INERT ON ALGEBRA, DISRUPTIVE IN-CONTEXT: A CAUSAL LEDGER FOR RANK COLLAPSE IN GROKKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective-rank collapse is one of the most reproducible structural correlates of grokking. We ask whether preventing it changes generalisation, using a rank pin on the token embedding that is magnitude-matched online to the force driving the natural collapse, and we run the same pin on both sides of the comparison. On modular arithmetic the collapse is dispensable: pins at 0.9 times initialisation leave grokking on time under two pinned rank estimators and under a weight-decay-free optimiser, and still generalise on the attention-only architecture our in-context tasks use. Capping the embedding's Fourier concentration only paces grokking, with a pacing threshold of 0.118 at a 60k-step horizon (profile interval 0.110 to 0.127). On two sibling in-context circuits the same pin has the opposite effect. On bursty in-context labelling (BICL) none of 5 pinned seeds generalises by 1.3M steps, 45 to 97 times each seed's own unpinned grokking step, with floors held; on 1-hop induction the circuit does not form in any of 8 pinned seeds, against 8 of 8 paired baselines (exact paired p = 0.004). Each circuit has its own stable-rank threshold, 0.600 at 200k steps (bootstrap 0.550 to 0.666) and 0.845 at 75k (profile 0.804 to 0.895), and neither pre-registered normalisation carries one to the other, so the threshold is task-local. The block is not carried by the value of the global rank statistic, since matched random pushes keep rank near initialisation and generalise; it tracks how firmly the constraint binds. Four pre-registered mechanism attempts returned negative, so the mechanism remains open. Under standard softmax cross-entropy, weight decay drives the post-memorisation trajectory, and more than a quarter of the collapse happens after generalisation. Finally, the same statistic is timing-blind when read from the input embedding and tracks the grokking step when read from the unembedding, which explains how the correlate was misread.

1 INTRODUCTION

Grokking, in which training accuracy saturates long before validation accuracy rises, is a model organism for phase transitions in learning (Power et al., 2022; Nanda et al., 2023). One of its most reproducible structural correlates is *effective-rank collapse*: as the network groks, the energy of its weights concentrates onto a few directions (Yunis et al., 2024), and the embedding has been causally implicated in the delay (AlquBoj et al., 2025). Is that collapse *necessary* for grokking, or a by-product of how these networks are trained?

We answer with a measured contrast, obtained by running one causal intervention on both sides of a dissociation. The intervention is a *pin* that holds a chosen rank statistic near its initial value, magnitude-matched on each task to the force that drives its natural collapse, and replicated penalty-free by hard projection (Fig. 1). **The same rank-preserving intervention is causally inert on algebra and causally disruptive, with a measured dose threshold, on two in-context circuits.** On modular arithmetic the pin leaves grokking on time with the rank of either W_E or W_U held high, even without weight decay, and nothing we tested blocks at an adequate horizon. Capping the embedding's frequency concentration only paces grokking; the block it appeared to cause at short horizons was a horizon artifact. On two in-context circuits the same pin blocks generalisation on one and prevents circuit formation on the other, while matched random pushes do not. Cross-

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.

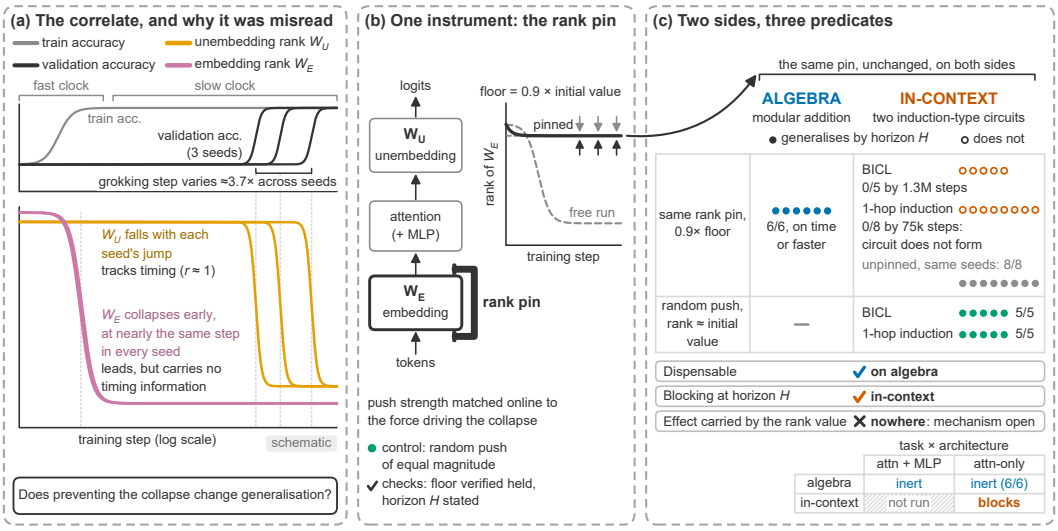


Figure 1: **One instrument, two sides, three predicates.** *(a)* Schematic, not data: W_E's rank collapses early at a nearly seed-invariant step, so it leads yet is timing-blind, while W_U's falls with each seed's jump. *(b)* The pin: floors on W_E's rank statistics at $0.9\times$ init, the push matched online to the force driving the collapse; control, a magnitude-matched random push; every verdict read with the floor checked and its horizon H stated. *(c)* Registered cell-level readouts, one dot per seed (Sec. 4). Algebra: the same-object pin generalises $6/6$ by 60k, on time or faster (statistic at $\approx 0.87\times$ init against 0.14–$0.21\times$ unpinned, floor strictly held in $2/6$; App. L). In-context, two sibling circuits of one induction family, floors held: BICL $0/5$ by 1.3M, and on 1-hop induction the circuit does not form ($0/8$ by 75k against $8/8$ unpinned, same seeds). Matched pushes keep rank $\geq 0.91\times$ init and generalise $5/5$ on each. Inset: the MLP is not what makes the pin inert on algebra ($6/6$ without it, median $0.92\times$ paired); the MLP-equipped in-context cell is unrun, so architecture is not excluded (Sec. 8). The block is not carried by the value of the global rank statistic; its mechanism is open.

controlled as the same object, the pin leaves modular grokking on time or faster. Each circuit gets its own measured floor, and neither registered transfer of that floor survives. The effect survives confinement to the complement of the circuit subspace, is not carried by the global rank scalar, and has no mechanism we could identify (Sec. 4). Around the contrast sits the anatomy that explains how the correlate was misread. The *same statistic* plays *opposite roles* depending on which matrix it is read from (Sec. 3), and the difference turns on two properties that progress-measure readings conflate: having a *lead* and predicting *when*.

The anatomy rests on a separation made explicit in Sec. 2: a signal that fires at a seed-invariant time has a large lead and carries no timing information.

Contributions. **(i) A causal contrast with measured thresholds (Sec. 4).** On algebra, pins on W_E or W_U leave grokking on time, robust to method, estimator and driver, and the one pin we ran with the MLP removed still generalises; nothing tested blocks at an adequate horizon, and caps only pace. On bursty in-context labelling (BICL) and 1-hop induction ("v2") the same pin, cross-controlled as the same object, blocks generalisation on the first and prevents circuit formation on the second, each at a stated horizon and with its own dose threshold, which transfers under neither registered normalisation. **(ii) A negative identification result (Sec. 4).** A blocking intervention does not identify the statistic it constrains. Matched random pushes generalise at near-initial rank, weakening the complement-confined pin removes its block, and four registered mechanism attempts returned negative. We therefore claim dispensability on algebra, blocking in-context, and operative status for the value of the rank statistic *nowhere*. **(iii) The collapse's driver and cost (Secs. 4, 5).** Under standard softmax cross-entropy, weight decay drives the whole post-memorisation trajectory; 27% of the collapse is post-generalisation, a trade that a label-free trigger makes explicit. **(iv) Same statistic, opposite roles (Secs. 2, 3).** W_E's collapse has the largest lead and is timing-blind (TOST within $|r|<0.3$, $n=130$), while W_U and the Fourier concentration track at $r\approx 1$. Blindness attaches

to the clock a detector point inherits, not to earliness. **(v) Basis-generality, an off-island boundary, and a calibrated predictor (Secs. 5, 6).** On S_5/A_5 the readout moves to a basis-specific irrep meter. Off the island, tracking transfers while strict blindness fails on 1-hop induction and holds on BICL. Delayed generalisation exists only where the pool size fits a measured window $[N_{\mathrm{gen}}, N_{\mathrm{mem}}]$. An online threshold reaches 7.3% median error at a 15% lead (n=30), where prior early warnings lead 5–6$\times$ longer and time near predict-the-mean: lead is not warning value.

Rank collapse at grokking is known, as are leading signals (Sec. 7); our contribution is the causal, cross-controlled contrast with measured, horizon-indexed thresholds. All evidence is from small transformers on algorithmic and synthetic in-context tasks (Sec. 8).

2 SETUP

Task and model. A one-layer attention-plus-MLP transformer (no LayerNorm; d_{model}=128, 4 heads, d_{mlp}=512) computes $(a \circ b) \bmod p$ at the final position of $[a, b, =]$, following Nanda et al. (2023): $p \in \{97, 113\}$, $\circ \in \{+, -, \times\}$, 30% of pairs as training set, full-batch AdamW (η=10^{-3}, β=$(0.9, 0.98)$), weight decay $\lambda \in \{0.5, 1.0, 2.0\}$; unless noted p=113, $\circ$=$+$, λ=1.0.

Signals and definitions. At log-spaced steps we record weight norms, the embedding Fourier spectrum and its concentration (Gini, entropy, top-5 fraction), and the participation-ratio *effective rank* of W_E and W_U, which measures energy concentration rather than hard rank. The causal floors also use the head-sensitive *stable rank*, in whose units the in-context thresholds are reported (formulas and all three estimators: App. J). A signal's *transition* is the midpoint crossing between its plateaus (also onset 10%, late 90%); T_{grok} is validation accuracy's midpoint crossing; *lead* is T_{grok} minus transition; *timing correlation* is the cross-seed correlation between them; completion fractions φ are means of per-seed ratios, never ratios of means (App. E). A detector point is *timing-blind* when its transition carries no timing information, certified by a permutation floor and, where powered, a TOST bound. The ratio $\mathrm{Var}(\mathrm{lead})/\mathrm{Var}(T_{\mathrm{grok}})$ measures earliness rather than blindness and is reported descriptively only (Sec. 6). Measurement detail and robustness grids: App. E.

Three predicates, kept distinct. A statistic is *dispensable* when holding it near initialisation still permits generalisation. An intervention is *blocking at horizon H* when, with its constraint verified held, no run generalises by H (horizons: Sec. 8). A statistic is *causally operative* when an intervention's effect is carried by that statistic. We establish the first on algebra, the second in-context, and the third nowhere.

Leading versus predicting. A signal *leads* if it completes part of its change before the validation jump; it carries *timing information* if the step at which it reaches a fixed milestone co-varies across runs with the grokking step. Training splits into a fast phase whose milestones are nearly seed-invariant, completing by T_0, and a slow phase of seed-variable duration $D(s)$, so $T_{\mathrm{grok}}(s) = T_0 + D(s)$. A *detector point* $\tau(s)$ is the step at which a signal completes a fixed fraction of its change. If τ falls in the fast phase, and is therefore independent of D, three things follow: τ carries zero timing correlation, its lead is large, and no predictor built on it beats predicting the mean (Obs. 1; a detector-noise version is in App. F). The derivation is elementary, which is why the blindness is structural rather than a small-sample accident. The most prominent leading indicators behave as this class predicts: the published commutator defect fires at a near-fixed step, yet it is read as a predictor because it leads. The empirical content lies where the arithmetic is silent, in which detector points inherit the fast clock. Fig. A8 catalogues this for every signal and detector point across three powered task families (App. F, where a stronger pre-registered monotone-φ reading is refuted).

3 SAME STATISTIC, OPPOSITE ROLES: W_E LEADS BLIND, W_U TRACKS

All 48 runs (6 conditions $\times$8 seeds) grok, and the grokking step varies widely at fixed hyperparameters ($\approx$3.7$\times$ across seeds at the base setting), which is what makes the timing question meaningful. W_E's effective rank has the largest mid-transition lead of all signals, yet its transition step is nearly seed-invariant (std 408 against $\sim$4000 for T_{grok}). At n=130 (130/130 grok) both W_E detector points are equivalence-confirmed within a pre-registered $|r| < 0.3$ bound (onset Spearman 0.12;

midpoint 0.06). The Fourier-concentration signals and W_U's rank instead reach their transitions in lockstep with T_{grok} (r=0.99–1.00) with small lead: they say that the network is at the transition, not ahead of it. The dividing line is W_E versus W_U, not effective rank as a statistic, and a functional witness ties it to functional progress (Fig. A5). The run has two clocks. W_E's collapse runs on the fast, seed-invariant one (φ=0.43 at midpoint, though its own completion point at $\varphi\approx1.5$ tracks); it is an early, decay-driven compaction of the input representation. W_U and the Fourier spectrum index the output-readout circuit whose formation is grokking (Figs. A3–A6; App. C).

4 THE CAUSAL CONTRAST: INERT ON ALGEBRA, DISRUPTIVE IN-CONTEXT

The dissociation is correlational, so we intervene. A differentiable auxiliary loss prevents a chosen signal from forming, with its gradient magnitude matched online to the cross-entropy gradient on the same matrix and its target set at the signal's natural value (one unsatisfiable early ceiling is disclosed in App. B). Two matched controls, a large non-readout matrix and a magnitude-matched random push, together with a mechanism-fidelity gate exclude degenerate shortcuts (App. B, Fig. A2). A floor on a rank statistic $\mathrm{stat} \in \{\mathrm{PR}, \mathrm{SR}\}$ adds $\lambda_t A$ to the loss, with

$$A = \mathrm{relu}\big(0.9\,\mathrm{stat}_0 - \mathrm{stat}(M)\big)^2, \qquad \lambda_t = \beta\,\frac{\|\nabla_M \mathcal{L}_{\mathrm{CE}}\|_F}{\|\nabla_M A\|_F},$$

where stat_0 is the value at initialisation and λ_t is recomputed at every step on the constrained matrix M. In-context, the cross-entropy gradient is negligible after memorisation, so the pin matches the weight-decay force instead, replacing the numerator with $\lambda_{\mathrm{wd}}\|W_E\|_F$ (β=10).

Rank pins leave algebraic grokking on time under two pinned estimators, penalty-free, and without weight decay. Preventing the effective-rank collapse of either W_E or W_U leaves grokking on time ($\Delta T_{\mathrm{grok}} = +50/+100$ steps, $10/10$), as does the matched non-readout control (the matched random push is on time at the median, -100, but speeds up two of ten seeds by over a third). The pattern holds across the weight-decay range, subtraction, p=97, a two-layer and a wider model ($6/6$ each), and replicates point-for-point under penalty-free hard projection (App. H). With the MLP removed, the one pin we ran, the same-object pin below, is inert too: on the two-layer attention-only in-context architecture it leaves modular addition generalising $6/6$ at a median $0.92\times$ its paired unpinned step (pre-registered; $4/6$ inside the registered on-time window; train fraction 0.4; App. L). The claim is estimator-robust, which our first floor was not: the participation-ratio floor is tail-dominated, and under it the head-sensitive *stable rank* still completed $\approx80\%$ of its natural collapse. Imposing the floor on the stable rank itself (at $3\times$ matched strength; the matched-strength arm lost to the CE gradient), or on both floors at once, still groks $6/6$ (paired median $+550/+50$ steps, against $+2850$ for the matched random push). The W_E-floored arms grok at the concentration their floors geometrically permit. Capped-but-grokking networks are not "still high-rank", as a first reading suggested. At ceilings 0.11–0.20 they end at the free-running stable rank, and looser ones keep more (0.225: median 19.5 vs. 6.9), so the cap paces grokking without reliably preserving rank, and rank-dispensability rests on the pin arms alone (App. J). Nor is the result a weight-decay artifact. Under $\perp$Grad (Prieto et al., 2025), which projects each gradient orthogonal to its own weight vector, the network generalises at λ=0 and collapses the rank as fully as decay does, and adding both floors still groks $4/4$ with rank held high. That driver-independence holds on the algebraic side only: the same $\perp$Grad diverges on the in-context task below ($0/4$; App. K). Refuting necessity and establishing it are not symmetric. One clean arm refutes necessity, because the outcome is positive and assumes nothing else about what the intervention does. Establishing necessity requires the intervention to act on the constrained statistic and on nothing else, which our own controls below deny. We therefore report the algebraic side as a refutation and the in-context side as a blocking effect whose operative variable is unknown.

Concentration caps pace; nothing we tested blocks at horizon. Preventing the embedding's *Fourier concentration* does not stop grokking at any satisfiable ceiling we extended; it paces it. At the original 15k–28k horizons this arm produced the $0/10$ and $0/6$ outcomes we first read as a categorical block. The $0/10$ is the original hinge, whose ceiling lies below the statistic's floor; never extended, it remains a 20k null. For satisfiable caps, our own horizon extensions dissolved the block (per-cap tallies: Fig. 2, left): no satisfiable cap we tested (0.095–0.225) categorically blocks, and the slowest grok arrives at step $254{,}100$. The dose therefore measures a horizon-conditional

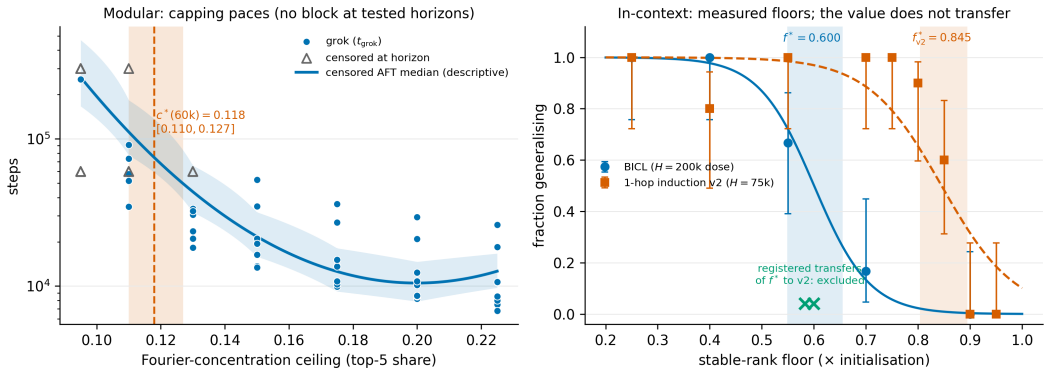


Figure 2: **The two sides of the contrast, measured.** *Left:* no satisfiable ceiling we tested categorically blocks. Each satisfiable cap's at-horizon null dissolved on extension, and divergence at some ceiling is unresolved (dots: t_{grok}; triangles: censored; curve: censored log-normal AFT, descriptive only). *Right:* the same pin on the two in-context circuits, with Wilson 95% intervals and censoring-aware fits. The v2 curve marks the fitted *location* only, since its slope is inflated by a $0.4\times$ meter artifact, and both registered transfers of f^* to v2 ($\times$) are excluded: the threshold exists on both circuits and is task-local. Bands are $c^*(60\text{k})$, $f^*(200\text{k})$, $f^*_{\text{v2}}(75\text{k})$; every fit, interval and horizon is in Sec. 4.

pacing threshold, fitted with the in-context side's censoring-aware model and always written with its horizon: $c^*(60\text{k}) = 0.118$, profile $[0.110, 0.127]$. Pacing is steep (median T_{grok} 9.6k $\to$ 57.5k as the ceiling tightens $0.225 \to 0.11$), and the transition sits above the statistic's uniform floor, so it is not a satisfiability tautology. Whether T_{grok} diverges at some ceiling or merely grows very large is not resolvable at our horizons. The constraint is basis-specific, and capped networks reach the task through a more distributed solution; on S_5 the same design paces in the irrep basis (Sec. 6). An "all-or-nothing necessity" reading of the abelian case therefore does not hold (cells, scoring and the frequency counts: App. B).

Off the algebraic island, the same pin blocks. Two synthetic in-context tasks carry the blocking side, both introduced in Sec. 6: *BICL*, a fixed-pool transplant of Reddy's task whose item$\to$label map is resampled every sequence, and *v2 1-hop induction*. "Generalise" means held-out-sequence accuracy, read alongside a circuit-strength meter. A same-object cross-control closes the instrument confound. The off-island pin (both floors, β=10, matched by a registered rule to the weight-decay force rather than to the CE gradient, which is negligible after memorisation there, ratio 6.2×10^{-4} on v2) was run unchanged on modular arithmetic, with the same statistic, floors, β, rule and code path. There it leaves grokking intact ($6/6$), faster than unpinned on $5/6$ seeds (the sixth at $1.04\times$). Its one registered prediction failed benignly: on time $4/6$ against a $\geq 5/6$ bar, with both misses faster than baseline (App. L).

On BICL, that pin at $0.9\times$ floors (both estimators) blocks: $0/5$ seeds generalise at a 1.3M-step horizon. That horizon is $6.5\times$ the dose horizon, $5.3\times$ the slowest grok in the whole unpinned cohort, and 45.5–$97\times$ each pinned seed's own unpinned grokking step, which on four of five seeds is past the $\approx 45\times$ paired multiple at which our last algebraic null dissolved and level with it on seed 0 ($45.5\times$ vs. $45.4\times$, within logging resolution). The floor held over the full window in all five cells and in the nine runs of the earlier 400k certification (5 at $0.9\times$ plus 4 at $0.7\times$), the circuit meter is TOST-flat over the final 200k steps, and the re-run replays its own archived 400k verdicts (App. L). The registered wording is *the $0.9\times$ floor lies above the fitted f^* CI upper bound*, a horizon-indexed claim. Paired on seed, the delay is graded in the floor, from a median $1.0\times$ their own unpinned step at $0.25\times$ to $5.3\times$ at $0.7\times$ (App. L); the paired figures are post-hoc, and the pinned seeds are among the fastest 12 of 16 baseline candidates, which lowers every denominator. So *even read as extreme pacing rather than arrest, the contrast stands*: the same pin leaves modular grokking on time or faster. Our last algebraic null dissolved at $\approx 16\times$ the slowest baseline grok and $\approx 45\times$ its own seed's, against $5.3\times$ and 45.5–$97\times$ here, so the precedent is passed on the paired scale (on four of five seeds, level on the fifth) and not on the cohort scale. A matched-magnitude random

push groks $5/5$ at 0.91–$1.05\times$ initialisation rank, so the block is specific to the floor constraint and not to the high global rank the push also produces. Nor is the push mild: measured on the v2 cells, it replaces nearly all of W_E's energy (App. L), so a near-totally replaced embedding at near-initial rank generalises while the pin blocks. A pin that lets the network fully memorise still blocks generalisation, so it stops the circuit, not memorisation. The block is graded in the dose over the stable-rank floor (12 seeds per floor, $12/12$ at $\leq 0.4\times$ falling through $8/12$ and $2/12$ to $0/12$ at $0.9\times$, bound in every cell; per-floor tallies in App. L). Fitted censoring-aware over all 60 runs, the dose locates $f^*(200\text{k}) = 0.600$ (boot $[0.550, 0.666]$, profile $[0.549, 0.655]$; Fig. 2, right), about $3\times$ the natural collapse endpoint.

A second circuit: the threshold exists again, and its value does not transfer. On v2 1-hop induction at pool size $N{=}15360$ (a registered escalation after the gate failed on the baseline side at 12288), the same pin construction prevents the circuit from forming at its 75k horizon: pin $0/8$ against baseline $8/8$ on the same eight seeds (exact paired $p{=}0.0039$, registered direction; the registered Fisher value assumed independent runs), matched push $5/5$, floors held $8/8$. We say "prevents the circuit from forming (no generalisation at all)" rather than "blocks grokking" because raising N converts stuck seeds into immediate generalisers, never delayed ones (the delayed count is exactly 3 at every N probed). A registered memorisation-decay disclosure travels with the pin cells: every pinned seed fully fitted the training set and then decayed off it, so "stops the circuit, not memorisation" holds here only in the memorisation-intact-window form (per-cell numbers, App. I). The dose gives this circuit its own threshold: $f^*_{\text{v2}}(75\text{k}) = 0.845$ (90 cells, floor held $90/90$; profile $[0.804, 0.895]$, boot $[0.811, 0.868]$), graded from $10/10$ at $0.25\times$ to $0/10$ at 0.9/$0.95\times$. The $8/10$ dip at $0.4\times$ is a meter artifact, which is why the fitted slope is never quoted. The two pre-registered transfers of BICL's threshold, init-normalised (0.600) and endpoint-normalised (0.583), are both excluded (half-widths 0.046/0.028): *the threshold transfers as an existence claim and fails as a quantitative one*, so any account of the floor must explain a task-specific constant.

Mechanism: open. Three further pre-registered results (four registered attempts, since (iii) covers two derivations) locate the blocking effect without explaining it. *(i) The operative variable is not the global rank scalar.* The matched push generalises at 0.91–$1.05\times$ init rank on BICL and at or above it on v2. At the $0.55\times$ floor the same dose both groks and blocks at overlapping final stable ranks (19.95–21.87 against 20.76–21.71; the strictly rank-matched contrast is underpowered, $p{=}0.165$, so the claim rests on the overlap itself). The circuit occupies a tight subspace on both tasks (medians 15 and 11 of 128 dimensions), and a pin confined entirely to the complement of BICL's subspace, with the complement statistics floored at $0.9\times$ their own init values and the circuit directions left free, blocks $5/5$ (held $5/5$; seed 4 still blocked at 400k, the other four at 300k; multiples in Sec. 8). That block leaves open a deflationary reading, that the total rank never got low enough, and we tested it. Re-running the identical construction on the identical five seeds with the complement floor weakened to $0.55\times$ drops the total stable rank to 9.2–11.5, the free-running endpoint, and the block disappears: $5/5$ blocked at $0.9\times$ against $0/5$ at $0.55\times$, floors held in every cell, with the achieved total rank tracking the outcome on every seed (exact paired $p{=}0.031$ one-sided, 0.0625 two-sided). The contrast spans an engine change, so a registered drift gate re-ran the archived cell closest to flipping before it was read; it was still blocked, to three decimals (App. L). Within this construction the block tracks how firmly the floor binds, not which directions it holds: the same directions, held less firmly, do not block. "How firmly" is a property of the intervention rather than the achieved value of a rank statistic, which across arms is not monotone in the outcome. Because the sweep fixes the subspace, it cannot show that direction identity does no work (App. L). *(ii) No evidence of direction identity, at two floors.* Confining the pin's *corrective gradient* to the maximally cross-talking off-circuit subspace, against a matched random one, separates neither arm from the other. Both grok at a $0.4\times$ total floor, which an isotropic gate shows is too weak to discriminate, and both generalise at $0.55\times$ ($5/5$ and $4/5$), where the unconfined dose gives both outcomes (Fisher $p{=}1.000$). Two pre-registered scorings of "the pin stayed intact" disagree on how many cells are evaluable, so that readout fails our own rule and is reported as unsupported (App. L). The blocking and non-blocking constructions differ in floor magnitude, subspace dimension and floor-carrying statistic at once, so no single-factor attribution is available. *(iii) Deriving the threshold failed twice.* A registered estimator (interference occupancy in the measured circuit subspace) postdicted f_{pred} median 0.180 against a registered $[0.55, 0.70]$ bar with a $7.8\times$ per-seed spread, and its one registered prediction also failed. A second, pin-independent derivation, with every term measured on freely trained networks and no intervention in it, was pre-registered in full before any held-out quantity was

computed and was refuted, in the wrong direction: 0.547 predicted against 0.845 measured, while v2 has the smaller circuit and the higher threshold, so no re-normalisation repairs the ordering and the family is excluded (arithmetic and riders, App. L). The measured thresholds stand as measurements; no derived threshold is claimed in this paper. What the collapse builds is measured descriptively in App. L: partial role structure without generalisation.

The driver and its cost. The network memorises at step $\sim$250. At λ=0 the training gradient then dies of Softmax Collapse, and nothing moves at any initialisation scale we tried. At λ>0, under standard softmax cross-entropy, weight decay drives the post-memorisation trajectory and regenerates the small live cross-entropy gradient that the modular pins are matched to. Regulariser-free generalisation is recoverable only by preventing the logit scaling that kills that gradient ($\perp$Grad, not loss-numerics fixes; App. K). Weight decay minimises norm, not error, so it does not stop when the task is solved: 27% of the rank collapse (38% of concentration) happens after generalisation, roughly three quarters of that is decay, and switching λ off there trades 1.8 validation points for nearly half the remaining rank (33.1 against 18.9). On modular arithmetic none of the observed collapse is necessary, while 73% of it is concurrent with generalisation and its tail is a norm-minimiser that does not know the job is finished. We make no sufficiency claim, because forcing a signal toward its grokked value is confounded with installing the answer. A grafting probe supplies the direction: a transplantable, basis-specific accelerant that never generalises alone (App. H).

5 From a post-hoc diagnostic to an online predictor

The fraction-of-rise diagnostic presupposes T_{grok} on both axes, so it is post-hoc only (App. A). What is observable online is an absolute threshold on Fourier-Gini. Under nested leave-one-seed-out calibration it predicts T_{grok} to 7.3% median error at a 15% lead on the powered n=30 set (7.8% at 18% on the n=8 base set), against 27–35% for cheap baselines, and calibrated on the base condition alone it carries to four held-out modular conditions and to S_5, but not to D_{60}. Sweeping the threshold traces the full lead-versus-error frontier (Fig. A9). Every prior early warning sits inside it at its published threshold, and at long lead every method approaches predict-the-mean: lead is not warning value. The live corner, completion, supports the ledger's corollary: fire at θ of the calibrated rise, then set $\lambda := 0$, which turns the always-on decay default into an explicit accuracy-versus-retained-rank choice (tallies, the θ grid, transfer percentages and the corollary's four caveats: App. M).

6 Beyond modular arithmetic: non-abelian groups and the off-island boundary

Non-abelian groups: the readout moves to irreps. Within the algebraic family the dissociation holds across all five additive conditions ($40/40$) and survives the meter change that multiplication forces; the quadratic family is a degenerate, underpowered boundary (App. C, D). On S_5 composition (same transformer, enumerable Cayley table) the structural meter is *irrep* concentration, computed from a character spectrum that reduces exactly to the DFT Gini for cyclic groups (App. C). The dissociation transfers. W_U rank tracks (r=1.00), and the irrep concentration tracks (r=0.97; Fig. A4) and is *basis-specific*, since a random relabelling destroys tracking (modular's specificity is instead established causally). W_E's rank *onset* is blind at every n we powered, while its *completion* migrates onto the grokking clock at n=30: onset-blindness is the common algebraic anchor, and completion behaviour is group-dependent. The replication extends to A_5 ($15/15$; irrep meter r=0.94, basis-specific; relative onset blind). A pre-registered advance-prediction test on dihedral D_{60} held its recipe and tracker predictions while its relabel control failed its bar, so we make no basis-specificity claim there (seed counts, both statistics and the D_{60} cells: App. N, G). Causally, capping the irrep concentration delays S_5 monotonically in the cap, with the relabelled-basis control only mildly slowed and rank prevention on time $6/6$. Capping therefore paces on both algebraic substrates, and the substrate-redundancy reading survives only as an interpretation (App. N).

Off the island: two tasks, one circuit, opposite ends of a measured spectrum. On a 2-layer attention-only model forming a known induction circuit (Olsson et al., 2022), a first design with

a memorisable lookup table never grokked ($0/12$, the in-weights/in-context competition); removing the shortcut makes it grok at an intermediate pool size predicted in advance. The *tracking* half transfers exactly (circuit-strength completion r=1.00), and effective rank does collapse. The *strict blindness* half fails there: the earliest input-side onsets rank-correlate with T_{grok} (0.87/0.79 at n=19 against a 0.45 floor; $\mathrm{onset} \propto T_{\mathrm{grok}}^{\gamma}$, γ=0.28 $[0.19, 0.36]$ on the drop-from-init detector), robustly across 15 detector definitions and the registered artifact checks (App. I). A pre-registered variance conjunct passes in all ten task×detector cells, coupled ones included, so we retire it; no positive blindness claim on this task survives (App. N). A second non-algebraic task we make grok, BICL, whose item→label map is resampled every sequence, comes out blind instead: γ=0.031 $[0.001, 0.074]$ on the same detector (the registered primary), excluding induction's interval and inside the algebraic range, where modular and A_5 show positive elasticities (App. N). Its W_E onsets fire at steps 200–250 while T_{grok} spans 8.6×, so an unstructured in-context task can carry a seed-invariant clock, and two pre-registered ablations fail to restore coupling. The boundary is therefore not algebraic versus non-algebraic but whether the W_E-rank collapse runs on a seed-invariant fast clock, and two tasks sharing one circuit span the range from coupled to blind.

Why some tasks cannot grok at all: the window $[N_{\mathrm{gen}}, N_{\mathrm{mem}}]$. Delayed generalisation needs a pool size at which memorisation still succeeds and the circuit is preferred. Measured on three non-algebraic tasks, the window is non-empty on 1-hop induction and BICL and measurably empty at this model size on a group-matching task. Where a data-hungry circuit pushes N_{gen} past N_{mem}, the delay disappears because the regime does not exist, not because hidden progress is slow (Barak et al., 2022) (endpoints, ceiling control, weight-decay direction: App. N).

7 Related work

Progress measures and rank. Nanda et al. (2023) define progress measures; Yunis et al. (2024) show that low-rank weight solutions coincide with generalisation; and AlquBoj et al. (2025) give the embedding a causal role while reporting its rank as stable. We reconcile the two: the matrix stays full *hard* rank ($113/113$ throughout) while its energy concentrates. Sakamoto & Sato (2026) tie grokking to within-class variance contraction and Clauw et al. (2024) give an information-theoretic order parameter; our character meter presupposes neither the irrep (Chughtai et al., 2023) nor the coset (Stander et al., 2024) reading. Other accounts of grokking dynamics include Varma et al. (2023); Kumar et al. (2024); Merrill et al. (2023); Thilak et al. (2024); Liu et al. (2022; 2023); He et al. (2026).

Early warnings. Prior early warnings (Notsawo et al., 2023; Xu, 2026a; Salah & Yevick, 2026) neither state the lead–timing dissociation nor separate online from post-hoc readouts. The commutator defect and its companion (Xu, 2026b) both lead by construction, which is the property Sec. 2 separates from timing information. Intervention work related to the pin and the decay corollary is discussed in App. O.

In-context circuits. Our off-island tasks build on the induction-circuit and in-weights/in-context literature (Olsson et al., 2022; Chan et al., 2022; Singh et al., 2023; Reddy, 2024; Bietti et al., 2023). Wang et al. (2024) observe fixed-pool 2-hop grokking near 7×10^5 steps under atomic-fact mixing, beyond our horizons, so our compositional cells are inconclusive rather than negative. Our novelty is not that rank drops, but that the *same constraint* has opposite causal effects across task families while neither effect is carried by the global rank statistic. This literature draws neither dividing line (W_E versus W_U; statistic versus intervention).

8 Limitations

Scope. The dissociation survives depth and width (at n=6, as non-rejections rather than n=130-grade equivalence; Fig. A7); each tested negative removes one precondition of the estimand (Table 2); and three compositional tasks are horizon-limited and inconclusive. Off the island the trackers still track but W_E's blindness fails on 1-hop induction, so the two-clock anatomy does not transfer whole, and Omnigrok-MNIST groks with no low-rank solution at all (App. D). We make no claim for vision networks (Humayun et al., 2024) or natural-language benchmarks.

Task and architecture. The algebraic side runs on attention-plus-MLP transformers and the in-context circuits on a two-layer attention-only model, so the contrast could be architectural: an MLP might supply the slack that lets a pinned network route around the constraint. We tested that from the algebraic side, pre-registered. On the in-context architecture itself the same pin leaves modular addition generalising $6/6$ at a median $0.92\times$ its paired unpinned step, floors held, so *the MLP is not what makes the pin inert on algebra*. Three riders apply: that model needed train fraction 0.4 (the one seed probed at the 0.3 used elsewhere memorises and does not generalise by 60k); this is one algebraic task under one recipe; and the other crossed cell, an MLP-equipped in-context model, is unrun, so architecture is not excluded (App. L).

The in-context side: two circuits, one family; mechanism open; every threshold horizon-indexed. *(a)* The in-context side rests on two circuits, each with a gate and a fitted threshold CI, but the tasks are siblings (one exact-match induction circuit under two input distributions, one engine), so the evidence is correlated, and a structurally different low-rank circuit is untested. We tried to add one: a further pre-registered in-context task was killed at its ceiling control, its meter far above naive chance with no induction-shaped attention at all (App. P, rule 9). *(b)* The mechanism is open. Four pre-registered attempts returned negative (Sec. 4; the complement-floor weakening closes an objection we had raised against ourselves rather than supplying a mechanism), and quantitative transfer is excluded under both registered normalisations. *(c)* Every blocking arm is a rank hinge. Three pre-registered attempts at a non-rank control matched to it in magnitude and duty each failed a validity gate before yielding an interpretable outcome: the first ran, but its floor was slack through the grokking window, so its cells were identical to their unpinned baselines until t_{grok}; the others were abandoned at their pilots (App. L). A persistent counter-force of that size is therefore not excluded as the operative thing. The same-object cross-control constrains that reading without closing it, since after memorisation only weight decay is live in-context. *(d)* Every "block" verdict and every threshold is certified at a stated horizon ($c^*(60\mathrm{k})$; $f^*(200\mathrm{k})$, block certified at 1.3M; $f^*_{\mathrm{v2}}(75\mathrm{k})$), because our own satisfiable-cap "blocks" dissolved at 3–$15\times$ their original horizons. The BICL block carries the strongest certification we could give ($6.5\times$ the dose horizon, 45.5–$97\times$ paired, flat meters, held floors). The complement-confined block is thinner: four cells at 300k, which is $1.22\times$ the cohort's slowest grok and 4.6–$10.8\times$ paired, and seed 4 at 400k, which is $1.63\times$ and $14.6\times$ paired, stationary; the paired figures are post-hoc. Our last algebraic null dissolved at $\approx 45\times$ its own seed's grokking step. The BICL block passes that multiple on four of five seeds and is level with it on the fifth; the complement-confined one does not. A flat meter is weak evidence against a transition whose signature is a plateau. Both bound, but cannot close, delay at scales we did not run, and the contrast we claim, on time against a roughly forty-five-fold or larger paired delay, does not need them to. *(e)* The programme was pre-registered with per-cell readouts fixed before each run (App. L). Its author-prediction record is 4 held of 11 (itemised, with three further predictions outside that tally, in App. L); every failure was caught by a registered bar, and *no surviving claim depends on an author prediction having been right*.

What would close these. The circuit-projected rank at blocked endpoints (executable on the cells whose weights are saved) separates a drowned circuit from one that never formed; an MLP-equipped in-context model separates task from architecture; and a persistent non-rank counter-force matched in magnitude and duty, which three attempts failed to build, separates "a rank floor blocks" from "any held constraint of that size blocks".

Statistics. At small n we rely on permutation floors and TOST bounds rather than r alone; precision on onset-blindness comes from the $n{=}130$ TOST rather than from a pre-registered pooling that contains a disclosed failure; and the predictor forecasts a new seed, not a new task (Apps. C, N, G).

Conclusion. Grokking's rank-collapse correlate is real but conditional, and the condition is measured on both sides with one instrument. On algebra the collapse is dispensable and nothing we tested blocks at an adequate horizon; caps pace ($c^*(60\mathrm{k}){=}0.118$), and the block the cap appeared to cause at short horizons was a horizon artifact. On the two in-context circuits the same pin blocks at a $6.5\times$-dose-horizon certification and prevents circuit formation, each with a measured, task-local dose threshold, yet that block is not carried by the global rank scalar and its mechanism is open. Where dispensable, the collapse is produced by the dynamics that generalise the network rather than required by the task. Around it sits the anatomy: the same statistic leads but cannot time on W_E, and times but does not lead on W_U. The methodology is collected as twelve rules in App. P.

AI USE STATEMENT

Generative AI tools were used in this work for tasks in every category for which ICLR requires disclosure except translation: generating the synthetic datasets, developing the conceptual framework, formulating and proving the claim of Sec. 2, proposing and refining hypotheses, designing the experimental methodology, implementing the methods, analysing the data, and interpreting the results. They were also used for tasks in the recommended-disclosure categories: identifying and summarising related work, writing and editing code, producing figures, suggesting experimental parameters, and drafting and editing this paper. Concretely, the research programme was executed by an autonomous coding agent operating under the direction of the authors, who set the research goals, approved or declined each proposed experiment, and made every compute-allocation decision.

We describe the workflow in this detail because it is also what makes the results checkable. Every experiment reported here was pre-registered before it was run, in a protocol carried in the supplementary material (70 filed amendments, numbered to A73): the cells, the readout, the pass/fail bar, the kill criterion and the budget were fixed in writing before any GPU time was spent, and amendments are appended rather than rewritten, so the record shows what was registered and when. Only cell-level registered readouts are treated as confirmatory; every other quantity in the paper is labelled descriptive. The consequences are visible throughout the text. The author-prediction record is 4 held of 11 (Sec. 8); four registered mechanism attempts returned negative and are reported as negative (Sec. 4); a registered readout on which two pre-specified scorings disagreed is reported as unsupported rather than quoted at the value that favoured us; and two conclusions we reached at earlier stages of this project, that concentration caps block grokking and that capped-but-grokking networks remain high-rank (true only at loose caps), were refuted by our own later runs, and the paper reports the corrected results.

We have reviewed all AI-assisted work. Load-bearing empirical claims were re-checked against the archived run artifacts in separate adversarial verification passes, which caught and corrected errors introduced by earlier passes over the same text. We take responsibility for the final content of this work, including its text, its claims, and the artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work trains small transformers (roughly 10^5–10^6 parameters) on synthetic algorithmic and in-context tasks that we generate ourselves. It involves no human subjects, no personal, scraped or licensed data, and releases no model with plausible misuse potential. All experiments ran on a single consumer GPU (NVIDIA RTX 3080, $10\,\text{GB}$); the upgrade programme that produced the causal contrast consumed approximately 195 GPU-hours. The integrity risk specific to work of this kind is selective reporting of interventions that did not behave as predicted; the pre-registration described in the AI use statement exists to constrain it, and registered negatives are reported as negatives, in the main text wherever they bear on a claim.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code for every experiment reported here, the complete pre-registration protocol (A1–A73, 70 filed) that fixes each experiment's cells, readouts and pass/fail bars before it was run, and the run logs from which the reported numbers are computed. In the paper itself, the intervention protocol (the online gradient-matching rule, the capacity-preserving cap, the floor definitions for the two pinned rank estimators, the third being logged only, and the mechanism-fidelity gate) is specified in App. B, with the penalty-free hard-projection replication in App. H; the two in-context tasks and their generators are specified in App. I and App. N; the censoring-aware threshold fits, each reported with its method label and interval, are in App. L; the weight-decay-free arms are in App. K; and the proof of Observation 1 is in App. F. Every "block" verdict in this paper is indexed by the horizon at which it was certified, and that horizon is stated wherever the verdict is quoted, so a replication can be budgeted to match it. App. P distils the methodology into twelve rules, each stated with the incident in this project that produced it and what that incident cost; it is the part of this work that transfers beyond grokking.

REFERENCES

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.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764. Curran Associates, Inc., 2022. doi: 10.52202/068431-1581. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588. Curran Associates, Inc., 2023. doi: 10.52202/075280-0077. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/056173 8a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022. doi: 10.52202/068431-1371. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Kenzo Clauw, Sebastiano Stramaglia, and Daniele Marinazzo. Information-theoretic progress measures reveal grokking is an emergent phase transition, 2024. URL https://arxiv.org/abs/2408.08944.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024. doi: 10.1038/s41586-024-07711-7. URL https://doi.org/10.1038/s41586-024-07711-7.

Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking, 2026. URL https://arxiv.org/abs/2602.16849.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20722–20745. PMLR, 2024. URL https://proceedings.mlr.press/v235/humayun24a.html.

Maksim A. Kazanskii. How to tame grokking: Representation geometry as a control signal, 2026. URL https://arxiv.org/abs/2607.11666.

Tanishq Kumar, Blake Bordelon, Samuel Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, pp. 23010–23035, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/63ed15a46a143ff57484b38cd6b85d91-Abstract-Conference.html.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients, 2024. URL https://arxiv.org/abs/2405.20233.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34651–34663. Curran Associates, Inc., 2022. doi: 10.52202/068431-2511. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfc310e81992d2e4cedc09ac47eff13e-Abstract-Conference.html.

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

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks, 2023. URL https://arxiv.org/abs/2303.11873. ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo).

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 (ICLR)*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Pascal Jr. Tikeng Notsawo, Hattie Zhou, Mohammad Pezeshki, Irina Rish, and Guillaume Dumas. Predicting grokking long before it happens: A look into the loss landscape of models which grok, 2023. URL https://arxiv.org/abs/2306.13253.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

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.

Hari K. Prakash and Charles H. Martin. Late-stage generalization collapse in grokking: Detecting anti-grokking with Weightwatcher, 2026. URL https://arxiv.org/abs/2602.02859.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *International Conference on Learning Representations*, pp. 81151–81168, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9e6ac15e689e06139d7b39e1667b165-Abstract-Conference.html.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007. URL https://ieeexplore.ieee.org/document/7098875.

Keitaro Sakamoto and Issei Sato. Explaining grokking and information bottleneck through neural collapse emergence. In *International Conference on Learning Representations*, pp. 107660–107690, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/af5509c72a244497c999ac39ba068ff4-Abstract-Conference.html.

Ahmed Salah and David Yevick. Tracing the path to grokking: Dropout, embeddings, and network activation. *Neural Processing Letters*, 58(3):36, 2026. doi: 10.1007/s11063-026-11843-4. URL https://doi.org/10.1007/s11063-026-11843-4.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27801–27819. Curran Associates, Inc., 2023. doi: 10.52202/075280-1208. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58692a1701314e09cbd7a5f5f3871cc9-Abstract-Conference.html.

Achyuthan Sivasankar. Circuit synchronization precedes generalization: A causal precursor to grokking, 2026. URL https://arxiv.org/abs/2606.12966.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46441–46467. PMLR, 2024. URL https://proceedings.mlr.press/v235/stander24a.html.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua M. Susskind. The slingshot effect: A late-stage optimization anomaly in adaptive gradient methods. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=OZbn8ULouY.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023. URL https://arxiv.org/abs/2309.02390.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95238–95265. Curran Associates, Inc., 2024. doi: 10.52202/079017-3017. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad217e0c7fecc71bdf48660ad6714b07-Abstract-Conference.html.

Hongyi Wang, Saurabh Agarwal, Pongsakorn U-chupala, Yoshiki Tanaka, Eric Xing, and Dimitris Papailiopoulos. Cuttlefish: Low-rank model training without all the tuning. In *Proceedings of Machine Learning and Systems*, volume 5, pp. 578–605, 2023. URL https://proceedings.mlsys.org/paper_files/paper/2023/hash/c2db3ef0b1ad4c5ec7c3a0a0c6f6c832-Abstract-mlsys2023.html.

Yongzhong Xu. Early-warning signals of grokking via loss-landscape geometry, 2026a. URL https://arxiv.org/abs/2602.16967.

Yongzhong Xu. Low-dimensional and transversely curved optimization dynamics in grokking, 2026b. URL https://arxiv.org/abs/2602.16746.

Yongzhong Xu. The lifecycle of the spectral edge: From gradient learning to weight-decay compression, 2026c. URL https://arxiv.org/abs/2604.07380.

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model. In *International Conference on Learning Representations*, pp. 100828–100856, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa5ddd6bac0d665c72969d79221b680a-Abstract-Conference.html.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Henrique Pamplona Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew Walter. Grokking, rank minimization and generalization in deep learning. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=6NHnsjsYXH.

A THE POST-HOC FRACTION-OF-RISE DIAGNOSTIC

Measuring the fraction $f(t)$ of the Fourier-concentration rise completed maps, across seeds, onto proximity $\tau = t/T_{\mathrm{grok}}$ with cross-seed collapse std 0.039 (e.g. $f{=}0.5$ at $\tau = 0.58 \pm 0.07$), whereas W_E's fraction saturates by $\tau \approx 0.37$. This characterises the mechanism (the normalised rise has a seed-invariant profile) but is not an online quantity: computing f needs the final plateau, and the τ

axis needs T_{grok}. Because T_{grok} enters on *both* axes, the diagnostic is partly circular and we do not use it as evidence for timing-tracking anywhere in the paper. By contrast, Fig. A1 shows the online predictor of Sec. 5, whose plateau anchors, threshold, and regression use no information from the held-out run.

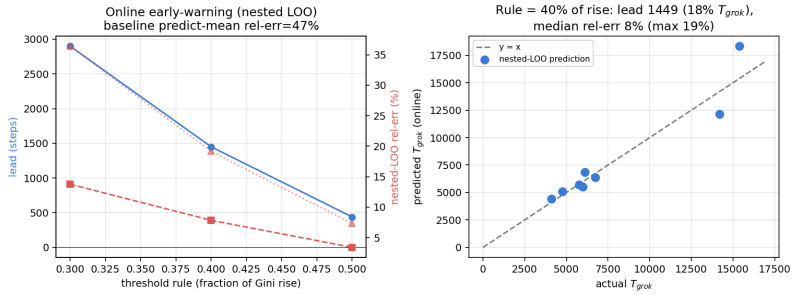


Figure A1: **The online predictor (nested LOO).** Left: as the absolute Gini threshold rises, lead shrinks and accuracy improves; predict-the-mean baseline 47% on this base $n{=}8$ set ($\sim$30% on the powered $n{=}30$ set of Sec. 5). Right: at the "40%-of-rise" rule, online predictions of T_{grok} (anchors, threshold, and regression from training seeds only) land near $y{=}x$ with median error 8% at a $\sim 18\%$ lead.

B Causal intervention protocol details

Arms (all 10 seeds, base setting; 6 seeds per hardening condition): PREV-Erank / PREV-Urank (hinge penalties holding the participation-ratio effective rank of W_E / W_U at its natural pre-grokking value), PREV-Efourier (hinge on the top-5 additive-DFT power fraction of W_E), PREV-large (MLP W_{in} rank, a large non-readout control), CTRL-mag (magnitude-matched random perturbation), CTRL-up (wrong-direction push). CTRL-up is not an independent control: it was registered as the wrong-direction control for a sufficiency arm (pushing W_U's rank *down*) that this paper no longer reports, it is the same hinge on the same matrix as PREV-Urank, and its cells are numerically identical to that arm's. The matched controls of Sec. 4 are therefore PREV-large and CTRL-mag. The auxiliary gradient norm is re-matched to the CE gradient on the same matrix at every step ($\beta{=}1$; the CE-gradient norm used for the match is computed in-run, at scale ~ 0.04 post-memorisation, and is not among the released per-step columns). TOST margin: the registered margin is the largest generic control shift ($\sim$100 steps). Under it no arm is equivalence-confirmed at $n{=}10$ (90% bootstrap intervals: PREV-Erank $[-30, 190]$, PREV-Urank $[-10, 250]$, PREV-large $[-90, 480]$). After seeing those intervals we adopted, post hoc, a margin of 10% of baseline T_{grok} (615 steps), under which the three are equivalent. CTRL-mag is equivalent under neither margin ($[-2300, 110]$: median -100, but two of ten seeds grok more than a third earlier). The rank-dispensability reading therefore rests on the effect sizes ($+50$/$+100$ steps, $\leq 2\%$ of T_{grok}) and on $10/10$ grokking, not on the equivalence test. The mechanism-fidelity gate reads the staggered per-frequency emergence signature from saved spectra checkpoints; for the Fourier-prevention arm this signature is defined on the suppressed quantity, so that arm's at-horizon verdict is adjudicated by the validation-accuracy gate plus the independent output-side witnesses of Sec. 4 (W_U effective rank and W_U Fourier content, which the constraint never touches; censored (no-grok-by-horizon) runs fail all witnesses, dose-restored runs pass them in lockstep with the ceiling).

Intervention definitions. Let s_i be the singular values of a matrix. The three rank estimators are $\text{PR} = (\sum_i s_i)^2 / \sum_i s_i^2$, $\text{SR} = \sum_i s_i^2 / \max_i s_i^2$ and the energy participation ratio $\text{EPR} = (\sum_i s_i^2)^2 / \sum_i s_i^4$; EPR is logged only and never pinned. The Fourier statistic is computed on the p number-token rows of W_E: with $F = \text{rfft}(W_E[1{:}p])$ along the token axis and $P_k = \sum_j |F_{kj}|^2$ for the non-DC bins $k = 1, \ldots, (p-1)/2$, C_5 is the share of $\sum_k P_k$ held by the five largest P_k, re-selected at every step. On the modular side every arm adds $\lambda_t A$ to the cross-entropy loss, with

$$\lambda_t = \beta \, \frac{\|\nabla_M \mathcal{L}_{\text{CE}}\|_F}{\|\nabla_M A\|_F} \qquad (\lambda_t = 0 \text{ if } \|\nabla_M A\|_F \leq 10^{-8}),$$

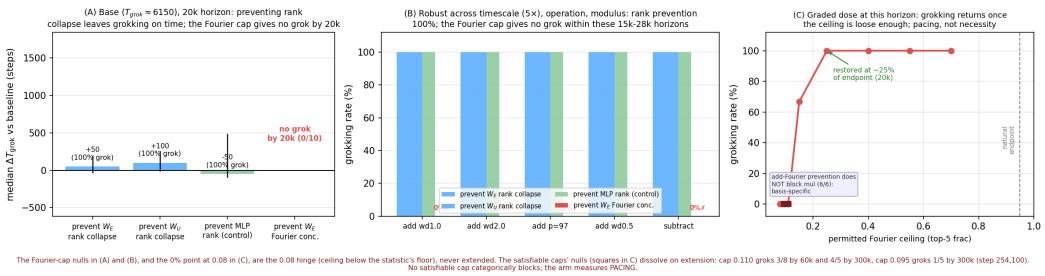


Figure A2: **Causal dissociation at the original 15k–28k horizons (10 seeds + dose; gradient-matched prevention).** (A) Preventing rank collapse of W_E, W_U, or a large non-readout matrix leaves grokking on time (ΔT_{grok} within $\sim$2%, 100% grok); preventing the embedding's Fourier concentration yields no grok within this horizon (0/10). That arm is the unsatisfiable 0.08 hinge, never extended; the satisfiable caps' at-horizon nulls are *horizon-censored, not a categorical block*: at 60k the 0.110 ceiling groks 3/8 (and 4/5 by 300k), and at 300k even 0.095 groks 1/5 (Sec. 4). (B) Holds across the weight-decay timescale span (5× in these 6-seed medians; 3.5× in 8-seed means), subtraction, and p=97. (C) Dose-response at this horizon: on time by ceiling 0.25 (6/6; no grok-by-horizon near initialisation), the same prevention inert on multiplication. Read jointly with the horizon extensions, this is a *graded, basis-specific pacing* threshold $c^*(H)$, not removal of the installed solution and not an all-or-nothing necessity.

recomputed at every step and treated as a constant, where M is the constrained parameter (all of W_E, including the "=" row, or W_U). The floor arms use the squared hinge $A = \text{relu}(0.9\,\text{stat}_0 - \text{stat})^2$, with $\text{stat} \in \{\text{PR}, \text{SR}\}$ evaluated on $W_E[1{:}p]$ or on W_U and stat_0 its value at initialisation; the joint arm sums the PR and SR hinges. The Fourier arm uses $A = \text{relu}(C_5 - c)^2$ with ceiling c. CTRL-mag uses $A = \langle \text{vec}\, W_E, u \rangle$ with u a unit Gaussian direction drawn once per seed, so its push has norm exactly $\beta \|\nabla_{W_E} \mathcal{L}_{\text{CE}}\|_F$. We use $\beta = 1$, except $\beta = 3$ for the stable-rank floor, the joint floor and their random-push control (App. J). The in-context pin is matched to the weight-decay force instead, because the cross-entropy gradient is negligible there after memorisation:

$$\lambda_t = \beta\, \lambda_{\text{wd}}\, \frac{\|W_E\|_F}{\|\nabla_{W_E} A\|_F} \qquad (\lambda_t = 0 \text{ if } \|\nabla_{W_E} A\|_F \le 10^{-12}), \qquad \beta = 10,$$

with A the sum of the PR and SR hinges at 0.9× init on the whole of W_E. The same-object cross-control runs this coefficient rule and this hinge on modular arithmetic, with the hinge evaluated on $W_E[1{:}p]$ and $\|W_E\|_F$ taken over the full matrix; a self-check asserts that its hinge equals the modular joint hinge exactly.

Hinge satisfiability disclosure, and the horizon corrections. The original blocked arm's ceiling (0.08) lies *below* the statistic's uniform floor ($5/56 \approx 0.089$ for the top-5 share over 56 frequencies) and below its initialisation value (≈0.104), so that hinge could never switch off: it acted as a perpetual flattening force, not a capacity-preserving cap (the logged spectra confirm those runs sat at ≈0.0895–0.0905, the floor). A first corrective re-run at *satisfiable* ceilings (0.110 and 0.095) produced 0/6 each at its 20k horizon, and an earlier draft of this paper read that as "holding the concentration at its initialisation level suffices to prevent grokking." That reading was false, and our own re-runs falsified it: re-running the satisfiable ceilings at adequate horizons, 0.110 groks 3/8 by 60k (t_{grok} 34,700–57,500) and 4/5 by 300k, and 0.095 groks 1/5 by 300k (step 254,100). This is the same horizon-censoring failure mode, one level up. Scored by *achieved* concentration (post-hoc, labelled), the dose is monotone with the transition between achieved 0.099 and 0.110, i.e. 0.95–1.06× the initialisation value and comfortably above the uniform floor:

Table 1: Grokking by 60k against the achieved top-5 concentration (post-hoc scoring).

achieved top-5	0.0951	0.0992	0.110	0.130	0.150	0.175	0.200	0.225
grok (60k)	0/8	0/8	3/8	6/8	8/8	8/8	8/8	8/8

The pooled censoring-aware fit runs over the 56 projection-cap cells whose cap actually held and gives $c^*(60\text{k}) = 0.1179$, profile CI $[0.110, 0.127]$, seed-clustered bootstrap $[0.109, 0.130]$ at

B=10,000; we print it as 0.118. The eight cells at cap 0.08 are excluded because that cap was never achieved in any of them, so they measure something other than the cap they are labelled with; the pacing curve makes the same exclusion. Two sensitivities: adding those eight back gives 0.1180 over 64 cells with an interval identical to three decimals, so the exclusion decides nothing; and the unpooled 48-cell fit gives 0.116, $[0.098, 0.126]$. A censored log-normal accelerated-failure-time model over all 56 deduplicated cells (each censored at its own horizon) strongly prefers curvature in the cap (LR $= 28.0$, $p < 10^{-4}$). The fitted curve is descriptive only, since a left-edge-out refit still overpredicts the 0.11 tally (4.7 expected against 3 observed groks by 60k), so the paper's quantitative content is the tallies and the fitted threshold, never an extrapolation. Two scope notes: the original *hinge* arm (achieved at the uniform floor 0.0895–0.0905) was never horizon-extended and remains a 20k null; and whether T_{grok} diverges at some ceiling above the floor or merely grows is unresolved at our horizons. Three details behind Sec. 4's summary: the median-T_{grok} span from ceiling 0.225 to the single 0.095 grok is $\gtrsim$25×, a lower bound under censoring; scored by *achieved* concentration (post-hoc, labelled) the transition sits above the uniform floor $5/56 \approx 0.089$; and the censored cells are moving, not arrested (six of seven still climbing at 60k), while untouched output-side witnesses track generalisation rather than the cap.

C Additional figures and statistical detail

The base cohort. All 48 runs (6 conditions × 8 seeds) grok: training accuracy saturates by step ∼150 while validation rises thousands of steps later, a gap ratio of 24–78× at the base setting, where T_{grok} itself ranges 4.1k–15.4k (≈3.7× across seeds at fixed hyperparameters). That spread is what makes the timing question answerable at all.

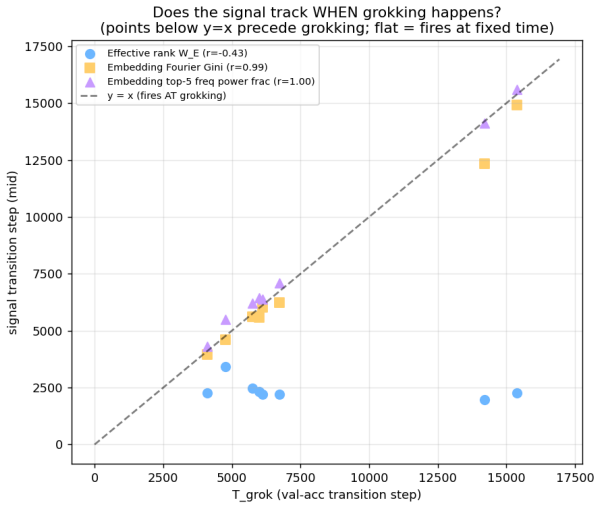


Figure A3: W_E **fires at a near-fixed step; the concentration signals track** T_{grok}**.** Each point is one of 8 seeds; x is the grokking step T_{grok}, y is the signal's transition step. Embedding effective rank (blue) is a flat band near 2.4k regardless of T_{grok}: it fires at a near-fixed early time and so is uncorrelated with timing ($r = -0.43$). Fourier-concentration signals (orange/purple) lie along $y = x$, tracking grokking timing ($r \approx 0.99$).

We avoid the regression slope β as a separate statistic since $\beta = r \cdot \text{std}_{\text{trans}}/\text{std}_{T_{\text{grok}}}$ is algebraically dependent on r, and we do not rely on the percentile bootstrap, which at n=8 can exclude 0 for W_E (single-seed leverage) when the permutation test does not. Across the six conditions (the five additive plus multiplication, whose W_E sits at the $+0.37$ endpoint) W_E's correlation flips sign (-0.43 to $+0.37$), all permutation-non-significant. Within the additive family this is uniform: W_E has the largest lead in all 40 runs with timing correlation inside the permutation floor everywhere, while Fourier-Gini and W_U track; multiplication breaks the additive-Fourier meter but the basis-free W_U rank still tracks (r=1.00).

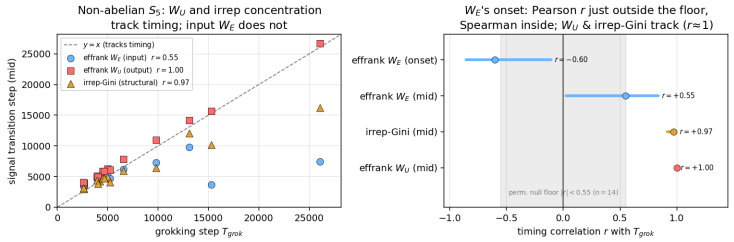


Figure A4: **The dissociation transfers to a non-abelian group (S_5, $|G|$=120, 14 grokking seeds).** Left: transition step vs. T_{grok}; the unembedding effective rank (W_U) and the *irrep*-concentration (the non-Fourier structural meter) track timing (r=1.00, 0.97), while the input embedding W_E does not. Right: timing correlations against the permutation null floor. W_E's onset is blind on the rank-based Spearman statistic; the *Pearson* value plotted here (-0.60) nominally crosses the drawn floor, which is the disclosed discrepancy of App. N, and both statistics are reported there rather than selected. W_U and irrep-Gini pin at $r \approx 1$. The structural timing-tracker is the irrep concentration, not Fourier. (At n=14; the powered n=30 replication shows W_E's *completion* tracks while its onset remains blind; see text.)

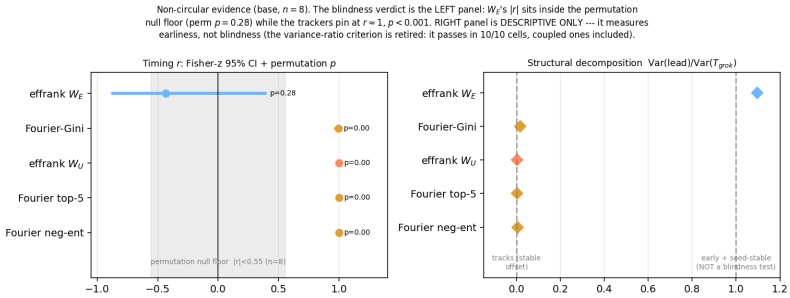


Figure A5: **Timing correlation and variance decomposition (n=8, base).** Left: timing correlation r with Fisher-z 95% CIs and permutation p; W_E's interval brackets 0 and lies inside the permutation null floor (grey), while the timing trackers pin at $r \approx 1$ with $p < 10^{-3}$. Right: the structural decomposition $\text{Var}(\text{lead})/\text{Var}(T_{\text{grok}})$, reported *descriptively*. Near 1 means the onset is early and seed-stable, which blindness implies but which does not imply blindness (the blindness verdict is the left panel's permutation floor); near 0 means the transition co-varies with T_{grok} (tracks).

Pooled estimands across conditions (pre-registered random-effects meta-analysis). Fisher-z with DerSimonian–Laird pooling of the per-condition Spearman correlations over the 6-condition $\times$ 8-seed modular set (Fourier-Gini over the 5 additive conditions): W_E *mid* pooled $r = -0.03$ $[-0.44, +0.40]$ (per-condition -0.71 to $+0.57$; the sign flips across conditions, which is itself the blindness signature: no consistent timing information exists to pool); W_E *onset* pooled $r = -0.42$ $[-0.75, +0.08]$, which brackets zero but fails the strict pre-registered ± 0.5 width bound, leaning negative (per-condition -0.86 to $+0.45$; the -0.86 extreme is the base condition at n=8, which the n=130 replication of that same condition shrinks to $+0.12$ with TOST equivalence within ± 0.3, so precision on onset-blindness comes from n=130, not from pooling 6×8); W_U mid pooled $r = 1.00$ and Fourier-Gini mid 0.999 $[0.92, 1.00]$ (per-condition 0.93–1.00; Fisher-z saturates as $r \to 1$, so the boundary CIs are degenerate and the per-condition range is the informative summary).

Independent validation of the irrep meter. Beyond Plancherel and the character-table battery, we rebuilt the spectral decomposition through a *different* construction: the isotypic projectors of the regular representation, $P_\rho = (d_\rho/|G|) \sum_s \overline{\chi_\rho(s)}\, L_s$, assembled from left-translation permutation matrices. On S_3, S_4, A_4, D_6, and D_{60} the projector algebra is exact (idempotent, mutually orthogonal, summing to the identity) and $\|P_\rho W\|_F^2$ matches the character meter's $\text{power}_\rho(W)$ to relative error $\leq 3 \times 10^{-15}$ on random W; hardcoded explicit S_3 standard-irrep matrices place 100% of the aligned W's power in the standard irrep under both code paths.

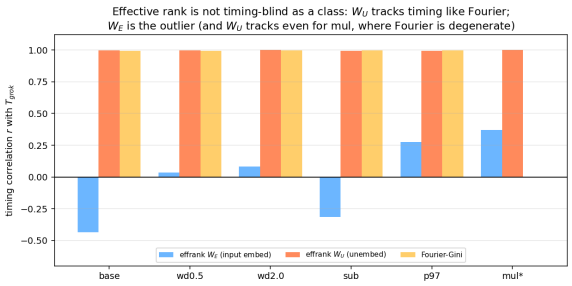


Figure A6: **Generality (8 seeds/condition).** Timing correlation with T_{grok} for the input embedding W_E (blue), the unembedding W_U (red), and Fourier-Gini (yellow). W_U tracks timing ($r \approx 1$) in *every* condition, including `mul`, where the additive-Fourier signal is degenerate, while W_E is timing-blind throughout. The dividing line is W_E, not effective rank.

Table 2: Every tested negative removes exactly one precondition of the lead–timing estimand.

Missing precondition	Task	Outcome
retained delayed-gen. gap	sparse parity (n=40, k=3)	never groks in our 13 runs (memorise-only / immediate / transient); lead undefined
non-empty grok window	group-matching induction	N_{gen}≈41k > N_{mem}≈25–33k: memorises below, generalises *immediately* above, no delayed regime exists at this model size (contrast the non-empty $[6144, 8192]$ of BICL, which groks; Sec. 6)
low-rank structured solution	Omnigrok-MNIST	effranks decouple; weight norm is the order parameter
input/output role asymmetry	teacher–student MLP	no W_E/W_U split to dissociate; no clean grok + tunable rank jointly
correct symmetry basis	mul, a^2+b^2 (additive DFT)	meter degenerate; W_U still tracks for mul (r=1.00) but *degrades* for a^2+b^2 (r=0.63, n=6)
powered timing variance	Omnigrok-MNIST (seed CV <0.04)	cross-seed correlation test unpowered
grokking at all (task size)	S_4 ($\lvert G\rvert$=24, 576 pairs)	no train fraction generalises by 30k (one seed each)

D BOUNDARY DETAILS

S_4 composition ($|G|$=24, 576 enumerable pairs). With one seed at each training fraction in $\{0.5, 0.6, 0.7, 0.8\}$ and a 30k-step horizon, every run memorises by step 100 and none generalises (final validation 0.04–0.40, still rising at 30k at fraction 0.7), so T_{grok} and the lead are undefined. We read this as a probable task-size boundary of this fixed-width transformer (over-parameterised for so small a group), not a dissociation failure; with one seed per fraction and one horizon, a longer run could still show delayed generalisation.

Quadratic family ($a^2+b^2 \bmod p$, 6 seeds, train fraction 0.20). Groks (val >0.99) but the additive-DFT spectrum never concentrates (final Gini 0.065); W_E loses its largest-lead property (0/6), its midpoint correlation is +0.70 (non-significant only against the small-n floor), and W_U's tracking degrades to 0.63.

Sparse parity (n=40, k=3, same tokenisation, 13 runs). Never groks: small data memorises without generalising (val ≈0.5); enough data generalises immediately (train and val rise together); the boundary regime (weight decay 2.0) generalises transiently then collapses to chance. No separated validation jump exists for a lead to precede.

Teacher–student MLP (tunable-rank labels, ∼44 probe configs). A two-layer MLP does not give clean grokking and rank-controllable input compression simultaneously: the label family that groks cannot dial solution rank, and the rank-clean family generalises immediately. The architecture lacks the W_E/W_U role asymmetry the dissociation is defined over.

Omnigrok-MNIST (6 seeds). Groks cleanly (train 1.0, test ∼0.92) but by T_{grok} the weight norm has completed ≈0.47 of its change (co-timing) while every weight- and representation-effective-rank is ≈0 or wrong-signed (they rise through the transition and compress only post-grok, weight-decay-driven). Cross-seed timing variance is also tiny (std ≈4% of the mean; a dedicated probe

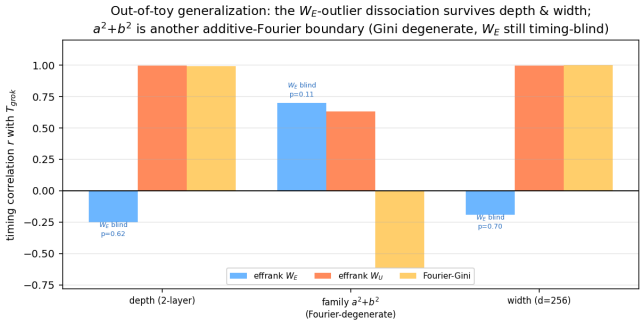


Figure A7: **Architecture generality and the quadratic boundary (6 seeds/condition).** The W_E-outlier dissociation survives depth (two layers) and width (d=256); for a^2+b^2 the additive-Fourier signal is degenerate and the dissociation does not cleanly survive. The in-image "W_E blind" labels mark non-rejections against the small-n permutation floor at n=6, not the n=130-grade equivalence of Sec. 3.

of three edge recipes found seed-only CV(T_{grok})<0.04 everywhere), so the correlation estimand is unpowered at the task level.

E MEASUREMENT-LAYER ROBUSTNESS

Detector sensitivity. Across a grid of detector hyperparameters (plateau window $\in \{3, 5, 8\}$, smoothing $\in \{3, 5, 9\}$, crossing-persistence $\in \{1, 3\}$; 18 configurations) on the n=130 set, the verdicts are unchanged in every cell: W_E onset/mid Spearman $\in [-0.15, +0.19]$ (all inside the $\pm$0.3 SESOI) while W_U and Fourier-Gini midpoints stay $\geq$ 0.992. A single "lead time" is also threshold-fragile (under a late 90% threshold every signal lags), which is why verdicts are stated per detector point rather than as one lead number.

Completion fractions. $\varphi = \mathbb{E}_s[\tau(s)/T_{\text{grok}}(s)]$, the mean of per-seed ratios, not $\mathbb{E}[\tau]/\mathbb{E}[T_{\text{grok}}]$: right-skewed T_{grok} makes the first exceed the second (W_E's midpoint: 0.43 vs. 0.30), and mixing the two would misorder detector points.

Effective-rank estimator. Re-logging 8 base seeds with three *measurement* estimators (participation ratio, Roy–Vetterli exponential-entropy rank (Roy & Vetterli, 2007), stable rank; not the causal triple of App. J, which replaces the second with the energy participation ratio). At this n three W_E detector points reach or cross their permutation floor of 0.71, all with negative sign (caveat (i) below; the participation-ratio midpoint sits at the floor, -0.71), the n=130 TOST equivalence settles the onset, and W_U's midpoint tracks under all three (the variance decomposition, $\text{Var}(\text{lead})/\text{Var}(T_{\text{grok}}) = 0.94$–$1.15$ at onset and mid, is reported descriptively). Two caveats: (i) at n=8 the participation-ratio and Roy–Vetterli W_E *onsets* cross the floor with *negative* rank correlations (-0.86/-0.76, permutation p=0.011/0.036); these onsets span only 32 and 64 steps against a T_{grok} range of 4.1k–15.4k, so this is sampling noise of a near-degenerate window, and the n=130 powered set settles it (Spearman $+0.12$, TOST-equivalent within $\pm$0.3); (ii) tracking *strength* is estimator-dependent: stable rank, dominated by the top singular value, tracks at 0.79 rather than 1.00.

Functional witness. Evaluating the network with W_E's numeric block projected onto its top-5 DFT components (a behavioural analogue of restricted loss for this one-layer model): the restricted validation accuracy's midpoint tracks T_{grok} at r=1.000 ($\text{Var}(\text{lead})/\text{Var}(T_{\text{grok}}) = 0.00$, completing at 0.96 T_{grok}; onset 0.98 at 0.60 T_{grok}), while the *excluded* remainder's accuracy collapses at a fixed early step (onset $r = -0.03$, Var ratio 1.00). The dissociation is a fact about functional progress toward the key-frequency solution, not an artifact of weight statistics.

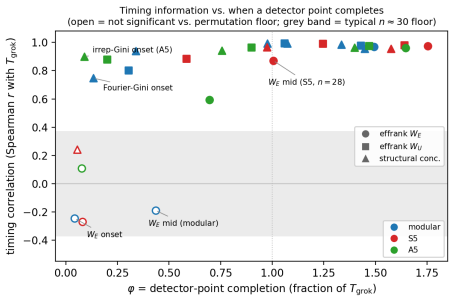


Figure A8: **Timing information attaches to the clock a detector point inherits, not to its earliness.** Each marker: one signal $\times$ detector point (onset/mid/late) on one powered task family; y = Spearman correlation with T_{grok}, x = mean completion fraction. Open markers: not significant vs. that point's permutation floor. Blindness is read off the floor, not off the variance ratio, which is $\approx$1 for every early point here, blind or coupled, and is therefore retired as a criterion (Sec. 6). Note the blind W_E midpoint at φ=0.43 and the tracking concentration onsets at $\varphi \approx 0.1$.

F THE PRE-REGISTERED r-VS-φ TEST BEHIND FIG. A8

The framing result of Sec. 2, stated and proved.

Observation 1 (variance separation) *Write $T_{\text{grok}}(s) = T_0 + D(s)$ for the fast-phase completion time T_0 and the seed-variable slow-phase duration $D(s)$, and let $\tau(s)$ be a detector point. Suppose (A1) $\tau(s) \le T_0$ almost surely, and (A2) τ is determined by the fast-phase dynamics, so that τ is independent of D. Then (i) $\text{corr}(\tau, T_{\text{grok}}) = 0$ exactly; (ii) the point's lead $T_{\text{grok}} - \tau$ is large, with mean exceeding $\mathbb{E}[D]$, while $\text{Var}(T_{\text{grok}} - \tau)/\text{Var}(T_{\text{grok}}) \approx 1$; and (iii) no predictor of T_{grok} from τ improves on predicting the mean.*

Proof. (i) $\text{cov}(\tau, T_{\text{grok}}) = \text{cov}(\tau, T_0 + D) = \text{cov}(\tau, D) = 0$ by (A2), and T_0 is constant across seeds by construction. (ii) $\mathbb{E}[T_{\text{grok}} - \tau] \ge \mathbb{E}[T_0 - \tau] + \mathbb{E}[D] \ge \mathbb{E}[D]$ by (A1), and $\text{Var}(T_{\text{grok}} - \tau) = \text{Var}(D) + \text{Var}(\tau) \ge \text{Var}(D) = \text{Var}(T_{\text{grok}})$, with equality as $\text{Var}(\tau) \to 0$. (iii) follows from (i) for any linear predictor and from independence for any measurable one. □

With detector noise. If $\tau = \tau_0 + \varepsilon$ with ε independent of D, every conclusion survives verbatim: added independent noise cannot create covariance with D, it only inflates $\text{Var}(\tau)$ and pushes the variance ratio further above 1. This is why the variance ratio is useless as a blindness criterion and why we read blindness off permutation floors instead.

What Fig. A8 shows. The blind class is exactly the W_E rank onsets, the modular W_E *midpoint* (φ=0.43, blind although far from early), and the S_5 irrep onset, while concentration onsets at $\varphi \approx 0.1$ already track on two of three task families.

We pre-registered (protocol locked before analysis) the stronger graded prediction that the completion fraction φ alone monotonically orders timing correlation across all signal $\times$ detector points, with blindness for $\varphi \le 0.2$ and tracking for $\varphi \ge 0.8$. Outcome on 30 points across modular (n=30), S_5 (n=28), and A_5 (n=15): the tracker-class prediction holds (17/17 points with $\varphi \ge 0.8$ track at $r \ge 0.87$), but monotonicity fails (rank-correlation $0.73 < 0.8$) and two $\varphi \le 0.2$ points already track: the modular Fourier-Gini onset (φ=0.13, r=0.75) and the A_5 irrep-Gini onset (φ=0.09, r=0.90), both jackknife-stable. We therefore do not claim the monotone-φ law; the surviving statement, reported in Sec. 2, is that blindness attaches to which clock a signal $\times$ detector point inherits, not to earliness. The early-tracking concentration onsets are an exploratory observation of independent practical interest (an early point that both leads by $\sim$90% of T_{grok} and carries timing information), pending online evaluation.

Framing-lemma detail (from Sec. 2). A detector-noise version needs no independence assumption: if the measured crossing is $\hat{\tau} = \tau + \varepsilon$ with detector noise $\varepsilon \perp (\tau, T_{\text{grok}})$ of scale σ_ε, then $|\text{corr}(\hat{\tau}, T_{\text{grok}})| \le \sigma_\tau / \sqrt{\sigma_\tau^2 + \sigma_\varepsilon^2}$: a point whose true crossing varies by less than the detector's

resolution is observably near-uncorrelated whatever the fast phase knows. Any dissociation built on this is arithmetic rather than discovery: a near-zero-variance transition cannot correlate with a grokking step that varies $3.7\times$. Timing-blindness is therefore a *structural* property of a signal class, not a sampling accident, and the empirical content is which detector points inherit which clock, which the lemma cannot decide.

G DIHEDRAL D_{60}: THE PRE-REGISTERED PREDICTION TEST

Provenance. Protocol, decision rules, and all analysis code were committed *before* the recipe probe ran; predictions, contingencies, interim rules, and claim boundaries were fixed in that commit. Two pre-data amendments are on record: A1 (after a 300-step pipeline smoke test at train fraction 0.7, run to validate the dataset/meter integration, showed near-immediate generalisation there, the probe grid was widened to $\{0.3, \ldots, 0.8\}$ and the recipe prediction sharpened from "groks at 0.7 like S_5/A_5" to "groks at some $\mathrm{tf} \in [0.3, 0.6]$, *below* S_5/A_5"; the smoke test also previewed the meters' qualitative behaviour over 300 steps and its trace is archived); and A2 (a 36-finding adversarial design review, all fixes adopted pre-launch: amplitude-validity gates, one-sided exact permutation tests, a verdict ladder that cannot label an unpassable test "refuted", positive-control gating for null-acceptance claims, and the structured-relabel attribution plan). The character-based D_n meter is generated from closed-form characters and gated by orthogonality, permutation-character, regular representation, and Plancherel checks plus a functional-alignment check (each irrep's matrix-entry-aligned W must place $\geq$99.9% of its power in that irrep), which provably catches rotation-class column misassignments the orthogonality battery cannot see.

Recipe probe (single seed per cell, raw first-crossing clock). tf 0.3: groks, gap 140.5 (T_{mem}=200, T_{grok}=28,100); 0.4: gap 27.5; 0.5: gap 24.0; 0.6: gap 4.3; 0.7/0.8: generalise immediately (final val 1.0, gap 2.3/2.6), so the window's upper end closes by immediate generalisation rather than by memorisation failure. The pre-registered recipe prediction ($\mathrm{tf} \in [0.3, 0.6]$ at weight decay 1.0) held; the binding rule selected tf 0.3.

Stage 2. 17 runs (seeds 0–14 plus pre-registered confirmation seeds 1000/1001, all at tf 0.3, 90k steps): $17/17$ grok, 0 censored, 0 excluded; $T_{\mathrm{mem}} \in [150, 175]$; $T_{\mathrm{grok}} \in [4.8\mathrm{k}, 53.6\mathrm{k}]$ ($11.2\times$; CV 0.50). Primary clock = raw first crossings (val $>$0.9 / train $>$0.99), pre-registered for D_{60} because its index-2 structure invites staged generalisation; indeed the reflection (coset) bit is learned *first* in every seed (reflection-part accuracy crosses 0.9 before rotation-part accuracy in $17/17$ seeds, at a median $0.56\,T_{\mathrm{grok}}$; in $2/17$ seeds within $\sim$200 steps), and the rotation part jumps only at grok. All tracking results are robust to the mid-rise clock (irrep-Gini 0.78, W_U 0.96).

Verdicts (pre-registered ladder). P1 (confirmatory, one-sided exact/MC permutation): irrep-Gini completion tracks, r=0.74, p=7×10^{-4}, transitioning at $0.55\,T_{\mathrm{grok}}$, $\mathrm{Var(lead)}/\mathrm{Var}(T_{\mathrm{grok}})$=0.51; the pre-registered saturation and coset-confound diagnostics did not fire (median one-dimensional-irrep power share at the crossing 0.10; the two-dimensional-irrep-only meter tracks at r=0.93). P3: W_U rank tracks, r=0.94. P4: W_E onset at $0.007\,T_{\mathrm{grok}}$, blind in the pre-registered sense of *no detectable tracking at this n's power* ($r = -0.46$ against an $|r|$ floor of 0.48, a narrow margin; $\mathrm{Var(lead)}/\mathrm{Var}(T_{\mathrm{grok}})$=1.00). Exploratory E1: W_E completion $r = -0.27$ (not detectably tracking; moderate tracking not excluded at n=17). P2 (uniform-relabel control, positive-control-gated): *fails* the basis-specificity bar. The run-time control (median over 5 fixed relabels per seed) rises above the pre-registered 0.10 amplitude floor in only $5/17$ seeds; on those five its completion tracks (r=1.00, exact p=0.0083, at $0.89\,T_{\mathrm{grok}}$; descriptive, since it is below the $n\geq6$ evaluability line), and the paired $\Delta r = r_{\mathrm{irrep}} - r_{\mathrm{relabel}}$ bootstrap CI $[-2.0, 0.0]$ does not exclude zero. The pre-registered "rank-echo" escape reading (the relabeled meter merely echoing W_E's rank collapse) is *rejected*: the control's tracking survives partialling out the W_E-completion transition (partial r=1.00). A single-relabel snapshot recomputation gives the same picture ($14/17$ below the floor; the draw-count difference explains 12 vs 14).

Structured-relabel attribution (pre-registered as exploratory; no confirmatory weight). We recompute the meter from weight snapshots under relabel families that preserve increasing amounts of group structure. Uniform relabel has rise amplitude 0.09 (median) and is not evaluable as a tracker; within-coset relabel (preserves only the rotation/reflection partition) has amplitude 0.19

and tracks at r=0.69; coset-scramble (preserves the partition and cyclic order via an automorphism $k \mapsto uk{+}c$) has amplitude 0.50 and tracks at r=0.89; the true labelling has amplitude $\sim$0.55 and r=0.74 (0.93 on two-dimensional irreps only). Tracking quality and rise amplitude are ordered by preserved structure, which we read (exploratorily) as a graded correlational counterpart of the causal redundancy-scaling of Sec. 6: modular/abelian at the basis-generic endpoint, D_{60} intermediate, S_5/A_5 basis-specific.

Exploratory negative: predictor transfer. The modular-calibrated 40%-rule with the fixed scale ratio, applied to D_{60}'s irrep-Gini as it was to S_5's, gives 40.3% median timing error vs. 32.0% predict-the-mean (D_{60}-self-calibrated ceiling 29.7%), so the calibrated point-predictor does not transfer here (vs. 17% on S_5), even though the meter's completion correlates with timing. Transferability of the calibrated rule is group-dependent; we report this as a limitation of the predictor, not of the dissociation.

H PROJECTION-BASED PREVENTION AND CHECKPOINT GRAFTING

Provenance. Protocol, arms, decision rules, and analysis code committed before any run; projection and splice operators unit-tested (the Fourier cap converges exactly to its ceiling while preserving the non-DC energy to $3{\times}10^{-5}$ and leaving DC untouched; the random-basis cap leaves the true DFT fraction unchanged; the effective-rank floor is Frobenius-preserving; the surgery splice is an identity to 10^{-6}). One deviation is disclosed: the registered dose expectation for the 0.25 ceiling was a drafting error that contradicted our own penalty dose-response (which already restored grokking at 0.25); the projection data match the *penalty* landscape, not the erroneous sentence, and are scored accordingly (protocol D1).

Operators. Each projection is applied at initialisation and after every optimiser step. *Fourier cap:* with $F = \mathrm{rfft}(W_E[1{:}p])$, T the total non-DC power and S the power in the current top-5 bins, if $S/T > c$ the top-5 bins are multiplied by $a = \sqrt{cT/S}$ and the other non-DC bins by $b = \sqrt{(1-c)T/(T-S)}$, with DC untouched, and the result is inverted with irfft. This keeps the non-DC energy and sets the share to exactly c; because the top-5 set can change, it is repeated at most five times. *Rank floor:* with $W = U\,\mathrm{diag}(s)\,V^\top$ and $s(\alpha) = (1-\alpha)s + \alpha\bar{s}$, the smallest α with $\mathrm{PR} \geq 0.9\,\mathrm{PR}_0$ is found by 40 bisection steps and the result is rescaled to the original Frobenius norm; the step is skipped when PR is within 0.05 of the target, and it acts on $W_E[1{:}p]$ or on W_U. *Random-basis cap:* the same rescaling on $C = R^\top W_E[1{:}p]$, with R a fixed orthogonal matrix from the QR factorisation of a Gaussian matrix and row 0 of C in the role of DC, followed by $W_E[1{:}p] \leftarrow RC$. *Graft:* with K the donor's five largest non-DC bins, the recipient's coefficients F_k are replaced by the donor's for $k \in K$ across all 128 columns of $W_E[1{:}p]$. Everything else (the "=" row, the positional embedding, W_U, attention, the MLP and the AdamW state) is the recipient's own, at its step-2000 checkpoint. The donor is the grokked step-20,000 checkpoint of the same seed (self-donor) or of the next seed (cross-seed). The surgery control runs the same splice with the recipient as its own donor, and the random-basis splice makes the same replacement in a fixed random orthogonal basis.

Projection grid (Part A; 6 paired seeds, 20k steps, setup identical to the penalty grid). Projections applied in-place after every optimiser step. Outcomes (grok = final val >0.9 and raw first-crossing gap ≥ 3; on-time = $6/6$ grok and median paired T_{grok} ratio vs. baseline in $[0.8, 1.25]$):

Table 3: Projection grid (Part A), outcomes by arm at the grid's 20k horizon.

arm	grok	median paired ratio	note
baseline	$6/6$	—	T_{grok} 4.6k–16.1k
proj. Fourier cap 0.110	$0/6$	—	no grok by *this grid's* 20k horizon (median achieved top-5 $= 0.1100$, at ceiling); horizon-censored: $3/8$ by 60k, $4/5$ by 300k (App. B)
proj. Fourier cap 0.25	$6/6$	1.239	restored, $\sim$24% slowdown (rule-edge on-time)
proj. Fourier cap 0.40	$6/6$	1.033	on time
proj. rank floor W_E	$6/6$	0.983	on time; groks with top-5 only $\sim$0.58
proj. rank floor W_U	$6/6$	1.015	on time
proj. cap in random basis	$6/6$	1.000	slack, not a basis test: a free W_E's random-basis top-5 share is 0.06–0.07 at step 2k and stays below the cap until after grokking, so T_{grok} equals the baseline's on every seed; true DFT fraction reaches 0.91–0.98

Grafting (Part B; 8 recipients, resume from step 2000, estimation framing). Every arm is compared with the surgery control (the identity splice through the same code path); a plain resume matches the surgery control (median $+100$ steps, CI $[-100, +200]$), so splice mechanics are clean. Median ΔT_{grok} vs. surgery, 10k-resample bootstrap CIs: **self-donor graft** -950 $[-2800, -600]$ ($8/8$ accelerated, per-recipient -500 to -4200, the largest absolute effects on the slowest recipients); cross-seed donor -300 $[-500, -200]$ ($8/8$); random-basis splice -100 $[-400, 0]$. Post-graft validation accuracy is 0.02–0.11 in every arm, so the transplanted structure never generalises by itself. Statistics field note: in the released `p21_stats.json` the `constraint_held` flag for the random-basis arm refers to the (intentionally unconstrained) true-DFT fraction; the random-basis constraint itself is enforced by construction.

I In-context induction circuits: from a shortcut boundary to a generalisation

Task and the two pre-registered designs. Length-16 sequences with a trigger bigram (a, b) at a uniformly random position in $[0, 12]$ and the query token a at the final position; the target is b, with cross-entropy at the final position only. A 2-layer attention-only transformer (d=128, no MLP, no LayerNorm) must form an induction circuit (previous-token head then match-and-copy) to solve held-out sequences. *Design v1* (pre-registered, closed): training targets came from a fixed random table $b{=}\pi(a)$, memorisable in-weights; across 12 recipes ($0/12$) the table wins in $\sim$100 steps and the circuit never forms (validation below chance, induction score $\sim$0). We first read this as a precondition-1 boundary; the diagnosis was that the table (64–128 entries) is *cheaper* than the circuit, so weight decay never prefers the circuit. *Design v2* (pre-registered separately, two changes only): each training pair is drawn independently on a fixed finite train set (so the only in-weights solution is per-example memorisation, whose cost grows with the train-set size N, which is the modular ordering), and a, b, and fillers share one vocabulary with fillers drawn from $V\backslash\{a\}$ and $b{\neq}a$ (killing a singleton-type shortcut). The random trigger position, architecture, and meters are unchanged.

Phase diagram (advance prediction). As predicted, the grok window sits at intermediate N: memorise-only for $N \leq 8192$ (circuit never forms), immediate generalisation for $N{=}16384$ (circuit at $\sim$500 steps, gap $<$3), and grokking at $N{=}12288$ (located by a pre-registered two-step bisection). At the chosen recipe grokking is *seed-fragile*: $\sim$27% of seeds grok cleanly (8 of 35 in the base cohort plus 11 of 35 in the powered extension, i.e. $19/70$; plus one transient), and the weight-decay-0.3 confirmation seeds failed, which triggered the pre-registered fallback to weight decay 1.0 (itself first evidence of the fragility). We read the fragility as expected near a memorise/generalise phase boundary. Reaching $n{=}8$ used a pre-registered seed extension (the strict $n{<}4$ case fell below the original escalation rule, a pre-registration gap patched before the extra runs).

Dissociation verdicts ($n{=}8$, T_{grok} 975–23,550, CV 0.99; raw clock, one-sided exact permutation tests; robust to the mid-rise clock). *Trackers.* The induction-circuit strength completes at $0.96\,T_{\text{grok}}$ (Spearman $r{=}1.00$, $p{<}10^{-4}$, $\text{Var}(\text{lead})/\text{Var}(T_{\text{grok}}){=}0.00$), the pre-registered confirmatory test; the unembedding rank ($1.14\,T_{\text{grok}}$, $r{=}1.00$) and the embedding rank *completion* ($1.05\,T_{\text{grok}}$, $r{=}1.00$) track as well, and effective rank collapses (W_E 93$\to$26, W_U 92$\to$14), so

the low-rank precondition holds here. *Leading signals.* The embedding-rank onset (per-seed ratio to T_{grok}: mean 0.23, median 0.23; absolute spread $3.5\times$ vs. T_{grok}'s $24\times$) and the previous-token-head onset (mean 0.12, median 0.11) are early and seed-stable (variance ratio 0.93 and 0.89, which measures earliness rather than blindness; the ratio is retired as a criterion in Sec. 6), but their ranks *do* correlate with T_{grok} ($r{=}0.95$ and 0.86, above the $n{=}8$ permutation floor 0.71; confirmed at power, Spearman 0.87/0.79 vs. a 0.45 null). The pre-registered blind test was a conjunction of rank-blindness and a high variance ratio, and it fails on the rank conjunct. On the induction task the onset is therefore compressed-early but rank-ordered, not the fixed-absolute-time, rank-blind onset of modular arithmetic. We do not read blindness into the surviving variance conjunct: off the algebraic island the strict test fails. What generalises off the algebraic island is the *tracking* half, in which structural completion (circuit strength and output-/input-side rank completion) tracks the grokking step, together with the low-rank precondition; the controlled v1-vs-v2 comparison shows precondition-1 is governed by the relative cost of the memorise vs. in-context solutions, connecting the framework to the in-weights/in-context learning-mode competition (Chan et al., 2022; Singh et al., 2023; Reddy, 2024).

Adjudicating the strict failure (pre-registered re-analyses of the archived runs). Three artifact accounts of the rank coupling were tested and excluded on the archived 35-run cohort, each with its decision rule fixed before the outputs were seen. *Detector choice:* 15 onset definitions (relative mid-crossing, six absolute effective-rank thresholds, four drop-from-init thresholds, four previous-token-head thresholds) give Spearman 0.95–1.00 throughout. *Clock origin:* re-anchoring every detector to the end of memorisation (delay $\mathrm{onset}{-}T_{\mathrm{mem}}$, ratio $\mathrm{onset}/T_{\mathrm{mem}}$, and partial correlation given T_{mem}) changes nothing ($0/3$ primary detectors pass), because T_{mem} is itself near seed-invariant (300–325 steps, $1.1\times$ spread, against a $24\times$ T_{grok} spread): memorisation is the fixed-time phase even off algebra, and what varies is the plateau escape. *Selection:* a 10^4-cohort null simulation of the 8-of-35 inclusion pipeline puts the selection-inflated 95th-percentile null at 0.78 ($P(|r|{\geq}0.95) \approx 0.4\%$). A model comparison (leave-one-out CV in raw steps) then prefers a zero-intercept power law $\mathrm{onset} \approx c\,T_{\mathrm{grok}}^{\gamma}$ with $\gamma{=}0.33$–0.36 over both an affine additive-clock model and pure proportionality on all three primary detectors. This is a predictive preference only: the affine intercept is itself non-zero (jackknife CI $[184, 316]$ steps on drop-from-init), so a small seed-invariant component is not excluded, merely dominated. On the two *absolute* detectors the same onsets fire 2.2–$2.3\times$ later (KS $p{<}10^{-3}$) in the never-grokking runs of the same recipe that clear the inclusion pipeline (25 of them), while the relative-midpoint detector reverses sign there (median 96 vs. 960 steps), a detector-shape artifact on trajectories that never leave the plateau and the reason the absolute detectors carry this reading. Together these say that the onset is race-gated with elasticity $\approx 1/3$, which reproduces both observed statistics at once ($24^{1/3} \approx 3$ compression $\Rightarrow$ variance ratio ≈ 0.93; monotonicity $\Rightarrow$ Spearman ≈ 1). The pre-registered powered extension (seeds 200–234, no early termination per an offline kill-rule audit that found no safe rule; one 60k-step extension arm) pools to $n{=}19$ grokkers with T_{grok} spanning $46\times$ and gives the statistics quoted in Sec. 6: Spearman 0.87/0.79 vs. a 0.45 null (**rank coupling confirmed at power**: the strict test fails on the powered cohort as it did on the small one), $\gamma = 0.28$ $[0.19, 0.36]$ on the drop-from-init detector and 0.29 $[0.20, 0.39]$ on the relative midpoint, and a variance ratio of 0.96 $[0.93, 0.98]$ / 0.98 $[0.97, 0.99]$ that clears the pre-registered 0.6 bar (modular matched-n reference ≥ 0.995). That pass is not used: the conjunct is retired as non-discriminating (App. N). There is therefore no certified blindness on this task, only a well-powered negative; the off-island blind positive (BICL) rests on the surviving criteria (App. N). The confirmatory *trackers* were re-run on the same pooled $n{=}19$ so that both halves are certified at equal power (circuit-strength completion Spearman 0.998, unembedding-rank completion 0.998, embedding-rank completion 0.997; permutation floor 0.46; variance ratios 0.03–0.05). The cross-task elasticity spectrum and the injection-exclusion test appear in App. N.

The causal gate on this task (the in-context side's second circuit). The pin gate of Sec. 4 runs at pool size $N{=}15360$, adopted by a registered escalation after the gate *failed on the baseline side* at $N{=}12288$ (baseline $3/8$) and a probe at 14336 passed only one of two bars. At 15360: pin $0/8$, baseline $8/8$, push $5/5$, floors held $8/8$, exact paired $p = 0.0039$ one-sided against the registered direction, the arms being the same eight seeds (the registered Fisher statistic, 1.6×10^{-4}, assumed sixteen independent runs). The N-sweep behind the escalation carries its own finding: raising N converts stuck seeds into immediate generalisers, never into delayed grokkers (the delayed-grokking count is exactly 3 at every N tried, $12288/14336/15360$), so the v2 grok window has no interior

where most seeds grok with a delay, and the registered wording for the gate is *"the pin prevents the in-context circuit from forming (no generalisation at all)"*, never "blocks grokking." A standing disclosure travels with the pin cells, in the wording a registered re-measurement fixed *at the adopted* $N{=}15360$: every pinned seed fully fitted the training set (t_{mem} 575–1400) and then spent a median 16% of the remaining trajectory at full training accuracy (minimum aggregate 8.8k steps), in repeated stretches of up to 3.8k steps, ending at 0.40–0.90; $0/8$ ended fully fitted. At the neighbouring $N{=}12288$ the same measurement reads median 33%, stretches up to 6.3k, $1/8$ ending fitted. Raising N to rescue the baseline arm therefore cost memorisation stability on the pin arm, and "the pin blocks generalisation in a fully-fitted network" is a less clean sentence at 15360 than at 12288. So "stops the circuit, not memorisation" is available on this task only in the restricted memorisation-intact-window form (BICL's fully-memorising pin carries the unrestricted form).

The f^*_{v2} dose and the transfer verdict. The registered dose (9 floors $\times$ 10 seeds, $90/90$ held; densified by registered amendments that added the $0.75\times$, $0.8\times$, $0.85\times$ and $0.95\times$ floors) tallies $10/10$ ($0.25\times$), $8/10$ ($0.4\times$; a meter artifact at this floor inflates the fitted slope, which we therefore never quote), $10/10$ ($0.55\times$, $0.7\times$, $0.75\times$), $9/10$ ($0.8\times$), $6/10$ ($0.85\times$), $0/10$ ($0.9\times$, $0.95\times$). The same censoring-aware latent-threshold model as BICL's gives $f^*_{\mathrm{v2}} = 0.845$, profile CI $[0.804, 0.895]$, cluster-bootstrap $[0.811, 0.868]$; the registered ≤ 0.07 half-width bar is met on both methods. An apparent step in the pre-densification grid was a grid artifact (the filled-in dose is graded $10/9/6/0$), and a "v2 is a step" characterisation we had briefly recorded is withdrawn. Both registered transfer predictions from BICL's f^*, init-normalised 0.600 and endpoint-normalised 0.583, are excluded under both interval methods: *the threshold transfers as an existence claim and fails as a quantitative one*.

J Estimator-robustness of the causal rank result (C1)

Pre-registered before any run, with the advance predictions and the two logged deviations, in `protocol_c1_rank_estimator.md`. Recipe, seeds and matched-gradient machinery are those of the causal grid (Sec. 4); nothing was re-tuned. All three rank estimators are logged on *every* arm at every step, so none can be chosen after the fact: $\mathrm{PR} = (\sum_i s_i)^2 / \sum_i s_i^2$ (the statistic used throughout the paper; tail-dominated), the stable rank $\sum_i s_i^2 / s_{\max}^2$ (head-sensitive), and the energy participation ratio $(\sum_i s_i^2)^2 / \sum_i s_i^4$. Floors are the paper's convention, $0.9\times$ the value at initialisation. Medians over 6 seeds; W_E statistics; ΔT_{grok} is the median of per-seed differences from BASELINE.

Table 4: Rank-estimator robustness: medians over 6 seeds, W_E statistics; bold marks a floored statistic.

arm	bound?	grok	ΔT_{grok}	PR	stable	energy PR	top-5
BASELINE (unconstrained)	—	6/6	—	21.7	6.90	9.4	0.946
PREV-ERANK (PR floor ≥ 77.4)	yes	6/6	-50	**76.9**	11.5	23.6	0.583
PREV-SRANK $\beta{=}3$ (stable floor ≥ 27.9)	yes	6/6	$+550$	55.2	**27.8**	35.5	0.343
PREV-BOTH $\beta{=}3$ (*both* floors)	yes	6/6	$+50$	**77.1**	**27.8**	43.0	0.356
CTRL-GRAD $\beta{=}3$ (random dir., matched)	—	6/6	$+2850$	82.1	21.8	49.4	0.327

(initialisation: PR 86.1, stable rank 31.0, top-5 0.101. All arms reach train accuracy 1.000; the constrained grokking runs pass the independent staggered-emergence gate with 30–56 key frequencies, versus the bar of 2.)

Deviation D1 (logged). At the pre-registered matched strength $\beta^\star{=}1$ the stable-rank hinge is *active but loses* to the cross-entropy gradient: PREV-SRANK ends at stable rank 10.5 against its 27.9 floor (while the PR-floor arms bind exactly). That arm groks $6/6$, but its intervention never happened, so we treat the pass as **void** rather than as confirmation, as a failed control elsewhere in this project voided a 931k-step run. The $\beta{=}3$ calibration in the table is the arm that implements the test. Holding the stable rank up requires suppressing the largest singular direction, which is what the loss is building; the tail-weighted floor asks for nothing so expensive.

Deviation D2 (the disclosure). The original prevention arm therefore prevents the collapse *only in the tail-weighted sense*: under it the head-sensitive stable rank still completes $\approx 80\%$ of its natural collapse. The claim survives because the head-sensitive and joint floors, once they bind, still leave grokking intact.

Why the constrained runs stop where they do. A floor on the stable rank at $0.9\times$ init requires $s_{\max}^2 \leq 1/27.9 \approx 0.036$ of the embedding's energy, which permits a top-5 Fourier fraction of at most ≈ 0.36, far above the measured pacing threshold $c^*(60\text{k}) = 0.118$ $[0.110, 0.127]$ (App. B), so the geometry never forces the constrained runs into the strongly-paced regime. The prediction that the head-sensitive floor cannot block grokking was derived and registered before the runs (originally against the since-superseded 0.25 bracket; the corrected threshold only widens its margin), and the constrained arms land at top-5 0.33–0.36: at the ceiling the geometry allows, and no further. The concentration a network uses is solution-family-dependent rather than a universal constant: a capped network reaches generalisation by spreading over *more* frequencies (56 key frequencies under the 0.25 cap, against 4 in the free-running baseline), which is why the free network runs its top-5 to ≈ 0.635 before reaching the validation accuracy a capped one reaches at 0.25.

Two numerical details behind Sec. 4. The arms that floor W_E's stable rank grok with top-5 fractions 0.34–0.36 (arm medians; per-seed 0.30–0.36), the ceiling their floors geometrically permit; the arm that floors W_U's instead leaves W_E's concentration unconstrained at 0.91, as it must. In numbers, capped-but-grokking networks do not stay high-rank; at ceilings 0.11–0.20 their final stable rank is 6.3–9.6 by cell median (per-seed 5.0–17.2), the free-running endpoint. Looser ceilings keep more: at 0.225 ($8/8$ grok) the final stable rank is 19.5 by median (per-seed 10.5–23.6), and the penalty caps at 0.25 and 0.40 end at 16.5 and 14.5, against 6.9 free-running.

K WHERE THE COLLAPSE COMES FROM: A WEIGHT-DECAY-FREE CONTROL AND A SWITCH-OFF (WD0)

Pre-registered in `protocol_wd0.md` before any run, with two logged deviations. Same recipe as Sec. 4. All three rank estimators, the top-5 Fourier fraction and the total weight norm are logged throughout.

$\lambda = 0$ **(deviation D1: the pre-registered design fails, and the failure supplies the premise below).** We intended to separate "the collapse is weight decay" from "the collapse is intrinsic" by running grokking at λ=0 with an Omnigrok-style large initialisation (all parameters $\times\alpha$). Nothing groks ($\alpha \in \{1, 3, 6, 10\}$, 50k steps, seed 0):

Table 5: Training at λ=0 with an initialisation scaled by α (seed 0, 50k steps).

α	T_{mem}	final val	PR(W_E)	stable		top-5
1	247	0.047	68.0	24.4		0.163
3	996	0.072	86.2	32.3		0.118
6	—	0.009	86.0	31.1	0.101	(never memorises)
10	—	0.009	80.7	28.4	0.102	(never memorises)
ref. λ=1	247	1.000	21.7	6.90		0.946

At λ=0 the network memorises perfectly and then sits still: no rank collapse, no frequency concentration, no generalisation. So λ=0 removes the collapse and grokking together: they are confounded, and this design cannot separate them. The stall is not a matter of horizon but of a dead gradient. It is Softmax Collapse (Prieto et al., 2025): after memorisation the logits grow until float32 absorption zeroes the correct-class gradient (measured on this recipe at α=1: correct-class gradient proxy 0.0000, SC fraction 1.00, median logit margin 82–93 nats against the 16-nat absorption threshold, all 5 seeds, to 100k; the same freeze holds at $\alpha \in \{0.5, 0.7\}$), a hard stop that no additional horizon can undo. What the design does establish is the premise for everything below: with the gradient dead, **under standard softmax cross-entropy weight decay is the only force still acting after memorisation**, and it drives the entire post-memorisation trajectory.

Softmax Collapse and generalisation without weight decay (Gate B). Because the $\lambda{=}0$ stall is a numerical hard stop, the test of whether weight decay is uniquely the compressive force is to remove Softmax Collapse and see what happens, pre-registered in `protocol_gateB.md`. Removing SC alone is not enough in this recipe: float64 softmax (it only delays SC) and StableMax (it removes SC but not the logit-scaling "naïve loss minimisation", whose weight norm blows up to $\sim$300) both fail to generalise at $\lambda{=}0$ in 300k steps ($0/4$ each). Preventing the logit scaling directly, with $\perp$Grad, which projects each gradient orthogonal to its weight vector (Prieto et al., 2025), generalises at $\lambda{=}0$ in all 4 seeds, and fast ($T_{\text{gen}}/T_{\text{mem}}{\approx}2$, i.e. without a grokking delay). That weight-decay-free generalisation **collapses the rank fully** (PR(W_E) 85 $\rightarrow$ 13–15, at or below the $\lambda{=}1$ endpoint, with top-5 Fourier 0.92–0.98). This wd-free generalisation is not addition-specific: it replicates on the multiplicative group ($4/4$ seeds, PR$\rightarrow$ 11–17) and on the quadratic family $a^2{+}b^2$ ($4/4$ seeds, fast, PR$\rightarrow$ 14–17). The one family that does not reach it, $a^2{+}ab{+}b^2$ ($0/4$ within 300k), turns out not to grok in this architecture even under standard weight-decay training (a diagnostic control: validation plateaus at $\sim$0.2 at train fraction 0.3 and $\sim$0.5 at 0.5, while its separable sibling $a^2{+}b^2$ groks cleanly), so its failure is a property of the task here, not evidence against the weight-decay-free mechanism. So the collapse is the signature of the concentrated Fourier circuit under either mechanism that generalises this architecture, weight decay (delayed) or $\perp$Grad (fast), rather than a weight-decay artifact; and, as Sec. 4 shows causally, it remains unnecessary (pinned high, a distributed solution groks). The grokking delay is specific to the weight-decay recipe; the low-rank circuit is not.

$\perp$Grad off the island: the rescue is one-sided. The same $\perp$Grad at $\lambda{=}0$, applied to unpinned BICL (registered in advance as a two-branch boundary measurement with no favoured outcome), does not generalise ($0/4$ seeds), and the logs show that it does not stall but diverges: $|\text{logits}| \sim 10^7$ with margins -270 to -520, where the identical method groks $4/4$ on modular. The pinned stage of that experiment therefore never opened. A registered limitation applies: $\perp$Grad combined with AdamW does not preserve the weight norm, because Adam's per-coordinate rescaling rotates the update out of the plane orthogonal to the weights, and even an exactly orthogonal discrete step grows the norm at second order. This is therefore a statement about the recipe as run, not about idealised orthogonal-gradient flow. Consequence for the main text: driver-independence of the algebraic-side non-necessity is established, and no weight-decay-free route is claimed on the in-context side.

The switch-off (deviation D2: prediction P5 fails its own bar). $\lambda{=}1$ until a run has generalised (validation > 0.95; median step 6750), then $\lambda := 0$, continuing to 50k; against a matched $\lambda{=}1$ control. Six paired seeds, medians:

Table 6: The switch-off: medians over six paired seeds.

	PR(W_E)	stable	top-5	weight norm	val
initialisation	85.4	30.9	0.104	37.2	—
at switch (val > 0.95)	37.2	10.87	0.635	—	0.972
S0 control ($\lambda{=}1 \rightarrow$ 50k)	**18.9**	6.65	**0.962**	34.3 (decays)	**1.000**
S1 switch-off ($\lambda{:=}0 \rightarrow$ 50k)	**33.1**	9.93	**0.720**	**62.7** (does not)	**0.982**

We had predicted the compression would freeze ($< 5\%$ further movement). It does not: it moves 9–13% (participation ratio -11%, stable rank -9%, top-5 $+13\%$), driven by the residual cross-entropy gradient while validation is still climbing (0.97 $\rightarrow$ 0.98). **P5 fails as stated.** What the compression does is stall:

- 73% of the total rank collapse (85.4 $\rightarrow$ 18.9) happens *before/at* generalisation; 27% after it. For the Fourier concentration: 62% / 38%.
- Of that post-generalisation compression, only 22% (rank) and 26% (top-5) survives with λ switched off, so roughly three quarters of it is weight decay.
- It buys 1.8 points of validation accuracy (0.982 vs 1.000), at the cost of nearly half the remaining effective rank (37 $\rightarrow$ 19).
- The weight norm confirms the intervention bound: 62.7 (no decay) against the control's 34.3.

A wording this corrected. The causal result of Sec. 4 on modular arithmetic (pin the rank, it still groks) shows that none of the modular collapse is necessary, although off the algebraic island the same pin blocks BICL generalisation at its 1.3M horizon and prevents the v2 circuit from forming, so the intervention's effect is conditional. It does not show that most of the modular collapse happens after generalisation; most of it does not, since 73% is concurrent. The main text keeps the two claims separate.

L THE PRE-REGISTERED UPGRADE PROGRAMME: PROVENANCE, HORIZON AUDIT, AND PREDICTION LEDGER

Provenance. Every result in this paper (the horizon corrections, the c^*/f^* fits, the second-circuit gate and dose, the cross-control, the subspace arms, and the failed mechanism attempts) comes from a single append-only pre-registration file (amendments numbered to A73; 70 filed, three numbers unused), each fixing estimators, bars, horizons, exclusion rules, and verbatim outcome wordings before the corresponding cells ran. Per-cell registered readouts are confirmatory; programme-level narrative (which experiment to run next) is exploratory and is labelled so. Engine extensions used by the subspace arms are verified additive on full runs through memorisation, collapse and grokking (under a 300k budget with the engine's standard $t_{\text{grok}}+5$k early-stop), not only unit tests: with the extensions present but off, three isotropic gate cells reproduce archived grokking steps exactly (t_{grok} 38,400/30,600/42,400, and initial stable rank to six decimals) across two engine changes and a horizon change.

Horizon audit of every causal cell family. "Block" is only ever asserted at a stated horizon with held floors and, where available, a flat order-parameter; "$k\times$" is the horizon as a multiple of the slowest unconstrained *grokking step* in the configuration-matched cohort (245,600 for the BICL baseline cohort; our first computation of these ratios divided by that run's last logged step, 265,600, and so understated every margin). That cohort maximum belongs to seed 9, which is in no pinned set, so Table 7 also gives the *paired* multiple: the horizon over each constrained seed's *own* unconstrained grokking step. Paired multiples are post-hoc (no amendment registered them) and descriptive, and one selection rider travels with them: the BICL dose seeds are the 12 fastest grokkers of 16 baseline candidates, chosen by the registered rule, which lowers every paired denominator (and the candidate list itself omits seeds 1 and 8, seed 1 as a known slow grokker). The unpinned base rate belongs here too: 18 of the 20-seed cohort grok by 400k and two never do, so "0/5" is read against the five pinned seeds' *own* unpinned runs, all of which grok by 28.6k.

Thresholds inherit the horizons of the cells behind them and are written horizon-indexed throughout: c^*(60k), f^* (200k dose, 1.3M block certification), f^*_{v2} (75k).

E2: the $0.9\times$ block re-certified at 1.3M (registered A73). The five $0.9\times$ cells were re-run from initialisation under the same engine to 1.3M steps, with both readouts registered before the run. They were re-run rather than continued because P0.5's checkpoints hold no optimiser state, and a continuation would have restarted Adam cold on a system whose claim is that it sits still. Each readout fired branch (i): 0 of 5 valid cells generalises, floors held over the whole window (5/5 at the registered $\geq 0.98\times$ tolerance), the circuit meter TOST-flat over the final 200k (max per-seed $|\Delta|$ 0.006 at 1.2M and 0.004 at 1.3M, 90% interval inside ± 0.02), and maximum validation accuracy 0.26–0.27 across the campaign. The first 400k steps replay the archived cells and reproduce every 400k verdict, so the earlier 400k certification is reproducible. Paired on seed the horizons are 42.0–89.6$\times$ (1.2M) and 45.5–97.0$\times$ (1.3M); the second passes the $\approx 45.4\times$ at which our last algebraic null dissolved on four of five seeds and on seed 0 is level with it within logging resolution (45.45 against 45.38). Passing it on every seed is what the extra 100k steps were registered to buy, so that registered aim is met on four seeds only. On the cohort scale, however, 1.3M is $5.3\times$ the slowest unpinned grok against that null's $15.8\times$: the precedent is passed on the paired scale only. Cost ≈ 35 hours on one GPU.

The BICL dose, paired on seed (post-hoc, descriptive). Every dose cell has an unpinned run of the same seed in the same campaign, so each pinned grokking step can be divided by that seed's own. The five 400k cells are seeds $0, 2, 3, 4, 5$, whose unpinned runs grok at 28,600, 17,400, 13,400, 27,400 and 27,800 steps: the 400k horizon is 14.0, 23.0, 29.9, 14.6 and $14.4\times$ those, and the 1.3M

Table 7: Horizon audit of every causal cell family.

cell family	constraint	n	horizon	held	verdict
modular rank pins (W_E/W_U/stable/both)	floors $0.9\times$ init	$10/10/6/6$	20k	yes	grok on time
modular $\perp$Grad $\lambda{=}0$ pin	both floors	4	—	yes	grok, fast
modular, *attention-only* two-layer model, same-object pin (train fraction 0.4)	both floors	6	55k (4.1–$10.4\times$ paired; never binding)	$6/6$ at $0.95\times$ ($3/6$ at $0.98\times$)	grok $6/6$, median $0.92\times$ paired
modular same-object wd-matched pin	both floors	6	60k	$2/6$ strict (4 sub-logging transients $\geq 0.96\times$, disclosed; statistic held $\sim 0.87\times$ init vs. baselines 0.14–$0.21\times$)	grok, *faster*
modular hinge Fourier (achieved ≈ 0.090)	uniform-floor cap	10	20k only	yes	no grok by 20k (never extended)
modular projection cap 0.095	cap	8	60k/300k	yes	delay: $0/8$ then $1/5$
modular projection cap 0.110	cap	8	60k/300k	yes	delay: $3/8$ then $4/5$
modular projection caps 0.13–0.225	cap	8 each	60k	yes	delay/grok ($6/8 \to 8/8$)
BICL pin $0.9\times$ (E2 re-run, A73)	both floors	5	1.3M ($6.5\times$ dose horizon; $5.3\times$ cohort's slowest grok; 45.5–$97.0\times$ paired; its first 400k replays the archived cells)	$5/5$ full-window	**block at H**, meter TOST-flat
BICL pin $0.7\times$	both floors	4	400k (14.0–$29.9\times$ paired)	yes	no grok by H
BICL dose 0.25–$0.9\times$ (5 floors)	stable-rank floor	12 each	200k	all bound	graded $12/12$, $12/12$, $8/12$, $2/12$, $0/12$ across the five floors; $f^*(200\text{k})$
v2 gate $N{=}15360$ (pin/base/push)	both floors	$8/8/5$	75k	$8/8$	**formation prevented**: $0/8$ vs $8/8$
v2 dose (f^*_{v2}), 9 floors	stable-rank floor	10 each	75k	$90/90$	graded $10/10$ to $0/10$, per-floor tally in App. I ($8/10$ dip at $0.4\times$ disclosed); $f^*_{\text{v2}}(75\text{k})$
Arm A (circuit-complement pin; seed 4's max validation 0.279 is identical to three decimals at 300k and 400k)	complement floors $0.9\times$	5	four cells 300k ($1.22\times$; 4.6–$10.8\times$ paired); s4 400k ($1.63\times$; $14.6\times$ paired)	$5/5$	**block at H** (seed-4 confirmed)
Arm A, same seeds, weaker floor	complement floors $0.55\times$	5	300k	$5/5$	grok $5/5$ (t_{grok} 27.6k–113.4k): block tracks how firmly the floor binds (subspace held fixed; direction untested)
isotropic comparators for the row above	total floors $0.30/0.35$	$8{+}8$	200k	$16/16$	grok $16/16$ (total stable rank 11.0–15.9)
Arm B (grad-confined, max-crosstalk vs random)	total floor $0.4\times$	$4{+}4$	300k	$4/5$ each	both grok (null; floor too weak to discriminate)
Arm B at the discriminating floor	total floor $0.55\times$	$5{+}5$	300k	$10/10$ / $5/10$ (two scorings)	no specificity; readout not supported (scorings disagree)
isotropic $0.4\times$ drift gate	total floor	3	300k budget	yes	grok $3/3$
$\perp$Grad $\lambda{=}0$ on BICL (unpinned)	—	4	300k	—	diverges, $0/4$

horizon of the E2 re-run is 45.5, 74.7, 97.0, 47.4 and 46.8× them. Among cells that grok by 200k the paired ratio is graded in the floor:

Table 8: The BICL dose paired on seed (post-hoc, descriptive).

stable-rank floor	0.25×	0.4×	0.55×	0.7×	0.9×
grok by 200k	12/12	12/12	8/12	2/12	0/12
median paired ratio (grokkers)	1.01	1.24	2.23	5.32	—
range (grokkers)	1.00–1.04	1.05–1.92	1.27–6.35	4.56, 6.08	—
censored cells, paired lower bound	—	—	>3.8–14.9	>3.6–14.9	>3.6–16.1

The comparison that matters for reading "block" is on the algebraic side: the cap-0.095 cell that grokked at step 254,100 is seed 0, whose unconstrained run groks at 5,600, a paired multiple of ≈45×. The BICL certification passes it on four of five seeds and is level with it on seed 0 (45.5–97.0× at 1.3M, registered A73); the complement-confined one does not (4.6–10.8×, seed 4 at 14.6×). On the cohort scale the algebraic precedent is still the larger of the two (15.8× against 5.3×), which is why the main text claims a block *at a horizon* and a contrast (on time against a roughly forty-five-fold or larger paired delay), and never arrest.

The architecture-crossed probe (registered A72). Every algebraic cell above runs on an attention-plus-MLP transformer and every in-context cell on a two-layer attention-only one, so "the pin is inert on algebra" might have been a fact about MLPs. We removed the MLP from the algebraic side: the in-context model class itself (two layers, attention only, no LayerNorm, d=128, 4 heads) trained on $(a + b) \bmod 113$ under the same-object pin (both floors at 0.9× init, wd-force matched, β=10; the pin's code is imported, not re-typed), paired on seed with its own unpinned baseline. The recipe rule, horizon rule, gates, branches and scoring script were fixed before any pinned cell. *Recipe.* At train fraction 0.3 the one seed probed (seed 0) memorises and does not generalise by 60k (our registered prediction that it would is scored failed); the registered fallback adopts the smallest fraction that groks with a delay, 0.4, at which 6/6 baselines grok (t_{grok} 5.3k–13.5k) with the usual collapse (stable rank 30–32 → 3–5). *Pins* (H=55k, 4× the slowest baseline, never binding; the last row is read at the step where the *unpinned* run generalises):

Table 9: The attention-only modular probe (A72), per seed.

seed	0	1	2	3	4	5
unpinned t_{grok}	13,400	13,500	9,300	8,700	5,900	5,300
pinned t_{grok}	10,900	9,600	8,600	11,000	7,000	4,800
paired ratio	0.81	0.71	0.92	1.26	1.19	0.91
min stable rank / floor	0.992	0.986	0.965	0.970	0.994	0.973
stable rank then: pinned / free	29.2/8.1	28.3/4.3	29.0/5.6	29.7/3.2	28.5/6.6	29.0/9.2

All six cells pass both registered gates (floors held at the pre-declared 0.95× tolerance, 3/6 at the strict 0.98×, the sub-logging transient this pin already showed on the MLP model; and the pin verifiably on at the step where the free network generalises), all six generalise, and the median paired ratio is 0.92: the registered branch is inert. Riders: one task, one recipe, a train fraction that differs from the rest of the paper's algebraic cells, and the other crossed cell (an MLP-equipped in-context model, which would move the pool-size window and is a campaign rather than a probe) unrun. 61 GPU-minutes.

Prediction ledger. Registered author predictions and the estimator's, scored as registered. The "4 of 11" quoted in Limitations counts the predictions attached to the results reported here; the pre-registration keeps its own running tally over every arm it ever registered, including arms not reported in this paper, and that tally read 4 held of 15 at A65. *Held (4):* caps 0.11 and 0.095 grok at 300k (registered with a rationale that was separately withdrawn as refuted out-of-sample before scoring; the prediction was kept, its justification was not); that the complement pin's own dose would return a null, registered as "Arm A at 0.55× groks ≥4/5", and it grokked 5/5; that the same pin is inert on the attention-only algebraic model (A72); and that E2's long-horizon re-run would return branch (i) at both registered horizons (P-E2, ∼70% confidence: 0/5 at 1.2M and

at 1.3M, floors held, meter flat). The second, third and fourth carry a qualification: predicting a null, after four mechanism-side negatives and after a same-object cross-control that had already shown one, is the easy direction to be right in. E2's alternative branch was live enough that A73 registered its wording in advance. *Failed (*$7{+}1$*):* that a two-layer attention-only model would grok modular addition at train fraction 0.3 (the one seed probed memorises and does not generalise by 60k; the registered fallback adopted 0.4); that the matched random push's high measured rank was an artifact (projected stable rank predicted $<0.6\times$ init; measured 0.98–$1.04\times$, so it generalised at near-init rank); the circuit-subspace gate's dispersion prediction; the 60k prediction that cap 0.095 would grok 1–$3/8$ (it was $0/8$; the block claim briefly built on that $0/8$ was itself retracted a day later when 0.095 grokked at 300k); the AFT rationale behind the held prediction (withdrawn as refuted out-of-sample; scored here because it was a registered piece of reasoning); Arm A at the $0.9\times$ complement floor groks $\geq 4/5$ (it blocked $0/5$); Arm B blocks $\geq 4/5$ (it grokked $4/4$); and the estimator E-b's own prediction that the gradient-confined arm blocks (it grokked; the estimator had already missed its postdiction bar, median 0.180 vs. $[0.55, 0.70]$). Every failure was caught by a registered bar before it could reach a claim; after these corrections, **no surviving claim depends on an author prediction having been right**. Two further registered cell-level predictions sit outside that running tally and also failed: the cross-control's on-time prediction ($\geq 5/6$; it was $4/6$, with both misses faster than baseline), and the registered $c^* \in (0.13, 0.20)$ (the fitted CI lies entirely below 0.13, a quantitative miss). A third sits outside the tally and held: that the 0.30/0.35 gap-filling cells would grok $16/16$ (they did). It stays outside because it was labelled low-information when it was registered, since those floors lie far below every threshold we measured. A post-hoc "v2 is a step" characterisation of the coarse dose grid was withdrawn when densification showed the dose graded; it was a characterisation, not a registered prediction, and is listed here for completeness.

Subspace-arm construction and riders. *Arm A*: per-seed circuit basis V_c = top-d_{circ} right singular vectors of that seed's own grokked endpoint W_E (d_{circ} from the registered G2 measurement: circuit dimension 10–28, median 15 of 128, IQR 18%; an oracle target from the free run, disclosed); the hinge is the BICL pin of Sec. 4 evaluated on $W_E(I - V_c^\top V_c)$ with floors $0.9\times$ the *complement's* own init statistics; held gate on the complement's post-memorisation stable rank. A registered feasibility check (satisfiability, like-for-like magnitude: complement/full init stable-rank ratios 0.91–0.99) passed on all seeds before launch. Rider: the registered follow-up measurement (circuit-projected rank at the endpoints, to separate "consolidated but drowned" from "never consolidated") was unexecutable on the cells as they ran, because the pin engine saved no endpoint weights at that time, so what the block does to the circuit directions themselves is not measured here. Endpoint saving was made mandatory afterwards and exists for the five weakened-floor cells below and for the re-run drift-gate cell; we register no readout on it in this paper and report none. What the block is a function of is answered instead by the arm's own dose, below. *Arm B*: pinned statistic is the *whole-matrix* stable rank at the registered $0.4\times$ total floor (byte-identical to the control); only the auxiliary *corrective gradient* is confined, to I = top-d_{circ} left singular vectors of the stacked four-head cross-talk matrix (a registered fix after the per-head choice proved a near-four-way tie; per-head subspaces overlap only 0.44–0.60), versus a matched random off-circuit subspace. Riders, carried verbatim: the subspace comes from an estimator that failed its own registered bar; I is 33–47% aligned with the top-energy subspace; seed 4 (smallest $m{=}10$) went under-floor in *both* arms and is excluded by the registered symmetric rule (including it gives $5/5$ vs. $5/5$, the same null); the estimator's failed arm-prediction was computed under the earlier single-head operator (the stacked recompute predicts the same way and fails the same way); and the ARMB-faster-in-4-of-5 pattern is post-hoc and not significant (sign test $p{=}0.375$), so it is not a trend. A first Arm B construction was withdrawn before running when an adversarial design review found it algebraically impossible (the cross-talk operator's rank is bounded by d_{circ}, leaving most of its "maximally cross-talking" directions with zero cross-talk); the withdrawal is on the record.

The two failed threshold derivations, in full. *(a) Interference occupancy* (registered): f_{pred} median 0.180 against a registered $[0.55, 0.70]$ bar, per-seed spread 0.075–0.585 (a factor of 7.8, so the estimator is unstable, not merely off), and its one registered prediction (that the gradient-confined arm would block) also failed. The estimator selects an induction head by argmax, and there is no single induction head (three to four layer-2 heads score ≈ 0.85 within 0.02 of each other), so a post-hoc, zero-weight recompute stacking all four heads was run: median 0.16, the same verdict, which is why the head-selection artifact does not rescue it. *(b) A pin-independent derivation*, pre-registered in full before any held-out quantity was computed: $f^* = (s_\infty + d_{\mathrm{circ}})/s_0$, with the free-run end-

point stable rank s_∞, the measured circuit dimension d_{circ} and the initial stable rank s_0 all read off freely trained networks, no intervention anywhere in it. It agrees on the fitting task (0.612 against a measured 0.600) and is *refuted* on the held-out one (0.547 against 0.845), in the wrong *direction*: v2 has the *smaller* circuit (11 vs 15 dimensions) and the *higher* threshold, so no re-normalisation repairs the ordering and the family is excluded. The pre-registration said in advance that agreement on the fitting task alone would count as a coincidence.

The matched push is not a mild perturbation. Measured on the v2 cells, projecting out the injected direction together with its sign pattern removes 99.6–100% of W_E's energy at t_{grok}. The networks that generalise under the push do so with an embedding that has been almost entirely replaced, at near-initial rank.

Three failed attempts at a non-rank control (registered A66–A71). Every blocking arm in this paper is a rank hinge, so "a rank floor blocks" is not yet separated from "any persistent restoring force of that magnitude blocks". We tried three times to build the control, each pre-registered, each wd-force matched at $\beta{=}10$ on the BICL engine, and none produced an interpretable outcome. *(1) A floor on* $\|W_E\|_F$ at $0.9\times$ init, seeds $0, 2, 3, 4, 5$: all five generalised, but $\|W_E\|_F$ *rises* to $\approx 1.5\times$ init while the network memorises and reaches that floor only at or after the grokking step, so every logged column of every cell is identical to its unpinned baseline until then and t_{grok} is identical to the step: the arm is void, not a null. Its one registered re-specification (floor $1.6\times$ init) was disqualified at a single-seed pilot: a radial hinge's gradient is anti-parallel to weight decay, and the pilot held stable rank at 38.6 at 20k steps, at or above what the rank pin itself retains there (33.2–38.0), so it could not serve as a *non-rank* control. *(2) A floor on a fixed random linear functional of* W_E: abandoned at its pilot on two registered gates: the hinge fired once at step 0 and never again in the window (logged fire fraction 0.0000), and rank neutrality failed (stable rank 17.1 at 15k against a 15.0 bar). *(3) A floor on the energy share in a fixed random* 32*-dimensional subspace* (scale-invariant, chosen by a zero-GPU screen): abandoned at its pilot on the same two gates (fire fraction 0.0000; 16.0 against 15.0). All three author predictions are scored void, neither held nor failed; ≈ 1.3 GPU-hours in all. The control therefore remains unrun in any valid form.

Headline proportions. Exact Clopper–Pearson 95% intervals for every headline tally are in the released statistics file; the ones a reader should weigh when a cell is small: v2 pin $0/8$ $[0, 0.369]$; BICL $0.9\times$ $0/5$ $[0, 0.522]$; Arm A $0/5$ $[0, 0.522]$; Arm B $4/4$ $[0.398, 1.000]$ per arm, which is why the Arm B null supports only "no evidence of direction-specificity" and nothing stronger.

What the collapse builds (descriptive, no causal claim). Grokked BICL endpoints install an item-vs-label role separation: separability in the top-8 singular directions is 0.53–0.59 at memorisation and ≥ 0.90 in every grokked seed. The blocked-and-held 400k endpoints sit strictly between those bands (0.61–0.79), and no blocked cell reaches any grokked cell. This is partial role structure without generalisation; the numbers do not support "the pin prevents installation".

The complement pin's own dose (registered before it ran). The $0.9\times$ complement arm was open to a deflationary reading we had raised against ourselves: its endpoints stop at total stable rank 18.3–23.1 while a free run collapses to ≈ 10, so an isotropic dose could explain the block with no reference to subspace identity. We re-ran the identical construction (same five seeds, same per-seed circuit basis, same horizon, same held gate) with the complement floor weakened from $0.9\times$ to $0.55\times$, registering the branches, the bar and our own prediction first. All five cells held their floors, driven to within 0.9989–1.0002 of them, and all five generalised, at total stable rank 9.2–11.5 (the projection before they ran was 11.2–14.1):

Table 10: The complement-confined pin at two floor strengths on the same seeds.

seed	complement floor $0.90\times$	complement floor $0.55\times$
4	total 23.1, block	total 11.3, grok
5	total 22.2, block	total 11.5, grok
6	total 18.3, block	total 9.2, grok
7	total 21.1, block	total 11.4, grok
8	total 20.8, block	total 11.4, grok

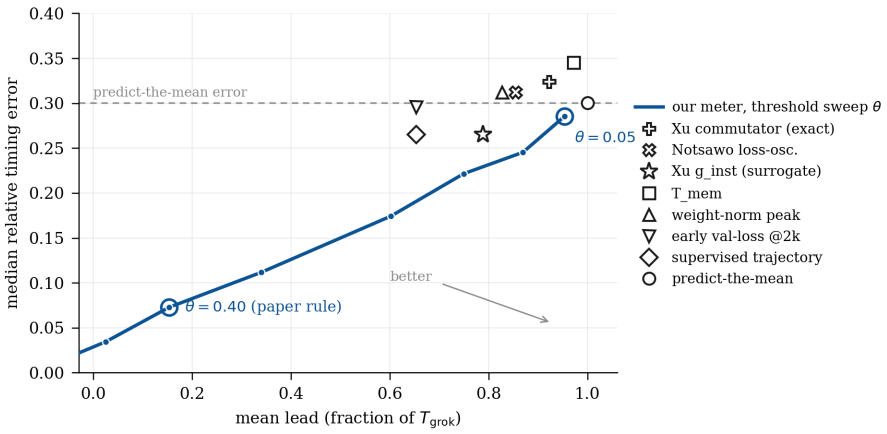


Figure A9: **Lead versus timing accuracy on one plane (n=30, identical nested LOO).** Sweeping our meter's firing threshold θ traces the curve; prior early-warning signals and cheap baselines are single points at their published/natural lead. Lower-right is better. The trade-off is intrinsic to a *single* signal (the curve); every baseline lies inside the frontier *at its published threshold* (a post-hoc sweep of two baselines' thresholds finds one point outside by 1.5 points; App. M); and every method, ours included, approaches the predict-the-mean error (dashed) as the lead approaches T_{grok}.

$5/5$ against $0/5$, perfect separation on paired seeds: exact paired test p=0.031 one-sided (the direction was registered in A63(9) before the cells ran), 0.0625 two-sided. The registered statistic was Fisher's exact on the 2×2 block table, p=0.0079; that null treats the ten runs as independent while the design pairs them on seed and circuit basis, so we report the paired value as primary and keep the registered one visible. The same correction applies to the v2 gate, whose arms are also the same eight seeds (0.0039 paired one-sided vs. a registered 1.6×10^{-4}). Neither claim turns on the p: both tables are perfectly separated with floors held in every cell. Sixteen further isotropic cells at floors 0.30/0.35 were run so that the registered matched-rank comparison would have comparators inside the achieved band at all; they generalise $16/16$ at total stable rank 11.0–15.9, and the registered contrast is $0/5$ against $0/14$ inside the band, p=1.000, reported descriptively because matching on achieved rank conditions on a post-treatment variable. Two disclosures travel with this. First, the paired contrast spans an engine change, so a registered drift gate re-ran the archived cell closest to flipping (seed 6, the lowest archived total rank) on the current engine: it still blocks and reproduces 18.307 against an archived 18.307. Second, the engine's duty-cycle statistic scored all five cells "slack" while the floors were bound to within 0.1%, the fourth time in this programme that a duty statistic has mis-scored a working pin, and the reason no duty-cycle bar gates any verdict in this paper.

What the sixteen new cells did and did not buy. They were registered to make the matched-rank comparison satisfiable at the *projected* Arm A band (11.2–14.1). The arm landed lower, at 9.2–11.5, so only 3 of the 16 fall inside the band that was finally matched on; the 14 in-band comparators are those 3 and 11 archived isotropic cells, and the 11 archived cells alone already meet the rule for the comparison (at least five cells inside the achieved band). The new cells were needed for the projected band, not the achieved one; what they add is a $16/16$ generalisation result at floors 0.30/0.35 and 3 of the 14 comparators.

M Predictor detail: the lead–error frontier and the corollary's caveats

The frontier. Comparing methods at their own natural lead is not a like-for-like comparison, so we sweep *our own* rule's threshold θ and trace the full lead-versus-error curve under the identical nested protocol (Fig. A9). As θ falls from 0.40 to 0.05 the same meter buys lead ($0.15 \rightarrow 0.95$ of T_{grok}) and pays in error (7.3% $\rightarrow$ 28.5%). Every prior early warning sits inside this frontier at its published threshold (commutator defect 32.4% at lead 0.92 vs. ≈24% for our meter at lead 0.87;

loss-oscillation 31.2% at 0.85). In the long-lead regime every method, ours included, approaches predict-the-mean (28.5% vs. 30.0% at lead 0.95; with 130 calibration seeds the gap widens only modestly, to 19.8% vs. 25.9%). A post-hoc, descriptive sweep of two baselines' own thresholds (`baseline_threshold_sweep.py`) qualifies this reading. The commutator defect lies inside the frontier at every multiple we tried ($k \in \{2, 3, 5, 10, 20, 50, 100, 200\}$, LOO error 25.7–35.7%; at k=50, lead 0.78, it reaches 26.1% against 22.8% for our meter), and for $k \leq 20$ every seed fires at the same one or two logged steps, so the rule reduces to predict-the-mean. The gradient-instability surrogate, stored at three fractions, lies inside at 0.5 and 0.7 (26.5% against 22.9%; 23.7% against 22.2%) and outside at 0.3, where it reaches 25.6% at lead 0.92 against 27.1% for our meter: 1.5 points better, and 4.5 points below predict-the-mean. Frontier-dominance is therefore claimed only at the published thresholds. Prior early warnings on the same powered n=30 runs: loss-oscillation (Notsawo et al., 2023) 31%; a gradient-instability surrogate 27%; the exact commutator defect of Xu (2026a) (Definition 2 and dual onset rule, verbatim) fires at 447 ± 85 steps against a T_{grok} std of ∼3200 (lead 92% of T_{grok}; Spearman 0.10, p=0.63; LOO error 32%, i.e. predict-the-mean).

Held-out modular conditions. Calibrating anchors, threshold and regression on the base condition only (θ=0.408) and predicting every other additive condition (8 seeds each) gives median relative errors of 6.6% (weight decay 0.5), 15.7% (weight decay 2.0), 4.5% (subtraction) and 4.1% (p=97); mean T_{grok} differs 3.5× between the fastest and slowest of those conditions. This is one calibration condition, not a rotating leave-one-condition-out, and it is transfer *within* modular arithmetic.

Calibrated transfer is group-dependent. The 40% rule and scale ratio calibrated on modular Fourier-Gini, applied unchanged to S_5's irrep-Gini, predict S_5's grokking step with 17% median error (vs. 58% predict-the-mean), but do *not* transfer to D_{60} (40% vs. 32% predict-the-mean; App. G).

The corollary's four caveats, in full. First, it is measured on modular arithmetic. Second, retained rank is not a downstream benefit we claim. A pre-registered probe fine-tuning the two endpoints on a second task found that the rank-preserved endpoint adapts no faster (0/6 seeds faster on the near task, negative transfer on a far one), consistent with norm rather than rank governing downstream plasticity (Dohare et al., 2024). Third, prolonged λ=0 training after grokking has a late-stage hazard (test-accuracy collapse near 10^7 steps, which weight decay suppresses; Prakash & Martin, 2026); on this recipe the post-switch endpoint is itself measured to *freeze* (Softmax Collapse zeroes the surviving gradient, 3/3 seeds, W_E gradient $\sim 10^{-12}$, validation flat to the horizon), so the hazard cannot engage here; for recipes where the gradient survives the switch, the safe form may be a floor $\lambda_{\min} > 0$. Fourth, the trigger pattern has neighbours of opposite sign (a synchronisation precursor used to *raise* λ (Sivasankar, 2026); post-grok decay removal as an analysis probe (Xu, 2026c); stable-rank-triggered regime switches (Wang et al., 2023)); the completion-gated decay-*off* direction and its measured price are what the ledger adds. We have not trained a trigger-gated arm end-to-end; the trigger fires a median ∼5% later along the same trajectory, so the rank its switch would retain is bounded by the measured endpoints (18.9 control, 33.1 oracle switch). At θ=0.60 the trigger fires at median validation 0.755, 29/30 early: the threshold must be written on the statistic that fires, against the event claimed, the same lesson as the estimator disclosure of Sec. 4. The two usable settings in full: θ=0.75 fires at median validation 0.987 with 2/30 early, θ=0.80 at 0.997 with 0/30 early. θ itself is chosen on this cohort and the full sweep is reported; only the *anchors* are nested-LOO, not θ.

N Off-island and non-abelian detail

Reconciling with reported irrep synchrony. Chughtai et al. (2023) report that embedding and unembedding concentrate onto irreps *synchronously*; our leading signal is W_E's *rank*, a distinct observable, and on the same runs W_E's rank collapse and its irrep concentration are essentially synchronous (10% onsets within a median 110 steps). What separates them is timing *fidelity*, not precedence, so the dissociation lives between two co-timed observables of one matrix, consistent with the reported synchrony.

S_5 onset disclosure (from Sec. 6). Seed counts: 14 grokking seeds originally, 30 in the powered replication, and 28–29 where an analysis excludes cells. The onset step's *Pearson* correlation is

weakly negative and nominally crosses the permutation floor (-0.60, p=0.025 at n=14; -0.42, p=0.028 at n=30) while the rank-based Spearman does not (-0.27, p=0.16); with an onset std of $\sim$28 steps (range $\sim$134) against a T_{grok} std of $\sim$5500, such an association carries no usable timing information. Both statistics are reported.

S_5 causal probe: full detail and a weaker control. Forbidding the unembedding's rank collapse (held at effective rank $\gtrsim 80$ vs. baseline $\sim$27) leaves grokking on time (T_{grok} median 4425 vs. 4366; 6/6). Capping W_E's irrep concentration delays monotonically (T_{grok} median 6.4k/10.2k/16.6k at caps 0.60/0.48/0.40; magnitude matched across the dose); the same penalty in a randomly relabelled basis (cap 0.30; a control added after the initial dose run, a disclosed protocol addition) yields at most a mild slowdown ($+24\%$, 6/6) with the true irrep concentration still rising (Gini 0.46 vs. 0.08). One control is weaker here: the matched generic freeze of W_E is itself strongly disruptive on S_5 (4/6 grok, median $\approx$15k; and a linear freeze, not the treatment arms' hinge form), so specificity rests on the relabelled-basis control and the within-dose monotonicity, not on the generic contrast. The substrate-redundancy reading (modular's one-dimensional Fourier modes are easier to obstruct than higher-dimensional irreps the optimiser can partially route around) survives only as an *interpretation* of why modular's pacing is steeper; the two doses are not commensurable across meters and we do not quantify the comparison.

The retired variance criterion, in full. Alongside the rank conjunct we had pre-registered a variance conjunct, $\text{Var}(\text{lead})/\text{Var}(T_{\text{grok}}) \geq 0.6$, and on the powered induction cohort it *passes*: 0.96 [0.93, 0.98] (0.98 [0.97, 0.99] on the drop-from-init detector). We report the pass and decline to read blindness into it: across all ten task $\times$ detector cells of the elasticity analysis the ratio lies in $[0.963, 1.004]$, so every cell clears the bar, including cells we ourselves classify as *coupled* (A_5 drop-from-init: Spearman 0.82, γ=0.20, ratio 0.994). The reason is arithmetic: with $\text{lead} = T_{\text{grok}} - \text{onset}$, the ratio is driven to 1 whenever the onset is early and low-variance, *however much timing information its rank ordering carries*. Consistently, the pre-registered scale-robust form of the same conjunct *fails* (log-T_{grok} MAD analogue 0.18/0.43, below the same bar). We therefore claim no surviving positive blindness result on the induction task, and none from the variance criterion; the off-island blind positive is BICL, on the surviving criteria.

Elasticity spectrum: full numbers. Uniform inclusion rule (memorise, then grok with $T_{\text{grok}}/T_{\text{mem}} \geq 3$; no task-specific filters). Two detectors, labelled as the protocol that registered this analysis labels them: *primary*, the drop-from-init onset (first step at which W_E's effective rank has fallen by 5), on which the registered injection test also runs; *secondary*, the relative midpoint at 0.1. (A later protocol for the powered induction cohort called the relative midpoint primary; we follow the registration of this analysis and report both.) Primary: modular $\gamma = 0.115$ [0.03, 0.19] (n=130); S_5 onsets identical across all 29 seeds at logging resolution ($\gamma = 0$); A_5 0.20 [0.14, 0.30] (Spearman 0.82); D_{60} -0.05 [-0.14, 0.00]; 1-hop induction 0.27 [0.19, 0.36] (pooled, n=20 observations; the grokked-cohort fit quoted in Sec. 6 is 0.28 [0.19, 0.36] at n=19); BICL $+0.031$ [0.001, 0.074] (n=18; Spearman 0.34 inside its 0.48 floor), excluding induction's interval. On the primary detector modular arithmetic and A_5 therefore carry real positive couplings at high power, and A_5's interval overlaps induction's; the onset-blindness verdicts of Secs. 3 and 6 for these groups rest on the relative-onset and midpoint detectors, not on this one. BICL's interval also lies just above zero, so by the protocol's sign rule it is not blind on γ alone; we call it blind because its rank correlation stays inside the permutation floor and its elasticity is a ninth of induction's. Injecting the induction elasticity into the algebraic tasks' own T_{grok} values predicts rank correlations that D_{60} excludes at P=10^{-4} and modular at P=0.011, while S_5's identical onsets carry no rank ordering at all; A_5 does *not* exclude it (P=0.86). Secondary: modular 0.013 [-0.013, 0.039], S_5 -0.03 [-0.08, 0.01], A_5 0.01 [-0.05, 0.05], D_{60} -0.08 [-0.19, 0.02], induction 0.29 [0.20, 0.39], BICL 0.021 [-0.008, 0.065]; the separation between algebra and induction is clean only on this detector. The claim is therefore graded: $\gamma \lesssim 0.2$ on algebra on the primary detector and $\approx$0 on the secondary, while the two off-island tasks split on both (blind 0.031 vs. coupled 0.27 on the primary).

BICL's seed-invariant clock and the ablations. BICL's item$\to$label bijection is resampled every sequence, so W_E has no global algebraic basis, yet its four W_E-rank onsets fire at steps 200/250/225/250 while their T_{grok} span 13.4k–115k (8.6$\times$), riding a large, non-degenerate collapse (participation ratio 101$\to$32, stable rank 37$\to$6–10) that completes well before T_{mem}. Two

pre-registered ablations attack the four features most obviously distinguishing BICL from 1-hop induction, and none restores coupling: roles off ($\gamma = +0.033$ $[-0.111, 0.241]$), raised answer entropy (L=128; $\gamma = -0.12$), fillers with fixed positions ($+0.012$ $[-0.030, 0.044]$), randomised positions (-0.11 $[-0.20, 0.00]$; markedly harder, grok rate 0.27 vs. 0.78, yet just as blind). In every arm the circuit-strength tracker still couples (completion Spearman ≥ 0.97), so the detectors retain power. The one axis not varied, the density of the in-context signal (eight exemplar pairs vs. a single trigger), is the open candidate.

Window criterion: detail. On the group-matching task the ceiling control rules out the trivial explanation: with fresh data each step the circuit forms in 2k steps (validation 0.995) and the embedding acquires the group structure (between-group variance 0.42 vs. 0.14 when memorising). The task is easy and has no grokking regime. Nor can the window be pried open with the obvious lever: weight decay prefers efficient circuits (Varma et al., 2023), whose theory predicts the critical dataset size does not move with weight decay, and our measurement agrees: raising it (0.5, 0.7) leaves the window shut and lowers N_{mem} *faster* ($4/4$ cells learn nothing). For 2-hop composition and in-context modular arithmetic we did not map the phase diagram and our horizons were short; with Wang et al. (2024) observing fixed-pool 2-hop grokking at $\sim$700k steps under atomic-fact mixing, we claim nothing beyond "no grokking in our runs" and call those cells inconclusive.

O Related interventions

Interventions. Gradient filtering (Lee et al., 2024) and a spectral regulariser on hidden-*representation* dimensionality that accelerates algebraic grokking up to $52\times$ (Kazanskii, 2026) ask a control question motivated by collapse *preceding* grokking; ours is a necessity question. Loss-numerics and gradient-projection fixes (Prieto et al., 2025) are relatives of our weight-decay-free arms, and embedding-transfer initialisation (Xu et al., 2025) is our grafting probe's closest relative (ours adds a wrong-basis null, mid-training). Closest to our corollary are a synchronisation precursor that triggers a weight-decay *increase* (Sivasankar, 2026), post-grok decay removal (Xu, 2026c) and online stable-rank triggers (Wang et al., 2023); ours runs the completion-gated decay-*off* direction and measures its cost (Prakash & Martin, 2026's hazard does not engage here).

P A transferable checklist for causal claims about training-dynamics correlates

Each rule below is stated with the incident in this project that produced it and what that incident cost. Nothing here is specific to grokking; all of it applies to any claim of the form "statistic X is (or is not) needed for behaviour Y".

1. **Horizon-index every "it blocks" verdict, and treat an at-horizon null as censored until extended.** We first read a concentration cap as a categorical block at 20k steps. Our own extensions dissolved it: 0.110 groks $3/8$ by 60k and $4/5$ by 300k, and 0.095 groks $1/5$, at step $254{,}100$. The correction came from us, about six weeks later, and we retracted a second block claim, registered at 60k for cap 0.095, one day after registering it. Together they are the reason every block in this paper carries the horizon at which it was certified and its ratio to the slowest matched unconstrained run.
2. **Write the constraint on the statistic your claim is about.** Our first rank floor constrained the participation ratio, which is *tail*-dominated. It bound exactly, and the head-sensitive stable rank collapsed anyway, $\approx 80\%$ of the natural amount. A cheap floor on a cheap statistic is satisfiable without touching the structure the claim is about. Cost: a result already written into our draft had to be redone under three estimators.
3. **Verify that the constrained quantity actually ended at its floor; do not infer it from the penalty being active.** A gradient-matched hinge can fire throughout and still lose to the cross-entropy gradient. One arm of ours grokked with the intervention effectively absent, which is a false pass rather than a result, and it was voided. Log the achieved value of the constrained statistic per cell and gate on it.

4. **Never gate a verdict on a duty-cycle-style statistic.** The fraction of steps at which the penalty fires has now mis-scored a working pin *four* times in this project: all five cells of one arm read "slack" with duty 0.000–0.003 while their floors were bound to within 0.1%. A pre-registered duty-cycle bar would have discarded the cells that carry our main null. Gate on the achieved floor.
5. **A blocking intervention needs a matched perturbation control, and the control can refute your interpretation rather than support it.** Our matched random push generalises $5/5$ at 0.91–$1.05\times$ initial rank, higher than the floor at which the pin blocks. That single control is why this paper claims a blocking effect but claims nothing about the statistic being operative, and why we dropped a "required" framing we had already drafted.
6. **Refuting necessity and establishing it are asymmetric.** One clean arm refutes necessity, because the outcome is positive and needs no assumption about what else the intervention did. Establishing necessity requires the intervention to act on the constrained statistic and nothing else. Do not let a symmetric-looking experimental design hide that asymmetry.
7. **Match comparator cells on the achieved readout, not on the knob.** We registered a matched-rank comparison at floors 0.6/$0.65\times$ because the *floor grid* was sparse there. The quantity being matched was the achieved total rank, and those floors land at $\approx$22–26 while the arm being matched was projected at 11–14 and landed at 9.2–11.5: the comparison had no comparators at all. Re-costing the registered design put it at 49.5 GPU-hours against a 20-hour bar, which fired the bar; the re-specified design at 0.30/0.35 was projected at 19.2 hours, cost 6.1, *and* produced comparators that exist. A sparse region of the knob is not a sparse region of the readout.
8. **Drift-gate any contrast that spans a code change, on the cell closest to flipping.** Our paired contrast crossed an engine extension, so we re-ran the archived cell with the lowest margin, the one most likely to flip, before reading the contrast: it reproduced 18.307 against an archived 18.307. The value of the gate lies in choosing the hardest cell rather than a convenient one.
9. **Compute a meter's *effective* chance rate, not $1/|V|$.** A third task of ours passed its generalisation meter while containing no circuit at all: the meter read 0.46 against a naive chance rate of 0.009, but the *effective* chance rate, guessing uniformly inside the small in-context candidate set, is 0.245, $27\times$ naive, and an attention probe found no induction-shaped attention at all. The task was killed at its ceiling control. Any accuracy-style meter needs its null computed under the actual candidate distribution.
10. **Sweep the whole $[N_{\mathrm{gen}}, N_{\mathrm{mem}}]$ window before declaring a task incapable.** Twice we called a task "memorisation-only" from cells on one side of the window. One of those tasks generalises *immediately* above a larger pool size; the correct statement was that its window is empty, which is a different and stronger claim. Sweep N across the full phase diagram.
11. **Treat a rewrite as an experiment on the text.** Two adversarial verification passes over our own corrections found four blockers, and *two of them were planted by the correction passes themselves*: an inverted claim, and a set of decay numbers attached to the wrong cohort because a restoration used a table and missed its addendum. Re-derive every headline number from the artifact after any rewrite. The last such audit, run one week before submission, found a factor-of-two error (400k described as $4\times$ rather than $2\times$ a 200k dose horizon) that had reached the abstract.
12. **A correction must reach the figures.** Our pre-compile grep for forbidden wording reads the LaTeX source, and cannot see inside a `.png`. Two appendix figures carried claims we had already corrected in their own titles ("BLOCKS grok $0/10$" and a variance ratio labelled as a blindness verdict) while their captions, in the source, correctly disclaimed them. After any correction, re-read every figure as an image.