DONOR-DERIVED TOKEN ORGANIZATION CHANGES WHAT A NEURAL NETWORK CAN REACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned weights are increasingly reused: generated, merged, transplanted. Successful reuse shows that a transferred tensor helps, not which of its properties suffice, because a learned tensor differs from a fresh one along many dimensions at once. We study that identification problem in a setting small enough for matched substitutions to be exact. A rank-4 token carrier is harvested from a network trained on modular addition and transplanted into an independently initialised recipient, which then trains to a frozen horizon; the endpoint is whether the recipient reaches a prespecified generalisation criterion at all. Donor carriers reach it in 45 of 64 donor–recipient cells, and a carrier preserving the donor's singular values, right basis and recipient coupling, with the token frame replaced by a centred random orthonormal frame, in none. On a fresh recipient cohort the effect reappears, 47 of 64, and further substitutions narrow what the carrier has to preserve. Projecting a donor carrier onto the conjugate Fourier modes holding 90% of its non-constant energy — one or two modes, a rank-2 object for five of the eight donors — still reaches the criterion in 35 of 64 cells. In a further prospectively specified test, a fully synthetic carrier matched donor by donor to those projections in rank, Fourier-support complexity, dominant-mode concentration and singular spectrum reaches it in 36 of 64 cells, against none for a norm-matched random carrier; the prespecified comparison between the projection and its synthetic analogue does not resolve a difference, which is not an equivalence result. Permuting that carrier's rows across token identities reaches it in none, although the permuted carrier lies in the same row-permutation orbit as the original, so every descriptor that ignores which row belongs to which token takes the same value on both. Equality under such descriptors therefore does not imply equality of what the recipient can subsequently reach. Each substitution, and the reading permitted of it, was fixed before its outcomes were observed.

1 INTRODUCTION

Learned parameters are increasingly treated as transferable objects. Generative methods produce parameter sets without training them, adapters move between base models, and merging composes networks that were never trained together. Each line evaluates a transfer by whether the result works.

That establishes reuse, not identification. A tensor taken from a trained network differs from a fresh one in its scale, its spectrum, its effective rank, its conditioning and the arrangement of its rows, all at once, and at model scale those properties cannot be varied one at a time. We therefore ask the question where it can be answered exactly: in a small algorithmic setting where a substitution can be built that matches a learned tensor on every property one chooses to name.

A donor network is trained on modular addition modulo $p = 113$. A rank-4 carrier is harvested from its token embedding at a preregistered moment and transplanted into a freshly initialised recipient, which then trains to a frozen horizon of 50,000 steps. The endpoint is *reachability*: does the recipient arrive at a prespecified generalisation criterion at all, rather than how quickly it arrives at one it would have reached anyway.

Transplanted organization does change it. In a leakage-separated experiment whose specification was sealed before any carrier existed, donor carriers reach the criterion in 45 of 64 donor–recipient cells,

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 a substitution preserving the donor's singular values, right basis, norms, rank and recipient coupling — replacing only the token frame — reaches it in none. On eight fresh recipient initialisations the donor effect reappears at 47 of 64.

We then ask what can be removed while the effect survives. Projecting each donor carrier onto the conjugate Fourier modes that carry 90% of its non-constant energy leaves one or two modes, and a rank-2 object for five of the eight donors. That projection reaches the criterion in 35 of 64 cells.

The sharpest result comes from a different substitution. Permuting the donor carrier's rows across token identities takes reachability from 47 of 64 to none. The permuted carrier lies in *exactly* the same row-permutation orbit as the original, so every descriptor invariant to reassigning rows to tokens takes the same value on both.

> Equality under every descriptor that factors through the row-permutation quotient does not imply equality of what the recipient can subsequently reach.

We do not identify a mediator, and one competing explanation was tested rather than left standing. A synthetic carrier matched to each projection on four prescribed structural targets, with its frequencies, phases and orientation in carrier space drawn independently of the donor, is itself active. That narrows what the earlier synthetic comparison can be read as showing (§5).

Contributions. **(i)** A transplanted rank-4 learned carrier changes whether an independently initialised recipient reaches a generalisation criterion within a fixed horizon; a substitution matched on spectrum, right basis and coupling does not. **(ii)** The effect reappears on a disjoint recipient cohort, 45 of 64 to 47 of 64. **(iii)** Constructive compression and synthesis: a severe donor-derived Fourier projection retains substantial reachability, 35 of 64 and at rank 2 for five of the eight donors, and a synthetic carrier reconstructed from selected structural targets is also active against a norm-matched random reference, 36 of 64. **(iv)** Exact row-permutation-orbit equality is insufficient for reachability at this operating point, which rules out a class of descriptions rather than a list of statistics.

The three substitutions ask one question: what must a learned carrier preserve to keep its effect on subsequent learning? One preserves selected lower-order properties while replacing the token frame, one removes most of the carrier's content, and one changes nothing that a row-assignment-blind descriptor can see.

2 Related work

Grokking and its mechanisms. Prior work explains delayed generalization on algorithmic tasks through several mechanisms: the weight norm and the mismatch between train and test landscapes, implicit bias across training phases, circuit efficiency, and regularised properties beyond the Euclidean norm (Power et al. 2022; Liu, Michaud, and Tegmark 2023; Varma et al. 2023; Lyu et al. 2024; Tikeng Notsawo, Dumas, and Rabusseau 2025). For modular addition the trained solution is described mechanistically: the embedding maps tokens to sines and cosines at a sparse set of key frequencies (Nanda et al. 2023), and margin maximisation alone specifies Fourier features for this task (Morwani et al. 2024). Two accounts bear on our design. Kumar et al. (2024) tie the delay to the alignment between initial features and the target function; Mohamadi et al. (2024) show that in the kernel regime no permutation-equivariant model reaches small population error on modular addition without a constant fraction of the data, so a carrier that specifies which token maps where breaks an equivariance that summary statistics do not see.

Injected structure during training. Two recent studies give causal evidence, from a different intervention, that the content of a representation controls grokking. Howe (2026a) injects contrastive representational priors into a one-layer transformer learning modular addition at $p = 97$, holding the loss form, strength, class sizes and geometry fixed while varying only what the positives encode, and finds a graded outcome across the task's true structure, a coherent-but-wrong sibling and a random partition; a control that replays the resulting weight-norm trajectory onto plain cross-entropy does not reproduce it. Howe (2026b) characterises when such a prior works, reporting that a coherent prior built from a foreign feature family blocks generalization much as a random partition does, that a label-free invariance prior suffices, and that the prior's effect is concentrated early in training. Those

interventions change the training objective and act while the network learns. Ours happens once, at initialisation, and transplants a fixed object harvested from another network; the question here is which properties of that object have to be preserved.

The nearest neighbour. GrokTransfer (Xu et al. 2025) trains a weaker model, extracts its input embedding, and uses it to initialize a stronger one; delayed generalization disappears, with a proof on a synthetic task and demonstrations on modular addition and parity. Their controls already answer one objection: a randomly initialized factorization of the same low-rank form does not reproduce the effect, so the low-rank structure is not what does the work. They also compare hand-built Fourier embeddings at the same modulus and the same 25% training fraction, constructed from $k = 7$ frequencies. Their question is acceleration; ours is reachability at an operating point where a matched random carrier does not arrive at all. What their design does not contain is the substitution our question needs: a carrier preserving the donor-derived spectrum, the donor-derived right-side structure and the recipient coupling while replacing only the donor-specific token organization.

Permutation symmetry and transferred parts. Git Re-Basin (Ainsworth, Hayase, and Srinivasa 2023) aligns hidden-unit permutations before merging, REPAIR (Jordan et al. 2023) shows alignment alone is insufficient, and mechanistic mode connectivity (Lubana et al. 2023) defines similarity through shared invariances. That literature makes a tighter substitution salient than a random frame: permuting the donor carrier across token identities, which we run. Li et al. (2026) move a frozen external memory between backbones and report that its learned content and correct addressing both matter, but that the transferred table becomes useful only through a reader aligned to the target model. Grant (2025) asks when two systems perform a task the same way and answers it causally, by transferring activity between networks and reading functional similarity off the resulting behaviour rather than off a correlation. Weight generation (Bedionita et al. 2024), task arithmetic (Ortiz-Jimenez, Favero, and Frossard 2023) and merging (Crisostomi 2026) evaluate transfer by whether the result works; none isolates which property of the transferred tensor makes it work, because at their scale the candidates are not separable.

Interpretability has instruments for the identification question we do not use here. Causal tracing and model editing localise a behaviour to components of a trained network (Meng et al. 2022), and causal abstraction formalises when a high-level variable stands in a causal relation to a network's internals (Geiger et al. 2021). Both intervene on a network while it runs and ask what its current behaviour depends on. Our intervention happens once, at initialisation, and the endpoint is what the recipient later becomes able to reach, so those instruments would answer a different question about a different object. They are the natural tools for probing what the recipient's internal state becomes, which our endpoint does not do.

3 Experimental design

Task, carrier and endpoint. Modular addition modulo $p = 113$ over all 12,769 ordered pairs. A single frozen permutation fixes 25% of the pairs for training; the complement is partitioned so that donor selection and recipient evaluation use disjoint, certified slices (Appendix A). The recipient is a bias-free three-layer network whose input embedding is not free but factorised, $E = AB$ with $A \in \mathbb{R}^{113\times4}$ and $B \in \mathbb{R}^{4\times128}$, so the transplant occupies a rank-4 interface. A carries the content under test; B is the recipient's own interface, identical across arms. Training is AdamW, full batch, float32, evaluated every 100 steps to a horizon of 50,000 with early stopping disabled. The criterion G2 is a sustained crossing: validation accuracy at or above 0.95 for three consecutive evaluations. G2 measures finite-horizon attainment, not persistent terminal occupancy (Appendix F).

Operating point. The cell $(\mathrm{wd} = 0.1, \mathrm{lr} = 0.01)$ was selected by a rule frozen before any donor existed, using donor-free arms only: the first cell in declared order at which a full-rank interface reaches the criterion and a matched random rank-4 carrier does not. The random arm's later failure rate is therefore not an independent finding. What makes the regime informative is the other half of the rule: a full-rank interface does reach the criterion there, so the task is learnable at this cell.

Arms. Every carrier is token-centred and scaled to the same Frobenius norm, and within each donor–recipient pair the recipient interface and downstream weights are cloned bitwise across arms, so the arms differ only in the content of A.

arm	carrier	role
DD	the donor's harvested carrier	the object under test
GM	donor spectrum, right basis and coupling preserved; token frame replaced by a centred random orthonormal frame	primary substitution, sealed before any GM carrier existed
RN	matched on norms alone	floor reference
FP	the donor's carrier projected onto the modes carrying 90% of its non-constant Fourier energy	what can be removed
FS	a synthetic Fourier carrier at seven frequencies fixed before any donor was trained, compressed to rank 4	synthetic structural prior
PD	a uniform row permutation of the donor carrier	unlabelled-geometry test
FSM	a synthetic carrier matched, donor by donor, to that donor's projection in rank, Fourier-support complexity, dominant-mode concentration and singular spectrum, with frequencies, phases and carrier-space orientation drawn independently	what a reconstruction from structural targets can do

The first three arms form the primary experiment on one recipient cohort; the first six form the follow-up on a fresh cohort of eight recipient initialisations disjoint from every earlier one, 8×8 per arm, 384 units. The last arm belongs to a second follow-up on another disjoint cohort of eight initialisations, run with DD, FP, FSM and RN, 256 units; its specification was written after the first follow-up's results were read and frozen before any of its own carriers existed, and it is labelled as such throughout. The follow-up reuses the same donors, so it replicates in the recipient direction only. Its provenance is mixed and stated as such: the three added constructions were defined in a specification written before any donor used here was trained, while the execution plan — cohort, endpoint, inference, reporting — was sealed after the primary result existed and before any follow-up carrier was generated (Appendices C and D).

Inference. The estimand is the paired difference in reachability, $Z_{dr} = Y^a_{dr} - Y^b_{dr}$ on the 8×8 panel. The primary test is a restricted wild cluster bootstrap-t with two-way clustering on donor and recipient, bootstrapping the donor dimension; with eight clusters the implementation enumerates all 2^8 Rademacher sign vectors rather than sampling, and sign symmetry leaves $2^7 = 128$ distinct p-values, so the finest reportable resolution is $1/128 = 0.0078125$. A donor-level t-test and a two-way CV1 interval are prespecified secondaries; the interval is an uncertainty summary, not a second test. Everything about the implementation, its certification, and the follow-up's Holm families is in Appendix B.

4 Learned organization changes reachability, and it replicates

Donor carriers reached the criterion in 45 of 64 cells. Neither control produced a single criterion-reaching unit in 64 (Figure 1). For the primary contrast, $\hat{\Delta}_U = 0.703$, the wild cluster bootstrap gives $p < 0.0078$, the donor-level test $t = 4.83$ on 7 degrees of freedom ($p = 0.00095$), and the secondary interval $[0.376, 1.030]$, which lies entirely above the preregistered substantive threshold of 0.25.

On the fresh recipient cohort the donor arm reached the criterion in 47 of 64 cells, against none for RN and none for GM; both replication contrasts are rejected at Holm-adjusted $p = 0.016$ with interval $[0.457, 1.012]$. This is a replication in the recipient direction, with the same eight donors, and it says nothing about donors in general.

The same comparison was run at a second operating cell, chosen from donor-free records without reading either of its arms, with the recipient's weight decay raised thirtyfold to 3.0. The donor arm reached the criterion in 8 of 64 cells and the matched substitution in 25 of 64, so the observed difference runs the other way at this cell ($\hat{\Delta} = -0.266$, interval $[-0.565, 0.033]$; the prespecified analysis does not resolve it). What the primary experiment supports is therefore specific to its operating cell, and the recipient's optimisation regime is part of the claim rather than a detail of it

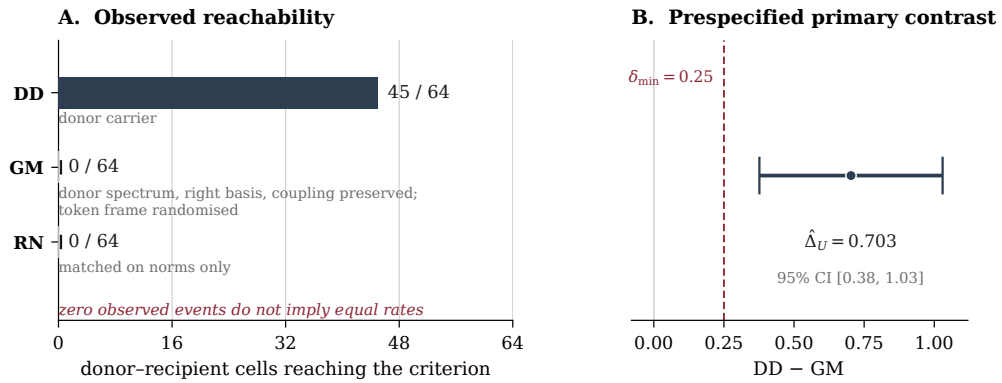

Figure 1: **A.** Criterion reachability across 64 donor–recipient cells. Zero observed events in GM and RN does not imply equal underlying rates; the comparison between them is censored at the floor. **B.** The prespecified primary contrast against the preregistered substantive threshold.

Table 1: Follow-up: reachability per arm (64 cells each), and the prespecified contrasts. p is Holm-adjusted within its family; a raw value with zero exceedances enters Holm at $1/128$.

	RN	FS	FP	DD	PD	GM
cells reaching the criterion	0	0	35	47	0	0

	contrast	$\hat{\Delta}$	p		95% interval
R1	DD − RN	0.734	0.016	rejected	$[0.457,\ 1.012]$
R2	DD − GM	0.734	0.016	rejected	$[0.457,\ 1.012]$
L1	FS − RN	0.000	—	not tested	zero events in both arms
L2	FP − FS	0.547	0.023	rejected	$[0.257,\ 0.837]$
L3	DD − FP	0.188	0.188	not rejected	$[-0.087,\ 0.462]$
A1	DD − PD	0.734	0.008	rejected	$[0.457,\ 1.012]$

(Appendix A). The norm-matched random carrier reached the criterion in 18 of 64 cells at this cell, where the donor-free screen that made the cell eligible had recorded none in three calibration seeds. A screen of that size was never a claim about a rate, and the primary contrast above is unaffected by it.

Two readings that the design does not support are worth stating once. The GM–RN comparison is censored: at $n = 64$ a rate of 0.03 leaves an arm empty 14% of the time, so zero against zero does not establish equal rates. And the four zero-event arms of the follow-up are not four independent failures — they are distinct interventions evaluated in a regime selected, before any donor existed, partly because an unstructured rank-4 carrier fails in it. That selection was made on the random arm alone, so it does not make the other three zeroes automatic.

4.1 What the criterion measures

Reachability is attainment within a horizon, and the follow-through matters for how the primary number should be read. Of the 45 donor runs that reached the criterion, 43 later fell below it at least once; 11 met the same three-evaluation criterion at the end of training, and 2 stayed at or above threshold at every evaluation from 300 steps after attainment onward. Those crossings are not an artefact of a threshold: across 1,317 observed transitions the summed distance of the two adjacent evaluations from the 0.95 boundary had median 0.194, with 6.3% within 0.01 and 79.6% beyond 0.05 (Appendix F). G2 therefore denotes finite-horizon reachability, not persistent terminal occupancy, and the paper's claims are stated in those terms.

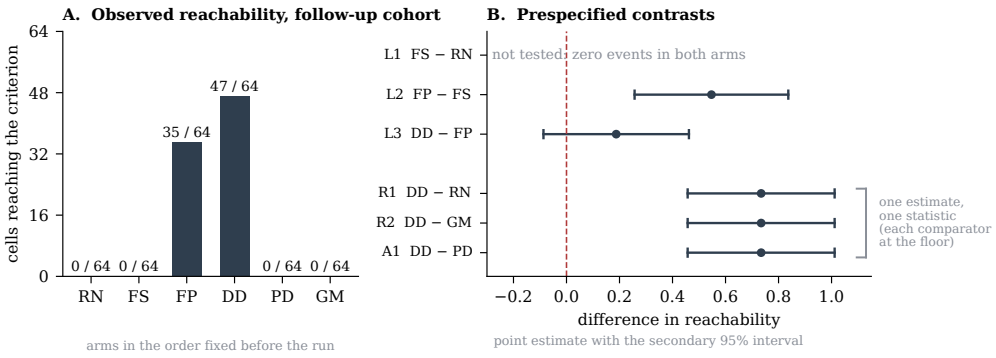

Figure 2: The follow-up reproduces the transplant effect on a fresh recipient cohort and narrows what the transplanted carrier has to keep. **A.** Criterion reachability, 64 cells per arm, in the order fixed before the run. RN, FS, PD and GM each produced zero events and are not ordered against one another. **B.** The prespecified contrasts with secondary 95% intervals. The bracketed group at the bottom shares one estimate and one test statistic, because each of its comparators sits at the floor: three interventions read against that floor, not three independent replications. L1 is not tested, both of its arms being empty.

5 What the carrier has to keep

5.1 A severe donor-derived compression still reaches the criterion

Taking the DFT of a token-centred carrier along the token axis, excluding the constant mode and combining conjugate pairs, let m_{90} be the number of modes needed to reach 90% of the energy. Every donor carrier here has $m_{90} \in \{1, 2\}$, with the dominant mode holding 71–94% of the non-constant energy and the two leading singular values nearly equal, as a single cosine–sine pair would give; the frequency differs between donors (Appendix A). The frozen diagnostic's random-matrix reference for m_{90} at this shape, from 20,000 draws taken before any donor was trained, is 40–47.

The FP arm keeps only those modes, discarding the rest and, for the five donors with $m_{90} = 1$, two of the four dimensions. It reached the criterion in 35 of 64 cells, against none for the synthetic Fourier carrier committed in advance; the registered contrast is rejected (Figure 2 and Table 1). What follows is that the donor-derived realisation outperformed that prespecified synthetic realisation. What does not follow is that donor specificity as such is what matters, or that these frequencies are privileged: the synthetic carrier differs from the donor carriers in which frequencies it uses, in concentration ($m_{90} = 6$ against 1 or 2) and in spectral shape at once.

The observed rates order as FS $<$ FP $<$ DD, but the contrast between the donor carrier and its own projection did not resolve, and its interval is wide.

5.2 A matched synthetic carrier is also active

A low-rank, concentrated synthetic carrier was the competing explanation the arm above cannot test, and it was tested. On a further disjoint recipient cohort, a synthetic carrier matched to each donor's projection in rank, Fourier-support complexity, dominant-mode concentration and singular spectrum — with its frequencies, phases and carrier-space orientation drawn independently of the donor — reached the criterion in 36 of 64 cells, against none for the norm-matched random carrier. The prespecified secondary contrast is rejected ($\hat{\Delta} = 0.563$, $p < 0.0078$, interval $[0.283, 0.842]$). The projection reached 35 of 64 cells in the same cohort, and the prespecified primary contrast between the projection and its synthetic analogue did not resolve a difference ($\hat{\Delta} = -0.016$, $p = 0.693$, interval $[-0.081, 0.050]$). Non-rejection here is not evidence of equivalence.

This narrows the reading of the earlier comparison. The projection still exceeds the seven-frequency synthetic carrier of the first follow-up, but that difference can no longer be attributed to the carrier's being donor-derived as such: a construction matched only on four prescribed structural targets of

the projection, drawing everything else independently, is itself active, and the prespecified analysis does not resolve it from the projection. Which matched property, or combination of them, carries the effect is not identified; the two carriers differ in frequency, phase and carrier-space orientation at once. The donor arm of that cohort is reported as an observed rate, 51 of 64, and enters no test, so no comparison with it is made here.

5.3 Exact unlabelled geometry is not enough

The PD arm permutes the rows of the donor carrier across token identities. It preserves the spectrum, the right singular basis, the multiset of row norms and the multiset of recipient-facing pairwise distances — and it reached the criterion in none of the 64 cells, against 47 for the carrier it was built from.

The strength of that comparison is not that a list of statistics was matched. It is that the two carriers are the same point of a quotient. Writing A for the donor carrier and P_ς for a permutation matrix, A and $P_\varsigma A$ lie in the same S_p-orbit exactly, not to a tolerance. Any descriptor that factors through that quotient — any descriptor which ignores the assignment of rows to tokens — therefore takes the same value on both, and the outcomes differ by 47 cells.

> Equality under every descriptor that factors through the row-permutation quotient does not imply equality of what the recipient can subsequently reach.

The experiment therefore rules out a class of descriptions, not a particular statistic; similarity measures built on such invariances cannot on their own determine reachability here. What it does not do is isolate token assignment: a uniform permutation also destroys the correspondence between the preserved geometry and the modular relations among tokens.

6 Where the substitutions sit

The four substitutions form a ladder of orbits rather than an unordered set of controls. With $A = U\Sigma V^\top$ token-centred, consider three groups acting on its rows,

$$\mathrm{AGL}(1,p) \;\subset\; S_p \;\subset\; O(p)\big|_{\mathbf{1}^\perp},$$

the affine relabellings $x \mapsto ux + v$ of $\mathbb{Z}_p$ inside the permutations inside the orthogonal maps of the centred subspace. GM is QA for an orthogonal Q by construction, up to the tolerances its specification enforces on the replacement frame. PD is $P_\varsigma A$ exactly, with ς uniform on S_p: of the $113!$ permutations the $113 \cdot 112 = 12{,}656$ affine ones are the only task-preserving ones (Appendix E), so a uniform draw is task-compatible with probability about 10^{-180}. This is organising mathematics, not a contribution; the statements are elementary and are collected in the appendix because §5 uses them.

Two consequences are worth carrying. The first is the exclusion above: narrowing the allowed transformations from the orthogonal orbit to the row-permutation orbit adds the row-norm and pairwise-distance multisets to what is preserved, and did not restore observed reachability in this experiment. The second concerns the rung the experiments leave open. Relabelling tokens by $x \mapsto ux + v$ moves a carrier's dominant Fourier mode from k to uk while leaving its concentration, spectrum and singular values unchanged, and the relabelled problem is the original problem with the data partition relabelled. For a carrier drawn independently of the partition, the identity of its non-zero dominant frequency can therefore affect reachability only through the induced relabelling of that partition. Our donor carriers do not satisfy that independence condition: they were trained, and their harvest epochs selected, on slices of the same frozen partition (Appendix E). The argument therefore constrains, rather than eliminates, frequency identity as an explanation of the observed runs, and it does so by argument rather than by another arm. Before the matched-synthetic experiment, that left concentration, spectral shape and the interaction with a finite partition as the immediate alternatives. The matched carrier subsequently holds the first two at the prescribed resolution while drawing frequency, phase and carrier-space orientation independently, and is active; it does not identify which matched property carries the effect, and it does not remove the dependence on the fixed partition.

7 DISCUSSION

The phenomenon reported here is positive: a transplanted learned carrier changes what a fresh network reaches, on disjoint recipient cohorts, against substitutions built to be as favourable as the design allows. What remains unresolved is one level down — which property of that carrier carries the effect — and the paper is explicit about that separation throughout.

What the experiments support. Viewed together, the substitutions form one identification sequence: progressively different notions of what it means to preserve a learned carrier are tested against the same subsequent-learning endpoint. A compact learned carrier, transplanted at initialisation, changes what an independently initialised recipient can reach within a fixed horizon, on disjoint cohorts of recipients. A severe donor-derived Fourier compression still reaches the criterion in 35 of 64 cells, though the design does not resolve that arm from the donor's own carrier in either direction. Equality under descriptions that ignore row-to-token assignment does not survive at all.

The ladder is constructive as well as eliminative: substantial reachability survives not only the projection of a learned carrier but the construction of a new one from a small set of structural targets. That same result narrows an earlier reading of ours — the advantage of the projection over the diffuse seven-frequency synthetic carrier cannot be assigned to donor origin itself.

What they do not. No mediator is identified: each substitution alters several properties at once. The matched synthetic carrier shows that a carrier constructed from the four prescribed structural targets can be active against the random reference, but the comparison between the projection and that carrier was a superiority test, not an equivalence test. Its non-rejection therefore neither establishes those targets as sufficient nor shows that donor-derived realisation is irrelevant; the two carriers also differ in frequency, phase and carrier-space orientation together. We do not show that the recipient reuses the donor's organization rather than becoming able to find its own; the endpoint is what it reaches, not how. And G2 is attainment within a horizon: of the 45 donor runs that reached it, 43 later fell below the criterion at least once, and 11 met the same three-evaluation criterion at the end of training (Appendix F).

Scope. One task, one architecture family, one operating cell — selected, before any donor existed, as a point where a matched random carrier fails — and a second at which the observed ordering is reversed, one data partition, and the same donor block across the replication. How much reachability varies across independently drawn partitions has not been measured. A confirmation at the preregistered effect size is not feasible under the frozen resource envelope, and that verdict stands beside the result rather than being replaced by it.

What would move this forward. A second operating cell, which needs no new apparatus, is the largest open boundary. Separating the matched descriptors from one another — a synthetic carrier matching all of them but one, run against the full match — is what would turn the non-resolution reported in §5 into an identification. Whether the recipient reuses the donor's organization or merely becomes able to find its own is a different measurement problem again: it requires reading the recipient's internal state rather than its endpoint, and the stitching and alignment literature has the instruments for it. A permutation drawn from the automorphisms of the token group would separate relabelling from the loss of the modular relations that a uniform permutation also destroys.

REPRODUCIBILITY STATEMENT

The supplementary archive contains the frozen specifications and their amendments in the order they were sealed, the resolved configuration of every run, the code that trained donors and recipients, the carrier generators and certificates, every analyzer, and the per-unit result records for every experiment reported here. Running `verify.py` in that archive recomputes the arm counts, the terminal-state counts and the contrast estimates directly from those records using only the Python standard library; the interval estimates and p-values additionally require R with `fixest` and `fwildclusterboot`. Each certificate records the SHA-256 of every carrier that was trained on, so a regenerated carrier can be checked against the one used, and Appendix G lists the chain those hashes belong to. Donor checkpoints and recipient trajectories are not included, for size; the configurations fix the seeds,

the operating cell, the horizon and the evaluation cadence needed to regenerate them. Donor and recipient identifiers in the released records are replaced by labels under an unpublished permutation, which preserves every statistic reported here, including the donor-clustered inference, while keeping per-donor outcomes unattributable; the archive documents that transformation, records release-side hashes and provenance sufficient to check integrity, and publishes no crosswalk that would reconnect the independently relabelled namespaces. A linkage audit that looks for such a crosswalk is included and can be re-run.

USE OF AI TOOLS

Generative AI tools assisted with drafting and editing, literature retrieval, research ideation and critique, mathematical reasoning, and software development. AI-assisted text and literature claims were reviewed by the authors against the underlying artifacts and primary sources. AI-assisted code included analysis, execution, and certificate-checking scripts; these were tested against frozen specifications and targeted failure cases, and the released analyzers refuse on specification or integrity mismatches. AI tools also assisted in developing the mathematical statements in Appendix E; the authors checked the arguments directly, including exhaustive verification at small p and numerical checks at $p = 113$, as documented there.

The authors made and approved the scientific decisions, including the experimental programme, specifications, endpoints, inference procedures, and decision rules, and take responsibility for all claims, code, analyses, and mathematical statements in this work. No experimental outcome was selected, altered, or discarded on the basis of AI output.

REFERENCES

Ainsworth, Samuel K., Jonathan Hayase, and Siddhartha Srinivasa. 2023. "Git Re-Basin: Merging Models Modulo Permutation Symmetries." In *International Conference on Learning Representations (ICLR)*.

Bedionita, Soro, Bruno Andreis, Hayeon Lee, Wonyong Jeong, Song Chong, Frank Hutter, and Sung Ju Hwang. 2024. "Diffusion-Based Neural Network Weights Generation."

Crisostomi, Donato. 2026. "Model Merging: Foundations and Algorithms." PhD thesis, Sapienza University of Rome.

Geiger, Atticus, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. "Causal Abstractions of Neural Networks." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Grant, Satchel. 2025. "Model Alignment Search."

Howe, Gunner Levi. 2026a. "Structure-Specific Representational Priors Causally Control the Grokking Delay." arXiv:2607.04333.

Howe, Gunner Levi. 2026b. "What Makes a Representational Prior Work? Feature Families, Label-Free Invariances, and Critical Windows in Grokking." arXiv:2607.12735.

Jordan, Keller, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. 2023. "REPAIR: REnormalizing Permuted Activations for Interpolation Repair."

Kumar, Tanishq, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. 2024. "Grokking as the Transition from Lazy to Rich Training Dynamics." In *International Conference on Learning Representations (ICLR)*.

Li, Mingyuan, Guangsheng Yu, Xu Wang, and Shaoxiong Ji. 2026. "Cross-Model Memory Transfer via Target-Side Reader Adaptation."

Liu, Ziming, Eric J. Michaud, and Max Tegmark. 2023. "Omnigrok: Grokking Beyond Algorithmic Data." In *International Conference on Learning Representations (ICLR)*.

Lubana, Ekdeep Singh, Eric J. Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. 2023. "Mechanistic Mode Connectivity."

Lyu, Kaifeng, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. 2024. "Dichotomy of Early and Late Phase Implicit Biases Can Provably Induce Grokking." In *International Conference on Learning Representations (ICLR)*.

Mohamadi, Mohamad Amin, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. 2024. "Why Do You Grok? A Theoretical Analysis of Grokking Modular Addition."

Morwani, Depen, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. 2024. "Feature Emergence via Margin Maximization: Case Studies in Algebraic Tasks." In *International Conference on Learning Representations (ICLR)*.

Nanda, Neel, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. "Progress Measures for Grokking via Mechanistic Interpretability." In *International Conference on Learning Representations (ICLR)*.

Ortiz-Jimenez, Guillermo, Alessandro Favero, and Pascal Frossard. 2023. "Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models."

Power, Alethea, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets."

Tikeng Notsawo, Pascal Jr, Guillaume Dumas, and Guillaume Rabusseau. 2025. "Grokking Beyond the Euclidean Norm of Model Parameters."

Varma, Vikrant, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. "Explaining Grokking Through Circuit Efficiency."

Xu, Zhiwei, Zhiyu Ni, Yixin Wang, and Wei Hu. 2025. "Let Me Grok for You: Accelerating Grokking via Embedding Transfer from a Weaker Model." In *International Conference on Learning Representations (ICLR)*.

A Design, partitions and provenance

Partitions. A single permutation under a frozen seed fixes the shared training set at 25% of the 12,769 pairs (3,192). The complement is carved, under a second frozen seed, into 1,915 recipient-validation pairs where the criterion is measured and which no donor sees, 1,915 donor-only validation pairs where donor validity and the harvest moment are decided, and 5,747 reserved pairs read by nobody. A certificate rebuilds the split through the runner's own construction and records cardinalities, order-independent index-set hashes, all six pairwise intersections, coverage as a set identity, and the proposition the replication exists for, $\mathrm{donor_val} \cap \mathrm{recipient_val} = \emptyset$. It also records what it does not certify: that donor training never read recipient-validation data is a claim about what the training loop consumes, for which disjoint index sets are necessary but not sufficient.

An earlier pilot, and why it is not the evidence. A first experiment decided the donor harvest moment on a held-out partition that, under the shared split seed, was exactly the recipient's validation set plus its reserved set. No label information moved — the donor never trained on those pairs and the labels are a deterministic function — but the choice of checkpoint was made on the space the recipient would later be scored on. The programme's own amendment had ruled this out prospectively; the defect was a failure to apply an existing rule, and it was found after that pilot had produced a strong result (donor 31 of 64 against 0 of 64). The main evidence in this paper is therefore the leakage-separated replication, and the pilot is reported rather than repaired.

Cohort and pairing. Donors and recipients are drawn from disjoint seed namespaces by a frozen traversal that admits the first eight seeds passing a preregistered validity gate and stops there; donors generated after the boundary were recorded as post-boundary and their descriptors were not consulted. The panel holds eight recipient initialisations, each paired with all eight donors; it does not hold 64 independent recipients. Within a pair, the recipient bundle is built once from its seed and cloned bitwise into every arm rather than re-drawn, and before training each pair is verified: interface and downstream weights identical across arms by hash, carriers differing, and both carrier norms at their analytic targets.

Carrier constants. $c_A = 21.158474$ and $c_E = 119.726617$ are the expected Frobenius norms a standard initialisation would give an unfactorised embedding of the same shape, so a transplanted carrier enters at the scale the recipient would otherwise have had; with Q orthonormal, $B = s_B Q^\top$ and $s_B = c_E/c_A$ make $\|AB\|_F = c_E$ exactly whenever $\|A\|_F = c_A$.

The substitution. For the carrier actually consumed, one decomposition gives $A = UR$ with $R = U^\top A$, and with the recipient interface B_r the coupling is $K = RB_r$. The substitution replaces U by a frame U_G drawn inside the subspace orthogonal to the all-ones vector, so the carrier is centred by construction; building in the full space and centring afterwards would change the singular values, because centring is a projection. R is taken from the same decomposition rather than re-materialised, since a second decomposition may return different signs or column order. The right factor is not a gauge freedom here: B_r is fixed and shared bitwise across arms.

Fourier description of the donor carriers. Taking the DFT along the token axis, excluding the constant mode and combining conjugate pairs:

donor	t_D	m_{90}	modes	dominant share	σ_2/σ_1
1	550	2	1, 37	0.819	0.992
2	2270	1	32	0.933	0.954
3	800	1	20	0.937	0.914
4	2900	1	54	0.910	0.970
5	580	2	28, 42	0.895	0.882
6	420	2	25, 32	0.713	0.921
7	1400	1	40	0.934	0.979
8	830	1	25	0.938	0.879

The synthetic arm uses seven frequencies fixed before any donor was trained, compressed to rank 4, which is why its energy is spread ($m_{90} = 6$) and its spectrum flat: a carrier of k balanced conjugate pairs has rank $2k$, so a rank-4 interface cannot carry seven of them.

A property deliberately left unmatched. For fixed coupling, pairwise token distances in the recipient-facing space are a functional of the token frame. Diagnostic analysis before the substitution was sealed found that donor carriers have a systematically denser lower tail of normalised pairwise distances than centred random frames with the same preserved factors: the donor value of (first percentile)/(median) lies below all 200 random draws for all eight donors, with the two ranges disjoint, 0.115–0.246 against 0.210–0.307. Minimum distance gave no evidence of pathological collision. The substitution does not match that structure and was not modified to match it: it is a consequence and possible mediator of replacing the frame, and matching it would condition on a mediator of the path under test.

A second operating cell. The recipient's weight decay was changed from 0.1 to 3.0, with the learning rate, the criterion, the horizon, the evaluation cadence, the data partition, the architecture and the pairing held as before, and the eight transplanted carriers byte-identical to those of the primary experiment. The cell was selected after the primary result was known but before any of its own outcomes existed, from the donor-free calibration records alone: eligibility was inherited from the frozen 50,000-step screen, which requires the full-rank interface to reach the criterion for all three calibration seeds and the norm-matched random carrier for none, and among the eligible cells a rule written at that point took the largest single-axis weight-decay perturbation. Neither arm of the primary contrast was read at any stage of that selection. Eight fresh recipient initialisations, 192 units, and one prespecified primary contrast with one secondary, Holm-corrected within the family of two.

B INFERENCE

The primary test is a restricted wild cluster bootstrap-t on $Z \sim 1$ with two-way clustering on donor and recipient, bootstrapping the donor dimension, Rademacher weights and the null imposed. It is not a permutation test: the arms are constructed deterministically, and independent within-pair sign flips would generate configurations the experiment cannot produce, because the donor carrier is shared across recipients while a random carrier is not.

With eight clusters the implementation enumerates all $2^8 = 256$ sign vectors; sign symmetry leaves $2^7 = 128$ distinct p-values, so the finest attainable resolution is $1/128$. The package returns 0.0 on zero exceedances under a strict-tail convention; that is reported throughout as $p < 0.0078125$, and the raw artifacts are released as produced with the correction recorded rather than by regenerating them. Under exhaustive enumeration the bootstrap seed is inert, which was verified rather than assumed.

The implementation was certified against seven targeted failure modes before use, including whether the multiway clustering argument is honoured rather than silently accepted and ignored, and whether enumeration actually occurs at this cluster count; the multiway test uses a deterministic fixture in which both clustering dimensions carry structure by construction. A donor-level one-sample t-test on the eight per-donor mean differences and a two-way CV1 interval are prespecified secondaries. A third sensitivity was declared optional and never uniquely operationalised; writing an implementation for it after the data existed would have been inventing a test after seeing the outcome, so it is recorded

as not run. Donor-specific means were computed inside the first sensitivity but omitted from the analysis artifacts and not used for stratification or interpretation; that restricts their analytic use and is not blinding, since the execution runner displayed unit-level outcomes during the runs.

In the follow-up, contrasts are grouped into two prespecified families with Holm correction inside each — $\{\text{DD} - \text{RN}, \text{DD} - \text{GM}\}$ and $\{\text{FS} - \text{RN}, \text{FP} - \text{FS}, \text{DD} - \text{FP}\}$ — plus the standalone DD − PD, all one-sided at $\alpha = 0.05$. A contrast whose paired differences are all zero is not tested and is reported as unresolved. No equivalence margin was preregistered, so no contrast is read as showing equivalence, and the substantive threshold is applied only to the two replication contrasts.

Feasibility. A separate frozen procedure asked whether any design inside a preregistered resource ceiling could reach 80% power at the preregistered minimum effect of 0.25, under a calibrated conservative bound on the variance components. None of 32 candidate designs could; the best attainable analytic power was 0.66. The verdict concerns prospective power at that minimum effect under a conservative bound; it is not a power calculation at the observed effect near 0.70, and nothing in the frozen design was changed in response to it.

C THE FOLLOW-UP

Six carriers occupy the same rank-4 interface, with the recipient bundle cloned bitwise across arms within each pair; every carrier is token-centred and scaled to the same Frobenius norm. The projection arm inherits the rank of its projection, so for the five donors with $m_{90} = 1$ its rank is 2 rather than 4, a difference recorded before the follow-up ran. Carriers were generated and certified by a program that runs no optimizer step; the training consumer verifies the file hash, the loaded tensor and the parameter held at step 0 against the certificate, together with the recipient bundle, and unit outcomes were not displayed during execution. The analysis ran once, after all 384 units were complete.

Readings were fixed before the outcome, one per contrast and outcome, and the analyzer emits the activated rows with the result. Two examples, quoted from the frozen table: the row activated by a rejected FP − FS permits "the donor-derived Fourier realisation outperformed the prespecified synthetic realisation" and forbids "donor specificity matters" and any claim about privileged frequencies; the row activated by a rejected DD − PD permits "preserving the carrier's global geometry while permuting its assignment to token identities did not preserve the reachability advantage" and forbids "relational organization" in general and any claim about which token-level property matters.

D THE SECOND FOLLOW-UP: A MATCHED SYNTHETIC CARRIER

Its provenance is stated plainly: the construction was chosen after the first follow-up's outcomes were read, in response to an explanation those outcomes made visible, and the specification was frozen before any of its carriers existed. It is not a preregistration independent of the findings that motivated it.

Construction. For each donor, four prescribed structural targets of its projection are read and nothing else: the rank r, the Fourier-support complexity m_{90}, the dominant-mode share, and the singular values σ. Then, per donor–recipient pair and from a frozen generator, $r/2$ distinct frequency classes are drawn uniformly from the 56 non-zero classes, a phase is drawn per mode, and an orientation $Q \in \mathbb{R}^{4\times r}$ with orthonormal columns is drawn from the Haar measure; the carrier is $FR\,\mathrm{diag}(\sigma)\,Q^\top$, with F the normalised cosine–sine columns at the drawn frequencies and R the block rotation the phases define, token-centred and scaled like every other arm. One draw per pair; the generator contains no retry loop and refuses if one is present in its own source.

Copying the orientation and phases from the projection instead of drawing them would make the construction $A_{\mathrm{FSM}}(x) = FP(ux)$ for the drawn u whenever the projection is a single mode — an affine relabelling of the rows, which Appendix E shows has the same reachability distribution marginally over partitions. The orientation and the phases are therefore what this arm draws independently.

Gates. Before any optimizer step, and per pair: rank and m_{90} equal to the projection's, dominant share within 0.005 of it, singular values within $10^{-6}\sigma_1$, norm and centring within their tolerances. A

failure fails the whole cohort. The dominant share of this construction is $(\sigma_1^2 + \sigma_2^2)/\sum_i \sigma_i^2$, fixed by the spectrum alone and independent of the drawn phase and orientation, so the share gate is a check on the implementation rather than an acceptance test on the draw. One generator run was refused before this one, for padding rank-2 constructions with two spurious zero columns; the refused certificate is retained.

Inference and result. Paired on (donor, recipient), one-sided at $\alpha = 0.05$, with the inference block used throughout. One primary contrast, the projection against the synthetic carrier, and one secondary, the synthetic carrier against the random one; no multiplicity adjustment, there being a single primary test. The donor arm is an observed rate and enters no test. Observed: random 0 of 64, projection 35, synthetic 36, donor 51. Primary $\hat{\Delta} = -0.016$, $p = 0.693$, interval $[-0.081, 0.050]$: not resolved. Secondary $\hat{\Delta} = 0.563$, $p < 0.0078$, interval $[0.283, 0.842]$: rejected. The frozen table provides no reading for a non-rejected secondary with events in both arms; the analyzer records that no row applies rather than writing one.

E TOKEN RELABELLING: STATEMENTS AND PROOFS

Nothing here is new mathematics. The statements are collected because the constants they fix are otherwise easy to get wrong. Throughout, p is prime, the task is $z = x + y \bmod p$, the recipient reads class z off column z of an output matrix with no bias, and the carrier enters only through the rows of the embedding.

Task-preserving relabellings. Call a bijection π of $\mathbb{Z}_p$ task-preserving if some bijection ρ of the labels satisfies $\rho(x+y) = \pi(x)+\pi(y)$ for all x, y. Then π is task-preserving iff $\pi(x) = ux+v$ with $u \neq 0$, and $\rho(z) = uz + 2v$ is unique. *Proof.* If $\pi(x) = ux + v$ then $\pi(x) + \pi(y) = u(x + y) + 2v$, a function of $x + y$. Conversely $y = 0$ gives $\rho(x) = \pi(x) + \pi(0)$; substituting, $\psi(x) := \pi(x) - \pi(0)$ satisfies $\psi(x + y) = \psi(x) + \psi(y)$, so ψ is an endomorphism of $(\mathbb{Z}_p, +)$ and $\psi(x) = ux$ with $u = \psi(1)$; bijectivity forces $u \neq 0$. □ The group is $\mathrm{AGL}(1, p)$, of order $p(p - 1) = 12{,}656$ at $p = 113$. Note $\rho = \mathrm{id}$ forces $u = 1, v = 0$: no non-trivial token relabelling leaves the output labelling alone.

What a relabelling does to a run. Since tokens enter only through rows, $P_\pi(AB) = (P_\pi A)B$, so the action lives on the carrier. For affine π, a run with carrier $P_\pi A$ on partition (S, V) and a run with carrier A on $(\pi S, \pi V)$ have the same distribution of validation accuracy: evaluating the first at (x, y) is evaluating the second at $(\pi x, \pi y)$, and the relabelled targets are absorbed by permuting the columns of the output matrix. It needs three properties of this apparatus: no class-dependent bias, output columns identically distributed across classes, and an optimizer and weight decay acting elementwise. A tied embedding and unembedding, or an output bias, would break it.

No affine-invariant partition of this size. Acting on pairs by $(x, y) \mapsto (ux + v, uy + v)$, the affine group has exactly two orbits: the diagonal, of size p, and its complement, of size $p(p - 1)$. (The difference $y - x$ scales by u; given two off-diagonal pairs, $u = (y' - x')(y - x)^{-1}$ and $v = x' - ux$ carry one to the other.) Invariant subsets therefore have size 0, p, $p(p - 1)$ or p^2, so a training set of the size used here cannot be invariant, and a comparison between a carrier and its affine relabelling at a fixed partition is confounded with a change of partition.

The frequency action. For $\pi(x) = ux + v$ and $\omega = e^{2\pi i/p}$, $\widehat{P_\pi A}(k) = \omega^{ku^{-1}v}\, \widehat{A}(u^{-1}k)$, so magnitudes are preserved and a carrier concentrated at k_0 becomes one concentrated at uk_0. Since u ranges over $\mathbb{Z}_p^\times$, which acts transitively on the non-zero frequencies, any dominant frequency can be moved to any other with concentration, spectrum and singular values unchanged.

Partition-marginal invariance, and why it does not apply here. The partition is drawn by a uniform permutation of the p^2 indices, so its law is exchangeable and $(\pi S, \pi V) \overset{d}{=} (S, V)$. With the statement above this would give $R(P_\pi A, S, V) \overset{d}{=} R(A, S, V)$ marginally over partitions — but only for a carrier drawn independently of the partition. The donors here were trained, and their

harvest epochs selected, on slices of the same frozen partition, so the carrier is a function of it and the marginal statement does not describe these runs.

Checks. The characterisation of the task-preserving relabellings was verified by exhaustion at $p = 5$ and $p = 7$: of the 120 and 5040 permutations, exactly 20 and 42 are task-preserving, and they are exactly the affine families. The frequency action was verified numerically on random centred carriers at $p = 113$ for several (u, v), agreeing to 2×10^{-13}.

F SECONDARY DESCRIPTIVE ANALYSES

These were specified in a separate frozen audit of the existing data, each run once against a hashed specification, and they are descriptive: no test and no promotion rule.

Terminal and post-attainment states. Counts out of 64 cells per arm:

	DD	GM	RN
criterion reached within the horizon (primary endpoint)	45	0	0
training accuracy ≥ 0.95 sustained within the horizon	64	18	48
validation accuracy ≥ 0.95 at each of the final three evaluations	11	0	0
validation accuracy ≥ 0.95 at the final evaluation only	16	0	0
no evaluation below 0.95 from $t_{\mathsf{G2}} + 300$ to the horizon	2	–	–
final training accuracy below 0.02 (chance 0.0088)	21	11	24

The two controls differ on the training-accuracy landmark — 18 against 48, and in paired cells the random arm reached it where the substitution did not 33 times against 3 — so the substitution is not a neutral replacement for the random control in training dynamics. That does not account for the primary contrast, since the random arm reached that landmark in 48 cells and the criterion in none.

Post-attainment dynamics. Of the 45 donor runs that reached the criterion, 43 later fell below it at least once. Across 1,317 observed state transitions, seven runs accounted for 597 of the 673 downward transitions. The summed distance of the two adjacent evaluations from the 0.95 boundary had median 0.194; 6.3% of transitions were within 0.01 and 79.6% exceeded 0.05, so the crossings are not predominantly near-threshold chatter under this check. Pooled occupancy above the criterion after attainment was 0.606, with a per-run median of 0.57.

A prespecified prediction attempt. Whether the three evaluations immediately after the qualifying window carry information about terminal persistence was asked with two features — the least-squares slope across those evaluations and their minimum margin above 0.95 — and an L2 logistic model under leave-one-donor-out evaluation. The pooled Brier score was 0.236 against 0.223 for the fold-wise prevalence baseline, a skill score of -0.057.

An exploratory question, recorded as killed. Whether first-loss times concentrate when measured from attainment rather than on the absolute training clock was asked with a prespecified dispersion ratio; the ratio was 1.55, so the alignment did not concentrate them, and no candidate law was locked.

G ARTIFACT CHAIN

Every stage records the hash of its inputs, so a reported number can be traced to the donor that produced it. The released archive contains the frozen specifications and amendments, the resolved configurations, the split-identity and carrier certificates, the analyzers, and the per-unit records; identifiers in the records are relabelled under an unpublished permutation, as described in the reproducibility statement.

Four defects in this programme were found only because the chain existed: a cohort specification that was passed, parsed and then ignored while the generator used hard-coded constants; the selection leakage of Appendix A, found after a strong result was already in hand; an extinction diagnostic whose factorized-carrier predicate was applied only in the arm for which it was first written, so that in

the later arms a fallback branch returned a negative result with a "not measured" flag that the tracker discarded, recording it indistinguishably from a measured non-extinction; and a p-value resolution label contradicted by a warning printed beside it, the package reporting the finest resolution as $1/256$ while warning in the same run that sign symmetry leaves 128 distinct values. The affected fields are treated as non-measurements and enter no claim. A float32 norm diagnostic that underflows to zero for subnormal carriers was found in the same audit and is likewise not used. None of these would have surfaced from a code repository and a random seed.

The apparatus is support structure, not a contribution of this paper.