Mechanistic Interpretability of Large-Scale Counting in LLMs through a System-2 Strategy

Anonymous ACL submission

Abstract

Large language models (LLMs), despite strong performance on complex mathematical problems, exhibit systematic limitations in counting tasks. This issue arises from architectural limits of transformers, where counting is performed across layers, leading to degraded precision for larger counting problems due to depth constraints. To address this limitation, we propose a simple test-time strategy inspired by System-2 cognitive processes that decomposes large counting tasks into smaller, independent sub-problems that the model can reliably solve. We evaluate this approach using observational and causal mediation analyses to understand the underlying mechanism of this System-2-like strategy. Our mechanistic analysis identifies key components: latent counts are computed and stored in the final item representations of each part, transferred to intermediate steps via dedicated attention heads, and aggregated in the final stage to produce the total count. Experimental results demonstrate that this strategy enables LLMs to surpass architectural limitations and achieve high accuracy on large-scale counting tasks. This work provides mechanistic insight into System-2 counting in LLMs and presents a generalizable approach for improving and understanding their reasoning behavior.

1 Introduction

Counting is a fundamental cognitive operation that underpins a wide range of reasoning tasks, from basic arithmetic to more complex forms of quantitative analysis (Feigenson et al., 2004; Dehaene, 2011). In Large Language Models (LLMs), the ability to count is important for controlled generation such as length-constrained summarization, sequential enumeration, and broader numerical/arithmetic reasoning (Retkowski and Waibel, 2025; Hou et al., 2025; Yang et al., 2024b; Gambardella et al., 2024).

Recent research has provided insights into the counting mechanism of LLMs, demonstrating that

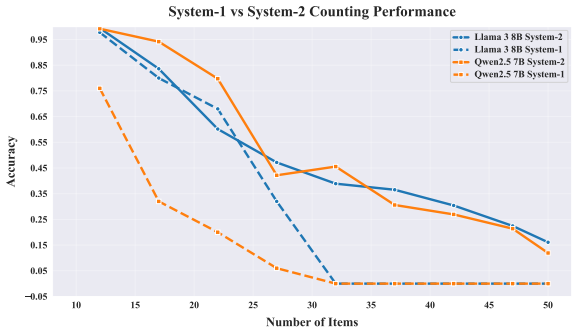


Figure 1: System-1 vs. System-2 counting performance as a function of problem size. System-1 performance degrades rapidly and collapses beyond approximately 30 items, reflecting the bounded capacity of the model's internal counter. In contrast, System-2 counting maintains high accuracy across the entire range by decomposing the task into small solvable sub-problems and aggregating the results.

these models use a layerwise internal counting process where numerical information is progressively accumulated across transformer layers (Hasani et al., 2025b). However, due to the depth limits, this progressive counting mechanism becomes saturated as the number of items increases. Complementary work on numerical representations suggests that LLMs encode numbers in a compressed, sublinear manner, similar to the human logarithmic mental number line (AlquBoj et al., 2025; Dehaene, 2011). While numerical order is preserved, representational resolution decreases with magnitude, explaining the reduced precision for large values.

These findings suggest that LLMs struggle to accurately count large numbers of items, with performance typically degrading for two- and three-digit counts (Zhang et al., 2024; Yehudai et al., 2024; Fu et al., 2024). This limitation reflects a fundamental architectural constraint rather than a lack of training data or supervision, as more layers are required to count a larger number of items (Vaswani et al., 2017; Yehudai et al., 2024; Golkar et al., 2024). We argue that this failure stems from the model's

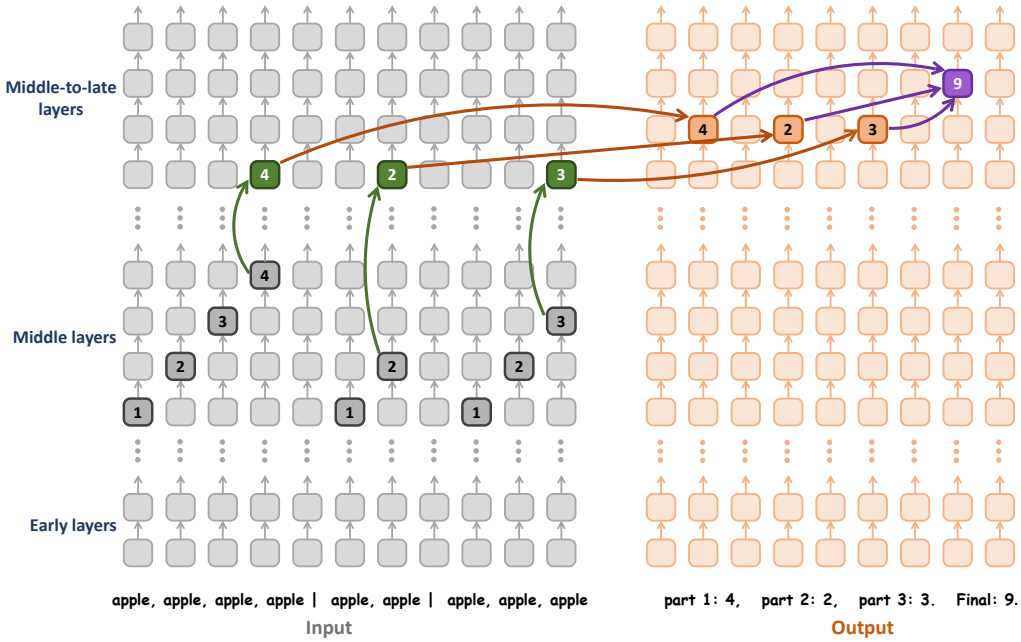


Figure 2: **Internal mechanism of System-2 test-time counting in LLMs.** A large counting task is divided into smaller partitions using an external separator (|). Within each partition, the model performs implicit System-1 counting, where count information accumulates token-by-token and is localized at the final item or separator token (gray blocks). The final count information is transferred via residual streams (green arrows) and stored in the middle-to-late layers (green blocks). These partition-level counts (e.g., 4, 2, 3) are then transferred (orange arrows) through attention pathways to explicit reasoning tokens that report intermediate results (orange blocks). Finally, the intermediate counts are aggregated (purple arrows) to produce the final answer. By keeping each sub-task within the model's reliable counting range, this System-2 procedure removes the upper bound imposed by the model's architectural limitations.

reliance on a System-1–like processing approach, which is fast, automatic, and capacity-limited (Kahneman, 2011; Dehaene, 2011).

To address this issue, we propose a simple test-time strategy that adopts a System-2–like approach to counting (Figure 1) (Kahneman, 2011). Instead of relying on the model's internal counting mechanisms, which are constrained by architectural limits, our approach decomposes large counting tasks into smaller, independent sub-tasks (Radhakrishnan et al., 2023; Qharabagh et al., 2024). Each sub-task, containing a manageable number of items, can be reliably counted by the model. The results from each sub-task are then aggregated to produce the final count. This approach mirrors human cognitive strategies, where complex problems are broken down into simpler, easier-to-solve sub-problems (Kahneman, 2011; Dehaene, 2011).

Our behavioral experiments on various LLMs demonstrate the effectiveness of this strategy in overcoming architectural limitations without requiring model modifications or fine-tuning. Additionally, we provide a detailed mechanistic interpretation of how it functions within LLMs. Using attention analysis and causal mediation techniques (Heimersheim and Nanda, 2024; Geiger et al., 2021; Ghandeharioun et al., 2024; Zhang and Nanda, 2024), we trace the flow of numerical information across the model and identify the mechanisms mediating the System-2 counting process. Figure 2 provides an overview of the main components of the System-2 counting mechanism and its internal information flow. This work offers a new perspective on both improving and understanding LLMs' reasoning capabilities, with a focus on a fundamental cognitive task.

2 Problem Setup and Methodology

We follow the standard counting framework used in prior research (Hasani et al., 2025b). In this setup, a list of repeated items, such as fruits or animals, is presented to the model, and the task is to report the total number of items. For example, given a context like "apple, apple, apple, . . . ", the model must output the total number of items in the list. Previous work has shown that LLMs exhibit high accuracy for small counts (fewer than 10 items), but that performance deteriorates as the number of items

Model	Input	Output	Accuracy				MAE			
			11-20	21-30	31-40	41-50	11-20	21-30	31-40	41-50
Qwen2.5 7B	Unstructured	w/o steps	0.38	0.13	0.06	0.00	0.88	2.19	5.29	10.50
		w/ steps	0.45	0.11	0.03	0.00	0.72	2.44	5.97	9.68
	Structured	w/o steps	0.20	0.13	0.05	0.01	3.56	3.54	4.33	6.35
		w/ steps	**0.95**	**0.61**	**0.38**	**0.24**	**0.07**	**1.36**	**1.53**	**2.18**
Llama 3 8B	Unstructured	w/o steps	0.80	0.49	0.02	0.00	**0.20**	**0.65**	4.93	11.92
		w/ steps	0.29	0.34	0.00	0.00	0.74	0.82	5.00	11.44
	Structured	w/o steps	0.08	0.05	0.03	0.01	6.62	5.75	5.96	6.62
		w/ steps	**0.84**	**0.54**	**0.38**	**0.26**	0.30	0.70	**1.21**	**2.20**
Gemma 3 27B	Unstructured	w/o steps	**1.00**	0.70	0.40	0.00	**0.00**	**0.30**	**1.00**	4.90
		w/ steps	**1.00**	0.50	0.30	0.00	**0.00**	0.60	1.40	4.70
	Structured	w/o steps	0.30	0.05	0.00	0.17	2.10	10.95	7.33	12.67
		w/ steps	**1.00**	**0.85**	**0.55**	**0.50**	**0.00**	0.35	2.15	**2.25**

Table 1: Average accuracy and mean absolute error (MAE) across **open-source models** for context sizes from 11 to 50. Each model is evaluated on structured and unstructured inputs, with and without intermediate reasoning steps.

Model	Input	Output	Accuracy					MAE				
			51-60	61-70	71-80	81-90	91-100	51-60	61-70	71-80	81-90	91-100
GPT-4o	Unstructured	w/o steps	0.70	0.54	0.56	0.18	0.24	0.36	1.42	0.72	3.62	4.26
		w/ steps	0.58	0.53	0.40	0.16	0.26	1.10	1.78	1.72	3.20	3.64
	Structured	w/o steps	0.37	0.31	0.10	0.11	0.11	1.04	1.53	3.22	2.64	3.03
		w/ steps	**0.96**	**0.91**	**0.87**	**0.83**	**0.86**	**0.04**	**0.10**	**0.16**	**0.22**	**0.18**
Gemini-2.5-Pro	Unstructured	w/o steps	0.52	0.50	0.44	0.42	0.20	0.60	0.50	3.17	4.67	2.70
		w/ steps	0.95	0.80	0.72	0.60	0.60	**0.05**	1.10	1.78	1.30	1.30
	Structured	w/o steps	0.82	0.80	0.78	0.75	0.79	0.25	0.08	0.18	0.30	0.10
		w/ steps	**0.97**	**0.95**	**0.95**	**0.91**	**0.91**	0.10	**0.05**	**0.05**	**0.06**	**0.07**

Table 2: Average accuracy and MAE across **closed-source models** for context sizes from 51 to 100. Each model is evaluated on structured and unstructured inputs, with and without intermediate reasoning steps.

increases beyond this range (Zhang et al., 2024; Fu et al., 2024). This suggests that the model's counting ability is limited by the depth of the transformer architecture and its internal counting mechanism (Yehudai et al., 2024).

To overcome this limitation, motivated by prior work on partitioning images in visual reasoning tasks (Izadi et al., 2025), we introduce a simple strategy that explicitly partitions the input list into smaller sub-problems. We use an external separator (|) to divide the list into smaller partitions. For instance, a structured context with three partitions is: "apple, apple, apple, apple, apple | apple, apple, apple | apple, apple, apple, apple, apple, apple".

The model is then instructed to first count the items in each partition and to aggregate the partial counts to produce the final result. This ensures that each sub-problem remains within the model's reliable counting regime. The number of items in each partition is chosen randomly from a range that the model can handle accurately. Models with deeper architectures can reliably process larger partitions.

Based on the input structure and output format, we consider four baselines in our study. Two input formats are used: an unstructured context with comma-separated items and a structured context with partitions separated by vertical bar symbol (|). For generation, we evaluate two output variants: a short-answer format, where only the final count is produced (corresponding to an immediate System-1-like process), and a Chain-of-Thought (CoT) format, where intermediate reasoning steps are included before the final answer (corresponding to a System-2-like process).

3 Behavioral Results

To illustrate how LLMs lose counting precision in long contexts, we first measure the average prediction probability of Qwen2.5 7B (Yang et al., 2024a) across different context sizes. As shown in Figure 3, model confidence decreases as the number of items exceeds 10. We also observe systematic biases toward more frequent numbers. For example, target counts of 16 and 21 are often predicted as 15 and 20.

We then conduct more systematic experiments on both open-source and closed-source models. For open-source models, we evaluate Qwen2.5 7B, Llama 3 8B (Grattafiori et al., 2024), and Gemma 3 27B (Team et al., 2025), which have 28, 32, and 62 layers, respectively. In addition, we evaluate GPT-4o (Achiam et al., 2023) and Gemini-2.5-Pro (Team, 2025) as stronger proprietary models on longer contexts. The corresponding results are

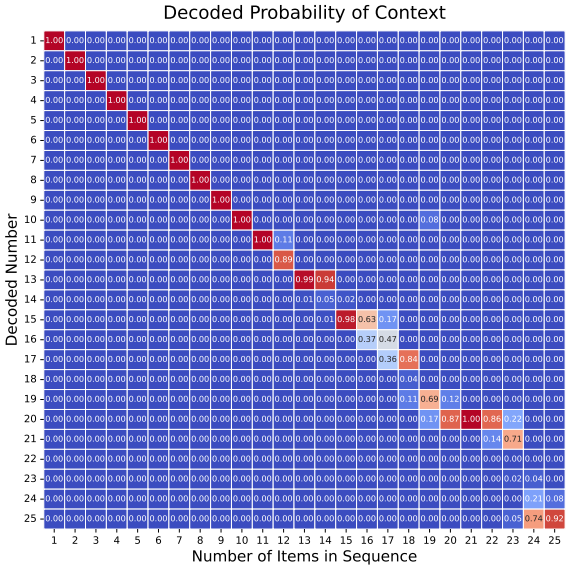


Figure 3: Decoded output probabilities for the unstructured baseline method on Qwen2.5 7B. The heatmap shows the decoded probabilities of model outputs, averaged over different item types, for target counts ranging from 1 to 25. As the count increases beyond 10, the diagonal entries gradually fade, indicating reduced model confidence.

reported in Tables 1 and 2. Across all models, we find that only one configuration consistently succeeds on long contexts: structured inputs combined with intermediate reasoning steps.

Surprisingly, encouraging CoT reasoning alone, without structured input, does not yield a notable improvement. Furthermore, structured input without intermediate steps is also ineffective and can be harmful in some cases. A likely explanation is that models must first consolidate partial results in intermediate steps before aggregating them into the final answer through a two-stage process. Our observational and causal analyses provide further insight into why models cannot simultaneously gather partial counts from the context and add them up without loss. Overall, these results show that neither external structure nor reasoning alone is sufficient. Their combination is necessary to overcome large-scale counting failures.

4 Attention Analysis

We analyze attention patterns of Qwen2.5 7B (Yang et al., 2024a) to understand how the proposed System-2 strategy enables large-scale counting. We first identify which tokens are attended to when the model generates (i) intermediate partition-level counts and (ii) the final aggregated answer. We then localize the layers and heads that contribute most to these processes.

Figure 4a shows attention patterns to the input items when the model generates intermediate reasoning steps. We consider a structured context with three partitions containing 3, 4, and 5 items. During the generation of a reasoning token such as "part 2: 4", the attention weight of the generated number peaks sharply on the final item and the final comma separator of the corresponding partition. This pattern is consistent across prompts and is most pronounced in middle to late layers, typically around layers 19 to 23. Earlier items within the same partition receive substantially lower attention.

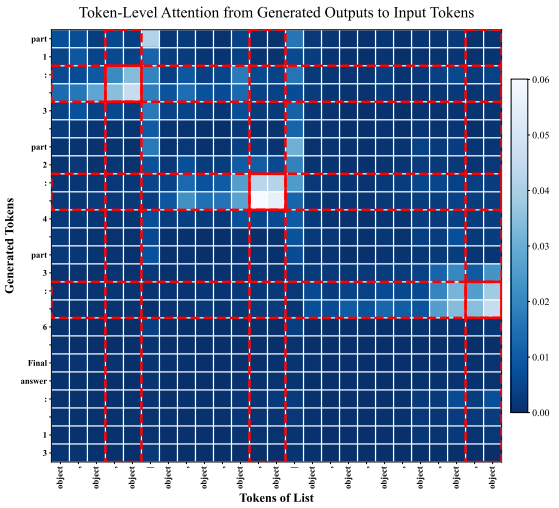
Figure 4b shows attention patterns when the model generates the final answer. The ":" token and the whitespace token immediately preceding the final number attend strongly to all intermediate reasoning numbers produced earlier (e.g., 3, 4, and 5). The attention mass is distributed across these tokens, with a clear concentration in the same layers identified for intermediate steps. Attention to the original input items is weak at this stage.

To further localize the attention pathways involved in intermediate reasoning and final aggregation, we focus on the most relevant tokens identified in Figure 4. We average attention values across different configurations, including item types and partition sizes. Figure 5 shows that attention peaks in layers 19 to 23, highlighting the central role of these layers in information transfer and aggregation. Figure 6 further shows that the most influential heads are concentrated in layers 21, 22, and 23. For example, head 13 in layer 22 consistently exhibits high attention to the selected tokens.

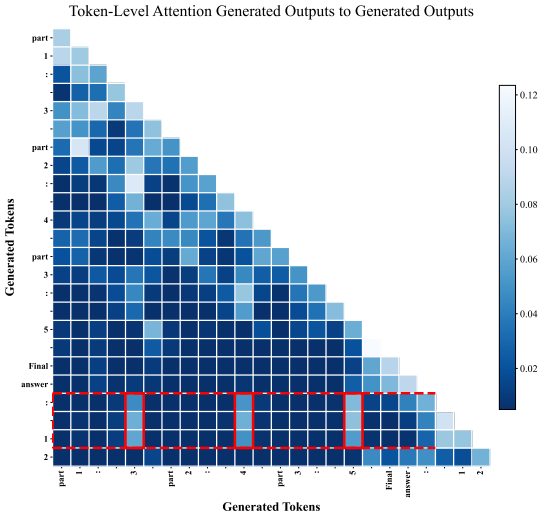
Together, these results suggest a staged computation. First, each partition is counted independently, with the final tokens of the partition encoding the local count. Second, these local counts are written into intermediate reasoning tokens. Finally, the model attends to these intermediate tokens and aggregates their values to produce the final answer. These observations are consistent with a hypothetical mechanism that separates counting, information transfer, and aggregation into distinct components.

5 Causal Mediation Analysis

To assess the hypothesis of a multi-stage System-2 counting mechanism suggested by the attention analysis, we perform a set of causal mediation experiments based on activation patching (Heimersheim and Nanda, 2024; Zhang and Nanda, 2024),



(a) Attention of generated tokens to input tokens.



(b) Attention of generated tokens to generated tokens.

Figure 4: Attention patterns of selected tokens under the System-2 counting strategy. Attention values are averaged across layers 19 to 23, all heads, and all item types (e.g., different fruit names).

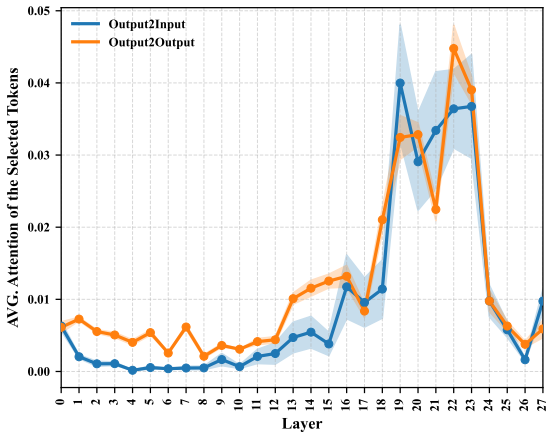


Figure 5: Average attention of correctly generated intermediate counts to the final item of their partition, alongside attention of the final answer to partition-level counts, across layers. Higher attention values are observed from layer 19 to 23 for both paths.

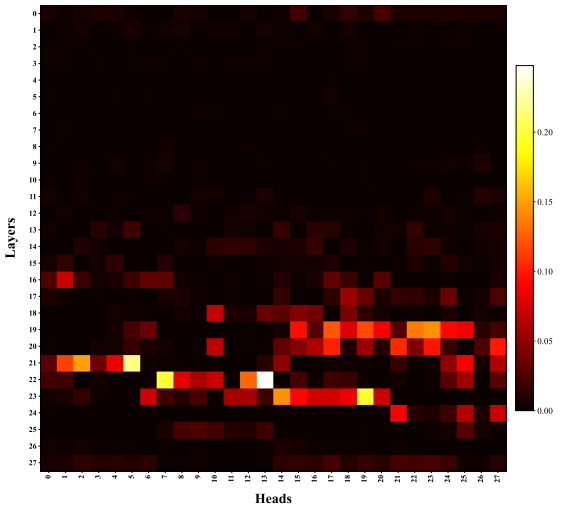


Figure 6: Heatmap of average attention to final items of partitions across layers and heads for intermediate counting steps. Layers and heads responsible for transferring numerical information from input partitions to intermediate counts appear with higher intensity.

masking ablation, and attention knockout (Geva et al., 2023). The goal is to identify where partition-level count information is stored, how it is transferred and consolidated, and which components are causally required for correct aggregation.

5.1 Token-Level Information Probing

We first examine where partition-level counts are encoded in the input context. This analysis requires vocabulary projection and probing tools. We observe that existing tools, such as logit-lens (NostAlgebraist, 2020) and tuned-lens (Belrose et al., 2023), are not reliable for decoding numerical information. We therefore use the CountScope method (Hasani et al., 2025b) to decode the implicit count associated with tokens. CountScope is a causal probing method that patches the activation of a target token into a blank counting context and decodes the model's output as the implied count.

Figure 7 shows the average decoding probability of ground-truth numbers across different token types. Ground-truth numbers correspond to the number of items in each partition. The results show that the ground-truth value is encoded with high confidence primarily at the final item and the final comma separator of each partition. For the first partition, the confidence at the final item is high; however, for subsequent partitions, the confidence

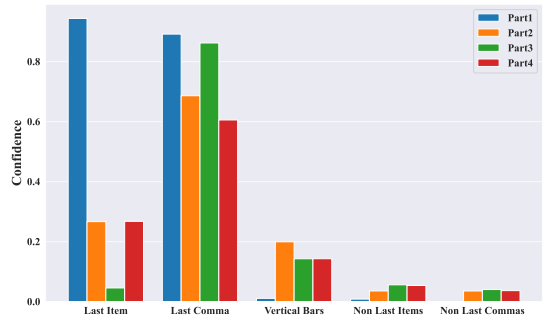


Figure 7: Ground-truth probabilities across different tokens of partitions decoded by CountScope. Probabilities are averaged over different item types and configurations (3 to 9 items per partition).

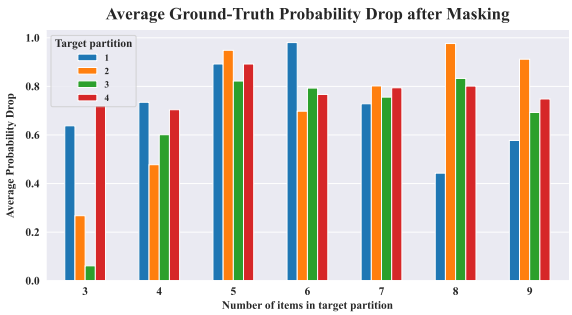


Figure 8: Average ground-truth probability drop after masking the final item and comma (e.g., ‘, apple’) from each partition, showing the effect on the target count for digit sizes ranging from 3 to 9.

at this token decreases, while the confidence at the vertical bar character increases. Nevertheless, the final comma separator reliably stores the latent count of the corresponding partition. These findings are consistent with the attention results in Section 4, which showed that decoded numbers of intermediate steps attend most strongly to these final partition tokens.

More interestingly, this experiment reveals that at each partition boundary, the count resets and begins again. As a result, the final item of each partition encodes the number of items counted since the beginning of that partition. This observation explains why the System-2 strategy avoids the large-number failure observed in unstructured inputs. Each partition is counted independently, and its size remains within the range where the model’s implicit counter is accurate.

5.2 Token-Level Causal Interventions

To test whether the identified tokens causally mediate the model’s behavior, we perform zero ablation on the final items and separators of each partition. Specifically, after processing the entire input sequence and extracting token embeddings, we replace the activation values of selected tokens with zeros across all layers.

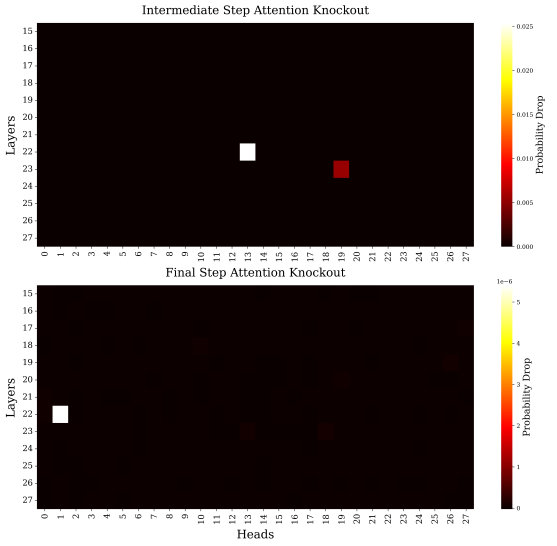


Figure 9: Heatmap of the average probability drop after attention knockout across layers and heads for intermediate (top) and final counting steps (bottom). Attention knockout is applied only to the identified effective tokens from the input context and intermediate steps.

Figure 8 shows the average probability drop after zero ablation of the final item and final comma separator. This intervention leads to a sharp drop in the probability of generating the correct partition-level count in the corresponding reasoning step. For example, ablating the final “, apple” token of a four-item partition significantly reduces the likelihood of generating the token “4”.

5.3 Information Pathway Localization

To perform fine-grained circuit localization, we analyze the pathways responsible for System-2 counting using attention knockout (Geva et al., 2023). We selectively block individual attention heads and layers and measure the resulting drop in counting accuracy for intermediate and final steps. Figure 9 shows that for Qwen2.5 7B, the most influential components are concentrated in layer 22. In particular, attention head 13 is most important for intermediate steps, while head 1 is most important for the final aggregation step.

Notably, head 13 in layer 22 also exhibits the highest attention values in the attention heatmap shown in Figure 6. Compared to Figure 6, the pattern in Figure 9 is considerably sparser. A plausible explanation is the presence of parallel information pathways, where knocking out a single pathway does not fully disrupt the computation because

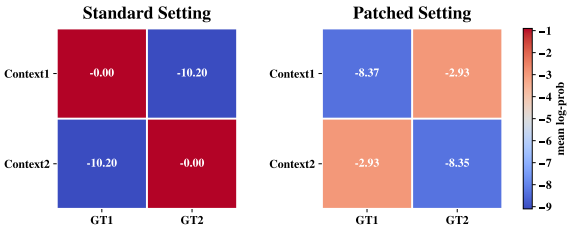


Figure 10: Average log probabilities of the first and second contexts before (left) and after (right) activation patching. The first and second columns correspond to the predicted outputs (total sums) of the first and second contexts. After swapping the layer embeddings of the selected tokens from a given partition between the two contexts, the output no longer follows the original total sum and instead reflects the transferred number. Values are averaged across different configurations (selected partitions, partition sizes, and item types).

other pathways can partially compensate. Finally, we observe that the heads responsible for transferring partition-level information from the input context to the intermediate steps are not necessarily the same as those responsible for transferring information from the intermediate reasoning steps to the final answer, even when they are located in the same layer.

5.4 Cross-Context Activation Patching

Finally, we perform cross-context activation patching to test how partition-level information is combined to form the final answer. We focus on intermediate responses and final aggregation steps. Our attention and masking analyses indicate that transferred numerical information from the context to intermediate steps is consolidated in the specific tokens ":", whitespace, and the partition number. Here, we investigate this behavior using a causal intervention setup (Geiger et al., 2021).

To this end, we sample two different contexts, each containing four partitions, with partition sizes randomly chosen between 3 and 9. After generating the intermediate steps, we select one partition and transfer the embeddings from layers 18 to 24 of the target tokens between the two responses. For example, let "part 1: 7, part 2: 4, part 3: 8" be the intermediate steps of the first context and "part 1: 5, part 2: 6, part 3: 3" be those of the second context. We swap the intermediate activations of the tokens highlighted in blue between the two responses. After this intervention, we allow the model to generate the final response.

This manipulation causally affects the corresponding intermediate steps and changes the final count accordingly. In the given example, the total sums of the first and second contexts are 19 and 14 before activation patching. After swapping the numerical contents of their second steps (shown in blue), the total for the first context becomes 21 and the total for the second context becomes 12. This shows that the final sum is causally mediated by the intermediate-layer embeddings of the tokens ":", whitespace, and the partition number. Figure 10 shows the log probabilities for this experiment, averaged over various configurations.

6 Related Work

Counting in LLMs. A growing body of work shows that counting in LLMs is brittle: performance can hinge on surface form and segmentation rather than a stable procedure (Fu et al., 2024). Subword tokenization can blur item boundaries and measurably affect counting accuracy (Zhang et al., 2024). More broadly, deterministic-task evaluations (including counting) can flip under minor prompt/content changes, complicating extrapolation from narrow setups (Ball et al., 2024). On the theory side, transformers can exactly count token frequencies only in specific regimes, with feasibility tied to representation capacity and context-length scaling (Yehudai et al., 2024).

Numerical representations and mechanistic views. Cognitive work argues for specialized number systems and structured magnitude representations (Feigenson et al., 2004; Dehaene, 2011), motivating analogous questions in neural models. Representation analyses study how numerical magnitude is organized in LLM hidden spaces and relate these structures to human-like effects (AlquBoj et al., 2025). Mechanistic studies localize how counting signals emerge and update across items in controlled settings (Golkar et al., 2024). Closest to our setting, Hasani et al. (2025b) combine a target-context probe with layerwise/token-level analyses and causal interventions to identify latent counter-like signals across both LLMs and LVLMs.

Decomposition and reasoning traces. Chain-of-Thought can improve multi-step reasoning by eliciting intermediate computations (Wei et al., 2022), but explicit traces are not necessarily faithful. Question decomposition work studies when decomposition increases faithfulness versus mainly serving as scaffolding (Radhakrishnan et al., 2023).

Counting in LVLMs. Counting remains challenging for LVLMs, particularly under clutter, layout variation, and compositional settings that stress binding between category and quantity (Guo et al., 2025; Campbell et al., 2024). Diagnostics and targeted modifications analyze where LVLM counting breaks and how performance can be recovered (Alghisi et al., 2025), while structured visual inputs or supervision can improve counting (Izadi et al., 2025; Qharabagh et al., 2024; Hasani et al., 2025a). Mechanistic evidence further suggests layerwise latent counter signals distributed across visual tokens (Hasani et al., 2025b).

Interpretability tools. Causal interventions such as activation patching help identify computation-critical states (Heimersheim and Nanda, 2024; Geiger et al., 2021), with best-practice guidance for reliable metrics and methods (Zhang and Nanda, 2024). Patch-based inspection frameworks further systematize intervention/inspection configurations (Ghandeharioun et al., 2024). Readout-style probes provide a complementary perspective by tracking when information becomes linearly decodable across layers, including the Logit Lens (NostAlgebraist, 2020) and learned variants such as the Tuned Lens (Belrose et al., 2023).

Our work. Our setup is closest to System-2-style prompting via intermediate steps: like CoT, we elicit explicit intermediate computations (Wei et al., 2022), and like task decomposition, we solve the instance by breaking it into sequential subproblems (Radhakrishnan et al., 2023). The key difference is that the decomposition is tied to a *structured input format*: we first partition the input into marked segments and then use a CoT protocol that is aligned with this structure (produce segment-level subcounts, then aggregate). This creates explicit stage boundaries that are absent in monolithic counting prompts and lets us study how counting representations evolve across stages. Mechanistically, our analysis aligns with prior work that localizes latent counter-like signals and tests them with interventions (Hasani et al., 2025b), while shifting the question to how the *structured CoT decomposition* reshapes internal computation on long contexts. In this way, we connect behavioral brittleness observations (Ball et al., 2024; Zhang et al., 2024) to a stage-wise mechanistic account of how count information is formed, routed, and combined under structured prompting (Heimersheim and Nanda, 2024; Zhang and Nanda, 2024).

7 Conclusion

This paper showed that the failure of large language models on large-scale counting tasks arises from reliance on System-1–like implicit counting mechanisms with limited capacity (Kahneman, 2011; Zhang et al., 2024). This problem reflects architectural constraints of transformer models rather than fundamental limits of numerical reasoning. We introduced a simple System-2 test-time strategy that decomposes large counting problems into smaller sub-tasks and aggregates their results, enabling accurate counting over long contexts without modifying model parameters or training procedures.

In addition to behavioral improvements, we provided a mechanistic explanation of how this strategy operates internally. Using attention analysis and causal mediation methods (Hasani et al., 2025b; Heimersheim and Nanda, 2024), we showed that partition-level counts are encoded at boundary tokens, transferred through specific attention pathways to intermediate reasoning steps, and aggregated in middle-to-late layers to form the final answer. Intermediate reasoning tokens play a crucial role in this process, mediating the flow of numerical information from input partitions to the final output.

In principle, this procedure removes fixed upper bound on countable size, as long as each sub-problem remains within the model's reliable regime. While our experiments focus on counting, the same analysis framework applies to other reasoning tasks where implicit representations saturate and explicit decomposition enables correct behavior. This study highlights how structured test-time strategies can reveal and extend the computational abilities of existing models, offering a path toward both improved performance and deeper interpretability. Future research could explore further applications of this strategy to complex reasoning tasks in language and beyond.

Limitations

This work focuses on a narrow counting task with limited object diversity, using simple repeated nouns in synthetic contexts. The approach depends on structured prompts and assumes prior knowledge of the model's reliable counting range, which may limit generalization. In addition, the effects of tokenization and number representation are not isolated and may independently influence the results.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, and et al. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Simone Alghisi, Gabriel Roccabruna, Massimo Rizzoli, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2025. [del re]constructing vlms' reasoning in counting. *Preprint*, arXiv:2510.19555.

H. V. AlquBoj, Hilal AlQuabeh, Velibor Bojkovic, Tatsuya Hiraoka, Ahmed Oumar El-Shangiti, Munachiso Nwadike, and Kentaro Inui. 2025. Number representations in llms: A computational parallel to human perception. In *International Conference on Learning Representations (ICLR)*.

Thomas Ball, Shuo Chen, and Cormac Herley. 2024. Can we count on llms? the fixed-effect fallacy and claims of gpt-4 capabilities. *arXiv preprint arXiv:2409.07638*.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. 2024. Understanding the limits of vision language models through the lens of the binding problem. *arXiv preprint arXiv:2411.00238*.

Stanislas Dehaene. 2011. *The Number Sense: How the Mind Creates Mathematics*, 2nd edition. Oxford University Press, New York, NY.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314.

Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. 2024. Why do large language models (llms) struggle to count letters? *arXiv preprint arXiv:2412.18626*.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Bangkok, Thailand. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning (ICML)*.

Siavash Golkar, Alberto Bietti, Mariel Pettee, Michael Eickenberg, Miles Cranmer, Keiya Hirashima, Geraud Krawezik, Nicholas Lourie, Michael McCabe, Rudy Morel, Ruben Ohana, Liam Holden Parker, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, and Shirley Ho. 2024. Contextual counting: A mechanistic study of transformers on a quantitative task. *arXiv preprint arXiv:2406.02585*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xuyang Guo, Zekai Huang, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2025. Your vision-language model can't even count to 20: Exposing the failures of vlms in compositional counting. *Preprint*, arXiv:2510.04401.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. 2025a. Uncovering grounding ids: How external cues shape multimodal binding. *Preprint*, arXiv:2509.24072.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. 2025b. Understanding counting mechanisms in large language and vision-language models. *arXiv preprint arXiv:2511.17699*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255.

Kuinan Hou, Marco Zorzi, and Alberto Testolin. 2025. Sequential enumeration in large language models. *Preprint*, arXiv:2512.04727.

Amirmohammad Izadi, Mohammad Ali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. 2025. Visual structures helps visual reasoning: Addressing the binding problem in VLMs. *arXiv:2506.22146*. Accepted to NeurIPS (poster).

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY.

NostAlgebraist. 2020. Interpreting gpt: The logit lens. LessWrong post.

Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. 2024. LVLM-COUNT: Enhancing the counting ability of large vision-language models. *arXiv preprint arXiv:2412.00686*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilē Lukošiūtē, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, and 5 others. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Fabian Retkowski and Alexander Waibel. 2025. Zero-shot strategies for length-controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 551–572, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and 1 others. 2024a. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2024b. Number cookbook: Number understanding of language models and how to improve it. *Preprint*, arXiv:2411.03766. ICLR 2025 poster.

Gilad Yehudai, Haim Kaplan, Asma Ghandeharioun, Mor Geva, and Amir Globerson. 2024. When can transformers count to n? *arXiv preprint arXiv:2407.15160*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*.

Xiang Zhang, Juntai Cao, and Chenyu You. 2024. Counting ability of large language models and impact of tokenization. *arXiv preprint arXiv:2410.19730*.

A Appendix

A.1 Details of the Behavioral Setup

We evaluate counting performance under controlled variations of input format and prompting strategy. Each example consists of a list of repeated single-word items, and the model is required to output the total count. Following the evaluation protocol summarized in Tables 1 and 2, we vary (i) whether the input is *structured* or *unstructured*, and (ii) whether the model is explicitly encouraged to produce intermediate reasoning steps (*with steps* vs. *without steps*). Performance is reported using exact-match accuracy and mean absolute error (MAE) across different context-length ranges.

Item types. All inputs are constructed from simple, common nouns drawn from two semantic categories: *fruits* and *animals*. The complete candidate set is shown below.

Items

```
apple, orange, peach, fig, mango,
pear, coconut, cherry, plum,
cat, dog, horse, rabbit, whale, cow,
frog
```

Input formats. In the **unstructured** setting, the input is a flat, comma-separated list of n identical items. In the **structured** setting, the same multiset is divided into multiple partitions separated by a vertical bar (|), enabling explicit local counting followed by summation.

Partition sizes. For **open-source models** (Table 1), each partition contains between **6 and 9 items**. For **closed-source models** (Table 2), which are evaluated on larger context lengths, partition sizes range from **15 to 25 items**. In all cases, partitions are constructed such that their sizes sum exactly to the target length n.

Prompting strategies. For each input format, we evaluate two prompting strategies: *without intermediate steps* and *with intermediate steps*. These settings correspond directly to the "w/o steps" and "w/ steps" rows reported in Tables 1 and 2.

Unstructured Input (w/o steps)

```
You will be given a list of items.
Count the total number of objects.
Output the final result exactly in the
following format:
Final answer: [x]
```

Unstructured Input (w/ steps)

```
You will be given a list of items.
Count the total number of objects.
Let's count step by step.
Output the final result exactly in the
following format:
Final answer: [x]
```

Structured Input (w/o steps)

```
You will be given multiple partitions
of content separated by "|".
For each partition, count the number
of items it contains.
After counting all partitions,
compute the total by summing the
counts.
Output the final result exactly in
the following format:
Final answer: [x]
```

Structured Input (w/ steps)

```
You will be given multiple partitions
of content separated by "|".
For each partition:
- Count the number of items it
contains.
- Report the count separately using
the format:
part1: [x1]
part2: [x2]
...
After counting all partitions:
- Compute the total by summing all
individual counts.
- Output the final total exactly in
this format:
Final answer: [x]
```

A.2 Attention Analysis

This section presents detailed methodological information regarding the attention analyses conducted

in this study. We perform our analysis using various large language models, Qwen 2.5 7B (Yang et al., 2024a), Llama 3.2 8B (Grattafiori et al., 2024), and Gemma 3 4B(Team et al., 2025). These models differ in architectural depth and layer organization, which allows for a more comprehensive and comparative investigation of attention behaviors. The analysis is divided into two parts. In the first part, we describe the experimental setup in detail, including the selection of specific layers for analysis and the prompts used in each experiment. In the second part, we analyze the attention patterns of the generated outputs with respect to the full input sequences for both models.

A.2.1 Input Details

The exact prompt used in all experiments is shown below:

```
You will be given a list of items, where groups
    (partitions) are separated by the "|"
    character.

For each partition:
- Count the number of items it contains.
- Report the count separately using the format:
  part1: x1
  part2: x2
  part3: x3
  ...

After counting all partitions:
- Compute the total by summing all individual
    counts.
- Output the final total exactly in this format:
Final answer: x
Just answer in this format without any extra
    things and follow the instructions.

apple, apple, apple | apple, apple, apple,
    apple | apple, apple, apple, apple, apple
```

To evaluate robustness across varying input structures, the final list provided to the model is randomized across experiments. Additionally, to reduce potential token-specific biases, we vary the item labels used in the lists, drawing from the Appendix A.1

A.2.2 Full Attention Analysis

This section presents a full attention analysis (following the methodology in Fig. 4) for Qwen2.5 7B, Gemma 3 4B, and Llama 3.2 8B. Our goal is to explore a little more of the layers in which attention most strongly concentrates on the tokens relevant to partition prediction, and to visualize the resulting full attention patterns.

As shown in Fig. 11, we analyze attention scores across different layer ranges for each model. For Gemma 3 4B, we examine layers 21 to 23, while for Llama 3.2 8B we consider layers 13 to 18. These layer intervals are selected based on the average attention magnitude from output tokens to key input tokens (specifically, the last item and the trailing comma of the predicted partition), which consistently peak within these ranges.

Figures 12, 13, and 14 illustrate the full attention maps for Qwen2.5 7B, Llama 3.2 8B, and Gemma 3 4B, respectively. In each figure, the x-axis corresponds to output tokens and the y-axis corresponds to input tokens. Across all models, we observe a consistent pattern: the token immediately preceding the generated number of parts exhibits the highest attention weights toward the last item and the final comma of the desired partition. This behavior supports our hypothesis that models rely heavily on attention to the final item–comma structure of each part when determining and generating the number of items for each part.

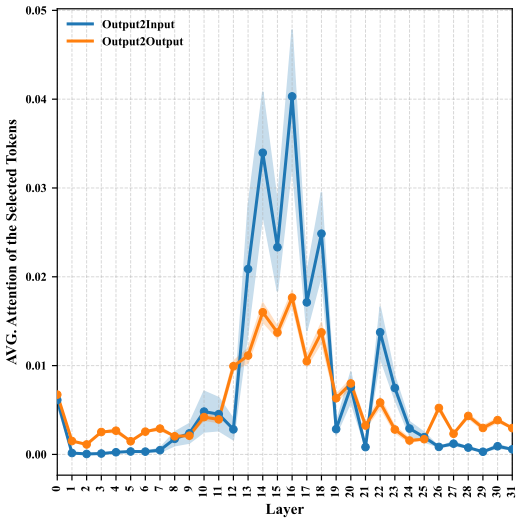
A.3 Causal Mediation Analysis

Token-Level Information Probing To decode the latent numbers of specific tokens, we employed CountScope. We modified the original source prompt to facilitate accurate counting by partitioning the input list using vertical bar delimiters (|). Additionally, we utilized a monotypic, question-first configuration. The exact prompt structure is provided below:

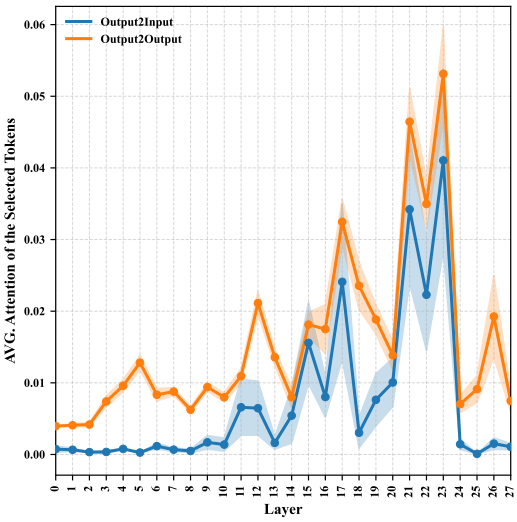
```
Answer the question with just a number only
    (We've separated each group of items with
    "|" so you can calculate the final count
    easier).
Question: How many fruits are there in the
    following sentence?
 apple, apple, apple | apple, apple, apple,
    apple, apple | apple, apple, apple, apple
```

To mitigate token-specific artifacts, we performed all experiments using a set of distinct items. (see Appendix A.1)

Information Pathway Localization To localize the specific attention heads and layers mediating the flow of count information, we performed attention knockout experiments as outlined in Section 5.3. This analysis utilizes a prompt that elicits specific System-2 behavior by explicitly separating partition counting from final aggregation. We employed a strict instructional format forcing the model to output intermediate counts before the final total. This allows us to measure the causal impact of blocking specific attention heads on the accuracy



(a) Average attention weights across layers for Llama 3.2 8B. We report attention from selected output tokens to both input and output tokens. Attention to the key partition-related tokens peaks consistently between layers 13 and 18, motivating our choice of this interval for full attention visualization.



(b) Average attention weights across layers for Gemma 3 4B. We report attention from selected output tokens to both input and output tokens (as defined in Fig. 4a). Attention magnitude peaks between layers 21 and 23.

Figure 11: Layer-wise attention analysis for Llama 3.2 8B and Gemma 3 4B. For each model, we visualize the average attention from selected output tokens to salient input tokens (notably the final item and comma of the partition). These trends are used to identify the layer ranges with the strongest partition-relevant attention.

of both intermediate transfer and final summation (the exact prompt template is provided below).

```
I will provide an input text containing fruits
    separated by the '|' delimiter. Your goal
    is to count the items in each section and
    provide a summation.

Input Text:
 apple, apple, apple | apple, apple, apple,
    apple, apple | apple, apple, apple, apple

Task:
1. Identify each partition separated by '|'.
2. Count the number of fruits in each specific
    partition.
3. Sum the counts of all partitions.

You must strictly follow this format, adapting
    the number of parts to the actual input:
part1: <count1>, part2: <count2>, ... , final
    count: <total_count>
```

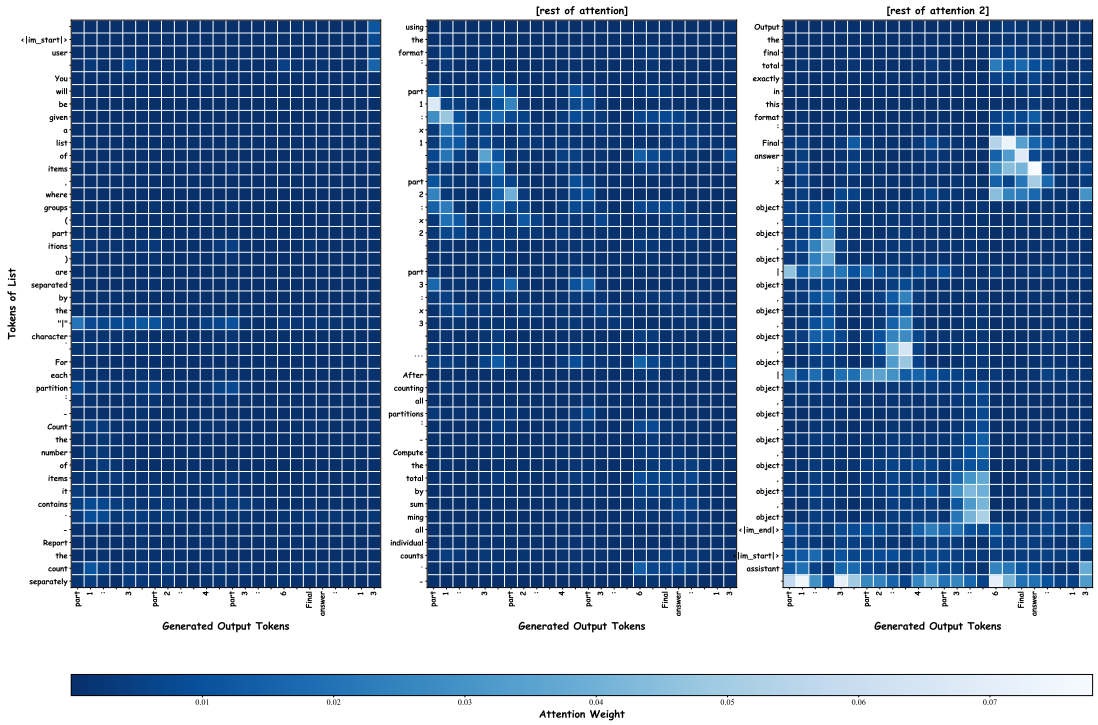


Figure 12: Full attention map for Qwen2.5-7B. The x-axis corresponds to output tokens and the y-axis to input tokens. The token immediately preceding the generated number of parts exhibits the strongest attention toward the final item and trailing comma of each partition matches its own segment.

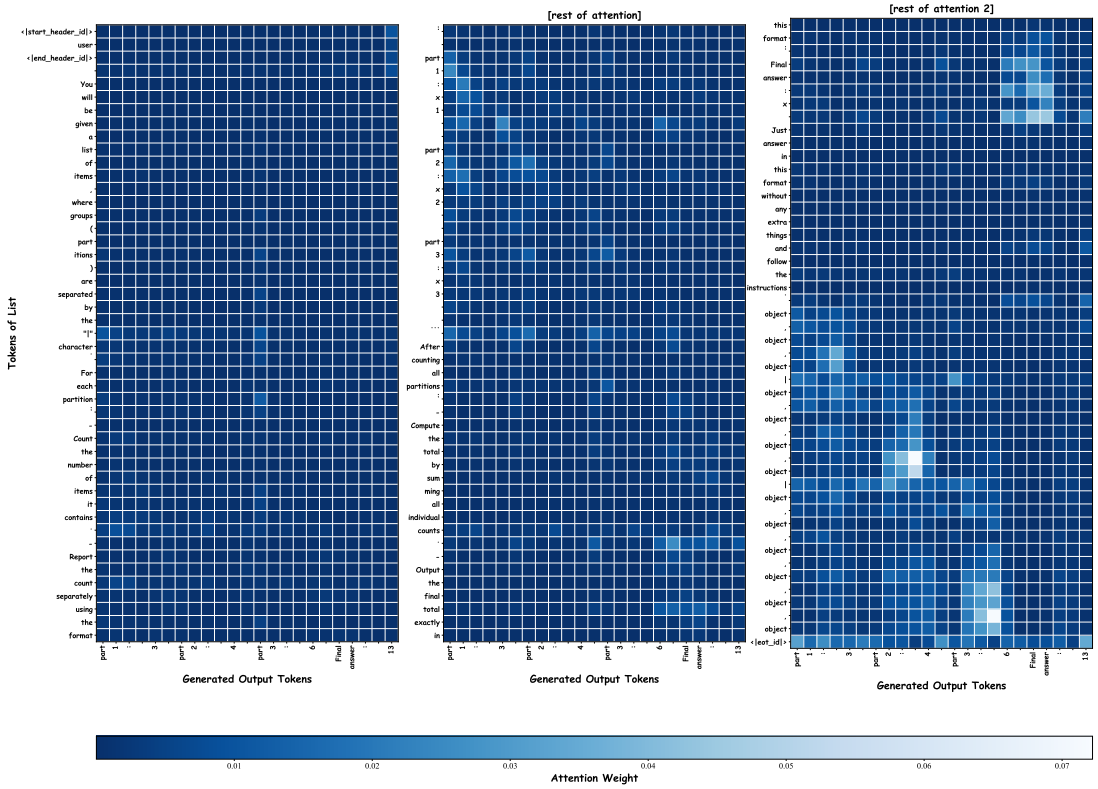


Figure 13: Full attention map for Llama 3.2 8B. The x-axis corresponds to output tokens and the y-axis to input tokens. The token immediately preceding the generated number of parts exhibits the strongest attention toward the final item and trailing comma of each partition matches its own segment.

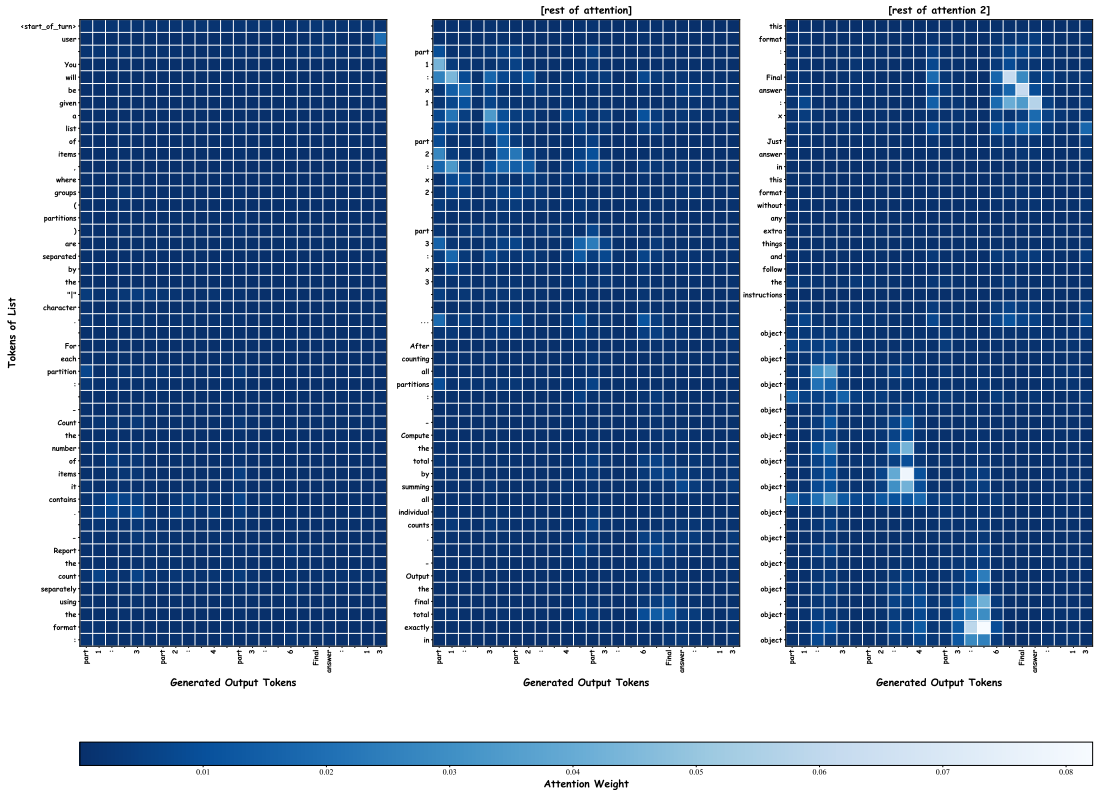
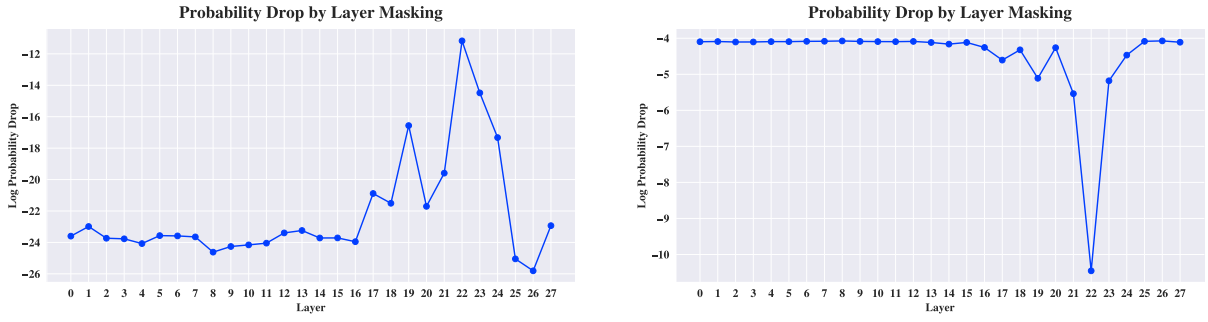


Figure 14: Full attention map for Gemma 3 4B. The x-axis corresponds to output tokens and the y-axis to input tokens. The token immediately preceding the generated partition-level count exhibits the strongest attention toward the final item and trailing comma of each partition matches its own segment.



(a) (b)

Figure 15: Average change in log probability of intermediate counts after layer masking (a) and unmasking (b) of the final item and separator of each partition. In the masking experiment, embeddings from the target layer are zero-ablated. In the unmasking experiment, embeddings of the selected tokens are zero-ablated across all layers, and only the target-layer embeddings are restored from the clean run.