CURRICULUM ORDER CONTROLS GROKKING THROUGH ITS HAND-OFF WEIGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A neural network can generalize long after memorizing its training data; this late generalization is called grokking. A curriculum often helps by training the model on a chain of simpler tasks first, each starting from the weights the last one left. It is not known whether this chain passes on skills, or only a better initialization that another model could simply copy. We study a small transformer that solves a linear differential equation symbolically after a chain of negation, addition and reading off the solution's two coefficients. On the target alone, runs differing only in their random starting weights sometimes generalize and sometimes fail. Ordered easiest to hardest, the chain lifts average test accuracy from 0.48 to 0.88. And nearly every run now generalizes. But run in reverse order, ending on the easiest, the same tasks do no better than the target alone. To see which it is, we copy part of the chain's final weights into a fresh model trained only on the target. We split those weights in two: the token identity, which maps tokens in and out, and everything else. On its own, the token identity recovers 97% of the gain, although the fresh model has never seen the simpler tasks. And copying only the rest does worse than no curriculum at all. We repeat this on a second problem, symbolic differentiation, with its own chain, model and vocabulary. There the rest of the network carries the gain instead. So we find that a curriculum passes on an initialization, set by the last task before the target. The benefit of a curriculum is thus something one model can hand to another rather than earn by training.

1 INTRODUCTION

Power et al. (2022) first reported grokking: networks that, trained past the point of memorization (a perfect fit on the training set; Zhang et al., 2017), eventually transitioned to generalization. At small scale and low weight decay (0.1) the seed alone decides. With architecture, data, and budget held fixed, the outcome splits: a fraction of runs generalize while the rest plateau. This per-seed variability is commonly read as an initialization lottery (an irreducible property of where in weight space a run begins). If so, no control variable exists short of a change of architecture or loss: in practice one can only train more seeds and keep the ones that grok.

There is one: the order the tasks are presented in. It acts through the state the last precursor hands to the target. On symbolic solution (Lample & Charton, 2020) of a first-order linear ODE at low weight decay, training from a cold start reproduces the lottery across seeds. A sequential easy-to-hard curriculum of three precursor stages (compute-matched on target-stage gradient steps) compresses it to a tight single peak. A long last precursor that admits no generalizing solution in the target's tokens drives the outcome below cold start (§4.2). With the same three stages in reverse order, on identical content and budget, the cold-start distribution returns. In the ODE family the effect localizes to that hand-off (the splice of §4.3).

We therefore study the state that a schedule leaves behind. A curriculum trains a model on progressively harder tasks (Bengio et al., 2009; Wu et al., 2021), so the natural reading is that the target reuses computation the earlier stages built. In the ODE family, two interventions that share no mechanism say otherwise (§4.3, Appx. C.1). Splicing a curriculum run's hand-off into a cold initialization one group at a time (Fig. 1) puts the benefit in token identity, the input embedding and readout (W_E, W_U, b_U), which recover 97% of the cold-to-curriculum gap. Without them, the body (all other weights the curriculum trained) lands below cold start (§4.3, Fig. 4). Rewriting the

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

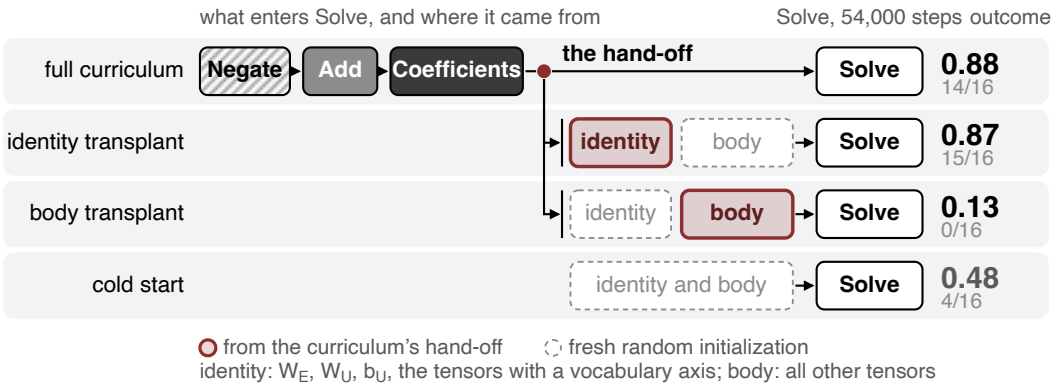


Figure 1: The design. The curriculum hands the target a state, and each transplant takes one parameter group from it and a fresh initialization for the other before the same 54,000 steps of Solve. Means over 16 seeds, with the seeds at or above 0.80.

same stages into a numeral alphabet the target never sees reproduces that boundary from the data side without touching a parameter: the curriculum learns the mathematics and, written in the wrong alphabet, transfers none of it.

The paper makes three contributions. First, stage order controls the per-seed distribution of a grokking lottery, and it does so through the state the last precursor hands over (§4.2). Second, a component-swap splice localizes that hand-off to one parameter group (the token identity in the ODE family), and a disjoint-alphabet curriculum draws the same boundary from the data side (§4.3). Third, the identity is a portable initialization: one seed's identity carries every other seed into the grokking basin (§4.4). A second family at another depth and vocabulary repeats the rule while reversing the carrier to the body (§4.5). The claim is about the per-seed *distribution* and what carries its control, not about raising a ceiling.

2 Background and Related Work

Stochastic grokking. Power et al. (2022) introduced grokking as delayed generalization on algorithmic tasks, and Nanda et al. (2023) tied it to structured circuits. Later work ties it to competition between memorizing and generalizing solutions (Wang et al., 2024) and to the geometry of multi-task training (Xu, 2026). Most work treats its per-seed variability at fixed budget as an initialization lottery. Ootani (2026) reframes it in a fully-tractable multi-seed study as a conditional, fragile phase transition: floating-point summation order alone flips a sixth of seeds. It turns on numerical nondeterminism and a task split (ours is stage order within one model on a shared vocabulary).

Curriculum and task ordering. Curriculum learning orders examples easy-to-hard to aid optimization (Bengio et al., 2009). Such schedules lower the sample complexity of compositional targets (Wang et al., 2025), improve generalization on reasoning targets (Abbe et al., 2023), reduce the capacity a small transformer needs (Both et al., 2026), and speed up language-model training when the stages are ordered prerequisite skills (Chen et al., 2023). Both et al. (2026) transfer numeral representations, the nearest antecedent of the identity we splice (without a complement arm showing what does not transfer). In two results, structural alignment with the target (not its difficulty) governs whether a model learns a hard task: reinforcement-learning post-training unlocks a target direct RL cannot when the earlier stage aligns structurally with the target family rather than merely in difficulty (Sun et al., 2026), and a structure-compatible representational prior causally controls the grokking delay on modular addition (Howe, 2026).

LeDoux (2026) comes closest, ordering *individual examples* within one task: on modular arithmetic, fixed example orderings reach high test accuracy from a fraction of the input space where an i.i.d. baseline does not. An adversarial ordering suppresses learning, and the Fourier features the model learns carry frequencies matching the ordering, so the sequence transmits information no single example contains. That mechanism cannot operate in our complementary regime: our unit is the

stage, we draw each stage's examples i.i.d. and shuffle them per seed, and the stages are separate datasets, so no signal can ride on the within-stage sequence.

Embedding transfer. The nearest method to our splice transfers an input embedding from a weaker model trained on the same target and removes the grokking delay (Xu et al., 2025), and embeddings are causally central to grokking through rare-token gradient stagnation (AlquBoj et al., 2025). We perform that operation but source it from a different and easier stage, and find the readout carries the effect as well as the embedding does, which transferring a representation of the target's inputs would not predict. Grokked tickets seed a fresh run with a sparse mask taken from a run that has generalized, and rule out weight norm with norm-matched controls (Minegishi et al., 2025); the splice moves a dense parameter group from an easier task and also runs its complement. A concurrent transplant study of arithmetic grokking (Gomezjurado Gonzalez, 2026) reports the same asymmetry within one task: a trained encoder moved into a fresh model accelerates grokking, while a trained decoder hurts. Our complement arm moves the whole body across a curriculum and withholds the token identity. Recombining components of two independently trained runs is the nearest methodological antecedent: Truong (2026) splits a tensor into magnitude and direction and finds direction donor-specific.

3 Experimental setup

Task and tokenizer. The curriculum trains three simpler stages and then the target, in this order, over one set of integer tokens:

$$\text{Negate}: a \mapsto -a, \qquad \text{Add}: (a, b) \mapsto a + b,$$

$$\text{Coefficients}: y' + ay = e^{bx} \mapsto (-a,\ a + b), \qquad \text{Solve}: y' + ay = e^{bx} \mapsto Ce^{-ax} + e^{bx}/(a + b).$$

Every example is a prefix-Polish token sequence, and the target generates its full symbolic solution after a separator. Coefficients's outputs are exactly the two coefficients that solution requires, a designed token overlap we test directly in §4.2. Fig. 6 (Appx. A.1) shows one example from each stage. Vocabulary: one vocabulary of 264 tokens for all stages, one per integer in $[-100, 100]$ plus the structural tokens. Split: a fixed 50/50 train/validation split, drawn once with seed 42, so the data partition is identical across model seeds.

Model and training. We use a decoder-only transformer (Vaswani et al., 2017) with 4 layers, 4 heads, $d = 128$, $d_{\text{ff}} = 512$ and context 64. Training: full-batch AdamW (Loshchilov & Hutter, 2019) at weight decay 0.1 unless noted, with learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and gradient clip 1.0. We evaluate exact-sequence accuracy on the held-out split every 1000 steps under teacher forcing and report the final value; since a sequence counts correct only if the argmax is right at every position, and feeding those same tokens back reproduces them, greedy free-running decoding returns the identical number on all 31 runs we can decode, as Appx. C.3 shows. Every run passes a build-time causality check over the full context, and trained checkpoints at both depths pass it with a leak of exactly zero, as Appx. C.4 reports.

Terms. A stage's *hand-off* is the parameter state it leaves the next: *structured* when the stage admits a generalizing solution in the target's tokens, *degenerate* when it admits only a memorized one. A *splice* copies one parameter group of a trained run into a fresh initialization, then trains the target as a cold run does. The groups partition the model into two: the *identity*, the input embedding and readout, and the *body*, everything else. The group whose splice moves the outcome is the *carrier*.

Conditions. The three baselines, named as the tables and figures name them, are *cold start*, which trains Solve alone for 54k steps; the *full curriculum*, which trains Negate, Add and Coefficients for 8k steps each and then Solve for 54k; and the *interleaved* mixture, which samples all stages jointly for 54k steps. All conditions except the 78k cold start are matched at 54k target-stage steps on Solve, and ablations hold that budget fixed and vary one factor at a time: for *ordering*, reverse, bookend and a random permutation of the easy stages; for the *penultimate stage*, single-precursor and skip variants. Two interventions act on the hand-off itself rather than on the schedule: the splice and a *disjoint-alphabet* curriculum, both in §4.3. Further controls isolate per-stage gradient mass

Table 1: The arms the argument rests on; all other conditions are in Appx. B. Gain share is the share of the gap between no curriculum and the tinted row that the arm recovers on shared seeds, from unrounded means.

Condition	accuracy, mean ± sd	seeds ≥ 0.80	gain share
ODE family: 4 layers, target Solve			
cold start	0.48 ± 0.31	4/16	+0.00
full curriculum: Negate, Add, Coefficients, then Solve	0.88 ± 0.05	14/16	+1.00
reverse order: Coefficients, Add, Negate, then Solve	0.50 ± 0.29	4/16	+0.05
Negate only for 24k steps, then Solve	0.12 ± 0.10	0/16	-0.91
identity transplant: W_E, W_U, b_U into a fresh body	0.87 ± 0.08	15/16	**+0.97**
body transplant: everything else, under a fresh identity	0.13 ± 0.19	0/16	-0.90
identity transplant from one donor seed into the other fifteen	0.90 ± 0.03	15/15	+1.07
disjoint alphabet, all precursors	0.07 ± 0.07	0/16	-1.04
Second family: 2 layers, target Differentiate			
cold start	0.66 ± 0.22	6/16	+0.00
three-stage ladder: Sum, Product-sum, then Differentiate	0.77 ± 0.20	8/16	+0.44
compatible precursor alone: Product-sum, then Differentiate	0.90 ± 0.11	14/16	+1.00
identity transplant	0.64 ± 0.18	3/16	-0.12
body transplant	0.85 ± 0.22	14/16	**+0.80**

with a weighted sampler, intermediate-stage label overlap with a disjoint-range Coefficients variant, the optimizer reset, weight decay, architecture at 8 heads and context 512, and experience replay. A second family, described in Appx. B.2, applies the same two splices on a 2-layer model over a 205-token vocabulary whose target, Differentiate, evaluates the derivative $f'(x_0) = 2ax_0 + b$, after two precursors: Sum, a three-term addition, and Product-sum, $ab+c$. We call a precursor compatible when it shares the target's rule shape in the target's own tokens, here a product plus a sum.

Metrics. In Tab. 1 the tinted row is each family's compressing schedule and the bold share marks the transplant that carries it. Reported: per-condition mean and population standard deviation of final accuracy and the grok rate, the share of seeds at or above 0.80, with Wilson intervals (Wilson, 1927); the appendix tables add the rates at 0.85 and 0.95. Tests: paired t (Student, 1908), permutation (Pitman, 1937) and McNemar exact (McNemar, 1947), over the seeds both conditions completed (sixteen seeds unless the arm's table gives another n; Appx. B).

4 Results

4.1 Cold start is a high-variance lottery; the full curriculum compresses it

Cold-start training on Solve at weight decay 0.1 spreads widely across sixteen seeds. The mean is 0.48, the standard deviation 0.31, and outcomes run from 0.04 to 0.86 (on identical architecture, data and budget). The full curriculum compresses that to a tight single peak (mean 0.88, standard deviation 0.05, 14/16 above 0.80; Tab. 1, Appx. B). The spread is not a fitting failure. Training accuracy reaches 1.000 within three evaluations in all 282 target-stage logs of the schedule and weight-decay arms, so a seed ending at 0.04 has memorized its equations, not failed to fit them. The standard deviation shrinks $6.4\times$ and all sixteen seeds improve, at paired $p = 2.8 \times 10^{-4}$. The ten seeds that pass 0.80 in only one arm all pass it with the curriculum, at exact McNemar $p = 0.002$ (Fig. 11).

The interleaved mixture does not compress it (0.59, 5/15 above 0.80), and its spread has a structural cause the sequential arms never face: every such seed's training accuracy ends at 0.927 (the ceiling that conflicting Coefficients and Solve labels impose), so it is a lottery over which label family wins, not over grokking. Dropping Coefficients removes the spread and yields the highest arm in the paper (0.95, 8/8): with no stage to conflict, every step hands the target a state a generalizing solution is reachable from, whereas a schedule with an order must arrange that in its last stage alone (Tab. 6, Appx. B.1). Budget does not explain the curriculum's effect from either direction: six single-precursor and reordered schedules spend a curriculum arm's same 78k yet span 0.83 down to 0.12 (Tab. 3), and cold start given the full 78k is still a lottery (0.53, 5/16).

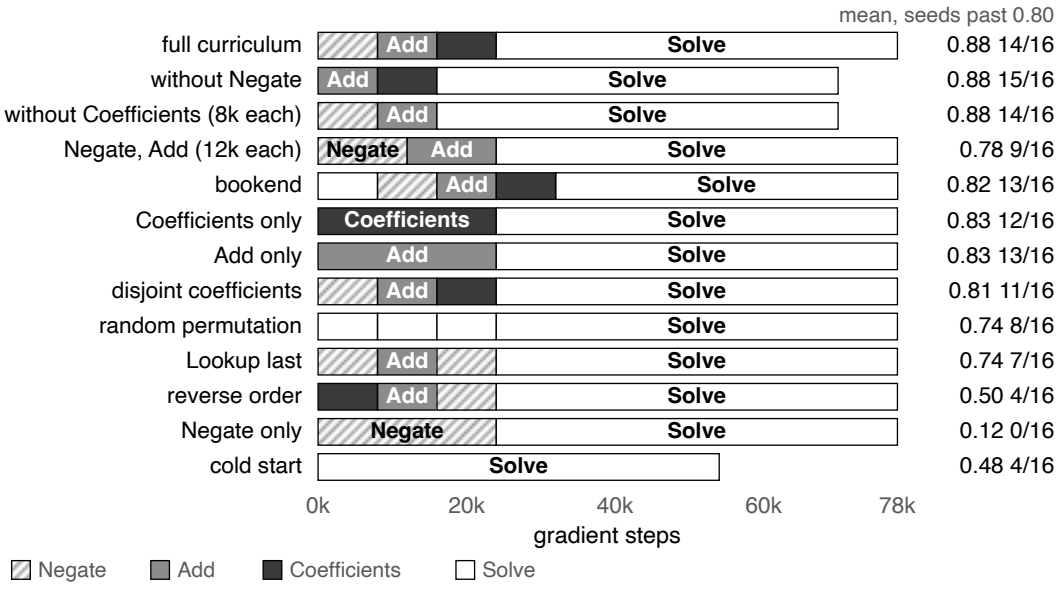


Figure 2: The schedules, drawn to scale in gradient steps. Each block is one stage, shaded by whether it admits a generalizing solution in Solve's own tokens; every schedule spends the same 54,000 steps on Solve.

4.2 Order controls the distribution by choosing which hand-off the target inherits

The ordering tests ask whether sequence matters or only that training ends on the target (Fig. 2 draws every schedule to scale). Reversed so that Negate comes last, the four stages lose the effect: reverse order is indistinguishable from cold start, 0.50 against 0.48, on identical stages and budget. A bookend order, with 8k steps of Solve before the easy stages and the remaining 46k after them, still compresses, and a random permutation of the easy stages is intermediate (Fig. 3). Two further controls: a sampler that gives each stage equal gradient mass does not compress, and keeping the optimizer state across stages leaves the curriculum's effect intact (Appx. B).

Order controls the per-seed distribution with the stage set fixed. Single-precursor experiments isolate which position carries it (Tab. 3). A structured stage immediately before the target compresses the lottery whether it is Coefficients or Add, as does a two-stage ramp ending on Add. A degenerate single-token precursor, Negate, in that position instead drives the outcome down to 0.12, an anti-curriculum effect. The registered size control is partial: a random lookup table over Add's 5,100 input pairs, in Coefficients's slot of the curriculum, reaches 0.74, above cold start yet short of the curriculum. The fall tracks the budget: 24,000 steps of Negate alone leave the target at 0.12, and the 8,000 that close reverse order, after Coefficients and Add, leave it at 0.50. What preceded the last precursor then sets how far above cold start a structured hand-off lands (Fig. 3).

The stage immediately before the target matters most. That rule refutes a monotone easy-to-hard ramp and explains the reverse-order failure, which ends on Negate. With the stages held fixed at Negate then Add, 8k steps each lands at 0.88 and 12k each at 0.78, a difference of level inside the structured side that falls short of significance.

Inside a single condition, the random-permutation arm tests the rule. It cycles through the six orders of the three easy stages across seeds, so its runs differ only in order, and the stage that lands immediately before the target alone reproduces the ordering. All six seeds that drew Coefficients last pass 0.80, two of the five on Add do, and none of the five on Negate. Coefficients beats Negate at Fisher exact $p = 0.002$ (Fisher, 1935), the rule's primary form, since condition, budget and content are fixed and only position varies with seed.

Structured and degenerate are measured properties of the stages. Measured on held-out data as each hands off, Add ends at 0.998 and Coefficients at 0.949 in the full sequence, while Negate ends at exactly 0.000 in all 325 placements that train it. What predicts the outcome among these stages is

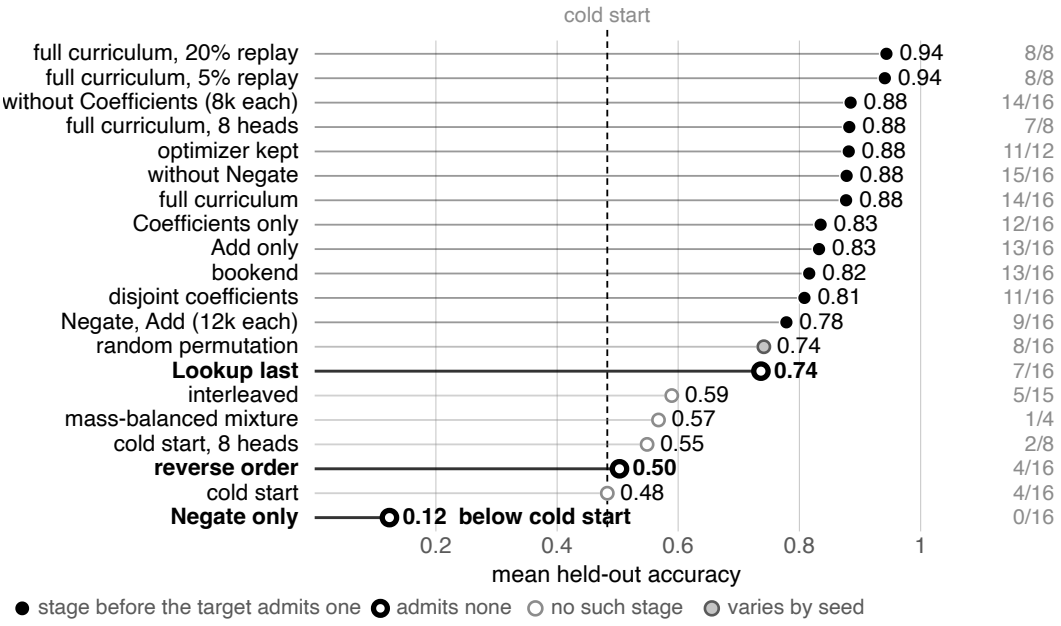


Figure 3: ODE-family schedules at weight decay 0.1, on one axis and ordered by outcome. Fill marks whether the stage trained immediately before Solve admits a generalizing solution in Solve's own tokens; the random-permutation arm cycles that stage across seeds.

whether the stage admits a generalizing solution, not how well it has learned one: trained first, as in Coefficients only, Coefficients hands off at 0.356 rather than 0.949 and still compresses the target. Coefficients compresses at the target's own 442 items, under a tenth of Add's 5,100 (Tab. 2).

The hand-off is not answer disclosure. Coefficients shares answer-key tokens with the target split, so the ordering result and the token-identity claim of §4.3, which rests on it, share one objection: a precursor hands over the answer rather than the alphabet it is written in. Removing Coefficients entirely leaves compression unchanged, at the full curriculum's 0.88. Keeping the Coefficients task but drawing its coefficients from $a \in [31, 60]$, so that no answer key overlaps the target split, still compresses, to 0.81. The effect survives removing every shared answer key while the numeral alphabet stays shared (Tab. 3).

4.3 Splicing the hand-off reproduces the effect

Every result above varies a schedule and reads an outcome. So we also test the hand-off directly (Fig. 1 draws the partition). We splice a curriculum run's parameters at the end of Coefficients into a cold initialization one group at a time, then train the target as a cold run does (seed, budget and optimizer fixed). The partition is structural: W_E, W_U and b_U are the only tensors with a vocabulary axis, so the positional embedding sits in the body.

The identity alone suffices, and its complement, the body, does not (Tab. 1, Fig. 4). The identity transplant reaches 0.87 with 15 of 16 seeds above 0.80, level with a full curriculum run, and recovers 97% of the gap between cold start and the curriculum. Either half of the identity suffices on its own, the readout at 0.88 and the embedding at 0.85 (Tab. 4). The body transplant reaches 0.13 with no seed above 0.80, below cold start. In this family the body carries nothing the target can use within the budget, and a positive control rules out the alternative reading, a warm-start penalty (Ash & Adams, 2020) in which a usable body cannot be addressed by a fresh identity. Freezing the body and training only the token identity, which leaves the identity free to adapt to it as Artetxe et al. (2020) do across languages, gives 0.005 over eight seeds against a pre-registered recovery bar of 0.70. The same test on the body of one cold seed that grokked on its own recovers it in 7 of 8 seeds (Appx. C.7). An identity can learn to address a body that generalizes, and finds nothing to address in the curriculum's. With everything training, the same body is no more portable than a fresh one: portability is the identity's, and where a body does carry, as in the second family of §4.5,

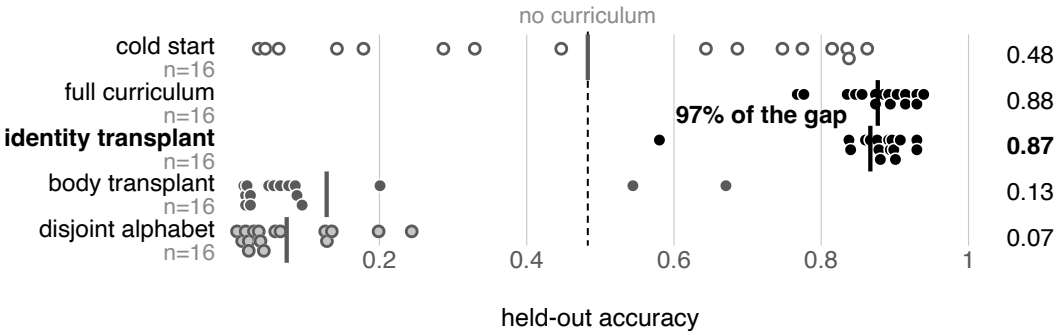


Figure 4: What carries the effect, seed by seed (means in Tab. 1). A full curriculum and a splice of only its token identity are indistinguishable; the complement and a curriculum learned in a disjoint alphabet both land far below cold start (dashed line).

the transplant reads it. The identity transplant reproduces the compression, not only the mean: all five seeds on which cold start collapses reach 0.84 or above (Appx. D.3).

Which hand-off is spliced decides the outcome, and two hand-offs from the same schedules show it from each side. Taken after the curriculum's 8,000 Negate steps instead of after Coefficients, with seed and procedure unchanged, the identity lands near cold start, 0.63 against 0.48, and the spread across seeds comes back with it. This identity sits below both a full curriculum run and the identity transplant. The Coefficients and 8k Negate identities differ in rank and scale in every seed (Appx. D.2). The 24k Negate-only schedule, which lands below cold start, localizes to the other group: its body under a fresh identity stays at the schedule's own 0.12, and its identity in a fresh body reaches 0.70, no worse than cold start.

Changing the alphabet instead of the parameters draws the same boundary. For the same test on the data, we rewrite the precursors into a disjoint numeral alphabet, so the curriculum learns identical functions on 201 numerals the target never sees, with parameter count matched by a control carrying the same unused rows (Appx. C.1). The precursor still learned, handing off at its control's accuracy. The curriculum learned the mathematics and transferred none of it: it reaches 0.07, with no seed above 0.25, against the control's 0.87, and lands below cold start as well. Its accuracy matches the body transplant's, reached by a harsher route: the target's own rows receive no gradient through the precursors, so weight decay shrinks them, yet a cold start with those rows pre-shrunk matches cold start (Appx. C.1).

4.4 The identity is a portable initialization

A published construction close to the splice, GrokTransfer (Xu et al., 2025), takes an input embedding from a weaker model trained on the target itself; we ran its embedding transfer and reran our splice on the same sixteen seeds, each seed's pair on one machine. It recovers 73% of the gap, against 95% for our splice. And five of its sixteen land below 0.80, none of ours. The arms separate on spread, their standard deviations differing 4.8-fold, and not on the mean or the grok rate. None of these contrasts is confirmatory, and the ten-seed version in the family of Appx. A.3 survives no correction; the one on spread is the strongest (Appx. C.2).

The identity is portable. That makes the splice a procedure: take one trained run's identity, put it into another seed's fresh initialization, and train the target at the cold budget. With one curriculum seed as the donor, all 15 other seeds reach the basin at 0.90, more than the 13 of 15 that reach it running the curriculum themselves (Tab. 7). A second donor repeats it (Appx. D.3). Nor need the donor be a curriculum run: we find the same with one cold seed that grokked on its own, whose identity carries 15 of 16 seeds to 0.90, level with the identity transplant of §4.3 (Tab. 7). Curriculum runs grok in 14 of 16 seeds and cold starts in 4 of 16. So what the curriculum adds is reliability, not uniqueness. What it installs is an arrangement of the target's tokens, not its answer statistics: the identity of a donor whose Coefficients stage never saw the target's equations reaches 0.87, level with the identity transplant of §4.3. We find that eight of its 73 directions carry the arrangement: kept to those before it is written, the identity is level with the unedited transplant. A rotated hidden basis leaves it level too, and permuting which numeral holds which row costs a little without separating

from it. Only redrawing all 73 directions at the donor's own spectrum falls below it (Appx. C.6). It also survives a change of the recipient. Written into a model with twice the heads and eight times the context, each seed's own identity reaches 0.88 in all of seeds 1 to 8. That configuration's cold start reaches 0.55 and its curriculum 0.88 (Tab. 6 adds later seeds).

4.5 A SECOND FAMILY REPRODUCES THE HAND-OFF RESULT AND REVERSES THE CARRIER

A second family reproduces the hand-off result and reverses the carrier. We run the same seed-paired design on the 2-layer model of §3, targeting symbolic differentiation (the derivative of a quadratic at a point) at its harness's own settings: a 205-token vocabulary, context 32 and weight decay 1.0 (Appx. B.2). A compatible precursor (a product plus a sum) before the target lifts the lottery from 0.66 to 0.90, 6/16 to 14/16 ($p = 0.0015$). The two splices again put the hand-off in one group. Here it is the body (0.85, 14/16), not the token identity (0.64, 3/16, below cold start's 6/16; Fig. 8). So which group carries a hand-off differs between the families: the identity there, the body here. The same pair of splices measures it in either family. Schedule against stage accuracy. The family's three-stage schedule (Sum, then Product-sum, then Differentiate) also separates the hand-off state from the stage's own accuracy. Prepending Sum hands the same stage over better learned yet loses the compression. We find the ladder reaches 0.77 with 8/16 against the single precursor's 0.90 with 14/16 ($p = 0.029$). The ODE family's schedules could not make that separation (Appx. D.1). At two layers on the paper's own task, the identity transplant again reproduces a full curriculum run and the body transplant again collapses. Cold start is not a lottery at that depth, though, so by the rule we fixed in advance the arm is uninformative about the carrier (Appx. C.5).

5 DISCUSSION

Which part of the starting point controls the outcome differs by family. In the ODE family it is not the penultimate stage's structural compatibility on its own, since a curriculum that learns the target's own sub-functions in the wrong numerals lands below cold start (§4.3). It is the token identity that stage leaves behind. In the second family it is the body (§4.5). This is basin selection, the reading curriculum learning was first proposed under (Bengio et al., 2009).

The basin-selection reading, and what it predicts. Read the target loss as bistable (Liu et al., 2023; Han et al., 2025): a generalizing basin and a memorizing one, separated by a separatrix and nearly indistinguishable under the training loss. Reaching the generalizing basin comes after an optimization delay (Kumar et al., 2024) on a timescale set by weight decay (Xu et al., 2026), and a target-only run leaves to chance which side its trajectory starts on. The outcome is then set by which basin contains the state the target inherits, θ_{pen}, and everything upstream matters only through it. Appx. D.1 proves this as Proposition 1 on a double well. The second family's ladder tests the reduction directly: prepending a third stage (Sum) moves the Product-sum hand-off, and the outcome follows the hand-off rather than the stage (§4.5). Basin selection further predicts that runs from a structured hand-off generalize sooner, as well as more often. It does. Among runs finishing at or above 0.80, those with a structured penultimate stage cross 0.80 after a median 10,000 target-stage steps, and cold start after 38,000 (Figs. 5 and 10). And the hand-off starts from a higher target loss than cold start does, so the gain is in time, not in the starting loss.

Two competing explanations. Neither account that would predict our ordering without basin selection survives: extra budget (§4.1) and the weight norm at initialization (Lyu et al., 2024), ruled out by two arms of sixteen norm-matched cold starts (Appx. D.4).

Where the effect holds. The effect also holds with twice the heads and eight times the context, and with a leaky schedule that mixes earlier-stage data into the target's batches. At 8 heads and context 512 only 2 of 8 cold starts pass 0.80 against 7 of 8 curricula, and drawing 5 to 20% of each final-stage batch from earlier stages does as well as the full curriculum or better (Tab. 6).

The effect is bounded by weight decay, because cold start's fragility is itself a function of decay (Fig. 7). Cold start's mean rises steadily with decay, from 0.37 at 0.05 to 0.82 at 1.0, and where it is weakest the curriculum lifts every seed. Above 0.3 cold start clears the threshold unaided, the arms converge and then cross, and the rule of §4.2 predicts the crossing: at decay 1.0 the Coefficients

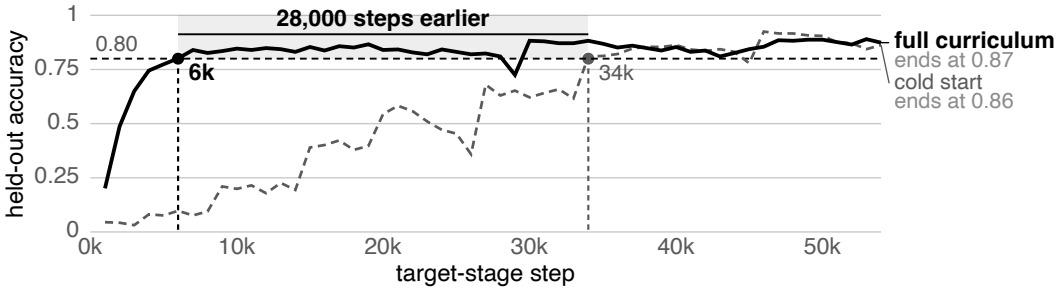


Figure 5: One seed, both ways, budget matched, so the runs differ only in what preceded the target. Both arms arrive, 0.011 apart, and what separates them is the 28,000 steps between their crossings of 0.80.

stage itself generalizes within its budget in 2/8 runs, and those are the two seeds that clear 0.80 on the target. The effect is established where cold start is fragile and the hand-off structured; by decay 1.0 there is little left to compress (Tab. 5).

Limitations. The evidence comes from two symbolic-math families on small models of four and two layers, with sixteen seeds each. Each family runs at the weight decay where its cold start is unreliable, 0.1 and 1.0 respectively. Which group carries the hand-off differs between the families, and we measure it rather than predict it. Whether large models trained in stages, as language models now are, inherit an initialization from their last stage in the same way is untested.

6 Conclusion

Whether a run generalizes is controlled by the weights the last precursor leaves the target, not by the data it saw. One parameter group carries the hand-off in each family: the token identity in the ODE family, the body in the second. For practice, curricula transfer an initialization rather than computation. One run's identity is enough for every seed, putting all 15 others into the basin at the cold budget. A curriculum's value thus lives in the state it leaves, which makes an unreliable training outcome something to inspect rather than retry.

AI Use Statement

Generative AI was used to aid and polish the writing of the paper; every task instance comes from the procedural generators of §3, and every reported result is computed from stored per-seed run artifacts. The authors take full responsibility for the content, including any text written or polished with the aid of generative AI.

References

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31–60. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/abbe23a.html.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

H. V. AlquBoj, Hilal AlQuabeh, Velibor Bojkovic, Munachiso Nwadike, and Kentaro Inui. Mechanistic insights into grokking from the embedding layer, 2025. URL https://arxiv.org/abs/2505.15624.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Jordan Ash and Ryan Adams. On warm-starting neural network training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3884–3894. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/288cd2567953f06e460a33951f55daaf-Paper.pdf.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48. Association for Computing Machinery, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Csaba Both, Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Daniel Karl I. Weidele, Mauro Martino, and Nima Dehmamy. Small models, smarter learning: The power of joint task training, 2026. URL https://arxiv.org/abs/2505.18369.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue WANG, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36000–36040. Curran Associates, Inc., 2023. doi: 10.52202/075280-1562. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/70b8505ac79e3e131756f793cd80eb8d-Paper-Conference.pdf.

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. A basin-selection perspective on grokking via singular learning theory, 2026. URL https://arxiv.org/abs/2603.01192.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935. doi: 10.2307/2342435. URL https://www.jstor.org/stable/2342435.

Laura Gomezjurado Gonzalez. The long delay to arithmetic generalization: When learned representations outrun behavior, 2026. URL https://arxiv.org/abs/2604.13082.

Ting Han, Linara Adilova, Henning Petzka, Jens Kleesiek, and Michael Kamp. Flatness is necessary, neural collapse is not: Rethinking generalization via grokking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=lbtOctHDQ3.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Gunner Levi Howe. Structure-specific representational priors causally control the grokking delay, 2026. URL https://arxiv.org/abs/2607.04333.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vt5mnLVIVo.

Guillaume Lample and François Charton. Deep learning for symbolic mathematics. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1eZYeHFDS.

Jordan LeDoux. The order is the message, 2026. URL https://arxiv.org/abs/2603.25047.

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XsHqr9dEGH.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

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. ISSN 2835-8856. 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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Yoshiyuki Ootani. Grokking is conditional and fragile: A fully-tractable, multi-seed study at 12k parameters, 2026. URL https://arxiv.org/abs/2607.05104.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347–352):240–242, 1895. doi: 10.1098/rspl.1895.0041. URL https://doi.org/10.1098/rspl.1895.0041.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937. doi: 10.2307/2984124. URL https://doi.org/10.2307/2984124.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007. doi: 10.5281/zenodo.40328. URL https://doi.org/10.5281/zenodo.40328.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159. URL https://www.jstor.org/stable/1412159.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. doi: 10.1093/biomet/6.1.1. URL https://www.jstor.org/stable/2331554.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9lycwRxAOI.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. RL grokking recipe: How does RL unlock and transfer new algorithms in LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CJJ8VxOWbG.

Xuan Khanh Truong. Cross-trajectory chimera interventions reveal dissociable roles of weight magnitude and direction in grokking, 2026. URL https://arxiv.org/abs/2607.06628.

Xuan Khanh Truong, Hoang Viet Doan, Duc Trung Luu, and Thanh Duc Phan. The weight norm sets the grokking timescale: A causal delay law, 2026. URL https://arxiv.org/abs/2606.13753.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=ns8IH5Sn5y.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5632–5711. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/wang25a.html.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28. URL https://doi.org/10.1093/biomet/34.1-2.28.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270. URL https://doi.org/10.1162/neco.1989.1.2.270.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://doi.org/10.1080/01621459.1927.10502953.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tW4QEInpni.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. Oolong: Investigating what makes transfer learning hard with controlled studies. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3280–3289, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.198. URL https://aclanthology.org/2023.emnlp-main.198/.

Mingyue Xu, Gal Vardi, and Itay Safran. To grok grokking: Provable grokking in ridge regression. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5nNNVY8NW4.

Yongzhong Xu. The geometry of multi-task grokking: Transverse instability, superposition, and weight decay phase structure, 2026. URL https://arxiv.org/abs/2602.18523.

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4rEI2JdHH6.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pp. 3320–3328. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

A EXPERIMENTAL SETUP

A.1 STAGE DATASETS AND VOCABULARY

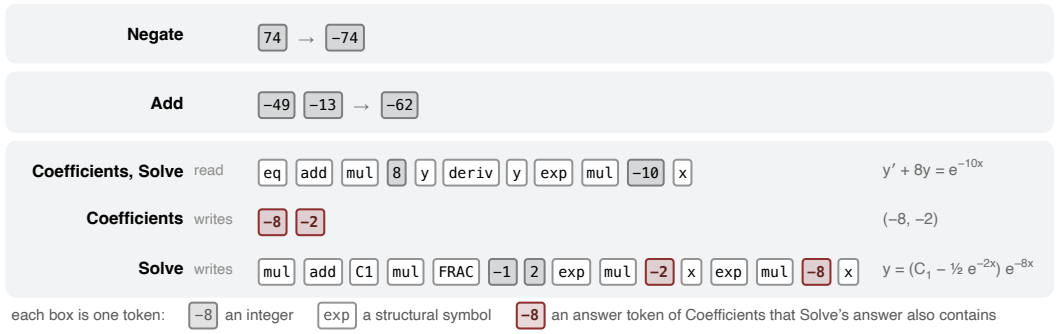


Figure 6: One real training example from each stage, as the tokens the model reads and writes, each followed by the numbers or equation it encodes.

Vocabulary: a single integer-token vocabulary of size 264: one token per integer in $[-100, 100]$, plus the structural tokens the prefix-Polish encoding needs (equation, derivative, fraction, inverse, negation, separator, padding). Because the vocabulary is shared, a curriculum stage never introduces an unseen embedding, and the model's output head is identical across stages.

Tab. 2 gives the size and sequence length of each stage. Split: drawn once with split seed 42 at a 0.5 train fraction, fixed across every condition and every model seed of the ODE family, so no comparison within it is confounded by a different partition. Disjoint-coefficients control: Coefficients regenerated on the coefficient range $a \in [31, 60]$, so that no answer key is shared with the target split.

Table 2: Stage datasets. Sequence length is in tokens after prefix-Polish encoding, including the separator and the answer span.

Stage	Map	train / val	seq. length	role
Negate	$N\{a\} \to N\{-a\}$	100 / 101	5	degenerate single-token
Add	$N\{a\}\, N\{b\} \to N\{a+b\}$	5100 / 5101	6	structured
Coefficients	equation prefix $\to [N\{-a\}, N\{a+b\}]$	442 / 443	14–16 (15.8)	structured, coefficients
Solve	$y' + ay = e^{bx} \to$ symbolic solution	442 / 443	22–29 (28.6)	target

The target stage is small on purpose: with 442 training equations and a 443-equation held-out split drawn from the same generator, a network can fit the training set long before it generalizes, which is where grokking is observed at all.

A.2 ARCHITECTURE, OPTIMIZER, AND BUDGET

Model: a decoder-only transformer, 4 layers, 4 heads, $d = 128$, $d_{\text{ff}} = 512$, context 64, trained from scratch for all runs except the transplant arms, which start from the spliced state of §4.3. Architecture-generality arm: 8 heads and context 512, all other settings unchanged (Tab. 6). Optimizer: full-batch AdamW, learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, gradient clipping 1.0, weight decay 0.1 unless a sweep states otherwise (Fig. 7); its state is reset at each stage boundary by default (a fresh moment estimate), except in the optimizer-kept curriculum, which carries it across (Tab. 3). Precision: fp32, except bfloat16 in the eight conflict-free mixture runs of Appx. B.1. Budget: 54,000 gradient steps on the target stage for all conditions except the 78k cold start; in the sequential arm Negate, Add and Coefficients run 8000 steps each before it. Determinism: on fixed hardware the trainer is bit-deterministic. A same-box rerun reproduces cold start and the full curriculum seed for seed on ten seeds, and across hardware it diverges, as Ootani (2026) documents. So we reran the headline transplant contrast alongside its baselines, each seed's pair on one machine (Appx. C.2), where it reproduces. Seed: full-batch training removes minibatch sampling noise as a source of per-seed variance, so the seed varies the initialization only; the replay arms (which sample each batch) are the exception.

Reported value: the final evaluation rather than a best-so-far or smoothed value, at a known cost: in 14 of the 924 runs with at least six evaluations (1.5%) the final evaluation lands more than 0.15 below the best of the preceding five, a transient optimizer collapse. And those 14 runs are spread across ten conditions, so the cost is not concentrated in any one condition.

A.3 Multiple comparisons

We registered 30 confirmatory tests across roughly twenty conditions in code, seven of them before their runs. And the tiers below are recomputed from that registry.

Strict tier: eleven survive Holm-Bonferroni (Holm, 1979) control of the family-wise error rate at 0.05: the token-identity and body-only splices against cold, the primary comparison on both its tests, the random-permutation Coefficients-against-Negate split, the disjoint-coefficient Coefficients control, the Negate splice against a full sequential run, the disjoint-alphabet arm against both its width-matched control and cold, the GrokTransfer arm against cold, and the two cold starts rescaled to the hand-off norms, against each other (Appx. D.4). Weaker tier: eight more survive Benjamini-Hochberg (Benjamini & Hochberg, 1995) control of the false discovery rate but not the family-wise one: weight decay 0.05 and 0.1, the Negate splice against the Coefficients splice, the step-2000 Spearman correlation (Spearman, 1904), the structured-against-degenerate Fisher test, and the three frozen-splice contrasts of Appx. C.7. Neither tier: the remaining eleven are weight decay 0.3, 0.5 and 1.0, the input-embedding splice against a full sequential run, the readout splice against a full sequential run, the readout splice against the input-embedding splice, the token-identity splice against a full sequential run, the GrokTransfer arm against a full sequential run, that same arm against our own splice, its partial-accuracy source against its converged one, and reverse order against cold. Every positive claim the paper's headline rests on, except cross-seed portability, is in the strict tier, and tests reported as nulls are included in the family, since correcting only over the significant ones is the selection the correction exists to prevent.

Further contrasts quoted in the paper are outside this family. One registered test is outside it too: a third family, $(a+b) \bmod 97$ trained before $(a \cdot b) \bmod 97$ on the same input pairs, does not replicate, landing at 0.09 against cold start's 0.26, because that precursor gives the target's own inputs a different answer, the label conflict of Appx. B.1. Tested after the family closed: the two cross-seed transplants against a full curriculum run, the two Add-last schedules against each other, the frozen identity's two mirror contrasts (Appx. C.7) and the four arms added afterwards, the disjoint-coefficients donor, the pre-shrunk cold start and the two 24k Negate splices (Appxs. C.1 and D.3), each read against a rule fixed before its first seed finished; all are reported uncorrected as the exploratory contrasts they are, and we do not add to a declared family after seeing the data. Second family: the replication of §4.5 is a family of its own, four tests corrected within themselves (Appx. B.2).

B Per-condition results, both families

Tabs. 3 to 6 report the conditions behind the main text; the second family is in Tab. 1, and the arms of Appx. C are stated where they are read. Four conditions have fewer completed seeds than their nominal count and are reported at their actual n, which each table gives per row. At $n = 4$ a single seed moves the grok rate by 25 points, so the mass-balanced mixture is read as failing to reproduce the compression rather than as a bound on gradient mass. Every table below reports exact-sequence validation accuracy at the end of the target stage, population sd over seeds, and grok rates at 0.80, 0.85 and 0.95 with a Wilson 95% interval on the first. Rows are unpaired per-condition summaries over all completed seeds; the paired comparisons in the main text are computed on the seeds present in both arms, which is why the weight-decay 0.3 contrast there is stated over seven paired seeds rather than the eight cold seeds listed here. The splice arms train the target stage for 54k steps from a hand-off-derived initialization, so the precursor's steps are embodied in that initialization rather than re-spent; the equal-total-budget cold arm instead spends a cold run's own budget up to a curriculum arm's 78k total.

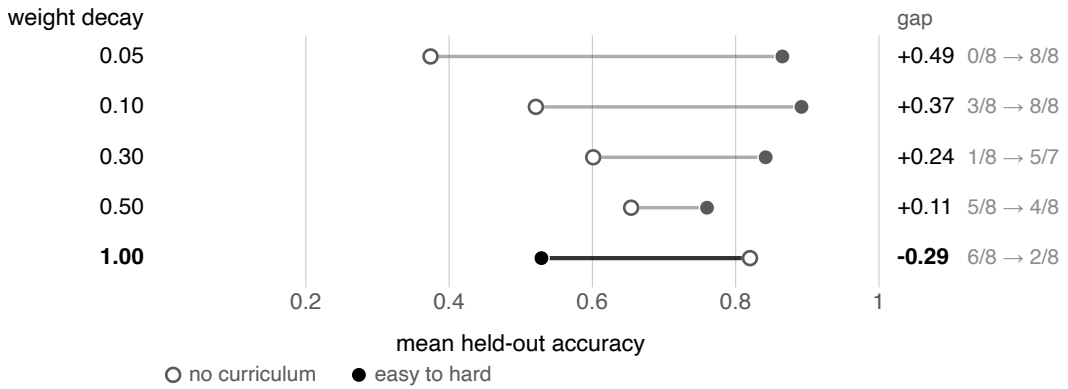


Figure 7: The ordering effect tracks cold-start fragility across weight decay (eight seeds per cell, seven for the curriculum at 0.3). Above decay 0.3 the arms do not separate at this sample size (§5).

Table 3: Primary comparison, ordering, and controls (weight decay 0.1). Tinted rows are the full curriculum, as in Tab. 1.

Condition	penultimate	n	mean	sd	grok@0.80	95% CI	grok@0.85	grok@0.95
cold start	none	16	0.48	0.31	4/16	[0.10, 0.49]	1/16	0/16
full curriculum	Coefficients	16	0.88	0.05	14/16	[0.64, 0.97]	12/16	0/16
interleaved	mixed	15	0.59	0.28	5/15	[0.15, 0.58]	5/15	2/15
mass-balanced mixture	mixed	4	0.57	0.17	1/4	[0.05, 0.70]	0/4	0/4
bookend	Coefficients	16	0.82	0.09	13/16	[0.57, 0.93]	6/16	0/16
random permutation	varies	16	0.74	0.19	8/16	[0.28, 0.72]	8/16	2/16
reverse order	Negate	16	0.50	0.29	4/16	[0.10, 0.49]	2/16	0/16
Coefficients only	Coefficients	16	0.83	0.08	12/16	[0.51, 0.90]	8/16	0/16
Add only	Add	16	0.83	0.11	13/16	[0.57, 0.93]	9/16	0/16
Negate, Add, Solve	Add	16	0.78	0.15	9/16	[0.33, 0.77]	6/16	0/16
Negate only	Negate	16	0.12	0.10	0/16	[0.00, 0.19]	0/16	0/16
curriculum without Negate	Coefficients	16	0.88	0.06	15/16	[0.72, 0.99]	11/16	1/16
curriculum without Coefficients	Add	16	0.88	0.09	14/16	[0.64, 0.97]	13/16	1/16
disjoint coefficients	Coefficients, disjoint	16	0.81	0.14	11/16	[0.44, 0.86]	8/16	1/16
curriculum, optimizer kept	Coefficients	12	0.88	0.05	11/12	[0.65, 0.99]	8/12	1/12

Table 4: Component-swap splice arms (§4.3) and the equal-total-budget cold arm.

Condition	n	mean	sd	grok@0.80	95% CI	grok@0.85	grok@0.95
identity transplant	16	0.87	0.08	15/16	[0.72, 0.99]	13/16	0/16
body transplant	16	0.13	0.19	0/16	[0.00, 0.19]	0/16	0/16
identity transplant, Negate	16	0.63	0.28	8/16	[0.28, 0.72]	5/16	0/16
input-side only	16	0.85	0.07	13/16	[0.57, 0.93]	11/16	0/16
readout only	16	0.88	0.04	15/16	[0.72, 0.99]	13/16	0/16
cold start, 78k total	16	0.53	0.31	5/16	[0.14, 0.56]	1/16	0/16

Table 5: Weight-decay sweep. Paired t against cold at each setting: $p = 0.004$ at 0.05, 0.022 at 0.1, 0.075 at 0.3, 0.42 at 0.5 and 0.083 at 1.0, all members of the confirmatory family of Appx. A.3. Tinted rows are the full curriculum, as in Tab. 1.

Condition	n	mean	sd	grok@0.80	95% CI	grok@0.85	grok@0.95
cold start, 0.05	8	0.37	0.29	0/8	[0.00, 0.32]	0/8	0/8
full curriculum, 0.05	8	0.87	0.04	8/8	[0.68, 1.00]	5/8	0/8
cold start, 0.1	8	0.52	0.34	3/8	[0.14, 0.69]	1/8	0/8
full curriculum, 0.1	8	0.89	0.03	8/8	[0.68, 1.00]	8/8	1/8
cold start, 0.3	8	0.60	0.26	1/8	[0.02, 0.47]	0/8	0/8
full curriculum, 0.3	7	0.84	0.09	5/7	[0.36, 0.92]	4/7	0/7
cold start, 0.5	8	0.65	0.31	5/8	[0.31, 0.86]	3/8	0/8
full curriculum, 0.5	8	0.76	0.15	4/8	[0.22, 0.78]	4/8	0/8
cold start, 1.0	8	0.82	0.13	6/8	[0.41, 0.93]	6/8	0/8
full curriculum, 1.0	8	0.53	0.34	2/8	[0.07, 0.59]	1/8	0/8

Table 6: Arms quoted in the text that no other table carries: the architecture and replay controls of §5, and the arms added after the confirmatory family closed, read as exploratory (§4.3, Appxs. C.1, C.7 and D.3); the grokked cold donor is seed 12, the best of the cold arm, and the conflict-free mixture is the paper's only arm trained in bfloat16.

Condition	n	mean	sd	grok@0.80	95% CI	grok@0.85	grok@0.95
cold start, 8 heads	13	0.48	0.26	3/13	[0.08, 0.50]	0/13	0/13
full curriculum, 8 heads	8	0.88	0.04	7/8	[0.53, 0.98]	7/8	0/8
curriculum with 5% replay	8	0.94	0.02	8/8	[0.68, 1.00]	8/8	3/8
curriculum with 20% replay	8	0.94	0.01	8/8	[0.68, 1.00]	8/8	2/8
identity transplant, disjoint donor	16	0.87	0.06	13/16	[0.57, 0.93]	10/16	1/16
cold start, numerals pre-shrunk	16	0.53	0.30	2/16	[0.03, 0.36]	1/16	0/16
identity transplant, 24k Negate	16	0.70	0.22	9/16	[0.33, 0.77]	5/16	0/16
body transplant, 24k Negate	16	0.12	0.13	0/16	[0.00, 0.19]	0/16	0/16
identity transplant, grokked cold donor	16	0.90	0.04	15/16	[0.72, 0.99]	15/16	2/16
body transplant, grokked cold donor	16	0.32	0.33	2/16	[0.03, 0.36]	2/16	1/16
grokked cold body frozen	8	0.73	0.27	7/8	[0.53, 0.98]	1/8	0/8
identity transplant, basis rotated	16	0.89	0.04	16/16	[0.81, 1.00]	14/16	1/16
identity transplant, leading 8 directions	16	0.89	0.03	16/16	[0.81, 1.00]	15/16	0/16
identity transplant, leading 32 directions	16	0.89	0.04	16/16	[0.81, 1.00]	14/16	0/16
identity transplant, numerals permuted	16	0.81	0.09	12/16	[0.51, 0.90]	7/16	0/16
identity transplant, singular values only	16	0.81	0.08	9/16	[0.33, 0.77]	8/16	0/16
Lookup last	16	0.74	0.16	7/16	[0.23, 0.67]	0/16	0/16
conflict-free mixture	8	0.95	0.02	8/8	[0.68, 1.00]	8/8	5/8
identity transplant, 8 heads and context 512	12	0.88	0.03	12/12	[0.76, 1.00]	10/12	1/12

B.1 The mixed-training arm is a label-conflict lottery

The interleaved arm does not compress the lottery, reaching 0.59 on the target with 5 of its 15 seeds above 0.80. The Coefficients and Solve stages ask different questions of the same 442 equations, but no token in the encoding says which one is being asked. Pooled, they put two different labels on the same prompt, so pooled training accuracy cannot exceed $5642/6084 = 0.927$, and every interleaved seed lands exactly at that cap: the arm fits as much of the pool as the conflict allows. Winners: the five seeds clearing 0.80 on the target are those where the target's labels won, with Coefficients validation accuracy below 0.03. The interleaved arm's spread is thus a lottery over which label wins the conflict, not over grokking. Without the conflict: the mixture without Coefficients, sampled jointly for the same 54,000 steps as the interleaved arm (no prompt carries two labels), reaches 0.951 ± 0.021 on the target with 8 of 8 seeds above 0.80 and a minimum of 0.916; joint training on compatible stages is not a lottery at all. Unlike every other arm in the paper, including the interleaved one they are read against, these eight runs were trained in bfloat16 rather than fp32.

B.2 A second task family

Harness: the 2-layer transformer and 205-token vocabulary of an existing symbolic-math harness, at its own settings: weight decay 1.0, context 32, no gradient clipping. Pairing: the harness draws the train/validation partition per model seed. So the second family's arms are paired on the seed but not on the split. Task: symbolic differentiation (the derivative of a quadratic at a point) as the target (100,000 steps, evaluated every 500). Two precursors: three-term addition (Sum, 15,000 steps) and a product-sum (Product-sum, 15,000 steps), at a 0.5 train fraction, sixteen seeds per arm. Schedules: the three-stage ladder from easy to hard (Sum, Product-sum, Differentiate), and the compatible precursor alone, one stage (Product-sum, Differentiate). And the two transplants take the Product-sum hand-off's token identity (its embedding and readout) or everything else into the same seed's own cold initialization, as in §4.3. No precursor is solved at hand-off (Product-sum spans 0.04 to 0.49 across seeds). Trained first, the paper's own Coefficients is in the same regime, a partial hand-off (0.356; §4.2).

The compatible precursor alone lifts accuracy from cold start's 0.66 to 0.90, and from 6 to 14 of 16 seeds (Tab. 1). And it beats the ladder directly, by 0.13, while the ladder does not separate from cold start. That is despite handing over the same penultimate stage better learned, a mean hand-off of 0.31 against 0.20, higher in 15 of 16 seeds. The carrier is the body: the body transplant reaches 0.85,

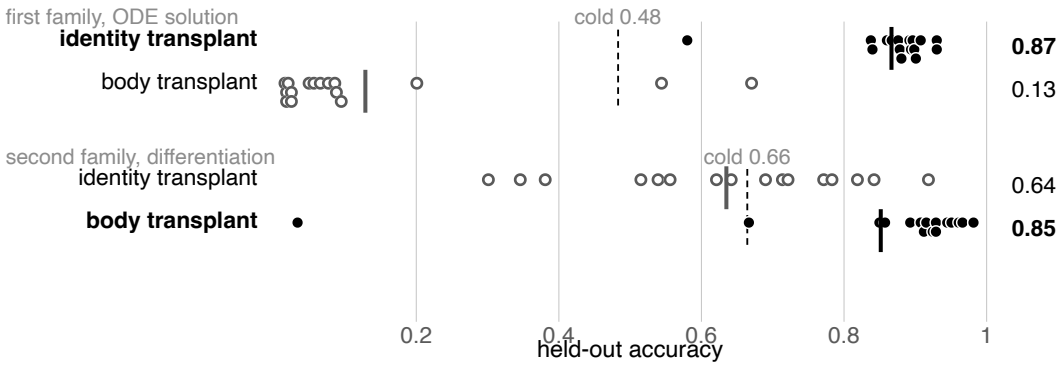


Figure 8: The same two transplants in both families, both against their own cold start (dashed). Sixteen seeds per arm, filled dots for the group that lands above its cold reference.

above cold start ($p = 0.007$), and the identity transplant 0.64, level with it. That reverses the carrier of §4.3. The ladder's shortfall with the same penultimate stage sharpens the rule of §4.2: what the target inherits is the hand-off state, not the stage's name, and prepending Sum changes that state while raising the stage's own accuracy, a separation the ODE family's schedules could not make. This family changes the precursor, the depth, the vocabulary and the training settings at once; depth and the precursor are each held alone in Appx. C.5: the precursor does not move the carrier, and the depth arm, where cold start is not a lottery, is uninformative about it. Multiplicity: the four tests against cold are corrected as their own family, and the compatible-precursor and body-transplant contrasts survive Holm within it.

C Competing readings

C.1 The transfer needs the target's own numerals

The splice of §4.3 localizes the effect to the token-identity parameters by moving them. This appendix reaches the same place by editing the data instead.

Data edit: the precursor stages are rewritten into a disjoint numeral alphabet. And the 201 single-token numerals, the integers -100 to 100, move from ids 63–263 to 264–464, widening d_{vocab} to 465. There is zero numeral overlap, and the 63 structural tokens, operators and separators, are shared by design. The target stage stays in the base alphabet, so it reads numerals the curriculum never trained.

Width-matched control: it carries the same 201 unused embedding rows and readout columns, ids 264–464, which none of its stages uses, and remaps nothing. A remapped arm would otherwise differ from a narrower baseline in parameter count and in per-token gradient frequency, which is itself a grokking lever. The remapped and width-matched arms are identical at 921,041 parameters.

Target result (§4.3, Tab. 1): seed for seed the disjoint arm falls 0.795 below its width-matched control and 0.409 below cold start, both paired. The width-matched control reaches 13/16 above 0.85.

Decay: the target's numerals are not, however, left at initialization. All parameters are subject to AdamW's decoupled weight decay in our recipe. So rows of W_E that receive no gradient for the 24,000 precursor steps are shrunk by $(1 - 10^{-4})^{24000} \approx 0.09$. And the readout columns and biases of the base numerals receive the softmax's negative gradient at every precursor step, since their logits enter every normalizer, and are trained never to be emitted while the readout norm grows. Measured at seed 1's hand-off, the W_E rows of the target's numerals sit at 0.09 of a cold initialization's norm, the pure-decay prediction. Their W_U columns have grown to 4.9 times cold with a mean bias of -0.11. The width-matched control's trained rows and columns sit at 0.67 and 7.0. The arm therefore hands the target a token identity worse than fresh. The hand-off is a trained body under a degraded identity.

Pre-shrunk control: a cold start with the target's numeral rows and columns pre-shrunk by the same factor, and nothing else, reaches 0.53, indistinguishable from cold start. The pre-shrunk control is far above the arm's 0.07, so the arm measures the trained body, not the degraded identity.

A remapped arm is only interpretable if its precursor still learned, else its result is vacuous. So each is gated on its own stage-3 hand-off, the Coefficients stage's final accuracy.

Remapping Coefficients alone fails that gate and is excluded: that stage then meets its numerals for the first time and 8000 steps do not suffice, leaving a hand-off of 0.035 against the control's 0.947. Remapping Negate, Add and Coefficients together passes it, because Add trains the same alphabet Coefficients reuses: hand-offs of 0.998 at Add and 0.942 at Coefficients, against the base-alphabet control's 0.997 and 0.947.

This result is a statement about a disjoint numeral alphabet in this task family, not about natural languages or writing systems; three cross-lingual results bear on it.

Two results report the opposite sign, and each differs from ours in a way that our design removes on purpose. Artetxe et al. (2020) transfer a monolingual model to a new language by learning a fresh embedding matrix with every other layer frozen, and reach competitive downstream performance with no shared vocabulary; their body is already fully formed when the new alphabet arrives, whereas ours is being shaped by the very stage that speaks the alien one. K et al. (2020) construct a Fake-English corpus with strictly zero character overlap and find lexical overlap negligible for cross-lingual transfer; both alphabets are present throughout their joint pretraining, so a shared representation can form, which a curriculum by construction forbids. A third result agrees with ours: Wu et al. (2023) perturb one axis of cross-lingual variation at a time and find models resilient to word-order shifts but vulnerable to vocabulary misalignment and embedding re-initialization, unrepaired by roughly 15M tokens of continued pretraining.

The dividing line those three suggest is whether the body was trained with access to the alphabet it must later read, which is exactly the variable our precursors control and the one that makes the body-only splice and this arm land together.

C.2 The competing method

The contrast with Xu et al. (2025) drawn in Appx. D.3 is a design argument. So we also ran their construction on our target. Source: what the recipe specifies, a weaker model on the target itself; the one we train has two layers instead of four and keeps width at 128. That source is competent (reaching 0.852 with 13 of 16 seeds above 0.80), so what follows is not a failure to train it.

Result: splicing its input embedding into a cold four-layer initialization leaves the target at 0.784 over sixteen seeds, recovering 72.8% of the cold-to-sequential gap. On the same sixteen seeds, each seed's two arms on one machine, our own splice recovers 95.1%. The means sit close together. Ours is ahead by 0.090 seed for seed, short of significance.

What the arms separate on. The distributions differ in their spread. Ours: all sixteen seeds above 0.80, a minimum of 0.817 and a standard deviation of 0.036. Theirs: eleven above and five below 0.80, a minimum of 0.269 and a standard deviation of 0.174, 4.8 times ours (one-sided permutation $p = 0.023$). And all five seeds on which the arms disagree at 0.80 favor ours, though five such seeds cannot reach significance. Both close most of the gap in the mean, but ours alone empties the lower tail.

The operating-point objection. Sourcing at their operating point does not help. Their method is argued for from a source that only *partially* generalizes, on the claim that such a cheap source suffices, whereas their arm above sources at convergence. We committed a test of it before the runs finished, sourcing each seed at about 0.70 (the accuracy their parity experiment uses); a gain of 0.10 or more over the converged arm, at $p < 0.05$, would count as a rescue. And on the ten seeds the two share, the partial-source arm reaches 0.777 against the converged arm's 0.792, no rescue.

Two deviations from the published method. The first is deliberate: we hold width at 128 and copy the embedding directly, so the arm tests their source without the trainable factorization their method puts on the target's embedding, which they report as a favorable initialization, and without

their change of width. The second is a limit of ours, not a change to their method: the partial-source arm has ten seeds where the comparison above has sixteen.

C.3 Teacher forcing and free-running decoding

Metric: whole-sequence exact match under teacher forcing (Williams & Zipser, 1989), for all accuracies in this paper. Greedy decoding of the held-out split from each final checkpoint returns the teacher-forced number to within 5×10^{-8} on all 31 runs whose final checkpoint we retain. These are sixteen cold starts and fifteen curricula. Sequences count correct only when the argmax is right at all supervised positions. And a decoder fed those same argmaxes emits them again, so sequences with any wrong position fail under both. This holds for exact-sequence match under greedy decoding; the paper reports neither token-level accuracy nor sampled decoding.

C.4 The causality gate

The target is trained under teacher forcing, with its answer tokens inside the input sequence. So a model with any non-causal path could copy the answer, and its accuracy would look like generalization. We check every model at build time, before it trains. For an input x that fills the model's context, let $x^{(t)}$ replace the token at position t with the next token in the vocabulary; the logits z at every earlier position must not move:

$$\max_{t} \max_{s<t} \left\| z_s\big(x^{(t)}\big) - z_s(x) \right\|_\infty \le 10^{-5}. \tag{1}$$

And a model that fails the check raises an error before the first gradient step. After training, two checkpoints (a four-layer curriculum run and a two-layer one) are checked again, so the check covers the answer span: we find that both satisfy (1) exactly, with every earlier logit bitwise unchanged and a maximum leak of zero at both depths.

C.5 Depth and the precursor, held alone

Each family is read at the weight decay where its lottery exists (0.1 and 1.0). Above 0.3 the ODE family's cold start clears the threshold unaided. And by decay 1.0 the curriculum's effect crosses to negative (Tab. 5). At 0.1 the second family's target averages 0.28 to 0.30 across conditions, 0 of 33 runs across its four schedules above 0.80, best single run 0.38. So no setting holds both families' lotteries. Task, depth and vocabulary do vary together, and two arms change one of them at a time. Depth arm: the paper's own Solve target, tokenizer, recipe, schedule and 54,000-step budget at two layers instead of four, with the same four conditions at eight seeds; see §4.5. The two transplants take the two-layer curriculum's own Coefficients hand-off, so nothing crosses depths. We read the transplants against the curriculum they were taken from, not against cold start, because that is the comparison the carrier question asks (§4.3). The question of what a network carries along depth is older: Yosinski et al. (2014) find that transferability falls with layer, and Neyshabur et al. (2020) that a fine-tuned model stays in its pretrained basin with reuse near the input. Our partition runs along the vocabulary axis; see Appx. C.1. Registration: we committed the decision rule before the first target finished.

condition	n	target accuracy	grok@0.80	paired p
depth alone: the paper's target and recipe at two layers				
curriculum	8	0.892 ± 0.038	8/8	
identity transplant	8	0.885 ± 0.025	8/8	0.72 vs curriculum
body transplant	8	0.179 ± 0.255	0/8	0.0002 vs curriculum
cold start	8	0.786 ± 0.128	5/8	0.067 vs curriculum
precursor alone: the Solve-const hand-off at the paper's depth				
cold start	16	0.41	2/16	
hand-off carried whole	16	0.23	0/16	0.13 vs cold
body under a fresh identity	16	0.10	0/16	0.001 vs cold
identity in a fresh body	16	0.84	13/16	$< 10^{-4}$ vs cold, vs body

Result: the transplants give the four-layer answer. The reversal the second family shows does not appear at this depth (the gap between the two groups is 0.71 at this depth against 0.74 at four layers).

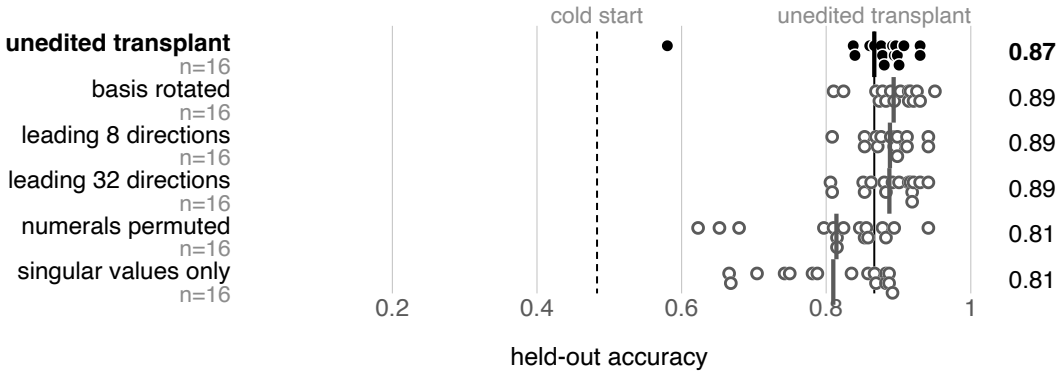


Figure 9: The transplanted identity with one property destroyed per row, against the unedited transplant (solid line) and cold start (dashed). Sixteen seeds per arm, all trained on the target with the cold recipe.

At this depth cold start is less fragile, and the curriculum's compression over it does not separate at eight seeds, so by the pre-registered rule that arm is uninformative. Multiplicity: both contrasts are exploratory and uncorrected.

Precursor arm: only the precursor changes, from reading the two coefficients off the equation to solving a sibling equation (Appx. A.1). Solve-const ($y' + ay = b$, in the target's own tokens) is learned to between 0.86 and 0.93 in its 8,000 steps in every seed; its hand-off holds a generalizing circuit (Olah et al., 2020) where Coefficients's holds a read-off. The two splices carry that hand-off into Solve at the paper's depth, recipe and budget, read against a cold start on the same machine; see Appx. A.2. Multiplicity: the four contrasts are corrected within themselves (Appx. A.3). And all but the whole hand-off against cold survive Holm. Carried whole, that hand-off lands with §4.2's degenerate schedules, not its structured ones (its identity in a fresh body is above the whole hand-off in every seed). A circuit in the precursor does not move the carrier (this hand-off splits as §4.3's does, as for the Negate hand-off of Appx. D.3). §4.2's rule covers stages that compute the target's own sub-quantities. A sibling equation with its own solution is outside it.

C.6 What the transplanted identity carries

The identity transplant of §4.3 moves three tensors (W_E, W_U and b_U), and this appendix asks which property of them the target reads. Each arm takes the same per-seed transplant and destroys one property of it over the 73 numerals the target uses. Everything else, including the cold body, is left untouched. And each arm trains on Solve for the same 54,000 steps. Rotation: the identity times a random orthogonal matrix on the hidden axis, $W_E Q$ and $Q^\top W_U$. This preserves everything up to a change of basis (hidden-unit permutations are a special case, Entezari et al., 2022; Ainsworth et al., 2023), which is meaningful because the body is a fresh draw with no preferred basis. Permutation: the 73 embedding rows, readout columns and biases reassigned among themselves. The set survives; which numeral holds which does not. Spectral: each block's singular values kept, the scale profile, while its singular vectors are redrawn at random. Rank: only the leading 8 or 32 of the 73 singular directions kept (a truncation). We committed the rule, the runner and the analysis before the first target finished.

Fig. 9 gives the result: the rotation and the truncation to eight directions are level with the unedited transplant, so neither the coordinates nor the 65 trailing directions are read. The permutation costs a little and is not separated from it; only the spectral arm is, which places what the target reads in the subspace the leading directions span rather than in the numeral each row belongs to or in the scale profile alone. None of the five contrasts is corrected: they are exploratory and sit outside the confirmatory family of Appx. A.3.

C.7 The test that could have refuted the headline

Our sharpest negative result is that splicing everything the curriculum trained *except* token identity lands below a cold start. We read that as the computation not transferring, but a narrower reading

(co-adaptation, in the sense of Yosinski et al., 2014) predicts all arms in the paper equally well. A body trained beside one embedding may be unable to use a random one, and that would be a fact about joint training, not about what the body holds. The readings separate if the body is frozen and only the token identity trains: the identity can then adapt to the body, instead of both adapting to each other. We fixed the reading before the runs: recovery toward a full sequential run would mean the body carries reusable computation that a cold token identity could not address, and the headline would have to narrow from "does not transfer" to "does not transfer without a matching token identity"; staying near the un-frozen splice would rule the co-adaptation reading out. We committed the decision rule, the analysis script and the runner before the first seed ran.

Result: over eight seeds the frozen arm reaches 0.005 with a bootstrap interval (Efron, 1979) of $[0.002, 0.010]$, below the un-frozen body splice's 0.128 and nowhere near the 0.70 that would have narrowed the claim. The frozen arm rules out the strong form of the co-adaptation reading (that the body suffices once an identity is fitted to it); the weaker form, that fine-tuning would exploit partially reusable computation, is the un-frozen body transplant of §4.3 itself, which lands at 0.13.

Capacity. We find it fits: all eight reach training accuracy 1.000 with training loss below 10^{-3} while held-out accuracy stays at or below 0.020. And held-out loss *rises* throughout.

Positive control: the body of the best cold seed (seed 12, 0.862), frozen under each of eight seeds' fresh token identities with only the identity trained, is recovered in seven of the eight, at a mean of 0.73; seed 7 stays at 0.016. And the same body, transplanted whole under a fresh identity with everything training, lands near cold start, at 0.32. So the un-frozen transplant of §4.3 measures what a fresh seed can use, not what a body holds. And seed 12's identity in a fresh body carries 15 of 16 seeds to 0.898 ± 0.041 (Tab. 6).

Mirror (whether freezing itself breaks the arm): freezing the spliced token identity and training the body instead reaches 0.809 over the same eight seeds. That closes 73% of the cold-to-sequential gap, close to but below both a full sequential run and the un-frozen identity splice. The same protocol on the same spliced state recovers when token identity is held and fails when the body is. This is the asymmetry the headline claims, and freezing does not produce it.

D THE BASIN ACCOUNT: PROOF AND NULL MEASUREMENTS

A singular-learning account ranks a run's competing basins by their local learning coefficient and reads grokking as the transition between them (Cullen et al., 2026); that is a selection inside one run, where the selection here is which basin the hand-off starts in.

D.1 PROOF OF PROPOSITION 1

We make this reduction precise using a minimal dynamical model. The target loss L_k reduces to a bistable potential whose two wells are the memorizing and generalizing basins,

$$U(x) = (x^2 - 1)^2, \qquad B_{\mathrm{gen}} = \{x < 0\}, \qquad B_{\mathrm{mem}} = \{x > 0\}, \qquad \Sigma = \{x = 0\}, \tag{2}$$

and each stage D_i acts by gradient flow on a convex surrogate centered at its own solution c_i, the position of D_i's attractor relative to Σ: structured stages such as Add and Coefficients, whose solutions sit in the target's generalizing basin, have $c = -1 \in B_{\mathrm{gen}}$, while the degenerate single-token stage Negate has $c = +1 \in B_{\mathrm{mem}}$. The statement below is elementary; it pins down what the account commits to and is not claimed as a technical contribution: once the reduction is granted, nothing upstream of θ_{pen} can affect the outcome. Its premise in (ii) is that the hand-off lies in B_{gen}, which depends on the hand-off and not on the stage's name: in the second family's ladder (Appx. B.2), prepending Sum moves the Product-sum hand-off across, and the outcome follows the hand-off, as (ii) says it must.

Proposition 1 (Penultimate-basin reduction)**.** *Under gradient flow on U from a start θ_{pen}, the outcome is $\mathbf{1}[\theta_{\mathrm{pen}} \in B_{\mathrm{gen}}]$. Hence (i) a cold start symmetric about Σ groks with probability one-half; (ii) a structured penultimate stage of sufficient duration places θ_{pen} in B_{gen}, so the run groks; (iii) a degenerate one places it in B_{mem}, below the cold baseline.*

Proof. The gradient flow on (2),

$$\dot{x} = -U'(x) = -4x(x^2 - 1), \tag{3}$$

has an unstable fixed point at $x{=}0$ and stable fixed points at ± 1, so the sign of x is preserved and no trajectory crosses Σ: $x(t) \to -1$ iff $x(0) < 0$, giving the indicator. Then (i) a θ_0 symmetric about Σ has $\Pr[\theta_0 < 0] = \frac{1}{2}$; (ii) repeated steps $x \leftarrow x - \eta(x - c)$ on the structured surrogate, centered at $c = -1$, with step size $\eta \in (0, 1)$ converge geometrically to c, so any bounded start crosses Σ after finitely many of them (the duration condition), and $\theta_{\text{pen}} \in B_{\text{gen}}$; (iii) the degenerate surrogate centered at $c = +1$ likewise carries any bounded start to $x > 0$ in finitely many steps, so $\Pr[\text{grok}] = 0 < \frac{1}{2}$. □

D.2 Nulls on what the hand-off carries

These analyses, and those in the three subsections that follow, did not support the account; they bound what the paper claims.

The carrier is not a numerically ordered embedding. Cheapest mechanistic guess: an arithmetic precursor installs a numerically ordered geometry over the numeral rows, and that geometry is the carrier. We score it per hand-off by the Spearman correlation between pairwise row distance and pairwise absolute value difference, over the 73 numerals the target actually uses rather than all 201. Negate trains 155 distinct numerals and Add 199, while Coefficients trains only the target's 73. Scoring all rows would therefore penalize Coefficients for rows that never left initialization. The guess fails on its own terms. We find Add orders the input embedding at $+0.574 \pm 0.161$ and Coefficients at $+0.061 \pm 0.069$, barely above the cold reference of $+0.029$, a roughly tenfold difference between two stages that compress the target to the same 0.83 when either sits immediately before it (Tab. 3). If ordering in W_E were the carrier, Add would dominate Coefficients; it does not. Readout: structured stages separate from the degenerate one, $+0.13$ to $+0.17$ against Negate's $+0.03$, but at a magnitude far too small to carry a 0.48 to 0.87 effect. This measures spliced states, not a training outcome, so it is not a member of the confirmatory family (Appx. A.3).

What does separate them is rank and scale. Over the same 73 numerals, the identities the splice of §4.3 compares differ in how much of the embedding they occupy. Coefficients hand-off: the target's numeral rows of W_E sit at 2.9 times the norm of the numerals the target never uses, compressed to an effective rank of 43.6 ± 2.5 (the exponential of the singular-value entropy, out of 73; Roy & Vetterli, 2007), while their readout columns in W_U spread to 49.3 ± 2.0. 8k Negate hand-off: the rows stay at those numerals' norm (ratio 1.0) and at 61.7 ± 1.9, near the 66.6 of a random matrix of that shape, with the readout at 32.7 ± 1.0. Add hand-off: between the two on the embedding (52.3 ± 2.5) and with Coefficients on the readout (50.0 ± 1.5). We find each contrast holds in all 16 seeds. We find the same rank tracks the outcome where no curriculum was run; across the sixteen cold runs, final accuracy falls with the embedding's end-of-run rank over the same numerals, at a rank correlation of -0.85. This is a description of the hand-off, not a mechanism.

D.3 Nulls on where the splice acts

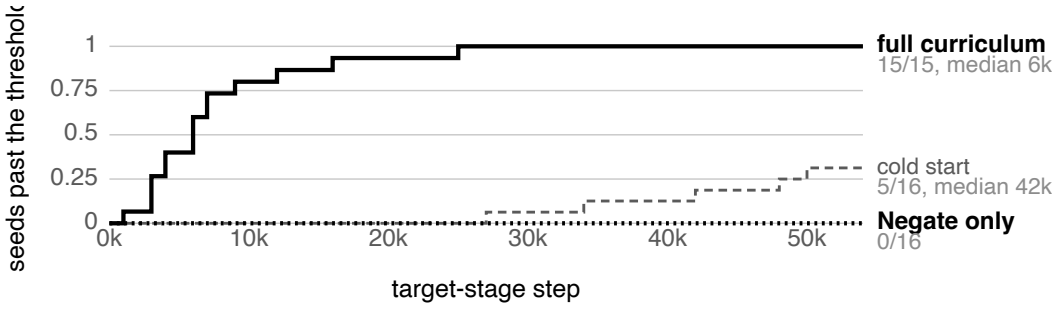


Figure 10: When each run generalizes, not only whether. The curve steps at each run's first crossing of 0.80 and stops at the budget, so runs that never cross set the ceiling. Counts can exceed Tab. 3's rates by the seeds that cross and fall back, and one curriculum seed has no step log. By this definition the medians over the runs that cross are 6,000 target-stage steps for the curriculum and 42,000 for cold start; the 10,000 against 38,000 in §5 are medians over runs finishing at or above 0.80, the first across all nine structured-penultimate conditions and the second for cold start.

The hand-off is not a warm start on the target. The token-identity, body, input-embedding and readout splices were each evaluated on the target stage before any optimizer step (eight seeds, full train and validation batches). We find initial target loss and final outcome run in opposite directions. The input-embedding splice starts at 5.624 with a logit scale of 1.02, indistinguishable from a cold initialization's 5.624 and 1.01, both near the uniform-prior value $\ln 264 = 5.576$, and still ends 0.37 higher in accuracy. The token-identity and readout splices start *worse* than cold, at 6.594 and 6.488, and end far better. The sequential hand-off itself starts at 19.451 with a logit scale of 45.91 and still reaches 0.877. The same measurement rules out the simplest artifact reading of the body-only arm: it starts at 5.640, a cold initialization's own loss, though at a logit scale of 4.43 against cold's 1.01, so the splice leaves the loss the target starts from unchanged and only the scale of its logits differs.

Splice pairing. Every splice is seed-paired unless stated otherwise. The donor for seed s is that same seed's own sequential run, and it is spliced into the initialization that seed s's own cold start draws, reproduced by the same seeding sequence the trainer uses. The sixteen splice runs are therefore sixteen independent donor-recipient pairings rather than one donor reused across recipients, so the paired tests against cold start and a full curriculum run are legitimate.

The splice arm, per seed. We find the five seeds on which cold start collapses below 0.20, at 0.04, 0.05, 0.06, 0.14 and 0.18, reach 0.93, 0.84, 0.88, 0.93 and 0.88 under the token-identity splice. And the arm's per-seed rank correlation with cold is only $+0.07$: the splice does not add a constant to each seed's luck; it removes the dependence on it. Seed dependence is reported as a Spearman rank correlation with a permutation null throughout, because the outcome distribution is bimodal at these budgets; Pearson's correlation (Pearson, 1895) agrees here ($+0.14$).

The cross-seed transplant. The exceptions to seed pairing are deliberate. The procedure of §4.4 also splices a single donor's Coefficients-stage token identity, seed 1's or seed 9's, into the cold initializations of the fifteen seeds other than the donor: thirty runs at the cold budget, listed per seed in Tab. 7. The grokked cold seed 12 is reused across all sixteen seeds as well, its identity in §4.4 and its body in §4.3. We find both donors carry all fifteen other seeds above 0.80 and average 0.90, level with those seeds' own curriculum runs, which reach 0.87 with 13 of 15. Because one donor is reused, the recipients share a token identity; the test against a full curriculum run came after the family closed and is exploratory (Appx. A.3). The transplant is the causal counterpart of measuring interpretive equivalence between independently trained models (Sun & Toneva, 2026), which compares representations without moving them.

Table 7: Held-out accuracy per recipient seed for the two cross-seed transplants and for the identity of seed 12, the grokked cold donor, beside each seed's cold start and its own curriculum run. A dash marks each cross-seed donor's own seed.

seed	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
cold start	.04	.69	.77	.84	.84	.75	.06	.81	.14	.18	.29	.86	.33	.05	.45	.64
own curriculum	.93	.89	.85	.90	.87	.77	.93	.94	.91	.89	.84	.87	.78	.86	.91	.88
identity of seed 1	–	.92	.91	.89	.88	.94	.85	.90	.90	.92	.83	.91	.90	.90	.92	.89
identity of seed 9	.90	.90	.94	.89	.85	.89	.93	.95	–	.86	.92	.89	.82	.90	.93	.92
grokked cold identity	.92	.93	.91	.85	.93	.87	.79	.88	.90	.89	.92	.91	.96	.92	.95	.86

Which Negate state is spliced. The Negate hand-off of §4.3 is the sequential run's state at the end of its 8,000-step Negate stage, as many Negate steps as end reverse order, and not the 24,000-step Negate-only state that lands at 0.12. The comparison of §4.3 therefore tests the favorable direction, the one the claim needs, and at 0.63 the 8k Negate token identity carries the lottery back with it; the same pair of splices on the 24k Negate-only state, sixteen seeds each, localizes its fall below cold start. Its body under a fresh identity stays at the schedule's own level, 0.12 with no seed above 0.80, below cold start; its token identity in a fresh body reaches 0.70 with 9 of 16, not below cold start and far above the schedule. This hand-off splits as the curriculum's does: both bodies under a fresh identity land below cold start, at 0.12 and 0.13, so the identities, at 0.70 and 0.87, separate the two schedules. Decoupled decay is not a second account of that fall. It needs the target's rows to receive no gradient while the precursor runs, which the disjoint alphabet of Appx. C.1 arranges by construction; Negate's own training data covers 59 of the target's 73 answer numerals, so most of those rows are trained rather than decayed. A cold start with all 201 numeral rows and

columns pre-shrunk by the same 24,000-step factor, more than this arm leaves unvisited, reaches 0.53, indistinguishable from cold start and far above the 0.12 here. The two accounts also disagree on where the harm sits, and the splices put it in the body while decay acts on the identity.

The splice does not localize below the token-identity group. The splice of §4.3 transfers the input embedding and the readout together. Splitting them does not resolve which one carries the effect: the input embedding alone recovers 93% of the gap between cold start and the curriculum and the readout alone all of it, the two do not separate over sixteen seeds, and their per-seed ranks are uncorrelated. Both are individually sufficient and sixteen seeds cannot separate them, so we read them as redundant carriers and do not claim a single one.

This bears on Xu et al. (2025), who transfer an input embedding from a weaker model on the same target: that operation is sufficient here as well, so the channel does not differentiate the two. What differs is the source, an easier task rather than a weaker model, and the complement arm, which lands below cold start.

D.4 Nulls on scale

Logit scale is not the carrier either. An account in which the hand-off acts through the starting scale of the output logits rather than through the norm of the weights that produce them is ruled out by the step-0 probe of Appx. D.3.

Readout scale is not the carrier. The curriculum drives the readout to roughly eight times a cold initialization's norm, so one may ask whether the token-identity splice hands over a scale and not an identity. The comparison of the 8k Negate and Coefficients hand-offs in §4.3 answers it without a new run, because the two splices it compares are not matched on that scale and the mismatch runs the wrong way: over sixteen seeds the degenerate Negate hand-off ends with $\|W_U\| = 39.4$ against the structured Coefficients hand-off's 29.2, a factor of 1.35, while $\|W_E\|$ moves the other way, 2.19 against 2.78. We find the arm carrying the larger readout, the 8k Negate hand-off's, is the one whose splice returns the lottery, so the effect is not monotone in readout scale.

Hand-off weight norm, in full. We find stage-end $\|\theta\|$ falls monotonically along the chain over 16 seeds, 52.1 ± 3.0 after the degenerate Negate, 45.9 ± 1.4 after Add and 40.7 ± 2.3 after Coefficients, and the degenerate stage's hand-off matches where a cold run ends (51.6 ± 2.1). The provable dependence of the grokking delay on initialization scale (Lyu et al., 2024) therefore predicts our ordering without any basin selection. We therefore initialize sixteen cold starts at each of the two hand-off norms, rescaling every parameter of the cold initialization θ_0 by one factor,

$$\theta_0 \leftarrow \frac{n}{\|\theta_0\|_2}\,\theta_0, \qquad n \in \{40.7,\ 52.1\}, \tag{4}$$

and changing nothing else. The favorable 40.7 arm reaches 0.450 ± 0.323 with 3/16 above 0.80, level with the cold-start distribution (0.483 ± 0.313, 4/16). The unfavorable 52.1 arm reaches 0.793 ± 0.134 with 11/16, short of the full curriculum's 14/16 on every statistic. And the paired difference between the arms runs opposite to that prediction (-0.343, permutation $p = 0.001$). End-of-run norms are consistent with the 40.7 arm's null: the two arms finish at 51.14 ± 1.77 and 51.67 ± 3.25, having started 11.4 apart, so optimization re-equilibrates the norm regardless of where initialization put it, as it does after the one-shot rescale of Truong et al. (2026).

D.5 Nulls on the end state

Linear mode connectivity is the wrong instrument. The natural test of a basin claim is to interpolate between solutions and look for a barrier, and it fails here for a known reason. Between independently initialized solutions, linear interpolation is dominated by permutation symmetry and shows a barrier whether or not the endpoints occupy the same basin (Entezari et al., 2022; Ainsworth et al., 2023). Result: the barrier between two generalizing solutions (2.78) is the *largest* in the set, above generalizing-to-failing (2.23) and failing-to-failing (2.12), which reproduces that signature rather than the naive prediction. Verdict: uninformative, and not used as evidence; a permutation-aligned measurement (Ainsworth et al., 2023) would be the correct instrument.

The end state does not reveal the outcome. We ran two separate probes of the trained network's end state; neither separates outcomes. We find the first principal component of the endpoint weights does not separate generalizing from failing runs. The second probe examines attention at the separator position across all 32 primary runs and finds no metric that distinguishes condition or outcome: separator-attention entropy, grand-mean attention to integer positions, the largest single-head attention to them, and the count of heads specialized on them; under a Welch test (Welch, 1947), none separates cold start from the full curriculum or generalizing runs from failing ones. Refuted: an earlier single-checkpoint observation had suggested a head-specialization signal, and it did not survive the full sweep. Both nulls point the same way and are consistent with §5's account: the outcome is fixed early, by the trajectory, and, beyond the embedding's end-of-run rank (Appx. D.2), is no longer legible in the geometry of the point the run ends at.

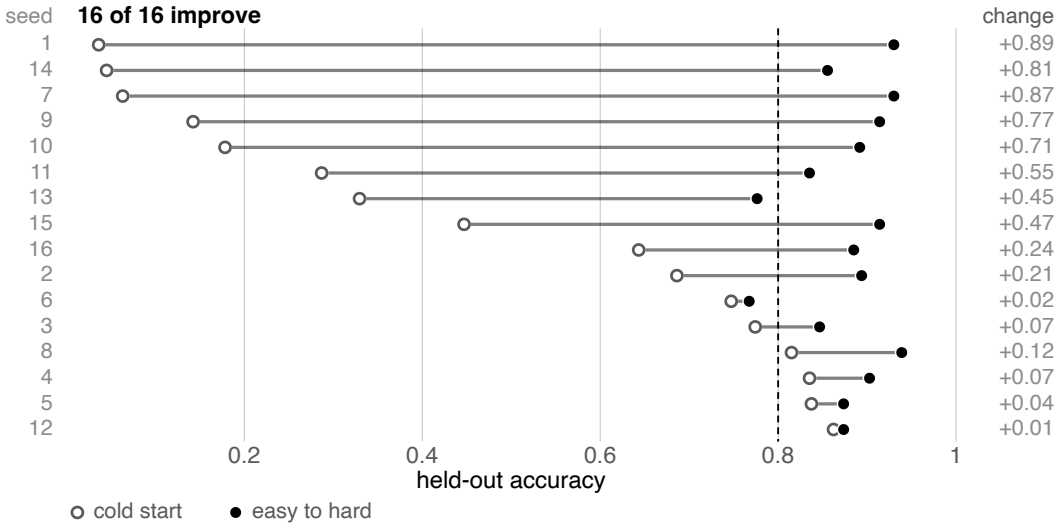


Figure 11: The same sixteen seeds before and after, ordered by where each started, with target-stage gradient steps matched. No seed is worse with the curriculum, and the seeds that start lowest move furthest, which is what collapses the spread rather than shifting it.