FULL-PATH GRADIENTS FOR MECHANISTIC DATA ATTRIBUTION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic data attribution (MDA) links interpretable units in language models to influential training documents, enabling data-based interventions on a chosen mechanism. Existing methods often restrict attribution to selected parameters. The derivative paths required to predict the resulting update remain underexplored. Later attention heads and multilayer perceptrons transmit changes in a target head's output even when their weights are fixed. We show how these downstream derivatives affect predictions of head-only updates. Under local smoothness and a fixed update direction, full-path prediction has error quadratic in the step size, whereas omitting downstream derivatives can introduce a first-order error. A selective-head backward pass matching released MDA code preserves forward values while blocking derivatives through most downstream attention heads. To compare these paths under a common update construction, we introduce Full-Path Target-Response Matching (FP-TRM), which weights document losses so that one descent step is predicted to increase the target head's contribution. We define this contribution as the observed next-token logit drop under target-head ablation, with a fixed upstream head ablated in both evaluations. On Pythia-1B, full-path predictions correlate with measured contribution changes at 0.99 versus 0.35 for selective predictions at one target, where the two paths also select nearly disjoint document sets. Across two targets, full-path gradients yield larger mean contribution changes on held-out text in every evaluated condition. FP-TRM also produces positive mean contribution changes at the evaluated targets in Pythia, OLMo-2 and Qwen3.5 models spanning 31M to 4B parameters. These results show that selecting data to modify a mechanism benefits from retaining the downstream computations that transmit its effects.

1 INTRODUCTION

Mechanistic data attribution (MDA) links interpretable units in language models to influential training documents (Chen et al., 2026). These links make it possible to study how data interventions change internal computations. For example, removing or augmenting documents can delay or accelerate induction heads that copy recurring patterns (Chen et al., 2026). At an existing checkpoint, this motivates a concrete intervention problem: which documents should receive weight in a head-only update intended to increase a chosen head's contribution?

Answering this question requires connecting a document-based update to a measurable mechanism response. Training data attribution estimates how training examples influence model outputs (Koh & Liang, 2017). Scalable approaches use curvature approximations (Grosse et al., 2023), projected gradient features (Park et al., 2023), or comparisons restricted to selected layers and parameter groups (Pruthi et al., 2020; Li et al., 2026). MDA applies parameter restriction to an interpretable unit (Chen et al., 2026). For measuring such units, activation patching replaces internal activations and measures the resulting output changes (Wang et al., 2023; Zhang & Nanda, 2023; Heimersheim & Nanda, 2024). AttentionInfluence, AIR and CircuitSeer also use head or circuit signals to guide data selection (Hua et al., 2025; Liu et al., 2026; Wang et al., 2025).

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.

Parameter restriction and derivative propagation serve different roles in this intervention. The former specifies which weights change. The latter determines how those changes reach the measured output. Later attention heads and multilayer perceptrons continue to read the target head's output even when their own weights remain fixed (Elhage et al., 2021). MDA's parameter-subspace objective admits complete differentiation through these components. A selective-head backward pass matching released MDA code, however, preserves forward values while blocking derivatives through most downstream attention heads (Chen et al., 2026). Consequently, identical forward computations can produce different predictions of which documents will increase the target head's contribution.

We analyze this difference for one head-only update at a fixed checkpoint. Contribution is the drop in the observed next-token logit when the target head is ablated. A specified upstream head remains ablated in both evaluations, fixing the context in which the target contributes. For a fixed update direction and a locally smooth contribution, the full-path linear prediction has an error quadratic in the step size. Omitting downstream derivatives can add an error proportional to the step size (Section 3). This analysis identifies a source of response-prediction error that persists even when the forward values agree.

To translate the contribution objective into a data intervention, we introduce Full-Path Target-Response Matching (FP-TRM). Target-gradient similarity and weighted gradient matching provide established tools for data selection (Xia et al., 2024; Killamsetty et al., 2021). We fit weighted document-loss gradients to the negative gradient of mean head contribution on fitting text, so descent on the weighted loss is predicted to increase that contribution. The same construction accepts either derivative path, allowing matched comparisons that hold the checkpoint, target, candidate documents and update magnitude fixed (Section 4). This checkpoint-level intervention follows the further-training perspective of final-model-only attribution (Wei et al., 2025).

The experiments connect the prediction error to the resulting document choices. At one Pythia-1B target, full-path predictions correlate with measured contribution changes above 0.99, compared with approximately 0.35 for selective predictions (Table 11), and the paths select nearly disjoint document sets. Across two targets, full-path gradients yield larger mean contribution changes on held-out text in every evaluated condition, with comparable time and memory in our implementation (Section 5.3). Across the evaluated targets in Pythia, OLMo-2 and Qwen3.5 models spanning 31M to 4B parameters, FP-TRM produces positive mean contribution changes (Section 5.5).

Our contributions are:

- We separate the parameter subspace being updated from the downstream derivative paths needed to predict its effect.
- We characterize the fixed-action prediction error, distinguishing the quadratic full-path remainder from the possible first-order error introduced by truncation (Proposition 1).
- We introduce FP-TRM as a common document-update construction and show how retaining downstream derivatives changes document selection and increases the measured head response.

2 Related Work

Mechanism-targeted attribution. Mechanistic interpretability localizes computations through induction-head analysis of how attention copies continuations of repeated token patterns (Olsson et al., 2022) and activation interventions (Wang et al., 2023) that measure output changes after replacing or removing internal states (Vig et al., 2020; Geiger et al., 2021; Vaidyanathan et al., 2026). Automated circuit discovery extends this analysis (Conmy et al., 2023; Syed et al., 2024; Kramár et al., 2024), while MDA links interpretable units to training documents (Chen et al., 2026). Concept Influence uses semantic or sparse-autoencoder features (Kowal et al., 2026), with related attribution targets including symbolic policies (Habibi et al., 2026) and persona features (Vetter et al., 2026). However, describing a circuit or attributing a feature leaves unspecified which downstream derivatives predict its response to a parameter update. Probe controls measure selectivity (Hewitt & Liang, 2019), while circuit benchmarks assess behavioral faithfulness (Hanna et al., 2024; Mueller et al., 2025). We complement these descriptions with document-based target-head updates that retain downstream derivatives and evaluate the resulting change in the head's ablation-defined contribution.

Influence estimation in language models. Influence estimation links data to output predictions such as observed-token likelihood through inverse-curvature estimates (Koh & Liang, 2017) with scalable EK-FAC approximations (Grosse et al., 2023) or checkpoint gradient comparisons (Pruthi et al., 2020) for language data influence (Yeh et al., 2022). Refinements examine curvature reliability (Li et al., 2025) and parameter weighting (Li et al., 2026), while SOURCE approximates training updates (Bae et al., 2024) and dattri benchmarks attribution estimates (Deng et al., 2024). ExPLAIND studies training trajectories (Eichin et al., 2026) and step-resolved attribution separates recurrent steps (Kaissis et al., 2026), whereas Dual Path Attribution follows downstream computation for input and component explanations (Jantsch et al., 2026). These approaches primarily explain output behavior or existing computations rather than defining a document update for a specified head contribution. We instead differentiate the head contribution through downstream components while restricting the updated parameter coordinates.

Gradient and circuit-guided data selection. Data selection approximates training gradients with coresets that reproduce a larger corpus's gradient using small weighted subsets (Mirzasoleiman et al., 2020) or selects documents through LESS target-gradient similarity (Xia et al., 2024) and GRAD-MATCH subset fitting (Killamsetty et al., 2021). Extensions constrain minibatches (Agrawal et al., 2026) or account for optimizer weighting (Wang et al., 2026), with group interactions (Yu et al., 2025) and diversity (Zhang et al., 2025a) guiding selection and subset perturbations recovering data effects (Dagli et al., 2026). AttentionInfluence (Hua et al., 2025) and AIR (Liu et al., 2026) use head signals, whereas CircuitSeer probes reasoning circuits (Wang et al., 2025). These signals guide data selection without specifying the derivative path from an updated head to its contribution. Sparse interpretability methods use matching pursuit with learned (Costa et al., 2025) or vocabulary (Basile et al., 2025) dictionaries. We adapt gradient matching to that contribution and compare target-subspace GRAD-MATCH/OMP, which selects documents by orthogonal matching pursuit (Tropp & Gilbert, 2007), under the same document-based update endpoint.

Behavior correction. Behavior correction changes model outputs through parameter editing (Meng et al., 2022) or LANCET's influence ranking and Bregman optimization (Zhang et al., 2025b). For factual edits, causal-tracing strength need not predict editing success (Hase et al., 2023). MDA additionally tests document removal and augmentation during mechanism development (Chen et al., 2026), while contrastive activation addition steers generation using representation differences (Rimsky et al., 2024). These interventions act on training dynamics or desired outputs rather than the effect of a document-weighted parameter update on a specified head's contribution. We instead select documents using full-path gradients of a carrier-conditioned head contribution and update only that head.

3 Problem and Path-Truncation Failure

We study how weighted documents can increase a target head's contribution through one update at a fixed checkpoint. We first define the intervention and its response, then isolate the prediction error caused by removing downstream derivatives.

3.1 Mechanism-targeted data actions

Let f_θ be a pretrained autoregressive language model and $\theta_{\mathcal{U}} \in \mathbb{R}^p$ the p selected parameters of a target head u. The subspace contains its query, key, value and output projection coordinates, abbreviated QKVO, with the architecture-specific mapping in Appendix G.4. This mapping also covers the grouped value heads evaluated in linear attention. All other weights remain fixed.

For an event $e = (x, y)$ consisting of a text prefix and its observed next token, let $z_y(x)$ be the token's unnormalized logit. We measure contribution by ablating the target output while keeping a specified upstream head c, called the carrier, ablated in both evaluations:

$$C_e(\theta_{\mathcal{U}}) = z_y(x \mid \mathrm{do}(h_c = 0)) - z_y(x \mid \mathrm{do}(h_c = 0, h_u = 0)). \tag{1}$$

Here h_c, h_u are head outputs and $\mathrm{do}(h = 0)$ sets an activation to zero. Positive C_e means that retaining the target raises the observed-token logit. The fixed carrier gives all candidate updates the same contribution context.

For a document-based direction $v \in \mathbb{R}^p$ and step size $\eta > 0$, define the response to one update as

$$R_e(v) = C_e(\theta_{\mathcal{U}} - \eta v) - C_e(\theta_{\mathcal{U}}). \tag{2}$$

Thus positive R_e denotes increased contribution, and the parameter displacement has norm $\eta \|v\|_2$. For candidate document d_i, the loss $\ell(d_i, \theta)$ sums next-token cross-entropies over its token transitions. Its gradient is $a_i = \nabla_{\theta_{\mathcal{U}}} \ell(d_i, \theta)$, and the contribution gradient is $g_e = \nabla_{\theta_{\mathcal{U}}} C_e$. Selecting K document indices S and weights $w \in \mathbb{R}^{|S|}$ gives $v = \sum_{i \in S} w_i a_i$, the gradient of the weighted document loss. Document selection and weighting therefore determine the update direction.

3.2 Parameter locality and derivative paths

To expose the downstream dependence, collect the residual states after the target layer in the two ablation evaluations into $h = H(\theta_{\mathcal{U}}, x) \in \mathbb{R}^d$. Let $F(h, x) \in \mathbb{R}^2$ return their observed-token logits and $b = (1, -1)^\top$ take their difference. The chain rule then gives

$$g_e = J_H^\top J_F^\top b, \qquad J_H \in \mathbb{R}^{d \times p}, \quad J_F \in \mathbb{R}^{2 \times d}.$$

Fixed downstream weights still transmit changes in h through J_F. The selective backward control preserves forward values and replaces downstream attention derivatives by a restricted operator $\widetilde{J}_F$. It therefore predicts with $\widetilde{g}_e = J_H^\top \widetilde{J}_F^\top b$, omitting $\delta_e = g_e - \widetilde{g}_e$. MLPs, residual connections and LayerNorm remain differentiable in this control (Appendix G.4).

We first hold v fixed to separate prediction error from changes in document selection. Define error as the measured response minus its linear prediction, using $-\eta\langle g_e, v\rangle$ or $-\eta\langle \widetilde{g}_e, v\rangle$ for the two paths.

Proposition 1 (fixed-action path error)**.** *Assume C_e has an L-Lipschitz gradient on the segment $[\theta_{\mathcal{U}} - \eta v, \theta_{\mathcal{U}}]$. For any fixed direction v,*

$$R_e(v) = -\eta \langle g_e, v\rangle + r_e(v), \qquad |r_e(v)| \le \frac{L\eta^2}{2} \|v\|_2^2 . \tag{3}$$

The full-path prediction error is $r_e(v)$. The selective prediction error is $r_e(v) - \eta \langle \delta_e, v\rangle$.

The omitted projection can introduce first-order error in the step size, whereas the full-path remainder is quadratic (proof in Appendix C.1). Both document gradients a_i and contribution gradients g_e depend on the derivative path. We next let each path construct its own action to measure how these changes affect document selection and the deployed response.

4 Full-Path Target-Response Matching

Full-Path Target-Response Matching (FP-TRM) takes candidate documents and a contribution objective and returns a document support, its weights and a target-only update. It computes gradient predictions, searches for a sparse weighted direction, and scales the selected direction to a prescribed displacement.

4.1 A full-path mechanism predictor

Data roles. We fit an action separately for each corpus, called an environment, and choose one coefficient family across the environments of a route. A route groups events by newline status and uses a common carrier. Within each environment, fit events E_f determine supports and weights, while validation events E_v choose between the completed coefficient families. Target discovery and confirmation precede this construction. Final held-out events measure the deployed response. Candidate documents and these event partitions have disjoint identities, with their sizes specified in Appendix G.4.

Gradient predictions. We retain the full computation graph while recording gradients only in $\theta_{\mathcal{U}}$. After extraction, fitting requires only gradient inner products. Stack the N document gradients as rows of $A \in \mathbb{R}^{N \times p}$ and form $G = AA^\top$. For event e, document descent $-a_i$ has linear response per unit step size $q_{e,i} = -\langle g_e, a_i\rangle$. Positive $q_{e,i}$ predicts increased contribution. The fitting target is $t = -\bar{g}$, where $\bar{g} = |E_f|^{-1} \sum_{e \in E_f} g_e$. Consequently, $m_i = \langle a_i, t\rangle$ is the mean of $q_{e,i}$ over E_f.

Let $Q \in \mathbb{R}^{|E| \times N}$ contain the unweighted event–document scores for either fit or validation events. To compare nonzero directions per unit parameter displacement, use

$$\widehat{\boldsymbol{r}}(S, w) = \frac{Q_{:,S} w}{\sqrt{w^\top G_{SS} w}}. \tag{4}$$

The numerator is the weighted eventwise prediction and the denominator is $\left\| A_S^\top w \right\|_2 = \|v\|_2$. Thus each output entry predicts the contribution change per unit displacement. For a fixed v, predictions from the two contribution-gradient arrays also give the omitted projection in Proposition 1 directly.

4.2 Sparse actions and target-only deployment

Fit coefficients, then compare supports. For each support S of size K, we minimize $\left\| A_S^\top w - t \right\|_2^2$, following gradient matching (Killamsetty et al., 2021) with a contribution-gradient target. The nonnegative fit uses NNLS, while the signed fit uses a Moore–Penrose pseudoinverse. Positive coefficients supply document-loss descent directions and negative coefficients reverse them. At a common support, the signed linear span contains the nonnegative cone and attains no larger unnormalized fitting residual (Appendix C.2).

We then compare the fitted supports by their mean normalized prediction on E_f in Eq. (4). Initializations include inner-product and norm-normalized alignments, followed by single-document swaps that refit the coefficients and improve this objective. The best fitted support across starts supplies one action per coefficient family and environment. Zero-norm fits are ineligible for this comparison. Appendix G.1 gives the deterministic search, numerical conventions and initialization scores.

Choose a coefficient family. We choose signed actions throughout a route when every environment's signed fit has all coefficients nonzero with both signs represented and strictly larger mean predictions than its nonnegative fit on both E_f and E_v. Otherwise the route uses its nonnegative actions. Validation changes this family choice without refitting supports or weights. Algorithm 1 summarizes these environment-level fits and the route-level decision.

Deploy at a common displacement. For the selected $v = A_S^\top w$, apply

$$\Delta\theta_{\mathcal{U}} = -\rho \frac{\|\theta_{\mathcal{U}}\|_2}{\|v\|_2} v. \tag{5}$$

Here ρ is the update norm relative to the target-parameter norm. For actions on the same target, fixing ρ holds their displacement magnitude constant. This corresponds to $\eta = \rho \|\theta_{\mathcal{U}}\|_2 / \|v\|_2$ in Eq. (2), so the linear prediction is $\rho \|\theta_{\mathcal{U}}\|_2 \widehat{\boldsymbol{r}}(S, w)$. The principal comparisons use full-rank updates with $K = 10$ and $\rho = 3 \times 10^{-4}$.

The optional rank study retains leading singular factors of each logical QKVO block and renormalizes the resulting direction to the same dose. The corresponding linear prediction uses the compressed direction, which can differ from $A_S^\top w$. Appendix G.1 specifies this construction and Appendix A illustrates it. Applying the same fitting and deployment rules to full-path and selective-path gradients provides the matched path comparison.

5 Experiments

We first test the path-error mechanism with matched Pythia-1B interventions, then compare document-selection methods and examine their behavior across model settings and deployment choices. These experiments distinguish prediction fidelity, contribution improvement and output prediction quality.

5.1 Shared evaluation setup

Models and documents. Table 1 separates the roles of the evaluated checkpoints. Pythia (Biderman et al., 2023) provides matched path controls and training-seed variation. OLMo-2 (OLMo Team et al., 2025) and Qwen3.5 (Qwen Team, 2026) extend the evaluation to other softmax and gated linear-attention parameterizations. Each checkpoint has independently discovered and confirmed targets and carriers. Principal comparisons use 256 candidate documents of 128 tokens, selecting $K = 10$ for full-rank deployment at $\rho = 3 \times 10^{-4}$.

Table 1: **Models, targets and experimental roles.** Layer/head indices are zero-based. Qwen candidates are external documents evaluated at the available checkpoint.

Checkpoint	Target	Candidate source	Principal role
Pythia-31M	L5H6	Pile	Dense method comparisons
Pythia-410M	Per seed	Pile	Seed and pool variation
Pythia-1B	L2H4, L7H4	Pile	Matched derivative paths
OLMo-2 1B	L8H12	OLMo-mix	Softmax attention transfer
Qwen3.5-2B	L22H5	External OLMo-mix	Linear attention and replay
Qwen3.5-4B	L30H0	External OLMo-mix	Larger model and deployment

Pythia candidates come from its public Pile training source (Gao et al., 2021), and OLMo-2 candidates from its versioned training release. The Qwen candidates are identity-disjoint continuations from OLMo-mix. Pythia evaluation environments are C4 (Raffel et al., 2020), FineWeb (Penedo et al., 2024) and WikiText-103 (Merity et al., 2016). Other models use OLMo-mix Wiki and ArXiv text plus OpenWebMath (Paster et al., 2023). Each has three non-newline (NN) and two newline (NL) corpus conditions.

Data flow. Target discovery and confirmation precede action fitting and validation, which precede final response measurement. Candidate and curvature documents are also identity-disjoint from these event partitions. The carrier must rank highest on both fit and validation text within its route. Table 19 gives partition sizes and NN/NL response budgets, including the expanded Pythia-31M and Qwen3.5-4B comparisons.

Comparators and their roles. The path control changes only the backward computation while fixing the checkpoint, target, carrier, pool, solver, coefficient rule, rank and dose. Target-subspace GRAD-MATCH/OMP (Killamsetty et al., 2021) instead changes support construction at the same Gram matrix, matching target and ten-document budget. For alternative selection signals, MDA's QKVO EK-FAC influence scores (Chen et al., 2026; Grosse et al., 2023) use 64 curvature sequences and a query without carrier ablation. Its highest-ranked documents receive equal positive weights and the common normalized dose. GradDot and GradSim (Xia et al., 2024) compare raw and norm-normalized alignment, while AIR (Liu et al., 2026), AttentionInfluence (Hua et al., 2025) and CircuitSeer (Wang et al., 2025) supply head- and circuit-based scores. BM25 (Robertson & Zaragoza, 2009), token frequency, reliability and random controls test simpler selection signals. Appendix B specifies these adaptations.

5.2 Response measurements and statistical summaries

We measure contribution change R_e from the same restored checkpoint for each action. Positive paired differences favor FP-TRM. Cosine similarity measures gradient or update-direction agreement, and support Jaccard measures selected-set intersection divided by union. Output log-probability changes separately assess next-token prediction. Mean next-token loss over the final 65 tokens must increase by less than 1%, a criterion satisfied by all reported action cells (Appendix G.4).

Paired event means and sample standard errors describe responses on shared held-out text. One-sided 95% normal bounds use Bonferroni adjustment within the families in Table 16. They quantify event-level uncertainty conditional on the selected actions. For comparison-table summaries, Δ weights fixed comparison entries equally, and $d_{z,w}$ standardizes the resulting difference using within- and between-condition variation (Appendix G.2). Comparison counts describe this fixed set of actions and conditions, not independent model replications.

5.3 Full-path gradients increase the matched response

The matched Pythia-1B comparison tests whether the omitted derivatives in Proposition 1 change useful document directions. Figure 1a shows reduced directional agreement between the two paths' L2H4 document gradients. This changes the selected documents: the three non-newline supports are disjoint (panel b). The resulting full-path actions yield larger mean contribution changes in all ten target–corpus–route conditions, with positive simultaneous lower bounds (panel c and Table 12).

To separate prediction error from action selection, we also hold each deployed full-path action fixed and predict its response using either contribution gradient. Table 2 shows that the full-path

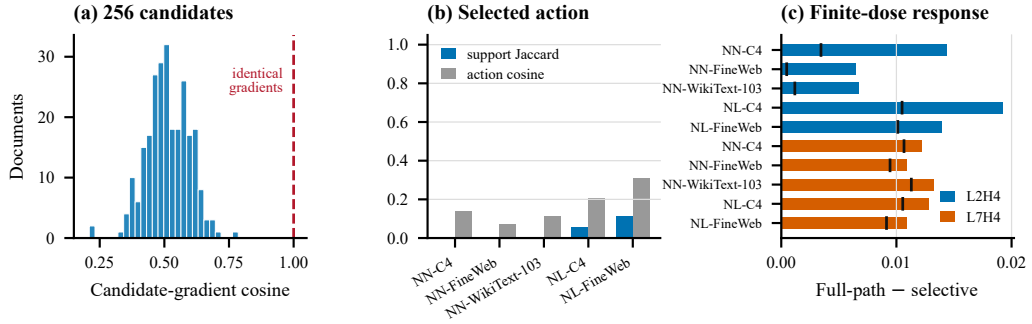


Figure 1: **Path changes propagate from gradients to actions and responses.** (a) Candidate-gradient cosine and (b) support Jaccard/update cosine at Pythia-1B L2H4. (c) Colored bars give full-path minus selective response means. Black ticks mark simultaneous lower bounds for family sizes 25 (L2H4, blue) and 5 (L7H4, orange). NN/NL denote non-newline/newline conditions.

Table 2: **Two distinct effects of derivative paths.** Upper rows compare paths over 256 documents and 640 fit/validation events per target. Norm ratios are selective/full. Lower rows hold deployed full-path actions fixed and correlate their predictions with measured R_e over 1,024 held-out events. Details and bootstrap intervals appear in Table 11.

Quantity	L2H4	L7H4
Mean document-gradient norm ratio	0.425	0.559
Mean eventwise candidate-score norm ratio	0.134	0.234
Mean correlation between candidate scores from the two paths	0.206	0.352
Prediction–response correlation, full path	0.990	> 0.999
Prediction–response correlation, selective path	0.350	0.733

prediction tracks the measured response at both targets. At L2H4, its correlation is 0.99, compared with 0.35 for the selective prediction. The candidate-score disagreement and weaker fixed-action prediction therefore connect the removed derivatives to both selection and response error. The two paths have comparable measured costs: selective/full backward times range from 0.98 to 1.03 in this implementation and workload (Table 21).

5.4 Contribution matching and alternative selection methods

We next ask whether contribution matching improves on other document-selection constructions at a common deployment endpoint. Table 3 separates signal adaptations from OMP support search. All MDA comparisons have positive simultaneous lower bounds, with checkpoint-level mean advantages ranging from 0.0019 to 0.017 logits. The common dose removes update magnitude as an explanation, while query context, curvature weighting and support construction distinguish the methods. Thus these comparisons support the complete contribution-matching construction across the evaluated adaptations.

The broader families cross fixed comparators with corpus–route conditions: 25 variants over five Pythia-31M conditions, 20 AIR comparisons, and 12 comparators over five Qwen3.5-4B conditions. Their positive point differences and adjusted-bound counts appear separately in Table 3. The variant family includes MDA, and both Qwen budgets reuse the same construction, so these rows share actions and evidence.

Support search has a different outcome. OMP shares the matching target and Gram matrix with FP-TRM, and neither solver dominates across settings. At Pythia-1B non-newline WikiText-103, signed OMP gives a larger response with a negative upper bound for the paired difference. Both environment-conditioned GradSim variants also outperform FP-TRM in this condition (Appendix F). These outcomes concern support construction within full-path geometry, distinct from the matched derivative-path effect. The Pythia-31M OMP row includes ten entries but nine distinct comparisons because signed and nonnegative OMP coincide on newline C4. Its three differences below the

Table 3: **Paired comparisons at a common endpoint.** n counts fixed comparison entries. Point counts are positive/zero/negative means, and bound counts are positive lower/negative upper bounds. Δ is in 10^{-3} logits. Row intervals appear in Table 14, and adjustment families in Table 16. † marks three raw differences smaller than the cross-model numerical reference scale.

Comparator setting	n	Δ	$d_{z,w}$	Points $+/0/-$	Bounds $+/-$
Selection-signal adaptations and expanded comparator families					
MDA, Pythia-31M	5	1.89	1.074	5/0/0	5/0
MDA, Pythia-1B	5	16.9	0.461	5/0/0	5/0
MDA, OLMo-2 1B	5	6.63	0.703	5/0/0	5/0
MDA, Qwen3.5-2B	5	5.53	1.108	5/0/0	5/0
AIR-score adaptations	20	17.7	0.629	20/0/0	18/0
Pythia-31M variants	125	2.23	1.024	125/0/0	125/0
Qwen3.5-4B, 256/128 NN/NL	60	2.73	0.933	60/0/0	58/0
Qwen3.5-4B, 768/384 NN/NL	60	2.53	0.763	60/0/0	59/0
Support search at a common Gram matrix and matching target					
GRAD-MATCH/OMP, Pythia-31M†	10	0.0242	0.156	8/0/2	3/0
GRAD-MATCH/OMP, Pythia-1B	10	−0.0875	−0.014	3/3/4	1/1

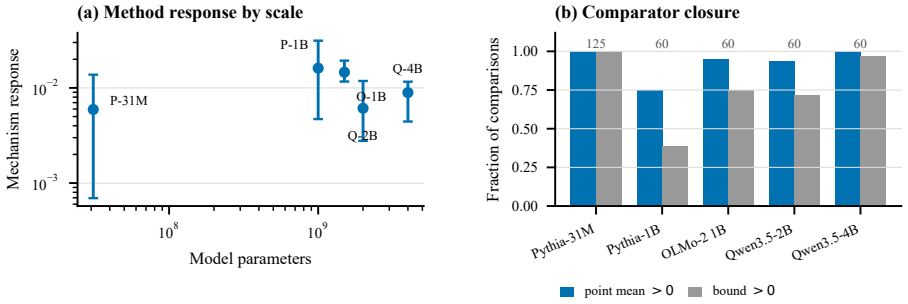


Figure 2: **Responses across independently selected model settings.** (a) Equal-condition means and minimum–maximum ranges on logarithmic axes. (b) Fractions of positive paired means (blue) and adjusted lower bounds (grey), with comparison counts. Qwen3.5-4B uses the principal budget. Targets and environments differ between checkpoints.

6.19×10^{-7} Pythia-1B batching reference retain their raw signs. That cross-model scale is a sensitivity reference, with no matching Pythia-31M batching measurement (Appendix E.2).

5.5 Positive responses across model settings

The cross-model study asks whether positive contribution changes extend to independently selected targets and architectures. Figure 2a shows positive mean responses in every principal condition, and panel b summarizes paired advantages over the comparison methods. Each checkpoint uses its own target, carriers and environments, making this a test of operating settings. For scale, a hypothetical 0.01-logit change corresponds to approximately 2–15% of the median absolute pre-update contribution (Table 17).

Training seeds. Of four screened Pythia-410M seeds, three satisfy the common-carrier criterion. Their 15 condition means are positive, and the five seed-averaged responses have positive joint lower bounds (Table 27). The result extends positive responses across the qualifying training realizations.

Candidate pools. To vary available document combinations, we hold the target, carriers, ten-document budget, dose and evaluation events fixed while enlarging nested pools. Table 4 shows that more conditions have positive response lower bounds as the pool expands. This is consistent with additional useful gradient combinations becoming available. The table reports threshold counts, which do not establish monotonic response magnitudes.

Table 4: **Nested candidate pools at a fixed Pythia-410M target.** Counts refer to five corpus–route conditions. Bounds use a joint Bonferroni family of 195 statistics across pool sizes. The last column counts positive means in 60 paired comparisons.

Candidates	Positive responses	Positive bounds	Positive paired differences
256	5/5	2/5	55/60
1,024	5/5	4/5	59/60
4,096	5/5	5/5	58/60

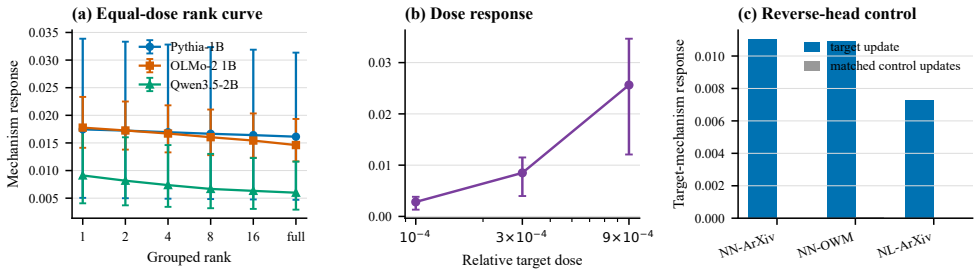


Figure 3: **Deployment behavior.** (a) Equal-condition responses at independently renormalized block ranks. (b) Qwen3.5-4B responses over a ninefold dose range. (c) The same target is measured after updating either that head or three matched control heads. Bars compare target-head response with mean absolute target response under the control updates. Error bars give ranges across fixed corpus–route conditions.

5.6 Deployment and output behavior

We first test whether signed fitting translates into a larger measured response. All five signed-minus-nonnegative bounds on Qwen3.5-4B are positive (Table 24, final column). We then vary deployment rank and dose, and the head receiving the update. At a common dose, independently renormalized low-rank directions retain positive responses, with the smallest tested rank performing best (Figure 3a). This supports compact directions at the evaluated ranks. Stored-coordinate counts appear in Table 28.

Across doses 10^{-4} to 9×10^{-4}, each threefold increase yields approximately threefold response growth, within 1.2% of linear scaling (panel b and Table 29). Panel c instead holds the measured target fixed and compares updates to that head with updates to three same-layer controls. The intended-target updates produce larger target responses, supporting the role of the selected update location (Table 26).

Finally, a separate Qwen3.5-2B replay study recomputes document gradients over three updates to test output effects beyond a single stored direction. At the higher cumulative dose, contribution increases coexist with lower observed-token log probabilities in every environment (Table 8). The output improvement therefore requires direct evaluation even when the internal contribution improves.

6 Conclusion

Predicting the effect of a local update to an attention head depends on derivatives through the layers that read its output, and a selective backward pass matching released attribution code removes most of them while leaving forward values intact. We measured what that removal costs, using Full-Path Target-Response Matching to build document actions from either gradient. The blocked paths carry most of the document-selection signal, and restoring them changes the selected actions and raises their measured contribution responses at comparable measured computational cost in our implementation. Evaluations across model families connect positive contribution changes to the document-selection objective at a common update endpoint. These findings make the downstream derivative path an explicit part of mechanism-targeted data selection. Our evaluation concerns normalized target-only updates of a single head under a specified carrier context, so extending the construction to interacting head groups and optimizer-dependent training trajectories, and connecting predicted internal responses to the effects of longer interventions through replay or retraining, remain open.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

The evaluated checkpoints and text corpora are public. Section 5 describes the shared evaluation. Appendix G.4 identifies the targets and carriers. It also specifies the candidate and event partitions. The solver settings and document budget are given there. The same appendix reports the dose and rank schedule. Appendix G.2 defines the comparison families and their Bonferroni adjustments. The supplementary file `cell_statistics.csv` contains per-condition estimates and their event counts. Its result-file names and keys identify the underlying statistics. Appendix J describes the scope of numerical reconstruction from stored results.

REFERENCES

Prayas Agrawal, Prateek Chanda, Ishita Khatri, Ganesh Ramakrishnan, Bamdev Mishra, and Pratik Jawanpuria. Minibatch selection for language models via partition matroid constrained gradient matching. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training data attribution via approximate unrolling. In *38th Conference on Neural Information Processing Systems*, 2024.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head Pursuit: Probing Attention Specialization in Multimodal Transformers, 2025. URL `https://arxiv.org/abs/2510.21518`.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023. URL `https://proceedings.mlr.press/v202/biderman23a.html`.

Jianhui Chen, Yuzhang Luo, and Liangming Pan. Mechanistic data attribution: Tracing the training origins of interpretable LLM units. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. Evaluating Sparse Autoencoders: From Shallow Design to Matching Pursuit, 2025. URL `https://arxiv.org/abs/2506.05239`.

Rishit Dagli, Abir Harrasse, Luke Zhang, Florent Draye, Amirali Abdullah, Bernhard Schölkopf, and Zhijing Jin. STRIDE: Training data attribution via sparse recovery from subset perturbations. *arXiv preprint arXiv:2606.05165*, 2026. URL `https://arxiv.org/abs/2606.05165`.

Junwei Deng, Ting-Wei Li, Shiyuan Zhang, Shixuan Liu, Yijun Pan, Hao Huang, Xinhe Wang, Pingbang Hu, Xingjian Zhang, and Jiaqi W. Ma. dattri: A library for efficient data attribution. In *38th Conference on Neural Information Processing Systems, Track on Datasets and Benchmarks*, 2024.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Florian Eichin, Yupei Du, Philipp Mondorf, Maria Matveev, Barbara Plank, and Michael A. Hedderich. ExPLAIND: Unifying model, data, and training attribution to study model behavior. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

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

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks, 2021. URL https://arxiv.org/abs/2106.02997.

Roger Grosse et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023. URL https://arxiv.org/abs/2308.03296.

Reza Habibi, Darian Lee, and Magy Seif El-Nasr. Symbolic mechanistic data attribution: Tracing training influence to learned behavioral policies. *arXiv preprint arXiv:2606.29171*, 2026. URL https://arxiv.org/abs/2606.29171.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms, 2024. URL https://arxiv.org/abs/2403.17806.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models, 2023. URL https://arxiv.org/abs/2301.04213.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

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 (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/d19-1275. URL https://doi.org/10.18653/v1/d19-1275.

Kai Hua, Steven Wu, Ge Zhang, and Ke Shen. Attentioninfluence: Adopting attention head influence for weak-to-strong pretraining data selection. *arXiv preprint arXiv:2505.07293*, 2025. URL https://arxiv.org/abs/2505.07293.

Lasse Marten Jantsch, Dong-Jae Koh, Seonghyeon Lee, and Young-Kyoon Suh. Dual path attribution: Efficient attribution for SwiGLU-transformers through layer-wise target propagation. *arXiv preprint arXiv:2603.19742*, 2026. URL https://arxiv.org/abs/2603.19742.

Georgios Kaissis, David Mildenberger, Juan Felipe Gomez, Martin J. Menten, and Eleni Triantafillou. Step-resolved data attribution for looped transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Matthew Kowal, Goncalo Paulo, Louis Jaburi, Tom Tseng, Lev E. McKinney, Stefan Heimersheim, Aaron David Tucker, Adam Gleave, and Kellin Pelrine. Concept influence: Leveraging interpretability to improve performance and efficiency in training data attribution. *arXiv preprint arXiv:2602.14869*, 2026. URL https://arxiv.org/abs/2602.14869.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Learning to weight parameters for training data attribution. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EhUkQp9Yah.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. Do influence functions work on large language models? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14367–14382, 2025. URL https://aclanthology.org/2025.findings-emnlp.775/.

Jinrui Liu, Kai Hua, Xuanguang Pan, Ge Zhang, Yong Wang, Shuai Ma, and Chongyang Tao. AIR: Post-training data selection for reasoning via attention head influence. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960, 2020.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

OLMo Team et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Keiran Paster et al. OpenWebMath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023. URL https://arxiv.org/abs/2310.06786.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Qwen Team. Qwen3.5-4B-Base model card. Hugging Face model documentation, 2026. URL https://huggingface.co/Qwen/Qwen3.5-4B-Base.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://doi.org/10.18653/v1/2024.blackboxnlp-1.25.

Joel A. Tropp and Anna C. Gilbert. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. doi: 10.1109/tit.2007.909108. URL https://doi.org/10.1109/tit.2007.909108.

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/2606.27510.

Clemens Vetter, David Kaczér, Lucie Flek, and Florian Mai. Data attribution of emergent misalignment with persona features. *arXiv preprint arXiv:2608.11025*, 2026. URL https://arxiv.org/abs/2608.11025. Accepted at BlackboxNLP 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Fangxin Wang, Peyman Baghershahi, Langzhou He, Henry Peng Zou, Sourav Medya, and Philip S. Yu. Filter-then-weight: Online data selection and reweighting for LLM fine-tuning. *arXiv preprint arXiv:2604.00001*, 2026. URL https://arxiv.org/abs/2604.00001.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Shaobo Wang, Yongliang Miao, Yuancheng Liu, Qianli Ma, Ning Liao, and Linfeng Zhang. Circuitseer: Mining high-quality data by probing mathematical reasoning circuits in LLMs. *arXiv preprint arXiv:2510.18470*, 2025. URL https://arxiv.org/abs/2510.18470.

Dennis Wei, Inkit Padhi, Soumya Ghosh, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy, and Maria Chang. Final-model-only data attribution with a unifying view of gradient-based methods. In *Advances in Neural Information Processing Systems*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, and Pradeep Ravikumar. First is better than last for language data influence. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32285–32298, 2022.

Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Scott Yih, and Chenyan Xiong. Group-level data selection for efficient pretraining. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Chi Zhang et al. Harnessing diversity for important data selection in pretraining large language models. In *International Conference on Learning Representations*, 2025a.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods, 2023. URL https://arxiv.org/abs/2309.16042.

Han Zhang, Zhuo Zhang, Yi Zhang, Yuanzhao Zhai, Hanyang Peng, and Yu Lei. Correcting large language model behavior via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14477–14485, 2025b. doi: 10.1609/aaai.v39i13.33586.

A Design schematic and checkpoint table

This section supports the target definitions in Section 3 and the model and deployment comparisons in Section 5. The checkpoint table identifies the evaluated targets. The diagram distinguishes the derivative paths summarized by the deployment results of Section 5.6 and Figure 3.

Table 5: Principal checkpoints and intervention handles. The last column gives the largest per-condition held-out event count reached at each checkpoint. For Qwen3.5-4B that entry is the expanded non-newline budget, while its principal comparisons use 256 non-newline and 128 newline events. Qwen documents are controlled continuation examples, not claimed Qwen pretraining records.

Checkpoint	Target	Candidate role	Largest causal sample
Pythia-31M	L5H6	Pile train	2,560
Pythia-410M	per-seed	Pile train	256
Pythia-1B	L2H4 (+L7H4)	Pile train	256
OLMo-2 1B	L8H12	OLMo-mix train	256
Qwen3.5-2B	L22H5	controlled continuation	256
Qwen3.5-4B	L30H0	controlled continuation	768

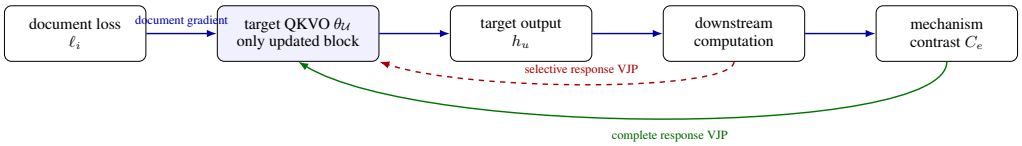


Figure 4: **Parameter locality and full derivative paths are independent.** FP-TRM changes only $\theta_{\mathcal{U}}$ while retaining every downstream vector–Jacobian product from the target output to the mechanism contrast. The dashed route illustrates the matched selective-backpropagation control. Forward activations are preserved in both cases.

B Comparator identity

This section specifies the comparison methods used in Sections 5 and 5.4. The adaptations share the document-based target update endpoint.

Table 6: Identity of the closest comparisons. "Adapted" means that a published signal is mapped to our common $K = 10$ target-subspace action endpoint.

Comparator	Identity in this paper
MDA	Same-task formal signal and fixed-action adaptation. Selective-head code behavior is a separate implementation control.
GRAD-MATCH	Nonnegative and signed OMP adapted from training-subset selection to the same target Gram matrix, K, and dose.
LESS family	GradDot and GradSim actions adapted to the same candidates and target subspace.
AIR	Retrieval-head score adapted to the same target action and causal events.
Our method	Full-path response gradients, validation-selected action family, and target-only finite-dose intervention.

C Additional Mathematical Details

This section supports the path-error analysis in Section 3 and the normalized gradient matching in Section 4.

C.1 Proof of Proposition 1

This subsection gives the fixed-action argument underlying Proposition 1 and its response interpretation.

Proposition 1 (fixed-action path error). Assume C_e has an L-Lipschitz gradient on the segment $[\theta_{\mathcal{U}} - \eta v, \theta_{\mathcal{U}}]$. For any fixed direction v,

$$R_e(v) = -\eta \langle g_e, v \rangle + r_e(v), \qquad |r_e(v)| \le \frac{L\eta^2}{2} \|v\|_2^2 ,$$

where L is the gradient Lipschitz constant on the stated segment and $r_e(v)$ is the Taylor remainder. The linear term predicts the response and the bound controls its deviation from the actual response. Taking error as measured response minus prediction, the selective error is $r_e(v) - \eta \langle \delta_e, v \rangle$.

Proof. Apply the descent lemma to $C_e(\theta_{\mathcal{U}} - \eta v)$. Substituting $g_e = \widetilde{g}_e + \delta_e$ gives the second statement. □

C.2 Fixed-support signed containment

This subsection supports the fixed-support fitting comparison in Section 4.

Let $\mathcal{C}_S = \{A_S^\top w : w \geq 0\}$ and $\mathcal{L}_S = \{A_S^\top w : w \in \mathbb{R}^{|S|}\}$. Since $\mathcal{C}_S \subseteq \mathcal{L}_S$, the signed fit attains no larger Euclidean residual to t on this fixed support. This statement compares the raw NNLS and pseudoinverse solutions. For storage, nonnegative coefficients are divided by their sum and signed coefficients by their absolute sum. This positive rescaling preserves Eq. (4) and the fixed-dose update, while the unnormalized least-squares residual pertains to the raw solution. Separately searched supports can have different finite-step response orderings.

C.3 Why equal dose is required

Equation (5) implements Eq. (2) with $\eta = \rho \left\|\theta_{\mathcal{U}}\right\|_2 / \left\|v\right\|_2$.

This subsection explains the normalization used in Eq. (5).

The normalized predicted response in Eq. (4) is invariant to positive rescaling of w. The causal update should retain the same invariance. Equation (5) fixes the update norm relative to the current target block and makes actions comparable even when their candidate-gradient norms differ. Rank-truncated actions are renormalized after truncation. Without this step, a rank curve would combine directional change with a smaller update norm.

D Qwen Intervention-Faithfulness Results

This section supports the comparison of contribution and output objectives in Section 5.6. It reports direct output measurements and document-replay actions at the same target endpoint.

D.1 Contribution and output measurements

Contribution and output measurements are reported together with the document-replay comparison below, separating the mechanism response from output behavior (Section 5.6).

To relate contribution changes to output prediction, we evaluate the observed token's log probability after the Pythia-1B actions. Three of five mean changes are positive, and one has a positive familywise lower bound. The contribution in Eq. (1) measures the effect of ablating a head under a fixed carrier context. The output metric instead evaluates the observed token directly. Thus these measurements distinguish contribution improvement from the output quantity used to assess prediction quality.

We further compare output changes when the selected documents are used in the Qwen3.5-2B replay experiment. At the primary dose, FP-TRM ranks third among 19 actions by mean correct-token log-probability change. At the higher dose it ranks eighteenth, with increased internal contributions and decreased output log probabilities in each environment (Table 8). For FP-TRM, replay executes documents chosen for contribution matching. The output evaluation instead ranks actions by their observed-token log probabilities. Direct output measurements therefore determine action choice when prediction quality is the intended outcome. In the matched derivative-path experiments, full-path contribution matching increases measured target-contribution responses.

This separate boundary experiment executes the selected document supports rather than applying a stored parameter vector. We partition each $K = 10$ support into three identity-ordered groups

S_1, S_2, S_3. At step t, we recompute

$$v_t = \sum_{i \in S_t} w_i \nabla_{\theta_{\mathcal{U}}} \ell(d_i, \theta_t),$$
$$\theta_{\mathcal{U},t+1} = \theta_{\mathcal{U},t} - \frac{\rho}{3} \frac{\|\theta_{\mathcal{U},0}\|_2}{\|v_t\|_2} v_t. \tag{6}$$

Candidate documents, behavior-calibration events, and the 768 evaluation events are identity-disjoint. All 19 actions share the support size and replay order. They also share the target-parameter scope and cumulative dose.

Table 7: Complete mechanism-ranking decision results for Qwen3.5-2B. "Supported" means that the realized best action minus the mechanism-predicted action has a two-sided 95% source-bootstrap interval above zero.

Dose	Environment	Mechanism-predicted top	ρ_S	τ_b	Regret	Realized top / supported
3×10^{-4}	Wiki	env. signed GradDot	$-.3656$	$-.2455$	.000727	repetition density / no
	ArXiv	env. signed GradDot	$-.1062$	$-.0941$	.001575	Wiki GradDot / yes
	OpenWebMath	env. signed GradDot	.0167	.0118	.002522	pooled GradSim / yes
9×10^{-4}	Wiki	env. signed GradDot	.1137	.0778	.001614	pooled GradDot / no
	ArXiv	env. signed GradDot	$-.2467$	$-.1529$	.002940	behavior signed GradSim / yes
	OpenWebMath	env. signed GradDot	.1185	.0941	.001367	pooled GradSim / no

Table 8: All 19 Qwen3.5-2B replay actions. Values are correct-token log-probability changes in 10^{-3}, averaged equally across Wiki, ArXiv and OpenWebMath (256 events each). Ranks use unrounded means with tied actions sharing rank. Output-targeted actions use output-gradient calibration, and contribution-targeted actions use the mechanism objective.

	$\rho = 3 \times 10^{-4}$		$\rho = 9 \times 10^{-4}$	
Action	Mean	Rank	Mean	Rank
Pooled GradSim	0.614	1	0.888	2
Output-targeted GradDot	0.428	2	0.197	5
FP-TRM	0.351	3	-0.849	18
Output-targeted GradSim	0.142	4	-0.894	19
Environment GradSim	0.079	5	0.131	8
Random equal weights	-0.022	6	-0.157	10
Environment signed GradSim	-0.072	7	-0.290	13
Environment GradDot	-0.111	8	-0.755	16
Output-targeted OMP	-0.131	9	-0.398	15
Repetition density	-0.146	10	-0.248	12
Wiki lower-bound search	-0.155	11	0.171	6
Wiki GradSim	-0.155	11	0.171	6
Adapted clean MDA	-0.270	13	0.124	9
Pooled signed GradSim	-0.279	14	-0.191	11
Pooled GradDot	-0.308	15	0.214	4
Wiki GradDot	-0.309	16	-0.821	17
Output-targeted signed GradSim	-0.351	17	0.971	1
Pooled signed GradDot	-0.364	18	0.254	3
Environment signed GradDot	-0.521	19	-0.364	14

The primary-dose point estimate places FP-TRM third, while the higher dose places it eighteenth. Table 9 separates these descriptive ranks from uncertainty in output gains. None of the six one-sided lower bounds is positive.

E Matched Path Geometry

This section supports the matched derivative-path comparison in Section 5.3. It gives the document and action geometry behind Figure 1 and the paired responses for both targets.

Across all 256 candidate documents, the path-gradient cosine has mean 0.513956 and standard deviation 0.082321. Its median is 0.510133. Its minimum and maximum are 0.209575 and 0.785851. The 5th and 95th percentiles are 0.378812 and 0.631884. These values describe the gradient geometry. The paired finite-dose responses in Table 12 measure the causal effect of the path change.

Table 9: Output-gain uncertainty for replay. One-sided 95% source-bootstrap lower bounds are in 10^{-3} log-probability units (20,000 draws). Holm adjustment applies to the three output claims within each dose.

Dose	Output contrast	Lower bound	Holm p
3×10^{-4}	FP-TRM absolute gain	-0.597	0.545
3×10^{-4}	FP-TRM minus MDA	-0.303	0.409
3×10^{-4}	FP-TRM minus pooled GradSim	-1.277	0.663
9×10^{-4}	FP-TRM absolute gain	-1.893	1.000
9×10^{-4}	FP-TRM minus MDA	-1.985	1.000
9×10^{-4}	FP-TRM minus pooled GradSim	-2.739	1.000

Table 10: Geometry of the Pythia-1B L2H4 full-path and selective-backprop actions.

Route	Environment	Jaccard	Cosine
non-newline	C4	0.0000	0.1376
non-newline	FineWeb	0.0000	0.0715
non-newline	WikiText-103	0.0000	0.1129
newline	C4	0.0526	0.2037
newline	FineWeb	0.1111	0.3079

Predicted-response gap between the two paths. Table 11 compares the stored attribution matrices of the two backward passes at both Pythia-1B targets. Each matrix holds the Gram matrix of the 256 candidate gradients and the per-event predicted responses $q_{e,i} = -\langle g_e, a_i \rangle$ for the five corpus–route conditions on their fit and validation splits, giving 640 events per target. The candidate norm ratio is $\sqrt{G^{\text{sel}}_{ii}/G^{\text{full}}_{ii}}$ averaged over documents, and the score norm ratio is $\left\|q^{\text{sel}}_{e,:}\right\|_2 / \left\|q^{\text{full}}_{e,:}\right\|_2$. The score correlation, rank correlation, score norm ratio and overlap of the ten highest-scoring documents are computed per event and then averaged. The two runs of each pair share the same candidate documents in the same order, which their identical stored candidate losses confirm, and the same fit and validation partitions. Chance overlap for ten documents drawn from 256 is 0.039. The deeper L7H4 target, which has five fewer layers reading its output, shows the smaller gap on every measure. Percentile bootstrap intervals from $20{,}000$ replicates lie within 0.012 of each reported mean, resampling documents for the candidate norm ratio and events within each condition for the other rows.

Table 11: Full-path against selective-backprop predictions at the two Pythia-1B targets. Ratios are selective over full path. The lower block compares each prediction with the measured response of the deployed update. Correlation intervals use 10,000 event bootstrap resamples; extra precision distinguishes values near one.

Quantity	L2H4	L7H4
Candidate norm ratio	0.4248	0.5593
Score correlation	0.2063	0.3522
Rank correlation	0.1967	0.3387
Score norm ratio	0.1343	0.2336
Top-ten overlap	0.1156	0.1644
Response correlation, full path	0.9904	0.9999
95% bootstrap interval, full path	[0.9839, 0.9957]	[0.999933, 0.999955]
Response correlation, selective path	0.3495	0.7334
95% bootstrap interval, selective path	[0.2687, 0.4305]	[0.7003, 0.7642]
Response MAE ratio	10.56	93.35

The rank correlation is the Spearman coefficient, and its per-condition means range from 0.149 to 0.240 at L2H4 and from 0.302 to 0.375 at L7H4. At the level of the fit-event mean vector that defines the matching target, the two paths correlate between 0.103 and 0.363 at L2H4 and between 0.337 and 0.585 at L7H4, with norm ratios between 0.131 and 0.202 and between 0.190 and 0.274. The candidate norm ratio at L2H4 has median 0.4160 with extremes 0.2327 and 0.8403, and at L7H4 median 0.5528 with extremes 0.3663 and 0.8224. The ratio is below one for all 256 candidates at both targets and below 0.5 for 221 of them at L2H4 and 45 at L7H4.

Calibration against the measured responses. For each condition we take the deployed full-path action and recompute the target-head contribution gradient on the unmodified checkpoint along both derivative paths. The two linear terms $-\eta \langle g_e, v\rangle$ and $-\eta \langle \widetilde{g}_e, v\rangle$ then predict the same measured response R_e and differ only in that path. The recomputed contributions agree with the stored values to a relative error below 2×10^{-5}, and the selective pass leaves every forward logit unchanged. For the 1,024 pooled events at each target, the correlation intervals in Table 11 are separated between paths. The full-path term also has the smaller mean absolute error in all ten conditions, by factors between 7.8 and 17.4 at L2H4 and between 58.0 and 124.6 at L7H4. The omitted projection $\langle \delta_e, v\rangle$ has a positive mean in all ten conditions and is positive for 71.6% of the L2H4 events and 95.3% of the L7H4 events. Across the pooled events the selective prediction vector has 0.2048 and 0.2779 times the norm of the full-path prediction vector.

E.1 Complete two-target finite-dose results

These paired responses support the two-target comparison in Section 5.3.

Table 12: All full-rank, matched full-path minus selective-backprop responses. The two targets use separate Bonferroni families. SE is the paired standard error and d_z is the paired mean divided by its sample standard deviation. At L7H4 non-newline C4 the full-path and selective means are 0.023569 and 0.011375.

Route	Environment	Events	Mean	SE	Lower bound	Upper bound	d_z
Pythia-1B L2H4							
non-newline	C4	256	0.014350	0.003787	0.003450	0.025250	0.237
non-newline	FineWeb	256	0.006497	0.002093	0.000472	0.012522	0.194
non-newline	WikiText-103	256	0.006759	0.001939	0.001178	0.012340	0.218
newline	C4	128	0.019245	0.003042	0.010491	0.027999	0.559
newline	FineWeb	128	0.013906	0.001314	0.010124	0.017688	0.935
Pythia-1B L7H4							
non-newline	C4	256	0.012194	0.000660	0.010659	0.013730	1.155
non-newline	FineWeb	256	0.010940	0.000642	0.009446	0.012433	1.065
non-newline	WikiText-103	256	0.013245	0.000838	0.011295	0.015195	0.988
newline	C4	128	0.012795	0.000965	0.010550	0.015041	1.172
newline	FineWeb	128	0.010930	0.000769	0.009140	0.012719	1.256

E.2 Source-cluster sensitivity and numerical tolerances

This subsection supports the uncertainty interpretation in Section 5.2 and the comparisons in Section 5.4.

Each event in the ten matched path conditions has a distinct source-unit identifier. The event count therefore equals the source-unit count and each source contributes one event. Source resampling is consequently equivalent to an event-level nonparametric bootstrap. This analysis concerns repeated-source and normal-approximation sensitivity. It retains the original effective sample size and assumes independence across source units. We resample source units with replacement 50,000 times per condition. We use NumPy PCG64 with seed 930093 and its linear quantile rule. The lower percentiles are $0.05/25$ for L2H4 and $0.05/5$ for L7H4. These choices preserve the primary families' one-sided tail probabilities. All ten sensitivity bounds in Table 13 remain positive.

Table 13: Source-cluster bootstrap sensitivity for full-path minus selective-backprop responses. Every event is its own source cluster.

Target	Route	Environment	Events	Source units	Bootstrap lower bound
L2H4	non-newline	C4	256	256	0.004476
L2H4	non-newline	FineWeb	256	256	0.000552
L2H4	non-newline	WikiText-103	256	256	0.001027
L2H4	newline	C4	128	128	0.012196
L2H4	newline	FineWeb	128	128	0.010400
L7H4	non-newline	C4	256	256	0.010662
L7H4	non-newline	FineWeb	256	256	0.009463
L7H4	non-newline	WikiText-103	256	256	0.011324
L7H4	newline	C4	128	128	0.010607
L7H4	newline	FineWeb	128	128	0.009157

Numerical tolerances of the reported responses. L7H4 uses event batch size one and satisfies the contribution-response, loss and numerical criteria at that setting. The batch-size-two comparison holds the target and actions fixed. It also holds the events and dose fixed under the same statistical

family. Its empty-mask consistency error is 5.72×10^{-6} and its batched-loss error is 1.43×10^{-6}. Both exceed the 10^{-6} tolerance. The contribution-response and loss criteria remain satisfied. The largest change in an environment mean is 6.19×10^{-7}. We use this Pythia-1B change as a conservative reference scale for a resolvable paired margin. Pythia-31M responses are two orders of magnitude smaller and have no corresponding batching measurement. The Pythia-1B comparisons of Section 5 have empty-mask and batched-loss errors of 7.63×10^{-6} and 1.43×10^{-6}. Their action-norm and off-target errors are 0. Restoration error is also 0 and relative dose error is 1.47×10^{-7}.

Reproduction comprises three levels. Stored identities specify the saved targets and actions together with the events and responses. A separate program reconstructs all five means and their errors. It also reconstructs the Bonferroni bounds and ten action norms. The maximum statistical discrepancy is zero and the largest norm discrepancy is 7.11×10^{-15}. The third level regenerates model gradients and interventions. The reported reconstruction covers the stored quantities. Independent Pythia-1B reconstruction on CPU reproduces every stored record and loss summary with zero maximum error. It therefore covers the identity and numerical-reconstruction levels for this checkpoint. Its intervention-side numerical differences are the tolerance exceedances reported above. The contribution-response results remain unchanged.

Qwen3.5-4B event budgets. The twelve-comparator Qwen3.5-4B family uses two event budgets. The principal budget has 256 non-newline and 128 newline events per condition. The larger budget has 768 and 384 events respectively. Both use the same checkpoint weights and candidate matrix as reported in Section 5.4. They estimate the same quantity from different event budgets. The 60 conditions correspond one to one and all their paired means differ. Changes in condition means have median 8.2% and mean 10.7%. The maximum change is 33.1%. The largest change occurs for non-newline OpenWebMath against pooled signed GradSim, from 2.21714×10^{-4} to 1.48254×10^{-4}. This condition has the only nonpositive simultaneous lower bound at the larger budget. The principal and larger budgets yield 58 and 59 positive bounds out of 60 respectively.

F Attribution and Gradient-Matching Comparisons

This section expands the OMP comparisons in Section 5.4 and Table 3.

Table 14 gives the two means behind each row of Table 3 together with an interval for their difference. Row means weight the original comparison entries equally, and each entry first averages the comparator responses of a shared event. The listed intervals are two-sided and per row, so they describe the current event sample of that row rather than the simultaneous families of Table 16.

Table 14: Method and comparator means behind the rows of Table 3, in logits. Δ is their mean paired difference. The interval is a two-sided 95% percentile interval from 20,000 resamples of source units with NumPy PCG64 and seed 20260921, drawing actions that share a source unit together.

Comparator setting	FP-TRM mean	Comparator mean	Δ	95% interval
MDA (EK-FAC), Pythia-31M	5.935×10^{-3}	4.046×10^{-3}	1.889×10^{-3}	$[1.841, 1.936] \times 10^{-3}$
MDA (EK-FAC), Pythia-1B	1.613×10^{-2}	-7.508×10^{-4}	1.688×10^{-2}	$[1.464, 1.939] \times 10^{-2}$
MDA (EK-FAC), OLMo-2 1B	1.462×10^{-2}	7.989×10^{-3}	6.627×10^{-3}	$[6.038, 7.222] \times 10^{-3}$
MDA (EK-FAC), Qwen3.5-2B	6.107×10^{-3}	5.738×10^{-4}	5.533×10^{-3}	$[5.264, 5.801] \times 10^{-3}$
GRAD-MATCH/OMP, Pythia-31M	5.935×10^{-3}	5.910×10^{-3}	2.422×10^{-5}	$[1.978, 2.859] \times 10^{-5}$
GRAD-MATCH/OMP, Pythia-1B	1.613×10^{-2}	1.622×10^{-2}	-8.749×10^{-5}	$[-4.560, 2.727] \times 10^{-4}$
AIR-score adaptation	1.537×10^{-2}	-2.287×10^{-3}	1.766×10^{-2}	$[1.644, 1.899] \times 10^{-2}$
Matched variants, Pythia-31M	5.935×10^{-3}	3.705×10^{-3}	2.230×10^{-3}	$[2.195, 2.266] \times 10^{-3}$
Qwen3.5-4B, 256/128 events	8.922×10^{-3}	6.190×10^{-3}	2.732×10^{-3}	$[2.575, 2.891] \times 10^{-3}$
Qwen3.5-4B, 768/384 events	8.681×10^{-3}	6.153×10^{-3}	2.528×10^{-3}	$[2.420, 2.630] \times 10^{-3}$

Among the Pythia-31M entries, the signed FineWeb NN and signed WikiText-103 NN differences, and the nonnegative FineWeb NL difference, have absolute values below 6.19×10^{-7}. This reference is the largest mean change in the Pythia-1B batching check, not a measured resolution for Pythia-31M. Their raw signs remain included in Table 3. The signed and nonnegative OMP actions coincide on Pythia-31M newline C4, so its two family entries correspond to one distinct comparison.

Three comparisons are exact ties because the selected method and OMP actions are identical under the relevant family. Four intervals have a positive lower bound for FP-TRM. One interval has a

Table 15: All 20 target-subspace GRAD-MATCH/OMP comparisons, with the paired event count n and standard error of each cell. Values are FP-TRM minus OMP under the common 25-comparison one-sided Bonferroni family. "NN" and "S" denote nonnegative and signed OMP. "NN = S" marks identical actions from the two solvers, represented by one distinct comparison.

Route	Environment	OMP	n	Mean	SE	Lower bound	Upper bound
Pythia-1B							
non-newline	C4	NN	256	0	0	0	0
non-newline	C4	S	256	-1.0221×10^{-3}	8.622×10^{-4}	-3.5037×10^{-3}	1.4595×10^{-3}
non-newline	FineWeb	NN	256	-1.2209×10^{-4}	5.175×10^{-4}	-1.6114×10^{-3}	1.3673×10^{-3}
non-newline	FineWeb	S	256	-1.7728×10^{-3}	9.745×10^{-4}	-4.5776×10^{-3}	1.0319×10^{-3}
non-newline	WikiText-103	NN	256	0	0	0	0
non-newline	WikiText-103	S	256	-2.4957×10^{-3}	5.872×10^{-4}	-4.1856×10^{-3}	-8.0571×10^{-4}
newline	C4	NN	128	1.5359×10^{-3}	5.670×10^{-4}	-9.6096×10^{-5}	3.1680×10^{-3}
newline	C4	S	128	4.8582×10^{-4}	3.340×10^{-4}	-4.7538×10^{-4}	1.4470×10^{-3}
newline	FineWeb	NN	128	2.5160×10^{-3}	2.487×10^{-4}	1.8002×10^{-3}	3.2318×10^{-3}
newline	FineWeb	S	128	0	0	0	0
Pythia-31M							
non-newline	C4	NN	2,560	6.0005×10^{-5}	5.983×10^{-6}	4.2786×10^{-5}	7.7225×10^{-5}
non-newline	C4	S	2,560	-5.8786×10^{-6}	3.134×10^{-6}	-1.4900×10^{-5}	3.1426×10^{-6}
non-newline	FineWeb	NN	763	1.3590×10^{-4}	1.587×10^{-5}	9.0212×10^{-5}	1.8159×10^{-4}
non-newline	FineWeb	S	763	2.5877×10^{-7}	1.298×10^{-7}	-1.1489×10^{-7}	6.3243×10^{-7}
non-newline	WikiText-103	NN	1,024	2.9003×10^{-5}	7.826×10^{-6}	6.4769×10^{-6}	5.1529×10^{-5}
non-newline	WikiText-103	S	1,024	-4.5916×10^{-7}	4.837×10^{-6}	-1.4380×10^{-5}	1.3461×10^{-5}
newline	C4	NN = S	768	3.7253×10^{-6}	3.142×10^{-6}	-5.3177×10^{-6}	1.2768×10^{-5}
newline	FineWeb	NN	261	4.0193×10^{-8}	2.862×10^{-7}	-7.8360×10^{-7}	8.6399×10^{-7}
newline	FineWeb	S	261	1.5832×10^{-5}	8.186×10^{-6}	-7.7283×10^{-6}	3.9393×10^{-5}

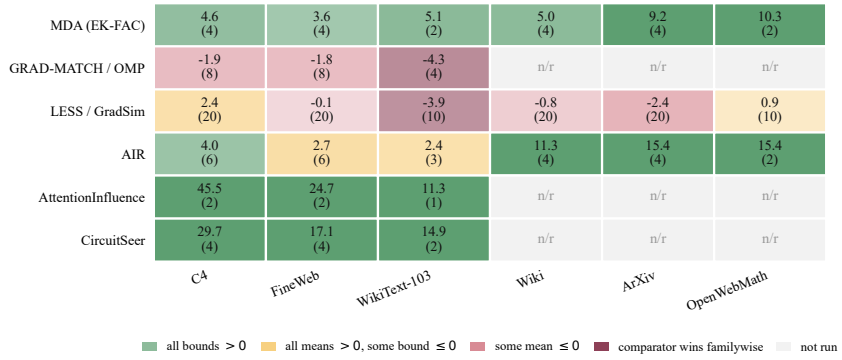


Figure 5: **Contribution-targeted matching improves on the evaluated MDA and AIR scores.** Each square gives the smallest paired mean difference divided by its standard error across measured routes and checkpoints for that comparator–corpus pair. Parentheses give the number of comparisons. Larger values favor FP-TRM. Green denotes positive simultaneous lower bounds throughout, and yellow denotes positive means throughout. Pink marks a nonpositive mean and dark red marks a comparator advantage at its family's critical value. Grey denotes an unevaluated pair. The aggregation retains the least favorable measured comparison for each pair.

negative upper bound and therefore favors signed OMP. The remaining comparisons are unresolved under the common family.

To compare document-selection signals, we evaluate their actions at the same target-subspace endpoint and dose. Figure 5 summarizes the smallest standardized paired difference within each comparator and corpus pair. Every MDA (Chen et al., 2026; Grosse et al., 2023) comparison represented in that figure has a positive simultaneous lower bound. Table 3 separates point-estimate outcomes from lower-bound counts for the other close comparisons. For AIR-score adaptations (Liu et al., 2026), all 20 point differences favor FP-TRM and 18 have positive lower bounds. Matching the contribution gradient directly aligns selection with the quantity subsequently measured after the update. These results support contribution-targeted matching over the evaluated MDA and AIR score adaptations.

G Additional Experimental Detail

This section supports the construction in Section 4 and the shared evaluation in Section 5.

G.1 Candidate-gradient computation and support search

This subsection specifies the implementation of the document-gradient construction and support search in Section 4.

Support search begins with documents ranked by their alignment with the target. GradDot ranks m_i, whereas GradSim ranks $m_i/\sqrt{G_{ii}}$ to remove document gradient magnitude. These scores use the target-gradient similarity principle of LESS (Xia et al., 2024). Appendix G.4 specifies their pooled and signed variants along with the remaining initializations. An optional swap step refines each initial support by replacing one selected document with one unselected document and refitting its coefficients. A swap is accepted only when the mean fit score in Eq. (4) improves by more than 10^{-12}, the numerical gain tolerance. Candidate-index order resolves ties. We stop when no swap exceeds this tolerance and retain the action with the highest fit score across initializations.

The two families can select different supports, so we compare their predicted responses after fitting. We choose the signed family for a route only when every environment satisfies the same conditions. Its coefficients must all be nonzero and contain both signs, and its mean predicted response must strictly exceed the nonnegative action on both fit and validation events. Otherwise we use the separately fitted nonnegative actions throughout that route. Validation selects between completed fits and leaves their supports and coefficients unchanged. Support and coefficient fitting use the gradient predictions before any finite-dose responses are measured.

To reduce the number of coordinates needed to represent an update, we also consider grouped rank reduction. We reshape v into its logical QKVO matrices and apply a singular value decomposition (SVD) to each block. SVD expresses a matrix as a sum of rank-one factors ordered by singular value. We retain the leading factors at the specified rank and renormalize the concatenated blocks to the same joint dose in Eq. (5). This separates the effect of changing direction from that of reducing the update norm. Retaining every factor recovers the original full-rank action. The rank is specified before response measurement, with the evaluated choices listed in Appendix G.4.

Algorithm 1 gives the complete construction. Gradient extraction first computes the $N \times p$ document-gradient cache A and the eventwise contribution gradients. Their inner products produce G with $O(N^2)$ entries and Q_f, Q_v with $O((|E_f| + |E_v|)N)$ entries per environment. Fitting and support search then run on CPU using these arrays without further model evaluations. The selected rows of A reconstruct the deployment direction, so the Gram matrix alone is not the full gradient-storage requirement. The principal $N = 256$ setting uses a 256×256 Gram matrix.

Numerically zero fits receive search score $-\infty$. For NNLS this includes a zero Gram subspace or no coefficient above its numerical positivity tolerance. The signed fit is ineligible when its aggregate squared norm is nonpositive or nonfinite. Response evaluation and deployment require a finite positive aggregate norm. The implementation uses a relative pseudoinverse cutoff of 10^{-10} and counts coefficient signs at 10^{-12}.

Different weighted combinations can have different norms even when they point in the same direction. We compare their predicted effects per unit parameter displacement so that selection is consistent with the normalized deployment. Let Q be the matrix with one row per fit or validation event and entries $q_{e,i}$. For a support S with a nonzero aggregate direction, the normalized linear response vector is

$$\widehat{\boldsymbol{r}}(S, w) = \frac{Q_{:,S} w}{\sqrt{w^\top G_{SS} w}},$$

where $Q_{:,S}$ selects the document columns in S and G_{SS} is the corresponding principal submatrix of G. The numerator sums their weighted eventwise predictions, and the denominator equals the aggregate norm $\|v\|_2$. Thus $\widehat{\boldsymbol{r}}$ records a response per unit displacement for each event. Its mean over E_f is the support-search objective, while its mean over E_v is used for coefficient-family selection. Multiplying it by the deployed update norm gives the linear prediction at the chosen dose.

Table 16: **Simultaneous bounds use experiment-specific comparison families.** For each paired response, the one-sided normal lower bound is its mean minus z times its sample standard error. Bonferroni adjustment gives 95% confidence within the listed family. Larger positive lower bounds indicate stronger support for a positive mean difference.

Comparison family	Family size	Critical value z
L2H4 path and OMP comparisons	25	2.8782
L7H4 path comparisons	5	2.3263
Twelve-comparator checkpoint studies	65	3.1673
Pythia-31M non-newline variants	75	3.2087
Pythia-31M newline variants	50	3.0902
AIR adaptations	20	2.8070

G.2 Comparison families and statistical summaries

This subsection gives the comparison families used for the simultaneous bounds in Section 5.2.

We use Pythia attention heads (Biderman et al., 2023) to compare derivative paths within the same architecture. We use Pythia-31M for the detailed baseline comparison and Pythia-1B for matched comparisons at L2H4 and L7H4. To examine architecture dependence, we evaluate a softmax-attention head in OLMo-2 (OLMo Team et al., 2025) and Gated DeltaNet (Qwen Team, 2026) value heads in Qwen3.5 (Qwen Team, 2026). The principal checkpoints and targets appear in Table 5. We additionally evaluate four public Pythia-410M training seeds to examine sensitivity to pretraining variation. Each checkpoint has its own target and carriers selected on separate discovery and confirmation text. These settings test the same contribution objective across distinct target parameterizations.

We separate newline and non-newline events as defined in Section 4 to test the objective in distinct token regimes. Each route uses the same upstream carrier across its environments. We require that carrier to have the highest score on both fit and validation text, keeping the contribution context consistent across those partitions. Candidate and curvature documents have disjoint identities from the event partitions. Discovery and confirmation events are also separate from action fitting and validation events. Held-out response events are disjoint from all these partitions, so the reported responses measure effects on text outside action construction. Appendix G.4 specifies partition sizes and target-selection criteria.

We measure the paired contribution change R_e in Eq. (2), where positive values mean that the update increases the target contribution. A positive paired difference between methods favors FP-TRM. We apply the target-only update of Eq. (5) at $\rho = 3 \times 10^{-4}$ so that directions are compared at a common relative norm. The normalized linear response vector in Eq. (4) scores directions per unit displacement. Multiplying it by $\rho \|\theta_{\mathcal{U}}\|_2$ gives the prediction for the uncompressed update. We measure the contribution before and after each action and restore the checkpoint between actions. This evaluates each direction against the same initial contribution.

Standardized paired effects. For a single comparison the standardized paired effect is $d_z = \bar{d}/s(d)$, where $\bar{d}$ is the paired mean and $s(d)$ its sample standard deviation with one degree of freedom removed. A row of Table 3 pools several fixed comparisons, so we first average the comparator responses of each event and then weight the resulting condition means $\bar{d}_g$ by their share w_g of the row's comparison entries. Writing $\Delta = \sum_g w_g \bar{d}_g$, the row effect is

$$d_{z,w} = \frac{\Delta}{\sqrt{\sum_g w_g \left[s_g^2 + (\bar{d}_g - \Delta)^2\right]}},$$

which reduces to d_z when the row has one condition. It summarizes the fixed comparator set of that row and not any single comparator.

Contribution magnitude before the update. The responses of Section 5 are changes in a logit, so their interpretation needs the size of the quantity being changed. Table 17 pools the held-out events of each target and reports the distribution of C_e at the unmodified checkpoint. The signed median crosses zero across targets, so we use the median absolute contribution as the magnitude reference

and read 0.01 logit against it. The resulting percentages describe the internal scale of the measured quantity and leave the effect on model outputs to the measurements of Appendix D.

Table 17: Contribution distribution before the update, in logits, pooled over the held-out events of each target. Qwen3.5-4B uses its principal event budget. The last column expresses 0.01 logit as a percentage of the median absolute contribution.

Target	Events	Median C_e	C_e $[Q_{25}, Q_{75}]$	Median $\lvert C_e \rvert$	0.01 as % of median $\lvert C_e \rvert$
Pythia-31M	5,376	-1.278×10^{-4}	$[-0.07344, 0.06366]$	0.06726	14.87
Pythia-1B L2H4	1,024	0.07309	$[-0.09565, 0.3199]$	0.1869	5.351
Pythia-1B L7H4	1,024	0.1029	$[-0.05797, 0.2903]$	0.1842	5.428
OLMo-2 1B	1,024	0.4958	$[0.1108, 1.050]$	0.5767	1.734
Qwen3.5-2B	1,024	0.03621	$[-0.02452, 0.09678]$	0.07059	14.17
Qwen3.5-4B	1,024	0.2501	$[0.1267, 0.4809]$	0.2587	3.865

To explain how paths affect the resulting action, we measure cosine similarity for paired candidate gradients and update directions. Larger cosine values indicate stronger directional agreement. Support Jaccard is the size of the intersection of selected document sets divided by the size of their union. Larger values indicate greater document overlap. We also measure mean next-token loss over the final 65 tokens to quantify changes in language-model performance. The relative-loss criterion permits an increase of at most 1%. For output evaluation, positive observed-token log-probability changes indicate improved next-token prediction. The document-replay comparison executes selected documents in three updates to measure their effects beyond the stored single-step direction (Appendix D). These measurements separate contribution change from document overlap and overall prediction loss.

We use paired natural events as the unit of analysis and report their mean response and sample standard error. One-sided 95% normal lower bounds use Bonferroni adjustment within each experiment's comparison family. A bound above zero supports a positive mean contribution change or paired advantage at that family's level. This section gives the comparison families and their critical values. Intervals describe event-level sampling uncertainty conditional on the selected action and evaluated event set. Appendix E.2 gives the source-resampling sensitivity analysis and numerical details. This pairing measures differences between actions on the same text events.

G.3 Complete action-construction procedure

This algorithm collects the input, fitting and deployment steps of Section 4.

Algorithm 1 FP-TRM action construction

Require: Checkpoint and target $\theta_{\mathcal{U}}$, candidate documents, routes with fixed carriers and corpus environments, disjoint E_f, E_v per environment, document budget K, dose ρ, deployment rank (full by default)

Extract full-path document gradients A and form $G = AA^\top$
for each route **do**
  for each environment in this route **do**
    Compute Q_f, Q_v (event rows, document columns) and $m = \mathrm{mean}_{e \in E_f} Q_f$
    for each coefficient family: nonnegative, signed **do**
      Initialize supports, fit coefficients, and apply strict improving one-document swaps using Q_f, G
      Retain the nonzero action with the largest mean normalized fit prediction across starts
    end for
  end for
  Choose signed for the route if every environment has all coefficients nonzero with both signs represented and strict mean fit and validation gains over its nonnegative action. Otherwise choose nonnegative
  for each selected environment action **do**
    Form $v = A_S^\top w$ and apply the specified block rank if smaller than full
    Normalize the resulting direction to relative dose ρ using Eq. (5)
  end for
end for
return one support, coefficient vector and deployment update per route–environment condition

G.4 ACTION-CONSTRUCTION DETAILS

Support initializations. Table 18 lists the deterministic starts. Here m_i is the mean fit-event prediction and $n_i = \sqrt{G_{ii}}$ is the document-gradient norm. Pooled scores average environment means with equal weight. The reference environment is C4 for Pythia and Wiki for OLMo-2 and Qwen3.5. "Clean MDA" uses the target-head contribution without carrier ablation and EK-FAC curvature fitted on a separate document partition, as detailed below.

Table 18: Support initializations before coefficient fitting. All scores precede finite-step response measurements.

Start	Ranking or construction
Environment alignment	Largest m_i (GradDot) or m_i/n_i (GradSim).
Pooled alignment	Largest equal-environment mean score, with or without document-norm normalization.
Signed alignment	Largest absolute alignment scores, retaining their signs for the initial action.
Clean MDA	Largest $-\langle a_i, \widehat{H}^{-1}_{\mathrm{EKFAC}} g_{\mathrm{clean}} \rangle$, initially with equal positive weights.
Reference lower bound	Strict one-swap ascent on the weakest lower bound for normalized fit-event prediction differences against reference GradDot, GradSim and clean-MDA supports.
Forward cone/span	Add the document maximizing the fitted normalized mean response at each of K steps.

The reference-bound construction uses mean difference minus a one-sided 95% Bonferroni normal critical value times its event-level standard error, across the three reference supports. Its output supplies an initialization to each environment's subsequent mean-response search. Every start receives the strict one-swap refinement in the reported runs, with an improvement tolerance of 10^{-12}. Candidate-index order resolves ties. After fitting, coefficient normalization fixes a representation scale as described in Appendix C.2.

Data partitions and architecture mapping. Table 19 collects the data partitions. Candidate and curvature documents are 128-token sequences, identity-disjoint from the event partitions within a run. OLMo-mix candidates cycle through Wiki, ArXiv and OpenWebMath. Their identities are chosen before finite-step response measurements.

Table 19: Data roles and held-out response budgets. NN/NL are non-newline/newline conditions. Dashes mark a stage shared across routes or an unevaluated condition.

Stage or response study	Unit	NN	NL
Target discovery / confirmation	Events	64 / 128	
Candidate / curvature pool	Documents	256 / 64	
Action fit / validation	Events per environment	64 / 64	
Principal response budget	Events per condition	256	128
Pythia-31M expanded	C4 events	2,560	768
	FineWeb events	763	261
	WikiText-103 events	1,024	–
Qwen3.5-4B expanded	Events per condition	768	384

Table 20 records the model-specific intervention settings. "NN" and "NL" denote non-newline and newline routes. Every row uses ten selected documents and relative dose 3×10^{-4}. Pythia uses its public Pile training release. OLMo-2 uses the Wiki, ArXiv, and OpenWebMath configurations of OLMo-mix-1124. Qwen uses identity-disjoint controlled continuations from the same release because its pretraining corpus is unavailable.

For GPT-NeoX and OLMo-2 attention, the selected coordinates contain the target head's query and key output rows together with its value output rows. They also contain the matching input columns of the output projection. For Qwen3.5 Gated DeltaNet, the selected value-head block uses Q and K matrices of shape $128 \times d_{\mathrm{model}}$. Its V matrix has shape $256 \times d_{\mathrm{model}}$ and its output matrix has shape $d_{\mathrm{model}} \times 256$. Head-specific gates and convolution parameters remain fixed. The decay parameters and shared normalization also remain fixed. The differentiated and deployed coordinates are therefore identical.

Table 20: Model-specific run configuration. Coefficient families are shown as carrier/family. Full denotes the complete target block, which every paired comparison of Section 5.4 deploys. The blockwise rank-1 deployment of Qwen3.5-4B belongs to the rank study of Table 24 and does not enter those comparisons.

Checkpoint	Target	NN carrier/family	NL carrier/family	Deployment	Gradient/intervention precision
Pythia-31M	L5H6	L3H0/signed	L1H4/nonnegative	Full	FP32/FP32
Pythia-1B	L2H4	L0H4/nonnegative	L0H3/signed	Full	FP32/FP32
Pythia-1B	L7H4	L5H4/signed	L5H4/signed	Full	FP32/FP32
OLMo-2 1B	L8H12	L3H4/signed	L3H4/signed	Full	BF16/BF16
Qwen3.5-2B	L22H5	L3H7/signed	L3H7/nonnegative	Full	BF16/FP32
Qwen3.5-4B	L30H0	L27H13/signed	L3H4/signed	Full	BF16/FP32

Selective-head backward pass and its cost. We construct the selective path as a matched control and measure its properties directly. For the evaluated 16-layer, eight-head Pythia-1B model, the control keeps H4 differentiable and detaches the other seven heads in each layer. The control registers a hook at the input of every GPT-NeoX attention output projection, splits the concatenated head dimension, replaces the seven non-selected heads by detached copies of the same values and reconcatenates them. Forward values are therefore preserved exactly. No separate detach is applied to embeddings, output projections, multilayer perceptrons, residual connections or LayerNorm operations, so the control removes the derivative branches through the non-selected heads rather than whole layers. The same wrapper serves the candidate gradients, the contribution gradients and the curvature observations. Table 21 separates isolated backward measurements from one complete attribution run. The isolated benchmark uses 100 synchronized repetitions after ten warmup calls, with peak allocation measured separately. Costs are comparable between paths for this implementation and workload. The selective implementation includes one hook per layer, so the timing comparison includes that overhead.

Table 21: Backward cost on Pythia-1B. Isolated times are mean $\pm$ SD in milliseconds over 100 repetitions. Complete-run time is in seconds for one run per path. Peak allocated memory is in MiB.

Workload	Target	Full time	Selective time	Full peak	Selective peak
Document backward	L2H4	36.82 ± 0.76	37.93 ± 0.26	4235.31	4236.31
Document backward	L7H4	31.6 ± 1.1	32.63 ± 0.43	4190.28	4191.28
Contribution backward	L2H4	60.1 ± 2.8	60.1 ± 3.3	4278.71	4278.72
Contribution backward	L7H4	51.4 ± 3.8	50.6 ± 2.0	4188.65	4188.66
Complete run (seconds)	L7H4	261.6	261.9	6580	6608

The clean-MDA comparator fits QK factors and separate V and O EK-FAC factors on the 64 curvature sequences. Its g_{clean} is the mean raw target-subspace gradient of the target-head contribution over the separate 128-event selection-confirmation partition, taken without carrier ablation as in the probe gradient of MDA. Our own fitting target $t = -\bar{g}$ instead averages carrier-conditioned contribution gradients over the fit events, so the two differ in query context as well as in curvature weighting. The anonymous supplementary file `cell_statistics.csv` contains all 270 per-condition statistics. These cover 125 dense endpoint comparisons and 20 AIR endpoint adaptations. They also cover both 60-condition Qwen3.5-4B event budgets and five correct-token conditions. Each row gives the paired event count and mean. It gives the sample standard deviation and standard error. The row also specifies its one-sided Bonferroni critical value and resulting bound. Its result-file location and key identify the underlying data.

Target and carrier selection. Let $C_e^{(0)}$ denote target-head contribution with no carrier ablation and $C_e^{(c)}$ the contribution under carrier c in Eq. (1). Discovery retains at most eight targets with mean $C_e^{(0)} \geq 0.02$ and positive-event fraction at least 0.55. Confirmation evaluates the nominated target on 128 new events using the same two thresholds. For an upstream head, define its disruption ratio by

$$\kappa(c) = \frac{\sum_e |C_e^{(c)} - C_e^{(0)}|}{\sum_e |C_e^{(0)}|}. \tag{7}$$

A material upstream head has $\kappa(c) \geq 0.25$ for standard attention targets, or $\kappa(c) \geq 0.10$ for the Qwen3.5-2B cross-operator selection. The confirmation criterion requires at least two such heads. For Qwen3.5-4B, confirmation instead retains the two largest finite positive carrier ratios. For each

route we rank these carriers by their minimum environment ratio, then by their equal-environment mean. The same carrier must rank first on the separate 64-event fit and validation splits.

Qwen3.5-4B comparator configurations. Table 22 lists the twelve comparators of Section 5.4. All of them use ten documents and the common normalized dose of Eq. (5). The Wiki lower-bound configuration includes its own one-swap search. We deploy the resulting comparator actions without the optional swap step of FP-TRM. Wiki denotes the Wiki configuration of OLMo-mix and is the reference environment of both routes.

Table 22: The twelve Qwen3.5-4B comparator configurations. m_i^E is the mean fit-event score m_i in environment E of the current route, $n_i = \sqrt{G_{ii}}$, and $\bar{m}_i$ is the equal-weight mean of m_i^E over the route's environments. Signed configurations use the signed score as the weight.

Configuration	Support selection	Weights
`wiki_lcga`	lower-bound one-swap search, Wiki reference	equal
`wiki_grad_dot`	top ten m_i^{Wiki}	equal
`wiki_grad_sim`	top ten m_i^{Wiki}/n_i	equal
`environment_grad_dot`	top ten m_i^E	equal
`environment_grad_sim`	top ten m_i^E/n_i	equal
`adapted_clean_mda`	top ten clean-MDA influence scores	equal
`pooled_grad_dot`	top ten $\bar{m}_i$	equal
`pooled_grad_sim`	top ten $\bar{m}_i/n_i$	equal
`environment_signed_grad_dot`	top ten $\lvert m_i^E\rvert$	m_i^E
`environment_signed_grad_sim`	top ten $\lvert m_i^E/n_i\rvert$	m_i^E/n_i
`pooled_signed_grad_dot`	top ten $\lvert\bar{m}_i\rvert$	$\bar{m}_i$
`pooled_signed_grad_sim`	top ten $\lvert\bar{m}_i/n_i\rvert$	$\bar{m}_i/n_i$

Only the clean-MDA comparator uses curvature. It fits QK, V, and O EK-FAC blocks on 64 sequences of 128 tokens. For eigenvalue element λ, its denominator is $\max\{\lambda + 0.1\,\overline{\lambda}, 10^{-5}\}$. The score is $-\left\langle a_i, \widehat{H}^{-1}_{\text{EKFAC}} g_{\text{clean}} \right\rangle$, and its top ten documents receive equal positive weights. In contrast, FP-TRM's Gram matrix and response dot products use raw gradients. The fitted coefficients and deployed direction also use the raw gradients defined in Section 3.1.

Action families. Each environment starts from clean, environment-specific, and pooled GradDot or GradSim supports. The nonnegative and signed solvers also include their forward greedy support. The optional strict one-swap step accepts only an objective increase. Candidate-index order resolves exact ties. A route falls back to its separately optimized nonnegative actions if any environment fails the signed validation rule.

Rank construction. We reshape the aggregate target gradient into query and key matrices together with value and output matrices. Each logical QKVO matrix is truncated independently. We renormalize the concatenated update to the same joint target-relative dose. The evaluated ranks are 1 and 2 followed by 4 and 8, then 16 and the original full-rank action. The descending response order is an empirical result. We specify the rank before evaluating responses.

Causal controls. The forward specificity comparison applies a target action and measures the designated target plus three same-layer controls. The reverse comparison builds matched actions for the controls, applies each update only to its control QKVO slice, and remeasures the target mechanism. This design separates readout leakage from dependence on the intended target update.

Intervention accuracy. Every action is reconstructed from its stored support and coefficients before intervention. Relative dose error is required to remain below 10^{-5}. Parameters outside the target slice must remain unchanged, and the target slice must return exactly to its starting state after measurement. Mean language-model loss over the final 65 tokens is measured before and after every action. All reported action cells remain below the 1% relative-loss threshold.

Resource use. Candidate gradients are computed in small batches and stored only for the target subspace. Model forwards for causal events are batched according to available memory. The 4B run uses full-precision target interventions and stays below 18 GiB peak allocation. Subsequent optimization and statistical reconstruction operate on the candidate Gram matrix and event-response arrays.

Table 23: Complete rank-curve means. NN and NL denote non-newline and newline routes. The Qwen3.5-2B rows come from a separate rank comparison with 128 events per condition, so their Full column differs slightly from the Qwen3.5-2B cell means of Figure 3a, which use the primary causal partitions of 256 non-newline events per cell.

Checkpoint	Route	Environment	Rank 1	Rank 2	Rank 4	Rank 8	Rank 16	Full
Pythia-1B	NN	C4	.009876	.009801	.009679	.009526	.009425	.009301
Pythia-1B	NN	FineWeb	.005055	.004991	.004902	.004838	.004770	.004706
Pythia-1B	NN	WikiText-103	.009296	.009223	.009030	.008835	.008694	.008529
Pythia-1B	NL	C4	.033868	.033328	.032810	.032347	.031882	.031348
Pythia-1B	NL	FineWeb	.029195	.028824	.028300	.027750	.027304	.026767
OLMo-2 1B	NN	Wiki	.014119	.013791	.013292	.012788	.012326	.011665
OLMo-2 1B	NN	ArXiv	.015637	.015026	.014527	.013872	.013284	.012553
OLMo-2 1B	NN	OpenWebMath	.014575	.014481	.014122	.013623	.013171	.012508
OLMo-2 1B	NL	Wiki	.021197	.020480	.019761	.018856	.018022	.017014
OLMo-2 1B	NL	ArXiv	.023332	.022501	.021807	.021045	.020341	.019341
Qwen3.5-2B	NN	Wiki	.004069	.003705	.003428	.003182	.003056	.002922
Qwen3.5-2B	NN	ArXiv	.005325	.004749	.004200	.003779	.003573	.003387
Qwen3.5-2B	NN	OpenWebMath	.005533	.004966	.004415	.004127	.003935	.003736
Qwen3.5-2B	NL	Wiki	.018226	.016022	.014613	.013018	.012265	.011617
Qwen3.5-2B	NL	ArXiv	.012417	.011383	.010187	.009264	.008791	.008294

Table 24: Complete Qwen3.5-4B rank-1 and signed-family statistics. LB denotes the familywise one-sided 95% lower bound.

Route	Environment	R1 mean	R1 LB	R1-0.5Full mean	LB	Signed mean	Signed LB	Signed$-$NNLS LB
NN	Wiki	.010541	.009137	.006132	.005316	.009329	.008617	.001268
NN	ArXiv	.012772	.011258	.007314	.006438	.011450	.010669	.001539
NN	OpenWebMath	.012947	.011208	.007464	.006460	.010497	.009749	.000927
NL	Wiki	.005097	.004389	.002959	.002549	.004043	.003770	.000107
NL	ArXiv	.008549	.007317	.004864	.004160	.006884	.006496	.000663

Table 25: Complete Qwen3.5-4B dose-family lower bounds. The first three columns correspond to doses 10^{-4}, 3×10^{-4}, and 9×10^{-4}.

Route	Environment	Low	Nominal	High	Nominal$-$low	High$-$nominal
NN	Wiki	.002647	.007957	.024013	.005310	.016056
NN	ArXiv	.003307	.009940	.029984	.006633	.020044
NN	OpenWebMath	.003413	.010257	.030922	.006844	.020665
NL	Wiki	.001202	.003623	.011006	.002420	.007383
NL	ArXiv	.002195	.006602	.019968	.004407	.013366

Table 26: Forward and reverse Qwen3.5-4B target controls. The forward contrast subtracts the mean absolute response of three same-layer controls. For reverse rows, "response" is the target response caused by a matched control-head action, and "contrast" is target-action response minus that value.

Design	Route	Environment/control	Response mean	Response LB	Contrast mean	Contrast LB
Forward	NN	ArXiv	.011199	.009777	.010938	.009558
Forward	NN	OpenWebMath	.009879	.008625	.009649	.008429
Forward	NL	Wiki	.004020	.003507	.003956	.003457
Forward	NL	ArXiv	.006908	.006369	.006773	.006246
Reverse	NN	ArXiv/H1	$-.000008$	$-.000013$	.011050	.009482
Reverse	NN	ArXiv/H8	.000009	$-.000004$	.011034	.009466
Reverse	NN	ArXiv/H15	$-.000012$	$-.000030$	.011054	.009484
Reverse	NN	OpenWebMath/H1	$-.000004$	$-.000008$	.010904	.009301
Reverse	NN	OpenWebMath/H8	$-.000013$	$-.000024$	.010913	.009310
Reverse	NN	OpenWebMath/H15	.000065	.000040	.010835	.009242
Reverse	NL	ArXiv/H1	$-.000005$	$-.000008$	.007258	.006643
Reverse	NL	ArXiv/H8	$-.000003$	$-.000007$	.007255	.006642
Reverse	NL	ArXiv/H15	$-.000009$	$-.000014$	.007262	.006646

H Training Variation and Candidate-Pool Results

This section expands the training-seed and candidate-pool comparisons summarized in Section 5.5.

The seed study screens four public Pythia-410M training realizations using independently selected targets and carriers. Three satisfy the common-carrier criterion on fit and validation text. Table 27 distinguishes screening, per-seed point estimates and seed-averaged uncertainty. Actions use the principal candidate budget, with joint adjustment across response and comparator statistics after averaging seeds.

Table 27: Training-seed evidence on Pythia-410M. Five corpus–route conditions are evaluated for each qualifying seed. The bound result concerns seed-averaged responses.

Quantity	Count
Screened training seeds	4
Seeds satisfying the common-carrier condition	3
Positive per-seed condition means	15/15
Positive joint bounds for seed-averaged responses	5/5

Candidate-pool size. To test whether a larger document pool improves action construction, we vary only the pool while keeping the target and document budget fixed. The carriers and dose are also fixed, as are the evaluation events. Table 4 shows that more method-response lower bounds become positive as the nested pool grows. Every mean response is positive and exceeds clean MDA in the corresponding setting. A larger pool makes additional document-gradient combinations available for matching the contribution target. The larger pools support positive contribution changes in more evaluated conditions.

I Signed Actions and Deployment Results

This section supports the signed-action, rank and dose comparisons in Section 5.6. It also reports the target-specificity measurements.

Signed coefficient family. To evaluate the signed construction, we compare it with the separately fitted nonnegative family at the same candidates and dose on Qwen3.5-4B. All five signed-minus-nonnegative lower bounds are positive, with minimum 1.07×10^{-4} (Table 24). The signed construction can combine opposing document directions when matching the contribution gradient. Since support selection accompanies coefficient fitting, this result supports the complete signed action construction.

Grouped rank reduction (Eckart & Young, 1936). To evaluate compact update representations, we compare the grouped ranks at the same relative norm. The rank curves in Figure 3a have the largest responses at the smallest evaluated rank. For Qwen3.5-4B, both the response at this rank and its difference from half the full-rank response have positive bounds (Table 24). Table 28 gives the coordinate counts for this rank study.

Table 28: Stored update coordinates for Qwen3.5-4B. Each direction is independently normalized to the same relative dose.

Representation	Stored coordinates
Blockwise rank 1 factors	11,008
Full target update	1,966,080

Dose dependence. To assess how update magnitude affects contribution, we evaluate relative doses 10^{-4}, 3×10^{-4} and 9×10^{-4} on Qwen3.5-4B. The direction and events stay fixed across doses. Every absolute response and adjacent-dose difference has a positive adjusted lower bound (Table 25). Table 29 compares response ratios with the linear prediction of threefold growth. The small departures from three are resolved by the paired intervals, while the largest departure remains below 1.2%. Thus dose controls response magnitude approximately linearly over the tested range.

Table 29: Response ratios for adjacent threefold dose increases on Qwen3.5-4B. Brackets give pointwise two-sided 95% paired source-bootstrap intervals from 20,000 draws. Each event has a distinct source identity. Low, nominal and high doses are 10^{-4}, 3×10^{-4} and 9×10^{-4}. Extra decimals resolve the small departures from a ratio of three.

Route	Environment	Events	Nominal/low	High/nominal
NN	Wiki	384	3.0054 [3.0045, 3.0062]	3.0155 [3.0133, 3.0178]
NN	ArXiv	512	3.0053 [3.0048, 3.0058]	3.0154 [3.0141, 3.0166]
NN	OpenWebMath	512	3.0044 [3.0039, 3.0050]	3.0133 [3.0119, 3.0148]
NL	Wiki	384	3.0121 [3.0111, 3.0131]	3.0355 [3.0336, 3.0373]
NL	ArXiv	512	3.0078 [3.0073, 3.0083]	3.0235 [3.0226, 3.0244]

Target specificity. To determine whether the response depends on updating the intended head, we compare target-head updates with matched updates to other same-layer heads. Under the target update, its response exceeds the mean absolute response of three other heads in all four evaluated conditions. Matched updates to those heads produce much smaller target responses, with all nine paired differences having positive lower bounds (Table 26 and Figure 3c). Changing which parameters receive the action therefore changes the response of the designated target contribution. These comparisons connect the measured response to the intended target-head update.

J Scope of the Evaluated Interventions

This section details the intervention scope of the contribution objective in Section 3 and the comparisons in Section 5.

Scope. Our matched Pythia-1B studies cover two fixed layer targets at head index H4 in one checkpoint. Every other architecture evaluates the full construction. MDA's formal parameter-subspace objective permits complete downstream differentiation. The matched finding concerns the selective backward computation. A target head is an intervention handle rather than a complete circuit identity. The normalized target-only update differs from deletion and retraining. Qwen3.5 uses controlled-continuation documents, so these experiments concern checkpoint-level data selection rather than provenance recovery. The correct-token and document-replay results in Section D.1 distinguish mechanism-response improvements from output behavior.

Comparisons and reproducibility. Three comparisons favor their baselines at the method's familywise level. All occur at Pythia-1B non-newline WikiText-103. They concern signed OMP and the two environment-conditioned GradSim variants. The comparative result therefore depends on the evaluated method and setting. Signed and rank-1 results concern the measured finite doses. Positive simultaneous lower bounds occur in 23 of 60 Pythia-1B comparisons. The two Qwen3.5-4B event budgets yield 58 and 59 positive bounds out of 60. The candidate-pool comparison has 79 of 195 positive bounds. The three-seed Pythia-410M family has 48 of 60. Group interactions and optimizer-dependent trajectories remain outside scope. Every matched path condition has at most one event per source identity. The 50,000 bootstrap replicates therefore resample events. Reproduction here reconstructs numerical quantities from stored results. It does not regenerate gradients or interventions.