DOWNSHIFT: INDUCTION CIRCUITS RECRUIT THE FASTEST SURVIVING ROPE BAND

Anonymous authors
Paper under double-blind review

ABSTRACT

In pretrained Pythia-70M, disabling the fastest rotary position embedding (RoPE) band in one previous-token head reproduces 87% of the induction-accuracy drop caused by disabling it model-wide. Yet induction forms and recruits band 1 with this rotation disabled throughout training, across three data-order seeds from one public 70M initialization. Head-set knockouts at 70M and 160M verify the recruited band's contribution to synthetic induction and held-out text prediction. Separate early-checkpoint tests at 70M support fastest-surviving-frequency allocation over fixed-index selection. This *band downshift* links frequency-based recruitment to acquired functional dependence, illustrating why strong endpoint dependence need not imply developmental necessity.

1 INTRODUCTION

An endpoint ablation reveals whether a trained model depends on a mechanism, not whether that mechanism was required for learning. We distinguish *dependence* (does the trained circuit use it?), *developmental preference* (does training select it by default?), and *developmental necessity* (can the function emerge without it?). These questions need not have the same answer: a function that is fragile to ablation after training may admit another implementation when the preferred mechanism is unavailable from the start.

RoPE (Su et al., 2024) makes this distinction testable one frequency at a time. It rotates query–key dimension pairs at geometrically spaced frequencies: fast rotations support local positional attention, while slow rotations can carry content with little direct phase change (Barbero et al., 2025; Urrutia et al., 2026b). Prior constrained-training work already shows that learned spectral signatures need not be required for induction to emerge (Li, 2026). We ask a more specific circuit-level question: which remaining frequency takes over positional routing, and does the circuit become functionally dependent on it?

We study the canonical two-stage induction circuit (Elhage et al., 2021; Olsson et al., 2022). In a sequence $[A][B] \dots [A]$, a previous-token head carries information about the first A to the position of B; a later induction head uses that information to attend to B and predict it after the second A. This division of labor predicts that previous-token routing, rather than the induction head's content lookup, bears the circuit's direct dependence on fast RoPE rotations. An additive query–key contrast localizes the two roles, and head-scoped ablations test their dependence separately. At the trained Pythia-70M endpoint, disabling band 0 in the previous-token head alone reproduces 87% of the model-wide induction-accuracy drop; the same intervention in the induction head changes accuracy by less than 0.001.

Intervention during learning gives a different answer (Figure 1). With band 0 disabled before the first optimizer update, Pythia-70M forms induction from its public initialization and recruits band 1, the fastest surviving rotation. Three paired data-order seeds share the same starting weights.

The circuit also becomes dependent on the recruited band. At the 70M initialization-trained endpoints and matched 160M continuations from step 512, disabling band 1 in independently selected previous-token heads sharply reduces induction accuracy, with much smaller effects in layer-matched random heads. The same intervention impairs held-out natural-text prediction, particularly on repeated contexts.

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

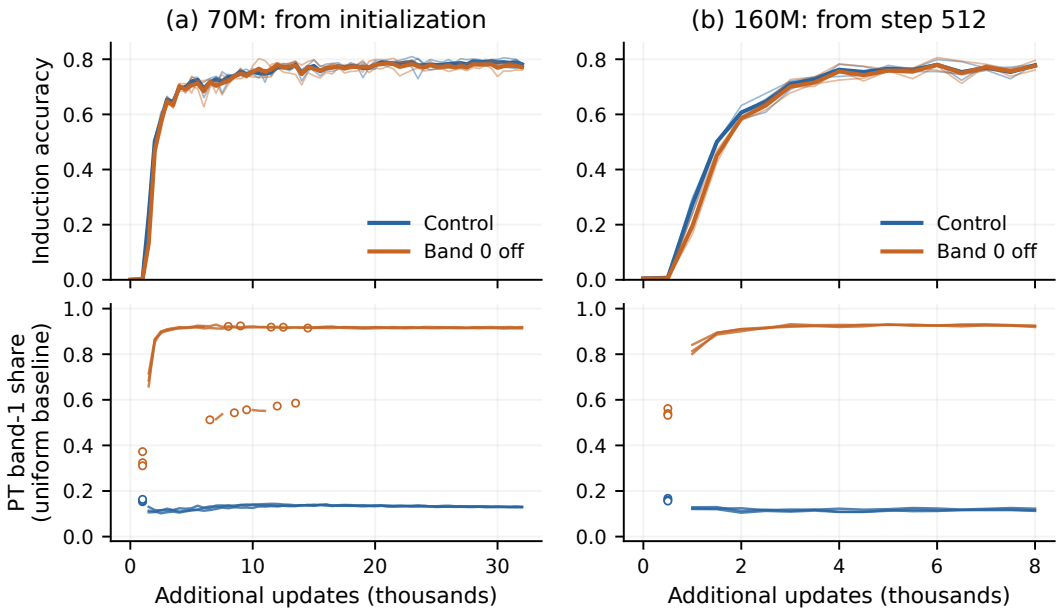


Figure 1: Induction forms with band 0 disabled from initialization at 70M (E1) and from step 512 at 160M (E2). Top: thick lines show means over three data-order seeds; thin lines show individual runs. Bottom: individual runs' band-1 shares for the reselected previous-token head, using the uniform baseline. Shares with predecessor attention below 0.30 or total margin at most 10^{-6} are omitted; head changes break the line and are marked by open circles. Horizontal axes count additional optimizer updates.

We call this *band downshift*: previous-token matching moves to the fastest surviving rotation—slower than the removed bands, but fastest among those still available. Complementary step-512 allocation tests support this rule under a removal ladder and a reversed frequency-to-index mapping.

- We demonstrate that induction forms from a public 70M initialization and recruits band 1 with band 0's rotation disabled throughout training (Sections 3 and 4.1).
- We establish the recruited band's contribution to synthetic induction and natural-text prediction with head-set knockouts at 70M and 160M (Sections 4.2 and 4.3).
- We show that previous-token margin allocation follows the fastest surviving rotation rather than a fixed dimension-pair index in step-512 interventions (Section 5.1).

2 Background and Setup

RoPE bands. Each RoPE band is a query–key dimension pair rotated by angle $\theta_b m$ at position m. In Pythia (Biderman et al., 2023), $\theta_b = 10000^{-2b/d_r}$ for rotary dimension d_r, so band 0 is fastest unless we explicitly permute the frequencies. Pythia rotates one quarter of each head: 16 of 64 dimensions at 70M–410M (8 bands, 48 unrotated dimensions), and 32 of 128 at 1.4B (16 bands, 96 unrotated dimensions). The full-rotary models in our descriptive survey retain their own frequency schedules (Appendix A.7). An ablation replaces the selected pair's rotation with the identity—cosine one, sine zero. It removes that pair's *direct phase change*, not its dimensions or content-carrying capacity; other bands and causal masking remain available. A global ablation applies to every head, whereas a head-scoped knockout applies only to the specified heads.

An additive matching contrast. To separate target matching from score magnitude, we use a mean-referenced query–key contrast. For query position m and correct key n^*,

$$\mathrm{margin}_b(m) \;=\; c_b(m, n^*) \;-\; \frac{1}{|\{n \le m\}|} \sum_{n \le m} c_b(m, n), \tag{1}$$

Table 1: Which experiment supports which claim? E1/E2 vary data order at fixed public starting weights; E3 evaluates their twelve endpoints without further training. The complementary step-512 experiments use separate protocols and do not supply initialization evidence.

Question	Intervention and setup	Readout and inference
Formation without band 0	E1: 70M from public step 0, with band 0 off before the first update; paired controls.	Induction accuracy and band allocation: formation without prior band-0 training in this 70M protocol.
Acquired band dependence	E1/E2 endpoints: 70M step-0 and 160M step-512 training; calibrated head-set knockouts.	Induction-accuracy drops: causal use of band 1 at both scales, not formation from initialization at 160M.
Natural-text contribution	E3: the same twelve endpoints; held-out documents and fixed matched-target masks.	NLL increases: contribution to text prediction, not full induction-path mediation.
Frequency choice	Separate 70M step-512 runs: removal ladder and frequency permutation.	Previous-token margin shares: allocation by frequency; deeper and permuted replacements lack causal tests.
Downstream connection	Separate resumed 70M endpoints: key/query residual-source patches.	Target attention: shared dominant key source; no output-logit mediation test.

where $c_b(m, n)$ is band b's contribution to the unscaled query–key score and the mean includes all causally accessible keys, including the target. Summing over bands and the *unrotated tail*, when present, exactly reconstructs the same contrast for the full score. This is a matching statistic, not the correct-versus-largest-competitor decision margin. We average each contribution over sequences and query positions, then define its share as $s_b = \overline{\text{margin}}_b / |\overline{\text{margin}}_{\text{total}}|$. The usual headwise attention scale cancels from this ratio. Shares are signed and can exceed one; a large share does not establish dependence on the corresponding rotation. An equal-per-dimension reference assigns $2/d_h$ to one band and $(d_h - d_r)/d_h$ to the tail, where d_h is head width.

At fixed pre-RoPE activations, a rotation knockout changes a band's score contribution as

$$c_b(m, n) = q_{m,b}^\top R(\theta_b(n - m))\, k_{n,b} \quad \xrightarrow{\text{rotation off}} \quad q_{m,b}^\top k_{n,b}, \tag{2}$$

where $q_{m,b}$ and $k_{n,b}$ are two-dimensional pairs and R is a two-dimensional rotation. The unrotated inner product remains, including contextual information already encoded in the activations. A pair can therefore carry matching information even if its rotation is dispensable. Our knockout evaluations use full forward passes, allowing the intervention to propagate through subsequent computations.

Unqualified shares use the uniform baseline above. For sensitivity analysis, a second baseline replaces the key mean with $\sum_{n \le m} a_{mn} c_b(m, n)$, where a_{mn} is the model's attention weight. For induction, the target key is $T - 1$ positions back on a repeated block of length T; for previous-token routing, it is one position back. For E1/E2 trajectories, we omit previous-token shares when attention to the predecessor is below 0.30 or the total margin is at most 10^{-6}. Where scale matters, we also report absolute margins and the query–key norm product $\mathbb{E}[\|q_m\|_2 \|k_{n^*}\|_2]$. Validation checks both additive reconstruction (maximum relative error 1.1×10^{-7}) and attention reconstructed from captured queries and keys (maximum absolute error below 10^{-4} in float32). Appendix A.7 reports the separate bfloat16 survey checks.

Behavioral probe. We feed a random 64-token block twice and measure argmax accuracy when predicting the tokens of the second copy. The probe averages four seeds of 64 sequences, sampled from the first 1,000 token IDs; argmax is over the model's full output vocabulary. Head scores measure attention to the immediate predecessor or the induction target, not output accuracy. Appendix A.3 identifies the smaller probe used in the original repair experiment.

Intervention protocol. Our primary comparisons pair control and global band-0-off training at 70M from public step 0 (E1) and at 160M from step 512 (E2), each with three data-order seeds and fixed shared starting weights. Both use fresh AdamW, effective batch 128, and length 512. E1 runs 32k updates with a 1k-step warmup to 10^{-3} followed by cosine decay; E2 runs 8k updates at

3×10^{-4}. The respective budgets are 2.097B and 0.524B token positions per arm, each comprising four passes over a finite Pile-derived buffer. Paired conditions see identical examples in identical order; these are not reconstructions of the original pretraining stream. Both sets of endpoints undergo synthetic causal tests and held-out natural-text evaluation (E3). The training-time mask remains active throughout evaluation. Separate step-512 runs are used for the ladder, permutation, and path tests; we report them separately rather than pooling them with E1/E2. Full protocols and numerical validation are in Appendices A.1 and A.3.

Table 1 maps the experiments and their readouts to the claims they support.

3 WHERE DIRECT ROPE-PHASE DEPENDENCE LIVES

We first locate the trained circuit's direct dependence on RoPE phase, then ask whether that dependence constrains formation.

Induction matching concentrates in unrotated channels. At Pythia-70M's final checkpoint the top induction head (L3H6, induction score 0.93) draws 97.0% of its matching margin from the 48 unrotated dimensions. The equal-per-dimension reference is 0.750: a rotating dimension contributes roughly 9% as much matching contrast as an unrotated dimension on average. Across the six heads with induction score above 0.25, the mean tail share is 0.973 (sd 0.039); the top-head endpoint value ranges from 0.956 to 0.987 across four Pythia scales. An analogous descriptive split appears in 11 full-rotary models, where induction matching concentrates in slow, *quasi-static* bands and previous-token matching in fast bands (Appendix A.7, Figure 6). The causal tests below concern Pythia.

The previous-token head carries the band-0 dependence. At offset 1, the top previous-token head (L2H1) assigns band 0 a share of 0.534 under the uniform baseline and 0.779 under the softmax baseline. Across all 48 heads, previous-token attention correlates positively with fast-band shares, but this association does not distinguish band 0 from band 1 (Appendix A.4). The head-scoped intervention supplies that distinction.

Head-scoped causal test. At 70M, disabling band 0 model-wide reduces induction accuracy from 0.775 to 0.208. Disabling it in L2H1 alone reduces accuracy to 0.283, reproducing 87% of the global drop; disabling it in all 47 other heads together reproduces only 3%. The induction-head and fixed comparison-head interventions each change accuracy by less than 0.001, as they do at 160M and 410M. These intervention effects are nonadditive: the 87% and 3% are not shares that must sum to 100%. Single-head concentration does not generalize across scale: the top previous-token head reproduces 24% of the global drop at 160M and 20% at 410M. At 160M, jointly intervening in the top three previous-token heads reaches 74%, and the top twelve reach 90%, motivating the head-set tests below.

The concentration is learned. Across four public Pythia checkpoint grids, previous-token band-0 concentration rises in the same measured interval as induction and remains preferred at the endpoint. The checkpoint spacing does not resolve which rise comes first; Figure 4 and the real-text ablation diagnostic are in Appendix A.4.

4 BAND DOWNSHIFT

4.1 FORMATION WITHOUT BAND 0

Intervening at step 512 leaves open whether band 0 helped establish earlier precursors. E1 addresses this gap by disabling the rotation before the first optimizer update of the public 70M initialization and keeping it disabled throughout learning. All three data-order seeds form induction (Figure 1). Endpoint accuracy is 0.771 ± 0.011 against 0.781 ± 0.006 for controls (mean $\pm$ sample sd); a nonrepeated random-token probe remains near chance, at mean 0.00083 in the ablated runs. The selected previous-token head assigns 0.916 ± 0.003 of its uniform-baseline margin to band 1, compared with 0.130 in the controls. We call the recruited band's endpoint share its *takeover share*. Thus band 0's rotation is not required for formation under this 70M protocol.

E2 repeats the matched comparison at 160M from step 512. It reaches induction accuracy 0.776 ± 0.018 against 0.778 ± 0.006, with band-1 share 0.923 ± 0.003.

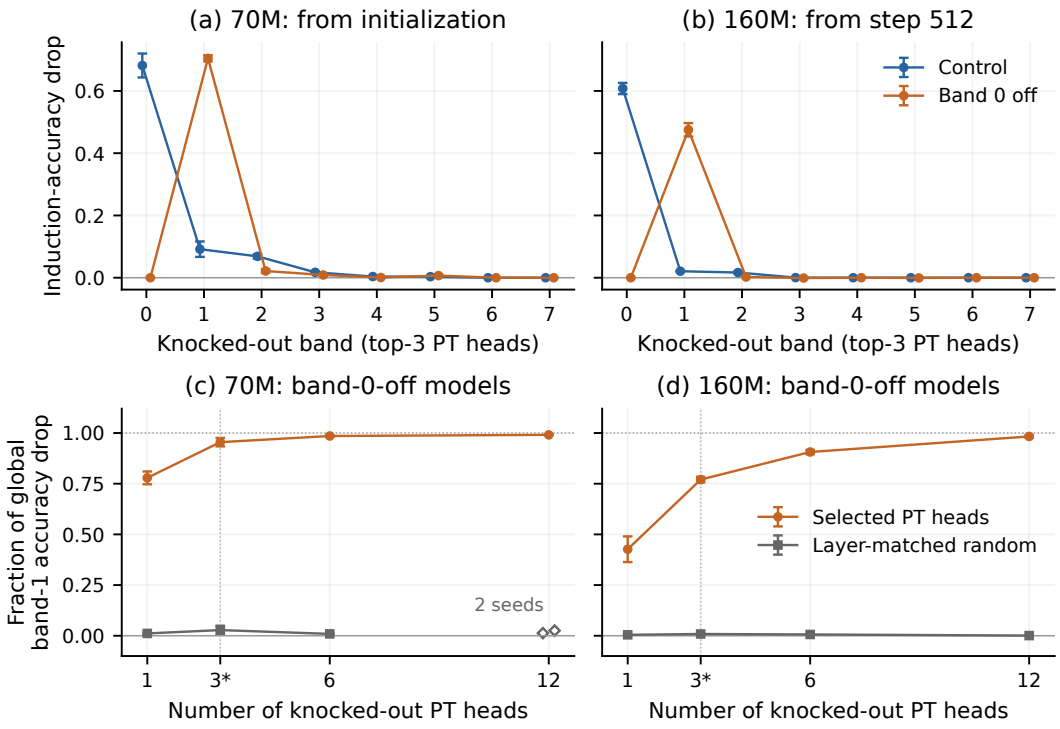


Figure 2: Band profiles and head-set sensitivity. (a,b) Absolute induction-accuracy drops from top-3 rotation knockouts; the training mask makes band 0 redundant in band-0-off models. (c,d) Band-1 drops in band-0-off models as fractions of each seed's global band-1 effect. Filled points are means $\pm$ sample sd over three data-order seeds; random effects first average ten layer-matched sets per seed. At 70M and k=12, open diamonds show each available seed's random-control mean (two of three seeds). The asterisk marks the primary k=3 analysis.

Allocation is baseline-sensitive: the softmax baseline gives band-1 shares of 0.765–0.921 at 70M and 0.200–0.545 at 160M. We therefore test dependence with knockouts (Section 4.2). The replacement can also involve a different head: E1 seed 2 selects L2H3 under ablation and L2H1 in control (Appendix A.2).

4.2 Causal dependence moves

Does band-1 allocation correspond to a new functional dependence? We evaluate E1/E2 with a head-set knockout protocol fixed before training. An independent calibration probe ranks heads by predecessor attention; the primary set contains the top three, with top-1/6/12 as secondary analyses. Selection is repeated for each endpoint using attention scores, not knockout outcomes, so paired models can assign the role to different heads. For each endpoint, ten random sets match the primary set's head count in each layer and exclude its selected heads and the induction head. Layer matching controls the intervention's depth distribution while testing whether the functionally selected heads are unusually consequential. All interventions retain the training-time global mask, so an additional band-0 knockout in a band-0-off model is redundant.

In the band-0-off models, top-3 band-1 knockout reduces 70M accuracy from 0.771 to 0.066 and 160M accuracy from 0.776 to 0.301 (Table 2). It reproduces 95.5% and 77.0% of the corresponding global band-1 effects, respectively (mean of seedwise ratios). Random-head costs are much smaller in every seed, whereas the matched controls remain most sensitive to band 0. The full profiles show that band 1 also has the largest knockout cost among surviving bands in the selected heads (Figure 2a,b). Dependence therefore shifts at both scales, without requiring single-head concentration at 160M.

Table 2: Causal dependence shifts from band 0 to band 1 in E1/E2. Each knockout disables one band's rotation in the prespecified top three previous-token heads. Drops are absolute accuracy reductions; positive values indicate impairment. Values are means ± sample sd over three data-order seeds. The random column first averages ten layer-matched head sets within each seed; dashes mark band 0 already disabled throughout training.

Model	Training	Native acc.	Drop: b=0	Drop: b=1	Random b=1
70M	Control	0.781 ± 0.006	0.682 ± 0.038	0.092 ± 0.025	0.0099 ± 0.0047
70M	Band 0 off	0.771 ± 0.011	—	0.705 ± 0.010	0.0207 ± 0.0161
160M	Control	0.778 ± 0.006	0.608 ± 0.018	0.021 ± 0.001	0.0016 ± 0.0004
160M	Band 0 off	0.776 ± 0.018	—	0.475 ± 0.022	0.0053 ± 0.0010

Table 3: Natural-text dependence shifts from band 0 to band 1. Additional rotation knockouts target the top three previous-token heads; entries are NLL increases in nats/token (mean ± sample sd over three data-order seeds). R and N are token- and position-bin-matched repeated and nonrepeated targets. For band 1, Random R first averages ten layer-matched random-head interventions per seed; dashes indicate that this text comparison was not measured for bands 0 and 3. Redundant band-0 rows are omitted for band-0-off models.

Model	Training	Band	All	Matched R	Matched N	Random R
70M	Control	0	0.352 ± 0.054	1.218 ± 0.202	0.298 ± 0.056	—
70M	Control	1	0.026 ± 0.007	0.144 ± 0.035	0.005 ± 0.003	0.020 ± 0.009
70M	Band 0 off	1	0.392 ± 0.016	1.326 ± 0.070	0.324 ± 0.014	0.038 ± 0.028
70M	Band 0 off	3	0.009 ± 0.005	0.031 ± 0.011	0.001 ± 0.003	—
160M	Control	0	0.275 ± 0.019	1.029 ± 0.046	0.201 ± 0.020	—
160M	Control	1	0.007 ± 0.001	0.047 ± 0.003	0.000 ± 0.002	0.009 ± 0.001
160M	Band 0 off	1	0.221 ± 0.007	0.837 ± 0.004	0.166 ± 0.016	0.022 ± 0.004
160M	Band 0 off	3	0.001 ± 0.001	0.008 ± 0.003	0.001 ± 0.002	—

For E2 seed 1 under band-0-off training, the selected previous-token head assigns band 6 a larger softmax-baseline share than band 1 (0.312 versus 0.200). Yet top-3 knockout of band 6's rotation produces no measured accuracy loss, while band 1 costs 0.494.

The secondary head-set sweep shows that this dependence is not specific to k=3 (Figure 2c,d). In the band-0-off 160M runs, top-1/3/6/12 band-1 knockouts reproduce 42.7%, 77.0%, 90.6%, and 98.3% of the global effect, respectively. At 70M, the top head already reproduces 77.9%, rising to 95.5% for the top three. Layer-matched random effects remain much smaller. These profiles are consistent with more distributed band-1 dependence at 160M: a weaker single-head effect need not imply weaker band dependence. The differing starting points and training budgets prevent attributing this difference to model size alone; joint knockout effects remain nonadditive.

4.3 Contribution to natural-text prediction

To test whether the recruited band contributes beyond the synthetic probe, we evaluate all twelve E1/E2 endpoints on 2,048 held-out Pile-derived documents. Each document contributes one hash-selected window of at most 512 tokens. Both conditions' training buffers exclude these documents and the validation sample by exact document hash. We identify repeated targets by whether the same two-token context and correct continuation occurred at least eight positions earlier in the window. Repeated and nonrepeated targets are then matched by target-token identity and 64-position bin. The text-only masks are fixed across models and interventions: 757,825 total targets, 100,310 repeated targets, and 84,663 in each matched group. Matching does not equate baseline predictability. In band-0-off models, native NLL on matched repeated/nonrepeated targets is $0.532/2.723$ at 70M and $0.629/2.968$ at 160M (nats/token, three-seed means).

The control models are more sensitive to band 0 than to band 1 on both matched target groups (Table 3). After band-0-off training, band-1 knockout produces large NLL increases, whereas band-3 effects remain small. Thus the change in band dependence extends beyond the synthetic probe to held-out text prediction.

In the band-0-off models, the band-1 effect is larger on matched repeated than nonrepeated targets: the difference is 1.001 ± 0.060 nats/token at 70M and 0.671 ± 0.012 at 160M. After subtracting the mean random-head effect, this repeated-minus-nonrepeated contrast is 0.964 nats/token at 70M and 0.651 at 160M. It is positive in every seed, and each seed's paired-document bootstrap 95% interval excludes zero. Each interval reflects document-sampling uncertainty conditional on one trained model (Appendix A.2). The group contrast compares knockout effects; it does not isolate repetition from other contextual differences or establish full-path mediation.

Without an additional knockout, overall NLL is 3.362 in control and 3.362 under band-0-off training at 70M, and 3.604 and 3.609 at 160M. These native losses are reported separately from the knockout effects; their proximity does not establish improved LM quality or statistical equivalence.

5 Frequency choice and downstream connectivity

Separate step-512 experiments test how positional matching is reassigned and how the previous-token head connects to the induction head. Their resume protocols differ from E1/E2 (Appendix A.3).

5.1 Frequency, not index

The band-1 result motivates a candidate allocation rule: the fastest *surviving* rotation takes over previous-token matching. We compare a removal ladder, non-preferred-band controls, and a frequency permutation. The stock-RoPE band-0-off runs yield takeover shares of 0.926 at 70M and 0.925 at 160M (Table 8). The 70M shift is not explained by renormalizing away band 0: band 1 accounts for 27.1% of the controls' remaining margin on average, compared with 93% after intervened training.

Specificity. Removing band 1 or band 3 alone leaves band 0 dominant at both scales (three seeds per condition; Table 8). Among these conditions, recruitment follows removal of the preferred band, not ablation of any band.

The ladder. At 70M, ablating bands 0 and 1 recruits band 2, and ablating bands 0 through 2 recruits band 3, with L2H1 remaining the selected previous-token head (Table 13). The rule describes *which* band is recruited, not its concentration: takeover shares are non-monotone at 0.93, 0.58, and 0.66. Concentration and margin-inflation diagnostics are in Appendix A.12.

Permutation. In stock RoPE, the fastest frequency also occupies the lowest dimension-pair index, so the ladder alone cannot distinguish frequency selection from index selection. At step 512, before detectable induction, we reverse `inv_freq` so that the fastest rotation sits at index 7. Across two seeds and three conditions, all six runs follow the frequency (the permuted rows of Figure 3). Without ablation, the largest share moves to *index 7*; ablating that index recruits *index 6*, now the second-fastest frequency, with shares of 0.922 and 0.926. Ablating index 0, now the slowest frequency, leaves index 7 dominant.

Together, these allocation measurements support selection by frequency rather than index. Functional dependence on the band-2, band-3, and permuted replacements remains untested.

5.2 Downstream connectivity

In the separate 70M step-512 runs, band-1 knockout in a single previous-token head reduces accuracy by 0.480 (sd 0.032; 83.8% of the global effect). This output-level effect localizes dependence to the head's rotation, but does not by itself identify the downstream connection. We therefore patch individual residual-stream sources feeding the induction head. Each patch substitutes a source's value from a nonrepeated corrupt sequence, recomputing either the key or query input while holding the other input clean. Path patching (Wang et al., 2023; Goldowsky-Dill et al., 2023) identifies that same head as the dominant source of the induction head's key input in both control and downshifted models, with MLP0 dominant on the query side. The key-source intervention reproduces 76–96% of the attention-score drop caused by corrupting both inputs. In these resumed runs, spectral recruitment thus coexists with a retained dominant previous-token-to-induction key connection. The

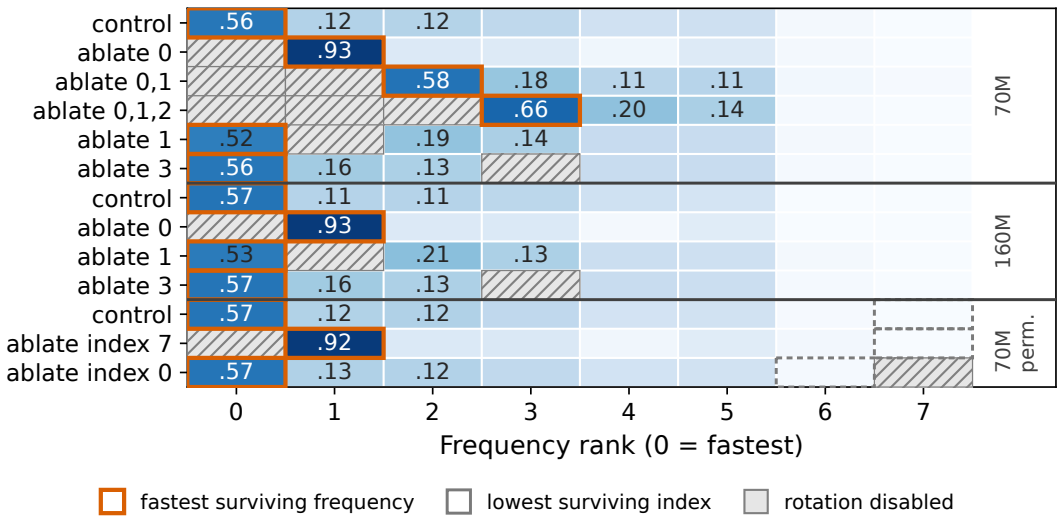


Figure 3: Previous-token margin allocation follows frequency, not dimension-pair index. Endpoint shares are averaged over seeds and ordered from fastest to slowest frequency; this order follows index in stock RoPE and reverses it in the permuted 70M rows. Colors show positive shares; values at least 0.10 are printed. Hatching marks ablated bands; orange outlines mark the fastest-surviving-frequency rule's prediction, and gray outlines mark the lowest-surviving-index alternative. These are allocation measurements, distinct from the E1/E2 causal profiles in Figure 2.

response is attention, not output logits; the patches do not establish full-path mediation or extend the connectivity result to the E1/E2 endpoints (Appendix A.6).

6 Discussion

An implementation choice can become an endpoint dependency. The combined evidence separates three observations: band 0 is preferred in the controls, the trained circuit is sensitive to its removal, and induction can nevertheless form without that rotation. The initialization experiment excludes a prior band-0 training history for the observed 70M formation. Two complementary results make this distinction concrete: endpoint knockouts establish acquired band-1 dependence, while the ladder and permutation constrain which frequency receives the previous-token margin. This is flexibility in how positional information is used, not evidence that positional information is dispensable. An account based only on a privileged dimension-pair index cannot explain the permutation result. Why optimization favors the fastest surviving rotation, and whether the deeper replacements acquire comparable functional dependence, remain separate questions.

Control and band-0-off E1/E2 endpoints have close induction accuracies and overall text losses, despite sharply different band-specific knockout profiles (Tables 2 and 3). Aggregate performance alone would miss this difference in internal dependence and does not establish agreement in token-level predictions.

The text controls also constrain a generic-fragility explanation. In the band-0-off models, band-3 and layer-matched random-head knockouts have much smaller effects than band-1 knockout in the selected heads (Table 3). The impairment is therefore concentrated in the recruited rotation and functionally selected heads within the tested comparisons.

Recovery windows do not identify formation constraints. The same distinction changes the interpretation of post-formation repair. In a three-seed 12k-step rerun, the control-minus-ablated induction-accuracy residual falls from 0.508 at 500 steps to 0.123 at 12k; we do not infer an asymptote from that trajectory. The original short run, its extensions, and a separate before-onset comparison are reported with their distinct protocols in Appendix A.9. A finite repair residual describes adaptation of an existing solution within a particular budget. It does not establish that an alternative solution cannot be learned from initialization.

7 Related work

Induction circuit formation. Elhage et al. (2021) described the induction circuit's two-stage computation, and Olsson et al. (2022) connected its emergence to changes in in-context learning. Controlled studies analyze the learning dynamics of induction and related in-context mechanisms (Bietti et al., 2023; Singh et al., 2024; Reddy, 2024). We study how the positional component is implemented in RoPE and whether its learned band preference constrains formation.

RoPE frequency structure. Prior work links faster rotations to local position and slower rotations to content or longer distances (Hong et al., 2024; Barbero et al., 2025), formalizes positional versus symbolic attention (Urrutia et al., 2026b;a), and incorporates that distinction into positional encodings (Chen et al., 2025). Barbero et al. (2025) also train Gemma 2B models from scratch with high- or low-frequency rotations removed and evaluate perplexity. Li (2026) reports spectral reconfiguration under constrained query–key operators in toy-scale training, providing prior evidence that a learned spectral signature need not be necessary for formation. We connect individual-frequency recruitment during training to acquired functional dependence in public language models. Our evidence combines band-0-off formation from a public 70M initialization, 160M continuation, and head-set rotation knockouts on synthetic induction and held-out text; ladder and permutation tests separately distinguish frequency-based from index-based allocation.

Reconfiguration after positional intervention. Transformers can learn positions without explicit positional embeddings (Haviv et al., 2022; Kazemnejad et al., 2023); DroPE adapts pretrained language models after removing them (Gelberg et al., 2026). Backup heads, self-repair, and fine-tuning also illustrate circuit reconfiguration (Wang et al., 2023; McGrath et al., 2023; Prakash et al., 2024). We distinguish inference-time compensation, repair through further training, and formation under intervention. E1 tests formation without prior band-0 training; the step-512 experiments examine frequency choice and downstream connectivity under resumed training.

8 Limitations

Initialization evidence covers one public 70M initialization and three data-order seeds; 160M starts at step 512. Head-set causal evidence covers both scales, but the ladder, permutation, and key-path patching remain resumed 70M tests. The deeper-rung and permuted replacements are supported by allocation measurements, not matched endpoint knockouts. Head identity need not be preserved, and the 11-model survey is descriptive attribution. OLMo accuracy declines in both conditions for an unresolved reason, leaving only directional evidence (Appendix A.8). This limits cross-family generalization; the formation and causal claims rest on the separate Pythia experiments.

Matched controls isolate the intervention within each protocol, but the schedules, repeated buffers, and optimization budgets limit generalization to other training regimes. Natural-text evaluation uses one corpus and windows of at most 512 tokens; exact-document exclusion covers new training, not near-duplicates or E2's public pretraining history.

Our margin is mean-referenced, not an argmax decision margin; signed shares are unstable near zero and baseline-sensitive even at trained endpoints, particularly at 160M. Knockouts establish dependence, not exclusive mediation by the full induction path in natural text. Margin inflation remains unexplained, finite repair does not identify an asymptote, and the scaling checks measure margin and norm adaptation rather than long-context quality.

9 Conclusion

In the tested 70M protocol, induction forms from initialization without band 0's rotation and acquires dependence on band 1. Head-set knockouts at 70M and 160M verify band 1's contribution to synthetic induction and natural-text prediction. Separate 70M frequency interventions support fastest-surviving-frequency allocation rather than fixed-index selection. *Band downshift* thus links frequency-based recruitment to acquired dependence, illustrating why strong endpoint dependence need not imply developmental necessity.

AI USE STATEMENT

We used generative AI tools for editorial critique, consistency checking, and copy-editing. The authors made the experimental and interpretive decisions, reviewed every suggested change, and verified all claims, numbers, and references against the archived artifacts or primary records. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work analyzes public model checkpoints and text corpora and involves no human participants. Our experiments concern circuit dependence and formation, not deployment safety; they do not remove biases or other risks inherited from the underlying models and data. The planned artifact release includes code, aggregate results, and research checkpoints, but not reproduced corpus text.

REPRODUCIBILITY STATEMENT

All code, per-run JSON artifacts, figure scripts, and circuit-analysis weights will be released. Included tests verify additive decomposition and attention reconstruction. Appendices A.1 and A.3 specify protocols, data construction, and uncertainty. E1/E2 tables and figures are generated from completed endpoint artifacts with protocol, code, data, and weight-hash checks.

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92408–92438, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e6d58fc68c0f3c36ae6e0e64478a69c0-Paper-Conference.pdf.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a Transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 1560–1588. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=3X2EbBLNsk.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv:2306.15595, 2023.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23044–23056. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1123. URL https://aclanthology.org/2025.acl-long.1123/.

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.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. Extending the context of pretrained LLMs by dropping their positional embedding. In

International Conference on Learning Representations (ICLR), 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/62133b95efaa6f10a5395eb10ca087cd-Abstract-Conference.html.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. arXiv:2304.05969, 2023.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, 2022. URL https://aclanthology.org/2022.findings-emnlp.99/.

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5877–5888. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.338. URL https://aclanthology.org/2024.findings-emnlp.338/.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 24892–24928. Curran Associates, Inc., 2023. URL https://dl.acm.org/doi/10.5555/3666122.3667204.

Hengyu Li. Fingerprint, not blueprint: How positional schemes set the default spectral algebra of attention. arXiv:2607.06621, 2026.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. arXiv:2307.15771, 2023.

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/540a3bf4863672cdaebfed69fa2ce335-Paper-Conference.pdf.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662, 2024. URL https://dl.acm.org/doi/abs/10.5555/3692070.3693925.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2 OLMo 2 Furious. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Felipe Urrutia, Juan José Alegría, Cinthia Sanchez Macias, Jorge Salas, Cristian B. Calderon, and Cristobal Rojas. Positional versus symbolic attention heads: Learning dynamics, RoPE geometry, and length generalization. arXiv:2605.31558, 2026a.

Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling positional and symbolic attention behavior in transformers. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=V38yAoqddQ.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

A Appendix

A.1 Initialization and matched-endpoint protocols

Starting points and matched conditions. E1 uses EleutherAI/pythia-70m at public step0, documented by Pythia as the initialization checkpoint.[1] Its pinned model revision is 87c8ac0ba129db9b084502d341c6ca8dec0fa01f; E2 uses EleutherAI/pythia-160m at step512, revision 700e8c2b860f97ff8fc55b237e75e92ada629553. E1 and E2 each comprise three paired data-order seeds (0, 1, 2), with the same public starting weights within each experiment. They are not independent weight initializations. The intervention replaces band 0's rotation with identity before the first optimizer update of each intervened run and preserves it throughout training and endpoint evaluation. Numerical no-gradient checks that temporarily toggle masks do not update weights. These endpoints are distinct from the resumed models used for the ladder, permutation, specificity, and path tests.

Optimization and budget. Both use fresh AdamW (weight decay 0.01, betas 0.9/0.95), sequence length 512, microbatch 32 and four-way accumulation for effective batch 128. E1 takes 32,000 updates, linearly warms the learning rate to 10^{-3} over 1,000 updates, then applies cosine decay to one tenth of that peak; gradient norm is clipped at 1. E2 takes 8,000 updates at constant learning rate 3×10^{-4}, without warmup or clipping. Each E1/E2 update contains 65,536 token positions, one thirty-second of the original Pythia update's 2,097,152. E1 cycles 8,000 materialized batches four times (524,288,000 buffer positions, 2,097,152,000 training positions per arm). E2 uses the first 2,000 batches of that seed's buffer four times (131,072,000 buffer positions, 524,288,000 additional training positions). Repeated positions are not distinct training data, and the two models' total histories and schedules are not matched across model sizes. Model states, AdamW moments, scheduler position, data index, RNG state, and intervention settings are checkpointed every 4,000 updates in E1 and 2,000 in E2. Probes run every 500 updates and at the starting checkpoint; the validation trajectory uses the first 128 of the 512 validation documents, never the test sample. All settings were fixed before primary optimization; all prescribed seeds and budgets are reported.

Corpus and exact-document exclusion. The corpus is monology/pile-uncopyrighted, pinned to revision 3be90335b66f24456a5d6659d9c8d208c0357119, with its explicitly published training, validation, and test files. Tokenization uses the Pythia-70M tokenizer at pinned revision a39f36b100fe8a5377810d56c3f4789b9c53ac42. We sample 512 validation and 2,048 test documents, require at least 64 tokens, and keep one deterministic document-hash-selected window of at most 512 tokens from each. Original UTF-8 document SHA256 hashes remove exact duplicates within and between these samples, and their union is excluded from every E1/E2 training buffer. Training data orders use seeds 0/1/2 and a shuffle buffer of 10,000 documents. This excludes

[1] https://github.com/EleutherAI/pythia

exact held-out documents from all E1 learning and from E2's continuation, not from the public checkpoint's earlier pretraining; it does not guarantee near-duplicate exclusion. Dataset/model revisions, buffer hashes, test masks, and per-run settings are recorded in the artifacts.

Head selection and causal evaluation. The independent calibration probe uses seed 7 and 16 repeated sequences; previous-token attention ranks heads, with ties broken by layer and head index. The primary head set is top-3, and top-1/6/12 are secondary analyses fixed in advance. For each size, ten random sets use seed $271828 + 1000k + i$, matching the selected head count in each layer and excluding selected heads and the chosen induction head. If such a set cannot be constructed, the random comparison is missing rather than relaxed: this affects top-12 in the two E1 seed-1 arms, but no primary top-3 comparison. We evaluate every rotary band globally and in each selected set, bands 0/1 in random sets, and bands 0/1 in the induction head. The probe uses seeds 11/22/33/44, 64 sequences per seed, a 64-token block repeated twice, and a vocabulary of the first 1,000 token IDs; argmax uses the full model output vocabulary. Accuracy reduction and correct-token NLL increase are paired across the same probe draws. Fractions of a global accuracy drop are interpreted only when that drop is at least 0.01; ratios are computed per training seed before averaging. All endpoint interventions retain the original global training mask. Redundant band-0 knockouts on band-0-off models reuse the native measurement after numerical equivalence checks, and are not counted as independent causal evidence.

Natural-text targets and uncertainty. At logit position p, a target is repeated if the triple (x_{p-1}, x_p, x_{p+1}) equals (x_{q-1}, x_q, x_{q+1}) for some q with $p - q \geq 8$ in the same window: the two-token context and its correct continuation have both occurred before. Matched repeated and nonrepeated targets balance target token ID and $\lfloor p/64 \rfloor$ across the test pool, using deterministic hash ordering within each stratum. Matching is across the pool, not necessarily within each document; target masks are fixed before model evaluation and do not depend on ablation outcomes. We evaluate native, global band 0/1, top-3 band 0/1/3, and ten layer-matched random top-3 band-1 conditions. Token-weighted NLL is reported on all, repeated, matched repeated, and matched nonrepeated targets. For paired effects, 2,000 bootstrap draws (seed 314159) resample entire documents with replacement, using the same document multiplicities for every group and paired intervention. Each draw divides the resampled loss difference by the corresponding resampled target count; repeated-minus-nonrepeated effects are then differenced within that draw. Random-control adjustment averages each document's ten random-head loss sums before the same paired bootstrap. Percentile 95% intervals condition on one trained seed; sample standard deviations in the main tables instead summarize the three data-order seeds.

Execution and validation. Each condition uses one A100 80GB GPU, PyTorch 2.6.0, Transformers 4.53.3, Datasets 3.6.0, and NumPy 1.26.4; training uses bfloat16 autocast with float32 parameters, and endpoint evaluations use float32 eager attention. We verify identity rotations, unchanged nontarget heads/bands, agreement of global and all-head masks, attention reconstruction, production-shape gradients, and exact optimizer/RNG resumption. An archived 70M endpoint reproduces induction accuracy 0.6975708 under this environment. A separate storage check exposed loss of a final unsynchronized metadata replacement; two setup runs were stopped and preserved, and all primary seeds restarted from their prescribed public checkpoints and identical data with synchronized unique staging files. Their results are not pooled with the primary suite, and the scientific protocol was not changed by this correction. All twelve primary endpoints completed the fixed training and evaluation budgets. Checkpoint hashes are verified before endpoint evaluation; the paper generator checks completion, protocol and source hashes, paired identities, and per-document NLL aggregation before updating the E1/E2 tables.

A.2 Seedwise initialization and matched-endpoint results

The initialization-trained 70M runs select L2H1 in both conditions for seeds 0/1, but seed 2 selects L2H3 after ablation against L2H1 in control. All E2 runs select L3H0 as their top previous-token head, while their top-3 sets can differ. Head identities from the separate earlier runs are reported in Appendix A.13.

Both scales show positive matched repetition selectivity in every seed, including after random-head adjustment (Table 5). This does not make the effect exclusive to repetition: matched nonrepeated tar-

Table 4: All E1/E2 endpoints. s_1 is the selected previous-token head's band-1 margin share, reported under both baselines; its size is not treated as causal evidence. The last column is the absolute accuracy reduction from the prespecified top-3 band-1 knockout, not the top single head. E1 starts at initialization and E2 at step 512; seeds vary data order while keeping the starting weights fixed within each model.

Model	Seed	Training	Acc.	PT head	s_1 (unif.)	s_1 (soft.)	Top-3 b_1 drop
70M	0	Control	0.785	L2H1	0.129	0.067	0.069
70M	1	Control	0.774	L2H1	0.129	0.111	0.088
70M	2	Control	0.784	L2H1	0.131	0.105	0.118
70M	0	Band 0 off	0.761	L2H1	0.919	0.785	0.708
70M	1	Band 0 off	0.783	L2H1	0.913	0.921	0.693
70M	2	Band 0 off	0.768	L2H3	0.917	0.765	0.713
160M	0	Control	0.782	L3H0	0.113	0.117	0.020
160M	1	Control	0.772	L3H0	0.115	0.059	0.021
160M	2	Control	0.781	L3H0	0.123	0.079	0.022
160M	0	Band 0 off	0.771	L3H0	0.919	0.352	0.452
160M	1	Band 0 off	0.761	L3H0	0.925	0.200	0.494
160M	2	Band 0 off	0.796	L3H0	0.924	0.545	0.480

Table 5: Natural-text effects of top-3 band-1 knockout in every band-0-off endpoint (nats/token). R/N denote the matched repeated/nonrepeated targets. Intervals are paired-document bootstrap 95% intervals conditional on the indicated trained seed. The final column subtracts the mean random-head intervention before forming the repeated-minus-nonrepeated contrast.

Model	Seed	Δ R	Δ N	R minus N [95% CI]	Random-adjusted CI
70M	0	1.310	0.311	0.999 [0.962, 1.037]	[0.954, 1.028]
70M	1	1.265	0.322	0.943 [0.909, 0.978]	[0.873, 0.941]
70M	2	1.402	0.340	1.062 [1.024, 1.103]	[0.960, 1.034]
160M	0	0.841	0.184	0.658 [0.630, 0.688]	[0.607, 0.664]
160M	1	0.834	0.161	0.673 [0.642, 0.705]	[0.627, 0.689]
160M	2	0.834	0.153	0.682 [0.654, 0.712]	[0.634, 0.691]

gets also worsen. Table 6 reports native LM-quality differences separately from these acute knockout effects.

A.3 Complementary checkpoint-resumption configurations

Training. The complementary experiments resume public checkpoints with a fresh AdamW (weight decay 0.01, betas 0.9/0.95) and sequence length 512; their configurations differ from E1/E2 (Appendix A.1). Batch size is 128 sequences unless noted. The 12k-step OLMo-2-1B runs use micro-batch 16 with four-step gradient accumulation, an effective batch of 64; the earlier 16k-step OLMo specificity checks, budget ladder, and original disrupt-and-repair run use batch 16 without accumulation. Learning rates are constant at the model-specific values in Table 7 (2e-5 for runs resuming the trained endpoint). Text is tokenized once into a finite buffer and cycled. The standard Pythia setting stores 2,000 batches, or 131.1M token positions at batch 128; an 8k-step run traverses it four times and the 12k recovery rerun six times. The three OLMo-mix-1124 data-order seeds per condition store 4,000 batch-16 batches (32.8M token positions) and, with four-way accumulation, traverse each buffer 12 times in 12k optimizer steps.[2] Within a data-order seed, control and ablation conditions use the same materialized examples in the same order. Table 7 lists the resume settings per experiment.

Probes. Induction accuracy uses a repeated block of length 64, with tokens drawn uniformly from the first 1,000 vocabulary ids, averaged over 4 probe seeds of 64 sequences each. The original 2,500-step disrupt-and-repair run predates that probe and uses a single batch of 16. The public Pythia checkpoint grids use a single batch of 16 sequences for margin attribution; the auxiliary 1.4B

[2]The matched corpus is `allenai/olmo-mix-1124`: `https://huggingface.co/datasets/allenai/olmo-mix-1124`.

Table 6: Native overall held-out NLL for every matched pair. The difference is band-0-off minus control; lower NLL is better. These are different trained endpoints, not acute knockouts, and all share the same 2,048-document evaluation sample.

Model	Seed	Control NLL	Band-0-off NLL	Paired difference
70M	0	3.36384	3.37175	+0.00791
70M	1	3.36405	3.35836	-0.00568
70M	2	3.35861	3.35602	-0.00259
160M	0	3.61083	3.61692	+0.00609
160M	1	3.60510	3.61223	+0.00712
160M	2	3.59478	3.59668	+0.00190

grid that supplies the step-143k endpoint uses 4. The complementary downshift, specificity, ladder, permutation, scaling, and OLMo trajectories use 4 sequences for margin attribution, with probe seed 7. Public-endpoint head localization and the separate 70M crossover tests use 8 sequences for attribution; E1/E2 use 16 for attribution and head calibration (Appendix A.1). The standalone global band-sweep and distance-by-context diagnostics average 60 probe batches of 16 sequences. Probe cadence is every 50 steps in the original 2,500-step disrupt-and-repair run, 250 in the before-onset comparison runs, 500 in the permutation, scaling, ladder, specificity, 160M downshift, and budget-ladder runs, and 1,000 in the 70M downshift and OLMo runs.

Seeds. Data-order seeds 0, 1, and 2, plus 3 and 4 for the five-seed condition.

Table 7: Resume configurations. Seeds per condition: "1+3" = one Pile-proxy + three OLMo-mix-1124 runs, each with a matched control. The earlier batch-16 OLMo specificity conditions remove band 1 or 20 on the Pile proxy; the lower-learning-rate check uses OLMo-mix-1124 and effective batch 64. "3/5" = three seeds at 70M, five at 160M. Full per-run JSON configurations, probe settings, and per-step trajectories ship with the code release.

Experiment	Model	Resume ckpt	LR	Steps	Seeds
Downshift, specificity, ladder	Pythia-70M	step512	5e-4	8,000	3/cond.
Downshift, specificity	Pythia-160M	step512	3e-4	8,000	3/cond.
Downshift (2 corpora)	OLMo-2-1B	stage1-300	5e-4	12,000	1+3
Specificity (bs 16)	OLMo-2-1B	stage1-300	5e-4	16,000	1/cond.
Lower-LR check	OLMo-2-1B	stage1-300	1e-4	12,000	1/cond.
Permutation control	Pythia-70M	step512	5e-4	8,000	2/cond.
Disrupt-and-repair, original	Pythia-70M	main (final)	2e-5	2,500	3/cond.
Disrupt-and-repair rerun	Pythia-70M	main (final)	2e-5	12,000	3/cond.
Budget ladder (bs 16)	Pythia-70M	main (final)	2e-5	16,000	3/cond.
Before-onset comparison	Pythia-70M/160M	step512	5e-4/3e-4	8k/12k	3/5
Scaling (PI, NTK)	Pythia-70M	step512	5e-4	8,000	3/cond.

The abbreviated OLMo resume checkpoint in Table 7 is public revision `stage1-step300-tokens1B`.

Pythia-1.4B is the exception to the grid settings above: that grid stops at step 96k, and its step-143k endpoint comes from an earlier grid measured at batch 4. The two 1.4B grids share seven checkpoints, the last at step 32k, and across the five post-transition ones their tail shares agree to within 0.006, which is the only cross-batch comparison our runs support. The two pre-transition shared checkpoints differ by 0.152 and 0.059, where the top induction head scores below 0.04 and the two batches can select different heads. The earlier 70M circuit analyses use saved endpoint weights in float32 with eager attention: the crossover knockout uses three downshifted and three control seeds; path patching uses three downshifted seeds and one control seed; both add the public final checkpoint. The path-patching decomposition reconstructs the residual stream with maximum absolute error below 1.4×10^{-5} on every run.

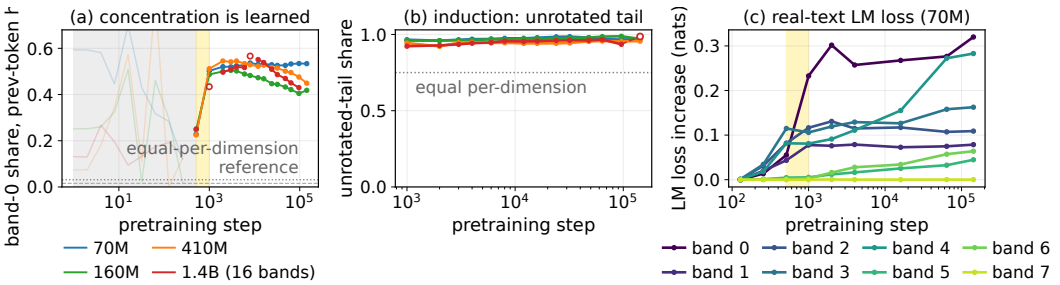


Figure 4: Public-checkpoint localization across scale; panels (a) and (b) share the model legend. **(a)** Band-0 share of the top previous-token head's offset-1 margin; faint traces mark previous-token scores below 0.30, where the share is not interpreted, while open markers denote a change in selected head. The gray background covers steps before 512; yellow in (a) and (c) marks the 512–1000 checkpoint interval. The dotted and dashed lines mark the equal-per-dimension references $2/64 = 0.03125$ at 70M–410M and $2/128 = 0.015625$ at 1.4B. **(b)** Unrotated-tail share of the top induction head's matching margin, against the even-split baseline of 0.75; the 1.4B endpoint comes from a separate grid. **(c)** LM-loss increase from ablating each band at 70M.

A.4 Public-checkpoint localization and diagnostics

The public-endpoint localization test uses one fixed comparison head per model: L2H2 at 70M, L1H5 at 160M, and L1H1 at 410M. Each is the 18th entry in a layer-major, head-minor list after excluding the selected previous-token and induction heads. This deterministic comparison is distinct from the random single-head controls in the saved-endpoint crossover test and the layer-matched random head sets in E1/E2.

Across the public checkpoint grids of all four Pythia scales, the previous-token head's band-0 share rises sharply across the induction transition, and band 0 stays the top band to the end of training (Figure 4). At 70M the share moves from 0.23 to 0.50 between steps 512 and 1000, the same interval in which induction accuracy moves from 0.003 to 0.761. The equal-per-dimension reference is $2/64 = 0.03125$ for each rotating pair, with 0.75 assigned to the unrotated tail; an equal split over only the rotating pairs would omit the tail from the denominator. After the transition it holds at 70M but drifts down 0.09 to 0.14 from its peak at the other three scales, which we do not explain. The rise survives both margin baselines, though whether the pre-transition state is genuinely uniform is baseline-dependent, and we claim no uniformity. A third measurement, independent of both the decomposition and the induction probe, agrees: from step 1000 onward, ablating band 0 costs more real-text language-modeling loss than any other band, where at step 512 it ranked fourth. Because the public checkpoints jump from step 512 straight to step 1000, we can place both rises in the same interval but not order them, consistent with Li (2026).

At the final 70M checkpoint, Spearman correlations across all 48 heads are $\rho = 0.716$ between previous-token attention and band-0 share, $\rho = 0.761$ for band 1, and $\rho = -0.392$ between induction attention and band-0 share. These descriptive correlations use the saved batch-16 checkpoint grid and average ranks for ties. Shares use the corresponding target offset: 1 for previous-token attention and $T - 1$ for induction attention. Both fast bands correlate positively with previous-token attention; the association alone does not single out band 0.

A magnitude-only diagnostic. For comparison, we compute the mean absolute score contribution $\mathbb{E}|c_b|$. This is inspired by, but not identical to, the per-pair query and key norm diagnostics of Barbero et al. (2025). On Pythia-70M's top induction head it ranks bands 6 and 7, the slowest, highest among rotating bands, yet disabling either rotation changes induction accuracy by at most 0.002. Magnitude, target contrast, and rotation dependence answer different questions. In particular, a near-identity slow rotation can have a small ablation effect even when its dimensions carry useful content; the null does not establish that those dimensions are unused. The margin contrast measures target matching, but its baseline sensitivity (Table 4) likewise requires separate causal checks.

Table 8: Original step-512 downshift and specificity experiments. Takeover is the recruited band's uniform-baseline margin share (mean ± sd, three seeds). The final column gives band-0 share in the matched controls, except in non-preferred-band specificity rows, where it describes the intervened run itself.

Model	Condition (n)	Recruited band	Takeover share	Band-0 share
Pythia-70M	ablate $\{0\}$ (3)	1	0.926 ± 0.003	0.555–0.564
Pythia-70M	ablate $\{1\}$ (3)	none	—	0.513–0.522
Pythia-70M	ablate $\{3\}$ (3)	none	—	0.556–0.559
Pythia-160M	ablate $\{0\}$ (3)	1	0.925 ± 0.004	0.568–0.573
Pythia-160M	ablate $\{1\}$ (3)	none	—	0.528–0.537
Pythia-160M	ablate $\{3\}$ (3)	none	—	0.562–0.572

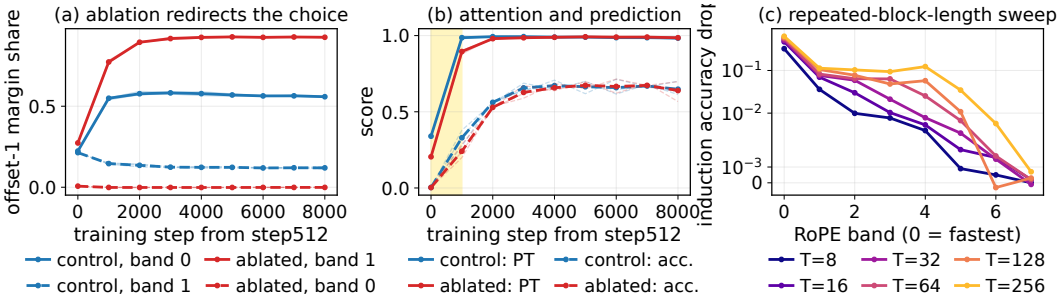


Figure 5: Panels (a) and (b) show the original Pythia-70M step-512 downshift runs, distinct from E1; thick lines are three-seed means, thin traces individual runs. **(a)** Previous-token offset-1 margin shares on bands 0 and 1. **(b)** Previous-token attention scores and output-level induction accuracy. **(c)** A separate repeated-block-length sweep at the public final checkpoint (symlog); each point averages 60 probe batches of 16 sequences. Varying block length T changes both the induction target offset $T-1$ and total sequence length $2T$; it does not isolate distance from context length. Band 0 leads at every tested block length.

A.5 Complementary step-512 downshift and specificity experiments

These are the original resumed runs, not E1 or E2. Band 0 is disabled at Pythia-70M step 512, where induction accuracy is 0.003, followed by 8k updates on Pile text. The same head (L2H1) that carries band 0 in the control runs develops its offset-1 margin on band 1, whose share rises from 0.27 to 0.77 by step 1k and ends at 0.926 across three data-order seeds (sd 0.003; Figure 5). Control runs concentrate on band 0 (0.555–0.564), close to the public final checkpoint's 0.534; agreement in band assignment constrains but does not eliminate resume-protocol confounds. Endpoint behavior is similar to controls: induction accuracy 0.639 against 0.648 and head score 0.986 against 0.984, three-seed means (per-seed values in Appendix A.13). Cross-entropy on a separately shuffled Pile sample is 3.906 downshifted against 3.910 control, every paired-seed difference within 0.014 of zero. That sample came from the training split without document-level exclusion and is not a guaranteed disjoint test set. Held-out prediction is assessed with E3 (Section 4.3).

The original Pythia-160M runs give band-1 takeover 0.925 (sd 0.004), with the same specificity pattern (Table 8). One downshifted run selects L3H2 against the controls' L3H0, so even this original replication concerns band assignment rather than universal head-identity preservation. The exploratory OLMo extension is reported separately in Appendix A.8, including its unresolved behavioral decline.

A.6 Single-head crossover and path patching at 70M

These circuit tests use the saved original 70M resumed endpoints: three downshifted and three control runs for the knockout, three downshifted and one control for path patching, plus the public final checkpoint. Knocking out band 1 in the downshifted previous-token head alone costs 0.480 induction accuracy (sd 0.032; 83.8% of the global band-1 effect). The public checkpoint's band-0 cost

Table 9: Original 70M single-head crossover and key/query-path evidence, distinct from the E1/E2 head-set tests. Accuracy change is intervention minus native; negative means impairment. Key/query columns give each source's fraction of the attention-score drop from corrupting both inputs. Knockout values use three seeds for resumed control and downshift; path values use one control and three downshift seeds. Random single-head knockout nulls over 6–10 other heads are centered at zero (max $|\Delta| = 0.054$).

Run	Band	Δ acc.	% global	Key (PT head)	Query (MLP0)
Public checkpoint	0	-0.492	86.9%	96.4%	52.6%
Resumed control	0	-0.410 ± 0.010	69–77%	76.2%	81.2%
Downshifted (n=3)	1	-0.480 ± 0.032	83.8%	83.8–85.4%	80.8–83.4%
Downshifted, other bands	$\neq 1$	$\lvert\Delta\rvert \leq 0.005$	—	—	—
Induction head, both arms	0, 1	$\lvert\Delta\rvert < 0.001$	—	—	—

is 0.492, and the saved resumed controls' cost is 0.410 (Table 9). Every other band in the downshifted previous-token head costs at most 0.005, and induction-head knockouts of bands 0 and 1 change accuracy by less than 0.001 in both arms. Within these tested interventions, the results support functional recruitment of band 1 in the previous-token head, not merely a change in attribution share.

Path patching uses the induction head's target attention score as the response and an independently sampled random sequence without the repeated-block structure as corruption. Pythia's parallel residual decomposes the stream feeding that head into 28 input-dependent source terms plus three constant attention-output biases; their sum reconstructs the stream with maximum absolute error below 1.4×10^{-5}. For each input-dependent source, we replace its clean residual contribution with its corrupted counterpart, recompute either the key or query input, and hold the other input clean. The previous-token head is the dominant key source (76–96% of the clean-to-both-inputs-corrupted attention drop in every run); MLP0 is the dominant query source. These fractions are not additive pathway shares. They identify a shared dominant source in these runs, not an unchanged circuit; we did not propagate patches to logits or perform causal scrubbing.

A.7 The full-rotary model survey

The 11 models below rotate every dimension and have no unrotated tail. The analogous channel group is *quasi-static*: its bands undergo little direct phase change over the queried offset, although their query and key content can still depend on context. We group a band as quasi-static when its phase rotation over the queried offset is below 0.1 radians. Figure 6 plots the induction-head share for all 15 models; this table adds the head identities, the head scores, and the share of the previous-token head's offset-1 margin that the eight fastest bands hold. Every model shows the same descriptive split as Pythia: the selected induction head's matching contrast concentrates where direct phase change is small, and the selected previous-token head concentrates on the fast end. The quasi-static bands take 0.80–0.99 of the induction head's matching margin against even-split baselines of 0.50–0.59. The eight fastest bands contribute 0.58–0.88 of the selected previous-token head's offset-1 margin.

The full-rotary survey uses one batch of four sequences, each formed by repeating a 64-token block drawn from the first 1,000 token IDs with probe seed 7, and runs in bfloat16 for memory. The reconstruction check compares attention entries at one middle layer (zero-indexed $\lfloor L/2 \rfloor$ for depth L); its maximum absolute error stays within 0.047 for every model. The six models we also ran in float32 (Qwen3-0.6B/1.7B/8B/14B, SmolLM3-3B, Granite-3.1-8B) reconstruct to within 0.001, select the same two heads, and agree with the bfloat16 shares to within 0.001. Head-scoped ablation was not run at these scales; these are attribution measurements, not causal ones.

A.8 OLMo: directional attribution with declining probe performance

OLMo-2-1B (Team OLMo et al., 2025) differs from Pythia in organization, corpus, tokenizer, and architecture. Starting from a checkpoint before detectable induction, band-0 ablation shifts the margin onto band 1 on both a Pile proxy and OLMo-mix-1124, while the matched controls concentrate

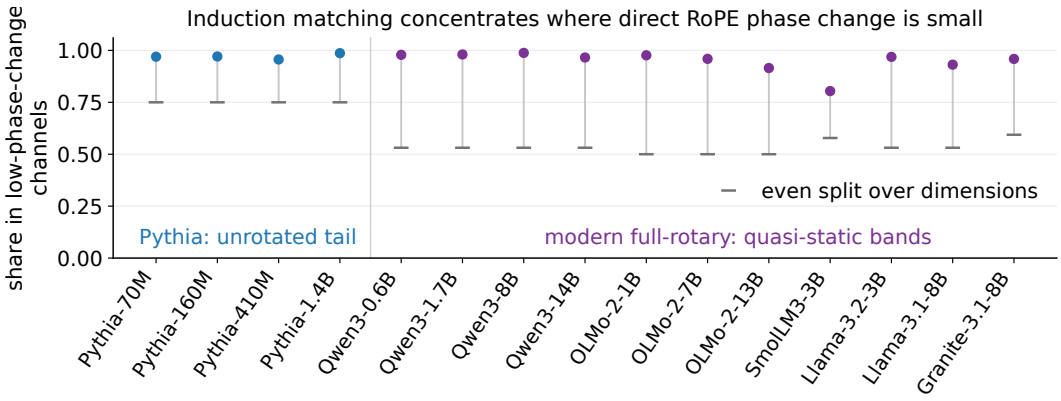


Figure 6: For every model measured, the selected induction head's mean-referenced matching contrast concentrates in channels with little or no direct RoPE phase change. The dot is that channel group's share of the top induction head's matching margin—the unrotated tail in Pythia, the quasi-static bands in the full-rotary models—and the tick below it is that model's own even-split baseline. This attribution survey is descriptive rather than an ablation test; head identities for the 11 full-rotary models are in Table 10.

Table 10: The full-rotary model survey (11 models, five families, all 64 bands). "Quasi-static" is the share of the selected induction head's matching margin carried by bands with less than 0.1 radians of phase change over the queried offset; "fastest 8" is the selected previous-token head's corresponding fast-band share. "Recon." is the maximum absolute attention-entry reconstruction error in the checked middle layer, not a maximum across all layers.

	induction head			previous-token head			
Model	head	score	quasi-static	head	score	fastest 8	recon.
Qwen3-0.6B	L16H14	0.97	0.979	L15H3	0.82	0.857	0.037
Qwen3-1.7B	L16H14	0.99	0.980	L15H3	0.81	0.875	0.035
Qwen3-8B	L22H23	0.98	0.988	L1H4	0.83	0.739	0.032
Qwen3-14B	L28H32	1.00	0.966	L16H0	0.84	0.740	0.046
OLMo-2-1B	L8H1	0.92	0.976	L7H10	0.64	0.575	0.020
OLMo-2-7B	L9H2	0.97	0.959	L7H8	0.83	0.865	0.015
OLMo-2-13B	L27H26	0.94	0.915	L9H30	0.76	0.734	0.016
SmolLM3-3B	L23H9	0.96	0.804	L0H4	0.78	0.801	0.011
Llama-3.2-3B	L14H22	0.95	0.969	L13H6	0.66	0.707	0.013
Llama-3.1-8B	L15H30	0.98	0.931	L14H26	0.68	0.758	0.011
Granite-3.1-8B	L12H27	0.97	0.959	L35H4	0.76	0.832	0.001

on band 0 (Table 11). The separate 16k-step, batch-16 specificity conditions removing band 1 or band 20 retain band-0 preference (one seed each). The OLMo takeover shares are smaller than at Pythia-70M and are flat over their last 5,000 steps; a finite plateau does not establish that adaptation has finished. The standard-learning-rate pairs in Table 11 select different previous-token heads across conditions; the lower-learning-rate pair instead selects L4H13 in both arms (Appendix A.13).

At the standard learning rate, induction accuracy peaks between 3k and 7k steps and falls 2.5- to 5.4-fold by 12k in both conditions, for a reason we could not establish. The archived lower-learning-rate check on the matched corpus (one seed per condition, 10^{-4} instead of 5×10^{-4}) also declines: control accuracy falls from 0.266 to 0.108 and band-0-ablated accuracy from 0.273 to 0.109, both peaking at step 4k. At the standard learning rate, the six OLMo-mix-1124 runs retain best-head previous-token attention scores of 0.902–0.982 and induction attention scores of 0.836–0.901 at step 12k, despite the behavioral decline; these heads are reselected at each measurement, not tracked as a fixed circuit. Neither corpus matching nor this single lower-learning-rate check eliminates the decline, and strong attention patterns alone do not establish a functional copying mechanism. The archived trajectories lack held-out language-model loss, correct-token probabilities, and restricted-vocabulary probe predictions, so they cannot distinguish general language-model degradation from

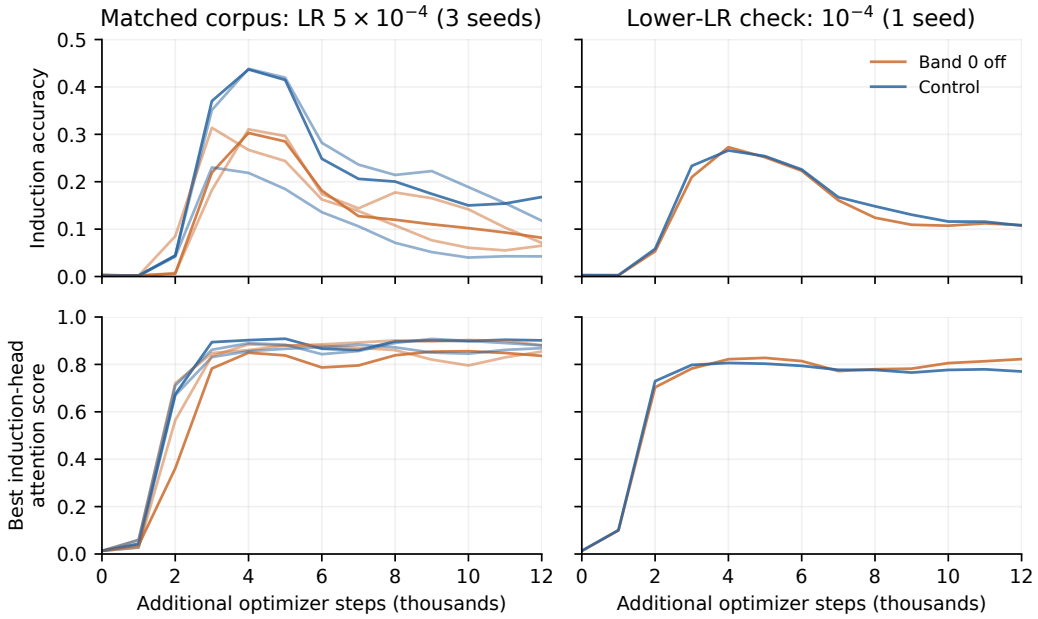


Figure 7: Archived OLMo runs on OLMo-mix-1124. Top: output-level induction accuracy peaks and declines in both conditions, at the standard learning rate (three seeds) and in a lower-learning-rate check (one seed). Bottom: best induction-head attention scores remain high. Heads are re-elected at each measurement. These descriptive trajectories do not identify why output accuracy declines.

probe-specific output changes. Figure 7 makes the pattern–behavior separation explicit. The decline in both conditions does not validate the training protocol or establish functional recovery. We therefore report the attribution direction, without interpreting the endpoints as a successful behavioral replication across model families.

Table 11: Exploratory OLMo band-1 recruitment after 12k updates at learning rate 5×10^{-4} and effective batch 64. Values are per seed; magnitudes are protocol-dependent and induction accuracy declines in both conditions.

Corpus	Seeds	Band-1 takeover share	Control band-0 share
Pile proxy	1	0.599 (flat from 7k)	0.489
OLMo-mix-1124	3	0.469/0.540/0.541	0.444/0.586/0.494

A.9 Post-formation repair and the before-onset comparison

A post-formation disrupt-and-repair experiment measures how quickly an existing circuit adapts; it does not directly test whether that circuit could have formed under the intervention. In an initial Pythia-70M run, induction accuracy remained 0.153 below control after 2,500 steps of band-0 ablation (three seeds, sd 0.014), while Gaussian noise calibrated to the same immediate accuracy loss had recovered. That finite observation window does not establish a persistent barrier.

The measured residual depends on the observation window. In a three-seed, batch-128 rerun extended to 12,000 steps, the residual—control accuracy minus ablated accuracy—falls from 0.508 at 500 steps to 0.253 at 4k, 0.172 at 8k, and 0.123 at 12k (Figure 8). An uninvolved-band condition supplies an empirical null around zero; by 12k its per-seed spread interleaves with that of the band-0 residual even though the latter's mean remains positive. A separate batch-16 budget run shows the same qualitative point, declining from 0.346 at 2.5k to 0.134 at 16k. Absolute values differ substantially across protocols, so we neither pool these trajectories nor fit an asymptotic floor. The apparent amount of non-recovery is still changing at the endpoints we observed.

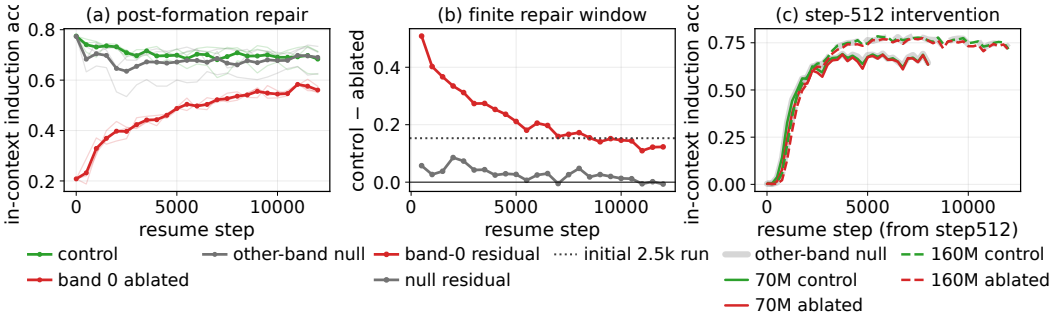


Figure 8: Post-formation repair and intervention before detectable onset answer different questions (Pythia-70M unless marked). Thick lines are seed means. **(a)** Disrupt-and-repair trajectories after band-0 ablation. **(b)** Within one 12k-step protocol, the residual falls from 0.508 to 0.123; no asymptote is inferred. The dotted line marks the residual from the separate initial 2.5k-step protocol. **(c)** Separate before-onset comparison trajectories, with three seeds at 70M and five at 160M, distinct from E1/E2. Table 12 compares late-window and final-checkpoint summaries; the exploratory equivalence verdicts depend on that choice.

A separate comparison of subsequent development. The before-onset comparison runs are distinct from both the downshift runs in Appendix A.5 and E1/E2; Table 7 specifies their configurations. The earlier exploratory analysis averaged each condition's accuracy over its last three saved checkpoints before forming per-seed control-minus-ablated residuals: steps 7500/7750/8000 at 70M and 11500/11750/12000 at 160M. Table 12 reports both this late-window average and the final checkpoint alone. For each summary, two one-sided tests (TOST, $\alpha = 0.05$) use a symmetric margin equal to twice the sample standard deviation of the corresponding control-minus-uninvolved-band residuals; the margin is therefore recomputed for each row. The last-three-checkpoint rule passes at 70M and is inconclusive at 160M, while the final-checkpoint-only rule reverses both verdicts. Both the residual distribution and the data-dependent margin change with the aggregation window. These exploratory results do not establish statistical equivalence under prespecified bounds or a stable contrast in equivalence between scales. Exact equality is not required to establish that induction emerges after the shared early checkpoint with band 0 disabled.

Table 12: Aggregation-window sensitivity in the separate before-onset comparison. Residuals are control minus band-0-ablated accuracy (mean ± sample sd across data-order seeds), after averaging within each seed over the listed additional training steps. The symmetric TOST bounds are $[-\delta, \delta]$; δ is estimated separately for each row from the uninvolved-band null. "Pass" denotes the exploratory rule's verdict, not equivalence under a prespecified margin.

Model (n)	Resume steps	Residual	δ	TOST rule
70M (3)	7500/7750/8000	0.0020 ± 0.0076	0.0252	Pass
70M (3)	8000	0.0005 ± 0.0134	0.0167	Inconclusive
160M (5)	11500/11750/12000	0.0120 ± 0.0178	0.0173	Inconclusive
160M (5)	12000	0.0039 ± 0.0210	0.0325	Pass

A.10 Boundary conditions: long-context scaling interventions

Downshift occurs when the preferred rotation is removed and another discrete band becomes the fastest survivor. Position interpolation (Chen et al., 2023) at a factor of four instead slows every band while preserving their order, leaving no faster alternative to band 0. At Pythia-70M the selected previous-token head stays on band 0; its mean query–key norm product increases 7.3-fold, and mean endpoint induction accuracy is within 0.013 of control. NTK-aware scaling raises the rotary base, leaves band 0's frequency unchanged, and keeps mean margin and norm product within 2% of control (Appendix A.13). The contrast is consistent with the motivation for preserving fast rotations in YaRN (Peng et al., 2024). Training stays at 512 tokens, and we did not measure long-context

task performance: the result concerns margin and norm adaptation, not computational cost or long-context quality.

A.11 Alternative hypotheses and counterevidence

The following seven candidate explanations are unsupported or overstated given the measurements. We collect them here to delimit what the evidence supports, without claiming prospective registration.

1. **The gating hypothesis** (band 0 is necessary for induction to form). E1 shows formation from initialization with band 0's rotation disabled at 70M (Section 4.1); the separate step-512 runs show subsequent emergence under the same rotation mask. That is a counterexample under the tested protocol, not a guarantee over all initializations or models.
2. **The wavelength–distance "aliasing" principle** (a band supports matching only at distances below its wavelength). Confounded with probe context length: at fixed query distance 7, band 4's ablation cost rises 60-fold as context grows from 16 to 512; the apparent distance dependence tracked context, not wavelength.
3. **"Band-0 dominance grows with repeat distance."** A normalization artifact: the raw drop grows from 0.324 to 0.635 across repeat distances, but baseline accuracy grows from 0.412 to 0.794 alongside it, leaving the ratio flat (0.79 shortest, 0.80 longest).
4. **Single-head localization generalizes across scale.** The top head reproduces 87% of the global band-0 effect at 70M but 24% at 160M and 20% at 410M; top-head localization alone does not capture the larger models' dependence (Section 3).
5. **Ladder monotonicity** (takeover concentration decreases smoothly down the ladder). Observed 0.93, then 0.58, then 0.66: the rule predicts the recruited band, not its concentration.
6. **"The pre-transition state is uniform across bands."** The appearance of a uniform starting state depends on the baseline and on weak, early head scores; we claim the measured rise in band-0 concentration, not initial uniformity.
7. **The phase-resolution cost prediction** (margin inflation should scale as the ratio of angular resolutions, 3.2-, 10-, and 32-fold down the ladder). Observed 4.3-, 3.8-, and 6.9-fold: the quantitative prediction is unsupported and the trend is non-monotone where the account requires a rise (Section 4).

A.12 The ablation ladder

Table 13 gives the per-rung ladder values behind Section 5.1. Figure 9 shows that recruitment concentration does not decline monotonically down the ladder. Margin inflation likewise does not follow a simple inverse-frequency account: the observed ratios are $4.3\times$, $3.8\times$, and $6.9\times$, against frequency ratios of $3.2\times$, $10\times$, and $32\times$. L2H1 remains the selected previous-token head at every rung, but the selected induction head changes from L3H6 on the first rung to L3H1 in all six runs on the two deeper rungs. Thus the ladder supports a band-recruitment rule, not preservation of every head in the circuit.

Table 13: The ablation ladder at Pythia-70M (three seeds per rung). The fastest-surviving-frequency rule matches the recruited band, but neither monotonic concentration nor inverse-frequency margin scaling matches the observations. Margin ratios use the three-control-run mean margin (181); the last column is the inverse-frequency ratio for the recruited band b.

Ablated set	Rule / observed	Takeover share (seeds)	Margin ratio	θ_0/θ_b
$\{0\}$	band 1 / band 1	0.928/0.922/0.927	$4.3\times$	$3.2\times$
$\{0,1\}$	band 2 / band 2	0.589/0.583/0.571	$3.8\times$	$10\times$
$\{0,1,2\}$	band 3 / band 3	0.661/0.648/0.657	$6.9\times$	$32\times$

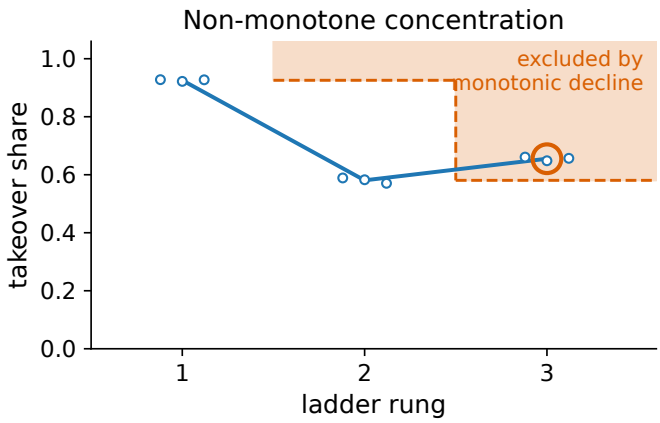


Figure 9: Non-monotone concentration in the 70M ladder. Rungs 1–3 ablate $\{0\}$, $\{0, 1\}$, and $\{0, 1, 2\}$, respectively; open circles show individual seeds with small horizontal offsets for visibility, and the line joins their means at integer rungs. The shaded region lies above the preceding rung's mean, contrary to a monotonic-decline hypothesis. The selection rule identifies the recruited band but does not predict these concentration levels.

A.13 Robustness of the complementary experiments

Per-seed endpoint induction accuracy at 70M. The original downshifted runs give 0.698, 0.567, and 0.653; controls give 0.700, 0.613, and 0.630. The OLMo behavioral decline is addressed separately in Appendix A.8. Across three seeds per scaling condition, position interpolation leaves mean induction accuracy within 0.013 of control and increases the mean margin 6.0-fold and mean query–key norm product 7.3-fold; NTK-aware scaling leaves both latter means within 2% of control.

Top-head identities per run. In the earlier Pythia-160M attribution experiments, all three control runs select L3H0, as do downshifted seeds 1 and 2; downshifted seed 0 selects L3H2. The standard-learning-rate OLMo-mix-1124 ablated runs select L4H8/L2H7/L3H11 (seeds 0/1/2); their controls select L1H13/L2H5/L2H8. The corresponding Pile-proxy ablated run selects L4H14, its control L2H4. These four standard-learning-rate pairs select different heads across conditions, whereas the lower-learning-rate OLMo-mix-1124 pair selects L4H13 in both arms. Head identity is therefore not generally preserved in the OLMo comparisons.

Other checks. Four further checks address measurement stability in the earlier Pythia runs. (i) The tested margin shares vary by about 0.01 across probe seeds and probe batch sizes, apart from the weak-head 1.4B checkpoints discussed in Appendix A.3. (ii) Tracking one fixed head backward through training shows the developmental transition is not an artifact of re-selecting the top head at each checkpoint. (iii) Both margin baselines are recorded in the artifacts; Table 4 reports the E1/E2 endpoints' baseline sensitivity. (iv) A distance-by-context-length grid separates query distance from probe context length.

Reproducing the exploratory equivalence analysis. `tost_equivalence.py` in the code release reads the archived run artifacts and reproduces both summaries in Table 12 with `--sensitivity`, pooling the four files that split the 160M seeds. The default `--tail 3` retains the earlier late-window calculation; `--tail 1` selects the final checkpoint alone. The table is generated from the same records and checked against their hashes.