INTERPRETABILITY-GUIDED DESIGN FROM A HEAD-BINDING FAILURE IN GROKKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers often generalise long after fitting their training data, a delayed transition known as grokking. Yet on some tasks the transition never comes, and it has been unclear what fails inside a model that does not generalise. We study two-layer transformers trained on math tasks, such as taking a derivative, whose answer combines several inputs. There the failure lies in variable binding: which input each first-layer head attends to. On the derivative task, most models that fail have one head bound to the constant term, which differentiation ignores, and no model that succeeds has one. Ablating that head costs far less than ablating the others, because the input it reads plays no part in the answer. Transplanting only a generalising model's embedding and readout layers, 12.5% of its weights, repairs all six failed models we tried. With the inputs shuffled, all 19 bound heads read their input through a learned tag instead of its position. The wrong binding is usually set within the first few hundred of fifty thousand steps, as each head's attention sharpens onto one input. **SoftBind** keeps the first layer's attention soft over the first 500 steps, so no head locks onto an input before the task can steer it. On a sum task this more than doubles the grok rate, and keeping the softening for the whole run raises it from 26.2% to 85.2%. Interpretability here does more than explain a failure: it can tell us where to intervene.

1 INTRODUCTION

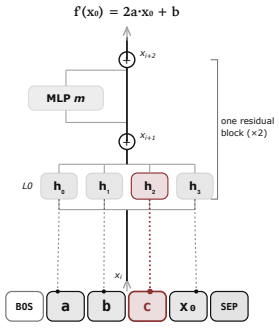
Grokking (Power et al., 2022), in which a transformer first overfits and only much later generalises, is the setting in which circuit formation during training has been studied most closely. Mechanistic accounts come largely from modular arithmetic: Fourier circuits underlie generalisation (Nanda et al., 2023), weight norm parametrizes the memorisation-to-generalisation transition (Liu et al., 2023), lottery-ticket subnetworks accelerate the phase change (Minegishi et al., 2025), and the embedding layer sets its delay (AlquBoj et al., 2025). Closest to our reading, Hidajat et al. (2026) formalise generalisation as attention placing mass on every informative token and relieve the delay with a KL-based intervention, and Gopalani & Hu (2025) and Baherwani et al. (2026) tie a plateau's end and a capability's emergence to when the attention map is learned. All three read the population; at the level of individual heads and seeds, we causally locate that coverage and relieve its failure without reading which tokens are informative.

That view treats grokking as broadly deterministic: architecture decides whether the circuit forms, dynamics decides when. It frays beyond modular arithmetic, on two counts. The per-head story is unreported: compositional and skill-ordering studies (Wang et al., 2024; Murty et al., 2023; Liu et al., 2025; Zhao et al., 2026) ask which skills emerge, not which components carry them; and single-seed reporting hides the variability that cross-seed claims rest on (Ootani, 2026).

We study both at the smallest non-trivial compositional unit, a single-step rule routing several inputs through the first attention layer: five symbolic-math tasks (*deriv*, *disc*, *det*, *dot*, *poly*) plus *disc_pad*, disc with one input its rule ignores, which serves as a positive control for the failure we describe, all sharing one tiny transformer and differing only in the rule and the number of relevant inputs (§2).

At one fixed configuration, tasks either never grok or grok on only a fraction of their seeds, and single-seed studies systematically misclassify that split (§3.1). A seed then seldom groks if one of its attention heads is lost to an input the rule ignores (Fig. 1): a per-head variable-binding problem that we localise causally and relieve by acting on the binding layer during the resolution window.

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.



$f(x) = ax^2 + bx + c$, read $f'(x_0) = 2ax_0 + b$.

Each first-layer (L0) head settles on one input position over the first few hundred steps, and the model can only generalise if the heads between them supply every reading the rule needs.

Differentiation drops c, the trap, so a head that settles there contributes nothing and leaves the rule three heads, which never suffice: the grokked runs spend one head on b and three on the (a, x_0) pair the product term needs. Which heads settle where varies by seed at one fixed configuration.

L0 bindings drawn here: $H_0 \to a$, $H_1 \to b$, $H_2 \to c$, $H_3 \to x_0$.

The seed drawn here fails: H_2 is on the trap.

Figure 1: Grokking is gated by which input each first-layer head comes to read. The seed drawn here fails the way failures actually occur: a head settled on the input the rule ignores.

Contributions. *(1) The gating event is per-head variable binding.* On the derivative rule the failure has one locus, a first-layer head bound to the one input the rule ignores: most failed runs carry exactly one; no grokked run carries one. And ablating it costs far less than ablating any other head. The finding replicates on a second rule (§3.2, §3.3), and with each row's inputs shuffled the head follows the tag that names its input, not the position (§3.4).

(2) The decision is held in the embedding: donating a grokked model's embedding block rescues failed runs. And donating the complementary parameters, seven times as many, rescues none (§3.4).

(3) The diagnosis prescribes a layer and a direction: none of 161 intervention runs chosen without it moves the failure; SoftBind softens the binding layer's attention and reads nothing about which inputs the task uses. It more than doubles the grok rate on the two tasks with headroom at pre-registered sample sizes, and held for the whole run it reaches 85.2% on sum3, where holding it at every layer does as well and at L1 alone does not (§3.5, Tab. 11). The same per-head measurement runs unchanged on Pythia models to 1B parameters (§4).

2 METHOD

Tasks: We use five math tasks plus a positive control with a padding input, distinguished by analytic rule and number of relevant inputs, and two single-trap rules introduced in §3.4:

$$\begin{aligned} \textit{deriv} &: (a,b,c,x_0) \mapsto 2ax_0 + b, & \textit{disc} &: (a,b) \mapsto a^2 - 4b, \\ \textit{det} &: (a,b,c,d) \mapsto ad - bc, & \textit{dot} &: (a_{1:3}, b_{1:3}) \mapsto \textstyle\sum_i a_i b_i, \\ \textit{poly} &: (a,b,c,x) \mapsto ax^2 + bx + c, & \textit{disc_pad} &: (a,b,c) \mapsto a^2 - 4b, \\ \textit{prodsum} &: (a,b,c,d) \mapsto ab + c, & \textit{sum3} &: (a,b,c,d) \mapsto a + b + c. \end{aligned} \tag{1}$$

Each variable occupies a fixed input slot, so binding a variable and binding its position are one event, and the draw is over which slot each head reads; Appx. B.3.3 shuffles the slots to separate the two.

Throughout, k is a task's number of relevant input variables and V the number of input positions visible to the query at the separator token SEP, from which the answer is read. H is the head count, indexed H_0 to H_{H-1} (Fig. 1), and a task's *trap* is an input its rule ignores; the full notation table is Appx. E.4. The trap is c on deriv and disc_pad and d on prodsum and sum3, whose roles are asymmetric and symmetric; the full table is in Appx. E.5. We generate (input, output) pairs from input ranges chosen so the output stays in $[-100, 100]$, then take a uniform sample of 5000 of them per task, or all of them for disc, whose range holds only 611 (Appx. E). Deriv is the informative case: its trap c is the constant term that differentiation drops.

Two splits, two capabilities: the *grouped* split holds out whole tuples of relevant inputs and the *row-wise* split single rows (Appx. E.2, which also defines the cohorts). We use the grouped split for every rate we report and reserve the word grokking for it, while the checkpoint-level analyses of §3.2–§3.4 run on the row-wise split by design, since passing it requires ignoring the trap. Every checkpoint-level number is scored on the rows its own run held out.

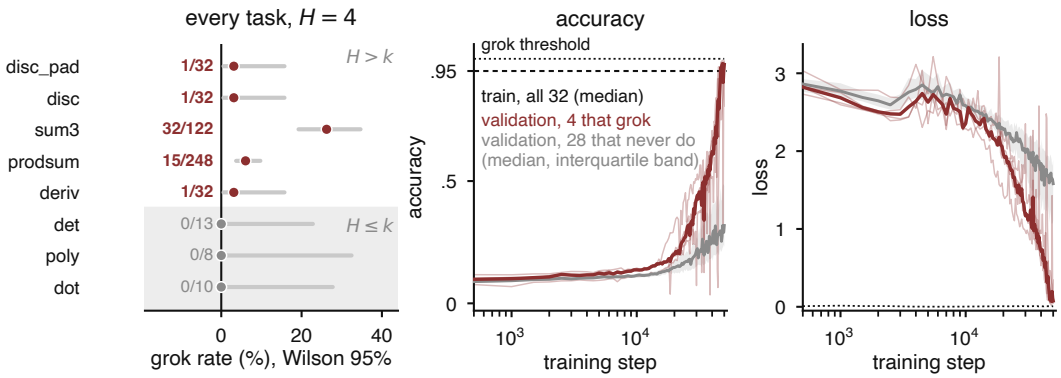


Figure 2: Left: grok rate per task at $H{=}4$ with Wilson 95% intervals (trap tasks on the grouped split; det, poly and dot row-wise); in the shaded region the heads cannot cover the relevant inputs. Centre and right: accuracy and loss for prodsum's 32 seeds, the 4 that grok drawn individually, the 28 that never do as a median with interquartile band. All memorise early, and nothing before step 20k separates the groups.

Grokking criterion: a two-layer (L0, L1), four-head, $d_{\text{model}}{=}128$ softmax-attention transformer (Vaswani et al., 2017) is trained with AdamW (Loshchilov & Hutter, 2019) for 50k or 200k steps (architecture, tokenizer and training details in Appx. E.3). A check before every run rules out copying the teacher-forced answer (Appx. E.3). A seed is *grokked* if its max validation accuracy reaches 0.95, and a task *groks stochastically* if its multi-seed grok rate falls strictly between 0 and 1 at this threshold.

Mechanistic analysis pipeline: for each saved checkpoint we extract attention patterns at the SEP token across all layers and heads (Appx. E.3). For each (layer, head) we sum the SEP-row attention $A_{\text{SEP},j}$ over the relevant inputs R and over the trap positions T,

$$\text{rel-attn} = \sum_{j \in R} A_{\text{SEP},j}, \qquad \text{trap-attn} = \sum_{j \in T} A_{\text{SEP},j}, \tag{2}$$

and classify the head **mis-bound** if trap-attn $> \theta \cdot$ rel-attn, default $\theta{=}1$; Tab. 4 applies it per head and per seed. Robustness across $\theta \in [0.5, 2.0]$ is in Appx. A.7.

Causal interventions: Appx. E.1 defines zero-ablation, weight and embedding transplantation, and the Xavier-init reset. *SoftBind* divides the L0 attention logits by a temperature that falls to 1 over the first T_{warm} steps,

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{\text{head}}}\,\tau(t)}\right), \qquad \tau(t) = 1 + (\tau_0 - 1)\left(1 - \min\left(1, \tfrac{t}{T_{\text{warm}}}\right)\right), \tag{3}$$

so from step T_{warm} the expression is the baseline's exactly; the *held* arm keeps $\tau{=}\tau_0$ for the whole run instead.

3 Results

3.1 Two regimes in math grokking

Rates follow the grouped-split protocol of §2, at $n{=}32$ except sum3 and prodsum, measured at $n{=}122$ and $n{=}248$ in §3.5; the three no-grok floor tasks are read on the row-wise cohorts of Appx. C, which the grouped split can only lower. Two regimes emerge, and a single seed cannot tell them apart. A **no-grok floor** at $H \le k$: det, dot and poly never grok. A **stochastic regime** at $H > k$, where the same architecture and setting generalises on some seeds and plateaus on others: deriv $1/32$, disc $1/32$, disc_pad $1/32$, prodsum $15/248$ and sum3 $32/122$ (Fig. 2, left).

Inside that regime, sum3 at 26.2% sits outside deriv's interval at 3.1% (Fig. 2, left), and the difference tracks role symmetry: where several heads can redundantly cover the same variables, more

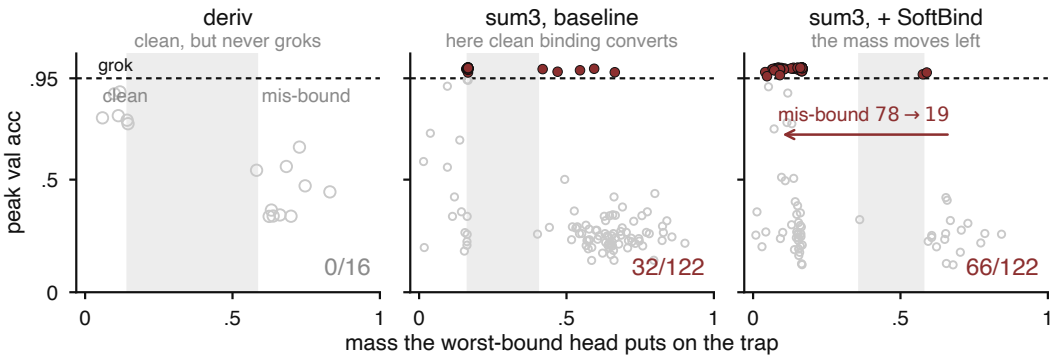


Figure 3: One dot per run at 50k: the trap mass of its worst-bound L0 head against the accuracy it reached, filled where the run groks. The quantity is bimodal with an empty band, and which side a run lands on gates the outcome.

assignments work, which is also what makes sum3's failures causally distributed rather than single-locus (§3.4).

Extending training four-fold to 200k unlocks no failed seed of the no-grok floor tasks (Appx. C).

3.2 Head specialisation, and the mis-bound head failure mode

Checkpointing deriv across the transition (200k steps), the grokked seed develops clean L0 specialisation: each relevant input (a, b, x_0) is the dominant target of at least one head, and the irrelevant c is suppressed across all heads (Appx. A.1).

On the checkpoint cohort the failed and grokked seeds differ by one head. At step 50k every one of the six seeds that fail the row-wise split (s2 to s4, s6 to s8) has exactly one L0 head meeting the mis-bound criterion of §2, with at most 0.23 of its attention on the relevant inputs and at least 0.64 on c, and neither seed that passes (s1, s5) has any (Tab. 4); this is the per-head granularity at which Wu et al. (2025) and Davies et al. (2023) document binding when it succeeds. Which head is mis-bound varies across seeds (H0×2, H1×1, H2×1, H3×2), so no head is predisposed to mis-bind. The classification is robust to the threshold θ, and the mis-bound head count predicts validation accuracy monotonically on a two-trap replication (Appx. A.7).

The binding predicts the outcome on the split where grokking is defined. On the row-wise split the criterion asks whether the model ignores the trap. Binding and outcome agreeing there proves little; on the grouped split they are different questions, tested over 16 seeds at 50k (Fig. 3). The classes separate completely, with no overlap between the two peak ranges (exact rank-sum p=2.5×10^{-4}): 0.337 to 0.644 for the ten mis-bound seeds against 0.748 to 0.890 for the six clean. Clean binding is worth about 0.3 in peak (max validation) accuracy on deriv but not grokking: no clean seed crosses 0.95 ($0/6$, best 0.890), the ceiling that bounds any intervention acting through this locus here (§3.5).

Empirical signature of the lock-in. The mis-bound head's SEP-row softmax saturates within 50 steps (Fig. 5), the per-head instance of entropy collapse (Zhai et al., 2023) that we call **QK pinning**; in multi-step tasks Hoffmann et al. (2024) find the reverse, attention too uniform to learn.

What decides which head is mis-bound. Holding the initialisation fixed and varying only the train/validation split separates the two candidate causes; training is full batch (Appx. E), so the split is the only data-side randomness. The split decides the assignment: across 3 initialisations × 8 splits, no run ends on the assignment it started with ($0/24$). At step 0 the dominant targets are BOS and SEP, not input positions, and heads migrate onto inputs over the first 25 to 50 steps (Fig. 5).

Whether the draw is ever revisited. At fixed initialisation no clean run acquires a mis-bound head later ($0/13$), while across seeds $3/5$ do (Appx. A.7).

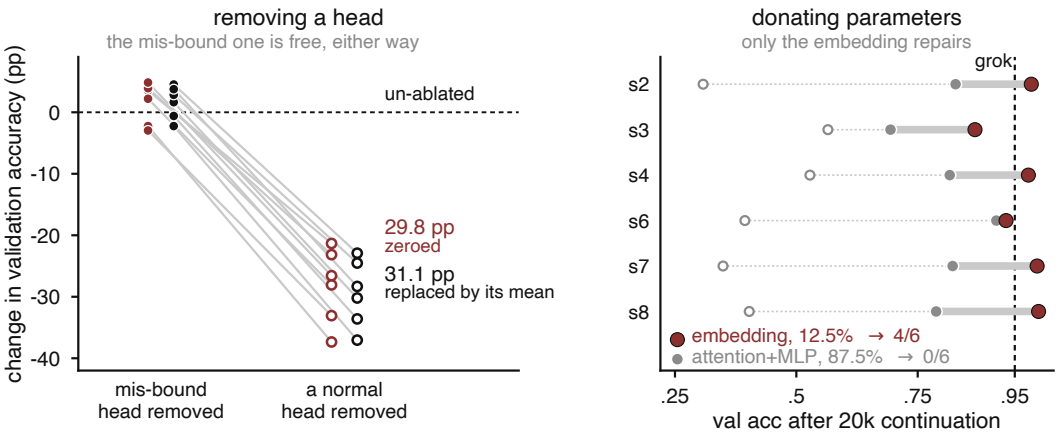


Figure 4: Failed deriv seeds. Left: removing the mis-bound head costs almost nothing and removing any other head is costly, under zeroing and under mean replacement. Right: after 20k steps of continued training, only the embedding donated from a grokked seed repairs; the legend counts runs above 0.95 at the final step (4/6), and by the max criterion of §2 all six cross (per-seed numbers in Appx. A.4 and A.5).

3.3 Causal validation: differential ablation

We zero-ablate single heads (Voita et al., 2019) and re-evaluate each failed seed of §3.2 on the rows its own run held out. Both arms share the model and the held-out rows; only the ablated head differs.

Deriv, six failed seeds. Ablating the mis-bound head moves accuracy -3.0 to $+4.8$ pp. But ablating one of the other three costs 21 to 37 pp on average. The differential, ablate-normal minus ablate-mis-bound, averages 29.8 pp (Fig. 4, left; Appx. A.4).

Grouped cohort of §3.1, n=29 failed seeds. The differential is 34.7 pp under zeroing and 42.9 under mean replacement. The mis-bound head's own mean replacement moves accuracy by a median of -0.6 pp. On all 29 seeds the flagged head is the cheaper one to replace (Tab. 6).

What the ablations say. On most seeds the mis-bound head is functionally inert. The other three carry the computation, and the absence of a fourth caps accuracy. Appx. A.2 reads which input each head settles on at 50k, and so what their specialisation is. Both grokked seeds spend one head on b and three on the (a, x_0) pair the product term $2ax_0$ needs. But no failed seed forms more than two such heads, and accuracy at 50k orders with their number. With one head spent on the trap only three remain, and three cannot be split one-and-three: the $H \leq k$ floor of §3.1 seen inside one run.

The differential is not an out-of-distribution artifact. Replacing the head with its dataset mean keeps the activation in distribution and *widens* the gap instead of closing it. With the same checkpoints and heads throughout, the differential is 31.1 pp against zeroing's 29.8 on the six seeds above, 18.5 against 17.0 on a second cohort of seven, and 42.9 against 34.7 on the grouped cohort of 29 (mean-replacement columns, Appx. A.4).

3.4 Causal rescue: train-time transplantation locates the binding

We now test *recoverability*, transplanting from a grokked donor into each failed seed at 50k and continuing 20k steps on that seed's own partition. Rescue is scored by the criterion of §2, with the final-step value quoted beside it.

The donor's embedding block rescues 6/6 failed deriv seeds. Transplanting only the token and positional embeddings, final LayerNorm and unembedding from a separately grokked donor (s9) is 12.5% of parameters, but *all* attention and MLP weights stay the failed target's. All six cross

Table 1: Grok rates on the grouped split; differential and rescue on the row-wise checkpoint cohort. Differential is [ablate a normal head] minus [ablate the mis-bound head], mean and range over 6, 5, 6 and 16 failed seeds; rescue counts seeds crossing 0.95 after the embedding transplant (controls $0/10$, Appx. A.8). The single locus holds where each relevant input needs its own head; rescue follows V against k instead.

task	rule	roles	grok	differential (pp)	rescue	failure mode
deriv	$2ax_0{+}b$	asymmetric	$1/32$	29.8 (24.9–35.1)	$6/6$	single locus
prodsum	$ab{+}c$	asymmetric	$15/248$	21.7 (13.2–27.7)	$5/5$	single locus (5 of 6)
sum3	$a{+}b{+}c$	symmetric	$32/122$	4.0 (0.2–6.1)	$5/5$	distributed
disc	$a^2{-}4b$	no trap	$1/32$	4.9 (0.2–16.4)	$0/5$	distributed, $V{=}k$

0.95, four of them still above it at the final step and the other two at 0.868 and 0.932 (Fig. 4, right; Appx. A.5). GrokTransfer (Xu et al., 2025) uses this operation to accelerate grokking from a weaker donor; here it locates the binding.

What the donated block has to be, and which half acts: a *failed* seed's block lifts all six targets by $+0.30$ to $+0.53$ and a fresh block moves them both ways, but neither crosses on any target. The rescue turns on which geometry is donated (Appx. A.5). Splitting the block locates the half: the embeddings alone clear the mis-bound head on all six targets, nothing above 0.16 left on the trap, while the final LayerNorm and unembedding alone leave it pinned on all six at 0.45 to 0.74; neither half crosses 0.95 on its own. Inside the embeddings, the attention score puts the whole trap preference in the positional term (Appx. D.1). With each row's inputs shuffled so that a learned tag, not the position, names each one, the same split puts it in the tag term on all 19 mis-bound heads (Appx. B.3.3): the head binds the input, whichever embedding names it.

Inverse transplant rescues $0/6$: the asymmetry pins the binding to the embedding. Donating instead the other 87.5%, every attention and MLP weight from the same donor, while keeping the target's own embedding, leaves mean final accuracy at 0.81 but crosses 0.95 on none of the six. Smaller attention-side blocks cross on at most one of the six, a run that touches 0.956 and ends at 0.930. And donating only the mis-bound head's own weights is bounded below grok at every intervention time (Appx. A.5).

But disc fails by a distinct, non-mis-bound mode: its ablation differential is 4.9pp against deriv's 29.8, no failed seed carries the deriv signature, the transplant that rescues deriv *lowers* its accuracy by mean 0.25 (Tab. 1, Appx. A.6).

A second rule confirms the mechanism, and Tab. 1 is its scope. We apply the identical protocol to two new single-trap tasks in deriv's regime, **prodsum** ($ab{+}c$, asymmetric roles) and **sum3** ($a{+}b{+}c$, symmetric roles); both grok stochastically. On prodsum's row-wise cohort of eight seeds, the one the checkpoint analyses use (§2), the mechanism reproduces: five of the six failed seeds carry exactly one trap-locked L0 head, its identity varying by seed (the sixth, at peak 0.72, carries none). On sum3 the signature is present but the failure is distributed, because the symmetric target lets the remaining heads cover a, b, c redundantly. So the account holds where each relevant input needs a dedicated head, and gives way under symmetric redundancy. The rescue axis is independent of the locus axis: the embedding transplant rescues sum3 as well ($5/5$ against $0/5$ matched controls at peak, Appx. A.8), and the boundary it tracks is $V{>}k$ against $V{=}k$, not single-locus against distributed. It also holds on the field's benchmark (Power et al., 2022; Nanda et al., 2023) once an operand can be ignored: on $(a{+}b) \bmod 97$ with two further inputs, $1/24$ mis-bound runs grok against $23/24$ clean ones (Tab. 14).

3.5 SoftBind: acting inside the window, without an oracle

A head's target resolves over roughly the first 250 steps by attention sharpening onto an input (Fig. 5, Appx. A.3). SoftBind's anneal (§2) runs over the first 500 steps, twice the measured window. SoftBind targets whether each head binds a relevant input, which every task shares, and is tested where

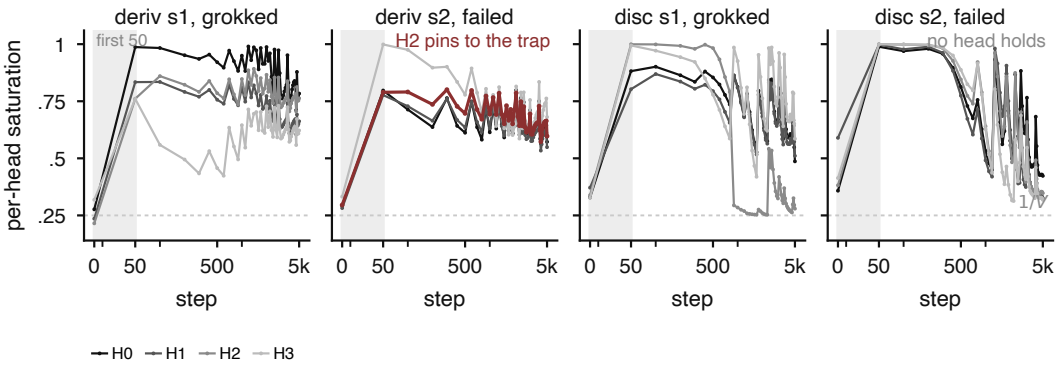


Figure 5: Per-head saturation, the maximum softmax weight in the SEP attention row averaged over the validation set, over the first 5000 steps; the dashed line is the uniform value $1/V$, shading marks the first 50 steps, and the axis is linear to step 50 and logarithmic after.

the rate has *headroom*: deriv's clean-bound runs top out at 0.890 at 50k (§3.2), while on sum3 clean binding converts at 61.4%. Tab. 2 gives it: 122 seeds in each arm on sum3 and 248 on prodsum, the arms matched by seed, with sample size and tests fixed in writing before the runs finished (protocol, mediation and controls in Appx. B).

Table 2: SoftBind on the two tasks with headroom, read at 50k on the grouped split, seeds matched across arms (Eq. 3). p is Fisher exact between the baseline and the anneal; held against anneal is 2×10^{-7} on the rate. Conditioning on binding status the anneal and the baseline do not differ, and applying the anneal's binding mix to the baseline's two conditional rates predicts 52.8% against the observed 54.1% (57.3% for the held form; 10.2% against 13.3% on prodsum). Prodsum's pre-registered endpoint, the paired counts and deriv's 0/16 in both arms are in Appx. B, B.2 and B.3.1.

	sum3, n=122 per arm				prodsum, n=248 per arm		
	baseline	+anneal	+held	p	baseline	+anneal	p
mis-bound rate	63.9%	15.6%	7.4%	7e-15	58.5%	29.8%	2e-10
grok rate, Wilson 95%	26.2 [19.2, 34.7]	54.1 [45.3, 62.7]	**85.2** [77.8, 90.4]	1e-5	6.0 [3.7, 9.7]	**13.3** [9.6, 18.1]	0.009
P(grok | clean)	61.4%	62.1%	89.4%	1.00	14.6%	19.0%	0.41
P(grok | mis-bound)	6.4%	10.5%	33.3%	0.62	0%	0%	1.00

A second task with headroom. Prodsum fails at a single locus where sum3 fails distributed (Tab. 1), and prodsum's pre-registered pool of 248 seeds per arm (Appx. B.1) gives the same picture (Tab. 2): the mis-bound rate, the pre-registered primary endpoint, halves; the grok rate more than doubles; and conditioning on binding status the arms do not differ.

The binding is upstream of the outcome. Reading the binding at 50k, the step the outcome is read at, leaves open that grokking suppresses the trap rather than the reverse. So we read it again at step 1000 in the same runs, after the anneal ends, where both arms are the same model (32 seeds per arm, none near threshold; Appx. B.3.2). No run mis-bound at 1000 groks: 0/14, against 27/50 of those clean there (p=1×10^{-4}). None recovers by 50k in either arm. By the time SoftBind stops acting the binding has already moved, 9.4% mis-bound against 34.4% (p=0.03, Tab. 13). Read as a stopping rule, ending every run mis-bound at 1000 saves 21.4% of this cohort's steps and loses none of the 27 that grok (Appx. B.3.2).

With the inputs shuffled. On sum3 with each row's inputs in a shuffled order, the held form lifts the grok rate from 3/32 to 14/32 (p=0.004) and leaves 2/32 runs mis-bound against 13/32. The anneal lowers the mis-bound rate to 4/32 but not the grok rate (1/32): on that layout a clean run seldom groks by 50k, and the binding still moves after step 1000 (Appx. B.3.3).

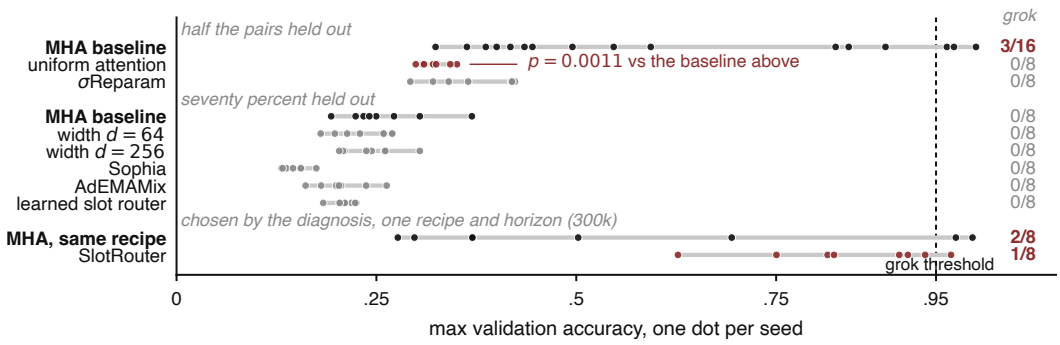


Figure 6: Interventions on deriv, one dot per seed, each group against the baseline trained beside it at the same split; no variant crosses the threshold at 50k. The last two rows put SlotRouter and MHA on one recipe and one 300k horizon: routing lifts the median seed from 0.437 to 0.863 without a rate gain (Appx. B.4).

4 Discussion

What carries the gain: six control arms separate the softening from its timing, its direction and its layer, and from early gentleness; the counts are in Tab. 11. Three change training without touching the binding, but none reproduces its shift: a learning-rate warmup over the same 500 steps, the sharp-early heat schedule of Hoffmann et al. (2024), and GrokFast (Lee et al., 2024) each leave the mis-bound rate at or above the baseline's, which is what isolates the binding from generic acceleration. And two locate it: the same softening applied at L1 rather than L0 leaves the mis-binding in place as well, and NormSoftmax, whose rows cannot sharpen, shows neither mis-binding nor grokking. But the sixth, the softening held for the whole run rather than annealed, is the only arm that goes further than SoftBind: it groks 104 of the same 122 seeds against 66, and 5 of 32, 24 of 32 and 24 of 31 when held at L1 only, at both layers, or only after step 1000 (Tab. 11). On prodsum it holds the same ordering over the 32 seeds it was run on (Appx. B.2).

Where the softening has something to act on: on sum3 no run mis-bound at step 1000 ever recovers, and that is where the gain is; deriv's binding resolves with training, so by a 200k horizon its runs cross and there is nothing left to soften (Appx. B.3.1). The two forms differ in channel (Appx. B.2).

No intervention chosen without the diagnosis moves the failure: such choices fall into three families, attention regularisation away from irrelevant variables (Zhai et al., 2023), attention restructuring (Gu & Dao, 2024; Gu et al., 2022; Pan et al., 2025) and optimizer changes (Liu et al., 2024; Pagliardini et al., 2025); we tested all three, plus a fourth derived from the mechanism. Every cell yields $0/8$ ($0/17$ for the training-time arm), ceilings 0.22 to 0.44 against the MHA baseline trained beside them at the same split, across 112 architecture, 32 optimizer and 17 direct-intervention runs (Fig. 6, Appx. B.4). Across tasks the response to an intervention follows failure mode and head margin (Tab. 1, Appx. B.4).

Two published anti-collapse fixes, one worse than nothing: σReparam (Zhai et al., 2023) and uniform attention (Yıldırım, 2026) both yield $0/8$ on deriv at the $50/50$ split. And over the same seeds uniform attention is *worse* than the baseline (exact rank-sum p=0.0011): flattening every row permanently removes the content conditioning a correct binding needs, and the useful change is later saturation, not less of it (§3.5). Prieto et al. (2025) trace a stall to saturation of the output softmax instead; here the saturated rows belong to one attention head, and its target rather than its magnitude decides the outcome.

Bypassing the draw: anchored-init vs. deterministic routing: *Anchored-QK init* biases each L0 query toward a relevant input's K-projection at initialisation, breaking the random-target symmetry without changing the architecture. It yields $0/8$ on deriv. So a bias placed at t=0 does not hold, as the fixed-initialisation measurement of §3.2 predicts. *SlotRouter* instead hard-wires each L0 head to

one input position in $\{a, b, x_0\}$ (no Q, no K), with the assignment supplied at construction time: a probe, not a prescription.

On deriv $3/8$ seeds cross the threshold cleanly under SlotRouter's own recipe, so fixing *which* input each head reads is necessary but not sufficient: in the compute-matched comparison routing lifts the median seed into the clean-bound band of §3.2 without lifting the rate (Fig. 6): on deriv the binding sets attainment, not the rate, as SoftBind's own deriv arm also shows (Tab. 12).

LLM-scale diagnostic transfer: the QK-pinning measurement runs unchanged on Pythia (Biderman et al., 2023): pinned heads reach $11/48$ at convergence. They are systematic BOS-sinks (Xiao et al., 2024; Gu et al., 2025), closer to dormant heads (Guo et al., 2024; Fu et al., 2026) than to a mis-bound head. A quarter of them are induction heads (Olsson et al., 2022) waiting on BOS until the context repeats (Appx. D.2). The diagnostic transfers where the failure does not (Appx. B.6).

5 Limitations

The account is scoped by the setting it was measured in: two-layer, four-head softmax-attention transformers on single-step rules with single-token outputs, where a first-layer head carries about one role. A $30\times$ capacity scale-up does not grok deriv either (Appx. B.5). The rate gains of §3.5 are shown on the two tasks whose correctly bound runs can grok at the horizon studied; on deriv the binding resolves with training and the softening has nothing left to act on: about half of its runs cross by 200k (Appx. B.3.1). Every rate is reported within one software stack. And paired tests pair runs within it.

6 Conclusion

We find that whether a run will fail this way is usually visible long before its accuracy moves, in what its first-layer heads attend to (§3.2). So a per-seed reading of the bindings says more than a grok rate averaged over seeds (§3.1). A stalled run can be checked for this before it is trained longer (Power et al., 2022) or moved to another optimiser (Liu et al., 2024; Pagliardini et al., 2025): does any first-layer head read an input the rule ignores? And if one does, the fix belongs to that head's layer and not to the optimiser (Tab. 11, Appx. B.4). The next setting is larger models and multi-step reasoning: there a first-layer head need not carry a single role, and a rule can leave more than one input unread.

Reproducibility Statement

All runs use deterministic kernels (Paszke et al., 2019), so a seed reproduces its trajectory exactly on one software stack (Appx. E). And the task generators, tokenizer, architecture, optimiser and training schedule are specified in §2 and Appx. E. Every rate in the paper is computed from stored per-seed run rows, and every checkpoint-level number from stored checkpoints, by scripts that were committed before their results were read.

AI Use Statement

Generative AI was used to aid and polish the writing of the paper; every task instance comes from the procedural generators of §2, and every reported number is computed from stored per-seed run artifacts. The authors take full responsibility for the content, including any text written or polished with the aid of generative AI.

References

Sarwan Ali. Critical windows of complexity control: When transformers decide to reason or memorize, 2026. URL https://arxiv.org/abs/2605.04396.

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.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Vatsal Baherwani, Zixi Chen, Shikai Qiu, Andrew Gordon Wilson, and Pavel Izmailov. Emergent capabilities arise randomly from learning sparse attention patterns, 2026. URL https://arxiv.org/abs/2606.25010.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata, 2023. URL https://arxiv.org/abs/2307.03637.

W. J. Dixon and A. M. Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, December 1946. ISSN 1537-274X. doi: 10.1080/01621459.1946.10501898. URL http://dx.doi.org/10.1080/01621459.1946.10501898.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention sink forges native moe in attention layers: Sink-aware training to address head collapse, 2026. URL https://arxiv.org/abs/2602.01203.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

Pulkit Gopalani and Wei Hu. What happens during the loss plateau? understanding abrupt learning in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=tnTM6JJuLi.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour, 2018. URL https://arxiv.org/abs/1706.02677.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yJE7iQSAep.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=78Nn4QJTEN.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms, 2024. URL https://arxiv.org/abs/2410.13835.

Kai Hidajat, Solden Stoll, and Joseph An. Grokking as structural inference: Transformers need bayesian lottery tickets, 2026. URL https://arxiv.org/abs/2605.15787.

David T. Hoffmann, Simon Schrodi, Jelena Bratulić, Nadine Behrmann, Volker Fischer, and Thomas Brox. Eureka-moments in transformers: Multi-step tasks reveal softmax induced optimization problems, 2024. URL https://arxiv.org/abs/2310.12956.

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

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.

Hong Liu, Zhiyuan Li, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3xHDeA8Noi.

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

Ziming Liu, Yizhou Liu, Eric J. Michaud, Jeff Gore, and Max Tegmark. Physics of skill learning, 2025. URL https://arxiv.org/abs/2501.12391.

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

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

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Bridging lottery ticket and grokking: Understanding grokking from inner structure of networks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=eQeYyup1tm.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. Grokking of hierarchical structure in vanilla transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 439–448, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.38. URL https://aclanthology.org/2023.acl-short.38/.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

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. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

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

Matteo Pagliardini, Pierre Ablin, and David Grangier. The adEMAMix optimizer: Better, faster, older. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jj7b3p5kLY.

Yuqi Pan, Yupeng Feng, Jinghao Zhuang, Siyu Ding, Han Xu, Zehao Liu, Bohan Sun, Yuhong Chou, Xuerui Qiu, Anlin Deng, Anjie Hu, Shurong Wang, Peng Zhou, Man Yao, Jibin Wu, Jian Yang, Guoliang Sun, Bo Xu, and Guoqi Li. Spikingbrain: Spiking brain-inspired large models, 2025. URL https://arxiv.org/abs/2509.05276.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

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.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TvfkSyHZRA.

Kiran Raja, Arav Maheria, Andrew Bae, and Alan Sun. Understanding knowledge acquisition and release in language models via circuits. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026. URL https://openreview.net/forum?id=MYgw7DBJye.

M. Sagitova, O. Duranthon, and L. Zdeborová. Specialization of softmax attention heads: insights from the high-dimensional single-location model. arXiv:2603.03993, 2026. URL https://arxiv.org/abs/2603.03993.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Alan Sun. Circuit stability characterizes language model generalization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9025–9040, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.442. URL https://aclanthology.org/2025.acl-long.442/.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lucky Verma. Weight decay regimes in grokking transformers: Cheap online diagnostics, 2026. URL https://arxiv.org/abs/2605.20441.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization, 2024. URL https://arxiv.org/abs/2405.15071.

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

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

Yiwei Wu, Atticus Geiger, and Raphaël Millière. How do transformers learn variable binding in symbolic programs? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kVtyv7bpnw.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model, 2025. URL https://arxiv.org/abs/2504.13292.

Alper Yıldırım. The geometric inductive bias of grokking: Bypassing phase transitions via architectural topology, 2026. URL https://arxiv.org/abs/2603.05228.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Xingyu Zhao, Darsh Sharma, Rheeya Uppaal, and Yiqiao Zhong. Shattered compositionality: Counterintuitive learning dynamics of transformers for arithmetic, 2026. URL https://arxiv.org/abs/2601.22510.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

A MECHANISTIC EVIDENCE: DERIV AND DISC

Deriv and disc both grok stochastically, but for mechanistically *distinct* reasons. This section pins down each in turn. On deriv the failure is discrete and localised: exactly one L0 head binds the irrelevant trap c and stays there, so ablating that head barely moves accuracy while ablating a correctly-bound head is costly (§A.2, per seed in Tab. 5). On disc there is no trap to mis-bind, and failure is instead *distributed*. All four heads under-specialise with no single load-bearing head, the qualitative inverse of the deriv signature (§A.6). Both diagnostics are robust to the threshold and replicate on a held-out task (§A.7).

A.1 DERIV: POSITION ATTRIBUTION IN THE GROKKED SEED

Position attribution across the transition. We measure per-input-position attribution (Appx. E.4 defines the symbols) for the grokked deriv seed s1 at 200k by ablating each input position via substitution with the padding token and recording the logit drop on the correct output token. Aggregated over 100 validation samples, the position effect (logit drop) trajectory across the grokking transition is given in Tab. 3.

Table 3: Position attribution trajectory across the deriv grokking transition. Effects on relevant inputs (a, b, x_0) grow from $\sim$2 to $\sim$8 across the transition; effect on c stays at $\sim$0 throughout. The model learns variable suppression early (by step 10k) and develops increasingly sharp routing of relevant variables as accuracy improves. val_acc is the reading at that checkpoint itself: this run first crosses 0.95 at step 45k (maximum 0.966 over the first 50k), reads 0.86 at the 50k checkpoint, and its late checkpoints climb to 0.98.

step	val. acc.	BOS	a	b	c	x_0
10000	0.36	−0.01	2.08	2.50	−0.32	2.72
20000	0.60	−0.01	4.31	4.32	−0.02	4.60
30000	0.86	−0.01	6.59	6.97	−0.04	6.16
50000	0.86	−0.00	6.14	7.02	−0.01	6.08
100000	0.91	−0.00	7.36	7.67	−0.04	7.38
200000	0.98	0.00	7.83	9.07	−0.01	7.82

The clean separation between relevant and irrelevant variables emerges before grokking is complete. Even at validation accuracy 0.36 (step 10k), the model has correctly identified c as irrelevant.

At the attention level, the same seed shows all four L0 heads stably attending to the relevant inputs (a, b, x_0) and suppressing the irrelevant c from early training: per-head relevant-attention stays ≥ 0.83 and c-attention ≤ 0.08 throughout.

A.2 PER-HEAD ATTENTION AT STEP 50K, FAILED SEEDS

Table 4 is the static fingerprint of the failure: in every failed seed exactly one head has flipped from attending its relevant input to attending c (the bold ***rel/c*** cells), and *which* head it is varies seed to seed while both grokked seeds have none. When it happens is read off the saturation trajectory of Fig. 5, measured in Appx. A.3.

A.3 DIRECT QK PINNING MEASUREMENT AT EARLY TRAINING STEPS

To verify that QK pinning is a fast early-training transition rather than a gradual late-training pattern, we directly measure per-head saturation (max softmax row at the SEP position, averaged over the validation set) every 100 training steps for the first 5000 steps, with additional reads at steps 50, 200 and 300, on two deriv seeds. The grokked-baseline seed s1 (first panel of Fig. 5) shows all four heads jumping from saturation ~ 0.25 at init to saturation ≥ 0.76 within 50 training steps, with c-attention ≤ 0.08 across all four heads, so correct binding to relevant inputs is established at once. The failed seed s2 (second panel) shows the mis-bound head H2 pinning to irrelevant input c at step 50 (c-attention 0.41, saturation 0.79), and remains there for the rest of training; the other three heads reach saturation 0.78 to 0.999 with c-attention ≤ 0.06. Each seed's binding is therefore decided before the optimiser has seen a full 1% of the standard 50000-step training horizon.

Table 4: Attention from SEP, summed over relevant inputs (a, b, x_0). Non-mis-bound cells show rel-sum only; mis-bound cells (shaded, bold) show *rel/c* where c exceeds rel. G=grokked, F=failed. Across all six failed seeds the mis-bound head is exactly one, but its identity (H0, H1, H2, or H3) is set by the seed; both grokked seeds have zero mis-bound heads. Lower block: the dominant SEP-row target of each head, $a{+}x_0$ marking a head that splits its mass between a and x_0 (both ≥ 0.25). Both grokked seeds spend one head on b and three on the pair; no failed seed forms more than two pair heads, and accuracy at 50k orders with their number: 0.31 to 0.41 with none, 0.53 with one, 0.58 with two, 0.96 with three.

Head	s1 (G)	s5 (G)	s2 (F)	s3 (F)	s4 (F)	s6 (F)	s7 (F)	s8 (F)
L0H0	0.93	0.72	0.94	0.74	**0.23/0.68**	0.86	0.90	**0.14/0.83**
L0H1	0.83	0.81	0.95	**0.23/0.64**	0.75	0.83	0.88	0.93
L0H2	0.91	0.69	**0.10/0.89**	0.80	0.68	0.85	0.90	0.90
L0H3	0.92	0.77	0.93	0.66	0.68	**0.19/0.74**	**0.20/0.76**	0.91
L0H0 reads	b	$a{+}x_0$	x_0	$a{+}x_0$	c	a	x_0	c
L0H1 reads	$a{+}x_0$	$a{+}x_0$	b	c	$a{+}x_0$	x_0	b	b
L0H2 reads	$a{+}x_0$	$a{+}x_0$	c	b	b	b	a	a
L0H3 reads	$a{+}x_0$	b	a	$a{+}x_0$	b	c	c	x_0

Disc shows qualitatively different early-training dynamics. Repeating the same trajectory measurement on disc reveals a different pattern (the two disc panels of Fig. 5): all heads briefly saturate within 50 steps as on deriv, but saturation then *decreases* over the next 5000 steps rather than stabilising. On disc s2 every head's saturation falls from ~ 1.0 at step 50 to ~ 0.4 at step 5000; on disc s1 head $H2$'s saturation collapses around step 1000 to ~ 0.3. The qualitative dynamics, an early lock-in followed by gradual unpinning rather than persistent commitment, visually confirms the under-specialisation framing for disc and contrasts cleanly with the persistent-pin pattern of deriv.

A.4 Per-Seed Causal Evidence Tables

Table 5 is the per-seed breakdown behind the differential-ablation claim of §3.3: for each failed seed it contrasts ablating the mis-bound head against the mean over the three normal heads, a near-zero effect against a large drop, and the rightmost column is the differential, the actual quantity of interest.

Table 5: Differential ablation, all six failed deriv seeds, each scored at step 50k on the 2500 rows its own run held out (row-wise $50/50$ split). "Mean ablate normal" averages over zero-ablating each of the three non-mis-bound heads individually.

Seed	Original	Ablate mis-bound	Mean ablate normal	Range (normals)	Differential
s2 (mis-bound=L0H2)	0.308	0.344 (+3.6)	0.095 (−21.3)	0.032, 0.139	**24.9 pp**
s3 (mis-bound=L0H1)	0.565	0.543 (−2.2)	0.192 (−37.4)	0.022, 0.500	**35.1 pp**
s4 (mis-bound=L0H0)	0.528	0.499 (−3.0)	0.198 (−33.1)	0.012, 0.383	**30.1 pp**
s6 (mis-bound=L0H3)	0.394	0.416 (+2.2)	0.113 (−28.1)	0.026, 0.179	**30.3 pp**
s7 (mis-bound=L0H3)	0.349	0.388 (+3.9)	0.118 (−23.1)	0.023, 0.196	**27.1 pp**
s8 (mis-bound=L0H0)	0.403	0.451 (+4.8)	0.137 (−26.6)	0.028, 0.222	**31.4 pp**

A.5 Causal Rescue per Seed

Table 7 walks up a ladder of train-time transplants from a grokked donor into each failed seed (from the smallest attention-side graft to the largest) and then inverts the protocol; it has four rungs, from a single L0 head at $\sim$5% of parameters through all four L0 heads and all eight L0+L1 heads to the full L0+L1 layer at 87.5% of the donor. Replacing attention and MLP weights lifts mean final val acc only from 0.45 to 0.81 but no seed clears the 0.95 threshold at the final step (one L0+L1 seed touches 0.956 mid-run). But inverting the graft is decisive. Donating only the token and positional embeddings, final LayerNorm (Ba et al., 2016), and unembedding (12.5% of parameters), while keeping *every* attention and MLP weight from the failed target, crosses 0.95 on $\mathbf{6/6}$ seeds by the max-validation criterion of §2, $4/6$ of them still above it at the final step (mean final $\mathbf{0.96}$, mean max-in-run 0.998). What pins the binding to the embedding geometry, rather than to any learned

Table 6: The differential ablation on the grouped deriv cohort of §3.1, seeds 101 to 132 (Fig. 3 plots a separate cohort, seeds 1 to 16), one row per failed seed (n=29; the cohort's grokked seed and two seeds whose held-out rows the memorisation gate could not verify are excluded). Each checkpoint is scored on its own held-out rows; mb is the mis-bound head under the θ=1 criterion, "normal" the mean over the other three heads, and Diff. is ablate-mis-bound minus ablate-normal in pp. Mean replacement of the mis-bound head moves accuracy by a median of -0.6 pp ($24/29$ within 3 pp) against a median -50.7 pp for a normal head.

Seed	Orig.	Mis-bound	Zero mb	Zero normal	Diff.	Mean mb	Mean normal	Diff.
s101	0.489	H0	0.277	0.145	13.2	0.493	0.160	**33.3**
s102	0.801	H1	0.542	0.130	41.2	0.773	0.164	**60.9**
s103	0.446	H1	0.464	0.106	35.8	0.452	0.100	**35.3**
s104	0.767	H0	0.674	0.240	43.4	0.712	0.215	**49.7**
s105	0.396	H3	0.369	0.243	12.6	0.392	0.233	**15.9**
s106	0.519	H2	0.525	0.174	35.1	0.502	0.183	**31.9**
s107	0.804	H3	0.440	0.187	25.3	0.775	0.176	**59.9**
s108	0.575	H3	0.479	0.142	33.7	0.572	0.149	**42.3**
s109	0.446	H1	0.441	0.154	28.7	0.440	0.159	**28.2**
s110	0.810	H1	0.655	0.371	28.4	0.651	0.304	**34.7**
s111	0.693	H3	0.325	0.123	20.3	0.671	0.127	**54.4**
s112	0.695	H2	0.647	0.230	41.7	0.692	0.173	**51.9**
s113	0.912	H0	0.886	0.241	64.5	0.909	0.235	**67.4**
s114	0.837	H1	0.772	0.251	52.1	0.834	0.232	**60.2**
s115	0.456	H2	0.439	0.146	29.3	0.442	0.158	**28.4**
s116	0.798	H3	0.598	0.144	45.4	0.796	0.150	**64.6**
s117	0.322	H3	0.340	0.116	22.4	0.330	0.098	**23.3**
s118	0.524	H0	0.515	0.139	37.6	0.537	0.097	**44.0**
s119	0.689	H0	0.394	0.165	23.0	0.665	0.140	**52.4**
s121	0.561	H1	0.420	0.153	26.7	0.560	0.132	**42.8**
s122	0.845	H3	0.840	0.205	63.6	0.838	0.153	**68.5**
s123	0.363	H0	0.381	0.107	27.4	0.360	0.088	**27.3**
s125	0.891	H3	0.854	0.222	63.2	0.781	0.200	**58.1**
s126	0.672	H0	0.619	0.153	46.6	0.678	0.122	**55.7**
s127	0.664	H2	0.604	0.191	41.3	0.611	0.136	**47.6**
s129	0.469	H3	0.388	0.155	23.3	0.466	0.165	**30.1**
s130	0.577	H1	0.551	0.145	40.7	0.571	0.113	**45.8**
s131	0.831	H0	0.455	0.269	18.6	0.294	0.267	**2.7**
s132	0.360	H3	0.294	0.071	22.3	0.343	0.080	**26.3**

attention weight, is that the large graft in the wrong place fails while the small graft in the right place succeeds.

A.6 DISC: DISTRIBUTED UNDER-SPECIALISATION

Disc groks as rarely as deriv on the grouped split ($1/32$ each) and more rarely on the row-wise cohort ($1/16$ against $3/16$, Appx. E.2), but it fails a different way: with only relevant inputs and no trap token, there is nothing for a head to *mis*-bind, so the deriv mis-bound-head signature (§3.2) cannot appear. The three probes below (a single-head transplant null, a differential-ablation spread contrast, and a per-head attention grid) converge on the same picture: disc failure is spread across all four under-specialised heads rather than concentrated in one wrong target.

Disc has 2 input variables (a, b) and 4 heads available per L0. The grokked s1 model develops attention specialisation where two heads attend predominantly to a (one with $\sim$0.85 weight) and two to b; the remaining residual stream computes a^2 via the MLP and subtracts $4b$. With no irrelevant input to lock onto, failed disc seeds plateau through under-specialisation rather than the discrete mis-bound head signature.

Single-head transplant on disc. We tested whether single-head transplant from grokked disc s1 rescues failed disc seeds across n=12 (target seed, head) combinations: 3 target seeds (s3, s5, s7) $\times$

Table 7: Causal rescue on the failed seed's own train/val partition, from single-head up to the 12.5% embedding-block transplant; each block names its donor (s9 peaks at 1.000). "Δ" is final val acc minus the step-50k baseline, brackets give the peak, and shaded cells cross 0.95 at the final step. Three blocks control the single-head rung: a second donor (s5), a Xavier-init reset with no donor, and a compute-matched extension that transplants nothing.

Seed	Mis-bound head	Initial (step 50k)	After 20k$^+$ intervention	Δ
Single mis-bound head, donor s1, ∼5% of parameters (n=6, mean 0.45):				
s2	L0H2	0.308	0.387	+0.079
s3	L0H1	0.565	0.612	+0.047
s4	L0H0	0.528	0.337	−0.192
s6	L0H3	0.394	0.457	+0.063
s7	L0H3	0.349	0.423	+0.074
s8	L0H0	0.403	0.468	+0.065
Single mis-bound head, donor s5, ∼5% (n=3, mean 0.56):				
s3 (donor s5)	L0H1	0.565	0.663	+0.098
s4 (donor s5)	L0H0	0.528	0.565	+0.037
s6 (donor s5)	L0H3	0.394	0.448	+0.054
Single mis-bound head, Xavier reset, no donor (n=3, mean 0.53):				
s2 + xavier L0H2	L0H2	0.308	0.409	+0.100
s3 + xavier L0H1	L0H1	0.565	0.613	+0.048
s4 + xavier L0H0	L0H0	0.528	0.561	+0.033
Vanilla-extension control, no intervention (n=3, mean 0.51):				
s2 (vanilla 20k)	n/a	0.308	0.367	+0.059
s3 (vanilla 20k)	n/a	0.565	0.579	+0.014
s4 (vanilla 20k)	n/a	0.528	0.568	+0.039
Whole L0, 4 heads (n=6, mean final 0.65):				
s2 (donor s9)	n/a	0.308	0.874 [max 0.874]	+0.566
s3 (donor s9)	n/a	0.565	0.369 [max 0.838]	−0.196
s4 (donor s9)	n/a	0.528	0.833 [max 0.850]	+0.305
s6 (donor s9)	n/a	0.394	0.466 [max 0.645]	+0.072
s7 (donor s9)	n/a	0.349	0.535 [max 0.535]	+0.186
s8 (donor s9)	n/a	0.403	0.840 [max 0.855]	+0.437
Whole L0+L1 attention, 8 heads (n=6, mean final 0.72):				
s2 (donor s9)	n/a	0.308	0.736 [max 0.752]	+0.428
s3 (donor s9)	n/a	0.565	0.746 [max 0.839]	+0.181
s4 (donor s9)	n/a	0.528	**0.930** [max **0.956**]	+0.402
s6 (donor s9)	n/a	0.394	0.862 [max 0.896]	+0.468
s7 (donor s9)	n/a	0.349	0.635 [max 0.776]	+0.286
s8 (donor s9)	n/a	0.403	0.384 [max 0.828]	−0.019
Full L0+L1 layer, attn + MLP + LN (n=6, mean final 0.81):				
s2 (donor s9)	n/a	0.308	0.828 [max 0.850]	+0.520
s3 (donor s9)	n/a	0.565	0.694 [max 0.778]	+0.129
s4 (donor s9)	n/a	0.528	0.816 [max 0.902]	+0.288
s6 (donor s9)	n/a	0.394	0.912 [max 0.912]	+0.518
s7 (donor s9)	n/a	0.349	0.822 [max 0.822]	+0.473
s8 (donor s9)	n/a	0.403	0.788 [max 0.790]	+0.385
*Embedding block, 12.5% donor (n=6, **6/6 grok** by max, 4/6 at the final step):*				
s2 (donor s9)	n/a	0.308	**0.984** [max **1.000**]	+0.676
s3 (donor s9)	n/a	0.565	0.868 [max 0.991]	+0.303
s4 (donor s9)	n/a	0.528	**0.978** [max **1.000**]	+0.450
s6 (donor s9)	n/a	0.394	0.932 [max 0.998]	+0.538
s7 (donor s9)	n/a	0.349	**0.996** [max **1.000**]	+0.647
s8 (donor s9)	n/a	0.403	**0.999** [max **1.000**]	+0.596

4 head positions (H0 to H3). Every combination fails; post-transplant validation accuracy stays at or below baseline (Tab. 8).

Table 8: Single-head transplant from grokked disc s1 into three failed seeds, all four head positions each. Validation accuracy before the transplant, and the range over the four positions after it.

target seed	before	after (over H0 to H3)
s3	0.187	0.18 to 0.20
s5	0.089	0.05 to 0.18
s7	0.147	0.07 to 0.19

The 12-cell null falsifies the single-point hypothesis of §3.2 for disc.

Differential ablation on disc: a distributed failure. As the direct positive complement, we apply the same per-head zero-ablation protocol used for deriv (§3.3) to all n=16 disc failed seeds at the final 50k checkpoint. For each seed we ablate L0H$_h$ for $h \in \{0, 1, 2, 3\}$ in turn and report Δ_h = baseline − ablated.

Table 9: Disc differential ablation, n=16 failed seeds (seeds 17–32). Δ_h = baseline val acc − val acc after zero-ablating L0Hh in pp. "spread" = $\max_h \Delta_h - \min_h \Delta_h$, the per-seed disagreement across heads (large means one head is much more load-bearing; small means the four heads contribute comparably). Mean spread is 4.9pp (range 0.2 to 16.4pp) versus deriv's 29.8pp (range 24.9 to 35.1pp), a 6.1× separation. No disc seed reaches deriv's minimum spread of 24.9pp.

seed	baseline	Δ_0	Δ_1	Δ_2	Δ_3	spread
s17	0.217	+0.002	+0.002	+0.007	+0.000	0.007
s18	0.227	+0.009	+0.002	+0.002	+0.000	0.009
s19	0.154	+0.030	+0.000	+0.091	+0.051	0.091
s20	0.229	+0.000	−0.002	+0.005	+0.002	0.007
s21	0.243	+0.033	+0.000	+0.164	+0.056	0.164
s22	0.173	+0.049	+0.047	+0.044	+0.040	0.009
s23	0.086	−0.007	+0.002	+0.012	+0.016	0.023
s24	0.229	−0.002	+0.007	+0.002	+0.002	0.009
s25	0.199	+0.096	+0.145	+0.037	+0.026	0.119
s26	0.175	+0.040	+0.037	+0.115	+0.033	0.082
s27	0.143	+0.040	+0.040	+0.086	+0.040	0.047
s28	0.143	+0.019	+0.009	+0.028	+0.023	0.019
s29	0.152	−0.005	+0.002	+0.000	−0.007	0.009
s30	0.140	+0.007	+0.051	+0.056	+0.042	0.049
s31	0.194	+0.002	+0.002	+0.000	+0.002	0.002
s32	0.252	+0.096	+0.131	+0.009	−0.009	0.140
mean	0.185	+0.026	+0.030	+0.041	+0.021	**0.049**

Tab. 9 separates the two tasks with no overlap in spread. A Welch two-sample t-test (Welch, 1947) on the spread distributions yields $t \approx 12.7$, $p \approx 5{\times}10^{-9}$. The two failure modes are statistically distinct in their per-head ablation signature: deriv has one near-inert head and three load-bearing heads; disc has all four heads contributing partially with no clean separation. Combined with the n=12 transplant null, this is direct causal evidence that disc's failure mode is distributed under-specialisation and not deriv's mis-bound head.

Per-head SEP-row attention on disc. Reading the L0 SEP-row per-head attention across six failed disc checkpoints (a six-seed subset, disjoint from the 16-seed sweep behind Tab. 9) gives the qualitative inverse of deriv's mis-bound grid: only $4/24$ (seed, head) cells exceed a 0.5 specialisation threshold, $15/24$ remain near the uniform baseline of 0.25, and three of the six seeds (s2, s3, s8) show *all* four heads at exactly 0.25 across the input positions. No disc failed seed exhibits the deriv-style profile of one sharp mis-bound head plus three clean-bound specialists.

A.7 Diagnostic robustness and cross-task replication

Is the binding revisited after step 1k? Holding the initialisation fixed and varying only the split, the mis-bound heads of 16 runs are mis-bound continuously from step 250 and no run clean at 1k acquires one later ($0/13$), at every $\theta \in [0.5, 1.5]$. Varying the seed, so that the initialisation moves as well as the split, 3 of the 5 runs clean at 1k acquire a mis-bound head afterwards. The pinning also follows the irrelevant input rather than the slot beside SEP: deriv's trap sits third of four, with x_0 adjacent to SEP. Its mis-bound heads sit on c (Appx. A.2), while on sum3 and prodsum, whose trap is the last slot, they sit on d.

Threshold robustness. The classification is insensitive to the threshold θ in c-attn $> \theta \times$ rel-attn: across $\theta \in \{0.5, 0.7, 0.9, 1.0, 1.5, 2.0\}$, every grokked seed ($n$=2) has zero mis-bound heads and every failed seed (n=6) exactly one. Heads are bimodal, mis-bound heads at rel-attn ≤ 0.23 with c-attn ≥ 0.64 and every other head at rel-attn ≥ 0.66 (Tab. 4).

Cross-task replication with two traps. On deriv with a second ignored input d, $(a, b, c, d, x_0) \mapsto 2ax_0+b$ (k=3, V=5; Appx. E.4), the diagnostic finds 0 to 2 mis-bound heads per seed across n=8, and the count orders validation accuracy: 0 mis-bound $\rightarrow$ 34% (n=1), 1 $\rightarrow$ mean 24% (n=4), 2 $\rightarrow$ mean 18% (n=3). Each mis-bound head removes one specialisation slot.

A.8 Second-task replication: prodsum and sum3

The same protocol on two new rules. To test whether the mis-bound-head mechanism (§3.2) is specific to deriv's construction, we ran the same diagnostic and differential-ablation protocol on two new single-trap rules in deriv's regime (k=3, V=4, one irrelevant input at position 4; Appx. E.4): **prodsum** ($ab+c$, asymmetric variable roles) and **sum3** ($a+b+c$, symmetric roles). Both grok stochastically, $2/8$ each on a row-wise $50/50$ split (Appx. E.2) at 50k steps; the cohort predates the grouped split, and every number in this section is scored on the rows each run held out.

Prodsum: one head on the trap, and inert. The per-head SEP-row attention audit of §3.2 at step 50k finds both grokked seeds with zero mis-bound heads and five of the six failed seeds with exactly one head locked onto the trap d (Tab. 1), its identity varying by seed; the sixth failed seed, at peak 0.72, carries none. Differential zero-ablation of single heads (Michel et al., 2019; Tab. 1) shows that head is inert while the normal heads carry the computation: per-seed differentials $25.0, 21.6, 13.2, 27.7, 20.8$ pp (mean $\mathbf{21.7}$ pp over n=5), against deriv's 29.8 pp and disc's 4.9 pp.

Sum3 fails distributed: the same head, and not inert. The attention audit again flags a single mis-bound head in $5/6$ failed seeds, but ablation (Tab. 1) shows it is *not* inert: differentials $0.3, 5.9, 5.7, 5.7, 0.2, 6.1$ pp (mean $\mathbf{4.0}$ pp), disc-like. Because $a+b+c$ is symmetric, the three non-trap heads redundantly cover a, b, c, so no single head is load-bearing. The mechanism tracks binding *asymmetry*, holding where each relevant input needs a dedicated head (deriv, prodsum) and giving way to distributed failure under symmetric redundancy (sum3, disc; Tab. 1).

The embedding transplant rescues both new rules, dissociating rescue from causal locus. We applied the embedding-block transplant of §3.4 to n=5 failed seeds per task, each against a compute-matched vanilla extension (the same 20k continuation from the same step-50k checkpoint, no transplant). The two conditions separate: every transplanted run peaks above 0.97 (prodsum ≥ 0.983, sum3 ≥ 0.972) and no control run exceeds 0.46 (prodsum ≤ 0.462, sum3 ≤ 0.271), all 10 matched pairs moving the same way (two-sided sign test, Dixon & Mood, 1946, $p = 0.002$). We quote peak because the continuation runs at constant LR under wd=1.0 and oscillates once past threshold.

What the two axes separate. sum3 fails *distributed*, yet the embedding transplant rescues it as it does the single-locus rules, while the same protocol *lowers* disc accuracy by 0.25 ($0/5$, Tab. 1). The two axes are independent (§3.4, Tab. 1), and disc_pad (a^2-4b with an irrelevant third input, k=2, V=3) is the cell that holds the rule fixed while raising V above k.

Head targets at initialisation. Across 500 random initialisations, each L0 head's dominant SEP-row target among the V=4 input positions (BOS and SEP excluded) is uniform over the four (χ^2 test $p \in \{0.42, 0.10, 0.34, 0.22\}$ for $H0$–$H3$; per-position frequencies within $[0.22, 0.30]$) and the four heads are mutually independent (pairwise mutual information ≤ 0.027 bits).

B Interventions: SoftBind, and What Does Not Move the Failure

B.1 SoftBind: the module and the protocol

Its forward pass is the baseline's with the divisor τ of Eq. 3 at L0, a non-persistent buffer, so it does not enter the checkpoint.

Pre-registration. The sample size and all three tests were fixed in writing before the runs finished, at the 80%-power point for the effect seen in a smaller pilot. The pilot is not reported as evidence and no interim analysis was run.

The powered prodsum cohort. Two smaller prodsum cohorts preceded it: the pilot, seeds 1 to 32, which the headline excludes, and seeds 33 to 64, which set the power and open the pool. Fixed in writing before seeds 65 to 280 ran: the pool, seeds 33 to 280 (n=248 per arm, power 0.80 at the second cohort's observed difference), and the primary endpoint, the mis-bound rate (two-sided Fisher exact test, Fisher, 1935). Read once at completion (Tab. 2): the mis-bound rate falls by 28.6 pp (95% CI $[20.0, 36.6]$); the grok rate is descriptive, since resolving the second cohort's rate difference at 80% power would need n=1157 per arm.

Hosts. The 244 baseline and anneal runs on sum3 were spread over two hosts on one software stack under deterministic kernels, the baseline arm mostly on one and the treated arm mostly on the other. Seed n is the same initialisation and the same split in each arm, so the McNemar (1947) tests of Tab. 11 are paired. On one stack a run is a function of its seed (we ran the first sixteen baseline seeds on both, and their peak accuracies agree on every seed to four decimals).

B.2 What carries the gain

§4 reads these results; the strata are in Tab. 10 and the control counts in Tab. 11.

Mediation. Conditioning on binding status, the arms are indistinguishable (Tab. 10). Applying the intervention's binding distribution to the baseline's two stratum rates predicts 52.8% against 54.1% observed.

Where this sits among accounts of the transition. Explanations of when a run transitions include circuit efficiency (Varma et al., 2023), a shift from lazy to rich training (Kumar et al., 2024), and staged head alignment in a solvable model (Sagitova et al., 2026), and the circuit a converged run implements varies across settings (Zhong et al., 2023). The quantity here is which runs transition at all, read per head, per seed, before the outcome exists (§3.5).

Table 10: sum3 at 50k, n=122 per arm, seeds matched. The intervention moves the binding rate and leaves both conditional grok rates where they were, which is what mediation by the binding predicts and what lets the account state the effect's size rather than only its sign.

stratum	baseline	+ SoftBind	Fisher p
P(grok | clean)	$27/44 = 61.4\%$	$64/103 = 62.1\%$	1.00
P(grok | mis-bound)	$5/78 = 6.4\%$	$2/19 = 10.5\%$	0.62
mis-bound rate	$78/122 = 63.9\%$	$19/122 = 15.6\%$	7.1×10^{-15}
grok rate	$32/122 = 26.2\%$	$66/122 = \mathbf{54.1}\%$	1.4×10^{-5}

Controls on the window. Tab. 11 gives the counts (on seeds 1 to 32 the anneal stands at $19/32$ against the baseline's $11/32$, discordant 10:2, p=0.039). Two controls, pre-registered before any of their runs existed (exact McNemar (1947) against SoftBind as the primary endpoint, against baseline as the secondary, the mis-bound rate descriptive), and three further arms (§4): the same softening applied at L1 with L0 standard, the heat schedule (τ_0=0.5 rising to 1 by 25k), and NormSoftmax, whose per-row standardisation keeps attention from sharpening at any point (the last two from Hoffmann et al., 2024). *LR warmup* (Goyal et al., 2018) keeps ordinary attention and warms the learning rate linearly over the same 500 steps. *Held* keeps SoftBind's τ_0=4 for the whole run instead of annealing it to 1 (§2, Appx. E.4). Both run on seeds 1 to 32 of the sum3 cohort on one host and one stack (every run pairs with the baseline and SoftBind runs of its seed); the warmup was then run to the headline's n=122 under a second pre-registration (seeds 33 to 122, Fisher exact against the anneal as the primary endpoint) and enters the table as rates. Row by row, LR warmup leaves the binding where the baseline has it; holding the softening groks nine seeds the anneal does not; sharpening early and softening the other layer leave the binding and the rate at the baseline's; rows that cannot sharpen show no mis-binding and no grokking (§4, Tab. 11).

Which half of the block. Splitting the rescue block of §3.4 (Tab. 7) and donating each half alone separates them completely (same donor, same six targets, same protocol). The embedding half (token and positional embeddings) clears the mis-bound head on $6/6$ (§3.4) and reaches 0.749 to

Table 11: The control arms around SoftBind, read at 50k on the grouped split; "vs" columns are exact McNemar where discordant pairs are shown and Fisher exact between the arms' rates otherwise, and ref. marks an arm's own column. The warmup's seeds 33 to 122 join seeds 1 to 32 as rates. †Pre-registered controls on the held form, run on seeds 1 to 32 on another stack (the late arm's seed 21 had not finished at the pre-registered cutoff) and compared by rate with the n=122 arms; against the held form's 104/122, their primary endpoint, they give p=2×10^{-13}, 0.19 and 0.29 (α=0.05/3).

sum3, seeds 1–32, seed-matched	grok	mis-bound	P(grok | clean)	P(grok | mis-b.)	vs SoftBind	vs baseline
baseline	11/32	18/32	10/14	1/18	2:10, p=0.039	ref.
SoftBind, anneal to 1 by step 500	19/32	4/32	19/28	0/4	ref.	10:2, p=0.039
LR warmup, plain attention (n=122)	47/122	74/122	39/48	8/74	p=0.021	p=0.055
τ=4 held (n=122)	**104/122**	9/122	101/113	3/9	2×10^{-7}	3×10^{-21}
τ=4 held at L1, L0 standard†	5/32	20/32	5/12	0/20	1×10^{-4}	p=0.25
τ=4 held at L0 and L1†	24/32	3/32	24/29	0/3	p=0.043	7×10^{-7}
τ=4 held from step 1000 on†	24/31	10/31	19/21	5/10	p=0.024	3×10^{-7}
heat, τ_0=0.5 → 1 by 25k	9/32	21/32	8/11	1/21	1:11, p=0.006	3:5, p=0.727
softened at L1, L0 standard	7/32	19/32	7/13	0/19	1:13, p=0.002	1:5, p=0.219
NormSoftmax	0/13	0/13	0/13	0/0	0:9, p=0.004	0:6, p=0.031
GrokFast, λ=2.0	6/16	11/16	4/5	2/11	p=0.21	p=0.70

0.889 (peak trap mass 0.03 to 0.16, against 0.57 to 0.74 before); the readout half (final LayerNorm and unembedding; Ba et al., 2016) clears it on 0/6 and reaches 0.683 to 0.763 (0.45 to 0.74 left on the trap). Neither half crosses on its own (on s3 the two halves differ by 0.034 in accuracy while differing completely in binding), so the binding is read directly rather than inferred from the rate.

Which channel each form uses. The anneal's gain is the binding shift alone (its binding distribution applied to the baseline's stratum rates predicts 52.8% against 54.1% observed; Tab. 2). The held form runs 27.9 pp above that prediction (57.3% predicted, 85.2% observed; clean-bound runs also convert more often, 89.4% against 61.4%; Tab. 2). That binding is read through the held form's own divisor, and on the shuffled layout, where it was also read at τ=1, all 32 held runs flag the same heads either way (Appx. B.3.3). Nor does the held form's gain need the early window: switched on only after step 1000, it groks 24/31 (Tab. 11). GrokFast (Lee et al., 2024) runs 15 pp above its own prediction.

The held form on the second task. On prodsum the held form keeps sum3's ordering of Tab. 2 (seeds 33 to 64: grok 5/32 against SoftBind's 3/32 and the baseline's 2/32; mis-bound 11/32 against 12/32 and 16/32), an ordering rather than a test at this base rate (§4).

B.3 Where the softening acts, and where it does not

B.3.1 Deriv at a 200k horizon

To test whether deriv's 50k null is the treatment or the horizon, since generalisation on these tasks can arrive long after the training loss has settled (Power et al., 2022). §3.5's two SoftBind arms run on seeds 1 to 16, as at 50k. A third arm, also pre-registered, holds the softening for the whole run; it ran on a second host, next to a baseline of its own. The pre-registered secondary endpoint is the rise from 50k (Tab. 12). The binding still separates the outcome four times further out: no mis-bound run groks in any arm.

Eight of the 10 runs mis-bound at 50k carry no mis-bound head at 200k, their worst-head trap mass falling from 0.58–0.83 to 0.09–0.15, so on deriv a mis-bound head can leave the trap between 50k and 200k. Appx. B.3.2 and Tab. 13 show that on sum3 no run mis-bound at step 1000 is clean at 50k.

B.3.2 Reading the binding early

This arm of §3.5 reads the binding twice in the same run; we pre-registered it before it existed.

Setup. We reran both arms on seeds 1 to 32, on another host, and probed every run once at step 1000 and once at 50k, with the same probe batch and the same θ (Appx. E.4) as every other reading.

Table 12: Deriv at the 200k horizon, seeds 1 to 16; the second host's baseline finished seeds 1 to 14. Each arm is tested against the baseline on its own host by exact McNemar over shared seeds: at 200k SoftBind is 3:4 discordant (p=1.00) and the held form 5:3 over 14 seeds (p=0.73), and against the first host's two arms the held form is 3:3 and 5:4 (p=1.00 each). The rise from 50k is p=0.0039 and 0.0078 on the first host, 0.031 and 0.0039 on the second.

	first host		second host	
	baseline	SoftBind	baseline	held
grok at 50k	0/16	0/16	0/14	0/16
grok at 200k	9/16	8/16	6/14	9/16
mis-bound at 200k	2/16	1/16	1/14	1/16
grok among mis-bound	0/2	0/1	0/1	0/1
grok among clean	9/14	8/15	6/13	9/15

We read at step 1000 rather than 500 because the anneal ends at 500: a reading taken there would still see the treated arm's softened rows, and a flatter row lowers the trap mass without any head having bound elsewhere. From step 500 on the two arms are the same model, so at 1000 we compare like with like. Tab. 13 gives the transition tallies.

Table 13: Transition tallies on sum3, 32 seeds per arm: the binding read at step 1000 against the binding read at 50k in the same run.

	clean → clean	clean → mis	mis → clean	mis → mis
baseline	14	7	0	11
SoftBind	26	3	0	3

No head leaves the trap after step 1000. Across all 64 runs, no head that sits on the trap at step 1000 has left it by 50k (Tab. 13). Runs that are clean at step 1000 but mis-bound at 50k grok at 2/10. When we run the mediation decomposition (Pearl, 2001) on the step-1000 reading instead of the 50k one (Tab. 10), it predicts 47.5% against the 50.0% we observe. At t=0 no row is saturated, so an initial bias cannot lock a head's target: 7 baseline runs and 3 treated ones still go from clean to mis-bound after step 1000, and none go back (Tab. 13).

The two edges of the class. The account is about tasks that grok stochastically (§2) and whose rule leaves an input unread; we pre-registered one arm on each edge of that description. Disc has no trap and so serves as a negative control: before the runs we wrote down a null, because there is nothing for the softening to unbind, and both arms are 0/16 at the grouped split (Appx. E.2, p=1.00). On modular addition, the benchmark this literature is built on (Power et al., 2022; Nanda et al., 2023), both arms grok 16/16 at the same split and horizon: with k=V there is nothing to mis-bind to.

Table 14: Modular addition with an ignored operand, $(a+b) \bmod 97$ with two further inputs. The binding read at 50k against the outcome, on a 16-run screen, a disjoint 32-run cohort, and the two pooled; every mis-bound head on the screen sits on an ignored operand, at 0.79 to 0.91 of its SEP-row mass. The screen and the cohort were registered before they ran.

	screen (n=16)	cohort (n=32)	pooled (n=48)
grok rate	9/16	15/32	24/48
grok among mis-bound	0/7	1/17	1/24
grok among clean	9/9	14/15	23/24
Fisher exact	p=9×10^{-5}	p=7×10^{-7}	p=4×10^{-11}

B.3.3 SUM3 WITH ITS INPUTS SHUFFLED

On every task above, input i sits at position i+1 in every row, so a head bound to the trap could be bound to a position rather than to an input, and Tab. 18 does find the trap preference in the positional

term. Each row lists its four inputs in an order of its own, one of the 24, and a learned tag added to each input's token embedding says which input it is. Under each seed the rows, answers and grouped split are sum3's. We read the binding with the probe and θ of Appx. E.4 after putting each row's attention back in input order, since a head bound to the trap would otherwise read as spread over four positions. Eight baseline seeds first checked that the task still groks stochastically ($1/8$), and the three arms then ran on seeds 1 to 32, with the tests in Tab. 15's caption fixed in writing before any of them started.

The mis-bound head survives the shuffle. At 50k, 13 of the 32 baseline runs carry one, and none of them groks, against 3 of the 19 clean runs. Each of the 19 mis-bound heads across the three arms holds its whole trap preference in the tag term, with the position and token terms near zero, and the other 365 heads prefer the relevant inputs through the tag as well.

Held for the whole run, the softening lifts the grok rate from $3/32$ to $14/32$ and lowers the mis-bound rate (Tab. 15), and the lift holds when a run must reach 0.95 again within its last 5000 steps. The anneal lowers the mis-bound rate too, to $4/32$, but not the grok rate ($1/32$). On this layout a clean run seldom groks by 50k, so clearing the binding has little rate to add; the held form lifts the clean runs themselves ($13/30$ against $3/19$). Nor does the binding settle by step 1000 here: 8 baseline runs clean at 1000 are mis-bound at 50k and 3 move the other way, one of which groks at step 41,500. Pooled over the arms, the diagnosis points the way it does on the fixed layout without reaching significance at this rate: $1/19$ mis-bound runs grok against $17/77$ clean ones (Fisher exact, Fisher, 1935; p=0.11).

Table 15: Sum3 with its inputs shuffled, 32 seeds per arm with the seeds shared, read at 50k on the grouped split. Registered tests, Fisher exact against the baseline: the held form's grok rate p=0.004 and mis-bound rate p=0.002, the anneal's grok rate p=0.61. Late re-cross counts runs that reach 0.95 again within their last 5000 steps; the last row counts mis-bound heads whose trap preference sits in the tag term.

	baseline	held	anneal
grok rate	3/32	**14/32**	1/32
Wilson 95%	[3.2, 24.2]	[28.2, 60.7]	[0.6, 15.7]
late re-cross	3/32	11/32	1/32
mis-bound at step 1000	8/32	1/32	1/32
mis-bound at 50k	13/32	2/32	4/32
grok among mis-bound	0/13	1/2	0/4
grok among clean	3/19	13/30	1/28
preference in the tag term	13/13	2/2	4/4

B.4 What does not move the failure

Cross-head statistics. Cross-head similarity and entropy statistics serve as online diagnostics of weight-decay regimes on modular arithmetic (Verma, 2026); they read coordination across heads at a swept hyperparameter, where the mis-bound criterion reads the target of one head at a fixed one. Ali (2026) finds a window in training, placed by the initialisation scale, in which the regularisation budget decides whether a transformer memorises or generalises; the resolution window here is the same kind of object read at the level of a head's target, with the temperature rather than the regulariser as the lever. The mis-bound head also supplies the per-head mechanism the circuit-stability frameworks of Sun (2025) and Raja et al. (2026) need: in a stochastic-grokking seed exactly one head stabilises onto an irrelevant variable.

What the sweep contains. The SlotRouter configuration that reaches $3/8$ is tuned (4 layers, GLU (Shazeer, 2020), cosine, init scale 0.5, GrokFast (Lee et al., 2024) λ=2.0, 300k steps). The optimiser rows of Fig. 6 are Sophia (Liu et al., 2024) and AdEMAMix (Pagliardini et al., 2025). If the failure were a capacity or architecture problem rather than a binding failure, giving the model more parameters, more layers, a different regulariser, or a non-softmax sequence mixer should unlock the failed seeds. This section rules out each of these alternatives in turn: four author-designed binding-aids, a 30×-larger model, a seven-configuration capacity sweep, and three non-softmax mixers (Mamba, S4D, SpikingBrain), none of which unlocks a single failed seed.

Author-designed architecture variants. The four regularizers referenced in §4 are: *bridge*, a learned skip that bypasses L0 self-attention as a residual route; *bypass*, direct routing of relevant input positions to L1 leaving L0 attention untouched; *gate*, a learned per-token sigmoid gate that lets the network suppress individual mis-bound heads at run time; and *temp*, a per-head learnable softmax temperature that lets attention sharpness adapt. All four yield $0/8$ grokked on deriv. The remaining architecture rows of Fig. 6 are two widths of the baseline, d=64 and d=256, and a *learned slot router*, an L0 mixer in which each head reads through a learned, input-independent softmax over positions, so that SGD is offered the assignment SlotRouter is given; all three are $0/8$. The 17 direct-intervention runs are an anti-lock auxiliary loss and one-shot, periodic and late resets of the mis-bound head across the six failed row-wise seeds, $0/17$ with peaks 0.23 to 0.35.

B.5 Capacity, other sequence mixers and automated repair

Capacity is not the bottleneck: a 4L/8H/d=512 deriv sweep, with $\sim$30$\times$ the parameters of the 2L/4H/d=128 baseline, groks $0/7$ at 50k steps (one seed excluded for a loss spike at step 25k).

Beyond softmax-MHA: alternative attention paradigms. We re-implement three sequence-mixing modules from recent literature as drop-in L0 replacements: *Mamba* (Gu & Dao, 2024), *S4D* (Gu et al., 2022) and *SpikingBrain* (Pan et al., 2025). All three use the same L1 (standard MHA; Vaswani et al., 2017) and downstream architecture; only L0 changes. Per-seed grok rate at n=8 on deriv: Mamba $0/8$ (max-val 0.19 to 0.26), S4D $0/8$ (max-val 0.19 to 0.24), SpikingBrain $0/8$ (max-val 0.20 to 0.23). Nothing without softmax QK groks deriv either, so what the sweep isolates is a binding problem.

No-grokking tasks. For det/dot/poly, the $4\times$-horizon controls (n=13, 10, 8) keep validation accuracy below 15% at every capacity in the sweep below, consistent with structural failure.

Multi-seed capacity sweep on deriv. Across seven non-baseline (L, H, d) configurations and five seeds each (n=35 runs), no seed groks, with max-val 0.20 to 0.37 (Tab. 16).

Multi-seed capacity sweep on disc. Three non-baseline configurations $\times$5 seeds = 15 runs, all $0/15$ grokked (max-val 0.21 at L=2, H=2; 0.22 at L=2, H=8; 0.21 at L=2, H=16).

Table 16: Multi-seed deriv capacity sweep at 50k on a row-wise split with 70% of rows held out, the split of the middle band of Fig. 6. At this split the baseline L=2, H=4, d=128 does not grok either ($0/8$, max-val 0.19 to 0.37), so the sweep bounds each configuration's ceiling, not its rate.

Config	n	grok	max-val
L=1, H=4, d=128	5	0/5	0.239
L=2, H=2, d=128	5	0/5	0.371
L=2, H=3, d=96	5	0/5	0.259
L=2, H=5, d=160	5	0/5	0.257
L=2, H=8, d=128	5	0/5	0.271
L=2, H=16, d=128	5	0/5	0.262
L=3, H=4, d=128	5	0/5	0.317

Automation gap: ARMA timing sweep. Our manual rescue protocols (single-head transplant, Xavier-reset (Glorot & Bengio, 2010) at step 50k under matched-split evaluation) lift validation accuracy by about 5pp over the vanilla-extension control (Tab. 7) but do not cross the grok threshold. Automating this lift through detection-and-Xavier-reset (ARMA) at {10k, 30k, 50k, 70k} also yields $0/8$ on deriv across all four timings (max-val 0.29 to 0.36).

B.6 LLM-scale diagnostic transfer

This section asks whether the phenomenon behaves the same way in real language models: whether pinning emerges over training and how it scales with model size.

Training trajectory. The pinned-relaxed rate (high saturation with input-independent argmax target) on Pythia-70m-deduped (Biderman et al., 2023) rises with training but with non-monotone intermediate fluctuations: $\sim 0\%$ at random init, 2% at step 1000, 2% at step 10k, 8% at step 20k, 4% at step 30k, 8% at step 50k, 10% at step 70k, 19% at step 100k, and 23% at the final step 143k checkpoint. Under an i.i.d. Gaussian-logit null, $P(\text{saturation} > 0.7)$ is 0.85% per head (~ 0.4 heads in expectation across 48); the observed 11 pinned heads at step 143k is 25 to 40σ above this null. Pinning emerges late and concentrates in deep layers (layer-5 mean saturation ramps from 0.42 at step 10k to 0.92 at step 143k while layer-0 plateaus around 0.39). Of the 11 pinned heads at convergence, 9 are BOS-sinks (Xiao et al., 2024; Gu et al., 2025) and the remaining are previous-token / self-attention pins.

Size sweep. Across Pythia 70M / 160M / 410M / 1B, the pinned-head rate is $23\%/24\%/17\%/11\%$ respectively. Pinning is increasingly confined to mid-depth layers as the model deepens: last-layer mean saturation drops from 0.92 at 70M to 0.39 at 1B as upper layers shift from saturating sinks to diffuse mixing.

B.7 Task-by-task scope of the router

The alternative non-softmax sequence mixers (Mamba, Gu & Dao, 2024; S4D, Gu et al., 2022; SpikingBrain, Pan et al., 2025; $0/8$ each), the ARMA timing sweep and the capacity-sweep null are in Appx. B.5. SlotRouter (§4) helps where the failure is single-point mis-binding with a head to spare, reaching $3/8$ on deriv ($\mu_H{=}1$, Appx. E.4; compute-matched reading in §4, $1/8$ against $2/8$). The head budget accounts for the rest: det and poly have no head to spare, dot has more relevant inputs than L0 heads, so no assignment covers them all, and disc has 2 spare heads yet does not respond, as its distributed failure predicts (Tab. 1). Because the router leaves det and poly unchanged, binding is ruled out as their constraint.

C Per-Task Results on the Row-Wise Split

This section holds the per-seed evidence behind the regime picture of §3.

Which split these tables measure. Every rate in this section is on the *row-wise* split (§2, Appx. E.2), so on trap tasks it reads far higher than the grok rates the paper reports: disc_pad is $8/8$ here and $1/32$ on the grouped split, deriv $3/16$ against $1/32$.

The per-seed record. Multi-seed sweep (50k steps, 2L/4H/128d unless noted): on this row-wise cohort deriv groks $3/16$ (seeds 1, 5 and 9, max val 0.962, 0.959, 1.000) and disc $1/16$ (seed 1, 0.951); the failed seeds sit at 0.31 to 0.89 on deriv and 0.45 to 0.87 on disc. Re-running failed seeds for $4\times$ longer never unlocks them: at 200k deriv s2's max val moves from 0.308 to 0.530 and s3's from 0.577 to 0.685, disc s2 from 0.526 to 0.536 and s5 from 0.471 to 0.601. A fresh deriv seed run only at 200k (s42) groks at 0.985. The no-grokking tasks (det, dot, poly) are summarised in Tab. 17.

Table 17: Aggregate per-seed results for the no-grokking tasks at 50k; none of the $4\times$ longer (200k-step) re-runs unlocked grokking on a previously failed seed.

task	n (50k)	range	grokked	200k re-runs
det	13	0.022 to 0.060	0/13	0/3 (s1 to s3, max 0.064)
dot	10	0.038 to 0.108	0/10	0/3 (s1 to s3, max 0.138)
poly	8	0.110 to 0.136	0/8	0/3 (s1 to s3, max 0.148)

C.1 Wilson Confidence Intervals

Wilson 95% intervals (Wilson, 1927) on the row-wise rates, counting both the 50k and 200k seeds since neither produced a grokked seed on det, dot or poly: disc_pad $[67.6, 100]$, deriv $[6.6, 43.0]$, disc $[1.1, 28.3]$, and upper bounds of 22.8 to 32.4 on the three tasks that never grok.

The regime distinction, some seeds grokking against none, is what these intervals separate at this sample size; the per-task rates behind them are in Appx. C.

D Circuit Readings: the Mis-Bound Head, and Pinned Heads

D.1 The mis-bound head's circuit

Tab. 18 reads the flagged heads of the row-wise cohorts of Appx. A.8, and of eight deriv seeds from the same sweep, through the circuit measurements of Elhage et al. (2021) and Olsson et al. (2022). Fixing each token's scale makes LayerNorm linear, and the head's attention score from SEP then separates exactly into what its key reads from the token embedding, what it reads from the positional embedding, and what it takes from the bias. We report the trap preference as that score on the trap less its mean over the relevant inputs, in logits; the bias term is the same in both and drops out. We ablate a head with every attention pattern held frozen, as Olsson et al. (2022) do. Freezing changes no per-seed differential by more than 0.2 points. The last two rows swap one of the seven other runs' tables into the run being read, whichever run donates it, and count the flagged heads still bound afterwards.

Table 18: The flagged L0 head against the other three heads of the same failed runs, five runs per task, each read at step 50k or at the last checkpoint that still fits its training rows.

	deriv		prodsum		sum3	
	flagged	others	flagged	others	flagged	others
trap preference	2.60	-0.92	2.20	-0.78	2.31	-1.31
position term	2.70	-0.94	2.23	-0.78	2.31	-1.32
token term	-0.09	0.02	-0.03	-0.01	-0.01	0.01
rows with the trap as argmax	97%	0%	100%	0%	100%	0%
ablation, loss rise (nats)	0.04	2.39	0.32	2.46	0.20	1.32
ablation, accuracy drop (pp)	0.4	29.7	3.6	25.3	3.3	7.6
still bound, token embedding swapped	$34/35$		$35/35$		$29/35$	
still bound, positional table swapped	$0/35$		$0/35$		$0/35$	

D.2 Pinned LLM heads include dormant induction heads

Tab. 19 crosses the pinned heads of Appx. B.6 with induction heads as Olsson et al. (2022) define them: prefix matching above 0.4 on 25 random tokens repeated four times after BOS, with positive eigenvalue mass in the OV circuit (Elhage et al., 2021). Expected is the overlap pinning would produce if it ignored induction, and the p beside it is hypergeometric. In Pythia-70M the pinned count first rises at step 20k, and three of the four heads pinned there are induction heads (p=0.002). Every pinned induction head is also a BOS sink on the prompts, which is what the last two columns read: BOS mass is the mean attention to BOS from the second half of the first copy, where nothing earlier matches, and from the three repeats.

Table 19: Induction heads among the pinned heads of each Pythia model at its final checkpoint.

					BOS mass, first copy $\to$ repeats	
model	pinned	induction	both	expected (p)	pinned induction	other pinned
70M	11	$5/48$	3	1.2 (0.07)	$0.36 \to 0.04$	$0.26 \to 0.13$
160M	34	$18/144$	11	4.3 (3×10^{-4})	$0.51 \to 0.15$	$0.47 \to 0.57$
410M	65	$19/384$	13	3.2 (4×10^{-7})	$0.84 \to 0.20$	$0.72 \to 0.58$
1B	14	$9/128$	4	1.0 (0.008)	$0.72 \to 0.13$	$0.46 \to 0.43$

E SETUP

E.1 THE CAUSAL INTERVENTIONS

Zero-ablation: zero the output-projection columns of a single head; mis-bound vs. normal head as control. *Weight transplantation*: copy a single head's Q, K, V, and output projection slices from a grokked checkpoint into a failed checkpoint, then continue training, the protocol of §3.4; *embedding transplantation* copies instead the token and positional embeddings, final LayerNorm and unembedding, leaving every attention and MLP weight in place. *Xavier-init reset* replaces those donor weights with Xavier random init (Glorot & Bengio, 2010).

E.2 THE TWO SPLITS, AND THE COHORTS

A row-wise split can place the same relevant-variable tuple (Tab. 21) in train and validation under different trap values, so passing it requires ignoring the trap but not extrapolating the rule to unseen relevant tuples: grouping by relevant tuple closes that channel. Every rate we report is on the grouped split (§2); the row-wise rates are in Appx. C.

A *checkpoint cohort* of 8 to 16 seeds is read head by head at the saved steps (Appx. E.4); a *rate cohort* of 32 to 248 seeds per arm is read only as a grok rate with a Wilson (1927) interval. An *arm* is one training condition, run on the same seeds as the others.

E.3 MODEL, TRAINING AND MEASUREMENT

Architecture. Pre-LN transformer with causal masking and learned positional embeddings. Default config: 2 layers, 4 heads, d_{model}=128, d_{ff}=512, $\sim$ 453K parameters.

Tokenizer and sequence format. Vocabulary of 205 tokens: 201 number tokens covering integers in $[-100, 100]$, plus 4 special tokens (PAD, SEP, BOS, EOS). Single-token number encoding: each integer maps to a unique vocabulary token. Sequences are formatted as $[\text{BOS}, x_1, \ldots, x_n, \text{SEP}, y, \text{EOS}]$ with cross-entropy loss on the output token only.

Training. AdamW (Loshchilov & Hutter, 2019) with lr=10^{-3}, wd=1.0, $(\beta_1, \beta_2) = (0.9, 0.98)$, ϵ=10^{-8}, fused. Constant learning rate (no warmup or scheduler). Full-batch updates (2500 examples per step under 50/50 train/val split). 50,000 steps throughout, 200,000 for the checkpointed deriv trajectory. Evaluation every 500 steps. No dropout, no gradient clipping.

Causality gate. Teacher forcing places the answer token inside the input, so any L0 module that lets an earlier position read the last position can copy it and fake grokking. The training script therefore runs a probe before every run: it overwrites the last input position on random inputs and requires the outputs at all earlier positions to be identical to within 10^{-6}, raising an error and refusing to train otherwise. The probe does not trust per-module masking, so a module that ignores the mask fails it. The same corruption is re-applied to trained checkpoints as a forensic check (intact and corrupted validation accuracy must agree), and every grokked checkpoint quoted from a non-standard L0 module passed it.

Causal ablation. Zero-ablation sets the output projection columns corresponding to a single head to zero.

Manual attention extraction. Because the fused SDPA path does not expose attention weights, we re-implement the forward pass using the same projection weights to extract per-head attention patterns. Forward equivalence verified to within 10^{-7}.

E.4 NOTATION AND COHORTS

Table 20: Symbols and cohort labels used across the paper.

Symbol	Meaning (where introduced)
H, H_0 to H_{H-1}	number of first-layer (L0) heads and their indices (§2)
V	attention-accessible input positions (§2)
k	relevant inputs the rule reads; $H \leq k$ is the no-grok floor (§3.1)
trap	the input the rule ignores: c on deriv, d on prodsum and sum3 (§2)
rel-attn, trap-attn	a head's SEP-row attention mass on the relevant inputs, on the trap (§2)
mis-bound head, θ	trap-attn $> \theta \cdot$ rel-attn, θ=1; sweep in Appx. A.7 (§2)
row-wise, grouped split	rows held out at random; whole relevant tuples held out (§2)
grokked	max validation accuracy ≥ 0.95 (§2)
τ, τ_0	SoftBind temperature, annealed from τ_0=4 to 1 over the first 500 steps (§3.5)
μ_H	head margin $H-k$
s1 to s16; s101 to s132	deriv row-wise cohort; deriv grouped cohort (labels are per cohort and task)
checkpoint cohort	8 to 16 seeds, read head by head at the saved steps (Appx. E.2)
rate cohort	32 to 248 seeds per arm, read only as a grok rate (Appx. E.2)
arm	one training condition, baseline or intervention, over the same seeds (Appx. E.2)

E.5 TASK GENERATION

Table 21: The five symbolic-math tasks (deriv, disc, det, dot, poly), the binding-lottery positive control disc_pad, and the two single-trap rules of §3.4, prodsum and sum3.

Task	Inputs	Output	# relevant vars	Analytic rule
deriv	(a, b, c, x_0)	$f'(x_0)$	3 (a, b, x_0)	$2ax_0 + b$
disc	(a, b)	discriminant	2 (a, b)	$a^2 - 4b$
det	(a, b, c, d)	determinant	4 (all)	$ad - bc$
dot	$(a_1{:}a_3, b_1{:}b_3)$	inner product	6 (all)	$\sum_i a_i b_i$
poly	(a, b, c, x)	poly value	4 (all)	$ax^2 + bx + c$
disc_pad	(a, b, c)	discriminant	2 (a, b)	$a^2 - 4b$ (c irrelevant pad)
prodsum	(a, b, c, d)	$ab + c$	3 (a, b, c)	$ab + c$ (d irrelevant)
sum3	(a, b, c, d)	$a + b + c$	3 (a, b, c)	$a + b + c$ (d irrelevant)

For each task (Tab. 21) we enumerate all (input, output) tuples within the input range, filter those whose output stays in $[-100, 100]$, and shuffle to obtain ∼5000 unique pairs; disc's range holds only 611, and it uses all of them. The filtered spaces are far larger, so the sample is about 1% of deriv's and 0.5% of sum3's; the collision the grouped split removes is common rather than marginal, 25.4% of a 5000-pair deriv sample sharing its relevant tuple with at least one other pair (15.7% on sum3). Every input lies in $[-15, 15]$, except dot's $a_i, b_i \in [-5, 5]$.

E.6 TOKENIZATION × WEIGHT-DECAY CONTROL

Table 22: 2×2 ablation of tokenization and weight decay on modular addition, $(a+b) \bmod 97$ with 30% of its 9409 pairs in training and 50k steps, on the 2L/4H/d=128 model and optimiser of the deriv and disc experiments. Digit-level encoding splits each number into digit tokens ($17 \to [1, 7]$), single-token encoding gives each value its own token. Only the (single-token, wd=1.0) corner groks after a delay; the (single-token, wd=0) corner memorises and never generalises, and both digit-level corners generalise at once.

	wd $= 0$	wd $= 1.0$
digit-level numbers	100% (instant)	100% (instant)
single-token numbers	3% (memorised)	**100% (grokking)**