LAGGED COUPLING: DECODABILITY AND PROBE-DIRECTION INTERVENTIONS ACROSS TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Information that is linearly decodable from a language model need not be accessible through an intervention along the decoding direction. We track this distinction through training across six Pythia models (160M–12B parameters), eight checkpoints, and four synthetic task families. At each model–checkpoint–task combination, we measure internal probe AUROC, output-level answer-score AUROC, and the effect of a single-site additive edit along the probe direction relative to norm- and site-matched random directions. The three measurements dissociate. Internal decodability is near ceiling (probe AUROC ≥ 0.98) from the first measured checkpoint; output-level discrimination, tabulated for the four largest models, rises by up to 0.43 AUROC over training; and 43 of 48 intervention summaries fall within the random-direction band $|z_e| < 2$, with no sustained positive effect and a consistently negative alignment at the two earliest checkpoints. Over the same period, the RMS activation along the probe direction grows by up to 56.8-fold without a corresponding intervention benefit. We call this dissociation *lagged coupling*. A first-order analysis explains why it is expected: the standardized intervention effect equals $\sqrt{d}$ times the cosine between the probe direction and the model's mean readout gradient, independently of the edit norm, and near-perfect decodability places no lower bound on that cosine when the representation is anisotropic. Probe accuracy is therefore not a proxy for steerability, and decodability and intervention efficacy should be evaluated separately throughout training.

1 INTRODUCTION

Linear probes can recover information from a language model's hidden states, while activation interventions can change its behavior. These observations motivate an appealing question: can the direction learned by a probe also serve as a useful intervention direction? Activation addition, representation engineering, inference-time intervention, and representation fine-tuning demonstrate that selected or learned internal directions can influence model outputs (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Wu et al., 2024). Their success, however, does not imply that every accurate decoding direction is effective for steering, and steering effects vary substantially across inputs, concepts, and models (Tan et al., 2024; Braun et al., 2025; Silva et al., 2025). The probing literature has long asked what probe accuracy establishes (Hewitt & Liang, 2019; Pimentel et al., 2020; Kumar et al., 2022) and whether the model uses the decoded information (Elazar et al., 2021; Lasri et al., 2022).

Training adds a temporal dimension to this distinction. A variable may become decodable before the model produces outputs that discriminate that variable, and neither measurement determines the response to an additive edit. Studies of grokking and syntax acquisition show that internal structure and behavior can develop on different schedules (Power et al., 2022; Nanda et al., 2023; Chen et al., 2024), whereas induction heads form together with the corresponding behavioral change (Olsson et al., 2022). For probe-derived directions, the corresponding question is whether decodability and intervention efficacy become aligned over training, and whether this alignment changes with model size. Measurements at a final checkpoint alone cannot characterize that trajectory; checkpoint studies have tracked probing accuracy, learned features, and circuits over pretraining (Liu et al., 2021; Tigges et al., 2024; Bayazit et al., 2026), and early-checkpoint probes and steering vectors have been reported for trustworthiness concepts (Qian et al., 2024).

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.

We investigate this question using six models from the Pythia suite (Biderman et al., 2023), ranging from 160M to 12B parameters, at eight checkpoints and on four synthetic task families. For each model–checkpoint–task combination, we measure three quantities: **internal decodability**, assessed by a logistic probe; **output-level discrimination**, assessed by the AUROC of the model's answer score; and **intervention efficacy**, assessed by adding the unit-norm probe direction at the probed site and comparing the effect with matched random-direction edits. The probe is fitted separately for each combination and frozen for its evaluation.

The resulting trajectories are dissociated. Internal probe performance is near ceiling from step 1,000, the earliest checkpoint examined (probe AUROC at least 0.990 on development and 0.982 on test data at every size). Output-level discrimination improves over much of training, with non-monotonic intermediate values and changing rankings across model sizes. Most intervention summaries remain within the random-direction reference band, and at the two earliest checkpoints every summary is negative. We use **lagged coupling** to describe this empirical separation between what a probe can decode and what an edit along its direction can change.

The dissociation has a simple geometric explanation (Section 4): to first order, z_e is $\sqrt{d}$ times the cosine between the probe direction and the mean readout gradient, and under anisotropy the logistic discriminant can be nearly orthogonal to that gradient while decoding almost perfectly.

Our contributions are as follows.

1. We introduce a developmental protocol that jointly measures internal decodability, output-level discrimination, and probe-direction intervention efficacy on a 6-size, 8-checkpoint, 4-task grid of Pythia models, comprising 192 evaluation units.
2. We show that the three measurements dissociate throughout training: decodability is near ceiling from the first checkpoint, output discrimination rises by up to 0.43 AUROC, and 43 of 48 intervention summaries remain within the random-direction band while projected activation grows by up to 56.8-fold, with all twelve summaries at the two earliest checkpoints negative.
3. We give a first-order account of the dissociation: the standardized effect measures alignment between the probe direction and the mean readout gradient independently of dose, and anisotropy permits near-perfect decodability with vanishing alignment, which predicts that difference-in-means directions should behave differently at the same sites.

The practical implication is that monitoring and steering rest on different quantities. A probe used as a sensor requires only decodability, which the grid shows early and at every size; a probe direction used as a lever requires alignment with the readout, which the grid does not show in a sustained way at any stage of training.

2 Related Work

Decodability and causal use. Probing methods recover linguistic and other structured information from hidden representations (Hewitt & Manning, 2019; Belinkov, 2022). Control tasks and information-theoretic analyses clarify what probe performance can establish (Hewitt & Liang, 2019; Pimentel et al., 2020), while amnesic probing and related interventions examine whether decoded information contributes to behavior (Elazar et al., 2021; Kumar et al., 2022; Lasri et al., 2022). Our study applies this distinction across training checkpoints. An additive edit along a discriminative probe is one operational test of intervention efficacy; it is not an exhaustive test of whether the model uses the corresponding variable.

Activation interventions and model editing. Activation addition, representation engineering, inference-time intervention, and representation fine-tuning use different procedures to select or learn interventions (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Wu et al., 2024). Contrastive activation addition and refusal-direction interventions provide further evidence that suitable directions can alter behavior (Rimsky et al., 2024; Arditi et al., 2024), although systematic evaluations find steering effects unreliable across inputs, concepts, and model families (Tan et al., 2024; Braun et al., 2025; Silva et al., 2025). Our intervention is narrower: a single-site additive edit along a logistic-probe direction. We do not compare these steering algorithms or evaluate their best achievable performance, which recent evaluations address directly, including a benchmark that separates concept detection from steering (Wu et al., 2025; Im & Li, 2025). Weight editing (Meng et al., 2022), causal

Table 1: Position of related lines of work. "Property relied on" names the quantity each line of work measures or presupposes; "Across training" asks whether that quantity is tracked over checkpoints. Citations are given in the surrounding text.

Line of work	Representative methods	Property relied on	Across training
Probing	structural, control-task, and amnesic probes	internal decodability	rarely
Activation steering	ActAdd, RepE, ITI, CAA, ReFT	intervention efficacy	rarely
Weight editing	ROME and causal tracing	intervention efficacy (weights)	no
Developmental circuits	induction heads, grokking progress measures	internal structure and behavior	yes, without steering
This work	per-unit probes with matched random-direction edits	all three jointly	**yes (6 sizes $\times$ 8 checkpoints)**

mediation (Vig et al., 2020), and causal-abstraction methods (Geiger et al., 2021; 2022) address related questions using different intervention families. Activation-patching methodology also shows that conclusions depend on intervention and metric choices (Zhang & Nanda, 2024), and that subspace interventions can alter outputs through dormant pathways rather than the targeted variable (Makelov et al., 2024).

Development and scale. Scaling studies characterize relationships among model size, training resources, and performance (Kaplan et al., 2020; Hoffmann et al., 2022; Ruan et al., 2024); analyses of emergence emphasize the role of measurement in apparent transitions (Wei et al., 2022; Schaeffer et al., 2023). Grokking and induction-head studies examine how internal computations develop during training (Power et al., 2022; Nanda et al., 2023; Olsson et al., 2022), complementing circuit analyses at individual checkpoints (Elhage et al., 2021; Wang et al., 2023) and across Pythia checkpoints and sizes (Tigges et al., 2024); syntactic structure can also appear abruptly and precede the capabilities it supports (Chen et al., 2024). Linear relation decoding and the linear representation hypothesis, which distinguishes representations used for measurement from those used for intervention, motivate studying directions in representation space (Hernandez et al., 2024; Park et al., 2024; Marks & Tegmark, 2024), and the emergence of such linear encodings during training has begun to be characterized (Ravfogel et al., 2025). Pythia and OLMo (Biderman et al., 2023; OLMo et al., 2025) supply intermediate checkpoints for developmental comparisons, and seed replicates allow run-to-run variability to be assessed (van der Wal et al., 2025). Our focus is the joint trajectory of decoding performance and a specified probe-direction intervention. Table 1 summarizes where these lines of work sit with respect to the three measurements and to training time.

3 Problem Formulation and Setup

Problem statement. Fix a model family, a training checkpoint t, and a binary task variable y that the model must express in its answer. Three quantities can be attached to the pair (checkpoint, task): the decodability I_t of y from a residual-stream site, the output-level discrimination B_t of y by the model's own answer score, and the efficacy $z_{e,t}$ of an additive edit along the decoding direction at that site. Practice that uses probes as steering handles relies on an implication from the first quantity to the third: a site and direction at which y is decodable should also be a site and direction at which moving the activation moves the answer. We ask whether that implication holds along training trajectories and across model sizes, and we use *lagged coupling* for the observed pattern in which I_t is already high while $z_{e,t}$ remains inside the random-direction band.

Grid and evaluation units. We study Pythia-160M, 410M, 1B, 2.8B, 6.9B, and 12B at steps 1k, 2k, 4k, 8k, 16k, 32k, 64k, and 143k. Four synthetic task families probe different ways in which a decodable variable can fail to reach the answer (Table 2; Appendix A).

An evaluation *unit* is one model–checkpoint–task combination, yielding 192 units. A *cell* denotes one of the 48 model–checkpoint combinations. For intervention effects, each cell summarizes the median standardized effect across the four task families.

Table 2: Task families. Each family is a binary answer-format contrast; they differ in what competes with the target variable for control of the answer.

Family	Contrast	What competes with the target
T	template contrast	nothing beyond the answer format itself
D	distractor robustness	competing surface cues in the prompt
P	prior-bias override	a prior learned from pretraining data
N	negative transfer	related but incorrect structure

3.1 THREE MEASUREMENTS

For prompt x, let $z_t(x)$ denote the logit of token t at the answer position, and let t^* be the task-designated answer token. The answer score is the vocabulary-wide logit margin, which measures whether the designated answer is the model's top prediction and therefore also registers answer-format errors that a margin between two candidate tokens would ignore:

$$s(x) = z_{t^*}(x) - \max_{t \neq t^*} z_t(x). \tag{1}$$

Let $y \in \{0, 1\}$ denote the task label and $a_\ell(x)$ the residual-stream activation at the selected site ℓ.

Internal decodability. A logistic probe predicts the label from the selected activation:

$$\hat{y}(x) = \sigma\big(w^\top a_\ell(x) + b\big), \qquad I = \mathrm{AUROC}\big(y, \hat{y}(x)\big), \tag{2}$$

where σ is the logistic sigmoid. Probes are fitted separately for each unit on held-apart fitting items. The site-selection procedure and probe are fixed before test evaluation. Development and test AUROCs are tracked separately. Throughout, "frozen probe" means frozen within an evaluated unit, rather than shared across model sizes or training checkpoints.

Output-level discrimination. Independently of the probe, we measure

$$B = \mathrm{AUROC}\big(y, s(x)\big). \tag{3}$$

The behavioral series use the development split. B measures how well the answer score ranks the binary labels. Differences between I and B may reflect an easily decoded input regularity before the model learns to express it in its output.

Probe-direction intervention. At the probed site, we add $\lambda\hat{w}$, where $\hat{w} = w/\|w\|_2$ and λ is the intervention strength. Let $s_\lambda(x)$ denote the answer score from this intervened forward pass. The average paired effect is

$$e(\lambda) = \mathbb{E}_x\left[s_\lambda(x) - s_0(x)\right]. \tag{4}$$

The expectation compares edited and unedited scores for the same evaluation items. Random-direction interventions matched for site and norm provide a reference distribution of paired effects. For a unit-strength intervention, we report

$$z_e = \frac{e(1) - \mu_{\mathrm{null}}}{\sigma_{\mathrm{null}}}, \tag{5}$$

where μ_{null} and σ_{null} are the mean and standard deviation of the random-direction effects.

We use $|z_e| \geq 2$ to flag effects that stand out from the random-direction reference. Similarly, a negative z_e means an effect below the null mean, not necessarily a negative raw effect $e(1)$.

3.2 ACTIVATION MAGNITUDE AND REFERENCE VARIABILITY

The projected RMS activation, which we call *headroom*, is

$$C = \left(\mathbb{E}_x\left[(\hat{w}^\top a_\ell(x))^2\right]\right)^{1/2}. \tag{6}$$

We also report the ratio $\hat{c} = e(1)/C$. This ratio has units of logit margin per activation unit; it is not a percentage of represented information used by the model. A comparison at a fixed fraction

of activation magnitude would require separate interventions with a corresponding dose rule, such as $\lambda = \alpha C$. Projected RMS magnitude can also increase because of changes in activation mean or variance, without an increase in label-specific information.

The analysis records a reference scale ε_m, defined as the median absolute random-direction effect within a unit. Per-size summaries are given in Table 4.

3.3 Scope and analysis protocol

The experiment evaluates a specified single-site probe-direction edit. We study the logistic direction because it is the direction a probe supplies; Section 4 shows why difference-in-means directions, which have been found more causally implicated (Marks & Tegmark, 2024; Im & Li, 2025), behave differently.

The grid, estimands, thresholds, and random-direction comparisons were fixed before scoring; auxiliary onset analyses are in Appendix E, and checkpoint provenance is documented in Appendix F.

4 Theoretical Analysis

The three measurements of Section 3.1 are defined independently of one another. This section relates them in a local linear model of the answer score. The analysis establishes three facts that govern how the results should be read: the standardized effect z_e is, to first order, an *alignment* statistic that does not depend on the edit norm; the band $|z_e| < 2$ corresponds to near-orthogonality between the probe direction and the model's mean readout gradient; and near-perfect decodability places no lower bound on that alignment when the representation is anisotropic.

Local model. Let $g(x) = \nabla_{a_\ell} s(x)$ be the gradient of the answer score (Eq. (1)) with respect to the activation at the probed site, and let $\bar{g} = \mathbb{E}_x[g(x)]$ be its average over evaluation items. For an edit of norm λ along a unit direction v, a first-order expansion gives $s_\lambda(x) = s_0(x) + \lambda\, g(x)^\top v + O(\lambda^2)$, and therefore

$$e(\lambda; v) = \lambda\, \bar{g}^\top v + O(\lambda^2). \tag{7}$$

The approximation is accurate when the edit is small relative to the scale over which the downstream computation curves.

Proposition 1 (The standardized effect measures alignment)**.** *Suppose the reference directions are distributed uniformly on the unit sphere S^{d-1}, equivalently as normalized isotropic Gaussians, and that the second-order term in Eq.* (7) *is negligible. Then $\mu_{\text{null}} = 0$, $\sigma_{\text{null}} = \lambda\|\bar{g}\|/\sqrt{d}$, and*

$$z_e = \sqrt{d}\, \cos(\bar{g}, \hat{w}). \tag{8}$$

In particular, z_e does not depend on λ, and for large d the median absolute reference effect is $\varepsilon_m \approx \Phi^{-1}(3/4)\, \lambda\|\bar{g}\|/\sqrt{d} \approx 0.674\, \lambda\|\bar{g}\|/\sqrt{d}$.

Proposition 1 (proved, with Proposition 2, in Appendix B) has three consequences for the grid. *First*, dose: in the regime where the response is linear in λ, a larger edit scales the probe-direction effect and the random-direction spread together, so it cannot move z_e. A small standardized effect can therefore be attributed to an insufficient fixed dose only if the response is nonlinear at the tested strengths. *Second*, the meaning of the band: $|z_e| < 2$ holds exactly when $|\cos(\bar{g}, \hat{w})| < 2/\sqrt{d}$. Table 3 converts this threshold for each Pythia width (Biderman et al., 2023); a unit inside the band is one in which the probe direction lies within a few degrees of orthogonality to the average readout gradient, and the tolerance tightens as models widen. *Third*, the reference scale: under the proposition, $\varepsilon_m\sqrt{d}/0.674$ estimates $\lambda\|\bar{g}\|$. Applied to the per-size summaries in Table 4, this quantity grows by a factor of about 21 from 160M to 12B, compared with a factor of 8.2 for ε_m itself. As a size-level summary pooled over checkpoints and tasks, it shows that a unit-norm edit is not negligible in the larger models, where arbitrary directions move the answer score more rather than less.

Reading versus writing. A probe weight w defines a functional $a \mapsto w^\top a$; an intervention requires a displacement in activation space. Identifying the two through the Euclidean inner product is a choice, and it is not the choice under which reading and writing coincide (Park et al., 2024). The following idealized model makes the consequence explicit.

Table 3: The band $|z_e| < 2$ expressed as an alignment tolerance via Eq. (8). d is the residual-stream width of each Pythia model; the angle is the smallest angle between $\hat{w}$ and $\bar{g}$ compatible with a cell inside the band.

Model	160M	410M	1B	2.8B	6.9B	12B
Width d	768	1024	2048	2560	4096	5120
$2/\sqrt{d}$	0.072	0.063	0.044	0.040	0.031	0.028
Angle to $\bar{g}$	85.9°	86.4°	87.5°	87.7°	88.2°	88.4°

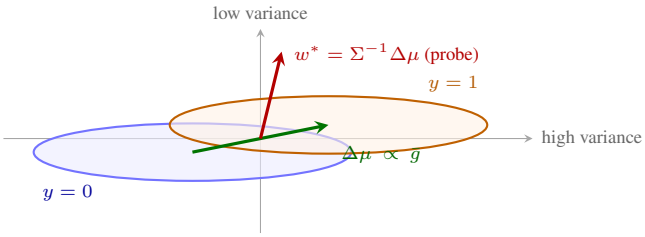

Figure 1: Proposition 2 in two dimensions. Within-class variation is large along the horizontal axis, so the class-mean difference $\Delta\mu$ is mostly horizontal, but the discriminant $\Sigma^{-1}\Delta\mu$ tilts toward the low-variance vertical axis, where a fixed separation is most detectable. A readout that responds to the average class difference is nearly orthogonal to the probe direction, although the probe separates the classes well.

Proposition 2 (Decodability without alignment)**.** *Let $a \mid y \sim \mathcal{N}(\mu_y, \Sigma)$ with a shared covariance $\Sigma \succ 0$ of condition number κ, let $\Delta\mu = \mu_1 - \mu_0$, and let $w^* = \Sigma^{-1}\Delta\mu$, the direction of the population log-odds to which unregularized logistic regression converges in this model. The probe $w^{*\top}a$ has* AUROC $= \Phi(\Delta_M/\sqrt{2})$*, where $\Delta_M^2 = \Delta\mu^\top\Sigma^{-1}\Delta\mu$. If the readout responds to the class-mean difference, $\bar{g} = c\,\Delta\mu$ with $c > 0$, then*

$$\frac{2\sqrt{\kappa}}{1+\kappa} \;\le\; \cos(\bar{g}, w^*) \;\le\; 1, \tag{9}$$

and the lower bound is attained. Moreover, for every AUROC $\in (\frac{1}{2}, 1)$ *and every $\epsilon > 0$ there exist $(\Delta\mu, \Sigma)$ attaining that AUROC with $\cos(\bar{g}, w^*) < \epsilon$, whereas the difference-in-means direction satisfies $\cos(\bar{g}, \Delta\mu) = 1$.*

Whitening by Σ^{-1} rotates the discriminant toward directions of low within-class variance, because a fixed mean difference is most detectable there (Figure 1). A readout that responds to the average difference between classes weights directions by their mean separation, not by their signal-to-noise ratio. Transformer representations are strongly anisotropic, with a few high-variance dimensions dominating inner products (Ethayarajh, 2019; Timkey & van Schijndel, 2021), so this configuration is plausible at the probed sites. Combined with Proposition 1, it shows how a probe can be nearly perfect while its weight vector lies inside the random-direction band. It also explains why mean-difference directions, which correspond to Σw^* rather than w^*, are the usual choice for steering (Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024) and have been found more causally implicated than logistic-regression directions (Marks & Tegmark, 2024).

In this model the difference-in-means direction attains $\cos(\bar{g}, \Delta\mu) = 1$, and hence by Eq. (8) a standardized effect of $\sqrt{d}$, between 28 and 72 reference standard deviations at the Pythia widths, so the contrast with the logistic direction is large wherever the idealization holds. The model thus makes the observed dissociation expected rather than paradoxical, and it accounts for the finding that difference-in-means directions are more causally implicated than logistic-regression directions (Marks & Tegmark, 2024): the two coincide only when the site covariance is close to isotropic.

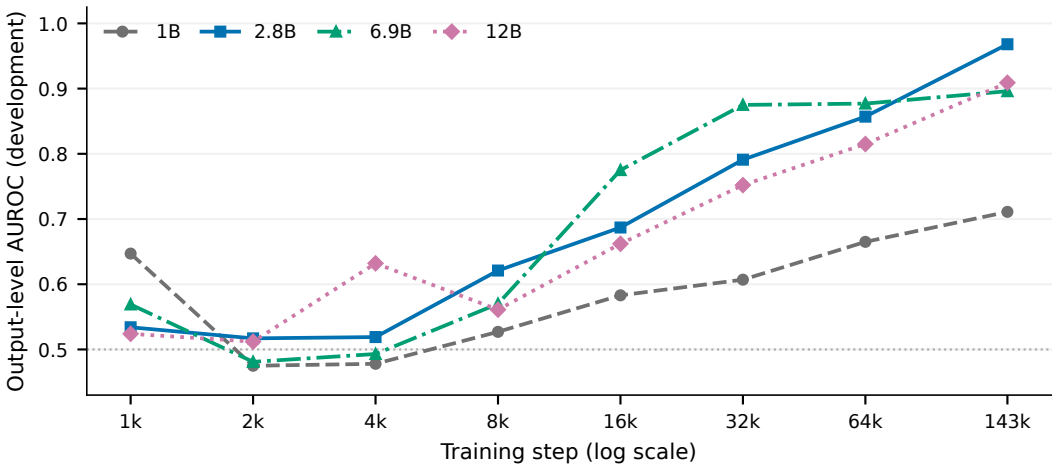

Figure 2: Output-level discrimination over training for the four model sizes with complete tabulated series. Points reproduce the development-split AUROCs in Table 7. The horizontal reference is chance AUROC. The trajectories change substantially after the earliest checkpoint, when internal decodability is already near ceiling.

5 Results

5.1 Decodability and output discrimination

Internal probe performance is near ceiling from step 1,000 across the six model sizes, so decodability is established by the earliest measured checkpoint. The output-level series show substantial subsequent change (Figure 2; complete values in Table 7). For the four sizes with tabulated behavioral trajectories, AUROC increases between the first and last checkpoint: from 0.647 to 0.711 for 1B, 0.534 to 0.968 for 2.8B, 0.569 to 0.896 for 6.9B, and 0.524 to 0.909 for 12B. The paths are not monotonic, and their ordering varies with training stage. At step 16k, for example, 6.9B (0.775) exceeds both 2.8B (0.687) and 12B (0.662), whereas 2.8B has the highest value at the final checkpoint. All four tabulated series still increase from 64k to 143k.

All 11 units with sufficient milestone information to classify a temporal ordering have internal readability leading the other milestones, and the lag shows no clear dependence on scale (Appendix E).

5.2 Intervention effects

Figure 3 reports the full 48-cell intervention grid. Of these cells, 43 have $|z_e| < 2$. Four lie below -2: 410M at 2k (-2.65), 2.8B at 1k (-2.68) and 2k (-2.48), and 12B at 2k (-2.43). The only value above $+2$ is 12B at 8k ($+2.49$); its value at the next checkpoint is -0.31. The development-phase counterpart of the 8k value is $+2.04$. These are isolated deviations from the random-direction reference, with no sustained positive pattern.

The grid also has a clear temporal structure. All twelve cells at steps 1k and 2k are negative, with mean -1.39, and they contain all four negative threshold crossings; the 36 later cells average $+0.27$, and the single positive crossing does not persist to the next checkpoint. Each of the six independently trained models is negative at both early checkpoints; counting only one early cell per model, a sign-symmetric null gives probability $2^{-6} \approx 0.016$. The median of four independent standard normal z-values has a null standard deviation below 0.6, so the threshold $|z_e| \geq 2$ is conservative for single cells and the early mean of -1.39 is a robust signal. Because the same sign convention applies at every checkpoint, this change from negative to near-zero alignment is a property of training rather than of the measurement: early in training the probe direction points against the readout, and the anti-alignment disappears as output discrimination develops. Proposition 2 assumes a readout that responds to the class-mean difference with $c > 0$; at steps 1k and 2k output discrimination is still near chance (AUROC 0.52–0.65 at 1k and 0.48–0.52 at 2k; Table 7), so $c > 0$ need not hold: within that

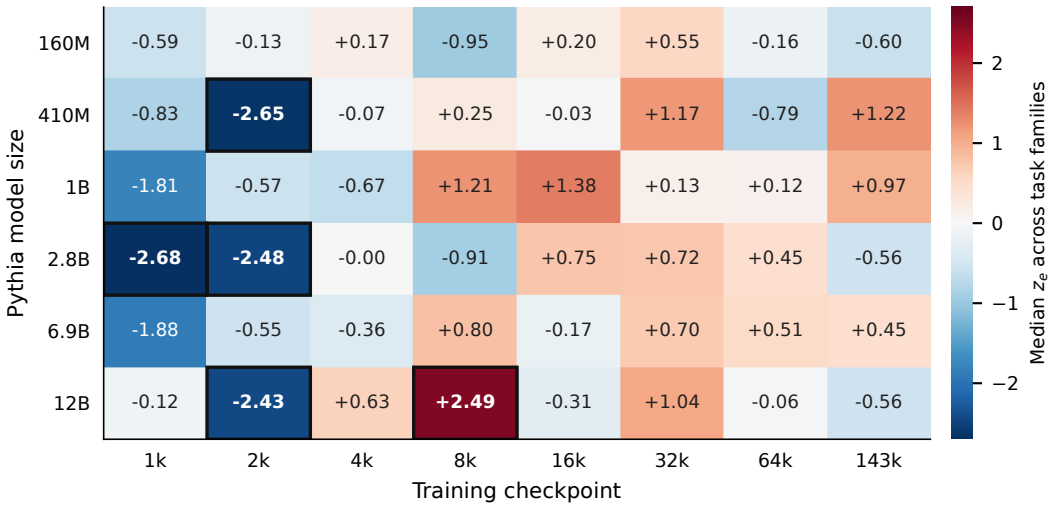

Figure 3: Probe-direction intervention effects relative to matched random directions. Each entry is the test-phase median z_e across four task families; all values are reproduced in Table 6. Boxes indicate the threshold $|z_e| \geq 2$, met by five of 48 cells.

Table 4: Per-size random-direction reference scale ε_m, growth of projected RMS activation C, and final-checkpoint absolute effect-to-magnitude ratio. The last column is in logit-margin units per activation unit, not percent.

Model	ε_m	C growth (1k to 143k)	$\lvert e(1)/C \rvert$ at 143k
160M	0.00398	2.0×	—
410M	0.00597	8.1×	—
1B	0.00619	3.0×	—
2.8B	0.01179	46.9×	0.00070
6.9B	0.0286	37.7×	0.00109
12B	0.0328	56.8×	0.00108

model the negative cells correspond to a readout gradient that points against the class-mean difference early in training, and the assumption becomes appropriate once output discrimination develops.

Read through Proposition 1, the grid says that after the earliest checkpoints the probe direction is nearly orthogonal to the average readout gradient: in 43 of 48 cells the task-median value of $\sqrt{d}\cos(\bar{g}, \hat{w})$ lies inside ± 2, a cosine tolerance between 0.028 and 0.072 (Table 3). Because the first-order statistic is invariant to the edit norm, this conclusion concerns the direction rather than the dose.

5.3 Activation magnitude and sensitivity

The projected RMS activation C increases from step 1k to 143k by factors of 2.0, 8.1, 3.0, 46.9, 37.7, and 56.8 for the six model sizes, respectively (Figure 4a). For 12B, the endpoints are approximately 0.038 and 2.16. The absolute cell-level effects at unit strength do not exceed 0.016 logit-margin units. At the final checkpoint, the ratios $|e(1)/C|$ are 7.0×10^{-4}, 1.09×10^{-3}, and 1.08×10^{-3} for 2.8B, 6.9B, and 12B.

The reference scale ε_m increases with model size, from 0.00398 at 160M to 0.0328 at 12B (Table 4).

Its values for 6.9B (0.0286) and 12B (0.0328) exceed the largest absolute cell-level effect, while the 2.8B value (0.01179) does not. The two trends point in the same direction. Random unit-norm edits move the answer score more in larger models, which by Proposition 1 reflects a larger mean readout gradient at the probed site, while the probe direction's own effect stays within the reference band. Growth in the RMS projection along $\hat{w}$, which C measures, and growth in the output's sensitivity

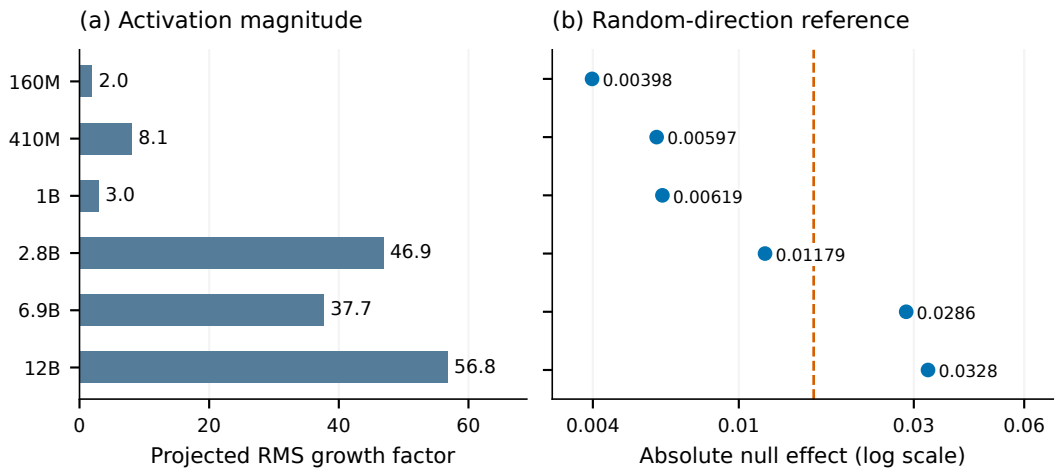

Figure 4: Activation magnitude and random-direction reference scale. (a) Growth of projected RMS activation C from step 1k to 143k. (b) Per-size summaries of the median absolute random-direction effect ε_m (logarithmic axis). The dashed reference is the largest absolute cell-level effect, 0.016. Only the 6.9B and 12B reference scales exceed this value; the 2.8B value is 0.01179.

along $\hat{w}$, which $e(1)$ measures, are governed by different quantities (Proposition 2), and the grid shows the first without the second.

6 Discussion

Our principal observation is a developmental dissociation: strong internal decodability coexists with changing output-level discrimination, with early anti-alignment and then near-orthogonality between the decoding direction and the readout. This extends the distinction between information availability and intervention efficacy from final checkpoints to the whole training trajectory, across a roughly 70-fold range of model size, and Section 4 shows why it should be expected: decodability constrains separation along the probe direction, not its alignment with the readout.

The analysis identifies the mechanism: a probe's weight vector is a discriminant, which anisotropy rotates away from the direction the readout responds to (Proposition 2), so accurate decoding and effective editing are governed by different geometric quantities. This account is consistent with the greater causal efficacy of difference-in-means directions reported in prior work (Marks & Tegmark, 2024) and extends it along the training trajectory. Because z_e depends only on the angle between the edit and the readout, the grid of Figure 3 is a map of alignment over training: anti-aligned at the first two checkpoints and near-orthogonal afterwards at every size. The early anti-alignment is itself a developmental finding: before the model's output discriminates the variable, an edit along the decoding direction moves the answer score against the target answer more than a random edit does.

Limitations. The study uses one model suite, with a single training run per size, and four controlled synthetic task families, which isolate the variables of interest and leave open-ended generation and other steering methods for future work. Internal decodability is already high at the first measured checkpoint, so its onset lies before the observed window. The measurements use unit-norm edits, for which Proposition 1 makes the standardized effect dose-invariant within the linear regime.

Conclusion. Across six model sizes and eight checkpoints, early decodability does not identify a direction with sustained intervention benefit, and a first-order analysis explains why: control depends on the alignment of the edit direction with the readout.

ETHICS STATEMENT

The evaluation uses publicly released pretrained models and synthetic task families; it involves no human subjects and collects no personal data, although the models' pretraining corpora may contain personal or sensitive material. Activation steering can be used to change model behavior, including safety-relevant behavior, and probes are increasingly proposed as monitors. Our results bear on both uses: they indicate that probe accuracy should not be taken as evidence that a probe direction can control a model, which argues for evaluating monitoring and control separately before relying on either.

REPRODUCIBILITY STATEMENT

The main text defines the evaluation grid and the three measurements and states the assumptions of the theoretical analysis. The appendix gives complete proofs, the model sizes, checkpoints, task families, and configuration, all intervention summaries together with the behavioral and activation-magnitude aggregates, the provenance of the checkpoints, and the seed assignment. Code, configurations, manifests, and unit-level results will be released with the camera-ready version.

AI USE STATEMENT

We used generative AI tools to assist with writing and editing the manuscript, searching for and verifying related literature, developing and debugging pipeline code, and preparing figures from tabulated results. AI tools also assisted in deriving and writing up the propositions of the theoretical analysis and in relating them to the measured grid; the authors checked every derivation and its connection to the data. We take responsibility for the final content of this paper, including text, claims, proofs, code, and figures produced with the aid of generative AI.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, 2024.

Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding Through Time: Tracking Emergence & Consolidation Of Linguistic Representations Throughout LLM Pretraining. In *Annual Meeting of the Association for Computational Linguistics*, 2026.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022.

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 *International Conference on Machine Learning*, 2023.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)Reliability of Steering Vectors in Language Models. *arXiv preprint arXiv:2505.22637*, 2025.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. In *International Conference on Learning Representations*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation With Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

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`.

Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

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

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing Causal Structure for Interpretable Neural Networks. In *International Conference on Machine Learning*, 2022.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of Relation Decoding in Transformer Language Models. In *International Conference on Learning Representations*, 2024.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

HJXA. Pythia-2.8B PT CheckPoints is same between steps 0 to steps 6300. GitHub issue 195, 2026. URL `https://github.com/EleutherAI/pythia/issues/195`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*, 2022.

Shawn Im and Sharon Li. A Unified Understanding and Evaluation of Steering Methods. *arXiv preprint arXiv:2502.02716*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing Classifiers are Unreliable for Concept Removal and Detection. In *Advances in Neural Information Processing Systems*, 2022.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the Usage of Grammatical Number. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, 2023.

William Li. Learning (to Reproduce Pythia 2.8B) Pretraining. Blog post, 2026. URL `https://zywilliamli.com/writings/writings-learning-pretraining/`.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. Probing Across Time: What Does RoBERTa Know and When? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *International Conference on Learning Representations*, 2024.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling*, 2024.

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

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

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. *arXiv preprint arXiv:2209.11895*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *International Conference on Machine Learning*, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-Theoretic Probing for Linguistic Structure. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards Tracing Trustworthiness Dynamics: Revisiting Pre-training Period of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Shauli Ravfogel, Gilad Yehudai, Tal Linzen, Joan Bruna, and Alberto Bietti. Emergence of Linear Truth Encodings in Language Models. In *Advances in Neural Information Processing Systems*, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance. In *Advances in Neural Information Processing Systems*, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? In *Advances in Neural Information Processing Systems*, 2023.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off Course: Reliability Challenges in Steering Language Models. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analyzing the Generalization and Reliability of Steering Vectors. *arXiv preprint arXiv:2407.12404*, 2024.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM Circuit Analyses Are Consistent Across Training and Scale. In *Advances in Neural Information Processing Systems*, 2024.

William Timkey and Marten van Schijndel. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem H. Zuidema, and Stella Biderman. PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs. In *International Conference on Learning Representations*, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, 2020.

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*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation Finetuning for Language Models. In *Advances in Neural Information Processing Systems*, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *International Conference on Machine Learning*, 2025.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Evaluation Grid and Configuration

The grid crosses six Pythia sizes (Biderman et al., 2023), eight training checkpoints (1k, 2k, 4k, 8k, 16k, 32k, 64k, 143k), and four synthetic task families, for 192 evaluation units. Table 5 gives the parameter counts. Total compute is approximately 240 instance-hours on instances with two A100-SXM4-40GB GPUs.

Table 5: Model sizes and parameter counts.

Model	Parameters
Pythia-160M	1.62×10^8
Pythia-410M	4.05×10^8
Pythia-1B	1.01×10^9
Pythia-2.8B	2.78×10^9
Pythia-6.9B	6.85×10^9
Pythia-12B	1.18×10^{10}

The task families are T (template contrast), D (distractor robustness), P (prior-bias override), and N (negative transfer). T isolates an answer-format contrast; D adds competing surface cues; P contrasts the target with a learned prior; N tests interference from related but incorrect structure.

The configuration uses logistic probes selected at a best site per unit and frozen after development; single-site additive interventions; a fixed grid of intervention strengths; and random directions matched for site and norm. Per-unit seeds derive from MD5 digests of unit identifiers.

B Proofs for Section 4

Proof of Proposition 1. For v uniform on S^{d-1}, $\mathbb{E}[v] = 0$ and $\mathbb{E}[vv^\top] = I/d$, so $\bar{g}^\top v$ has mean zero and variance $\|\bar{g}\|^2/d$. Substituting Eq. (7) into Eq. (5), the factor λ cancels. As d grows, $\sqrt{d}\,\bar{g}^\top v/\|\bar{g}\|$ converges in distribution to a standard normal variable, whose median absolute value is $\Phi^{-1}(3/4)$. □

Proof of Proposition 2. Within each class, $w^{*\top}a$ is Gaussian with variance $w^{*\top}\Sigma w^* = \Delta_M^2$, and the class means differ by $w^{*\top}\Delta\mu = \Delta_M^2$; hence $P(w^{*\top}a_1 > w^{*\top}a_0) = \Phi(\Delta_M/\sqrt{2})$. The cosine bound is the Kantorovich inequality for the positive definite matrix Σ^{-1}; equality holds when $\Delta\mu$ combines the eigenvectors of the largest and smallest eigenvalues, with squared weights proportional to the opposite eigenvalue. For the construction, take $d = 2$, $\Sigma = \mathrm{diag}(1, \sigma^2)$, and $\Delta\mu = (t, \sigma c)$ with $t, c > 0$ and $t^2 + c^2 = \Delta_M^2$. Then Δ_M, and therefore the AUROC, is fixed, while $\cos(\Delta\mu, \Sigma^{-1}\Delta\mu) = (t^2 + c^2)/\left(\sqrt{t^2 + \sigma^2 c^2}\,\sqrt{t^2 + c^2/\sigma^2}\right) \to 0$ as $\sigma \to 0$. □

C Complete Intervention Grid

Table 6: Test-phase median standardized effect z_e across the four task families in each model–checkpoint cell. Bold entries cross the threshold $|z_e| \geq 2$.

Model	1k	2k	4k	8k	16k	32k	64k	143k
160M	-0.59	-0.13	$+0.17$	-0.95	$+0.20$	$+0.55$	-0.16	-0.60
410M	-0.83	$\mathbf{-2.65}$	-0.07	$+0.25$	-0.03	$+1.17$	-0.79	$+1.22$
1B	-1.81	-0.57	-0.67	$+1.21$	$+1.38$	$+0.13$	$+0.12$	$+0.97$
2.8B	$\mathbf{-2.68}$	$\mathbf{-2.48}$	-0.00	-0.91	$+0.75$	$+0.72$	$+0.45$	-0.56
6.9B	-1.88	-0.55	-0.36	$+0.80$	-0.17	$+0.70$	$+0.51$	$+0.45$
12B	-0.12	$\mathbf{-2.43}$	$+0.63$	$\mathbf{+2.49}$	-0.31	$+1.04$	-0.06	-0.56

The standardized effect is defined in Equation 5 from paired edited and unedited scores (Equation 4). The development-phase counterpart of the 12B step-8k summary is $+2.04$.

D Behavioral Scores and Activation-Magnitude Summaries

Table 7: Output-level answer-score AUROC on the development split, for the four model sizes with complete tabulated trajectories.

Model	1k	2k	4k	8k	16k	32k	64k	143k
1B	0.647	0.475	0.478	0.527	0.583	0.607	0.665	0.711
2.8B	0.534	0.517	0.519	0.621	0.687	0.791	0.857	0.968
6.9B	0.569	0.481	0.493	0.570	0.775	0.875	0.877	0.896
12B	0.524	0.512	0.632	0.561	0.662	0.752	0.815	0.909

For 12B, the projected RMS activation grows from approximately 0.038 at step 1k to 2.16 at step 143k, the 56.8-fold increase in Table 4.

E Auxiliary Onset and Ordering Analyses

The onset analysis regresses onset step on $\log_{10}$ parameter count using weighted least squares with weights $1/\mathrm{se}^2$. It includes a main analysis over six sizes and a secondary analysis over five sizes. For the time analysis, it regresses a milestone index on log step for each size and applies a majority rule to the resulting interval-based classifications. Table 8 summarizes the results.

Table 8: Auxiliary analysis outputs. Both scale intervals include zero; the time rule produces no majority.

Analysis	Variant	Estimate	n	95% CI / outcome
Scale	main	$+0.240$	6	$[-0.600, +0.866]$
Scale	secondary	$+0.325$	5	$[-0.121, +1.182]$
Time	majority rule	—	6	3 negative / 3 unresolved

The three time classifications that are negative are 410M, 2.8B, and 6.9B. The 2.8B classification has an upper confidence limit of -5.8×10^{-12}, and the time rule yields no majority whether it is counted as negative or unresolved.

Eleven units have the internal-leading milestone ordering, and the catch-up regression has a 95% interval of $[-0.722, +0.438]$.

F Checkpoint Provenance

A developmental comparison requires each training label to correspond to distinct checkpoint content. The pipeline uses a content-hash check that compares weight payloads for duplicate content across training steps, quarantines affected lineages, and pins the revision used for evaluation. This check flagged a Pythia-2.8B lineage, and all measurements use a replacement revision that passed the check.

External reports document related problems. Several Pythia-2.8B repository generations can be distinguished (Li, 2026): some files repeat a checkpoint across training labels, while sharded files in one generation match the deduplicated branch at the corresponding step. A community issue reports identical weights across early steps 0–6,300 (HJXA, 2026). These reports underline that developmental comparisons require both checkpoint-content checks and explicit repository and revision provenance.

G Prior-Bias Analysis

The auxiliary prior-bias analysis fits bias surfaces over intervention strengths and aggregates 24 size–task combinations. The weighted regression of bias magnitude on log parameters has a 95% slope interval of $[0.00062, 0.291]$, with intercept -0.719. The interval is positive, indicating that the prior bias grows with model size.