FUNCTIONAL PORTABILITY, GEOMETRIC LIMITS: CROSS-CONSTRUCTION NUMBER INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-construction transfer reveals reusable output control, but which properties of the underlying mechanism does it identify? We fit grammatical-number directions on subject–verb agreement and apply them to disjoint reflexive outputs. Across 21 held-out Pythia size–seed models, all three target constructions reach persistent onset (63/63 target trajectories), versus 0/63 for a fixed orthogonal control; this original comparison uses unmatched edit amplitudes. A separate equal-edit-norm follow-up on exposed original materials retains an advantage over random and orthogonal directions, but not a uniform advantage over a simple lexical-number direction. Further boundary tests show that selected supports concentrate the effect while broader random supports recover much of it, and that changing the writing site changes a fixed direction's structural response. On a prospective BLiMP subset, target gain meets its requirement but joint collateral preservation does not; lowering dose reduces both gain and measured collateral. Supplementary reorientation on exposed materials offers no demonstrated advantage over a diagonal approximation. Together, these studies distinguish portable number control from evidence for a uniquely syntactic, compact, position-invariant, or selectively low-collateral mechanism.

1 INTRODUCTION

When a direction fitted to subject–verb agreement also changes a reflexive pronoun, what has transferred? An *is/are* direction that alters *himself/themselves* works beyond its fitted output labels. This excludes literal reuse of one output-token pair, but leaves reusable syntax, a generic number signal, and broader output perturbations compatible with the effect.

We ask what cross-construction transfer establishes beyond output control. After testing disjoint output labels, we examine three interpretive boundaries: coordinate support, writing site, and collateral change. Transfer alone need not identify unique compact support, site-independent responses, or low non-target cost.

Number studies link representations to agreement (Finlayson et al., 2021; Ravfogel et al., 2021; Lasri et al., 2022; Hao & Linzen, 2023); interchange methods test aligned variables through output effects (Geiger et al., 2022; 2024; Arora et al., 2024). We use difference-in-means directions to test what number transfer identifies, rather than propose another estimator.

Our contribution is an empirical separation of these properties, using a held-out transfer inventory and subsequent tests in separate settings, not one fully matched experiment (Figure 1):

- **Number control transfers to disjoint outputs.** All three reflexive targets reach matched onset in each of 21 held-out Pythia models. A separate same-task equal-norm comparison separates sentence directions from random controls, but not uniformly from lexical-number directions (Section 3).
- **Concentration does not establish unique support.** Selected 8.3% Pythia support succeeds in 14/14 models versus 0/14 for budget-matched random support, but broader random supports recover many effects. Matching norm and source cosine leaves a small local response difference, not an orientation-matched comparison (Sections 4–5).

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.

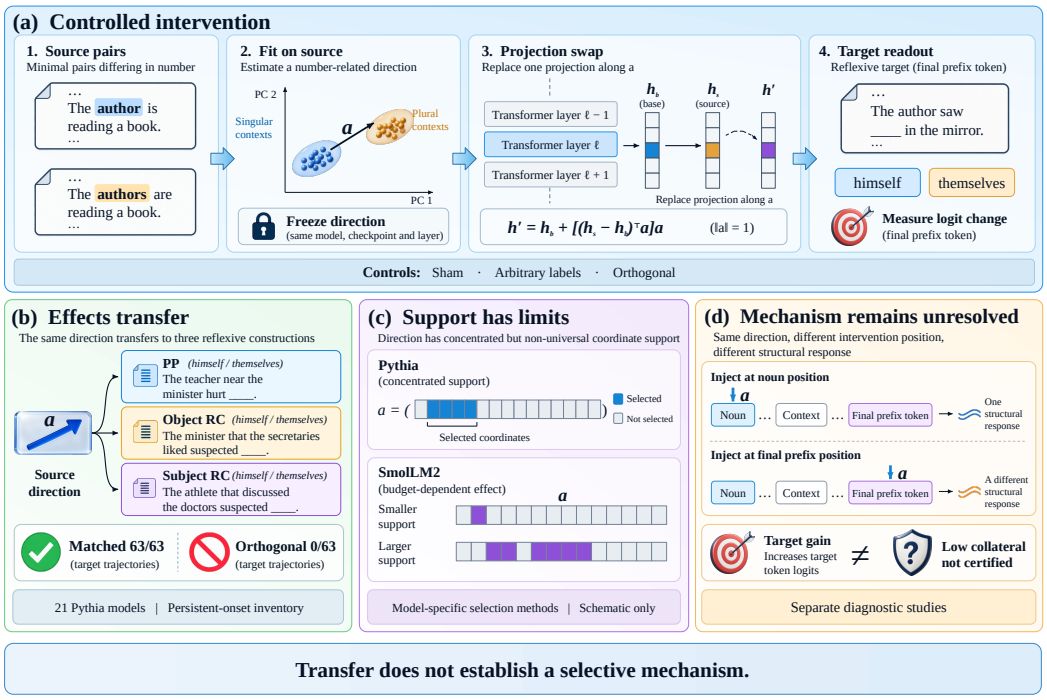

Figure 1: Readout control and its interpretation. (a) Rank-one interchange replaces a base projection with a source-donor projection. (b) Matched onset reaches 63/63 targets versus 0/63 orthogonal, nested in 21 Pythia models. (c) Support limits are study-specific: Pythia and SmolLM2 use different selectors; the grids are schematic. (d) Writing-site and collateral diagnostics test explanations left open by transfer. These separate studies do not identify a selective syntactic mechanism.

- **Structural response and collateral require separate tests.** Lexical directions retain structural interactions; moving a fixed direction changes its response. Lower dose reduces both gain and collateral, while supplementary reorientation shows no advantage over a diagonal approximation (Section 5).

The 63 targets share 21 models; they are not 63 independent model replications. Native transfer amplitudes are unmatched, and the later equal-dose and exposed-material diagnostics do not retrospectively repair that comparison. Table 1 states the unit and evidence status of each study.

2 Problem Setup and Study Design

2.1 Intervention and Output Measurements

The source tasks score subject–verb agreement using *is/are*, *has/have*, or *was/were*; targets score the disjoint reflexive pair *himself/themselves*. Exact examples appear in Appendix Table 7. Base and source sentences are fixed number-class counterparts within a lexical cluster, not necessarily same-lemma substitutions. The endpoint is reflexive-form *number choice*, not binding or licensing. Gender, frequency, and the availability of singular *themselves* also distinguish the output forms.

At a fixed model and layer, let $\mathbf{h}_b$ and $\mathbf{h}_s$ be hidden states at the chosen positions in a base sentence and its number-swapped source donor. We estimate a class-conditional difference-in-means vector from training activations and normalize it to obtain $\mathbf{a}$. Rank-one distributed interchange replaces only the projection onto this direction:

$$\mathbf{h}_b^* = \mathbf{h}_b + \left(\mathbf{h}_s^\top \mathbf{a} - \mathbf{h}_b^\top \mathbf{a}\right) \mathbf{a}. \tag{1}$$

This is the one-dimensional DII operator benchmarked by CausalGym (Arora et al., 2024); DAS learns an alignment for the same intervention family (Geiger et al., 2024). Using a simple estimator lets us study its effect across settings. Appendix B gives implementation details.

Comparing directions also changes the amount written. For unit $\mathbf{a}$, the native edit has length

$$\|\delta\mathbf{h}\|_2 = |(\mathbf{h}_s - \mathbf{h}_b)^\top \mathbf{a}|. \tag{2}$$

Changing direction or donor therefore changes both orientation and applied amplitude. Unit-normalizing the orthogonal control does not norm-match its edits to matched transfer. The later equal-dose diagnostics address this confound locally; they do not retrospectively amplitude-match the original checkpoint grid.

For each number orientation, N is the intervention-induced change in source-versus-base log-odds. Repeating this calculation with deterministically assigned arbitrary labels gives A, and the control-adjusted effect is $S = N - A$. A same-class donor supplies a second sham effect. The strong score averages $S_{\text{target}} - S_{\text{sham}}$ within each lexical cluster. Unmodified preference measures behavior; N measures intended-label movement; the adjusted score tests whether that movement survives both nuisance controls.

A passing score concerns the output effect of Equation 1. It does not establish necessity, natural mediation, or a common semantic interpretation for all retained coordinates. Table 1 separates the measurements and their evidence units.

2.2 Data, layers, and onset

Each training/transfer task has 50 test lexical clusters with two number orientations. Directions use training activations only; development data select the region by median arbitrary-label-adjusted S, not the later sham-adjusted score. Six depth-matched layer anchors prevent one favorable layer from defining onset. Splits, tie-breaking, and layer mappings appear in Appendix R.

The source-development region is mapped to the corresponding target region. Disjoint outputs exclude reuse of the fitted output-token pair, but remain compatible with generic number morphology. Arbitrary-label and same-class-sham controls address nuisance responses associated with the fixed lexical variation.

Exact one-sided sign tests use 50 paired cluster scores. A checkpoint must pass the natural-label effect, control-adjusted sham contrast, within-family Holm correction, and six-layer coverage. Output-effective onset is the start of the first adjacent pair satisfying these conditions. Behavioral onset applies adjacent persistence to unmodified preference. Strict lag requires an earlier output-effective index; ties fail. Persistence rejects isolated spikes but limits resolution to the sampled grid.

The detectors share persistence, but not their statistic, threshold, or power. "Earlier" therefore means an earlier sampled detector crossing, not earlier latent acquisition; unequal sensitivity to a shared signal remains a viable explanation.

Lexical clusters are nested within targets and models (Appendix Table 11). Pythia seeds select distinct pretrained repositories. Strong uses 29 unexposed trajectories; Transfer and support use overlapping subsets of 21 and 14, not additional weight replications. Targets share models and test schedules. Model-cluster summaries respect this nesting but do not establish independent training-data or model-population sampling. Appendix A gives model identities and checkpoint grids.

Intervention position. Matched Transfer edits the final token of the source-selected region in the target. Many selected sites precede the intervening phrase; none is at the final target matrix verb. Early-site transfer can carry noun number forward without encoding the future construction. The site inventory is not a position-matched causal comparison; Appendix A reports every selected site, and Appendix R gives the count breakdown.

2.3 Frozen analysis and negative evidence

Confirmatory materials and analyses precede held-out evaluation; "frozen" means project records, not external preregistration. The initial study failed its same-class-sham comparison. That negative is preserved and its outcomes are not reused as confirmation data (Table 1; Appendix R).

Measurement	Question	Evidence unit	Status and main limitation
Transfer	Control of disjoint reflexive outputs?	21 Pythia models; three targets each	Held-out confirmation; native dose is not matched
Equal-norm follow-up	Does direction matter at matched actual L2?	Eight snapshots of four Pythia trajectories; 24 target cells	Original exposed materials; descriptive only; not an onset replication
Support	Concentration relative to another selector?	14 overlapping Pythia models; one run per other family/size	Within-study comparisons; selectors differ between families
Source and site	Does structural response depend on fitting or writing location?	Shared lexical blocks within fixed checkpoints	Development/exposed diagnostics; no natural circuit identification
Matched support / reorientation	Do cosine and norm explain response; can collateral be reduced?	One Pythia checkpoint; 24 blocks per material, 8 used for fitting	Exposed supplementary tests; no full-Fisher advantage
Collateral	Does intended gain meet non-target tolerances?	22 sentences, 16 held-out head lemmas	Prospective external subset; one teacher-forced continuation

Table 1: Different questions require different measurements. Study units cannot be added as independent replications. The support studies measure native-basis concentration, not basis-invariant geometry; the collateral study does not certify natural mediation.

3 Cross-construction transfer

3.1 Held-out confirmation design

We test 21 held-out size–seed models, each on three reflexive-number targets paired with subject–verb sources. Seven seeds per size are evaluated at four checkpoints (512, 1000, 2000, and 3000), six layers, and 50 paired clusters per condition. The 63 target trajectories contain 252 model-checkpoint conditions. Pilot seeds 0–2 are excluded; the design precedes all target forward passes.

We compare three conditions. *Matched transfer* applies the source subject–verb direction to the corresponding reflexive target. *Within-target* estimates the direction on the reflexive construction itself. A deterministic hash-orthogonal vector is the negative control. Every condition also subtracts a same-class sham. The primary criterion requires persistent matched onset in at least 48/63 trajectories, coverage across every size and target, no more than 5/63 orthogonal onsets, measurable within-target onset, and a persistent matched-over-orthogonal advantage. Orthogonality is a deliberately sharp negative control for this rank-one geometry; it does not exhaust all alternative subspaces or prove unique semantic specificity.

3.2 Matched transfer succeeds in every tested model

All 21 size–seed models reach matched onset on all three targets. The corresponding inventory is 63/63 target trajectories: 3 at step 512 and 60 at step 1000. Within-target directions also reach onset in 63/63, whereas the hash-orthogonal control reaches onset in 0/63 (Figure 2a). The paired matched-minus-orthogonal advantage succeeds in all 63 trajectories. Every size and target meets its frozen coverage criterion. These are finite inventories with three dependent targets per model; historical target-level probabilities are retained only in Appendix Table 6, not used as calibrated independent-replication evidence.

Continuous effects agree with the positive onset separation. After averaging the three targets within each model, all 21 matched-minus-orthogonal effects are positive at step 1000: median 1.915, mean 1.995, and 95% model-bootstrap interval $[1.735, 2.276]$. This post-hoc summary complements the frozen onset test; later checkpoints appear in Appendix Figure 7.

A post-hoc sweep of 528 threshold and layer-gate settings retains 63/63 matched versus 0/63 orthogonal onsets across all tested score thresholds at the original four-of-six layer requirement. Stricter five- and six-layer requirements reduce coverage, and thresholds change some onset times (Appendix E). This is score sensitivity, not evidence of criterion-independent onset or matched edit amplitude.

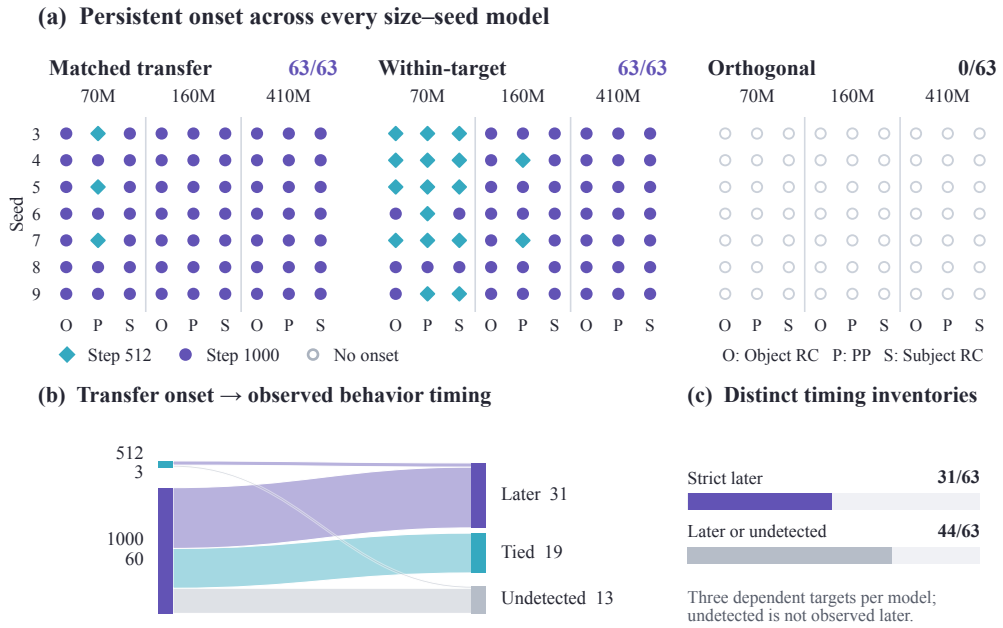

Figure 2: Matched source directions transfer to disjoint reflexive-form outputs. (a) Each marker is one target trajectory: seven seeds, three sizes, and three targets per condition. Shape and fill encode onset at step 512, step 1000, or nondetection. (b) Ribbons connect matched-onset checkpoints to observed behavior timing; widths encode exact counts, totaling 31 later, 19 tied, and 13 undetected. (c) Strict and strict-or-undetected inventories remain distinct. Three targets share each of 21 models, not 63 independent trials. Unequal detector sensitivity and censoring preclude latent-time ordering.

3.3 Transfer is distinct from temporal precedence

Behavior is later than matched onset in 31 trajectories, tied in 19, and undetected through step 3000 in 13. Strict lag is therefore 31/63; adding the 13 undetected cases yields the separate horizon-specific inventory 44/63 (Figure 2b–c). Undetected cases do not locate a later behavioral onset, and neither inventory establishes latent-time ordering.

Model-cluster sensitivities likewise leave strict precedence unresolved. Historical target arithmetic, exploratory model tests, and the detector-sensitivity matrix remain in Appendices C and R.

3.4 Which explanations remain?

An equal-actual-L2 follow-up revisits the original reflexive tasks at eight snapshots of four Pythia trajectories. With source-selected sites and fixed polarities, the native-aligned sentence direction exceeds random and orthogonal ensembles in all 24 snapshot–target cells (mean differences 2.250 and 2.254 nats). Its mean difference from a lexical-number direction is -0.005 nats, positive in only 15/24 cells. Thus, matched editing magnitude preserves the random-control separation but does not establish a distinctively syntactic source representation. These are exposed-material, finite-grid descriptives, not a new holdout or equal-dose onset replication (Appendix F). We next test coordinate support.

4 Support concentration and selector dependence

How much of a direction must be retained, and does its selected support have a distinctive advantage? We test native-basis geometry within each model. Coordinates are not transported between models, and these changed budgets and operators do not reproduce the original transfer design unchanged.

Magnitude-ranked support concentrates the Pythia effect. The coordinate-dose study covers 42 target trajectories nested in 14 Pythia size–seed models spanning 70M and 410M. We rank absolute training mean differences, equivalently $|a_i|$ after unit normalization. Selected 8.3% support reaches onset in 42/42 targets under both coordinate interchange and sparse rank-one intervention. Same-class shams and budget-matched hash-random supports reach 0/42. At the model unit, selected support succeeds on all three targets in 14/14 models, versus 0/14 for random support.

The budget curve locates the effect more precisely (Figure 3a). Selected support reaches 11/42 targets at 2.1% and 42/42 at 8.3%: 31 target gains, no losses, and 11 ties. All-three-target success changes from 0/14 to 14/14 models. The smallest universally passing tested budget is therefore 8.3%, with the transition lying in $(2.1\%, 8.3\%]$ on this grid. Post-hoc model and seed-block sensitivities are reported in Appendix R.

Concentration does not make the selected coordinates unique. Broad random 33% support recovers 39/42 targets. Pythia's magnitude ranking is the direction-energy selector used below; this study does not compare it against causal-leverage ranking. The measured advantage is over budget-matched random support.

The Pythia criterion does not reproduce on one OLMo-2 trajectory. We apply the same fractional doses, targets, layer anchors, shams, and random-support controls to 12 public checkpoints from one OLMo-2 1B stage-1 training trajectory. Selected 8.3%, full-coordinate, and full sparse anchors fail for all three targets; same-class and random controls also remain at zero (Figure 3b). This is a scoped non-replication of the Pythia criterion. It does not show that OLMo-2 lacks grammatical-number structure under other estimators, layer mappings, seeds, or interventions.

Removal changes outputs on fresh lexical material. After fitting source directions as before, we evaluate 50 new lexical clusters per target. Removing the full source-derived projection changes the frozen contrast in 42/42 targets; removing its selected 8.3% coordinate band does so in 39/42. Under the all-three-target model criterion, the counts are 14/14 for full removal, 12/14 for selected removal, and 0/14 for random removal. Post-hoc paired sensitivities appear in Appendix R.

The encoding-onset criterion is a separate measurement: no target trajectory passes (0/42). This is nondetection, not representational absence. The joint criterion is unmet, and off-target degradation was not measured. Removal therefore shows an operator response, not selective necessity or natural path mediation.

An implementation-specific SmolLM2 comparison. We test three checkpoint triplets from single SmolLM2-135M, 360M, and 1.7B runs (Allal et al., 2025). The causal-leverage selector ranks $|a_i g_i|$, using a batch-weighted source margin gradient that gives the final four examples four times the per-example weight (Appendix B). Its negative result does not establish failure of uniform-gradient selection.

Dense anchors pass every target at every size. Causal-leverage 8.3% support passes 3/3 targets at 135M and 360M, but 0/3 at 1.7B; 33.3% restores 3/3. Magnitude selection passes 3/3 at both smaller sizes and 2/3 and 3/3 at the two 1.7B budgets. A separate corrected-gradient follow-up on three exposed 1.7B checkpoints changes the supports but leaves all 18 layer-gate decisions unchanged. Its natural-label effect is below magnitude ranking in all 18 fixed conditions (Appendix G); this is not a full cross-scale onset replication.

At 1.7B, larger support adds a median 32.9 percentage points of retained energy as onset recovers (Figure 6). Budget also changes orientation and dose; this is neither a causal decomposition nor, with one run per size, a population scaling law.

These are distinct boundaries: OLMo-2 lacks a passing dense anchor under the tested protocol, whereas SmolLM2 retains dense effects but loses them at smaller supports. Neither identifies a population-level cause of cross-family variation.

5 Direction source, writing site, and collateral change

Separate follow-ups test sentence-derived fitting, writing site, and non-target cost. Their development, exposed-material, and prospective evidence statuses remain distinct.

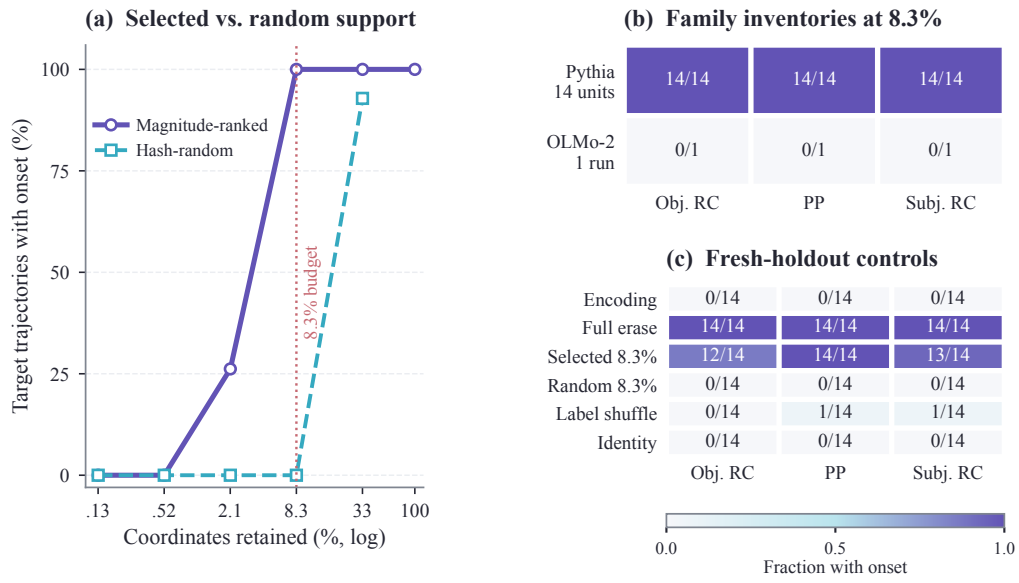

Figure 3: Selector- and operator-specific boundaries. (a) Pythia magnitude-ranked support reaches 42/42 targets at 8.3%, versus 0/42 for budget-matched hash-random support. Magnitude ranking is direction-energy ranking; causal leverage is not compared here. Lines connect tested budgets, not an inferred continuous threshold. (b) Unequal family coverage is explicit: 14 Pythia units versus one OLMo-2 trajectory per target. (c) Fresh-holdout removal and encoding inventories. Encoding onset is undetected under the frozen rule; removal is an operator response, not selective necessity. Selected counts sum to 39/42 targets but only 12/14 models pass on all targets. Colors denote finite inventory fractions, not population probabilities.

A lexical direction retains the role–order interaction. At Pythia-410M step 143000, we compare sentence-fitted and vocabulary-frame lexical directions on 12 new lexical pairs. The design crosses controller/attractor roles, both PP orders, and all four number assignments. All eight source-fitted directions receive the same applied $L_2 = 6.936$. To measure whether the controller advantage changes with order, define $R_o = G_{\text{controller},o} - G_{\text{attractor},o}$ and $I = R_{\text{fronted}} - R_{\text{postposed}}$.

DiM has $I = 1.634$ and lexical DiM has $I = 1.373$. Before evaluating Qwen2.5-0.5B, we fix the interaction prediction and refit every direction on its source activations. Qwen uses its own reference norm of 5.960. It gives $I = 0.954$ for DiM and 0.787 for lexical DiM (Table 27). Every DiM lexical-pair interaction is positive in each model. The local interaction persists at equal applied length and does not require sentence-derived fitting. The lexical direction mixes morphological number and distributional noun identity; it does not isolate either.

Moving a fixed direction changes its structural response. Holding a noun-derived direction and dose fixed, we move the edit from the first noun to the final prefix token. The site–scope interaction is 0.203 nats on relative-clause materials and 0.239 on after-gerund materials. Both sets were exposed. Direction identity alone therefore does not explain this local difference. A post-hoc analysis finds that the embedded/subject response matrices also deviate from exact multiplicative site gain on every observed lexical block at the lower dose. Dose-normalized responses vary across the two measured amplitudes, so the finite-dose interaction should not be interpreted as a measured derivative (Appendix T). Propagation distance, nonlinearity, and available context remain alternatives. Fixed L17/L23 follow-ups do not identify a selective circuit (Appendix Q); training-criterion timing is separate (Appendix D).

Equal length does not match direction. On the same Pythia checkpoint, selected and random truncated directions are normalized to the same applied norm. Selected 85/1,024 coordinates yield controller-site gains of 0.606/1.610 nats in postposed/fronted order, versus 0.183–0.326/0.440–0.799

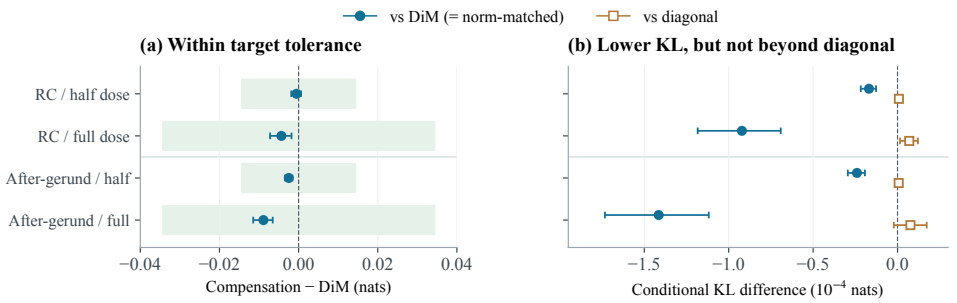

Figure 4: A constrained edit reduces conditional KL without establishing full-Fisher superiority. (a) Paired central-response intervals lie inside the calibration-defined tolerance bands. (b) Conditional-KL differences favor compensation over DiM, not over the diagonal approximation. RC: 16 fit-excluded lexical blocks; after-gerund: 24. Half/full denote fixed scales 0.5/1. Intervals are pointwise 95% block-bootstrap intervals, conditional on the fitted edit and tolerance. No untouched-data or per-example guarantee is implied.

for three fixed random supports. This rules out applied length alone for this local difference. It does not equate orientation: for unit dense a and normalized truncation $b = a_S/\|a_S\|$, $a^\top b = \|a_S\|$. The selected 85- and 341-coordinate supports have cosine 0.624 and 0.896 to dense DiM. Broader random supports recover much of the effect (Appendix N); coordinate uniqueness remains unestablished.

A difference remains after matching source cosine. A supplementary test constructs 64- and 256-coordinate edits at common cosine 0.174 and applied norm 1.734. Across relative-clause and after-gerund materials, the four selected-minus-random average response differences are 0.012–0.034 nats, with all four pointwise intervals above zero. The 64-coordinate subject-family contrast is unresolved in both sets. Source cosine and norm therefore do not determine response, but full orientation remains unmatched (Appendix U).

Target gain and collateral depend on the operating point. A prospective, text-audited BLiMP relative-clause distractor subset (Warstadt et al., 2020) meets its target-gain criterion but not the joint collateral criterion. It contains 22 test sentences from 16 held-out head lemmas. At reference dose, DiM's mean counterfactual gain is 2.125 nats, above the prespecified 0.5-nat criterion. Conditional non-target KL and absolute lexical ΔNLL average 0.04288 and 0.10809, respectively. Table 28 places each mean beside its criterion. Failed upper-bound certification alone does not prove a population violation; the exploratory opposite-tail lexical analysis is reported separately in Appendix R.

All eight methods receive the same 16 calibration doses. At DiM's calibration-selected dose of 0.375, mean target gain falls from 2.12528 to 0.76736 nats, conditional KL from 0.04288 to 0.00446, and absolute lexical ΔNLL from 0.10809 to 0.03214. These lower-dose sample means are not a joint-certification result. DiM also misses the required equal-benefit match with DAS, so this comparison establishes neither method superiority nor an unavoidable collateral tradeoff. Appendix P.2 retains every operating point; no dose was selected from test outcomes. The lexical probe covers one teacher-forced continuation, not generation quality.

Constrained reorientation modestly reduces collateral. Fitting a 16-dimensional conditional-Fisher proxy on 96 relative-clause rows yields a frozen edit that retains 97.8–99.6% of mean central response and reduces conditional non-target KL by 10.4–11.7% across two fit-excluded material sets and two doses. Absolute reductions are 1.69×10^{-5}–1.41×10^{-4} nats (Figure 4). A diagonal approximation has slightly lower mean KL in all four cells. Both material sets were exposed, and the criterion concerns mean response, not each example. This supplementary result establishes neither full-Fisher superiority nor repair of the BLiMP certificate; Appendix U gives the complete comparison.

6 Related work

Behavioral syntax and mechanistic number studies. Targeted and factorial evaluations test grammatical alternatives and structural generalization (Marvin & Linzen, 2018; Hu et al., 2020). Number/syntax units in LSTMs and causal number encodings in BERT already link representations to agreement (Lakretz et al., 2019; Hao & Linzen, 2023); AlterRep tests RC-boundary representations across RC types (Ravfogel et al., 2021). Cross-language studies compare agreement circuits (Ferrando & Costa-jussà, 2024; Gidi et al., 2026) and transfer DAS subspaces, including other-phenomenon and novel-label controls (Boguraev et al., 2026). Thus neither transfer nor disjoint-label testing alone is our novelty claim. We test auxiliary-to-reflexive number transfer together with selector, dose, lexical, and collateral limits; this is not a role- or licensing-specific generalization test.

Intervention benchmarking and interpretive validity. CausalGym benchmarks linguistic causal efficacy (Arora et al., 2024); DAS learns interchange alignments (Geiger et al., 2022; 2024). Grant et al. (2026) investigate representation divergence and a Counterfactual Latent loss. Our output KL differs from their hidden-state measure, and we do not test that mitigation. We separate support, source, site, and collateral empirically, without claiming estimator superiority or novelty for the general interpretive warning.

Training and support. Public checkpoints make it possible to follow representations through training (Biderman et al., 2023; Allal et al., 2025; Müller-Eberstein et al., 2023; Duan et al., 2025; Bayazit et al., 2026). Probe success need not imply use (Hewitt & Liang, 2019; Elazar et al., 2021; Lasri et al., 2022), and causal mediation (Finlayson et al., 2021) asks a stronger question than output controllability. These distinctions, together with sparse-steering evidence (Arad et al., 2025), motivate measuring support size separately from selector quality. Our family-specific failures delimit the tested native-basis criteria; they do not rule out all compact geometries or all syntactic mechanisms.

7 Discussion and limitations

Portability and its boundaries. In the equal-norm subset, sentence directions separate from random controls but not uniformly from lexical-number directions: surviving a dose control is not the same as identifying syntactic abstraction. Support concentration, site dependence, and dose-dependent collateral likewise require distinct explanations, not one common mechanism.

What the controls distinguish. Edit norm and source cosine alone do not explain the local support difference; full orientation and target-gradient alignment remain unmatched. A fixed noun-derived direction has site-dependent responses that deviate from exact multiplicative gain (Appendix T). Nonlinearity, propagation distance, and distributional damage remain alternatives. The L17/L23 follow-ups detect no additional joint scope effect and identify no selective two-node circuit.

Scope. Synthetic English number preference is not binding or generation quality; disjoint labels do not exclude generic morphology. Native-basis geometry, unmatched donor amplitude, and unmeasured removal collateral limit mechanism claims. Unequal detector power and unresolved strict lag (31/63) constrain timing; the conservative Strong seed block is 8/10 ($p = 0.0547$). Discrete checkpoints and single OLMo-2 and SmolLM2 runs limit inference about training runs. No result establishes training improvement.

Vaidyanathan et al. (2026) distinguish noising natural indirect effects from denoising pure indirect effects and isolate an interaction with bypass activations. Our writing-site contrast does not estimate this decomposition or identify that interaction as its cause. The matched-cosine differences of 0.012–0.034 nats are local diagnostic separations, not established practical steering improvements.

Implication. Transfer reports should specify labels, actual edit norm, support, writing site, and collateral. Our linear proxy permits target-invisible compensation (Appendix S), but the measured reduction does not outperform a diagonal approximation or establish natural selectivity. This 16-dimensional conditional-Fisher diagnostic is neither a full implementation nor a head-to-head comparison with FishBack's pullback-Fisher steering (Wang et al., 2026).

Reproducibility statement

The companion reconstructs 252 Transfer cells and 504 gates from raw score exports, and supplies a pinned-model fitting/intervention example executed on one Pythia-70M layer (400 source rows, 10 target orientations). Separate CPU entries reaggregate the equal-norm and corrected-gradient follow-ups from their supplied scores; weights and the bridge's full-vocabulary arrays are excluded. These routes do not replay every study or retrain models. Appendices A and R specify inputs and coverage.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259, 2025. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding through time: Tracking emergence and consolidation of linguistic representations throughout LLM pretraining. *arXiv preprint arXiv:2509.05291*, 2026. doi: 10.48550/arXiv.2509.05291. URL https://arxiv.org/abs/2509.05291. Accepted to ACL 2026.

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.

Sasha Boguraev, Toshiki Nakai, Kyle Mahowald, and Julius Steuer. Causal interventions reveal typologically organized syntactic mechanisms in multilingual language models. *arXiv preprint arXiv:2608.28924*, 2026. URL https://arxiv.org/abs/2608.28924. Version 1, August 28, 2026.

Xufeng Duan, Zhaoqian Yao, Yunhao Zhang, Shaonan Wang, and Zhenguang G. Cai. How syntax specialization emerges in language models. *arXiv preprint arXiv:2505.19548*, 2025. URL https://arxiv.org/abs/2505.19548.

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. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Javier Ferrando and Marta R. Costa-jussà. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10115–10125, 2024. doi: 10.18653/v1/2024.findings-emnlp.591. URL https://aclanthology.org/2024.findings-emnlp.591/.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Isabella Gidi, Antonio Almudévar, Core Francisco Park, Naomi Saphra, and Ricard Marxer. Shared circuits for shared grammar: Tracing subject-verb agreement across languages. *arXiv preprint arXiv:2608.18545*, 2026. URL https://arxiv.org/abs/2608.18545. Version 1, August 19, 2026.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cZrTMqYVL6.

Sophie Hao and Tal Linzen. Verb conjugation in transformers is determined by linear encodings of subject number. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4531–4539. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.300. URL https://aclanthology.org/2023.findings-emnlp.300/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725–1744, 2020. doi: 10.18653/v1/2020.acl-main.158. URL https://aclanthology.org/2020.acl-main.158/.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. In *Proceedings of NAACL-HLT*, pp. 11–20, 2019. doi: 10.18653/v1/N19-1002. URL https://aclanthology.org/N19-1002/.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8818–8831. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.603. URL https://aclanthology.org/2022.acl-long.603/.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2018. doi: 10.18653/v1/D18-1151. URL https://aclanthology.org/D18-1151/.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13190–13208.

Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.879. URL https://aclanthology.org/2023.findings-emnlp.879/.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.15. URL https://aclanthology.org/2021.conll-1.15/.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching, 2026. URL https://arxiv.org/abs/2606.27510v1. arXiv preprint, v1.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback fisher geometry for optimal activation steering in transformers, 2026. URL https://arxiv.org/abs/2605.17231v2. arXiv preprint, v2.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

AI USE STATEMENT

Generative AI assisted hypothesis refinement, methodological review, implementation and testing, data inspection, figure creation, and drafting. Assisted code was tested; separate numerical checks and their raw-versus-derived coverage are documented in the supplement. Such checks are not independent scientific replication. The authors reviewed and take responsibility for the final work.

A AUDITED UNITS, CHECKPOINT TESTS, AND INTERVENTION SITES

Weight identities and reuse. The pretrained-seed index selects a model repository, not merely an evaluation random-number seed. Seed 0 uses `EleutherAI/pythia-<size>`; seeds 1–9 append `-seed<n>`. The registry contains 30 size–seed trajectories and ten revisions per trajectory. Its 300 source-weight SHA-256 digests are distinct; the archived prefetch manifest also records 300 distinct loaded-safetensor digests. These are checks of archived identities, not a fresh rehash of every model file or proof of independent training data/order. Transfer and M4 reuse the corresponding Strong weights (Table 2); they must not be added as new training replications. Each Transfer target uses the same 100 test rows, paired into 50 lexical clusters, across all its models and checkpoints. Targets, layers, checkpoints, and reused studies are repeated measurements, not independent model draws.

Study	Weight trajectories	Inclusion and relationship
Strong	30 registered: 3 sizes × seeds 0–9	Primary $n = 29$: omit 70M seed 0 because PP and object-RC were previously examined; its subject-RC trajectory remains in the template analysis.
Transfer	21: 3 sizes × seeds 3–9	Subset of the same pretrained weights; pilot seeds 0–2 excluded. Three reflexive targets share each model.
M4	14: 70M/410M × seeds 3–9	Subset of Strong/Transfer weights; held-out-scale evaluation of the hypothesis developed on 160M, not new independent training.

Table 2: Experimental identities. Sizes are 70M, 160M, and 410M unless restricted. Exact repositories, commits, and weight digests are supplied in the audit CSVs. Distinct archived weight identities do not establish an IID training-population model.

Raw-logit and cluster reductions. For lexical cluster c, number orientation $o \in \{1, 2\}$, and selected layer $\ell \in \{1, \ldots, 6\}$, write b, s for the natural base and counterfactual source label tokens. With baseline logits z and intervened logits z', define

$$\begin{aligned} N_{co\ell} &= (z'_s - z'_b) - (z_s - z_b), \\ A_{co\ell} &= (z'_{s_a} - z'_{b_a}) - (z_{s_a} - z_{b_a}), \\ S_{co\ell} &= N_{co\ell} - A_{co\ell}, \qquad R_{co\ell} = S^{\text{active}}_{co\ell} - S^{\text{sham}}_{co\ell}. \end{aligned} \tag{3}$$

Here b_a, s_a are the fixed arbitrary-label pair. The nested logit subtractions producing N, A use CPU float32; their subsequent subtraction uses float64. The active and same-class-sham records are paired by cluster, orientation, layer, and selected region. Each checkpoint contributes exactly 50 observations

$$\begin{aligned} \bar{N}_c &= \frac{1}{12} \sum_{o,\ell} N^{\text{active}}_{co\ell}, \qquad \bar{R}_c = \frac{1}{12} \sum_{o,\ell} R_{co\ell}, \\ \bar{R}_{c\ell} &= \frac{1}{2} \sum_{o} R_{co\ell}. \end{aligned} \tag{4}$$

Thus the checkpoint test averages all twelve orientation–layer values within a cluster; it does not select a favorable layer or take a layerwise maximum.

Checkpoint tests, thresholds, and equalities. Let $Q(k) = 2^{-50} \sum_{j=k}^{50} \binom{50}{j}$. The natural and adjusted component tests are

$$\begin{aligned} p_N &= Q\left(\sum_c \mathbf{1}\{\bar{N}_c > \log 1.1\} \right), \\ p_R &= Q\left(\sum_c \mathbf{1}\{\bar{R}_c > \log 1.5\} \right), \qquad p_I = \max(p_N, p_R). \end{aligned} \tag{5}$$

The intersection–union intervention test uses p_I. The binomial calibration is conditional on independent cluster indicators whose null success probabilities are at most one half; distinct row hashes do

Study / family	Complete chronological checkpoint grid
Strong	0, 128, 256, 512, 1000, 2000, 3000, 4000, 10000, 143000
Transfer; M4; M6	512, 1000, 2000, 3000
Pythia	
OLMo-2 1B stage 1	0/0, 10000/21, 20000/42, 40000/84, 100000/210, 200000/420, 400000/839, 600000/1259, 800000/1678, 1200000/2517, 1600000/3356, 1907359/4001
SmolLM2-135M	240000, 960000, 1920000
SmolLM2-360M	160000, 1280000, 2560000
SmolLM2-1.7B	125000, 2500000, 5125000

Table 3: Complete primary/geometry grids. Pythia revisions are `step<n>`; SmolLM2 revisions are `step-<n>`. OLMo entries are step/token-billions pairs, naming revisions `stage1-step<n>-tokens<b>B`. No Strong checkpoint is dropped; its exposure exclusions concern trajectories, not checkpoints. The final entry is the observation endpoint, and the penultimate entry is the latest possible onset under adjacency. OLMo checkpoints belong to one training run; each SmolLM2 size likewise contributes one run, not three training replicates.

not verify this sampling assumption. Without it, the computed tails and decisions remain reproducible finite-schedule summaries, not an established population error guarantee. Equality to a component threshold is a failure retained among all 50 trials, not a discarded sign-test tie. Behavior instead uses

$$B_c = \mathbf{1}\{z_{b,c1} > z_{s,c1} \text{ and } z_{b,c2} > z_{s,c2}\}, \qquad p_B = Q\left(\sum_c B_c\right). \tag{6}$$

Both number orientations must prefer the correct base label. The margin threshold is zero; equality in either orientation fails the cluster. No extra positive log-odds threshold is imposed on behavior.

The additional layer gate requires at least four medians $m_\ell = \text{median}_c\, \bar{R}_{c\ell}$ to satisfy $m_\ell \geq \log 1.2$, and all six to satisfy $m_\ell \geq -\log 1.2$. These two boundaries are inclusive. There are no six layerwise significance tests: the layer gate is separate from the checkpoint p_I.

Holm, persistence, and the advantage family. For each fixed size–seed–task–condition, Holm correction at $\alpha = .05$ covers the complete checkpoint grid in Table 3. Intervention and behavior have separate families, using p_I and p_B, respectively. Sort ascending by p, breaking exact ties by chronological checkpoint order. At ordered rank r in a family of size m, reject while $p_{(r)} \leq .05/(m-r+1)$ and stop at the first failure. After correction, intervention eligibility is Holm rejection *and* the layer gate; behavior eligibility is its Holm rejection alone. Onset is the left endpoint of the first adjacent eligible pair. Persistence does not require every subsequent checkpoint to pass, and an isolated final passing checkpoint cannot define onset. Strict lag requires both detected onsets and a strictly smaller intervention index; ties and nondetection fail strict lag. The separately reported Transfer right-censored composite additionally counts detected transfer with behavior still undetected through step3000.

The matched-over-orthogonal comparison is a different family: at each of four Transfer checkpoints, pair the 50 cluster scores and test $\bar{R}_c^{\text{matched}} - \bar{R}_c^{\text{orthogonal}} > 0$ with the same exact upper tail, keeping zero differences as failures. Apply Holm and adjacent persistence to these four p-values. This advantage test does not additionally impose the natural/rescue magnitude thresholds or the six-layer median gate.

Primary source and target positions. Source development data choose a region separately for every size–seed–checkpoint–layer by maximizing the median lexical-cluster $S = N - A$, *before* same-class-sham subtraction. Exact median ties follow the frozen region order in Table 4. The source-trained direction and source-selected region are imported together into primary matched Transfer: the target uses that same named region, not a separately optimized target site. Within-target directions have their own development selection, whose candidate order appends the matrix verb. At every named region, the site is its final token, `span.stop - 1`; output logits use the final input token.

Across all $21 \times 3 \times 4 \times 6 = 1512$ primary selected sites, subject accounts for 1051 (69.5%), object 266, RC verb 122, *the* 39, *that* 27, and preposition 7. None uses the target's final matrix-verb/pre-output

Construction	Source candidate order	Within-target addition
PP	subject, preposition, the, object	matrix verb, last
Object RC	subject, that, the, object, RC verb	matrix verb, last
Subject RC	subject, that, RC verb, the, object	matrix verb, last

Table 4: Candidate sites. Primary matched target sites inherit the chosen source region; the final target matrix verb is not a source candidate. Selection is layer- and checkpoint-specific.

Target	Before marker	Partial phrase	Full phrase
PP	360 / 4.024	27 / 0.004	117 / 0.000
Object RC	347 / 2.615	41 / 0.001	116 / 0.000
Subject RC	344 / 3.798	37 / 0.009	123 / 0.000
Total sites	1051	105	356

Table 5: Descriptive primary selected-position strata. Each task entry is site count / median of the layerwise 50-cluster median R (nats), pooled over the tested models and checkpoints. These selected-site summaries are neither independent replicates nor a position-matched causal contrast; model, layer, checkpoint, and selection differ across strata. The original onset gate combines all six selected layers, so its trajectory onset count cannot be assigned as an independent outcome to each stratum.

site. Table 5 separates causal-prefix availability from the observed effect. A subject site precedes the construction marker. "Full phrase" means the object in PP/subject-RC and the RC verb in object-RC; the remaining non-subject sites have only partial intervening material. This is a prefix-availability classification, not a claim that the corresponding syntactic variable is encoded.

Some primary selected positions therefore have post-context information available, but the archive does not provide a matched before/after-position intervention comparison or a primary matched pre-output-site comparison. The separate fixed-subject 3×3 diagnostic matrix does not supply those missing contrasts.

Conditional normalized bands. For the continuous Transfer diagnostic, average the three target-level medians within each model, then normalize each measure by its observed full-sample 21-model mean at step3000. The existing pointwise 95% percentile bands use 100,000 resamples of the 21 model values *conditional on that fixed observed late reference*; the denominator is not re-estimated per resample. Each checkpoint/measure is bootstrapped separately, so these are not simultaneous trajectory bands or paired-ratio uncertainty intervals. Normalizing the two different statistics does not equate detector power.

Audit coverage and reusable files. The accompanying repository supplement contains the audit under `reviews/repair_20260906/units/`. File `unit_identity.csv` lists the 30 weight-trajectory identities and study membership; `checkpoint_weight_identity.csv` supplies all 300 full commits and source-weight digests. File `primary_transfer_positions.csv` gives all 1512 source/target choices by model, checkpoint, and layer, linked to source-cell digests; `primary_transfer_context_strata.csv` contains descriptive effect ranges. File `transfer_onset_reconstruction.csv` records the 63 reconstructed trajectories. These paths are relative to that audit directory.

The CPU audit verified all 252 Transfer cell hashes against the terminal and reproduced all 504 matched/orthogonal checkpoint gates from archived orientation effects, sham records, and behavior logits, with no discrepancies. It recovered matched onset in 63/63, orthogonal onset in 0/63, strict lag in 31/63, and behavioral nondetection in 13/63. Within-target onsets were reconstructed from stored gates. It did not regenerate logits, refit directions, or load model weights. The locally inspected Strong key-results archive contains trajectory reports and provenance but not its full 900 raw cells; this audit therefore does not claim a new full raw-score reproduction of Strong. The implementation, coverage limits, and original source-code locators are in `audit_units.py`, `AUDIT.json`, and `REPORT.md`.

B INTERVENTION AND CONTROL SPECIFICATION

Hook and readout. Original Strong/Transfer interventions replace a vector in the decoder block's output residual tensor. Layers are zero-indexed: Pythia uses `gpt_neox.layers`, whereas the SmolLM2 adapter exposes `model.layers`. The original six Pythia anchors are $(0, 1, 2, 3, 4, 5)$, $(1, 3, 5, 7, 9, 11)$, and $(3, 7, 11, 15, 19, 23)$ at 70M, 160M, and 410M, respectively. A named region selects its last token, not the following token; predictions are read at the final active input position. Models remain frozen in evaluation mode. This is not attention-head replacement or an edit to the final layer-normalized state.

Three distinct operators. Write h and h_s for base and donor states, d for the source-training mean difference, and P_S for the diagonal mask of support S. Dense rank-one interchange uses $a = d/\|d\|_2$ and $h' = h + [(h_s - h)^\top a]a$. Coordinate interchange and sparse rank-one interchange are, respectively,

$$h'_{\text{coord}} = h + P_S(h_s - h), \qquad b_S = \frac{P_S d}{\|P_S d\|_2}, \qquad h'_{\text{rank1}} = h + [(h_s - h)^\top b_S]b_S. \tag{7}$$

Coordinate interchange directly copies the selected donor coordinates, with no normalization. Sparse rank-one interchange normalizes *after* masking. The Pythia coordinate-dose study evaluates both operators; SmolLM2 fractional-support comparisons use coordinate interchange, with rank-one interchange additionally evaluated at full width. Neither native operator equalizes edit length across supports or control directions. Magnitude supports rank $(-|d_i|, i)$; every registered width is evaluated, not selected by the best target score. Hash-random supports include width in their identity and need not be nested across widths.

Centered removal. The fresh-holdout removal study fits a normalized direction b_S separately for each support and sets its scalar projection to the source-class midpoint:

$$c_S = \tfrac{1}{2}\left(\mathbb{E}_{\text{train},0}[h^\top b_S] + \mathbb{E}_{\text{train},1}[h^\top b_S]\right), \qquad h' = h + (c_S - h^\top b_S)b_S. \tag{8}$$

Thus removal eliminates the centered scalar channel, not every coordinate in the selected band, and generally does not set $h'^\top b_S$ to zero. The implementation uses CPU-float32 replacement and bounded residual corrections along the same direction; coordinates outside the support are unchanged. These details distinguish removal from coordinate interchange and from an identity/no-op control.

Original controls. Arbitrary output labels are the fixed leading-space strings `dog` for plural and `give` for singular; only scored labels change. They are not freshly sampled labels on each trial. Same-class sham donors are selected by sorting canonical row hashes, then scanning cyclically to the first row with the same base class and a distinct base endpoint. Its base-state activation supplies the donor, retaining the evaluated row's label contrast. This is neither a no-op nor necessarily an edit-length-matched control. The Transfer orthogonal vector starts from hash-generated Rademacher signs r, removes $(r^\top a)a$, and normalizes the remainder before float32 quantization. One deterministic vector is used per size–seed–checkpoint–target–layer identity, not an ensemble. The original class convention orients DiM singular-minus-plural, but its sign does not affect projection replacement.

Implemented SmolLM2 gradient weighting. A source audit identified an averaging discrepancy in the frozen selector. For batch B, its code differentiates the mean margin and then averages the site gradients across rows. The outer accumulator weights that result by batch size and divides by $N = 100$, giving, apart from roundoff,

$$g_{\text{impl}} = \frac{1}{N}\sum_B \frac{1}{|B|}\sum_{r\in B} \nabla_{h_r} m_r, \qquad m_r = z_{r,1} - z_{r,0}. \tag{9}$$

Six batches have 16 rows and the final batch has four. Consequently, the last four rows receive four times the per-row weight of the preceding rows; this need not leave support rankings unchanged. The archived interventions and target measurements remain those of $\text{topk}\,|a_i g_{\text{impl},i}|$. Appendix G reports a separate uniform-gradient sensitivity on three 1.7B checkpoints; corrected cross-scale onset remains untested. No historical support, threshold, or outcome was replaced after discovering this discrepancy.

Later fixed-dose diagnostics. The development panels use their own source-fitted directions, shuffled-source-label and Gaussian-random controls, and signed additive edits rather than the original arbitrary-output-label and orthogonal controls. Their reference amplitude is the lower median of 16 absolute source-pair projections onto DiM. Each panel's sign, edit site, and normalization rule must be read with its design: these diagnostics are not interchangeable replications of native projection replacement. Their results cannot retroactively establish equal-dose control in the original confirmation studies.

Specifically, in the equal-dose Pythia and Qwen role–order panels, let $n_{\mathrm{site}} = 0$ denote a singular edited noun and $n_{\mathrm{site}} = 1$ a plural one. For multiplier $d > 0$, reference r, and method unit vector v, the edit is

$$s = 1 - 2n_{\mathrm{site}}, \qquad \delta_0 = sdrv, \qquad \delta = \frac{dr}{\|\delta_0\|_2}\delta_0, \qquad h' = h + \delta. \tag{10}$$

The final float32 norm is checked against dr to absolute tolerance 2×10^{-5}. The Pythia panel reuses and renormalizes its frozen development-source vectors and inherits their reference amplitude; Qwen refits on its own source examples. The earlier unequal-dose pilot instead used the controller's number to set the sign at both sites, and is not substituted for these corrected site-specific panels.

Removal-study encoding detector. This detector reads the selected-support scalar projection $q = h^\top b_S$ and predicts class one when $q > c_S$ (a tie predicts class zero). For each of 50 lexical clusters, the correctness indicator is averaged over two balanced orientations and six layers to obtain A_c. The checkpoint component is the one-sided binomial tail for $K = \sum_c \mathbb{1}[A_c > 0.75]$ out of 50 at null probability one half. A separate gate requires at least four layer balanced accuracies strictly above 0.75 and all six at least 0.5. The detector then applies checkpoint Holm correction and adjacent-pair persistence. Its inferential interpretation requires the same component-validity assumptions stated in the statistical decision rules; it is not a claim of calibrated equal power against the behavior detector.

C HISTORICAL TARGET-LEVEL CALCULATIONS RETAINED FOR AUDIT

Table 6 preserves outputs of the original frozen arithmetic. Three targets share each size–seed model and use fixed schedules. These probabilities are therefore not used as calibrated independent-model inference in the revised main text. No record, threshold, or authoritative terminal has been changed. The paired continuous effects and separately labeled model/seed sensitivities retain their own scope; they are not new confirmatory replacements.

Behavior onset occurs at step 1000 in 20 targets and 2000 in 30, with 13 undetected. The strict-or-undetected composite is 20/21 at 70M and for object-RC, 11/21 at 410M, and 10/21 for PP. Among the 50 detected cases, 31 have strict lag. The historical independent-Bernoulli calculation would require 39/63 successes to cross 5%, hence at least 8 of 13 undetected cases resolving as later. This is arithmetic conditional on a hypothetical resolution, not calibrated power or an independence-supported population test.

Historical calculation	Target inventory	Original one-sided p
Transfer matched onset	63/63	1.084×10^{-19}
Transfer paired advantage	63/63	1.084×10^{-19}
Transfer strict timing	31/63	0.599
Transfer strict-or-undetected composite	44/63	0.001114
Strict timing among observed behavior	31/50	0.0595
Pythia 2.1% to 8.3% support	31 gains, 0 losses	4.66×10^{-10}

Table 6: Historical dependent-target probabilities, reported solely for traceability. The composite's Bernoulli null is success probability at most 0.5, not equal onset distributions; that arithmetic additionally requires an independence assumption not justified by counting targets. Ties and horizon-undetected cases are retained as originally defined.

D A SEPARATE TIMING TEST ACROSS PYTHIA TRAINING

D.1 MODELS, TASKS, AND HELD-OUT TRAJECTORIES

We use the 70M, 160M, and 410M Pythia model families because their public intermediate checkpoints permit controlled comparisons across training and scale (Biderman et al., 2023). The strong study crosses ten seeds, three subject–verb agreement templates (PP, object relative clause, and subject relative clause), and ten checkpoints from initialization through step 143000. Six approximately depth-matched layers are evaluated at every checkpoint. Each cell contains 50 paired lexical clusters in two number orientations. The resulting design contains 900 cells and 300 model-checkpoint groups.

The primary unit is a size–seed trajectory. Two previously examined 70M-seed0 template trajectories are excluded, leaving 29 complete unseen trajectories. Subject relative clauses are a forced holdout. These exclusions and all outcome rules were fixed before any strong-study model forward pass.

D.2 THE FROZEN INTERVENTION CRITERION USUALLY CROSSES FIRST

The frozen trajectory analysis yields 27 successes among 29 complete unseen trajectories ($p = 8.121 \times 10^{-7}$; Figure 5d), with a descriptive Wilson interval of $[0.780, 0.981]$. All three scale tests pass after Holm correction: 70M has 9/9 successes ($p_{\mathrm{Holm}} = 0.00586$), 160M has 9/10 ($p_{\mathrm{Holm}} = 0.0215$), and 410M has 9/10 ($p_{\mathrm{Holm}} = 0.0215$; Figure 5b). The two failures, 160M-seed7 and 410M-seed4, are sampled-onset ties in two templates rather than cases in which behavior crosses first. The rule counts these ties as failures.

This evidence is strongest as a finite-grid operational inventory. The frozen test treats each size–seed trajectory as a unit, but the registry does not establish independence between sizes sharing the same numeric seed. A conservative post-hoc block requiring every available size for a seed to pass yields 8/10 blocks ($p = 0.0547$); a majority-of-sizes rule yields 10/10 ($p = 9.77 \times 10^{-4}$). Leave-one-seed-out counts remain 24/26–25/27. The trajectory result is therefore robust to omitting any seed, while its pooled p-value is not a seed-blocked population test.

The continuous onset inventory shows where the aggregate separation arises. Across all 90 size–seed–template trajectories, including the two exposed trajectories only for this descriptive summary, output-effective onset occurs at step 512 for 4 trajectories, step 1000 for 79, and step 2000 for 7. Behavioral onset is more dispersed: 12 at step 1000, 64 at step 2000, 5 at step 3000, 7 at step 4000, and 2 at step 10000 (Figure 5a). The result is thus not driven by a few very late behavioral outliers; it mostly reflects a one-grid-step difference between concentrated intervention and behavioral crossings.

Template family provides the clearest boundary. Object relative clauses have 26/29 successes ($p_{\mathrm{Holm}} = 1.52 \times 10^{-5}$), and the subject-relative holdout has 27/30 ($p_{\mathrm{Holm}} = 1.26 \times 10^{-5}$). PP has 18/29 and does not reject after correction ($p_{\mathrm{Holm}} = 0.132$; Figure 5c). Both RC criteria pass and PP does not. Because the protocol did not freeze a direct between-template contrast, this pattern does not establish that PP differs significantly from either RC condition.

D.3 WHY THE TRAINING RESULT AND TRANSFER TIMING DIFFER

The two studies answer different timing questions. In the training study, output-effect crossings concentrate at step 1000 while behavioral crossings concentrate at step 2000. In Transfer, matched onset still occurs by step 1000 for every target; what varies is whether the behavioral criterion has already crossed. The right-censored composite decreases across 70M/160M/410M (20/21, 13/21, 11/21), whereas matched onset remains 21/21 within each size. This pattern is compatible with a shorter observed interval before behavioral crossing, not loss of transfer. Because the detectors are not equal-power calibrated, it does not identify an earlier latent representation.

E SCORE THRESHOLDS, LAYER COVERAGE, AND ONSET SENSITIVITY

We recompute Transfer from its original per-orientation scores over a complete post-hoc grid of 528 settings. This is a descriptive sensitivity analysis, not another held-out experiment. Natural-effect thresholds are $\log r_N$, $r_N \in \{1, 1.05, 1.1, 1.2\}$; adjusted sham-contrast thresholds are $\log r_R$,

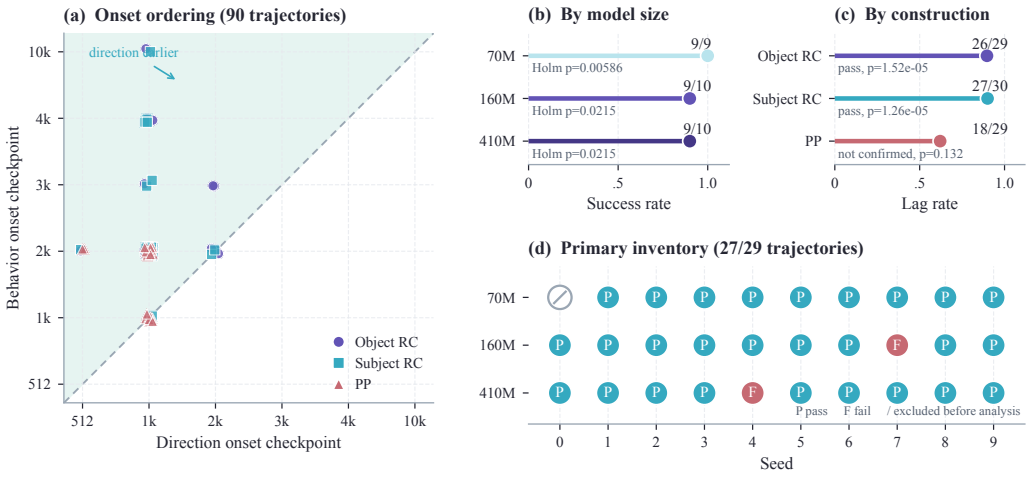

Figure 5: Operational criterion ordering is widespread on the sampled Pythia grid and bounded by template family. (a) Onset checkpoints for all 90 size–seed–template trajectories; points above the diagonal have an earlier output-effect crossing. (b) Exact frozen success counts by size; displayed p-values are Holm adjusted across the three sizes. (c) Template-specific exact tests, Holm adjusted across three templates; PP is not confirmed. (d) Exact inventory for the 29-trajectory primary test. P denotes success, F failure, and the slashed 70M-seed0 unit is the pre-evaluation exclusion. The detectors are not equal-power calibrated, so the figure does not establish latent-time ordering.

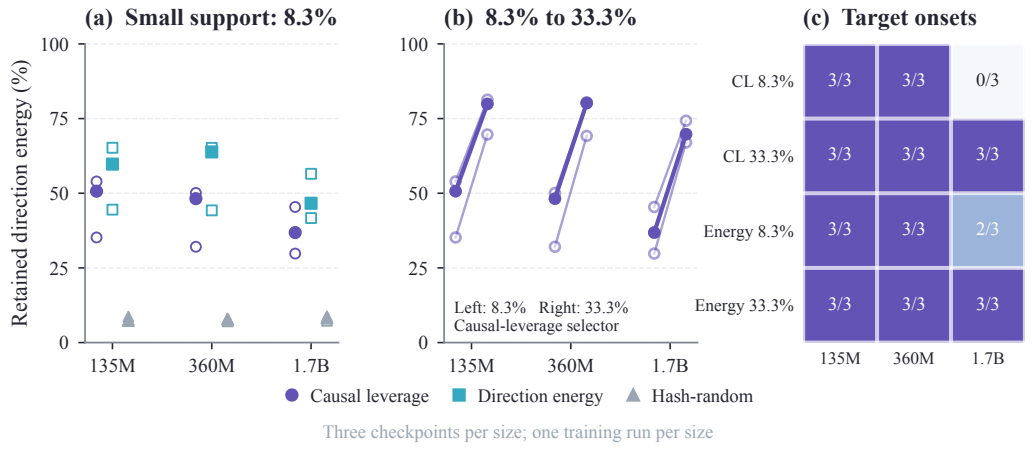

Figure 6: CPU-only SmolLM2 support analysis; no new model evaluations. (a) At 8.3% support, open symbols denote three checkpoints per size and filled symbols their medians. Both the implemented batch-weighted causal-leverage and direction-energy selectors retain less energy at 1.7B; random support stays near its nominal fraction. (b) Raising support to 33.3% increases retained energy at every size, by a median 32.9 percentage points at 1.7B. Thin lines show checkpoint pairs. The plotted $k = 64/256$ are *reference* widths: actual supports are $48/192$, $80/320$, and $171/683$ at 135M, 360M, and 1.7B, respectively. (c) Functional recovery accompanies this shift, but the energy selector's similar performance preserves the negative specificity conclusion. This post-hoc analysis is not a general scaling-law estimate.

$r_R \in \{1, 1.2, 1.5, 2\}$. Positive-layer thresholds are $\log r_L$, $r_L \in \{1, 1.1, 1.2, 1.5\}$, with a minimum of 3, 4, 5, or 6 positive layers. We either retain the original negative-layer floor $-\log(1.2)$ or disable that floor. These choices yield 512 combinations; 16 additional combinations omit the layer gate

Pair	Subject–verb source prompts	Reflexive target prompts	Scored continuations
PP	"The customers across from the customers" ↔ "The secretary across from the customers"	"The clerks near the minister hurt" ↔ "The teacher near the minister hurt"	SV: is/are, has/have, was/were; Refl.: himself/themselves
Object RC	"The taxi drivers that the customers hurt" ↔ "The executive that the customers hurt"	"The minister that the secretaries liked suspected" ↔ "The farmers that the secretaries liked suspected"	SV: is/are, has/have, was/were; Refl.: himself/themselves
Subject RC	"The actors that injured the architects" ↔ "The executive that injured the architects"	"The consultants that discussed the doctors suspected" ↔ "The athlete that discussed the doctors suspected"	SV: is/are, has/have, was/were; Refl.: himself/themselves

Table 7: Illustrative rows drawn exactly from the fixed test schedules. Each arrow joins two number orientations in a lexical cluster. The source tasks span three auxiliary families, whereas the reflexive targets use a disjoint output vocabulary. This blocks literal output-token reuse but does not distinguish syntax-specific structure from generic number morphology.

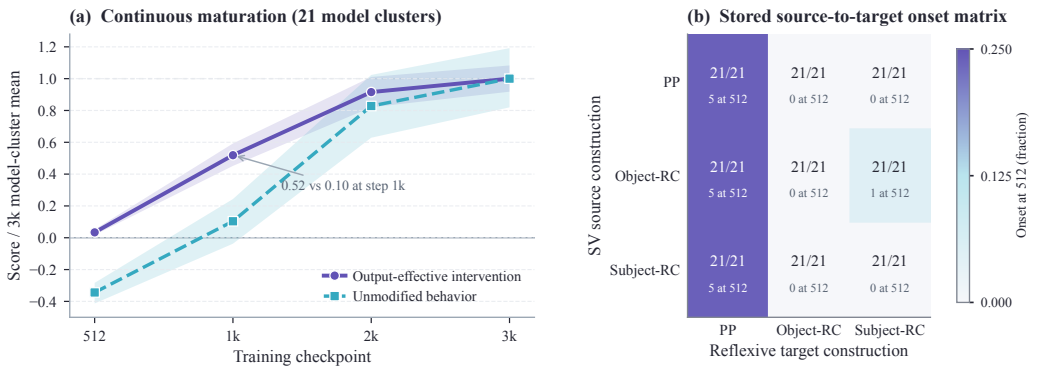

Figure 7: Post-hoc summaries over 21 size–seed models. (a) Scores relative to the observed step-3000 model mean: intervention uses the per-target median of 50 adjusted cluster scores; behavior uses the median of the weaker natural margin across two orientations. Shading is pointwise 95% percentile uncertainty from 100,000 model resamples, conditional on fixed observed late denominators, not full ratio or simultaneous-curve uncertainty. (b) The diagnostic fixed-subject matrix: all nine combinations have adjacent onset in 21/21 models; color is the fraction beginning at 512 rather than 1000. This matrix is separate from the mixed-site primary design. Neither panel calibrates detector power or establishes mediation.

entirely. Four-checkpoint Holm correction at 0.05 and adjacent-checkpoint persistence are unchanged. All combinations are reported; none replaces the original detector.

Table 8: Complete descriptive sensitivity grid. Counts are target trajectories nested three per model, not independent replications. Ranges span settings, not confidence intervals.

Layer requirement	Settings	Matched onset	Orthogonal onset
Disabled	16	63/63	0/63
At least 3 of 6	128	63/63	0/63
At least 4 of 6	128	63/63	0/63
At least 5 of 6	128	56–63/63	0/63
All 6	128	0–63/63	0/63

At the original positive-layer threshold $\log(1.2)$, changing only the required layer count from four to five reduces coverage to 59/63; requiring all six gives 0/63. Across the entire grid, matched counts are 63 in 336 settings, 61 in 32, 59 in 32, 56 in 32, and zero in 96. These frequencies do not estimate a probability of robustness. They distinguish persistent output effects under partial depth coverage from the stronger requirement of positive effects at every sampled layer.

Timing also changes. Holding other original choices fixed, reducing the sham-contrast ratio from 1.5 to 1 or 1.2 moves the number of step-512 onsets from 3 to 12. Increasing it to 2 gives 0 step-512, 60 step-1000, and 3 step-2000 onsets. Thus eventual coverage at the original layer gate is not confined to the original score thresholds, whereas exact onset remains detector-dependent.

The CPU implementation reproduces all 252 original cells and 504 original matched/orthogonal gates. A separately implemented reduction, binomial-tail calculation, and Holm adjustment agrees on all 66,528 trajectory-onset decisions. The companion contains the configuration, full per-setting counts, per-trajectory onsets, and checker. This computation neither matches actual edit norm nor calibrates behavioral and intervention detectors to equal power.

F EQUAL-EDIT-NORM COMPARISON ON THE ORIGINAL REFLEXIVE MATERIALS

Purpose and fixed design. This comparison asks whether the source direction retains an advantage when alternative directions receive the same actual editing magnitude on the original reflexive tasks. The design uses Pythia-70M and Pythia-410M, repository seeds 3 and 4, at training steps 2,000 and 3,000: eight fixed snapshots of four training trajectories, not eight independent models. Each snapshot is evaluated on the PP, object-relative-clause, and subject-relative-clause targets, using the original 100 rows (50 paired clusters) per construction. The six layers are 0–5 for 70M and 3, 7, 11, 15, 19, 23 for 410M (zero-indexed). For each source task, snapshot, and layer, the direction and writing region are inherited from the original source-development selection. We write at the last token of that region's span and read logits at the final prefix token. No target outcome selects a layer, region, direction, or sign.

Directions and matched magnitude. Let h and h_s denote the base and original counterfactual-donor residuals at the selected position. We first apply the original projection replacement, using its archived direction a_0 and float32 arithmetic, and obtain h_{nat}. Its *executed* edit norm,

$$\rho = \|h_{\mathrm{nat}} - h\|_2,$$

is the reference magnitude for that row and layer; it is not a global dose. For each unit direction v and each predeclared polarity $p \in \{+1, -1\}$, the requested displacement is

$$\delta_{v,p} = p\ \mathrm{sgn}(q)\,\rho v, \qquad q = (h_s - h)^\top a_0, \qquad \mathrm{sgn}(0) = 1.$$

The eight directions are the normalized source sentence direction, a lexical direction, three fixed random directions, and three fixed directions orthogonal to the sentence direction. Random vectors use a fixed hashed Rademacher construction; orthogonal vectors are projected away from the sentence direction, not selected by their responses. The lexical direction averages singular-minus-plural residual differences for the 19 noun families in the matched source-training set, using the standalone prefixes `The {noun}` and the final noun token at the same layer. Its fit assigns equal weight to each noun family.

Float32 addition can change the realized magnitude. We therefore adjust only a scalar multiplying $\delta_{v,p}$, with at most eight deterministic corrections, until the actual norm differs from ρ by no more than $\max(2 \times 10^{-5}, 10^{-5}\rho)$. This correction never uses logits. Zero-norm examples remain exact no-ops. Both polarities are retained separately; the negative arm reverses the displacement rather than replacing a_0 by $-a_0$ inside a projector. The positive sentence-direction arm must agree with native replacement within 2×10^{-4} in maximum absolute full-vocabulary logit difference. Source-only numerical and resource checks precede target execution; a failed numerical check is retained and blocks interpretation of the affected cell.

Readout and descriptive comparisons. The primary response is the change in the counterfactual-versus-base reflexive logit contrast relative to the common unedited baseline. Within each polarity, we pair the sentence direction with the lexical direction and with the mean of each three-direction control ensemble. The two number orientations and six layers are averaged equally within a cluster; the 24 snapshot–target cells are then summarized with equal cell weight. Neither layers nor control directions count as independent model replications. Complete vocabulary logits and executed edits are retained, allowing numerical checks and descriptive KL, entropy, and natural-label probability diagnostics.

The original native matched/orthogonal and same-class-sham responses are kept separately: their arbitrary-label and sham corrections are not subtracted from the equal-norm responses, because those native controls have their own editing magnitudes.

Material dependence and scope. These are previously exposed original materials, not a new lexical holdout. An outcome-blind inventory finds 19 noun and 12 verb families across the 150 target clusters. Linking clusters that share a noun family yields a single connected component, even when linkage is restricted to subject nouns or requires the same grammatical role. Consequently, the planned shared-family block bootstrap has only one resampling block and cannot provide a non-degenerate confidence interval. We restrict this comparison to finite-grid descriptive summaries, without confidence intervals or an invented single-family random-effect grouping. This conservatively treats shared vocabulary as a linkage rule; it does not assert that every pair of examples has the same statistical dependence. The separate new-lexicon cohort is not ready and is not included. This bounded comparison neither repeats the full 63-target trajectory inventory nor establishes its onset timing under equal-edit-norm controls.

Results on the fixed grid. All 144 layer cells completed and the independent CPU verifier reconstructed 288,000 intervention records, checked the executed edit norms and native full-logit anchors, and verified the stored artifact hashes. Table 9 reports the six fixed primary contrasts. Under the native-aligned polarity, the sentence direction exceeds the random and orthogonal ensembles in every snapshot–target cell. The mean advantages are 2.2503 and 2.2537 nats, respectively. By contrast, its mean difference from the lexical direction is -0.0048 nats, with both positive and negative cell differences. This is not an equivalence test, but it supplies no uniform sentence-direction advantage over the simpler lexical estimator. Reversing the displacement reverses the aggregate advantage over random and orthogonal controls; both polarities are retained rather than selecting the favorable one.

Comparator	Polarity	Mean difference	Cell range	Positive cells
Lexical	Native-aligned	-0.0048	$[-0.1314, 0.1055]$	15/24
Random ensemble	Native-aligned	2.2503	$[0.9911, 3.3241]$	24/24
Orthogonal ensemble	Native-aligned	2.2537	$[0.9798, 3.3154]$	24/24
Lexical	Reversed	0.0154	$[-0.0513, 0.0903]$	17/24
Random ensemble	Reversed	-0.5137	$[-0.9910, -0.0382]$	0/24
Orthogonal ensemble	Reversed	-0.5140	$[-1.0917, -0.0270]$	0/24

Table 9: Equal-actual-L2 sentence-minus-comparator target-logit changes (nats). Means weight the 24 fixed snapshot–target cells equally; each cell averages 50 paired clusters and six layers. Ranges show variation across cells, not confidence intervals. The cells share four training trajectories and exposed lexical materials.

The bounded result narrows the dose explanation: editing magnitude alone does not explain the contrast with these random controls on these tasks. It does not distinguish syntactic abstraction from reusable lexical number information. Neither this comparison nor its numerical verifier retroactively matches the original onset grid, validates the unavailable new-lexicon arm, or establishes generalization to independent materials.

G UNIFORM-GRADIENT CORRECTION SENSITIVITY

Correction and source-only check. We separately evaluate uniform per-example weighting on one SmolLM2-1.7B trajectory. With $N = 100$ source-development examples and class-1-minus-class-0 margin m_r, the corrected gradient is

$$g_{\text{uniform}} = \frac{1}{N}\sum_{r=1}^{N} \nabla_{h_r} m_r, \qquad S_{\text{uniform}}(k) = \text{topk}_i\, |a_i g_{\text{uniform},i}|.$$

The archived hook returns $\widetilde{g}_b = n_b^{-2}\sum_{r\in b}\nabla_{h_r} m_r$ for batch size n_b. The correction aggregates $N^{-1}\sum_b n_b^2 \widetilde{g}_b$, rather than $N^{-1}\sum_b n_b \widetilde{g}_b$. It uses the actual final batch size of four. Model and hook computation remain float32 with eager attention; the first multiplication by n_b is float32, followed by

float64 accumulation and a final float32 gradient. Ties use ascending coordinate index. Before target evaluation, batches of 16 and batches of one must produce exactly the same selected support sets at both budgets for every source task and layer. All three checkpoints passed this source-only gate. This checks support membership, not bitwise equality of gradients or ordering within sets.

Fixed scope. The checkpoints are steps 125,000, 2,500,000, and 5,125,000, with layers $(3, 7, 11, 15, 19, 23)$ and PP, object-RC, and subject-RC targets. Directions use 400 source-training rows; writing-site selection uses 100 development rows. Reference budgets 64 and 256 select 171 and 683 of 2,048 coordinates. Each target retains 100 orientation rows in 50 paired clusters. At each budget, coordinate interchange compares uniform-gradient, historical batch-weighted, magnitude, and fixed hash-random selectors. Uniform-gradient and full-coordinate/full sparse rank-one interventions retain same-class shams. These are native interventions, not another equal-edit-norm comparison. All 75,600 target records completed; three numerical verifiers reconstructed support indices, score reductions and stored component tests. The collector retains 126 intervention cells and 54 paired comparisons. A separate packaged CPU entry reaggregates supplied scores without model inference.

Results. Uniform weighting changes the selected set in all 108 source–layer–budget comparisons against historical weighting. Nevertheless, the layer-gate boolean is unchanged in all 18 checkpoint–target–budget comparisons. Table 10 reports continuous natural-label contrasts and nominal checkpoint eligibility, not persistent onset. Across the 18 fixed conditions, uniform-minus-historical mean natural effects range from -0.0543 to 0.0412 nats (seven positive, eleven negative); this is not an equivalence test. Uniform-minus-magnitude effects are negative in all 18, ranging from -0.3849 to -0.0540 nats. Correcting the weighting therefore does not establish an advantage over magnitude ranking on this grid.

Checkpoint	Coordinates	ΔN: U−H	ΔN: U−E	Eligible H/U/E
125,000	171	0.0041	−0.0837	0/0/0
125,000	683	0.0006	−0.3291	3/3/3
2,500,000	171	−0.0023	−0.1853	0/0/2
2,500,000	683	−0.0144	−0.1988	3/3/3
5,125,000	171	0.0130	−0.1896	3/3/3
5,125,000	683	−0.0011	−0.0781	3/3/3

Table 10: Corrected-selector sensitivity on exposed materials. U: uniform gradient; H: historical weighting; E: magnitude ranking. Differences are nats, equally averaged over the three targets after paired-cluster reduction. Eligibility counts targets out of three with $\max(p_N, p_S) \leq 0.05$ and the inherited layer gate, using thresholds $\log 1.1$ for N and $\log 1.5$ for S. These nominal component decisions omit Holm adjustment and adjacent-checkpoint persistence and must not be read as onset counts.

Interpretation boundary. These dependent checkpoints and targets belong to one model-size trajectory and historically exposed materials. They neither revise the preserved cross-scale results nor test corrected selectors at 135M or 360M. The nominal gate comparison does not rerun the full original M9 criterion. The evidence removes the global "uniform weighting untested" limitation only for this bounded follow-up; it does not establish a general selector ranking.

H SUPPLEMENTARY SENSITIVITY DETAILS

The primary transfer statistic treats each size–seed–target trajectory as its target-specific unit, although three targets share each model. The model-level sensitivity therefore requires matched onset for all three targets and timing success for at least two of three. Matched onset remains 21/21 models and orthogonal onset occurs in 0/21. Right-censored timing holds in 15/21 models (uncorrected exact one-sided $p = 0.0392$), whereas strict timing holds in 12/21 ($p = 0.332$). This coarser dependence unit is consistent with the primary result while reinforcing that strict lag alone is not the positive temporal result.

The continuous analysis keeps targets clustered within the same 21 models. Each target contributes the median of 50 matched same-class-sham-adjusted cluster scores and the median of the weaker

natural-logit margin across its two orientations; each measure is divided by its 21-model mean at step 3000. At step 1000, intervention reaches 0.519 of its late reference (model-cluster bootstrap 95% interval $[0.452, 0.592]$), whereas behavior reaches 0.104 ($[-0.037, 0.243]$). At step 2000, the values are 0.916 ($[0.820, 1.011]$) and 0.828 ($[0.629, 1.023]$). These pointwise intervals condition on the observed late-reference denominator, which is not recomputed within resamples; they are not ratio-uncertainty or simultaneous bands. The matched-minus-orthogonal effect is positive in 19/21 model clusters at step 512 and 21/21 thereafter. This post-hoc sensitivity does not calibrate the two detectors to equal power.

Finally, every completed condition stores a diagnostic subject-region direction from each source construction. Applying the same Holm, layer, and adjacent-persistence rule after the primary analysis yields onset in 21/21 models for all nine source–target combinations. This matrix was not a primary criterion. It is compatible with construction-general output effectiveness on the synthetic tasks, but does not distinguish generic number morphology from syntax-specific structure.

Dependence and leave-one-group sensitivities. The frozen Strong analysis counts 29 size–seed trajectories. A stricter same-numeric-seed block that requires every available size to pass yields 8/10 blocks ($p = 0.0547$), while majority-size blocking yields 10/10 ($p = 9.77 \times 10^{-4}$). Omitting any one seed leaves 24/26–25/27 primary successes. Template omission is less stable: omitting PP preserves 24/29 ($p = 2.73 \times 10^{-4}$), whereas omitting either RC family does not reject. This pattern indicates that the positive aggregation depends on the two RC templates rather than three interchangeable replications. For M4, requiring both sizes at each numeric seed to show selected-over-random gains gives 7/7 blocks ($p = 0.00781$). Row and cluster identifiers are unique within each schedule and disjoint across tasks, but hashed identifiers cannot establish semantic lexical independence. All exact tests therefore concern the finite frozen schedules unless a model- or seed-blocked sensitivity is stated.

I Statistical decision rules

Auditable pipeline. Appendix A gives the complete reduction, strict versus inclusive boundaries, behavior event, Holm ordering, and adjacent-pair endpoints. Source development selects on $S = N - A$, while the held-out test uses sham-adjusted R. The six layers and two orientations are averaged into each checkpoint's 50 cluster observations before component tests; there are no six layerwise significance tests. The layer-median gate is separate. A selected support ranks absolute training mean differences with coordinate-index tie breaks; hash-random support ranks frozen SHA-256 digests.

SmolLM2 support selectors. For each size–checkpoint, source task, and selected layer/region, the direction a is fit on 400 source-training rows. The implemented gradient g of the class-1-minus-class-0 source margin uses 100 source-development rows with batch-dependent weighting, not a uniform example mean (Appendix B). Causal-leverage and energy supports are then

$$S_{\mathrm{CL}}(k) = \mathrm{topk}_i\, |a_i g_i|, \qquad S_{\mathrm{E}}(k) = \mathrm{topk}_i\, |a_i|, \tag{11}$$

with coordinate-index tie breaks. Stored inputs are float32; ranking multiplies their values after conversion to Python floats. The random selector orders coordinates by SHA-256 of the frozen namespace, group, target, layer, width, and index. Supports are canonicalized before any held-out target row is loaded; target-test forward count at selection is zero. Reference widths $(64, 256)$ scale to $(48, 192)$ at 135M, $(80, 320)$ at 360M, and $(171, 683)$ at 1.7B. Exact held-out target measurements, not gradient or energy scores, determine onset. This historical comparison remains distinct from the corrected 1.7B follow-up in Appendix G.

Sequential error control. Conditional on valid component p-values, any detected onset under a global no-effect null requires a false rejection in its ten-checkpoint Holm family. The per-detector probability of any onset is then bounded by 0.05; the layer gate and adjacent-pair requirement can only reduce that bound. The exact binomial components require independent cluster indicators with null success probabilities at most one half. Holm permits dependence between checkpoints, but does not repair dependence that invalidates a component test. Lexical uniqueness alone does not establish these within-checkpoint assumptions for our fixed synthetic schedules. We therefore distinguish this conditional analytic bound from the descriptive onset inventory; neither establishes equal detector power or calibrates their first-crossing difference. Strict lag is not interpreted as latent-time separation.

Frozen chronology and protocol dictionary. Strong is the Pythia timing study; Transfer is the original protected cross-construction confirmation (distinct from the later R4 estimator comparison); M4 is the Pythia coordinate-dose study; M6 is the fresh-holdout removal study; and M9 is the SmolLM2 cross-scale support test. The support-energy, model-clustering, and dependence-blocking analyses are explicitly post-hoc CPU-only sensitivities. Exact repository revisions are pinned in the supplemental model registries; the main text does not abbreviate them as model names alone. Full SHA-256 values are in the hash DAG; protocol prefixes are Transfer `6dd9dbcefbee`, M4 `e86535f469cf`, M6 `26525046cb7`, and authoritative final M9 `9f4739395f98`. Table 12 lists invalid or unresolved studies. No post-hoc analysis replaces a primary decision.

Null map. The Strong pooled, size, and template tests are exact one-sided binomial tests of a trajectory success probability at most 0.5, with onset ties counted as failures and Holm correction within each frozen size or template family. Transfer onset coverage is a multicomponent design gate rather than a population-frequency test; the matched-minus-orthogonal comparison is an exact paired sign test of a nonpositive median advantage. The historical right-censored composite assumes independent Bernoulli trials with success probability at most 0.5; its dependent target inventory is not calibrated current inference or a test of equal onset-time distributions. Coordinate-dose and erasure target counts are exact inventories; the post-hoc inferential contrasts use 14 size–seed models and an all-three-target cluster criterion. The OLMo-2 and SmolLM2 outcomes are exact gate inventories over the tested runs and targets, not population-level family or scaling estimates.

The strong study defines output-effective onset and behavior onset as the first adjacent pair of Holm-eligible checkpoints. Output-effect eligibility requires both the checkpoint family and the six-layer criterion. A complete size–seed trajectory succeeds when strict lag holds for at least two of the three templates. The primary exact sign test contains 29 previously unseen complete trajectories; size tests contain 9, 10, and 10 trajectories.

The transfer study requires matched onset, matched-over-orthogonal advantage, within-target measurability, size and target coverage, and a maximum of five orthogonal-control onsets. Right-censored lag is reported separately from strict lag. It counts a matched onset before detected behavior or a matched onset when behavior remains undetected through step 3000.

J Evidence and claim boundaries

The initial broad result is a scientific negative because it did not separate from the same-class sham. The first follow-up is invalid and contributes no scientific evidence. The 900-cell timing and 252-condition transfer studies preserve the original operational and transfer inventories. The coordinate-dose study separately establishes the Pythia support boundary; the OLMo-2 and fresh-holdout studies bound cross-family, encoding, and selectivity claims while permitting exact component outcomes to be reported. The SmolLM2 study establishes a negative result for largest-scale reference-k=64 generalization under the implemented batch-weighted selector: dense effects and exact support counts remain reportable, but this selector fails the compact-support criterion. The separate uniform-gradient sensitivity in Appendix G covers one 1.7B trajectory and does not replace this historical cross-scale criterion. The estimator comparison remains unresolved and does not support a ranking. Its later R5 CPU numerical validation cannot retroactively supply the missing original R4 verifier receipt or promote that comparison to verified confirmation.

K Figure source data

The overview is an editable schematic, not a measured plot. Quantitative figures use the accompanying CSV source data. Figure 5 uses the six size/template tests and all 90 trajectory onset reports. Figure 2 uses all 63 held-out trajectory reports. Figure 3 uses the coordinate-dose, OLMo-2, and fresh-holdout summaries. Figure 7 uses all 252 transfer conditions. Checksums identify plotted results and derived source tables.

Claim and unit	Success criterion	Contradicting outcome	Observed outcome	Scope boundary
Strong temporal onset; 29 unseen size–seed trajectories	Strict lag on $\geq 2/3$ templates, followed by the primary exact test	Insufficient trajectory successes, a failed scale family, or failed layer/control criteria	27/29, $p = 8.12 \times 10^{-7}$; all three scales pass	Same-seed all-size block: 8/10, $p = 0.0547$; two examined trajectories excluded
Template family; one trajectory per size–seed–template	Holm-corrected exact test within the three-template family	A template fails its corrected criterion	Object RC 26/29; subject RC 27/30; PP 18/29	Pass/fail pattern is not a direct between-template test
Held-out transfer; 63 size–seed–target trajectories	Matched and within-target onset, matched advantage, size/target coverage, and orthogonal ceiling	Any required component or coverage condition fails	Matched 63/63; within-target 63/63; advantage 63/63; orthogonal 0/63	Three targets are nested in each of 21 models
Transfer timing; 63 targets nested in 21 models	Horizon-specific exact test of the right-censored composite	Too few strict-or-undetected-behavior cases	31 strict + 13 behavior-undetected $= 44/63$; model sensitivity 15/21	Strict lag is 31/63 and 12/21; unequal detector power and censoring preclude latent-time ordering
Coordinate dose; 42 targets nested in 14 models	Selected 8.3% and full anchors succeed; matched random 8.3% and shams remain zero; both sizes/targets covered	Any required anchor/control/coverage component fails	Target inventory 42/42 vs 0/42; model analysis 14/14 vs 0/14; seed block 7/7	Transition lies in $(2.1\%, 8.3\%]$ on the tested grid
Cross-family replication; one OLMo-2 trajectory, three targets	Pythia-style selected/full/sparse anchors succeed for every target	Any primary anchor absent	No primary anchor succeeds	One trajectory; scoped non-replication
Fresh holdout; 42 targets nested in 14 Pythia models plus OLMo-2	Joint encoding, selectivity, controls, and cross-family criteria	Any joint criterion fails	Full erase 14/14 models; selected 12/14; random/encoding 0/14; joint criterion unmet	Operator response only; off-target degradation unmeasured
SmolLM2 budget boundary; one run per size, three targets	Dense anchors succeed; approximately 8.3% causal support remains portable and specific at 1.7B	Largest-size portability or any selector-specificity criterion fails	Dense 3/3 sizes; causal 8.3% support 2/3 sizes; specificity 0/3	Dense target-logit effect survives; compact support is not scale-stable under this selector/basis

Table 11: Inferential units, success criteria, falsifiers, and outcomes. Counts are exact. Supplementary model-cluster sensitivities leave the primary decisions unchanged; component results are not converted into unsupported joint claims.

L Development diagnostics: number directions, dose, and sham donors

Two development panels use one Pythia-410M checkpoint (step 143000), block index 11, float32 inference, and constructed English prefixes. They neither reconstruct the original confirmation's missing full-vocabulary outputs nor add independent model replications. All rows are development data; the second panel was designed after examining the first. We report all directions descriptively, without superiority tests or replacing any frozen decision.

Study	Why it is excluded	Outcomes examined?	Relation to reported evidence	Resolution
Initial broad study	Did not separate from the same-class sham	Yes	Scientific negative; not reused as positive evidence	Narrower hypothesis with examined trajectories excluded
First agreement follow-up	Invalid cached-file reference	No valid scientific result	Cannot support or invalidate the later positive result	Rebuilt independently with explicit path validation
Estimator pilot	Balanced logistic probe did not converge	Pilot estimators only	No estimator ranking enters the paper	Retained as a failed pilot
First full estimator comparison	Quantized orthogonal control was not unit length	Partial outputs only	Separate from the 900-cell timing and 252-condition transfer studies	Excluded from scientific comparisons
Second full estimator comparison	Independent linear refit disagreed with the original computation	All 84 groups	CPU reanalysis checks the stored numerics but cannot establish a ranking	Estimator comparison remains unresolved
Fresh-holdout numerical check	Historical JSON and BLAS-dependent scalar differences	Scientific outcomes unchanged	6,120 refits were repeated under one deterministic CPU setting	Component counts retained; joint claim remains unsupported

Table 12: Excluded and unresolved studies. These outcomes remain part of the study history but do not enter positive scientific statistics. Separate analysis scripts rederive the reported raw-logit reductions and onset decisions.

Source-only directions and dose. We fit normalized mean-difference directions on 16 singular/plural noun pairs in *The [noun] near the gate* (DiM) and *Vocabulary: [noun]* (lexical DiM). The vocabulary frame removes the agreement-sentence context but retains noun-number morphology and distributional noun-class differences; it is neither a pure inflection direction nor a control removing all lexical semantics. We also fit a centered ridge direction ($\lambda = 1$), one rank-one DAS direction with 80 Adam updates (learning rate 0.02, batch size 16; final iterate), a balanced shuffled-label direction, and three fixed random directions. Eight disjoint source-development noun pairs supply positive checks; target outputs do not select the layer, direction, or dose. The DAS backbone is frozen. These fixed implementations do not constitute an exhaustive tuned baseline comparison.

We use 12 other lexical pairs, three PP/RC structures, and two role assignments, giving 72 prefixes per continuation type. Each pair exchanges controller and attractor words while preserving their multiset. We intervene at the final token of either noun and record the full output vocabulary for *blamed* and for the separate continuation prefix *walked across the*. The latter uses the prescribed next token *street* as a narrow collateral probe. Native rank-one interchange is compared with signed edits at 0.25, 1, and 4 times a source-derived reference norm of 6.936. Equal multiplier means equal actual L_2 in this panel; the native operator generally has a different norm. The first panel contains 9,216 edited rows across 384 cells, plus 288 zero-edit comparisons.

What the first panel shows. At reference dose, DiM and rank-one DAS yield similar mean target gains, while the lexical direction retains much of the gain with less full-vocabulary change (Table 13). Target probability mass changes are reported separately in the source data: -0.001108 for DiM, -0.001084 for lexical DiM, -0.000923 for ridge, and -0.000972 for DAS. Conditional KL alone would miss this dimension of the change. For each lexical pair, we average the three structures and two assignments before contrasting controller-site with attractor-site gain, then subtract the lexical direction's same contrast. DiM's mean interaction is 0.0409 at reference dose (9/12 positive lexical clusters), but -0.5653 at fourfold dose (0/12 positive). Native interchange gives 0.7835 (12/12

Direction	Target gain	Full KL	Conditional KL	Collateral ΔNLL
DiM	0.512	0.01133	0.01112	−0.017
Lexical DiM	0.464	0.00735	0.00712	−0.002
Ridge	0.472	0.01152	0.01137	−0.028
Rank-one DAS	0.511	0.01363	0.01345	−0.013
Shuffled labels	0.022	0.00225	0.00223	−0.027
Random 0	−0.027	0.00159	0.00158	−0.022
Random 1	−0.046	0.00139	0.00138	−0.004
Random 2	0.022	0.00141	0.00142	−0.023

Table 13: Single-checkpoint development means at the controller site and equal actual $L_2 = 6.936$. Target gain is the log-odds change toward the diagnostic counterfactual number, not natural accuracy. KL is $D_{\mathrm{KL}}(p\|q)$ in nats; conditional KL excludes the two reflexive tokens and renormalizes. Collateral ΔNLL is measured on separate prefixes; positive values indicate harm to the prescribed token. These are not equal-gain comparisons or evidence of general generation quality.

positive), but is not norm matched across directions. Thus neither a favorable native contrast nor one dose supports an invariant structural advantage.

A causal-position limitation. At the source noun position, a causal transformer cannot use the future phrase *near the gate*. Consequently, a source noun direction extracted in this sentence frame is not by itself evidence that the complete agreement dependency has been represented. Role exchange also changes serial position in the first panel. These observations motivate the separate role/order and real-donor controls below; they are not a proof that the model lacks syntactic representations elsewhere.

Real same-class donors and role/order crossing. The second panel reuses all eight frozen directions and the 12 exposed lexical pairs without fitting or selecting parameters. It crosses all four controller/attractor number assignments with *The [controller] near the [attractor] blamed* and *Near the [attractor], the [controller] blamed*. There are 96 prefixes per continuation type. An opposite-class donor changes only the edited noun's number within the same lexical pair and order; a same-class donor uses the next lexical pair cyclically, preserving both number assignments and order. Both use native projection interchange. Their actual norms are recorded, not claimed to be matched. We evaluate 6,144 edits in 128 cells, with 384 separate zero-edit comparisons.

For this panel, signed gain follows the edited noun's hypothetical number flip. Thus a positive attractor-site gain denotes output sensitivity to the attractor, not correct reflexive behavior; a same-class donor's signed gain is drift, not a desired label change. Within each lexical pair we average the four number configurations, subtract same-class drift at each site, and contrast controller with attractor. Table 14 reports both orders rather than pooling away their difference.

	Postposed PP		Fronted PP	
Direction	Role contrast	Positive pairs	Role contrast	Positive pairs
DiM	−0.253	2/12	1.579	12/12
Lexical DiM	−0.149	2/12	1.042	12/12
Ridge	−0.195	2/12	1.506	12/12
Rank-one DAS	−0.250	2/12	1.596	12/12
Shuffled labels	0.0033	7/12	0.0018	5/12
Random 0	0.0021	7/12	−0.0007	5/12
Random 1	0.0004	7/12	0.0015	7/12
Random 2	0.0015	8/12	0.0022	6/12

Table 14: Development role contrasts: $(G_{\mathrm{controller,opposite}} - G_{\mathrm{controller,same}}) - (G_{\mathrm{attractor,opposite}} - G_{\mathrm{attractor,same}})$. Each row uses all 12 lexical pairs, not 12 models. Positive-pair counts have no inferential threshold. Native amplitudes differ by donor, direction, and site. The sign reversal does not isolate syntax from context, serial position, or dose.

Sensitivity follows order as well as role. For DiM, opposite-class mean gain is 0.842 at the controller and 1.099 at the attractor in the postposed PP, compared with 2.185 and 0.608 in the fronted PP. Same-class signed means are 0.0056/0.0097 and 0.0060/0.0084, respectively. Averaged over orders at the controller site, DiM's same-class absolute gain is 0.0558; native edit norms average 5.973 for opposite-class donors and 0.669 for same-class donors. A small signed sham mean therefore neither means zero drift nor establishes an equal-dose advantage.

In this native-dose panel, the four number-sensitive directions have negative role contrasts in the postposed frame and positive contrasts in the fronted frame. The more responsive site is the later noun in both frames, compatible with order/context-sensitive intervention leverage. Preceding context and native dose also change. The equal-dose successors in Appendix M address amplitude and refit the directions on Qwen; the latter does not reproduce the negative postposed role contrast. We therefore do not identify a universal later-noun rule, isolate serial position as the cause, or infer the absence of syntax.

Numerical audit and scope. Separate CPU formulas reconstruct every edit and recompute every row from saved hidden states, vectors, and full-vocabulary logits. Both audits pass: zero-edit differences are zero and metric errors below 3.2×10^{-6}. These are same-author numerical checks, not scientific reviews. GPU runner times are 136.66 and 92.72 seconds on one V100, with peak allocated memory below 1.71 GiB; the first includes fitting and evaluation, the second has no backward steps. Runner time excludes transfer, audit, and other rental costs.

Synthetic gender-neutral occupations with masculine reflexive candidates limit natural-language validity. The panels diagnose direction-, dose-, and order-dependent reflexive control, not general generation quality, method superiority, selective natural mediation, training improvement, or cross-model generalization.

M EQUAL-DOSE ROLE–ORDER DIAGNOSTICS ACROSS TWO MODELS

We first address the native-amplitude confound without reinterpreting the earlier panels as confirmation. The Pythia-410M successor crosses 12 new lexical pairs, both role assignments, both PP orders, and all four number configurations: 192 prefixes per continuation type. An explicit *male* modifier fixes the intended singular gender condition in these constructed reflexive prompts. All eight source directions from the first development panel are retained. Edits use 0.25, 1, and 4 times the source reference $L_2 = 6.936$, with float32 renormalization so that actual norm is matched across methods, sites, and orders. The 192 cells contain 18,432 edits and 768 noops.

For each lexical pair, define $R_o = G_{\text{controller},o} - G_{\text{attractor},o}$ and the order interaction $I = R_{\text{fronted}} - R_{\text{postposed}}$. The predeclared descriptive comparison is DiM at reference dose on the reflexive prefixes. Pythia gives $I = 1.634$, positive in all 12 clusters; the lexical direction gives 1.373. Equal dose thus retains an order interaction, but not an exclusively sentence-derived effect.

A prospective model replication with a narrower conclusion. Before running Qwen2.5-0.5B, we fixed the positive DiM interaction prediction and retained every method and dose. We refit all eight estimators on Qwen source activations, use midpoint block index 11 of 24, and apply Qwen's own source reference $L_2 = 5.960$. The source fit includes the fixed 80-step DAS procedure. The target inventory matches Pythia's; this is a prospective descriptive model replication on exposed templates, not a fresh confirmatory lexical sample.

Qwen yields $I = 0.954$ with 12/12 positive clusters (range 0.052–1.721); lexical DiM, ridge, and DAS give 0.787, 1.033, and 0.968, respectively. Qwen's DiM role contrast is positive in both orders: 1.007 postposed and 1.961 fronted. Pythia instead gives -0.088 and 1.546. The replicated interaction does not imply a universal preference for the later noun. Dose sensitivity also matters: Qwen DAS has $I = -0.803$ at fourfold dose, with only 2/12 positive clusters. Complete method–dose tables remain in the source data; no direction is selected by these outputs.

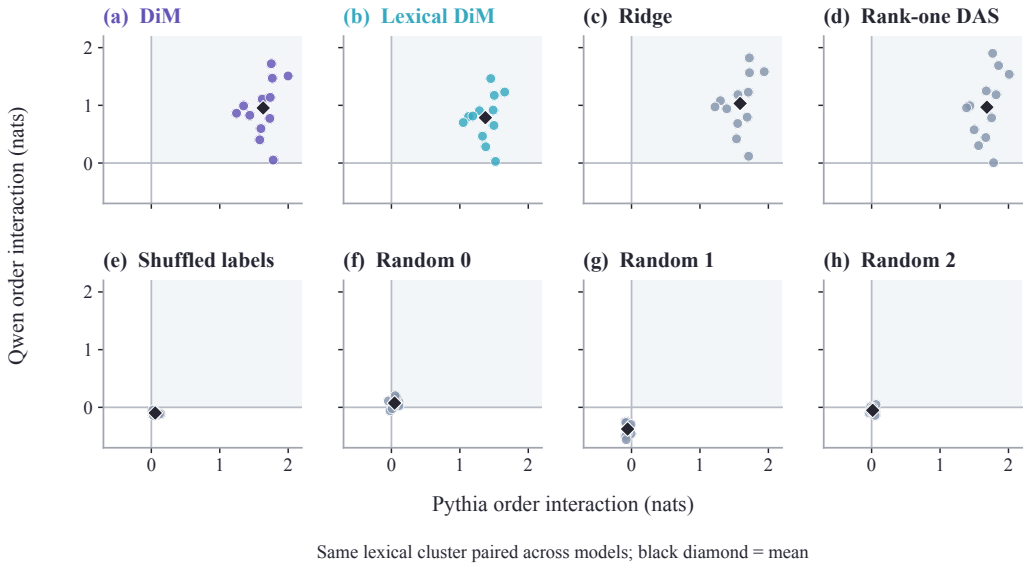

Figure 8: Paired reference-dose order interactions in the development Pythia diagnostic and its prospective Qwen replication. Each point pairs the same lexical cluster across models; each of eight facets retains all 12 clusters. Black diamonds mark the two model means, not uncertainty intervals. Shading marks positive interaction in both models. Models use their own source-derived reference norm and independently estimated directions. Shared stimuli are not two independent lexical samples.

N PYTHIA SUPPORT ORIENTATION AT MATCHED INTERVENTION ENERGY

The original coordinate study uses a different intervention setting and retains its original decision. Here we ask a narrower follow-up question on the same Pythia-410M checkpoint and crossed development prefixes: do selected and random supports differ when their *applied* vector lengths are equal? We restrict the frozen DiM to its 85 or 341 largest absolute coordinates out of 1,024, then normalize. Each width has three fixed uniform-without-replacement random supports. All nine directions, including dense DiM, receive the same actual $L_2 = 6.936$. No target output chooses the support. The study contains 72 cells, 6,912 edits, and 768 noops.

At 85 coordinates, selected controller gains are 0.606 postposed and 1.610 fronted, versus random-support ranges 0.183–0.326 and 0.440–0.799. At 341 coordinates, selected gains recover 0.766 and 1.942, close to dense gains 0.773 and 1.979; random supports recover 0.494–0.670 and 1.368–1.753. Retained unit-DiM norms are 0.624 and 0.896, equal to the normalized supports' cosines with dense DiM; their squares, 0.389 and 0.803, are retained direction energy. Post-normalization applied length is identical. Appendix P lists these distinct quantities for all nine supports alongside all measured external operating points.

This contrast rules out *applied length alone* as an explanation of the local selected-over-random result. It does not equate angular alignment to dense DiM, identify unique necessary coordinates, or establish robustness across checkpoints. Broader random recovery remains visible rather than being reduced to a binary failure label. All attractor-site, lexical-collateral, and full-vocabulary summaries are retained alongside the displayed controller comparison.

O A PROSPECTIVE EXTERNAL SUBSET: EFFICACY WITHOUT THE JOINT SELECTIVITY CRITERION

Data and pre-model audit. We use the author-released BLiMP relative-clause distractor-agreement data (Warstadt et al., 2020), pinned to commit `3e56b06`. Before any model forward, we restrict to an unambiguous common-noun head, a number-invariant RC auxiliary, and same-tense matrix

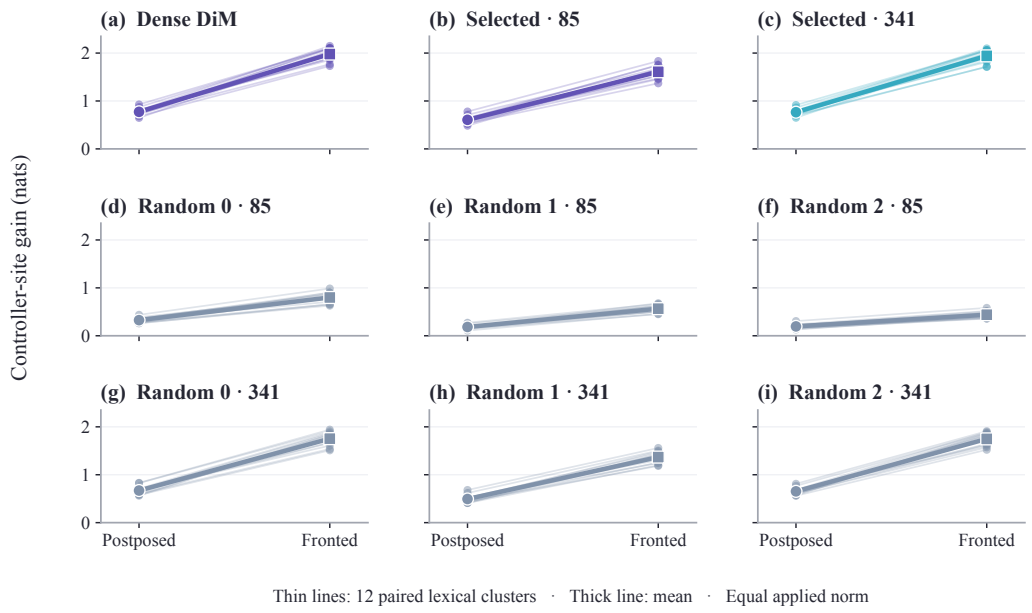

Figure 9: Paired-order controller-site gains at equal applied energy. Each thin line joins the same lexical cluster in postposed and fronted order; the thick line joins the means, with circle and square endpoints. All 12 clusters and all nine supports are retained, including three fixed random supports per width. Points are lexical-pair means, not independent support draws. This development diagnostic does not rerun the earlier support-success criterion.

auxiliary pairs: *has/have*, *does/do*, *is/are*, or *was/were*. We exclude overlap with the recorded source and prior development noun sets. Hash-ranked head lemmas define calibration and test splits; at most two hash-ranked sentences per lemma are selected. We inspected all 37 selected good sentences and bad auxiliary alternatives with AI assistance, excluded two problematic matrix-complement constructions, and then excluded three rows whose candidate or continuation was not one exact token. Every disposition precedes the experimental clock.

The final calibration split has 10 sentences and eight head lemmas; the test has 22 sentences and 16 disjoint head lemmas. Base matrix-auxiliary accuracy is 20/22, measured after freezing and never used for inclusion. For example, the prefix *This customer who had visited most children* contrasts *has/have*. Its prescribed lexical continuation is *worn*, scored after the original correct auxiliary. This is an AI-assisted text audit of generated data, not fresh human annotation, natural-corpus coverage, or a claim of unknown pretraining exposure. Semantic oddity remains, and the task still concerns grammatical number.

Fixed protocol and matched calibration opportunities. We retain the eight Pythia source directions, block, and reference norm from the earlier diagnostic. Each estimator receives the same 16 fixed dose candidates on the calibration split. We choose the dose whose equal-head-lemma mean counterfactual gain is closest to 0.5 nat, breaking ties toward smaller dose, and save all choices before any test forward. No interpolation or held-out recalibration is allowed. This matches *dose-validation opportunities*, not the training cost or complete hyperparameter search of DiM, ridge, and DAS. Every method is tested at both reference and selected doses.

We internally prespecify four one-sided endpoint slots at $\alpha = 0.0125$: reference DiM gain lower bound above 0.5; reference conditional non-target KL upper bound below 0.02; reference absolute lexical ΔNLL upper bound below $\log(1.05)$; and calibrated collateral improvement over fixed DAS, only if *both* methods actually meet 0.5 ± 0.05 in calibration and test. These are explicit study tolerances, not universal fluency or safety standards. A half-nat multiplies two-candidate odds by 1.649. Lexical drift measures one teacher-forced word, not full-sentence quality.

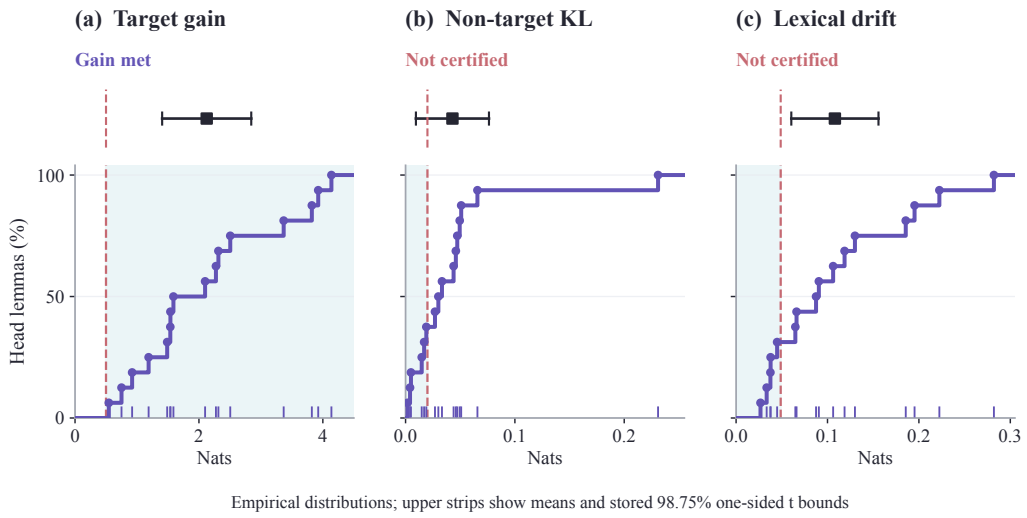

Figure 10: Empirical cumulative distributions of all 16 held-out head-lemma means for reference DiM on the restricted BLiMP subset; rugs retain every value. Upper strips share the horizontal scale: black squares are means and whiskers join the two one-sided 98.75% t bounds. Dashed lines mark pretest thresholds; shading marks acceptable values, not confidence regions. Only the lower gain and upper KL/drift bounds enter their registered slots. Gain is met; non-target criteria are not. Counterfactual control is not grammatical-accuracy improvement.

Results and the infeasible equal-benefit comparison. Mean reference gain is 2.125 nats, with lower bound 1.406; all 16 cluster means are positive. Conditional non-target KL is 0.04288 (upper bound 0.07629), and absolute lexical drift is 0.10809 (upper bound 0.15577). The three primary dispositions are gain met, low-KL certification not met, and low-drift certification not met. An upper bound failing to fall below a tolerance does not establish a population violation. Here both observed means exceed their tolerances; the lexical-drift lower bound, 0.06041, also exceeds its tolerance, but that opposite-tail comparison is exploratory rather than a primary slot.

Direction	Chosen dose	Calibration gain	Within 0.5 ± 0.05
DiM	0.375	0.676	No
Lexical DiM	0.375	0.520	Yes
Ridge	0.375	0.660	No
Rank-one DAS	0.250	0.533	Yes
Shuffled labels	4.000	0.521	Yes
Random 0	3.000	0.506	Yes
Random 1	4.000	0.544	Yes
Random 2	2.500	0.479	Yes

Table 15: All target-domain dose-calibration choices using source-fitted directions. "Yes" describes only calibration feasibility, not held-out success, low collateral, or method superiority. Dose is relative to the fixed Pythia source norm. Large random doses can reach the calibration gain, which makes target gain alone an insufficient quality measure.

DiM misses calibration tolerance even at its nearest grid point. On test, calibrated DiM and DAS gains are 0.767 and 0.648, both outside the required interval. Their mean absolute collateral difference, DAS minus DiM, is -0.01170, but unequal benefit invalidates the intended superiority interpretation. We retain the comparison as infeasible, rather than selecting a new dose from the test curve. The joint claim is not established. This is an informative limit on a selective-control claim, not a numerical failure and not a test of whether grammatical number can be manipulated at all.

Statistical and resource scope. All within-head observations are averaged before analysis, with equal weights across head lemmas. We use one-sided t bounds under an approximate independent-cluster sampling assumption and a fixed-seed 20,000-resample cluster bootstrap sensitivity. Bootstrap bounds agree with the registered dispositions: gain lower bound 1.526, KL upper bound 0.07739, and absolute-drift upper bound 0.15268. Before the test, a Pythia-development variance-envelope sensitivity calculation found that the non-target endpoints did not attain 80% power within its candidate grid of 12–192 clusters. The realized 16-cluster generated subset is therefore not a high-powered population confirmation. No missing endpoint reduces the four-slot family.

CPU verification covers every raw-logit record and actual edit: Pythia equal-dose and Qwen each contribute 18,432 edits; support contributes 6,912; external calibration/test contributes 1,984. Zero-edit differences are zero throughout. Qwen source directions are independently refit on CPU where linear; its saved DAS optimizer trajectory is replayed from recorded gradients, not a second neural-gradient computation. External calibration choices and statistical reductions are independently recomputed from raw outputs. All failures, exposed development data, and unsuccessful criteria remain separate from the earlier confirmation.

The four new GPU runners take 267.35, 760.24, 117.22, and 38.76 seconds, respectively, on the existing V100. These are runner times, not end-to-end rental cost. The largest full raw package remains on the remote research instance; local archival coverage is specified by filename and hash, not described as a complete redundant backup. No result here establishes SOTA, a training benefit, or an equal-compute estimator advantage.

P Post-review CPU accounting of archived interventions

This appendix was added after review using existing local archives only. It adds no model forward pass, intervention, direction fit, test-set dose selection, or independent training replication. The original protocols and failed criteria are unchanged. The CPU audit binds inputs by SHA-256, preserves every declared operating point, and distinguishes missing measurement primitives from zero effects.

P.1 Native amplitudes in the original Transfer archive

For a unit direction a, the native rank-one intervention has applied length

$$\|\Delta h\|_2 = \left|(h_s - h_b)^\top a\right|. \tag{12}$$

Changing the direction therefore changes both its orientation and the donor-projection amplitude; unit-normalized directions do not imply equal-length edits. We scanned all 252 original Transfer cell files and verified their archived hash sidecars. The matched, within-target, and orthogonal conditions, together with each condition's same-class sham, contain $252 \times 3 \times 2 \times 600 = 907{,}200$ orientation–layer records. None records the target base/source hidden-state pair, donor projections, actual edit vector, or applied norm needed to recover the original-grid amplitude distribution. Directions and output logits alone do not supply these missing primitives. Layer-level effect and field-availability tables are retained in the audit package. The later local equal-dose diagnostics do not retroactively amplitude-match the original Transfer study.

P.2 Complete held-out operating-point inventory

Tables 16 and 17 show all eight source-fitted directions at the reference dose and the previously selected calibration dose. Calibration used the target-domain BLiMP calibration split (10 sentences, eight head lemmas), not the source-direction training split. Each held-out point uses the same 22 sentences and 16 disjoint head lemmas: first average within each head lemma, then weight the 16 lemma means equally. All 163 raw R6 files are present locally and hash-bound; all 704 held-out edits across the two text fields were independently reduced from stored full-vocabulary logits and edit vectors, agreeing with the archived CPU records to 4.4×10^{-15} maximum metric error. This is numerical reconstruction, not new scientific replication.

None of the eight previously selected doses yields a held-out mean gain within 0.5 ± 0.05; no test interpolation or dose reselection is performed. DiM and DAS therefore still do not meet the intended

Direction	Regime	Dose	Gain	Full KL	Conditional KL
DiM	Reference	1	2.12528	0.04460	0.04288
DiM	Calibrated	0.375	0.76736	0.00462	0.00446
Lexical DiM	Reference	1	1.94271	0.03762	0.03600
Lexical DiM	Calibrated	0.375	0.65098	0.00364	0.00353
Ridge	Reference	1	2.08461	0.04275	0.04108
Ridge	Calibrated	0.375	0.73910	0.00425	0.00411
Rank-one DAS	Reference	1	2.91624	0.11893	0.11612
Rank-one DAS	Calibrated	0.25	0.64828	0.00296	0.00285
Shuffled labels	Reference	1	0.06947	0.00284	0.00281
Shuffled labels	Calibrated	4	0.59363	0.04302	0.04289
Random 0	Reference	1	0.01743	0.00202	0.00201
Random 0	Calibrated	3	0.32604	0.01822	0.01817
Random 1	Reference	1	0.03400	0.00195	0.00194
Random 1	Calibrated	4	0.36756	0.02362	0.02344
Random 2	Reference	1	0.14047	0.00292	0.00288
Random 2	Calibrated	2.5	0.41011	0.01665	0.01646

Table 16: All 16 predeclared held-out operating points, evaluated at the matrix-auxiliary prefix. Gain is counterfactual two-candidate log-odds change, not grammatical-accuracy improvement. Both KL measures use $D_{\mathrm{KL}}(p\|q)$ in nats; conditional KL excludes and renormalizes away the two target candidates. Dose multiplies the frozen source reference length, approximately 6.935975. These descriptive means are not equal-benefit method comparisons.

Direction	Regime	Edited mass ($\times 10^{-3}$)	Δmass ($\times 10^{-4}$)	Signed lexical ΔNLL	Absolute lexical drift
DiM	Reference	3.74179	-2.59504	-0.01357	0.10809
DiM	Calibrated	3.95773	-0.43565	-0.00064	0.03214
Lexical DiM	Reference	3.83525	-1.66047	-0.00900	0.09366
Lexical DiM	Calibrated	4.00908	0.07781	0.00005	0.02411
Ridge	Reference	3.71920	-2.82098	-0.00022	0.09766
Ridge	Calibrated	3.95033	-0.50972	0.00416	0.02996
Rank-one DAS	Reference	3.51654	-4.84756	-0.03771	0.11196
Rank-one DAS	Calibrated	3.91266	-0.88636	-0.00529	0.02044
Shuffled labels	Reference	4.30237	3.01067	-0.00940	0.09990
Shuffled labels	Calibrated	4.70775	7.06451	-0.02817	0.29724
Random 0	Reference	3.86832	-1.32978	-0.01792	0.06961
Random 0	Calibrated	3.47629	-5.25012	0.00276	0.20607
Random 1	Reference	3.98160	-0.19697	-0.03507	0.07751
Random 1	Calibrated	3.72594	-2.75362	-0.08098	0.20919
Random 2	Reference	3.76111	-2.40187	-0.01311	0.08012
Random 2	Calibrated	3.34852	-6.52778	-0.04505	0.19034

Table 17: Target-mass and lexical accounting for the same points as Table 16. The common unmodified target-mass mean is 0.0040012984; edited mass and its change are measured at the matrix-auxiliary prefix. Lexical changes are measured on the separate teacher-forced prefix ending after the original correct auxiliary; they concern the next prescribed lexical word, not whole-sentence quality. Positive signed ΔNLL indicates reduced probability of that word. Absolute drift is the mean of rowwise absolute changes before cluster aggregation, not the absolute value of the signed mean. Both lexical columns are in nats.

equal-benefit comparison. Reference DiM passes the registered gain lower-bound criterion, while the registered upper-bound certifications for conditional KL and absolute lexical drift are not established. Failure of an upper-bound certification is not, by itself, proof of the opposite inequality. The observed means exceed the stated budgets; the separate lexical-drift lower-bound comparison above its budget remains exploratory. The family retains all four endpoint slots, including the infeasible superiority slot.

P.3 Retained orientation in the equal-energy support diagnostic

For normalized dense direction a, retained support S, and normalized truncation $b = a_S/\|a_S\|_2$,

$$a^\top b = \|a_S\|_2, \qquad \text{retained direction energy} = \|a_S\|_2^2. \tag{13}$$

The audit rederives all nine R5 support memberships from the archived source direction and checks direct cosine against retained norm within 2×10^{-7}. The selected supports rank the absolute source DiM coordinates. Table 18 separates orientation, retained energy, and the actual applied length. The two coordinate fractions are $85/1024$ and $341/1024$; the three random supports per width are fixed draws, not three independent models. These quantities do not establish target-gradient alignment or uniquely necessary coordinates.

Support	Coordinates	Cosine / retained norm	Retained energy	Applied L_2
Dense	1024	1.000000	1.000000	6.935975
Selected 85	85	0.623803	0.389131	6.935975
Random 85 / 0	85	0.341027	0.116300	6.935975
Random 85 / 1	85	0.291698	0.085087	6.935975
Random 85 / 2	85	0.313889	0.098526	6.935975
Selected 341	341	0.896352	0.803447	6.935975
Random 341 / 0	341	0.603498	0.364210	6.935975
Random 341 / 1	341	0.569388	0.324203	6.935975
Random 341 / 2	341	0.620714	0.385286	6.935975

Table 18: All nine supports in the single-checkpoint Pythia-410M equal-energy development diagnostic. Applied lengths agree at the displayed precision, while dense-direction alignment differs. This is post-review accounting of the archived diagnostic, not a norm-matched rerun of the earlier Pythia coordinate-dose confirmation.

P.4 All three SmolLM2-1.7B support-budget pairs

The archived CPU analysis reduces each size–checkpoint–selector–budget group by the median across its layer–task entries. We recompute all 54 group medians and retain all checkpoint pairs. Table 19 gives the three 1.7B causal-leverage pairs underlying the reported budget gain. Their median paired change is 32.940507 percentage points, with a finite-list range of 28.881144–37.192337. These are selected checkpoints from one training run, not independent training replicates. The historical bootstrap resampled the three 1.7B checkpoint IDs for the budget contrast, and separately resampled three large-model versus six pooled smaller-model checkpoints for the scale contrast. Those intervals are conditional finite-list sensitivities, not justified independent-training population uncertainty.

Phase	Training step	Energy at 8.3%	Energy at 33.3%	Paired gain (pp)
Early	125,000	0.298072	0.669995	37.192337
Middle	2,500,000	0.368328	0.697733	32.940507
Late	5,125,000	0.453966	0.742778	28.881144

Table 19: All three selected SmolLM2-1.7B checkpoints, paired by checkpoint. Direction-energy entries are fractions; the last column is their difference in percentage points. Actual supports are 171 and 683 coordinates out of 2048. The full audit retains all 27 size–checkpoint–selector pairs; no size or selector is omitted from its structured tables.

P.5 Recovered power assumptions and a new 16-cluster sensitivity

The original pretest policy, R3 development records, receipt, and planning script match their archived hashes. Its alternative means were 0.75 for gain, 0.01 for conditional KL, and $\log(1.05)/2$ for absolute lexical drift. Distances from the corresponding decision boundaries were 0.25, 0.01, and $\log(1.05)/2$, respectively. The noise envelope takes the largest sample standard deviation of 12 lexical-pair means over all 96 method–dose–order–site configurations for each endpoint. The maximizing configurations are DiM/fourfold/postposed/attractor for gain, DAS/fourfold/postposed/attractor for KL, and random direction 2/fourfold/fronted/attractor on the collateral prefix for lexical drift. No held-out mean determines these assumed alternatives.

All 45 original scenarios (three endpoints, three SD multipliers, and cluster counts $12, 24, 48, 96, 192$) are reproduced. The original grid did *not* include 16. Table 20 adds $n = 16$ only as a post-review sensitivity using the same assumptions. It is not backdated as pretest planning or interpreted as achieved power. For assumed standard deviation $c\sigma$, the calculation is a one-sided noncentral-t rejection probability with noncentrality $\sqrt{16}\,|\mu_1 - b|/(c\sigma)$ and per-slot $\alpha = 0.0125$.

Endpoint	μ_1	Boundary b	SD envelope σ	$P_{16}(\sigma)$	$P_{16}(2\sigma)$	$P_{16}(4\sigma)$
Gain	0.75000	0.50000	0.22668	0.96284	0.41315	0.10969
Conditional KL	0.01000	0.02000	0.05831	0.05364	0.02707	0.01860
Absolute lexical drift	0.02440	0.04879	0.14821	0.05092	0.02628	0.01831

Table 20: Recovered alternative/noise assumptions and post-review $n = 16$ power scenarios. Probabilities depend on independent approximately normal lexical-cluster means and transporting the development SD envelope to the restricted BLiMP subset; neither assumption is established by this numerical recovery. Values are not posterior or realized power estimates.

The fourth, equal-benefit superiority slot had no established matched-benefit baseline variance and remains unpowered in this planning record. The four-slot family is not reduced. Calibration infeasibility, imprecision of an upper-bound certification, and an observed mean on the unfavorable side of its budget are distinct issues; this recovery neither changes the thresholds nor authorizes a new sample-size choice or model run.

Reproducible accounting interface. The deterministic analysis and formatting scripts, field inventories, per-head-lemma values, all operating points, and hashed input manifest are retained with the supplementary evidence. The analysis script is `repair_metrics_20260906.py`; this appendix is rendered from its CSV outputs by `repair_metrics_20260906_format_tex.py`. No unavailable original amplitude is imputed, no unfavorable operating point is dropped, and no new method-superiority or population-level claim is derived from these tables.

Q Writing-site dependence and a fixed two-MLP follow-up

Question and scope. We ask whether a fixed direction has the same structural response at different writing sites, and whether two downstream components account for it. These Pythia-410M step-143000 diagnostics do not replace protected Transfer evidence or add training replications. Layer indices are zero-based. Runs use float32, eager attention, batch size eight, and no parameter updates.

Materials and response. Each set contains 24 lexical blocks, four first/second-noun number configurations, and three families (288 prefixes). In subject coordination, both nouns form the matrix subject: "The passenger and the commuter, after greeting the visitor,". In embedded coordination, only the first noun controls the matrix continuation: "The passenger, after greeting the commuter and the visitor,". A single-subject control is "The passenger near the commuter". Relative-clause materials instead use "who greeted" and the corresponding subject/embedded coordination placement. The target contrast is *have* minus *has*. We add a fixed noun-derived unit DiM vector with signed amplitude ± 1.734 at the layer-11 output, either at the last token of the first noun or at the final prefix token. For margin m and writing site s, define

$$\begin{aligned} R_{f,s} &= \tfrac{1}{2}\{m_{f,s}(+\alpha a) - m_{f,s}(-\alpha a)\}, \\ D_s &= R_{\text{embedded},s} - R_{\text{subject},s}, \qquad J = D_{\text{noun}} - D_{\text{final}}. \end{aligned} \tag{14}$$

These are finite-dose logit responses, not derivatives or grammatical-accuracy improvements. We average the four number configurations within lexical block before taking the family contrast. Descriptive 95% percentile intervals resample the 24 paired blocks 4,000 times; they are unadjusted for the broader exploratory search and do not represent model-population uncertainty.

Holding direction identity fixed. The earlier crossed-site experiment crosses two source-estimated directions, two sites, both signs, two doses, and fixed orthogonal controls. The lower-dose noun-derived direction gives positive site–scope interactions on both exposed sets (Table 21). After-gerund final-site responses are 0.752 and 0.746 nats: both interventions remain effective. The final-site-derived direction is less consistent, with interactions of 0.008 [−0.038, 0.056] on relative clauses and 0.130 [0.057, 0.195] on after-gerund materials. We do not infer a direction-invariant rule. Site changes also alter propagation distance and available context; equal length does not match distributional damage.

Material set	D_{noun}	D_{final}	J [95% interval]
Relative clause	0.237	0.035	**0.203 [0.167, 0.239]**
After-gerund	0.245	0.006	**0.239 [0.179, 0.296]**

Table 21: Same noun-derived direction and applied dose, different writing sites; units are nats. Both material sets were exposed before the crossed-site follow-up. These are within-model diagnostics, not two independent model replications.

Component selection and the failed capability gate. An earlier development screen examined attention and MLP outputs at layers 12, 17, and 23. It selected L17 and L23 MLPs for follow-up; the selection was not blind. Before testing their joint effect, a new lexical/*while*-clause set failed the fixed native-capability gate: embedded coordination was correct on 61/96 prefixes, subject coordination on 94/96, and the single-subject control on 96/96. The run stopped after its 288 baseline evaluations, with no component interventions. A baseline-only lexical-by-template audit found 96/96 embedded accuracy for the new lexicon with the original template, and 77/96 for the old lexicon with the new template. The failed set is retained. This diagnosis concerns a template bundle, not the isolated word *while*, and does not retroactively validate that set.

The final two runs. We fixed two further runs before either outcome was observed: new lexical items with the original relative-clause templates (A), and the previously evaluated after-gerund materials (B). Their baselines were known; neither is a fresh blind confirmation. Each has 288/288 correct unmodified preferences, including every number/family cell. Both use the same noun-site writer and the two selected MLPs, with no layer or dose search. At the final prefix token, we replace the L17 output, the L23 output, or both with their own unmodified values. The replacement is absolute; the second node is clamped after any upstream intervention has propagated. These operations remove an intervention-induced component response, rather than delete the component from natural computation. With b_1, b_2 indicating the two clamps, define

$$I_f = R_f^{11} - R_f^{10} - R_f^{01} + R_f^{00}, \qquad \Delta I = I_{\text{embedded}} - I_{\text{subject}}. \tag{15}$$

The predeclared primary diagnostic is ΔI. Seeded random additions match each node's actual clamp-displacement norm, separately for each sign and mask, but not its direction or damage. We also patch unmodified donor outputs without the artificial writer: same-structure opposite-first-noun-number donors, next-block same-number donors, and opposite-number donors from the other coordination family. The last control changes multiple features and is not a pure structure intervention. Native effects are oriented toward the donor's first-noun number, not necessarily the grammatically correct matrix continuation. Each run evaluates 24 intervention arms and 7,792 sequences including baseline, writer captures, and no-op checks; these counts are not independent samples.

Results and interpretation. L23 clamping reduces the artificial scope contrast in both runs; adding L17 yields no detected additional nonadditive scope effect (Table 22). This does not prove additivity or absence of a smaller interaction. Natural-donor patches produce different patterns (Table 23). Their nonadditivity cannot be substituted for the artificial-writer result: the interventions differ, and wrong-family donors also produce a negative interaction on A. A native patch may change other

Artificial-writer scope contrast	A: relative clause	B: after-gerund
No clamp, D^{00}	0.24164	0.24518
L17 clamp, D^{10}	0.24413	0.24248
L23 clamp, D^{01}	0.17233	0.17071
Both clamps, D^{11}	0.17474	0.16840
Joint interaction, ΔI	-0.00008	0.00038
95% interval for ΔI	$[-0.00144, 0.00123]$	$[-0.00246, 0.00337]$
Random-adjusted ΔI	-0.00009	0.00044
95% interval	$[-0.00148, 0.00123]$	$[-0.00239, 0.00343]$

Table 22: Fixed two-MLP follow-ups in nats. L23 clamping reduces the artificial scope contrast but leaves a substantial remainder. The interaction intervals cross zero in both runs; no equivalence bound was specified. Rows retain all number configurations.

Native-donor joint scope interaction	A [95% interval]	B [95% interval]
Same structure, opposite N1 number	-0.0192 $[-0.0354, -0.0026]$	-0.0741 $[-0.0955, -0.0527]$
Other lexical block, same number	0.0041 $[-0.0135, 0.0210]$	0.0068 $[-0.0150, 0.0276]$
Other family, opposite N1 number	-0.0217 $[-0.0399, -0.0038]$	0.0157 $[-0.0036, 0.0340]$
Norm-matched random addition	0.0035 $[-0.0004, 0.0079]$	0.0039 $[-0.0012, 0.0095]$

Table 23: For native donor patches, E is the donor-number-oriented margin change and the interaction is $E^{11} - E^{10} - E^{01}$; the table reports its embedded-minus-subject contrast. Intervals describe lexical-block variability, not multiplicity-corrected circuit discovery.

features or downstream computation. The complete tables retain the single-subject family, all number cells, original-gold accuracy, actual edit norms, and merged-target KL; none is an outcome-based exclusion. These diagnostics constrain the candidate mechanism but do not identify a selective serial circuit, a rescue of grammatical computation, or an independent content/routing decomposition.

Integrity, exposure, and limits. CPU checks cover baseline full-logit replay, no-op outputs, replacements, norm controls, code/parent/raw hashes, and all-condition reductions. Final protocol SHA-256 prefixes are `36c3005813bd` and `1299d4bc3fdd`; the failed capability run is `3f97584856cf`. All outcomes are retained; internal freezing is not external preregistration. Full raw arrays are locally archived outside the compact companion, not established as publicly available. One checkpoint, selected components, synthetic prefixes, and no independent human stimulus review preclude a model-general mechanism or method ranking.

R Implementation, provenance, and supplementary sensitivities

Implementation and study chronology. Difference-in-means directions in the original grids are estimated from the complete training activation set and unit-normalized on CPU float32. Each orientation's raw intervention contrast uses four stored float32 logits. Confirmatory studies fix rows, checkpoints, model revisions, layers, directions, thresholds, and analysis code before their held-out evaluations. Separate scripts recompute raw-logit reductions and onset decisions without further model runs. Internal freezing is not external preregistration; the initial sham-negative study remains negative, and exposed diagnostics retain their separate status.

Selected sites and model identity. Of 1,512 selected Transfer sites, 1,051 are at the subject before the PP/RC marker, 105 have partial intervening context, and 356 have the full intervening phrase. None is at the final target matrix verb. The inventory describes the original selection, not a controlled comparison of writing sites. Model manifests identify 300 distinct checkpoint-weight digests over 30 size–seed trajectories. Strong, Transfer, and Pythia support use overlapping subsets, not additional independent weight replications.

Splits, selection, and timing sensitivities. Original tasks contain 400 training, 100 development, and 100 test rows after tokenization checks. The same task schedule is reused across models and

checkpoints. Development-region ties follow the fixed region order. Layer anchors include all six blocks for 70M and evenly spaced anchors for 160M and 410M. Transfer's later median matched-minus-orthogonal model effects are 3.627 at step 2000 and 3.751 at step 3000. Right-censored timing holds in 15/21 models (post-hoc one-sided $p = 0.0392$), versus strict timing in 12/21 ($p = 0.332$). These unadjusted exploratory tests are not new confirmatory findings. The nine-combination detector matrix in Figure 7 conditions its normalized bands on fixed observed late denominators, not full ratio uncertainty.

Availability detail. The separate Transfer archive contains 252 raw-score cells and metadata for replaying 504 matched/orthogonal gates. Neither it nor the compact companion includes weights, credentials, or external-test logits/edit arrays. Strong coverage comprises manifests and derived trajectories, not all 900 raw cells. The additional 45,760 edits have mixed local, sampled, and remote-only coverage; full replay requires further inputs. Availability claims apply to named packages, not every local research file.

Pythia model-level sensitivities. Selected support passes all three targets in 14/14 models versus 0/14 for random support (post-hoc paired sign $p = 6.10 \times 10^{-5}$). Blocking both sizes by shared numeric seed gives 7/7 ($p = 0.00781$). These are unadjusted sensitivity analyses, not additional confirmatory families. In the removal study, full removal passes in 14/14 models and selected removal in 12/14, versus 0/14 for random removal; the corresponding unadjusted paired sign values are 6.10×10^{-5} and 2.44×10^{-4}. The encoding criterion has 0/42 passing target trajectories, and the joint claim remains unsupported.

SmolLM2 coverage. The three actual small supports are 48/576, 80/960, and 171/2048 coordinates; the large supports contain 192, 320, and 683 coordinates. The nine groups contain 275,400 registered records, not independent replications. The focal gradient is batch weighted; a uniform-gradient variant was not evaluated in this original cross-scale study. The separate 1.7B follow-up is reported in Appendix G. These details distinguish the implemented selector from a claim about all causal-leverage estimators.

Retained direction energy changes with the budget. In a post-hoc analysis of the nine completed groups, small causal-leverage support captures 50.7%, 48.2%, and 36.8% median direction energy at 135M, 360M, and 1.7B. At 1.7B, expansion adds a median 32.9 percentage points as onset recovers from 0/3 to 3/3 targets (Figure 6). These changes co-occur; budget also changes orientation and applied dose. All three checkpoint pairs per size are shown. They are points on one training run, not independent training replications, and any checkpoint-resampling interval describes this finite list only.

External bound decisions. DiM's reference-dose target gain has a one-sided 98.75% lower bound of 1.406 nats. Conditional non-target KL and absolute lexical ΔNLL have upper bounds of 0.07629 and 0.15577, compared with tolerances 0.02 and $\log(1.05)$. The observed means exceed the tolerances, and an exploratory opposite-tail lexical-drift bound does too. The latter is not the prospective certification test. The joint criterion remains unmet; a failed upper-bound certificate alone does not prove a population violation. All eight methods and their 16 target-domain calibration doses remain part of the design, including the unsuccessful equal-benefit comparison.

S What the intervention measurements can identify

We fix a model, layer, hidden-state basis, and collection of materials and readouts indexed by e. The analysis states elementary linear-algebra consequences of the measured intervention families. It introduces no new estimator and assumes neither a globally linear language model nor an empirically identified compensation mechanism.

Additive edits and native projection swaps. Let $F_e(\delta)$ be a differentiable readout after an additive edit at a specified site, and let $A_e = DF_e(0)$. In the local linear model, edits with the same stacked response $B\delta = y$ form

$$\{\delta : B\delta = y\} = B^{\dagger}y + \ker B, \qquad y \in \operatorname{range}(B). \tag{16}$$

Intersect this set with the allowed norm or support constraints; it need not contain multiple admissible edits. If F_e has Hessian norm bounded by L_e along the edit, its remainder is at most $L_e\|\delta\|^2/2$.

Native interchange instead uses $\delta_e = aa^\top d_e$, where $d_e = h_{s,e} - h_{b,e}$. Holding donors fixed, its linearized measurements act on the projector:

$$\mathcal{L}(P) = (A_e P d_e)_e, \qquad P = aa^\top. \tag{17}$$

Two unit directions a, b are response-equivalent in this linear model exactly when $\mathcal{L}(aa^\top - bb^\top) = 0$. A projector is identifiable only if the kernel contains no nonzero admissible projector difference. The swap is invariant to $a \mapsto -a$. Equal onset decisions are coarser than continuous response equivalence and cannot establish this condition.

Alignment matching has a feasibility constraint. For a unit dense direction a and any unit vector v supported on S, Cauchy–Schwarz gives

$$|a^\top v| \le \|a_S\|. \tag{18}$$

Normalized truncation $a_S/\|a_S\|$ attains the positive bound. Magnitude top-k maximizes this bound over k-coordinate supports. With a strict gap between the kth and $(k+1)$th squared magnitudes, another support cannot match the top-k truncation's cosine. For $|S| \ge 2$ and $a_S \ne 0$, a common feasible cosine c can instead be constructed as

$$v_S = \frac{c}{\|a_S\|^2} a_S + \sqrt{1 - \frac{c^2}{\|a_S\|^2}}\, w_S, \quad w_S \perp a_S, \quad \|w_S\| = 1, \quad |c| \le \|a_S\|. \tag{19}$$

Here w_S is supported on S. Orthogonality verifies unit norm and cosine c. This changes the truncation rule. Appendix U evaluates the construction on previously exposed materials at common cosine and applied norm. These constraints do not equate full orientation, damage, or target-gradient alignment.

Finite-dose response and separability. Write the symmetric response as $R_{f,s}(\alpha) = [f_{f,s}(\alpha) - f_{f,s}(-\alpha)]/2$, where $f_{f,s}(t) = m_{f,s}(ta)$ is the scalar margin for family f and site s. If its third derivative is bounded by $M_{f,s}$ on $[-\alpha, \alpha]$, Taylor expansion gives

$$\left|R_{f,s}(\alpha) - \alpha f'_{f,s}(0)\right| \le M_{f,s}|\alpha|^3/6. \tag{20}$$

The even-order terms cancel. Without a verified bound, a finite-dose interaction is not a derivative measurement: $f(t) = t^3$ has zero derivative at zero but nonzero central response. Also, $R_{f,s} = u_f v_s$ can yield nonzero $J = (u_E - u_S)(v_N - v_F)$. A more specific diagnostic is the minor

$$Q = R_{E,N} R_{S,F} - R_{E,F} R_{S,N}, \tag{21}$$

which is zero for exact multiplicative separability. Averages of rank-one block matrices need not be rank one, so block-wise minors and the minor of a mean answer different questions. Appendix T reports both.

Target-preserving collateral compensation. In an exact linear proxy, fix target response $B\delta = y$, collateral $\|C\delta\|_2$, and norm budget ρ. Let $\delta_0 = B^\dagger y$ and let the columns of N be an orthonormal basis of $\ker B$. For feasible y with $\|\delta_0\| \le \rho$, write $\delta = \delta_0 + Nz$. Orthogonality gives $\|\delta\|^2 = \|\delta_0\|^2 + \|z\|^2$, and hence

$$\epsilon_* = \min_{\|z\| \le \sqrt{\rho^2 - \|\delta_0\|^2}} \|C\delta_0 + CNz\|. \tag{22}$$

Zero collateral is possible exactly when $-C\delta_0 \in \mathrm{range}(CN)$ and its minimum-norm preimage $-(CN)^\dagger C\delta_0$ lies within this radius. If y is outside $\mathrm{range}(B)$ or $\|\delta_0\| > \rho$, the target itself is infeasible. These statements follow from the full solution of the linear target equation and the minimum-norm solution of the collateral equation.

For example, take $B = (1, 0)$, $C = (1, 1)$, and $y = 1/2$. Edits $(1/2, 1/2)$ and $(1/2, -1/2)$ have the same norm and target response, but collateral 1 and 0. Restricting support to the first coordinate forces $(1/2, 0)$, with collateral $1/2$. A target-invisible coordinate can therefore enable compensation. More generally, nested supports enlarge the feasible edit set, so optimal collateral cannot increase with support at fixed y and ρ. This statement concerns the optimum over all allowed edits, not the behavior of a particular truncated direction.

The proxy does not certify the nonlinear model. Conditional KL and absolute lexical NLL require their own local metrics and remainder bounds; neither is automatically a linear $C\delta$. Appendix U estimates a finite-difference target gradient and a binned conditional-Fisher surrogate in a fixed 16-dimensional basis. Its finite-dose test evaluates this local construction, not a general compensation mechanism. It does not estimate a lexical-NLL operator or validate on untouched material.

T POST-HOC RESPONSE-SHAPE AND DOSE DIAGNOSTICS

Scope and complete inclusion. We reanalyze the stored crossed-writer responses without model calls. All 9,216 rows are retained: two exposed material sets, two direction origins, two writing sites, DiM and orthogonal directions, two doses, 24 lexical blocks, four number assignments, and three families. Within each block we average the four number assignments. The actual amplitudes are 1.733993649482727 and 3.467987298965454, rather than the dose indices 0 and 1. All 48 family-pair minors and 48 dose comparisons are retained in the accompanying tables. Pointwise 95% percentile intervals use 20,000 paired block resamples with seed 2026092107. They are descriptive, not simultaneous intervals or a new confirmatory testing family.

Within-block separability diagnostic. For the original lower-dose noun-derived DiM comparison, the embedded/subject minor Q is positive in 24/24 relative-clause blocks and 24/24 after-gerund blocks. Table 24 reports both the average within-block minor and the minor of the average response. This distinction avoids treating aggregate rank as a test of block-specific rank. The result describes deviation from an exact gain model on observed stimuli, not rejection of an approximate model with a prespecified tolerance or evidence for a routing module. Orthogonal controls and all other families, origins, and doses remain in the full tables.

Materials	Mean block Q [95% interval]	Q of means	Positive blocks
Relative clause	0.0669 [0.0565, 0.0775]	0.0661	24/24
After-gerund	0.1822 [0.1547, 0.2107]	0.1815	24/24

Table 24: Post-hoc lower-dose noun-derived DiM diagnostic. Minors have units of squared nats. Two exposed sets share one model checkpoint; the block counts are not independent model replications. No significance decision or equivalence claim is based on these pointwise intervals.

Finite-dose behavior is not uniformly linear. For the noun-derived direction, relative-clause embedded responses at the noun site have $R(\alpha)/\alpha = 0.2108$ and 0.2962 at the two amplitudes. Their paired difference is 0.0853 [0.0715, 0.0982]. After-gerund values at the same site are 0.2603 and 0.3347, with difference 0.0743 [0.0638, 0.0858]. Final-site changes are smaller and can have a different sign. These finite-dose results do not establish a zero-dose derivative or a uniform Taylor bound. Every dose/source/site/family comparison is retained, including intervals spanning zero.

Support and operating-point checks. We retain all nine support choices in both orders and verify the normalized-truncation identity between squared cosine and retained energy. Their actual edit norms agree to within 10^{-5}. We also cross-check all 16 external method/regime operating points against all 32 field-specific records: target gain and conditional KL use the prefix field, while lexical drift uses the collateral-text field. These checks validate the organization of existing audited summaries, not a fresh replay of all raw logits. Neither all-point display nor mean-level tolerance comparisons establish a jointly certified low-collateral method.

U MATCHED SUPPORT AND CONSTRAINED COLLATERAL REDUCTION

Two supplementary tests address gaps left by the earlier diagnostics. First, equal edit norm does not control alignment with the source direction. Second, a failure to certify collateral preservation does not establish that collateral is irreducible. The tests below evaluate the matched-cosine construction and a local compensation proxy from Appendix S. They do not replace the original transfer or BLiMP experiments.

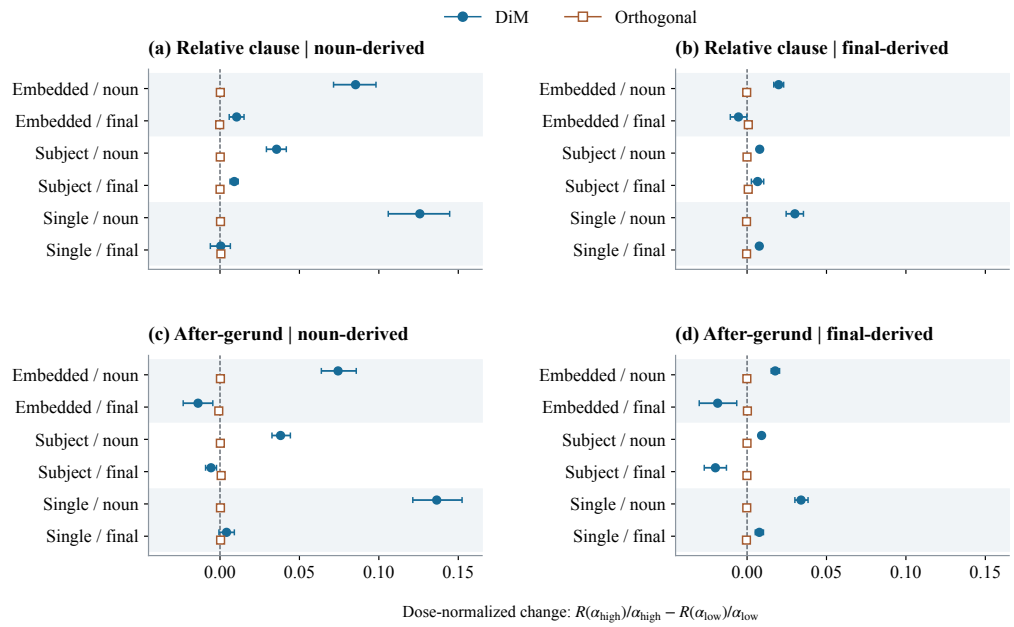

Figure 11: All 48 dose-normalized contrasts from the stored response table. Panels cross material set and direction origin; each retains all families, both writing sites, DiM, and the orthogonal control. Points are means over 24 paired lexical blocks; bars are pointwise 95% bootstrap intervals, not simultaneous intervals. The common horizontal scale includes zero and every interval. Nonzero changes describe finite-dose nonlinearity, not a measured zero-dose derivative. Both material sets were exposed and share one model checkpoint.

Shared setting and exposure. We fix Pythia-410M at step 143000, the output of layer 11 (zero-based), and the first noun position. The frozen noun-derived DiM direction and reference amplitude $\alpha = 1.733994$ come from the completed phrase-scope diagnostic. Each material set has 24 lexical blocks, four number assignments, and three families: 288 relative-clause and 288 after-gerund rows. Both sets were examined in earlier work. The fitting split below is disjoint only within this supplementary run, not an untouched test set. The response is control of the *have-minus-has* logit margin, not accuracy on grammatical continuations.

For edit δ, we measure central response and collateral as

$$R_e(\delta) = \frac{m_e(\delta) - m_e(-\delta)}{2}, \qquad K_e(\delta) = \frac{1}{2} \sum_{s \in \{-1,1\}} D_{\mathrm{KL}}(q_e(0) \,\|\, q_e(s\delta)) \,, \tag{23}$$

where q_e is the full next-token distribution conditioned on excluding exactly *has* and *have*. This definition permits the excluded target mass to change. It is neither merged-target KL nor lexical continuation ΔNLL. We retain each signed margin change, conditional KL, and excluded target mass in the accompanying row table.

U.1 Matching cosine and applied norm

For each budget $k \in \{64, 256\}$, we use the magnitude top-k support and two seeded uniform random supports. Each support receives two seeded source-orthogonal orientations. Two dense common-cosine vectors, dense DiM, and the existing orthogonal control complete the 16-vector inventory. Using only the source direction, we set $c = 0.8 \min_S \|a_S\| = 0.174050$ over the selected support inventory and construct the unit vectors from Appendix S. Both signs receive the same actual additive norm α. Dense DiM and the orthogonal reference have cosine 1 and 0; they are not members of the common-cosine comparison.

Figure 12 shows every vector on both materials. The prespecified contrast averages the two top-support orientations and subtracts the mean over the four random-support orientations, pairing within lexical block. The four aggregate response differences are 0.0124 [0.0095, 0.0153] and 0.0343 [0.0296, 0.0390] nats on relative clauses, and 0.0154 [0.0118, 0.0194] and 0.0297 [0.0244, 0.0358] on after-gerund material, at 64 and 256 coordinates respectively. Brackets denote pointwise 95% lexical-block intervals.

These contrasts condition on fixed support and orientation draws. At 64 coordinates, the subject-coordination contrast is unresolved in both materials; at 256, each family's response contrast is positive. Conditional-KL differences do not uniformly favor selected support: the after-gerund, 256-coordinate aggregate interval crosses zero (Table 25). The two same-cosine dense orientations also have substantially different responses. Thus source cosine and edit norm are insufficient to determine response, but the experiment does not isolate unique causal coordinates or establish superiority over dense DiM.

U.2 A calibration-fitted reorientation

We fit on relative-clause blocks 0–7 (96 rows), then evaluate on blocks 8–23 (192 rows) and all 24 after-gerund blocks (288 rows). Calibration and evaluation texts are disjoint within the run. Let U contain the frozen DiM direction followed by 15 seeded orthonormal random vectors. Central finite differences at zero, with step 0.05α, estimate the mean target gradient b in this basis. For collateral fitting, each calibration row has 64 baseline-selected non-target tokens and a residual bin. If $p_{e,r}$ is a conditional bin probability and $g_{e,r}$ its finite-difference log-probability gradient, we use

$$H = \frac{1}{96} \sum_e \sum_{r=1}^{65} p_{e,r} g_{e,r} g_{e,r}^\top. \tag{24}$$

This binned Fisher is a local surrogate; evaluation uses the full conditional vocabulary. With $\theta_0 = \alpha e_1$ and $\lambda = \max(10^{-10}, 10^{-3} \operatorname{tr}(H)/16)$, we solve

$$\min_\theta \; \theta^\top H \theta + \lambda \|\theta\|^2 \quad \text{subject to} \quad b^\top \theta = b^\top \theta_0, \qquad \|\theta\| \le \alpha. \tag{25}$$

A deterministic multiplier search enforces the norm constraint; it performs no model-evaluated hyperparameter tuning. We compare the full-H solution with the diagonal-H solution under the same

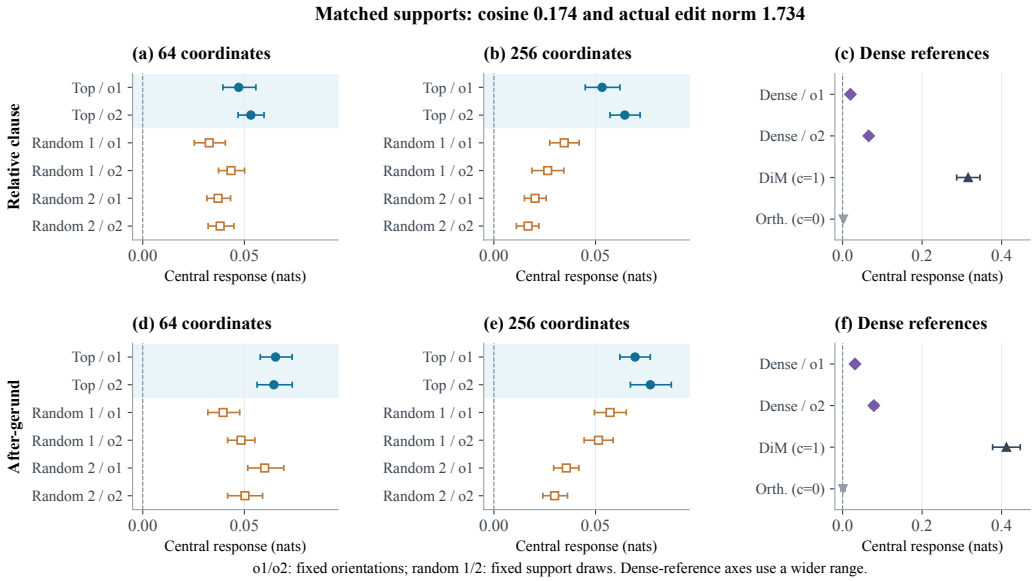

Figure 12: Every frozen support/orientation under matched source cosine and actual additive norm. The first two columns share a response scale; the dense-reference column uses a wider, explicitly labeled scale. Each point averages 24 lexical blocks; bars are pointwise 95% block-bootstrap intervals. Two top orientations and four random orientations per budget remain visible, rather than only their favorable averages. Dense DiM and the orthogonal reference are not cosine-matched controls. Neither support draws nor repeated conditions are independent model replications.

Material	k	Family	Response difference [95% CI]	KL difference [95% CI]
RC	64	All	0.0124 [0.0095, 0.0153]	-0.462 [-0.673, -0.255]
RC	64	Embedded	0.0072 [0.0046, 0.0097]	-0.211 [-0.389, -0.047]
RC	64	Subject	0.0015 [-0.0014, 0.0045]	-0.195 [-0.315, -0.086]
RC	64	Single	0.0284 [0.0228, 0.0342]	-0.980 [-1.467, -0.518]
RC	256	All	0.0343 [0.0296, 0.0390]	-0.181 [-0.333, -0.015]
RC	256	Embedded	0.0316 [0.0270, 0.0360]	-0.267 [-0.424, -0.118]
RC	256	Subject	0.0280 [0.0226, 0.0335]	-0.057 [-0.144, 0.031]
RC	256	Single	0.0433 [0.0353, 0.0514]	-0.217 [-0.586, 0.162]
After-gerund	64	All	0.0154 [0.0118, 0.0194]	-0.541 [-0.813, -0.294]
After-gerund	64	Embedded	0.0116 [0.0061, 0.0169]	-0.499 [-0.773, -0.235]
After-gerund	64	Subject	-0.0013 [-0.0049, 0.0022]	-0.106 [-0.249, 0.031]
After-gerund	64	Single	0.0358 [0.0295, 0.0426]	-1.019 [-1.596, -0.449]
After-gerund	256	All	0.0297 [0.0244, 0.0358]	-0.013 [-0.170, 0.145]
After-gerund	256	Embedded	0.0156 [0.0110, 0.0212]	-0.100 [-0.290, 0.089]
After-gerund	256	Subject	0.0269 [0.0209, 0.0339]	0.067 [-0.058, 0.198]
After-gerund	256	Single	0.0465 [0.0371, 0.0562]	-0.007 [-0.293, 0.305]

Table 25: All support contrasts: the average of two top-support orientations minus the average of four random-support orientations, within each material and budget. Response is in nats; KL differences are in 10^{-4} nats. Intervals use 24 lexical blocks, conditional on these fixed support/orientation draws. Aggregate rows average all three families. At $k = 64$, the subject-family response contrast is unresolved in both materials.

constraints, DiM, and DiM scaled to the full solution's norm. All coefficients are frozen before either finite-dose test, at scales 0.5 and 1 with both signs. No test-dependent refitting is allowed.

Criterion and observed operating points. For each material and scale, the frozen joint diagnostic requires the entire paired central-response difference interval against DiM to lie inside $\pm 10\%$ of the calibration DiM mean, and the paired KL-difference upper bound to be below zero. Calibration and

		Central response (nats)			Conditional KL (10^{-4} nats)		
Material	Scale	DiM	Full	Diagonal	DiM	Full	Diagonal
RC	0.5	0.1227	0.1221	0.1219	1.445	1.276	1.268
RC	1	0.3019	0.2975	0.2962	8.232	7.310	7.241
After-gerund	0.5	0.1741	0.1717	0.1715	2.319	2.079	2.072
After-gerund	1	0.4121	0.4033	0.4020	13.376	11.962	11.885

Table 26: All fit-excluded aggregate operating points. RC: 16 blocks/192 rows; after-gerund: 24 blocks/288 rows. Full and Diagonal denote the full- and diagonal-Fisher constrained solutions. Norm-matched DiM is exactly identical to DiM because the norm constraint is active. The diagonal solution has slightly smaller KL in every displayed cell; there is no full-Fisher superiority claim.

evaluation DiM means must both exceed 0.05 nats. The absolute response tolerances are 0.014363 and 0.034361 nats at the two scales. They are conditional on calibration, not recomputed as 10% of each evaluation mean. Intervals use 10,000 paired lexical-block bootstrap resamples, are pointwise rather than simultaneous, and do not propagate calibration uncertainty.

All four aggregate evaluation checks pass (Figure 4). Point means retain 97.8–99.6% of central response and reduce KL by 10.4–11.7%, corresponding to absolute reductions of 1.69×10^{-5}–1.41×10^{-4} nats. Table 26 gives every aggregate operating point. The norm constraint is active at 1.733994, so norm-matched DiM is identical to DiM, not a second independent baseline. The observed reduction is not obtained by using a smaller edit norm.

The simpler comparator limits the method claim. Full-H has slightly higher mean KL than diagonal-H in all four cells. On full-dose relative clauses, their difference is 6.95×10^{-6} nats, with interval $[1.46 \times 10^{-6}, 1.21 \times 10^{-5}]$; the other three intervals include zero. All four target-response difference intervals between the solutions include zero. This is no evidence of full-Fisher superiority, and not proof of exact equality. The useful positive comparison supports modest constrained reorientation, which a diagonal approximation also achieves.

Finite-dose and per-example limits. The local target equality does not imply exact finite-dose preservation. Full-dose calibration DiM response is 0.343607 nats versus a local prediction of 0.267242; compensation gives 0.338427 against the same prediction. The finite-dose checks, not the equality constraint alone, support approximate mean preservation. Three aggregate target-difference intervals against DiM exclude zero on the negative side, while remaining within the frozen tolerance. Central-response retention also does not certify either sign or each example. Conditional collateral, excluded target mass, and lexical or generation quality are distinct endpoints. Single-checkpoint, exposed-material evidence does not establish a general compensation mechanism or repair the earlier BLiMP certification failure.

Coverage and numerical checks. The frozen run records 31,312 scored sequences: 18,432 support evaluations, 9,216 reorientation evaluations, 3,072 calibration finite differences, 576 baselines, and 16 no-op checks. An independent CPU implementation reconstructs metrics from full-vocabulary FP32 logits, re-estimates the calibration matrices, solves the constrained problem independently, reconstructs support vectors, and checks checkpoint and artifact hashes. It passes 1,571 checks. The 448 condition summaries include all families, calibration/evaluation splits, and methods; per-sign metrics and all 88 contrasts are retained. Reported hidden-state application errors are instrumentation checks, not an independent model replay. Raw logits remain in the associated remote run archive; the compact companion includes derived records and their hashes, not weights or every raw array. Numerical audit does not itself imply a positive scientific result.

V DETAILED TABLES FOR THE SOURCE AND EXTERNAL-COLLATERAL DIAGNOSTICS

The tables below retain the exact source/site and external-collateral summaries cited in Section 5. They are separate from the supplementary conditional-Fisher experiment and use their original materials and decision rules.

Model	DiM R_{post}	DiM R_{front}	DiM I	Lexical I
Pythia-410M	-0.088	1.546	**1.634**	**1.373**
Qwen2.5-0.5B	1.007	1.961	**0.954**	**0.787**

Table 27: Equal-dose diagnostic means in nats; 12 shared lexical pairs, not 12 models. Pythia is a development analysis; Qwen prospectively tests a fixed prediction on exposed stimuli. Each model has independently source-fitted directions. The order interaction replicates, but the negative postposed contrast does not. Full method/dose data and all lexical points are in Appendix M.

Endpoint	Mean	Threshold	Prespecified bound decision
Target gain	2.125	≥ 0.5	**Pass**
Conditional non-target KL	0.04288	≤ 0.02	**Not certified**
Absolute lexical ΔNLL	0.10809	$\leq \log(1.05)$	**Not certified**

Table 28: Reference-dose DiM on 16 held-out head lemmas; all quantities are in nats. Decisions use prespecified one-sided bounds, not the displayed means alone. Lower-dose measurements are retained in Appendix P.2.