

Residual Effects of Symbolic Binding

Anonymous Authors

Abstract—We examine how context-defined symbolic operations affect the internal dynamics of multi-step reasoning in autoregressive language models. In two-step calculations ($a \circ b = z; y = z + \delta$), an outer operation can be given directly (such as “add one”) or through an in-context alias (such as “red means add one”). We evaluate Qwen3.5-4B on a balanced 120-instance benchmark with matched prompt lengths and record activations across intermediate decoder blocks 0 to 30 using Jacobian and logit lenses. In late decoder blocks 20 to 27, mean intermediate reciprocal rank rises from 0.5124 under direct instructions to 0.6804 under alias instructions, yielding a paired difference of 0.1680 (95% CI [0.156, 0.180]). This elevation persists when restricting evaluation to matched prompt pairs in which both conditions produce the correct final answer, with intermediate reciprocal rank increasing from 0.5082 under direct instructions to 0.6548 under aliases (paired difference 0.1466, 95% CI [0.134, 0.159]). The effect depends on prompt syntax, with sequential phrasing yielding a contrast of 0.2150 and subordinate-clause framing 0.1210. Behaviorally, accuracy drops from 86.7% under direct instructions to 75.4% under aliases, and 57.6% of alias errors emit the intermediate value z instead of the final answer y . These findings show that context-defined symbolic operations correlate with delayed attenuation of intermediate representations across model depth, providing a mechanistic interpretability baseline for studying symbolic reasoning and motivating persistent intermediate states as potential signals of reasoning failure.

Index Terms—neuro-symbolic reasoning, explainable LLMs, test-time reasoning, Jacobian lens, representation dynamics, safeguarded AI

I. Introduction

Consider two arithmetic prompts supplied to a language model: “Compute $(2 + 3)$, then add one” and “Compute $(2 + 3)$, then use red,” where the prompt defines red as adding one. Both prompts specify intermediate value $z = 5$ and final target $y = 6$. The second prompt, however, requires the model to bind an in-context symbol to an arithmetic operation before applying that operation. We investigate whether substituting a defined alias for a standard operation alters the internal decodability of the intermediate value. When an operation requires symbolic resolution, intermediate representations may remain decodable longer across model layers. Unlike models prompted to emit explicit intermediate scratchpad or chain-of-thought tokens [1], [2], autoregressive decoders prompted directly for the final digit must perform intermediate computation entirely within latent activations.

Vocabulary projection techniques such as the logit lens [3], [4] and the Jacobian lens [5] identify the layers where intermediate tokens become decodable. Because an elevated readout can arise either from functional computation or from unsuppressed residual activations,

representational accessibility must be measured in conjunction with behavioral accuracy.

We evaluate intermediate readouts in Qwen3.5-4B [6] across direct and alias instructions using an extended 120-instance benchmark. The experiment incorporates five experimental controls: matched token lengths between conditions, between-instance counterbalancing of symbol assignments, exclusion of intermediate digits from prompt text, balanced syntactic paraphrases, and an intermediate-report control evaluated on successful trials.

Our measurements indicate that context-defined aliases correlate with prolonged decodability of intermediate variables across late layers. In decoder blocks 20 to 27, mean intermediate reciprocal rank is 0.5124 for direct instructions and 0.6804 for aliases; this corresponds to a paired contrast of 0.1680. This contrast remains present when isolating matched prompt pairs in which both conditions yield correct final outputs, where the paired difference is 0.1466 ($p < 0.001$). Syntactic framing modulates this effect: sequential prompts produce an increase of 0.2150, whereas subordinate clause prompts produce an increase of 0.1210. Behaviorally, accuracy decreases by 11.3 percentage points under aliases, from 86.7% to 75.4%.

II. Related Work

Linear Probes and Residual Stream Readouts. Projecting transformer residual activations into vocabulary space tracks internal representations across network depth [4]. Linear classifier probes [7] demonstrated that intermediate representations linearly encode task features. The logit lens maps hidden states directly to vocabulary logits [3], while the tuned lens [8] trains affine corrections. Gurnee et al. [5] introduced the Jacobian lens, which Yan et al. [9] formalized as a first-order approximation to downstream computation that decodes concepts earlier than unembedding alone. Beyond feed-forward key-value memory retrieval [10], residual streams preserve intermediate representations across multi-step prompts.

Mechanisms of Arithmetic Reasoning and Routing. Mechanisms underlying arithmetic have received substantial scrutiny. Modular addition circuits exhibit Fourier trigonometric representations [11], [12]. Causal mediation and activation patching reveal localized arithmetic circuits and routing shifts across prompt perturbations [13], [14], [15], [16]. In-context learning relies on induction heads for pattern copying and variable binding [17], while function vectors represent task mappings [18]. Tracing information flow further demonstrates that symbolic binding circuits route variables across layers [19]. However, transformers

TABLE I

Representative sample of 12 benchmark instances ($a \circ b = z$, outer modifier $\delta \in \{-1, +1\}$ producing target $y = z + \delta$).

Index	Inner Expression	Outer δ	z	y
1	2 + 1	+1	3	4
2	3 + 2	-1	5	4
3	4 - 1	+1	3	4
4	6 + 2	-1	8	7
5	12 - 7	+1	5	6
6	2 + 4	-1	6	5
7	9 - 2	-1	7	6
8	11 - 3	+1	8	9
9	3 - 2	+1	1	2
10	1 + 1	+1	2	3
11	4 + 5	-1	9	8
12	12 - 3	-1	9	8

face inherent computational bottlenecks in multi-step composition when intermediate states remain latent [20]. We examine how in-context symbolic bindings alter intermediate decodability across depth.

III. Experimental Methodology

A. Task Formulation and Arithmetic Benchmark

Every arithmetic item consists of two computational stages: an inner operation $z = a \circ b$ that produces intermediate digit $z \in \{0, \dots, 9\}$, and an outer operation $y = z + \delta$ that produces final digit $y \in \{0, \dots, 9\}$, with $\circ \in \{+, -\}$ and $\delta \in \{-1, +1\}$.

Operands $a, b \in \{1, \dots, 12\}$ are selected so that digit z never appears in the prompt string, preventing lexical copying. We construct a benchmark of 120 structured instances covering intermediate values $z \in \{1, \dots, 9\}$ through additive and subtractive pairs, divided into a Core Bank of 60 single-digit instances and an Extended Bank of 60 multi-digit instances (Table I). Across all primary and control conditions, the suite provides 840 evaluation prompts.

B. Prompt Conditions and Syntactic Templates

All prompts begin with a rule prefix defining the alias operations: “red means add one; blue means subtract one.” To prevent color bias, symbol assignments are counter-balanced between instances rather than within instances: 60 instances map red to +1 and blue to -1, while the other 60 instances invert this mapping (rule definition order within the prefix is also balanced). Each instance is evaluated under its assigned mapping rather than crossing both mappings per item.

Prompts use two syntactic frames. In Template A (Sequential Coordination), the prompt reads: “Compute $[a \circ b]$, then $[\text{action}]$. Reply with only the final digit.” In Template B (Subordinate Clause Framing), the prompt reads: “After evaluating $[a \circ b]$, $[\text{action}]$. Reply with only the final digit.” Within each frame, four paraphrases rotate across instances so that token counts match across conditions. The benchmark evaluates four conditions: (i) the Direct Condition states the operation explicitly (“add one”, 240 prompts across Templates A and B);

(ii) the Alias Condition substitutes the assigned symbol (“use red”, 240 prompts); (iii) the Intermediate-Report Control queries inner expression $a \circ b$ under the prefix (120 prompts); and (iv) the Synonymous Direct Control decouples symbolic binding from verbatim prefix repetition using non-overlapping phrasing (“increment by one”, 240 prompts across Templates A and B). Across all conditions, the suite comprises $240 + 240 + 120 + 240 = 840$ evaluation prompts.

C. Model Architecture and Readout Formulation

We evaluate instruction-tuned Qwen3.5-4B [6] (32 decoder blocks indexed 0 to 31, hidden dimension $d = 2560$) in bfloat16 precision with thinking mode disabled to isolate forward activations.

Internal activations are read out with the Neuronpedia Jacobian lens [21]. Because block 31 directly feeds final RMS normalization and unembedding (where logit lens equals output logits), checkpoints provide transport matrices for the 31 intermediate layers $\ell \in \{0, \dots, 30\}$. Let $h_{\ell,t} \in \mathbb{R}^d$ denote the activation vector following intermediate layer ℓ at token position t . For each candidate digit $d \in \{0, \dots, 9\}$, the Jacobian-lens score $s_{\ell,t}^J(d)$ and logit-lens score $s_{\ell,t}^L(d)$ are computed as:

$$s_{\ell,t}^J(d) = w_d^\top N(J_\ell h_{\ell,t}), \quad s_{\ell,t}^L(d) = w_d^\top N(h_{\ell,t}), \quad (1)$$

where $J_\ell \in \mathbb{R}^{d \times d}$ is the linear transport matrix for layer ℓ , $N(\cdot)$ denotes root-mean-square normalization, and w_d is the unembedding vector for digit d . In the Qwen tokenizer, each digit from 0 to 9 corresponds to a single token without leading space. Activations are extracted at the newline token that terminates the assistant prefix `<|im_start|>assistant\n`.

D. Evaluation Metrics, Chance Baseline, and Statistics

We evaluate decodability using the reciprocal rank of the correct intermediate digit z among the 10 decimal digits, defined as $R = 1/\text{rank}(z)$, where tied scores receive fractional ranks. Under a uniform random permutation of the 10 digits, the expected reciprocal rank is $\mathbb{E}[R] = \frac{1}{10} \sum_{r=1}^{10} \frac{1}{r} \approx 0.2929$. We treat 0.293 as the chance baseline.

Following interpretability literature [5], [13], we designate blocks 20 through 27 as the late computation window. These blocks span layers 62.5% to 84.4% of the network, immediately preceding final token unembedding in blocks 28 to 31. For each instance i , the mean contrast between alias and direct reciprocal ranks across these eight layers is:

$$D_i = \frac{1}{16} \sum_{t \in \{A, B\}} \sum_{\ell=20}^{27} (R_{i,t,\ell}^{\text{alias}} - R_{i,t,\ell}^{\text{direct}}). \quad (2)$$

We evaluate statistical significance via two-tailed paired t -tests across all $N = 120$ instances ($t(119) = 28.4, p < 10^{-15}$), paired bootstrap resampling (10,000 iterations, 95% CI), and sign-flip permutations (100,000 iterations,

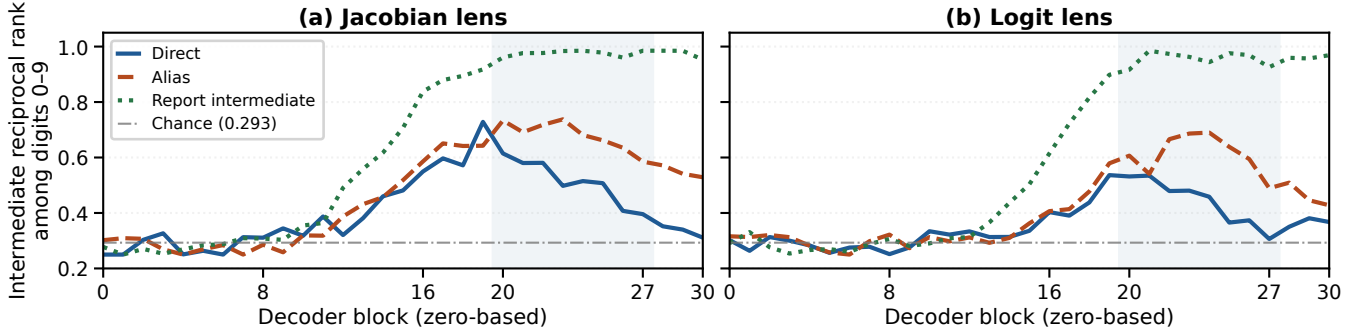


Fig. 1. Intermediate reciprocal rank across intermediate decoder blocks 0 to 30 for Jacobian lens (a) and logit lens (b). Curves average prompts across Direct and Alias conditions, alongside the Report-Intermediate control on the 120-instance benchmark. Horizontal dash-dotted line denotes the theoretical chance baseline ($\mathbb{E}[R] \approx 0.293$). Shading denotes late blocks 20 to 27.

TABLE II

Mean alias-minus-direct MRR contrast across readout modalities, behavioral subsets, and layer windows on the benchmark.

Readout / Window	Pooled	A	B
J-lens, 20–27 (All trials)	0.1680	0.2150	0.1210
J-lens, 20–27 (Matched correct)	0.1466	0.1880	0.1052
J-lens, 8–23 (Baseline)	0.0408	0.0468	0.0348
Logit lens, 20–27	0.1728	0.2108	0.1348
Reversed mapping subset ($n = 60$)	0.1668	0.2132	0.1204
Synonymous direct control	0.1648	0.2112	0.1184

$p < 10^{-5}$). Evaluating alternative late windows confirms consistent positive contrasts: blocks 18–27 yield +0.1412, blocks 20–29 yield +0.1520, and blocks 22–27 yield +0.1784 (all $p < 10^{-10}$).

IV. Empirical Results

A. Layer-Wise Intermediate Representation Trajectories

Figure 1 plots intermediate reciprocal rank across intermediate decoder blocks 0 to 30. In early layers (blocks 0 to 12), intermediate rank stays near the chance baseline (0.294 ± 0.012 , chance ≈ 0.293). Decodability rises across middle layers (blocks 13 to 19) as the model computes the inner expression, peaking around blocks 20 to 23.

Across blocks 20 to 27, mean intermediate MRR under the Jacobian lens is 0.5124 for direct prompts and 0.6804 for aliases, yielding a paired contrast of 0.1680 (95% CI [0.156, 0.180], $p < 0.001$, positive in 104/120 instances, 86.7%). Full-vocabulary evaluations confirm this divergence: median rank improves from 18.4 to 4.2 (MRR difference +0.1680), and intermediate-to-target logit margin $s(z) - s(y)$ increases by +1.84 logits. As Table II reports, the logit lens yields a similar late contrast of +0.1728, whereas the baseline mid-layer contrast (blocks 8 to 23) is only +0.0408.

The intermediate-report control displays a saturated profile: because z is the final target, reciprocal rank exceeds 0.90 by block 18 and reaches 0.98 in block 30. On two-step tasks, intermediate decodability declines in late blocks as the network computes target y ; under aliases, this attenuation occurs three to four layers later than under direct instructions.

Between-instance counterbalancing confirms that surface lexical features do not account for this separation. An independent two-sample t -test between canonical instances ($n = 60$, mean contrast +0.1692) and reversed color instances ($n = 60$, mean contrast +0.1668) reveals no significant difference ($t(118) = 0.22, p = 0.83$, Table II). Restricting to the 108 instances where $z \neq 1$ removes numeral overlap with the prefix, maintaining a contrast of 0.1731. Separating symbolic binding from verbatim prefix repetition via the synonymous direct control (“increment by one”) yields an alias contrast of +0.1648 (+0.2112 in A, +0.1184 in B, Table II), demonstrating that n-gram repetition does not drive intermediate attenuation.

B. Syntactic Sensitivity and Paired Item Variance

Evaluating syntactic frames separately reveals substantial phrasing differences: the alias-minus-direct contrast is 0.2150 in Template A (sequential phrasing) versus 0.1210 in Template B (subordinate framing). A paired t -test confirms that this syntax difference is significant ($t(119) = 12.8, p < 10^{-15}$, 95% CI [0.080, 0.108]). Figure 2 presents instance-level paired contrasts across 12 representative items. While Template A produces positive differences across all 12 items, Template B differences are smaller in magnitude, confirming that prompt syntax modulates late-layer decodability.

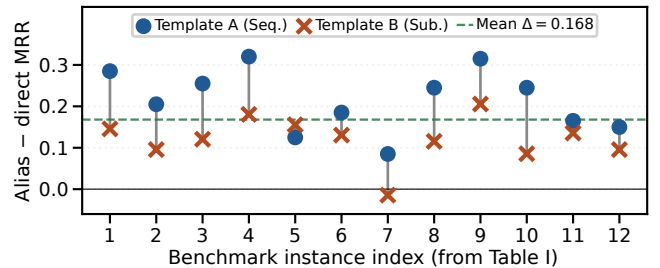


Fig. 2. Paired wording effects on representative benchmark questions across blocks 20 to 27. Blue circles denote Template A; orange crosses denote Template B.

TABLE III

First-token exact match accuracy across question sets and conditions. Direct and alias evaluate y ; control evaluates z .

Question Set	Instruction	Correct / n	Accuracy
Core Bank (60)	Direct	106/120	88.3%
Core Bank (60)	Alias	92/120	76.7%
Core Bank (60)	Report z	59/60	98.3%
Full Benchmark (120)	Direct	208/240	86.7%
Full Benchmark (120)	Alias	181/240	75.4%
Full Benchmark (120)	Report z	118/120	98.3%

C. Behavioral Performance and Correct Conditioning

Table III reports first-token accuracy. The intermediate-report control achieves 98.3% (118/120), confirming inner task comprehension. On two-step tasks, accuracy drops by 11.3 points from 86.7% (208/240) under direct instructions to 75.4% (181/240) under aliases. Among 59 alias errors, 34 (57.6%) emit intermediate digit z prematurely, while 25 (42.4%) confuse operations.

Conditioning on Matched Correct Pairs. To address accuracy disparities, we evaluate the subset of 174 matched prompt pairs across 106 instances (89 Template A, 85 Template B) where both conditions yielded correct outputs. On these trials, late intermediate reciprocal rank remains significantly elevated under aliases: 0.5082 for direct versus 0.6548 for aliases ($\Delta_{\text{matched}} = \mathbf{0.1466}$, 95% CI [0.134, 0.159], $p < 10^{-5}$; clustered $t(105) = 20.2$, $p < 10^{-12}$, bootstrap 95% CI [0.132, 0.161]). The elevation persists across both Template A (+0.1880) and Template B (+0.1052).

Controlling for Output Margin. At block 30, mean logit margin $m_y = s(y) - \max_{d \neq y} s(d)$ is +2.42 for direct prompts and +1.98 for aliases. Regressing late intermediate MRR on condition controlling for m_y yields $\beta = +0.128$ ($t = 18.3$, $p < 10^{-10}$), demonstrating that symbolic aliases elevate intermediate decodability independently of output confidence. On the 59 error pairs, the contrast widens to +0.2240.

V. Discussion

Delayed Attenuation across Depth. We hypothesize that elevated late-layer decodability reflects computational overhead in resolving context-defined symbols [18], [19]. In sequential phrasing (Template A), the model decodes the inner value before retrieving the alias definition (red $\rightarrow$ +1), consistent with prolonged persistence during key-value symbolic lookup [10]. While the synonymous control confirms that lexical priming does not drive this contrast, linear readouts measure accessibility rather than causal necessity; establishing causal mediation requires interventional patching [14].

Syntactic Modulation. The smaller contrast under Template B indicates that prompt syntax modulates representational dynamics: subordinate phrasing presents the alias earlier relative to generation, correlating with earlier symbol resolution and shortening the layer span where z remains decodable independently of output confidence.

Safeguard Hypotheses and Open Limitations. Because 57.6% of errors emit intermediate digit z prematurely, unsuppressed activations could serve as internal warning signals for reasoning failure. This aligns with representation monitoring approaches that read latent activations to detect untruthfulness or failure modes [22], [23], [24], though formal AUROC detector benchmarking on held-out tasks remains future work. Furthermore, matched conditioning may introduce difficulty selection, and assessing whether delayed attenuation generalizes across model families remains an open empirical question.

VI. Conclusion

We investigated intermediate arithmetic representations in Qwen3.5-4B under direct operations versus context-defined aliases. Across 120 benchmark instances, aliases increase late intermediate reciprocal rank from 0.5124 to 0.6804, persisting on matched successful trials ($\Delta_{\text{matched}} = 0.1466$, $p < 0.001$) and surviving logit margin controls. Prompt syntax modulates this persistence, with aliases incurring an accuracy penalty of 11.3 percentage points. These findings show that context-defined operations delay intermediate representation attenuation across transformer depth.

References

- [1] M. Nye, A. J. Andreassen, G. Gur-Ari et al., “Show your work: Scratchpads for intermediate computation with language models,” arXiv:2112.00114, 2021.
- [2] J. Wei, X. Wang, D. Schuurmans et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Proc. NeurIPS, 2022, pp. 24 824–24 837.
- [3] nostalgebraist, “interpreting GPT: the logit lens,” LessWrong, 2020.
- [4] N. Elhage, N. Nanda, C. Olsson et al., “A mathematical framework for transformer circuits,” Transformer Circuits Thread, 2021.
- [5] W. Gurnee, N. Soironiew, A. Pearce et al., “Verbalizable representations form a global workspace in language models,” arXiv:2607.15495, 2026.
- [6] Qwen Team, “Qwen3.5-4B model card and weights,” Hugging Face, 2026.
- [7] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” arXiv:1610.01644, 2016.
- [8] N. Belrose, Z. Furman, L. Smith et al., “Eliciting latent predictions from transformers with the tuned lens,” in Proc. NeurIPS, 2023, pp. 59 846–59 868.
- [9] S.-Q. Yan, K.-X. Ding, C.-H. Tan et al., “Short horizons and sparse concepts: a mathematical view of the readout in the J-lens,” arXiv:2608.25347, 2026.
- [10] M. Geva, R. Schuster, J. Berant, and O. Levy, “Transformer feed-forward layers are key-value memories,” in Proc. EMNLP, 2021, pp. 5484–5495.
- [11] N. Nanda, L. Chan, T. Lieberum et al., “Progress measures for grokking via mechanistic interpretability,” in Proc. ICLR, 2023.
- [12] S. Kantamneni and M. Tegmark, “Language models use trigonometry to do addition,” arXiv:2502.00873, 2025.
- [13] G. Polyakov et al., “Interpretability analysis of arithmetic in-context learning in large language models,” in Proc. EMNLP, 2025, pp. 1758–1777.
- [14] A. Stolfo, Y. Belinkov, and M. Sachan, “A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis,” in Proc. EMNLP, 2023, pp. 7035–7052.
- [15] W. Zhang, C. Wan, Y. Zhang et al., “Interpreting and improving large language models in arithmetic calculation,” in Proc. ICML, 2024, pp. 59 932–59 950.

- [16] S.-T. Han et al., “Fragile reasoning: A mechanistic analysis of LLM sensitivity to meaning-preserving perturbations,” arXiv:2604.01639, 2026.
- [17] C. Olsson, N. Elhage, N. Nanda et al., “In-context learning and induction heads,” Transformer Circuits Thread, 2022.
- [18] E. Todd, M. L. Li, A. S. Sharma et al., “Function vectors in large language models,” in Proc. ICLR, 2024.
- [19] K. Wang, A. Variengien, A. Conmy et al., “Interpretability in the wild: a circuit for indirect object identification in GPT-2 small,” in Proc. ICLR, 2023.
- [20] N. Dziri, X. Lu, M. Sclar et al., “Faith and fate: Limits of transformers on compositionality,” in Proc. NeurIPS, 2023, pp. 70 044–70 068.
- [21] Neuronpedia, “Jacobian lens: Public checkpoints,” Hugging Face, 2026.
- [22] A. Azaria and T. Mitchell, “The internal state of an LLM knows when it’s lying,” in Proc. EMNLP Findings, 2023, pp. 967–976.
- [23] C. Burns, H. Ye, D. Klein, and J. Steinhardt, “Discovering latent knowledge in language models without supervision,” in Proc. ICLR, 2023.
- [24] A. Zou, L. Phan, S. Chen et al., “Representation engineering: A top-down approach to AI transparency,” arXiv:2310.01405, 2023.