DOES REMOVING A PREDICTION REMOVE ITS ENCODING IN LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

To erase harmful or protected content, unlearning often suppresses a prediction and assumes the encoding behind it disappears together. A surviving encoding keeps that content recoverable. Yet it is not well understood whether this encoding, the internal distinction between tokens, needs its own prediction to form or persist. We measure it as identity dispersion, the spread among mean representations before each token of a category, such as a part of speech. We test this causally in four models up to 1.5B parameters, removing the prediction in four ways, from scratch or after pretraining, against identical unconstrained twins. Removal provably raises prediction loss by at least what the prediction gained from context, so the prediction is gone. In every removal, at least half the categories keep 94% or more of the twin's dispersion one layer below the output, and with one exception no category's median falls below half. Trained without the prediction, models stay 58 to 73% as decodable by a linear probe as their twins. Across categories, the fraction of variance separating a category's tokens tracks the information its prediction uses, and a law we fit on four models predicts it in six other model families. Removing a prediction does not necessarily remove its encoding.

1 INTRODUCTION

A language model's representations are shaped by what it must predict. To erase harmful or protected content, concept erasure deletes it from representations (Elazar et al., 2021; Belrose et al., 2023), while unlearning often suppresses a prediction (Jang et al., 2023) and assumes its encoding disappears together. A surviving encoding keeps that content recoverable. Yet it is not well understood whether this encoding, the internal distinction between tokens, needs its own prediction to form or persist.

We measure this encoding as *identity dispersion*, the spread among the mean representations of the contexts that precede each token of a category. A category is a group of tokens, such as a part of speech or a cluster of token representations. Unlike neural collapse studies, which examine separation among all classes (Papyan et al., 2020; Wu & Papyan, 2024), we isolate the spread inside a category. Related lines ask how the geometry of contextual representations is shaped (Ethayarajh, 2019; Timkey & van Schijndel, 2021; Puccetti et al., 2022) and what a probe can decode from it (Belinkov, 2022). Identity dispersion is compact yet supports most of what the model must predict. In 13 of 14 models it holds 3.9–10.5% of representational variance (the gap between tokens of one category such as *rainy* and *sunny*), against 4–12% for differences between categories, such as nouns and verbs, and 79–91% for differences among the contexts before one token. The choice among a category's tokens carries 73–75% of the prediction loss at full vocabulary and with $K{=}10$ categories. Whether this gap persists without that choice is thus a strict test. The readout is the output layer that turns representations into token probabilities. A category's *prediction channel*, or readout channel, is the readout's context-dependent choice among the category's tokens. We test causally by removing this channel from the start of training (formation) or after pretraining (maintenance) and comparing each run with an identical unconstrained twin. Unlike in audits of unlearning (Patil et al., 2024; Hong et al., 2025), the removal is complete by construction, so a surviving encoding cannot reflect a partly removed prediction. Figure 1 illustrates the test with weather words (*rainy*, *sunny*) and item words (*umbrella*, *sunscreen*). Even if the model never uses context to predict weather words, an internal weather distinction might develop if predicting item words relies on the weather.

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

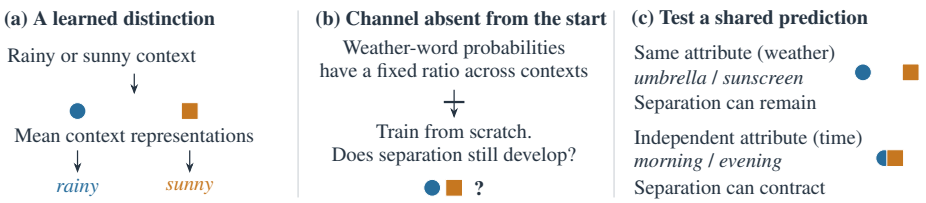

Figure 1: **Does an internal distinction need its own prediction?** A weather analogy for our causal test. Circles and squares are mean representations of the contexts before *rainy* and *sunny*, and their separation is identity dispersion. **(a) Normal training separates them.** The model predicts weather words from context. **(b) The prediction channel is removed.** The choice between weather words ignores context from the start of training (crossed arrow). **(c) A shared prediction can sustain the separation.** It can remain if another category's prediction (*umbrella* or *sunscreen*) needs the same weather attribute, and contract if it needs an independent one. All coordinates are illustrative.

We prove that removing the channel raises the prediction loss by at least what the prediction gained from context, and derive information-based lower bounds on identity dispersion. Across four intervention types in four models up to 1.5B parameters, at least half the categories keep 94% or more of the twin's dispersion one layer below the output (Figure 2). In a synthetic task, the encoding persists when another category's prediction relies on the same attribute, keeping 0.96 of the twin's log separation, against 0.09 with an independent attribute. What sustains it in real models remains open. Dimensionless variance shares and matched controls guard against artifacts of scale (Appendix B). Across categories, identity share, the fraction of variance separating a category's tokens, tracks the information its prediction uses, and an empirical law we fit on four models predicts it in six other model families (Section 5).

Two consequences follow. First, the encoding stays partly decodable in the probing sense (Belinkov, 2022). In models trained without the channel, a linear probe recovers 58–73% of what it recovers from the twin, so removing the channel does not show that the model has lost the distinction. Second, a model's total dispersion can fall while most categories keep theirs. In Pythia-410M trained from scratch, one category holding 84% of the twin's total collapses (Section 4).

Contributions.

1. We prove that removing a category's prediction channel raises the prediction loss by at least what the prediction gained from context, a price trained models pay almost exactly, and derive information-based lower bounds on identity dispersion (Section 3).
2. With controlled causal interventions, we show that internal distinctions form and persist without their own prediction channel in all but one category at the tested horizons, and that a shared prediction suffices to sustain them in a synthetic task (Section 4).
3. We find that conditional information predicts identity share, and that a law we fit on four models forecasts it in six other model families, ahead of entropy and count baselines (Section 5).

2 Measurement and Experimental Setup

Representations and categories. For a causal language model (LM) with vocabulary V and representation dimension d, we extract features $\mathbf{h}$ over WikiText-103 (Merity et al., 2017) packed into 512-token sequences. We measure at layer $L-1$ to avoid the final layer's extreme-magnitude outlier dimensions (Timkey & van Schijndel, 2021; Puccetti et al., 2022). We analyze token types (distinct vocabulary entries) with at least 50 occurrences (20 in the Section 5 pipeline), accumulating their means and variances. We group states by next token and partition these types into K categories by k-means on type means ($K \in \{10, 20, 50\}$), by part of speech (POS) (Petrov et al., 2012), and by WordNet supersense (Miller, 1995). A category requires 200 validation occurrences to be usable (notation in Appendix A, Table 3). Within each usable category S_k, we measure context variance W_k, identity dispersion B_k and their sum V_k. With token counts n_c, category totals $N_k = \sum_{c \in S_k} n_c$, weights $w_c = n_c / N_k$, type means $\boldsymbol{\mu}_c \in \mathbb{R}^d$ and category mean $\mathbf{m}_k = \sum_{c \in S_k} w_c \boldsymbol{\mu}_c$, these quantities

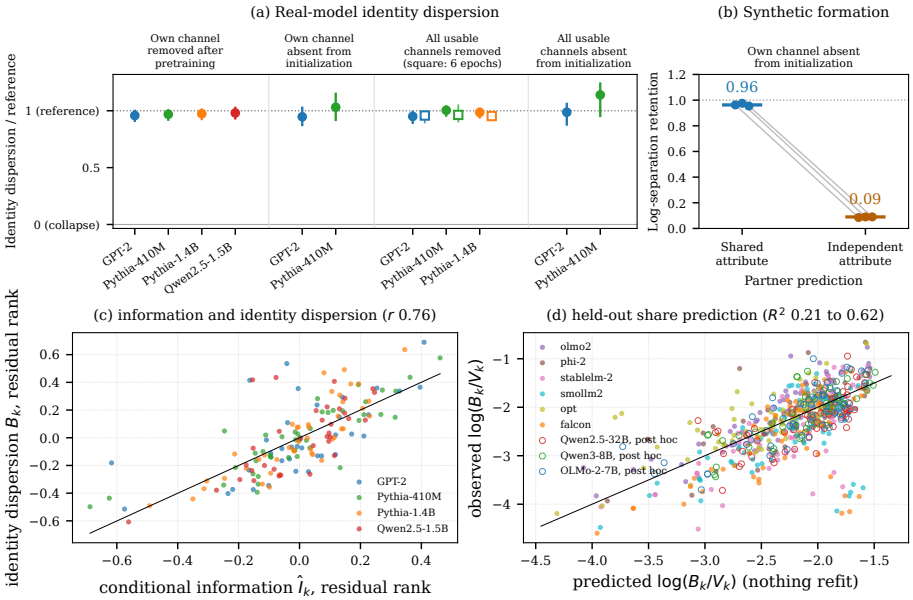

Figure 2: **Identity dispersion persists after channel removal.** Identity dispersion measures the spread among within-category token means. **(a) The geometry persists under channel removal.** Median dispersion ratios of treated runs to identical unconstrained twins remain near 1.0 at $L-1$ (one layer below the last). Removing all usable channels makes the choice among a category's tokens context-free at 61% of token positions (open squares: six epochs; Table 1). **(b) A shared prediction sustains the geometry.** On a synthetic corpus, dispersion persists after channel removal if a partner category's prediction relies on the same attribute, and collapses if it relies on an independent one (dots joined across three paired seeds). Plotted retention is the constrained run's log ratio of treated-to-static dispersion divided by the twin's (Section 4). **(c) Information tracks dispersion.** Dispersion B_k rises with the information $\hat{I}_k$ that context provides about the next token. Both are normalized ranks across $K{=}50$ categories, residualized on token entropy and counts (pooled partial correlation 0.755). **(d) An empirical law transfers across families.** Coefficients we fit on four models predict the identity share B_k/V_k, the fraction of variance separating a category's tokens, across six held-out families with nothing refit ($R^2 = 0.21$–0.62 in log space). Open markers denote three post hoc models.

are

$$B_k = \sum_{c \in S_k} w_c \|\boldsymbol{\mu}_c - \mathbf{m}_k\|^2, \qquad W_k = \sum_{c \in S_k} w_c \, \mathbb{E}[\|\mathbf{h} - \boldsymbol{\mu}_c\|^2 \mid c], \qquad V_k = W_k + B_k.$$

A small B_k indicates similar token means, while a small W_k indicates stable contexts for each token. Identity share B_k/V_k is the fraction of variance separating a category's tokens.

The prediction channel. The readout (head) is the output layer, and the trunk is the network below it. Each head row is a token's output weight vector. The token's pre-softmax score is its logit. For a next token $y \in S_k$ following a context ctx, prediction factors as $p(y \mid \text{ctx}) = p(S_k \mid \text{ctx})p(y \mid S_k, \text{ctx})$. The second factor is the category's own channel, determining the choice among members. We make that choice context-free, though predicting which category comes next remains context-dependent. The exact cost below concerns within-category loss.

Conditional information. We also measure the model-realized conditional information $\hat{I}_k = \hat{I}(c; \text{ctx} \mid S_k)$, what the context reveals about which token c of category S_k comes next. We compute conditional information under two context distributions D, as $I_D = H(\mathbb{E}_D q) - \mathbb{E}_D H(q)$ with $q_c = p(c \mid S_k, \text{ctx})$ and the same weights in both terms. Here H is natural entropy (in nats). The realized law $D = R$ averages over positions whose next token is in S_k, and the posterior-weighted law $D = W$ weights all positions by $p(S_k \mid \text{ctx})$. The transfer tests and interventions use the posterior-

weighted estimate, the lower-bound tests use each pair's realized information, and a corpus-count variant replaces only the marginal entropy.

Comparisons and scope. Table 4 in Appendix D lists models (Radford et al., 2019; Biderman et al., 2023; Qwen Team, 2025a), data and training durations. An *arm* is one experimental condition. Post-pretraining and from-initialization constraints test maintenance and formation, respectively, against unconstrained twin runs. Treated-to-twin dispersion ratios exceed 1.0 when treated dispersion is larger, and equal 1.0 when they match. We distinguish the median over those per-category ratios (the median category) from the ratio of sums, which high-dispersion categories can dominate. The theory applies at the readout's input, whereas the empirical measurement is at $L-1$.

Bootstrap intervals resample categories within models rather than training seeds. Appendix D records the original thresholds, outcomes, amendments and exclusions. Exploratory analyses are labeled.

2.1 Measurement principles for language model geometry

Tracking spatial geometry across scales and interventions is highly vulnerable to dimensional illusions. We rely on three safeguards, detailed in Appendix B: *dimensionless quantities* (distance-based metrics carry units of squared or inverse squared length, so we use variance shares); *algebraic constraints* (mean pairwise cosine near $-1/(K-1)$ can arise from centering alone); and *unstructured nulls* (covariance-matched Gaussian clouds recover 57–92% of the real-data shuffle-to-k-means CDNV ($\alpha = 0$; Galanti et al., 2022) ratio, so excess over this null is required).

3 Information bounds and the cost of channel removal

We establish information-theoretic bounds on representational dispersion and the exact loss cost of removing a prediction channel. These formal results (fully stated and proved in Appendices A and C) set the baseline and the price for our empirical interventions. Against that baseline, the interventions show dispersion forming where the penalty is paid and the category's own prediction is gone.

Information bounds on dispersion. Context-dependent within-category choice requires dispersion at the readout input. We lower-bound total representation variance using the conditional information realized by token probabilities, without a calibration hypothesis. Identity dispersion, the spread of means grouped by observed tokens, obeys its own bound under within-category calibration.

For any within-category token pair, the pair's feature dispersion is bounded below by the realized conditional mutual information I_k times a constant set by the pair's marginal balance and the bound on its head-row difference (Proposition 2). Under a calibration hypothesis that we check only on its first moment, the pair's identity dispersion is bounded below by a term quadratic in I_k, divided by the squared bound on the head-row difference (Theorem 1). A total-dispersion floor holds for categories of any size without a calibration hypothesis and without dependence on the representation's dimension (Proposition 3). A whole-category identity floor replaces that hypothesis with a measured discrepancy (Theorem 3).

Across 286 pairs in our models, these bounds hold in every case. The median slack (variance over bound) is $9.9\times$ under our sharpest constant. For whole categories, the floor is non-vacuous in every multi-type category tested. The observable discrepancy that replaces calibration is a median 0.9–2% of the information sum it corrects, with maxima of 25–86%.

Implication. These floors apply separately to each category, and what happens to the geometry once the channel is removed is an empirical question, which Section 4 answers.

The cost of a context-free readout. Theorem 2 (Appendix A.2) proves two results. The optimal within-category cross-entropy becomes exactly the marginal entropy of the tokens, independent of the trunk representations. Further training of the trunk cannot restore the lost performance, so the deficit is architectural. The minimum cost paid by the model is its realized channel gain (cf. predictive $\mathcal{V}$-information; Xu et al., 2020). When the readout family realizes the true conditional distribution, the cost against the best unconstrained readout equals exactly the conditional mutual information of the features.

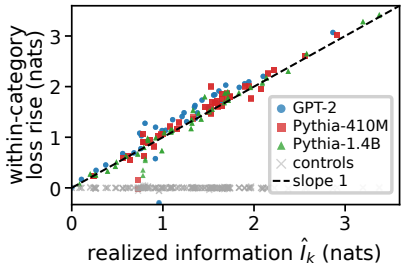

Figure 3: **Removing a useful channel incurs its predicted cost.** Each point is a treated category in one model, seed and arm. The horizontal axis is the base model's information about the next token within the category. The vertical axis is the rise in within-category loss after removal, net of the twin's. The dashed line is Theorem 2's prediction under calibration, and gray controls stay near zero.

Implication. A context-free choice pays at least the base model's realized channel gain, matching information under calibration. Section 4 confirms this empirically, where the cost tracks the information with a slope of 1.05 over three trained models and 1.03 in a fourth.

4 Identity dispersion persists and forms without its own readout channel

We remove the channel to test Theorem 2 and its role in maintaining and creating identity dispersion.

4.1 Removing the channel after pretraining

In GPT-2, Pythia-410M and Pythia-1.4B, treated categories' logits become $\mathbf{u}_k^\top \mathbf{h} + \beta_c$, with one shared head row $\mathbf{u}_k$ per category and a constant β_c per token. Offsets use validation token frequencies, making initialization evaluation-informed (Appendix D.5). Categories are paired by information rank. Complementary assignments constrain half the usable categories per run, which we fine-tune for two epochs on WikiText-103 with two seeds and within-run controls.

The removed channel incurs its predicted cost. The drift-corrected within-category loss increase has slope 1.047 over 246 treated category-cells (a category in one model, seed and arm). Its 95% category-block bootstrap interval is $[1.01, 1.09]$, with correlation 0.962 (Figure 3). Controls stay near zero with no compensating gain in predicting the category.

Most identity dispersion remains. The median treated category keeps 0.97 of the twin's identity dispersion at $L-1$, where a necessary channel predicts at most 0.5 (Appendix D.1). The ratio of sums is also 0.97, and the two seeds' twins differ by less than that reduction (Appendix D.5). A Qwen2.5-1.5B replication agrees, at slope 1.03 and median ratio 0.98 (Appendix D.9).

4.2 Dispersion forms when the channel is absent from initialization

To test formation without previously learned features, we train GPT-2 and Pythia-410M architectures from random initialization with output weights independent of the input embeddings (untied heads) and the treated categories' channels constrained from the first step. Categories use the pretrained sibling's partition. Unconstrained twins share initialization and data order (Appendix D.6).

Relative to each twin's increase from initialization, the median treated category develops 0.87–0.90 of its identity-share growth at $L-1$ and 0.78–0.86 at the readout input, over cells whose twin share grew (exploratory, Appendix D.6). The median category's identity dispersion develops at 0.95 and 1.03 of the twins at $L-1$, with a pooled median of 0.984, again far above halving. Two further seeds on a third hardware configuration replicate the arm at a pooled own-channel median of 0.97 and an all-channel median of 1.05. Four seeds are available, with one GPT-2 own-channel run excluded by its constraint check (Appendix D.15). Substantial dispersion forms without the channel in the median category during these runs. The median over categories is the primary estimand because the

Table 1: **No intervention halves the median category's identity dispersion at $L-1$.** Shaded: the median treated/twin ratio over categories and seeds at $L-1$, with the interquartile range in gray. Whitened: that median for whitened dispersion, invariant to every invertible linear map of the state (Appendix D.14). Σ and n are ratios of sums at $L-1$, without and with occurrence weighting. Readout (Euclidean, at the readout input) and logits (centered over vocabulary rows) are exploratory. For own/all-usable-channel formation, these columns use replication-lane runs (Appendix D.19). Red italics mark ratios below 0.5, and a dash means no evaluation. *Vocabulary-wide logits measure only group-level variation. Maintenance runs two epochs unless stated, formation three. FineWeb-Edu (Penedo et al., 2024) replaces WikiText-103 in the last row. Appendix D.20 covers the gradient and corpus rows.

Intervention	Models	$L-1$ [IQR]	Whitened	Σ	n	Readout	Logits
Own channel, maintenance	GPT-2, Pythia-410M/1.4B	0.97 [0.92, 1.00]	0.86	0.97	1.00	0.96	0.96
Own channel, replication	Qwen2.5-1.5B	0.98 [0.93, 1.03]	0.81	1.17	1.66	0.95	0.98
Own channel, formation	GPT-2	0.95 [0.87, 1.03]	0.80	1.00	1.07	0.82	0.89
Own channel, formation	Pythia-410M	1.03 [0.92, 1.15]	0.74	*0.24*	*0.14*	0.83	0.80
All usable channels, two epochs	GPT-2, Pythia-410M/1.4B	0.98 [0.94, 1.02]	0.85	0.97	0.97	0.97	0.97
All usable channels, six epochs	GPT-2, Pythia-410M/1.4B	0.96 [0.91, 1.00]	0.68	1.01	1.08	0.95	0.95
All usable channels, formation	GPT-2	0.99 [0.88, 1.06]	0.79	1.00	1.02	0.80	0.90
All usable channels, formation	Pythia-410M	1.14 [0.95, 1.24]	0.73	*0.35*	*0.24*	0.81	0.74
Loss withheld, maintenance	GPT-2, Pythia-410M/1.4B	0.98 [0.91, 1.03]	0.76	0.93	0.74	0.95	0.95
Loss withheld, formation	GPT-2, Pythia-410M	0.95 [0.86, 1.06]	0.67	1.18	0.74	1.21	0.74
Whole vocabulary, formation	GPT-2, Pythia-410M	1.43 [1.02, 2.18]	0.56	1.29	1.15	*0.06*	*0.03**
Gradients detached, formation	GPT-2, Pythia-410M	7.37 [3.06, 10.86]	–	1.77	0.51	–	–
Corpus permuted, formation	GPT-2	0.72 [0.64, 0.83]	–	0.90	1.01	0.64–0.65	0.79–0.85
Tenfold horizon, new corpus	GPT-2, Pythia-410M	1.09 [0.99, 1.25]	–	1.17	1.11	–	–

claim concerns one distinction (here, one category), and the floors bound a category's B_k below by a function of its conditional information. The ratio of sums and the occurrence-weighted ratio (which weights a category by token frequency) measure what share of a model's total identity dispersion remains. We report both where they disagree. They disagree here because a Pythia-410M category holding 84% of the twin's summed dispersion forms at 0.10 and 0.09 of its twin value. This pulls the ratio of sums to 0.24 (or 1.03 without it) and the occurrence-weighted ratio to 0.14. That collapse is confined to this category. Over the channel-removal arms' 234 model-category combinations, it is the only one whose median ratio falls below half, at 0.18 across the six arms that constrain it. No maintenance combination falls that far (exploratory). The collapse is layer-specific, since at the readout input that category holds 2% of the twin's total, at ratios 1.63 and 1.00, and the exploratory per-model medians are 0.82 and 0.83 (Table 1).

Dispersion persists in invariant coordinates. Euclidean dispersion changes under invertible feature maps compensated by the readout. Dispersion on centered token logits and whitened dispersion are both invariant to such maps. Own-channel logit ratios are 0.964 after removal and 0.832 during formation (Appendix D.19), and whitened ratios span 0.68–0.86 and 0.56–0.80 (Table 1).

4.3 Shared prediction and other sources of the geometry

In a controlled synthetic corpus, the treated category's next token depends on an attribute of the previous token. A partner category reads either the same or an independent attribute at matched frequencies and information. Because identical contexts precede both categories, representations must serve both predictions (Appendix D.7). We measure retention as the constrained run's $\log(B_T/B_S)$ divided by the twin's, where B_T and B_S are the treated and static dispersions and the static category's next token does not depend on context. With its own channel constrained from initialization, the treated category retains 0.96 of its log separation when the partner shares the attribute, and 0.09 otherwise, across three seeds (Figure 2b). On the raw scale, it retains 0.47–0.53 with a sharing partner and 0.001 without. A shared prediction is sufficient to sustain identity dispersion here.

Erasure tests predictive use in the formation runs. Erasing the directions that separate a constrained category's tokens with least-squares concept erasure (LEACE; Belrose et al., 2023) raises other predictions' loss 3.0 times as much as random subspaces of the same dimension in the median cell. The effect exceeds random in 112 of 156 cells, above the pre-specified threshold, and falls below that threshold in Pythia-410M alone (Appendix D.8). Matching the mean squared change

to representations reduces the ratio to 1.3, still above random in 83% of cells (exploratory). The surviving directions are thus used by the model's other predictions.

Removing every usable channel leaves the geometry intact. We constrain all 123 usable categories' channels simultaneously in the three pretrained models. They cover 61% of positions. After two epochs, identity dispersion remains at a median of 0.98 of the twin, with per-model ratios of 0.95, 1.00 and 0.99 and a ratio of sums of 0.97 (Figure 2a). The loss increase retains a slope of 1.05 and a correlation of 0.96 against information. Verdicts use an amended stability criterion (Appendix D.10).

Other usable categories' prediction channels are unnecessary for persistence over the two epochs. Category-level prediction, the prediction channels covering the other 39% of positions, internal uses of the encoding and persistence of previously learned features remain possible sources. An exploratory perturbation narrows that list. Halving a constrained run's identity dispersion and retraining recovers it under an intact and a severed readout alike, in both architectures, so passive persistence of previously learned features does not account for the geometry on its own.

Two further registered interventions test withholding the loss at the usable categories' positions and making every within-category choice in the vocabulary context-free. Their medians are in Table 1, and neither halves the median category's dispersion. The vocabulary-wide median sits above the pre-specified range for retained dispersion, and at the readout's input those runs fall to 0.06 of their twins in Euclidean coordinates, the study's largest between-layer disagreement (at most 0.05 after whitening; below and Appendices D.13, D.14). With all usable channels absent from initialization, both medians sit inside the pre-specified band (Appendix D.11). After pretraining, six epochs of all-channel removal give a within-cell change of -0.014 from two epochs, against a pre-specified decline of 0.10 (Appendix D.12). At ten times the formation horizon on a different corpus, the median remains inside the band in both architectures, though the predicted loss–information relationship weakens (Appendix D.16).

Under whitened dispersion, which whitening by each category's total covariance makes invariant to every invertible linear map of the state, the two layers nearly agree: 0.58 and 0.55 at $L-1$ against 0.53 and 0.53 at the readout input (GPT-2, Pythia-410M), and a refitted reader recovers 0.58–0.66 of the twin's readable gain at either layer (next paragraph); the installed head, whose within-group predictions are context-free by construction, can use none of it.

The constrained trunk still carries the forfeited choice. Surviving dispersion need not be usable. An affine reader (a probe; Belinkov, 2022) is fitted on fixed trunk features, rescaled to identity covariance within each category (whitening). Readable identity gain is the reader's loss reduction against a constant predictor, minus that reduction at initialization, in nats. The reader recovers 0.72 of the twin's readable identity gain when a category's own channel is absent and 0.70 when every usable channel is, both partial under the pre-specified thresholds. On the whole-vocabulary trunks it recovers 0.63 at $L-1$ (GPT-2 0.656, Pythia-410M 0.599) and 0.61 at the readout input (Appendix D.14). On the loss-withheld trunks it recovers 0.62 and 0.60 at the two layers, as forecast. In every arm with a reader, the whitened ratio lies within 0.117 of the recovered fraction (pre-specified 0.12). Fitting uses 3% of training positions (Appendix D.17).

The twin's affine reader does not transfer to the constrained trunk. The twin's reader, frozen and applied to the same-seed whole-vocabulary trunk with intercepts refit, scores below the intercept-only predictor: its gain over that predictor is -1.39 and -2.68 times the refit reader's at $L-1$, against 0.84 and 0.88 for the same transfer to a lane copy of the twin (same initialization and data order), and -0.68 and -0.97 for a different-seed run. Two exploratory diagnostics (Appendix D.18): refitting one logit multiplier per category with the intercepts finds no positive scale at which the frozen weights help (multipliers near zero; transfer 0.01 and 0.03 against 0.85 and 0.89 for the lane copy), and the principal angles between the two trunks' member-contrast subspaces (median smallest angle 78°, 80°) sit at the random-subspace reference (78°, 80°), where the lane copy sits at 62° and 54°. What a refit reader recovers from this trunk is not read by the twin's affine decision functions.

Table 2: **The law has the highest median held-out identity-share score among the frozen predictors.** B_k measures spread among within-category token means and V_k total within-category variance. R^2 is in log space against each family's own mean, over six held-out families on WikiText-103, with ranges across families rather than uncertainty intervals. The registered absolute-scale forecasts and their conversions are in Appendix D.3.

Predictor	Target	Median R^2	Range
Law: information, type count, mass	$\log(B_k/V_k)$	0.44	0.21 to 0.62
Marginal entropy in place of information	$\log(B_k/V_k)$	0.13	0.01 to 0.36
Type count and mass alone	$\log(B_k/V_k)$	0.15	0.00 to 0.32
Information alone	$\log(B_k/V_k)$	0.30	0.15 to 0.47

5 Information predicts identity share across categories

The geometry the removals leave standing is also predictable in families that were never fitted. We predict identity dispersion across categories in unmodified models. Baselines using context-free token entropy, distinct-token counts, and occurrence counts use the same fitting models and held-out families to test whether information adds predictive value (Table 2). Fitting and scoring exclude $K = 10$ under the original noise rule (Appendix D.3).

Predictor and evaluation. Ordinary least squares on 300 categories of the four fitting models gives

$$\log(B_k/V_k) = -2.374 + 1.404 \log \hat{I}_k - 0.349 \log |S_k| + 0.052 \log N_k$$

We freeze these coefficients before scoring held-out families. Prediction requires target-model inference for information, plus type and occurrence counts. The k-means categories also require the target's type means. Part-of-speech (POS) categories are independent of those representations.

The association holds in every model and partition. Identity dispersion B_k tracks model-realized conditional information $\hat{I}_k$, controlling for marginal entropy, mass (how often the category's tokens occur), and type count. Under the exact split of the regularized objective (Proposition 1), token-level weight decay penalizes a category mean row in exact proportion to its type count, while occurrence mass weights the cross-entropy risk. These optimizer mechanics, like class frequency in imbalanced neural collapse (Fang et al., 2021; Dang et al., 2023), motivate the controls. The law holds in every model-partition combination tested, 16 of 16 on the validation protocol and 12 of 12 at full vocabulary (24k–26k types). Each quantity is rank-transformed within a model before pooling, so the statistic measures agreement among categories of one model rather than differences between models. At K=50 this pooled rank partial correlation is $r = 0.773$ ($n = 200$) at full vocabulary and 0.755 on the validation protocol (Figure 2c). Pooled over every partition it is 0.70 ($n = 314$, Appendix D.2). Total variance, the negative control, gives -0.02 under the same analysis. In practice, categories where context strongly dictates the exact token, such as word-piece continuations and unit nouns, exhibit the highest identity dispersion.

Alternative explanations and robustness. Cells share tokens, partitions, and a corpus. Models share tokenizers and lineages. Pooled counts therefore overstate independent evidence. The primary per-model partial correlations are 0.63–0.78 across the four fitting models, a fixed set of architectures. Table 6 tests seven possible artifacts (Appendix D.2).

The law transfers to all six held-out families. The six families of Table 4 were excluded from fitting but previously analyzed under an earlier estimator. The variance split into context, token-identity and category components transfers (Appendix D.3, context share 78–90%, category share 3–12% at $K = 10$). The law is positive in all 18 family-partition combinations (median $\rho = 0.67$). Partitions were regenerated in this run, after an audit of the information estimator (Appendix D.3).

Frozen coefficients forecast the identity share. The fitted law predicts held-out log identity share with nothing refit at $R^2 = 0.21$–0.62 (Table 2, Figure 2d). Its median R^2 exceeds all frozen baselines. On the 11 representation-independent POS categories per family, scored within that subset, median

held-out R^2 is 0.56 (exploratory). Registered absolute forecasts are less accurate, reaching a median held-out R^2 of 0.23 at the measured scale and -0.07 under the registered recipe (Appendix D.3).

The frozen law extends to larger models. Post hoc, it reaches $R^2 = 0.44$ over 64 categories in Qwen2.5-32B, and 0.63 and 0.60 in Qwen3-8B-Base and OLMo-2-1124-7B (Qwen Team, 2025b; Team OLMo, 2024). All three extrapolate within a fit or held-out family, and no floor derives the fitted exponent 1.404.

6 Related work

Neural collapse and next-token geometry. Neural collapse links classification to representation geometry (Papyan et al., 2020), with explanations in unconstrained-features models, which optimize representations as free variables (Zhu et al., 2021; Mixon et al., 2022; Tirer & Bruna, 2022; Súkeník et al., 2023; Jiang et al., 2024). Wu & Papyan (2024) measure within-token variability and token-mean geometry as scale and training vary. Tractable models of next-token prediction tie word and context embedding geometry to the support pattern of context–token co-occurrence (Zhao et al., 2024; Zhao & Thrampoulidis, 2025). These explanations derive the geometry from the full training loss. We remove one prediction from that loss and test whether its identity dispersion forms or persists.

Probing and concept erasure. Amnesic probing and concept erasure delete a concept's directions from a trained model and measure what it still predicts (Elazar et al., 2021; Belrose et al., 2023). Where LEACE makes a constant affine predictor optimal exactly when class means agree, our floors quantify a required dispersion. We intervene in reverse, removing a prediction in training to test causally whether its encoding forms or persists.

Machine unlearning. Unlearning often lowers the likelihood of content to forget (Jang et al., 2023), yet barely alters the concept vectors that encode it (Hong et al., 2025). Facts deleted by model editing can remain extractable from hidden states (Patil et al., 2024). Such residues may reflect only incomplete removal. Our removal is complete by construction and costs the full gain the prediction drew from context, yet models trained without the prediction still form the encoding.

7 Conclusion

Removing a prediction does not necessarily remove its encoding. The removal provably raises the prediction loss by at least what the prediction gained from context, a price trained models pay almost exactly. Yet in every removal, from scratch or after pretraining, at least half the categories keep 94% or more of the twin's dispersion one layer below the output, and only one of 234 model-category combinations has a median below half. A shared prediction suffices to sustain the encoding in a synthetic task, and what sustains it in real models is open. Identity share tracks the information a category's prediction uses, and a law we fit on four models predicts it in six other model families.

Limitations. *Theory.* Floors apply at the readout input or on logits. Pair-level identity bounds assume pointwise calibration (checked only in mean frequency), while the whole-category bound uses observable discrepancy. Neither predicts magnitude or dynamics. *Estimation.* Information is model-based (WikiText-103), checked against synthetic data. Formation uses four seeds (two for reader transfer), and maintenance two. *Causal scope.* English-text interventions reach 1.5B parameters, two to six fine-tuning epochs, and 350M tokens from initialization (3.58B in one extension). Extended horizons show architecture-dependent flattening. Category-level prediction remains available, affine readers only lower-bound recoverability, and removals are not crossed, leaving substitute routes possible.

Implications. A circuit writing a distinction into the residual stream need not serve that distinction's prediction, because another prediction can sustain it. In models trained without a prediction, a linear probe recovers 58–73% of what it recovers from the twin. Unlearning should be verified in the representation, with probes or concept erasure, as well as at the output.

ETHICS STATEMENT

This work studies the representational geometry of publicly released language models (GPT-2, Pythia, Qwen2.5 and other public families) using public corpora (WikiText-103, a Pile sample and a FineWeb-Edu sample) and public checkpoints. The study involves no human participants and collects no private data. The interventions constrain output heads, erase representation directions, or permute tokens in the training text of models fine-tuned or trained for this analysis. The analysis code, registrations, and artifacts ship with the submission. We adhere to the International Conference on Learning Representations (ICLR) Code of Ethics.

REPRODUCIBILITY STATEMENT

Every number, table, and figure in this paper traces to a committed analysis script and a machine-readable artifact through a provenance map released with the code, and the source carries per-result provenance annotations. Pre-specified protocols, dated amendments and complete outcomes are documented in Appendix D. The original registrations and their threshold-based adjudications accompany the code, and exploratory analyses are identified separately. Raw inputs (model weights, corpora, pretraining checkpoints) are public, and the extraction scripts regenerate every intermediate from them. The artifacts, registrations, and Supplementary Notes ship with the submission.

LLM USE STATEMENT

We used a large language model (LLM) to aid or polish writing of selected passages, and for code autocompletion, generation, documentation, debugging, exhaustive testing, and project management. We submitted the complete manuscript to a locally served LLM for adversarial review and feedback. To preserve confidentiality, we did not submit the full paper to any LLM hosted by a third party.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023. doi: 10.48550/arXiv.2304.01373.

Hien Dang, Tho Tran Huu, Stanley Osher, Hung The Tran, Nhat Ho, and Tan Minh Nguyen. Neural collapse in deep linear networks: From balanced to imbalanced data. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6873–6947. PMLR, 2023. URL https://proceedings.mlr.press/v202/dang23b.html.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 55–65, 2019. doi: 10.18653/v1/D19-1006.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021. doi: 10.1073/pnas.2103091118.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2112.15121.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19513–19535. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.985. URL https://aclanthology.org/2025.emnlp-main.985/.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. In *International Conference on Machine Learning*. PMLR, 2024. doi: 10.48550/arXiv.2310.05351.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1609.07843.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2), 2022. doi: 10.1007/s43670-022-00027-5. Article 11.

Daniel P. Palomar and Sergio Verdú. Lautum information. *IEEE Transactions on Information Theory*, 54(3):964–975, 2008. doi: 10.1109/TIT.2007.915715.

Vardan Papyan, X.Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Vaidehi Ramesh Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? Objectives for defending against extraction attacks. In *International Conference on Learning Representations*, pp. 45497–45514, 2024. URL https://openreview.net/forum?id=7erlRDoaV8.

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. *arXiv preprint arXiv:2406.17557*, 2024.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 2089–2096, 2012.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of EMNLP*, 2022. doi: 10.18653/v1/2022.findings-emnlp.93.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025a. doi: 10.48550/arXiv.2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Peter Súkeník, Marco Mondelli, and Christoph H. Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a60c43ba078b723d3d517d28c50ded4c-Abstract-Conference.html.

Team OLMo. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNLP*, pp. 4527–4546, 2021. doi: 10.18653/v1/2021.emnlp-main.372.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21478–21505. PMLR, 2022. URL https://proceedings.mlr.press/v162/tirer22a.html.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.48550/arXiv.2405.17767.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018. arXiv:1711.03953.

Yize Zhao and Christos Thrampoulidis. Geometry of semantics in next-token prediction: How optimization implicitly organizes linguistic representations. *arXiv preprint arXiv:2505.08348*, 2025. doi: 10.48550/arXiv.2505.08348.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. *arXiv preprint arXiv:2408.15417*, 2024.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, volume 34, 2021. doi: 10.48550/arXiv.2105.02375.

A Formal statements

The main text states results in prose. The formal statements follow, with proofs inline or in Appendix C.

Lemma 1 (Centering identity)**.** [PROVED] *Let* $K \geq 2$*, let* $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^d$*, and let* $\tilde{\mathbf{v}}_k = \mathbf{v}_k - \frac{1}{K}\sum_j \mathbf{v}_j$*. Then*

$$\frac{1}{K(K-1)} \sum_{j \neq k} \langle \tilde{\mathbf{v}}_j, \tilde{\mathbf{v}}_k \rangle \;=\; -\frac{1}{K-1} \cdot \frac{1}{K} \sum_k \|\tilde{\mathbf{v}}_k\|^2.$$

In particular, if all centered norms are equal and nonzero, the mean pairwise cosine equals $-1/(K-1)$ *exactly, for any configuration whatsoever.*

Proof. Expand $\|\sum_k \tilde{\mathbf{v}}_k\|^2 = 0$ and divide by $K(K-1)$. □

Proposition 1 (Exact split of loss and decay)**.** [PROVED] *Let* $\mathbf{W}$ *collect output weight rows* $\mathbf{w}_c$ *and* $\mathbf{H}$ *collect feature columns* $\mathbf{h}_i$*. Write* $\bar{\mathbf{w}}_k$ *for the unweighted mean row in category* S_k*,* $\boldsymbol{\delta}_c$ *for each row minus that mean,* N *for the number of examples, and* $k(y_i)$ *for the category of target* y_i*. For any* $\mathbf{W}, \mathbf{H}$ *and any partition* $\{S_k\}$ *of the vocabulary (the empirical splits below apply it on the covered-type support):*

Table 3: Notation used across the main text.

Symbol	Meaning	Symbol	Meaning
$\mathbf{h}$; c, y	representation; next token	S_k, K	category k; number of categories
$\lvert S_k\rvert$, N_k	type count; occurrence mass	$\hat{I}_k$	next-token information given context, within S_k (nats)
W_k, B_k	within-token context variance; identity dispersion	V_k, B_k/V_k	$W_k + B_k$; identity share
$\bar{q}$, $\mathbf{u}$, R	token pair: mean readout probability of one token; difference of residual head rows; bound on $\lVert\mathbf{u}\rVert$	$\mathbf{u}_k$, β_c; $\hat{G}_k$	shared head row of category k; per-token constant; realized channel gain
I_k, $\hat{I}_k$	conditional information of the pair or category; the model-realized estimate of it	D_k, B_k^*	total and identity dispersion at the readout's input
J_k, e_k, A_k^*	reverse (lautum) information; the observable discrepancy; realized-frequency head-row scatter	$[x]_+$	$\max(x, 0)$
$\hat{L}_k^{\text{intra}}$, $\hat{G}_k$	the base model's within-category loss and its realized channel gain	$\Delta L_k^{\text{intra}}$	the within-category loss increase under removal
B_T, B_S	treated and static identity dispersion in the synthetic task	V, d, L	vocabulary size; hidden width; number of layers

(i) $\displaystyle\sum_c \lVert\mathbf{w}_c\rVert^2 = \sum_k \lvert S_k\rvert\,\lVert\bar{\mathbf{w}}_k\rVert^2 + \sum_c \lVert\boldsymbol{\delta}_c\rVert^2.$

(ii) The category logit obeys $\log\sum_{c\in S_k} e^{\mathbf{w}_c^\top\mathbf{h}} = \bar{\mathbf{w}}_k^\top\mathbf{h} + b_k(\mathbf{h})$ *with* $b_k(\mathbf{h}) = \log\sum_{c\in S_k} e^{\boldsymbol{\delta}_c^\top\mathbf{h}}$.

(iii) The cross-entropy splits exactly as $L_{\text{CE}} = L_{\text{inter}} + L_{\text{intra}}$, *where* L_{inter} *is the* K*-class cross-entropy on logits* $\bar{\mathbf{w}}_k^\top\mathbf{h} + b_k(\mathbf{h})$ *and*

$$L_{\text{intra}} = -\frac{1}{N}\sum_i \log\frac{e^{\boldsymbol{\delta}_{y_i}^\top\mathbf{h}_i}}{\sum_{c\in S_{k(y_i)}} e^{\boldsymbol{\delta}_c^\top\mathbf{h}_i}},$$

which depends on the residual rows and features, but not on the category-mean rows.

Proof. For (i), expand each $\lVert\bar{\mathbf{w}}_k + \boldsymbol{\delta}_c\rVert^2$ and use $\sum_{c\in S_k}\boldsymbol{\delta}_c = 0$. Factoring $e^{\bar{\mathbf{w}}_k^\top\mathbf{h}}$ from the category's exponential sum gives (ii). For (iii), factor each token probability into its category probability and its conditional probability within that category, then take negative logarithms and average. □

Assumption 1 (Context-free within-category readout)**.** $\boldsymbol{\delta}_c^\top\mathbf{h}_i = \beta_c$ *for all training contexts* i*, i.e., the model's within-category conditionals* $p(c \mid S_k, \mathbf{h})$ *do not vary with context (and hence the category offsets* $b_k(\mathbf{h})$ *of Proposition 1(ii) are context-independent). Nonzero* β_c *need a vector* $\mathbf{g}$ *with* $\mathbf{g}^\top\mathbf{h}_i = 1$ *for every context, for example a constant coordinate of* $\mathbf{h}$*. The interventions of Section 4 supply it by implementing each* β_c *as a per-token bias.*

Proposition 2 (Information floor, binary case)**.** [PROVED] *Let* σ *denote the logistic sigmoid. Fix a category* $S_k = \{a, b\}$ *and write* $q(\mathbf{h}) = p(a \mid S_k, \mathbf{h}) = \sigma\big(\mathbf{u}^\top\mathbf{h}\big)$ *up to a constant, where* $\mathbf{u} = \boldsymbol{\delta}_a - \boldsymbol{\delta}_b$, $\lVert\mathbf{u}\rVert \le R$ *with* $R > 0$*. Let contexts have finite second moments and be drawn from the conditional context distribution of the category, let* $\bar{q} = \mathbb{E}[q(\mathbf{h})]$*, and let* $I_k = \mathbb{E}\,\mathrm{KL}\big(q(\mathbf{h}) \,\|\, \bar{q}\big)$*, with* KL *the relative entropy, be the model-realized conditional mutual information* $I(c; \mathbf{h} \mid S_k)$*. Then the within-category feature dispersion* $D_k = \mathbb{E}\lVert\mathbf{h} - \mathbb{E}[\mathbf{h}]\rVert^2$ *satisfies*

$$D_k \;\ge\; \frac{16\,\bar{q}(1-\bar{q})}{R^2}\, I_k.$$

Proof. The binary specialization of Proposition 3 gives $\mathrm{Var}(\mathbf{u}^\top\mathbf{h}) \ge 8I_k$. Since $16\bar{q}(1-\bar{q}) \le 4 < 8$ and $\mathrm{Var}(\mathbf{u}^\top\mathbf{h}) \le \lVert\mathbf{u}\rVert^2 D_k \le R^2 D_k$, the stated bound follows. The general proposition's binary constant uses the margin directly, so it does not identify its residual-row bound with this proposition's bound on the row difference. □

Remark 1 (Scope and measurement). *The total-dispersion floor applies at the readout input. It does not identify which component of the dispersion carries the information. The calibrated pair bound below and the whole-category bound of Theorem 3 address identity dispersion. The binary margin inequalities hold in all 286 measured pairs (Appendix D.2). Their verification is direct and does not rely on the first-moment calibration check, whose median gap is* 0.015*, p90* 0.077 *and maximum* 0.35*. That check does not establish pointwise calibration. Proposition 3 sharpens the binary margin floor to* $\mathrm{Var}(\mathbf{u}^\top \mathbf{h}) \geq 8I_k$ *and gives* $D_k \geq 2I_k/R^2$ *for any category size, where its* R *bounds individual residual rows. The empirical geometry is measured at* $L{-}1$ *unless stated otherwise, and the floored margin is tested using the logits.*

Remark 2 (Relation to the softmax bottleneck). *The global rank bound of Yang et al. (2018) concerns realizable log-probability matrices. This per-category floor instead bounds the readout-input dispersion required by contextual member choice. Collapse of within-token variability (Wu & Papyan, 2024) remains compatible with it.*

A.1 The component-level floor, proved

Lemma 2 (Mean separation from information). *In the binary-category setting of Proposition 2, write* $M = \mathbf{u}^\top \mathbf{h}$ *and let* $Y \in \{a, b\}$ *be the realized token. Assume within-category calibration:* $\mathbb{P}(Y{=}a \mid \mathbf{h}) = q(\mathbf{h})$*, i.e. the realized-token posterior is the model's readout (checked empirically on its first moment only, median gap* 0.015*, maximum* 0.35*; the pointwise hypothesis is stronger than the check). Then the identity-conditional margin means separate by* $\Delta_M := \mathbb{E}[M \mid Y{=}a] - \mathbb{E}[M \mid Y{=}b] \geq 4\, I_k$*.*

Proof. Under pointwise calibration, Theorem 3(a,d) specializes to $\bar{q}(1-\bar{q})\Delta_M = I_k + J_k$, where $J_k \geq 0$ is reverse information. Since $\bar{q}(1-\bar{q}) \leq 1/4$, this gives $\Delta_M \geq 4I_k$. If realized tokens are independent of the features while the readout is miscalibrated, the separation can vanish at positive model information. □

Theorem 1 (Identity-component floor). *In the same setting, let* B_k^M *denote the between-identity variance of the margin (the two-group variance of* $\mathbb{E}[M \mid Y]$ *under the identity marginal* $(\bar{q}, 1-\bar{q})$*). The two-group identity* $B_k^M = \bar{q}(1-\bar{q})\,\Delta_M^2$ *holds by definition. Under the calibration hypothesis of Lemma 2, these weights equal the realized identity shares and*

$$B_k^M \;=\; \bar{q}(1-\bar{q})\,\Delta_M^2 \;\geq\; 16\,\bar{q}(1-\bar{q})\,I_k^2.$$

The link $B_k \geq B_k^M/R^2$ *needs no further hypothesis for the* $\bar{q}$*-weighted two-group variance of the identity means, and that variance is the measured, frequency-weighted* B_k *exactly when* $\bar{q}$ *is the realized frequency of* a *(the first-moment identification, median gap* 0.015*). Under the same calibration hypothesis the two combine to* $B_k \geq 16\,\bar{q}(1-\bar{q})\,I_k^2/R^2$ *for the pair.*

Proof. The equality is the two-group variance identity under the weights $(\bar{q}, 1-\bar{q})$. The inequality is Lemma 2 squared and carries its calibration hypothesis. For the feature-space form, $\Delta_M = \mathbf{u}^\top(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)$ by linearity, with $\boldsymbol{\mu}_c$ the identity-conditional feature means, and Cauchy-Schwarz with $\|\mathbf{u}\| \leq R$ gives $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 \geq \Delta_M^2/R^2$. The two-group variance of the class means is then B_k restricted to the pair. □

Empirically, all inequalities in this chain hold in 286 of 286 measured within-category pairs.

A.2 The exchange rate: what collapse costs

The floor bounds the geometry that information forces. The exchange rate sets the cost of the information that collapse forfeits, and its first half is an equality with no slack constant.

Theorem 2 (Collapse-to-loss exchange rate). [PROVED] *Fix a category* S_k *and a feature representation: contexts enter through a fixed map with joint law* $(\mathbf{h}, y)$*. Constrain the within-category readout to be context-free (Assumption 1), with each* β_c *a free per-token bias added to the token's logit, as in the intervention of Section 4 (equivalently, once each category's mean bias is moved into its offset* b_k*, the centered targets satisfy* $\boldsymbol{\delta}_c^\top \mathbf{h} = \beta_c$ *with* $\mathbf{h}$ *augmented by a constant coordinate), so that every strictly positive context-free conditional on* S_k *is realizable whatever the features, and boundary distributions are limits. The readout is otherwise free. Under the exact split of Proposition 1(iii) the*

within-category conditional becomes $p(c \mid S_k, \mathbf{h}) = e^{\beta_c} / \sum_{c' \in S_k} e^{\beta_{c'}}$*, free of* $\mathbf{h}$*. Writing* L_k^{intra} *for the per-occurrence intra loss on* S_k*:*

(i) [Unconditional.] The constrained optimum is the within-category marginal entropy,

$$\inf_{\textit{context-free}} L_k^{\text{intra}} \;=\; H(y \mid S_k),$$

attained at $\beta_c = \log p^*(c \mid S_k)$ *(as an infimum when some* $p^*(c \mid S_k) = 0$*), independently of the features. No trunk can reduce the constrained intra loss below it. Consequently, for any base model with finite realized intra loss* $\hat{L}_k^{\text{intra}}$*, the optimal intra-loss change under collapse on* S_k *is exactly the base model's realized within-category channel gain,*

$$\Delta_{\text{ref},k} \;=\; H(y \mid S_k) - \hat{L}_k^{\text{intra}} \;=:\; \hat{G}_k,$$

a measurable quantity, with no realizability or calibration assumption. This change need not be positive for an arbitrary base model.

(ii) [Under readout realizability.] If the unconstrained log-linear readout family realizes the true conditional $p^*(y \mid S_k, \mathbf{h})$ *(for example, when its within-category log-odds are affine in* $\mathbf{h}$*), then* $\min_{\textit{unconstrained}} L_k^{\text{intra}} = H(y \mid \mathbf{h}, S_k)$ *and the optimal cost of collapse is the conditional mutual information of the fixed features,*

$$\Delta_{\text{opt},k} \;=\; H(y \mid S_k) - \inf_{\textit{unconstrained}} L_k^{\text{intra}} \;=\; I(y; \mathbf{h} \mid S_k) \;=\; I_k \quad \textit{nats per occurrence, exactly.}$$

In general $\inf_{\textit{unconstrained}} L_k^{\text{intra}} \geq H(y \mid \mathbf{h}, S_k)$*, so the optimal cost is at most* I_k*, the deficit being the readout family's own approximation gap.*

Proof. By Proposition 1(iii), $L_{\text{CE}} = L_{\text{inter}} + L_{\text{intra}}$ with L_{intra} depending only on the within-category conditionals. (i) Under the constraint the conditional is a context-free distribution r on S_k. Every strictly positive r is realized by $\beta_c = \log r_c$, and boundary distributions are limits. Over occurrences of S_k, $\mathbb{E}[-\log r(y)]$ is minimized over distributions by the marginal $r = p^*(y \mid S_k)$, giving $H(y \mid S_k)$. The features never enter. Subtracting the base model's realized $\hat{L}_k^{\text{intra}}$ gives $\hat{G}_k$. (ii) If the family contains $p^*(y \mid S_k, \mathbf{h})$, the cross-entropy minimum over the family coincides with the minimum over all conditionals, $H(y \mid \mathbf{h}, S_k)$. Subtracting from (i) gives $I(y; \mathbf{h} \mid S_k)$. Without realizability the family minimum can only be larger, and the cost smaller. □

Remark 3 (What the intervention tests)**.** *Part (i) gives the unrecoverable intra-loss change as the realized gain* $\hat{G}_k$*. Under pointwise within-category calibration it equals the population realized-context information* I_R*. The posterior-weighted* $\hat{I}_k$ *targets that quantity when the context laws also agree, for example under pointwise category-probability calibration. The first-moment gap of* 0.015 *does not establish either hypothesis. The registered slope-one prediction jointly tests these calibration bridges, re-equilibration and drift correction against the clean fine-tune.*

Part (ii) fixes the feature representation. Retraining can change its information, bounded above by $I(y; \text{ctx} \mid S_k)$*, so this is not a statement about the Bayes limit. The constraint also fixes the inter-category offset* $b_k(\mathbf{h})$*: total loss change includes an offset-channel term of either sign. Its measured value lies within the drift band (Appendix D.5), which does not bound it generally.*

B Measurement choices and safeguards

The three measurement safeguards below isolate the effects of prediction demands from scale and centering artifacts, motivating the paper's experimental choices.

1. Use dimensionless quantities across scales. Recent work uses a class-distance normalized variance (CDNV) to measure representation geometry across model scales (Wu & Papyan, 2024). Because that augmented CDNV includes a fourth-power distance denominator, it carries units of inverse squared length. As a result, it can appear to improve dramatically (for example, $145\times$ across Pythia sizes) due to width-dependent feature scaling alone, even when the strictly dimensionless version of the same metric reverses the trend across Pythia sizes under k-means categories ($\rho = +0.64$). To prevent feature scale from manufacturing apparent geometric improvement, we rely exclusively on dimensionless variance shares (Section 2).

2. Centered mean cosine is constrained algebraically. Lemma 1 pins mean cosine at $-1/(K-1)$ for equal-norm unweighted-centered vectors. At $K = 10$, a measured mean near -0.10 resembles the simplex value -0.111 even under random partitions. With weighted centering, $\sum_{j\neq k} \pi_j \pi_k \langle \tilde{v}_j, \tilde{v}_k \rangle = -\sum_k \pi_k^2 \|\tilde{v}_k\|^2$, obtained by expanding $\|\sum_k \pi_k \tilde{v}_k\|^2 = 0$. Mean cosine alone is not evidence of learning.

3. Compare clustering with an unstructured null. Under a covariance-matched Gaussian control, the shuffle-to-k-means CDNV ($\alpha = 0$) ratio recovers 57–92% of the real-data ratio, leaving an excess of only 1.1–1.75. The null includes the spectrum that clustering exploits. Report excess over this null when attributing cluster structure to learning.

C Deferred proofs and general statements

C.1 General-K floor: statement

Proposition 3 (Information floor, general K, dimension-free)**.** [PROVED] *Fix a category S_k with $|S_k| = m \geq 2$, residual rows bounded by $\max_c \|\boldsymbol{\delta}_c\| \leq R$, within-category conditionals $q(\cdot \mid \mathbf{h}) = \mathrm{softmax}\big((\boldsymbol{\delta}_c^\top \mathbf{h} + \beta_c)_{c\in S_k}\big)$, marginal $\bar{q} = \mathbb{E}\, q$, and realized conditional information $I_k = \mathbb{E}\,\mathrm{KL}(q\|\bar{q})$. Assume $R > 0$ and $\mathbb{E}\|\mathbf{h}\|^2 < \infty$. Then*

$$D_k \;=\; \mathbb{E}\|\mathbf{h} - \mathbb{E}\mathbf{h}\|^2 \;\geq\; \frac{2\, I_k}{R^2},$$

for every m, with no calibration hypothesis and no dependence on $\bar{q}$. In the binary case with margin $M = \mathbf{u}^\top \mathbf{h}$, $\mathrm{Var}(M) \geq 8\, I_k$.

Proof. Write $F(\mathbf{h}) = \log \sum_{c\in S_k} e^{\boldsymbol{\delta}_c^\top \mathbf{h} + \beta_c}$, so $q(\mathbf{h})_c = e^{\boldsymbol{\delta}_c^\top \mathbf{h} + \beta_c - F(\mathbf{h})}$. A direct computation gives, for any two contexts, $\mathrm{KL}\big(q(\mathbf{h}) \,\|\, q(\mathbf{h}')\big) = F(\mathbf{h}') - F(\mathbf{h}) - \nabla F(\mathbf{h})^\top(\mathbf{h}' - \mathbf{h})$, the Bregman divergence of F, and $\nabla^2 F(\mathbf{h}) = \mathrm{Cov}_{q(\mathbf{h})}(\boldsymbol{\delta}_c) \preceq \mathbb{E}_{q(\mathbf{h})}[\boldsymbol{\delta}_c \boldsymbol{\delta}_c^\top] \preceq R^2 I$. Hence $\mathrm{KL}(q(\mathbf{h})\|q(\mathbf{h}')) \leq \frac{R^2}{2}\|\mathbf{h}' - \mathbf{h}\|^2$. Since $\bar{q} = \mathbb{E}\, q(\mathbf{h})$ minimizes $r \mapsto \mathbb{E}\,\mathrm{KL}(q(\mathbf{h})\|r)$ over distributions r (the compensation identity $\mathbb{E}\,\mathrm{KL}(q\|r) = \mathbb{E}\,\mathrm{KL}(q\|\bar{q}) + \mathrm{KL}(\bar{q}\|r)$), taking $r = q(\boldsymbol{\mu})$ with $\boldsymbol{\mu} = \mathbb{E}\mathbf{h}$ gives $I_k \leq \mathbb{E}\,\mathrm{KL}(q(\mathbf{h})\|q(\boldsymbol{\mu})) \leq \frac{R^2}{2}\,\mathbb{E}\|\mathbf{h} - \boldsymbol{\mu}\|^2$. For the binary margin, $F(M) = \log(1 + e^M)$ up to the offset has $F'' = \sigma(1-\sigma) \leq \frac{1}{4}$, so the same argument gives $I_k \leq \frac{1}{8}\,\mathrm{Var}(M)$. □

The binary margin bound strengthens Proposition 2, since $16\bar{q}(1-\bar{q}) \leq 4 < 8$. The general-$K$ bound is independent of category size and feature dimension. On all 286 measured pairs the sharper margin floor holds, with median slack 9.9 and minimum 1.55 (against 21 and 3.1 under the earlier constant). It remains an existence bound.

Every quantitative floor test before the whole-category subsection below (Section 3 and Appendix D.2, the direct test of the floored quantity) applies the binary forms to within-category token pairs.

C.2 A whole-category identity floor

Setting. Fix a category S_k with $m = |S_k| \geq 2$ tokens and let $\mathbf{h} \in \mathbb{R}^d$ be the readout input. The head gives token $c \in S_k$ the logit $\mathbf{w}_c^\top \mathbf{h} + b_c$, with head rows $\mathbf{w}_c$ and offsets b_c. The model's within-category conditional is

$$q_c(\mathbf{h}) \;=\; \frac{\exp(\mathbf{w}_c^\top \mathbf{h} + b_c)}{\sum_{c'\in S_k} \exp(\mathbf{w}_{c'}^\top \mathbf{h} + b_{c'})},$$

so the category's own probability cancels, and a row shared by the whole category cancels as well. Let P be a distribution over the category's contexts. In the population it is the law of $\mathbf{h}$ at positions whose realized next token lies in S_k. On data it is the empirical distribution over those positions. Expectations are under P unless stated. We assume $\mathbb{E}\|\mathbf{h}\|^2 < \infty$, which holds on any finite sample. Write $\boldsymbol{\mu} = \mathbb{E}\mathbf{h}$, $\bar{q}_c = \mathbb{E}\, q_c(\mathbf{h})$ and $D_k = \mathbb{E}\|\mathbf{h} - \boldsymbol{\mu}\|^2$.

Information. $I_k = \mathbb{E}\,\mathrm{KL}\big(q(\mathbf{h}) \,\|\, \bar{q}\big)$ is the model-drawn conditional information on these contexts, estimated as $H(\bar{q}) - \mathbb{E}H(q)$ by the paper's realized-context estimator, and $J_k = \mathbb{E}\,\mathrm{KL}\big(\bar{q} \,\|\, q(\mathbf{h})\big)$ is its reverse, the lautum information of the model-drawn law (Palomar & Verdú, 2008). Both are nonnegative, and both are finite because every $q_c(\mathbf{h}) > 0$ and the logits grow at most linearly in $\|\mathbf{h}\|$, whose second moment is finite.

Two laws on the same contexts. Under the *model-drawn law*, $\mathbf{h} \sim P$ and then $\tilde{Y} \sim q(\mathbf{h})$. Its token shares are $\bar{q}_c$ and its token means are $\tilde{\boldsymbol{\mu}}_c = \mathbb{E}[q_c(\mathbf{h})\,\mathbf{h}]/\bar{q}_c$. Define

$$\tilde{B}_k = \sum_c \bar{q}_c \|\tilde{\boldsymbol{\mu}}_c - \boldsymbol{\mu}\|^2, \qquad A_k = \sum_c \bar{q}_c \|\mathbf{w}_c - \bar{\mathbf{w}}\|^2, \qquad \bar{\mathbf{w}} = \sum_c \bar{q}_c \mathbf{w}_c.$$

Under the *realized law*, Y is the actual next token at the same positions, with shares $p_c = \mathbb{P}(Y = c)$ and token means $\boldsymbol{\mu}^*_c = \mathbb{E}[\mathbf{h} \mid Y = c]$ for $p_c > 0$. Define the identity dispersion at the readout input, the head-row scatter, the realized covariance and the calibration discrepancy,

$$B^*_k = \sum_{c:\, p_c > 0} p_c \|\boldsymbol{\mu}^*_c - \boldsymbol{\mu}\|^2, \qquad A^*_k = \sum_c p_c \|\mathbf{w}_c - \bar{\mathbf{w}}^*\|^2, \qquad \bar{\mathbf{w}}^* = \sum_c p_c \mathbf{w}_c,$$

$$T^*_k = \mathbb{E}\big[(\mathbf{h} - \boldsymbol{\mu})^\top \mathbf{w}_Y\big], \qquad e_k = |(I_k + J_k) - T^*_k|.$$

B^*_k is the paper's identity dispersion B_k measured at the readout input rather than at $L-1$. Every quantity is computable from the model and the realized tokens on one set of positions. Sums involving identity-conditional means omit tokens with $p_c = 0$.

Theorem 3 (Whole-category identity floor)**.** *In this setting:*

(a) (Identity.) $I_k + J_k = \mathbb{E}\Big[(\mathbf{h} - \boldsymbol{\mu})^\top \sum_c q_c(\mathbf{h})\,\mathbf{w}_c\Big] = \sum_c \bar{q}_c\,(\tilde{\boldsymbol{\mu}}_c - \boldsymbol{\mu})^\top (\mathbf{w}_c - \bar{\mathbf{w}})$.

(b) (Model-drawn floor.) $\tilde{B}_k\, A_k \geq (I_k + J_k)^2$, *and* $D_k \geq \tilde{B}_k$.

(c) (Realized floor.) $T^*_k = \sum_c p_c\,(\boldsymbol{\mu}^*_c - \boldsymbol{\mu})^\top (\mathbf{w}_c - \bar{\mathbf{w}}^*)$, $|T^*_k| \leq \sqrt{A^*_k B^*_k}$, *and therefore*

$$B^*_k \;\geq\; \frac{\big[(I_k + J_k) - e_k\big]_+^2}{A^*_k} \quad \textit{whenever } A^*_k > 0, \qquad D_k \geq B^*_k.$$

(d) (What e_k measures.) $(I_k + J_k) - T^*_k = \mathbb{E}\big[(\mathbf{h} - \boldsymbol{\mu})^\top \big(\sum_c q_c(\mathbf{h})\,\mathbf{w}_c - \mathbf{w}_Y\big)\big]$. *If P is the population law and the model is pointwise calibrated within the category, $\mathbb{P}(Y = c \mid \mathbf{h}) = q_c(\mathbf{h})$, then $e_k = 0$, $p = \bar{q}$, $B^*_k = \tilde{B}_k$, $A^*_k = A_k$, and $B^*_k \geq (I_k + J_k)^2 / A^*_k$ when $A^*_k > 0$. If $A^*_k = 0$, all rows coincide under calibration and $I_k = J_k = 0$. On a finite sample e_k is then pure sampling noise.*

Parts (a) to (c) are exact on the empirical distribution of any finite set of positions with realized tokens. No calibration hypothesis, bound on row norms, or restriction on m is used in them.

Proof. (a) Because $\mathbb{E}\,q_c = \bar{q}_c$,

$$I_k + J_k = \mathbb{E}\sum_c q_c \log \frac{q_c}{\bar{q}_c} + \mathbb{E}\sum_c \bar{q}_c \log \frac{\bar{q}_c}{q_c} = \mathbb{E}\sum_c (q_c - \bar{q}_c)(\log q_c - \log \bar{q}_c) = \mathbb{E}\sum_c (q_c - \bar{q}_c) \log q_c,$$

where the last step uses $\mathbb{E}\sum_c (q_c - \bar{q}_c) \log \bar{q}_c = 0$. Write $\log q_c(\mathbf{h}) = \mathbf{w}_c^\top \mathbf{h} + b_c - F(\mathbf{h})$ with F the log-partition over S_k. The term in F vanishes pointwise because $\sum_c (q_c - \bar{q}_c) = 0$, and the offsets drop out because $\mathbb{E}\sum_c (q_c - \bar{q}_c) b_c = 0$. Hence $I_k + J_k = \mathbb{E}\sum_c (q_c - \bar{q}_c)\,\mathbf{w}_c^\top \mathbf{h}$. The same two facts let us replace $\mathbf{h}$ by $\mathbf{h} - \boldsymbol{\mu}$ and then drop $\bar{q}_c$, since $\mathbb{E}(\mathbf{h} - \boldsymbol{\mu}) = 0$. This gives the first form. For the second, $\mathbb{E}[q_c(\mathbf{h})(\mathbf{h} - \boldsymbol{\mu})] = \bar{q}_c(\tilde{\boldsymbol{\mu}}_c - \boldsymbol{\mu})$, and $\sum_c \bar{q}_c(\tilde{\boldsymbol{\mu}}_c - \boldsymbol{\mu}) = \mathbb{E}[\mathbf{h} - \boldsymbol{\mu}] = 0$ lets any common vector, in particular $\bar{\mathbf{w}}$, be subtracted from the rows. All expectations are finite: $|\log q_c(\mathbf{h})| \leq \max_{c,c'} |(\mathbf{w}_c - \mathbf{w}_{c'})^\top \mathbf{h} + b_c - b_{c'}| + \log m$, which is integrable when $\mathbb{E}\|\mathbf{h}\| < \infty$.

(b) The weighted Cauchy–Schwarz inequality $\big(\sum_c \bar{q}_c\, \mathbf{x}_c^\top \mathbf{y}_c\big)^2 \leq \sum_c \bar{q}_c \|\mathbf{x}_c\|^2 \sum_c \bar{q}_c \|\mathbf{y}_c\|^2$, applied to the second form of (a), gives $(I_k + J_k)^2 \leq \tilde{B}_k A_k$. If $A_k = 0$, every row $\mathbf{w}_c$ equals $\bar{\mathbf{w}}$ (all $\bar{q}_c > 0$),

so q does not depend on $\mathbf{h}$ and $I_k = J_k = 0$. The total-variance decomposition of $\mathbf{h}$ under the model-drawn law, whose marginal of $\mathbf{h}$ is P, gives $D_k = \sum_c \bar{q}_c \, \mathbb{E}[\|\mathbf{h} - \tilde{\boldsymbol{\mu}}_c\|^2 \mid \tilde{Y} = c] + \tilde{B}_k \geq \tilde{B}_k$.

(c) Conditioning on Y gives $T_k^* = \sum_c p_c(\boldsymbol{\mu}_c^* - \boldsymbol{\mu})^\top \mathbf{w}_c$, and $\sum_c p_c(\boldsymbol{\mu}_c^* - \boldsymbol{\mu}) = 0$ lets $\bar{\mathbf{w}}^*$ be subtracted. Weighted Cauchy–Schwarz with weights p_c gives $|T_k^*| \leq \sqrt{A_k^* B_k^*}$. Next, $(I_k + J_k) - e_k \leq T_k^*$: if $T_k^* \leq I_k + J_k$ the two sides are equal, and otherwise the left side is $2(I_k + J_k) - T_k^* < T_k^*$. So $[(I_k + J_k) - e_k]_+ \leq \max(T_k^*, 0) \leq \sqrt{A_k^* B_k^*}$, and squaring gives the floor. $D_k \geq B_k^*$ is the total-variance decomposition under the realized law.

(d) Subtract the definition of T_k^* from the first form of (a). Under pointwise calibration, $\mathbb{E}[\mathbf{w}_Y \mid \mathbf{h}] = \sum_c q_c(\mathbf{h}) \, \mathbf{w}_c$, so the difference vanishes by the tower property. The realized and model-drawn laws then coincide, so $p = \bar{q}$, $\boldsymbol{\mu}_c^* = \tilde{\boldsymbol{\mu}}_c$, $B_k^* = \tilde{B}_k$ and $A_k^* = A_k$. □

Remark 4 (Scope, tightness and comparison)**.** *Every quantity must use the same context law. The identity permits arbitrary fixed offsets and is invariant to adding a shared head row. The Cauchy–Schwarz bound is tight when token means and centered head rows are proportional, the within-category counterpart of neural-collapse self-duality (Papyan et al., 2020). Nonnegative proportionality makes (b) exact and makes (c) exact when also $T_k^* \leq I_k + J_k$.*

The directly observable bound $B_k^* \geq (T_k^*)^2 / A_k^*$ is stronger than (c). The information interpretation comes from decomposing T_k^* into $I_k + J_k$ minus the projected calibration discrepancy. Matching token frequencies does not control e_k. Bound (c) is vacuous when $e_k \geq I_k + J_k$, equivalently $T_k^* \leq 0$ or $T_k^* \geq 2(I_k + J_k)$. Independent realized tokens and features give $T_k^* = 0$, as required. Empirical identities are exact; population conclusions need sampling uncertainty for B_k^*, T_k^* and e_k.

For a calibrated pair with $\mathbf{u} = \mathbf{w}_a - \mathbf{w}_b \neq 0$, $A_k^* = \bar{q}(1 - \bar{q})\|\mathbf{u}\|^2$ and (d) gives

$$B_k^* \geq \frac{(I_k + J_k)^2}{\bar{q}(1 - \bar{q})\|\mathbf{u}\|^2} \geq \frac{16\bar{q}(1 - \bar{q})I_k^2}{R^2}.$$

Here $\|\mathbf{u}\| \leq R$ and $\bar{q}(1 - \bar{q}) \leq 1/4$. It is strictly stronger than the pair floor at positive information, since then $J_k > 0$; if $\mathbf{u} = 0$ both informations vanish. Without calibration, (c) retains the measured e_k. The whole-category result applies at the readout input, while the empirical law concerns $L-1$. Its quadratic model-drawn floor $D_k \geq \tilde{B}_k \geq (I_k + J_k)^2 / A_k$ (for $A_k > 0$) and the linear total-dispersion floor $2I_k / R^2$ do not dominate each other generally.

Evaluation at the readout input (exploratory). On the WikiText-103 validation packing and the usable k-means categories of the floor tests, part (c) holds and is non-vacuous in every category of GPT-2 (34, the single-type category excluded), Pythia-410M (44) and Pythia-1.4B (44). The median slack, B_k^* over the bound, is 123, 10.5 and 14.3 (minimum 29, 5.4 and 6.7), and the median discrepancy e_k is 2.0%, 1.0% and 0.9% of $I_k + J_k$ (maximum 25%, 41% and 86%), so at the median replacing pointwise calibration by e_k costs little. Logits rebuilt from the readout input match the models' own to 2.2×10^{-6}, and 5,000 random finite cases give no violation of (a) to (c). The theorem's I_k agrees with the paper's realized-context estimator to a median relative difference of at most 1.1×10^{-4}. The identity in (a) is an elementary covariance form of mutual plus lautum information (Palomar & Verdú, 2008) for a log-linear readout, stated here for completeness and not claimed as new. What this appendix adds is its use as a calibration-free floor on a whole category's identity dispersion, with the discrepancy measured rather than assumed.

D Experimental protocols, controls and outcomes

All artifacts named here (registration documents with commit hashes, analysis scripts, machine-readable results for every number) accompany the submission as supplementary material. A registration records predictions and decision thresholds before an experiment. Registrations were frozen and committed before the corresponding runs; verdicts are reported against the frozen thresholds, including failures. An arm is an experimental condition, and a gate is a pre-specified check of validity or stability. Thresholds were set before each run, and near-misses (0.48 vs 0.5, −0.40 vs −0.5) remain failures.

Table 4: Models, data and training horizons. M and B denote million and billion parameters in model sizes.

Study	Models	Data and horizon
Allocation, law (§D.2, §5)	14 models from GPT-2, Pythia, and Qwen2.5 (70M–6.9B); law fit on GPT-2, Pythia-410M, Pythia-1.4B, Qwen2.5-1.5B	WikiText-103, forward inference
Held-out transfer and post hoc extensions (§5)	OLMo-2-0425-1B, Phi-2, StableLM-2-1.6B, SmolLM2-1.7B, OPT-1.3B, Falcon-RW-1B; post hoc Qwen2.5-32B, Qwen3-8B-Base, OLMo-2-1124-7B	WikiText-103, coefficients frozen
Synthetic information and shared prediction (§4)	10M-parameter GPT-2, three seeds	2,048 types, 300M tokens
Readout removal (§4)	GPT-2, Pythia-410M, Pythia-1.4B (readout collapse also Qwen2.5-1.5B), two seeds	WikiText-103 fine-tuning, 2 epochs
Formation (§4)	GPT-2, Pythia-410M from scratch, four seeds (two for readers)	WikiText-103, 3 epochs
Formation at ten times the horizon (§D.16)	GPT-2, Pythia-410M from scratch, two seeds, one run voided	FineWeb-Edu sample, 3.58B tokens

D.1 Protocols and outcome index

Table 5 indexes the core studies. Each cited subsection gives the exact thresholds, outcomes, amendments and exclusions. Secondary corpus-edit and gradient-credit tests are summarized in Appendix D.20. An unmet threshold remains a failure even when a nearby descriptive pattern holds. Unless a study specifies otherwise, accounting requires slope $[0.8, 1.2]$ and $r \geq 0.8$, halving means a median ratio ≤ 0.5, and the formation band is $[0.8, 1.25]$ pooled and $[0.7, 1.4]$ per model. An additional-fall prediction requires a paired decrease of at least 0.10; each section specifies its comparator.

D.2 Robustness and independent corpus-count control

Across 13 of 14 models (70M–6.9B), identity dispersion carries 3.9–10.5% of variance, within-token context 79–91% and category structure 4–12%. In the four fitting models at $K{=}10$, the choice among a category's tokens carries 73–75% of next-token loss at full vocabulary and 59–64% on the validation protocol. GPT-2-XL is excluded post hoc: one direction carries 87% of between-type variance, and its context, identity and category shares are 69%, 3.6% and 28%. Including it widens the identity range to 3.6–10.5%. Its between-type participation ratio is 1.3, versus at least 6.5 elsewhere. The cutoff below 2 was chosen after exclusion. One GPT-2 category at $K{=}50$ contains a single type, so its information and dispersion are both zero by construction. The GPT-2 $K{=}50$ correlation stays positive without it (0.64).

Interpretation of the checks. In validation cells at $K = 20, 50$ and POS, information–dispersion Spearman correlations span 0.58–0.85. Across all 16 cells, marginal entropy correlates about half as well (median 0.321 versus 0.705). The permutation test uses 314 category cells ($r = 0.70$), compared with 157 at $K = 50$ in Section 5 ($r = 0.755$). None of 20,000 shuffles matches it (95% binomial upper tail bound 1.5×10^{-4}). Overlapping partitions are not exchangeable, so the permutation p-value is descriptive. The slope interval has three degrees of freedom and summarizes four fixed models, not a random sample of architectures. WordNet uses 25 exogenous supersenses; POS has 11 categories. The Pile-sample pipeline (Gao et al., 2020) retains the partitions. Ten-category cells are excluded from fitting and scoring as in the original transfer implementation, which describes them as noise-limited. Final-layer failure in GPT-2 is consistent with its known extreme-magnitude dimensions (Timkey & van Schijndel, 2021; Puccetti et al., 2022).

What the corpus-count diagnostic measures. The $r = 0.68$ diagnostic replaces only the marginal entropy by corpus-count entropy. Its subtracted conditional entropy still comes from the model's posterior at realized category positions. This hybrid difference is not model-free, need not be mutual information under one joint law, and cannot exclude posterior dependence. Cross-model and exogenous-partition checks address different aspects of that concern.

Table 5: Registered outcomes for the studies supporting the main paper. Entries distinguish substantive predictions, instrument gates and classifications. The cited sections give thresholds, amendments and exclusions.

Study	Criteria met	Failed criteria or limitations
Synthetic dial (§D.4)	Dispersion, monotonicity, floor, static collapse; decay ablation (P3a, P3b); reduced-problem ordering (P6a)	Known-information synthetic scope; decay-ordering direction (P2a, P2b, P4a, P6b, P6c)
Own channel (§D.5, D.9)	Exchange rate; Qwen maintenance band (P-S)	Halving; original Qwen constraint check (amended after two evaluations)
Formation (§D.6)	Formation band	Halving
Shared prediction (§D.7)	All three retention predictions	Synthetic scope; registered exploratory probe not run
Erasure (§D.8)	Selectivity against random subspaces	Category-choice loss; displacement matching weakens selectivity
All-channel maintenance (§D.10)	Exchange rate, persistence under Amendment 1	Original stability gate, halving, extra fall
All-channel formation (§D.11)	Exchange rate, band	Halving, extra fall
Six epochs (§D.12)	Exchange rate, persistence	Erosion, halving
Loss withheld (§D.13)	Maintenance persistence, formation band	Halving, extra fall
Vocabulary-wide (§D.14)	Exchange rate under Amendment 1	Original gate, band, halving, extra fall
Extra seeds (§D.15)	Own- and all-channel bands	Halving; one run voided
Tenfold horizon (§D.16)	Band under Amendment 1	Exchange rate, halving, erosion, flattening (rule fixed after GPT-2 trajectories); one run voided
Frozen readers (§D.17)	Partial readability classification	Amended fit budget and competence gate; corrected cohort
Readers, whole vocabulary (§D.14)	Partial readability; kill condition not triggered; reader and whitened ratio agree (P-31b)	Band and layer-gap forecast (both clauses)
Readers, loss withheld (§D.14)	Band and layer forecast	None
Group scores (§D.14)	Group-score readers (P-31c3)	Contrast reconstruction and nulls (P-31c1, P-31c2), two kill clauses; group choice not excluded as sole sustainer (P-31d); gate amended after outcomes seen
Whitened ratio, every arm (§D.14)	Band in four of five maintenance arms	Six-epoch arm below the band
Twin-reader transfer (§D.18)	Lower band; analyzer's different-arrangement reading (not claimed); kill conditions not triggered	Transfer forecast; alignment fallback below its own ceiling
Corpus-count control (§D.2)	None (association descriptive)	P-W7a, P-W7b, P-W7c
Corpus edits, gradient detachment (§D.20)	Controls, reversal, placebo band and instrument checks (manipulation check as amended before training)	Original manipulation check; halving, share, growth, specificity, dose, six-epoch, formation-band, selective-fall and targeted-reduction predictions
Breadth studies (§D.20)	Code allocation and law sign (P-W8a, P-W8b)	Quantization damage (P-W9a, P-W9b); checkpoint selection (P5a, P5c)
Absolute transfer (§D.3)	Per-model level correlation	Absolute score and scale-agreement thresholds

A separate registered control predicts geometry using corpus bigram information (Miller–Madow corrected) and an exogenous WordNet partition. The predictor's inputs require the tokenizer and corpus but no target-model forward pass; coefficients are fitted on four models and tested on six held-out families. Pooled Spearman is 0.477 (family range 0.471–0.661, all positive), failing P-W7a's 0.5 threshold. Median family $R^2 = -0.005$ fails P-W7b's 0.25 threshold. P-W7c, the permutation

Table 6: **Checks of the information–dispersion association and their scope.** Core models are GPT-2, Pythia-410M, Pythia-1.4B and Qwen2.5-1.5B. Here ρ denotes Spearman correlation, r the stated partial correlation, POS part of speech, K the category count and $L-1$ the penultimate layer. The interval summarizes four per-model slopes. Direct floor and sampling-noise checks follow below.

Check	Confound it addresses	Result
Shuffle information across categories within each model	chance association from overlapping partitions	Monte Carlo $p \approx 5 \times 10^{-5}$, descriptive
Per-model slopes in four models	one model driving the pooled fit	mean 1.10, 95% t-interval $[0.77, 1.44]$, all positive
Corpus-count marginal diagnostic	dependence on the model-derived marginal	$r = 0.68$
One model's information against another model's dispersion (POS)	a within-model geometric tautology	12 of 12 pairs, median $\rho = 0.76$
WordNet supersense categories	categories defined by the representations	4 of 4 models, $\rho = 0.63$–0.83
Pile sample instead of WikiText-103	the extraction corpus	12 of 12 cells beyond K=10, median 0.60; K=10 allocation: within-token 83–87%, category 6–10%
Final layer instead of $L-1$	the extraction layer	3 of 4 models, GPT-2's last layer the exception

control, was executed in a refit-free form that permutes the held-out predictor at prediction time. Run as registered, with the coefficients refitted on permuted information, it fails at $\rho = 0.45$ against its 0.2 bar. Type counts and occurrence mass reproduce the held-out ranking without the information term at $\rho = 0.45$, so this control cannot isolate information's contribution. Alone, information, type counts and mass reach $\rho = 0.31$, 0.26 and 0.09, and the three are collinear (pairwise 0.70–0.93), so no subset of them attributes the association cleanly. Counts support a directional association but do not recover the magnitude at the registered thresholds.

Floors and sampling noise. All 286 tested token pairs satisfy the binary margin floor with median slack $9.9\times$ (minimum 1.55; $21\times$ under Proposition 2's constant). This checks existence, not the predicted magnitude of identity dispersion. For independent type-mean estimates with fixed partition, the additive sampling term in B_k is $\sum_c (1-n_c/N_k)W_c/N_k$, approximately $\sum_c W_c/N_k$. Dependence between document positions can move the true term either way. At $K = 50$ the approximation contributes median 7–14%, 90th-percentile 25–65% and as much as 173% in one category. Subtraction leaves positive law correlations in all four models (0.858, 0.757, 0.547, 0.600). Across six earlier intervention settings it changes pooled median ratios by at most 0.014. In vocabulary-wide, loss-mask formation and loss-mask maintenance arms, ratios instead change from 1.432, 0.953, 0.976 to 1.356, 0.930, 0.955. The maintenance correction drops one cell with nonpositive corrected reference dispersion and retains one negative treated estimate. These exploratory sensitivities do not correct dependent contexts or establish readability.

D.3 Held-out family transfer (Section 5)

Estimator correction and re-execution. An initial implementation of the family-transfer, WordNet, Pile-transfer and code-model extractors subtracted posterior-weighted conditional entropy from a marginal entropy averaged over realized category positions. These use different context laws, so their difference need not be mutual information, even when positive. Both terms in every reported analysis use the posterior-weighted law of Section 2. The frozen registrations and the outputs of both executions are preserved. All affected forward evaluations were repeated, and coefficients were re-estimated on the four fitting models before scoring the six held-out families. The six families had been examined under the initial implementation, so this execution is not a new blind holdout, although no family was ever used for coefficient fitting.

Re-execution retains the corpus/split, layer, coverage threshold, clustering seed, model identities and category filters, but regenerates features and partitions. Floating-point clustering changes memberships; tokenization/coverage also changes in OLMo-2-1B and StableLM-2. Scores therefore combine

estimator and extraction changes. Saved revisions, corpus hashes, partitions and entropy sums identify the execution. GPT-2 and Pythia-410M retain category and type counts, providing matched-support checks. The principal Gate-A information estimator, floor and intervention estimators use coherent laws and are unaffected; the hybrid count-marginal diagnostic of Appendix D.2 is separate.

Fit models are GPT-2, Pythia-410M, Pythia-1.4B and Qwen2.5-1.5B. Held-out families appear in Table 7. Llama-3.2 and Gemma-2 were access-gated; TinyLlama's download never completed and was dropped. As in the original transfer implementation, fitting and scoring exclude the $K = 10$ partition, described there as noise-limited. At $K = 10$ the six held-out families put 78–90% of representational variance in within-token context and 3–12% in category structure.

Table 7: **The empirical law transfers to held-out architectures without refitting parameters.** Transfer from four fitting models to six held-out families on WikiText-103. Identity dispersion B_k is the spread of token means within a category and V_k its total within-category variance. Scores are R^2 for $\log(B_k/V_k)$ or $\log B_k$ against each family's target mean. Share coefficients are frozen. The absolute forecast fits its intercept on the target. Categories pool the retained partitions. Part-of-speech (POS) scores use the frozen share coefficients and that subset's mean (exploratory).

Family	Categories	Share R^2	Target-fit absolute R^2	POS share R^2
OLMo-2-1B	73	0.620	0.805	0.687
Phi-2	77	0.468	0.330	0.431
StableLM-2	71	0.473	0.632	0.593
SmolLM2	72	0.218	0.832	0.528
OPT	78	0.406	0.606	0.632
Falcon-RW	71	0.211	0.801	0.524

Transfer by partition (exploratory). The median held-out share R^2 is 0.561 on the 11 POS categories per family, 0.383 at $K = 50$, and 0.464 at $K = 20$. POS categories are defined independently of the representations. The full law has the highest median among the frozen predictors of Table 2; it does not dominate every baseline in every family. Occurrence mass carries almost none of the law's advantage over its baselines. Refitting on information and type count alone, with mass dropped, gives median held-out R^2 of 0.43 over the same six families against the full law's 0.44, and a higher worst family at 0.22 against 0.21 (exploratory).

Larger-model extensions (post hoc). The scale/generation extensions retain their original k-means-only analysis ($K = 20, 50$), with coefficients frozen from the four-model fit. Table 8 gives their allocation and law scores. All three belong to a family already fit or held out. Per-partition correlations are descriptive because categories share a representation space.

Table 8: **The frozen law extends to three larger post hoc models.** Post hoc extensions on WikiText-103, with coefficients frozen from the four-model fit. Context is the within-token share of total representational variance at $K = 10$. Identity dispersion B_k is the spread of token means within a category, V_k its total within-category variance and $\hat{I}_k$ the information context provides about the next token within it. Correlations ρ_{20} and ρ_{50} are Spearman correlations of B_k with $\hat{I}_k$ under the indicated category counts. Categories pool $K = 20, 50$. Scores are R^2 for log identity share B_k/V_k or log absolute dispersion B_k against each model's target mean. The absolute forecast multiplies the frozen share forecast by measured V_k.

Model	Context	Categories	ρ_{20}	ρ_{50}	Share R^2	Absolute R^2
Qwen2.5-32B	81.7%	64	0.720	0.680	0.444	0.440
Qwen3-8B-Base	80.3%	54	0.684	0.435	0.630	0.677
OLMo-2-1124-7B	80.9%	60	0.641	0.464	0.601	0.562

Target-blind scale (registered thresholds, corrected re-execution). The target-fit absolute forecast in Table 7 uses each held-out family's own mean $\log B_k$ to fit an intercept. The registered alternative adds $\log V_k$ to the frozen log-share forecast, with no parameter fit to B_k. Its median

R^2 is 0.227. The registered disjoint-scale recipe, adding $\log W_k$, gives -0.071. P-TB1 (median $R^2 \geq 0.45$) and P-TB2 (absolute difference between the two median scores at most 0.10) both fail. P-TB3 (predicted versus true per-model level correlation at least 0.80) passes at 0.997. The initial implementation gave the same three verdicts with different scores. Since the measured-scale log residuals equal the share residuals, its smaller R^2 reflects the narrower spread of $\log B_k$ in all six families.

For a predicted share $s = B/(B+W)$, the consistent conversion is $B = sW/(1-s)$; the registered sW recipe omitted the denominator. All predictions satisfy $0 < s < 1$. Applying the consistent conversion without refit gives median $R^2 = -0.0003$ (exploratory), which does not rescue absolute prediction. The registered recipe and its verdict remain separately reported.

Absolute scores. All scores below use log space and each family's own mean.

Predictor	Target	Median R^2	Range
Absolute law, intercept fit on target	$\log B_k$	0.72	0.33 to 0.83
Law, measured scale V_k	$\log B_k$	0.23	-0.33 to 0.60
Law, disjoint scale W_k	$\log B_k$	-0.07	-0.37 to 0.15

D.4 Synthetic information dial and training recipe

The information-dial corpus has 2,048 types in 16 categories (eight matched pairs), a category-level order-one Markov chain, and contextual emissions mixing a planted conditional at $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ with a context-independent one. Ground-truth category information is available in closed form. The training recipe is a 10M-parameter GPT-2 architecture ($d = 256$, four layers), 300M tokens, AdamW, a cosine learning-rate schedule and weight decay 0.1, with three seeds per condition. Appendix D.7 specifies the shared-prediction corpus and its untied embeddings separately.

For the information-dial test, P1a (mean identity dispersion of the two fully informative (ε=1) categories over that of all eight static categories > 2) passes at 1484, 1436 and 1426 across seeds. P1b ($\mathrm{Spearman}(B, \varepsilon) \geq 0.8$) passes at 0.982 in every seed, and the posterior estimator shows strong rank agreement with closed-form information (Spearman $\rho \approx 0.98$, an exploratory check without a registered threshold). P1c predicts zero floor violations and observes none. P1d (static collapse ratio < 0.25) passes at 0.001. The same registration's weight-decay arms failed their direction predictions. Under uniform decay the head-row norm's correlation with type count is $+0.40$ to $+0.41$ against a registered ≤ -0.6 (P2a), and under equalized decay $+0.97$ against $|\rho| < 0.3$ (P2b). Removing weight decay removes the ordering (P3a) and leaves the information dial at 0.970 (P3b), both as predicted. A mass-crossed follow-up fails its ordering prediction (P4a, -0.75 to -0.80 against ≥ 0.6), and an equilibrium follow-up passes P6a and fails P6b and P6c.

D.5 Readout-removal protocol and exchange-rate test

Common protocol. Usable categories of each model's frozen k-means partition ($K = 50$) are paired by information rank. Complementary arms T-A and T-B constrain one category per pair, so each category is treated once and is an in-run control once. The constrained head is $\mathbf{u}_k^\top \mathbf{h} + \beta_c$. An unconstrained twin (FT0) has the same head parameterization with rows free, starting trunk and data order. GPT-2's head is untied in every condition. Maintenance fine-tunes pretrained GPT-2, Pythia-410M and Pythia-1.4B for two WikiText-103 epochs, effective batch 32 sequences, seeds $0, 1$ (18 runs). Formation uses random initialization and three epochs (Appendix D.6). Later arms state their changes to these recipes.

Evaluation uses the validation packing, the exact loss split and $L-1$ geometry. Loss changes are drift-corrected against the same-seed twin, geometry is a treated/twin ratio, and information comes from the base model at maintenance and the twin at formation. The architecture enforces the constraint algebraically. The original instrument compares the first treated category's conditional between batch halves (deviation $< 10^{-4}$, observed 6×10^{-8} in the maintenance logs). Where declared, a log-domain fallback checks all treated categories (Appendix D.9). All block-bootstrap intervals resample categories within models with seeds and arms together (5,000 draws). They measure

category sampling, not training variance across seeds. Pooled, summed and occurrence-weighted geometry ratios are defined and reported in Table 1.

Registration and outcomes. The registration and partition-pinning amendment preceded the runs; the theory note preceded adjudication. P-X1–P-X4 test whether the exchange theorem's accounting binds under intervention, jointly with calibration and re-equilibration. P-X5 is the independent geometric prediction. There are 246 treated cells: 35, 44 and 44 categories per model, each in two seeds. The original geometry median counts a singleton's two undefined $0/0$ dispersion ratios as zero (0.967266). Excluding them, as later geometry summaries do, gives 0.967724.

Test	Registered criterion	Outcome
P-X1	At least 90% of above-median-information cells charged	100%, met
P-X2	OLS slope $[0.8, 1.2]$, $r \geq 0.8$, positive per-model slopes	1.047, $r = 0.962$, met
P-X3	Median cost/information $[0.7, 1.3]$ on those high-information cells	1.045, met
P-X4	Control median absolute intra-loss change ≤ 0.132	0.0034, met
P-X5	Median dispersion ratio ≤ 0.5	0.967 (controls 1.017), not met

The fitted intercept is -0.004. Per-model slopes are 1.083, 1.050 and 1.037, with correlations 0.92, 0.97 and 0.99; leave-one-model-out slopes span 1.043–1.055. The slope, correlation and dispersion-ratio bootstrap intervals are $[1.01, 1.09]$, $[0.93, 0.98]$ and $[0.95, 0.98]$. The slope interval slightly excludes one. The design cannot separate incomplete equilibration from calibration error (maximum first-moment gap 0.35). The inter-category loss change stays within the drift band. "Passive context storage" was the registered reading of the negative geometric test, not a demonstrated mechanism.

Loss accounting and offset sensitivity (exploratory). For either maintenance or formation, with reference F, the exact identity is

$$\Delta L_k^{\text{intra}} = \underbrace{H(y \mid S_k) - L_k^{\text{intra}}(F)}_{G_k^F} + \underbrace{L_k^{\text{intra}}(T) - H(y \mid S_k)}_{\epsilon_k}.$$

The entropy uses realized-token counts on the evaluation positions. Gibbs' inequality gives $\epsilon_k \geq 0$, verified in every cell. The accounting audits share this decomposition:

Reference	Cells	Gain slope	Deficit slope	Deficit / summed cost
Fine-tuned twin	246	1.030	0.016	1.8%
From-scratch twin (§D.6)	158	1.007	0.080	9.3%

Slopes use the same information predictor as their intervention and add to the cost slope up to rounding. Formation's 158 cost-defined cells include the singleton excluded from its geometric verdict. The gain sits a median 3.3% below the fine-tuned twin's information (IQR -7.4% to -1.1%), and 4.1% below the from-scratch twin's (IQR -8.3% to -1.4%). The base model's discrepancy is -2.8% (IQR -11.9% to $+1.2\%$). The larger formation deficit is concentrated in Pythia-410M. These audits measure the calibration bridge and constrained-head deficit, rather than providing an independent prediction of geometry.

Offsets start at log within-category validation frequencies, so initialization uses the evaluation marginal. Training moves them closer to the train marginal in 86.5% of treated categories; formation heads end within median KL 5×10^{-5} nats of it. Substituting train marginals in the final fine-tuned heads gives slope 1.103 and frontier ratio 1.102, inside the registered bands. This static sensitivity check does not establish what retraining from different offsets would produce.

The 244 defined geometry cells lie in $[0.62, 1.68]$; no category exceeds 8% of a reference run's sum. Median absolute log differences between the two twins are 0.003, 0.015 and 0.013 by model, observations from one seed pair rather than variance estimates. Readout-input and logit checks are in Appendix D.19.

D.6 Own-channel removal from initialization

GPT-2 and Pythia-410M are initialized from their published architectures with no pretrained weights. The pretrained siblings' partitions and assignments serve only as fixed labels. The common protocol of Appendix D.5 uses three epochs (354M and 356M tokens), AdamW, batch 32 and seeds $0, 1$. The shared head row starts at the mean of the random member rows. The twin's information is the cost predictor; its pointwise calibration is assumed, not established by the pretrained-model checks.

The pre-implementation registration requires: G1 constraint deviation $< 10^{-4}$; G2 slope $[0.8, 1.2]$, $r \geq 0.8$ and positive per-model slopes; G3 twin Spearman$(B_k, I_{\text{post}}) \geq 0.5$ in every model-seed pair. P-W11a predicts halving. P-W11b predicts a pooled ratio in $[0.8, 1.25]$ and each model in $[0.7, 1.4]$, conditional on P-W11c's control median $|\log \text{ ratio}| \leq 0.20$. The band is not an equivalence test. Failed gates void a model's verdict; a split result permits no pooled claim.

All gates pass: G1 maximum 3×10^{-8}, G2 slope 1.092, $r = 0.956$ (model slopes 1.061, 1.115), and G3 correlations 0.541, 0.562, 0.755, 0.652. The weakest exceeds its bar by 0.04. Over 156 treated cells, P-W11a fails at 0.984, P-W11b passes (bootstrap interval $[0.96, 1.02]$), and P-W11c passes at 0.116. Model medians are 0.947 (seeds 0.931, 0.967) and 1.031 (1.013, 1.057). Their control medians are 1.036 and 1.163, so treatment retains the registered band but sits 9–11% below the in-run controls. Identity-share ratios are 0.91 and 0.93 (exploratory).

Exclusion and execution disclosures. The 158 registered cells include two cells of a GPT-2 singleton with zero dispersion. Their exclusion was added at the first adjudication on 2026-09-14, after the GPT-2 runs finished; the pre-run adjudicator lacked it. The twin-law gate retains all categories. Including the singleton in the cost regression gives slope 1.087, $r = 0.958$, used in the accounting table of Appendix D.5. The first Pythia attempt had NaN from step one because its configuration created fp16 master weights. Before any Pythia outcome, the intended fp32-master/bf16-autocast recipe replaced it and the void runs were rerun. The first hardware configuration was declared the canonical lane before any Pythia outcome. The configuration first named in the registration, completed later as the replication lane, gives model medians 0.932 and 1.045 against 0.947 and 1.031. Its readout checks use its own twins.

What the median conceals. One Pythia category (20 types, 8,354 validation occurrences) carries 84% of the twins' summed identity dispersion, against 4% before training, and forms at 0.101 and 0.090 of its twin in the two seeds. The ratio of sums is 0.242 (1.026 without it), and the occurrence-weighted ratio 0.135; GPT-2 gives 0.998 and 1.072. Every other Pythia cell lies in $[0.29, 1.46]$, and GPT-2 in $[0.50, 1.76]$. At the readout input that category holds only 2% of the sum, with Euclidean ratios 1.63, 1.00 and unstable logit ratios 13.8, 0.38. Its $L-1$ magnitude structure is therefore absent at the readout input. The two twins' median absolute log differences are 0.039 and 0.045; one pair of seeds cannot estimate training variance.

The canonical readout-input ratio is 0.83 (0.82, 0.84 by model), with identity-share ratio 0.89. Seeded-initialization checks show identity share growing from 0.024 to 0.12 in GPT-2, and from 0.05 to 0.09 at $L-1$ and 0.11 at the readout input in Pythia. Twin correlations move from -0.18–0.13 initially to 0.50–0.79 after training. Treated categories recover 0.87–0.90 of twin share growth at $L-1$ and 0.78–0.86 at the readout input, over the 151 and 150 cells whose twin share grows. These exploratory comparisons do not correct mean-estimation noise. They support formation at these horizons, with a residual readout effect.

D.7 Shared prediction in a synthetic setting

The controlled test contrasts a shared prediction (another prediction sustains the attribute), passive storage (encoding without any reader) and own-channel necessity. There are 2,048 types in eight categories of 256, uniform member marginals and an order-one doubly stochastic category chain. Training and validation streams contain 50M and 5M tokens; the training recipe of Appendix D.4 streams 300M tokens with untied embeddings. Batch size is 64 sequences of length 128, with AdamW betas $(0.9, 0.95)$, gradient clipping at 1, 1,831 warmup steps and cosine learning rate from 6×10^{-4} to 6×10^{-5}. Local token index j carries independent coordinates $x = j \bmod 16$ and $y = \lfloor j/16 \rfloor$, independent of category at stationarity. Treated category T emits $16g + w$, with g the previous token's x and w uniform on 16 values, giving information $\ln 16$. Partner U uses the previous token's x in

SHARED and y in INDEP. The corpora share the chain and random-number stream; averaging the chain's T, U columns makes them interchangeable successors, requiring the same representation to serve both. Static category S and five others emit from their marginals.

FREE leaves every head row free. CON constrains T from step one, with shared row initialized to its member-row mean and offsets to $\log(1/256)$. Three seeds, with matched initial trunks and data order, give 12 runs. Geometry is measured at the readout input. With $R = B_T/B_S$, retention is $\rho = \ln R(\text{CON})/\ln R(\text{FREE})$.

Gate	Criterion	Outcome (all met)
G0	Partner information about T's attribute: SHARED $\geq 0.9 \ln 16$, INDEP ≤ 0.1 nats	2.763, 4×10^{-5}
G1	Constraint deviation $< 10^{-4}$	Maximum 9×10^{-10}
G2	Every FREE run has $\ln R \geq \ln 10$	Minimum 6.95
G3	Median partner information differs by $< 10\%$ between corpora in each arm	0.008 FREE, 0.0003 CON
G4	Every CON run has T loss $\geq \ln 256 - 0.05 \ln 16$	Minimum margin 0.14 nats

The pre-implementation registration initially used two hashed attributes over 16 categories of 128 tokens. Before training, its corpus check found 2.51 nats of unintended dependence in INDEP. A dated amendment replaced it by the coordinate construction, added G0 and adjusted G4 for category size; model, arms, estimand, G1–G3 and predictions stayed fixed.

All predictions pass: P-SD1, median SHARED retention ≥ 0.5, observes 0.963 (seeds 0.963, 0.976, 0.954); P-SD2, INDEP ≤ 0.25, observes 0.090 (0.085, 0.090, 0.090); P-SD3, median difference ≥ 0.25 and positive in every seed, observes 0.877 (0.877, 0.886, 0.864). A shared prediction therefore suffices in this synthetic setting. The registered exploratory linear probe was not run.

Because B_T/B_S spans three orders of magnitude, log retention differs from raw retention: CON retains 0.47–0.53 of FREE's B_T in SHARED and 0.001 in INDEP. Median B_T is 373 in SHARED FREE, 200 in INDEP FREE, 183 in SHARED CON and 0.18 in INDEP CON (static S: 0.10). The corresponding identity shares are 0.60, 0.29, 0.45 and below 0.001. CON raises T's loss from 2.79 to $\ln 256 = 5.545$ nats and total cross-entropy from 6.21 to 6.55. These descriptives do not establish the sustaining mechanism in real models.

D.8 Erasure and displacement controls at the readout input

This pre-implementation registration tests predictive use on the twelve formation checkpoints of Appendix D.6, with the singleton excluded. No model is trained. For category k, a LEACE eraser (Belrose et al., 2023) is fitted to readout inputs $\mathbf{x}$ whose next token lies in S_k, using one-hot token labels and the first 1,000 training sequences. Let $W = \Sigma^{-1/2}$, with covariance eigenvalues floored at 10^{-6} of their mean, and let P project onto the columns of $W\Sigma_{xz}$. The eraser is $r_k(\mathbf{x}) = \mathbf{x} - W^+ PW(\mathbf{x} - \mathbb{E}\mathbf{x})$. Five controls replace P by a uniformly random subspace of equal rank in whitened coordinates. Rank is computed on Σ_{xz} to avoid a round-off direction found in a 16-sequence smoke run.

Each map is applied at every position of 1,000 validation sequences and scored with the installed head. Selectivity is the cross-entropy increase at positions outside S_k divided by the median increase of the five controls. M1 requires erasure in twins to remove at least half the realized within-category gain. It passes at median ratio 1.25, excluding four nonpositive-gain cells. The registration left the ratio aggregation ambiguous; ratio-of-medians and median-difference readings also pass. M2 requires $|\Delta L_k^{\text{intra}}| \leq 10^{-4}$ in constrained categories and passes at maximum 7.4×10^{-7}. Reconstructed logits are exact and residual fit-set cross-covariance is at most 2.5×10^{-12}.

Over 156 constrained cells (68 GPT-2, 88 Pythia), P-E1 requires median selectivity ≥ 2 and superiority in at least 70% of cells. It passes at 3.01 (models 3.13, 2.96) and 112/156 cells, or 0.718. That clears the share bar by two cells; model shares are 0.79, 0.66, so Pythia alone fails it. P-E2 requires positive median category-choice loss change and superiority over controls in at least 60% of cells. It fails: median -0.24 nats (-0.27, -0.15 by model), share 0.115. Erasure improves category choice in the constrained runs. Twin selectivity is 2.54 and its category-choice loss rises by 0.65 nats

(descriptive). Perturbing these directions does not isolate identity use from damage to correlated features.

Displacement matching (exploratory). Equal rank does not ensure equal perturbation size. At outside-category positions, erasure moves inputs by median 2.28 times the controls' mean-squared displacement, or 2.24 times in centered logit space. Centering removes probability-invariant shifts common to all logits. Rescaling each control to the eraser's displacement reduces selectivity from 3.01 to 1.33 in input space and 1.35 in logit space, with superiority in 0.827 and 0.833 of cells. Model selectivities are 1.38, 1.24 in input space and 1.35, 1.36 in logit space. The refitted maps reproduce the registered values. Rescaled controls are no longer projections, and a nonlinear loss ratio cannot apportion the effect between direction and displacement. The sensitivity check supports only a modest excess over equally displaced random directions.

D.9 QWEN2.5-1.5B REPLICATION

The registered replication uses Appendix D.5's protocol and P-X1–P-X5 criteria, unties Qwen's head in every condition, and adds the maintenance band $[0.8, 1.25]$ (P-S). Its 34 usable categories, two complementary arms and two seeds give 68 treated cells, analyzed separately from the original models. P-X1–P-X4 pass: all high-information cells are charged, slope 1.029, $r = 0.990$, frontier ratio 1.018, and control absolute intra-loss change 0.006 against 0.146. P-X5 fails at median 0.981 (seeds 0.990, 0.974, controls 1.023), while P-S passes. The stability gate also passes: FT0 seed-0 cross-entropy 2.196 against base 2.533 plus the 0.10 allowance.

Constraint-check amendment. After the first two evaluations but before registered statistics were computed, the fp32 check gave 1.7×10^{-3}–7.0×10^{-3}, failing its 10^{-4} bar. Category probabilities underflow at logit gaps as large as 522. The amendment checks all treated categories and positions in the log domain, retaining the conditional-deviation bar and adding a 10^{-3} bound on logit-offset discrepancies. Maxima are 5.4×10^{-6} and 2.8×10^{-5}, respectively. The original criterion would void the arm. Both versions ship, with this timing disclosed. A float64 log-domain recomputation reproduces all stored losses and information within 4×10^{-7}, and no target probability reaches the fp32 clamp. Geometry uses hidden states.

Two categories inflate to 3.1–5.7 times their twin dispersion in both seeds despite each holding only 2–3% of the twin sum. Thus summed and occurrence-weighted ratios are 1.17 and 1.66, above the median. Other cells lie in $[0.64, 1.89]$, and the twin seed difference is 0.008 in median absolute log ratio. The treated/twin identity-share ratio at the readout input is 0.957; Appendix D.19 reports its coordinate checks.

D.10 ALL USABLE CHANNELS REMOVED AFTER PRETRAINING

The registered arm constrains the union of T-A and T-B in all three models under Appendix D.5's recipe. Those categories cover 61% of positions; category choice, member prediction at the other 39%, and trunk computations remain available. P-A0 tests the usual exchange-rate band, P-A1 halving, P-A2 persistence (median ≥ 0.8), and P-A3 a fall of at least 0.10 relative to own-channel removal.

Stability amendment and instrument. The original 0.5-nat stability allowance fails every run, although the exchange rate predicts rises of 0.761, 0.770 and 0.873 nats by model. Amendment 1 (2026-09-18) replaces it by predicted rise plus 0.25 nats. It followed the first lane's gate output and preceded registered geometry statistics. Exploratory first-lane ratios 0.967 (GPT-2) and 0.978 (Pythia-410M) in logit space, and 2.63 and 0.95 Euclidean at the readout input, had already been seen. The allowance used the frontier band's upper edge, 1.3, on a 0.77-nat prediction plus noise. Maximum observed excess is 0.125, so every run passes the amended gate. Pythia-1.4B fails the fp32 constraint check at 1.5×10^{-4} and 2.4×10^{-4}; its predeclared log-domain fallback passes, with conditional deviations $\leq 1.3 \times 10^{-6}$ and offset discrepancies $\leq 6.7 \times 10^{-6}$. Under the original stability criterion the arm would have no registered verdict.

Over 244 cells after singleton exclusion, P-A0 and P-A2 pass: slope 1.047, $r = 0.961$, median ratio 0.981 (IQR 0.94–1.02), with model medians 0.951, 1.004, 0.986 and identity-share ratio 0.975. P-A1

fails. P-A3 fails at $+0.012$, with model changes -0.006, $+0.032$, $+0.016$. Descriptive bootstrap intervals are $[0.971, 0.995]$ for the median and $[1.01, 1.09]$ for the slope. Inter-category drift is 0.031 nats. Summed ratios by model are 0.962, 1.003, 0.984; every cell lies in $[0.64, 1.48]$, and no category exceeds 8% of a reference sum. Pooled readout/logit medians are in Appendix D.19; the two-seed per-model logit medians are 0.967 (GPT-2), 0.979 (Pythia-410M) and 0.973 (Pythia-1.4B).

The result excludes maintenance by the other usable categories' prediction channels at this horizon. It does not distinguish demand from the remaining positions, category-level prediction or the trunk's computations.

D.11 All usable channels removed from initialization

The registered arm applies Appendix D.10's constraint from step one under the two-model formation recipe of Appendix D.6. Its pre-run stability criterion allows the twin's cross-entropy plus the exchange-rate prediction and 0.15 nats. Every run passes, and constraint deviations are 2×10^{-9}–3×10^{-8}. No criterion changed.

Over 156 cells, P-F0's exchange-rate criterion passes at slope 1.092, $r = 0.956$. P-F1 passes the formation band: pooled median 1.045 (IQR 0.93–1.16), GPT-2 0.987 and Pythia 1.140. P-F2 halving fails. P-F3's additional fall of 0.10 fails at $+0.059$ relative to own-channel removal ($+0.050$, $+0.075$ by model). The near-identical own-channel and all-channel loss slopes follow the constrained optimum's independence from the trunk, not a new geometric law.

The Pythia category identified in Appendix D.6 again forms at 0.214 and 0.192 of its twin. Model summed ratios are 0.998 and 0.346 (1.117 for Pythia without that category), occurrence-weighted ratios 1.024 and 0.242, and identity-share ratios 0.910 and 0.932. Only the dominant category and one other Pythia cell (0.334) lie outside $[0.5, 2]$. The registered comparison pairs replication-lane runs with canonical-lane twins. Using same-lane twins and own-channel comparators preserves every verdict (pooled median 1.037, model medians 0.974, 1.154, paired change $+0.064$). Readout-input ratios are 0.804 Euclidean and 0.806 in logit space (model logit ratios 0.905, 0.741), below own-channel formation's 0.832 on that lane. Category choice, uncovered prediction channels and trunk computations remain unconstrained.

D.12 All-channel maintenance at six epochs

The pre-run registration extends Appendix D.10 to six epochs. Clean references train unconstrained pretrained GPT-2, Pythia-410M and Pythia-1.4B with the FT0 recipe of Appendix D.5, extended to six WikiText-103 epochs with seeds $0, 1$. These are the same-seed FT0L runs originally produced for the corpus-edit horizon test. The Pythia twins overfit (validation losses 2.84, 3.44 versus 2.56, 2.41 at two epochs), so comparison is always at matched horizon. The stability gate allows the predicted rise plus 0.15 nats and passes all runs. It is one-sided: Pythia's realized rises remain 0.10–0.36 nats below prediction. GPT-2 twins and treated runs use different hardware. Pythia uses matched hardware.

One fp32 constraint check fails (8.0×10^{-4}); the predeclared log-domain fallback passes (1.1×10^{-6} conditional deviation, 6.2×10^{-6} offset discrepancy). Before that diagnostic existed, run-level and model-level voiding left 200 and 156 cells. Both provisional analyses give the same four verdicts as the final 244 cells. No criterion was amended.

P-L0 passes at slope 1.044, $r = 0.918$. P-L1 persistence (≥ 0.8) passes at median 0.958 (IQR 0.91–1.00; models 0.958, 0.964, 0.954). P-L2's fall of at least 0.10 fails at -0.014 against two epochs (model changes $+0.004$, -0.028, -0.036). P-L3 halving fails. The summed ratio is 1.011, identity-share ratio 0.957, and only two cells lie outside $[0.5, 2]$, both above it (2.25, 2.52), with no dominant category. Twin seed differences are 0.004, 0.016, 0.015 in median absolute log ratio. Appendix D.19 reports readout ratios near 0.95.

D.13 Loss withheld at treated positions

The registered loss-mask arm excludes the 61% of positions whose next token belongs to a usable category, while retaining all tokens in the input stream and leaving heads free. It uses the three-model maintenance and two-model formation recipes of Appendix D.5, seeds $0, 1$, and unconstrained twins

on the same hardware lane. No constraint check applies. No cross-entropy gate was registered because no theorem predicts loss when only 39% of positions contribute gradients.

Maintenance (244 cells): P-N1 persistence ≥ 0.8 passes at 0.98; P-N2 halving fails; P-N3's additional fall of 0.10 from all-channel removal fails at -0.001. Formation (156 cells): P-M1's formation band passes at 0.95 (models 1.02, 0.91); P-M2 halving fails; P-M3's 0.10 fall fails at -0.064. Table 1 reports the aggregate and readout ratios. Per arm, three categories have medians below half (exploratory), GPT-2 categories 4 and 46 at maintenance (0.24 and 0.38 in both seeds) and Pythia-410M category 4 at formation (0.30). These arms withhold loss rather than remove a channel, so Section 4's count over channel-removal arms excludes them.

Identity-share ratios are 0.915 at maintenance and 0.846 at formation. Cell ratios span 0.24–21.19 and 0.15–60.71, with 7 and 11 outside $[0.5, 2]$. Two Pythia categories grow more than tenfold at formation, separating the summed and occurrence-weighted ratios. Maintenance loss on masked positions rises to 16.5–17.3 nats from twins' 1.72–1.94. Unmasked loss falls to 3.108–3.814 from 3.525–4.321. At these horizons, demand at a category's own positions is unnecessary for the median category's dispersion, despite these heterogeneous effects.

D.14 Whole-vocabulary within-group constraint

The registered formation arm constrains all 50 categories plus a residual group of remaining vocabulary tokens. Only choice among these 51 groups depends on context. Offsets start from add-one-smoothed training counts, without evaluation data. Both architectures use seeds $0, 1$ and same-lane unconstrained twins under the formation recipe.

Stability amendment. The initial gate omitted the divergence between training and evaluation marginals from the predicted held-out loss. The omitted corpus-count term is 0.2125 nats in GPT-2 and 0.1899 in Pythia, 87% from the residual group. Amendment 1 followed inspection of the gate block (realized and uncorrected predicted rises, cell counts and twin seed noise), before any geometry statistic. With the corrected prediction all four runs pass, exceeding prediction by 0.05–0.07 nats within the unchanged 0.15-nat allowance. Under the original gate every run fails. Original and amended gates remain in the registration.

Over 156 cells, P-V0 accounting passes at slope 1.091, $r = 0.953$. P-V1's formation band fails upward: pooled median 1.432, GPT-2 0.958 in band, Pythia 1.938 above it. P-V2 halving fails; P-V3's additional fall of 0.10 fails at 0.440 relative to usable-channel removal. The identity-share ratio is 0.695, and 46 cells lie outside $[0.5, 2]$, spanning 0.48–43.97. Table 1 retains the summed and weighted ratios.

The same runs have readout-input dispersion only 0.058 of their twins (models 0.039, 0.111), with identity-share ratio 0.639. Thus $L{-}1$ growth coexists with near collapse at the readout input in Euclidean coordinates; the share shows that scale alone does not explain the full discrepancy, and the paragraph below shows that neither the collapse nor the growth appears in whitened coordinates or in readability. The logit ratio 0.034 measures group-level output variation: within-group row tying makes every member contrast context-free. These measurements cannot identify preserved trunk dispersion with an active within-group output distinction.

Readers and the coordinate-free ruler. A later registration applies the reader protocol of Appendix D.17 to these trunks, every usable category a cell, with two bands fixed beforehand: footprint ($R \in [0.15, 0.40]$) and functional ($R \in [0.60, 0.85]$), $L{-}1$ primary. Competence on the twins is 0.918–0.938 at $L{-}1$ and 0.931–0.950 at the readout input, 3% budget. At $L{-}1$, $R = 0.628$ over 156 cells (GPT-2 0.656, functional; Pythia-410M 0.599, 0.001 below the functional band; partial by Appendix D.17's bands at both layers), per-cell median 0.690 over cells with a twin increment above 0.01 nats (two excluded at $L{-}1$, one at the readout input), range over budgets 0.620–0.698. At the readout input, $R = 0.613$ (models 0.645, 0.581). The pre-specified prediction, $L{-}1$ in $[0.60, 0.80]$ with the readout input at least 0.10 lower, failed on both clauses: Pythia-410M missed the band by 0.001, and the readout input reads within 0.02 of $L{-}1$; the kill condition ($L{-}1$ below 0.40) did not trigger. The registered readings anticipated a footprint, or a functional $L{-}1$ with a lower readout input; near-equal recovery at both layers was not among them, and the whitened comparison below is the interpretation offered for it. The fit-free whitened dispersion is computed on the same features.

Each category's states are whitened by their total covariance, and $Q = \mathrm{tr}(S_W^+ S_B)$, where S_B is the occurrence-weighted covariance of the token means in those coordinates and $S_W = I - S_B$ the within-token covariance. Q matches $\mathrm{tr}(\Sigma_W^{-1}\Sigma_B)$ in the original coordinates when the within-token covariance Σ_W has full rank. An invertible linear map of the state changes the whitened coordinates only by an orthogonal map, because both have identity covariance and are linear in the same states, so Q is invariant to every such map whatever the rank, in exact arithmetic on the covariance's support. That covers the readout input, where the final normalization confines every state to a hyperplane, and the 2 of 68 GPT-2 and 6 of 88 Pythia-410M cells with fewer fit positions than dimensions, where the raw pseudo-inverse $\mathrm{tr}(\Sigma_W^+\Sigma_B)$ would not be invariant. It gives treated/twin ratios of 0.576 and 0.552 at $L{-}1$ and 0.528 and 0.528 at the readout input, identical at eigenvalue floors 10^{-8}, 10^{-6} and 10^{-4}; a random invertible map of one category's features moved raw B by a factor of at least 4.9 and this ratio by less than 10^{-7} relative at both layers. Over the 16 (layer, arm, architecture) rows of this appendix, the loss-withheld arm below included, and Appendix D.17, R lies within 0.117 of the whitened ratio, meeting the pre-specified 0.12 with no architecture-ordering flip; the whitened ratio sits below R on this arm and above it on the others, and the maximum falls on this arm's readout-input row in GPT-2. The trace is a finite-sample estimate, biased upward with the ratio of dimension to occurrences and sensitive to the smallest within-covariance directions; the floor sweep bounds the second, not the first, and the treated/twin ratio cancels the bias only to first order. The Euclidean readout-input collapse above shrinks to a gap of at most 0.05 after whitening, and the Pythia-410M $L{-}1$ Euclidean excess (raw median 1.94) reverses to 0.552.

Readers on the loss-withheld trunks. A third amendment to that registration, frozen before any fit, applies the same protocol, bands and ruler to the loss-withheld formation trunks of Appendix D.13, every usable category a cell. At $L{-}1$, $R = 0.619$ over 156 cells (GPT-2 0.621, Pythia-410M 0.617, both functional), per-cell median 0.779; at the readout input, $R = 0.605$ (models 0.611, 0.599). The forecast, $L{-}1$ in $[0.55, 0.80]$ in both architectures with the readout input within 0.10 of it, was met, and the kill condition ($L{-}1$ below 0.40) did not trigger. The whitened ratios are in Table 9.

Group scores and the member contrasts. The same registration asks whether the whole-vocabulary trunk's 50 surviving group log-odds carry the member contrasts at $L{-}1$. One global affine map from the group scores to the $L{-}1$ state is fitted on the 3% draw, and F_k is one minus the occurrence-weighted squared error of category k's reconstructed member-mean contrasts on validation, relative to their energy. The median F_k is 0.46 and 0.46 in the two GPT-2 seeds and 0.32 and 0.32 in Pythia-410M, against the forecast of at least 0.75 (P-31c1 fails). The map beats a size-matched random grouping (95th percentiles 0.25–0.29) but not the top 50 principal directions of the $L{-}1$ covariance (0.56–0.59), so P-31c2 fails in every run, and two of the three kill clauses trigger. A reader fitted on the 50 group scores alone recovers 0.89 (GPT-2) and 0.94 (Pythia-410M) of the full-feature reader's gain above initialization (P-31c3 met). The registered reading is mixed and names no mechanism. The member contrasts do not lie mostly in the group-score subspace, yet the group scores carry most of the readable member information. The descriptive span fraction sits at its random baseline (0.056 in GPT-2 against 0.063, and 0.049–0.051 in Pythia-410M against 0.049), and Pythia-410M's dominant category reaches F_k of 0.72 and 0.80. Part D asks whether the same-position group choice carries too little about the member to sustain the geometry alone. It carries a median 0.68 (GPT-2) and 0.60 (Pythia-410M) nats, above the registered 0.3, so that choice is not excluded as the sole sustainer. Amendment 1 moved part C's instrument gate, an identity that holds by construction, from the float16 feature store to float32 evaluation after residuals of 4–6×10^{-3} failed the 10^{-3} bar, and all four runs then pass. By then the four runs' median F_k at both layers and one GPT-2 run's $L{-}1$ block had been seen. No outcome field was recomputed.

The whitened ruler on every arm. A fourth amendment, frozen before any maintenance fit, applies the same statistic to every arm whose checkpoints are on one lane: the five maintenance arms and, recomputed from their probe files, the four formation arms, whose 16 per-architecture values reproduce the adjudicated ones exactly. Table 9 gives the treated/twin ratio per arm at both layers. The forecast staked before these fits, a pooled $L{-}1$ median in $[0.75, 1.05]$ for each maintenance arm, every architecture above 0.5 and the readout input within 0.15, held for four of the five arms. The six-epoch arm reads 0.68, and 0.54 in Pythia-1.4B, the lowest maintenance values, while its Euclidean median is 0.96.

Table 9: **Whitened dispersion on every arm with checkpoints.** Median over cells of the treated/twin ratio of whitened dispersion, which no invertible linear map of the state changes, pooled over models and seeds with the interquartile range, per-architecture medians at $L-1$ (P: Pythia), and cells below 0.5 at $L-1$. The gradient-detachment, corpus-permutation and tenfold-horizon arms have no checkpoints on this lane.

Arm	$L-1$ median [IQR]	Readout input	Per architecture, $L-1$	Below 0.5
Own channel, maintenance	0.86 [0.79, 0.95]	0.86 [0.80, 0.95]	GPT-2 0.96, P-410M 0.85, P-1.4B 0.82	1 of 243
Own channel, replication	0.81 [0.73, 0.86]	0.83 [0.75, 0.88]	Qwen 0.81	0 of 68
Own channel, formation	0.76 [0.67, 0.83]	0.75 [0.64, 0.83]	GPT-2 0.80, P-410M 0.74	18 of 156
All usable channels, two epochs	0.85 [0.79, 0.96]	0.85 [0.80, 0.96]	GPT-2 0.96, P-410M 0.85, P-1.4B 0.81	1 of 243
All usable channels, six epochs	0.68 [0.56, 0.92]	0.65 [0.53, 0.92]	GPT-2 0.93, P-410M 0.67, P-1.4B 0.54	39 of 243
All usable channels, formation	0.75 [0.66, 0.83]	0.74 [0.62, 0.83]	GPT-2 0.79, P-410M 0.73	19 of 156
Loss withheld, maintenance	0.76 [0.67, 0.89]	0.77 [0.67, 0.86]	GPT-2 0.91, P-410M 0.76, P-1.4B 0.71	13 of 243
Loss withheld, formation	0.67 [0.54, 0.76]	0.68 [0.51, 0.76]	GPT-2 0.69, P-410M 0.66	31 of 156
Whole vocabulary, formation	0.56 [0.49, 0.67]	0.53 [0.47, 0.65]	GPT-2 0.58, P-410M 0.55	44 of 156

D.15 Two further seeds

The own- and all-channel formation arms (Appendices D.6, D.11) use two seeds. A registration frozen before any run repeats all four conditions on seeds 2 and 3, on a third hardware lane, with the gates of Appendix D.6 and, for the all-channel arm, the stability gate of Appendix D.11. One run failed the constraint check, at 0.0002 against the 10^{-4} bar and without the log-domain diagnostic that the pre-declared fallback needs, so the registered run-level rule voids its 17 cells.

Verdicts. Own channel, over 139 surviving cells: the accounting gate is met with slope 1.096 and correlation 0.948; P-R1 (band) met, median 0.973 (GPT-2 0.919, Pythia-410M 1.048); P-R2 (halving) not met. All usable channels, over 156 cells: every gate passed; P-R3 (band) met, median 1.052 (GPT-2 0.980, Pythia-410M 1.134); P-R4 (halving) not met. Both formation results therefore replicate on fresh seeds and fresh hardware, and four seeds now stand.

The pre-declared variance summary. The analyzer computes it over every cell before the run-level void above, so seed 2's GPT-2 own-channel entry includes the 17 voided cells. Across the four seeds the per-seed medians span 0.918 to 0.967 in GPT-2's own-channel arm and 1.013 to 1.057 in Pythia-410M's, with standard deviations 0.023 and 0.019; the all-channel arms give 0.050 and 0.026. The summary also asked whether the category that dominates the Pythia-410M twins' summed dispersion in seeds 0 and 1 does so again. It does, in all four seeds, holding 0.82 to 0.86 of the twin's sum, and it forms at 0.090 to 0.136 of its twin value with its own channel absent and 0.138 to 0.274 with every usable channel absent. That structure and its collapse are therefore a property of this architecture and recipe rather than of two seeds, and the median result holds beside it. Pooled over the channel-removal arms (exploratory), it is the only one of 234 model-category combinations whose median ratio falls below 0.5, at 0.183 across the six arms that constrain it, and none of the 156 maintenance combinations does.

D.16 Tenfold formation horizon on a new corpus

The registered all-channel arm trains both architectures from initialization on one pass over 3.58B FineWeb-Edu tokens, saving weights at 10%, 30%, 60% and 100%. Evaluation remains on WikiText-103 validation at $L-1$ against same-seed unconstrained twins. Before any run, Amendment 1 added the omitted train/evaluation marginal-divergence terms (0.647, 0.635 nats by model) to the stability prediction, retaining its 0.15-nat allowance.

One of four runs, GPT-2's second seed, fails stability: realized rise 2.425 nats against prediction 0.720. The registered rule voids it. Its twin ends 1.5 nats worse than the other seed, with 26 times the median context variance and a final treated/twin dispersion ratio of 479. The three survivors pass stability, constraint and twin-law instrument checks, with constraint deviation $\leq 10^{-8}$.

Over 122 cells, P-H1's formation band passes at pooled 1.089 (models 0.959, 1.179). P-H2 halving fails. P-H3's erosion threshold fails at 0.024, with opposite model directions (-0.011, 0.099). P-H0 accounting fails: slope 1.055 is in band but $r = 0.689 < 0.8$. Geometry verdicts carry that caveat,

and the long-horizon result relies on one valid GPT-2 seed and two Pythia seeds. Horizon and corpus change together, so the unmet erosion prediction cannot isolate a horizon effect.

All surviving cell ratios lie in $[0.5, 2]$; identity-share ratio is 0.927. The Pythia twin seed difference is 0.061 in median absolute log ratio; GPT-2 has only one valid pair. Aggregate ratios appear in Table 1. The twin-share trajectory flattens in GPT-2 but keeps rising most strongly in its final segment in both Pythia runs (last-segment changes 0.0056, 0.0029). The reading rule was fixed after GPT-2 trajectories were seen and before any Pythia trajectory had been seen. It requires flattening in a surviving Pythia run. That criterion fails. Treated/twin ratios decline toward their final values in all three runs, but intermediate checkpoints are descriptive. At this horizon no category dominates the twins' summed dispersion. In Pythia-410M the largest, category 24 in both seeds, holds 6.5% and 6.2% of the twin's sum and forms at 1.04 and 1.29 of its twin value.

D.17 Frozen-trunk readers

This registered arm tests recoverability, separately from dispersion and the installed head's use. An affine softmax reader is fitted within each category on frozen features, whitened by its trunk's within-category covariance. Intercepts are free; the ridge strength is selected on the twin and held fixed for treated and initialized trunks. Gain G is the held-out loss reduction from an intercept-only reader fitted on the same positions. The endpoint is the occurrence-weighted gain above initialization, aggregated within each architecture and pooled with equal architecture weights:

$$R_m = \frac{\sum_k n_k (G_{\text{treated},k} - G_{\text{init},k})}{\sum_k n_k (G_{\text{twin},k} - G_{\text{init},k})}, \quad R = \frac{1}{M} \sum_m R_m.$$

Classification bands are readable ($R \geq 0.8$), partial ($0.5 \leq R < 0.8$) and limited ($R < 0.5$) at this budget. Amendment 1 added a 3% fit budget and lowered the registered competence bars from 0.9 to a floor of 0.8, with the exact value reported, because readers see a fraction of one epoch whereas installed heads trained on three. Before it, a two-category debug run on GPT-2 seed 0 had been seen for the twin, initialization and constrained trunks, in which the twin's reader reached 0.875 and 0.853 of its installed head.

The checkpoints are same-lane own-channel and all-channel formation runs, seeds $0, 1$ (Appendices D.6, D.11), independent of the run voided in Appendix D.15; whole-vocabulary removal is covered in Appendix D.14. At the primary readout-input layer, the synthetic competence check recovers 0.951 of installed-head gain and separates SHARED from INDEP by 1.82 nats against 1.31. Real-twin competence is 0.931–0.950, above 0.8. The budget rule selects 3%, where initial trunks recover 0.183 of twin gain.

Both arms are partial: own-channel $R = 0.715$ over 156 cells (models 0.725, 0.706), all-channel $R = 0.704$ over 156 cells (models 0.715, 0.693). Per-cell medians are 0.857 and 0.846. The own-channel cohort was corrected on 2026-09-20: the first implementation included both constrained and free categories from both complementary trunks, giving 0.856 over 312 cells. Restricting each trunk to its registered constrained categories gives the endpoint above. Both executions and frozen thresholds remain disclosed.

At $L-1$, endpoints are 0.724 (GPT-2 0.733, Pythia-410M 0.716) and 0.712 (GPT-2 0.722, Pythia-410M 0.701). Reducing fit data from 3% to 0.01% gives ranges 0.695–0.715 and 0.677–0.704. The registered twin-seed ceiling pools swapped seeds and is exactly one by construction. The exploratory per-seed version ranges from 0.996 to 1.004. These checks support partial affine readability at the tested budgets, not an identified real-model sustaining mechanism.

D.18 Transfer of the twin's reader to the constrained trunk

The readers above show what a refitted reader recovers. This part of the registration in Appendix D.14 asks whether the twin's own reader reads the whole-vocabulary trunk. Every trunk's features are whitened by the same-seed twin's within-category covariance on the fit draw, so a whitened direction is the same raw direction in every trunk. The reader of Appendix D.17 is fitted on the twin with the twin's ridge and applied frozen to another trunk with intercepts refit on that trunk's fit set. The transfer fraction T is the frozen reader's held-out gain over the destination's intercept-only predictor, divided by the gain of the same reader refit on the destination; this gain is not initialization-subtracted, unlike

R above. Cells are pooled as ratios of occurrence-weighted sums; 8 (GPT-2) and 14 (Pythia-410M) of the cells have a refit gain at or below zero, and the per-cell median of T over cells with a positive refit gain is -1.65 and -3.02. The ceiling is the same transfer to a single copy of the twin trained on a second lane with the same initialization and data order, staged from the replication lane, so the copies diverge only through nondeterministic kernels over the run; the floor is the transfer to the other seed's twin, a run with a different initialization and data order. Bands were fixed beforehand: transfer at or above 0.7 of the ceiling with the floor at most half the ceiling reads as the same affine decision functions; at or below 0.3 of the ceiling, with the refit gain near the twin's, as a different arrangement; between as uninterpretable. The forecast staked before the fits, transfer at 0.7 of the ceiling or more, failed. The registered alignment fallback was an orthogonal Procrustes map between whitened member means fitted on the odd-indexed usable categories; whitened means of different categories share no basis (each category has its own whitener), so Amendment 1 (2026-09-24, after the transfer jobs crashed at the alignment step and before any transfer number existed) fits the map on raw member means and applies it before the twin's whitening. Intervals are category-resampled bootstrap 95% intervals.

At $L{-}1$, $T = -1.394$ [-1.71, -1.17] (GPT-2, 70 cells) and -2.676 [-3.52, -2.22] (Pythia-410M, 88 cells); ceilings 0.842 and 0.877; floors -0.682 and -0.965; the reverse direction (constrained trunk to twin) -0.401 and -0.393. At the readout input, $T = -0.102$ and -0.284 against ceilings 0.775 and 0.883 and floors -0.961 and -0.893; there the raw transfer sits above the floor. A negative T satisfies the lower band trivially, and every (architecture, layer) block falls in it. The aligned transfer is -0.165 and -0.672 at $L{-}1$, but the same alignment carries the lane copy only to 0.037 and 0.036, so the fallback fails its own ceiling. Its values are also unstable across hardware in cells where the frozen reader is far off. The reader concentrates 95% of its weight energy in 10 (GPT-2) and 8 (Pythia-410M) directions (per-cell range 2–35), yet transfer fails.

Two exploratory diagnostics were registered as Amendment 2 after six of the eight files had been read, in response to a calibration confound visible in them (the constrained trunk's within-category variance is about twice the twin's), and before any of their fields was computed; the registered statistic, bands and reading rule were unchanged and remain the sole basis of the verdict, so these descriptives characterize an adjudicated result and could not alter it. First, one logit multiplier per category is refit with the intercepts on the destination trunk ($z \mapsto \tau z + b$, $\tau > 0$). The fitted multipliers are near zero, median 0.03 and 0.03 at $L{-}1$, far below the 0.7 that a twofold variance inflation alone would give: the fit finds no positive scale at which the frozen weights help. The rescaled transfer is 0.014 [0.003, 0.029] and 0.028 [0.010, 0.050] against 0.847 and 0.885 for the lane copy. This rules out a purely scalar-calibration account of the negative raw transfer; it cannot exclude alignment confined to a subset of types that a single scale does not expose. Second, the principal angles between the twin's and each other trunk's whitened member-contrast subspaces on validation means: against the constrained trunk the median smallest angle is 78.1° [76.8, 79.7] (GPT-2) and 79.5° [78.5, 80.3] (Pythia-410M); against the other-run twin 78.6° and 79.7°; against the lane copy 61.8° [59.5, 63.9] and 53.9°; for two random subspaces of the cells' dimensions (2 to 35) in the trunk's width, 77.7° and 79.8°. The constrained trunk's angles are within the other-run twin's interval and at the random reference, 16 to 26° above the lane copy's. All eight files were recomputed with the amended script; the registered pooled values of the six earlier files moved by at most 0.011 and their per-cell registered fields by at most 0.003 nats, except the aligned transfer noted above. The twin's affine decision functions therefore do not transfer to the constrained trunk, and its member-contrast subspaces are no closer to the twin's than chance. The frozen analyzer returns the different-arrangement reading from the transfer clause alone. The registration also asks for a refit gain near the twin's without fixing a number, and the refit reader recovers 0.58–0.66 of the twin's gain above initialization, so we do not claim that the trunk builds a substitute for what the twin encodes. Neither registered kill condition triggered.

D.19 Logit-space identity dispersion (exploratory)

For member-mean covariance Σ_B, effective head W, and row-centering operator P_S, we measure $\mathrm{tr}(P_S W \Sigma_B W^\top P_S)$. It is invariant to invertible changes of readout-input coordinates compensated by the head. Rows S include the whole vocabulary or only other categories. The own-row value is zero in every constrained category. The instrument was specified before evaluation. Table 10 compares medians at the same layer. Formation uses the replication lane and its own twins.

Table 10: Readout-input dispersion ratios and their logit-space counterparts. Each entry is a median treated/twin ratio. The last column uses other categories' rows. These checks concern the logits that receive the geometry.

Arm	Cells	Euclidean	All logits	Other logits
Own-channel maintenance (§D.5)	244	0.963	0.964	0.967
Own-channel formation (§D.6)	156	0.826	0.832	0.835
Qwen replication (§D.9)	68	0.946	0.977	0.979
All-channel maintenance (§D.10)	244	0.973	0.974	0.978
All-channel formation (§D.11)	156	0.804	0.806	0.810
All-channel, six epochs (§D.12)	244	0.946	0.947	0.950

Own-channel formation has per-architecture Euclidean ratios 0.820 and 0.829, the source of Table 1's readout-input entries. Its pooled logit ratio of 0.83 is below the $L-1$ ratio of 0.98. GPT-2's Euclidean ratios of 1.393, 2.63 and 3.08 in own-channel, all-channel and six-epoch maintenance have logit counterparts of 0.9635, 0.967 and 0.968. The discrepancy is consistent with a coordinate effect; for the whole-vocabulary formation runs the whitened ratios of Appendix D.14 shrink it to at most 0.05, and in the three GPT-2 maintenance arms above the whitened readout-input ratios are 0.95, 0.95 and 0.93 (per-architecture values behind Table 9). Vocabulary-wide removal is assessed separately in Appendix D.14: its surviving logit variation can distinguish groups, while within-group output distinctions are constrained to be context-free.

D.20 Secondary interventions and limits of inference

These registered experiments do not resolve the sustaining mechanism or establish an information-specific causal effect of corpus edits. Their complete protocols, thresholds and results remain in the supplementary package.

Breadth studies (W3, W8 and W9). Three earlier registered studies bear on no claim here. On Python code, two code models reproduce the variance allocation (identity share 0.13 and 0.12, P-W8b) and the law's positive sign over eight lexical classes ($\rho = 0.52$ and 0.05, P-W8a). Under 4-bit quantization of Pythia-1.4B, per-category damage tracks the inverse floor slack at $\rho = 0.39$ against the registered 0.40 (P-W9a fails), and int8 damage is too small to order (P-W9b fails). Across Pythia pretraining checkpoints in two sizes, between-category share does not predict probe accuracy (P5a fails), and selecting the checkpoint of maximal share fails its registered criterion (P5c).

Corpus permutation (W10 and W25). Permuting tokens within categories preserves unigram counts. The registered manipulation check failed on its control leg (control information changed by 20–24%), because treated tokens condition control bigrams. Before any training, Amendment 1 replaced it with a check that excludes those positions, which passes. With free heads, two-epoch fine-tuning of GPT-2 and both Pythias leaves median dispersion at 0.809 of the clean reference, missing the halving prediction. The share prediction also fails. For this information-reducing arm, untouched-category controls pass, and dispersion recovers when fine-tuning returns to clean text (the registered reversal check). An information-increasing edit misses its growth prediction and fails specificity in GPT-2, whose total loss rises from 2.9 to 5.7. Dose medians decrease in all three models, but pooled correlation -0.397 misses the registered -0.5 threshold. The six-epoch ratio 0.784 misses 0.65, so the interpretation registered conditional on that threshold does not apply. The bigram-preserving placebo passes its registered band but also reduces model-realized information, preventing an information-only interpretation. We found no stronger placebo at the fraction of positions these edits must move. A min-cost assignment that matches each category's within-bin realized-information profile still misses the pre-stated within-category loss band in every treated category. The realized signal is tied to which position carries which member rather than to the bin profiles, so the limitation above is a property of the corpus rather than this edit (exploratory). From initialization, GPT-2 on the permuted corpus retains 0.718 of twin dispersion against 0.786 for the placebo. Its manipulation and untouched-category checks pass, but halving, formation-band and selective-fall predictions all fail. These results delimit the causal interpretation of the cross-category law without testing the algebraic readout constraint.

Gradient detachment (W26). From initialization in GPT-2 and Pythia-410M, loss masking plus detachment of keys and values at treated positions passes the zero-gradient instrument check. All three predictions fail upward: targeted reduction relative to both masking and sham, halving, and the formation band. Median dispersion is 7.37 of the twin while total within-category variance is 11.7 times the twin's. Identity share is 0.815 of the twin, versus 0.843 for the sham. Summed and occurrence-weighted dispersion ratios are 1.77 and 0.51, qualifying the median inflation. The sham matches the number of detached states but retains their own loss, so it does not match the training credit removed. No registered mechanism conclusion follows.