SPARSE-CIRCUIT RECOVERABILITY ACROSS GROKKING AND DECISION-MATCHED REALISATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does the circuit-recovery advantage associated with grokking depend on a network's original weights, or does it recur in a new realisation of its decisions? We test this by combining longitudinal analysis of behaviourally sufficient component masks with exact, full-domain decision matching. In a five-seed modular-addition study, greedy recovery at 99% decision agreement retains 510–515 of 516 components at delayed checkpoints, compared with 65–146 after stable generalisation. Targeted non-monotone and distribution-sensitive follow-ups retain a large continuous gap. We then train independently initialised students to reproduce each teacher's complete decision table, including its errors, without constraining logits or internal representations. Under greedy recovery against each model's own centred logits, the stable-minus-pre contrast is negative for teachers and student-cell medians in all fourteen paired addition seeds and, descriptively, all eight paired multiplication seeds. Mean reductions are 21.3 and 12.0 percentage points for addition teachers and students, respectively, and 8.8 points for both on multiplication. Absolute endpoints vary across decision-matched realisations, while the executed hard-concrete policy returns predominantly near-intact masks. Under the tested retraining and greedy-recovery procedures, the phase ordering therefore recurs without the original teacher weights, while its magnitude and absolute endpoints change.

1 INTRODUCTION

A circuit measurement along a training trajectory leaves an important question unresolved: does the observed regularity require the network's particular weights and learning history, or does it recur in a new network with the same decisions? This distinction matters when circuit analyses are used to characterise what a model has learned. A smaller sufficient subnetwork may reflect a reusable property of the learned behaviour, a feature of one implementation, or the interaction of that implementation with a recovery procedure. Following one set of weights cannot distinguish persistence beyond that set of weights. Here, a recovered circuit is a behaviourally sufficient component mask under zero ablation of head outputs and post-ReLU neuron activations, without retraining. The question is the repeatability of a recovery measurement, not whether networks share an internal algorithm.

Grokking provides a controlled setting for this question. Networks can fit a small algorithmic training set long before generalising to held-out inputs (Power et al., 2022). Mechanistic analyses connect this process to the development and cleanup of generalising circuits (Nanda et al., 2023), while compression and lottery-ticket studies associate generalisation with simpler representations and useful subnetworks (Merrill et al., 2023; Liu et al., 2023; Minegishi et al., 2025; DeMoss et al., 2025). These findings motivate a stronger comparison than another before–after pruning curve: replace the trained realisation while holding its decision behaviour exactly fixed.

Our study combines longitudinal checkpoint analysis with 241 exactly decision-matched students across two fully enumerable arithmetic tasks and five transformer configurations, evaluated using multiple fidelity levels and complementary discovery procedures. At selected teacher checkpoints, we enumerate the entire decision table and train fresh students to reproduce it. Each eligible student reproduces *every* teacher decision, including its errors, across the complete input domain—not merely its aggregate accuracy. We apply the same recovery protocol separately to each teacher

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.

and student, measuring fidelity against that model's own dense outputs. Students do not inherit the teacher's weights, masks or optimiser state: this is a decision-matched reimplementation, not further training of a pruned teacher.

The results reveal **persistence of the phase contrast together with variation in its magnitude and endpoints**. First, a five-seed longitudinal study establishes a large recovery gap across generalisation, reproduced by an independent search permitting additions and swaps and by distribution-sensitive evaluation. Second, the greedy phase contrast recurs under own-dense centred-logit fidelity in all fourteen paired addition seeds and all eight paired multiplication seeds after decision-matched reimplementation. Third, matching decisions does not reproduce the same recovered size: the addition contrast is attenuated in students, stable student endpoints are more concentrated across teacher groups, and different students of one teacher can still yield different endpoints. Selected changes in student depth, head count and MLP width preserve the direction of the contrast.

Together, the studies distinguish transfer of a recovery ordering from invariance of recovered size. Earlier and stable teachers compute different decision tables; we carry each table into fresh weights and test whether the ordering recurs. Exact decision matching preserves a stronger behavioural object than task accuracy, while leaving logits and representations free. The ordering survives this replacement, but the numerical recovery endpoints change. We report absolute endpoints and achieved fidelity for both discovery methods.

2 Related work

Grokking, compression and sparse subnetworks. Nanda et al. (2023) analyse memorisation, circuit formation and cleanup in modular addition. On sparse parity, Merrill et al. (2023) connect grokking to a sparse subnetwork becoming dominant over a poorly generalising dense subnetwork. Liu et al. (2023) and DeMoss et al. (2025) study associated complexity changes. Minegishi et al. (2025) connect grokking to lottery tickets: subnetworks from grokked models accelerate generalisation, and pruning can expose generalising structure in memorising networks. Sparse-subnetwork emergence is therefore established prior work, not the novelty claimed here. Unlike testing whether a pruned network can learn or perform the arithmetic task (Frankle & Carbin, 2019), we first reconstruct complete phase-specific decision tables in fresh dense networks, then test recurrence of the recovery contrast.

Distillation and matching. Knowledge distillation transfers teacher behaviour to students (Hinton et al., 2015); in grokking, Singh et al. (2025) use teacher transfer to support generalisation below a critical data threshold. Closer to our interpretability question, Haskins & Adams (2025) examine internal restructuring and component alignment in teacher–student language-model pairs. Wadhwa et al. (2025) instead train students with an objective aligning corresponding internal circuit representations. Our experiment imposes no internal alignment: it reconstructs every entry of each phase-specific teacher's decision table, including errors, and evaluates recovered masks against each realisation's own logits. It tests recurrence of a phase-indexed recovery ordering under exact decision equality, rather than internal alignment or improved held-out task performance. The complete-domain matching condition leaves no unobserved input on which an eligible student's decision can differ from its teacher's.

Circuit stability, multiplicity and evaluation. Zhong et al. (2023) show that modular-addition networks can implement different algorithms, including multiple solutions in parallel. Structural mask differences alone therefore do not identify algorithmic multiplicity. Tigges et al. (2024) find consistent high-level circuit analyses across language-model training and scale, despite changes in individual components. Sun (2025) formalises circuit stability across input subtasks. Schiffman (2026) extracts recurring low-dimensional activation subspaces across independently trained transformers, including modular-addition cores emerging at grokking. That work targets shared algorithmic structure; we test a phase-indexed recovery statistic after exact, error-preserving decision-table reconstruction. Subspace dimension and retained head/neuron count concern different intervention families. Our comparison imposes exact decision equality within teacher–student groups. Automated recovery (Conmy et al., 2023) also requires an intervention and a faithfulness criterion; Miller et al. (2024) demonstrate sensitivity to these choices. Mahale (2026) reports instability of circuit-derived claims under analytic variation on GPT-2 small and indirect object identification. Méloux

et al. (2026) further distinguish input-sampling instability from methodological sensitivity in circuit estimation. Complete-domain evaluation removes input-sampling uncertainty for each of our fixed masks, while teacher seeds and student realisations remain distinct sources of variation. These observations inform our robustness checks. We ask which aspects of a learning-phase contrast survive controlled replacement of the original weights.

3 EXPERIMENTAL DESIGN

3.1 MODELS, INTERVENTIONS AND FIDELITY

The canonical model is a one-layer causal transformer with residual width 128, four attention heads of width 32 and a 512-neuron ReLU MLP. The intervention universe has $N = 516$ components: four whole heads and 512 neurons. A binary mask zeros head outputs and post-ReLU neuron activations without retraining. Embeddings, unembedding, biases and other operations remain active. Retained proportion counts this declared basis, not parameters or FLOPs.

For logits $z_\theta(x)$ and masked logits $z_{\theta,m}(x)$, let $\widetilde{z}$ subtract the mean over classes separately for each input. Phase I's primary fidelity is complete-domain decision agreement,

$$F_{\mathrm{dec}}(m;\theta) = |\mathcal{X}|^{-1} \sum_{x\in\mathcal{X}} \mathbf{1}\{\arg\max z_{\theta,m}(x) = \arg\max z_\theta(x)\}, \tag{1}$$

with threshold 0.99. A distribution-sensitive follow-up constrains mean Jensen–Shannon divergence between masked and dense softmax outputs to at most 0.0005. Phase II uses centred-logit fidelity,

$$F_{\mathrm{logit}}(m;\theta) = 1 - \frac{\sum_x \|\widetilde{z}_{\theta,m}(x) - \widetilde{z}_\theta(x)\|_2^2}{\sum_x \|\widetilde{z}_\theta(x)\|_2^2}, \qquad F_{\mathrm{logit}} \geq 0.99. \tag{2}$$

Class centring removes additive logit shifts. Exact qualification uses ordered float64 accumulation and rejects nonfinite scores or a zero denominator. Each student is evaluated against *its own* dense logits. The two numerical 0.99 thresholds preserve different quantities and are not directly comparable.

Our endpoint is the smallest retained proportion among the procedure's evaluated qualifying masks, with the intact mask allowed. We call this *recovered size*: an achieved upper bound on the smallest sufficient mask in the specified intervention universe. This definition applies throughout. Phase I additionally uses the original half-model convention, at most 258 components, when reporting sparse families.

3.2 LONGITUDINAL ANALYSIS AND DECISION-MATCHED REIMPLEMENTATION

Phase I: a common checkpoint grid. Five addition-modulo-113 teachers share 3,830 training and 8,939 test input pairs. Full-batch AdamW training uses learning rate 0.001, weight decay 1 and 40,000 updates. Seven common checkpoints give 35 positions: nine delayed, eleven transitional and fifteen stable. Delayed positions precede the first saved checkpoint above 10% test accuracy; stable onset is the fifth checkpoint in the first five-consecutive-checkpoint sequence at or above 99%. The remaining positions are transitional. Four seeds reach stable onset in the analysed grid; seed 3 does not reach it within this window and is retained. Independent non-monotone searches use ten streams per position under each of decision and JS fidelity.

Phase II: complete decision tables. We train fifteen main teachers for addition modulo 113 and fifteen for multiplication modulo 59, whose complete domains contain 12,769 and 3,481 ordered pairs. The pre checkpoint is the latest saved point with training accuracy at least 99.9% and test accuracy at most 5%, before the first point reaching 10% test accuracy. Stable selection uses the sustained-99% rule. Addition also includes the intervening checkpoint closest to 50% test accuracy. Selection uses accuracy, not circuit outcomes, and unavailable phases remain unavailable.

For each available teacher/phase, independently initialised hard students minimise cross-entropy against teacher argmax labels over the entire domain. Eligibility requires exact agreement on every input. Addition has three eligible canonical students per cell, multiplication two. There are 43 available addition roles and twenty multiplication roles, giving 129 and forty canonical hard students.

Another 72 eligible students belong to four assigned architecture panels. The total of 241 students is nested within the thirty main teacher runs. Pre–stable pairing yields fourteen addition and eight multiplication teacher seeds (Appendix A). Student phase labels always refer to the source teacher: no claim about grokking during student training is required.

3.3 Recovery procedures and analysis units

Greedy search starts intact, ranks component deletions using full-domain gradients, and accepts exactly evaluated qualifying deletions. Phase I also uses simulated annealing with additions and swaps followed by deterministic deletion cleanup. Phase II compares greedy with stochastic hard-concrete gates (Louizos et al., 2018); model weights stay fixed in both. Hard-concrete uses four restarts of 5,000 Adam steps, an absolute centred-logit MSE objective plus 0.001 times mean expected L_0, and binary proposal extraction. The final qualification machinery is common. Greedy's canonical native allowance is 8,256 candidate requests; the final ledger allowance is 256. These are different budgets, and a shared final allowance does not mean equal compute. Appendix C specifies the executed gates, optimiser, fitting batches and extraction grids.

The population unit is the teacher seed. For students, we first take the median within each teacher/phase cell, then calculate stable-minus-pre differences within teacher and average across complete pairs. Addition inference uses the saved sign-flip reference test, 10,000 teacher-level bootstrap resamples and a four-comparison Holm family. Multiplication has eight pairs, below the previously declared minimum of ten for inferential release, so it is descriptive. Checkpoints, students, search streams and threshold views are not counted as independent teachers. Dispersion, transition and stopping-time summaries are descriptive reanalyses of saved records. The additional teacher–student attenuation interval is exploratory and separate from the original four-comparison family.

4 Results

4.1 Generalisation accompanies a large recovery gap

Search recovers much smaller decision-preserving masks at stable checkpoints than at earlier checkpoints. Figure 1 places recorded training and held-out accuracy beside recovered size for all five seeds. At 99% decision agreement, delayed and transitional positions each have median size 514/516, whereas stable positions have median 82. The respective ranges are 510–515, 297–515 and 65–146. Every stable position, and no earlier position, falls below the half-model convention. This is a large continuous change, not a small crossing of an arbitrary size boundary. These 35 positions are repeated observations within five teacher seeds; four reach stable onset in the circuit-evaluation window (all seed-specific endpoints: Table 6). Search streams are repeated recovery attempts, not additional trained models.

The separation survives a search that can undo deletions. The annealing and JS studies are targeted follow-ups designed after the original pattern was observed (Appendix B). Non-monotone annealing has phase medians 514, 513 and 71 components, with all 150 stable streams below half the model and all 200 earlier streams above it. Stable streams improve on the corresponding greedy endpoint in 137/150 cases. Under the distribution-sensitive JS criterion, the stable median rises to 276.5, but its largest endpoint (368) remains below the smallest earlier endpoint (432). Thus the continuous recovery gap survives distribution-sensitive fidelity, while the universal half-model label does not (Table 1).

Table 1: **The longitudinal gap survives changes in search and fidelity.** Entries are median retained components (range). Greedy has 9/11/15 delayed/transition/stable positions; each annealing row has ten streams per position. Streams are repeated searches, not additional teacher seeds.

Search / fidelity	Delayed	Transition	Stable
Greedy / decisions $\geq .99$	514 (510–515)	514 (297–515)	82 (65–146)
Annealing / decisions $\geq .99$	514 (508–515)	513 (264–515)	71 (58–136)
Annealing / JS $\leq .0005$	513 (509–515)	511 (432–515)	276.5 (209–368)

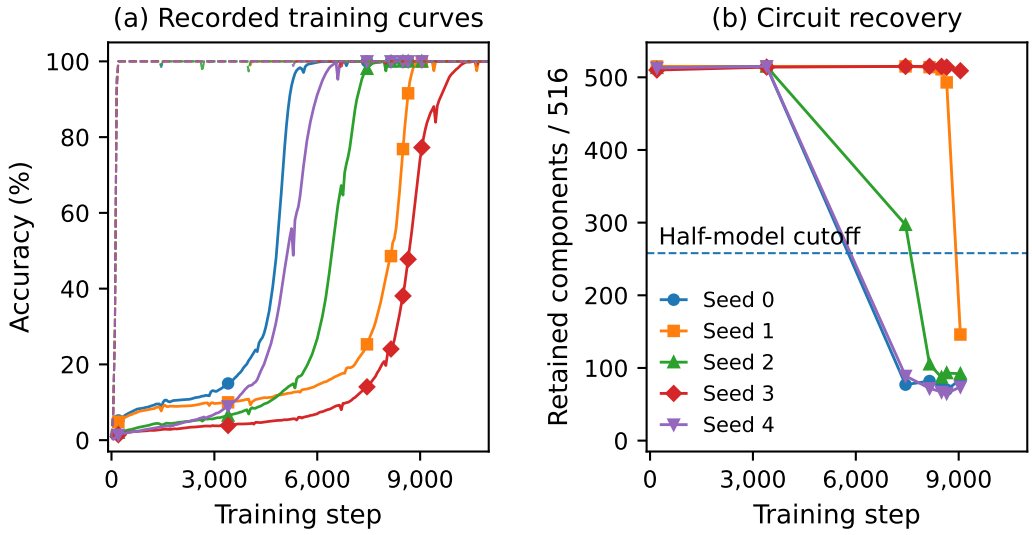

Figure 1: **Generalisation and circuit recovery along the same five trajectories.** Left: recorded accuracy every 50 updates (solid: test; dashed: train); markers identify the seven circuit-evaluation steps. Right: greedy retained size at full-domain decision agreement ≥ 0.99, measured only at those checkpoints. Seed encodings match across panels. Accuracy continues to step 11,000; circuit evaluation ends at 9,050, before seed 3 reaches stable generalisation. Connecting recovery markers does not add intermediate measurements. Full 40,000-step curves appear in Appendix B.

Recovered masks are also behaviourally specific. None of 4,200 size-matched or head/MLP-stratified random masks reaches 99% decision agreement. Independently generated annealing pools contain ten mutually qualifying masks at every stable position under the original overlap cap, without diversity-guided search. Appendix G reports overlap, shared-core and union interventions; these characterise structural alternatives without presuming distinct algorithms. The next experiment asks whether the recovery contrast itself extends beyond the original weights.

4.2 The greedy phase contrast recurs after decision-matched reimplementation

The stable-phase advantage recurs in fresh networks under greedy own-dense centred-logit recovery. Every one of the fourteen paired addition seeds has a negative stable-minus-pre contrast for both the teacher and its student-cell median. Mean reductions are 21.332 pp for teachers and 11.974 pp for students, with bootstrap intervals $[-26.204, -16.653]$ and $[-12.805, -11.143]$ pp. Both Holm-adjusted p-values are 0.0004883. Figure 2 shows every pair; Table 9 gives the numerical values.

The second task reproduces the direction. All eight paired multiplication seeds have negative greedy contrasts in both conditions, with mean changes -8.818 pp for teachers and -8.842 pp for students. Effect sizes vary: the seed-10 teacher decreases by only 0.78 pp, compared with 12.31 pp for its student-cell median. This comparison is conditional on eight of fifteen teacher initialisations having both required phase roles. On addition, the student reduction is smaller in twelve of fourteen seeds: mean attenuation $+9.36$ pp, with an exploratory 95% teacher-paired bootstrap interval $[4.93, 14.06]$ pp (Appendix D.1). On multiplication, seed-specific teacher–student differences have both signs. The greedy ordering thus recurs more consistently than the magnitude.

The same direction persists across acceptance thresholds applied to the saved candidate pools. All fourteen addition teacher and student contrasts remain negative from fidelity 0.90 through 0.995. Multiplication students retain eight negative contrasts throughout; multiplication teachers have seven negative and one positive at 0.995, versus eight negatives at lower thresholds. These are sensitivity views of existing candidate sets, not separately retuned searches (Appendix F).

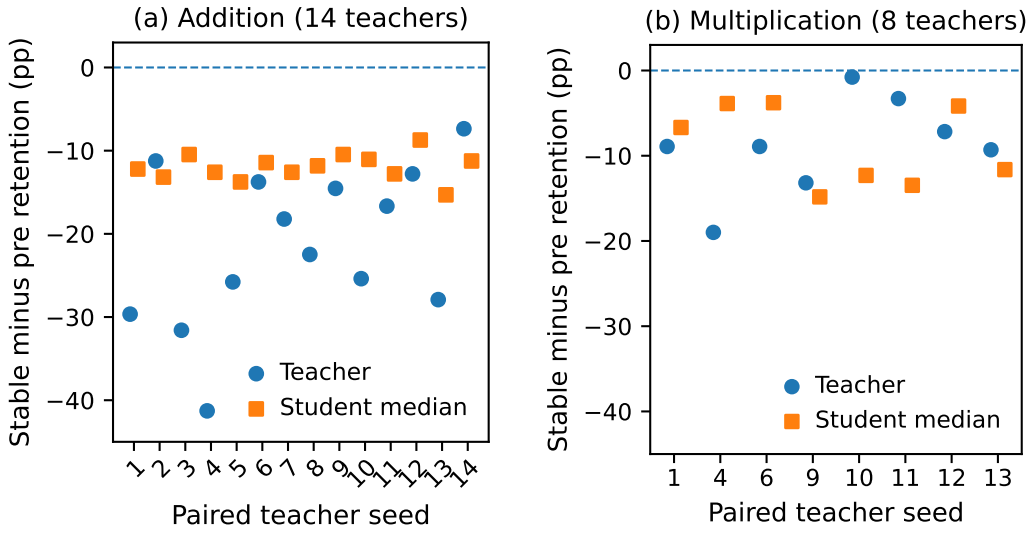

Figure 2: **The greedy phase contrast transfers on both tasks.** Stable-minus-pre retained proportions at own-dense centred-logit fidelity ≥ 0.99. All fourteen addition and all eight multiplication teacher pairs have negative changes for teachers and student-cell medians. Both panels use the same vertical scale. Multiplication is descriptive under the original minimum-sample rule. Exact decision matching holds within each teacher/phase group, not between the pre and stable phases.

Transition checkpoints show transfer of more than a two-endpoint difference. Addition teachers decrease by 7.849 pp from pre to transition and another 13.483 pp from transition to stable. Student-cell medians decrease by 4.790 and 7.184 pp, respectively. Thirteen of fourteen teacher trajectories and all fourteen student trajectories are strictly decreasing over the three phases. Figure 3a shows these absolute levels. Because students are separately trained to each target table, this three-point ordering does not require retaining the teacher's trajectory through weight space.

4.3 Decision-matched realisations share ordering, not the same endpoints

Stable addition students occupy a much narrower recovery range across teacher groups, even though students of one teacher still differ. Among the fourteen matched addition teachers, stable retained proportions range from 57.36% to 89.34%; the corresponding student-cell medians range from 83.53% to 88.76%. The across-teacher standard deviation is 9.11 pp for teachers and 1.32 pp for student medians (Figure 3b). Thus the smaller student phase contrast accompanies concentration at relatively high retained proportions, not universally improved compression.

Taking three-student medians could mechanically reduce dispersion. We therefore also consider uniform selection of *one* saved student per teacher, avoiding within-cell aggregation. The square root of the exact expected across-teacher sample variance is 2.63 pp, still well below the teachers' 9.11 pp; Appendix E gives the calculation and all phase summaries. This descriptive concentration is task-specific. On multiplication, stable teacher SD is 4.90 pp, student-median SD is 4.39 pp, and the one-student control is 5.28 pp. The evidence does not support a general claim that distillation reduces recovery dispersion.

Across-teacher concentration and within-teacher variation are different observations. The median range across individual addition students within a teacher/phase cell grows from 0.58 pp pre-generalisation to 5.62 pp at stable checkpoints; multiplication grows from 0.78 to 5.04 pp. Different eligible students therefore need not share an endpoint even when every decision agrees. The phase pattern recurs under decision-matched retraining while the recovered endpoint varies.

The contrast also survives selected changes in student architecture. On the same four matched addition anchor teachers, all four alternative architectures have negative greedy phase changes in every pair (Table 2). These conditional panels extend the result beyond the canonical student width

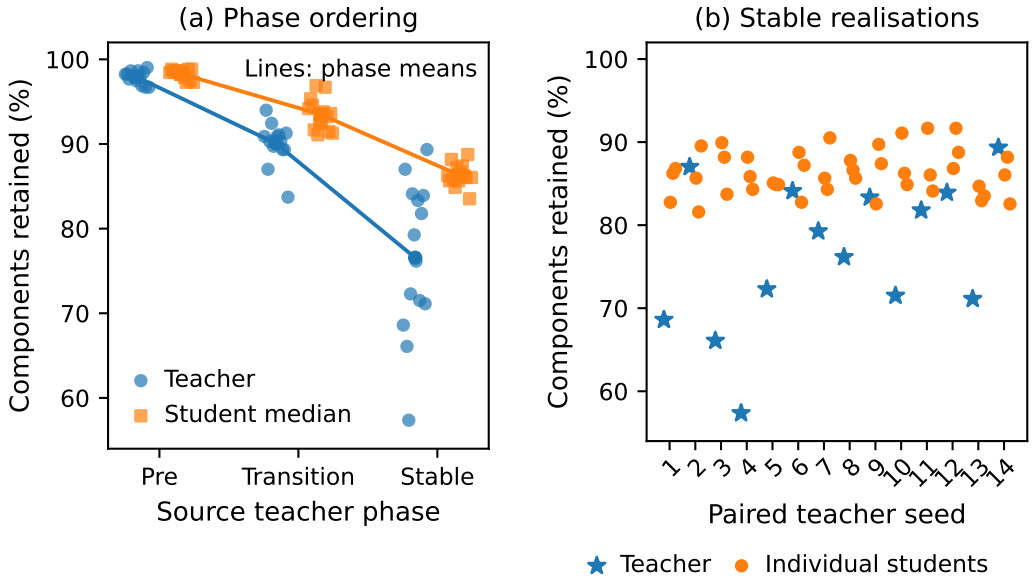

Figure 3: **Phase ordering transfers while absolute recovery changes.** Addition, fourteen matched teacher seeds. (a) Each dot is a teacher or student-cell median; connected symbols are phase means over the same fourteen seeds, not student training trajectories. (b) All three individual students are shown beside their source teacher at the stable checkpoint. The narrow student distribution is not a claim of identical realisations: conditional differences remain visible within teacher groups. Both panels use centred-logit fidelity ≥ 0.99 and the same retained-proportion scale.

and depth. They are four views of the same four teachers, not sixteen independent replications. Basis changes are more heterogeneous: attention-coordinate and rotated-MLP panels preserve the negative direction, whereas coarse neuron blocks yield two negative changes and two ties (Appendix F.1).

Table 2: **Decision-matched transfer across four student architectures.** Mean greedy stable-minus-pre changes, with all four individual teacher-paired values shown (pp). Two eligible students per phase; cell medians precede pairing. Each row uses the same four addition teacher seeds.

Architecture	Seed 1	Seed 2	Seed 3	Seed 4	Mean
Two layers	-7.99	-6.54	-5.96	-4.31	-6.20
Eight heads	-16.73	-17.69	-10.10	-10.48	-13.75
MLP width 256	-14.42	-14.81	-11.35	-17.69	-14.57
MLP width 1,024	-13.57	-10.07	-9.73	-14.59	-11.99

4.4 Hard-concrete endpoints use little of the allowed fidelity error

The executed procedures differ sharply in absolute recovery and achieved fidelity. Greedy typically reaches the acceptance boundary; hard-concrete predominantly returns near-intact masks far above it (Figure 4). For stable addition teachers, median achieved fidelity is 0.9900030 for greedy and 0.9999879 for hard-concrete. Multiplication medians are 0.9900046 and 1.0000000. At primary fidelity, the hard-concrete addition phase changes are only -0.305 pp for teachers and -0.415 pp for students, versus the much larger greedy effects. Multiplication changes are $+0.824$ and -2.350 pp, respectively. Near-intact recovery is not evidence of equally small circuits in both phases. All 464 canonical Phase II endpoint masks retain all four attention heads, so the counted removals in this panel are MLP-neuron removals.

The continuous gates support this diagnosis beyond binary extraction. Across all 240 multiplication optimiser traces, final mean gates have median 0.995791, with 168 at least 0.99. The optimiser minimises absolute logit MSE while final acceptance uses relative error, a mismatch that motivates

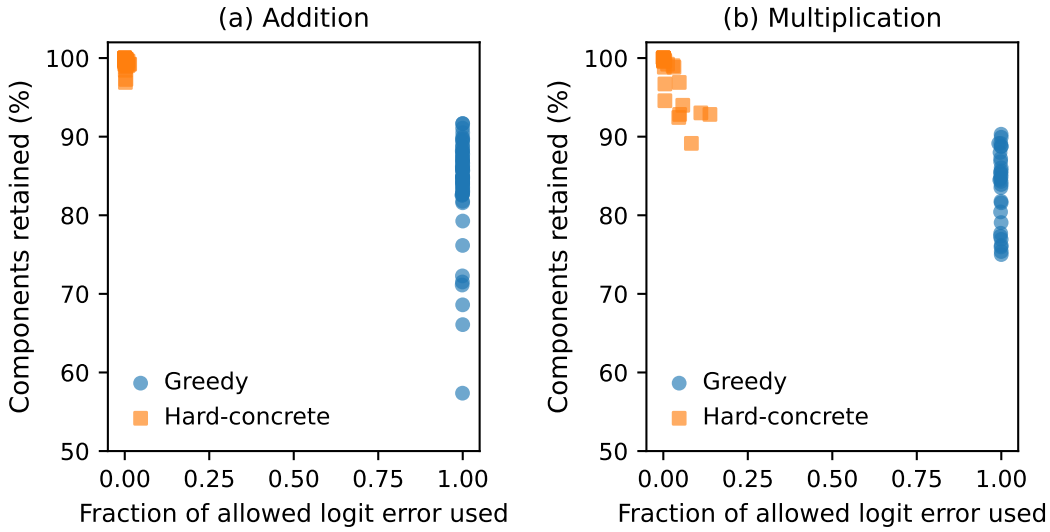

Figure 4: **Conservative hard-concrete endpoints versus boundary-following greedy recovery.** All available canonical stable teachers and individual hard students, with one point per model–method endpoint (120 addition and sixty multiplication points). Horizontal position is $(1 - F_{\text{logit}})/0.01$: zero uses none of the permitted error and one reaches the acceptance boundary. Near-intact hard-concrete masks preserve much more fidelity than required. Overlapping points do not denote additional independent teacher seeds.

a scale-aware optimisation audit. The available evidence identifies conservative recovery under the executed policy; it does not establish whether additional local improvement or a changed objective would remove the gap. Phase I supplies the positive alternative-search replication. Phase II establishes transfer specifically under greedy recovery, with both methods' full endpoints retained in Appendix D.

5 Discussion

The reimplementation experiment answers a question that a longitudinal recovery curve cannot: **the greedy contrast recurs without the original teacher weights under the tested reimplementation procedure**. Fresh networks trained to each complete decision table reproduce its direction on both arithmetic tasks, and addition students preserve the intermediate phase ordering. The accompanying endpoint changes are equally informative. The phase-level regularity recurs while the number of components retained by the procedure changes across realisations.

Decision-table structure and transfer. Stable arithmetic-like tables and earlier, partly incorrect tables are different targets. Their structure is a plausible explanation for the transferred ordering, not a variable held fixed between phases. What we hold fixed is each teacher's complete decision table during replacement of its weights. A compact implementation of a table need not be exposed as a small zero-ablation subnetwork in a trained model. The experiment tests this distinction directly: greedy recovery preserves the phase ordering across reimplementations, while its absolute endpoints and effect sizes change. It establishes transfer under the tested pipeline, rather than a causal decomposition of table structure, logits and representations.

Training duration and later cleanup. Students are measured at first eligibility, not after a subsequent cleanup phase. In the saved addition records, stable student-cell stopping is longer in five matched teacher groups, tied in three and shorter in six; multiplication has longer stable stopping in all eight groups (Appendix A.3). Uniformly longer stable training is absent on addition, but duration is coupled to phase on multiplication. Further training could change student endpoints while deci-

sions remain matched; testing that requires continued-training controls beyond the present stopping-point comparison.

Could student confidence explain the changed endpoints? Exact argmax agreement leaves probabilities, margins and relative logit geometry unconstrained. Equation 2 removes additive shifts and, for a fixed mask, is invariant to common scalar rescaling of both dense and masked logits. However, that invariance does not make the cross-entropy-based search ordering scale-invariant, nor does it remove per-input or class-dependent geometry changes. The current design therefore does not assign endpoint differences uniquely to internal representation rather than output geometry. A complete paired logit-matched cohort would sharpen that distinction; the preserved soft students are stable-only and cannot supply it (Appendix J).

How far does the result extend? The contrast survives non-monotone recovery and distribution-sensitive fidelity in the longitudinal experiment, and greedy reimplementation across two tasks and four small architecture panels. Phase II demonstrates component reduction, not the dramatic sparsity seen under Phase I's decision-agreement criterion; its canonical endpoints retain all four heads. Recovery is intervention- and procedure-relative: mean ablation and a scale-normalised hard-concrete objective with a sparsity-penalty sweep remain useful sensitivity tests. The controlled arithmetic domains enable exhaustive matching but do not establish broader task generality. Auxiliary packing and Fourier-interchange results do not establish broad multiplicity or a shared causal abstraction and are reported with their controls in Appendices H–I.

These results suggest a useful distinction in interpreting circuit measurements: *a comparative pattern can be stable across decision-equivalent realisations even when the individual measurements are not*. Exact decision-matched reimplementation supplies an experimental way to test that distinction. It complements training-time and cross-model circuit analyses by replacing approximate behavioural similarity with a complete, explicit matching condition.

6 Conclusion

The longitudinal recovery gap across grokking remains under non-monotone search and distribution-sensitive fidelity. Across modular addition and multiplication, independently initialised networks matching their teachers' complete decision tables reproduce the greedy phase contrast under own-dense centred-logit fidelity. **The ordering survives replacement of the original weights; the magnitude and absolute recovered sizes change.** Exact decision-matched reimplementation thus tests which aspects of a circuit-recovery result recur beyond the network in which they were first observed.

AI USE STATEMENT

Generative AI tools assisted with research planning and methodological feedback, code development, processing and analysis of recorded results, interpretation, visualisation, literature search, and manuscript drafting and revision. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The anonymous supplement provides two reproduction routes. The command `sh reproduce.sh` rebuilds the saved-result analyses and figures, including the dispersion control, stopping-time audit and exploratory attenuation interval. The separate `retraining/` package supplies executable Phase I training code and a Phase II training, discovery and endpoint replay route, with recorded splits, seeds, stopping rules, budgets and dependency lockfiles. It requires no original cluster account. Small CPU pipelines passed training, recovery, repeatability and resume checks; a complete rerun of the historical GPU experiment was not performed. Trained checkpoints are not included, and bitwise equivalence across hardware is not claimed. Appendix K distinguishes regeneration of saved analyses from new model execution.

REFERENCES

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.14997.

Branton DeMoss, Silvia Sapora, Jakob Foerster, Nick Hawes, and Ingmar Posner. The complexity dynamics of grokking. *Physica D: Nonlinear Phenomena*, 482:134859, 2025. doi: 10.1016/j.physd.2025.134859. URL https://arxiv.org/abs/2412.09810.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1803.03635.

Reilly Haskins and Benjamin Adams. Distilled circuits: A mechanistic study of internal restructuring in knowledge distillation, 2025. URL https://arxiv.org/abs/2505.10822v2. Version 2 revised 9 March 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ziming Liu, Ziqian Zhong, and Max Tegmark. Grokking as compression: A nonlinear complexity perspective. *arXiv preprint arXiv:2310.05918*, 2023. URL https://arxiv.org/abs/2310.05918.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1712.01312.

Ajay Pravin Mahale. Explanation multiplicity: Circuit-level interpretability evidence does not survive defensible analytic variation, 2026. URL https://arxiv.org/abs/2608.13754v1.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. *arXiv preprint arXiv:2510.00845v4*, 2026. URL https://arxiv.org/abs/2510.00845v4.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *Workshop on Understanding Foundation Models at ICLR*, 2023. URL https://arxiv.org/abs/2303.11873.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2407.08734.

Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Bridging lottery ticket and grokking: Understanding grokking from inner structure of networks. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=eQeYyup1tm.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217.

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. URL https://arxiv.org/abs/2201.02177.

Joshua S. Schiffman. Transformers converge to invariant algorithmic cores, 2026. URL https://arxiv.org/abs/2602.22600v2.

Vaibhav Singh, Eugene Belilovsky, and Rahaf Aljundi. When data falls short: Grokking below the critical threshold. *arXiv preprint arXiv:2511.04760*, 2025. URL https://arxiv.org/abs/2511.04760.

Alan Sun. Circuit stability characterizes language model generalization, 2025. URL https://arxiv.org/abs/2505.24731v1.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.10827.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Circuit distillation, 2025. URL https://arxiv.org/abs/2509.25002v1.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.17844.

A DESIGN, TRAINING AND REALISED COVERAGE

A.1 INPUTS, MODELS AND TRAINING TARGETS

The arithmetic domain is $\mathcal{X}_p = \{(a, b) : 0 \leq a, b < p\}$. Inputs are the three-token sequence $(a, b, =)$, where the final token uses the additional vocabulary index p; the final-position output has p classes. The canonical transformer has one causal attention layer, learned positional embeddings, residual width 128, four attention heads of width 32 and a 512-neuron ReLU MLP, without normalisation or dropout. A canonical mask acts at each head's output and each neuron's post-ReLU activation. It does not remove the residual stream or embedding/unembedding operations.

Phase I has five independent initialisations sharing a fixed split of 3,830 training and 8,939 test pairs. Full-batch AdamW training runs for 40,000 updates at learning rate 0.001 and weight decay 1. All five run manifests record CPU execution, checkpointing and evaluation every 50 updates, including step zero and the final step.

Phase II teachers train on explicit stored subsets with arithmetic labels: 3,830 of 12,769 addition pairs and 1,044 of 3,481 multiplication pairs. Students instead use the complete teacher-labelled domain in every update. The executed training settings are identical across both tasks and all reported student architectures (Table 3). Every candidate attempt feeding the 241 included hard-student selections uses this same numerical recipe; selecting a different eligible attempt therefore cannot change the reported recipe. Model weights are float32; ordered float64 accumulation is used for final circuit qualification. Hard training stops at the first saved evaluation achieving exact full-domain decisions, or at the 40,000-update horizon. Changed soft-rescue recipes remain a separate lineage.

Table 3: Executed Phase II teacher and hard-student training settings. Teacher and student updates are full-batch over their respective supports; evaluation chunks are not training minibatches. All architecture panels use the same student optimiser recipe.

Quantity	Teachers	Hard students
Training target	Arithmetic class	Teacher argmax, including errors
Addition support per update	3,830 stored pairs	All 12,769 ordered pairs
Multiplication support per update	1,044 stored pairs	All 3,481 ordered pairs
Optimiser	AdamW	AdamW
Learning rate / schedule	0.001 / constant, no warmup	0.001 / constant, no warmup
Adam coefficients / epsilon	$(0.9, 0.98)$ / 10^{-8}	$(0.9, 0.98)$ / 10^{-8}
Weight decay / precision	1 / float32	1 / float32
Update horizon	40,000	At most 40,000
Evaluation interval	Every 50 updates, including 0	Every 50 updates, including 0
Stopping rule	Declared horizon	First eligible evaluation or horizon
Initialisation	Main seeds 0–14 per task	Independent, fixed attempt-specific seeds; no teacher weights or optimiser state

Hard training minimises cross-entropy against $d_\theta(x)$; correct arithmetic answers are not substituted for mistaken teacher predictions. Soft training minimises per-input, per-class mean squared error between centred logits. Soft eligibility requires both exact full-domain decisions and centred-logit MSE at most 10^{-4}. Neither a high fraction of matching decisions nor a near-threshold soft loss qualifies. Circuit evaluation then uses each model's own dense logits as its reference, independently of its student-to-teacher eligibility calculation.

The architecture sensitivity panel retains residual width 128 and changes either depth to two layers, head count to eight width-16 heads, or MLP width to 256 or 1,024. The corresponding canonical component counts are 1,032, 520, 260 and 1,028. These counts follow the head-plus-neuron convention; the architectures are not parameter-matched. The calibration-only addition-modulo-7 model has residual width 16, two width-8 heads and 16 MLP neurons, for 18 components.

A.2 BEHAVIOURAL CHECKPOINT SELECTION

Phase I uses steps 200, 3,400, 7,450, 8,150, 8,500, 8,650 and 9,050. Positions before the first saved checkpoint above 10% test accuracy are delayed. Stable onset is the fifth checkpoint in the earliest five-consecutive-checkpoint sequence with test accuracy at least 99%; positions between these landmarks are transitional. The common grid was fixed after pilot work and before the multi-seed family outcomes. The use of a common grid means seeds contribute different numbers of positions to each phase.

Phase II uses stricter, selected roles. Pre is the latest saved checkpoint with training accuracy at least 99.9% and test accuracy at most 5%, before the first saved checkpoint reaching 10% test accuracy. Stable uses the sustained-99% rule above. Addition transition is the checkpoint strictly between selected pre and stable that is nearest 50% test accuracy, with earlier-step tie-breaking. A role is unavailable when its rule cannot be met. In particular, Phase I "delayed" and Phase II "pre" are not interchangeable sampling conditions. Students inherit the phase label of their target teacher checkpoint; it is not a measured student grokking phase.

Table 4: Phase II coverage. Available phase counts differ from matched populations. Canonical student counts include all available phases, not only paired endpoints.

Unit	Addition	Multiplication
Main teacher initialisations	15	15
Planned checkpoint roles	45	30
Available pre / transition / stable	14 / 14 / 15	10 / — / 10
Matched pre–stable teacher seeds	14	8
Canonical hard students, planned	135	60
Canonical hard students, available and preserved	129	40
Alternative-architecture hard students, planned	80	—
Alternative-architecture students, available and preserved	72	—
Eligible hard students, total	201	40

Addition seed 0 lacks pre and transition roles. Multiplication pre seeds are 0, 1, 4, 6, 7, 9, 10, 11, 12 and 13; stable seeds are 1, 3, 4, 5, 6, 9, 10, 11, 12 and 13. The matched set is therefore 1, 4, 6, 9, 10, 11, 12 and 13. Primary addition pairs use seeds 1–14. Architecture panels were assigned to seeds 0–4; missing seed-0 pre leaves four pairs.

There are 241 preserved eligible hard students from 275 intended slots. The other 34 slots have unavailable source phases, not failed distillation. The available attempt records separately contain five failed hard attempts before the secured cohort: two addition and three multiplication. Unused reserved attempts are not executed failures. The policy seeks three canonical addition successes from at most six fixed attempts per cell, and two successes from at most four attempts in the smaller panels. A separate calibration teacher is not a thirty-first main population replicate.

A.3 FIRST-ELIGIBILITY STOPPING IN THE HARD-STUDENT COHORT

An exploratory pre-submission audit linked each selected canonical hard student to its saved first-eligibility update, following the original attempt order and eligibility rank. It linked all 169 canonical students (129 addition and forty multiplication) and reproduced the 63 student-cell recovery medians and all 22 primary student phase contrasts. The matched cohorts in Table 5 use the same fourteen addition and eight multiplication teachers as the primary comparison; the remaining available students are retained in the accompanying data. Each complete eligible attempt's last update equals its eligible update and follows the recorded 50-update evaluation cadence.

Comparing the *within-cell median* stopping updates between stable and pre gives longer stable training in five addition teacher groups, ties in three and shorter training in six, with median paired difference zero. Multiplication has longer stable stopping in all eight groups and median paired difference 4,000 updates. The difference of the pooled individual-student medians in Table 5 need not equal the median of these teacher-paired differences.

The data do not support a uniform longer-training description for stable addition students, whereas multiplication couples longer training to stable targets. Within matched stable cohorts, descriptive

Table 5: First-eligibility stopping updates for selected hard students in the matched cohorts. Medians and ranges describe individual students nested within teacher groups; the students are not independent teacher replicates. Source phase labels refer to the teacher.

Task	Phase	Teachers	Students	Median	Range
Addition	Pre	14	42	400	350–450
Addition	Transition	14	42	350	300–400
Addition	Stable	14	42	350	200–9,000
Multiplication	Pre	8	16	200	200–250
Multiplication	Stable	8	16	2,225	300–28,950

Spearman correlations between cell-median stopping updates and recovered proportions are -0.161 on addition ($n = 14$) and $+0.905$ on multiplication ($n = 8$). Correlations between paired stopping changes and paired recovery changes are -0.393 and $+0.833$, respectively. These associations have no new significance tests and are not adjustments for training duration. They do not identify the separate effects of target-table structure, optimisation, accumulated weight decay or output geometry; no common-horizon continuation was performed.

B PHASE I TRAJECTORIES, FOLLOW-UP TIMING AND SEARCH STATUS

Table 6 gives every primary first-greedy endpoint with its phase, avoiding an inference from pooled phase medians alone. The original family grid crosses five seeds, seven positions, six decision thresholds and three overlap limits, giving 630 repeated protocol cells. All 15 stable primary cells have families of six or seven masks; the 20 earlier primary cells have no family below the 258-component cap. Across the full grid, 300 cells are nonempty and 330 are empty; 270/270 stable cells are nonempty, but they are not 270 separately trained models. The five seed-level primary family changes are $+6, +7, +7, 0, +7$, giving a nonzero-sign exact test of $p = 0.125$.

Table 6: Every primary Phase I first-greedy endpoint. D/T/S denote delayed, transition and stable. Each cell gives retained count and phase at decision fidelity ≥ 0.99.

Seed	200	3,400	7,450	8,150	8,500	8,650	9,050
0	513 D	515 T	77 S	82 S	75 S	70 S	84 S
1	515 D	515 D	515 T	514 T	511 T	493 T	146 S
2	513 D	515 D	297 T	105 S	87 S	93 S	92 S
3	510 D	514 D	515 T	515 T	515 T	514 T	509 T
4	513 D	515 D	89 S	72 S	67 S	65 S	74 S

The first-path analysis reconstructs 210 endpoints (35 positions times six decision thresholds) from saved searches. The first path is identical across overlap settings because it has no previous family member. The original family search subsequently favours deleting components reused in earlier members. It requests up to ten members, with 10,000 exact evaluations per requested member across restarts and 50,000 per cell. Observed family counts are lower bounds, not saturation estimates.

Matched-overlap nulls, random-mask nulls and first-path reconstruction were designed after the original results. The annealing and distribution-sensitive follow-ups had execution rules fixed before their new search outcomes, but after the original pattern was known. They are targeted follow-ups, not a fully independent preregistered replication of the whole programme.

The non-monotone top-one study has 350 streams, ten at each position, with 20,000 unique evaluations per stream and deterministic deletion cleanup. Stable streams accept median counts of 18 additions and 399 swaps. Stochastic-search censor flags are set for 86/90 delayed and 58/110 transitional streams, but no stable stream; all stable streams nevertheless exhaust their 20,000 unique-evaluation allowance. Budget exhaustion and the recorded censor criterion are different facts. Feasible-only movement and local cleanup do not establish a global minimum.

For the distribution-sensitive study, let P_x and Q_x be dense and masked softmax outputs and $M_x = (P_x + Q_x)/2$. The discrepancy is the domain mean of $\frac{1}{2}\mathrm{KL}(P_x \| M_x) + \frac{1}{2}\mathrm{KL}(Q_x \| M_x)$. The JS

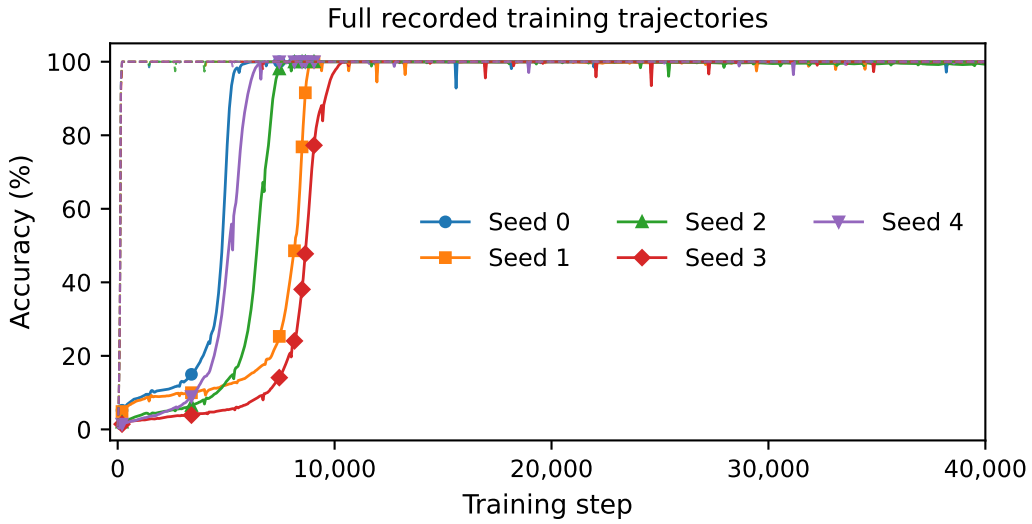

Figure 5: Recorded Phase I learning curves over the full training horizon: 801 accuracy observations per seed, from step 0 through 40,000 at intervals of 50. Solid lines are held-out test accuracy; dashed lines are training accuracy. Markers identify the original seven circuit-evaluation steps, all at or before 9,050. Later accuracy observations do not add later circuit measurements. The main figure enlarges the first 11,000 updates.

tolerance 0.0005 was chosen by a stated calibration rule at seed 0's final stable checkpoint; that checkpoint also appears in the full study with separate optimiser streams. Calibration does not equate the stringency of JS and decision agreement. The full JS study has ten streams at each of the same 35 positions. All delayed streams and 48/110 transitional streams carry the stochastic censor flag, whereas none of the stable streams does; every endpoint completes local cleanup.

Separate deterministic searches cover six JS tolerances, six decision thresholds and six normalised-margin tolerances. At JS 0.0005 the deterministic stable median is 281; at JS 0.005 it is 191, with all fifteen stable positions below half the model. The margin criterion failed its prospective calibration gate, so no full margin annealing campaign was performed. A missing campaign is not a null experimental outcome. The main text therefore relies on the completed JS study rather than substituting a favourable margin setting.

C Phase II search and measurement specification

The recovery measure is evaluated on the same complete finite domain used to define the dense reference, rather than on a Monte Carlo sample. For a fixed model and mask, there is consequently no input-sampling standard error in the reported agreement or logit sum of squares. This does not turn adaptive search into unbiased estimation of global minimum size or eliminate population uncertainty over models.

The canonical greedy routine starts intact and computes gradients of mean cross-entropy against the dense argmax labels, with one continuous gate per component. Signed deletion damage ranks the currently active components; canonical indices break ties. Each round scans this ordering in batches of at most sixteen. It exactly evaluates the entire batch, then accepts the highest-fidelity qualifying deletion in the first batch with any qualifying candidate, with canonical-index tie-breaking. It re-ranks after acceptance. It stops when no single deletion in the scanned ordering qualifies, the native allowance is exhausted, or execution is interrupted. An incompletely evaluated batch cannot supply an accepted deletion.

Its native allowance is $16N$: 8,256 candidate-mask requests for the canonical universe. Requests can include repeated masks. Persistent exact records distinguish requests from newly executed physical evaluations. The final ledger prepends the intact model, then prioritises the final native mask, the

reversed accepted trajectory and chronological proposals, deduplicating candidates. The final ledger allowance is 256. A truncated final ledger can carry a censor status without implying that the native search exhausted its own allowance; these statuses must remain separate. The same final candidate inventory is re-reduced at alternative thresholds.

Hard-concrete learns stochastic continuous gates while keeping model weights fixed. Four restarts each use 5,000 Adam steps with zero initial log-alpha; all endpoint-linked restart bindings share the gate and optimiser constants in Table 7. For log-alpha ℓ_i, temperature $\beta = 2/3$ and stretches $(\gamma, \zeta) = (-0.1, 1.1)$, a training sample is

$$g_i = \mathrm{clip}_{[0,1]}\left[(\zeta - \gamma)\,\sigma\left(\frac{\ell_i + \log u_i - \log(1 - u_i)}{\beta}\right) + \gamma\right], \qquad u_i \in [10^{-6}, 1 - 10^{-6}], \tag{3}$$

where uniform draws are clamped to the displayed interval. The expected nonzero probability entering the regulariser is $q_i = \sigma(\ell_i - \beta \log(-\gamma/\zeta))$. The objective is mean squared centred-logit error over the fitting examples and classes, plus $0.001N^{-1}\sum_i q_i$. At zero-based optimiser step t, the fitting batch is the contiguous domain slice starting at $(256t) \bmod |\mathcal{X}|$ and ending at the smaller of that start plus 256 and $|\mathcal{X}|$. The terminal slice can therefore be shorter than 256; batches are not randomly shuffled. Final qualification still evaluates the complete domain.

Binary extraction uses the different deterministic score $v_i = \mathrm{clip}_{[0,1]}[(\zeta - \gamma)\sigma(\ell_i) + \gamma]$, not q_i. A threshold mask retains $v_i \geq \tau$. Top-k masks rank v_i descending with canonical-index tie-breaking. The stochastic proposals independently retain components by Bernoulli draws with probabilities v_i, using identity-derived CPU random streams. Canonical extraction produces nine threshold, eleven top-k and 32 stochastic proposal records per restart before deduplication. The configured cap is 256 proposal records, so it does not truncate this 52-record list. The final combined exact ledger uses the same 256 allowance and qualification/reduction machinery as greedy. Equal final-ledger allowance does not equalise optimisation steps, gradients, proposal generation or all exact evaluations.

Table 7: Executed Phase II recovery settings. Numerical values come from the job-bound configurations linked to the completed endpoints; fitting and extraction semantics follow the corresponding implementation.

Quantity	Setting
Canonical intervention	Four whole heads and 512 post-ReLU MLP neurons
Qualification	Full domain; own-dense centred-logit fidelity ≥ 0.99
Greedy ranking / batch	Signed cross-entropy deletion gradient; canonical ties; at most 16 exact candidate checks per batch
Greedy native allowance	$16N$ candidate requests (8,256 for $N = 516$)
Final exact ledger	256 entries; intact mask, deduplication and declared priority order
Hard-concrete restarts	Four, each with 5,000 updates; initial log-alpha zero
Gate shape	Temperature $2/3$; stretch lower -0.1, upper 1.1; clamp $[0, 1]$; sampling epsilon 10^{-6}
Hard-concrete Adam	Learning rate 0.01, constant (decay factor 1); coefficients $(0.9, 0.999)$; epsilon 10^{-8}; no weight-decay term
Optimisation objective	Absolute centred-logit MSE plus 0.001 times mean expected L_0
Fitting / evaluation chunks	At most 256 inputs; fitting uses the deterministic slice schedule stated above; exact evaluation traverses the full domain
Extraction thresholds	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ applied to v_i
Canonical top-k list	$0, 8, 16, 32, 64, 96, 128, 192, 256, 384, 516$
Stochastic extraction	32 Bernoulli masks per restart using v_i
Alternative acceptance views	$0.90, 0.95, 0.975, 0.99, 0.995$ on the same candidate ledger

For a noncanonical universe of N components, each canonical k maps to $\lfloor (kN + 258)/516 \rfloor$, followed by deduplication and explicit inclusion of 0 and N. Table 8 gives every executed grid in the reported architecture and basis panels. These changes rescale the intervention universe, not the regularisation coefficient or fidelity threshold.

For multiplication, the 240 complete optimiser traces yield 11–52 unique proposals per restart (median 12). Final continuous gate means have median 0.995791 and range 0.912355–1.000000; 168/240 are at least 0.99 and 218/240 at least 0.95. Median minibatch fidelity loss changes from 177.7264 to 0.513586, while the final expected-L_0 penalty term has median 0.000994955. These

Table 8: Complete top-k extraction grids used in the endpoint-linked panels. Canonical $N = 516$ also applies to each rotated-MLP basis. All grids use the same thresholds and stochastic-draw count.

Panel	N	Retained counts
Canonical / MLP rotation	516	0, 8, 16, 32, 64, 96, 128, 192, 256, 384, 516
Depth two	1,032	0, 16, 32, 64, 128, 192, 256, 384, 512, 768, 1,032
Eight heads	520	0, 8, 16, 32, 64, 97, 129, 193, 258, 387, 520
MLP width 256	260	0, 4, 8, 16, 32, 48, 64, 97, 129, 193, 260
MLP width 1,024	1,028	0, 16, 32, 64, 128, 191, 255, 383, 510, 765, 1,028
Attention coordinates	640	0, 10, 20, 40, 79, 119, 159, 238, 318, 476, 640
Coarse MLP blocks	20	0, 1, 2, 4, 5, 7, 10, 15, 20

scalar values motivate investigating objective scale; they do not prove which gradients dominated or which retuning would succeed. A changed objective would be a post-result sensitivity experiment, not a silent correction of the primary method.

The 464 canonical primary endpoint records comprise 344 addition and 120 multiplication model–method evaluations. Saved masks, their retained counts and their recorded qualifying fidelity agree with their aggregate endpoints. Every mask retains the four attention heads. Attention was allowed to change under the canonical intervention; universal retention among these recovered masks does not prove global head necessity. Whole heads and neurons also have unequal parameter costs, so a retained-component fraction cannot be presented as a parameter-compression rate.

D Paired estimation, absolute levels and seed-specific outcomes

For teacher s and phase p, let r^T_{sp} denote its recovered proportion, and let r^S_{spj} be the corresponding value for eligible student j. The student-cell value is $\bar{r}^S_{sp} = \mathrm{median}_j\, r^S_{spj}$ using the required number of students. For each condition $q \in \{T, S\}$,

$$\Delta^q_s = 100(\bar{r}^q_{s,\mathrm{stable}} - \bar{r}^q_{s,\mathrm{pre}}), \qquad \widehat{\Delta}^q = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \Delta^q_s, \tag{4}$$

where the teacher has a single observation in each cell and $\mathcal{S}$ contains only complete teacher pairs. Student medians are calculated before pairing and averaging; pooling student rows directly would change the estimand and overweight more prolific cells.

The addition reference test enumerates all 2^{14} sign assignments to the observed teacher differences and compares the absolute summed difference with the observed absolute sum. This is an exact sign-flip reference distribution, not randomised assignment of learning phases or a causal test. Its interpretation requires the corresponding sign-symmetry null across teacher differences. Raw p-values in teacher-greedy, teacher-HC, student-greedy and student-HC order are 0.0001220703, 0.0008544922, 0.0001220703 and 0.0002441406. Holm adjustment uses the declared four-comparison family. Bootstrap intervals resample fourteen teacher differences with replacement 10,000 times and use the 2.5th and 97.5th percentiles with linear interpolation. They are pointwise, not simultaneous intervals. Identity-derived random seeds preserve the recorded bootstrap realisation.

Multiplication is reported descriptively because its eight matched seeds fall below the pre-existing inferential release rule. Figure 2 in the main text displays every paired greedy contrast for both tasks.

Table 10 reports all available cell-level medians. Different phase populations mean that subtracting these medians need not reproduce the paired mean difference. Likewise, within-cell student ranges are descriptive conditional spreads. For addition, their phase medians are 0.581, 1.163 and 5.620 pp (pre, transition, stable); maximum ranges are 0.775, 6.783 and 7.946 pp. Multiplication phase medians are 0.775 and 5.039 pp, and maxima are 1.357 and 8.915 pp. These are ranges over three or two eligible initialisations, not confidence intervals or variance components.

Student-minus-teacher greedy differences also depend on the task. Addition phase-specific medians are $+0.485$, $+2.519$, $+6.395$ pp. All fourteen transition cells and twelve of fifteen stable cells retain more components in the student median. Multiplication phase medians are -1.017 and -3.198 pp,

Table 9: Every matched greedy phase pair. Pre/stable entries are retained proportions; changes are percentage points. Student columns use cell medians. Display rounding does not change the full-precision calculation.

Task	Seed	T pre	T stable	Δ T	S pre	S stable	Δ S
Add	1	0.983	0.686	-29.65	0.984	0.862	-12.21
Add	2	0.983	0.870	-11.24	0.988	0.857	-13.18
Add	3	0.977	0.661	-31.59	0.986	0.882	-10.47
Add	4	0.986	0.574	-41.28	0.984	0.859	-12.60
Add	5	0.981	0.723	-25.78	0.986	0.849	-13.76
Add	6	0.979	0.841	-13.76	0.986	0.872	-11.43
Add	7	0.975	0.793	-18.22	0.983	0.857	-12.60
Add	8	0.986	0.762	-22.48	0.984	0.866	-11.82
Add	9	0.979	0.833	-14.53	0.979	0.874	-10.47
Add	10	0.969	0.715	-25.39	0.973	0.862	-11.05
Add	11	0.984	0.818	-16.67	0.988	0.860	-12.79
Add	12	0.967	0.839	-12.79	0.975	0.888	-8.72
Add	13	0.990	0.711	-27.91	0.988	0.835	-15.31
Add	14	0.967	0.893	-7.36	0.973	0.860	-11.24
Multiply	1	0.932	0.843	-8.91	0.921	0.854	-6.69
Multiply	4	0.940	0.750	-18.99	0.930	0.891	-3.88
Multiply	6	0.936	0.847	-8.91	0.914	0.876	-3.78
Multiply	9	0.922	0.791	-13.18	0.912	0.764	-14.83
Multiply	10	0.895	0.888	-0.78	0.927	0.804	-12.31
Multiply	11	0.932	0.899	-3.29	0.933	0.798	-13.47
Multiply	12	0.928	0.857	-7.17	0.900	0.859	-4.17
Multiply	13	0.946	0.853	-9.30	0.932	0.816	-11.63

Table 10: Absolute recovered proportions at centred-logit fidelity 0.99. Entries are medians over available teacher cells, followed by cell counts. Students are first reduced within cells.

Task	Model	Search	Pre	Transition	Stable
Add	Teacher	Greedy	0.980 (14)	0.903 (14)	0.793 (15)
Add	Teacher	HC	1.000 (14)	1.000 (14)	0.996 (15)
Add	Hard	Greedy	0.984 (14)	0.934 (14)	0.860 (15)
Add	Hard	HC	1.000 (14)	1.000 (14)	0.996 (15)
Multiply	Teacher	Greedy	0.932 (10)	—	0.855 (10)
Multiply	Teacher	HC	0.991 (10)	—	1.000 (10)
Multiply	Hard	Greedy	0.924 (10)	—	0.813 (10)
Multiply	Hard	HC	0.994 (10)	—	0.968 (10)

with six negative and four positive cells in each phase. Distillation therefore does not uniformly make the model easier or harder to recover sparsely.

D.1 Exploratory teacher–student attenuation interval

For the fourteen matched addition teachers, define the paired difference of phase contrasts

$$D_s = \Delta_s^S - \Delta_s^T, \qquad \overline{D} = \frac{1}{14}\sum_{s=1}^{14} D_s. \tag{5}$$

Positive values indicate a smaller stable-minus-pre reduction in student-cell medians. Twelve of fourteen D_s are positive. Their mean is $+9.3577$ pp. An exploratory teacher-paired percentile bootstrap gives a 95% interval $[4.9280, 14.0642]$ pp: independently resample fourteen D_s with replacement 100,000 times, average each resample, and take the 2.5th and 97.5th percentiles with linear interpolation. The random generator is NumPy PCG64 with seed 20260923. Resampling the already paired differences preserves the teacher–student and pre–stable pairing; the interval is not obtained by subtracting marginal confidence limits.

This interval was added in response to pre-submission review and is separate from the original four-comparison inference family. No primary p-value, Holm adjustment or saved bootstrap interval was changed. It describes uncertainty over the observed teacher-seed population under the specified task, split, selection and recovery procedures, not uncertainty over tasks or arbitrary reimplementation methods. Every D_s and its original component contrasts are supplied in `revision_results/attenuation_per_seed.csv`.

E Descriptive dispersion and transition analyses

These analyses reuse the saved canonical greedy endpoints at fidelity 0.99. They use exactly the primary complete-pair cohorts: addition seeds 1–14 and multiplication seeds 1, 4, 6, 9, 10, 11, 12 and 13. The addition seed with only a stable checkpoint is excluded from these paired-population dispersion summaries, but remains in all-available summaries elsewhere. Each phase is evaluated on the same teachers within a task.

For each phase we calculate sample standard deviation across dense teachers, across student-cell medians, and a one-student-per-teacher selection control. Let x_{sj} be the retained percentage for saved student j of teacher s, $\mu_s = J_s^{-1}\sum_j x_{sj}$, and $v_s = J_s^{-1}\sum_j (x_{sj} - \mu_s)^2$. Independently choose one saved student uniformly within each teacher. If X_s is its endpoint and S_X^2 the sample variance over the n teachers, then

$$\mathbb{E}[S_X^2] = \frac{1}{n-1}\sum_s (\mu_s - \bar{\mu})^2 + \frac{1}{n}\sum_s v_s. \tag{6}$$

This identity follows by expanding $\sum_s X_s^2 - n\bar{X}^2$ and using independence of the selections. It is an exact expectation over the finite inventory of saved students, not a fitted population variance decomposition. The reported control is $\sqrt{\mathbb{E}[S_X^2]}$ (an RMS SD), not $\mathbb{E}[\sqrt{S_X^2}]$. The analysis script checks the identity against exhaustive enumeration of a small artificial selection example; that example is a code test, not scientific data.

Table 11: Across-teacher dispersion in the matched cohorts, in percentage points. "One student" is the RMS SD under uniform selection of one saved student per teacher. No additional inferential tests are introduced.

Task	Phase	Teachers	Teacher SD	Median SD	One student
Addition	Pre	14	0.74	0.57	0.62
Addition	Transition	14	2.42	1.90	2.01
Addition	Stable	14	9.11	1.32	2.63
Multiplication	Pre	8	1.53	1.18	1.24
Multiplication	Stable	8	4.90	4.39	5.28

For stable addition, teacher and student-median interquartile ranges are 12.55 and 1.36 pp, respectively; their complete ranges are 31.98 and 5.23 pp. Thus the concentration is visible in both range and IQR, not only SD. The one-student control retains considerably less across-teacher dispersion than the teacher endpoints. Multiplication provides an important limit: its one-student RMS SD exceeds the teacher SD, so the addition pattern is not reported as a universal distillation effect. Within-teacher ranges and between-teacher SDs describe different sources of descriptive variation and should not be interchanged.

On the same fourteen addition teachers, successive mean changes are -7.849 and -13.483 pp for teachers, and -4.790 and -7.184 pp for student medians. Strictly decreasing pre–transition–stable sequences occur for thirteen teacher trajectories and all fourteen student trajectories. These three selected behavioural landmarks do not determine whether recovery changes smoothly or abruptly between checkpoints.

E.1 What logit scale invariance does and does not establish

For a fixed model–mask output pair and nonzero scalar a, multiplying both centred outputs by a cancels from the numerator and denominator of Eq. 2. Class-dependent geometry or input-dependent scaling need not cancel. Moreover, the cross-entropy gradient used to rank greedy deletions can change under output rescaling. The algebra therefore removes a trivial scale dependence of the fixed-mask score, not every confidence-related explanation of recovery endpoints. No confidence-adjusted causal analysis is inferred from this identity.

F Acceptance-threshold and intervention sensitivity

Table 12 includes every listed fidelity threshold. All entries re-reduce the existing final candidate sets; they are not five separately optimised studies. In particular, flat intervals and abrupt changes can arise because a discrete candidate pool contains few masks in some fidelity ranges. They do not identify a flat or discontinuous global optimum.

Table 12: Mean paired phase contrasts over the saved candidate frontier, in pp. Addition has 14 pairs; multiplication has 8. HC denotes hard-concrete.

Task	Fidelity	Teacher G	Hard G	Teacher HC	Hard HC
Add	0.900	-21.290	-11.877	-5.689	-20.183
Add	0.950	-21.290	-11.877	-0.305	-0.415
Add	0.975	-21.290	-11.877	-0.305	-0.415
Add	0.990	-21.332	-11.974	-0.305	-0.415
Add	0.995	-11.739	-6.658	-0.305	-0.415
Multiply	0.900	-8.818	-8.842	-8.769	-23.098
Multiply	0.950	-8.818	-8.842	+0.824	-8.261
Multiply	0.975	-8.818	-8.842	+0.824	-2.350
Multiply	0.990	-8.818	-8.842	+0.824	-2.350
Multiply	0.995	-5.160	-5.015	+0.824	-2.350

F.1 Architecture and basis panels

The attention-coordinate basis replaces four whole-head units by 128 head coordinates and retains 512 neuron units, for 640 components. Balanced blocks retain four head units and sixteen 32-neuron blocks, for 20 components. The neuron ordering is a seed-bound Fisher–Yates permutation. Rotated-MLP masks act in a seeded orthogonal basis of the post-activation state, returning the retained-coordinate displacement to the original basis. These operations change the allowable interventions, not the underlying unmasked model. Coarse blocks can produce ties because removing one block is a substantially larger intervention than removing one neuron.

Three HC architecture panels have four negative teacher-paired cell differences; the width-256 panel has two negatives and two ties. Equal aggregate rotation effects do not imply duplicate individual outcomes: for example, seed-0 stable greedy proportions are 0.8972868 and 0.9050388 in the two

Table 13: Descriptive sensitivity panels. Each row has four matched addition teacher seeds. Architecture rows use two hard students per available cell; basis rows below use direct teachers. Values are mean stable-minus-pre retained proportions in pp.

Panel	Greedy	HC	Greedy signs $-/0/+$
Two layers, matched widths	-6.202	-0.521	4/0/0
Eight heads, residual width 128	-13.750	-0.409	4/0/0
MLP width 256	-14.567	-0.240	4/0/0
MLP width 1,024	-11.989	-0.778	4/0/0
Attention coordinates	-21.719	-0.391	4/0/0
Balanced 32-neuron blocks	-2.500	0.000	2/2/0
Post-activation MLP rotation 0	-14.099	0.000	4/0/0
Post-activation MLP rotation 1	-14.099	0.000	4/0/0

rotations. The panels are too small and not sufficiently crossed to isolate independent causal effects of depth, width or head count. Raw retained proportions also do not give parameter-weighted equivalence across bases.

G Phase I multiplicity, matched nulls and shared structure

At the last analysed checkpoint, four stable seeds contribute 27 original family masks with sizes 66–180 and decision fidelities 0.9901–0.9918. Their 78 dependent within-family pairs have pooled median Jaccard overlap 0.1581, compared with median size-matched and head/MLP-stratified null expectations 0.0879 and 0.1085. The observed masks overlap more than uniform matched subsets on average, not less. Seed 1 is closer to its null expectation than the other stable seeds. Because pairs share masks and teachers, these pooled summaries are not 78 independent population observations.

The behavioural nulls use 21 unique seed/composition profiles, 100 draws per profile and null, giving 2,100 size-matched and 2,100 head/MLP-stratified masks. Median fidelities are 0.0261 and 0.2830; maxima are 0.4120 and 0.6361. No draw passes 0.99. The smallest profile-level gap between an original mask and the best stratified draw is 0.3542. Zero successes in 100 draws gives a two-sided exact 95% upper limit of 0.0362 per profile/null pass probability; it is not a search-adjusted significance comparison between equally optimised procedures.

Independent top-one annealing pools support families of ten at all fifteen stable positions when maximum pairwise Jaccard is 0.50. At overlap 0.25, exact maximum families over those saved pools range from one to ten. JS-qualified pools at tolerance 0.0005 support ten-mask families at three final positions and a four-mask family at seed 4, step 7,450, under overlap 0.50; under overlap 0.25, none contains a qualifying pair. Multiplicity therefore depends jointly on behavioural fidelity, allowed size, overlap and the candidate inventory.

Shared-core interventions further limit an interpretation as independent algorithms. The literal neuron intersections of the four original final-stable families contain 3, 8, 4 and 1 neurons. Keeping an intersection plus all heads gives fidelity only 0.0324–0.0458. Removing that intersection from each family member lowers fidelity to 0.6918–0.9874; all 27 members fail 0.99. This shows collective contribution of the tested shared set, not individual necessity of every shared neuron. The corresponding family unions have fidelities 0.982927, 0.999530, 0.859895 and 0.854100: three of four unions fail despite each constituent member passing. Faithfulness under these interventions is not monotone in component inclusion.

The same 27 masks have JS divergence 0.0252–0.2130, normalised reference-margin error 0.653–0.889 and relative centred-logit RMSE 0.560–0.841. None meets the calibrated JS tolerance. The stable dense-margin versus JS-recovered-size Spearman correlation is descriptively -0.949, with very few repeated positions per seed; it is not an adjusted explanation of the phase effect. A one-seed random-label control returns zero primary families at its seven checkpoints. The reduced-data no-generalisation control was unavailable because no candidate met its selection rule, rather than being an observed zero. A subset-transfer pilot groups seven structural masks into one coarse behavioural group at the tested cutoffs. These controls constrain interpretation but do not identify independent algorithms.

H PHASE II PACKING AND ITS CALIBRATION LAYERS

The packing endpoint asks for a mutually compatible subset of qualifying recovered masks under size and maximum pairwise Jaccard caps. The primary caps are 0.25 and 0.25; the extended grid crosses size caps 0.125, 0.25 and 0.375 with overlap caps 0, 0.125, 0.25 and 0.50 at five fidelity thresholds. All 37,920 non-null addition views and all 7,200 multiplication views have recovered packing zero. Each endpoint-reduction job expands to sixty views. These 45,120 views are repeated reductions, not separate trained models or searches.

The four calibration layers answer different questions. First, 36 matched-family combinatorial jobs each draw 10,000 times, but every source family has size zero. Those 360,000 empty-family draws provide no nontrivial multiplicity baseline and are not random-mask forward evaluations. Second, 36 ordinary-restart reductions, each with sixteen completed restarts, contain 5,063 exact records but no qualifying sparse record at the primary cap. Nominal greedy restarts should not be assumed to add stochastic search diversity because the primary greedy rule is deterministic. Third, 36 fresh-exact local-perturbation jobs evaluate 128 masks each: 3,828/4,608 meet fidelity 0.99, but none also meets size cap 0.25. Local behavioural preservation does not imply sparse packing.

Fourth, exhaustive enumeration is feasible for the separate 18-component addition-modulo-7 calibration model. All $2^{18} = 262{,}144$ masks were evaluated. At 0.99, sixteen masks qualify and the true minimum is fourteen components (Table 14). This certifies the absence of quarter-model qualifying masks for that small instance, not for the canonical model.

Table 14: Exhaustive calibration on one 18-component model. These minima are global in this small intervention universe only.

Fidelity threshold	Qualifying masks	Minimum retained / 18
0.900	11,552	9
0.950	2,480	10
0.975	400	12
0.990	16	14
0.995	8	15

Phase I families and Phase II zero packing are not a controlled contradiction. The fidelity criterion, size cap, model sampling and procedures differ. Phase I also has JS-qualified families under overlap 0.50, so its multiplicity cannot be described as existing only under top-one agreement. Zero recovered packing is an observed lower bound, not proof that no alternative masks exist.

I FOURIER INTERCHANGE: DESIGN AND RESULTS

The completed interchange arm covers fourteen pre and fifteen stable addition teacher/hard-student pairs. Each pair has one aligned condition and five controls; each condition has 256 changed-sum trials and 256 matched no-change preservation trials. The total is 174 completed condition jobs and 89,088 trial slots, nested within 29 pairs. No multiplication interchange result is available.

The intervention reads the final-token post-activation MLP state in layer 0. Source and recipient rank-eight bases use four sine/cosine Fourier mode pairs $(1,1),(2,2),(3,3),(4,4)$. An orthogonal Procrustes alignment is fitted on training-split activations without output logits or test outcomes as fitting inputs. For recipient base input and source donor input, the intervention removes the recipient projection onto its fitted subspace and inserts the aligned donor code, retaining the orthogonal remainder and downstream weights. The recipient counterfactual keeps its first operand and changes the second so that its sum equals the donor sum modulo 113.

For class-centred recipient logits, the recorded changed-sum displacement score is

$$S = 1 - \frac{\|z_{\text{intervened}} - z_{\text{counterfactual}}\|^2}{\|z_{\text{base}} - z_{\text{counterfactual}}\|^2}. \tag{7}$$

A score of one reaches the recipient counterfactual logits, zero has the same squared distance as the base output, and negative values are worse. Zero denominators are unavailable. Counterfactual

argmax success and no-change preservation are separate outcomes; displacement is not an accuracy percentage.

Table 15: Mean aligned-minus-control displacement contrasts across teacher pairs. Positive values favour aligned interchange on this outcome only. Units are score differences, not percentage points.

Control	Pre ($n = 14$)	Stable ($n = 15$)
Wrong Fourier mode	-0.684	-0.752
Shuffled coefficients	$+3.545$	$+2.932$
Mismatched input	$+0.115$	-0.236
Equal-norm random state	$+3.644$	$+2.993$
Unaligned ordinary subspace patch	-1.022	-4.018

Wrong-mode control uses $(1, 2), (2, 1), (1, -1), (2, -2)$. The other controls permute the eight coefficients, use another declared input, substitute an equal-norm random eight-dimensional code, or patch the donor activation directly into the recipient subspace without source-to-recipient alignment. The last is capacity-constrained subspace patching, not unrestricted replacement of the entire activation state. Capacity accounting fixes eight real degrees of freedom, rank eight, float32 writes and one hook write per trial.

Aligned interchange exceeds shuffled-coefficient and equal-norm controls on mean displacement, but not wrong-mode or ordinary patching. Stable aligned no-change preservation is also low, approximately 0.02–0.08 across pairs. The recorded claim status is therefore indeterminate: these conditions do not jointly establish a shared causal abstraction. They also do not prove different algorithms or exclude a better alignment, basis or intervention site. Ordinary subspace patching outperforms the aligned intervention on the recorded mean displacement in both phases.

J STRONGER MATCHING AND INCOMPLETE SOFT COVERAGE

The original soft recipe uses unnormalised centred-logit MSE, temperature one, tolerance 10^{-4} and exact decisions. Later rescue runs change optimisation after observing failure and must be distinguished from the original policy. Seven eligible stable multiplication students are preserved in the recovered soft-success bank (Table 16). Their saved verification reports show complete 3,481/3,481 agreement; no additional discovery or phase contrast follows from eligibility alone.

Table 16: Seven preserved stable multiplication soft-student lineages. Values are saved eligibility measurements, not fresh forward evaluations performed for this manuscript reanalysis. Lineage labels distinguish independently tracked rescue attempts.

Rescue lineage	Teacher seed	Slot	Centred-logit MSE
Expansion case 0	6	0	0.00009729
Expansion case 1	9	0	0.00009646
Expansion case 2	10	0	0.00009513
Transfer case 2	1	0	0.00009696
Transfer case 3	1	1	0.00008596
Transfer case 6	4	0	0.00009590
Transfer case 7	4	1	0.00009995

Two other preserved candidate reports remain ineligible: an addition stable candidate has MSE 0.00281927 despite matching all decisions, and a multiplication pre candidate has MSE 0.00057334 with six wrong decisions. There is no secured matched pre/stable soft cohort comparable to the hard study. Historical reports of additional node-only successes are not counted without recoverable evidence. The narrow successes show feasibility for some stable cases; failures elsewhere do not establish impossibility or an intrinsic obstruction to soft distillation.

K EVIDENCE PROVENANCE AND REPRODUCIBILITY BOUNDARY

The experimental evidence cut-off is 20 September 2026. All analyses, including the dispersion and matched-transition reductions, use saved results. Phase I derives from checkpoint and search tables; Phase II derives from recovered addition and multiplication archives authenticated by their declared file sizes and hashes.

The accompanying audit reconstructs 252 canonical teacher/phase/condition/method cells, all eight primary paired contrasts, the four addition raw and Holm-adjusted p-values, the saved bootstrap intervals and forty fidelity-frontier summaries. It links 464 primary mask-inspection records to endpoint sizes and saved qualifying fidelities. These checks reproduce the saved-result reductions, not the original training or fresh checkpoint forward passes. The pre-submission stopping audit additionally links 550 published hard-attempt records to bound job identities and reproduces the selected canonical student endpoints. Its integer stopping fields were extracted from the empirical reference and independently checked using a second text-order parser; both the extracted inputs and their provenance are included. The exploratory attenuation interval is regenerated from the unchanged teacher-paired differences.

The figures and tables are rebuilt from included numerical inputs. Figure 1 uses the original 4,005 time-indexed accuracy observations from the five main-study runs; the 35 circuit-evaluation positions match the checkpoint registry exactly. Run-manifest and checkpoint identities, final metrics, phase landmarks and the reference's grouped curve summaries were checked before plotting. The methods supplement links all 752 completed canonical and sensitivity endpoint jobs to their bound search settings and source models, including the 241 hard-student selections. It includes the numerical configurations and their full extraction grids. Teacher identities, phase pairing and individual student values are preserved.

The additional `retraining/` directory supplies separate Phase I and Phase II runtime packages. A new local Phase II wrapper reuses the recorded numerical workers and exported job settings, restores hash-checked training splits, and executes selected endpoint dependencies without the original provider scheduler. It records new results as independent reproductions; it does not relabel them as historical evidence or replace missing phases and failed attempts. The supplied route covers all 752 published Phase II endpoint bindings, but does not automatically rerun the ancillary Fourier, null-calibration or soft-rescue campaigns. Phase I dataset generation and five-update training were tested. Phase II passed 41 inherited runtime tests and two new wrapper tests, including identical repeated small CPU pipelines, resumption and rejection of corrupted evidence. These checks do not establish full-trajectory or cross-hardware bitwise equivalence. Full retraining creates the weights and dense references needed for new recovery; the original trained checkpoints are not in this supplement. Technical fixtures, hardware benchmarks and unrecovered jobs are excluded from the scientific results.