IDENTIFYING THE FUNCTIONAL DISTANCE BETWEEN ATTENTION HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention heads in transformers are routinely compared through the similarity of their attention rows. Most heads place their largest attention weight on a sink token, and analysts must choose whether to keep or remove the sink before comparing, without knowing what either choice measures about a head's effect on the model. We prove that attention rows do not determine how differently two heads act, while the value vectors do. We show this by writing the difference between two head outputs as the value matrix applied to the difference of their attention rows: the rows leave the distance free within an interval, and the Gram matrix of the values fixes it exactly. We then introduce the recipient-slot distance, the output change when one head's output passes through the output projection of another, and prove that this substitution changes the layer output less than removing the head exactly when the distance is below the head's output norm. On GPT-2, Pythia, and Llama, comparisons based only on attention rows agree with the value-aware comparison about as often as chance, and about nine in ten heads change their functional nearest neighbour. Head outputs within a layer are nearly orthogonal, so their distance is set mostly by their norms, which attention rows do not contain. Selecting heads for removal with the value-aware distance costs less perplexity than selection by attention rows in eight of nine settings, and on Llama, where heads share values, substitutions chosen by the recipient-slot distance cost less than every removal rule. We further split every standard dissimilarity exactly into a sink part and a content part. This split shows when keeping or removing the sink reverses conclusions. Under natural axioms, no dissimilarity can be exactly separable and also continuous at the boundary or monotone under processing. Practitioners can compare heads within a layer through their projected values and report sink mass and content distance separately.

1 INTRODUCTION

Attention heads are routinely compared by their attention patterns, for example by clustering heads with the Jensen–Shannon divergence between their attention distributions (Clark et al., 2019). A head, however, contributes $V^\top p$ to the residual stream, where p is its attention row and V holds its projected values (Elhage et al., 2021). The row fixes the weights of a sum whose summands it does not contain. For example, in "The cat sat because it was tired", two heads may both attend from "it" to "cat" and still write different outputs, since each copies information from "cat" through its own values. Their patterns are identical, yet replacing one head by the other can change what the model computes. On the three models studied below, the head with the most similar pattern is the head with the closest output for 2 to 9 percent of heads, about as often as chance (3 to 9 percent, Figure 1). This gap between similar patterns and similar outputs matters for interventions. On Llama-3.2-1B, replacing 10 percent of the heads of each layer by the partners that the best pattern score selects raises perplexity by 28 percent, against 6 percent for the partners selected by the recipient-slot distance introduced here (Table 2).

We ask two questions. What does a comparison of attention rows reveal about the functional distance between heads, the distance between their outputs, and which quantity determines it? Which measurable quantity predicts the cost of replacing one head by another, and when is replacement cheaper than removal? The answers decide whether similar patterns are evidence of similar function.

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.

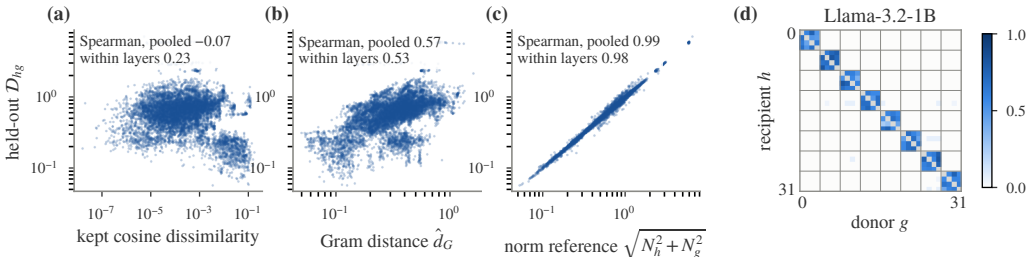


Figure 1: Llama-3.2-1B, seed 0 in (a) to (c) and all seeds in (d). (a) to (c): three training-window scores against the held-out complete-output distance $\mathcal{D}_{hg}$ for all 7,936 within-layer head pairs, with Spearman correlations pooled over layers and averaged within layers. Kept cosine dissimilarity of the attention signatures carries little information (a), the Gram distance $\hat{d}_G$ recovers part of the ordering (b), and the norm reference $\sqrt{N_h^2 + N_g^2}$ built from the two RMS (Root Mean Square) output norms, which is $\mathcal{D}$ without its cross term, recovers almost all of it (c). (d) Fraction of layers and seeds in which substituting donor g (column) into recipient h (row) changes the attention-sublayer output less than removing h on the training windows, the condition $S_{h\leftarrow g} < N_h$ of Corollary 1 (diagonal masked). It holds almost only inside the 4×4 blocks of query heads that share a key–value group (grid lines).

Five lines of work come closest. Elhage et al. (2021) separate the query–key circuit, which sets the pattern, from the output–value circuit, which sets what is written, and we ask what the former determines about the latter. Kobayashi et al. (2020) show that the norms $\|p_i v_i\|_2$ of the weighted value vectors change what attention weights mean, without comparing heads. Brunner et al. (2019) show that attention weights are not identifiable from outputs under a fixed value map, the converse of our question. Papamichalis & Ruane (2026) show that keeping or masking the sink position reverses row-based verdicts on 17 to 47 percent of head triples, without relating any row geometry to the outputs. Haklay et al. (2026) define head similarity by swapping circuits and measuring the change in next-token distributions, a downstream analogue of the local quantities here.

We separate three versions of the functional distance that head similarity conflates, all for heads of one layer. The *fixed-signature distance* d_G, also called the Gram distance, compares the outputs that two fixed attention rows, called signatures, produce through the projected values, and it is the quantity a row comparison identifies once one matrix is added. The *complete-output distance* $\mathcal{D}$ compares the actual outputs of two heads over an input distribution and answers whether two heads write the same thing. The *recipient-slot distance* $S_{h\leftarrow g}$ measures the output change when the unprojected output of a donor head g passes through the output projection of a recipient head h, and it is the local size of a head substitution, to be compared with the RMS (Root Mean Square) output norm N_h of the recipient. Scores computed from the rows alone are called pattern scores or row scores. Our contributions are as follows.

(i) *Identification.* We prove that attention rows fix only an interval of output distances, that the Gram matrix of the projected values identifies the distance, and that its plug-in estimator is rate-optimal, although loose at our sample size (Theorem 1). We prove that row-only dissimilarities weight the sink by factors that ignore the values, and that under natural axioms only multiples of the Aitchison distance separate every position from the others (Theorems 2 and 3).

(ii) *Intervention.* We bound the gap between the fixed-signature and complete-output distances and show when the resulting certificate can hold (Proposition 1). We introduce the recipient-slot distance, prove that it differs from $\mathcal{D}$ by at most the projection mismatch, and prove that substitution perturbs the attention-sublayer output less than removal exactly when $S_{h\leftarrow g} < N_h$, a condition that reduces to a Gram comparison of the two rows when the heads share values (Proposition 2, Corollary 1).

(iii) *Evidence.* On three language models, row scores recover nearest neighbours in $\mathcal{D}$ near chance, the norm reference $\sqrt{N_h^2 + N_g^2}$ recovers 65 to 86 percent, and masking the sink lowers the agreement of the pattern scores where the sink is strong. Within layers, $S_{h\leftarrow g}$ predicts substitution cost above $\mathcal{D}$ and the head norm in the point estimate of every seed, and selection

by $\mathcal{D}$ fails on Llama. The condition $S_{h\leftarrow g} < N_h$ holds for 62 percent of Llama pairs that share a key–value group and for at most 1.2 percent of all other pairs, on Llama the substitutions selected by $S_{h\leftarrow g}$ cost less than every removal rule, and on the other two models no substitution rule beats the best removal rule.

The scope is measurement at a fixed checkpoint, not how heads acquire rows or values.

2 Setup

An attention row over $D \geq 3$ common key positions is a point of the open simplex $\mathcal{S}^D = \{p \in \mathbb{R}^D_{>0} : \sum_i p_i = 1\}$. Coordinate 0 is a designated position, typically a sink, and *content* denotes the remaining positions. Every $p \in \mathcal{S}^D$ factors uniquely as $p = (s, (1-s)\pi)$ with sink mass $s = p_0$ and content $\pi = \mathcal{C}(p_1, \ldots, p_{D-1}) \in \mathcal{S}^{D-1}$, where $\mathcal{C}$ renormalizes to unit sum. We write $q = (t, (1-t)\rho)$, $\Delta s = s - t$, $\bar{s} = (s+t)/2$, and $p_s = (s, (1-s)\pi)$, $q_s = (s, (1-s)\rho)$ for matched sink mass. *Kept* scores use all D coordinates, and *masked* scores delete coordinate 0 and renormalize. Perturbation $p \oplus c = \mathcal{C}(p_0 c_0, \ldots, p_{D-1} c_{D-1})$ and powering $\alpha \odot p = \mathcal{C}(p_0^\alpha, \ldots, p_{D-1}^\alpha)$ make $\mathcal{S}^D$ a vector space (Aitchison, 1982), and $d_A(p,q) = \|\operatorname{clr} p - \operatorname{clr} q\|_2$ with $\operatorname{clr} p = \log p - \overline{\log p}\,\mathbf{1}$ is the Aitchison distance. The balance of coordinate j is $b_j(p) = \sqrt{(D-1)/D}\,\log(p_j / \operatorname{gm}(p_{-j}))$ with $\operatorname{gm}(p_{-j})$ the geometric mean of the other coordinates, and $p^{(-j)}$ is the subcomposition without coordinate j. We also use the Jensen–Shannon divergence (Lin, 1991) $\mathrm{JS}(p,q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m)$, with $m = (p+q)/2$ and KL the Kullback–Leibler divergence, and, for convex f with $f(1) = 0$, the f-divergence (Csiszár, 1967) $D_f(p\|q) = \sum_i q_i f(p_i/q_i)$. Logarithms are natural.

3 Identification of the functional distance

For a common value matrix $V \in \mathbb{R}^{D \times d_{\text{out}}}$ with rows $v_0, \ldots, v_{D-1}$, write $z_p = V^\top p$. Output projections are absorbed into V, so distinct heads are compared in a common output space, with $d_{\text{out}} = d_{\text{model}}$ for the heads of Section 4. For positive-semidefinite G, let $\|x\|_G^2 = x^\top G x$ and $\langle x, y\rangle_G = x^\top G y$.

Theorem 1 (Identification of the functional distance)**.** *Fix $p \neq q$ in $\mathcal{S}^D$ and $B > 0$, and let $\mathcal{V}_B = \{V : \max_i \|v_i\|_2 \leq B\}$.*

(a) Non-identifiability. $\{\|V^\top(p-q)\|_2 : V \in \mathcal{V}_B\} = [0, B\|p-q\|_1]$. *The supremum is attained by $v_i = B\,\mathrm{sign}(p_i - q_i)\,u$ with $\|u\|_2 = 1$ and the infimum by constant rows. For two pairs whose differences $x = p - q$ and $\tilde{x} = \tilde{p} - \tilde{q}$ are linearly independent, both strict orderings of their output distances are realizable in $\mathcal{V}_B$, so no attention-only score recovers this ordering uniformly over $\mathcal{V}_B$.*

(b) Identification by the value Gram. *With $G = VV^\top$ and $d_G(p,q) = \sqrt{(p-q)^\top G (p-q)}$, we have $\|z_p - z_q\|_2 = d_G(p,q)$. With e_0 the first standard basis vector, let $\boldsymbol{\nu} = e_0 - (0, (\pi+\rho)/2)$ be the sink direction and $\Delta_c = (0, \pi - \rho)$ the content difference. Then $p - q = \Delta s\,\boldsymbol{\nu} + (1-\bar{s})\Delta_c$ and*

$$d_G^2 = (\Delta s)^2 \|\boldsymbol{\nu}\|_G^2 + 2\Delta s(1-\bar{s})\langle \boldsymbol{\nu}, \Delta_c\rangle_G + (1-\bar{s})^2\|\Delta_c\|_G^2, \qquad \|\boldsymbol{\nu}\|_G^2 = \|v_0 - \bar{v}_c\|_2^2,$$

with $\bar{v}_c = \sum_{i\geq 1}(\pi_i + \rho_i)v_i/2$. For distinct heads h and g with value matrices V_h, V_g, $\|V_h^\top p - V_g^\top q\|_2^2 = p^\top G_{hh} p + q^\top G_{gg} q - 2p^\top G_{hg} q$ with $G_{ab} = V_a V_b^\top$, and if the two value matrices vary independently within $\mathcal{V}_B$, the attainable interval is $[0, 2B]$ for every p, q, including $p = q$.

(c) Uniform estimation. *Let $V^{(1)}, \ldots, V^{(n)}$ be i.i.d. copies of a random value matrix with rows of norm at most R almost surely, let $G = \mathbb{E}[VV^\top]$ and $\hat{G}_n = n^{-1}\sum_{i=1}^n V^{(i)}V^{(i)\top}$. For $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$\sup_{p,q\in\mathcal{S}^D} \left| d^2_{\hat{G}_n}(p,q) - d_G^2(p,q) \right| \;\leq\; 4R^2\sqrt{2\log(2D^2/\eta)/n}.$$

In expectation, this uniform error is at most $4R^2 \min\{1, \sqrt{2\log(2D^2)/n}\}$, and its minimax value is at least $(R^2/256)\min\{1, \sqrt{\log D/n}\}$ (Proposition 3 in Appendix A).

Part (a) says that the rows fix only the scale of the largest possible output change, and identical rows of distinct heads are compatible with every output distance in $[0, 2B]$. Part (b) shows that one

matrix determines the distance, and its three-term form shows that sink mass acts through the contrast $\|v_0 - \bar{v}_c\|_2$ between the sink value and the mean content value, so a small unprojected sink value does not make the sink channel small, and for a common value matrix d_G is a fixed multiple of Euclidean row distance if and only if $PGP = c_G P$ for a scalar $c_G \geq 0$, where $P = I - \mathbf{1}\mathbf{1}^\top/D$, that is, for projected values isotropic on sum-zero directions. Part (c) holds uniformly over all row pairs, including signatures selected from the sample, and is loose at our sample size $n = 96$, where at $\eta = 0.05$ it equals $2.0R^2$ to $2.5R^2$ against squared distances of at most $4R^2$. We write $\hat{d}_G = d_{\hat{G}_n}$ for the estimated Gram distance, which for distinct heads uses paired blocks $\hat{G}_{hg} = n^{-1}\sum_{i=1}^n V_h^{(i)} V_g^{(i)\top}$ on the same inputs, with HD in place of D in part (c) for H heads per layer. The Gram distance is not proposed as an estimator of $\mathcal{D}$, which the same forward passes measure directly: it is the quantity that a row comparison identifies once the values are added, and its three-term form is what the masking experiment tests. The proofs are in Appendix A.

4 Complete outputs and interventions

The next result bounds the gap to complete head outputs without independence assumptions.

Proposition 1 (Complete-output error and ordering certificate). *For a common input X and query position, let $a_h(X)$ be a nonnegative attention row summing to one over all available keys, and let $V_h(X)$ contain its projected values, with $\max_i \|v_{h,i}(X)\|_2 \leq R_h$ almost surely. Fix an ordered set K of W keys available to every compared head almost surely and a signature $p_h^\star \in \mathbb{R}^W_{\geq 0}$ summing to one. Let $y_h = V_h^\top a_h$ be the complete output, $\widetilde{y}_h = V_{h,K}^\top p_h^\star$ the fixed-signature output, $m_h = \sum_{i\in K} a_{h,i}$ the retained mass and $G_{hg}^K = \mathbb{E}[V_{h,K} V_{g,K}^\top]$ the paired Gram block, with both heads evaluated on the same input, and define $\mathcal{D}_{hg} = (\mathbb{E}\|y_h - y_g\|_2^2)^{1/2}$,*

$$d_G(h,g)^2 = (p_h^\star)^\top G_{hh}^K p_h^\star + (p_g^\star)^\top G_{gg}^K p_g^\star - 2(p_h^\star)^\top G_{hg}^K p_g^\star, \qquad \varepsilon_h = R_h\big(\|a_{h,K} - p_h^\star\|_1 + 1 - m_h\big),$$

and $\mathcal{E}_h = (\mathbb{E}\varepsilon_h^2)^{1/2}$. Then $\|y_h - \widetilde{y}_h\|_2 \leq \varepsilon_h$ pointwise, this radius is sharp over bounded projected values and

$$|\mathcal{D}_{hg} - d_G(h,g)| \leq \mathcal{E}_h + \mathcal{E}_g, \qquad d_G(h,k) - d_G(h,g) > 2\mathcal{E}_h + \mathcal{E}_g + \mathcal{E}_k \implies \mathcal{D}_{hg} < \mathcal{D}_{hk}.$$

The residual is $y_h - \widetilde{y}_h = V_{h,K}^\top(a_{h,K} - p_h^\star) + V_{h,K^c}^\top a_{h,K^c}$, the triangle inequality gives the radius, and its reverse form in L^2 gives the distance bound (proof and Figure 4 in Appendix A.4). Since $\mathcal{E}_h \geq 2R_h(1-\bar{m}_h)$ with $\bar{m}_h = \mathbb{E}m_h$ and $d_G(h,k) \leq R_h + R_k$, for a common value-norm bound the certificate requires $2\bar{m}_h + \bar{m}_g + \bar{m}_k > 3$, that is, $\bar{m} > 3/4$ for equal retained masses: no certificate can hold for triples whose heads place less than three quarters of their attention on the signature keys (Remark 2 in Appendix A.4, derived after the certificate failed).

From an output change to an intervention "Replacing head h by head g" conflates two operations. Replacing the projected contribution y_h by y_g changes the attention-sublayer output by $y_g - y_h$, of RMS size $\mathcal{D}_{hg}$. Head-slot substitution, the operation of our experiments, passes the donor's unprojected output ζ_g through the recipient's projection slice O_h and changes the sum by $O_h(\zeta_g - \zeta_h)$, and removal deletes $O_h\zeta_h$. Both edits save the same computation, since substitution deletes h and adds O_h to the slice of g.

Proposition 2 (Local distance and head-slot substitution). *Let X be a random input, let $\zeta_h(X), \zeta_g(X) \in \mathbb{R}^{d_v}$ be the complete unprojected outputs of two heads of one layer, let $O_h, O_g \in \mathbb{R}^{d_{\text{model}} \times d_v}$ be their output-projection slices, and assume finite second moments. Define $\Delta y_{h\leftarrow g} = O_h(\zeta_g - \zeta_h)$ and*

$$S_{h\leftarrow g} = \big(\mathbb{E}\|\Delta y_{h\leftarrow g}\|_2^2\big)^{1/2}, \qquad N_h = \big(\mathbb{E}\|O_h\zeta_h\|_2^2\big)^{1/2}, \qquad \Gamma_{h\leftarrow g} = \big(\mathbb{E}\|(O_h - O_g)\zeta_g\|_2^2\big)^{1/2},$$

and $\mathcal{D}_{hg} = (\mathbb{E}\|O_g\zeta_g - O_h\zeta_h\|_2^2)^{1/2}$. Then $|S_{h\leftarrow g} - \mathcal{D}_{hg}| \leq \Gamma_{h\leftarrow g}$, the bound is sharp, and $S_{h\leftarrow g} = \mathcal{D}_{hg}$ if $O_h = O_g$. If the loss ℓ_X of the state r entering the rest of the network changes by $\Delta\ell_{h\leftarrow g} = \ell_X(r + \Delta y_{h\leftarrow g}) - \ell_X(r)$ with $|\Delta\ell_{h\leftarrow g}| \leq L(X)\|\Delta y_{h\leftarrow g}\|_2$, then $\mathbb{E}|\Delta\ell_{h\leftarrow g}| \leq (\mathbb{E}L^2)^{1/2} S_{h\leftarrow g} \leq (\mathbb{E}L^2)^{1/2}(\mathcal{D}_{hg} + \Gamma_{h\leftarrow g})$. No universal lower bound or ranking preservation holds for local distances, even with equal projections and a fixed 1-Lipschitz loss.

We write S, $\mathcal{D}$, Γ and N when the pair is clear, and call $S_{h\leftarrow g}$ the local distance of the substitution. The distance bound is the reverse triangle inequality applied to $\Delta y_{h\leftarrow g} - (O_g\zeta_g - O_h\zeta_h) = (O_h - O_g)\zeta_g$, the loss bound is Cauchy–Schwarz, and the last claim follows from a two-dimensional cross-entropy example (Appendix A.5). Local measurement and intervention response remain distinct, since the loss bound has an unestimated factor L and the example rules out an ordering guarantee.

Corollary 1 (Substitution versus removal)**.** *In the setting of Proposition 2, the substitution $h \leftarrow g$ changes the attention-sublayer output by less than removing h, in RMS, if and only if $S_{h\leftarrow g} < N_h$, that is, if and only if $2\,\mathbb{E}\langle O_h\zeta_g, O_h\zeta_h\rangle > \mathbb{E}\|O_h\zeta_g\|_2^2$. If the two heads share their unprojected values, $\zeta_h = U^\top a_h$ and $\zeta_g = U^\top a_g$ with a common U, as the query heads of one key–value group do under grouped-query attention (Ainslie et al., 2023), then $\Delta y_{h\leftarrow g} = V_h^\top(a_g - a_h)$ with $V_h^\top = O_h U^\top$, and*

$$S_{h\leftarrow g}^2 = \mathbb{E}\big[(a_g - a_h)^\top G_{hh}(a_g - a_h)\big], \qquad G_{hh} = V_h V_h^\top.$$

The first claim expands $\mathbb{E}\|O_h(\zeta_g - \zeta_h)\|_2^2 < \mathbb{E}\|O_h\zeta_h\|_2^2$, and the second follows from $\zeta_g - \zeta_h = U^\top(a_g - a_h)$ and the pointwise identity of Theorem 1(b), so G_{hh} is the Gram matrix of the input inside the expectation, whereas G_{hg}^K in Proposition 1 is an expected block.

Remark 1 (Norm, cross term and shared values). Expanding the square gives $S_{h\leftarrow g}^2 = N_h^2 + \mathbb{E}\|O_h\zeta_g\|_2^2 - 2\,\mathbb{E}\langle O_h\zeta_g, O_h\zeta_h\rangle$, so the recipient norm enters S directly and what S adds are the donor's norm through the recipient's slot and the cross term. With head-specific unprojected values U_h, U_g, the displacement splits exactly as $\Delta y_{h\leftarrow g} = O_h U_h^\top(a_g - a_h) + O_h(U_g - U_h)^\top a_g$. The first term is a change of attention row measured in the recipient's Gram geometry and vanishes when $a_g = a_h$. The second is nonzero in general whenever the value maps differ, even for identical rows, and vanishes within a key–value group, where $S_{h\leftarrow g} < N_h$ reduces to a comparison of rows. The two terms can cancel, so the decomposition says only where the condition is easy to satisfy. S is defined in the native value coordinates and bounds only single substitutions (Remark 3 in Appendix A.5).

5 Row geometries and sink masking

Theorem 1(b) weights sink mass against content by value-dependent terms. Row-only dissimilarities impose their own weighting, which the next two results make explicit.

Theorem 2 (Sink–content calculus)**.** *Write $p = (s, (1-s)\pi)$, $q = (t, (1-t)\rho)$ and $\lambda = (1-s)/(2(1-\bar{s}))$. Then*

$$\mathrm{JS}(p, q) = \mathrm{JS}_b(s, t) + (1 - \bar{s})\mathrm{JS}_\lambda(\pi, \rho),$$

where $\mathrm{JS}_\lambda(\pi,\rho) = \lambda\mathrm{KL}(\pi\|\mu) + (1-\lambda)\mathrm{KL}(\rho\|\mu)$ with $\mu = \lambda\pi + (1-\lambda)\rho$ and the Bernoulli term JS_b is the divergence between $(s, 1-s)$ and $(t, 1-t)$. For matched sink masses and every f-divergence, $D_f(p_s\|q_s) = (1-s)D_f(\pi\|\rho)$, whereas $d_A(p_s, q_s)$ does not depend on s. Appendix E expands the statement into parts (a) to (f): chain rules, Hellinger and Fisher–Rao laws, attenuation exponents and partitions.

Masking deletes the Bernoulli term and replaces the weighted content term $(1-\bar{s})\mathrm{JS}_\lambda(\pi,\rho)$, with $\lambda - \frac{1}{2} = -\Delta s/(4(1-\bar{s}))$, by $\mathrm{JS}(\pi,\rho)$, so kept and masked pipelines report different functionals whenever the sink masses differ. Proposition 4 in Appendix E gives the sharp condition under which the sink masses alone force the kept JS ordering, and by the matched-sink identity a sink mass shared by the three heads of a triple scales every f-divergence by $1-s$, so masking changes such a verdict only through a sink mismatch.

Theorem 3 (Separability characterization)**.** *Let $D \geq 3$ and let $\delta : \mathcal{S}^D \times \mathcal{S}^D \to [0,\infty)$ satisfy (S1) $\delta(c \oplus p, c \oplus q) = \delta(p, q)$ for every $c \in \mathcal{S}^D$, (S2) $\delta(\alpha \odot p, \alpha \odot q) = |\alpha|\,\delta(p, q)$ for every $\alpha \in \mathbb{R}$, (S3) symmetry and $\delta(p, q) = 0$ if and only if $p = q$, and (S4) separability at every coordinate j (Sep-A in Appendix F), meaning $\delta(p, q)^2 = \varphi_j(b_j(p), b_j(q)) + \Psi_j(p^{(-j)}, q^{(-j)})$ for some functions φ_j, Ψ_j. Then $\delta = \beta\, d_A$ for some $\beta > 0$, and d_A satisfies (S1)–(S4). Permutation invariance, the triangle inequality and continuity are not assumed.*

The proof shows that separability makes δ^2 quadratic with a common coefficient along every balance direction and telescopes (Appendix F). The Gram distance lies outside this class, since (S1) forces matched-sink invariance whereas $d_G(p_s, q_s) = (1-s)\,d_{G,c}(\pi,\rho)$, with $d_{G,c}$ the content-block distance, varies with s. The disagreement between row metrics measured by Papamichalis & Ruane (2026) therefore has a candidate cause, and no choice among them supplies the missing value information (Table 1, and Table 13 in Appendix G for the properties of each row metric).

6 EXPERIMENTS

Each paragraph after the protocol tests named results: the first two test what rows can measure (Theorems 1 and 2 and a consequence of Theorem 3), and the last four test outputs and interventions (Propositions 1 and 2 and Corollary 1, with Remark 1, formulated after the first results, read against them). Numerical checks of the proved identities are in Appendix I. Table 12 in Appendix D records whether each prediction was fixed before the run that tested it and whether it held: of six fixed in advance, three held, two in part and one failed, and two of seven post hoc predictions failed.

Protocol and scope We compare heads within layers of GPT-2 (Radford et al., 2019), Pythia-160M (Biderman et al., 2023) and Llama-3.2-1B (Grattafiori et al., 2024) on WikiText-103 (Merity et al., 2016), with the full protocol and the compute (Table 3) in Appendix B. Signatures are attention rows renormalized over the first $W = 64$ keys and averaged over query positions 64 to 191 of 96 training windows of 192 tokens, and paired block Grams use the same windows. Complete outputs, $\mathcal{D}$, S and Γ are measured on 96 disjoint held-out windows at the same queries, whereas the head norms N_h, the norm reference, the alignment and every selection rule use the training windows. Interventions edit every position of a disjoint stream of 32,640 test tokens: the cost of a pair substitution is the mean change in negative log-likelihood per token over its two directions, and the cost of a rule is its increase in perplexity. Key 0 of each window is the designated coordinate, the first text token since no BOS token is prepended, and the largest key for 78 to 96 percent of heads (Table 4 in Appendix C). Three seeds redraw the windows, the evaluated pairs and the random baselines, and we report means and standard deviations over seeds. Each baseline has a role: kept cosine and kept JS are the pattern scores used in practice (Clark et al., 2019), computed here between averaged signatures, the masked variants their sink-aware form, Euclidean distance the isotropic reference of Theorem 1(b), d_A the compositional score of Papamichalis & Ruane (2026), the head norm a magnitude control (Remark 1), the norm reference $\sqrt{N_h^2 + N_g^2}$ the part of $\mathcal{D}$ without its cross term, $\mathcal{D}$ measured directly the upper reference and random selection the lower reference.

Attention rows and output distance By Theorem 1(a) and (b), fixed rows allow either ordering of output distances, and the Gram blocks determine the fixed-signature distance exactly, so we test how much of the ordering of $\mathcal{D}$ each score recovers. The six row scores, kept and masked cosine and JS, Aitchison and Euclidean distance (the last added after the first results), reach within-layer Spearman correlations of -0.06 to 0.23 with the held-out $\mathcal{D}$, agree with it on 48 to 56 percent of triple verdicts, a verdict being which of two candidates g and k a score places closer to h, and recover 2 to 9 percent of nearest neighbours, against chance levels of 9, 9 and 3 percent for nearest neighbours on the three models (Figure 1a, and Table 5 in Appendix C). The training Gram distance reaches 0.62, 0.42 and 0.53 within layers, 71, 65 and 70 percent of triple verdicts and 23 to 25 percent of nearest neighbours. Its training and held-out verdicts agree on 98.5 to 99.4 percent of triples, so the estimation error of Theorem 1(c) changes few verdicts at $n = 96$ and the gap to $\mathcal{D}$ is mostly approximation error (Proposition 1). The values, however, enter $\mathcal{D}$ mostly through the norms. The norm reference $\sqrt{N_h^2 + N_g^2}$ reaches 0.94, 0.96 and 0.98 within layers and recovers 65, 86 and 71 percent of nearest neighbours, the normalized alignment $A_{hg} = \mathbb{E}\langle y_h, y_g\rangle/(N_h N_g) = (N_h^2 + N_g^2 - \mathcal{D}_{hg}^2)/(2N_h N_g)$ of two complete outputs on the training windows has median 0.00, 0.01 and 0.00 with $|A_{hg}| < 0.1$ for 63 to 91 percent of pairs, and the nearest neighbour of a head in $\mathcal{D}$ is the smallest-norm head of its layer for 66, 86 and 77 percent of heads (Table 4 in Appendix C). Head outputs within a layer are thus close to orthogonal, and $\mathcal{D}$ orders pairs mostly by scale, which the rows do not contain.

Sink masking and row geometry By Theorem 2, masking removes from JS the Bernoulli term, the content weight $1 - \bar{s}$ and the skew $|\lambda - \frac{1}{2}| = |\Delta s|/(4(1 - \bar{s}))$, and by Theorem 1(b), for a common value matrix, the squared output distance contains the sink channel $(\Delta s)^2 \|v_0 - \bar{v}_c\|_2^2$, which every masked score discards. We therefore expected, after the first results, masking to lower agreement with the output-based targets where $|\lambda - \frac{1}{2}|$ is large. On GPT-2 and Llama, where the median within-layer $|\lambda - \frac{1}{2}|$ is 0.15 and 0.18, kept and masked JS give opposite verdicts on 21 and 26 percent of within-layer triples, the criterion of Proposition 4 in Appendix E certifies that the sink masses alone decide 18 and 15 percent, and masking lowers the within-layer correlation of cosine and JS with both $\mathcal{D}$ and the substitution cost in all 24 seed-level comparisons, by 0.05 to 0.27 (seed means

Table 1: Top: mean full-row attention on key 0, median content skew $|\lambda - \frac{1}{2}|$ of the training signatures over within-layer pairs, and percent of within-layer triples with opposite kept and masked JS verdicts or with a kept JS verdict that Proposition 4 in Appendix E certifies from the sink masses alone. Bottom: within-layer Spearman correlation of training-signature scores with the held-out $\mathcal{D}$ and with the substitution cost, means over three seeds with standard deviations at most 0.04 (Tables 5 and 7 in Appendix C). The loss from masking is the kept minus the masked correlation, averaged over cosine and JS before rounding.

	GPT-2		Pythia-160M		Llama-3.2-1B	
mean attention on key 0 (%)	39.5		14.1		51.3	
median content skew $\lvert\lambda - \frac{1}{2}\rvert$ (Theorem 2)	0.153		0.033		0.180	
kept and masked JS verdicts opposite (%)	21.4		8.0		26.0	
kept JS verdict certified by sink masses (%)	18.0		0.05		15.2	
within-layer Spearman with	$\mathcal{D}$	cost	$\mathcal{D}$	cost	$\mathcal{D}$	cost
kept cosine	0.20	0.20	0.18	0.23	0.23	0.27
masked cosine	−0.06	−0.01	0.18	0.20	0.05	0.18
kept JS	0.14	0.17	0.14	0.16	0.19	0.25
masked JS	−0.06	−0.01	0.11	0.11	0.04	0.17
loss from masking	0.23	0.19	0.01	0.03	0.16	0.09
Aitchison d_A	0.03	0.07	0.02	−0.02	0.06	0.18

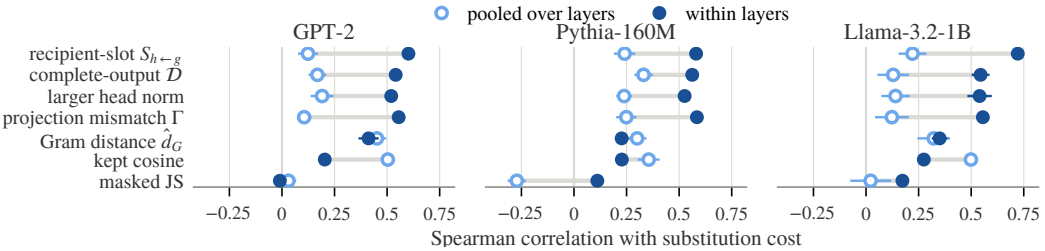


Figure 2: Spearman correlation between each predictor and the substitution cost, pooled over layers (open) and within layers (filled), mean ± standard deviation over three seeds, on 400 within-layer head pairs per seed (300 on Llama) in both directions. Pair costs and the directional predictors S and Γ are averaged over the two directions. Predictors are $S_{h\leftarrow g}$ and Γ of Proposition 2, $\mathcal{D}$, the larger head norm, $\hat{d}_G$ and two row scores (all predictors in Table 7 in Appendix C).

in Table 1). On Pythia, where $|\lambda - \frac{1}{2}|$ is 0.03, masking reverses 8 percent of verdicts, the criterion certifies 0.05 percent, the mean loss per score and target is at most 0.05 and 3 of the 12 seed-level differences are zero or negative. The larger skew on Llama does not bring a larger loss, so the skew separates weak from strong effects without ordering them. Aitchison distance, the unique separable geometry of Theorem 3, reaches at most 0.06 with $\mathcal{D}$, against 0.42 to 0.62 for the Gram distance, which weights the two channels by the values. Figure 5 and Table 14 in Appendix H reproduce the mechanism under controlled sink mismatch (Exp. 8).

Transfer to complete outputs At $W = 64$ the signature keys hold 53, 44 and 64 percent of the attention mass, below the three quarters that the certificate of Proposition 1 requires for equal retained masses. Accordingly, the realized residual of the fixed-signature output is 0.86, 1.01 and 1.00 times the RMS output norm and no nearest neighbour is certified at $W \in \{16, 32, 64\}$, with plug-in radii (R_h the sample maximum of the value norms) or with realized residuals in place of the radii (Table 6 in Appendix C), so the transfer to complete outputs is supported only by rank agreement.

Local distances and substitution cost Proposition 2 identifies $S_{h\leftarrow g}$ as the local size of a slot substitution and bounds its gap to $\mathcal{D}_{hg}$ by $\Gamma_{h\leftarrow g}$. On these heads the median $\Gamma/\mathcal{D}$ is 0.96 to 1.00, so the bound does not certify $\mathcal{D}$ as a proxy for S, although the two differ by a median of 5 to 6 percent only. Pooled over layers, S correlates with the pair cost at only 0.12 to 0.24, below kept cosine (0.36 to 0.50), and the cost falls with depth (Spearman -0.46, -0.78 and -0.33 with the layer index), which plausibly proxies the factor $L(X)$ of the loss bound. Every selection rule compares heads of one layer, so the informative comparison is within layers, an analysis chosen after the pooled result. There, S predicts the cost at 0.60, 0.58 and 0.72, above $\mathcal{D}$ (0.54 to 0.56), the larger head norm

Table 2: Increase in perplexity, in percent of the unedited perplexity (in parentheses), when 10, 20 and 30 percent of the heads of each layer (1, 2 and 4 of 12, or 3, 6 and 10 of 32) are substituted or removed, mean ± standard deviation over three seeds, with cells of 1000 percent or more rounded to two significant figures without standard deviation. Pair rules take the closest still-unused within-layer pair under the named score and overwrite or zero its smaller-norm member, the recipient-slot rule takes the still-unused ordered pair with the smallest training-window $S_{h\leftarrow g}$ and overwrites h with g, and norm removal zeroes the smallest-norm heads. Bold marks the smallest mean per column and mode among the non-random rules, including ties at the shown precision. Tables 10 and 11 in Appendix C give every rule in perplexity points.

rule	GPT-2 (53.7)			Pythia-160M (56.1)			Llama-3.2-1B (18.8)		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
substitution by $S_{h\leftarrow g}$	**+7±1**	+1800	+1900	**+7±0**	**+29±0**	**+192±2**	**+6±0**	**+30±24**	**+101±20**
substitution by $\hat{d}_G$	+44±1	+131±0	**+1200**	+16±2	+51±0	+216±13	+54±25	+204±52	+11000
substitution by $\mathcal{D}$	+15±1	**+89±24**	+2100	**+7±0**	+32±8	+231±26	+63±38	+22000	+60000
substitution by kept JS	+672±596	+4500	+17000	+42±3	+118±8	+340±57	+39±1	+117±9	+467±48
substitution by masked JS	+456±382	+774±181	+4300	+44±5	+122±7	+396±103	+28±6	+75±19	+224±38
random substitution	+179±103	+704±473	+5800	+36±3	+148±44	+639±217	+2000	+40000	+68000
removal by $\hat{d}_G$	+82±4	+179±0	**+312±5**	+7±1	+27±2	+108±15	+41±18	**+98±36**	**+532±122**
removal by $\mathcal{D}$	**+7±1**	**+36±2**	+847±232	**+2±0**	+19±7	+97±12	**+28±13**	+4400	+6100
removal by norm	**+7±1**	+3100	+26000	**+2±0**	**+11±0**	**+72±5**	+44±2	+141±79	+2200
removal by kept JS	+310±258	+844±900	+4600	+20±1	+52±2	+123±24	+70±0	+392±60	+9000
removal by masked JS	+483±423	+503±20	+2500	+21±3	+46±1	+120±31	+84±13	+492±174	+795±336
random removal	+312±462	+639±307	+4500	+26±5	+76±8	+266±11	+157±111	+2300	+15000

(0.52 to 0.54), which carries the norm term of S^2 in Remark 1, the Gram distance (0.23 to 0.41) and the row scores other than Euclidean (at most 0.27), and Γ exceeds S by 0.004 on Pythia and trails it elsewhere (Figure 2). In the point estimate of every seed S exceeds $\mathcal{D}$ and the head norm, and bootstrap intervals stratified by layer place the gap to $\mathcal{D}$ above zero on GPT-2 and Llama and include zero on Pythia (Table 8 in Appendix C). Per direction, $S_{h\leftarrow g}$ predicts the cost of $h \leftarrow g$ at 0.53 to 0.66 and $\mathcal{D}$ at 0.41 to 0.45, whereas the scale-free ratio $S_{h\leftarrow g}/N_h$ predicts the cost neither pooled over layers (-0.14 to 0.08) nor within layers (-0.16 to 0.09), so what predicts the cost is the size of S, not its ratio to the recipient norm (Table 7 in Appendix C).

Substitution against removal Corollary 1 states that a substitution perturbs the attention-sublayer output less than removing its recipient exactly when $S_{h\leftarrow g} < N_h$, and Remark 1 locates such pairs among heads that share values. In substitution mode (Table 2, and Table 11 in Appendix C), the recipient-slot rule, the only rule that also chooses the direction of each substitution, costs less than every row-score rule and random substitution in 8 of 9 model–budget settings and less than every other substitution rule in 6 of 9. On Llama its increases are 6, 30 and 101 percent, against 28, 75 and 224 percent for masked JS, the best row score, and 28, 98 and 532 percent for the best removal rule. Selection by the directly measured $\mathcal{D}$ costs 22,000 percent at 20 percent on Llama. The pairs it selects have a median $S_{h\leftarrow g}/N_h$ of 1.43 against 0.72 for the recipient-slot rule, but the Gram, kept JS and masked JS rules select medians of 1.62, 1.40 and 1.19 and cost 204, 118 and 75 percent at the same budget, so the ratio does not explain the failure, whose cause we have not identified. The recipient-slot rule fails on GPT-2 at 20 and 30 percent, above 1000 percent. On GPT-2 and Pythia at most 4.2 percent of the pairs it selects satisfy $S_{h\leftarrow g} < N_h$, whereas on Llama 78 to 85 percent do, which is consistent with Corollary 1 but does not predict the failures, since Pythia has the same low fractions and no failure (P13 in Appendix D, a post hoc reading). Every budget edits several heads and lies outside the single-substitution scope of the corollary. In removal mode, the pair rules test whether output redundancy, which rows cannot measure (Theorem 1(a)), lowers the cost below that of removing the smallest-norm heads. Removal by $\hat{d}_G$ costs less than every row-score rule in 8 of 9 settings, the exception being Pythia at 30 percent (Table 10 in Appendix C), and has the smallest worst case among the rules (12 times the cheapest rule of a setting, against 45 and 86 for $\mathcal{D}$ and norm removal), but norm removal is the cheapest or tied-cheapest rule at every Pythia budget. Of the 63 combinations of the seven substitution rules of Table 11 in Appendix C and nine settings, 8 beat the best removal rule, all on Llama, so the post hoc prediction that substitution beats removal only where the selected pairs satisfy $S < N$ is rejected as stated (Appendix D).

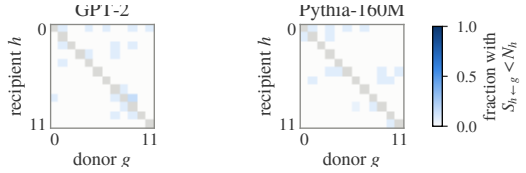


Figure 3: The condition $S_{h\leftarrow g} < N_h$ of Corollary 1 on the two models without shared values, as in Figure 1d (fraction of layers and seeds, training windows, diagonal masked). The condition holds for at most 1.2 percent of pairs, against 62 percent of the within-group pairs of Llama-3.2-1B.

Where substitution beats removal Remark 1 locates the pairs with $S_{h\leftarrow g} < N_h$ among heads that share values. On the training windows, the condition holds for 1.2 and 1.0 percent of ordered within-layer pairs on GPT-2 and Pythia (Figure 3), for 0.08 percent of Llama pairs from different key–value groups and for 62 percent of Llama pairs that share a group (Figure 1d). Near-orthogonality is consistent with the rarity: by Remark 1, $S < N_h$ needs a cross term in the recipient's slot above half the donor's slot norm squared, and the rare pairs that satisfy it have complete outputs with a median alignment A_{hg} of 0.69 on Pythia and 0.10 on GPT-2, against 0.01 and 0.00 for the other pairs. On Llama the pairs that satisfy it are not aligned in complete output (median A_{hg} of -0.08 against 0.00), which is consistent with differing projections: within a key–value group the donor shares the recipient's values, so the recipient's projection re-aligns it, as in Remark 1. The recipient-slot rule, which does not use the grouping, joins two heads of one group in 94 to 97 percent of its Llama substitutions, and a single within-group substitution costs a third as much as one between other pairs (median 0.0021 against 0.0067 nats per token). Within groups, where Remark 1 reduces S to a row comparison, S carries information: over the 192 within-group substitutions it correlates with the cost within layers at 0.51, against 0.12 for $\mathcal{D}$ and 0.33 for the recipient norm (Table 9 in Appendix C).

7 Related work

Work on attention sinks explains their formation and function (Xiao et al., 2024; Sun et al., 2024; Gu et al., 2025; Guo et al., 2024; Wong et al., 2025; Su et al., 2026), whereas we take rows and values as given. Centered kernel alignment (Kornblith et al., 2019) compares representations through centered Gram matrices, whereas d_G is an absolute output discrepancy between two heads. Bian et al. (2021) measure head redundancy with pattern and representation distances, and importance scores and learned gates (Michel et al., 2019; Voita et al., 2019) rank single heads by their effect on the loss, whereas our rules select pairs by output geometry. Jin et al. (2025) align heads by Procrustes transformations before merging them into key–value groups, which changes the value coordinates our substitutions hold fixed. Table 17 in Appendix J separates classical, restated and new results.

8 Conclusion

We proved that rows alone leave the output distance free within an interval, while the Gram matrix of the projected values fixes it exactly, and that substituting one head for another changes the attention-sublayer output less than removing it exactly when the recipient-slot distance is below the recipient's output norm. On three language models, pattern scores find a head's nearest neighbour about as often as chance, and head outputs within a layer are nearly orthogonal, so their distance is set mostly by norms the rows do not contain. The recipient-slot distance predicts substitution cost within layers, selection by the complete-output distance fails on Llama for reasons we have not identified, and substitution beats removal only on the grouped-query model. Our results therefore suggest comparing heads within a layer through their projected values and, at a fixed checkpoint, restricting substitution to heads that share values. These conclusions rest on three models of at most 1.24 billion parameters, one corpus and three seeds. The orthogonality, within-layer and key–value-group analyses came after the first results, the substitution advantage is shown on one grouped-query model and reverses on GPT-2 at 20 and 30 percent, and per-token row comparisons, gradient-based importance and mean ablation are not evaluated.

Reproducibility statement The supplementary material contains the code of the real-model study and of the verification experiments with their calibration statistics, the per-seed result files, the job logs and the script that regenerates every real-model table and figure.

AI-use statement Language-model assistants were used, under the authors' direction and review, for drafting and editing text, for implementing and debugging the experiment code. The authors retain responsibility for the theorems, proofs, experiments and claims.

Ethics statement The work analyses publicly released pretrained models on public datasets and involves no human subjects, personal data, or deployment. Head removal and substitution are diagnostic interventions and are not proposed as a compression method.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

John Aitchison. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4): 365–379, 1992.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*, 35:99–110, 1943.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pp. 930–945, 2021.

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.

Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

Peter J Bushell. Hilbert's metric and positive contraction mappings in a banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

L Lorne Campbell. An extended čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

Nikolai Nikolaevich Cencov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 2000.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3): 279–300, 2003.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Ionas Erb and Nihat Ay. The information-geometric perspective of compositional data analysis. In *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*, pp. 21–43. Springer, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024.

Tal Haklay, Nikhil Prakash, Sana Pandey, Antonio Torralba, Aaron Mueller, Jacob Andreas, Tamar Rott Shaham, and Yonatan Belinkov. Pitfalls in evaluating interpretability agents. *arXiv preprint arXiv:2603.20101*, 2026.

Sam Hilton-Jones, Timothy J Norman, and Zhanxing Zhu. Modelling attention with aitchison geometry: Token distinguishability and temperature scaling. In *Forty-third International Conference on Machine Learning*.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Qingyun Jin, Xiaohui Song, Feng Zhou, and Zengchang Qin. Align attention heads before merging them: An effective way for converting mha to gqa. In *EMNLP (Findings)*, pp. 8804–8816, 2025.

Pascual Jordan and J von Neumann. On inner products in linear, metric spaces. *Annals of Mathematics*, 36(3):719–723, 1935.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Elon Kohlberg and John W Pratt. The contraction mapping approach to the perron-frobenius theory: Why hilbert's metric? *Mathematics of Operations Research*, 7(2):198–210, 1982.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Frank Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2020.

Frank Nielsen and Ke Sun. Clustering in hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*, pp. 297–331. Springer, 2018.

Marios Papamichalis and Regina Ruane. Which question is your attention metric answering? attention rows as compositional data. *arXiv preprint arXiv:2608.14712*, 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2931–2951, 2019.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Jeffrey TH Wong, Cheng Zhang, Louis Mahon, Wayne Luk, Anton Isopoussu, and Yiren Zhao. On the existence and behavior of secondary attention sinks. *arXiv preprint arXiv:2512.22213*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Appendix

Roadmap. Appendix A proves the results of Sections 3 and 4. Appendix B gives the protocol of the real-model study and its sampling assumptions, Appendix C the complete result tables, and Appendix D the prediction ledger. Appendices E to G prove the results of Section 5 and the incompatibility theorem, Appendix H reports the synthetic regime map, Appendix I the numerical checks of the proved statements, and Appendix J the extended related work and the provenance of every result.

A Proofs for Sections 3 and 4 (identification, estimation, transfer and substitution)

A.1 (a) The range interval

Write $\boldsymbol{\delta} = p - q$. Upper bound: $\|V^\top \boldsymbol{\delta}\|_2 \le \sum_i |\boldsymbol{\delta}_i| \, \|v_i\|_2 \le B\|\boldsymbol{\delta}\|_1$, by the triangle inequality applied to $V^\top \boldsymbol{\delta} = \sum_i \boldsymbol{\delta}_i v_i$. Attainment: take $v_i = B \operatorname{sign}(\boldsymbol{\delta}_i)\, u$ for any unit $u \in \mathbb{R}^{d_{\text{out}}}$. Then $V^\top \boldsymbol{\delta} = B\|\boldsymbol{\delta}\|_1 u$. Infimum 0: constant rows $v_i \equiv Bu$ give $V^\top \boldsymbol{\delta} = B(\mathbf{1}^\top \boldsymbol{\delta})u = 0$ since $\mathbf{1}^\top \boldsymbol{\delta} = 0$. Fullness of the interval: $V_\theta = (1-\theta)V_{\text{inf}} + \theta V_{\text{sup}}$ stays in $\mathcal{V}_B$ (row-wise convexity of the norm ball) and $\theta \mapsto \|V_\theta^\top \boldsymbol{\delta}\|_2$ is continuous from 0 to $B\|\boldsymbol{\delta}\|_1$. The intermediate value theorem fills $[0, B\|\boldsymbol{\delta}\|_1]$. *Ranking inversion:* if $\boldsymbol{\delta} = p - q$ and $\tilde{\boldsymbol{\delta}} = \tilde{p} - \tilde{q}$ are linearly independent, pick $w \in \boldsymbol{\delta}^\perp \setminus \tilde{\boldsymbol{\delta}}^\perp \subset \mathbb{R}^D$ and set the rows of $V^{(1)}$ to $v_i = B\,(w_i/\|w\|_\infty)\, u$ for a fixed unit $u \in \mathbb{R}^{d_{\text{out}}}$: every row obeys the norm bound, $V^{(1)\top}\boldsymbol{\delta} = B\,(w^\top \boldsymbol{\delta})/\|w\|_\infty \, u = 0$, while $V^{(1)\top}\tilde{\boldsymbol{\delta}} \neq 0$ since $w \notin \tilde{\boldsymbol{\delta}}^\perp$. The symmetric choice gives $V^{(2)}$ and both strict orders are realized (Exp. 5: 8×10^{-16} vs 0.221 and 4.23 vs 5×10^{-17}). Since any attention-only score is a fixed function of $(p, q, \tilde{p}, \tilde{q})$, it cannot equal the sign of a quantity realizable with either sign. □

A.2 (b) Gram identity, channels, envelope

$\|z_p - z_q\|^2 = (p-q)^\top V V^\top (p-q) = \boldsymbol{\delta}^\top G \boldsymbol{\delta}$ with $\boldsymbol{\delta} = p - q$. For the channel identity, check coordinates. At the sink, $\boldsymbol{\delta}_0 = s - t = \Delta s = \Delta s\, \boldsymbol{\nu}_0$ and $\Delta_{c,0} = 0$. At content coordinate $i \ge 1$,

$$\Delta s\, \boldsymbol{\nu}_i + (1 - \bar{s})\Delta_{c,i} = -\tfrac{s-t}{2}(\pi_i + \rho_i) + \Big(1 - \tfrac{s+t}{2}\Big)(\pi_i - \rho_i) = \pi_i(1-s) - \rho_i(1-t) = \boldsymbol{\delta}_i,$$

collecting the π_i coefficient $-\frac{s-t}{2} + 1 - \frac{s+t}{2} = 1 - s$ and the ρ_i coefficient $-\frac{s-t}{2} - 1 + \frac{s+t}{2} = t - 1$. Expanding $\boldsymbol{\delta}^\top G \boldsymbol{\delta}$ bilinearly in the two channels gives the three-term formula. Also, $\|\boldsymbol{\nu}\|_G^2 = \boldsymbol{\nu}^\top V V^\top \boldsymbol{\nu} = \|V^\top \boldsymbol{\nu}\|^2 = \big\|v_0 - \sum_{i \ge 1} \frac{\pi_i + \rho_i}{2} v_i\big\|^2$ and $\langle x, y\rangle_G = x^\top G y$. The envelope is Cauchy–Schwarz on the cross term: $|\langle \boldsymbol{\nu}, \Delta_c\rangle_G| \le \|\boldsymbol{\nu}\|_G \|\Delta_c\|_G$, so d_G lies between $|a - b|$ and $a + b$ with $a = |\Delta s|\, \|\boldsymbol{\nu}\|_G$ and $b = (1 - \bar{s})\|\Delta_c\|_G$. □

Identification of G up to PGP Differences $p - q$ of interior simplex points contain a neighbourhood of 0 in $\mathbf{1}^\perp$. Indeed, for any $x \in \mathbf{1}^\perp$, the rows $p = D^{-1}\mathbf{1} + \varepsilon x/2$ and $q = D^{-1}\mathbf{1} - \varepsilon x/2$ are positive for sufficiently small $\varepsilon > 0$. Thus $d_G \equiv d_{G'}$ implies $x^\top (G - G')x = 0$ for every $x \in \mathbf{1}^\perp$, by homogeneity. Polarization gives $x^\top (G - G')y = 0$ for all $x, y \in \mathbf{1}^\perp$, i.e. $P(G - G')P = 0$ with $P = I - \frac{1}{D}\mathbf{1}\mathbf{1}^\top$. Conversely $P(G - G')P = 0$ makes the two pseudometrics equal since $P\boldsymbol{\delta} = \boldsymbol{\delta}$. Taking $G' = c_G I$ with $c_G \ge 0$ gives the isotropy statement of Section 3: $d_G = \sqrt{c_G}\, \|\cdot\|_2$ on row differences if and only if $P(G - c_G I)P = 0$, that is, $PGP = c_G P$. Metric versus pseudometric: $d_G(p, q) = 0$ for $p \neq q$ iff $\boldsymbol{\delta} \in \ker G \cap \mathbf{1}^\perp = \ker V^\top \cap \mathbf{1}^\perp$. □

Corollary 2 (Zero sink values can retain sink-mass effects)**.** *If $v_0 = 0$, then $\|\boldsymbol{\nu}\|_G^2 = \|\bar{v}_c\|_2^2$. In particular, for a fixed content composition π,*

$$\|V^\top p_s - V^\top p_t\|_2 = |s - t| \left\| \sum_{i \ge 1} \pi_i v_i \right\|_2 .$$

Thus changing the sink mass can change the output even when the sink value vanishes: it rescales the content contribution. This effect is nonzero exactly when $s \neq t$ and the content-weighted value sum is nonzero.

Cross-head block identity For projected values in a common output space,

$$\|V_h^\top p - V_g^\top q\|_2^2 = p^\top G_{hh} p + q^\top G_{gg} q - 2p^\top G_{hg} q, \qquad G_{ab} = V_a V_b^\top.$$

For random values and fixed signatures, replace each block by $G_{ab} = \mathbb{E}[V_a V_b^\top]$ to obtain the mean squared output distance. Its plug-in estimator uses paired observations $\hat{G}_{ab} = n^{-1} \sum_{i=1}^n V_{a,i} V_{b,i}^\top$.

More explicitly, stack the values of H heads into $\mathbb{V} \in \mathbb{R}^{HD \times d_{\text{out}}}$, write $\mathbb{G} = \mathbb{E}[\mathbb{V}\mathbb{V}^\top]$ and embed a signature as $\widetilde{p}_h = e_h \otimes p$. Then the squared distance is $(\widetilde{p}_h - \widetilde{q}_g)^\top \mathbb{G} (\widetilde{p}_h - \widetilde{q}_g)$. Because $\|\widetilde{p}_h - \widetilde{q}_g\|_1 \le 2$, the proof of part (c) applies with D replaced by HD. This requires independent sampled inputs, not independence between heads evaluated on the same input. For a single fixed pair of heads, $2D$ suffices.

The common-value interval in part (a) should be distinguished from the cross-head case: if V_h and V_g vary independently subject to the same row-norm bound B, then

$$\{\|V_h^\top p - V_g^\top q\|_2 : V_h, V_g \in \mathcal{V}_B\} = [0, 2B].$$

The triangle inequality gives the upper bound. Choosing every row of V_h as au and every row of V_g as $-au$, with $0 \le a \le B$ and $\|u\|_2 = 1$, attains the entire interval. □

A.3 (c) Uniform estimation of squared distances

Main bound (entrywise Hoeffding). Write $E = \hat{G}_n - G$. For any pair, $|d^2_{\hat{G}} - d^2_G| = |\boldsymbol{\delta}^\top E \boldsymbol{\delta}| \le \sum_{i,j} |\boldsymbol{\delta}_i||E_{ij}||\boldsymbol{\delta}_j| \le \|E\|_{\max} \|\boldsymbol{\delta}\|_1^2 \le 4\|E\|_{\max}$, uniformly over the simplex, since $\|\boldsymbol{\delta}\|_1 \le 2$. Each entry of $VV^\top$ is $\langle v_i, v_j \rangle \in [-R^2, R^2]$, so Hoeffding gives $\Pr(|E_{ij}| \ge \varepsilon) \le 2\exp(-n\varepsilon^2/(2R^4))$ and a union bound over the D^2 entries yields, with probability $1 - \eta$,

$$\|E\|_{\max} \le R^2 \sqrt{\tfrac{2\log(2D^2/\eta)}{n}}, \qquad \sup_{p,q} \left| d^2_{\hat{G}_n} - d^2_G \right| \le 4R^2 \sqrt{\tfrac{2\log(2D^2/\eta)}{n}}.$$

Every constant is explicit and no factor of D enters outside the logarithm. □

Distances. $|d_{\hat{G}} - d_G| \le |d^2_{\hat{G}} - d^2_G| / (d_{\hat{G}} + d_G) \le 4\|E\|_{\max}/d_G$ on pairs with $d_G > 0$ and $|d_{\hat{G}} - d_G| \le \sqrt{|d^2_{\hat{G}} - d^2_G|} \le 2\sqrt{\|E\|_{\max}}$ always: an $n^{-1/4}$ generic rate, $n^{-1/2}$ when d_G is bounded below.

Proposition 3 (Minimax lower bound)**.** *For every $n \ge 1$, $R > 0$ and $D \ge 3$,*

$$\inf_{\widehat{d^2}} \sup_{\mathcal{P}} \mathbb{E}_{\mathcal{P}} \left[\sup_{p,q \in \mathcal{S}^D} \left| \widehat{d^2}(p,q) - d^2_{G_{\mathcal{P}}}(p,q) \right| \right] \ge \frac{R^2}{256} \min\left\{ 1, \sqrt{\frac{\log D}{n}} \right\}.$$

Here $\widehat{d^2}$ ranges over estimators based on n i.i.d. value matrices, $\mathcal{P}$ ranges over distributions satisfying $\max_i \|v_i\|_2 \le R$ almost surely and $G_{\mathcal{P}} = \mathbb{E}_{\mathcal{P}}[VV^\top]$.

Proof. Use scalar values $v_0 = R$ and $v_j = R\xi_j$ for $1 \le j < D$, where the ξ_j are independent sign variables taking values in $\{-1, 1\}$. Multiplying these scalar values by a fixed unit vector embeds the construction in any output dimension $d_{\text{out}} \ge 1$. Under P_0 every ξ_j has mean zero. Under P_j only ξ_j has nonzero mean $\varepsilon \in (0, 1/2]$. For each j, take interior rows with

$$p_0^{(j)} = \tfrac{1}{4}, \quad p_j^{(j)} = \tfrac{1}{2}, \qquad q_0^{(j)} = \tfrac{1}{2}, \quad q_j^{(j)} = \tfrac{1}{4},$$

$$p_i^{(j)} = q_i^{(j)} = \frac{1}{4(D-2)} \quad (i \notin \{0, j\}).$$

Since $p^{(j)} - q^{(j)} = (e_j - e_0)/4$,

$$d^2_{G_P}(p^{(j)}, q^{(j)}) = \frac{R^2}{8}\big(1 - \mathbb{E}_P \xi_j\big).$$

The D resulting target vectors, indexed by $P_0, \ldots, P_{D-1}$, have pairwise ℓ_∞ separation $R^2\varepsilon/8$. The Bernoulli bound $\mathrm{KL}(P_j \| P_0) \le \chi^2(P_j \| P_0) = \varepsilon^2$ gives $\mathrm{KL}(P_j^{\otimes n} \| P_0^{\otimes n}) \le n\varepsilon^2$.

For $D \geq 4$, choose $\varepsilon = \min\{1/2, \sqrt{\log D/(8n)}\}$. Fano's inequality bounds the average error probability of every classifier of these D hypotheses below by

$$1 - \frac{n\varepsilon^2 + \log 2}{\log D} \geq \frac{3}{8}.$$

Apply this to the classifier that selects the closest target vector to the estimated distances on the displayed pairs. A classification error requires an ℓ_∞ estimation error of at least $R^2\varepsilon/16$. The worst-case expected uniform error is therefore at least $3R^2\varepsilon/128$. Since $\varepsilon \geq (2\sqrt{2})^{-1}\min\{1, \sqrt{\log D/n}\}$, the coefficient is at least $3/(256\sqrt{2}) > 1/256$.

For $D = 3$, retain only P_0 and P_1 and choose $\varepsilon = 1/(2\sqrt{n})$. Pinsker's inequality gives total variation at most $1/(2\sqrt{2})$. The two-point testing bound then gives expected error at least

$$\frac{R^2}{64\sqrt{n}}\left(1 - \frac{1}{2\sqrt{2}}\right),$$

which also exceeds the stated lower bound. □

The entrywise argument also gives

$$\mathbb{E}\sup_{p,q}\left|d^2_{\hat{G}_n}(p,q) - d^2_G(p,q)\right| \leq 4R^2\min\left\{1, \sqrt{\frac{2\log(2D^2)}{n}}\right\}.$$

Indeed, Hoeffding's lemma gives $\mathbb{E}e^{\lambda E_{ij}} \leq e^{\lambda^2 R^4/(2n)}$ for every real λ. For $\lambda > 0$, the maximum over both signs and all entries therefore satisfies

$$\mathbb{E}\|E\|_{\max} \leq \frac{\log(2D^2)}{\lambda} + \frac{\lambda R^4}{2n}.$$

Minimizing over λ gives the displayed stochastic bound after multiplication by 4. Both squared distances lie in $[0, 4R^2]$, which gives the deterministic cap. The plug-in estimator is therefore minimax rate-optimal in both n and D for expected uniform absolute error in squared distances, with loose constants: at $n = 96$ and $\eta = 0.05$ the high-probability bound equals $2.0R^2$ for $D = 64$ and $2.5R^2$ for $HD = 2048$, against the elementary cap $4R^2$.

A.4 Transfer from fixed signatures to complete outputs

Proof of Proposition 1. For each input, the remainder is exactly

$$r_h := y_h - \widetilde{y}_h = V_{h,K}^\top(a_{h,K} - p_h^\star) + V_{h,K^c}^\top a_{h,K^c}.$$

Since the attention weights are nonnegative,

$$\|r_h\|_2 \leq R_h\|a_{h,K} - p_h^\star\|_1 + R_h\sum_{i\notin K} a_{h,i} = \varepsilon_h.$$

The definition remains valid when $m_h = 0$. If $m_h > 0$, with $p_h = a_{h,K}/m_h$ the renormalized retained row, the identity $a_{h,K} - p_h^\star = m_h(p_h - p_h^\star) - (1 - m_h)p_h^\star$ gives

$$\|a_{h,K} - p_h^\star\|_1 \leq m_h\|p_h - p_h^\star\|_1 + 1 - m_h.$$

One can use any probability vector for p_h when $m_h = 0$ because its coefficient in this bound is zero. The bounds imply $0 \leq \varepsilon_h \leq 2R_h$ and hence finite second moments. Expanding the squared norm of $V_{h,K}^\top p_h^\star - V_{g,K}^\top p_g^\star$ and taking expectations gives $d_G(h,g) = \|\widetilde{y}_h - \widetilde{y}_g\|_{L^2}$. The reverse triangle inequality in $L^2(\mathbb{P}, \mathbb{R}^{d_{\text{out}}})$ and Minkowski's inequality then give

$$\begin{aligned}\left|\mathcal{D}_{hg} - d_G(h,g)\right| &\leq \|(y_h - y_g) - (\widetilde{y}_h - \widetilde{y}_g)\|_{L^2} \\ &= \|r_h - r_g\|_{L^2} \leq \|r_h\|_{L^2} + \|r_g\|_{L^2} \leq \mathcal{E}_h + \mathcal{E}_g.\end{aligned}$$

Thus $\mathcal{D}_{hg} \leq d_G(h,g) + \mathcal{E}_h + \mathcal{E}_g$ and $\mathcal{D}_{hk} \geq d_G(h,k) - \mathcal{E}_h - \mathcal{E}_k$. The stated gap condition makes the first upper bound smaller than the second lower bound. □

Sharpness of the approximation radius For fixed a_h and $p_h^\star$, the bound $\|r_h\|_2 \le \varepsilon_h$ is sharp over all projected value vectors satisfying the norm bound. Choose a unit vector u and set $v_{h,i} = R_h \operatorname{sign}(a_{h,i} - p^\star_{h,i})u$ on K and $v_{h,i} = R_h u$ off K. Zero coefficients can be assigned any admissible value. Every nonzero summand in r_h is then a nonnegative multiple of u, so equality holds. The coefficient 2 of the omitted mass cannot be reduced in the normalized-prefix bound. Indeed, when $p_h = p_h^\star$, setting all prefix values to $-R_h u$ and all omitted values to $R_h u$ gives $\|r_h\|_2 = 2R_h(1 - m_h)$.

Remark 2 (Scope of the certificate). Since $\|a_{h,K} - p_h^\star\|_1 \ge 1 - m_h$, the radius satisfies $\mathcal{E}_h \ge \mathbb{E}\varepsilon_h \ge 2R_h(1 - \bar{m}_h)$ with $\bar{m}_h = \mathbb{E}m_h$, whereas $d_G(h,k) \le R_h + R_k$. With a common value-norm bound R, the margin condition $d_G(h,k) - d_G(h,g) > 2\mathcal{E}_h + \mathcal{E}_g + \mathcal{E}_k$ can hold only if $2R > 2R\big[2(1-\bar{m}_h) + (1-\bar{m}_g) + (1-\bar{m}_k)\big]$, that is, only if $2\bar{m}_h + \bar{m}_g + \bar{m}_k > 3$. For equal mean retained masses this is $\bar{m} > 3/4$. The condition is necessary and not sufficient, and it was derived after the certificate failed on the three language models, where the mean retained mass at $W = 64$ is 0.44 to 0.64 (Table 6 in Appendix C).

What the bound certifies For a candidate set $\mathcal{J}_h$ excluding h, if $g \in \mathcal{J}_h$ satisfies the gap condition for every $k \in \mathcal{J}_h \setminus \{g\}$, it is the unique nearest neighbour of h over $\mathcal{J}_h$ in complete-output RMS distance. The expectation compares heads on the same input distribution. It does not compare values drawn independently for the two heads. The bound allows arbitrary dependence between the attention rows and projected values on each input. At the attention sublayer, replacing the projected contribution y_h by y_g while holding all other head contributions fixed changes the summed output by exactly $y_g - y_h$. Therefore $\mathcal{D}_{hg}$ is the RMS size of that immediate perturbation. Downstream loss and perplexity need not follow the same ordering.

Adding a simultaneous estimation event For $\hat{d}_G(h,g)^2 \ge 0$ and $\tau_{hg} \ge 0$, on any event where $|\hat{d}_G(h,g)^2 - d_G(h,g)^2| \le \tau_{hg}$, define

$$L_{hg} = \max\left\{0, \sqrt{\left(\hat{d}_G(h,g)^2 - \tau_{hg}\right)_+} - \mathcal{E}_h - \mathcal{E}_g\right\}, \qquad U_{hg} = \sqrt{\hat{d}_G(h,g)^2 + \tau_{hg}} + \mathcal{E}_h + \mathcal{E}_g,$$

where $x_+ = \max\{x, 0\}$. Monotonicity of the square root gives $\sqrt{(\hat{d}_G(h,g)^2 - \tau_{hg})_+} \le d_G(h,g) \le \sqrt{\hat{d}_G(h,g)^2 + \tau_{hg}}$, and combining these inequalities with Proposition 1 gives $\mathcal{D}_{hg} \in [L_{hg}, U_{hg}]$. For a candidate set $\mathcal{J}_h$ excluding h, if $g \in \mathcal{J}_h$ and $U_{hg} < L_{hk}$ for every $k \in \mathcal{J}_h \setminus \{g\}$, then g is the unique nearest neighbour of h over $\mathcal{J}_h$ in complete-output RMS distance (Figure 4). If the estimation event holds simultaneously over the candidates, so do the intervals. Replacing each $\mathcal{E}_h$ by any valid upper bound preserves validity. Fixed signatures chosen using an independent training sample can be treated conditionally on that sample. The uniform Gram bound also controls sample-selected signatures, but still requires the stipulated sampling assumptions and valid approximation-error bounds.

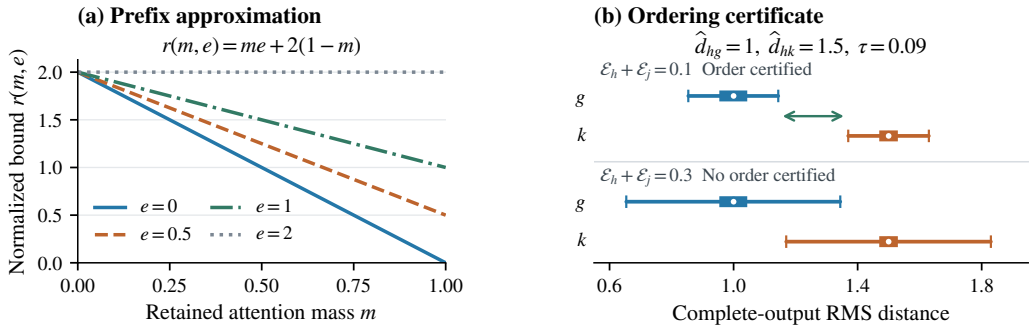


Figure 4: Analytical illustration of Proposition 1. (a) For $R_h > 0$, retained mass $m > 0$ and signature discrepancy $e = \|p_h - p_h^\star\|_1$, the function $r(m,e) = me + 2(1-m)$ bounds the pointwise output error divided by R_h. The endpoint $m = 0$ shows the limiting bound. (b) Circles are illustrative estimates $\hat{d}_G(h,j)$ for $j \in \{g,k\}$. Thick segments are the distance-scale intervals $\left[\sqrt{(\hat{d}_G^2 - \tau)_+}, \sqrt{\hat{d}_G^2 + \tau}\right]$ obtained from a squared-distance tolerance τ. Thin segments also add the RMS approximation radii $\mathcal{E}_h + \mathcal{E}_j$. Separated intervals certify the displayed ordering, and overlapping intervals are inconclusive. These are analytical bounds and illustrations, not model measurements.

A.5 Projection mismatch and downstream response

Proposition 2 (Local distance and head-slot substitution, expanded statement)**.** *Let X be a random input, let $\zeta_h(X), \zeta_g(X) \in \mathbb{R}^{d_v}$ be the complete unprojected outputs of two heads in the same layer and let $O_h, O_g \in \mathbb{R}^{d_{\text{model}} \times d_v}$ be their output-projection slices. Define*

$$\begin{aligned}
\Delta y_{hg}(X) &= O_g \zeta_g(X) - O_h \zeta_h(X), \\
\Delta y_{h\leftarrow g}(X) &= O_h\big(\zeta_g(X) - \zeta_h(X)\big), \\
\mathcal{D}_{hg} &= \big(\mathbb{E}\|\Delta y_{hg}(X)\|_2^2\big)^{1/2}, \\
S_{h\leftarrow g} &= \big(\mathbb{E}\|\Delta y_{h\leftarrow g}(X)\|_2^2\big)^{1/2}, \\
\Gamma_{h\leftarrow g} &= \big(\mathbb{E}\|(O_h - O_g)\zeta_g(X)\|_2^2\big)^{1/2}.
\end{aligned}$$

These definitions describe substitution at one query position. For simultaneous substitution at T positions, stack the unprojected outputs into $\mathbb{R}^{Td_v}$ and replace each projection slice by $I_T \otimes O_h$. The same argument applies with the Euclidean norm of the stacked state, equivalently the Frobenius norm before vectorization. Assume the stated second moments are finite. Then

$$\big|S_{h\leftarrow g} - \mathcal{D}_{hg}\big| \le \Gamma_{h\leftarrow g}. \tag{1}$$

The bound is sharp. In particular, $S_{h\leftarrow g} = \mathcal{D}_{hg}$ if $O_h = O_g$.

Let $r(X)$ be the state entering the rest of the network before substitution and let ℓ_X be its scalar loss as a function of that state, with all other inputs fixed. Suppose there is a measurable $L(X) \ge 0$ with $\mathbb{E}L(X)^2 < \infty$ such that

$$\big|\ell_X(r(X) + \Delta y_{h\leftarrow g}(X)) - \ell_X(r(X))\big| \le L(X)\|\Delta y_{h\leftarrow g}(X)\|_2.$$

Writing the loss difference inside the absolute value as $\Delta\ell_{h\leftarrow g}(X)$ gives

$$\mathbb{E}|\Delta\ell_{h\leftarrow g}(X)| \le \big(\mathbb{E}L(X)^2\big)^{1/2} S_{h\leftarrow g} \le \big(\mathbb{E}L(X)^2\big)^{1/2}\big(\mathcal{D}_{hg} + \Gamma_{h\leftarrow g}\big). \tag{2}$$

There is no universal positive lower bound or preservation of response rankings in terms of $\mathcal{D}_{hg}$, even when the two projections agree and the downstream loss is fixed.

The main text states the proposition with $N_h = (\mathbb{E}\|O_h\zeta_h\|_2^2)^{1/2}$ added for Corollary 1.

Proof. The two local changes satisfy

$$\Delta y_{h\leftarrow g}(X) - \Delta y_{hg}(X) = (O_h - O_g)\zeta_g(X).$$

The reverse triangle inequality in L^2 gives (1). Equality is possible in one dimension with $\zeta_h = 0$, $\zeta_g = 1$, $O_h = 0$ and $O_g = 1$, for which $S_{h\leftarrow g} = 0$ and $\mathcal{D}_{hg} = \Gamma_{h\leftarrow g} = 1$. Applying the assumed loss bound and Cauchy–Schwarz gives the first inequality in (2). The distance bound gives the second.

For the final claim, take deterministic inputs and three heads with identical projection slices I_2. Let the reference output be $\zeta_h = (0, 0)$ and the donor outputs be $\zeta_{g_1} = (0, 2)$ and $\zeta_{g_2} = (1, 0)$. Take the unchanged contribution to be zero, so the baseline state is $r = (0, 0)$. For the fixed binary cross-entropy loss $\ell(x) = \log(1+e^{x_1})$, the substitutions have norms 2 and 1, while their loss changes are 0 and $\log(1 + e) - \log 2 > 0$, respectively. This loss is globally 1-Lipschitz. Thus a positive local distance can have zero response and a larger distance can have a smaller response. These head outputs can be realized with strictly positive attention rows by assigning the same desired value to every key in each head. □

Remark 3 (Scope). $S_{h\leftarrow g}$ is defined in the native value coordinates of the checkpoint: negating U_g together with O_g leaves the model unchanged but changes the raw copy $O_h\zeta_g$, so S measures the specified substitution and not an intrinsic equivalence of heads, and a common change of coordinates cancels within a group. For simultaneous substitutions with distinct recipients in one layer, the displacement is the sum of the single displacements, so its RMS is at most $\sum_i S_{h_i\leftarrow g_i}$, whereas no such bound compares it with removal.

Interpretation For one common query position, combining the substitution bound with Proposition 1 gives

$$\mathbb{E}|\Delta\ell_{h\leftarrow g}| \le \|L\|_{L^2}\big(d_G(h, g) + \mathcal{E}_h + \mathcal{E}_g + \Gamma_{h\leftarrow g}\big).$$

For substitution at several query positions, every term must refer to the corresponding stacked state. Apply Proposition 1 at each substituted position t, using an available key set and a fixed signature for that position. If $d_{G,t}(h, g)$ and $\mathcal{E}_{h,t}$ are the per-position quantities, the stacked quantities are

$$d_G^{\mathrm{seq}}(h, g) = \left(\sum_t d_{G,t}(h, g)^2\right)^{1/2}, \qquad \mathcal{E}_h^{\mathrm{seq}} = \left(\sum_t \mathcal{E}_{h,t}^2\right)^{1/2}.$$

Use these with the stacked projection discrepancy and a loss-sensitivity bound for the stacked state. No independence between query positions is required. A single-query distance or an unspecified query average cannot be inserted unchanged into the sequence bound. For fixed T, a uniformly sampled-query RMS distance is the stacked RMS distance divided by $\sqrt{T}$.

On the three language models of Section 6, plug-in radii certify no nearest neighbour at $W \le 64$, and $\Gamma_{h\leftarrow g}$ is as large as $\mathcal{D}_{hg}$ (median ratio 0.96–1.00). The fixed-signature score is therefore not certified as a proxy for either complete-output quantity, and the bound cannot guarantee that $\mathcal{D}_{hg}$ approximates $S_{h\leftarrow g}$, although the two differ by a median of 5–6 percent on the evaluated pairs. The loss-sensitivity factor is not estimated. The bound states additional conditions for a response guarantee and does not establish one for the reported predictor rankings.

B Experimental protocol and scope

This appendix specifies the real-model study (Exp. 7) whose results are in Section 6 and Appendix C, states the sampling assumptions under which the guarantees apply, lists the compute, and describes the calibrated synthetic rows used by Exps. 1, 2 and 5.

Exp. 7 (real-model study): protocol *Data.* Raw WikiText-103 text is tokenized without special tokens and cut into non-overlapping windows of $T = 192$ tokens. Per seed, a seeded permutation of the validation windows gives 96 training and 96 disjoint held-out windows. The response stream consists of the first 170 windows of the test split (32,640 tokens). Queries are the positions $t \in [64, 192)$, so every key of the prefix $K = \{0, \dots, W - 1\}$, $W \le 64$, is available to every query.

Capture. For the analysis, models load in float32 with eager attention. Unprojected values are captured at the value projection (the fused QKV projection for GPT-2, the interleaved projection for NeoX and the value projection with grouped-query repetition for Llama), head outputs ζ_h at the input of the output projection, and projected values and outputs are formed with each head's

output-projection slice O_h. On the first batch, the pipeline checks in the first, middle and last layers that attention times values reproduces ζ_h and that the slices reproduce the output projection, and it aborts otherwise. The largest relative error over all runs was 2.3×10^{-6}. Second moments are accumulated in float64 without TF32.

Estimators. Each signature $p_h^\star$ averages, over training queries and windows, the rows renormalized over K (normalize, then average). Within one forward pass, the projected values $V_{h,K}(X_r)$ of the prefix of window X_r are shared by all queries, so the paired block Gram uses the window as the independent unit:

$$\hat{G}_{hg}^K = \frac{1}{n}\sum_{r=1}^{n} V_{h,K}(X_r) V_{g,K}(X_r)^\top,$$
$$\hat{d}_G(h,g)^2 = (p_h^\star)^\top \hat{G}_{hh}^K p_h^\star + (p_g^\star)^\top \hat{G}_{gg}^K p_g^\star - 2(p_h^\star)^\top \hat{G}_{hg}^K p_g^\star,$$

with $n = 96$ and both heads evaluated on the same windows. The complete outputs $y_h = O_h \zeta_h$ and the RMS quantities $\mathcal{D}_{hg}$, $S_{h\leftarrow g}$ and $\Gamma_{h\leftarrow g}$ are averaged over all queries of the held-out windows, or of the training windows when they drive a selection rule, and are computed exactly from accumulated second moments. The plug-in radius uses for R_h the largest projected value norm of head h over all keys of the held-out windows. Head norms are training RMS norms of y_h. The norm reference $\sqrt{N_h^2 + N_g^2}$ uses these training norms, and the normalized alignment $A_{hg} = (N_h^2 + N_g^2 - \mathcal{D}_{hg}^2)/(2N_hN_g) = \mathbb{E}\langle y_h, y_g\rangle/(N_hN_g)$ is computed with $\mathcal{D}_{hg}$ and both norms on the training windows, so that the identity holds exactly. Row scores use the training signatures, and masked scores delete key 0 and renormalize.

Metrics. Closer-of-two agreement uses all within-layer triples $(h; g, k)$ with $g, k \neq h$ and excludes ties. Spearman correlations over head pairs pool all within-layer pairs, and the nearest neighbour is the minimizer over the other heads of the layer.

Interventions. Evaluation reloads the model with scaled-dot-product attention and TF32 matrix multiplication and edits the input of the output projection. Removal sets $\zeta_h = 0$, and substitution $h \leftarrow g$ replaces ζ_h by ζ_g from the same forward pass at every position. With baseline mean NLL per predicted token $\mathcal{L}_0$ on the response stream and $\mathcal{L}_{h\leftarrow g}$ after substituting g into h, the pair response is the signed average

$$R_{hg} = \tfrac{1}{2}\big[(\mathcal{L}_{h\leftarrow g} - \mathcal{L}_0) + (\mathcal{L}_{g\leftarrow h} - \mathcal{L}_0)\big].$$

Per seed, 400 unordered within-layer pairs (300 on Llama) are drawn uniformly without replacement and evaluated in both directions, and every predictor is scored on the same pairs. Pair-level correlations use the mean of the two directional responses and, for the directional predictors S and Γ, the mean of the two directional values, and the directional analysis treats each direction as one observation. Bootstrap intervals resample the evaluated pairs with replacement within each layer 1,000 times per seed and report the 2.5 and 97.5 percentiles of the difference of within-layer correlations. The within-group analysis pools the directed substitutions of the three seeds whose two heads share a key–value group. Budgets round 10, 20 and 30 percent of the heads of each layer to the nearest integer (1, 2 and 4 of 12 heads and 3, 6 and 10 of 32). Random removal draws the heads of each layer uniformly, random substitution draws disjoint ordered pairs, and each uses five draws per seed. Llama-3.2-1B has 16 layers with 32 query heads and 8 key–value heads per layer. Values are shared within a key–value group and repeated to its query heads before projection, so each query head has its own attention row, output and projection slice, and budgets count query heads.

Scope of the estimand. Each signature is a 64-key prefix renormalization of the row, so the estimated distance concerns normalized prefix outputs. It does not include the mass assigned to that prefix or the contribution of later keys. To state the distinction precisely, let a be a full attention row, K its first 64 keys, $m_K = \sum_{i\in K} a_i$ and $p_K = a_K/m_K$. The complete output is

$$z = m_K V_K^\top p_K + V_{K^c}^\top a_{K^c}.$$

The normalized-prefix target retains $V_K^\top p_K$. Fixed signatures introduce a further distinction from the input-dependent p_K. The uniform bound of Theorem 1(c) permits signatures selected from the Gram-estimation sample, provided that its independent-sample and bounded-value assumptions hold. The empirical accuracy measurements do not by themselves verify these sampling assumptions.

Head-slot substitution also differs locally from the cross-head projected-value comparison. Let U_h contain the unprojected values and let O_h be the output-projection slice, so that $V_h^\top = O_h U_h^\top$. Substituting head g into head h's slot changes the residual contribution by

$$O_h(U_g^\top q - U_h^\top p),$$

whereas the cross-head Gram compares

$$O_g U_g^\top q - O_h U_h^\top p.$$

These expressions need not agree, and Proposition 2 bounds their difference by $\Gamma_{h\leftarrow g}$. Unless marked as seed 0 or as pooled over seeds, the reported summaries are means over three resampling seeds.

BOS at key 0. The main runs do not prepend BOS, so key 0 is the first text token of each window. Prepending BOS to every window (seed 0) makes key 0 the largest key in 96, 92 and 100 percent of head signatures (Table 4).

Reproducibility of the pipeline. A second run recomputed all training and held-out statistics and all intervention costs shared with the first run, adding the recipient-slot and complete-output rules. It reproduced every Gram estimate, row-score correlation and shared intervention cell exactly.

B.1 Sampling assumptions and scope of the empirical claims

The fixed-signature target requires a signature $p_h^\star$ for each head, projected values in a common output space and blocks $G_{hg}^K = \mathbb{E}[V_{h,K} V_{g,K}^\top]$ formed with the two heads evaluated on the same input. Its sample counterpart averages these paired products (estimators above). Sampling the two heads on different inputs would define a different target, and averaging output norms would differ from taking the square root of the average squared norm.

Theorem 1(c) requires independent samples of bounded projected values. When it is applied to our protocol, the independent unit is the sequence, not the query position, because queries within a sequence are dependent. The radius R is a population bound, and a sample maximum is not by itself an almost-sure bound.

The plug-in radii use the sample maximum of projected value norms on the held-out windows, which is not an almost-sure bound, and the projection discrepancies are estimated on the same windows. Loss-sensitivity factors are not estimated. The transfer and response bounds therefore remain sufficient analytical conditions, and the intervention results provide direct evidence only about the tested selection rules at the stated budgets.

Compute resources Hardware: two NVIDIA GeForce RTX 4090 GPUs (24 GB each) and 20 CPU cores on one server. Software: Python 3, PyTorch 2.4.1 with CUDA 12.1, Transformers 4.46.3, Datasets 3.1.0, NumPy 2.4 and SciPy 1.17. Table 3 lists the device and wall-clock time of every experiment, taken from the launcher and job logs. All randomness is seeded. The CPU suite regenerates every verification table and figure from a single entry point, and the GPU launcher writes one log and one result file per job, and both are part of the supplementary material.

Calibration protocol The release of Papamichalis & Ruane (2026) contains, per model and seed, per-layer mean sink mass s_ℓ and mean content entropy H_ℓ^c (nats) computed on $W{=}64$ sliding windows over held-out text (or patches for ViT). To generate a *calibrated row*, we first sample a record uniformly from the pool of 13 models with up to three seeds each (37 model–seed files), which contains 615 usable per-layer records. Three ViT layer-11 records with $H_\ell^c \geq \log(W{-}1)$ are excluded by this calibration procedure. Values strictly above $\log(W{-}1)$ are impossible for $W-1$ content positions, and equality corresponds to uniform content and is also excluded by the filter. We next draw standard normal logits $z \in \mathbb{R}^{63}$, centre them and bisect a scale σ so that $H\big(\mathcal{C}(e^{\sigma z})\big) = H_\ell^c$ to 10^{-10}. We form the pair partner by jittering the logits ($\tau = 0.15$) and re-centring. The jittered partner is not constrained to have the entropy H_ℓ^c, so the pair distribution is a synthetic perturbation model anchored to selected summary statistics, not a reconstruction of the original attention rows. We then compose each content with the measured s_ℓ. For Exp. 5, the two rows of a pair take *independent* sink masses from two records, so their sink masses can differ and the sink channel can be active. This places each sweep at empirically observed operating points without raw attention tensors. The identity, exponent and defect conclusions are distribution-free, while the Exp. 5 regime panel depends on the sampling distribution.

Table 3: Compute resources and wall-clock time per experiment, from the launcher and job logs.

experiment	device	wall-clock time
Exps. 1–5, verification suite	one CPU core	about 2 min in total
Exp. 8, factorial regime map (9 chunks in parallel)	9 CPU cores	about 1 min wall, about 10 CPU-minutes
Exp. 6, tables and figure from the result files	one CPU core	under 1 min
Exp. 7, one model and seed: analysis, 63 pruning evaluations and 400 substitution pairs (300 on Llama) in both directions	one RTX 4090	GPT-2 3.8 min, Pythia-160M 2.8 min, Llama-3.2-1B 15.3–15.6 min
Exp. 7, pruning-only rerun adding the recipient-slot and complete-output rules (72 evaluations per job)	two RTX 4090	5.2 min for all nine jobs, including smoke tests
BOS diagnostics (analysis and baseline, seed 0)	one RTX 4090	0.2–0.5 min per model
smoke tests (8 windows per model)	one RTX 4090	0.2–0.6 min per model
launcher for all main runs (smoke tests, nine jobs, three diagnostics, summaries)	two RTX 4090	34 min from launch to summaries

C RESULT TABLES FOR THE REAL-MODEL EXPERIMENTS

This appendix collects the complete result tables of the real-model experiments (Exp. 7). Every number quoted in the main text appears here, in Appendix D or in the number log of the supplementary material.

Prefix width. With $W = 16$, 32 and 64, the training Gram distance predicts the held-out $\mathcal{D}$ with pooled Spearman correlations of 0.53, 0.60 and 0.67 on GPT-2, 0.51, 0.52 and 0.52 on Pythia and 0.49, 0.52 and 0.57 on Llama. Over the same widths, the attention mass inside the prefix rises from 0.43 to 0.53, from 0.22 to 0.44 and from 0.55 to 0.64 (Table 6).

Table 4: Controls and sink statistics on the three language models, means over three seeds except the BOS rows (seed 0). Verdict agreement is the fraction of within-layer closer-of-two triples $(h; g, k)$, ties excluded, on which a score computed from the training signatures and the held-out Gram distance give the same verdict. The null rows compare the Gram distances of the training and held-out windows. Nearest-neighbour rows give the fraction of heads whose nearest neighbour changes. Key-0 rows use the training signatures ($W = 64$) and full attention rows. The reversal rows count all within-layer triples. The alignment rows use $A_{hg} = \mathbb{E}\langle y_h, y_g\rangle/(N_h N_g)$ with the complete outputs and norms of the training windows, the smallest-norm row uses $\mathcal{D}$ on the training windows, and these three rows pool the pairs and heads of the three seeds (median and fractions over the pool). Perplexities refer to the response stream of 170 non-overlapping 192-token windows. The BOS rows prepend the BOS token to every window, so each window holds one fewer text token, and their perplexity is computed with eager attention and without TF32.

quantity	GPT-2	Pythia-160M	Llama-3.2-1B
kept cosine verdict agreement with $\hat{d}_G$	0.51	0.54	0.55
kept JS verdict agreement with $\hat{d}_G$	0.49	0.54	0.55
masked cosine verdict agreement with $\hat{d}_G$	0.45	0.55	0.52
masked JS verdict agreement with $\hat{d}_G$	0.45	0.54	0.52
Aitchison d_A verdict agreement with $\hat{d}_G$	0.47	0.52	0.53
$\hat{d}_G$ training vs held-out verdict agreement (null)	0.992	0.985	0.994
nearest neighbour changed, kept cosine vs $\hat{d}_G$	0.97	0.85	0.96
nearest neighbour changed, masked JS vs $\hat{d}_G$	0.94	0.89	0.98
nearest neighbour changed, training vs held-out $\hat{d}_G$ (null)	0.01	0.05	0.04
heads whose signature is largest at key 0	0.96	0.78	0.90
with BOS at key 0 (seed 0)	0.96	0.92	1.00
key-0 weight / median other-key weight (median over heads)	254	43	492
with BOS at key 0 (seed 0)	240	57	607
mean full-row attention on key 0	0.40	0.14	0.51
triples whose kept JS verdict sink masses alone fix, percent (Prop. 4)	18.0	0.05	15.2
triples with opposite kept and masked JS verdicts, percent	21.4	8.0	26.0
median normalized alignment A_{hg} of complete head outputs	0.004	0.014	0.003
pairs with $\lvert A_{hg}\rvert < 0.1$, percent	63	84	91
heads whose nearest neighbour in $\mathcal{D}$ is the smallest-norm head, percent	66	86	77
unedited perplexity on the response stream	53.7	56.1	18.8
with BOS at the start of every window (seed 0)	56.0	62.5	20.6

Table 5: Predicting the held-out complete-output distance $\mathcal{D}$ from scores computed on the training windows, over within-layer head pairs, means over three seeds (standard deviations below 0.06). Pooled is the Spearman correlation over all within-layer pairs of a model, within is the mean over layers of the correlation within one layer, triples is the closer-of-two verdict agreement over all within-layer triples with ties excluded, and NN is the fraction of heads whose nearest neighbour under $\mathcal{D}$ is recovered, with chance levels $1/(H-1)$ of 0.09, 0.09 and 0.03. The two reference rows measure $\mathcal{D}$ directly and its norm-only part $\sqrt{N_h^2 + N_g^2}$ on the training windows, and $\lvert N_h - N_g\rvert$ is the difference of the training RMS output norms. Euclidean is the Euclidean distance between signatures, added after the first results.

score	GPT-2				Pythia-160M				Llama-3.2-1B			
	pooled	within	triples	NN	pooled	within	triples	NN	pooled	within	triples	NN
Gram distance $\hat{d}_G$ (training)	0.67	0.62	0.71	0.25	0.52	0.42	0.65	0.24	0.57	0.53	0.70	0.23
reference: $\mathcal{D}$ on training windows	1.00	1.00	0.98	0.93	1.00	0.99	0.98	0.97	1.00	1.00	0.98	0.86
reference: $\sqrt{N_h^2 + N_g^2}$ (training)	0.98	0.94	0.88	0.65	0.97	0.96	0.92	0.86	1.00	0.98	0.94	0.71
$\lvert N_h - N_g\rvert$ (training)	0.47	0.42	0.58	0.12	0.49	0.34	0.57	0.13	0.53	0.34	0.53	0.05
kept cosine	0.21	0.20	0.54	0.05	0.08	0.18	0.56	0.09	-0.08	0.23	0.56	0.03
kept JS	0.06	0.14	0.53	0.04	0.13	0.14	0.54	0.09	0.20	0.19	0.55	0.03
masked cosine	-0.26	-0.06	0.48	0.08	0.07	0.18	0.55	0.08	0.10	0.05	0.51	0.02
masked JS	-0.34	-0.06	0.48	0.09	0.05	0.11	0.54	0.09	0.01	0.04	0.51	0.02
Aitchison d_A	-0.35	0.03	0.51	0.09	0.02	0.02	0.52	0.07	-0.05	0.06	0.52	0.03
Euclidean	0.19	0.15	0.52	0.06	0.18	0.17	0.55	0.09	0.35	0.20	0.55	0.04

Table 6: Proposition 1 on real heads, held-out windows, means over three seeds. $\bar{m}$ is the mean attention mass on the prefix K of width W. The ratios are medians over heads of the realized RMS residual $\|r\|$ of the fixed-signature output, the plug-in radius $\mathcal{E}$, with R_h the sample maximum of projected value norms, and the RMS output norm $\|y\|$. Bound holds is the fraction of head pairs with $|\mathcal{D} - d_G| \le \mathcal{E}_h + \mathcal{E}_g$, which equals one by construction for in-sample radii and serves as an implementation check. The last two columns give the fraction of heads whose nearest neighbour the margin certifies with the radii or with the realized residuals.

model	W	$\bar{m}$	$\|r\|/\|y\|$	$\mathcal{E}/\|y\|$	$\|r\|/\mathcal{E}$	bound holds	cert. ($\mathcal{E}$)	cert. ($\|r\|$)
GPT-2	16	0.43	0.94	5.6	0.16	1.000	0.00	0.00
GPT-2	32	0.46	0.90	5.4	0.16	1.000	0.00	0.00
GPT-2	64	0.53	0.86	5.2	0.16	1.000	0.00	0.00
Pythia-160M	16	0.22	1.05	8.2	0.13	1.000	0.00	0.00
Pythia-160M	32	0.30	1.02	8.0	0.13	1.000	0.00	0.00
Pythia-160M	64	0.44	1.01	7.7	0.13	1.000	0.00	0.00
Llama-3.2-1B	16	0.55	1.07	7.0	0.15	1.000	0.00	0.00
Llama-3.2-1B	32	0.58	1.04	6.9	0.15	1.000	0.00	0.00
Llama-3.2-1B	64	0.64	1.00	6.6	0.15	1.000	0.00	0.00

Table 7: Spearman correlation between each predictor and the substitution cost on identical head pairs (400 per seed, 300 on Llama), mean and standard deviation over three seeds. Pooled correlations use all pairs. Within-layer correlations are computed inside each layer with at least eight evaluated pairs and averaged with weights equal to the number of pairs. S, $\mathcal{D}$ and Γ are measured on the held-out windows, the Gram distance and the row scores on the training windows, and head norms are training RMS output norms. The lower block treats the two directions of each pair separately, except for the last two rows.

predictor	GPT-2		Pythia-160M		Llama-3.2-1B	
	pooled	within layer	pooled	within layer	pooled	within layer
recipient-slot $S_{h\leftarrow g}$	$0.12_{\pm 0.05}$	$0.60_{\pm 0.02}$	$0.24_{\pm 0.05}$	$0.58_{\pm 0.02}$	$0.22_{\pm 0.07}$	$0.72_{\pm 0.02}$
complete output $\mathcal{D}$	$0.17_{\pm 0.04}$	$0.54_{\pm 0.02}$	$0.33_{\pm 0.04}$	$0.56_{\pm 0.02}$	$0.13_{\pm 0.08}$	$0.54_{\pm 0.04}$
larger head norm	$0.19_{\pm 0.05}$	$0.52_{\pm 0.02}$	$0.24_{\pm 0.04}$	$0.53_{\pm 0.02}$	$0.14_{\pm 0.07}$	$0.54_{\pm 0.06}$
projection mismatch Γ	$0.11_{\pm 0.03}$	$0.56_{\pm 0.02}$	$0.25_{\pm 0.05}$	$0.58_{\pm 0.02}$	$0.12_{\pm 0.08}$	$0.56_{\pm 0.03}$
Gram $\hat{d}_G$	$0.45_{\pm 0.04}$	$0.41_{\pm 0.05}$	$0.30_{\pm 0.04}$	$0.23_{\pm 0.02}$	$0.32_{\pm 0.08}$	$0.35_{\pm 0.04}$
kept cosine	$0.50_{\pm 0.01}$	$0.20_{\pm 0.02}$	$0.36_{\pm 0.05}$	$0.23_{\pm 0.02}$	$0.50_{\pm 0.03}$	$0.27_{\pm 0.01}$
kept JS	$0.33_{\pm 0.03}$	$0.17_{\pm 0.01}$	$-0.14_{\pm 0.05}$	$0.16_{\pm 0.02}$	$0.28_{\pm 0.09}$	$0.25_{\pm 0.02}$
masked cosine	$0.03_{\pm 0.02}$	$-0.01_{\pm 0.01}$	$-0.05_{\pm 0.05}$	$0.20_{\pm 0.02}$	$0.02_{\pm 0.09}$	$0.18_{\pm 0.02}$
masked JS	$0.03_{\pm 0.04}$	$-0.01_{\pm 0.03}$	$-0.27_{\pm 0.04}$	$0.11_{\pm 0.03}$	$0.02_{\pm 0.10}$	$0.17_{\pm 0.03}$
Aitchison d_A	$0.15_{\pm 0.06}$	$0.07_{\pm 0.03}$	$-0.51_{\pm 0.03}$	$-0.02_{\pm 0.04}$	$0.06_{\pm 0.09}$	$0.18_{\pm 0.04}$
per direction, pooled: $S_{h\leftarrow g}/N_h$	$-0.03_{\pm 0.02}$		$-0.14_{\pm 0.00}$		$0.08_{\pm 0.04}$	
per direction, within layer: $S_{h\leftarrow g}/N_h$	$0.08_{\pm 0.03}$		$-0.16_{\pm 0.02}$		$0.09_{\pm 0.05}$	
per direction, within layer: $S_{h\leftarrow g}$	$0.56_{\pm 0.02}$		$0.53_{\pm 0.02}$		$0.66_{\pm 0.03}$	
per direction, within layer: $\mathcal{D}$	$0.41_{\pm 0.00}$		$0.45_{\pm 0.03}$		$0.45_{\pm 0.01}$	
per direction, within layer: $\hat{d}_G$	$0.31_{\pm 0.03}$		$0.15_{\pm 0.02}$		$0.29_{\pm 0.04}$	
costlier direction predicted by S	$0.66_{\pm 0.02}$		$0.66_{\pm 0.01}$		$0.66_{\pm 0.01}$	
cost vs layer index	$-0.46_{\pm 0.04}$		$-0.78_{\pm 0.02}$		$-0.33_{\pm 0.03}$	
median $\Gamma/\mathcal{D}$	$0.96_{\pm 0.00}$		$0.96_{\pm 0.00}$		$1.00_{\pm 0.00}$	

Table 8: Bootstrap intervals for the within-layer Spearman correlation with substitution cost. For each seed, the evaluated pairs are resampled with replacement within each layer 1,000 times, the within-layer correlations of S, $\mathcal{D}$ and the larger head norm are recomputed, and the columns give the observed gaps with the 2.5 and 97.5 percentiles of the resampled gaps.

model	seed	S	$\mathcal{D}$	norm	$S - \mathcal{D}$ [95% interval]	$S-$norm [95% interval]
GPT-2	0	0.59	0.52	0.51	0.063 [0.022, 0.103]	0.076 [0.030, 0.123]
GPT-2	1	0.63	0.56	0.54	0.067 [0.020, 0.112]	0.084 [0.024, 0.137]
GPT-2	2	0.59	0.54	0.51	0.052 [0.003, 0.101]	0.086 [0.027, 0.146]
Pythia-160M	0	0.59	0.58	0.54	0.009 [-0.031, 0.057]	0.051 [0.002, 0.099]
Pythia-160M	1	0.56	0.54	0.52	0.021 [-0.021, 0.059]	0.037 [-0.019, 0.092]
Pythia-160M	2	0.59	0.56	0.51	0.025 [-0.017, 0.062]	0.076 [0.021, 0.122]
Llama-3.2-1B	0	0.73	0.50	0.49	0.229 [0.144, 0.331]	0.241 [0.157, 0.329]
Llama-3.2-1B	1	0.70	0.54	0.52	0.154 [0.074, 0.234]	0.173 [0.098, 0.244]
Llama-3.2-1B	2	0.74	0.59	0.60	0.147 [0.064, 0.220]	0.131 [0.044, 0.207]

Table 9: Single substitutions within key–value groups on Llama-3.2-1B, a post hoc analysis of the evaluated pairs whose two heads share a group (93 distinct pairs in 96 draws over the three seeds, both directions). Within-layer correlations pool the directed substitutions of the three seeds and use layers with at least eight of them. Costs are changes in NLL per token on the response stream. The quartiles partition the directed substitutions by $S_{h\leftarrow g}/N_h$, with S on the held-out windows and N_h the training RMS output norm.

quantity	value
directed substitutions (pairs per seed)	192 (38, 25, 33)
within-layer Spearman with cost: $S_{h\leftarrow g}$	0.51
within-layer Spearman with cost: $\mathcal{D}_{hg}$	0.12
within-layer Spearman with cost: recipient norm N_h	0.33
pooled Spearman with cost: $S_{h\leftarrow g}$	0.03
pooled Spearman with cost: layer index	-0.27
costlier direction predicted by S (mean over seeds)	0.73
median $S_{h\leftarrow g}/N_h$; fraction below one	0.89; 0.64
mean cost by quartile of $S_{h\leftarrow g}/N_h$ (nats per token)	0.0021, 0.0031, 0.0042, 0.0055
median cost by quartile of $S_{h\leftarrow g}/N_h$	0.0019, 0.0021, 0.0037, 0.0022
other pairs: within-layer Spearman with cost, $S_{h\leftarrow g}$	0.63
median cost, same group versus other pairs	0.0021 versus 0.0067

Table 10: Removal interventions: increase in perplexity over the unedited model when the outputs of the selected heads are zeroed at 10, 20 and 30 percent of heads per layer (1, 2 and 4 of 12 heads and 3, 6 and 10 of 32), mean and standard deviation over three seeds. Pair rules remove the smaller-norm member of the closest still-unused pair under the named score, with $\mathcal{D}$ measured on the training windows. Random rows average five draws per seed, with the standard deviation taken over the three seed means. Bold marks the smallest mean per column among the non-random criteria.

rule	10%	20%	30%
GPT-2 (unedited perplexity 53.7)			
Gram distance $\hat{d}_G$	$+44.2_{\pm 2.1}$	$+96.3_{\pm 0.0}$	$\mathbf{+167_{\pm 2.7}}$
complete-output distance $\mathcal{D}$	$+3.7_{\pm 0.4}$	$\mathbf{+19.2_{\pm 0.9}}$	$+455_{\pm 124}$
norm only	$\mathbf{+3.6_{\pm 0.4}}$	$+1656_{\pm 142}$	$+14092_{\pm 308}$
kept cosine	$+61.4_{\pm 85.7}$	$+178_{\pm 208}$	$+1176_{\pm 1337}$
kept JS	$+166_{\pm 138}$	$+453_{\pm 484}$	$+2475_{\pm 1830}$
masked cosine	$+448_{\pm 491}$	$+440_{\pm 235}$	$+1034_{\pm 69.6}$
masked JS	$+260_{\pm 227}$	$+270_{\pm 10.9}$	$+1357_{\pm 263}$
random (5 draws per seed)	$+168_{\pm 248}$	$+343_{\pm 165}$	$+2404_{\pm 2937}$
Pythia-160M (unedited perplexity 56.1)			
Gram distance $\hat{d}_G$	$+4.0_{\pm 0.4}$	$+15.3_{\pm 1.3}$	$+60.8_{\pm 8.3}$
complete-output distance $\mathcal{D}$	$\mathbf{+1.3_{\pm 0.0}}$	$+10.4_{\pm 3.8}$	$+54.5_{\pm 6.6}$
norm only	$\mathbf{+1.3_{\pm 0.0}}$	$\mathbf{+6.2_{\pm 0.0}}$	$\mathbf{+40.5_{\pm 2.7}}$
kept cosine	$+11.1_{\pm 0.6}$	$+29.4_{\pm 1.3}$	$+51.7_{\pm 9.5}$
kept JS	$+11.4_{\pm 0.7}$	$+29.1_{\pm 0.8}$	$+69.0_{\pm 13.7}$
masked cosine	$+10.6_{\pm 0.4}$	$+29.1_{\pm 2.3}$	$+54.6_{\pm 8.5}$
masked JS	$+12.0_{\pm 1.8}$	$+25.8_{\pm 0.4}$	$+67.3_{\pm 17.6}$
random (5 draws per seed)	$+14.8_{\pm 2.8}$	$+42.8_{\pm 4.7}$	$+149_{\pm 6.1}$
Llama-3.2-1B (unedited perplexity 18.8)			
Gram distance $\hat{d}_G$	$+7.7_{\pm 3.3}$	$\mathbf{+18.4_{\pm 6.9}}$	$\mathbf{+100_{\pm 23.0}}$
complete-output distance $\mathcal{D}$	$\mathbf{+5.3_{\pm 2.5}}$	$+824_{\pm 408}$	$+1139_{\pm 382}$
norm only	$+8.4_{\pm 0.4}$	$+26.6_{\pm 14.9}$	$+405_{\pm 11.6}$
kept cosine	$+15.2_{\pm 1.1}$	$+71.3_{\pm 16.0}$	$+365_{\pm 191}$
kept JS	$+13.3_{\pm 0.1}$	$+73.8_{\pm 11.3}$	$+1686_{\pm 2289}$
masked cosine	$+16.2_{\pm 1.4}$	$+64.8_{\pm 6.6}$	$+218_{\pm 72.7}$
masked JS	$+15.7_{\pm 2.4}$	$+92.5_{\pm 32.7}$	$+150_{\pm 63.3}$
random (5 draws per seed)	$+29.5_{\pm 20.8}$	$+431_{\pm 220}$	$+2801_{\pm 1901}$

Table 11: Substitution interventions: increase in perplexity when each selected head receives its partner's unprojected output, with the budgets and conventions of Table 10. Pair rules overwrite the smaller-norm member of each selected pair. The recipient-slot rule repeatedly takes the still-unused ordered pair with the smallest training-window $S_{h\leftarrow g}$ and overwrites h with g.

rule	10%	20%	30%
GPT-2 (unedited perplexity 53.7)			
Gram distance $\hat{d}_G$	$+23.9_{\pm 0.3}$	$+70.5_{\pm 0.0}$	$\mathbf{+634_{\pm 3.0}}$
recipient-slot distance $S_{h\leftarrow g}$	$\mathbf{+3.8_{\pm 0.5}}$	$+948_{\pm 29.3}$	$+1041_{\pm 168}$
complete-output distance $\mathcal{D}$	$+7.9_{\pm 0.4}$	$\mathbf{+47.6_{\pm 12.8}}$	$+1112_{\pm 479}$
kept cosine	$+146_{\pm 215}$	$+465_{\pm 504}$	$+4260_{\pm 4421}$
kept JS	$+361_{\pm 320}$	$+2430_{\pm 2133}$	$+8946_{\pm 6442}$
masked cosine	$+382_{\pm 382}$	$+569_{\pm 271}$	$+2233_{\pm 312}$
masked JS	$+245_{\pm 205}$	$+416_{\pm 97.3}$	$+2295_{\pm 330}$
random (5 draws per seed)	$+96.3_{\pm 55.4}$	$+378_{\pm 254}$	$+3125_{\pm 1977}$
Pythia-160M (unedited perplexity 56.1)			
Gram distance $\hat{d}_G$	$+8.8_{\pm 1.1}$	$+28.4_{\pm 0.3}$	$+121_{\pm 7.5}$
recipient-slot distance $S_{h\leftarrow g}$	$+3.8_{\pm 0.1}$	$\mathbf{+16.5_{\pm 0.2}}$	$\mathbf{+108_{\pm 1.4}}$
complete-output distance $\mathcal{D}$	$\mathbf{+3.8_{\pm 0.1}}$	$+18.2_{\pm 4.4}$	$+130_{\pm 14.3}$
kept cosine	$+19.6_{\pm 1.9}$	$+56.8_{\pm 3.8}$	$+156_{\pm 20.5}$
kept JS	$+23.6_{\pm 1.9}$	$+66.3_{\pm 4.5}$	$+191_{\pm 31.7}$
masked cosine	$+18.7_{\pm 1.3}$	$+61.6_{\pm 1.4}$	$+179_{\pm 26.2}$
masked JS	$+24.6_{\pm 2.7}$	$+68.6_{\pm 4.1}$	$+222_{\pm 57.8}$
random (5 draws per seed)	$+20.3_{\pm 1.9}$	$+82.9_{\pm 24.6}$	$+359_{\pm 122}$
Llama-3.2-1B (unedited perplexity 18.8)			
Gram distance $\hat{d}_G$	$+10.1_{\pm 4.6}$	$+38.4_{\pm 9.7}$	$+1980_{\pm 1387}$
recipient-slot distance $S_{h\leftarrow g}$	$\mathbf{+1.1_{\pm 0.1}}$	$\mathbf{+5.6_{\pm 4.5}}$	$\mathbf{+19.0_{\pm 3.7}}$
complete-output distance $\mathcal{D}$	$+11.9_{\pm 7.1}$	$+4222_{\pm 1678}$	$+11210_{\pm 6716}$
kept cosine	$+12.1_{\pm 1.3}$	$+31.1_{\pm 5.7}$	$+124_{\pm 25.9}$
kept JS	$+7.3_{\pm 0.2}$	$+22.1_{\pm 1.6}$	$+87.8_{\pm 9.1}$
masked cosine	$+6.4_{\pm 1.2}$	$+15.7_{\pm 2.2}$	$+60.1_{\pm 9.2}$
masked JS	$+5.4_{\pm 1.2}$	$+14.1_{\pm 3.7}$	$+42.2_{\pm 7.2}$
random (5 draws per seed)	$+384_{\pm 550}$	$+7581_{\pm 6722}$	$+12753_{\pm 5379}$

D PREDICTIONS AND OUTCOMES

Table 12 records every prediction, whether it was fixed before the run that tested it, its outcome and where the evidence is. "Before" means stated before the run whose result it concerns, which is not a public pre-registration. All numbers refer to the three language models of Section 6.

Table 12: Prediction ledger. Outcome "held in part" means that the prediction held after a change of analysis that was made after the first result, or held in most but not all settings.

	prediction	fixed	outcome	evidence
P1	Gram verdicts are stable across the training and held-out windows, and row-score verdicts disagree with them far beyond that null	before	held	Table 4
P2	key 0 receives the largest attention weight in most heads	before	held	Table 4
P3	$\hat{d}_G$ selection costs less than cosine and JS selection at 10 to 30 percent on at least two of three models	before	held	Tables 10, 11
P4	$S_{h\leftarrow g}$ predicts substitution cost better than $\mathcal{D}$ and $\hat{d}_G$	before	held in part	Table 7
P5	selecting substitutions by $S_{h\leftarrow g}$ costs less than Gram selection	before	held in part	Table 11
P6	the certificate of Proposition 1 certifies a nontrivial fraction of nearest neighbours at $W = 64$	before	failed	Table 6
P7	substitution beats removal only where the selected pairs satisfy $S < N_h$	after	failed as stated	Tables 10, 11, note P7
P8	Euclidean signature distance predicts $\mathcal{D}$ poorly	after	held	Table 5
P9	the certificate needs $2\bar{m}_h + \bar{m}_g + \bar{m}_k > 3$ (Remark 2), a derivation	after	consistent	Table 6
P10	masking lowers the agreement of cosine and JS with the output-based targets where $\lvert\lambda - \frac{1}{2}\rvert$ is large	after	held	Table 1
P11	$\mathcal{D}$ is set mostly by the two head norms, so head outputs within a layer are nearly orthogonal	after	held	Tables 4, 5, note P11
P12	the scale-free ratio $S_{h\leftarrow g}/N_h$ predicts substitution cost pooled over layers	after	failed	Table 7
P13	the recipient-slot rule fails where few of its selected pairs satisfy $S < N_h$	after	consistent	note P13

P1. Training and held-out Gram verdicts agree on 0.985 to 0.994 of triples, whereas the five row scores of Table 4 agree with the Gram verdict on 0.45 to 0.55 (Euclidean distance, added later, appears in Table 5). **P2.** Key 0 is the largest key in 96, 78 and 90 percent of signatures, and in 96, 92 and 100 percent when BOS occupies key 0. **P3.** In removal the prediction holds in 8 of 9 settings, all but Pythia at 30 percent, and in substitution it holds at every budget on GPT-2 and Pythia and at none on Llama. **P4.** Rejected for correlations pooled over layers (0.12 to 0.24) and supported within layers (0.58 to 0.72, against at most 0.56 for $\mathcal{D}$ and 0.41 for $\hat{d}_G$). The within-layer analysis was chosen after the pooled result. **P5.** Supported in 7 of 9 settings and rejected on GPT-2 at 20 and 30 percent. **P6.** No head is certified on any model. **P7.** Over seven substitution rules and nine settings, substitution beats the best removal rule in 8 of 63 cases, all on Llama. Three are the recipient-slot cases, whose selected pairs have a median $S_{h\leftarrow g}/N_h$ of 0.70 to 0.73. Five are row-score rules at 20 or 30 percent, with medians of 1.19 to 1.43. Masked JS at 10 percent has a median of 0.89 but costs $+5.4$ against $+5.3$ for removal. On GPT-2 and Pythia every rule has a median above one and none beats removal. A median below one thus coincides with beating removal in 57 of 63 cases overall, against 55 of 63 for the constant prediction that substitution never beats removal, and in only 15 of the 21 Llama cases. The exact local condition does not imply downstream superiority over a different removal rule. **P8.** Euclidean distance between signatures, which Theorem 1(b) makes exact for a common isotropic value matrix, predicts held-out $\mathcal{D}$ within layers at 0.15, 0.17 and 0.20, with triple agreement 0.52 to 0.55 and nearest-neighbour recovery of 4 to 9 percent. The idealized exact-recovery setting assumes a common value matrix and fixed rows, so this cross-head comparison reports the poor prediction without isolating anisotropy as its cause. **P9.** At $W = 64$ the mean retained mass is 0.53, 0.44 and 0.64, below three quarters. **P10.** A hypothesis motivated by the terms that the decompositions of Theorems 1(b) and 2 show masking discards, not a consequence of the theorems. Supported on

GPT-2 and Llama, with median skews of 0.15 and 0.18, in all 24 seed-level comparisons, by 0.05 to 0.27, whereas on Pythia, with a median skew of 0.03 and 14 percent of the attention on key 0, the mean losses are 0.006 to 0.046 and 3 of the 12 seed-level differences are zero or negative. The skew does not order the two models with strong sinks, since Llama has the larger skew and the smaller loss. **P11.** An observation made when the norm reference $\sqrt{N_h^2 + N_g^2}$ was added as a baseline: it predicts held-out $\mathcal{D}$ within layers at 0.94, 0.96 and 0.98, the normalized alignment A_{hg} has median 0.004, 0.014 and 0.003 with $|A_{hg}| < 0.1$ for 63, 84 and 91 percent of pairs, and the nearest neighbour in $\mathcal{D}$ is the smallest-norm head of the layer for 66, 86 and 77 percent of heads. On Llama, the within-group pairs have median alignment -0.03 and the pairs with $S < N_h$ have median alignment -0.08, so the substitution condition holds for pairs whose complete outputs are not aligned. On Pythia the 47 pairs with $S < N_h$ over three seeds have median alignment 0.69, and on GPT-2 the 57 such pairs have median alignment 0.10. **P12.** The ratio predicts the directed cost pooled over layers at -0.03, -0.14 and 0.08 and within layers at 0.08, -0.16 and 0.09, so it does not replace the within-layer analysis. **P13.** Among the pairs the recipient-slot rule selects at 10, 20 and 30 percent, the fraction satisfying $S_{h\leftarrow g} < N_h$ on the training windows is 0, 4.2 and 2.1 percent on GPT-2, 0, 0 and 1.4 percent on Pythia and 85, 78 and 79 percent on Llama. The rule fails on GPT-2 at 20 and 30 percent and succeeds on Llama at every budget. This is consistent with Corollary 1, which compares local perturbations, but the fractions do not predict the failures, since Pythia at 10 and 20 percent has no selected pair satisfying the condition and the rule costs 6.8 and 29.4 percent against 2.3 and 11.1 percent for the best removal rule, without failing.

E SINK–CONTENT IDENTITIES AND PROOFS

We now determine how standard row dissimilarities depend on sink mass and content. Throughout, $p = (s, (1-s)\pi)$ and $q = (t, (1-t)\rho)$ and H_b, KL_b and JS_b denote the entropy and divergences of the Bernoulli distributions $(s, 1-s)$ and $(t, 1-t)$. Throughout the appendices, $\mathrm{TV} = \frac{1}{2}\|p-q\|_1$, Pearson's $\chi^2(p\|q) = \sum_i (p_i - q_i)^2/q_i$, the Rényi divergences R_α and cosine dissimilarity $1-\cos(p,q)$ have their standard definitions, $\mathrm{BC}(p,q) = \sum_i \sqrt{p_i q_i}$ (Bhattacharyya, 1943), the Fisher–Rao distance $\mathrm{FR} = 2\arccos \mathrm{BC}$ (Rao et al., 1945; Amari & Nagaoka, 2000), $\mathrm{osc}(x) = \max_i x_i - \min_i x_i$, the congruent embedding $E_{D,k} : \mathcal{S}^D \to \mathcal{S}^{D+k-1}$ splits one coordinate into k equal parts, Markov morphisms between open simplices have no zero row (Campbell, 1986; Ay et al., 2017), and a family $\{\delta_D\}$ is *Markov-nonexpansive* if $\delta_{D'}(Tp, Tq) \le \delta_D(p,q)$ for every Markov morphism T. With $\Delta b = b(p) - b(q)$ and $d_A^\perp = d_A(\pi, \rho)$, $d_A(p,q)^2 = (\Delta b)^2 + (d_A^\perp)^2$ (Papamichalis & Ruane, 2026, Lemma 1).

Theorem 2 (Sink–content calculus, expanded statement)**.** (a) Chain rules.

$$H(p) = H_b(s) + (1-s)\, H(\pi),$$
$$\mathrm{KL}(p \,\|\, q) = \mathrm{KL}_b(s \,\|\, t) + (1-s)\, \mathrm{KL}(\pi \,\|\, \rho).$$

(b) Jensen–Shannon. *With* $\lambda = \frac{1-s}{2(1-\bar{s})} \in (0,1)$,

$$\mathrm{JS}(p,q) = \mathrm{JS}_b(s,t) + (1-\bar{s})\, \mathrm{JS}_\lambda(\pi, \rho),$$

where $\mathrm{JS}_\lambda(\pi,\rho) = \lambda \mathrm{KL}(\pi\|\mu) + (1-\lambda)\mathrm{KL}(\rho\|\mu)$, $\mu = \lambda\pi + (1-\lambda)\rho$, *is the* λ*-skew JS divergence (Lin, 1991; Nielsen, 2020).*

(c) Bhattacharyya, Hellinger, Fisher–Rao. $\mathrm{BC}(p,q) = \sqrt{st} + \sqrt{(1-s)(1-t)}\,\mathrm{BC}(\pi,\rho)$. *Consequently, with colatitudes* $\psi_p = \arccos\sqrt{s}$, $\psi_q = \arccos\sqrt{t}$,

$$\cos\tfrac{\mathrm{FR}(p,q)}{2} = \cos\psi_p \cos\psi_q + \sin\psi_p \sin\psi_q \cos\tfrac{\mathrm{FR}(\pi,\rho)}{2},$$

a spherical law of cosines: the sink masses are colatitudes on the Hellinger sphere and the content separation is the dihedral angle. Equivalently, $\mathrm{Hel}^2(p,q) = \mathrm{Hel}_b^2(s,t) + \sqrt{(1-s)(1-t)}\,\mathrm{Hel}^2(\pi,\rho)$ *with* $\mathrm{Hel}_b^2(s,t) = (\sqrt{s}-\sqrt{t})^2 + (\sqrt{1-s}-\sqrt{1-t})^2$*: an exact two-channel Hellinger law at arbitrary sink masses.*

(d) Matched sink. *For every* f*-divergence and all* s*,*

$$D_f(p_s, q_s) = (1-s)\, D_f(\pi, \rho),$$

hence $\mathrm{Hel}(p_s, q_s) = \sqrt{1-s}\,\mathrm{Hel}(\pi, \rho)$ *exactly. The Rényi divergences obey an exact log-mix law for orders* $\alpha \in (0,1)\cup(1,\infty)$*, whence* $R_\alpha(p_s, q_s) = C_\alpha(1-s) + O((1-s)^2)$ *with* $C_\alpha = \frac{e^{(\alpha-1)R_\alpha(\pi,\rho)}-1}{\alpha-1}$*. Since* $e^y \geq 1+y$*, with equality exactly at* $y=0$*,* $C_\alpha \geq R_\alpha(\pi,\rho)$ *for* $\alpha > 1$ *and* $C_\alpha \leq R_\alpha(\pi,\rho)$ *for* $\alpha \in (0,1)$*, with equality iff* $\pi = \rho$*, so the exponent is* 1 *but the constant differs from* $R_\alpha(\pi,\rho)$ *(proved below). In addition,* $d_A(p_s, q_s) = d_A(p_{s'}, q_{s'})$ *and* $d_H(p_s, q_s) = d_H(\pi, \rho)$ *for all* s, s' *(exact invariance).*

(e) Attenuation spectrum. *Fix* $\pi \neq \rho$*, assume* $0 < D_f(\pi,\rho) < \infty$ *for the* f*-divergences considered (excluding degenerate affine* f*) and let* $s \to 1$*. Then, for the scales listed,* $\delta(p_s, q_s) \sim C_\delta\,(1-s)^\gamma$ *with*

$\gamma = 2$	$\gamma = 1$	$\gamma = 1/2$	$\gamma = 0$
cosine	*Euclid,* TV*, all* D_f	Hel, FR, $\sqrt{\mathrm{JS}}$	$d_A,\ d_H$

with leading constants $C_{\cos} = \frac{1}{2}\|\pi-\rho\|_2^2$*,* $C_{D_f} = D_f(\pi,\rho)$*,* $C_{\mathrm{Hel}} = \mathrm{Hel}(\pi,\rho)$ *and* $\mathrm{FR}(p_s, q_s) = 4\sin\big(\frac{\mathrm{FR}(\pi,\rho)}{4}\big)\sqrt{1-s} + O\big((1-s)^{3/2}\big)$.

(f) Partitions (multiple sinks). H*,* KL *and* BC *obey exact grouped laws under any partition of positions (arithmetic group masses), while* d_A^2 *splits orthogonally with a between-group term computed from* geometric *group means (Egozcue & Pawlowsky-Glahn, 2005). The displayed identities are proved in the partition subsection below.*

Parts (a) and (f) are classical grouping identities. Parts (b)–(e) specialize standard divergence and spherical identities to sink–content coordinates and make their attenuation constants explicit. Proposition 5 treats unmatched sinks, and Remark 4 distinguishes attenuation of numerical values from changes in rankings. Numerical checks of the identities have maximum absolute residual below 4×10^{-12} (Exp. 1, Appendix I). Corollary 3 states, for the bounded divergences considered here, the matched-sink range collapse analysed in Theorem 3 of Papamichalis & Ruane (2026). Proposition 4 gives the sharp JS dominance criterion.

Corollary 3 (Dynamic-range collapse)**.** *On the matched-sink slice* $\{(p_s, q_s) : \pi, \rho \in \mathcal{S}^{D-1}\}$*, the diameter of any bounded* f*-divergence is exactly* $(1-s)$ *times its nominal diameter:* $\sup_{\pi,\rho} \mathrm{JS}(p_s, q_s) = (1-s)\log 2$*,* $\sup \mathrm{TV} = (1-s)$*,* $\sup \mathrm{Hel}^2 = 2(1-s)$*. At a common matched sink mass* $s = 0.99$*, the attainable divergence values on this slice occupy 1 percent of the divergence's nominal range.*

Proposition 4 (Keep-vs-mask reversal, sharp form)**.** *Consider two comparisons* $r \in \{1,2\}$ *with sink masses* (s_r, t_r)*, mean* $\bar{s}_r$ *and contents* (π_r, ρ_r)*. The kept pipeline reports* $J_r = \mathrm{JS}_b(s_r, t_r) + (1-\bar{s}_r)\mathrm{JS}_{\lambda_r}(\pi_r, \rho_r)$ *with* $\lambda_r = \frac{1-s_r}{2(1-\bar{s}_r)}$ *(Theorem 2(b)). The masked pipeline removes the sink from both rows and renormalizes and reports exactly* $\mathrm{JS}(\pi_r, \rho_r)$*. Write* $B_r = \mathrm{JS}_b(s_r, t_r)$*,* $a_r = 1-\bar{s}_r$ *and* $U_r = a_r\,H_b(\lambda_r)$ *with* H_b *the binary entropy. Then: (i)* $B_r \leq J_r < B_r + U_r \leq B_r + a_r\log 2$*. The minimum* B_r *is attained exactly at* $\pi_r = \rho_r$*, the supremum* $B_r + U_r$ *is approached by nearly disjoint contents but never attained on the open simplex and* $U_r = a_r \log 2$ *exactly when* $s_r = t_r$*. (ii)* $J_1 > J_2$ *for* every *choice of the four contents* if and only if $B_1 \geq B_2 + U_2$*. At equality the dominance is pointwise strict but the infimum gap is* 0*, so the sharp constant is* $H_b(\lambda_2)$*, not* $\log 2$*. (iii) Write* $M_r = \mathrm{JS}(\pi_r, \rho_r)$*. A strict keep-versus-mask reversal occurs exactly when* $(J_1 - J_2)(M_1 - M_2) < 0$*. Cases in which exactly one difference vanishes are tie changes, which we distinguish from strict reversals. Masking changes both terms: it deletes the Bernoulli channel and simultaneously replaces the weighted skew content term* $(1-\bar{s})\mathrm{JS}_\lambda$ *by ordinary* $\mathrm{JS}_{1/2}$ *with unit weight, so the two pipelines report different functionals of* (s, t, π, ρ) *whenever* $s \neq t$ *or sink masses differ across comparisons. Analogous statements hold per decomposition in Theorem 2, with metric-specific masked forms. The argument is not identical across metrics because the masked content terms differ (masked forms per metric, below).*

Proposition 5 (Unmatched-sink regimes)**.** *Let* $u = 1-s$ *and* $v = 1-t$ *both tend to* 0*, with* $\lambda = \frac{u}{u+v}$*. Then*

$$\mathrm{JS}(p, q) = (1-\bar{s})\big[\log 2 - H_b(\lambda) + \mathrm{JS}_\lambda(\pi, \rho)\big] + O\big((1-\bar{s})^2\big),$$

where H_b *is the binary entropy in nats. The Bernoulli channel is exact up to an explicit remainder:* $\mathrm{JS}_b(s, t) = (1-\bar{s})[\log 2 - H_b(\lambda)] + R$ *with* $0 \leq R \leq \frac{(1-\bar{s})^2(2\lambda-1)^2}{2(2\bar{s}-1)}$ *whenever* $\bar{s} > 1/2$*, so the* O*-constant is absolute. If the two sink masses approach* 1 *at comparable rates (*λ *bounded away from* $\{0,1\}$*), both channels are* $O(1-\bar{s})$*, and for fixed distinct contents the content channel is then*

$\Theta(1-\bar{s})$. *The Bernoulli channel is* $\Theta(1-\bar{s})$ *only when* $|\lambda - 1/2|$ *is also bounded away from zero. Its leading coefficient* $\log 2 - H_b(\lambda)$ *vanishes exactly at* $\lambda = 1/2$*, while the content coefficient vanishes exactly at* $\pi = \rho$*. If the rates are mismatched (*$\lambda \to 0$ *or* 1*), then* $\mathrm{JS}_\lambda(\pi, \rho) \le H_b(\lambda) \to 0$ uniformly over all content pairs*:* JS is asymptotically content-insensitive under rate-mismatched sinks*, reporting only* $(1-\bar{s})\log 2$ *at leading order.*

Remark 4 (What the exponents do and do not classify). The exponent γ is a property of the numerical scale: replacing δ by δ^r multiplies γ by r (JS versus $\sqrt{\mathrm{JS}}$, Hel^2 versus Hel), so part (e) classifies the conventional scales as they are used in practice, not equivalence classes of rankings. A continuous strictly increasing reparametrization fixing 0 preserves convergence to zero and exact matched-sink invariance, but need not preserve a power-law asymptotic. The invariant distinction is therefore attenuation versus exact invariance, not the existence or sign of a power-law exponent after arbitrary reparametrization. At matched sink each f-divergence's ranking of content pairs is preserved (a common sink mass multiplies every value by $1-s$) by (d). The exponent matters for absolute-scale uses (thresholds, entropy diagnostics, dynamic range) and for comparisons across unequal sink masses, where the decompositions in (a)–(c), not the exponents, govern.

Content variants For a common value matrix, removing the sink and cross terms from Theorem 1(b) and dividing by $(1-\bar{s})^2$ gives $d_{G,c}(\pi,\rho)^2 = (\pi-\rho)^\top G_{cc}(\pi-\rho)$, where $G_{cc} = (G_{ij})_{i,j=1}^{D-1}$. This is a pseudometric on content compositions, and multiplication by $1-\bar{s}$ restores the compression factor. For distinct heads, let $V_{h,c}$ and $V_{g,c}$ denote the projected values at the content positions, evaluated on the same input, and let $G_{ab,cc} = \mathbb{E}[V_{a,c}V_{b,c}^\top]$. For fixed content signatures π_h, π_g,

$$d_{G,c}(h,g)^2 = \pi_h^\top G_{hh,cc}\pi_h + \pi_g^\top G_{gg,cc}\pi_g - 2\pi_h^\top G_{hg,cc}\pi_g,$$

and the empirical analogue uses paired sample averages of these blocks. It reduces to the common-value form only when the two value matrices agree. The JS content share over the $(1-s, 1-t)$ plane is plotted in Fig. 6 (Appendix I).

E.1 Proofs of the identities

Throughout, $p = (s, (1-s)\pi)$, $q = (t, (1-t)\rho)$ and sums over i run over content coordinates $1, \dots, D-1$.

E.2 Chain rules (a)

$H(p) = -s\log s - \sum_i (1-s)\pi_i \log\big((1-s)\pi_i\big) = -s\log s - (1-s)\log(1-s) + (1-s)H(\pi) = H_b(s) + (1-s)H(\pi)$. For KL,

$$\mathrm{KL}(p\|q) = s\log\frac{s}{t} + \sum_i (1-s)\pi_i \log\frac{(1-s)\pi_i}{(1-t)\rho_i} = s\log\frac{s}{t} + (1-s)\log\frac{1-s}{1-t} + (1-s)\mathrm{KL}(\pi\|\rho).$$

Both are the classical grouping identities of information theory.

E.3 Jensen–Shannon as skew JS (b)

Let $m = \frac{p+q}{2} = (\bar{s}, (1-\bar{s})\mu)$ with $\mu = \lambda\pi + (1-\lambda)\rho$ and $\lambda = \frac{1-s}{2(1-\bar{s})}$, because $(1-\bar{s})\mu = \frac{(1-s)\pi + (1-t)\rho}{2}$. Apply the KL chain rule twice:

$$\mathrm{KL}(p\|m) = \mathrm{KL}_b(s\|\bar{s}) + (1-s)\mathrm{KL}(\pi\|\mu), \qquad \mathrm{KL}(q\|m) = \mathrm{KL}_b(t\|\bar{s}) + (1-t)\mathrm{KL}(\rho\|\mu).$$

Averaging, the Bernoulli parts sum to $\mathrm{JS}_b(s,t)$ and the content parts give $\frac{1-s}{2}\mathrm{KL}(\pi\|\mu) + \frac{1-t}{2}\mathrm{KL}(\rho\|\mu) = (1-\bar{s})\big[\lambda\mathrm{KL}(\pi\|\mu) + (1-\lambda)\mathrm{KL}(\rho\|\mu)\big] = (1-\bar{s})\mathrm{JS}_\lambda(\pi,\rho)$, which is Nielsen's λ-skew JS (Nielsen, 2020) since $\mu = \lambda\pi + (1-\lambda)\rho$. When $s = t$, $\lambda = \frac{1}{2}$ and the law reduces to $\mathrm{JS}(p,q) = (1-s)\mathrm{JS}(\pi,\rho)$, consistent with (d). □

E.4 Reversal (sharp form) and unmatched-sink proofs

Sharp bounds. $\mathrm{JS}_\lambda(\pi,\rho) = \lambda\mathrm{KL}(\pi\|\mu) + (1-\lambda)\mathrm{KL}(\rho\|\mu) \le \lambda\log\frac{1}{\lambda} + (1-\lambda)\log\frac{1}{1-\lambda} = H_b(\lambda)$, since $\mu \ge \lambda\pi$ and $\mu \ge (1-\lambda)\rho$ coordinatewise. The bound is approached by nearly disjoint contents

and never attained on the open simplex, where every coordinate of μ strictly exceeds both lower bounds somewhere. Hence $J_r = B_r + a_r \mathrm{JS}_{\lambda_r}(\pi_r, \rho_r) \in [B_r,\ B_r + U_r)$ with $U_r = a_r H_b(\lambda_r)$, the minimum attained exactly at $\pi_r = \rho_r$ and $H_b(\lambda_r) \le \log 2$ with equality iff $\lambda_r = 1/2$, i.e. $s_r = t_r$. *Dominance iff.* If $B_1 \ge B_2 + U_2$: $J_1 \ge B_1 \ge B_2 + U_2 > J_2$ for every content quadruple. If $B_1 < B_2 + U_2$: take $\pi_1 = \rho_1$ (so $J_1 = B_1$) and contents for the second comparison close enough to the supremum that $J_2 > B_1$. Dominance fails. At equality $B_1 = B_2 + U_2$ the first display still gives $J_1 > J_2$ pointwise, while the second construction drives $J_1 - J_2 \to 0$: strict pointwise dominance with zero infimum gap. *Bernoulli remainder.* With $u = 1 - s$, $v = 1 - t$, $a = \frac{u+v}{2}$, $\lambda = \frac{u}{u+v}$: the content-mass contribution to JS_b is $\frac{1}{2}[u \log \frac{u}{a} + v \log \frac{v}{a}] = a[\lambda \log(2\lambda) + (1-\lambda)\log(2(1-\lambda))] = a[\log 2 - H_b(\lambda)]$ exactly and the complementary half is $R = \frac{1}{2}[f(1-u) + f(1-v)] - f(1-a)$ with $f(x) = x \log x$, so $R \ge 0$ by convexity and a second-order Taylor bound with $f''(x) = 1/x$ on $[1 - a - h, 1 - a + h]$, $h = a|1 - 2\lambda|$, gives $R \le \frac{a^2(2\lambda-1)^2}{2(1-2a)}$ whenever $a < 1/2$. This proves the exact channel form and the absolute $O((1-\bar{s})^2)$ constant in Proposition 5. The uniformity over contents in the mismatched regime is the bound $\mathrm{JS}_\lambda \le H_b(\lambda)$ again. On the closed simplex the supremum $B_r + U_r$ is attained at disjointly supported contents, so at equality $B_1 = B_2 + U_2$ dominance becomes non-strict ($J_1 = J_2$ is attainable). Uniform strict dominance there requires $B_1 > B_2 + U_2$. Under comparable rates the content channel is also bounded below, since for $\pi \ne \rho$ the map $\lambda \mapsto \mathrm{JS}_\lambda(\pi, \rho)$ is continuous and positive on $[\epsilon, 1-\epsilon]$, and its vanishing as $\lambda \to \{0, 1\}$ is the bound $\mathrm{JS}_\lambda \le H_b(\lambda) \to 0$ again, which completes the proof of Proposition 5.

E.5 Bhattacharyya and the spherical law of cosines (c)

Directly, $\mathrm{BC}(p, q) = \sqrt{st} + \sum_i \sqrt{(1-s)\pi_i(1-t)\rho_i} = \sqrt{st} + \sqrt{(1-s)(1-t)}\,\mathrm{BC}(\pi, \rho)$. The Hellinger form follows by expanding $\mathrm{Hel}^2 = 2 - 2\mathrm{BC}$: $\mathrm{Hel}^2(p, q) = 2 - 2\sqrt{st} - 2\sqrt{(1-s)(1-t)}\,\mathrm{BC}(\pi, \rho) = \mathrm{Hel}_b^2(s, t) + \sqrt{(1-s)(1-t)}\,\mathrm{Hel}^2(\pi, \rho)$, since $\mathrm{Hel}_b^2(s, t) = 2 - 2\sqrt{st} - 2\sqrt{(1-s)(1-t)}$. Substituting $\sqrt{s} = \cos\psi_p$, $\sqrt{t} = \cos\psi_q$, $\mathrm{BC}(\pi, \rho) = \cos\frac{\mathrm{FR}(\pi,\rho)}{2}$ and $\mathrm{BC}(p, q) = \cos\frac{\mathrm{FR}(p,q)}{2}$ yields the displayed law of cosines. Geometrically, $\sqrt{p}$ lies on the unit sphere. The sink mass fixes the colatitude from the pole e_0 and the content angle is the azimuthal separation, so the identity is the spherical law of cosines on the Hellinger sphere (Amari & Nagaoka, 2000). □

E.6 Matched-sink laws (d)

For any f-divergence with $f(1) = 0$: the sink coordinate contributes $t\, f(s/t) = s f(1) = 0$ when $s = t$ and each content coordinate contributes $(1-s)\rho_i f\big(\frac{(1-s)\pi_i}{(1-s)\rho_i}\big) = (1-s)\rho_i f(\pi_i/\rho_i)$. Summing gives $D_f(p_s, q_s) = (1-s) D_f(\pi, \rho)$. The Hellinger form follows with $f(u) = (\sqrt{u} - 1)^2$ and a square root. For Rényi of order $\alpha \in (0, 1) \cup (1, \infty)$ (with $\alpha = 1$ defined by continuity as KL), $\sum_{i=0}^{D-1} (p_s)_i^\alpha (q_s)_i^{1-\alpha} = s + (1-s)\sum_{i=1}^{D-1} \pi_i^\alpha \rho_i^{1-\alpha}$, giving the log-mix law $R_\alpha(p_s, q_s) = \frac{1}{\alpha-1}\log\big(s + (1-s)e^{(\alpha-1)R_\alpha(\pi,\rho)}\big)$. Degenerate affine f with $D_f \equiv 0$ are excluded from (e) by the hypothesis $0 < D_f(\pi, \rho) < \infty$. For d_A, set $c = \mathcal{C}(\frac{s}{s'}, \frac{1-s}{1-s'}, \dots, \frac{1-s}{1-s'})$. Then $p_s = c \oplus p_{s'}$ and $q_s = c \oplus q_{s'}$, so $\oplus$-invariance gives invariance in s. For d_H: $\log p_s - \log q_s = (0, \log\pi_1 - \log\rho_1, \dots)$ and since π and ρ both sum to one and $\pi \ne \rho$, $\max_i \log(\pi_i/\rho_i) > 0 > \min_i \log(\pi_i/\rho_i)$ (Lemma 1), so the oscillation over all D coordinates equals the oscillation over the content coordinates, which is $d_H(\pi, \rho)$. In particular including the 0 entry changes nothing. □

Lemma 1. *If $\pi \ne \rho \in \mathcal{S}^{D-1}$ then $\max_i \log\frac{\pi_i}{\rho_i} > 0 > \min_i \log\frac{\pi_i}{\rho_i}$.*

Proof. If $\log(\pi_i/\rho_i) \le 0$ for all i, then $\pi_i \le \rho_i$ with strict inequality somewhere, contradicting $\sum \pi_i = \sum \rho_i = 1$. Symmetrically for the minimum. □

E.7 Attenuation spectrum (e)

f-divergences and TV are exactly linear in $(1-s)$ by (d), so $\gamma = 1$ with $C = D_f(\pi, \rho)$. Euclidean: $\|p_s - q_s\|_2 = (1-s)\|\pi - \rho\|_2$ exactly. Hellinger: exactly $\sqrt{1-s}\,\mathrm{Hel}(\pi, \rho)$ by (d). $\sqrt{\mathrm{JS}}$: square root of an exactly linear law. Fisher–Rao: by (c) at $s = t$, $\cos\frac{\mathrm{FR}}{2} = s + (1-s)\cos\gamma_c$ with $\gamma_c = \frac{\mathrm{FR}(\pi,\rho)}{2}$.

Writing $1-\cos\frac{\mathrm{FR}}{2}=(1-s)(1-\cos\gamma_c)$ and $1-\cos x=2\sin^2\frac{x}{2}$ gives $\sin\frac{\mathrm{FR}}{4}=\sqrt{1-s}\,\sin\frac{\gamma_c}{2}$ *exactly*, whence $\mathrm{FR}=4\arcsin\big(\sqrt{1-s}\sin\frac{\gamma_c}{2}\big)=4\sin\frac{\gamma_c}{2}\sqrt{1-s}+O((1-s)^{3/2})$. Cosine: with $r=(1-s)/s$, $\langle p_s,q_s\rangle=s^2\big(1+r^2\langle\pi,\rho\rangle\big)$ and $\|p_s\|_2^2=s^2\big(1+r^2\|\pi\|_2^2\big)$. Expanding $(1+x)^{-1/2}$ gives

$$1-\cos(p_s,q_s)=\frac{r^2}{2}\,\|\pi-\rho\|_2^2+O(r^4).$$

The leading constant as $s\to 1$ is $\frac{1}{2}\|\pi-\rho\|_2^2$. The factor s^{-2} in r^2 is the first-order finite-s correction. It raises the local log–log slope of the leading term to $2(1+(1-s)/s)$, consistent with the fitted slope 2.02 over $s\in[0.9,1-10^{-4}]$ in Figure 7 (Appendix I). d_A and d_H: exactly constant by (d). □

E.8 Partitions (f)

Let $\mathcal{B}$ be a partition of $\{0,\dots,D-1\}$ into nonempty blocks. For each $B\in\mathcal{B}$, write

$$a_B=\sum_{i\in B}p_i,\qquad b_B=\sum_{i\in B}q_i,\qquad p^B=(p_i/a_B)_{i\in B},\qquad q^B=(q_i/b_B)_{i\in B}.$$

The vectors $a=(a_B)_{B\in\mathcal{B}}$ and $b=(b_B)_{B\in\mathcal{B}}$ contain the arithmetic group masses. Expanding the definitions gives

$$H(p)=H(a)+\sum_{B\in\mathcal{B}}a_B H(p^B),$$
$$\mathrm{KL}(p\|q)=\mathrm{KL}(a\|b)+\sum_{B\in\mathcal{B}}a_B\mathrm{KL}(p^B\|q^B),$$
$$\mathrm{BC}(p,q)=\sum_{B\in\mathcal{B}}\sqrt{a_Bb_B}\,\mathrm{BC}(p^B,q^B).$$

For Aitchison distance, let

$$\ell_i=\log(p_i/q_i),\qquad n_B=|B|,$$
$$w_B=\frac{1}{n_B}\sum_{i\in B}\ell_i=\log\mathrm{gm}(p_B)-\log\mathrm{gm}(q_B),\qquad \bar{w}=\frac{1}{D}\sum_{B\in\mathcal{B}}n_Bw_B.$$

Here $p_B=(p_i)_{i\in B}$ and $q_B=(q_i)_{i\in B}$ are the unnormalized blocks. Since $\sum_{i\in B}(\ell_i-w_B)=0$, expanding $\ell_i-\bar{w}=(\ell_i-w_B)+(w_B-\bar{w})$ yields

$$d_A(p,q)^2=\sum_{B\in\mathcal{B}}d_A(p^B,q^B)^2+\sum_{B\in\mathcal{B}}n_B(w_B-\bar{w})^2.$$

Singleton blocks contribute zero within-block distance. This is the orthogonal within/between decomposition expressed by ILR balances (Egozcue et al., 2003; Egozcue & Pawlowsky-Glahn, 2005). □

E.9 Corollary 3

Corollary 3 is (d) plus the classical suprema $\sup\mathrm{JS}=\log 2$, $\sup\mathrm{TV}=1$, $\sup\mathrm{Hel}^2=2$, attained in the limit of disjointly supported contents (approached within the open simplex). □

Masked forms per metric (Prop. 4) Masking removes coordinate 0 and renormalizes both rows, so the masked pipeline evaluates the base dissimilarity on (π,ρ). The kept and masked content terms differ per metric: for JS, kept $(1-\bar{s})\mathrm{JS}_\lambda(\pi,\rho)$ versus masked $\mathrm{JS}_{1/2}(\pi,\rho)$. For f-divergences at matched sink, kept $(1-s)D_f(\pi,\rho)$ versus masked $D_f(\pi,\rho)$ (a pure rescaling). For FR, kept $2\arccos\big(\sqrt{st}+\sqrt{(1-s)(1-t)}\cos\frac{\mathrm{FR}(\pi,\rho)}{2}\big)$ versus masked $\mathrm{FR}(\pi,\rho)$ (nonlinear). For entropy diagnostics, kept $H_b(s)+(1-s)H(\pi)$ versus masked $H(\pi)$. For d_A, masking retains $d_A^{\perp}=d_A(\pi,\rho)$ and removes the squared balance term. For d_H, the equality $d_H(p_s,q_s)=d_H(\pi,\rho)$ holds at matched sink masses but need not hold when $s\neq t$. The sharp dominance criterion (ii) uses the exact range $[0,(1-\bar{s})H_b(\lambda))$ of the JS content term. The same range argument applies to another decomposition only after establishing its exact bounds and whether the endpoints are attained. The weaker bound $(1-\bar{s})\log 2$ does not yield the sharp "if and only if" criterion. The flip characterization (iii) is metric-specific through the masked form.

F Proof of Theorem 3 (separability characterization)

Step 0: notation The centered log-ratio clr maps $\mathcal{S}^D$ isometrically onto $\mathcal{H}_D = \{x \in \mathbb{R}^D : \sum_i x_i = 0\}$ with the Aitchison distance on the left and the Euclidean distance on the right (Aitchison, 1982). Let $u_j = \sqrt{D/(D-1)}\,(e_j - \frac{1}{D}\mathbf{1}) \in \mathcal{H}_D$, a unit vector. A direct computation using $\sum_i (\operatorname{clr} p)_i = 0$ gives

$$\langle \operatorname{clr} p, u_j \rangle = \sqrt{\tfrac{D}{D-1}}\,(\operatorname{clr} p)_j = \sqrt{\tfrac{D-1}{D}}\,\log \frac{p_j}{\operatorname{gm}(p_{-j})} = b_j(p),$$

so under our normalization the balance is *exactly* the u_j-coordinate of $\operatorname{clr} p$. We use the following lemma.

Lemma 2. *Let $p, q \in \mathcal{S}^D$, set $x = \operatorname{clr} p - \operatorname{clr} q$ and let $\Pi_j = I - u_j u_j^\top$ on $\mathcal{H}_D$. The zero-padded CLR difference of $p^{(-j)}$ and $q^{(-j)}$ equals $\Pi_j x$. In particular, if $\langle x, u_j \rangle = 0$, then $x_j = 0$ and the subcomposition CLR difference is obtained by deleting that zero coordinate from x.*

Proof. Write $L = \log p - \log q$, so $x = L - \bar{L}\mathbf{1}$ with $\bar{L}$ the mean of L. For $i \neq j$, the subcomposition CLR difference is $L_i - \frac{1}{D-1}\sum_{k\neq j} L_k = x_i - \frac{1}{D-1}\sum_{k\neq j} x_k = x_i + \frac{x_j}{D-1}$, because $\sum_k x_k = 0$. Since $\langle x, u_j\rangle u_j = \frac{D}{D-1} x_j \big(e_j - \frac{1}{D}\mathbf{1}\big)$, the i-th coordinate of $\Pi_j x$ is $x_i + \frac{x_j}{D-1}$ for $i \neq j$, and its j-th coordinate is $x_j - x_j = 0$. Hence the zero-padded subcomposition CLR difference equals $\Pi_j x$. Finally, $\langle x, u_j\rangle = \sqrt{D/(D-1)}\,x_j$, so $\langle x, u_j\rangle = 0$ implies $x_j = 0$ and $\Pi_j x = x$. □

Lemma 3. *$\{u_j\}_{j=0}^{D-1}$ spans $\mathcal{H}_D$ and $\langle u_j, u_k\rangle = -\frac{1}{D-1}$ for $j \neq k$.*

Proof. The vectors $e_j - \frac{1}{D}\mathbf{1}$ are the columns of the projection $I - \frac{1}{D}\mathbf{1}\mathbf{1}^\top$ onto $\mathcal{H}_D$, hence span it. The inner product is a direct computation as in the main text. □

Step 1: reduction to a norm-like function on $\mathcal{H}_D$ By (S1) with $c = (\ominus q)$ (the compositional inverse), $\delta(p, q) = \delta(p \ominus q,\, e)$ where e is the barycenter, so $\delta(p, q) = N(\operatorname{clr} p - \operatorname{clr} q)$ for some $N : \mathcal{H}_D \to [0, \infty)$. (S3) makes N even ($N(-x) = N(x)$, from symmetry) and definite ($N(x) = 0 \iff x = 0$). (S2) gives $N(\alpha x) = |\alpha| N(x)$ for all $\alpha \in \mathbb{R}$ (powering acts as scalar multiplication in CLR). Set $Q = N^2$. Then Q is even, definite and 2-homogeneous.

Step 2: the decomposition in linear coordinates Fix j. Every perturbation $c \in \mathcal{S}^D$ acts on both arguments simultaneously, leaving $x = \operatorname{clr} p - \operatorname{clr} q$ unchanged while shifting the *channel pair* of the individual points: $b_j(p)$ and $b_j(q)$ both shift by $b_j(c)$ and the subcompositions get perturbed by $c^{(-j)}$. Applying (S4) before and after such a shift and using that all combinations $(h, c') = (b_j(c), c^{(-j)})$ are realizable (choose the j-th part of c freely given its content), we conclude that φ_j is invariant under simultaneous translation of its two arguments and Ψ_j under simultaneous perturbation of its two arguments. By the same transitivity as Step 1 (now on the sub-simplex),

$$Q(x) = f_j\big(\langle x, u_j\rangle\big) + G_j\big(\Pi_j x\big) \qquad \text{for all } x \in \mathcal{H}_D,$$

where we used Lemma 2 to identify the sub-CLR difference with $\Pi_j x$ and Step 0 to identify the balance difference with $\langle x, u_j\rangle$. Evaluating at $x = t u_j$ and at $x = \Pi_j x$ and at 0 and replacing (f_j, G_j) by $(f_j - f_j(0),\, G_j - G_j(0))$ (which preserves the decomposition and, by $Q(0) = 0$, sets both to vanish at 0), we may assume $f_j(0) = G_j(0) = 0$. Then $f_j(t) = Q(t u_j)$ and $G_j(w) = Q(w)$ for $w \perp u_j$.

Step 3: the quadratic form on balance lines By 2-homogeneity, $f_j(\alpha t) = Q(\alpha t u_j) = \alpha^2 Q(t u_j) = \alpha^2 f_j(t)$ for all α, hence $f_j(t) = c_j\, t^2$ with $c_j = Q(u_j) \geq 0$ (f_j is even since Q is).

Step 3′: the coefficients agree without permutation invariance Fix $j \neq k$ and set $\mu = \langle u_j, u_k\rangle = -\frac{1}{D-1} \neq 0$. Expand $Q(a\,u_j + b\,u_k)$ along coordinate j: the j-channel is $a + b\mu$ and $\Pi_j(a\,u_j + b\,u_k) = b(u_k - \mu u_j)$, so by Step 2 and 2-homogeneity

$$Q(a\,u_j + b\,u_k) = c_j(a + b\mu)^2 + b^2\, Q(u_k - \mu u_j).$$

Expanding along coordinate k instead gives $c_k(b+a\mu)^2 + a^2\,Q(u_j - \mu u_k)$. Both are polynomial identities in $(a,b) \in \mathbb{R}^2$. Matching the ab coefficient yields $2c_j\mu = 2c_k\mu$, hence $c_j = c_k =: c$. (Matching a^2 and b^2 then gives $Q(u_k - \mu u_j) = Q(u_j - \mu u_k) = c(1-\mu^2)$, consistent with the conclusion.) Permutation invariance of δ is thereby a consequence of the axioms, not a hypothesis.

Step 4: telescoping For arbitrary $x \in \mathcal{H}_D$, $t \in \mathbb{R}$, decompose $x = \langle x, u_j\rangle u_j + \Pi_j x$. Then $x + tu_j$ has j-channel $\langle x, u_j\rangle + t$ and unchanged Π_j-part, so

$$Q(x + t\,u_j) - Q(x) = c\big(\langle x, u_j\rangle + t\big)^2 - c\langle x, u_j\rangle^2 = c\,t^2 + 2c\,t\,\langle x, u_j\rangle. \tag{3}$$

This is an exact identity valid for every x. No continuity is invoked. Now take any $y \in \mathcal{H}_D$ and write $y = \sum_{m=1}^{M} t_m u_{j_m}$ (possible by Lemma 3). Let $y_{<m} = \sum_{l<m} t_l u_{j_l}$ and telescope (3) along the partial sums starting from $x = 0$:

$$\begin{aligned} Q(y) &= \sum_m \Big(c\,t_m^2 + 2c\,t_m \langle y_{<m}, u_{j_m}\rangle \Big) \\ &= c\sum_m t_m^2 + 2c \sum_{l<m} t_l t_m \langle u_{j_l}, u_{j_m}\rangle \\ &= c\Big\| \sum_m t_m u_{j_m} \Big\|^2 = c\|y\|^2. \end{aligned}$$

Thus $N = \sqrt{c}\,\|\cdot\|_2$ on $\mathcal{H}_D$, i.e. $\delta = \sqrt{c}\,d_A$. Definiteness forces $c > 0$. The converse direction is the Pythagorean identity of Papamichalis & Ruane (2026) at each coordinate. □

The d_θ family and all-coordinate separability (Prop. 6) $d_\theta^2(p,q) = \theta\langle x, u_0\rangle^2 + \|\Pi_0 x\|^2 = \|x\|^2 + (\theta-1)\langle x, u_0\rangle^2$ is (S1)–(S3) and sink-0 separable for every $\theta > 0$. At a coordinate $j \neq 0$, write $u_0 = \alpha u_j + w$ with $w = \Pi_j u_0 \neq 0$ and $\alpha = \langle u_0, u_j\rangle = -\frac{1}{D-1}$: then $\langle x, u_0\rangle^2 = \alpha^2\langle x, u_j\rangle^2 + 2\alpha\langle x, u_j\rangle\langle \Pi_j x, w\rangle + \langle \Pi_j x, w\rangle^2$ contains the product term $\langle x, u_j\rangle \cdot \langle \Pi_j x, w\rangle$, which is not of the form $f(\langle x, u_j\rangle) + G(\Pi_j x)$ unless $\theta = 1$ (test the rectangle $x = t_1 u_j + t_2 w/\|w\|$ with $w = \Pi_j u_0$: its mixed second difference equals $2(\theta-1)\alpha\|w\|\,\Delta_1\Delta_2$ with $\alpha = \langle u_j, u_0\rangle = -\frac{1}{D-1}$ and $\|w\| = \sqrt{1-\alpha^2}$, nonzero for $\theta \neq 1$ and increments $\Delta_1, \Delta_2 \neq 0$). Hence all-coordinate separability forces $\theta = 1$. Note d_θ ($\theta \neq 1$) is also not permutation-invariant, consistent with Step 3′.

Transport under permutation invariance If δ additionally satisfies permutation invariance and is separable at one coordinate j_0, it is separable at every j: for the transposition $\sigma = (j_0\, j)$, $b_{j_0}(\sigma p) = b_j(p)$ and $(\sigma p)^{(-j_0)}$ is a fixed permutation of $p^{(-j)}$, so composing the coordinate-j_0 decomposition with σ yields a coordinate-j decomposition. Theorem 3 then applies. Thus either all-coordinate separability or permutation invariance with one-coordinate separability suffices to obtain βd_A.

Proposition 6 (Designated-sink characterization)**.** *Under (S1)–(S3) with sink-separability imposed only at coordinate* 0*, the solutions are exactly*

$$\delta(p,q)^2 = c\,\big(b_0(p) - b_0(q)\big)^2 + N_\perp\big(\operatorname{clr}\pi - \operatorname{clr}\rho\big)^2,$$

where, for each δ, the coefficient $c > 0$ and the function $N_\perp$ are uniquely determined. Across admissible dissimilarities, $N_\perp$ may be any even, definite, absolutely homogeneous function on the content CLR space, equivalently any content dissimilarity satisfying (S1)–(S3). The family $d_\theta^2 = \theta(\Delta b)^2 + (d_A^\perp)^2$, $\theta > 0$ (the d_λ of Papamichalis & Ruane (2026)), is the subfamily with $N_\perp$ Euclidean. It violates Sep-A at every non-sink coordinate for $\theta \neq 1$ (below). Two independent routes collapse the residual freedom: requiring Sep-A at every coordinate (Theorem 3), or adding permutation invariance, under which separability at a single coordinate transports to all coordinates and Theorem 3 applies.

Proof of Proposition 6 (designated sink) Sink-separability at coordinate 0 gives, by Steps 1–2, $Q(x) = f_0(\langle x, u_0\rangle) + Q(\Pi_0 x)$ with $f_0(t) = c\,t^2$ by Step 3. The restriction of Q to $u_0^\perp \cap \mathcal{H}_D$ is even, definite and 2-homogeneous (inherited) and $u_0^\perp \cap \mathcal{H}_D$ is isometric to the content CLR space by Lemma 2, so $Q|_{u_0^\perp} = N_\perp^2$ for an arbitrary function of that type. Conversely every such pair $(c, N_\perp)$ defines a dissimilarity satisfying (S1)–(S3) and sink-0 separability. Definiteness of δ forces $c > 0$ and $N_\perp$ definite. □

Remark 5 (Sep-A versus Sep-P). Call δ Sep-P if $\delta(p,q)^2 = \varphi(s,t) + \Psi(\pi,\rho)$ in probability coordinates. Exact separation is coordinate-system relative. In probability coordinates, d_A is *not* separable: writing $\Lambda = \operatorname{logit} s - \operatorname{logit} t$ and $\Lambda_c = \log \frac{\mathrm{gm}(\pi)}{\mathrm{gm}(\rho)}$,

$$d_A^2 = \tfrac{D-1}{D}\,(\Lambda - \Lambda_c)^2 + (d_A^{\perp})^2,$$

which follows from $\Delta b = \sqrt{(D-1)/D}\,(\Lambda - \Lambda_c)$ and the Pythagorean identity, so its Sep-P mixed second difference over a rectangle $\{s_0, s_1\} \times \{\pi_0, \pi_1\}$ equals $-2\frac{D-1}{D}\,\Delta\Lambda\,\Delta\Lambda_c$ exactly, confirmed to 4.3×10^{-14} in Exp. 4 (Appendix I). Conversely, the scalar metric

$$d_P(p,q) = \sqrt{(\operatorname{logit} s - \operatorname{logit} t)^2 + d_A(\pi,\rho)^2}$$

is Sep-P by construction. One may instead report its two components separately. None of the classical dissimilarities examined in Table 16 is Sep-P. A vector report in probability coordinates should be distinguished from a scalar dissimilarity that is Sep-P in the sense of Theorem 3.

Proposition 7 (Sep-P is impossible for attenuating dissimilarities)**.** *Suppose there are two content pairs whose squared dissimilarities differ at some common sink level and the difference of these squared dissimilarities tends to zero as $s \to 1$. Then δ is not Sep-P. This condition holds for the attenuating families in Theorem 2(e) whenever the selected pairs have different dissimilarities at some common sink level.*

Proof. Sep-P gives

$$\delta^2(p_s, q_s) = \varphi(s,s) + \Psi(\pi,\rho).$$

The difference across the two content pairs is therefore independent of s. By hypothesis it is nonzero at some s but tends to zero as $s \to 1$, a contradiction. □

Strength of the hypotheses The axioms are sufficient, not claimed minimal. Under (S1), condition (S2) at $\alpha = -1$ already implies symmetry: $\delta(p,q) = \delta(e, (\ominus p) \oplus q)$ by (S1) with $c = \ominus p$ (writing $\ominus p = (-1) \odot p$) and $\ominus e = e$ gives $\delta(e,z) = \delta(e, \ominus z)$, whence $\delta(p,q) = \delta(q,p)$. The forward implication uses separability only for pairs of the form (p,e) against the barycenter. Full pairwise separability is retained because it is the operational property of interest. No continuity, measurability, convexity, or triangle inequality is assumed anywhere in the proof.

G Incompatibility theorem and proofs

Table 13 lists the properties of the dissimilarities compared in this appendix and in Theorem 3. Section 5 refers to the following theorem through Table 13. Theorem 3 identifies the unique separable dissimilarity in the stated class. Within that class, separability is incompatible with boundary continuity and with Markov nonexpansiveness.

Theorem 4 (Incompatibility under compositional axioms)**.** (i) Compositional ⇒ boundary-singular. *If δ is $\oplus$-invariant, vanishes on the diagonal and is not identically zero, it admits no finite jointly continuous extension to the closed simplex: continuity forces the extension to vanish at (e_0, e_0) through the diagonal, while perturbations $c_r = \mathcal{C}(r, 1, \ldots, 1)$ carry any fixed pair to (e_0, e_0) at constant δ (writing $p_s^{\pi} = (s, (1-s)\pi)$). Consequently every nontrivial boundary-continuous dissimilarity that vanishes on the diagonal fails $\oplus$-invariance. Boundary continuity by itself forces only that matched-sink content sensitivity vanish as $s \to 1$,* uniformly over contents *(since $\|p_s^{\pi} - e_0\|_2 \le \sqrt{2}\,(1-s)$) and Theorem 2(e) supplies the polynomial rates $(1-s)^{\gamma}$, $\gamma > 0$, for the enumerated families.*

(ii) Sep-A and Markov nonexpansiveness under compositional axioms. *(a) If $\delta(p,p) = 0$, sink-separability at coordinate 0 forces every matched-sink curve $s \mapsto \delta^2(p_s, q_s)$ with permuted contents $\rho = \sigma\pi$ to be constant. Indeed, $\mathrm{gm}(\sigma\pi) = \mathrm{gm}(\pi)$ makes the balance terms for $(\pi, \sigma\pi)$ and (π,π) identical and the latter pair has zero dissimilarity (proof below). No invariance axiom is used. Consequently, sink-separability is impossible whenever there exist π, a permutation σ and $s_0 \in (0,1)$ with $\delta\big((s_0, (1-s_0)\pi), (s_0, (1-s_0)\sigma\pi)\big) > 0$ while this dissimilarity tends to zero as $s \to 1$. The condition holds for every f-divergence with a finite convex non-affine generator, every positive power of one, and the Fisher–Rao, Euclidean and cosine dissimilarities. For example, $(1-s)\mathrm{JS}(\pi, \sigma\pi)$ is*

nonconstant whenever $\sigma\pi \neq \pi$. None of these dissimilarities is therefore separable even at the single sink coordinate, and none is Sep-A. Separately, (S1) alone implies full matched-sink invariance for arbitrary contents: matched-sink pairs at different s are perturbation-equivalent. (b) In each fixed dimension, (S1)–(S4) force $\delta_D = \beta_D d_A$ with a possibly dimension-dependent scale (Theorem 3). Suppose the family $\{\delta_D\}_{D\geq 3}$ satisfies the axioms in every dimension and *is nonexpansive under every Markov morphism. Let $k \geq 2$, let $E = E_{D,k}$ split coordinate j into k equal parts and let C be the deterministic merge that sums them back. Both are Markov morphisms and $CE = \mathrm{Id}$, so nonexpansiveness in both directions forces $\delta_{D+k-1}(Ep, Eq) = \delta_D(p, q)$ for* every *pair. A pair whose CLR difference is supported on two unsplit coordinates has d_A ratio 1 under E, forcing $\beta_{D+k-1} = \beta_D$. The balance witness $\operatorname{clr} p - \operatorname{clr} q \propto e_j - \frac{1}{D}\mathbf{1}$ then satisfies*

$$\frac{d_A(E_{D,k}p,\, E_{D,k}q)}{d_A(p,q)} = \sqrt{\frac{kD}{D+k-1}} > 1$$

exactly. This witness value is the supremum of the ratio over all pairs, because the squared *ratio $kD/(D+k-1)$, not the ratio itself, is the top generalized eigenvalue of a two-block quadratic-form pencil (computed analytically below) and the distance ratio tends to $\sqrt{D}$ as $k \to \infty$: contradiction. The violation is not confined to dimension-changing morphisms. The strictly positive column-stochastic map and interior pair*

$$T = \begin{pmatrix} .01 & .98 & .98 \\ .61 & .01 & .01 \\ .38 & .01 & .01 \end{pmatrix}, \quad p = \begin{pmatrix} .68 \\ .31 \\ .01 \end{pmatrix},\ q = \begin{pmatrix} .26 \\ .73 \\ .01 \end{pmatrix},$$

give $d_A(p, q) = 1.28615$ and $d_A(Tp, Tq) = 1.41503$: a same-dimension expansion by $\times 1.1002$ (further examples below and in Exp. 3 of Appendix I). Every f-divergence and FR *are exactly invariant under $E_{D,k}$ (proved below) and monotone under all Markov morphisms (Csiszár, 1967; Cencov, 2000). Thus no family satisfying (S1)–(S4) in every dimension is nonexpansive under all Markov morphisms. The displayed map also shows that βd_A is not nonexpansive under all three-dimensional Markov maps.*

(iii) Properties of Hilbert's metric. *d_H is $\oplus$-invariant and $\odot$-homogeneous, is permutation-invariant, is exactly invariant under every congruent embedding and is nonexpansive under every stochastic map with no zero row, with uniform contraction under strictly positive maps (Birkhoff, 1957; Bushell, 1973). Yet it is not Sep-A (explicit four-point witness with defect -1, proof below) and not induced by an inner product (it fails the parallelogram law). The converse characterization of Kohlberg & Pratt (1982) requires its additional regularity and projectivity hypotheses, as discussed in Remark 6.*

G.1 (i) Boundary singularity

Suppose δ is $\oplus$-invariant and $\delta(p^\star, q^\star) = \varepsilon > 0$ for some interior pair. Let $c_r = \mathcal{C}(r, 1, \ldots, 1)$ and note $c_r \oplus p \to e_0$ and $c_r \oplus q \to e_0$ (componentwise) as $r \to \infty$ for *every* interior p, q. If δ extended continuously to the closed simplex with $\delta(x, x) = 0$, then $\delta(c_r \oplus p^\star, c_r \oplus q^\star) \to \delta(e_0, e_0) = 0$, contradicting $\oplus$-invariance which fixes the value at ε. Hence no such extension exists. Contrapositively, any nontrivial boundary-continuous δ vanishing on the diagonal fails $\oplus$-invariance and Theorem 2(d,e) gives matched-sink decay rates for the specific dissimilarities considered there. □

KL and χ^2 are not boundary-continuous. Part (i) does not apply to them because they are not $\oplus$-invariant, so we give a direct witness. For $0 < \varepsilon < 1/2$, let $p_\varepsilon = (1 - 2\varepsilon, \varepsilon, \varepsilon)$ and $q_\varepsilon = (1 - \varepsilon - \varepsilon e^{-1/\varepsilon}, \varepsilon e^{-1/\varepsilon}, \varepsilon)$. Both tend to e_0. The second coordinate contributes $\varepsilon \log e^{1/\varepsilon} = 1$ to $\mathrm{KL}(p_\varepsilon \| q_\varepsilon)$ and the first contribution tends to zero, so $\mathrm{KL}(p_\varepsilon \| q_\varepsilon) \to 1$. The second coordinate contributes $\varepsilon(1 - e^{-1/\varepsilon})^2 e^{1/\varepsilon} \to \infty$ to $\chi^2(p_\varepsilon \| q_\varepsilon)$. Since both divergences vanish on the diagonal, neither admits a finite jointly continuous extension at (e_0, e_0).

G.2 (ii)(a) Sink-separability excludes attenuating families

Part without invariance axioms. Assume $\delta(p, p) = 0$ and sink-separability at $j = 0$: $\delta(p, q)^2 = \varphi_0\big(b_0(p), b_0(q)\big) + \Psi_0(\pi, \rho)$. With contents fixed and matched sink mass s varying,

$$\big(b_0(p_s),\, b_0(q_s)\big) = \sqrt{\tfrac{D-1}{D}}\big(\operatorname{logit} s - \log \mathrm{gm}(\pi),\ \operatorname{logit} s - \log \mathrm{gm}(\rho)\big),$$

so the φ_0 term depends on the contents only through their geometric means. Compare two content pairs with identical geometric means: $(\pi, \rho) = (\pi, \sigma\pi)$ for a permutation σ (so $\mathrm{gm}(\rho) = \mathrm{gm}(\pi)$) and $(\pi', \rho') = (\pi, \pi)$. The φ_0 terms then agree for *every* s, so

$$\delta^2(p_s, q_s) - \delta^2(p'_s, q'_s) = \Psi_0(\pi, \sigma\pi) - \Psi_0(\pi, \pi)$$

is constant in s. Since $p'_s = q'_s$, the subtrahend is identically zero, hence $s \mapsto \delta^2\big((s, (1-s)\pi), (s, (1-s)\sigma\pi)\big)$ must be constant. For any family admitting a permuted-content pair whose dissimilarity is positive at some $s_0 \in (0, 1)$ and tends to zero as $s \to 1$, this curve cannot be constant, a contradiction. Such a pair exists for every finite convex non-affine generator f with $f(1) = 0$: choose $r > 1$ such that f is not affine on $[r^{-1}, r]$ and choose two content coordinates in the ratio r, with σ swapping them. Strict Jensen inequality gives

$$\frac{f(r) + rf(r^{-1})}{1 + r} > f(1) = 0.$$

For $D > 3$, assign equal positive values to the remaining coordinates in both compositions and scale the selected two coordinates by their remaining total mass. The divergence remains positive. This proves the exclusion for these f-divergences and every positive power of them. For FR, Euclidean and cosine, any pair with $\sigma\pi \neq \pi$ has positive dissimilarity.

Matched-sink invariance under (S1). For any $s, s' \in (0, 1)$, set

$$c = \mathcal{C}\left(\frac{s}{s'}, \frac{1-s}{1-s'}, \ldots, \frac{1-s}{1-s'}\right).$$

Then $p_s = c \oplus p_{s'}$ and $q_s = c \oplus q_{s'}$, so (S1) gives $\delta(p_s, q_s) = \delta(p_{s'}, q_{s'})$ for every content pair. This argument requires no separability assumption. □

Remark 6 (Scope of the Kohlberg–Pratt comparison). The Kohlberg–Pratt theorem, under its own regularity, projectivity and strongly-positive-map hypotheses, identifies a contraction-compatible projective metric only as $\mathfrak{d} = f \circ d_H$ with f continuous and strictly increasing. Contraction alone does not force $\mathfrak{d} = c\, d_H$. Adding powering homogeneity gives $f(\alpha t) = \alpha f(t)$ for $\alpha, t \geq 0$ and since every $t \geq 0$ is realized as a Hilbert distance, $f(t) = ct$. Scalar uniqueness therefore needs both the Kohlberg–Pratt hypotheses and $\odot$-homogeneity. The statement of Theorem 4(iii) is scoped accordingly.

G.3 (II)(B) The exact split expansion and extremality

Let $E = E_{D,k}$ split coordinate $j = D-1$ (wlog) into k equal parts. On CLR differences, if $x = \mathrm{clr}\, p - \mathrm{clr}\, q$ then $\mathrm{clr}(Ep) - \mathrm{clr}(Eq)$ equals the vector x^E with $x^E_i = x_i - \bar{\Delta}$ on the untouched coordinates and k copies of $x_{D-1} - \bar{\Delta}$ on the split block, where the recentring constant $\bar{\Delta}$ is irrelevant to differences of the quadratic form below. Concretely, let $L = \log p - \log q$. The following two quadratic forms are unchanged by replacing L with $L + c\mathbf{1}$:

$$d_A(p, q)^2 = \|L\|^2 - \tfrac{(\mathbf{1}^\top L)^2}{D}, \qquad d_A(Ep, Eq)^2 = \sum_{i<D-1} L_i^2 + kL_{D-1}^2 - \frac{\left(\sum_{i<D-1} L_i + kL_{D-1}\right)^2}{D+k-1}.$$

On the witness $L = e_{D-1}$ (realized by $p = \mathcal{C}(e^{e_{D-1}})$, q uniform): $d_A^2 = 1 - \frac{1}{D} = \frac{D-1}{D}$ and $d_A(E\cdot)^2 = k - \frac{k^2}{D+k-1} = \frac{k(D-1)}{D+k-1}$, whence the ratio

$$\frac{d_A(Ep, Eq)^2}{d_A(p, q)^2} = \frac{kD}{D + k - 1} \xrightarrow{k\to\infty} D.$$

An explicit cosine expansion. The strictly positive column-stochastic map and interior pair

$$T = \begin{pmatrix} .084 & .846 & .859 \\ .114 & .136 & .073 \\ .802 & .018 & .068 \end{pmatrix}, \quad p = \begin{pmatrix} 0.902 \\ 0.093 \\ 0.005 \end{pmatrix}, \; q = \begin{pmatrix} 0.714 \\ 0.200 \\ 0.086 \end{pmatrix},$$

give $1 - \cos(p, q) = 0.020420$ and $1 - \cos(Tp, Tq) = 0.034041$, an expansion of $1 - \cos$ by a factor 1.667. Cosine dissimilarity is therefore not Markov-nonexpansive, by explicit witness and not by sampling frequency alone.

Extremality, analytically. Take $j = D-1$ without loss of generality (relabel coordinates). Write both forms on the slice $\{\mathbf{1}^\top L = 0\}$ and decompose it orthogonally as $V_1 \oplus \mathrm{span}(w)$, where $V_1 = \{v : v_{D-1} = 0,\ \sum_{i<D-1} v_i = 0\}$ (dimension $D-2$) and $w = e_{D-1} - \frac{1}{D}\mathbf{1}$. For $v \in V_1$: the original-dimension form gives $\|v\|^2$ and the split-dimension form gives $\sum_{i<D-1} v_i^2 + k \cdot 0 - \big(\sum_{i<D-1} v_i\big)^2/(D+k-1) = \|v\|^2$, so the ratio is 1 on V_1. Both bilinear forms vanish between V_1 and w: with weights $c = (1, \ldots, 1, k)$ and $W = \mathrm{diag}(c)$, $\langle Wv, w\rangle - \frac{(c\cdot v)(c\cdot w)}{D+k-1} = 0$ and $\langle v, w\rangle - \frac{(\mathbf{1}\cdot v)(\mathbf{1}\cdot w)}{D} = 0$ because $v_{D-1} = 0$ and $\sum_{i<D-1} v_i = 0$. Hence the pencil is block-diagonal with eigenvalue 1 on V_1 and $kD/(D+k-1)$ on w, so the witness is extremal and, for $k \geq 2$, $D \geq 2$, the maximal ratio strictly exceeds 1. Exp. 3 confirms its square root to numerical precision.

Family-level incompatibility with dimension-dependent scales. Suppose $\{\delta_D\}_{D\geq 3}$ satisfies (S1)–(S4) in every dimension, so $\delta_D = \beta_D d_A$ by Theorem 3 and suppose the family is nonexpansive under every Markov morphism. Let $C : \mathcal{S}^{D+k-1} \to \mathcal{S}^D$ sum the k split coordinates back. Then C is deterministic column-stochastic and $CE_{D,k} = \mathrm{Id}$. Then $\delta_D(p,q) = \delta_D(CEp, CEq) \leq \delta_{D+k-1}(Ep, Eq) \leq \delta_D(p,q)$, so $\beta_{D+k-1} d_A(Ep, Eq) = \beta_D d_A(p,q)$ for every pair. Any pair with $L \in V_1$ (e.g. supported on two untouched coordinates) has $d_A(Ep, Eq) = d_A(p,q)$ by the block computation above, forcing $\beta_{D+k-1} = \beta_D$. The witness pair then violates the displayed equality. □

Same-dimension counterexamples for d_A. With

$$T = \begin{pmatrix} .01 & .98 & .98 \\ .61 & .01 & .01 \\ .38 & .01 & .01 \end{pmatrix}, \qquad p = (.68, .31, .01), \qquad q = (.26, .73, .01),$$

$$d_A(p,q) = 1.28615, \qquad d_A(Tp, Tq) = 1.41503,$$

an expansion by $\times 1.1002$ under a strictly positive column-stochastic same-dimension map. A second construction, $T = 0.98\,T_0 + 0.02\,U$ with T_0 columns $\big((0,1,0), (\frac{1}{2}, 0, \frac{1}{2}), (0,0,1)\big)$, $p = (.03, .969, .001)$, $q = (.77, .229, .001)$, expands by $\times 1.0690$. A local search around the first example reaches $\times 1.150$ (Exp. 3). The mechanism is near-deterministic transport of a small coordinate through a large log-ratio contrast. The witness also embeds in higher dimension: at $D = 16$, the strictly positive map $(1 - 10^{-3})\,\mathrm{diag}(T, I_{13}) + 10^{-3}\mathbf{1}\mathbf{1}^\top/16$ applied to $(0.9p,\ 0.1\mathbf{1}/13)$ and $(0.9q,\ 0.1\mathbf{1}/13)$ expands d_A by $\times 1.17$. Because the expansion ratio is continuous in (T, p, q), each strict witness has an open neighbourhood of expanding maps and pairs of the same dimension. The 4000 random strictly positive map–pair draws of Exp. 3 at $D = 16$, none of which expanded d_A, therefore missed a set of positive measure. They do not show that expansion is confined to special cases. In contrast, merging coordinates can never expand d_A. If coordinates i and j are merged, the merged log-ratio $\log \frac{p_i + p_j}{q_i + q_j}$ lies between $\log \frac{p_i}{q_i}$ and $\log \frac{p_j}{q_j}$, so for every centring constant its squared deviation is at most the sum of the two original squared deviations, and minimizing over the constant gives the claim. The absence of expansion over 2×10^5 random pairs under a fixed three-to-two amalgamation is therefore guaranteed.

Invariance of f-divergences and Fisher–Rao distance under $E_{D,k}$. Each split coordinate contributes $k \cdot \frac{q_j}{k} f\big(\frac{p_j/k}{q_j/k}\big) = q_j f(p_j/q_j)$: summing, $D_f(Ep \| Eq) = D_f(p\|q)$. Likewise $\mathrm{BC}(Ep, Eq) = \sum_{i\neq j} \sqrt{p_i q_i} + k\sqrt{\frac{p_j}{k}\frac{q_j}{k}} = \mathrm{BC}(p,q)$, hence FR invariant. Monotonicity under general Markov morphisms is Csiszár's DPI (Csiszár, 1967) and Čencov's theorem (Cencov, 2000; Campbell, 1986). d_H is invariant under $E_{D,k}$ because the log-ratio multiset gains only repeated values. Combining with Theorem 3: any δ satisfying (S1)–(S4) equals βd_A and hence expands under $E_{D,k}$. Each of the listed Markov-monotone families (f-divergences and FR) has attenuation exponent $\gamma > 0$, hence is matched-sink noninvariant and non-separable by (ii)(a). In contrast, d_H is Markov-nonexpansive and matched-sink invariant and fails separability by the explicit four-point mixed-difference witness in part (iii) below. □

G.4 (II) Supplement: Euclidean distance violates same-dimension DPI

Let T_0 have columns (e_1, e_1, e_2) and $T = 0.98\,T_0 + 0.02\,U$ with U the uniform column matrix (T strictly positive, column-stochastic). Take $p = (0.35, 0.35, 0.30)$, $q = (0.25, 0.25, 0.50)$, so $p - q = (0.1, 0.1, -0.2)$ and $T_0(p-q) = (0.2, -0.2, 0)$: $\|p - q\|_2 = 0.245$, $\|T(p-q)\|_2 = 0.98 \cdot 0.283 = 0.277 > 0.245$. Cosine dissimilarity likewise expands on a positive fraction of random

Table 13: The property matrix behind Theorem 4. Here "bdry" means a finite jointly continuous extension to the closed simplex and "Markov" means nonexpansiveness under the specified stochastic maps. KL and χ^2 lack boundary continuity by the explicit witness in this appendix. $\checkmark^*$: Hilbert's metric satisfies DPI for every stochastic map with no zero row and is invariant under congruent embeddings. Strictly positive maps additionally give uniform strict contraction. Euclidean distance violates DPI already for same-dimension stochastic maps (this appendix). Cosine admits an explicit strictly positive expanding map ($\times 1.67$ in $1 - \cos$, this appendix) and expanded in 3.7 percent of random map–pair draws and d_A expands under structured same-dimension maps by $\times 1.10$ (Thm. 4, this appendix). Characterizations of FR, d_H and d_A require their respective additional hypotheses. The displayed properties alone are not asserted to characterize them.

	bdry	$\oplus$-inv	$\odot$-hom	Sep-A	Markov
Euclidean, cosine	✓	×	×	×	×
JS, TV, Hel^2	✓	×	×	×	✓
KL, χ^2 (f-div, unbounded)	×	×	×	×	✓
Fisher–Rao	✓	×	×	×	✓
Hilbert d_H	×	✓	✓	×	$\checkmark^*$
Aitchison d_A	×	✓	✓	✓	×

stochastic maps (3.7 percent in Exp. 3), so neither raw-vector dissimilarity is monotone even without dimension changes.

G.5 (III) Hilbert's metric

$\oplus$-invariance and $\odot$-homogeneity are immediate from $d_H(p, q) = \mathrm{osc}(\log p - \log q)$. Invariance under $E_{D,k}$: repeated log-ratio values do not change the oscillation. Nonexpansiveness under every stochastic map with no zero row (these preserve the positive cone's interior) is Birkhoff's theorem (Birkhoff, 1957; Bushell, 1973), with uniform strict contraction when the map is strictly positive, the modulus governed by the projective diameter of the image. *Failure of Sep-A*: work at $D = 3$ in the channel parametrization $x = \alpha u_0 + (0, w, -w)$. Translating by a constant (which osc ignores), $x \simeq (A', w, -w)$ with $A' = \sqrt{3/2}\,\alpha$, so $d_H^2 = F(A', w)$ with $F(A', w) = \mathrm{osc}(A', w, -w)^2$. Sep-A at coordinate 0 would make F additive in (A', w), because A' determines the balance channel and w the content channel. On the rectangle $A' \in \{0, 1\}$, $w \in \{0, 1\}$,

$$F(1, 1) = 4, \quad F(0, 1) = 4, \quad F(1, 0) = 1, \quad F(0, 0) = 0,$$

so the mixed second difference is $4 - 4 - 1 + 0 = -1 \neq 0$. *Failure of the parallelogram law* (hence not induced by any inner product (Jordan & Neumann, 1935)): for $u \simeq (1, -1, 0)$, $v \simeq (1, 0, -1)$: $\mathrm{osc}(u) = \mathrm{osc}(v) = 2$, $\mathrm{osc}(u + v) = \mathrm{osc}(2, -1, -1) = 3$, $\mathrm{osc}(u - v) = \mathrm{osc}(0, -1, 1) = 2$. Thus $9 + 4 \neq 2(4 + 4)$. Kohlberg–Pratt's converse (Kohlberg & Pratt, 1982) shows that within projective metrics, contraction under all positive linear maps identifies d_H up to increasing reparametrization and $\odot$-homogeneity pins the scalar (Remark 6), so the Hilbert row of Table 13 has a characterization under these additional hypotheses. The Fisher–Rao and Aitchison rows have their respective characterizations under the hypotheses of Čencov's theorem and Theorem 3. □

Remark 7 (Sources of Aitchison expansion). Under the congruent split $E_{D,k}$, f-divergences, the Fisher–Rao distance and Hilbert's metric are unchanged. The split has a deterministic left inverse, the merge C with $CE_{D,k} = \mathrm{Id}$, so a dissimilarity family that is nonexpansive under Markov morphisms must preserve its value under the split. Unweighted Aitchison distance instead gives the repeated log-ratio coordinates additional geometric weight and expands by at most the factor $\sqrt{kD/(D + k - 1)}$ of Theorem 4. This is a property of the geometry, not an increase in statistical distinguishability. The expansion also follows from subcompositional dominance (Aitchison, 1992). Deleting all but one copy of the split coordinate gives the subcomposition $\mathcal{C}(p_{-j}, p_j/k)$, and the common perturbation that multiplies its last part by k recovers p, and likewise q. Subcompositional dominance and perturbation invariance therefore give $d_A(p, q) \leq d_A(E_{D,k}p, E_{D,k}q)$. The same-dimension examples above show that expansion is not limited to refinement. The information-geometric analysis of Erb & Ay (2021) concerns coarse-graining operations such as amalgamation and is compatible with these expansions, and see also the Hilbert–Aitchison comparisons of Nielsen & Sun (2018). If a change of representation is modeled by the same stochastic map on both rows and contraction under that map is required, a Markov-nonexpansive dissimilarity is appropriate. For a fixed set of named positions,

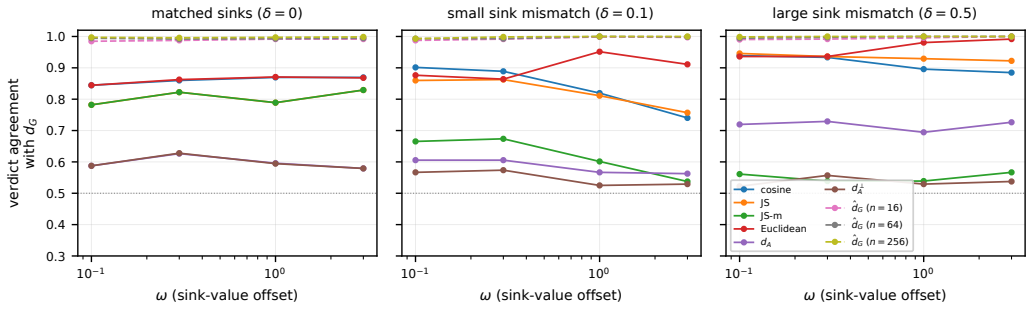


Figure 5: Exp. 8, factorial regime map. Verdict agreement with the output distance ("is q closer to p than r?") for each score, averaged over sink mass $s \in \{0.5, 0.8, 0.95\}$, content divergence and value anisotropy $\kappa \in \{1, 30\}$, against the sink-value offset ω, for matched, slightly mismatched and strongly mismatched sinks (sink logit jitter $\delta = 0$, 0.1 and 0.5). This offset is distinct from the norm ratio of Exp. 5, and the sink-channel weight itself is $\|\boldsymbol{\nu}\|_G^2$ in Theorem 1(b). Under a common matched sink mass, kept and masked JS induce identical verdicts because $\mathrm{JS}(p_s, q_s) = (1 - s)\mathrm{JS}(\pi, \rho)$. Under mismatch, the tested masked scores have lower verdict agreement with the output distance. The Gram plug-in (dashed) reaches 0.98–1.00 on average at $n = 16$ sampled value matrices. In the legend, JS-m is masked JS and $d_A^{\perp} = d_A(\pi, \rho)$ is the Aitchison distance between the contents.

one may instead impose the compositional and separability axioms. Theorem 4 shows that the two requirements cannot be imposed together. These conclusions concern the unweighted geometry, and weighted Aitchison geometries require a separate analysis.

H Comparison across regimes

Exp. 8 places the row dissimilarities, the Gram plug-in and the output distance on one synthetic grid in which the sink mass, the sink mismatch, the content divergence, the value anisotropy and the sink-value offset are controlled (Figure 5 and Table 14). Matched-sink attenuation changes numerical scale and does not, by itself, change an f-divergence ranking. The complete-output distance $\mathcal{D}$ and the recipient-slot distance S are not dissimilarities on rows, so the separability statements do not apply to them, and they are compared with the row scores as predictors of substitution cost and as selection criteria in Appendix C.

Scope of the Gram comparison The two-channel report makes the sink and content contributions explicit. Under a common value matrix, the Gram distance identifies the local output distance and extends continuously to the closed simplex for every fixed finite G. It does not avoid matched-sink attenuation, since

$$d_G(p_s, q_s) = (1 - s)\sqrt{(\pi - \rho)^\top G_{cc}(\pi - \rho)} = (1 - s)\, d_{G,c}(\pi, \rho).$$

The uncompressed content pseudometric removes this factor by changing the estimand. Remaining limitations include the distinction between fixed-signature and input-coupled targets, estimation from sampled values and the difference between local output distance and network response. The observed response correlations do not establish a ceiling on predictability.

Exp. 8 (factorial regime map). Figure 5 and Table 14 report this experiment. Synthetic triples (p, q, r) at $D = 64$ have sink mass $s \in \{0.5, 0.8, 0.95\}$, sink mismatch given by a logit jitter $\delta \in \{0, 0.1, 0.5\}$, the three panels of Figure 5 and columns of Table 14, and content divergence given by a clr jitter in $\{0.05, 0.2, 0.6\}$. Value matrices have a content spectrum of condition number $\kappa \in \{1, 30\}$, and the sink value is placed at the mean content value plus an offset $\omega \in \{0.1, 0.3, 1, 3\}$ times the mean content norm. We use 40 triples per cell across 216 cells and one per-sequence value-noise model for the sampled Grams.

With matched sinks, the aggregated row-only scores in Table 14 reach at most 0.87 agreement: cosine and Euclidean reach 0.84–0.87, JS 0.78–0.83 and d_A 0.58–0.63. Kept and masked JS induce identical

Table 14: Exp. 8: closer-of-two agreement with d_G, averaged over the tested sink masses, content perturbations and value-anisotropy settings, with the range over ω shown for each sink-mismatch setting δ. Bold marks ranges whose minimum is at least 0.90.

score	$\delta = 0.0$	$\delta = 0.1$	$\delta = 0.5$
cosine	0.84–0.87	0.74–0.90	0.88–0.94
JS	0.78–0.83	0.76–0.86	**0.92–0.95**
Hellinger	0.78–0.83	0.76–0.86	**0.92–0.95**
Fisher–Rao	0.78–0.83	0.76–0.86	**0.92–0.95**
TV	0.77–0.82	0.85–0.89	**0.94–0.97**
Euclidean	0.84–0.87	0.86–0.95	**0.94–0.99**
JS (masked)	0.78–0.83	0.54–0.67	0.54–0.57
cosine (masked)	0.82–0.86	0.53–0.70	0.52–0.55
d_A	0.58–0.63	0.56–0.61	0.69–0.73
d_H	0.53–0.57	0.57–0.60	0.72–0.72
$d_A^{\perp}$	0.58–0.63	0.53–0.57	0.52–0.56
$\lvert\Delta \operatorname{logit} s\rvert$	const.	0.74–0.94	**0.90–0.97**
$\hat{d}_G$, $n = 16$	**0.98–0.99**	**0.99–1.00**	**0.99–1.00**
$\hat{d}_G$, $n = 64$	**0.99–0.99**	**0.99–1.00**	**0.99–1.00**
$\hat{d}_G$, $n = 256$	**1.00–1.00**	**0.99–1.00**	**1.00–1.00**

verdicts under the common sink mass, as Theorem 2(d) predicts, but this equivalence does not hold for every score. With a small mismatch, kept cosine and JS degrade as ω grows ($0.90 \to 0.74$ and $0.86 \to 0.76$), Euclidean improves ($0.86 \to 0.95$) because verdicts become sink-driven, and masked cosine and JS fall to 0.53–0.70. With a large mismatch, kept cosine, Euclidean, JS, TV, Hellinger and Fisher–Rao reach 0.88–0.99, whereas d_A reaches 0.69–0.73 and d_H approximately 0.72. In this synthetic grid, Euclidean distance has the highest agreement among row-only scores under large mismatch, and $|\Delta \operatorname{logit} s|$ alone reaches 0.90–0.97. The masked scores reach 0.52–0.57, close to the 0.5 agreement of an independent fair binary verdict when ties are absent, and no independence test is made.

At $n = 16$, the aggregated Gram-verdict agreements in Table 14 range from 0.98 to 1.00. The table averages over operating points and does not report the worst of the 216 individual cells. At $D = 64$, $n = 16$ and $\eta = 0.05$, the entrywise bound of Theorem 1(c) equals $4.90R^2$, which exceeds the elementary maximum $4R^2$ of the absolute squared-distance error. The observed agreement therefore describes this generator, and it does not make the worst-case bound informative at $n = 16$.

For a common matched sink mass, exact isotropy of the content Gram on content contrasts, $C_c G_{cc} C_c = c_G C_c$ with $C_c = I_{D-1} - \mathbf{1}\mathbf{1}^\top/(D-1)$ and $c_G > 0$, implies $d_G(p_s, q_s) = \sqrt{c_G}\,\|p_s - q_s\|_2$ by Theorem 1(b), so Euclidean and Gram verdicts then agree exactly. The recorded Euclidean agreement of 0.89 at $\kappa = 1$ with matched sinks, against 0.84 at $\kappa = 30$, shows that the sampler labelled $\kappa = 1$ does not make G_{cc} exactly isotropic on content contrasts. These results describe agreement under the specified synthetic sampler. They do not establish a model-independent regime in which a row-only score recovers output-distance orderings over bounded value matrices.

I VERIFICATION EXPERIMENTS

Table 15 summarizes Exps. 1–6. Experiments 1–5 check numerical implementations of the proved identities, exponents, constants and counterexamples on synthetic or calibrated rows. They complement the proofs and check the implementation on the tested instances, and they are not independent evidence about the models. Experiment 6 checks the sharp reversal criterion on real head signatures. Throughout, the working dimension is $D = W = 64$ unless stated.

Exp. 1 (identities). The 21 listed identities and specializations of Theorem 2 were checked numerically on 1200 random pairs at $D \in \{3, 8, 64\}$ and 300 calibrated pairs. The maximum absolute residuals range from 2.2×10^{-16} to 3.6×10^{-12} over the 1,500 instances and the Exp. 1 script of the supplementary material writes the residual of each identity.

Exp. 2 (attenuation). Figure 7: fitted exponents over 40 calibrated content pairs match the spectrum with cross-pair variation below the displayed precision (1.000 ± 0.000, 0.500 ± 0.000, 0.000 ± 0.000

Table 15: Numerical checks of the proved statements (Exps. 1–5) and the implementation check of the reversal criterion on real rows (Exp. 6). Details, figures and tables are in this appendix.

Exp.	what is checked	result
1	21 identities and specializations of Thm. 2	residuals 2.2×10^{-16}–3.6×10^{-12} over 1500 instances
2	attenuation exponents and constants (Thm. 2(e))	fitted exponents $2.02/1.00/0.50/0.00$, constants to 1.0002 or better
3	split expansion and Markov witnesses (Thm. 4)	closed form to 2×10^{-15}. Witnesses $\times 1.10$, $\times 1.13$, $\times 1.67$
4	separability defects (Thm. 3)	machine zero where predicted, nonzero elsewhere
5	output-distance interval, Gram identity, envelope (Thm. 1)	endpoints to below 10^{-15}. Envelope holds at every n
6	sharp JS dominance condition (Prop. 4) on real signature triples	implementation check: kept verdict as implied on all 112,584 covered triples over nine runs; coverage 18.0, 0.05 and 15.2 percent of within-layer triples on GPT-2, Pythia and Llama

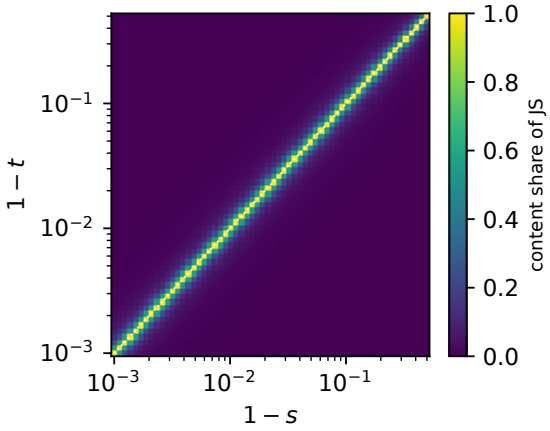


Figure 6: Exp. 2, unmatched sinks. Content share of the JS divergence between two calibrated rows as a function of their sink masses s and t (logarithmic axes in $1-s$ and $1-t$). On the diagonal of matched sinks the Bernoulli term vanishes and the content term carries the whole divergence. Proposition 5 guarantees asymptotic Bernoulli dominance when both sink masses tend to one at mismatched rates. It does not assert dominance at every point off the diagonal.

and cosine 2.020, consistent with the first-order finite-s correction) and the asymptotic constants of Theorem 2(e) are verified to ratios 1.0002, 1.00000001, 1.0000000002 at $1-s = 10^{-4}$. The exact content share of JS over the $(1-s, 1-t)$ plane (Prop. 5) is Fig. 6.

Exp. 3 (Markov morphisms). Figure 8: the measured d_A split-expansion equals $\sqrt{kD/(D+k-1)}$ to 2.2×10^{-15} and coincides with the square root of the top generalized eigenvalue of the two-block pencil (Appendix G). JS, TV, Hellinger, KL, FR and d_H are invariant under the split to 5×10^{-15}. Both displayed same-dimension counterexamples reproduce to the reported precision ($\times 1.1002$ and $\times 1.0690$) and a local search reaches $\times 1.150$. Over 4000 independent map–pair draws (one random strictly positive map, one random pair each) at $D = 16$: no DPI family expanded, and no expansion was observed for d_H (Birkhoff) or for d_A and Euclidean distance, although explicit counterexamples exist for both (Thm. 4, Appendix G), and cosine expanded in 3.7 percent of draws. Over 2×10^5 random input pairs under one fixed 3→2 amalgamation, no d_A expansion was observed (max ratio 0.99999), as guaranteed because merging coordinates cannot expand d_A (Appendix G).

Exp. 4 (separability). Table 16: over 200 random channel rectangles, the Sep-A defect of d_A is 2.1×10^{-14} while every other dissimilarity has defects of order 0.06–73. Every classical dissimilarity tested has a nonzero Sep-P defect. Proposition 7 explains the obstruction for attenuating families,

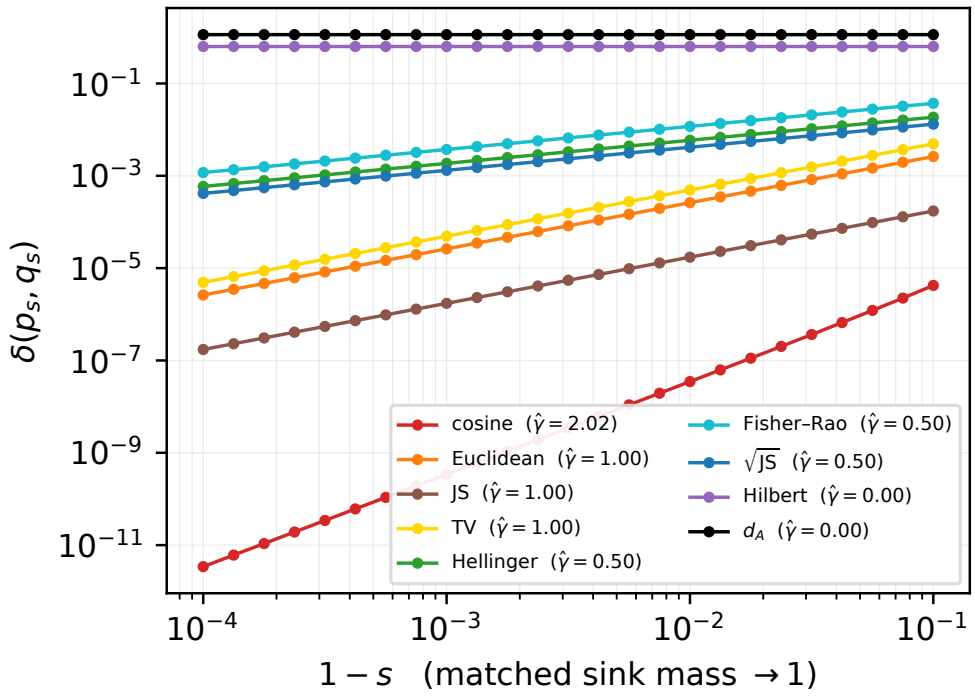


Figure 7: The attenuation spectrum (Thm. 2(e)) on a content pair calibrated to measured model statistics: log–log decay of nine dissimilarities as matched sink mass $s \to 1$. Fitted slopes $\hat{\gamma}$ (legend) reproduce the predicted exponents $2, 1, 1/2, 0$, and over 40 calibrated content pairs the fitted exponents are 1.000 ± 0.000, 0.500 ± 0.000, 0.000 ± 0.000 and 2.020 for cosine, whose first-order finite-s correction explains the offset from 2 (Appendix E). Fits use $s \in [0.9,\ 1 - 10^{-4}]$. The distances d_A and d_H are exactly flat.

while Remark 5 gives the exact Aitchison defect, matched numerically to 4.3×10^{-14}. The scalar metric d_P of that remark has a defect of 2.8×10^{-14}.

Exp. 5 (output distance). Figure 9 uses independently drawn sink masses, allowing $\Delta s \neq 0$. Constructed suprema and infima attain $[0, B\|p - q\|_1]$ to below 10^{-15}. Both orderings of a fixed pair of head pairs are realized by explicit value matrices (8×10^{-16} vs. 0.221, then 4.23 vs. 5×10^{-17}). The Gram and three-term channel identities hold to 2×10^{-13}. Across the sink-value sweep, the sink-retaining scores correlate 0.80–1.00 with d_G, whereas the content-only reporter $d_A^{\perp}$ has a correlation of -0.05 at both displayed endpoints ($\omega = 0.1$ and $\omega = 20$, over 120 calibrated pairs with independent sink masses and 25 Gram draws per ω). These correlations describe the calibrated sampler. The decomposition of Theorem 1(b) contains the content coefficient $(1 - \bar{s})^2$, but the relative contribution of the content channel also depends on $\|\Delta_c\|_G^2$, the sink term and the mixed term, which these correlations do not separate. Estimation uses bounded sphere-constructed value samples matching G exactly. The certified $4\|\hat{G}_n - G\|_{\max}$ envelope holds at every n. The vertex-difference maximum (a lower bound on the pairwise supremum, since quadratic errors can peak on faces) tracks it and both decay with single-realization fitted slope -0.64.

Table 16: Maximal four-point (rectangle) separability defects of δ^2 over 200 random channel rectangles at $D = 16$ (Exp. 4): Sep-A tests balance/content channels, Sep-P tests mass/content channels. Machine-precision zeros appear exactly where Theorem 3 and Remark 5 predict.

dissimilarity	Sep-A defect	Sep-P defect
Euclidean	0.46	0.12
cosine	0.39	0.26
JS	0.06	0.03
$\sqrt{\text{JS}}$	0.12	0.07
TV	0.26	0.13
Hellinger	0.26	0.14
Fisher–Rao	1.10	0.62
Hilbert d_H	7.73	14.44
clr-ℓ_1	72.89	68.25
Aitchison d_A	2.1×10^{-14}	2.79
$d_P = \sqrt{(\Delta \operatorname{logit} s)^2 + (d_A^{\perp})^2}$	n/a	2.8×10^{-14}

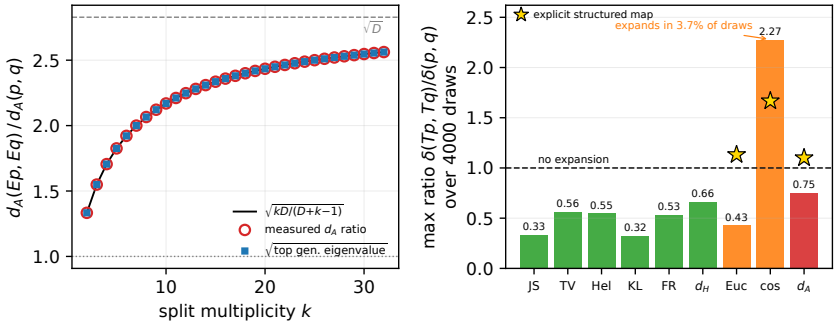


Figure 8: Left: d_A expansion under k-fold congruent splits at $D = 8$: measured ratio, closed form $\sqrt{kD/(D+k-1)}$ and the square root of the top generalized eigenvalue of the induced quadratic-form pencil coincide to 2×10^{-15}, approaching $\sqrt{D}$ (dashed). Right: maximum observed ratio $\delta(Tp, Tq)/\delta(p, q)$ over 4000 independent map–pair draws (one random strictly positive stochastic map, one random pair each, $D = 16$), dashed line at 1. The six provably nonexpansive scores stay below 1 in every draw. Cosine exceeds 1 in 3.7 percent of draws (maximum $\times 2.27$). Stars mark the explicit witnesses of Thm. 4 and Appendix G: cosine $\times 1.67$ (in $1 - \cos$), Euclidean $\times 1.13$ and d_A $\times 1.10$. Aitchison and Euclidean distances contract in these random draws but expand for the displayed counterexamples.

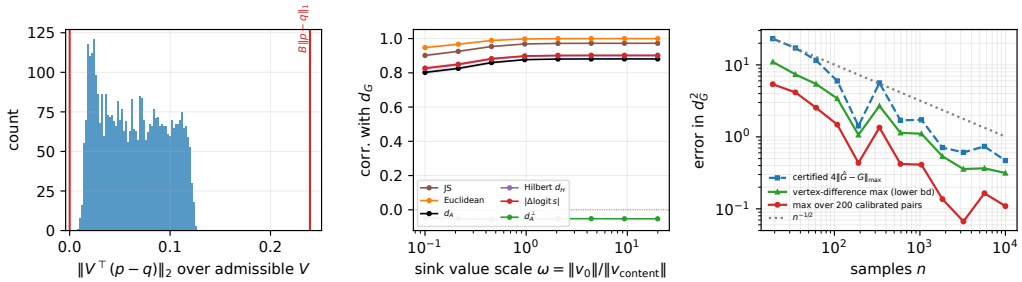


Figure 9: Numerical evaluation of Theorem 1, with independently drawn sink masses so the sink channel can be active. Left: realizable output distances for one calibrated row pair fill $[0, B\|p-q\|_1]$, and constructed endpoints attain the bounds to below 10^{-15}. Middle: correlation of each score with d_G as the sink-value norm ratio $\omega = \|v_0\|/\overline{\|v_{\text{content}}\|}$ varies (distinct from the offset of Exp. 8). Under this calibrated sampler, the sink-retaining scores correlate 0.80–1.00 and rise with ω, and the content-only reporter $d_A^\perp$ has a correlation of -0.05. The synthetic sweep agrees with the exact Gram and three-term channel identities, and the correlations depend on the sampler and do not establish channel dominance in real head populations. For pairs with linearly independent row differences, suitable values realize either strict ordering (part (a)). Right: uniform error in $d^2_{\hat{G}_n}$ with bounded value samples: the envelope $4\|\hat{G}_n - G\|_{\max}$ (Thm. 1(c)), the maximum over vertex-pair differences (a lower bound on the supremum over all row pairs) and the maximum over 200 calibrated pairs, with fitted slope -0.64 over the plotted range. The theorem supplies a high-probability upper bound, not a realization-wise decay rate.

Exp. 6 (the reversal criterion on real rows). On all 112,584 within-layer signature triples covered by the sharp JS dominance condition of Proposition 4 (4,274 on GPT-2, 11 on Pythia and 108,299 on Llama over three seeds), the kept verdict is the one the condition implies. Because the condition is sufficient by an identity, this checks the implementation. The condition covers 18.0, 0.05 and 15.2 percent of all within-layer triples, and kept and masked JS verdicts are opposite on 21.4, 8.0 and 26.0 percent of triples.

J Related work

Relation to Papamichalis & Ruane (2026) Table 17 distinguishes classical identities, results restated from Papamichalis & Ruane (2026) and results established here. The entropy chain rule is classical, and that paper uses it in sink–content coordinates. Papamichalis & Ruane (2026) measure reversals caused by keeping or removing the sink, prove the Pythagorean split of the Aitchison distance and characterize the Aitchison metric under a parallelogram law and permutation invariance. Theorem 3 replaces the parallelogram assumption by all-coordinate separability in Aitchison coordinates and derives permutation invariance. The identification, estimation, transfer and intervention results address a different target, the local output distance for specified attention signatures and values.

Attention sinks Attention sinks were characterized in streaming inference (Xiao et al., 2024), related to massive activations (Sun et al., 2024), studied during pretraining (Gu et al., 2025) and argued to help models avoid over-mixing (Barbero et al., 2025). Su et al. (2026) survey the area, and vision transformers exhibit analogous high-norm tokens that registers absorb (Darcet et al., 2024). Guo et al. (2024) analyse an active–dormant mechanism linking attention sinks and value-state drains, derive training dynamics on a simplified Bigram–Backcopy task and examine corresponding behaviour in pretrained Llama and OLMo models. Wong et al. (2025) identify secondary sinks with distinct layers of onset and lifetimes and trace their formation to middle-layer MLPs. We instead study the comparison problem conditional on the attention rows and projected values. When $v_0 = 0$ the output is $(1 - s) \sum_{i \geq 1} \pi_i v_i$, the content attenuation present in the active–dormant mechanism. Corollary 2 quantifies how a change of sink mass changes a comparison of outputs even when the sink contributes no value vector. It does not imply that a dormant head has a large absolute output or a large effect on the prediction. The grouped identities of Theorem 2(f) apply to any specified partition of positions, including several designated sinks. The present experiments do not detect secondary sinks or evaluate their layer-dependent onset and lifetime.

Attention weights, values and head comparison Kobayashi et al. (2020) analyse the norms of transformed input vectors, and a Gram formulation retains the pairwise inner products as well as these norms. Brunner et al. (2019) study whether attention weights can be recovered from outputs under a fixed value transformation. Our identification result instead fixes the rows and varies bounded values. The debate on attention as explanation (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019) concerns what individual attention rows reveal about a prediction. Head comparisons by attention similarity (Clark et al., 2019) supply the row scores evaluated in Section 6. Importance-based head pruning (Michel et al., 2019; Voita et al., 2019) estimates which heads the loss depends on, and representation similarity measures such as CKA (Kornblith et al., 2019) compare whole representations through their Gram matrices. Neither is evaluated here, and Section 7 states how the substitutability question differs from both. Hilton-Jones et al. use Aitchison geometry to study token distinguishability and temperature scaling in long contexts, a different question from head comparison. The quadratic identity and the concentration and testing arguments are standard tools. The contribution is their use to characterize an explicit output-distance estimand, determine its minimax estimation rate and test the consequences for head-comparison decisions.

Compositional geometry The grouping identities for entropy and KL are classical, as is the between/within balance decomposition in compositional geometry (Egozcue et al., 2003; Egozcue & Pawlowsky-Glahn, 2005). Theorem 2 organizes the metric-specific sink–content identities and their matched-sink attenuation laws. The characterization and incompatibility results concern the explicitly stated compositional axioms and distinguish row geometry from the value-dependent geometry of head outputs.

Table 17: Provenance of the results. "Classical" marks textbook identities, "restated" marks results taken from Papamichalis & Ruane (2026) and "new" marks results established in this paper.

result	status
entropy chain rule	classical, used in sink–content coordinates by Papamichalis & Ruane (2026)
Pythagorean split of d_A	standard ILR orthogonality, with the attention formulation restated from Papamichalis & Ruane (2026, Lemma 1)
keep-versus-mask reversal measurements	restated as motivation (Papamichalis & Ruane, 2026, Theorem 1 and experiments)
matched-sink range collapse (Cor. 3)	related to the dynamic-range collapse of Papamichalis & Ruane (2026, Theorem 3), and the present formulas follow from Theorem 2(d)
skew-JS, Hellinger and Fisher–Rao sink–content formulas	consequences of standard divergence and spherical identities, organized here for sink masking
attenuation constants, unmatched-sink asymptotics, sharp JS dominance (Props. 4, 5)	new, with explicit calculations for the sink–content comparison problem
separability characterization with derived permutation invariance (Thm. 3)	new, whereas Papamichalis & Ruane (2026, Theorem 2) assume permutation invariance and a parallelogram law
incompatibility theorems, split constant, expansion witnesses (Thm. 4)	new
identification interval and minimax estimation rate (Thm. 1, Prop. 3)	new for this estimand, using standard Gram, concentration and testing arguments
complete-output approximation and intervention bounds, substitution versus removal (Props. 1, 2, Cor. 1)	new, explicit consequences of norm inequalities for the stated targets
certificate scope, norm and cross term, shared values, value coordinates (Remark 2 in Appendix A.4, Remarks 1 and 3)	new, elementary consequences of Props. 1 and 2, derived after the first results
real-model comparisons, decision experiments, regime map	new
within-layer, sink-masking, key–value-group, within-group, Euclidean, norm-reference, alignment, ratio and bootstrap analyses	new, added after the first results (Appendix D, Tables 8 and 9)
training versus held-out verdict null	control of Papamichalis & Ruane (2026), adapted to the output distance