SPARKS OF GENERALIZATION: CAPACITY WINDOWS AND STEERING DURING TRANSFORMER PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a transformer learn rules while retaining facts under the same next-token objective? We study Qwen3 and Gemma3 trained from scratch on controlled corpora with native tokenizers, one head per layer, and one or two layers in both architectures (three in Qwen3). Single-digit addition and subtraction probe in-distribution generalization, requiring all 196 training and four withheld equations correct at the final evaluation. With facts included, success also requires recall of at least 48 of 50 facts at that evaluation (joint success). In one-layer arithmetic-only baselines of both architectures, generalization is most frequent within a limited hidden-width range. At the widest tested width, all 10 seeds per architecture fit the training equations and none generalizes. To steer generalization, we introduce answer consistency, an auxiliary loss on arithmetic representations grouped by operator and training answer. Across thirteen steering blocks (three arms on shared seeds) using arithmetic alone or adding facts and approximately 1,400 prose tokens, it raises success over the no-penalty baseline in nine and over a per-step shuffled-label control in five (nominal, unadjusted paired $p < .05$, all assigned seeds). In one-layer arithmetic-only models at the widest tested width, steering yields no detectable gain in success. With facts and this prose corpus, steered one-layer width-14 models reach joint success in 34/60 Qwen3 and 46/60 Gemma3 seeds, versus 20/60 and 32/60 for baseline and 17/60 and 33/60 for shuffled labels. Steering also tightens same-answer representations at the regularized site across the architectures and depths studied.

1 INTRODUCTION

We want a pretrained model to learn rules from examples while also remembering facts. Both demands enter through the same next-token objective, yet fitting the training tokens need not produce a solution that generalizes. We ask how model capacity and training content affect generalization, and whether an additional training signal can make it more reliable. In a controlled pretraining environment, we train Qwen3 and Gemma3 transformers from random initialization with native tokenizers and one attention head per layer. Fully specified corpora let us account for every token, and we vary hidden width, depth (one or two layers in both architectures, three in Qwen3), and content across repeated seeds.

Following Barron & White (2025), we use single-digit addition and subtraction with four equations withheld. *In-distribution generalization* (IDG) requires correct greedy answers to all 196 training and all four withheld equations. It is a behavioral probe of rule learning, not an identification of the learned algorithm. Success requires IDG at the final evaluation. With facts in the corpus, success is *joint success*, which also requires recall of at least 48 of 50 facts at that evaluation.

One-layer arithmetic-only models generalize most often within a limited range of hidden widths, whereas much wider models fit every training equation but fail on the withheld equations. Adding facts and then growing amounts of prose (about five times as much in the $5\times$ condition as in $1\times$) moves the arithmetic-only setting toward pretraining's mixed learning demands. Across the one-layer corpus progression, joint success declines while factual recall stays high, but arithmetic exposure and sampled widths also change, so the decline does not isolate an effect of prose.

Our central contribution is *answer consistency*, an auxiliary loss on arithmetic examples selected by a training-data flag. It acts on their hidden representations, grouped by operator and training answer,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

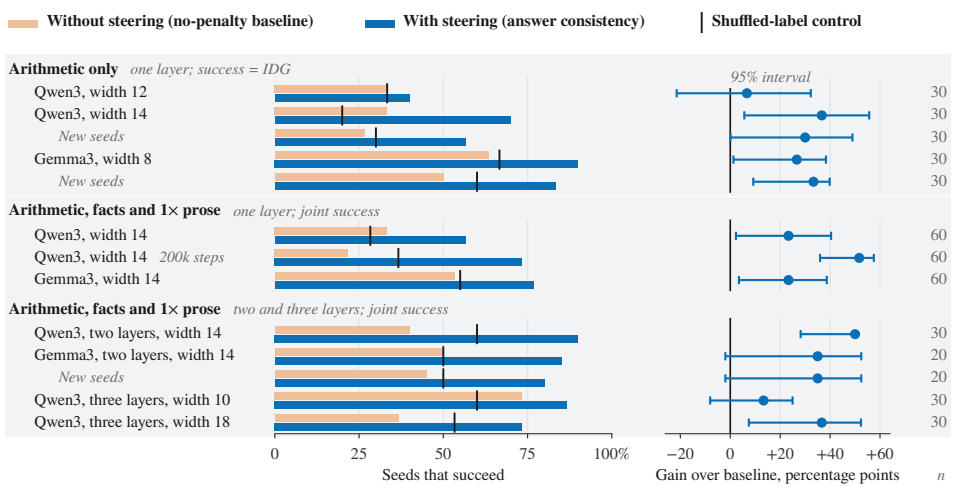

Figure 1: **Targeted steering makes success more frequent in most blocks.** Bars give the share of all assigned seeds that succeed in each of the thirteen steering blocks of Table 1 (n seeds per arm); success is final-evaluation IDG, plus at least 48 of 50 facts when facts are trained (joint success). Points: steered minus baseline on paired seeds, with 95% intervals from inverting the exact McNemar test (Appendix E); intervals exclude zero exactly when nominal, unadjusted $p < .05$.

while all content keeps the usual next-token loss. As with worked examples used to teach the chain rule in calculus, the aim is to learn a rule that applies to new examples. The flag marks the worked examples, and the objective supplies answer-derived groupings, not the rule itself, without requiring the model to forget the examples.

In one-layer arithmetic-only models, steering with answer consistency raises final-evaluation success from 10/30 baseline seeds to 21/30 in Qwen3 at width 14 and from 19/30 to 27/30 in Gemma3 at width 8. Both gains recur in second populations on new seeds. In all four comparisons, every assigned model completes training and is counted, and each gain over baseline has nominal, unadjusted $p < .05$ (Figure 1; Table 1).

With facts and $1\times$ prose, steered one-layer width-14 models of both architectures reach joint success more often than either the baseline or a shuffled-label control (trained with the same auxiliary loss but with group labels reshuffled at every step). Across thirteen *steering blocks* (three arms on shared assigned seeds; one to three layers), steering improves on the baseline in nine and on shuffled labels in five at nominal, unadjusted $p < .05$; three of the thirteen comparisons with the shuffled-label control remain after Holm correction. We detect no benefit with $5\times$ prose in one-layer Qwen3, or with arithmetic alone at the widest tested one-layer width (48) in either architecture.

Internally, swapping the regularized representation between equations with the same operator and answer gives steered models lower excess swap loss than either comparator; the average advantage almost vanishes at matched relative swap size, so the lower swap loss reflects smaller differences between same-answer representations.

2 Related Work

Arithmetic and generalization. Fitting and generalization can come apart on small algorithmic tasks: in grokking, generalization emerges after the model has already fit its training data (Power et al., 2022; Nanda et al., 2023). Small transformers trained from scratch can learn arithmetic; formatting changes and chain-of-thought training data improve their accuracy (Lee et al., 2023). Length-generalization studies test extrapolation beyond trained operand lengths (Jelassi et al., 2023; McLeish et al., 2024). Every operand and answer value in our withheld equations occurs elsewhere in training. We keep this probe fixed while varying capacity and co-trained content.

Joint learning of arithmetic and facts. We build on Barron & White (2025), who trained character-level transformers on single-digit arithmetic and 50 capital-city facts. Only their narrowest model reproduced the withheld equations, and no jointly trained model did. Here, 18/20 one-layer models trained with arithmetic and facts at widths 6–8 succeed in each architecture. Tokenizer, architecture, scoring, and schedule differ, so this contrast identifies no single cause. We also add prose, test training-time steering, and report generalization and factual recall separately and jointly.

Information capacity and coexistence. Networks can fit arbitrary labels (Zhang et al., 2016). In language models, memorization depends on scale, repetition, and example properties (Carlini et al., 2023; Tirumala et al., 2022), and studies of knowledge capacity and training-set memorization measure how much information a model can retain (Allen-Zhu & Li, 2025; Morris et al., 2026). Feldman (2020) argues that memorization can be necessary for learning long-tailed distributions. On corrupted algorithmic data, Doshi et al. (2024) show that memorizing and generalizing representations can coexist. We measure factual recall alongside IDG as depth, width, and training content vary.

Training-time steering and representation objectives. Deduplication and goldfish loss reduce memorization (Lee et al., 2022; Hans et al., 2024), and gradient routing and memorization sinks localize what is learned to designated parts of a network (Cloud et al., 2024; Ghosal et al., 2025). Memory Dial studies controllable memorization during language-model training (Zhang & Emami, 2026). Answer consistency does not target memorization directly. It adds a representation objective to selected arithmetic examples while retaining their next-token loss, penalizing within-answer scatter relative to total scatter at an intermediate prediction site. It relates to supervised contrastive learning, which uses labels to organize representations (Khosla et al., 2020), and to the classical comparison of within- and between-class scatter (Fisher, 1936).

3 Setting and Evaluation

Fitting the training arithmetic, generalizing to withheld equations, and recalling facts are separate outcomes in this study. Following the task of Barron & White (2025), the arithmetic grid contains all 200 additions and subtractions of two single-digit operands. Withholding $5 + 7$, $7 + 5$, $5 - 7$, and $7 - 5$ leaves 196 training equations, with every withheld operand and answer occurring elsewhere in training. The 50 capital-city facts also follow that study.

The corpora progress from arithmetic alone to arithmetic with facts, then arithmetic and facts with approximately 1,400 ($1\times$) or 7,000 ($5\times$) prose tokens. The $1\times$ text is a 2015 UN Environment Programme press release, and $5\times$ adds Churchill's 1946 "Sinews of Peace" speech. Both texts name cities in the facts, covering ten capitals in total. The speech was chosen because it names many capitals. A sampler sets the expected arithmetic share of input tokens: 1.0 for arithmetic alone, .81/.84 for arithmetic with facts in Qwen3/Gemma3, .534 for $1\times$, and .208 for $5\times$. Fixed-mix $5\times$ comparisons also use .534. Each row is an equation, fact, or prose window (Appendix A). Prose contributes training content but is not scored.

Models and training. Qwen3 and Gemma3 models train from random initialization with their native tokenizers (Yang et al., 2025; Team et al., 2025), one attention head per layer, one or two decoder layers in both architectures, and three in Qwen3. Hidden width is the residual-stream dimension and also sets the head and MLP dimensions. All content receives next-token cross-entropy. We use AdamW without weight decay or dropout, 256 rows per step, warmup to a learning rate of .01, cosine decay to 10^{-4} at 50k steps, and 100k training steps. The extended schedule trains fresh models for 200k steps with decay ending at 100k. Evaluation occurs every 500 steps.

Outcomes and comparisons. Arithmetic is scored by greedy generation from the prefix before "=", parsing the first integer after it; facts, by greedy continuation of "Capital of X is", which must begin with the correct capital. Success requires IDG at the final evaluation and, when facts are present, recall of at least 48 of 50 facts at that evaluation (*joint success*). Training completion means 196/196 training equations correct at the final evaluation. All assigned seeds count, so incomplete models are failures; IDG, recall and completion are reported alongside joint success. Arms within a configuration share initialization and data-order seeds; width comparisons share data-order seeds and batch streams, though parameter initializations differ in size. Exact two-sided paired McNemar

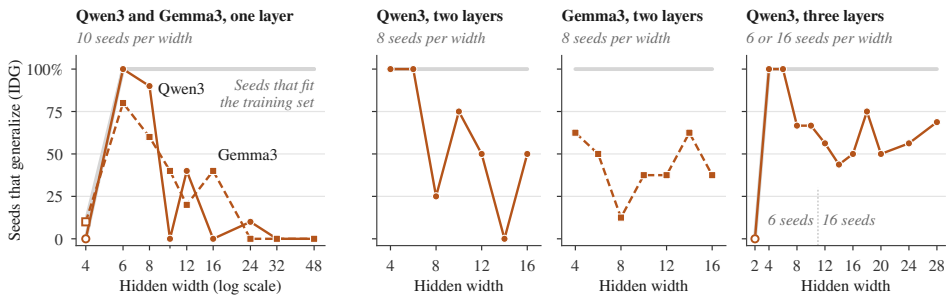

Figure 2: **Arithmetic-only IDG by hidden width at one to three layers.** Lines give the share of assigned seeds reaching IDG (orange; Qwen3 circles, Gemma3 squares) and fitting all 196 training equations (gray) at the final evaluation. Open markers flag widths where not every seed fit.

tests use these shared-seed subsets rather than full populations, and $+b/-c$ denotes b gains and c losses against the named comparator. Unless adjusted, p-values are nominal; analyses restricted to training-complete models are sensitivity analyses (Appendix E).

4 Capacity and Training Content

4.1 Arithmetic Alone: A One-Layer Capacity Window

Both architectures generalize most often within a limited range of hidden widths (Figure 2, left). At width 4, only 0/10 Qwen3 and 1/10 Gemma3 models complete training, whereas every model completes from width 6 onward. Success peaks at the narrowest widths that fit: Qwen3 succeeds in 10/10 and 9/10 seeds at widths 6 and 8, and Gemma3 in 8/10 and 6/10. At widths 24–48, success is at most 1/10 per tested width. At width 48, all ten seeds fit and none generalizes in either architecture. Within the window, success varies non-monotonically (Figure 2, left).

4.2 Deeper Models With Arithmetic Alone

IDG remains present at the widest widths tested (Figure 2, right three panels). Although all two-layer models complete training, success is most frequent at the narrowest tested widths in Qwen3 and varies non-monotonically in Gemma3. At width 16, success is 4/8 in Qwen3 and 3/8 in Gemma3. In three-layer Qwen3, no model completes training at width 2 (a fitting edge), while IDG remains present through width 28. Figure 2 and Appendix B give the distributions and denominators.

4.3 Mixed Corpora at a Fixed Data Mix

Holding content and arithmetic fraction fixed, wider models can fit more reliably while succeeding less often. At one layer with $1\times$ prose and fraction .534, joint success at widths 14/16/18/20 is 13/10/3/4 in Qwen3 and 20/15/4/3 in Gemma3, out of 40 assigned seeds per width and architecture (Figure 3). Completion rises from 34/40 to 40/40 in Qwen3 and from 39/40 to 40/40 in Gemma3; at least 39/40 models meet the 48-fact recall threshold at each width (Appendix B). These width patterns are descriptive, and the widths were chosen from pilot scans (Appendix B.3).

With $5\times$ prose at the same fraction, increasing width from 14 to 28 reduces joint success and improves perfect factual recall in both architectures (Figure 3, right). The paired joint comparisons use 18 shared Qwen3 seeds ($6/18 \to 0/18$, $p = .031$) and 20 Gemma3 seeds ($+0/-10$, $p = .0020$). Nine of Qwen3's paired seeds were selected because their width-14 model fit the $1\times$ corpus. Raising the arithmetic fraction from .208 to .534 at width 28 does not restore Qwen3 success (1/20 and 0/20).

Lower success at wider widths under a fixed data mix also appears descriptively in three-layer Qwen3 with $1\times$ prose. On the same 16 seeds at each width, every model completes training and recalls at least 48 facts. Joint success is 10/16 at width 10 and lower at wider widths on average: 34/64 model

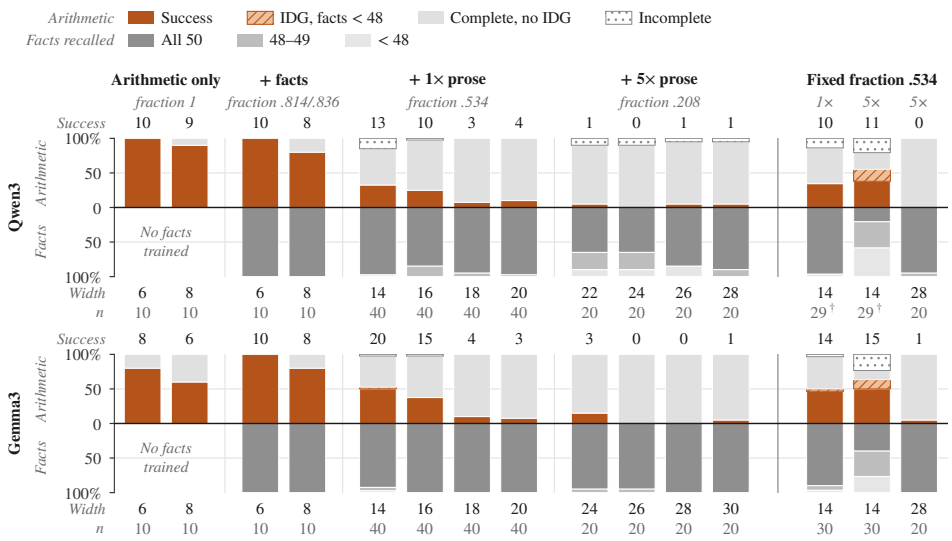

Figure 3: **One-layer success and its components across training content and width.** Each bar is one population of assigned runs at the final evaluation; success is IDG, plus at least 48 of 50 facts when facts are trained, and incomplete runs did not fit all 196 training equations. Group headings give the arithmetic fraction (arithmetic's share of tokens; Qwen3/Gemma3 where they differ). Arithmetic-only bars are Figure 2's one-layer runs at widths 6 and 8, the widths used with facts; mixed-corpus widths were chosen per condition (Appendix B.3). The right group fixes the fraction at .534: width 14 with $1\times$ (a subset of the $1\times$ group's runs) and $5\times$ on the same seeds, and width 28 with $5\times$. †Includes nine seeds selected for fitting the $1\times$ corpus at width 14 (Appendix B.4).

runs at widths 10–16 versus 20/64 at widths 18–28, with non-monotonic variation. Only the one-layer $5\times$ width contrasts above receive paired significance tests here.

4.4 Across corpora

In the one-layer corpus progression, joint success remains frequent with arithmetic and facts, then declines across the prose conditions while recall stays high (Figure 3, left four groups). Because sampled widths and arithmetic exposure change with the corpus, this descriptive progression does not isolate an effect of prose. Adding facts at widths 6–8 changes success from 19/20 to 18/20 in Qwen3 and from 14/20 to 18/20 in Gemma3 (untested; the arithmetic fraction also changes). At matched width 14 and fraction .534, moving from $1\times$ to $5\times$ prose changes joint success little, with many seeds switching in both directions, but perfect recall falls from 28/29 to 6/29 in Qwen3 and from 27/30 to 12/30 in Gemma3 (paired $p = 3.0 \times 10^{-6}$ and .00027; Figure 3, right) while IDG does not decrease. The added speech introduces four fact capitals and nine fact countries absent from the press release, so this recall decline cannot be attributed to volume alone. Appendix B extends the comparison of arithmetic alone with arithmetic, facts and $1\times$ prose to two layers in both architectures and three in Qwen3.

5 Steering with answer consistency

5.1 Objective, targeting, and design

A per-example training-data flag marks arithmetic examples for an auxiliary loss. This flag is dataset metadata, not an input token, and has no role at inference. For flagged examples in a minibatch, let z_i be the last layer's post-attention residual, measured after adding the attention output and before the MLP, at the *units-prediction position* immediately before the answer's final digit. Examples are grouped by operator and signed answer. With group label g_i, group mean $\bar{z}_{g_i}$, and overall mean $\bar{z}$,

training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{next-token}} + \lambda \frac{\sum_{i \in \mathcal{M}} \|z_i - \bar{z}_{g_i}\|_2^2}{\max\{\sum_{i \in \mathcal{M}} \|z_i - \bar{z}\|_2^2, \epsilon\}}, \tag{1}$$

where $\mathcal{M}$ is the minibatch's arithmetic subset and $\epsilon = 10^{-8}$. All content retains next-token loss. Facts and prose receive no direct auxiliary penalty, although shared-weight updates can affect them. The withheld equations contribute to neither loss.

A *steering block* comprises three arms with the same model configuration and data mix, trained on the same assigned seeds: no-penalty baseline, answer consistency, and shuffled labels. The control uses the same loss and weight but freshly permutes labels among arithmetic examples at every step, preserving group sizes. Improvement over this control supports an answer-specific benefit relative to per-step reshuffling. The default λ is .1, with exceptions and training-side selection described in Appendix A.5.

5.2 Arithmetic alone, one layer

Answer consistency increases final-evaluation success over baseline in both architectures, including on new seeds (Table 1). For Qwen3 at width 14, success rises from 10/30 to 21/30, and from 8/30 to 17/30 in a second population. Gemma3 at width 8 rises from 19/30 to 27/30, and from 15/30 to 25/30 on new seeds. Every model completes training, but the Qwen3 width-12 comparison shows no significant gain (10/30 versus 12/30). Gains over the shuffled-label control are nominally significant only in the first Qwen3 width-14 block.

A separate test on 40 fresh Qwen3 seeds at width 14 compares answer-derived with per-step reshuffled labels: success is 25/40 versus 5/40 ($+23/-3$, $p = 8.8 \times 10^{-5}$), with every model complete. This establishes superiority to that resampled control at this configuration, not to persistent groupings.

5.3 With facts and prose, one layer

At width 14 with $1\times$ prose, joint success increases over both comparators in both architectures (Table 1). Qwen3 gives 20/34/17 successes and Gemma3 32/46/33, in baseline/answer/shuffled order, each out of 60 assigned seeds. Recall remains near ceiling: 51/60 steered Qwen3 models recall all 50 facts, versus 55/60 baselines, as do 59/60 steered Gemma3 models. No steered model recalls over three fewer facts than its paired baseline, but near-ceiling recall does not establish zero cost.

On a 200k-step schedule (fresh models on the same seeds, decay ending at 100k), joint success is 13/44/22 of 60 and completion 58/56/60; baseline joint success also falls relative to 100k, so the larger contrast reflects both arms. All 60 steered models recall at least 48 facts, and 58 recall all 50.

5.4 Deeper models

The five deeper-model steering blocks all use arithmetic, facts, and $1\times$ prose. Every model recalls at least 48 facts, so joint success and IDG coincide. At two Qwen3 layers, success is 12/27/18 of 30, improving over both comparators at nominal significance. At three layers, the gain over baseline reaches nominal significance at width 18 (11/30 versus 22/30) but not at width 10 (22/30 versus 26/30); neither comparison with shuffled labels reaches nominal significance. Both Gemma3 two-layer populations favor steering, with 10/17/10 and 9/16/10 successes out of 20, but neither comparison with either control reaches nominal significance. Completion-selected sensitivities appear in Appendix C.

At nominal, unadjusted $p < .05$, steering improves success over baseline in nine of the thirteen blocks (Figure 1) and over shuffled labels in five (Table 1). The latter five are four one-layer blocks in both architectures and the Qwen3 two-layer block; all nominal gains beyond one layer are in Qwen3. Shared configurations and seeds mean the blocks are not thirteen independent confirmations.

5.5 Limits

With $5\times$ prose, steering does not establish a benefit in either tested one-layer Qwen3 configuration. At width 14 and fraction .534, joint success is 11/7/9 of 29, IDG 16/11/13, facts ≥ 48 in 17/21/20, and

Table 1: Steering on all assigned seeds. Triples are baseline/answer/shuffled (B/A/S), and n is seeds per arm. Rows 1–13 are every steering block with the $(5, 7)$ withheld pair, arithmetic alone or the $1\times$ corpus, and the penalty from the start, except the width-48 plateau tests (§5.5); row 7 uses 200k steps, the others 100k. Success is final-evaluation IDG, plus $\geq 48/50$ facts when facts are trained. $p_{A,B}$ and $p_{A,S}$ are nominal exact two-sided paired McNemar tests; row 6 has $p_{A,B} = .058$ on IDG alone. Holm-adjusted over the thirteen $p_{A,S}$, rows 2, 6 and 7 remain significant (.00079, .037, .0038). The separate labels test has no baseline and is outside the tally and Holm family.

Configuration	Corpus	n	Success	IDG	Facts ≥ 48	Complete	$p_{A,B}$	$p_{A,S}$
1. Qwen3 1L w12	Arithmetic	30	10/12/10	10/12/10	–	30/30/30	.80	.80
2. Qwen3 1L w14	Arithmetic	30	10/21/6	10/21/6	–	30/30/30	.019	6.1×10^{-5}
3. Qwen3 1L w14, new seeds	Arithmetic	30	8/17/9	8/17/9	–	30/30/30	.049	.057
4. Gemma3 1L w8	Arithmetic	30	19/27/20	19/27/20	–	30/30/30	.039	.092
5. Gemma3 1L w8, new seeds	Arithmetic	30	15/25/18	15/25/18	–	30/30/30	.0063	.092
6. Qwen3 1L w14	$1\times$	60	20/34/17	23/35/19	56/57/55	54/48/54	.029	.0033
7. Qwen3 1L w14	$1\times$	60	13/44/22	15/44/24	58/60/57	58/56/60	3.7×10^{-8}	.00031
8. Gemma3 1L w14	$1\times$	60	32/46/33	33/46/33	59/60/60	58/59/60	.02	.029
9. Qwen3 2L w14	$1\times$	30	12/27/18	12/27/18	30/30/30	29/30/30	6.1×10^{-5}	.022
10. Gemma3 2L w14	$1\times$	20	10/17/10	10/17/10	20/20/20	20/19/20	.065	.092
11. Gemma3 2L w14, new seeds	$1\times$	20	9/16/10	9/16/10	20/20/20	20/20/20	.065	.15
12. Qwen3 3L w10	$1\times$	30	22/26/18	22/26/18	30/30/30	30/30/28	.29	.057
13. Qwen3 3L w18	$1\times$	30	11/22/16	11/22/16	30/30/30	30/30/30	.013	.18
Separate labels test: Qwen3 1L w14	Arithmetic	40	–/25/5	–/25/5	–	–/40/40	–	8.8×10^{-5}

completion 23/15/19. Steered models complete less often than baseline ($+3/-11$, $p = .057$), as do shuffled-label models, so an answer-specific cost is not established; under a symmetric location-shift model, an average facts cost of four or more is rejected, but smaller costs are not (Appendix E). At width 28 and fraction .208, joint success is 1/0/0 of 30, too few successes to assess harm to generalization; completion falls from 29/30 baseline to 23/30 steered ($p = .031$), and all models recall at least 48 facts. At width 48 with arithmetic alone, success is 0/0/0 in Qwen3 and 0/1/0 in Gemma3, each out of 20 with all models complete: steering did not open the wide plateau. Steering also delays fitting: on paired seeds where both arms complete, later first fits to 196/196 under steering outnumber earlier ones relative to baseline in ten of thirteen blocks and never the reverse (nominally significant in five). For Qwen3 $1\times$ at 100k, the baseline and steered medians are 19.5k and 24.5k steps (32 paired seeds later, 10 earlier; paired sign $p = .00094$), while the shuffled control has no nominal delay. No Table 1 block shows a nominal final-completion difference against either comparator, although Qwen3's 100k steered arm completes in 48/60 versus 54/60 baselines. Appendix C reports the fitting-time distributions and the second withheld pair $(4, 9)$, on which a steering benefit was not established.

6 Internal changes under steering

6.1 Populations and measurements

The one-layer cohort contains 48 final checkpoints: eight paired seeds per architecture and arm, width 14, $1\times$ prose, and 100k steps. The deeper cohort contains 240 models: Qwen3 at two layers and at three layers with width 18 (30 seeds each), and one Gemma3 two-layer population (20 seeds), all three arms. These retrospective assays reuse checkpoints, so they are not independent replications.

At the units-prediction position, write the final layer's state before output normalization as $h_i = A_i + B_i$, where A_i is the post-attention residual and B_i the MLP output. For training equations i, j sharing an operator and signed answer, a post-attention swap replaces A_i by A_j while holding B_i fixed. The modified state is passed through the final output normalization and vocabulary readout. *Excess swap loss* is the units-token loss under this replacement minus the recipient's unmodified loss, conditional on the correct prefix in both cases. This teacher-forced measurement is distinct from greedy-generation IDG. We also measure same-answer component differences, raw and relative to the state norm. Appendix D defines pair selection and averaging.

Table 2: Full-size and size-matched post-attention swaps, followed by geometry. Full advantage is comparator minus steered excess loss in nats on the pairs retained for matching. Ranges span both comparators (baseline and shuffled labels); remainder also spans matching variants. B/S give paired-seed denominators. Distance ratio is the mean steered/comparator same-answer distance ratio. Shared shrinkage compares the contractions of different- and same-answer distances within operator/preceding-token groups; negative values mean different-answer distances grew. All use $1\times$ prose, 100k steps, and width 14 (three layers: 18). Details and per-seed results: Appendix D.

Population	Seeds B/S	Full advantage	Remainder (%)	Distance ratio	Shared shrinkage
Qwen3 1L	7/8	3.23 to 3.68	-12 to $+3$	0.29 to 0.34	0.97
Gemma3 1L	8/8	2.78 to 3.02	-4 to -2	0.25 to 0.30	0.91 to 0.92
Qwen3 2L	30/30	0.75 to 1.17	-11 to $+1$	0.27 to 0.31	0.39
Gemma3 2L	20/20	1.14 to 1.51	-5 to -1	0.19 to 0.21	0.67 to 0.69
Qwen3 3L	30/30	0.38 to 0.41	-4 to $+2$	0.31 to 0.33	-0.77 to -0.76

6.2 Component swaps and size-matched controls

Steered models have lower excess swap loss than both comparators at one layer and at depth (Table 2). At one layer we examine both post-attention and MLP-output swaps. Exchanging B_i for B_j gives $A_i + B_j$, the same state as the reverse-direction post-attention swap. Their symmetric average losses supply one measurement, rather than independent evidence.

To test whether lower swap loss reflects a smaller perturbation, we match the post-attention displacement relative to the recipient state norm, pair by pair across steered and comparator models. Each comparator's swap loss is evaluated along its own swap path at the steered model's relative size. Almost none of the steered models' average advantage remains: -12% to $+3\%$ across the one-layer contrasts and matching variants, where a negative remainder means steered swaps cost more at matched size. Every Gemma3 seed retains at most 10% of its own advantage, while Qwen3 seeds vary. Without the four training-incomplete models, the lower swap loss holds and the Qwen3 matched-size remainder is negative in every variant (signs untested; Appendix D).

At depth, lower full-size excess swap loss occurs in at least 27/30 Qwen3 and 19/20 Gemma3 paired seeds per contrast (Holm-adjusted over six sign tests, $p < 10^{-4}$). Matching relative size leaves -11% to $+2\%$ of the mean advantage across contrasts and matching rules, without a consistent sign across seeds. Thus, across the studied depths, the lower swap loss reflects smaller post-attention differences.

Equal-displacement random perturbations provide a second control. At full size, the swaps' advantage over random perturbations is smaller under steering. This between-arm contrast becomes much smaller at matched size at one layer and nearly vanishes at depth, without establishing equivalence between swaps and random perturbations. Low swap loss also does not certify successful arithmetic: a one-layer baseline fitting only 28/196 equations and a steered two-layer Gemma3 model fitting 64/196 have near-zero excess swap loss.

6.3 Tighter same-answer representations

Steered models have smaller same-answer post-attention differences, raw and relative to the state norm: at one layer in every defined paired comparison (seven against Qwen3 baseline, eight otherwise), and at depth in at least 29/30 Qwen3 and 20/20 Gemma3 seeds per contrast. This is an expected manipulation check: the objective directly penalizes within-group scatter relative to total scatter at this site.

At one layer, most of the contraction also occurs between different-answer equations sharing the operator and the token immediately before the units digit, though same-answer pairs tighten somewhat more, most consistently in Gemma3. Gemma3 at two layers shows a similar pattern, less uniformly across seeds. In Qwen3, less than half of the shrinkage is shared at two layers, and different-answer distances grow at three layers, mostly among negative answers. Architecture, depth, λ, and width vary together, so this difference cannot be attributed to any one of them.

6.4 Relation to success

Within the one-layer arms, generalizing models show no tendency toward lower swap loss. At depth, pooled within-arm comparisons find no detectable association between same-answer differences relative to the state norm and final-evaluation IDG, although a descriptive Gemma3 two-layer training-complete sensitivity is an exception (one-sided $p = .013$, threshold .025). A separate focused one-layer analysis finds a normalization-dependent geometry–IDG association in steered Gemma3 models that remains with all assigned steered models. The pooled all-assigned and Qwen3 analyses detect none (Appendix D). Steering changes representations at the objective's site, and lower swap loss reflects smaller differences. These measurements do not establish why success improves.

7 Discussion

Above the underfitting regime, restricting hidden width in one-layer models can make IDG more frequent even though wider models also fit every training equation. One interpretation follows the traditional capacity-control intuition: a restricted representation may favor a shared solution over fitting examples separately, and a rule can describe its examples more compactly, the idea behind minimum description length (Rissanen, 1978). This interpretation does not imply that the narrow models were incapable of memorizing the arithmetic table. Width also changes the optimization problem, and differences in which solutions training reaches could contribute to this capacity window. In deeper arithmetic-only models, however, IDG remains present at the widest widths tested.

The decline in joint success across the corpus progression motivates a competition hypothesis: learning facts and prose may compete with acquiring a shared arithmetic solution. Other explanations remain plausible because a smaller arithmetic token fraction changes exposure to the equations, and learning several kinds of content in shared parameters may change optimization difficulty. Sampled widths also differ across corpora, and matched-width comparisons run in both directions without isolating suppression. Separating scored components matters: high factual recall can coexist with failed IDG, while at matched content perfect recall declines without a decline in IDG.

Evidence for steering's benefit does not depend on the competition hypothesis: the comparisons hold model configuration and data mix fixed, and gains also occur with arithmetic alone. Within each operator, answer consistency connects different operand combinations through a shared signed training answer. One hypothesis is that this coupling makes predictive features shared across equations easier to acquire or stabilize during optimization. The answer-versus-shuffled comparisons ask whether meaningful groupings add a benefit beyond the same loss with resampled labels, and support that distinction in several configurations. The broad contraction at one layer suggests that changes extending beyond individual answer groups may also matter. In deeper Qwen3 models, steering raises success even though the contraction it produces is more specific to same-answer pairs (architecture, depth, λ, and width covary here), so broad contraction may not be necessary.

8 Conclusion

In Qwen3 and Gemma3 trained from scratch on fully specified corpora, fitting arithmetic examples does not guarantee generalization. One-layer arithmetic-only IDG is most frequent within a limited hidden-width range, while deeper arithmetic-only models also generalize at the widest widths tested (two layers in both architectures, three in Qwen3). Across the one-layer corpus progression, joint success declines while factual recall stays high, but changing arithmetic exposure and sampled widths prevent isolating a prose effect. Answer consistency improves success over baseline and the shuffled-label control in several configurations at nominal, unadjusted $p < .05$, including joint success in both architectures at one layer and width 14 with facts and $1\times$ prose. No gain in success is detected at the one-layer arithmetic-only width-48 plateau or in the tested one-layer Qwen3 configurations with $5\times$ prose. Steered models have tighter same-answer representations at the regularized site and lower excess swap loss than both comparators, an average advantage that almost disappears when swap displacement is matched in size. These findings concern single-digit addition and subtraction, one head per layer, and the tested widths and depths, with steering benefits established on the $(5, 7)$ withheld pair. Whether answer-derived training signals help with other rules, larger models, and heterogeneous pretraining corpora remains open.

AI USE STATEMENT

Generative AI tools assisted writing and polishing, retrieval and discovery of related work, research ideation and execution, and drafting sections of this paper. These uses included assistance with code and analysis as well as manuscript preparation. Numerical claims were checked against stored experimental records, and computational checks were used to validate reported counts and comparisons. The authors are responsible for the paper's claims, interpretation, and final content.

ETHICS STATEMENT

The study uses synthetic arithmetic, capital-city facts, and identified published prose to investigate learning in small models trained from scratch. It involves no human-subject study. The experiments characterize a controlled setting and do not establish how the intervention would affect privacy, memorization of personal information, or performance in deployed language models.

REPRODUCIBILITY STATEMENT

Appendix A specifies the corpora, model configurations, training recipe, scoring, and auxiliary objective. Appendices B–E provide population definitions, detailed outcomes, internal-measurement procedures, and statistical methods. Code will be released after peer review. The two prose texts are identified by source so that readers can obtain them.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2404.05405.

Joshua Barron and Devin White. Too big to think: Capacity, memorization, and generalization in pre-trained transformers. *arXiv preprint arXiv:2506.09099*, 2025.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv preprint arXiv:2410.04332*, 2024.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. In *International Conference on Learning Representations*, volume 2024, pp. 3782–3810, 2024.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pp. 954–959, 2020.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Gaurav R Ghosal, Pratyush Maini, and Aditi Raghunathan. Memorization sinks: Isolating memorization during llm training. *arXiv preprint arXiv:2507.09937*, 2025.

Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, et al. Be like a goldfish, don't memorize! mitigating memorization in generative llms. *Advances in Neural Information Processing Systems*, 37:24022–24045, 2024.

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. *arXiv preprint arXiv:2307.03381*, 2023.

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, et al. Transformers can do arithmetic with the right embeddings. *Advances in Neural Information Processing Systems*, 37: 108012–108041, 2024.

John Xavier Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much can language models memorize? In *Forty-third International Conference on Machine Learning*, 2026.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Xiangbo Zhang and Ali Emami. Memory dial: A training framework for controllable memorization in language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3654–3673, 2026.

A METHODS

A.1 ARITHMETIC, FACTS, AND PROSE

The arithmetic grid consists of $a + b$ and $a - b$ for every $a, b \in \{0, \ldots, 9\}$. Equations are written as `<a+b=c>` or `<a-b=c>` followed by a newline. The four equations involving both 5 and 7 are excluded from the training corpus. The 196 remaining equations and 50 statements of the form "Capital of X is Y" follow Barron & White (2025). For the second-pair experiment we instead withhold both operand orders under both operators for $(4, 9)$.

The $1\times$ prose corpus is the UN Environment Programme press release beginning "Tallinn, 30 April 2015", announcing the regional launch of the Global District Energy in Cities Initiative. It contains 1,412 Qwen3 tokens and 1,378 Gemma3 tokens. The $5\times$ corpus adds Winston Churchill's "The Sinews of Peace", delivered at Westminster College in Fulton, Missouri, on 5 March 1946, for totals of 7,007 and 6,903 tokens. The speech was chosen because it names many capital cities. Of the 50 fact capitals, the press release names Paris, Tokyo, London, Copenhagen, Helsinki, and Warsaw; the speech names Berlin, Moscow, Warsaw, Vienna, and Athens. The texts overlap at Warsaw and cover ten fact capitals in total. The speech also adds nine fact countries absent from the press release.

A.2 TOKENIZATION AND SAMPLING

Each architecture uses its full native tokenizer and vocabulary: 151,669 Qwen3 entries and 262,145 Gemma3 entries. Text rows are encoded without added special tokens. Every arithmetic digit is a separate token, and both tokenizers merge "=-". Each prose line is tokenized and divided into stride-one windows matching the arithmetic input-row width, eight tokens for Qwen3 and nine for Gemma3. Examples are right-padded within batches, with padding excluded from next-token loss. Each training step samples 256 rows with replacement.

The mixture is controlled in input-token units. For partitions k (arithmetic, facts, prose) with desired token shares t_k and mean row lengths $\bar{\ell}_k$, the partition sampling probabilities are

$$q_k = \frac{t_k/\bar{\ell}_k}{\sum_j t_j/\bar{\ell}_j}. \tag{2}$$

Sampling is uniform within each partition. Arithmetic has token share f, and the remaining share is split between facts and prose in proportion to their raw token counts, before prose windowing. Arithmetic-only training uses $f = 1$; arithmetic with facts uses the natural row mixture, yielding fractions .814 and .836 in Qwen3 and Gemma3. The prose conditions use .534 and .208, with .534 also used in the fixed-mix $5\times$ comparisons. The batch sampler has its own generator seeded by the data-order seed, permitting matching batch streams across widths.

A.3 ARCHITECTURE AND OPTIMIZATION

Models are decoder-only Qwen3 or Gemma3 architectures initialized randomly, with tied input/output embeddings. Each layer has one attention head, and the residual width, head dimension, and MLP intermediate dimension are equal. At one layer and width 14, the total/non-embedding parameter counts are 2,124,808/1,442 for Qwen3 and 3,671,500/1,470 for Gemma3. Thus, total parameter count alone obscures how little computation is available outside the embeddings.

AdamW uses $\beta_1 = .9$, $\beta_2 = .99$, gradient clipping at norm 1, no weight decay, and no dropout. The learning rate warms up over 200 steps to .01, follows cosine decay to 10^{-4} at step 50,000, and stays there until step 100,000. Extended-schedule models start from random initialization on the same seeds, decay until step 100,000, and train for 200,000 steps. All training used one NVIDIA RTX 5090 GPU in FP32 with PyTorch and Hugging Face Transformers; a 100k-step run took approximately 20–40 minutes depending on architecture and depth.

Grokking studies typically train far beyond fitting the training set, often with weight decay (Power et al., 2022). Such horizons were not tested here: the longest reported schedule is 200k steps, and whether much longer training would produce generalization is untested.

Table 3: Main experiment families. Q/G denote Qwen3/Gemma3; widths list tested ranges, with exact grids in Appendix B. All models have one head per layer. Success is IDG at the final evaluation for arithmetic alone and joint success whenever facts are present. The default schedule is 100k steps.

Family	Model/layers	Widths	Arith. frac.	Seeds/width or arm	Steps
Arithmetic	Q/G, 1	4–48	1	10	100k
Arithmetic	Q/G, 2	4–16	1	8	100k
Arithmetic	Q, 3	2–28	1	6 or 16	100k
Facts	Q/G, 1	6, 8	.81/.84	10	100k
1× prose	Q/G, 1	14–20	.534	40	100k
5× prose	Q/G, 1	22–30	.208	20	100k
5×, fixed mix	Q/G, 1	14, 28	.534	20–30	100k
1× prose	Q, 3	10–28	.534	16	100k
Steering	Q/G, 1–3	8–18	1/.534	20–60	100k/200k

A.4 Scoring and evaluation

Evaluation occurs every 500 steps. Arithmetic generation begins before the equals sign, produces at most six new tokens greedily, and scores the first integer parsed after "=". This is a value parser, rather than exact string matching. Facts are generated from "Capital of X is" for at most eight new tokens; a continuation is correct when it begins with the right capital. Prose is not evaluated.

Success is always scored at the final evaluation: all 196 training and all four withheld equations must be correct; where facts are present, at least 48 of 50 must also be recalled. We also report perfect recall separately. Training completion refers to arithmetic fitting at the final evaluation, not to whether a process reached its final scheduled step. Incomplete models remain in the assigned denominator. Their recall is still reported from the measured factual score, even though they necessarily fail success.

The appendix robustness outcome is late-window stable IDG: all training and withheld equations correct in at least 30% of evaluations from 60% of training onward. At 100k steps this is at least 25 of 81 evaluations; at 200k it is at least 49 of 161. Facts are not assessed by this outcome, so its mixed-corpus comparison is with final IDG rather than joint success.

A.5 Answer consistency

Equation 1 uses arithmetic examples selected by dataset metadata, including when the minibatch also contains facts and prose. The units-prediction position immediately precedes the answer's final digit under the actual tokenizer. Group labels combine the operator, token at that position, and following units digit; on this task this is equivalent to operator and signed answer. The residual is measured after the attention contribution is added and before the MLP. Gradients from the auxiliary term propagate through the shared model. If fewer than two arithmetic examples are sampled, the auxiliary loss is zero. The denominator is clamped at 10^{-8}.

For shuffled labels, a fresh permutation is drawn among each minibatch's arithmetic examples at every training step, preserving the multiset of group sizes. The loss form and weight remain the same. The permutation has a separate random generator, so it does not consume the batch sampler's random stream. Neither withheld equation labels nor their representations enter either training loss.

We use $\lambda = .1$ unless stated (.05 for the Qwen3 two-layer and three-layer width-18 experiments, the $5\times$ width-28 experiment and the second withheld pair; .025 for the Qwen3 three-layer width-10 experiment). Where we compared values, we chose λ from training-set outcomes only.

B Capacity and training content: detailed results

B.1 One-layer distributions

Table 4 reports each tested one-layer width under the corpus progression. Its arithmetic-only rows appear in Figure 2 and its mixed-corpus rows in Figure 3; Table 5 pools each mixed-corpus condition over its widths and gives the fixed-fraction populations. Counts refer to assigned model runs, not

distinct seed labels across widths. Success and IDG are final-evaluation outcomes; stable IDG is the separate late-window robustness outcome in Appendix A. Widths absent from the figures were not tested.

B.2 Depth and cross-corpus comparisons

The two-layer sweeps end at width 16, below the one-layer plateau widths (24–48); the three-layer sweep extends to width 28. Table 6 gives arithmetic-only outcomes at both depths and the three-layer fixed-mix sweep. The latter uses seeds 0–15 at every width; the wider steering populations at widths 10 and 18 do not replace those sweep populations.

For two-layer cross-corpus comparisons, Table 7 distinguishes full mixed-corpus counts from the eight shared seeds used for each paired test. Outcomes vary with width and architecture: some narrow mixed-corpus models do not fit, whereas every tested two-layer arithmetic-only model does. At three layers, the shared seeds 0–15 at widths 12/14/16/18/20/24/28 yield 64/112 arithmetic-only IDG successes and 44/112 mixed-corpus joint successes. Every model fits, and every mixed-corpus model recalls at least 48 facts, so its joint and IDG counts coincide. The totals pool model runs across seven widths. Counts agree at widths 12–16 and are lower with mixed content at 18–28, but no individual-width difference reaches nominal significance (Table 8). These comparisons change the arithmetic fraction from 1.0 to .534 and neither isolate prose nor cover the full one-layer progression.

B.3 Width selection

With facts only, we used widths 6 and 8, where arithmetic-only models generalized most often. For each prose condition and architecture, we chose the widths from a one-seed pilot scan: from near the smallest width at which the pilot model fit the training arithmetic and, where it generalized, spanning the widths where it did. With $5\times$ prose it generalized at no scanned width, so those widths follow fitting alone. The pilot seed is among the counted seeds.

B.4 Fixed mix and matched content

The Qwen3 width-14 baseline reuses 29 existing runs, 9 of which were chosen because they had fit the $1\times$ corpus. This is the $5\times$ baseline at fraction .534; its selected seeds are 0, 1, 3, 4, 5, 6, 7, 8, and 10, with the other twenty seeds 11–30. All nine selected seeds belong to the 18 shared seeds in the Qwen3 width comparison. Full-population counts are given in Table 5; component tests on shared seeds are in Table 9. For Gemma3 facts ≥ 48, the paired test gives $p = .0625$, whereas an unpaired Fisher test gives .033. The paired comparison is primary because the widths share data-order seeds and batch streams.

At width 14 and fraction .534, the paired $1\times \to 5\times$ changes are: Qwen3 joint $+7/-6$ ($p = 1$), IDG $+11/-5$ ($p = .21$), all-50 recall $+1/-23$ ($p = 3.0 \times 10^{-6}$), and facts ≥ 48 $+1/-12$ ($p = .0034$), over 29 seeds. Gemma3 gives joint $+9/-8$ ($p = 1$), IDG $+8/-4$ ($p = .39$), all-50 recall $+1/-16$ ($p = .00027$), and facts ≥ 48 $+1/-7$ ($p = .070$), over 30 seeds. Completion is 25/29 versus 23/29 in Qwen3 and 29/30 versus 23/30 in Gemma3. The Qwen3 selection above applies here too.

For one-layer Qwen3 width 14 on seeds 0–29, arithmetic-only success and $1\times$ joint success are both 10/30, with six successes gained and six lost. Completion is 30/30 versus 24/30. This width is near the fitting edge once facts and prose are added, so the comparison mixes completion failures with failure to generalize after fitting; it is weak evidence either way about suppression, and equal counts do not establish equivalence.

B.5 Second withheld pair

Table 10 gives the one-layer Qwen3 arithmetic-only sweep with the $(4, 9)$ equations withheld. It includes fitting and both arithmetic endpoints; the corresponding mixed-corpus steering test is in Appendix C.

Table 4: One-layer results by width. Arithmetic-only success is IDG; mixed-corpus success is joint. Recall columns count models recalling at least 48 or all 50 facts. Completion is training arithmetic 196/196 at the final evaluation. Stable IDG does not assess facts.

						Facts recalled			
Model	Corpus	Width	n	Success	IDG	≥ 48	All 50	Complete	Stable
Qwen3	Arithmetic	4	10	0	0	–	–	0	0
Qwen3	Arithmetic	6	10	10	10	–	–	10	10
Qwen3	Arithmetic	8	10	9	9	–	–	10	9
Qwen3	Arithmetic	10	10	0	0	–	–	10	0
Qwen3	Arithmetic	12	10	4	4	–	–	10	4
Qwen3	Arithmetic	16	10	0	0	–	–	10	0
Qwen3	Arithmetic	24	10	1	1	–	–	10	1
Qwen3	Arithmetic	32	10	0	0	–	–	10	0
Qwen3	Arithmetic	48	10	0	0	–	–	10	0
Qwen3	+ facts	6	10	10	10	10	10	10	10
Qwen3	+ facts	8	10	8	8	10	10	10	8
Qwen3	1×	14	40	13	13	39	39	34	13
Qwen3	1×	16	40	10	10	40	34	39	10
Qwen3	1×	18	40	3	3	40	38	40	3
Qwen3	1×	20	40	4	4	40	39	40	4
Qwen3	5×	22	20	1	1	18	13	18	1
Qwen3	5×	24	20	0	0	18	13	18	0
Qwen3	5×	26	20	1	1	17	17	19	1
Qwen3	5×	28	20	1	1	20	18	19	1
Gemma3	Arithmetic	4	10	1	1	–	–	1	1
Gemma3	Arithmetic	6	10	8	8	–	–	10	8
Gemma3	Arithmetic	8	10	6	6	–	–	10	8
Gemma3	Arithmetic	10	10	4	4	–	–	10	4
Gemma3	Arithmetic	12	10	2	2	–	–	10	2
Gemma3	Arithmetic	16	10	4	4	–	–	10	4
Gemma3	Arithmetic	24	10	0	0	–	–	10	0
Gemma3	Arithmetic	32	10	0	0	–	–	10	0
Gemma3	Arithmetic	48	10	0	0	–	–	10	0
Gemma3	+ facts	6	10	10	10	10	10	10	10
Gemma3	+ facts	8	10	8	8	10	10	10	8
Gemma3	1×	14	40	20	21	39	37	39	22
Gemma3	1×	16	40	15	15	40	40	39	16
Gemma3	1×	18	40	4	4	40	40	40	5
Gemma3	1×	20	40	3	3	40	40	40	4
Gemma3	5×	24	20	3	3	20	19	20	3
Gemma3	5×	26	20	0	0	20	19	20	0
Gemma3	5×	28	20	0	0	20	20	20	0
Gemma3	5×	30	20	1	1	20	20	20	1

Table 5: Success and its components under changes in training content and width. All counts are over assigned model runs at the final evaluation; Facts ≥ 48 and "All 50 facts" count models meeting those recall thresholds. (a) The one-layer progression pools the sampled widths and changes arithmetic fraction. (b) Width varies at fixed $5\times$ content and fraction .534. (c) Content varies at fixed width 14 and fraction .534. In (b,c), the Qwen3 width-14 population contains nine seeds selected for fitting the $1\times$ corpus; all nine are among the 18 shared Qwen3 seeds used for the paired width tests. Gemma3's width test uses 20 shared seeds. Component tests are in Appendix B.4.

(a) Corpus progression

Model	Corpus	Arith. frac.	n	Joint	IDG	Facts ≥ 48	All 50 facts	Complete
Qwen3	+ facts	.814	20	18	18	20	20	20
Qwen3	$1\times$	.534	160	30	30	159	150	153
Qwen3	$5\times$	.208	80	3	3	73	61	74
Gemma3	+ facts	.836	20	18	18	20	20	20
Gemma3	$1\times$	.534	160	42	43	159	157	158
Gemma3	$5\times$	.208	80	4	4	80	78	80

(b) Fixed-mix width comparison

Model	Width	n	Joint	IDG	Facts ≥ 48	All 50 facts	Complete
Qwen3	14	29	11	16	17	6	23
Qwen3	28	20	0	0	20	19	20
Gemma3	14	30	15	19	23	12	23
Gemma3	28	20	1	1	20	20	20

(c) Matched content

Model	Prose	n	Joint	IDG	Facts ≥ 48	All 50 facts	Complete
Qwen3	$1\times$	29	10	10	28	28	25
Qwen3	$5\times$	29	11	16	17	6	23
Gemma3	$1\times$	30	14	15	29	27	29
Gemma3	$5\times$	30	15	19	23	12	23

C STEERING: CONTROLS, ROBUSTNESS, AND LIMITS

C.1 PAIRED COMPARISONS AND MULTIPLICITY

Table 11 gives the discordant pairs underlying Table 1. Each plus denotes success under answer consistency only, and each minus denotes success under its comparator only. The family contains exactly the thirteen three-arm blocks defined in the main table's caption. Three comparisons with shuffled labels remain significant after Holm adjustment across all thirteen.

C.2 LATE-WINDOW ROBUSTNESS

Table 12 uses late-window stable IDG, which does not score facts. Gemma3's arithmetic-only answer counts change little relative to the final evaluation (27 to 27 and 25 to 26), while its baselines increase (19 to 21 and 15 to 20); both baseline comparisons cease to reach nominal significance. Comparing these endpoints does not establish when the benefit emerges.

C.3 COMPLETION-SELECTED SENSITIVITIES AND FITTING TIME

On the 19 Gemma3 two-layer seeds where all three arms complete training, answer versus shuffled labels gives $+10/-2$, $p = .0386$; the all-assigned comparison has $p = .092$. In the second Gemma3 population, all three arms complete training on all twenty seeds, giving $p = .146$. For Qwen3 three-layer width 10, the 28 seeds where all three arms complete give $+9/-3$, $p = .146$, compared with $+11/-3$, $p = .057$ over all thirty assigned seeds.

Table 13 reports the first evaluation reaching full training accuracy on pairs whose two arms are training-complete at the final evaluation. Nonzero delays enter the exact paired sign test. Medians are calculated on the same paired population. No steered-versus-comparator final-completion comparison

Table 6: Per-width depth results. All arithmetic-only models complete training except three-layer width 2. Every three-layer mixed-corpus model completes and recalls at least 48 facts. The late-window column assesses arithmetic only.

Model	Layers	Corpus	Width	n	Success	IDG	Facts ≥ 48	Complete	Stable
Qwen3	2	Arithmetic	4	8	8	8	–	8	8
Qwen3	2	Arithmetic	6	8	8	8	–	8	8
Qwen3	2	Arithmetic	8	8	2	2	–	8	2
Qwen3	2	Arithmetic	10	8	6	6	–	8	7
Qwen3	2	Arithmetic	12	8	4	4	–	8	6
Qwen3	2	Arithmetic	14	8	0	0	–	8	1
Qwen3	2	Arithmetic	16	8	4	4	–	8	4
Gemma3	2	Arithmetic	4	8	5	5	–	8	6
Gemma3	2	Arithmetic	6	8	4	4	–	8	4
Gemma3	2	Arithmetic	8	8	1	1	–	8	1
Gemma3	2	Arithmetic	10	8	3	3	–	8	3
Gemma3	2	Arithmetic	12	8	3	3	–	8	4
Gemma3	2	Arithmetic	14	8	5	5	–	8	5
Gemma3	2	Arithmetic	16	8	3	3	–	8	3
Qwen3	3	Arithmetic	2	6	0	0	–	0	0
Qwen3	3	Arithmetic	4	6	6	6	–	6	6
Qwen3	3	Arithmetic	6	6	6	6	–	6	5
Qwen3	3	Arithmetic	8	6	4	4	–	6	4
Qwen3	3	Arithmetic	10	6	4	4	–	6	4
Qwen3	3	Arithmetic	12	16	9	9	–	16	10
Qwen3	3	Arithmetic	14	16	7	7	–	16	8
Qwen3	3	Arithmetic	16	16	8	8	–	16	7
Qwen3	3	Arithmetic	18	16	12	12	–	16	13
Qwen3	3	Arithmetic	20	16	8	8	–	16	9
Qwen3	3	Arithmetic	24	16	9	9	–	16	9
Qwen3	3	Arithmetic	28	16	11	11	–	16	11
Qwen3	3	1×	10	16	10	10	16	16	10
Qwen3	3	1×	12	16	9	9	16	16	9
Qwen3	3	1×	14	16	7	7	16	16	6
Qwen3	3	1×	16	16	8	8	16	16	9
Qwen3	3	1×	18	16	6	6	16	16	6
Qwen3	3	1×	20	16	2	2	16	16	3
Qwen3	3	1×	24	16	7	7	16	16	7
Qwen3	3	1×	28	16	5	5	16	16	5

Table 7: Two-layer arithmetic-only versus $1\times$ mixed-corpus training. The full-population columns describe every $1\times$ model; the shared-seed columns compare arithmetic-only IDG with $1\times$ joint success on the eight seeds used in both corpora, giving discordant seeds (success with $1\times$ only/arithmetic only) and exact paired p. All arithmetic-only models complete training.

		Full $1\times$ population					Eight shared seeds			
Model	Width	n	Joint	IDG	Facts ≥ 48	Complete	Arith. IDG	$1\times$ joint	$1\times$ vs arith.	p
Qwen3	4	8	0	0	0	0	8	0	$+0/-8$	.0078
Qwen3	6	8	0	0	5	0	8	0	$+0/-8$	.0078
Qwen3	8	8	1	1	7	2	2	1	$+1/-2$	1.00
Qwen3	10	8	5	5	8	7	6	5	$+2/-3$	1.00
Qwen3	12	8	6	6	8	7	4	6	$+3/-1$	.62
Qwen3	14	30	12	12	30	29	0	4	$+4/-0$	.12
Qwen3	16	8	1	1	8	8	4	1	$+0/-3$	.25
Gemma3	8	16	8	9	15	11	1	4	$+4/-1$	.38
Gemma3	10	16	10	10	16	15	3	3	$+1/-1$	1.00
Gemma3	12	16	11	11	16	16	3	5	$+4/-2$	.69
Gemma3	14	40	19	19	40	40	5	4	$+2/-3$	1.00
Gemma3	16	16	3	3	16	16	3	2	$+1/-2$	1.00

Table 8: Three-layer Qwen3, arithmetic-only versus $1\times$ mixed-corpus training on the same 16 seeds at each width. All models complete training; all mixed-corpus models recall at least 48 facts, so their joint and IDG counts coincide. Tests are nominal paired McNemar comparisons.

Width	Shared seeds	Arith. IDG	$1\times$ joint	$1\times$ vs arith.	p
12	16	9	9	$+3/-3$	1.00
14	16	7	7	$+5/-5$	1.00
16	16	8	8	$+4/-4$	1.00
18	16	12	6	$+2/-8$	.11
20	16	8	2	$+1/-7$	.07
24	16	9	7	$+2/-4$	.69
28	16	11	5	$+2/-8$	.11

among the thirteen blocks reaches nominal significance. Eight of the twelve incomplete steered Qwen3 $1\times$ models at 100k end at 160–161/196, a plateau also seen without steering.

C.4 Separate answer-label test

The separate Qwen3 one-layer, width-14 arithmetic-only test uses forty fresh paired seeds, $\lambda = .1$, and answer versus shuffled-label arms without a baseline. All eighty models complete training. Final IDG is 25/40 versus 5/40, with 23 gains, three losses, two successes in both arms, and twelve failures in both. The exact paired p is 8.8×10^{-5}. Late-window stable IDG is 24/40 versus 8/40 ($+20/-4$, $p = .0015$). This population is analyzed separately from the other width-14 populations. It tests training-answer groupings against per-step resampling, without comparing persistent groupings.

C.5 Recall and the larger prose corpus

For the one-layer $1\times$ blocks, facts ≥ 48 counts in baseline/answer/shuffled order are 56/57/55 of 60 for Qwen3 100k, 58/60/57 for Qwen3 200k, and 59/60/60 for Gemma3. All-50 counts are 55/51/46, 53/58/48, and 57/59/58. Minimum steered recall is 46, 49, and 49 respectively; 8, 2, and 1 paired steered models recall fewer facts than their baselines, with no drop exceeding three facts.

With $5\times$ prose at width 14 and fraction .534, the all-assigned Qwen3 counts are joint 11/7/9, IDG 16/11/13, facts ≥ 48 17/21/20, and completion 23/15/19, each out of 29. The population selection in Appendix B.4 applies. The mean paired answer-minus-baseline recall difference is $+0.21$ facts; an exact sign test on 24 nonzero differences gives $p = .31$. Sign-flip inversion bounds the mean difference under a symmetric-shift model, excluding an average cost of at least four facts. It does not exclude smaller costs or establish equivalence, and it concerns recall rather than training completion (Appendix E).

Table 9: Paired width-28 minus width-14 comparisons on the $5\times$ corpus at fraction .534. Tests use 18 shared Qwen3 seeds and 20 Gemma3 seeds. A plus is a success or recall criterion met at width 28 only; a minus is met at width 14 only.

Model	Outcome	Shared seeds	w28 vs w14	p
Qwen3	Joint	18	$+0/-6$	.031
Qwen3	IDG	18	$+0/-9$	.0039
Qwen3	Facts ≥ 48	18	$+7/-0$	.016
Qwen3	All 50 facts	18	$+13/-0$	.00024
Qwen3	Complete	18	$+4/-0$	.12
Gemma3	Joint	20	$+0/-10$	.002
Gemma3	IDG	20	$+0/-12$	.00049
Gemma3	Facts ≥ 48	20	$+5/-0$	.062
Gemma3	All 50 facts	20	$+11/-0$	.00098
Gemma3	Complete	20	$+4/-0$	.12

Table 10: Arithmetic-only Qwen3, one layer, with $(4, 9)$ withheld. Ten assigned seeds per width.

Width	n	Final IDG	Complete	Stable IDG
4	10	0	0	0
6	10	7	9	7
8	10	3	10	3
10	10	1	10	1
12	10	0	10	0
14	10	1	10	2
16	10	2	10	2
18	10	0	10	0

At width 28, fraction .208, $\lambda = .05$, thirty assigned seeds give joint 1/0/0 and completion 29/23/30; every model recalls at least 48 facts. The steered-versus-baseline completion comparison is $+0/-6$, $p = .031$. The larger tested weights .1 and .2 failed the training-completion check. At the arithmetic-only width-48 plateau, twenty seeds per arm yield 0/0/0 IDG in Qwen3 and 0/1/0 in Gemma3, with every model complete.

C.6 Second withheld pair

For the $(4, 9)$ withheld pair, Qwen3 one-layer width 14 with $1\times$ prose and $\lambda = .05$ gives joint success 1/2/1 of twenty assigned seeds. IDG counts coincide, facts ≥ 48 counts are 18/19/18, and completion 19/18/19. A steering benefit was not established on this second pair.

D Internal measurements and controls

D.1 Populations, pair selection, and readout

The one-layer cohort consists of seeds 0–7 in each arm of the width-14, $1\times$, 100k populations, yielding 48 models. Four are training-incomplete: Qwen3 baseline seed 2 (28/196), Qwen3 answer seeds 3 and 6 (161/196 and 155/196), and Gemma3 baseline seed 3 (161/196). They remain in the main analysis. The deeper cohort has 240 models: thirty seeds per arm for Qwen3 two layers, thirty per arm for Qwen3 three layers at width 18, and twenty per arm for Gemma3 two layers. Only one of the two Gemma3 two-layer populations has the saved checkpoints used here. Qwen3 two-layer baseline seed 12 and Gemma3 two-layer answer seed 13 are training-incomplete. The three-layer width-10 steering population is not included in these measurements.

The component swaps use 188 training equations as recipients and 666 ordered donor–recipient pairs per model. Equations are partitioned into three folds; donors come from another fold, share the operator, signed answer, preceding token and units-prediction position, and are neither the recipient nor its operand-reversed partner. Eight training equations lack such a donor and are not swapped. We

Table 11: Paired success comparisons and Holm adjustment. Row numbers identify Table 1. Success is joint whenever facts are present. A, B and S denote answer consistency, baseline and shuffled labels.

Configuration	A vs B	p	A vs S	p	Holm A vs S
1. Qwen3 1L w12	$+9/-7$	.80	$+9/-7$	.80	.80
2. Qwen3 1L w14	$+15/-4$	.019	$+15/-0$	6.1×10^{-5}	.00079
3. Qwen3 1L w14, new seeds	$+13/-4$	.049	$+11/-3$	.057	.46
4. Gemma3 1L w8	$+10/-2$	.039	$+10/-3$	.092	.55
5. Gemma3 1L w8, new seeds	$+11/-1$	.0063	$+10/-3$	.092	.55
6. Qwen3 1L w14	$+25/-11$	.029	$+24/-7$	.0033	.037
7. Qwen3 1L w14	$+33/-2$	3.7×10^{-8}	$+29/-7$	.00031	.0038
8. Gemma3 1L w14	$+23/-9$	.02	$+22/-9$	.029	.27
9. Qwen3 2L w14	$+15/-0$	6.1×10^{-5}	$+11/-2$	.022	.22
10. Gemma3 2L w14	$+9/-2$	.065	$+10/-3$	.092	.55
11. Gemma3 2L w14, new seeds	$+9/-2$	.065	$+9/-3$	.15	.55
12. Qwen3 3L w10	$+6/-2$	.29	$+11/-3$	.057	.46
13. Qwen3 3L w18	$+14/-3$	.013	$+10/-4$	.18	.55

Table 12: Late-window stable IDG on all assigned seeds. B/A/S counts and paired tests assess only arithmetic, including for mixed-corpus blocks. This robustness comparison uses final IDG as its reference.

Configuration	n	Stable IDG B/A/S	$p_{A,B}$	$p_{A,S}$
1. Qwen3 1L w12	30	10/13/12	.61	1.00
2. Qwen3 1L w14	30	9/21/6	.0075	.00027
3. Qwen3 1L w14, new seeds	30	8/16/8	.057	.057
4. Gemma3 1L w8	30	21/27/21	.11	.15
5. Gemma3 1L w8, new seeds	30	20/26/20	.11	.18
6. Qwen3 1L w14	60	23/35/19	.058	.0052
7. Qwen3 1L w14	60	14/44/28	6.9×10^{-8}	.011
8. Gemma3 1L w14	60	34/46/33	.043	.029
9. Qwen3 2L w14	30	13/28/18	6.1×10^{-5}	.0063
10. Gemma3 2L w14	20	11/17/10	.11	.092
11. Gemma3 2L w14, new seeds	20	9/16/13	.065	.51
12. Qwen3 3L w10	30	22/26/21	.29	.27
13. Qwen3 3L w18	30	12/23/19	.0074	.34

average over donors within each recipient, then over recipients, then equally over models. Equation pairs are not treated as independent replications.

At the units-prediction position, A is the residual after the final attention addition and before the MLP. B is the final MLP branch contribution (including Gemma3's branch normalization), and $h = A + B$ enters the final output normalization. A post-attention intervention at dose α reads out

$$h_i(\alpha) = h_i + \alpha(A_j - A_i), \qquad X_{ij}(\alpha) = \ell(h_i(\alpha), y_i) - \ell(h_i, y_i). \tag{3}$$

With the MLP held fixed, the readout applies the final normalization, unembedding and native logit transformation to this state. Loss is conditional on the correct prefix and scored on its units token, even when greedy arithmetic generation would fail. The MLP-output intervention replaces $A_j - A_i$ by $B_j - B_i$. At full dose, $A_i + B_j$ equals the reverse post-attention-swap state, and targets agree for the paired equations; thus the complete-swap measurements are algebraically linked.

D.2 Size matching and random controls

Relative displacement is $r_{ij}(\alpha) = \|\alpha(A_j - A_i)\|_2 / \|h_i\|_2$. For each paired seed and equation pair, we read the comparator's excess loss at the steered model's full-swap relative displacement. The primary interpolation is piecewise linear in r, with measurements at doses (swap scales α) 0, .25, .5 and 1; beyond the measured range we extend the final segment linearly. Sensitivities interpolate in r^2 or retain only pairs within the comparator's measured range. At depth, a further variant uses

Table 13: Time to first reach 196/196 training equations. Rows refer to Table 1; each comparison pairs the named arm with baseline. Times are thousands of training steps. Only pairs with both arms complete at the final evaluation enter; later/earlier/tied counts refer to the named arm. Tests are nominal two-sided sign tests on non-ties.

Row	Arm	Pairs	B median (k)	Arm median (k)	Later/earlier/tie	p
1	answer	30	4.5	6	22/5/3	.0015
1	shuffled	30	4.5	4.5	13/12/5	1.00
2	answer	30	3.5	4.25	19/10/1	.14
2	shuffled	30	3.5	4	13/8/9	.38
3	answer	30	4.5	4	13/13/4	1.00
3	shuffled	30	4.5	4	11/17/2	.34
4	answer	30	7	8.75	18/12/0	.36
4	shuffled	30	7	7.75	16/11/3	.44
5	answer	30	6.75	8.25	15/15/0	1.00
5	shuffled	30	6.75	8.5	15/14/1	1.00
6	answer	43	19.5	24.5	32/10/1	.00094
6	shuffled	48	19.5	19.5	24/21/3	.77
7	answer	54	21.25	29.25	40/14/0	.00054
7	shuffled	58	21.25	24.5	34/23/1	.18
8	answer	57	19.5	23.5	35/19/3	.04
8	shuffled	58	19.75	18.5	29/27/2	.89
9	answer	29	7	7	13/13/3	1.00
9	shuffled	29	7	6	7/21/1	.013
10	answer	19	7	8	14/4/1	.031
10	shuffled	20	7	6	9/11/0	.82
11	answer	20	8	8.75	11/8/1	.65
11	shuffled	20	8	6.5	7/13/0	.26
12	answer	30	10	11.5	19/8/3	.052
12	shuffled	28	9.5	12.5	16/12/0	.57
13	answer	30	2.5	3	13/10/7	.68
13	shuffled	30	2.5	2.5	11/9/10	.82

measured doses 2, 4 and 8. Equation pairs whose comparator full-swap $r \leq 10^{-4}$ are omitted from every matching variant. The one-layer Qwen3 baseline seed 2 has essentially no post-attention spread, so it contributes to unscaled swap summaries but has no defined size-matched contrast or geometry ratio. The remaining one-layer matched denominators are seven against Qwen3 baseline and eight otherwise; deeper contrasts retain thirty Qwen3 and twenty Gemma3 paired seeds.

For each comparison, the full advantage is comparator minus steered excess loss, averaged with the weighting above on the retained pairs. The matched advantage uses the comparator's matched-size loss instead; the reported remainder is the ratio of these two mean advantages, not the mean of the per-seed ratios. Negative remainders mean steered swaps cost more at matched size. Table 14 reports all matching rules. At depth, comparator swaps at the steered models' relative size themselves cost 0.01–0.08 nats, 2–7% of the full-size advantage. Extrapolation can matter: the one-layer Qwen3 baseline contrast reaches 11.2 times a comparator's largest stored displacement on some pairs, which motivates the in-range sensitivity.

Random controls add four isotropic unit directions scaled to the same displacement norm as the swap, with matched draws across arms. Their losses are averaged per pair. Define E as swap excess loss minus random-perturbation excess loss; negative E means the swap is cheaper. Table 15 gives the steered-minus-comparator change in E, at full and matched relative size. A positive change means steering reduced the swaps' advantage over random perturbations. The full-to-matched reduction concerns this between-arm contrast; it is not a test of equivalence between swaps and random perturbations.

D.3 Geometry and shared contraction

Same-answer geometry is the mean Euclidean difference between training equations sharing operator and signed answer, excluding operand-reversed partners. Different-answer comparisons hold operator and the token before the units digit fixed. These measurements use 501 same-answer and 4,220 different-answer unordered pairs per model. The scaled measurement divides the mean difference by the mean norm of $h = A + B$ over the training recipients. For paired models let R_s and R_d be steered/comparator distance ratios for same- and different-answer pairs. The shared fraction of contraction is $(1 - R_d)/(1 - R_s)$. In Table 2, distance ratios are arithmetic means across paired seeds; the shared fraction uses geometric-mean ratios, as a population description. Per-seed fractions can be unstable when same-answer contraction is near zero; negative values indicate expansion of different-answer distances.

At one layer, every defined paired comparison has smaller raw and scaled same-answer differences under steering. Most contraction is shared with different-answer pairs in every such seed. Same-answer pairs tighten more in every Gemma3 seed and in 6/7 and 5/8 Qwen3 seeds against baseline and shuffled labels. At Gemma3 two layers, the aggregate contraction is mostly shared, with more variable per-seed sharing. In Qwen3, less than half is shared at two layers, while different-answer distances increase at three layers, predominantly among negative-answer equations. These are measurements at the objective's own site, where lower within-answer scatter is expected. Architecture, depth, width, and λ vary together across these comparisons.

D.4 Training-complete sensitivities

Table 16 removes the four one-layer training-incomplete models and, for each between-arm contrast, their paired seeds. The lower swap loss, within-group sharing, and within-arm null persist. The Qwen3 size-matched mean advantage is negative in every variant: -28% against baseline and -14% against shuffled labels when all retained pairs may be extrapolated, with five and six paired seeds respectively and one positive seed in each. Keeping only pairs within the measured range yields -2% and -8%. The sign is not tested. The failing-baseline example cannot be repeated in this filtered population because that model is excluded.

Removing the two training-incomplete deeper models leaves the lower swap loss and the size-matched interpretation unchanged.

Table 14: Size-matching results. B/S denote baseline/shuffled comparators, and Seeds counts paired seeds. Matching interpolates linearly in the relative displacement r (primary) or in r^2, extrapolating beyond the measured doses; "in-range pairs" applies the primary rule only to equation pairs inside the comparator's measured range, and "doses up to 8" adds the depth measurements at doses 2, 4 and 8. Full and matched advantages are comparator minus steered excess loss in nats. Remainder is their ratio. Swaps counts retained ordered donor–recipient pairs summed over seeds; Extrapolated counts those beyond the comparator's measured range, which the in-range rule omits.

Population	Comparator	Matching	Seeds	Full (nats)	Matched (nats)	Remainder	Swaps	Extrapolated
Qwen3 1L	B	linear in r	7	3.683	-0.331	-9.0%	4662	844
Qwen3 1L	B	linear in r^2	7	3.683	-0.327	-8.9%	4662	844
Qwen3 1L	B	in-range pairs	7	4.663	0.146	+3.1%	3818	0
Qwen3 1L	S	linear in r	8	3.233	-0.374	-11.6%	5328	792
Qwen3 1L	S	linear in r^2	8	3.233	-0.362	-11.2%	5328	792
Qwen3 1L	S	in-range pairs	8	3.859	-0.169	-4.4%	4536	0
Gemma3 1L	B	linear in r	8	2.778	-0.102	-3.7%	5328	226
Gemma3 1L	B	linear in r^2	8	2.778	-0.113	-4.1%	5328	226
Gemma3 1L	B	in-range pairs	8	2.892	-0.080	-2.8%	5102	0
Gemma3 1L	S	linear in r	8	3.016	-0.079	-2.6%	5328	299
Gemma3 1L	S	linear in r^2	8	3.016	-0.100	-3.3%	5328	299
Gemma3 1L	S	in-range pairs	8	3.169	-0.062	-2.0%	5029	0
Qwen3 2L	B	linear in r	30	1.166	-0.043	-3.7%	19980	1745
Qwen3 2L	B	linear in r^2	30	1.166	-0.041	-3.5%	19980	1745
Qwen3 2L	B	in-range pairs	30	1.293	-0.022	-1.7%	18235	0
Qwen3 2L	B	doses up to 8	30	1.166	0.015	+1.3%	19980	0
Qwen3 2L	S	linear in r	30	0.748	-0.085	-11.4%	19980	2153
Qwen3 2L	S	linear in r^2	30	0.748	-0.085	-11.4%	19980	2153
Qwen3 2L	S	in-range pairs	30	0.873	-0.026	-3.0%	17827	0
Qwen3 2L	S	doses up to 8	30	0.748	-0.056	-7.5%	19980	3
Gemma3 2L	B	linear in r	20	1.509	-0.045	-3.0%	13320	608
Gemma3 2L	B	linear in r^2	20	1.509	-0.052	-3.5%	13320	608
Gemma3 2L	B	in-range pairs	20	1.554	-0.021	-1.4%	12712	0
Gemma3 2L	B	doses up to 8	20	1.509	-0.039	-2.6%	13320	6
Gemma3 2L	S	linear in r	20	1.135	-0.051	-4.5%	13320	361
Gemma3 2L	S	linear in r^2	20	1.135	-0.053	-4.6%	13320	361
Gemma3 2L	S	in-range pairs	20	1.167	-0.038	-3.3%	12959	0
Gemma3 2L	S	doses up to 8	20	1.135	-0.045	-4.0%	13320	0
Qwen3 3L	B	linear in r	30	0.384	-0.015	-3.9%	19980	1835
Qwen3 3L	B	linear in r^2	30	0.384	-0.016	-4.1%	19980	1835
Qwen3 3L	B	in-range pairs	30	0.413	-0.003	-0.8%	18145	0
Qwen3 3L	B	doses up to 8	30	0.384	0.007	+1.9%	19980	4
Qwen3 3L	S	linear in r	30	0.410	-0.006	-1.4%	19980	1817
Qwen3 3L	S	linear in r^2	30	0.410	-0.007	-1.8%	19980	1817
Qwen3 3L	S	in-range pairs	30	0.442	0.003	+0.7%	18163	0
Qwen3 3L	S	doses up to 8	30	0.410	0.007	+1.8%	19980	8

Table 15: Change under steering in the swaps' advantage over equal-displacement random perturbations: steered minus comparator in E = swap loss − random loss, in nats. Matching uses the primary linear-in-r rule. Full-size contrasts retain all assigned paired seeds; one undefined Qwen3 baseline matching contrast is omitted only at matched size.

		Full size		Matched size	
Population	Comparator	Seeds	Contrast	Seeds	Contrast
Qwen3 1L	B	8	+1.923	7	+0.246
Qwen3 1L	S	8	+1.509	8	+0.061
Gemma3 1L	B	8	+2.646	8	+0.207
Gemma3 1L	S	8	+1.100	8	+0.057
Qwen3 2L	B	30	+1.264	30	+0.021
Qwen3 2L	S	30	+1.013	30	+0.012
Gemma3 2L	B	20	+2.069	20	+0.053
Gemma3 2L	S	20	+2.292	20	+0.034
Qwen3 3L	B	30	+0.934	30	-0.007
Qwen3 3L	S	30	+0.869	30	-0.015

Table 16: One-layer size-matching sensitivity after removing training-incomplete models and their paired seeds. Advantages are in nats; B/S denote baseline/shuffled, and matching rules are as in Table 14. The last column is the ratio of mean matched to mean full advantage.

Population	Comparator	Matching	Seeds	Full	Matched	Remainder
Qwen3 1L	B	linear in r	5	2.751	-0.769	-27.9%
Qwen3 1L	B	linear in r^2	5	2.751	-0.760	-27.6%
Qwen3 1L	B	in-range pairs	5	3.834	-0.079	-2.1%
Qwen3 1L	S	linear in r	6	3.440	-0.498	-14.5%
Qwen3 1L	S	linear in r^2	6	3.440	-0.494	-14.4%
Qwen3 1L	S	in-range pairs	6	3.948	-0.313	-7.9%
Gemma3 1L	B	linear in r	7	2.895	-0.143	-4.9%
Gemma3 1L	B	linear in r^2	7	2.895	-0.152	-5.2%
Gemma3 1L	B	in-range pairs	7	3.018	-0.110	-3.6%
Gemma3 1L	S	linear in r	8	3.016	-0.079	-2.6%
Gemma3 1L	S	linear in r^2	8	3.016	-0.100	-3.3%
Gemma3 1L	S	in-range pairs	8	3.169	-0.062	-2.0%

D.5 Within-arm associations and outcome membership

All association analyses retain final-evaluation IDG as their outcome. In the one-layer cohort, no generalizing model recalls fewer than 48 facts; joint-success and IDG group memberships coincide. They also coincide at depth, where every model recalls at least 48 facts.

Within the one-layer architecture–arm cells, generalizing models show no tendency toward lower excess swap loss. An opposite-direction estimate reaches one-sided $p = .022$ for the symmetric swap statistic but depends on the statistic and individual seeds, so no opposite-direction finding is claimed. At depth, the pooled within-arm comparison of same-answer difference divided by state norm yields no detected association (one-sided $p = .071$, threshold .025). The descriptive Gemma3 two-layer training-complete sensitivity is an exception ($p = .013$); it does not establish an overall association direction.

The focused one-layer analysis uses steered models at width 14, $1\times$ prose, 100k steps, seeds 8–59. Among training-complete models, Qwen3 has 30 IDG successes and 12 failures, and Gemma3 39 successes and 12 failures. The pooled area under the comparison curve (AUC, oriented toward smaller distances in successes) is .65, with .75 in Gemma3 and .53 in Qwen3. With all assigned models, the pooled AUC is .52 and no association is detected; Qwen3's .34 also does not reach the two-direction threshold (opposite-direction $p = .026$ against .025). Gemma3 retains an AUC of .75 over 39 successes and 13 failures (descriptive one-sided $p = .0033$).

This association depends on normalization, with no pooled association detected in the unscaled measurements. The final-state norm alone separates Gemma3 outcomes in the opposite direction, while scaling distances by the post-attention norm yields associations in both the steered and shuffled-label arms; specificity to answer labels is therefore not established. Under a joint-success grouping, one steered Qwen3 model and six comparator-arm models in the focused analyses would change membership. The steered Gemma3 result is unaffected.

D.6 PER-SEED SWAP COMPARISONS

Table 17 gives the primary matching comparison for every retained paired seed. The random contrast is steered minus comparator in E at matched size. Individual remainders can have either sign even when the aggregate matched advantage is close to zero.

Table 17: Per-seed post-attention swap comparisons in nats. B/S are the baseline and shuffled-label comparators; matched values use linear interpolation in relative displacement.

Population	Comparator	Seed	Full advantage	Matched advantage	Random contrast
Qwen3 1L	B	0	-0.006	-2.381	+0.785
Qwen3 1L	B	1	7.451	+0.485	-0.409
Qwen3 1L	B	3	0.423	+0.494	-0.138
Qwen3 1L	B	4	3.902	-1.239	+0.137
Qwen3 1L	B	5	2.553	-0.047	+0.042
Qwen3 1L	B	6	11.607	+1.032	+0.701
Qwen3 1L	B	7	-0.146	-0.663	+0.601
Qwen3 1L	S	0	2.014	-1.198	+0.236
Qwen3 1L	S	1	2.682	+0.263	-0.651
Qwen3 1L	S	2	0.749	-0.385	+0.534
Qwen3 1L	S	3	2.228	+0.114	+0.028
Qwen3 1L	S	4	-0.867	-1.449	-0.069
Qwen3 1L	S	5	1.716	-0.174	+0.428
Qwen3 1L	S	6	2.992	-0.115	-0.187
Qwen3 1L	S	7	14.348	-0.045	+0.167
Gemma3 1L	B	0	4.440	-0.250	+0.302
Gemma3 1L	B	1	0.551	-0.241	+0.003
Gemma3 1L	B	2	2.877	-0.044	+0.108
Gemma3 1L	B	3	1.959	+0.186	-0.094
Gemma3 1L	B	4	4.810	-0.121	+0.102
Gemma3 1L	B	5	5.061	-0.150	+0.226
Gemma3 1L	B	6	1.959	+0.021	+0.453
Gemma3 1L	B	7	0.564	-0.214	+0.558
Gemma3 1L	S	0	0.795	-0.309	+0.199
Gemma3 1L	S	1	0.968	-0.358	-0.069
Gemma3 1L	S	2	3.295	+0.002	+0.038
Gemma3 1L	S	3	12.090	+0.282	-0.074
Gemma3 1L	S	4	1.143	-0.043	+0.040
Gemma3 1L	S	5	1.656	+0.050	-0.116
Gemma3 1L	S	6	0.927	-0.239	+0.362
Gemma3 1L	S	7	3.252	-0.016	+0.074
Qwen3 2L	B	0	1.236	+0.024	-0.041
Qwen3 2L	B	1	1.028	+0.021	+0.013
Qwen3 2L	B	2	0.872	+0.011	+0.041
Qwen3 2L	B	3	0.758	+0.002	+0.001
Qwen3 2L	B	4	0.447	-0.001	-0.015
Qwen3 2L	B	5	0.440	-0.031	+0.043
Qwen3 2L	B	6	0.040	+0.000	-0.001
Qwen3 2L	B	7	1.070	+0.002	+0.007
Qwen3 2L	B	8	0.303	-0.005	-0.002
Qwen3 2L	B	9	0.739	-0.189	+0.161
Qwen3 2L	B	10	1.931	-0.003	+0.051
Qwen3 2L	B	11	0.068	-0.001	-0.004
Qwen3 2L	B	12	0.619	+0.117	+0.204
Qwen3 2L	B	13	-0.401	-0.789	-0.073

Population	Comparator	Seed	Full advantage	Matched advantage	Random contrast
Qwen3 2L	B	14	1.773	+0.063	-0.022
Qwen3 2L	B	15	0.150	-0.038	+0.019
Qwen3 2L	B	16	1.946	+0.014	+0.056
Qwen3 2L	B	17	0.826	-0.056	+0.061
Qwen3 2L	B	18	2.405	-0.398	+0.056
Qwen3 2L	B	19	2.864	+0.050	-0.013
Qwen3 2L	B	20	4.695	+0.075	+0.078
Qwen3 2L	B	21	0.200	+0.000	+0.003
Qwen3 2L	B	22	1.136	+0.003	+0.019
Qwen3 2L	B	23	1.400	-0.321	+0.062
Qwen3 2L	B	24	2.921	+0.067	+0.080
Qwen3 2L	B	25	0.573	+0.020	-0.004
Qwen3 2L	B	26	0.590	+0.005	+0.016
Qwen3 2L	B	27	2.613	+0.077	-0.184
Qwen3 2L	B	28	0.784	+0.002	+0.006
Qwen3 2L	B	29	0.943	+0.001	+0.003
Qwen3 2L	S	0	1.611	+0.091	+0.066
Qwen3 2L	S	1	0.892	+0.001	+0.006
Qwen3 2L	S	2	0.716	+0.005	+0.013
Qwen3 2L	S	3	0.202	+0.000	-0.001
Qwen3 2L	S	4	0.063	-0.004	-0.028
Qwen3 2L	S	5	0.466	-0.024	+0.033
Qwen3 2L	S	6	0.412	-0.000	-0.001
Qwen3 2L	S	7	0.324	+0.000	+0.002
Qwen3 2L	S	8	0.790	-0.008	+0.020
Qwen3 2L	S	9	1.173	-0.283	+0.206
Qwen3 2L	S	10	0.796	-0.009	-0.009
Qwen3 2L	S	11	0.155	+0.000	-0.003
Qwen3 2L	S	12	0.350	-0.004	+0.026
Qwen3 2L	S	13	-1.622	-1.616	+0.070
Qwen3 2L	S	14	1.816	+0.031	+0.055
Qwen3 2L	S	15	2.144	-0.042	+0.127
Qwen3 2L	S	16	0.461	+0.000	+0.002
Qwen3 2L	S	17	1.168	-0.042	+0.047
Qwen3 2L	S	18	2.586	-0.259	-0.133
Qwen3 2L	S	19	0.643	-0.010	-0.019
Qwen3 2L	S	20	1.000	+0.005	+0.020
Qwen3 2L	S	21	0.207	+0.000	+0.004
Qwen3 2L	S	22	0.801	-0.001	+0.036
Qwen3 2L	S	23	0.362	-0.209	-0.314
Qwen3 2L	S	24	0.396	-0.098	-0.048
Qwen3 2L	S	25	0.449	+0.013	-0.011
Qwen3 2L	S	26	1.452	+0.019	+0.045
Qwen3 2L	S	27	0.821	-0.118	+0.156
Qwen3 2L	S	28	0.523	+0.001	-0.002
Qwen3 2L	S	29	1.295	+0.002	+0.004
Gemma3 2L	B	0	3.143	+0.000	+0.065
Gemma3 2L	B	1	0.379	-0.038	+0.086
Gemma3 2L	B	2	0.138	-0.004	-0.013
Gemma3 2L	B	3	0.547	+0.001	+0.005
Gemma3 2L	B	4	6.526	+0.165	+0.178
Gemma3 2L	B	5	1.583	-0.284	+0.278
Gemma3 2L	B	6	2.439	-0.018	+0.069
Gemma3 2L	B	7	-0.058	-0.190	-0.194
Gemma3 2L	B	8	3.877	-0.031	+0.061
Gemma3 2L	B	9	0.561	-0.000	+0.024
Gemma3 2L	B	10	1.097	-0.037	+0.062
Gemma3 2L	B	11	0.593	+0.002	+0.003
Gemma3 2L	B	12	0.488	+0.003	+0.017
Gemma3 2L	B	13	0.357	+0.003	-0.001
Gemma3 2L	B	14	2.920	-0.090	+0.150
Gemma3 2L	B	15	2.584	+0.017	+0.014
Gemma3 2L	B	16	0.672	+0.003	-0.004

Population	Comparator	Seed	Full advantage	Matched advantage	Random contrast
Gemma3 2L	B	17	0.843	-0.401	+0.249
Gemma3 2L	B	18	0.536	+0.001	+0.000
Gemma3 2L	B	19	0.961	+0.001	+0.004
Gemma3 2L	S	0	1.029	+0.005	+0.040
Gemma3 2L	S	1	1.206	-0.046	+0.078
Gemma3 2L	S	2	0.423	-0.004	-0.014
Gemma3 2L	S	3	1.235	+0.002	+0.011
Gemma3 2L	S	4	0.444	-0.021	+0.132
Gemma3 2L	S	5	0.168	-0.315	+0.337
Gemma3 2L	S	6	3.020	-0.011	+0.075
Gemma3 2L	S	7	0.505	-0.144	-0.147
Gemma3 2L	S	8	0.865	-0.041	+0.024
Gemma3 2L	S	9	0.508	-0.002	+0.009
Gemma3 2L	S	10	0.153	-0.042	+0.048
Gemma3 2L	S	11	1.521	+0.004	+0.013
Gemma3 2L	S	12	1.275	+0.009	+0.014
Gemma3 2L	S	13	0.104	+0.000	-0.000
Gemma3 2L	S	14	0.736	-0.159	+0.161
Gemma3 2L	S	15	1.212	+0.003	+0.013
Gemma3 2L	S	16	1.788	+0.004	-0.004
Gemma3 2L	S	17	0.203	-0.267	-0.184
Gemma3 2L	S	18	1.360	+0.001	+0.027
Gemma3 2L	S	19	4.953	+0.008	+0.049
Qwen3 3L	B	0	0.814	+0.002	+0.027
Qwen3 3L	B	1	2.113	+0.000	+0.002
Qwen3 3L	B	2	1.062	+0.019	+0.006
Qwen3 3L	B	3	0.169	-0.001	+0.008
Qwen3 3L	B	4	0.099	-0.000	+0.002
Qwen3 3L	B	5	0.303	-0.000	+0.000
Qwen3 3L	B	6	0.110	-0.000	+0.002
Qwen3 3L	B	7	0.026	-0.000	+0.003
Qwen3 3L	B	8	1.566	+0.000	+0.009
Qwen3 3L	B	9	-0.099	-0.133	-0.098
Qwen3 3L	B	10	-0.036	-0.073	-0.021
Qwen3 3L	B	11	0.688	+0.033	-0.018
Qwen3 3L	B	12	0.118	-0.105	+0.063
Qwen3 3L	B	13	0.753	+0.001	+0.000
Qwen3 3L	B	14	0.287	-0.004	-0.039
Qwen3 3L	B	15	0.008	+0.002	-0.016
Qwen3 3L	B	16	0.124	+0.001	+0.003
Qwen3 3L	B	17	0.451	+0.077	-0.089
Qwen3 3L	B	18	0.179	-0.007	+0.020
Qwen3 3L	B	19	0.636	+0.000	-0.000
Qwen3 3L	B	20	0.019	-0.000	-0.000
Qwen3 3L	B	21	0.816	+0.000	+0.004
Qwen3 3L	B	22	-0.157	-0.175	-0.117
Qwen3 3L	B	23	0.212	-0.005	+0.008
Qwen3 3L	B	24	0.041	+0.000	-0.033
Qwen3 3L	B	25	0.274	-0.072	+0.055
Qwen3 3L	B	26	0.625	+0.000	+0.004
Qwen3 3L	B	27	0.167	+0.000	+0.004
Qwen3 3L	B	28	0.052	-0.000	+0.000
Qwen3 3L	B	29	0.096	-0.012	+0.008
Qwen3 3L	S	0	0.711	+0.001	+0.014
Qwen3 3L	S	1	0.076	-0.000	+0.000
Qwen3 3L	S	2	0.623	+0.000	+0.044
Qwen3 3L	S	3	0.063	+0.013	-0.007
Qwen3 3L	S	4	0.224	-0.000	+0.002
Qwen3 3L	S	5	0.316	-0.000	+0.007
Qwen3 3L	S	6	0.222	+0.000	+0.000
Qwen3 3L	S	7	0.052	+0.000	+0.000
Qwen3 3L	S	8	0.324	+0.000	+0.000
Qwen3 3L	S	9	0.754	-0.061	-0.107

Population	Comparator	Seed	Full advantage	Matched advantage	Random contrast
Qwen3 3L	S	10	0.537	-0.001	+0.188
Qwen3 3L	S	11	0.123	-0.000	-0.005
Qwen3 3L	S	12	1.025	-0.070	-0.014
Qwen3 3L	S	13	0.212	-0.000	+0.003
Qwen3 3L	S	14	0.413	-0.002	-0.083
Qwen3 3L	S	15	0.153	-0.001	-0.096
Qwen3 3L	S	16	0.333	-0.000	+0.007
Qwen3 3L	S	17	1.357	+0.134	-0.228
Qwen3 3L	S	18	0.380	+0.024	-0.026
Qwen3 3L	S	19	0.032	-0.001	-0.004
Qwen3 3L	S	20	0.356	+0.000	+0.001
Qwen3 3L	S	21	0.206	+0.000	+0.001
Qwen3 3L	S	22	-0.031	-0.149	-0.198
Qwen3 3L	S	23	1.240	+0.009	+0.002
Qwen3 3L	S	24	0.446	+0.011	-0.041
Qwen3 3L	S	25	0.510	-0.059	+0.065
Qwen3 3L	S	26	0.346	+0.000	+0.000
Qwen3 3L	S	27	0.147	+0.000	+0.001
Qwen3 3L	S	28	0.283	-0.000	+0.006
Qwen3 3L	S	29	0.855	-0.014	+0.006

E STATISTICAL METHODS AND POPULATIONS

Binary outcomes. Let b count paired seeds successful under answer consistency only and c count those successful under its comparator only. Exact two-sided McNemar tests condition on $b + c$ and use

$$p = \min\left\{1,\ 2\sum_{j=0}^{\min(b,c)}\binom{b+c}{j}2^{-(b+c)}\right\}, \tag{4}$$

with $p = 1$ when there are no discordant pairs. Counts include all assigned seeds, and success requires final training completion. Recall outcomes use the observed facts scores even for arithmetic-incomplete models. Tests of completion, facts ≥ 48, and perfect recall preserve pairing in the same way. Across widths, pairs share data-order seeds and batch streams, not identically sized initial parameter tensors. Paired tests use only the shared-seed subset.

Effect sizes and Figure 1. The paired difference in success frequency is $(b - c)/n$, where n includes concordant as well as discordant pairs. Figure 1 gives a 95% interval for this difference by inverting the exact McNemar test. With the $d = b + c$ discordant seeds held fixed, b is binomial with probability π that a discordant seed favors answer consistency, and the difference is $d(2\pi - 1)/n$; mapping the Clopper–Pearson interval for π gives the interval. It therefore excludes zero exactly when $p < .05$. Because d is held fixed, the interval reflects how the discordant seeds split, not how many are discordant; when $c = 0$, the estimate lies at its upper end. Intervals are pointwise and not multiplicity-adjusted, and no interval treats different blocks as independent confirmations.

Multiplicity. All reported tests are nominal unless explicitly adjusted. For the thirteen answer-versus-shuffled comparisons in Table 1, Holm adjustment sorts the p-values and compares the ith to $.05/(14 - i)$, equivalently reporting the running maximum of $(14 - i)p_{(i)}$ capped at one. The separate two-arm labels test, the $5\times$ tests, the width-48 plateau tests, and the test on the second withheld pair are outside that family. The separate labels test is not pooled with other populations. For deeper swaps, Holm adjustment is applied separately over six comparator contrasts for full-size loss and for the random-perturbation contrast. These are different families from the behavioral tally.

Fitting time. Time to fit is the first evaluation with 196/196 training equations correct. Comparisons use only seeds where both named arms are training-complete at the final evaluation; the table reports this denominator, both medians, and later/earlier/tied counts. Exact two-sided sign tests discard ties from the test statistic. A shortfall in final completion alone is not interpreted as delayed fitting.

Recall differences and bounds. For $5\times$ width 14, factual recall is available for all 29 paired seeds, including arithmetic-incomplete models. The sign test uses the 24 nonzero answer-minus-baseline differences. The mean-difference bound is a separate inversion of sign-flip tests under a symmetric location-shift model for paired differences. Exact enumeration is used when feasible, otherwise fixed-draw Monte Carlo; this assumption and the small paired population limit the interpretation of the bound. It excludes an average loss of at least four facts under that model, not every possible recall cost. A small number of severe failures could be missed by an aggregate comparison.

Internal measurements and associations. Between-arm internal contrasts use paired seeds, with equation pairs treated as nested measurements within models. At depth, exact two-sided sign tests compare per-seed loss differences. Association tests compare successful and unsuccessful models within architecture–arm cells and permute outcome labels within each cell, preserving the number of successes. AUC gives half credit to ties and is oriented toward smaller measurements in successful models; the pooled statistics preserve their defined strata. Each of the two directions is assessed against .025. Architecture-specific focused tests and completion-selected checks are descriptive sensitivities. The association endpoint is final IDG, including when the behavioral steering outcome is joint success.

Population accounting. The one-layer corpus progression pools runs across widths, so its denominators are model runs rather than distinct random seeds. The three-layer fixed-mix sweep uses the same sixteen seeds at every width, even where larger steering populations exist. Width-14 Qwen3 comparisons involving the $5\times$ corpus retain their explicitly identified completion-selected seeds. Late-window stable IDG is an arithmetic-only robustness check and does not replace the final-evaluation success rule.