findings in human cognition that individual differences in WM capacity predict performance on RPM tasks (Harrison et al., 2015; Wiley et al., 2011).

Notably, this correlation emerges despite the very different surface characteristics of the two tasks: BDS requires only simple digit reversal with no semantic content, while Raven's matrices require abstract pattern recognition and rule induction. The shared variance is consistent with an account wherein both tasks invoke the capacity to maintain and manipulate multiple representations during inference. Future work aims to establish whether this association reflects a shared underlying computational mechanism, for example by testing whether common internal representations or attention patterns support performance across both tasks, or whether the relationship is instead explained by broader differences in model capability.

We observe even stronger results for ARC-AGI-1. As shown in Figure 3, backward span capacity (Threshold_{50}) correlates strongly with ARC-AGI-1 performance ($r = 0.73$, $p = 0.007$), a benchmark requiring few-shot learning of abstract geometric transformations. Related analyses using alternative BDS measures (area under the curve and capacity estimate K) yield similar positive relationships but are not shown.

Discussion

We employed backward digit span as a minimal benchmark for the working component of working memory in LLMs, focusing on inference-time manipulation under conditions where maintenance is effectively unconstrained. Our evaluation of 15 OpenAI models revealed three key findings. First, all models show near-perfect forward span but systematic degradation on backward span, indicating that strong capacity limits arise even when the relevant information is fully available. Second, capacity is improved by reasoning optimization but still degrades at predictable sequence lengths. Third, backward digit span capacity is positively associated with performance on text-based Raven's Progressive Matrices and ARC-AGI-1, paralleling the well-established relationship between WM and fluid intelligence in humans.

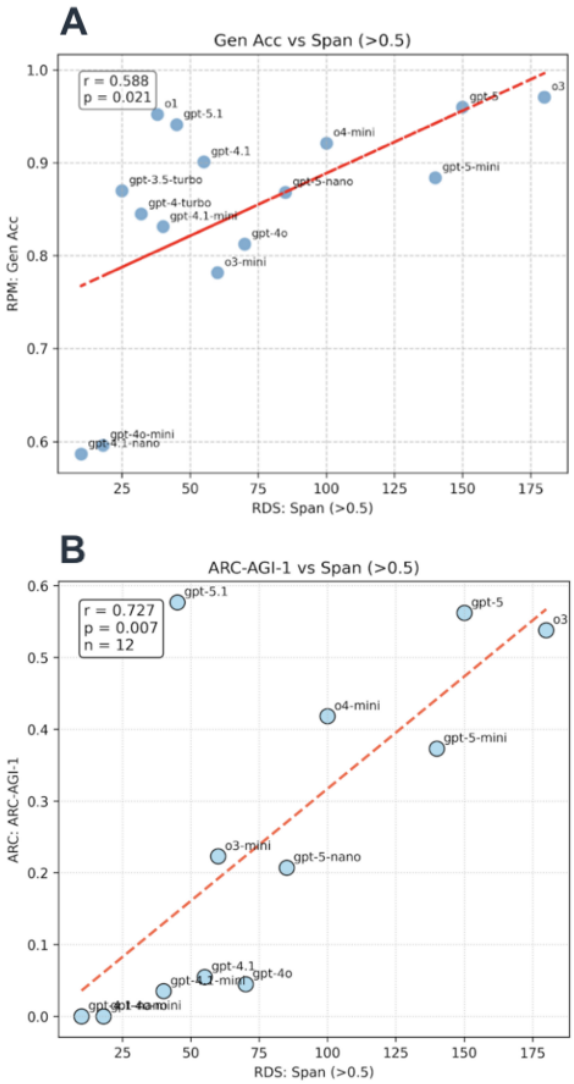

Figure 3: Relationship between backward digit span capacity and measures of fluid intellience. Scatterplot showing Threshold_{50} (the largest sequence length at which models maintain ≥50% accuracy) versus (A) Raven's Progressive Matrices and (B)ARC-AGI-1 benchmark score across OpenAI models. Backward span capacity shows significant positive correlation with both measures of abstract reasoning performance (RPM: $r = 0.588$, $p = 0.02$, ARC: $r = 0.73$, $p = 0.007$), paralleling the relationship between working memory and fluid intelligence in humans.

The Maintenance-Manipulation Dissociation

The strong asymmetry between forward and backward span—near-perfect forward performance versus Span_{90} values ranging from just 9 to 60 digits for standard models—confirms that LLMs face no meaningful limitation in *maintaining* information from the context window. The bottleneck lies in *manipulating* that information during inference. This dissociation is theoretically important because it locates capacity limits at the level of manipulation rather than maintenance, paralleling the distinction between short-term storage and central executive function in human WM models (Baddeley, 1992).

Chain-of-Thought May Change the Task

Models that use "chain-of-thought" (CoT) effectively have an external scratchpad that they can use to write down the partial products of computation. We therefore expect these models to fare better than non-reasoning (or "base") models. This would be analogous to allowing a human participant in a backward span task to use pen-and-paper to write down and cross off steps as they go. Indeed, CoT models clearly outperform non-reasoning models: median Span_{90} rises from 26 to 120, and median Threshold_{50} from 38 to 220. However, reasoning models are still subject to capacity limits orders of magnitude smaller than the size of the context window. They merely shift the degradation threshold to larger sequence lengths. Explicit reasoning traces extend effective capacity, but whether or not they fundamentally alter the underlying representational constraints remains unclear.

In experiments with open-source models with auditable chains-of-thought including DeepSeek-R1 and GPT-OSS (not reported), models often use indexing strategies in their chains-of-thought to unwind and reverse a sequence, effectively externalizing the computational demands into the context window and treating it as a forward span task.

Working Memory and Fluid Intelligence in LLMs

The correlation between BDS capacity and Raven's Progressive Matrices performance ($r = 0.588$, $p = 0.021$) parallels one of the most robust findings in human cognitive psychology: that working memory capacity predicts fluid intelligence (Engle, 2002). In humans, this relationship has been attributed to shared demands on executive attention—the ability to maintain goal-relevant information in the face of interference and to flexibly update mental representations (Wiley et al., 2011).

Our findings suggest that an analogous relationship exists in LLMs. Models with greater capacity to manipulate information during inference (as indexed by BDS) also show superior performance on abstract reasoning tasks. This convergence is notable because it emerges from architectural and training differences rather than the genetic and environmental determinants of individual differences in humans.

The correlation also provides construct validity for BDS as a measure of a general computational resource. If backward span merely reflected task-specific abilities (e.g., string manipulation), we would not expect it to predict performance on a reasoning task with entirely different surface characteristics. The observed correlation suggests that both tasks tap into a shared capacity for maintaining and manipulating multiple representations during inference.

Comparing LLM and Human Backward Span

Although this question is clearly of interest, the current results only bear on it indirectly. While the OpenAI models we study here appear to perform above human-levels, there are a number of reasons this is not a perfect comparison. First, most obviously, all models have perfect maintenance of the forward sequence. Although our results suggest that representational interference may be sufficient to strongly limit capacity, any additional deleterious effects of decay limiting human performance are lost. Second, the best performing models, those with CoT effectively have a "scratchpad" on which they can write the partial products of their computation. Humans are not generally afforded this luxury in tests of backward span.

A fairer comparison might involve human experiments in which the forward sequence has been memorized (e.g., 867-5309). If memorized, the sequence can be accessed easily, again placing the demands on manipulation. With memorized forward sequences, we would expect that human backward span is presumably considerably greater than 5 or 6, paralleling increases in forward span when there is chunked conceptual structure. We also note anecdotally that experiments with simpler models such as GPT-2 were unable to perform the task, while models outside the OpenAI family (Mistral, Qwen-2.5 20B, LLAMA 7B) exhibit even more stringent capacity-limits than those reported here. We are currently undertaking a more systematic evaluation of a broader class of models.

Candidate Sources of the Capacity Limit

Several mechanisms could contribute to the length-dependent limit observed here. One possibility is representational interference: as more digit-position bindings must be maintained and re-indexed, overlapping representations may interfere with one another (Frankland et al., 2021; Oberauer et al., 2016). A second possibility is that backward span reflects the training and architectural asymmetries of autoregressive LLMs. Forward reproduction aligns with next-token prediction and with common continuation patterns in text, whereas reversing a sequence requires using later input positions to determine earlier output positions. This connects backward span to prior work on autoregressive biases and reversal failures (Berglund et al., 2024; McCoy et al., 2024). However, backward digit span differs from the Reversal Curse in an important respect: the relevant sequence is available in-context, and models often succeed up to substantial lengths. The phenomenon studied here is therefore not a failure to infer a reversal from parameterized data, but a capacity limit on in-context sequence manipulation. More generally, although the autoregressive structure training structure might be necessary for such a difference, it fails to offer an explanation for why capacity is limited.

A system can avoid interference by learning task-specific representations that do not generalize (e.g., a mapping from *3,1,6* to *6,1,3*). However, we assume there is no pressure for either LLMs or humans to acquire dedicated representations for backward span: an LLM's basic objective is to continually predict the next token in a vast corpus of text, and humans did not evolve in a setting in which reversing explicit numerals had adaptive benefit. By contrast, for both systems, it does seem adaptive to acquire representations and mechanisms that generalize across many different contexts — e.g., representations of number identity (*3,5,2*), representations of position (*first, second, last*), and some way of binding the former to latter dynamically, given that they are subject to change from context to context. Whenever these representations must be redeployed without any specialized mechanisms for separating them, they are likely to induce interference and limit processing capacity (Frankland et al., 2021). For forward span, we expect these representations do not interfere because there is a general pressure on LLMs to use recent associations to predict future tokens. This underlies "in-context" learning, with a plausible circuit mechanism identified in "induction heads" (Olsson et al., 2022). By contrast, there is no pressure to use a subsequent token to retrieve a preceding token. This means that we should expect backward span to rely upon re-used representations of digits and position with no dedicated circuitry for performing the reversal task, putting generalization ability and processing capacity in tension. Evaluating whether this is, in fact, the source of the capacity-limits we observe here in LLMs remains an important area of future work.

Limitations and Future Directions

Several limitations should be noted. First, our evaluation focused on digit sequences; future work should examine whether capacity limits generalize to other token types (words, symbols, mixed sequences). Second, we did not systematically vary prompt structure or few-shot examples, which may influence effective capacity. Third, while the correlation between BDS and RPM performance is significant, the sample of 15 models limits statistical power; replication with a larger and more diverse set of models would strengthen these conclusions. Fourth, there is clearly a close relationship between "scaling" properties of the models and processing capacity. Future work should specifically target the specific internal representations that support success/failure in reversal that may co-vary with network size. The strong correlation between backward span and matrix reasoning performance suggests that similar mechanisms may underlie both capabilities – such as re-indexing a shared set of representational pieces.

Acknowledgments

LLMs assisted in the production of code for data generation, analysis, and visualization. All code was audited by the authors, who assume full responsibility for the veridicality of the results.

References

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

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23.

Frankland, S. M., Webb, T., Lewis, R., & Cohen, J. D. (2021). No coincidence, george: Processing limits in cognitive function reflect the curse of generalization.

Gong, D., Wan, X., & Wang, D. (2024). Working memory capacity of ChatGPT: An empirical study. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition*, *43*(3), 389–396.

Hu, X., & Lewis, R. (2025). Do language models understand the cognitive tasks given to them? investigations with the n-back paradigm. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: Acl 2025* (pp. 2665–2677). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.136

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? [Epub 2016 Mar 7]. *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L., Galbreath, N., & Talbert, W. (1915). The stanford revision of the binet-simon scale and some results from its application to 1000 non-selected children. *Journal of Educational Psychology*, *6*(9), 551.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wechsler, D. (2008). Wechsler adult intelligence scale–fourth edition (wais-iv)[database record]. *APA PsycTests*, *10*.

Wilde, N., & Strauss, E. (2002). Functional equivalence of WAIS-III/WMS-III digit and spatial span under forward and backward recall conditions. *The Clinical Neuropsychologist*, *16*(3), 322–330. https://doi.org/10.1076/clin.16.3.322.13858

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and raven's advanced progressive matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 256–263.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of clinical and experimental neuropsychology*, *33*(1), 101–111.

Zhang, C., Jian, Y., Ouyang, Z., & Vosoughi, S. (2024). Working memory identifies reasoning limits in language models. *arXiv*.