PREDICTIVE GEOMETRY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

For a given token, a language model's prediction may be more or less robust to changes in the preceding context. We ask how training exposure changes this context robustness. To capture that change, we trace the full next-token distribution as the preceding context grows. We apply Fisher–Rao geometry to this path of predictive distributions, viewing context growth as motion on a statistical manifold. We develop a theory of these paths that measures the total predictive change and decomposes it into the part associated with the observed token's probability and the remaining change in the distribution over alternative tokens. Across four model–data settings, we show that the theoretical decomposition captures systematic effects of repeated exposure: both components of the predictive change increase with exposure. We then use the theory in two ways. For privacy auditing, we use Fisher–Rao context paths to distinguish documents presented during continued training from those omitted from it. For context-scoped editing, we use the geometry to strengthen a continuation after a short context while keeping longer-context predictions near their pre-edit values. It rises from 0.002% to 6–9%, while prediction changes on held-out longer extensions are 41–$58\times$ smaller than without protection. Together, these results show that training exposure leaves a measurable geometric signature in context-dependent predictions that can be used both to detect exposure and to control where its effects appear.

1 INTRODUCTION

Whether a document is included in language-model training, and how often it is repeated, can affect prediction, memorization, and generalization (Kandpal et al., 2022; Carlini et al., 2023; Hernandez et al., 2022). Training exposure, however, does not determine a prediction by itself. At inference time, the model's next-token distribution also depends on the preceding tokens supplied as context.

Consider a document whose training exposure is controlled: during continued training it is omitted, presented once, or presented repeatedly. After training, withhold one of its tokens and query the fixed model with progressively longer portions of the token's preceding text. This produces a context-indexed path of next-token distributions. A full-context evaluation sees only its endpoint, and hence cannot describe how the predictive distribution changes as context accumulates. Throughout, *document exposure* refers to presentations during training, whereas *inference-time context* refers to the preceding tokens supplied to the trained model.

The target token's probability captures only part of this evolution. With more context, mass can move toward the target and be redistributed among alternatives. Because all vocabulary probabilities sum to one, target and alternative probabilities are coupled. Predictions with the same target probability may distribute the remaining mass differently. Conversely, identical redistribution among alternatives can have different geometric effects at different target probabilities.

Analyses of document exposure commonly summarize each prediction through the probability or loss of the observed token, and recent context-dependent membership inference methods follow this scalar view. Existing work does not characterize full-distribution motion across nested contexts.

Our contribution. We develop an information-geometric account of what scalar target probability captures and misses across nested contexts. We test it under randomized document exposure, then study two scoped uses: membership inference as a final-checkpoint readability probe and contextual editing as a locality test. Our main contributions include:

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

- We represent next-token distributions under accumulating inference-time context as paths in the probability simplex, introduce a token-directed projection of the Fisher–Rao displacement between predictions, and derive its exact decomposition into binary target-versus-rest motion and a nonnegative gap that vanishes precisely when the conditional distribution over alternatives is unchanged.
- We test this geometry through randomized continued training, comparing documents that are omitted, presented once, or repeated under a fixed presentation budget. Across four model–data settings, both validation and confirmation evaluations reject the hypothesis that exposure leaves the measured outcomes unchanged, and every observed exposure-response slope is positive. Binary target-versus-rest motion explains most of the retained mean response, while the additional full-distribution gap is smaller but positive.
- We apply the framework to supervised membership inference and use Fisher–Rao context paths to distinguish included from held-out documents using only the final checkpoint. We achieve performance competitive with Min-K%++ (Zhang et al., 2025) in two settings, recover most of the ranking signal from the observed token's likelihood across context lengths, and obtain additional gains from Fisher–Rao features in some settings.
- We use our predictive-geometry framework to define a context-scoped edit of the selected token's output weights, projecting the update to preserve predictions under longer and unrelated contexts. In a case study, we raise the continuation's probability from approximately 0.002% to 6–9% and reduce prediction changes under held-out longer contexts by factors of 41–58 relative to a matched unprotected edit.

Together, these results show that document exposure leaves a context-dependent trace in the predictive distribution. In our experiments, scalar target-probability profiles capture most of the measured response, while full-distribution geometry isolates the additional component and provides a common language for detecting exposure and evaluating localized interventions.

2 Related work

Document inclusion and repetition. Repeated exposure to training examples can increase memorization and extraction (Kandpal et al., 2022; Carlini et al., 2023), and alter generalization (Hernandez et al., 2022). Carlini et al. (2023) also showed that memorization depends on the amount of context available at inference. Studies of counterfactual memorization (Zhang et al., 2023) and randomized training data (Lin et al., 2022) further measure how predictions change when particular examples are included or excluded. However, the mentioned works do not characterize how exposure changes the full next-token distribution across nested contexts, as we study here.

Information geometry. The Fisher metric gives the canonical local geometry of regular statistical models (Rao, 1945; Amari, 2016). For categorical distributions, the coordinate-wise square-root map identifies the Fisher-Rao geometry with spherical geometry (Davis et al., 2024). We apply this geometry to pairs of next-token predictions obtained from nested contexts, introducing a token-directed projection of the Fisher-Rao displacement between two predictions. Relying on classical convexity properties of spherical distance (Afsari et al., 2013; Ferreira et al., 2014), we also derive meaningful decompositions of the displacement.

Membership inference. Membership inference asks whether a record appeared in training data (Shokri et al., 2017). For language models, recent methods extract inference signals beyond average token loss: Min-K%++ (Zhang et al., 2025) calibrates observed-token likelihoods using vocabulary-wide statistics, while Chang et al. (2025) and ReCall (Xie et al., 2024) exploit relative conditional log-likelihoods and perplexity under varying contexts. Closest to us, likelihood-array regression (Sun & Cai, 2026) also uses nested contexts but learns from scalar likelihood-derived features; we analyze trajectory of full categorical next-token distributions.

Model editing. ROME localizes factual associations to feed-forward computations and performs targeted parameter updates (Meng et al., 2022). MEMIT extends this approach to many simultaneous edits (Meng et al., 2023). AlphaEdit projects updates onto a preserving null space to reduce interference with existing knowledge (Fang et al., 2025). We use a related null-space principle for a more restricted purpose: while editing a single output row under exact linear constraints, we ask whether one continuation strengthens under a short trigger while full distributions stay stable elsewhere.

3 PREDICTIVE GEOMETRY OF CONTEXT-DEPENDENT PREDICTION

In this section, we ask how prediction for a fixed target changes as more context is supplied and what the target probability alone reveals. We distinguish variation across inference-time contexts from variation across model states, then compare predictive distributions geometrically to separate movement toward the observed token from redistribution among alternatives. The geometric framework developed here is used in Sec. 4 to measure and decompose how randomized document exposure changes context-dependent prediction, then applied to membership inference from the final checkpoint in Sec. 5.1 and to context-scoped editing in Sec. 5.2. Here, *document exposure* occurs during training; *inference-time context* is supplied after training.

3.1 PREDICTIONS ACROSS CONTEXTS AND MODEL STATES

Let $X = (x_1, \ldots, x_T)$ be a tokenized document over vocabulary $\mathcal{V}$, and let $p_{\boldsymbol{\theta}}$ be a language model with parameters $\boldsymbol{\theta}$. Fix a position $t \in \{2, \ldots, T\}$, with observed token x_t. For a context length $k \in \{1, \ldots, t-1\}$, define the preceding context

$$c_{t,k}(X) = x_{t-k:t-1} = (x_{t-k}, \ldots, x_{t-1}). \tag{1}$$

Increasing k extends the context to previous tokens while keeping the target token fixed.

The model's full next-token distribution under this context is

$$\boldsymbol{P}_{\boldsymbol{\theta},t}(X; k) = \big(p_{\boldsymbol{\theta}}(j \,|\, c_{t,k}(X))\big)_{j \in \mathcal{V}}, \tag{2}$$

and the probability assigned to the observed token is its x_t-coordinate, $\big[\boldsymbol{P}_{\boldsymbol{\theta},t}(X; k)\big]_{x_t} = p_{\boldsymbol{\theta}}(x_t \,|\, c_{t,k}(X))$. As k increases, we provide the model with progressively longer portions of the preceding text while keeping the target position fixed. This produces a sequence of next-token distributions. The observed-token probability need not vary monotonically: additional context may raise or lower it while also redistributing probability among the alternative tokens.

We make two comparisons. For a fixed model, we increase k and track how the next-token distribution changes as more preceding text is supplied; we call this sequence its *context path*. For a fixed k, we compare matched model states before and after a parameter update, thereby measuring the associated change in prediction. We use the first comparison to study context dependence under randomized exposure. In Sec. 5.2, we use the second to assess whether a targeted parameter update strengthens the intended continuation without disturbing predictions under contexts designated for protection.

3.2 FISHER–RAO DISPLACEMENT AND TOKEN-DIRECTED PROJECTION

We now compare a shorter- and a longer-context prediction. Our goal is to separate their total geometric displacement into motion toward or away from the observed token and motion that redistributes probability among the alternatives. Fix the model $\boldsymbol{\theta}$, document X, and target position t. For two context lengths $k_1 < k_2$, define

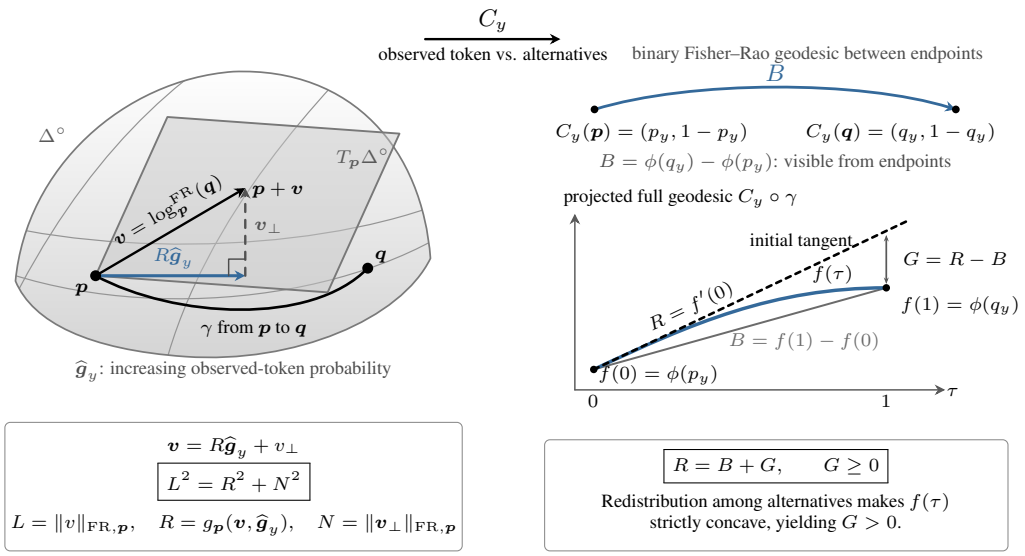
$$\boldsymbol{p} = \boldsymbol{P}_{\boldsymbol{\theta},t}(X; k_1), \qquad \boldsymbol{q} = \boldsymbol{P}_{\boldsymbol{\theta},t}(X; k_2), \qquad y = x_t. \tag{3}$$

Thus, $\boldsymbol{p} = (p_j)_{j \in \mathcal{V}}$ and $\boldsymbol{q} = (q_j)_{j \in \mathcal{V}}$ are the distributions produced by the shorter and longer contexts, respectively. Softmax with finite logits assigns every token strictly positive probability, so $\boldsymbol{p}, \boldsymbol{q} \in \Delta^\circ$, the interior of the vocabulary simplex. The corresponding observed-token probabilities are p_y and q_y. We compare the endpoints using the constant-speed Fisher–Rao geodesic $\gamma : [0, 1] \to \Delta^\circ$ from $\boldsymbol{p}$ to $\boldsymbol{q}$. The context path consists of the model predictions obtained at different values of k; the geodesic provides the intrinsic comparison between the selected endpoints $\boldsymbol{p}$ and $\boldsymbol{q}$.

An infinitesimal change in a probability distribution must have coordinate sum zero. Thus, the tangent space $T_{\boldsymbol{p}}\Delta^\circ$ consists of $|\mathcal{V}|$-dimensional zero-sum vectors. For $\boldsymbol{u}, \boldsymbol{w} \in T_{\boldsymbol{p}}\Delta^\circ$, the Fisher metric is

$$g_{\boldsymbol{p}}(\boldsymbol{u}, \boldsymbol{w}) = \sum_{j=1}^{|\mathcal{V}|} \frac{u_j w_j}{p_j}. \tag{4}$$

The coordinate-wise square-root map $\psi(\boldsymbol{p}) = 2\sqrt{\boldsymbol{p}}$ is an isometry that sends the simplex to the positive orthant of a radius-two sphere (Rao, 1945; Amari, 2016; Davis et al., 2024). This allows



(a) Full-distribution Fisher–Rao displacement

(b) What observed-token probabilities retain

Figure 1: **Full-distribution and observed-token views of predictive change.** (a) The Fisher–Rao log-map $\boldsymbol{v} = \log_{\boldsymbol{p}}^{\mathrm{FR}}(\boldsymbol{q})$ decomposes into the token-directed component $R\widehat{\boldsymbol{g}}_y$ and orthogonal residual $\boldsymbol{v}_\perp$, giving $L^2 = R^2 + N^2$. (b) Coarse-graining $\boldsymbol{p}, \boldsymbol{q}$ to the observed token versus all alternatives yields the signed binary endpoint displacement B. Along the projected full geodesic, $f(\tau) = \phi(\gamma_y(\tau))$ has initial slope R; redistribution among alternatives makes f concave, producing the nonnegative gap $G = R - B$. The schematic shows $q_y > p_y$; $\boldsymbol{p} + \boldsymbol{v}$ is only an affine tangent endpoint.

the Fisher–Rao geometry to be understood through ordinary spherical geometry. Fig. 1 previews the two decompositions developed in this section: full-distribution tangent motion and its binary coarse-grained view.

The logarithmic map represents the finite change from $\boldsymbol{p}$ to $\boldsymbol{q}$ as a tangent vector at $\boldsymbol{p}$, analogous to the Euclidean difference $\boldsymbol{q} - \boldsymbol{p}$. It equals the initial velocity of the endpoint geodesic:

$$\boldsymbol{v} = \log_{\boldsymbol{p}}^{\mathrm{FR}}(\boldsymbol{q}) = \dot{\gamma}(0), \tag{5}$$

where $\dot{\gamma}(\tau) = \mathrm{d}\gamma(\tau)/\mathrm{d}\tau$. Its direction indicates how the distribution moves from $\boldsymbol{p}$ toward $\boldsymbol{q}$. Its norm is the total Fisher–Rao displacement:

$$L = \|\boldsymbol{v}\|_{\mathrm{FR},\boldsymbol{p}} = \sqrt{g_{\boldsymbol{p}}(\boldsymbol{v}, \boldsymbol{v})} = \sqrt{\sum_{j=1}^{|\mathcal{V}|} \frac{v_j^2}{p_j}}. \tag{6}$$

The Fisher metric weights each squared coordinate change v_j^2 by $1/p_j$, so a fixed absolute change receives greater weight when token j has lower probability under $\boldsymbol{p}$. Since γ is a constant-speed geodesic parameterized over $[0, 1]$, $L = \|\dot{\gamma}(0)\|_{\mathrm{FR},\boldsymbol{p}}$ equals the length of γ, and hence the Fisher–Rao distance between $\boldsymbol{p}$ and $\boldsymbol{q}$. When the base point is clear, we shorten $\|\cdot\|_{\mathrm{FR},\boldsymbol{p}}$ to $\|\cdot\|_{\mathrm{FR}}$.

We next identify the tangent direction that most directly increases the observed-token probability. Let $\boldsymbol{e}_y$ denote the one-hot vector for y. The Fisher–Rao gradient of $\boldsymbol{p} \mapsto \log p_y$ points in the direction of steepest local increase per unit Fisher–Rao length (Amari, 2016):

$$\nabla^{\mathrm{FR}} \log p_y = \boldsymbol{e}_y - \boldsymbol{p}, \qquad \left\|\nabla^{\mathrm{FR}} \log p_y\right\|_{\mathrm{FR},\boldsymbol{p}} = \sqrt{\frac{1 - p_y}{p_y}}. \tag{7}$$

Any differentiable increasing reparameterization $f(p_y)$ with positive derivative at p_y has a Fisher-Rao gradient that is a positive scalar multiple of $\boldsymbol{e}_y - \boldsymbol{p}$, and thus yields the same direction after normalization. The resulting unit direction and the signed projection of $\boldsymbol{v}$ onto it are

$$\widehat{\boldsymbol{g}}_y = \sqrt{\frac{p_y}{1 - p_y}}(\boldsymbol{e}_y - \boldsymbol{p}), \qquad R = g_{\boldsymbol{p}}(\boldsymbol{v}, \widehat{\boldsymbol{g}}_y), \tag{8}$$

respectively. The signed magnitude R is positive when the endpoint geodesic initially moves toward higher observed-token probability and negative when it moves away. The residual vector $\boldsymbol{v}_\perp = \boldsymbol{v} - R\widehat{\boldsymbol{g}}_y$ has magnitude $N = \|\boldsymbol{v}_\perp\|_{\mathrm{FR},\boldsymbol{p}}$. For $L > 0$, define the normalized alignment $A = R/L$. It is the cosine of the angle between the total displacement and the direction toward the observed token.

To connect this decomposition to the scalar target probability, we define the *binary token-versus-rest view* by collapsing the vocabulary into two outcomes, the observed token and the aggregate of all alternatives, via a coarse-graining $C_y \colon \Delta^\circ \to (0,1)^2$ that maps a categorical distribution $\boldsymbol{r}$ to

$$C_y(\boldsymbol{r}) = (r_y, 1 - r_y). \tag{9}$$

The first coordinate retains the observed-token probability, while the second sums the probabilities of all alternative tokens. Its differential maps $\boldsymbol{v}$ to $\mathrm{d}C_y(\boldsymbol{v}) = (v_y, -v_y)$. The next proposition formalizes how the full displacement decomposes into the observed-token direction and its complement.

***Proposition* 1** (Token-directed Fisher–Rao decomposition)**.** *For interior categorical distributions $\boldsymbol{p}, \boldsymbol{q}$ and an observed token y, the token-directed and orthogonal components satisfy*

$$L^2 = R^2 + N^2. \tag{10}$$

Moreover, the orthogonal component has zero observed-token coordinate, and the binary coarse-graining C_y retains exactly the squared magnitude of the token-directed component

$$(\boldsymbol{v}_\perp)_y = 0, \qquad \|\mathrm{d}C_y(\boldsymbol{v})\|^2_{\mathrm{FR}} = R^2. \tag{11}$$

When $\boldsymbol{p} = \boldsymbol{q}$, we have $L = R = N = 0$, so $A = R/L$ is undefined. For $L > 0$, $A \in [-1, 1]$ and $A^2 = R^2/L^2$ is the fraction of squared displacement retained by the token-versus-rest view.

The proposition gives an exact Pythagorean decomposition: L is the total displacement, R is its signed token-directed component, and N is the remaining motion. Because $(\boldsymbol{v}_\perp)_y = 0$, N redistributes probability among the alternative tokens without changing the observed-token coordinate to first order. The binary token-versus-rest view therefore retains exactly the Fisher–Rao magnitude $|R|$ of the token-directed tangent component and discards N; the sign of R indicates whether the motion is toward or away from the observed token. The proof of Prop. 1 is given in App. A.1 and App. A.2.

3.3 Finite binary displacement and projection gap

In the previous subsection, we defined R by projecting the tangent displacement $\boldsymbol{v}$ at $\boldsymbol{p}$ onto the direction of increasing observed-token probability. Although this direction is defined by the observed token, R is computed from the full distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. We now compare it with the change captured by the observed-token probabilities p_y and q_y alone, thereby separating the contribution of token-directed motion from the effect of redistribution among alternative tokens.

Under coarse-graining C_y, the endpoints $\boldsymbol{p}$ and $\boldsymbol{q}$ become $C_y(\boldsymbol{p}) = (p_y, 1 - p_y)$ and $C_y(\boldsymbol{q}) = (q_y, 1 - q_y)$. For a binary distribution specified by $(s, 1 - s)$, let $\phi(s) = 2 \arcsin \sqrt{s}$ denote its Fisher–Rao arc coordinate. Along the full endpoint geodesic, define $f(\tau) = \phi(\gamma_y(\tau))$. Because $\gamma_y(0) = p_y$ and $\gamma_y(1) = q_y$, we have $f(0) = \phi(p_y)$ and $f(1) = \phi(q_y)$. Hence, the binary endpoint displacement B and the initial token-directed rate R can be written as

$$B = \phi(q_y) - \phi(p_y) = f(1) - f(0), \qquad R = f'(0) = \frac{\mathrm{d}}{\mathrm{d}\tau}\phi\big(\gamma_y(\tau)\big)\Big|_{\tau=0}. \tag{12}$$

The initial token-directed rate R is identical to the signed projection defined in Eq. (8), while B is the total signed Fisher–Rao displacement between binary endpoints. The quantity B depends only on p_y and q_y, and is positive when the longer context assigns greater probability to the observed token, see Fig. 1.

The difference between R and B arises from redistribution among the alternative tokens. To characterize this redistribution, we condition on the outcome not being y. The resulting alternative-token probabilities are $p_j/(1 - p_y)$ and $q_j/(1 - q_y)$ for $j \in \mathcal{V} \setminus \{y\}$. Their Hellinger affinity (equivalently, Bhattacharyya coefficient) h and normalized squared Hellinger distance D are

$$h = \sum_{j \in \mathcal{V}\setminus\{y\}} \sqrt{\frac{p_j}{1 - p_y}\frac{q_j}{1 - q_y}}, \qquad D = \frac{1}{2}\sum_{j \in \mathcal{V}\setminus\{y\}} \left(\sqrt{\frac{p_j}{1 - p_y}} - \sqrt{\frac{q_j}{1 - q_y}}\right)^2, \tag{13}$$

for which $D = 1 - h$. Here $h = 1$, or equivalently $D = 0$, occurs when the conditional distributions over alternative tokens agree. The full-distribution affinity, spherical angle, and distance are

$$\rho = \sum_{j=1}^{|\mathcal{V}|} \sqrt{p_j q_j} = \sqrt{p_y q_y} + \sqrt{(1-p_y)(1-q_y)}\, h, \qquad \omega = \arccos \rho, \qquad L = 2\omega, \tag{14}$$

respectively. As explained in App. A.1, the token-directed projection is then expressed as

$$R = \frac{2\omega}{\sin \omega} \frac{\sqrt{q}_y - \rho\sqrt{p}_y}{\sqrt{1-p_y}}. \tag{15}$$

At $\omega = 0$, we have $\boldsymbol{p} = \boldsymbol{q}$, and hence $R = 0$; equivalently, (15) extends continuously to this case. Define the *finite projection gap* as $G = R - B$. It is the remainder between two operations: computing the full-distribution displacement and projecting it toward the observed token, which gives R, and first collapsing the endpoints to the binary token-versus-rest view and then measuring their displacement, which gives B. Thus, whereas Prop. 1 describes the tangent-space effect of coarse-graining at $\boldsymbol{p}$, G compares this projected displacement with the *finite* displacement between the coarse-grained endpoints. The following theorem characterizes the sign of G and when it vanishes.

***Theorem* 1** (Binary displacement and projection gap)**.** *For interior categorical distributions $\boldsymbol{p}, \boldsymbol{q}$ and an observed token y, it holds that $G = R - B \geq 0$, and $G = 0$ if and only if $h = 1$. The gap G depends only on (p_y, q_y, h). At fixed $h < 1$, changes in the observed-token probabilities p_y and q_y can change the magnitude of G.*

In App. A.3, we prove Thm. 1 and derive an integral representation of G. The theorem relates the two views: the full-distribution projection R is at least as large as the binary displacement B. When $h = 1$, the conditional alternatives are unchanged and f is affine, so $R = B$. When $h < 1$, the projected path $C_y \circ \gamma$ is not the constant-speed binary geodesic joining $C_y(\boldsymbol{p})$ and $C_y(\boldsymbol{q})$. In this case, redistribution among the alternatives makes f concave (Afsari et al., 2013; Ferreira et al., 2014), so its initial slope exceeds its endpoint secant and $G = f'(0) - [f(1) - f(0)] > 0$. Thus alternative-token redistribution is necessary and sufficient for a positive gap. At a fixed alternative-token affinity h, the magnitude of G also depends on the endpoint probabilities p_y and q_y, which we illustrate via a concrete example in App. A.4.

For nearby endpoints, App. A.5 shows that G is second order in the orthogonal displacement, with leading term determined by alternative-token motion. App. A.6 further shows that G is a different geometric quantity from Hellinger-distance deficit, although they share the same zero set.

4 Predictive geometry in a randomized exposure study

In this section, we present our primary experimental study of how repeated document exposure changes the context-dependent predictive motion defined in Sec. 3, and decompose the exposure response into binary target-versus-rest motion and the additional full-distribution gap.

4.1 Experimental design and geometric outcome

For each candidate document i and continued-training run m, we evaluate the same observed tokens at two model states: the initial checkpoint before run m and the final checkpoint after the run. As in Sec. 3.1, each token is evaluated under nested portions of its preceding text. Every pair of adjacent context lengths that remain distinct after rounding yields a transition $(\boldsymbol{p}, \boldsymbol{q}, y)$ and therefore one token-directed projection R. The resulting R values form an empirical distribution at each checkpoint. To capture concentrated exposure effects without relying on the maximum R, we use its 90th percentile, $S_R = Q_{0.90}(R)$, as our primary summary. The observed checkpoint change is

$$\Delta S_{R,im} = S_R(\text{final checkpoint } m, i) - S_R(\text{initial checkpoint}, i). \tag{16}$$

A positive $\Delta S_{R,im}$ means that the upper tail of its token-directed context motion is stronger at the final checkpoint. Because the document, tokens, and context pairs are held fixed, Eq. (16) compares the same measurement before and after continued training. App. B gives the detailed procedure for constructing the transitions and computing R and S_R at the checkpoints.

Each candidate document i receives a presentation count $D_{im} \in \{0, 1, 2, 4, 8, 16\}$ in training run m. Candidate documents are grouped in sets of six matched on source or topic, tokenized length, and

Table 1: **Primary upper-tail response.** Slopes are in 10^{-3} Fisher–Rao arc units per $\log_2(D_{im}+1)$; ranges span targets. V/C: validation/confirmation. Initial-70M p values are selection-corrected (App. C.4).

Setting and phase	Phase slope ($\widehat{\beta}_{S_R}$)	Run range	p
Pythia 70M study 1 V	9.506	[8.984, 10.309]	3×10^{-5}
Pythia 70M study 1 C	9.753	[8.691, 10.334]	3×10^{-5}
Pythia 70M study 2 V	9.853	[9.543, 10.278]	10^{-5}
Pythia 70M study 2 C	9.999	[9.129, 10.498]	10^{-5}
Pythia 160M V	15.606	[14.784, 16.537]	10^{-5}
Pythia 160M C	15.967	[14.981, 16.462]	10^{-5}
SmolLM2 V	2.808	[2.245, 3.124]	10^{-5}
SmolLM2 C	2.951	[2.610, 3.184]	10^{-5}

initial-model difficulty. Within each group and training run, the six presentation counts are assigned once each; their assignment is rotated across runs according to a prespecified schedule. Every group therefore contributes 31 candidate-document presentations to each run, while the total presentation budget and number of updates remain fixed. Because all candidates assigned to a run contribute to the same final model, a document's outcome can also depend on the counts assigned to other documents. The analysis therefore compares complete randomized assignments of presentation counts across documents. App. C.2 gives the scheduling and allocation details.

We summarize the observed presentation-count pattern on the transformed scale $d_{im} = \log_2(D_{im}+1)$. For any checkpoint summary F, let ΔF_{im} denote its final-minus-initial change, defined as in Eq. (16). The within-run slope and its average across training runs are

$$\widehat{\beta}_{F,m} = \frac{\sum_{i\in\mathcal{I}}(d_{im}-\bar{d})\Delta F_{im}}{\sum_{i\in\mathcal{I}}(d_{im}-\bar{d})^2}, \qquad \widehat{\beta}_F = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\widehat{\beta}_{F,m}, \tag{17}$$

respectively, where $\bar{d}$ is the fixed transformed-count mean, $\mathcal{I}$ is the document set, and $\mathcal{M}$ is the training-run set in the phase. A positive slope means that documents assigned larger presentation counts tend to show larger checkpoint that increases in F within the randomized assignment. This centered slope provides a single summary of the overall trend across the six presentation counts.

4.2 Exposure changes token-directed motion

We evaluate four continued-training settings spanning three pretrained language-model checkpoints. Pythia-70M studies 1 (S1) and 2 (S2) (Biderman et al., 2023) start from the same checkpoint but use distinct FineWeb-Edu candidate documents (Penedo et al., 2024) and training runs. Pythia-160M uses the S2 assignments, whereas SmolLM2-135M uses arXiv title-abstract records (Ben Allal et al., 2025). See App. C.1 for details on the documents. Each setting uses disjoint validation (V) and confirmation (C) document pools, yielding eight phases across document samples, model scales, model families, and sources. App. C.3 details checkpoint sharing and the randomization test.

Specializing Eq. (17) to $F = S_R$, Table 1 reports the phase slope $\widehat{\beta}_{S_R}$, the range of the three training-run slopes $\widehat{\beta}_{S_R,m}$, and the randomization-test p value for each V and C phase. All reported slope values are multiplied by 10^3; the unscaled slopes are in Fisher–Rao units per unit of d_{im}. Every phase and training-run slope is positive, and in all eight phases the randomization test rejects the hypothesis that every measured document outcome would remain unchanged under any allowed assignment of presentation counts. S1 p values include a factor-of-three correction for selection among three candidate document-sample sizes; the others are unadjusted p values for this test.

Table 1 shows that documents assigned more presentations have larger final-minus-initial increases in S_R. In the framework of Sec. 3, repeated exposure therefore strengthens the model's movement toward the observed token as additional preceding context is supplied. This response persists across document samples, model scales, and model families. The decomposition of R then separates the token-versus-rest component from a gap determined jointly by alternative-token redistribution and endpoint probabilities. App. E reports supplementary checks: observed-token probability summaries have positive slopes in all eight phases, while mean, median, and 75th-percentile summaries of R have positive slopes in both phases of Pythia-70M S1.

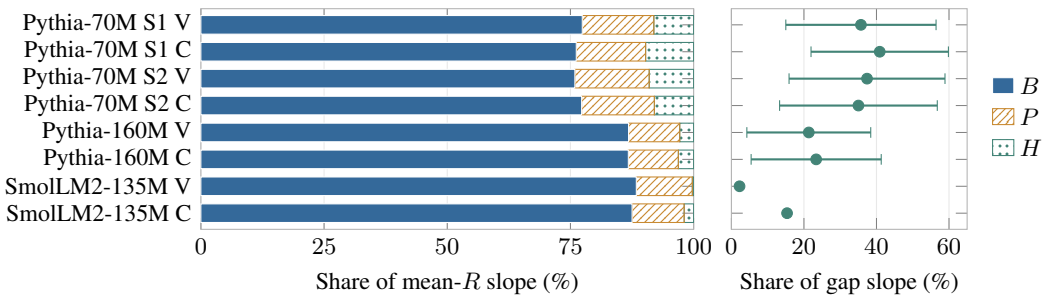


Figure 2: Mean-response composition (left) and affinity replacement orders (right) in the validation and confirmation phases of four continued-training settings spanning three pretrained language models.

4.3 Decomposing the exposure response

The primary upper-tail score S_R does not admit an additive decomposition, so we use arithmetic means to examine the allocation-associated response. For $Z \in \{R, B, G\}$, let M_Z be its checkpoint mean over the same retained target-token and context-length pairs. Since $R = B + G$ for every transition (Thm. 1), averaging, taking the initial-to-final difference, and applying Eq. (17) give

$$\Delta M_R = \Delta M_B + \Delta M_G, \qquad \widehat{\beta}_{M_R} = \widehat{\beta}_{M_B} + \widehat{\beta}_{M_G}. \tag{18}$$

This decomposes the mean response, not the primary change in the 90th percentile of R.

All ratios below use a common numerical mask requiring at least 0.01 total alternative-token probability at both context endpoints and both checkpoints. It retains 97.44–98.78% of transitions in each phase. Because the mask depends partly on final-checkpoint predictions, the ratios describe the retained instances. App. D gives the numerical checks.

The left plot of Fig. 2 shows that binary endpoint displacement B accounts for 75.90–88.37% of the phase-level mean-R slope. The finite gap G accounts for the remaining 11.63–24.10%. Thm. 1 shows that G vanishes exactly when the conditional alternative-token distributions agree across the two context endpoints, but its magnitude also depends on their observed-token probabilities.

To describe the gap response, App. D allocates each transition's change in G between endpoint-probability changes (P) and alternative-token affinity changes (H), averaging the two replacement orders. The left plot displays the slopes of B, P, and H as shares of the mean-R slope, while the right plot shows the affinity share of the gap slope and its two replacement orders. Endpoint contributions exceed averaged affinity contributions in every target and numerical variant. The affinity slope is positive in 23 of 24 target–phase evaluations, although direct alternative-token dissimilarity has a negative slope in three (App. E). These retrospective attributions show that the observed gap response combines endpoint weighting with changes among alternatives.

5 Applications of predictive geometry

In this section, we apply the predictive-geometry framework developed in Sec. 3 and validated in Sec. 4 to membership inference (Sec. 5.1) and context-scoped editing (Sec. 5.2).

5.1 Reading exposure from the final checkpoint

Sec. 4 showed that exposure changes Fisher–Rao context paths. We now test whether we can infer exposure from the final checkpoint alone. For the membership classifier $f_{\boldsymbol{\eta}}$, we consider logistic regression and histogram gradient boosting and select the classifier type and hyperparameters by grouped cross-validation. For a candidate document X with label $M = \mathbf{1}[D_{im} > 0]$, we query each eligible token under 24 nested contexts and form

$$\boldsymbol{z}_{\mathrm{FR}}(X) = \mathrm{Agg}_{t,k}(A, L, R, B, G, N, D, \text{path structure}), \qquad s_{\mathrm{FR}}(X) = f_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathrm{FR}}(X)), \tag{19}$$

where we aggregate token-level statistics and path summaries, then train $f_{\boldsymbol{\eta}}$ on labeled auxiliary members ($M = 1$) and nonmembers ($M = 0$). We compare with schemes that use Fisher–Rao paths, observed-token likelihood paths, and their concatenation as inference signal. We also compare with

Table 2: Membership inference AUC from final-checkpoint context paths

Score	Initial 70M	Fresh 70M	160M	SmolLM2
Min-K%++	0.724	0.734	0.818	0.603
Fisher–Rao paths	0.703	0.737	0.829	0.562
Likelihood paths	0.746	0.761	**0.857**	**0.628**
Likelihood + Fisher–Rao	**0.751**	**0.763**	0.854	0.605

Min-K%++ (Zhang et al., 2025), a strong established reference-free baseline that averages the lowest $K\%$ of vocabulary-standardized observed-token log probabilities. We evaluate held-out document groups. App. F gives the complete details.

In Table 2, we show membership-inference AUC, averaged equally over groups pairing a final checkpoint with a held-out document fold. The results show that Fisher–Rao paths alone carry membership information in every setting and match or exceed Min-K%++ in AUC on Pythia-70M study 2 and Pythia-160M. By combining Fisher–Rao and likelihood features, we obtain the best AUC in both 70M studies. We obtain the remaining best results with likelihood paths. Using Thm. 1, we explain this pattern: a likelihood path recovers B exactly, and Sec. 4.3 shows that B dominates the exposure response. The final checkpoint therefore reveals an exposure-induced context signature.

5.2 Context-scoped editing

We use predictive geometry to define context-scoped editing: increase B toward a chosen continuation in one short context while keeping L small in longer and unrelated contexts. This compares intended and collateral effects in the same units. A one-row *protected* update realizes this objective by preserving every logit on selected construction contexts.

Let H_{protect} stack the output-layer hidden states of the protected contexts, and let $\boldsymbol{a}_s = \nabla_{\boldsymbol{w}_y} \phi(p_{\boldsymbol{\theta}}(y \mid c_s))$ for the target row $\boldsymbol{w}_y$. The minimum-norm unit-gain protected update is $\boldsymbol{d}^\star = \frac{\Pi_\perp \boldsymbol{a}_s}{\boldsymbol{a}_s^\top \Pi_\perp \boldsymbol{a}_s}$, where $\Pi_\perp$ is the Euclidean orthogonal projector onto the null space $\ker H_{\mathrm{protect}} = \{\boldsymbol{d} : H_{\mathrm{protect}} \boldsymbol{d} = 0\}$. The projection removes directions affecting protected contexts. When $\Pi_\perp \boldsymbol{a}_s \neq 0$, $\boldsymbol{d}^\star$ is the minimum-norm unit-gain direction and preserves their logits for any finite step, following the null-space principle of Fang et al. (2025). Any differentiable coordinate of p_s with positive derivative preserves the direction of $\boldsymbol{d}^\star$. The Fisher–Rao coordinate makes B directly comparable with L.

We use B and L to select the step size, requiring strong short-context motion and small longer- and unrelated-context changes, and test the same criteria on disjoint held-out contexts. On Pythia-70M, editing `The Virelion code is` → `blue` raises target probability from about 0.002% to 9.15% and 6.44% in two fresh seeds, while held-out longer-context $Q_{0.90}(L)$ remains 0.012–0.016 and unrelated-context RMS L remains 0.002–0.006. A strength-matched unprotected update produces 41–$58\times$ larger longer-context changes and 49–$131\times$ larger unrelated-context changes.

For comparison, after fixing the protocol, we train the same association with ordinary next-token updates in five new runs. Mean target probability reaches 0.0516 after 64 presentations, but longer-context movement exceeds the short-context effect and unrelated-context RMS L reaches 0.333. No tested document at any presentation count meets the locality criteria, whereas the protected edit achieves the target change and limits longer- and unrelated-context changes more effectively. App. G gives the derivation, protocol, and full results.

6 Conclusion

Using Fisher–Rao predictive geometry, we analyzed how next-token distributions move with inference context and how repeated training exposure changes that motion. We derived an exact split into binary token-versus-rest displacement and a nonnegative full-distribution gap determined by alternative-token redistribution and endpoint probabilities. Through randomized continued training, we show that exposure strengthens both, predominantly the binary term. The same context paths retain a readable final-checkpoint exposure signature and define target and locality criteria for scoped editing. Predictive geometry thus connects exposure, auditability, and control. Our evidence covers controlled continued training of small models and one editing association; extending these findings to original pretraining at scale and broader editing targets remains open.

Ethics statement. The work studies controlled continued-training inclusion and may inform both training-data audits and attacks. The membership analysis is an observational probe, but the same signals could be used adversarially. Model editing can also create unintended trigger behavior if applied without authorization. Experiments use public models and public or licensed corpora; released artifacts should exclude raw material that cannot be redistributed. The results do not identify membership in original pretraining.

Reproducibility statement. The appendices state the mathematical conventions and an anonymous release bundle containing the code to reproduce the results is located at https://anonymous.4open.science/r/predictive-geometry-in-language-models-private-archive-12FE.

AI use statement. Generative AI assisted code, verification, literature organization, and editing. The authors checked the resulting derivations, analyses, citations, and prose and take responsibility for the paper.

REFERENCES

Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Japan, 2016.

Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big—data-centric training of a small language model. In *Conference on Language Modeling (COLM)*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

O. P. Ferreira, A. N. Iusem, and S. Z. Németh. Concepts and techniques of optimization on the sphere. *TOP*, 22(3):1148–1170, 2014.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Hugging Face Smol Models Research. SmolLM2-135M model repository. Hugging Face model repository, revision `93efa2f097d58c2a74874c7e644dbc9b0cee75a2`, 2025. Accessed 2026-09-02.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Dimitri Kountourogiannis and Paul Loya. A derivation of Taylor's formula with integral remainder. *Mathematics magazine*, 76(3):217–219, 2003.

Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the effect of training data on deep learning predictions via randomized experiments. In *International Conference on Machine Learning (ICML)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICML)*, 2023.

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 (NeurIPS)*, 2024.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

Jiajun Sun and Zhanrui Cai. Token-level likelihood-array regression for membership inference and AI-generated text detection, 2026. arXiv preprint arXiv:2608.22179.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. ReCaLL: Membership inference via relative conditional log-likelihoods. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

A GEOMETRIC DERIVATIONS

This section supplies proof completion, computational formulas, and boundary cases for the local decomposition in Proposition 1 and the finite-gap result in Theorem 1.

A.1 THE PROJECTION IN CATEGORICAL COORDINATES

We present the projection from the simplex into the positive orthant of the radius-two sphere, and partly prove Prop. 1.

The categorical Fisher–Rao metric at an interior distribution $\boldsymbol{p}$ is $g_{\boldsymbol{p}}(\boldsymbol{v}, \boldsymbol{w}) = \sum_{j=1}^{|\mathcal{V}|} v_j w_j / p_j$ for tangent vectors with zero coordinate sum. The map[1] $\boldsymbol{p} \mapsto \boldsymbol{s} = 2\sqrt{\boldsymbol{p}}$ is an isometry into the positive orthant of the radius-two sphere: its differential sends $\boldsymbol{v}$ to $(v_j/\sqrt{p_j})_j$, whose Euclidean inner product is $g_{\boldsymbol{p}}$ (Rao, 1945; Amari, 2016). This supplies a direct route to the formulas, using ordinary spherical geometry. Throughout, Fisher-Rao geodesics refer to the geodesics induced by the Levi–Civita connection of the Fisher metric, i.e., the unique torsion-free connection compatible with this metric (Amari, 2016, Sec. 5.9).

Let

$$\rho = \sum_{j=1}^{|\mathcal{V}|} \sqrt{p_j q_j}, \qquad \omega = \arccos \rho, \qquad L = 2\omega. \tag{20}$$

The spherical displacement at $\boldsymbol{s} = 2\sqrt{\boldsymbol{p}}$ toward $2\sqrt{\boldsymbol{q}}$ is

$$\widetilde{\boldsymbol{v}} = \frac{2\omega}{\sin \omega}(\sqrt{\boldsymbol{q}} - \rho\sqrt{\boldsymbol{p}}). \tag{21}$$

At $\boldsymbol{p} = \boldsymbol{q}$, the displacement is zero and $\omega / \sin \omega$ takes its continuous limit 1.

The unit tangent direction that increases p_y is

$$\boldsymbol{u}_y = \frac{\boldsymbol{e}_y - \langle \boldsymbol{e}_y, \sqrt{\boldsymbol{p}} \rangle \sqrt{\boldsymbol{p}}}{\|\boldsymbol{e}_y - \langle \boldsymbol{e}_y, \sqrt{\boldsymbol{p}} \rangle \sqrt{\boldsymbol{p}}\|} = \frac{\boldsymbol{e}_y - \sqrt{p_y}\sqrt{\boldsymbol{p}}}{\sqrt{1 - p_y}}, \tag{22}$$

where $\boldsymbol{e}_y$ is the one-hot vector with coordinate y equal to one and all other coordinates zero. Indeed, $\boldsymbol{u}_y$ is orthogonal to $\sqrt{\boldsymbol{p}}$ and has Euclidean norm one. Taking the inner product with Eq. (21) gives the computational formula for the signed projection of $\widetilde{\boldsymbol{v}}$ onto $\boldsymbol{u}_y$

$$R = \langle \widetilde{\boldsymbol{v}}, \boldsymbol{u}_y \rangle = \frac{2\omega}{\sin \omega} \frac{\sqrt{q}_y - \rho\sqrt{p_y}}{\sqrt{1 - p_y}}. \tag{23}$$

Note that this is the same R defined in (8), expressed in spherical coordinates. The gradient of $\phi(p_y) = 2 \arcsin \sqrt{p_y}$ on the sphere is also $\boldsymbol{u}_y$. Therefore its derivative along the constant-speed endpoint geodesic is R at $t = 0$, proving the equivalence with Eq. (12).

For the orthogonal identity in Prop. 1, the square-root isometry gives $N = \|\widetilde{\boldsymbol{v}} - R\boldsymbol{u}_y\|_2$. The two vectors $R\boldsymbol{u}_y$ and $\widetilde{\boldsymbol{v}} - R\boldsymbol{u}_y$ are orthogonal because $\langle \widetilde{\boldsymbol{v}}, \boldsymbol{u}_y \rangle = R$ and $\|\boldsymbol{u}_y\| = 1$. The spherical logarithmic map has norm $\|\widetilde{\boldsymbol{v}}\| = L$. Pythagoras therefore gives $L^2 = R^2 + N^2$ and $|R| \le L$.

This is a decomposition of the initial tangent vector at the shorter-context prediction $\boldsymbol{p}$. It does not assert that the binary coarse-grained endpoint displacement B equals the token-directed component R. The binary coarse-graining does imply $|B| \le L$ by contraction, but the signed comparison $B \le R$ needs the curvature argument below. When $L = 0$, the projection and orthogonal norm vanish; a numerical value for $A = R/L$ requires an explicit convention in any implementation that uses it.

A.2 INTRINSIC PROJECTION AND BINARY COARSE-GRAINING

Proof of Prop. 1. The preceding spherical-coordinate calculation proves the orthogonal identity $L^2 = R^2 + N^2$ and the bound $A \in [-1, 1]$. At $\boldsymbol{p} = \boldsymbol{q}$, the displacement and both components vanish,

[1] All square roots are coordinatewise.

so A is undefined. For $L > 0$, the retained fraction of tangent energy is $R^2/L^2 = A^2$. It remains to establish the intrinsic and coarse-grained statements in Eq. (11).

The main-text vector $\boldsymbol{v} = \log_{\boldsymbol{p}}^{\mathrm{FR}} \boldsymbol{q}$ is in probability coordinates; its image under $\mathrm{d}\psi_{\boldsymbol{p}}$ is the spherical log vector $\widetilde{\boldsymbol{v}}$ defined in Eq. (21). For every tangent $\boldsymbol{w}$, the zero-sum constraint gives

$$g_{\boldsymbol{p}}(\boldsymbol{e}_y - \boldsymbol{p}, \boldsymbol{w}) = w_y/p_y = \mathrm{d}\log p_y[\boldsymbol{w}].$$

Thus the natural gradient is $\boldsymbol{e}_y - \boldsymbol{p}$, with squared norm $(1-p_y)/p_y$. Its normalized form $\widehat{\boldsymbol{g}}_y$ maps to the unit spherical direction $\boldsymbol{u}_y$ in Eq. (22). Isometry proves the intrinsic projection identity in Eq. (8).

For $\boldsymbol{v}_\perp = \boldsymbol{v} - R\widehat{\boldsymbol{g}}_y$, orthogonality between $\boldsymbol{v}_\perp$ and $\widehat{\boldsymbol{g}}_y$ gives $g_{\boldsymbol{p}}(\boldsymbol{v}_\perp, \boldsymbol{e}_y - \boldsymbol{p}) = 0$, hence $(\boldsymbol{v}_\perp)_y = 0$. Since $(\widehat{\boldsymbol{g}}_y)_y = \sqrt{p_y(1-p_y)}$, we have $v_y = R\sqrt{p_y(1-p_y)}$. Therefore

$$\|\mathrm{d}C_y(\boldsymbol{v})\|_{\mathrm{FR}}^2 = \frac{v_y^2}{p_y} + \frac{v_y^2}{1-p_y} = R^2. \tag{24}$$

This completes the proof of Eq. (11). □

For fixed interior $\boldsymbol{p}$, a tangent $\boldsymbol{w}$, and a infititesimally displaced vector $\boldsymbol{q} = \boldsymbol{p} + \varepsilon\boldsymbol{w}$, smoothness of the logarithm gives $\log_{\boldsymbol{p}}^{\mathrm{FR}} \boldsymbol{q} = \varepsilon\boldsymbol{w} + O(\varepsilon^2)$. Hence, the local token-directed projection is

$$R(\boldsymbol{p}, \boldsymbol{p} + \varepsilon\boldsymbol{w}; y) = \varepsilon \frac{w_y}{\sqrt{p_y(1-p_y)}} + O(\varepsilon^2), \tag{25}$$

which is a standardized observed-token probability change. For finite endpoints, $\mathrm{d}C_y(\log_{\boldsymbol{p}}^{\mathrm{FR}} \boldsymbol{q})$ has signed binary arc coordinate R, while $\log_{C_y(\boldsymbol{p})}^{\mathrm{FR}} C_y(\boldsymbol{q})$ has coordinate B. Their difference G is the finite coarse-graining remainder characterized below.

A.3 Proof of Thm. 1 and explicit form of the finite gap

The purpose of the proof is to identify what bends the observed-token coordinate along the full geodesic. Define the distance from a distribution to the observed-token vertex by

$$\rho_y(\boldsymbol{p}) = 2\arccos\sqrt{p_y} = \pi - \phi(p_y). \tag{26}$$

This is unsquared geodesic distance on the completed radius-two simplex. Along the interior endpoint geodesic, it is smooth, since $0 < p_y < 1$.

Proof of Thm. 1. On a sphere of radius two, the Hessian of distance to a fixed point, in a region with distance $\rho \in (0, \pi)$, is

$$\mathrm{Hess}\,\rho(\boldsymbol{w}, \boldsymbol{w}) = \frac{1}{2}\cot(\rho/2)\left(\|\boldsymbol{w}\|^2 - \langle \boldsymbol{w}, \nabla\rho\rangle^2\right). \tag{27}$$

This is the radius-two form of the classical spherical distance Hessian (Ferreira et al., 2014). Its coefficient is nonnegative in the positive orthant. Let $f(t) = \phi(\gamma_y(t))$ along the constant-speed minimal geodesic. Since geodesic acceleration is covariantly zero, and $f = \pi - \rho_y \circ \gamma$, Eq. (27) yields

$$-f''(t) = \frac{1}{2}\sqrt{\frac{\gamma_y(t)}{1-\gamma_y(t)}}\|\dot\gamma_\perp(t)\|_{\mathrm{FR}}^2. \tag{28}$$

Here, the subscript $\perp$ removes the component along the observed-token ascent direction at the current point. Taylor's formula with integral remainder (Kountourogiannis & Loya, 2003) now gives

$$G = f'(0) - [f(1) - f(0)] = \frac{1}{2}\int_0^1 (1-t)\sqrt{\frac{\gamma_y(t)}{1-\gamma_y(t)}}\|\dot\gamma_\perp(t)\|_{\mathrm{FR}}^2\,\mathrm{d}t. \tag{29}$$

We conclude that $G \geq 0$ since every factor in the integrand is nonngenative. The endpoint geodesic remains in the interior, where the weight is strictly positive except for its $(1-t)$ factor at the last point. Equality therefore holds exactly when the orthogonal velocity vanishes throughout the path. In the square-root coordinates, that means the direction of the complement vector is constant. Its

squared normalized coordinates are the conditional alternative-token probabilities. Hence equality is equivalent to the two endpoint conditional alternatives being identical, or $h = 1$. This includes the case $\boldsymbol{p} = \boldsymbol{q}$ by continuity.

Finally, affinity factorizes as (see (14))

$$\rho = \sqrt{p_y q_y} + \sqrt{(1 - p_y)(1 - q_y)}\, h. \tag{30}$$

Substituting this into Eq. (23) and subtracting B expresses G as a function of (p_y, q_y, h). In App. A.4, we shall construct two pairs that have identical h and different gaps G, proving that fixed $h < 1$ does not fix G. □

For numerical attribution, denote ρ in (30) by $\rho(p_y, q_y, h)$, then the gap is expressed as

$$G(p_y, q_y, h) = \frac{2 \arccos \rho(p_y, q_y, h)}{\sqrt{1 - \rho(p_y, q_y, h)^2}} \frac{\sqrt{q_y} - \rho(p_y, q_y, h)\sqrt{p_y}}{\sqrt{1 - p_y}} - 2(\arcsin \sqrt{q_y} - \arcsin \sqrt{p_y}). \tag{31}$$

Hybrid inputs $\{(p_y, q_y)_i, h_j\}$ with interior endpoints and $h_j \in [0, 1]$ define the same function, whether or not that hybrid was an observed checkpoint pair.

Equation (29) also clarifies the attribution's scope. The integrand depends on the full geometric path and on probability-dependent weights. The initial orthogonal norm N^2 is only one endpoint quantity; finite G generally differs from a fixed multiple of N^2.

A.4 Constructed endpoint-weighting example

With y in the first coordinate, consider two endpoint pairs

$$\boldsymbol{p}^{(0)} = (0.2, 0.3, 0.5), \quad \boldsymbol{q}^{(0)} = (0.2, 0.6, 0.2), \quad G^{(0)} = 0.0607174, \tag{32}$$

and

$$\boldsymbol{p}^{(1)} = (0.4, 0.225, 0.375), \quad \boldsymbol{q}^{(1)} = (0.4, 0.45, 0.15), \quad G^{(1)} = 0.0739863. \tag{33}$$

Within each pair, $p_y = q_y$, so $B = 0$. Conditioning on the outcome not being y gives the truncated distribution over the alternative tokens

$$\frac{(0.3, 0.5)}{0.8} = \frac{(0.225, 0.375)}{0.6} = (0.375, 0.625), \qquad \frac{(0.6, 0.2)}{0.8} = \frac{(0.45, 0.15)}{0.6} = (0.75, 0.25).$$

Both pairs therefore have the same conditional alternative change and the same $h = 0.925615$, but their different common observed-token probabilities produce different gaps G. Fig. 3 plots the binary coordinate along their endpoint geodesics and makes this endpoint-weighting effect visible when binary endpoint motion is zero.

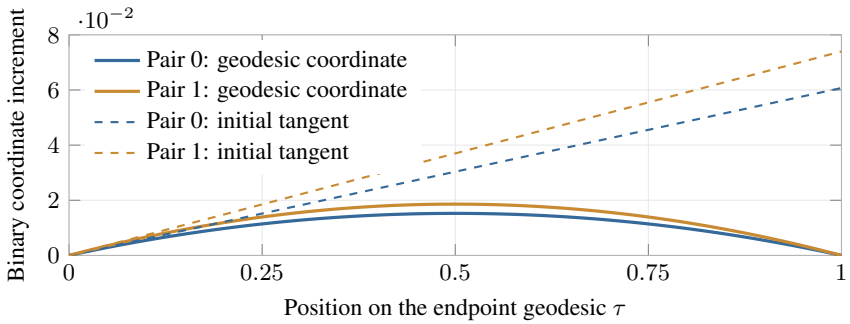


Figure 3: **An initial projection can be positive even when the binary endpoint change is zero.** The curves show $\phi(\gamma_y(\tau)) - \phi(p_y)$; dashed lines are their initial tangents. Each curve returns to zero at $\tau = 1$, while its tangent reaches $R = G > 0$. The second pair has a larger gap despite identical conditional alternatives. The horizontal axis is geometric interpolation, not context length or training time.

A.5 THE LOCAL LIMIT

For a tangent vector $\boldsymbol{w}$ at an interior $\boldsymbol{p}$, the binary/complement split can be checked without a path. Let $r_j = p_j/(1-p_y)$ for $j \neq y$ and write $w_j = -w_y r_j + (1-p)y)\dot{r}_j$. Since $\sum_{j\in\mathcal{V}\setminus\{y\}} \dot{r}_j = 0$,

$$\|\boldsymbol{w}\|_{\mathrm{FR}}^2 = \frac{w_y^2}{p_y(1-p_y)} + (1-p_y)\sum_{j\in\mathcal{V}\setminus\{y\}} \frac{\dot{r}_j^2}{r_j}. \tag{34}$$

The first term is binary observed-token/rest energy; the second is conditional alternative energy. It explains why the orthogonal tangent component has a natural local interpretation.

The next result analyzes the scaling of G for nearby endpoints.

***Corollary* 1** (Second-order gap for nearby endpoints)**.** *For fixed $\boldsymbol{p}$ and tangent $\boldsymbol{w}$, let $\boldsymbol{q}_\varepsilon = \exp_{\boldsymbol{p}}(\varepsilon\boldsymbol{w})$, remaining interior for sufficiently small ε. Then*

$$G(\boldsymbol{p}, \boldsymbol{q}_\varepsilon) = \frac{\varepsilon^2}{4}\sqrt{\frac{p_y}{1-p_y}}\|\boldsymbol{w}_\perp\|_{\mathrm{FR}}^2 + O(\varepsilon^3). \tag{35}$$

Proof. Along the rescaled geodesic, velocity is $\varepsilon\boldsymbol{w} + O(\varepsilon^2)$ in smooth local coordinates. The weight in Eq. (29) equals $\sqrt{p_y/(1-p_y)} + O(\varepsilon)$, uniformly on the unit parameter interval. Integration of $(1-t)/2$ gives $1/4$. □

The expansion (35) holds at a fixed interior point. Its coefficient grows near saturation; it is not a uniform approximation over the simplex boundary.

A.6 COMPARISON WITH THE HELLINGER DEFICIT

Define the normalized squared Hellinger distance by $\mathrm{Hel}^2(\boldsymbol{p}, \boldsymbol{q}) = 1 - \sum_{j=1}^{|\mathcal{V}|}\sqrt{p_j q_j}$. Equation (30) immediately gives

$$J = \mathrm{Hel}^2(\boldsymbol{p}, \boldsymbol{q}) - \mathrm{Hel}^2(C_y(\boldsymbol{p}), C_y(\boldsymbol{q})) = \sqrt{(1-p_y)(1-q_y)}(1-h). \tag{36}$$

For interior endpoints, J and G vanish on the same set. Both depend on endpoint probabilities as well as h. However, they have different numerical values and different additive roles: J compares Hellinger distances, whereas G measures the additional token-directed displacement induced by changes in the alternative-token distribution beyond what is captured by the binary endpoint probabilities alone. Neither equality condition establishes that one measurement predicts a downstream outcome better than another.

B MEASUREMENT ACQUISITION AND AGGREGATION

This section describes the original paired-checkpoint measurements. The new final-checkpoint trajectory acquisition and its explicit validity masks are described in App. F.

The acquisition rule uses zero-based token indices: its position r corresponds to $t = r + 1$ in Sec. 3.1. Write the retained sequence as $(x_0, \ldots, x_{T_{\mathrm{acq}}})$ in this convention. Acquisition retained at most 129 leading token IDs and therefore at most 128 predicted positions. At position $r \in \{1, \ldots, T_{\mathrm{acq}}\}$, the 24 nested context lengths were

$$\ell_{rg} = \min\Big\{r, \max\Big\{1, \mathrm{rint}((r+1)^{g/23} - 1)\Big\}\Big\}, \quad g = 0, \ldots, 23, \tag{37}$$

with $\ell_{r,23} = r$. NumPy's `rint` uses round-to-nearest with ties to even. A transition was eligible only when $\ell_{r,g+1} > \ell_{rg}$; both contexts were suffixes of the same left prefix, and the observed token was x_r. Base and final checkpoints used the same transition mask. Every eligible transition was scored at unit token stride rather than subsampling positions.

Documents contributed 758–2,000 eligible transitions in initial Pythia (median 1,676), 826–2,000 in fresh Pythia (median 1,694), and 1,874–2,000 in SmolLM2/arXiv (median 2,000). The respective predicted-position ranges were 58–128, 62–128, and 121–128. These counts are measurement

Table 3: Pinned continued-training systems. Candidate and background columns count presentations per target. All but Pythia-70M study 1 use 40,300 total presentations and 5,038 optimizer updates; study 1 uses 43,400 and 5,425.

System	Base checkpoint revision	Candidate	Background
Pythia-70M study 1	`e93a9faa9c77e5d09219f6c868bfc7a1bd65593c`	21,700	21,700
Pythia-70M study 2	`e93a9faa9c77e5d09219f6c868bfc7a1bd65593c`	18,600	21,700
Pythia-160M	`582159a2dfe3e712a8d47ae83dec95ae3bde8e7e`	18,600	21,700
SmolLM2-135M	`93efa2f097d58c2a74874c7e644dbc9b0cee75a2`	18,600	21,700

multiplicities, not independent experimental units; the randomization analysis operates on document–target outcomes.

Complete next-token distributions were converted to normalized square-root vectors and evaluated in float64. The implementation computes the spherical angle as $\operatorname{atan2}(\|\sqrt{\boldsymbol{q}} - c\sqrt{\boldsymbol{p}}\|_2, c)$ and assigns $R = A = 0$ when displacement or observed-token ascent norm is at most 10^{-12}. This is a finite-precision convention; the exact results assume interior, nondegenerate inputs. Historical production artifacts do not record how often this branch fired. Production roots and realized-token channels were retained with different rounding, which motivates the reconstruction residual and common-mass mask in App. D.

For the primary endpoint, transition values were aggregated separately at each checkpoint with NumPy's linear empirical 90th percentile, then differenced as in Eq. (16). The decomposition instead uses arithmetic means on identical retained transitions, because means preserve $R = B + G$. Neither the mean decomposition nor a transitionwise gap is an algebraic decomposition of the primary difference of quantiles.

C Training design and randomization inference

C.1 Documents, targets, and training

Pythia-70M study 1 (Biderman et al., 2023) formed 700 matched FineWeb-Edu blocks (Penedo et al., 2024): 100 development blocks and 300 in each held document pool. Its recorded power procedure selected the first 200 blocks for each primary held analysis. The Pythia-70M study 2 used 600 new blocks, with targets 1–3 evaluated on the validation pool and separately trained targets 4–6 on confirmation. Paired Pythia-160M targets 4–6 reused the study-2 documents, assignments, and presentation order. SmolLM2 (Ben Allal et al., 2025; Hugging Face Smol Models Research, 2025) used 600 blocks of 3,600 arXiv title–abstract records first published after the pinned base checkpoint, with three validation and three separately trained confirmation targets.

Blocks match six documents on source or topic, pinned-tokenizer length, and base-model difficulty. Pythia-70M study 1 excludes exact, normalized, and inter-document near duplicates. Study 2 and SmolLM2 additionally require word-5-shingle Jaccard and character-5-gram TF–IDF cosine below 0.80 against their declared background, candidate, prior-study, and older-reference pools. This screening limits measured overlap within the constructed studies; it does not establish absence from the models' original pretraining corpora.

All runs update the complete model with AdamW (Loshchilov & Hutter, 2019), peak learning rate 10^{-5}, 100 linear-warmup steps, weight decay 0.01, gradient clipping at 1.0, and effective batch size eight. When memory requires microbatching, token-weighted losses are accumulated before one clip, optimizer step, and schedule update. Pythia-70M study 2, Pythia-160M, and SmolLM2 have 5,037 full batches followed by one batch of four. Retained nonpadding predicted-token totals span 0.076% of their mean in study 2 and 0.0061% in SmolLM2; paired 160M runs use the same ledgers. Comparable token totals for study 1 and FLOPs for all settings were not retained. The design therefore fixes presentations and optimizer updates, not exact tokens or compute.

C.2 Complete Latin allocation and scheduling

Within every six-document block, each target assigns the doses $D_{im} \in \{0, 1, 2, 4, 8, 16\}$ once, keeping their sum equal to 31. A block draws a uniform dose permutation and targets receive cyclic

shifts of that permutation. Pythia-70M study 2 and SmolLM2 use all six shifts. Study 1 has eight targets and uses the six shifts plus two block-indexed repeated shifts; resampling for its three-target analyses preserves this complete eight-target construction. Pythia-160M inherits the shifts of paired study-2 targets 4–6. Thus neither training nor inference assigns document doses independently.

For dose $D_{im} > 0$, presentation $r \in \{0, \ldots, D_{im}-1\}$ receives ideal training progress $(r+1/2)/D_{im}$. Independently permuted background documents receive analogous evenly spaced progress values. Candidate and background events are jointly ordered with seeded tie breaking, then batched subject to document uniqueness within a batch. Timing is approximately spread over training, but exact spacing and last-presentation recency still vary with dose. Repetition, recency, and peer reallocation are therefore bundled in the allocation estimand.

Validation outcomes were opened before confirmation in each experiment. Pythia-70M study 2 and SmolLM2 use targets 1–3 only for validation and separately trained targets 4–6 only for confirmation. Their two phases still share one six-target Latin assignment. Pythia-160M evaluates both document pools on the same three targets. Immutable access logs record these openings, but sequential labels do not make the phase estimates independent.

C.3 Complete assignments and the sharp null

A single continued-training checkpoint is shared by many evaluated documents. Its predictions can consequently depend on the entire allocation Z. Write $Y_{im}(Z)$ for the outcome of document i in target m under that allocation. The phase-level global sharp null is

$$H_0^{\text{sharp}}: \quad Y_{im}(z) = Y_{im}(z') \quad \text{for every evaluated } (i,m) \text{ and allowed } z, z'. \tag{38}$$

Under this null, the observed outcomes impute the outcomes at all allowed allocations. The randomization distribution must therefore draw from the complete experimental assignment mechanism and then recompute the test statistic with those fixed outcomes. Permuting individual document labels independently would generally create assignments that the study could not have produced. Here the global sharp null itself supplies the required outcome imputation. For non-sharp hypotheses under network interference, Athey et al. (2018) construct artificial experiments in which the tested hypothesis becomes sharp; the present test uses the stated global sharp null directly. This complete-design analysis follows the design-based principle used in randomized training-data studies (Lin et al., 2022).

For a prespecified statistic T where larger values are more extreme, the plus-one Monte Carlo calculation has the form

$$p_{\text{MC}} = \frac{1 + \sum_{b=1}^{B_{\text{MC}}} \mathbf{1}\{T(Z^{(b)}, Y_{\text{obs}}) \geq T(Z_{\text{obs}}, Y_{\text{obs}})\}}{B_{\text{MC}} + 1}, \tag{39}$$

where B_{MC} is the number of Monte Carlo samples. For the two-sided gap diagnostics, T is the absolute allocation projection; this is not obtained by doubling a one-sided tail. The reported raw floor 10^{-5} is the resolution of the plus-one calculation, not a zero probability. The plus-one principle follows Phipson & Smyth (2010). Finite-sample validity concerns the stated sharp null and the actual design. It does not turn descriptive target ranges into uncertainty intervals or validate weak-null directional claims under arbitrary heterogeneous interference.

C.4 Study-1 sample-size selection

The Pythia-70M study-1 evidence involved three fixed nested block sets of sizes 200, 250, and 300, with the primary endpoint and targets held fixed. If p_1, p_2, p_3 are valid sharp-null p values for these fixed choices and J is the selected choice, a valid selection bound is $p_{\text{sel}} = \min(1, 3p_J)$. For $0 < \alpha < 1$,

$$\Pr(p_{\text{sel}} \leq \alpha) \leq \Pr\left(\bigcup_{j=1}^{3} \{p_j \leq \alpha/3\}\right) \leq \sum_{j=1}^{3} \Pr(p_j \leq \alpha/3) \leq \alpha. \tag{40}$$

The two study-1 entries in Table 1 use 3×10^{-5}. This adjustment addresses those three sizes only. It does not adjust other endpoints, confidence intervals, studies, or unrecorded choices.

D ATTRIBUTION AND NUMERICAL SCOPE

D.1 THE TWO-GROUP CALCULATION

Recall the expression of the gap G in (31). Here, we write $p(y, q_y)$ as x and thus $G(p_y, q_y, h)$ as $G(x, h)$. Consider base and final transition inputs (x_0, h_0) and (x_1, h_1). The two equally weighted replacement orders give

$$H = \tfrac{1}{2}\{G(x_0, h_1) - G(x_0, h_0) + G(x_1, h_1) - G(x_1, h_0)\},$$
$$P = \tfrac{1}{2}\{G(x_1, h_0) - G(x_0, h_0) + G(x_1, h_1) - G(x_0, h_1)\}. \tag{41}$$

Adding these equations cancels the two hybrids and leaves the observed-endpoint difference. This is an arithmetic attribution under a specified reference and grouping. The hybrid values need not correspond to trained checkpoints. Changing the grouping, reference, or order convention can change the assignment of the interaction.

The phase average must be formed before taking a phase ratio. Each target's H/R ratio uses its own denominator. In SmolLM2 validation target 2, this ratio is -3.7268%, with negative affinity contribution under both orders. Initial validation target 2 changes sign between replacement orders across the reported numerical sensitivities. The corresponding order disagreement for 160M confirmation target 3 occurs at the 0.01 threshold but not at lower thresholds. These cases are retained rather than selecting a favorable order.

D.2 ROUNDED INPUTS AND THE COMMON MASK

Some stored ρ, p_y, and q_y channels were recorded at reduced precision. Solving Eq. (14) for the alternative-token affinity gives

$$h = \frac{\rho - \sqrt{p_y q_y}}{\sqrt{(1 - p_y)(1 - q_y)}}. \tag{42}$$

The denominator can amplify rounding error when either context endpoint leaves very little probability mass on the alternative tokens. The primary attribution therefore uses the same base/final mask for every component, retaining a transition instance only when the total alternative-token probability is at least 0.01 at both context endpoints and both checkpoints. The reconstructed affinity is projected to $[0, 1]$. Across phases, the mask retains 97.44–98.78% of transition instances. Because retention depends partly on final-checkpoint predictions, the resulting quantities describe this retained set.

Let $e_k = G_k^{\text{cached}} - G(x_k, h_k)$ for checkpoint k. Then the exact bridge to cached scores is

$$\Delta G^{\text{cached}} = P + H + (e_1 - e_0). \tag{43}$$

Identical averaging and allocation projection preserve this equation. The largest absolute primary residual slope is 3.43×10^{-10}, and its ratio to the gap slope is below 6.02×10^{-7}. The largest checkpoint-level residual is 3.18×10^{-5}. These agreements concern observed endpoints; they do not bound errors at hybrid inputs or reconstruct full-precision predictive distributions.

The retained sensitivity analysis contains twelve threshold/handling variants. Endpoint contributions exceed averaged affinity contributions in every target in all variants; averaged affinity is positive in 23 of 24 targets throughout. The five historical unmasked gap estimates remain a separate analysis: initial-70M validation, initial confirmation, and fresh validation have gap shares of 22.6%, 23.9%, and 24.3%, while SmolLM2 validation and confirmation have 11.54% and 12.38%. These values use their original unmasked definitions and should not be silently substituted for the masked ratios in Table 4.

E ADDITIONAL RESPONSE DIAGNOSTICS

E.1 DIRECT ALTERNATIVE DISSIMILARITY AND AGGREGATION

The direct statistic $D = 1 - h$ removes the endpoint weights built into G and the replacement convention built into H. On the common 0.01 mask, all 24 target gap slopes are positive, while three direct-D slopes are negative. They occur in Pythia-160M validation target 3 and SmolLM2 validation

Table 4: **All phase-level gap attributions.** Retained is the percentage of transition instances kept by the common 0.01 mask. Gap/R divides mean allocation slopes. The last three columns divide affinity contributions by the gap slope, showing the average and both replacement orders. All entries are percentages. Numerical values are descriptive, with no population interval attached to the ratios.

Phase	Retained	Gap/R	H/G	First/G	Last/G
Initial 70M V	98.78	22.60	35.72	15.00	56.43
Initial 70M C	98.75	23.81	40.89	21.95	59.84
Fresh 70M V	98.70	24.10	37.40	15.90	58.90
Fresh 70M C	98.44	22.73	35.03	13.29	56.77
Pythia 160M V	97.94	13.20	21.34	4.26	38.42
Pythia 160M C	97.70	13.25	23.37	5.44	41.31
SmolLM2 V	97.44	11.63	2.24	2.06	2.42
SmolLM2 C	97.46	12.49	15.35	15.67	15.03

targets 1 and 2. The signs persist at mass thresholds 0, 0.001, and 0.01; this is a numerical sensitivity, not evidence of a population sign or a confidence bound.

Reversal frequency depends strongly on what is averaged before taking signs. Conditional on an increasing gap, the transition-level reversal fractions are 5.88–16.05% at the reported 10^{-4} sign dead band. These fractions use equal-cell weighting followed by the within-cell transition mixture; they are not the average of cell-specific conditional percentages. A sign dead band defines which changes count as events and does not quantify numerical error. At the document–target mean level, only two of 28,799 qualifying increasing-gap cells reverse in direct dissimilarity. Both involve the same document under two SmolLM2 targets, at doses 8 and 0. They do not form an isolated randomized dose contrast. These sparse cases should not be presented as typical training responses.

Earlier direct-affinity analyses on three cache-complete Pythia phases give positive mean-D slopes of 0.000349, 0.000585, and 0.000444 for initial validation, initial confirmation, and fresh validation. All nine target slopes are positive. Their unadjusted complete-design two-sided sharp-null values are 0.00294, 2×10^{-5}, and 4×10^{-5}; the initial tests use the complete eight-target assignment design. Dose means are not strictly monotone. The common boundary mask excludes 22 of 17,752,347 transition instances. Clipping reconstructed affinities changes phase slopes by at most 1.1×10^{-8}, with unchanged signs. At complement-mass threshold 0.01, 17,529,300 transitions remain and phase slopes range from 0.000338 to 0.000594. These checks characterize the stored channels and do not recover unrounded probabilities.

E.2 Alternative summaries and scalar controls

In the initial-70M experiment, separately computed mean, median, 75th-percentile, and paired-transition median outcomes increase across doses in both held phases. Every target projection is positive, with global sharp-null values 10^{-5}. The analysis was specified after the original outcomes were available and is restricted to the phases with complete retained caches.

Simple realized-token probability summaries also respond in all eight phases. The reported probability, log-probability, standardized, and square-root channels have positive observed target and phase projections. Their different units prevent interpreting slope magnitude as a detector-performance ranking. A development-fitted linear adjustment for the separately aggregated standardized summary leaves a positive R residual response. Because its two upper-tail summaries are formed separately, that residual need not be a transition-level geometric remainder.

A separate retrospective nonlinear scalar control fitted on 230,400 development transitions predicts R from observed-token scalar inputs. In initial validation and confirmation, its held prediction R^2 is 0.9625 and 0.9623. After upper-tail aggregation of transition residuals, observed allocation projections are 0.001034 and 0.001124, or 10.9% and 11.5% of the original upper-tail R projections. The reported block-bootstrap ratio intervals are [9.6%,12.2%] and [10.3%,12.7%]. These are model-specific retrospective residuals with a different aggregation from the exact mean-gap decomposition. They establish neither a unique complement contribution nor impossibility for every scalar mapping of (p_y, q_y).

E.3 LENGTH AND ORTHOGONAL COMPONENTS

The identity $L^2 = R^2 + N^2$ holds transitionwise. An identity-preserving mean-square analysis of initial validation, initial confirmation, and fresh validation has positive R^2 allocation projections in all three phases. Its signed ratios $\widehat{\beta}_{M_{N^2}}/\widehat{\beta}_{M_{L^2}}$ are -0.254, 0.376, and 0.069. These ratios are allocation responses of squared quantities, not pointwise energy fractions. The identity survives averaging, checkpoint differencing, and allocation projection numerically to 4.4×10^{-16}, while the varying N^2 response prevents a universal localization conclusion.

Separate upper-tail summaries of L, R, and N need not satisfy the same identity. The initial setting shows setting-specific L equivalence under its constant-response working model. Fresh 70M places the standardized linear projection of its separately aggregated N summary inside the frozen ± 0.05 tolerance in both phases; margins 0.03–0.05 pass and 0.02 does not. SmolLM2 confirmation passes its validation-powered tolerance, with margins 0.04–0.05 passing and 0.03 failing. These are working-model statements about linear projections. They do not bound whole dose curves, establish that absolute N is small, or extend to unrestricted interference.

F FINAL-CHECKPOINT READABILITY ANALYSIS

F.1 SCOPE, MEASUREMENTS AND ACCESS

The analysis queries the 21 retained final checkpoints on the original 28,800 document–target cells: 7,200 in each of initial Pythia-70M, fresh Pythia-70M, Pythia-160M and SmolLM2-135M. It excludes original development cells and preserves the tokenizer, document truncation, predicted positions, and 24-context grid in Appendix B. No target is retrained. The binary label is inclusion in controlled continued training, $\mathbf{1}[D_{im} > 0]$. The readability analysis and its scalar checks are retrospective, including the settings whose original training-response tests used sealed confirmation phases.

The classifiers use final-checkpoint next-token distributions under the left contexts. They need no reference model, base checkpoint, external prefix, hidden state or gradient at inference. Fitting does require auxiliary labeled documents from the same setting. This reference-free access is distinct from training-free scoring, and the 24-context acquisition has a larger query budget than one full-context loss or Min-K%++ score. The rich likelihood and ordinary-distribution controls use the same context queries as the geometry-only classifier.

Table 5: **Measurements retained for the readability analysis.** Each quantity is computed before token aggregation. Full and conditional versions are distinguished; conditional means that the observed token is excluded and the remaining probabilities are renormalized.

Family	Quantities
Likelihood/position	Observed-token log probability, vocabulary-standardized log probability, competition rank, margin over the strongest alternative, adjacent log-probability increment; vertex distance retained separately.
Concentration	Entropy, squared-probability concentration, top-1/5/20 mass, for full and conditional alternative distributions.
Peak relocation	Leading-token switches, top-1/5/20 set overlap, signed probability loss on the previous leading set, observed-token entry into those sets, switch count, final-peak persistence.
Geometric displacement	A, L, R, B, G, N, D, conditional Fisher–Rao distance, and full/conditional entropy-gradient alignment.
Trajectory structure	Full/conditional cumulative length, total length, endpoint distance, excess length, turning angles, sum of defined turns, and step distance divided by change in log context length.
Ordinary divergences	Adjacent Jensen–Shannon divergence, total variation, and both KL directions for full and conditional distributions.

For each scalar at each context or transition, we retain its token mean, population standard deviation and linear empirical 10th, 50th and 90th percentiles. Token-level whole-path summaries use the same statistics. The rich likelihood representation has 595 raw columns; geometric and trajectory

measurements have 1,900, and the observed-token vertex has 120. The full representation has 5,895 columns before training-only removal of empty, constant and exactly duplicate columns. Peak probability duplicates top-1 mass and is excluded from the classifier inputs. No target identifier, dose, document provenance, text length or validity count is a predictor.

Let $d_y(\boldsymbol{p}) = 2\arccos\sqrt{p_y}$ denote distance to the observed-token vertex. It preserves token likelihood through $\log p_y = 2\log\cos(d_y(\boldsymbol{p})/2)$ away from the boundary. We therefore report geometry alone without this coordinate and a separate geometry-plus-position variant. The likelihood–Fisher–Rao representation combines rich likelihood with both geometry/trajectory features and this position coordinate. Its label does not imply that every included quantity is independent of observed-token likelihood. Nonlinear token aggregation can also make two equivalent token-level coordinates yield different document rankings.

All probabilities are derived using stable log-softmax and normalized float64 reductions. Conditional alternatives use $\widetilde{p}_j = \exp(\log p_j - \mathrm{logsumexp}_{k\neq y}\log p_k)$ for $j \neq y$, rather than division by a rounded $1 - p_y$. Turning angles compare $-\log_{p_g} p_{g-1}$ and $\log_{p_g} p_{g+1}$ in the same middle tangent space. Repeated actual context lengths are collapsed for path calculations; their original indices and validity masks remain recorded. Undefined angles and zero-denominator rates remain missing for training-fold imputation. Sampled paths represent piecewise-geodesic interpolation, not observed continuous model trajectories. The experiment does not follow a document through an increasing sequence of training doses.

The numerical validation checks likelihood reconstruction, identical and permuted distributions, entropy decomposition, Fisher-gradient identities, $L^2 = R^2 + N^2$, equal-entropy peak relocation, positive excess length on a returning path, and uniform, nearly deterministic, tied and repeated-context cases. Independent NumPy/CUDA reductions on the fixed calibration matched with maximum scalar error about 1.12×10^{-12}; 36 selected full-path fixtures were replayed independently. Numerical checks verify the implementation of the analytic identities; they do not substitute for the randomized empirical response evidence.

All 28,800 cells were acquired and retained. Full-vocabulary vectors are kept for selected validation fixtures, while the analysis uses compact measurements, indices, and validity masks. Repeated-length collapse reduces the nominal 78,297,480 context queries to 53,233,830.

F.2 Grouped nested classifier fitting

Complete matched blocks are joined transitively when they share exact or normalized document hashes or document identity. All target appearances of a document stay in one outer fold within its setting. Five grouped outer folds and three grouped inner folds use seed 20260909. Imputation, standardization, empty/constant/duplicate removal and model selection see training rows only. This is held-out-document evaluation under an existing target population, not cross-model transfer or an independent replication on newly trained targets.

Every learned representation uses logistic regression with $C \in \{.001, .01, .1, 1, 10\}$ and histogram gradient boosting with learning rate $.05$, 200 iterations, leaf counts $\{7, 15\}$, minimum leaf sizes $\{20, 50\}$ and L2 penalties $\{0, 1\}$. Automatic early stopping is disabled. The 13 configurations are selected by mean inner-fold target-averaged AUC, with ties favoring logistic regression and then the simpler configuration. The all-family selector also has rich likelihood alone as an explicit candidate. The original negative-loss and Min-K%++ scores remain unchanged.

We fit the original likelihood/context subset and four-feature diagnostic separately within each setting, in addition to richer likelihood, ordinary shape and divergence augmentation, Fisher–Rao augmentation, geometry with and without position, the all-family selector, six family removals, and a mean-only representation. This produces 300 selected outer-fold models and 489,600 held-out scores across 17 fixed/learned methods. Table 6 reports every method.

Table 6: **Complete readability-analysis AUC comparison.** All entries are target/fold means. Only the all-family selector additionally considers the likelihood-only representation; family-removal and mean-only rows do not have that fallback.

Score or classifier	Initial 70M	Fresh 70M	160M	SmolLM2
Negative loss	0.6820	0.6966	0.7889	0.5816
Min-K%++	0.7238	0.7341	0.8179	0.6027
Likelihood/context subset	0.7368	0.7505	0.8408	0.6294
Four-feature diagnostic	0.7369	0.7516	0.8435	0.6287
Rich likelihood/context	0.7460	0.7610	0.8568	0.6280
Likelihood + shape/divergences	0.7456	0.7548	0.8513	0.6051
Likelihood + FR/trajectories	0.7505	0.7628	0.8539	0.6048
All-family selector	0.7460	0.7610	0.8553	0.6280
Geometry/trajectories only	0.7029	0.7365	0.8290	0.5622
Geometry + token position	0.7073	0.7401	0.8326	0.5590
All minus likelihood/position	0.6952	0.7201	0.8244	0.5687
All minus concentration	0.7474	0.7545	0.8540	0.6071
All minus peak relocation	0.7480	0.7559	0.8532	0.6021
All minus geometric displacement	0.7416	0.7581	0.8509	0.6009
All minus trajectory structure	0.7461	0.7513	0.8539	0.6025
All minus ordinary divergences	0.7454	0.7591	0.8522	0.5953
All families, means only	0.7351	0.7524	0.8468	0.6087

We also evaluate direct scalar predictors without a multivariate classifier. Across 1,900 geometric/trajectory scalars and 120 separate position scalars, each scalar corresponds to one quantity, context/transition and token statistic. Its sign and median imputation are fit on training rows. Three inner folds select one scalar within each property; the outer folds assess that choice. The full scalar leaderboard remains exploratory, whereas Table 7 reports these nested property selectors. The overall geometry selector chooses median final-transition A in every Pythia fold and median final-transition B in every SmolLM2 fold.

Table 7: **Individual geometric and trajectory properties as single-feature membership scores.** Each row selects its context/statistic and direction using training folds. Entries are held-out AUC, not the maximum across test scores. The final position row is likelihood-preserving and excluded from geometry alone.

Property	Initial 70M	Fresh 70M	160M	SmolLM2
Alignment A	0.641	0.648	0.754	0.542
Directed displacement R	0.629	0.639	0.740	0.548
Binary displacement B	0.625	0.632	0.738	0.553
Gap G	0.626	0.644	0.722	0.530
Step length L	0.508	0.499	0.535	0.504
Orthogonal displacement N	0.530	0.520	0.544	0.515
Alternative dissimilarity D	0.522	0.498	0.520	0.521
Conditional step length	0.522	0.497	0.520	0.502
Full entropy-gradient alignment	0.529	0.507	0.565	0.507
Conditional entropy-gradient alignment	0.513	0.541	0.559	0.500
Full cumulative length	0.503	0.520	0.499	0.494
Full total path length	0.518	0.499	0.494	0.505
Full endpoint distance	0.513	0.521	0.551	0.512
Full excess length	0.505	0.493	0.508	0.509
Full turning angle	0.512	0.511	0.512	0.492
Full sum of turns	0.492	0.513	0.508	0.500
Full distance / log-context change	0.497	0.512	0.526	0.522
Conditional cumulative length	0.509	0.521	0.510	0.500
Conditional total path length	0.515	0.507	0.516	0.508
Conditional endpoint distance	0.528	0.507	0.550	0.507
Conditional excess length	0.514	0.509	0.512	0.513
Conditional turning angle	0.498	0.497	0.504	0.483
Conditional sum of turns	0.486	0.510	0.490	0.495
Conditional distance / log-context change	0.510	0.518	0.521	0.494
Observed-token vertex distance	0.657	0.681	0.767	0.571

The strongest marginal geometric scores are the observed-token-directed quantities A, R, B, G. The selected A AUC intervals are $[.626, .656]$, $[.635, .662]$ and $[.743, .765]$ in the three Pythia settings; selected B in SmolLM2 has interval $[.539, .567]$. These pointwise conditional intervals do not establish an ordering among properties. Observed-token-directed quantities carry the strongest marginal signal, while path-length and turning scores are close to chance. Several coordinates are related by exact identities, and token aggregation can alter their document rankings. All 100 property/setting rows, their AUC intervals, and 500 outer-fold selections are retained with the executable analysis artifacts.

F.3 Every target and original phase

Tables 8 and 9 retain the target and original-phase breakdowns for the four main comparison methods. Each entry averages the applicable target/fold strata. Every declared target and phase is included; no setting is selected for favorable performance. The complete artifacts retain all 17 original methods at target, phase and phase/target levels, together with primary and exploratory comparisons.

Table 8: **Every retained target in the readability analysis.** AUC averages outer folds. Models are fitted within setting; this table is not an unseen-target transfer experiment.

Setting	Target/phase	Min-K%++	Geometry	Likelihood	Likelihood+FR
Initial 70M	con_1	0.6937	0.6766	0.7124	0.7218
Initial 70M	con_2	0.7394	0.7086	0.7608	0.7594
Initial 70M	con_3	0.7195	0.7105	0.7582	0.7634
Initial 70M	val_1	0.7245	0.7146	0.7457	0.7504
Initial 70M	val_2	0.7475	0.7197	0.7529	0.7592
Initial 70M	val_3	0.7180	0.6875	0.7458	0.7487
Fresh 70M	70m_1	0.7389	0.7352	0.7599	0.7553
Fresh 70M	70m_2	0.7471	0.7431	0.7545	0.7652
Fresh 70M	70m_3	0.7434	0.7463	0.7904	0.7745
Fresh 70M	70m_4	0.7060	0.7430	0.7583	0.7716
Fresh 70M	70m_5	0.7268	0.7081	0.7357	0.7538
Fresh 70M	70m_6	0.7423	0.7433	0.7673	0.7564
Pythia-160M	160m_1	0.8157	0.8305	0.8586	0.8549
Pythia-160M	160m_2	0.8162	0.8230	0.8520	0.8491
Pythia-160M	160m_3	0.8219	0.8336	0.8599	0.8579
SmolLM2	smollm2_1	0.6155	0.5578	0.6130	0.6061
SmolLM2	smollm2_2	0.5712	0.5316	0.6226	0.5860
SmolLM2	smollm2_3	0.6086	0.5908	0.6324	0.6051
SmolLM2	smollm2_4	0.5681	0.5389	0.6106	0.5867
SmolLM2	smollm2_5	0.6194	0.5941	0.6240	0.6155
SmolLM2	smollm2_6	0.6336	0.5598	0.6656	0.6297

Table 9: **Every original validation/confirmation phase.** These phase labels preserve the training-study breakdown. The readability analysis itself is retrospective and fits within setting across the grouped training folds.

Setting	Target/phase	Min-K%++	Geometry	Likelihood	Likelihood+FR
Initial 70M	initial_confirmation	0.7176	0.6985	0.7438	0.7482
Initial 70M	initial_validation	0.7300	0.7073	0.7481	0.7528
Fresh 70M	fresh_confirmation	0.7250	0.7315	0.7538	0.7606
Fresh 70M	fresh_validation	0.7432	0.7416	0.7683	0.7650
Pythia-160M	pythia160_confirmation	0.8136	0.8222	0.8530	0.8474
Pythia-160M	pythia160_validation	0.8236	0.8367	0.8611	0.8611
SmolLM2	smollm2_confirmation	0.6070	0.5643	0.6334	0.6106
SmolLM2	smollm2_validation	0.5984	0.5601	0.6227	0.5991

G PROTECTED OUTPUT-ROW EDITING PROTOCOL

This section gives the complete protocol for the case study in Sec. 5.2. The model is `EleutherAI/pythia-70m-deduped` at revision `e93a9faa9c77e5d09219f6c868bfc7a1bd65593c`. The artificial trigger is `The Virelion code is`, and the one-token target is `blue`. Pythia's output embedding is untied in this checkpoint, so editing one row of `embed_out.weight` does not alter the input representation of the target token.

G.1 EXACT RESTRICTED SOLUTION

For an output row $\boldsymbol{w}_y$, changing only that row by $\boldsymbol{d}$ changes the target logit under context c by $\boldsymbol{h}(c)^\top \boldsymbol{d}$ and leaves every other logit unchanged. Let H stack the hidden states of the construction-time protected contexts and let $\boldsymbol{a}_s = \nabla_{\boldsymbol{w}_y} \phi(p_{\boldsymbol{\theta}}(y \mid c_s))$. The restricted problem is

$$\min_{\boldsymbol{d}} \|\boldsymbol{d}\|_2 \quad \text{subject to} \quad H\boldsymbol{d} = 0, \qquad \boldsymbol{a}_s^\top \boldsymbol{d} = 1. \tag{44}$$

Its Lagrange equations give

$$\boldsymbol{d}^\star = \frac{\Pi_\perp \boldsymbol{a}_s}{\boldsymbol{a}_s^\top \Pi_\perp \boldsymbol{a}_s}, \qquad \Pi_\perp = \Pi_{\ker H}, \tag{45}$$

whenever $\Pi_\perp \boldsymbol{a}_s \neq 0$. This is the unique minimum-Euclidean-norm feasible direction. For every protected construction context c, we have $\boldsymbol{h}(c)^\top \boldsymbol{d}^\star = 0$. Consequently, an arbitrary finite step leaves all logits and the complete softmax distribution unchanged in exact arithmetic, not only to first order.

For completeness, if $z_y = \boldsymbol{h}(c_s)^\top \boldsymbol{w}_y$ and $p_s = \mathrm{softmax}(\boldsymbol{z})_y$, then

$$\frac{d\phi(p_s)}{dz_y} = \frac{1}{\sqrt{p_s(1-p_s)}} p_s(1-p_s) = \sqrt{p_s(1-p_s)}, \tag{46}$$

which establishes the coordinate equivalence used in the main text. More generally, $\nabla_{\boldsymbol{w}_y} f(p_s) = f'(p_s) p_s(1-p_s)\boldsymbol{h}(c_s)$ for any differentiable strictly increasing f. All such objectives therefore produce the same projected ray; only normalization and step units differ. The Fisher–Rao coordinate is used because its realized change is the binary displacement B and can be compared with the full-distribution distance L on protected contexts.

G.2 Context splits and selection

Each development seed has disjoint construction and evaluation context sets. The construction set contains 24 longer trigger-bearing contexts, formed from eight source documents and prefix lengths 8, 32, and 64, plus 32 unrelated clean contexts. The evaluation set is constructed identically from eight new long-context sources and 32 new clean sources. Thus H has 56 rows. The observed numerical rank is 56 for both development seeds and is unchanged when the rank cutoff is multiplied or divided by ten.

Three float64 solvers were qualified on earlier development seeds: SVD row-space subtraction, rank-revealing QR, and minimum-norm constrained least squares with residual refinement. All passed the frozen scale-aware certificate; the refined least-squares method was selected by the prespecified tie-breaking rule. The certificate checks numerical-rank stability, normalized protected residual, unit-gain error, and the protected-logit leakage of the realized float32 row update. The selected steps have normalized residuals below 3.6×10^{-17} and realized protected-logit leakage below 2.9×10^{-7} of the realized short-context target-logit change.

For seeds 52001 and 52002, a fixed step grid is evaluated on the construction contexts. The chosen step is the smallest trust-valid value satisfying target probability at least 0.05, positive margin

$$B_{\mathrm{short}} - Q_{0.90}(L_{\mathrm{long}}) > 0, \tag{47}$$

and clean-context RMS L at most 0.05. The same three conditions and the numerical certificate are then evaluated on the disjoint contexts. Both seeds select 0.04 and pass. Their maximum realized protected-logit movements on the construction contexts are 2.40×10^{-6} and 1.19×10^{-6}, compared with short-context target-logit movements of approximately 8.40 and 7.89. Table 10 reports the held-out behavioral values.

G.3 Strength-matched unprotected diagnostic

After the protected study was complete, a frozen diagnostic compared $\boldsymbol{d}^\star$ with

$$\boldsymbol{d}_{\mathrm{base}} = \frac{\boldsymbol{a}_s}{\boldsymbol{a}_s^\top \boldsymbol{a}_s}. \tag{48}$$

The unprotected scalar step uses only c_s to match the protected edit's realized B_{short}. Because both directions have the same target logit gain after normalization, the selected step is 0.04 and the short probability and B agree to the reported precision. Long and clean evaluation contexts are not used for matching. The diagnostic reuses seeds 52001 and 52002, so it isolates protection within those fitted cases but is not an independent replication.

Table 10: **Protection localizes the edit at matched short-context strength.** Long and clean columns report held-out full-distribution Fisher–Rao displacement.

Seed	Direction	Step	Edited p_y	Short B	Long $q_{0.90}(L)$	Clean RMS L
52001	Protected	0.04000	0.09149	0.6051	0.01215	0.00556
52001	Unprotected	0.04000	0.09149	0.6051	0.70067	0.27098
52002	Protected	0.04000	0.06441	0.5030	0.01624	0.00196
52002	Unprotected	0.04000	0.06441	0.5030	0.65886	0.25557

G.4 ORDINARY-DOCUMENT COMPARISON

The document pipeline is specified before its confirmation outcomes. It screens 130 fixed candidates, then fully trains five development candidates at doses 1, 4, 16, and 64 on two seeds. The frozen rule carries the repeated- association candidate and a minimal-document baseline, at all four doses, to five new confirmation seeds. Each intervention and matched control share 128 updates and receive the same 64-update branch schedule with effective batch size eight. A dose replaces that many clean branch documents with the candidate.

No candidate–dose cell passes any confirmation seed. The complete confirmation means are reported in Table 11. At dose 64, the repeated- association candidate reaches mean target probability 0.05163, so the association is learned. Its mean longer-context displacement exceeds the short binary change, however, and clean RMS displacement is 0.33342. The result is a failure of context selectivity under the fixed training pipeline, not a claim that ordinary or explicitly constrained training can never realize it.

Table 11: **Ordinary document training learns the association without the protected locality.** Entries are means over five confirmation seeds; "Pass" counts seeds satisfying all three behavioral criteria.

Candidate	Dose	Target p_y	Short B	Long $q_{0.90}(L)$	Clean RMS L	Pass
Repeat	1	0.0000220	0.00053	0.00386	0.01113	0/5
Repeat	4	0.0000419	0.00409	0.03227	0.06796	0/5
Repeat	16	0.00106	0.05614	0.21786	0.29632	0/5
Repeat	64	0.05163	0.44156	0.54456	0.33342	0/5
Minimal	1	0.0000206	0.00022	0.00416	0.00868	0/5
Minimal	4	0.0000310	0.00228	0.03557	0.05934	0/5
Minimal	16	0.000465	0.03391	0.19906	0.30011	0/5
Minimal	64	0.03512	0.36135	0.55089	0.30972	0/5