What Attention Control Can Reach: Exact Output Geometry

Anonymous authors
Paper under double-blind review

Abstract

Editing where a transformer's attention heads look is now a standard control, and the same patterns are read as evidence of what it computes. Both rest on an unchecked assumption: that pushing a head's attention produces the intended change in its output. For one head at a fixed input we settle an edit's reach and its cost. Reach is fixed before any budget is spent: a bias shared by a group of keys moves each row of the head's output only inside the hull of one average value vector per group, a region the head's values and unedited attention fix. Every interior point is reached by a closed-form bias, a target outside stays out at any budget, and a linear readout's largest displacement follows from one number per group on a single unedited pass. Cost is charged only to edits that cross the softmax: a row's distance from uniform and its first-order response to logit edits obey a sharp product bound, response vanishes as a row becomes deterministic, and a finite edit pays along its whole path, not at its start. In two live models the excluded directions do not move beyond the float32 floor of the head's own output; in 7 pretrained transformers the upper tail of trained rows sits nearer the bound than at random initialization; and on one model's pre-registered held-out split the routing law read at the spent budget, not differentiated at zero, orders real interventions its derivative misorders more often than not. An attention edit can therefore be screened before it is run; neither its reach nor its cost is legible in the pattern itself.

1 Introduction

Attention plays two roles. As a diagnostic, a concentrated row of attention weights marks a distinctive routing pattern, most of a head's weight on a few keys; as a control surface, the same row is edited, mass pushed onto chosen keys by a logit bias or an equivalent reweighting (Zhang et al., 2024a; Guardieiro et al., 2025). Both uses assume that pushing a head's attention produces the output change one intends, and the second raises two questions the first does not answer: which changes of the head output such an edit can produce at all, and how far the routing can move under the allowed budget.

Both are properties of the map from edit to head output, not of the method's name. For one head of scaled dot-product attention (Vaswani et al., 2017) that map is the chain

$$z \longmapsto \mathbf{L}(z) \longmapsto \mathbf{A}(z) = \mathrm{softmax}(\mathbf{L}(z)) \longmapsto Y(z) = \mathbf{A}(z)U(z) + \Delta(z),$$

with z any parameter an edit moves, $\mathbf{L}$ the pre-softmax logits, Y the head output, the attention $\mathbf{A} = (a_{ij})$ averaging the rows u_j of the value-output operator U, the value vectors after the output projection, and Δ any term added after the softmax. The last link, the averaging, limits reach. A control that adds one bias b to the logits of a key *span* S, any set of keys, moves the span's *routing mass* $\psi = \sum_{j\in S} a_{ij}$ to $\psi(b)$ and, with the *conditional value centroids* $\mu_S = \psi^{-1}\sum_{j\in S} a_{ij}u_j$ and $\mu_{\bar{S}}$, the attention-weighted mean value inside and outside the span, changes row i of the output by exactly

$$Y_i(b) - Y_i(0) = \underbrace{\big(\psi(b) - \psi\big)}_{\text{routing-mass gain}}\ \underbrace{\big(\mu_S - \mu_{\bar{S}}\big)}_{\text{payload contrast}}.$$

A vanishing contrast is therefore the only way the control can be inert at every budget; with one bias per group over k groups the row output is a k-way softmax over k fixed centroids and reaches exactly the relative interior of their convex hull.

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.

The softmax exposes a cost. With d_i the distance from uniform of a row on m keys and γ_i, its *aperture*, the operator norm of its softmax Jacobian, $d_i\gamma_i \le C(m)$ for a sharp constant $C(m) \le \sqrt{2}/4$: concentration can make a row distinctive while narrowing its local response to logit perturbations. The chain rule puts that cost on the logit-changing component alone, and a finite edit pays it integrated along its path, not at the base point, the unedited forward, where steering effects are usually predicted.

Where a head looks does not fix what an edit can change, because the values do and the pattern does not show them; how sharply it looks fixes the opposite of what it seems to, because distinctiveness and response obey a sharp product bound. The hull geometry itself is classical, the mean map of an exponentially tilted family (§3); what is new is the attention reading.

Contributions. We prove the exact set of outputs that k groupwise logit biases can reach: the relative interior of the hull of k centroids, of affine rank at most $k-1$, with a criterion for inertness that holds at every budget. Read through a fixed linear readout, the same factorisation gives a budget-free ceiling, inert under a condition strictly weaker than the vector one (Corollary C.3). A path law splits a general intervention into logit, value and post-softmax routes, putting the aperture on the logit route alone and closing the finite budget by integration along the path. The sharp bound $d_i\gamma_i \le C(m)$ locates where logit responsiveness vanishes (Figure 2). And the same routing law, read at the budget an edit spends rather than differentiated at zero, gives a predictor that orders real interventions a zero-bias derivative misorders more often than not, on a pre-registered held-out split of one model and one task. The identities the results rest on are machine-checked (Appendix M).

Scope. The control geometry is row-wise for fixed values and the interface theorem per head at a fixed input (aggregation in Appendix C); the aperture is a local Euclidean Jacobian norm, so other budget norms change its constants.

2 Attention as a Control Surface

Attention. A head receives the layer input $X \in \mathbb{R}^{n\times d_{\text{model}}}$, with n the sequence length and d_{model} the residual width, and forms $Q = XW_Q$, $K = XW_K$, $V = XW_V$ with $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}}\times d_h}$, with d_h the head width. The pre-softmax logits are $\mathbf{L} = QK^\top/\sqrt{d_h}$; the attention matrix $\mathbf{A} = \mathrm{softmax}(\mathbf{L})$ applies the softmax to each row over that row's unmasked support Σ_i of size m_i ($m_i = n$ without masking), so $\mathbf{A}$ is row-stochastic. The head output is

$$Y = \mathbf{A}\,U, \qquad U := VW_O = XP, \qquad P := W_VW_O, \tag{1}$$

with $W_O \in \mathbb{R}^{d_h\times d_{\text{model}}}$; U (an activation, rows u_j) is the value-output operator and P its parameter.

Softmax Jacobian and aperture. Row i has Jacobian $J_i = \mathrm{diag}(\mathbf{A}_i) - \mathbf{A}_i\mathbf{A}_i^\top$, symmetric positive semidefinite with $J_i\mathbf{1} = 0$ (Gao & Pavel, 2017); a first-order logit perturbation moves the row by $\delta\mathbf{A}_i = J_i\,\delta\mathbf{L}_i$. The row's *aperture* is $\gamma_i = \|J_i\|_{\text{op}}$, the largest first-order change of the row per unit change of its logits, and the head's is $\gamma(\mathbf{A}) = \max_i \gamma_i$.

Row distinctiveness. $d_i = \|\mathbf{A}_i - \frac{1}{n}\mathbf{1}\|_2$ is the distance from uniform in the ambient n-key simplex, including masked zero coordinates; the support-aware distance d_i^Σ uses uniform on Σ_i instead, with $(d_i^\Sigma)^2 = d_i^2 + 1/n - 1/m_i$.

Three intervention interfaces. An intervention is a parameter z on which the head output depends. It enters through the *pre-softmax interface* (the logit route) if z affects Y only through the logits, $Y(z) = \mathrm{softmax}(\mathbf{L}(z))\,U$: logit surgery, span steering (PASTA (Zhang et al., 2024a) and SpotLight (Venkateswaran & Contractor, 2026), §3), and edits to W_Q or W_K. It enters through the *post-softmax state interface* if it acts on the head output or the residual stream with $\mathbf{A}$ held fixed, $Y(z) = \mathbf{A}\,U + \Delta(z)$ (Li et al., 2023b; Turner et al., 2024; Panickssery et al., 2024; Zou et al., 2023), and through the *parameter interface* if it edits a weight downstream of the softmax, $Y(z) = \mathbf{A}\,X\,(P + \delta P(z))$, or an MLP weight (Meng et al., 2022).

Responses and certificates (bounds computable in advance) are first order and per head at the given input unless integrated along a path; heads and layers compose by explicit bounds (Appendix A).

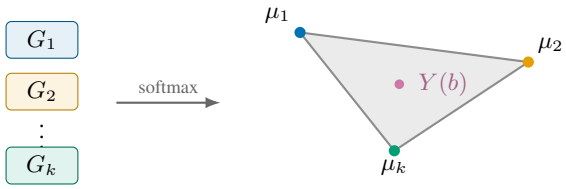

Figure 1: **The partition fixes the reachable set.** One bias per key group makes the row output a k-way softmax over the k centroids the values fix, $Y(b) = \sum_g w_g(b)\,\mu_g$, so it ranges over the relative interior of their hull, of affine rank at most $k-1$ (Theorem 1).

3 Reachable Output Geometry and Path Cost

What a bounded edit can reach. A bias shared by every key in a group multiplies the group's unnormalised weights by one common factor, so no proportion inside the group changes and the control can only move mass between groups; everything below follows. A span is the two-group case, S and its complement. Adding a constant to every logit changes nothing, so under a cap B on the edit's range the reachable rows are a ball of Hilbert's projective metric (Birkhoff, 1957; Karlsson et al., 2006) whose vertices bias one set of keys, and a linear readout is maximised by biasing one span (Proposition C.2).

Theorem 1 (Exact groupwise control geometry)**.** *Fix one row with entries a_j, its value rows u_j held fixed, and partition its positive support into k nonempty groups $G_1, \dots, G_k$. Let $\pi_g = \sum_{j\in G_g} a_j$ and $\mu_g = \pi_g^{-1} \sum_{j\in G_g} a_j u_j$ be the base mass and conditional value centroid. Adding bias b_g to every logit in G_g gives, exactly,*

$$Y(b) = \sum_{g=1}^{k} w_g(b)\mu_g, \qquad w_g(b) = \frac{\pi_g e^{b_g}}{\sum_{h=1}^{k} \pi_h e^{b_h}}. \tag{2}$$

Conversely every q in the relative interior Δ°_{k-1} of the $(k-1)$-simplex is attained by the bias $b_g = \log(q_g/\pi_g)$, unique up to one constant added to every group. Thus the coefficient space is exactly that relative interior, the reachable output set is exactly the relative interior of $\mathrm{conv}\{\mu_1, \dots, \mu_k\}$*, and its affine dimension is*

$$\mathrm{affdim}\{Y(b) : b \in \mathbb{R}^k\} = \dim \mathrm{span}\{\mu_g - \mu_1 : g > 1\} \le k-1. \tag{3}$$

Proof. Step 1 (the tilt leaves conditionals invariant; where the partition hypothesis enters). With unnormalised weights $a_j e^{b_g}$ for $j \in G_g$, regrouping the attention sum by group gives $Y(b) = \sum_g \left[\pi_g e^{b_g}/Z(b)\right] \sum_{j\in G_g} (a_j/\pi_g) u_j$ with $Z(b) = \sum_h \pi_h e^{b_h}$; the inner sum is μ_g, independent of b, which is (2). *Step 2 (the converse; routine).* Each $w_g(b) > 0$ and $\sum_g w_g(b) = 1$. Given q in the relative interior, $b_g = \log(q_g/\pi_g)$ gives $\pi_g e^{b_g} = q_g$ and $Z = 1$, so $w(b) = q$; $Z(b + c\mathbf{1}) = e^c Z(b)$ leaves w invariant, and conversely $w(b) = w(b')$ forces $e^{b'_g - b_g} = Z(b')/Z(b)$ for every g, so the fibre over q is exactly $b + c\mathbf{1}$. *Step 3 (the affine hull; analytical, the step the formalisation does not cover).* $Y(b) - \mu_1 = \sum_{g>1} w_g(b)(\mu_g - \mu_1)$. Since w is onto the relative interior of Δ_{k-1} and an affine map sends the relative interior of a convex set onto that of its image, the reachable set is exactly the relative interior of $\mathrm{conv}\{\mu_g\}$ (vertices are never attained at finite bias) and the first relation in (3) is an equality; the second is arithmetic, strict exactly when the centroids are affinely dependent. □

What this fixes, and what it leaves free, is as follows. The centroids depend on the values and on the unedited attention, so the reachable set is settled before any bias is chosen; the budget decides only where inside it the output lands. Pushing harder moves the output further from where it started and never enlarges the set. Refining the partition does enlarge it, but in a particular way: k groups expose at most $k-1$ directions, so a finer partition buys directions to move along rather than distance along the ones already there. The question a practitioner asks, whether this head can produce that change, is therefore a property of the head at this input, answered by one unedited pass rather than by a search over budgets.

The routing factor in (2) is classical: a groupwise bias is an exponential tilt of the k-atom categorical family with base measure π, so $b \mapsto w(b)$ is its natural-to-mean map (Amari, 2016), and the image of that simplex under the value map, the rank ceiling with it, is the family's mean-map image (Wainwright & Jordan, 2008); the closed-form bias is in Kang et al. (2026, App. A.1) and the $b = 0$ factorisation in Su et al. (2025), with further antecedents in Appendix L. What the value map adds is the payload contrast deciding whether a routing change transmits (Figure 1).

Corollary 2 (Binary span: routing × payload)**.** *Let S have mass $\psi \in (0, 1)$ and conditional value centroids $\mu_S, \mu_{\bar{S}}$. Adding $b \in \mathbb{R}$ to the logits in S gives*

$$\psi(b) = \frac{\psi e^b}{1 - \psi + \psi e^b}, \tag{4}$$

$$Y(b) - Y(0) = \big(\psi(b) - \psi\big)\big(\mu_S - \mu_{\bar{S}}\big), \tag{5}$$

so the displacement has norm $|\psi(b) - \psi| \, \|\mu_S - \mu_{\bar{S}}\|$. The finite path lies on the line through $Y(0)$ in direction $\mu_S - \mu_{\bar{S}}$, one-dimensional exactly when $\mu_S \neq \mu_{\bar{S}}$. The first-order gain is $\psi'(0) = \psi(1 - \psi) \leq \gamma_i |S|$, and reaching $\psi_\star$ costs $b = \log[\psi_\star(1 - \psi)/(\psi(1 - \psi_\star))]$.

Proof. Theorem 1 with $k = 2$ gives $w_S(b) = \psi(b)$ as in (4) and $Y(b) = \psi(b)\mu_S + (1 - \psi(b))\mu_{\bar{S}}$; subtracting $Y(0)$ gives (5). Differentiating (4) at $b = 0$ gives $\psi(1 - \psi) = \mathbf{1}_S^\top J_i \mathbf{1}_S$; inverting gives the cost. □

Corollary 3 (All-budget inertness)**.** *Under the hypotheses of Corollary 2, for every $b \neq 0$, $Y(b) = Y(0)$ if and only if $\mu_S = \mu_{\bar{S}}$: a vanishing contrast makes the control inert at every budget although $\psi(b)$ sweeps $(0, 1)$.*

Proof. By (5), since $\psi(b) - \psi = \psi(1-\psi)(e^b - 1)/(1 - \psi + \psi e^b) \neq 0$ for $b \neq 0$ and $\psi \in (0, 1)$. □

This is the constructive, exact form of attention non-identifiability (Brunner et al., 2020): the shift direction lies in the left nullspace of U exactly when the centroids coincide. Reading the same factorisation through a fixed linear readout φ (any linear function of the head output; the experiments use the gradient of an answer margin) replaces the vector condition by k scalars $c_g = \varphi(\mu_g)$: the readout's supremum is exactly $\sup_b \varphi(Y(b)) = \max_g c_g$, approached but not attained unless the c_g agree, a readout ceiling that one forward pass (the centroids) and one backward pass (the readout gradient) determine, and the readout is inert exactly when the c_g agree, which is strictly weaker than $\mu_S = \mu_{\bar{S}}$ (Corollary C.3).

Twin spans on one row. On the $m = 5$ row $\mathbf{A}_i = (0.28, 0.17, 0.26, 0.19, 0.10)$ with the value matrix of Appendix E, the spans $S_A = \{1, 2\}$ and $S_B = \{3, 4\}$ share the row, aperture, mass $\psi_A = \psi_B = 0.45$, size, first-order gain, budget and closed-form $\psi(b)$, and differ only in

$$\mu_{S_A} - \mu_{\bar{S}_A} = (0, 0), \qquad \mu_{S_B} - \mu_{\bar{S}_B} = -\tfrac{10}{99}(1, 17). \tag{6}$$

Span A is output-inert at every b while its mass runs over $(0, 1)$; span B moves the output by 0.95 at $b = 8$. Nothing in the attention row distinguishes them; the difference lives in U.

Control consequence. A target displacement outside span$\{\mu_g - \mu_1 : g > 1\}$ is unreachable at any budget: a statement about one head's immediate row output at a fixed input, not about model behaviour, which the all-layer methods below move through many such outputs. Inside that span, control reduces to a k-way softmax over the centroids, so a partition can be screened before optimization; a finer partition can enlarge the hull without raising the rank. Composition across spans, rows and heads is in Appendix C.

Where published span controls sit in this geometry. Corollary 3 sharpens the value-norm reading of Kobayashi et al. (2020): the contrast, not the norm, decides, and large-norm values with zero contrast stay inert at every budget. PASTA and SpotLight are (4) at biases of several nats, a bias being a log-odds shift (Appendix L), which the corollary requires: the first-order gain $\psi_i(1 - \psi_i)$ vanishes exactly on rows that ignore the span or attend to nothing else, and the first case is what steering is meant to fix.

	crosses the softmax	**does not**
activation-level	logit surgery, span steering	head-output shifts, activation addition, steering vectors, RepE
parameter-level	W_Q, W_K edits	$W_V W_O$ and MLP edits

Table 1: Intervention families by whether their causal path crosses the softmax, where Theorem 4 attaches the aperture.

What the geometry leaves reachable, a finite edit still has to pay for, at a cost set by how the intervention enters the head. Budget below is the Frobenius norm of the logit edit; the range B and the span bias b are other measures of the same edit (on row i a span bias costs $\|\delta\mathbf{L}_i\|_F = |b|\sqrt{|S|}$; Appendix F), and the three routes are the three interfaces of §2.

Theorem 4 (Softmax interface)**.** *Let $z \mapsto Y(z) = \mathrm{softmax}(\mathbf{L}(z))\,U(z) + \Delta(z)$ be a differentiable intervention on one head at a fixed input, z scalar. Then*

$$\partial_z Y \;=\; \mathcal{R} \circ \partial_z \mathbf{L} \;+\; \mathbf{A}\,\partial_z U \;+\; \partial_z \Delta, \qquad \|\mathcal{R}\|_{F\to F} = \max_i \sigma_1(J_i U) \;\le\; \gamma(\mathbf{A})\,\|U\|_{\mathrm{op}}, \quad (7)$$

where the response operator $\mathcal{R} : \delta\mathbf{L} \mapsto (J\,\delta\mathbf{L})\,U$ *acts row by row, σ_1 is the largest singular value and $\|\cdot\|_{F\to F}$ the Frobenius-induced norm. The aperture multiplies only the first term:*

*(i) (*Pre-softmax.*) If $\partial_z U = 0$ and $\partial_z \Delta = 0$, then*

$$\|\partial_z Y\|_F \;\le\; \|\mathcal{R}\|_{F\to F}\,\|\partial_z \mathbf{L}\|_F \;\le\; \gamma(\mathbf{A})\,\|U\|_{\mathrm{op}}\,\|\partial_z \mathbf{L}\|_F. \quad (8)$$

*(ii) (*Post-softmax.*) If $\partial_z \mathbf{L} = 0$, then $\|\partial_z Y\|_F \le \|\mathbf{A}\|_{\mathrm{op}}\|\partial_z U\|_F + \|\partial_z \Delta\|_F$: no factor $\gamma(\mathbf{A})$; the attention enters through $\|\mathbf{A}\|_{\mathrm{op}} \in [1, \sqrt{n}]$, not through its aperture, and a downstream weight edit $U(z) = X(P + \delta P(z))$ is the case $\|\mathbf{A}X\|_{\mathrm{op}}\,\|\partial_z \delta P\|_F$.*

Proof idea. The product rule in $(\mathbf{L}, U, \Delta)$, with row i of $\delta\mathbf{A}$ equal to $J_i\,\delta\mathbf{L}_i$; $\mathcal{R}$ is block diagonal over rows, and in the Frobenius norm, under which the blocks are orthogonal, its norm is the largest block norm $\max_i \sigma_1(J_i U) \le \gamma(\mathbf{A})\|U\|_{\mathrm{op}}$, attained by loading the budget on a maximising row $i^\star$ along the leading left singular vector of $J_{i^\star}U$. That identity, $\|\mathcal{R}\|_{F\to F} = \max_i \sigma_1(J_i U)$, fails in the spectral norm, where a row-wise action can gain up to $\sqrt{n}$. Clause (ii) is the triangle inequality with $\|\mathbf{A}\|_{\mathrm{op}}$. Both directions: Appendix D. □

Remark (The response can vanish). $\gamma(\mathbf{A}) > 0$ and $U \neq 0$ do not imply $\mathcal{R} \neq 0$: identical rows of U are annihilated by every J_i, leaving every first-order direction inert.

Table 1 sorts Section 2's methods by the decomposition rather than by the usual activation-versus-parameter axis (Ostermann et al., 2026): both rows split. An edit to the head input X moves two terms of (7), since $\mathbf{L}$ and $U = XP$ both depend on it; a residual-stream edit leaves the heads before it untouched and reaches those after it by both routes at once.

A bound frozen at the base point, $\gamma(\mathbf{A})\,\|\delta\mathbf{L}\|_F$, need not hold on a norm ball: a dispersing perturbation of a concentrated row raises the Jacobian norm along the path, so the base bound can underestimate the displacement, which §5 measures. The finite statement integrates the aperture along the path, as Integrated Gradients does for a base-point gradient (Sundararajan et al., 2017).

Corollary 5 (Finite path aperture)**.** *Let $\mathbf{L}_t = \mathbf{L}_0 + t\,\delta\mathbf{L}$, $\mathbf{A}_t = \mathrm{softmax}(\mathbf{L}_t)$, and $Y_t = \mathbf{A}_t U$ with U fixed. Writing $\mathcal{R}_t(H) = (J_t H)U$ and*

$$\Gamma_{\mathrm{path}}(\mathbf{L}_0, \delta\mathbf{L}) := \int_0^1 \gamma(\mathbf{A}_t)\,dt, \quad (9)$$

the finite displacement satisfies

$$Y_1 - Y_0 = \int_0^1 \mathcal{R}_t(\delta\mathbf{L})\,dt, \;\; \|Y_1 - Y_0\|_F \le \|\delta\mathbf{L}\|_F \int_0^1 \|\mathcal{R}_t\|_{F\to F}\,dt \le \|U\|_{\mathrm{op}}\|\delta\mathbf{L}\|_F \Gamma_{\mathrm{path}}(\mathbf{L}_0, \delta\mathbf{L}), \quad (10)$$

all three integrands being continuous in t, so the integrals exist.

Proof idea. Integrate the product rule along the path and bound $\|\mathcal{R}_t\|_{F\to F}$ by $\gamma(\mathbf{A}_t)\|U\|_{\text{op}}$. The same integration on a general path, with U and Δ also moving, gives a three-term bound (Appendix B). □

Reach and cost answer different questions, and they are computed differently. Reach is exact and budget-free: the centroids settle it from one unedited pass, and no bias brings a target inside a hull it was never in. Cost is local and is paid along the path: the aperture bounds the first-order response at a point, and a finite edit pays its integral. Neither is visible in the attention row, and the two fail independently, a control being unreachable, unaffordable, or both. The next section makes the second precise, bounding the aperture by the same distance from uniform that makes a row distinctive.

4 The Aperture: A Forbidden Region

A row that places nearly all of its mass on one key is the most distinctive row by Euclidean distance from uniform and also nearly frozen, since its Jacobian vanishes. Two upper bounds on $\|J_i\|_{\text{op}}$, its envelopes,

$$\|J_i\|_{\text{op}} \;\le\; \operatorname{tr} J_i \;=\; 1-\|\mathbf{A}_i\|_2^2 \;=\; 1-\tfrac{1}{m}-d_i^2, \qquad \text{and} \qquad \|J_i\|_{\text{op}} \;\le\; \tfrac{1}{2}, \tag{11}$$

the first an equality exactly when $\mathbf{A}_i$ has at most two nonzero entries (Lemma B.1), turn this into an exact constraint on the pair (d_i, γ_i). Both are classical: the first from the Jacobian's positive semidefiniteness (Gao & Pavel, 2017), the second a Popoviciu variance bound (Popoviciu, 1935) sharp in every ℓ_p (Nair, 2026). What is new is the coordinate: read in d_i the trace envelope becomes a curve through the $(d_i, \|J_i\|_{\text{op}})$ plane, tight along the two-key family and unmoved by a sharper row-adaptive envelope (Yudin et al., 2025).

Theorem 6 (Softmax Aperture Bound)**.** *For every probability row in the closed simplex on $m \ge 2$ coordinates (masked zeros allowed), with d_i and $C(m)$ computed in the same m coordinates,*

$$d_i\,\|J_i\|_{\text{op}} \;\le\; C(m) := \max_{0\le d\le\sqrt{1-1/m}} d\,\min\Big\{\tfrac{1}{2},\; 1-\tfrac{1}{m}-d^2\Big\} \;\le\; \tfrac{\sqrt{2}}{4} \approx 0.354. \tag{12}$$

For $m \ge 4$, $C(m) = \frac{1}{2}\sqrt{(m-2)/(2m)}$, with equality only at balanced two-key rows, which a masked row attains and a strictly positive row only approaches. At $m = 2$ and $m = 3$ the constant differs; Table I.1 lists every case.

Proof idea. Maximizing $d\min\{1/2, 1-1/m-d^2\}$ puts the $m \ge 4$ optimum at the crossing $d^2 = 1/2 - 1/m$, saturated by the balanced two-key row; for $m \in \{2, 3\}$ the cubic branch peaks later (both branches: Appendix I). □

Figure 2 draws the feasible region at $m = n = 128$: cap, trace envelope, and the *frontier* $d\,\|J\|_{\text{op}} = \sqrt{2}/4$, which $C(m)$ approaches from below and the two-key extremiser attains only as m grows. Near uniform the product is small because d_i is, near determinism because the trace envelope forces $\gamma_i \to 0$; the aperture is not monotone in d_i (the uniform row has $\gamma_i = 1/m$), so this is a forbidden region, not an anticorrelation.

In the concentrated regime the collapse is linear: at $a_{i,\max} = 1-\varepsilon$ with $\varepsilon \le \frac{1}{2}$ the aperture is at most $2\varepsilon(1-\varepsilon)$ while d_i stays within $O(\varepsilon)$ of its maximum (Corollary B.3). The constant does not soften under a sharper envelope (Appendix I).

The aperture γ_i is an upper bound, not a point predictor: a direction realises only a fraction of it, its utilization, and value-aligned directions collect far more of it than isotropic ones, so a small aperture obstructs every logit direction while a large one helps only an aligned direction (Appendix H). A finite perturbation pays the aperture integrated along its path (Corollary 5). Composition across heads and layers is in Appendix A.

5 Tests of the Theory

Design. Two claims carry the paper: reach is fixed by the values before any budget is spent, and distinctiveness is paid for in responsiveness. Each arm varies what one of them says should matter

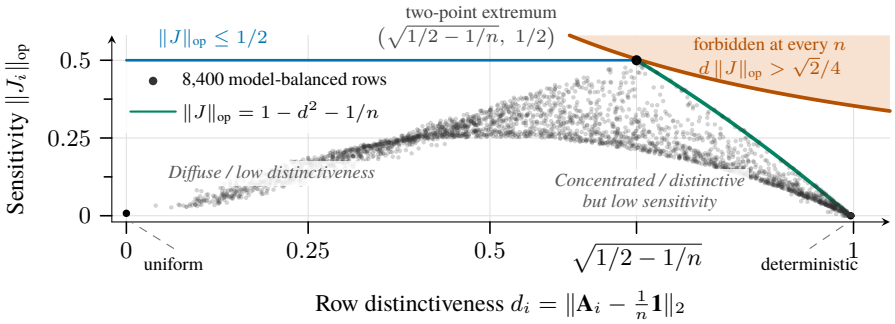

Figure 2: **The forbidden region.** Black marks are a seeded, equal-per-model sample of 8,400 trained attention rows from the 70.4M rows measured across 7 model variants, in each row's ambient sequence coordinates ($n \le 512$), masked zeros included. The shaded region $d\|J\|_{\text{op}} > \sqrt{2}/4$ is forbidden at every width; the green curve is the trace envelope at one width, $n = 128$, so rows of this variable-width sample with larger n can sit just above it. The boundary forces local logit sensitivity to vanish near determinism: a forbidden region, not a global anticorrelation. Ambient coordinates do not separate trained from random weights; Table E.1 is the support-aware comparison that does.

Arm	Tests	Models
reachable-set install (App. E)	Thm. 1, Cor. 3	OLMo-2, Qwen2.5
steering vs. bypass (App. F)	Thm. 4, Cor. 2	GPT-2, Qwen2, SmolLM2
path trajectories (App. B)	Cor. 5	GPT-2
saturation sweep (App. E)	Thm. 6	7 trained, untrained, ckpts
span choice (App. G)	contrast as ex-ante rule	Qwen2, SmolLM2, OLMo-2, Qwen2.5
context control (App. J)	centroid-aligned selection	OLMo-2, Qwen2.5
finite response (App. K)	finite law vs. derivative	Qwen2.5
logit perturbation (App. H)	direction dependence	GPT-2, Pythia, Qwen2, SmolLM2

Table 2: The evidence arms in the order the theory states them, then the three that apply it, each with the claim it tests and the appendix holding its protocol, which also gives its data and readout; model names are shortened to the family. The last arm is reported in its appendix alone.

and holds the rest fixed (Table 2). Every checkpoint is public and is evaluated without training, on WikiText-103 (Merity et al., 2017), HaluEval-QA (Li et al., 2023a), TruthfulQA (Lin et al., 2022) and NQ-Swap (Longpre et al., 2021). Sweep ranges and untrained controls are in each arm's appendix, and which arms were registered in advance in Appendix M.

What the geometry forbids stays forbidden at the head output. Installing the bias $b_g = \log(q_g/\pi_g)$ of Theorem 1 in a live forward lands the target routing and the predicted output to float32 tolerance, and, splitting a readout direction into its reachable part, in the centroid-contrast span, and the orthogonal remainder, mean orthogonal movement is smaller than mean reachable movement by a factor of 1.5×10^7 (OLMo-2) and 1.7×10^7 (Qwen2.5-3B) at $|b| \le 50$, the budget that drives the group masses to the simplex boundary. The theorem predicts exactly zero there. Of the registered criteria, the one on reach asks only that the orthogonal error not shrink as the budget grows, and it is met: the error grows with the bias magnitude, at the float32 scale of the run rather than the scale of reach; the two on rank and on overshooting the ceiling fail and stand as failures (Appendix E). Budget buys nothing in the forbidden direction: Corollary C.3 holds on the head output up to that overshoot, in a network to first order through the stack, and the ceiling's own ranking of which answers an edit flips carries measured content only within about one block after the edited layer, on the one model swept for depth (Appendices E and J). Routing and payload also separate directly: across 234 gain-matched records on GPT-2 (base span mass at most 0.95) the bias of Corollary 2 holds the routing gain at 0.05 to within 5.97e-16 while the realised output change still spreads by 2.3× between its 10th and 90th percentiles, by (5) the spread of the payload contrast across real spans; on 864 fixed-budget records routing gain alone ranks that change with Spearman $\rho = 0.983$ and payload alone with $\rho = 0.289$, and neither suffices (Appendix E).

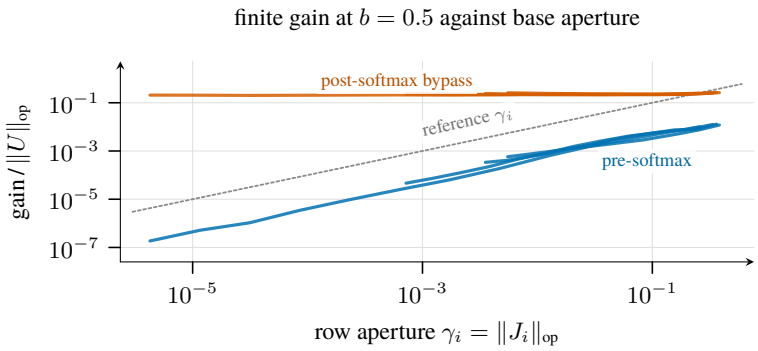

Figure 3: **Empirical interface separation under span steering.** A constant logit shift b on a span against a matched post-softmax bypass of the same head-output displacement, over 600 samples per model, three datasets and every head of one layer; medians in twelve logarithmic aperture bins. Each colour bundles four models (GPT-2, Qwen2-0.5B, Qwen2-1.5B, SmolLM2-1.7B), whose curves the bundle width shows to agree; the per-model slopes are in Table F.1. At $b = 0.5$ finite pre-softmax gain tracks the base-point aperture and the matched bypass does not. The diagonal is a base-point reference, not a finite-budget guarantee, which integrates the aperture along the path (Corollary 5).

The two interfaces separate under real steering. Adding a constant b to a span's logits is the primitive of Corollary 2; the control is a matched post-softmax bypass that adds the realised per-head change with the attention weights untouched, so only the pathway differs. Over 600 samples per model, the measured mass agrees with its closed form to at most 3.5e-07. At $b = 0.5$ the binned median pre-softmax gain rises with the base-point aperture, with descriptive log–log slope 0.563–0.987, whereas the matched bypass is flat (slope at most 0.017; Figure 3); the regressions pool correlated rows and are descriptive, not inferential (Appendix F, Table F.1).

Path aperture closes the finite-budget gap. Across 1728 real GPT-2 trajectories Corollary 5's integrated bound has zero violations, while the frozen base aperture fails on 29.46% of paths and underestimates the realised displacement by up to 22.53× (Appendix B).

Trained rows sit nearer the frontier than initialization. No measured ambient row violates $d_i\gamma_i \leq C(n)$, and every model, trained or untrained, reaches that bound to reported precision, because early causal masks supply the balanced two-coordinate equality case; only the support-aware distribution separates learning from mask structure. The *saturation* $s_i = d_i^{\Sigma}\gamma_i / C(m_i) \in [0, 1]$ on a row's m_i unmasked keys is the fraction of the sharp bound it uses in its own coordinates. Rows at position ≥ 8 with non-initial argmax key reach 99th-percentile $s_i \in [0.799, 0.949]$ in trained models versus ≤ 0.193 untrained, and it rises over training in both Pythia models (Appendix E). Typical rows sit well below the frontier (pooled median saturation 0.338): trained models populate it where attention is distinctive, not wholesale (Figure 2, Table E.1).

What the contrast ranks, and what it does not beat. On 256 HaluEval-QA examples per model the evidence passage and the question are each steered at the all-layer bias $b = 1.0$, and P_{contrast}, the routing-mass gain of (4) against the payload contrast summed at the final query row, exact in routing and first order downstream, ranks the two spans by realised gain in the margin, the logit gap between the answer options. It is above chance on every model and above span mass on 2 of the 3 registered models and on the 3B replication arm, and it does not beat always steering the question, whose paired interval covers zero: a score read at one row ranks spans and does not choose between them (Appendix G).

Where the budget goes decides what it buys. If a control acts through centroids, budget on heads whose contrast aligns with the readout should transmit more than budget spread elsewhere. It does. NQ-Swap edits its passages to contradict the model's memorised answer, so a context-faithful answer is distinguishable from a memorised one. Editing the passage span at one layer, the subset of heads whose centroid contrast aligns with the readout gradient beats the uniform edit at that layer

in context-faithful accuracy by $+4.02$ [$+2.26$, $+6.03$] and $+3.02$ [$+1.51$, $+4.77$] percentage points (95% example-level bootstrap intervals), moving the head output less. Holding scope fixed and varying only the selection rule, it also beats PASTA's profiling-based head ranking (Zhang et al., 2024a) inside that layer at the same head count and matched mass transfer, by $+3.52$ [$+1.51$, $+5.78$] and $+2.76$ [$+1.26$, $+4.52$] pp; 3 of the 4 comparisons, two models by solved dose or matched transfer, separate from zero; the exception and the counts of records that differ are in Appendix J. The advantage stops at the selection rule. Against heads chosen by the all-row derivative, the derivative of the margin with respect to the bias taken by autodiff through every query row the bias touches, contrast selection establishes none: $+0.25$ [-0.50, $+1.26$] pp on Qwen2.5-3B and the same answer from both on every OLMo-2 record (Appendix K).

Finite routing predicts what the zero-bias derivative misranks. The derivative is the better local direction and the worse finite one. On Qwen2.5-3B, where four attention layers follow the edited one, it is steeper at zero bias on 382 of 398 development records yet realises the smaller gain at the operating budget on 256 of them, so a zero-bias linear prediction misorders most pairs there. The two selectors pick different heads for a different reason: the all-row derivative counts query rows the final-row score P_{contrast} omits (Appendix K). Replacing the linearised routing by the exact finite law (4) at every affected query row, with the downstream gradient held at its base-point value, recovers most of the reversal. On a pre-registered held-out split of 398 records sharing no question with development, it cuts the paired-difference error, predicted minus realised, from 0.350 to 0.191 logits of margin and orders the pairs at 77.1% against 60.2% for the constant choice, always the contrast plan fixed on development, and, below chance, 39.8% for the linear prediction; the paired intervals are in Appendix K. The routing is exact and the downstream part first order, so this gain is the routing correction's, on one arm, one model, one layer, one budget and one task.

6 Related Work

Self-attention is not globally Lipschitz (Kim et al., 2021) and the sharp $1/2$ softmax cap holds in every ℓ_p norm (Nair, 2026); Yudin et al. (2025) read a row-adaptive Jacobian bound as an exclusion on the top mass, where we couple the norm to the distance from uniform and propagate the result along a path. Linear-response predictors of steering effects (Billa, 2026; Olivieri & Pérez Rodríguez, 2026; Piontkovskaia & Nikolenko, 2026) linearise the response at the base point; §5 measures what that costs at the budget. PASTA (Zhang et al., 2024a), AutoPASTA (Zhang et al., 2024b), SpotLight (Venkateswaran & Contractor, 2026) and InstABoost (Guardieiro et al., 2025) all implement Corollary 2's odds shift; we contribute the ex-ante certificate and the bypass control.

The routing factor's tilt geometry, classical and attributed in §3, is steered over the vocabulary by Park et al. (2026); ours is not the tilt but what the value map does to it: a payload contrast that decides whether a routing change transmits, and with it a reachable set and a rank ceiling. That puts Corollary 3 against attention non-identifiability, where Brunner et al. (2020) construct alternative attention rows in the left nullspace of the value map, and Sun & Marasović (2021) compute the surviving component and read it against the raw weights. Theirs is an existence result, generic once the sequence outruns the head width; ours is a criterion, $\mu_S = \mu_{\bar{S}}$, over the rows a bounded span bias can reach, and it never softens with budget (Appendix L).

7 Conclusion

Attention is a control surface whose reach and cost are computable before an edit is run, and whose effect the routing law predicted on the one model tested for it. Reach and cost together make attention a poor guide to itself: neither is legible in the pattern, the values fix the first, and sharpness buys distinctiveness at the price of the second. The rows that read as most interpretable are the ones a logit edit can move least. Reach belongs to the centroids and not to how hard the routing is pushed, k groupwise biases exposing at most $k-1$ row-output directions while the path sets the cost, so screen a control first: one unedited pass identifies the centroids and compares a row-output target with their hull. On NQ-Swap, centroid-contrast selection beats PASTA's selector at matched scope under matched transfer and, on Qwen2.5-3B, does not beat the all-row derivative: the derivative is the better local direction and the worse finite one. At the budgets steering spends, compute the routing change, not its slope at zero.

AI use statement

Generative AI tools (large language models) were used in this work in the following roles: prose editing and drafting; proposing and refining hypotheses and candidate theorem statements; drafting experimental designs, including the protocols registered in advance, and critiquing interpretations of results, with every design and interpretation decided by the authors; exploration of candidate proof strategies and drafting of proofs, each subsequently checked line by line by the authors; stress-testing of statements against candidate counterexamples; visualization and data-processing code; experimental analysis code, including generating and running the numerical experiments, with outputs reproduced deterministically by the accompanying scripts; and assistance with the formal (Lean) verification development. Pretrained transformers are additionally the *subject* of study in this paper, not merely a drafting aid. The authors reviewed and take responsibility for every theorem statement, proof, experiment, figure, and citation in this work. The aperture bound and the interface theorem are machine-checked in a proof assistant against the actual softmax Jacobian, which provides verification independent of any drafting assistance; Appendix M states which downstream constants enter that development as hypotheses rather than being derived in it.

Ethics statement

This work analyses publicly available pretrained models and releases no new models or datasets. Its practical content is a screening tool: the aperture is computable before an intervention is attempted and certifies, in exact arithmetic, when a pre-softmax edit cannot achieve a target (floating-point realizations of the screen can produce false exclusions), which serves interpretability and evaluation work more than it serves an attacker. The results do not reduce the effort required for weight-level attacks, which the paper shows are governed by a different interface entirely, and the steering experiments use published primitives on public benchmarks without introducing new capability.

Reproducibility statement

The experimental protocol (models, data, and evaluation) is summarised in §5, with full details, compute environment, and seed and tolerance settings in Appendix E and Appendix M. All models are publicly available, permissively licensed pretrained checkpoints evaluated on public data; no training is performed. Every theorem in the main text carries an explicit hypothesis block; proofs appear in the main text or in the appendix, starting at Appendix A. The headline theorems are additionally machine-verified in Lean 4 / Mathlib (zero `sorry`; axioms restricted to `propext`, `Classical.choice`, and `Quot.sound`), with the correspondence listed in Appendix M. Which experiments were registered in advance, and which two criteria were added after a first outcome had been read, is stated in Appendix M. An anonymised copy of the experimental pipeline, the frozen protocols and the Lean formalisation is available for review at `https://anonymous.4open.science/r/softmax-aperture-6D50`; it will be released on a permanent public repository upon acceptance.

References

Dyah Adila, John Cooper, Alexander Yun, Avi Trost, and Frederic Sala. Weight updates as activation shifts: A principled framework for steering. *arXiv preprint arXiv:2603.00425*, 2026.

Faruk Alpay and Taylan Alpay. Manipulating transformer-based models: Controllability, steerability, and robust interventions. *arXiv preprint arXiv:2509.04549*, 2025.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, 2016.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference*

on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Garrett Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85:219–227, 1957. doi: 10.1090/S0002-9947-1957-0087058-6.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. arXiv:1908.04211.

P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52:330–338, 1973. doi: 10.1007/BF00247467.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840. PMLR, 2024.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pp. 625–635. Springer, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. Anthropic technical report.

Seyed Morteza Emadi. Exact attention sensitivity and the geometry of transformer stability. *arXiv preprint arXiv:2602.18849*, 2026.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Joshua Goodman. Classes for fast maximum entropy training. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001. arXiv:cs/0108006.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction following by principled boosting attention of large language models. *arXiv preprint arXiv:2506.13734*, 2025.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668, 2023. doi: 10.52202/075280-0774.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

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*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357.

Dongkyung Kang, Jaeyeon Hwang, Junseo Park, Minji Kang, Yeryeong Lee, Beomseok Ko, Hanyoung Roh, Jeongmin Shin, et al. MAST: Mask-guided attention control for training-free regional-multi style transfer. *arXiv preprint arXiv:2604.12281*, 2026.

Anders Karlsson, Volker Metz, and Gennady A. Noskov. Horoballs in simplices and Minkowski spaces. *International Journal of Mathematics and Mathematical Sciences*, 2006(1):1–20, 2006. doi: 10.1155/IJMMS/2006/23656.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

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*, pp. 7057–7075. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.574.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6449–6464, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36*, pp. 41451–41530. Curran Associates, Inc., 2023b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023c. Phi-2 extends Phi-1.5; checkpoint: `microsoft/phi-2`.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, 2nd edition, 2011. ISBN 978-0-387-40087-7. doi: 10.1007/978-0-387-68276-1.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Erik Miehling, Karthikeyan Natesan Ramamurthy, Praveen Venkateswaran, Ching-Yun Ko, Pierre Dognin, Moninder Singh, Tejaswini Pedapati, Avinash Balakrishnan, et al. AI steerability 360: A toolkit for steering large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 436–444, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-demo.43.

Pravin Nair. Softmax is $1/2$-Lipschitz: A tight bound across all ℓ_p norms. *Transactions on Machine Learning Research*, 2026. arXiv:2510.23012.

David N. Olivieri and Antonio F. Pérez Rodríguez. Transformer field theory: a response-theoretic approach to mechanistic interpretability. *arXiv preprint arXiv:2605.25225*, 2026.

Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. From weights to activations: is steering the next frontier of adaptation? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29854–29879. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1377.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. arXiv:2311.03658.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. arXiv:2602.15293.

Irina Piontkovskaia and Sergey Nikolenko. First-order predictable but pairwise fragile: local task adaptation in trained transformers. *arXiv preprint arXiv:2607.16821*, 2026.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9: 129–145, 1935.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Navid Rezazadeh and Arash Gholami Davoodi. Vertex-softmax: Tight transformer verification via exact softmax optimization. *arXiv preprint arXiv:2605.10974*, 2026.

Yi Su, Quantong Qiu, Yuechi Zhou, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, et al. CaliDrop: KV cache compression with calibration. *arXiv preprint arXiv:2507.19906*, 2025.

Chung-En Sun, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Steer2Edit: From activation steering to component-level editing. *arXiv preprint arXiv:2602.09870*, 2026.

Kaiser Sun and Ana Marasović. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4126–4135. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.361.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017. arXiv:1703.01365.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Praveen Venkateswaran and Danish Contractor. Spotlight your instructions: Instruction-following with dynamic attention steering. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3752–3770. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.174.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

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*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Nikolay Yudin, Sergei Kudriashov, Alexander Gaponov, and Maxim Rakhuba. Pay attention to attention distribution: A new local Lipschitz bound for transformers. *arXiv preprint arXiv:2507.07814*, 2025.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *International Conference on Learning Representations*, 2024a.

Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, Jianfeng Gao, Tuo Zhao, Dan Roth, et al. Model tells itself where to attend: Faithfulness meets automatic attention steering. *arXiv preprint arXiv:2409.10790*, 2024b.

Günter M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, 1995. doi: 10.1007/978-1-4613-8431-1.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Reading guide. *Proofs.* Theorem 1 and Corollaries 2 and 3 are proved in §3; their extension to many rows and heads, Propositions C.1 and C.2 and the readout ceiling (Corollary C.3), in Appendix C; Theorem 4's norm identity in Appendix D; Corollary 5, the two envelopes (Lemma B.1), Lemma B.2 and Corollary B.3 in Appendix B; Theorem 6's maximisation over both branches, its small-m constants and the envelope comparison in Appendix I; the multi-layer bound in Appendix A. *Experiments* for §5, by appendix: the reachable-set install, its registered criteria, the gain-matched records and the saturation sweep in Appendix E; span steering against the bypass in Appendix F; path trajectories in Appendix B; span choice in Appendix G; NQ-Swap context control and the depth sweep in Appendix J; the selector comparison and the finite-response predictor in Appendix K; logit-space perturbations in Appendix H. What was registered in advance, compute, and the Lean correspondence are in Appendix M; extended related work is in Appendix L.

A L-Layer Composition Bound: Proof

Consider L transformer blocks, each with attention sublayer Jacobian norm $c_a^{(\ell)}$ and MLP Jacobian norm $c_m^{(\ell)}$ at the given input, per-head value-output operator $U_\ell = V_\ell W_O^{(\ell)}$, per-head aperture $\gamma^{(\ell)} = \max_i \|J_i^{(\ell)}\|_{\text{op}}$, and forward-propagation coefficient

$$x_\ell = \gamma^{(\ell)} \|U_\ell\|_{\text{op}} \, (1 + c_m^{(\ell)}) \prod_{j=\ell+1}^{L} (1 + c_a^{(j)})(1 + c_m^{(j)}), \tag{13}$$

the aperture-limited reach of a layer-ℓ logit intervention after first-order propagation through the subsequent blocks.

Across H heads the per-head bounds of §3 add,

$$\|\delta Y\|_F \;\le\; \sum_{h=1}^{H} \gamma(\mathbf{A}^{(h)}) \, \|U^{(h)}\|_{\text{op}} \, \|\delta \mathbf{L}^{(h)}\|_F. \tag{14}$$

Proposition A.1 (L-layer composition bound)**.** *Any coordinated pre-softmax intervention distributing total logit budget $\sum_\ell \|\delta \mathbf{L}^{(\ell)}\|_F^2 = B^2$ across the L blocks produces a first-order final-block output change of at most*

$$\|\delta Y^{(L)}\|_F \;\le\; B \sqrt{\sum_{\ell=1}^{L} x_\ell^2}. \tag{15}$$

Equality requires the budget to be allocated in proportion to x_ℓ and the propagated per-layer contributions to be aligned at the final block. Uniform coefficients give $\sqrt{L}$ scaling; an expansive stack ($\kappa = (1 + c_a)(1 + c_m) > 1$ per block) is dominated by the earliest layer through $\kappa^{L-\ell}$; $L = 1$ gives the single-head bound times the edited block's own $(1 + c_m^{(1)})$, the factor a head-output statement does not carry. The coefficient treats one head at a time and places the normalisation with the MLP; combining heads within a block, and architectures that normalise elsewhere, are not covered by this statement, which is among those listed as unformalised.

Proof. Three inequalities, none sharp in general. *Per layer:* Theorem 4(i) with the multi-head bound (14) controls the local change at block ℓ, $\|\delta \mathbf{A}^{(\ell)} U_\ell\|_F \le \gamma^{(\ell)} \|U_\ell\|_{\text{op}} \, \|\delta \mathbf{L}^{(\ell)}\|_F$. *Propagation:* the edit enters at block ℓ's attention and still passes that block's own MLP and residual, contributing $(1 + c_m^{(\ell)})$ before any later block acts; thereafter each residual block j is $(1 + c_a^{(j)})(1 + c_m^{(j)})$-Lipschitz in the Euclidean norm of the residual stream (the Frobenius norm over tokens), and Lipschitz constants multiply under composition, so the layer-ℓ contribution to the final output obeys

$$\|\delta Y^{(L)}_{\ell\text{-from}}\|_F \;\le\; \|\delta \mathbf{A}^{(\ell)} U_\ell\|_F \, (1 + c_m^{(\ell)}) \prod_{j=\ell+1}^{L} (1 + c_a^{(j)})(1 + c_m^{(j)}) \;\le\; x_\ell \, \|\delta \mathbf{L}^{(\ell)}\|_F. \tag{16}$$

Aggregation: the L contributions live in one Frobenius space, so the triangle inequality followed by Cauchy–Schwarz on $(x_\ell)_\ell$ against $(\|\delta \mathbf{L}^{(\ell)}\|_F)_\ell$ gives $\|\delta Y^{(L)}\|_F \le \sum_\ell x_\ell \|\delta \mathbf{L}^{(\ell)}\|_F \le B \sqrt{\sum_\ell x_\ell^2}$. Cauchy–Schwarz is tight iff $\|\delta \mathbf{L}^{(\ell)}\|_F \propto x_\ell$, and the triangle inequality iff the propagated contributions

are positively aligned at the final block, which no structure guarantees, so the bound is approached rather than generically attained. For $L = 2$ with uniform κ and $\gamma^{(1)}\|U_1\|_{\text{op}} = \gamma^{(2)}\|U_2\|_{\text{op}} = \alpha$ it reads $\alpha(1 + c_m)\sqrt{\kappa^2 + 1}\, B$. □

B Finite-Perturbation Extension

Corollary 5 is the primary finite-budget statement: it integrates the actual aperture along the intervention trajectory. This appendix supplies two useful corollaries, the uniform $1/2$ fallback and the exact span-shift response, and states precisely why a *base-point* aperture cannot be extrapolated over an arbitrary finite trajectory. The distinction matters: a case in which $a_{i,\max} = 1 - \epsilon_0$ for tiny ϵ_0 is perturbed with $\|\delta\mathbf{L}\|$ large enough to drive $a_{i,\max}$ across the concentrated/dispersed boundary does not violate any bound below; it merely exits the regime in which the base-point aperture is the informative quantity.

On a general path, with $Y_t = \mathbf{A}_t U_t + \Delta_t$ differentiable on $[0, 1]$ and the velocity terms integrable, the product rule and its integral give

$$\dot{Y}_t = \mathcal{R}_t(\dot{\mathbf{L}}_t) + \mathbf{A}_t\dot{U}_t + \dot{\Delta}_t, \quad \|Y_1 - Y_0\|_F \le \int_0^1 \Big(\|\mathcal{R}_t(\dot{\mathbf{L}}_t)\|_F + \|\mathbf{A}_t\dot{U}_t\|_F + \|\dot{\Delta}_t\|_F\Big)\, dt. \quad (17)$$

Proof of Corollary 5 (finite path aperture). The velocity identity in (17) is the product rule along a path differentiable on $[0, 1]$; integrating it, with each of the three velocity terms integrable, gives the bound. On the affine logit path $t \mapsto \mathbf{A}_t$ is smooth and $A \mapsto \sigma_1(A)$ is 1-Lipschitz, so each of the finitely many $t \mapsto \sigma_1(J_{i,t}U)$ and $t \mapsto \|J_{i,t}\|_{\text{op}}$ is continuous, hence so is their maximum, which discharges the integrability there. Further, on the affine logit path the velocity of $\mathbf{A}_t U$ is $\mathcal{R}_t(\delta\mathbf{L})$ (chain rule), and the norm-of-integral bound with $\|\mathcal{R}_t\|_{F\to F} \le \gamma(\mathbf{A}_t)\|U\|_{\text{op}}$ gives (10). □

Lemma B.1 (Classical envelopes)**.** *For a probability vector $\mathbf{A}_i$ on m keys with Jacobian J_i, the two envelopes of* (11) *hold, and the first is an equality exactly when $\mathbf{A}_i$ has at most two nonzero entries.*

Proof. Positive semidefiniteness gives $\|J_i\|_{\text{op}} \le \operatorname{tr} J_i$, with equality exactly at rank at most one; the second envelope is the sharp variance bound (Popoviciu, 1935) for $v^\top J_i v = \operatorname{Var}_{\mathbf{A}_i}(v)$. □

Lemma B.2 (Jacobian variance bound; classical)**.** *For every row $\mathbf{A}_i$ with $p = a_{i,\max}$, $\|J_i\|_{\text{op}} \le 2p(1-p) \le \frac{1}{2}$; for $p \ge \frac{1}{2}$ the first inequality is an equality exactly at the two-point rows. The bound is a weighted-variance inequality of Popoviciu type (Popoviciu, 1935); in its softmax-specific form it is implied by the sharper estimate $\|J_i\|_{\text{op}} \le a_{(1)}(1 - a_{(1)} + a_{(2)})$ of Yudin et al. (2025), with the same equality case (Appendix I).*

Proof. With $a = \mathbf{A}_i$ and $\sum_k a_k = 1$, the quadratic form is the weighted variance in its pairwise form (cf. Marshall et al., 2011), $v^\top J_i v = \sum_k a_k v_k^2 - (\sum_k a_k v_k)^2 = \sum_{r<s} a_r a_s (v_r - v_s)^2$. By $(x-y)^2 \le 2x^2+2y^2$ this is at most $2\sum_r a_r(1-a_r)v_r^2 \le 2p(1-p)\|v\|_2^2$, because $a_r(1-a_r) \le p(1-p)$ for every r: for $p \le \frac{1}{2}$ by monotonicity of $x(1-x)$ on $[0, \frac{1}{2}]$, and for $p \ge \frac{1}{2}$ because every other coordinate satisfies $a_r \le 1 - p \le \frac{1}{2}$. At the two-point row $(p, 1-p, 0, \dots, 0)$ the unit vector $(1, -1, 0, \dots, 0)/\sqrt{2}$ attains $2p(1-p)$. □

Corollary B.3 (Collapse in the concentrated regime)**.** *If $a_{i,\max} = 1 - \varepsilon$ with $\varepsilon \le \frac{1}{2}$, then $\|J_i\|_{\text{op}} \le 2\varepsilon(1-\varepsilon)$ while $d_i^2 \ge (1-\varepsilon)^2 - 1/m$: the row distinctiveness stays within $O(\varepsilon)$ of its maximum $\sqrt{1-1/m}$ while the aperture vanishes linearly, so $d_i\,\|J_i\|_{\text{op}} \le 2\varepsilon\sqrt{1-1/m}$. The bound $2\varepsilon(1-\varepsilon)$ is attained by the two-key row $(1-\varepsilon, \varepsilon, 0, \dots, 0)$.*

Proof of Corollary B.3 (linear collapse). Substitute $a_{i,\max} = 1-\varepsilon$ with $\varepsilon \in [0, 1/2]$ into the variance bound of Lemma B.2: $\|J_i\|_{\text{op}} \le 2(1-\varepsilon)\varepsilon \le 2\varepsilon$. The deviation satisfies $\sqrt{(1-\varepsilon)^2 - 1/m} \le d_i \le \sqrt{1-1/m}$ (the upper bound is the deterministic row $(1, 0, \dots, 0)$, the lower bound uses $\|\mathbf{A}_i\|_2^2 \ge a_{i,\max}^2$), so $d_i \cdot \|J_i\|_{\text{op}} \le 2\varepsilon\sqrt{1-1/m}$: the sensitivity vanishes as 2ε while the row distinctiveness stays within $O(\varepsilon)$ of its maximum, and the product vanishes linearly with ε. □

Proposition B.4 (Uniform $1/2$-Lipschitz softmax; classical)**.** *For any $\mathbf{L}, \delta\mathbf{L} \in \mathbb{R}^n$ and any $\varepsilon \geq 0$,*

$$\big\|\mathrm{softmax}(\mathbf{L} + \varepsilon\,\delta\mathbf{L}) - \mathrm{softmax}(\mathbf{L})\big\|_2 \;\leq\; \frac{1}{2}\,\varepsilon\,\|\delta\mathbf{L}\|_2. \tag{18}$$

Equivalently, as a map $\mathbb{R}^n \to \Delta_{n-1}$, softmax is $(1/2)$-Lipschitz in ℓ^2 uniformly in the starting point. The constant is not new: Nair (2026) sharpens Gao & Pavel (2017)'s bound and proves it holds uniformly over ℓ_p. It is restated here because the finite-perturbation argument uses it in this uniform form; it follows from Lemma B.2, whose cap $\|J\|_{\mathrm{op}} \leq \frac{1}{2}$ holds at every base point, by integrating the Jacobian along the segment from $\mathbf{L}$ to $\mathbf{L} + \varepsilon\,\delta\mathbf{L}$.

Path-aperture validation protocol. At GPT-2's midpoint layer we select, for every head and each of 12 WikiText-103 samples, the eligible row with largest attention mass (at least eight unmasked keys). We apply unit-Frobenius logit directions that disperse the dominant key, concentrate it, or are isotropic in the zero-sum tangent space, at budgets 0.1, 1, 3, and 4.6. Along the exponential-tilt path we integrate the exact secular-equation Jacobian norm with 24-point Gauss–Legendre quadrature and test both $\|\mathbf{A}_1 - \mathbf{A}_0\|_2$ and the realised $\|(\mathbf{A}_1 - \mathbf{A}_0)U\|_2$ using the model's recorded value-output operator. Across 1728 paths there are 0 and 0 violations, respectively. As a check on the path and the quadrature nodes, integrating the tangent velocity $J_t\,\delta\mathbf{L}$ with the same rule reproduces the endpoint attention row to a maximum relative error of 3.9e-10; this certifies the endpoint reconstruction, not Γ_{path} itself, whose accuracy is not established at that level. The zero-violation claim has the margin between the maximum bound ratio 0.9949 and one to absorb quadrature error. The frozen base-point aperture, by contrast, fails on 29.46% of all paths and 85.07% of the dispersing ones, underestimating the realised displacement by up to $22.53\times$, and at the largest budget the dispersing paths' median path aperture is $3.24\times$ the base-point value.

Exactly when the linearisation stops holding. For the span directions of Corollary 2 the question does not need a bound at all, because the response is available in closed form. Adding b to the logits of a key set of mass ψ is an exponential tilt, so the mass moves to $\psi e^b/(\psi e^b + 1 - \psi)$ exactly, and the ratio of the realised to the first-order response is

$$\rho(\psi, b) \;=\; \frac{e^b - 1}{b\,\big(1 + \psi\,(e^b - 1)\big)}. \tag{19}$$

A single factored identity organises everything that follows. Writing

$$\psi^\star(b) \;=\; \frac{1}{b} - \frac{1}{e^b - 1} \tag{20}$$

for the *zero-error mass*, a direct computation gives

$$\rho(\psi, b) - 1 \;=\; \frac{(e^b - 1)\,\big(\psi^\star(b) - \psi\big)}{\psi e^b + 1 - \psi}, \tag{21}$$

so for $b > 0$ the error vanishes exactly on $\psi^\star$, is positive below it and negative above it, and ρ is strictly decreasing in ψ; for $b < 0$ the factor $e^b - 1$ is negative and every sign reverses, so ρ is then strictly increasing in ψ and exceeds one above $\psi^\star$.

Three readings (Figure B.1). *Which side we are on.* Under a *promoting* shift $b > 0$ we have $\psi^\star(b) < \frac{1}{2}$, so every span carrying at least half the attention mass has $\rho < 1$: there the first-order quantity *over*-states the response, the conservative side when it is used inside an upper bound, and under-prediction is confined to spans lighter than $\psi^\star(b)$. The sign of the shift is part of the statement and cannot be dropped. Suppression is the mirror image: $\rho(1 - \psi, -b) = \rho(\psi, b)$, so a *heavy* span under a negative shift sits on the under-predicting side ($\rho = 1.20$ at $\psi = 0.75$, $b = -1$, and 2.19 at $\psi = 0.9$, $b = -3$). Promotion of a heavy span and suppression of a light one are the safe cases; the other two are not. For $b > 0$, $\psi^\star$ is pinned by the sandwich $\frac{1}{2} - b/12 \leq \psi^\star(b) < \frac{1}{2}$, sharp to leading order, with $\psi^\star(b) \to \frac{1}{2}$ as $b \to 0$ and $\psi^\star$ strictly decreasing.

How large the error is. The balanced row is the best case, not the worst: there $\rho = 2\tanh(b/2)/b$ is even in b, so the deviation from one carries no linear term and obeys

$$\big|\rho(\tfrac{1}{2}, b) - 1\big| \;\leq\; \frac{b^2}{12}, \qquad \text{hence} \qquad |b| \leq \sqrt{12\,\eta} \implies \rho(\tfrac{1}{2}, b) \geq 1 - \eta, \tag{22}$$

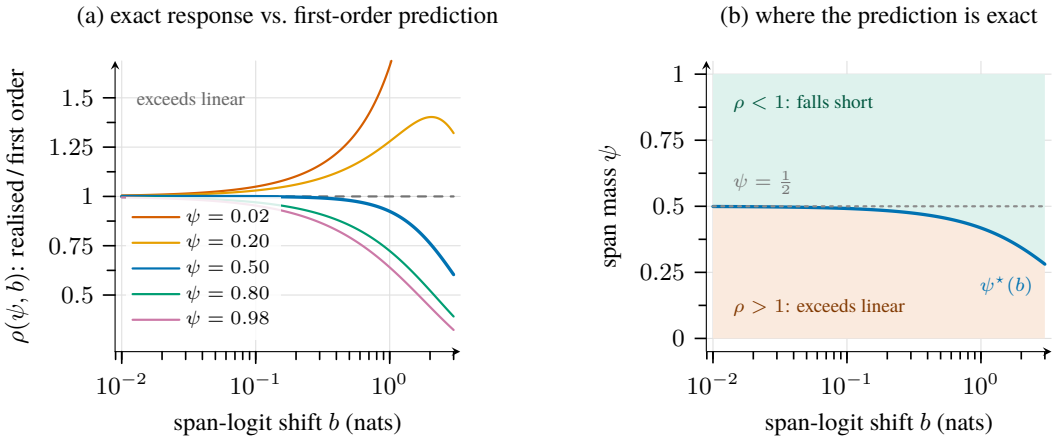

Figure B.1: **Exact finite-to-linear response ratio.** For a promoted span of base mass ψ, panel (a) plots the exact mass gain divided by its first-order prediction, $\rho(\psi, b) = (e^b - 1)/[b(1 + \psi(e^b - 1))]$. Values above one mean that the base-point linearization underpredicts the finite response. Panel (b) shows the zero-error locus $\psi^\star(b) = 1/b - 1/(e^b - 1)$. At $\psi = 1/2$ the error is second order, whereas for other fixed masses it is generally first order. Suppression follows by $\rho(1 - \psi, -b) = \rho(\psi, b)$. These analytic curves contain no measured data.

a validity radius: the prediction is accurate to within 1% for $|b| \le 0.34$ and within 10% for $|b| \le 1.09$. The exact minimiser of $|\rho - 1|$ at a given budget is $\psi^\star(b)$, where the error vanishes; what distinguishes $\psi = \frac{1}{2}$ is that it is the unique *budget-independent* mass with second-order accuracy, every other fixed ψ costing error linear in b by (21). That is the relevant comparison for Theorem 6, whose extremal row for $m \ge 4$ is the balanced two-key row: splitting it by either of its two keys gives $\psi = \frac{1}{2}$, so the configuration where the aperture bound is tight is also the one where the first-order description is stable across budgets rather than only at a single tuned budget.

What does not degrade. The prediction degrades with b; the bound does not. For a span shift the exact response satisfies

$$\big|\psi(b) - \psi\big| \;\le\; \tfrac{1}{4}\,|b| \qquad \text{for every } b, \tag{23}$$

with $\frac{1}{4} = \max_\psi \psi(1 - \psi)$ the balanced-row value of the Jacobian quadratic form: the exact response never outruns the aperture rate, at any budget. For a general direction the same role is played by the uniform constant $\frac{1}{2}$ (Proposition B.4). What (19) quantifies is the accuracy of the first-order *prediction*, not the validity of the bound.

C Reach at Scale

Theorem 1 and Corollary 2 are per row with the values fixed. This appendix states what follows for several biased rows or heads, for a range-capped budget, and for several capped spans. Every machine-checked step is named in Table M.1; the analytical steps are collected in the last paragraph of this appendix. We identify a linear functional with its Riesz vector and write $\varphi(y) = \langle \varphi, y \rangle$.

Aggregation. Heads write additively into the residual stream (Elhage et al., 2021), and the rows of one head are steered by one bias vector each, so the object is a finite family of rows indexed by $j \in \mathcal{J}$, each with its own positive base row $a^{(j)}$, partition into k_j groups with masses $\pi_g^{(j)} > 0$ and centroids $\mu_g^{(j)}$ as in Theorem 1, bias $b^{(j)} \in \mathbb{R}^{k_j}$, and a linear output map W_j into a common space F (the head's output projection, or the embedding of a row into the head output). Write $Y(b) = \sum_j W_j Y^{(j)}(b^{(j)})$ for the aggregate output and fix a reference group $g_0(j)$ in every row.

Proposition C.1 (Additive aggregation ceiling). *Let* $S = \sum_{j \in \mathcal{J}} W_j \operatorname{span}\{\mu_g^{(j)} - \mu_{g_0(j)}^{(j)} : g \le k_j\} \subseteq F$.

(i) $Y(b) - Y(b') \in S$ *for all bias assignments* b, b'.

(ii) $\dim S \le \sum_{j\in\mathcal{J}}(k_j-1)$, *hence* $\dim S \le \min\{\sum_j (k_j-1),\ \dim F\}$.

(iii) If every row has k groups and one bias $b \in \mathbb{R}^k$ is shared by all of them, then $Y(b,\dots,b) - Y(0) \in S$ *and* $\dim S \le |\mathcal{J}|\,(k-1)$.

Proof. (i) Step 1 of Theorem 1 applied to row j gives $Y^{(j)}(b^{(j)}) - \mu^{(j)}_{g_0(j)} = \sum_g w^{(j)}_g(b^{(j)})\big(\mu^{(j)}_g - \mu^{(j)}_{g_0(j)}\big)$; subtracting the same identity at $b'^{(j)}$, applying W_j and summing over j gives (i). (ii) S is spanned by the $\sum_j(k_j-1)$ vectors $W_j(\mu^{(j)}_g - \mu^{(j)}_{g_0(j)})$ with $g \neq g_0(j)$, and is a subspace of F. (iii) is (i) at $b^{(j)} = b$ and $b' = 0$, and (ii) at $k_j = k$. □

Equality in (ii) requires the projected contrasts to be linearly independent across rows, a property of the values. The set traced by a shared bias in (iii) is in general a curve inside S, because each row renormalises with its own masses $\pi^{(j)}$; when every row has the same mass vector all rows share one weight vector $w(b)$, and the set is the relative interior of $\mathrm{conv}\{\sum_j W_j\mu^{(j)}_g : g \le k\}$, inside a $(k-1)$-flat. Neither statement is part of the proposition.

Range-capped budget. A logit edit whose entries range over at most B moves a positive row a on m unmasked keys exactly to the rows in the radius-B ball of Hilbert's projective metric,

$$\mathcal{P}_B(a) = \left\{q \in \Delta^\circ_{m-1} : q_i a_j \le e^B q_j a_i \text{ for all } i,j\right\}, \tag{24}$$

without the factor $1/2$ (Birkhoff, 1957; Bushell, 1973), whose nontrivial vertices for $B > 0$ are the two-level reweightings $q^S_i \propto a_i e^{B\mathbf{1}\{i\in S\}}$, one common factor on a set of keys (Karlsson et al., 2006, Prop. 3.2), the extremality behind the Vertex-Softmax certificate (Rezazadeh & Gholami Davoodi, 2026). With one row and its values fixed, that range cap bounds the reachable outputs through a polytope.

Proposition C.2 (Budgeted reachable outputs)**.** *Fix a positive row $a = \mathrm{softmax}(\ell)$ on m keys, the value rows $u_1,\dots,u_m$ of U, and $B \ge 0$. Then*

$$\left\{\mathrm{softmax}(\ell+\delta\ell)^\top U :\ B(\delta\ell) \le B\right\} = \{q^\top U : q \in \mathcal{P}_B(a)\} = \mathrm{conv}\left\{(q^S)^\top U : S \subseteq [m]\right\}, \tag{25}$$

with $q^S_i \propto a_i e^{B\mathbf{1}\{i\in S\}}$ the two-level reweightings. Consequently a linear readout of the output attains its maximum over the budgeted outputs at some q^S, that is, at a binary span shift of bias B on S.

Proof. The first equality is the classical reachability (24): a range-bounded edit is an exponential tilt whose likelihood ratios obey the displayed inequalities, and conversely the log-ratio edit $\delta\ell_i = \log(q_i/a_i)$ realises any $q \in \mathcal{P}_B(a)$ within range B; both directions are machine-checked. For the second, $\mathcal{P}_B(a)$ is a compact polytope whose vertices are two-level reweightings (Karlsson et al., 2006, Prop. 3.2), and the image of a polytope under a linear map is the convex hull of the images of its vertices (Ziegler, 1995); this convex-geometry step is analytical. A linear functional on the convex hull of finitely many points attains its maximum at one of them, which is the consequence; optimality of a two-level point over a score box is the Vertex-Softmax certificate (Rezazadeh & Gholami Davoodi, 2026), and the development checks the algebraic certificate under which a balanced two-level candidate dominates every competitor whose likelihood ratios lie in a band. □

The readout ceiling. Proposition C.2 bounds a linear readout over the budgeted reachable set of an unrestricted per-key edit. The groupwise control of Theorem 1 is a far smaller class, and there the supremum is in closed form and is free of any budget.

Corollary C.3 (Readout ceiling)**.** *In the setting of Theorem 1, let φ be a linear functional on the head-output space and write $c_g = \varphi(\mu_g)$. Then for every bias $b \in \mathbb{R}^k$,*

$$\varphi\big(Y(b)\big) \le \max_g c_g, \qquad \max_g c_g - \varphi\big(Y(b)\big) \le \big(1 - w_{g^\star}(b)\big)\big(\max_g c_g - \min_g c_g\big), \tag{26}$$

where $g^\star$ attains $\max_g c_g$. The supremum is therefore exactly $\max_g c_g$, approached as the mass outside a maximising group vanishes, and attained at every finite bias exactly when the c_g all agree, the inert case below. Putting budget t on $g^\star$ and nothing elsewhere gives $1 - w_{g^\star} \le (\pi^{-1}_{g^\star}\sum_{g\neq g^\star}\pi_g)e^{-t}$, so the shortfall decays exponentially and reaching within ε of the ceiling costs a bias logarithmic in $1/\varepsilon$. Moreover $\varphi(Y(b))$ is constant in b if and only if the c_g are all equal.

Proof. By (2), $\varphi(Y(b)) = \sum_g w_g(b)c_g$ is a convex combination of the c_g, which gives the first bound, and since every $w_g(b) > 0$, equality there forces $c_g = \max_h c_h$ for every g, the attainment claim; the second follows by discarding the vanishing $g^\star$ term from $\max_g c_g - \sum_g w_g(b)c_g = \sum_{g\neq g^\star} w_g(b)(\max_h c_h - c_g)$ and bounding each remaining factor by the spread. For the last claim, if the c_g agree the combination is constant; conversely, every strictly positive q is realised by a bias, and a linear functional that is constant on the relative interior of the simplex has equal coefficients, as moving mass between two coordinates of the uniform weight shows. □

Two consequences are worth separating from the proof, which is elementary. The ceiling is computed from the unedited row and the centroids together with one readout direction, so whether a head can carry a prescribed margin is decided before any edit is applied and without any profiling pass, to the extent the map from the head output to the readout is affine (Appendix J). And inertness for a fixed readout is one scalar equation per group, $c_1 = \cdots = c_k$, rather than the coincidence of centroids in d_{model} dimensions required by Corollary 3: a head may move its own output substantially while leaving a given readout fixed at every budget. The converse direction of Corollary 3 is thus strictly stronger than what a downstream readout can detect.

Capped spans. For H span controls acting on separate rows or heads, with base masses $p_h \in (0,1)$, write $\psi_h(b) = p_h e^b/(p_h e^b + 1 - p_h)$ for the shifted mass of (4) and $t_h(b) = \psi_h(b) - p_h$ for the transfer.

Proposition C.4 (Interval-to-box composition and capacity)**.** *Assume $0 < p_h < 1$ for every h.*

- *(i) For endpoints $l_h \le u_h$, the transfer map $b \mapsto (t_h(b_h))_h$ sends the bias box $\prod_h[l_h, u_h]$ onto the transfer box $\prod_h[t_h(l_h), t_h(u_h)]$; in particular $\prod_h[0, B_h]$ onto $\prod_h[0, \psi_h(B_h) - p_h]$.*
- *(ii) For $0 \le t \le \psi_h(B_h) - p_h$ the bias $b_h(t) = \log\frac{p_h+t}{1-p_h-t} - \log\frac{p_h}{1-p_h}$ satisfies $\psi_h(b_h(t)) = p_h + t$ and $0 \le b_h(t) \le B_h$.*
- *(iii) For gains $g \in \mathbb{R}^H$, caps $B_h \ge 0$ and a threshold τ, biases $b_h \in [0, B_h]$ with $\sum_h g_h t_h(b_h) \ge \tau$ exist iff $\tau \le \sum_h \max\{0,\ g_h(\psi_h(B_h) - p_h)\}$; with signed caps $b_h \in [l_h, u_h]$, $l_h \le u_h$, iff $\tau \le \sum_h \max\{g_h t_h(l_h),\ g_h t_h(u_h)\}$.*
- *(iv) For directions d_h and a vector v in a real inner product space, and $l_h \le u_h$, $\max\{\langle v, \sum_h t_h d_h\rangle : t \in \prod_h[l_h, u_h]\} = \sum_h \max\{l_h\langle v, d_h\rangle,\ u_h\langle v, d_h\rangle\}$, attained.*

When the spans act on separate rows of one head, or on separate heads after their output projections, the aggregate displacement at biases b is $\sum_h t_h(b_h)\, d_h$ with d_h the embedded contrast $\mu_{S_h} - \mu_{\bar S_h}$ of span h, by (5) row by row and additivity over heads; so (iii) with $g_h = \langle \varphi, d_h\rangle$ decides whether a linear target $\langle \varphi, Y(b) - Y(0)\rangle \ge \tau$ is feasible under the caps, and (iv) is the support function of the reachable set of a signed transfer box.

Proof. (i) ψ_h is continuous and strictly increasing for $p_h \in (0,1)$, so the image of an interval is the interval between the endpoint values, and the box is the product of the images. (ii) $p_h e^{b_h(t)} = (p_h + t)(1 - p_h)/(1 - p_h - t)$ and $p_h e^{b_h(t)} + 1 - p_h = (1 - p_h)/(1 - p_h - t)$, whose ratio is $p_h + t$; the bounds follow from strict monotonicity, since $p_h \le \psi_h(b_h(t)) \le \psi_h(B_h)$. (iii) Put $t_h = \psi_h(B_h) - p_h$ when $g_h \ge 0$ and $t_h = 0$ otherwise, realised by the biases of (ii); conversely every feasible transfer obeys $g_h t_h \le \max\{0, g_h(\psi_h(B_h) - p_h)\}$ termwise. The signed case is the same with the two endpoints. (iv) The objective is a sum of independent linear functions of the coordinates, each maximised at an endpoint. □

What is analytical here. The convex-hull image step of Proposition C.2 is analytical, as is the composition of a finite capacity with the support function of Proposition C.4(iv). A nonlinear downstream readout needs a remainder bound on top of these exact linear statements.

D Softmax Interface: The Response Norm

Theorem 4's norm identity in both directions. A first-order logit perturbation gives $\delta Y = \mathcal{R}\,\delta\mathbf{L}$ with $(\mathcal{R}\,\delta\mathbf{L})_i = \delta\mathbf{L}_i^\top J_i U$ row by row, so $\mathcal{R}$ acts blockwise with blocks $J_i U \in \mathbb{R}^{n\times d_{\text{model}}}$, and the row blocks are orthogonal in the Frobenius inner product: $\|\mathcal{R}\,\delta\mathbf{L}\|_F^2 = \sum_i \|\delta\mathbf{L}_i^\top J_i U\|_2^2 \le$

$\max_i \sigma_1(J_iU)^2 \, \|\delta\mathbf{L}\|_F^2$, so $\|\mathcal{R}\|_{F\to F} \le \max_i \sigma_1(J_iU)$. The Frobenius norm is essential: in the spectral norm a row-wise action can gain up to $\sqrt{n}$ over its largest block, and the identity fails. The reverse inequality is attained by loading the whole budget on a row $i^\star$ maximising $\sigma_1(J_iU)$, along a leading left singular vector of $J_{i^\star}U$, which gives $\|\mathcal{R}\,\delta\mathbf{L}\|_F = \sigma_1(J_{i^\star}U)\,\|\delta\mathbf{L}\|_F$. Hence $\|\mathcal{R}\|_{F\to F} = \max_i \sigma_1(J_iU)$. Block by block, $\sigma_1(J_iU) \le \|J_i\|_{\text{op}}\|U\|_{\text{op}}$ (standard; cf. Horn & Johnson, 2013, Sec. 5.6.5), so $\|\mathcal{R}\|_{F\to F} \le \gamma(\mathbf{A})\|U\|_{\text{op}}$, with equality exactly when some row attains both $\|J_i\|_{\text{op}} = \gamma(\mathbf{A})$ and $\sigma_1(J_iU) = \|J_i\|_{\text{op}}\|U\|_{\text{op}}$, that is when its leading Jacobian eigenvector is a leading left singular vector of U.

E Experimental Details

Reachable-set protocol (the exclusion measurement). The groupwise control law is installed as a bias on the attention logits of a live forward, over $k = 2$ to 5 groups and 3072 head, layer and partition cells per model. The prescribed group masses are attained to 1.8×10^{-7} and 4.6×10^{-6}, and the predicted output to 4.6×10^{-7} and 6.1×10^{-6} relative, against a 1.0×10^{-4} tolerance. The reachable set's dimension is read the same way: the singular value at the predicted index exceeds the next by 2.1×10^6 and 2.2×10^5 in median, inside a 128-dimensional head-output space.

The registered criteria. Fixed before any outcome was read, and all read from the calibration run the protocol assigns them: (A) both errors above within 1.0×10^{-4} relative in float32, met; (B) the observed affine dimension never above $k - 1$, failed (below); (C) the orthogonal target's error not falling by more than 10^{-3} times the targets' common displacement between budgets 5 and 50, met; (D) whether the readout gradient's squared share in the reachable span predicts the realised margin movement, registered without a pass mark, whose correlation was not computed and is left open; and, from an addendum, (E) no realised movement exceeding the readout ceiling by more than 10^{-6} relative, failed (below); and, from a second addendum, (F) the distribution of the headroom ratio, also without a pass mark (below). The other half of (D) was computed. Against a random subspace of the same dimension r, whose expected share is $r/128$ and averages 0.0195 on both models, the share averages 0.0335 [0.0323, 0.0346] and 0.0272 [0.0262, 0.0282], 1.71 and 1.39 times the null, and exceeds it at the 0.05 level on 19.6% and 13.0% of cells against the 5% the null gives: the reachable span is better aligned with the readout direction than a random subspace of its size, and still holds only that small share of it. The protocol's decision rule has branches for A to C all holding and for C failing, and none for B failing, so none of its branches is claimed.

What the floor is. At the saturating budget, mean movement along the orthogonal part is 4.5×10^{-8} and 1.6×10^{-7} against mean reachable movements of 6.6×10^{-1} and 2.7: about one relative roundoff unit of float32 at that scale (2^{-24} times the value) in each model. It is not flat in budget, rising by 3.6×10^{-8} and 1.2×10^{-7} between budgets 5 and 50, which is what gate C is read on and four orders inside its tolerance. What identifies the residual as arithmetic rather than leaked reach is a control: shuffling the partition destroys the geometry and leaves the mean orthogonal movement at 0.98 and 1.02 times its unshuffled value, so the residual does not know whether the partition is real.

Two registered gates failed. Gate B, that the observed affine dimension never exceeds $k - 1$, holds on 99.97% and 96.84% of cells at a 10^{-5} singular-value threshold, with a largest excess of 1 and 11 dimensions; gate E, the no-overshoot condition, fails at 2.6×10^{-3} and 6.0×10^{-3} relative. Both stand as failures. The spectral gap above and the absolute movements here are reported beside them, not in place of them: on the offending rows the orthogonal movement is indistinguishable from rows that pass, and the largest relative overshoot sits in the smallest decile of ceilings. These are float32 runs; in bfloat16 the rounding is the same order as the effect under test, so the measurement would report arithmetic instead.

The ceiling and the headroom. The readout ceiling of a head is the exact supremum of the first-order margin movement its routing can produce, measured from its unedited readout as $\max_g c_g - \varphi(Y(0))$ in the notation of Corollary C.3, with median 0.036 and 0.107 logits of margin across heads. The realised movement, as a fraction of that ceiling, has median 0.422 and 0.087 at budget 1, 0.860 and 0.445 at budget 2, 1.000 and 0.997 at budget 5, and 1.000 and 1.000 at budget 50: in median the ceiling is reached to within a percent by a budget of five, the approach Corollary C.3 predicts. The headroom ratio (F) divides the sum of the ceilings of a layer's heads by the margin an edit must close, on the 1152 and 576 cells (example, layer and k, pooled over $k = 2$ to 5) whose clean margin is negative. Summing per-head ceilings assumes the heads' movements add, so the ratio is first order.

	trained			random init		
model	p_{50}	p_{90}	p_{99}	p_{50}	p_{90}	p_{99}
GPT-2	0.344	0.535	0.788	0.001	0.010	0.104
Pythia-160M	0.204	0.618	0.948	0.002	0.016	0.152
Pythia-1B	0.338	0.537	0.794	0.009	0.068	0.425
Qwen2-0.5B	0.325	0.580	0.856	0.002	0.014	0.131
Qwen2-1.5B	0.353	0.577	0.826	0.006	0.043	0.297
Phi-2	0.327	0.523	0.809	0.014	0.093	0.489
OLMo-1B	0.400	0.652	0.909	0.014	0.085	0.446

Table E.1: Support-aware saturation $s_i = d_i^{\Sigma}\gamma_i / C(m_i) \in [0, 1]$, the fraction of the sharp bound a row uses in its own unmasked coordinates, over all measured rows of each model (counts above). Trained weights against randomly initialised weights of the same architecture, same data, same code path. Every percentile is higher in the trained arm on every model. These are whole-distribution numbers; the position-filtered 99th percentiles quoted for the saturation sweep in §5 exclude early-causal and BOS-sink rows and are therefore not comparable with this table entry for entry.

Its median is 1.00 and 0.58, with 50.0% and 69.4% of cells below one and a 10th to 90th percentile range of 0.48 to 27.9 and 0.10 to 4.1. On about half the OLMo-2 cells and most Qwen2.5 cells, a single layer's k-group routing lacks the first-order reach to close the margin. These are calibration cells. Whether the ratio predicts which answers an edit flips was tested on the test split at each edited layer, and is reported with the depth sweep (Appendix J).

Phase-space sweep row counts. The structural sweep behind the saturation sweep's ambient arm and Figure 2 pools 70,350,928 rows over 200 samples per model, with 0 violations of the product bound: GPT-2 4,185,936, Pythia-160M 4,245,696, Pythia-1B 3,773,952, Qwen2-0.5B 10,389,792, Qwen2-1.5B 10,389,792, Phi-2 29,766,656, OLMo-1B 7,599,104.

Models and checkpoints. GPT-2 (124M): `openai-community/gpt2` (Radford et al., 2019). Pythia-160M: `EleutherAI/pythia-160m`, revision `step143000` (Biderman et al., 2023). Pythia-1B: `EleutherAI/pythia-1b`, revision `step143000` (Biderman et al., 2023). Qwen2-0.5B: `Qwen/Qwen2-0.5B` (Yang et al., 2024). Qwen2-1.5B: `Qwen/Qwen2-1.5B` (Yang et al., 2024). SmolLM2-1.7B: `HuggingFaceTB/SmolLM2-1.7B` (Ben Allal et al., 2025). Phi-2 (2.7B): `microsoft/phi-2` (Li et al., 2023c). OLMo-1B: `allenai/OLMo-1B-hf` (Groeneveld et al., 2024). OLMo-2-1B-Instruct: `allenai/OLMo-2-0425-1B-Instruct` (Team OLMo et al., 2025). Qwen2.5-3B-Instruct: `Qwen/Qwen2.5-3B-Instruct` (Qwen Team, 2024).

Ambient support in the saturation sweep. The saturation sweep tokenises without padding and truncates at 512 tokens, so each row has its actual ambient width $n \leq 512$. Its distance d_i is taken over that full n-key row, including masked zeros, and compared with $C(n)$. The intervention and steering experiments use a maximum length of 128; the support-aware statistic s_i is the one that normalises by the unmasked count m_i instead.

Trained against randomly initialised weights. The random arm measures the same rows on the same data with the same code path, on a model built from the architecture configuration alone: no pretrained weights are loaded, so every parameter comes from the model class's own default initialiser in the HuggingFace implementation, under seed 1319. Table E.1 gives the support-aware saturation distribution of both arms. On the rows at position ≥ 8 whose argmax is not the initial key, the 99th-percentile saturation is 0.799 to 0.949 across trained models and at most 0.193 across untrained ones; over all rows the median saturation, pooled as the median of the trained models' medians, is 0.338. The two separate at every reported percentile on every model, which the ambient statistic of Figure 2 does not: there both arms reach the width ceiling, because the two-key equality case is supplied by the causal mask at early positions rather than by learning. The trained extremisers are learned selectors: on Pythia-160M the maximising row is a balanced two-of-182-key row at position 181. Over the Pythia training checkpoints the position-filtered 99th percentile rises steeply in Pythia-160M ($0.699 \to 0.949$ over steps 1000–143000) and shallowly in Pythia-1B ($0.741 \to 0.799$).

Exact-control validation. The binary vertices match the likelihood-ratio-polytope optimum on 2,800 random objectives to 1.33e-15, and the routing–payload and centroid factorisations hold on 1,098 span interventions and 216 grouped trials to 8.90e-12 and 6.83e-16, with 0 excess over the $k-1$ ceiling. The synthetic check samples positive rows for $m = 2, \dots, 8$, budgets in $[0.02, 5]$, and compares every nontrivial binary vertex against linear-program optima for random scalar objectives. The real-model check uses six WikiText-103 samples of at most 64 tokens (two are 57 and 45 tokens long), GPT-2 layer 6, the final row of every head, and prefix/recent/top-attention/random spans covering a quarter of the row's support (16, 14, or 11 keys), at biases 0.1, 1, and 4.6. Matched controls increase span mass by 0.05 and are omitted for spans whose base mass exceeds 0.95, since the target must stay inside the simplex; this leaves 234 gain-matched records. On them the bias holds the routing gain at 0.05 to within 5.97e-16 while the realised output change spreads by $2.3\times$ between its 10th and 90th percentiles; on the 864 fixed-budget records routing gain alone ranks that change with Spearman $\rho = 0.983$ and payload alone with $\rho = 0.289$. Group tests use $k \in \{2, 3, 4\}$ and 216 independently sampled controls. No gradients, generation, or downstream Jacobians are required.

Twin spans on one row. This is the witness summarised in §3 at (6), in full: every routing quantity is shared and only the payload contrast differs. On $m = 5$ keys take

$$\begin{aligned} \mathbf{A}_i &= (0.28,\ 0.17,\ 0.26,\ 0.19,\ 0.10), \qquad Y(0) = \mathbf{A}_i U = (1, 1), \\ U^\top &= \begin{pmatrix} -0.70 & 3.80 & 2.00 & -0.50 & 1.25 \\ 0.15 & 2.40 & -1.00 & 1.50 & 5.25 \end{pmatrix}, \\ S_A &= \{1, 2\}, \quad S_B = \{3, 4\}, \qquad \psi_A = \psi_B = 0.45, \\ \mu_{S_A} - \mu_{\bar{S}_A} &= (1, 1) - (1, 1) = (0, 0), \\ \mu_{S_B} - \mu_{\bar{S}_B} &= \left(\tfrac{17}{18}, \tfrac{1}{18}\right) - \left(\tfrac{23}{22}, \tfrac{39}{22}\right) = -\tfrac{10}{99}\,(1, 17). \end{aligned} \tag{27}$$

The two spans share the row, the aperture, the mass, the span size, the first-order gain $\psi(1-\psi)$, the budget $|b|\sqrt{2}$ at every b, and the closed-form mass $\psi(b)$. Span A is output-inert at every b: $Y_A(b) = (1, 1)$ identically while its mass runs from 0.015 at $b = -4$ to 0.9996 at $b = 8$. Span B is injective in b: $Y_B(b) = (1, 1) + (\psi(b) - \psi)\,(\mu_{S_B} - \mu_{\bar{S}_B})$ moves by 0.21 at $b = 0.5$ and by 0.95 at $b = 8$. Nothing in the attention row distinguishes the two spans; the difference lives in U.

Compute budget. The structural measurements run on Apple Silicon (MPS backend) in float32, at a few GPU-hours across all 7 model variants; the pipeline also runs on CUDA. The pre-softmax intervention sweep and the downstream-factor estimation run on Apple Silicon in float32: the perplexity arm takes 22 minutes per model on GPT-2, and the downstream-factor estimate takes 11 minutes on GPT-2 and 75 on Qwen2-0.5B. Random seed: 1319 for all stochastic operations except the span-choice draw and its bootstrap (20260912, Appendix G).

F Span-Steering Protocol

The steering-interface comparison and Figure 3 come from a separate experiment that applies the published steering primitive itself rather than a synthetic direction.

Primitive. A constant b is added to the pre-softmax scores of the span columns $j \in S$ lying inside each row's causal support; masked keys are untouched and the model recomputes its own softmax. In the Frobenius budget of §3 that edit costs $\|\delta \mathbf{L}_i\|_2 = |b|\sqrt{|S|}$ on row i, which is how a span bias and a budget are compared below. This is exactly PASTA's rescaling of the non-highlighted post-softmax weights by α (Zhang et al., 2024a, §5.3), which is $b = \log(1/\alpha) = 4.6$ at that method's fixed $\alpha = 0.01$, and SpotLight's $b_i = \log(\psi_\star/\psi_i)$ (Venkateswaran & Contractor, 2026). Shifts $b \in \{0.5, 1, 2, 3, 4.6\}$ plus a SpotLight-adaptive arm with $\psi_\star = 0.1$, which computes ψ_i from the current pre-shift scores of that layer and leaves rows with $\psi_i \geq \psi_\star$ unshifted. Every shift applied is therefore non-negative, so the sign condition of §B is satisfied throughout and the first-order prediction is an upper bound exactly on the spans heavier than $\psi^\star(b)$. Two settings: the single-layer setting shifts layer $\lfloor L/2 \rfloor$ only (GPT-2 6, Qwen2-0.5B 12, Qwen2-1.5B 14, SmolLM2-1.7B 12), the all-layer setting shifts every layer as SpotLight does; all heads in both. The span is the question span where the dataset marks one, otherwise a contiguous window of at most eight tokens drawn in the first half of the sequence.

Post-softmax bypass control. The same output change is routed the other way: the realised per-head change $\delta a_h V_h$ is added to the value contraction with the attention weights left clean, so δY is identical by construction and only the pathway differs. The two forwards must agree downstream, and do, to a maximum relative error of 3.0e-04 (GPT-2) to 2.6e-03 (Qwen2-1.5B). This is the control that makes the slope separation in Figure 3 a statement about the interface rather than about two different perturbations.

Recorded quantities and checks. Per row: the span mass before and after, its closed-form value $\psi e^b/(\psi e^b + 1 - \psi)$, the first-order gain $\psi(1-\psi)$, $\|J_i\|_{\text{op}}$, $|S \cap \text{supp}_i|$, b_i, the budget $\|\delta \mathbf{L}_i\| = b_i\sqrt{|S \cap \text{supp}_i|}$, the realised $\|\delta a_i\|$ against its first-order value $\|J_i \delta \mathbf{L}_i\|$, the realised $\|\delta Y_i\|$, $\|V_h W_O^h\|_{\text{op}}$, and the two gains $g_{\text{pre}} = \|\delta Y_i\|/\|\delta \mathbf{L}_i\|$ and $g_{\text{post}} = \|\delta Y_i\|/\|\delta a_i\|$. The realised mass matches the closed form to at most 2.8e-07 (GPT-2) to 3.5e-07 (SmolLM2-1.7B), which is the check that the hook does what the corollary says. Rows are a fixed random subset per sample and head plus the final row, giving 634240 rows on GPT-2 to 1691277 on SmolLM2-1.7B. Models GPT-2, Qwen2-0.5B, Qwen2-1.5B and SmolLM2-1.7B; 600 samples per model over WikiText-103, HaluEval-QA and TruthfulQA; sequence length 128; seed 1319; wall clock 5–25 minutes per model.

Budget sweep. Table F.1 reports, per shift b and per model in the single-layer setting with datasets pooled, the descriptive log–log slope of the pre-softmax gain g_{pre} and of the post-softmax bypass gain g_{post} against the base-point aperture γ_i, with the R^2 of the pre-softmax fit. The pre-softmax slope decays with b as the base-point linearisation degrades, the bypass slope stays near zero at every b, and the R^2 values are small throughout: these are pooled row-wise regressions over correlated rows, and the slope separation is the finding, not the fit.

	GPT-2			Qwen2-0.5B			Qwen2-1.5B			SmolLM2-1.7B		
b	pre	post	R^2	pre	post	R^2	pre	post	R^2	pre	post	R^2
0.5	0.582	0.003	0.054	0.563	0.017	0.090	0.987	0.006	0.762	0.830	0.005	0.294
1	0.530	0.003	0.047	0.506	0.017	0.079	0.966	0.006	0.762	0.785	0.005	0.284
2	0.404	0.003	0.031	0.366	0.017	0.049	0.912	0.007	0.755	0.671	0.005	0.254
3	0.259	0.003	0.015	0.206	0.017	0.020	0.840	0.007	0.739	0.537	0.005	0.207
4.6	0.040	0.003	0.001	-0.033	0.017	0.001	0.695	0.007	0.684	0.316	0.005	0.112
adaptive	-0.029	0.043	0.001	-0.015	0.053	0.001	0.160	0.007	0.391	0.074	0.018	0.026

Table F.1: Span-steering budget sweep: descriptive log–log slopes of the pre-softmax gain (pre) and of the matched post-softmax bypass gain (post) against the base-point aperture, per shift b, single-layer setting, datasets pooled, with the R^2 of the pre-softmax fit. The adaptive row is the SpotLight rule $b_i = \log(\psi_\star/\psi_i)$.

G Span-Choice Protocol

The span-choice experiment tests the routing–payload factorisation (5) as an ex-ante decision rule rather than as an identity. The protocol (SHA-256 prefix `5c949389`) was frozen before any outcome was read, and no budget, span, layer set or threshold was changed afterwards.

Replication arm. A pre-registered addendum (SHA-256 prefix `054f4db0`), frozen before that run, repeats the protocol unchanged on Qwen2.5-3B-Instruct: the same architecture family as Qwen2-1.5B at twice the scale and instruction-tuned, with the same seed, data, exclusions, spans, budgets, predictors, endpoints and tie tolerance, and with the consequences of both outcomes fixed in advance. The arm replicates: $A_{\text{contrast}} = 0.613$ [0.555, 0.672] against $A_{\text{mass}} = 0.391$ [0.332, 0.449], paired difference +0.223 [+0.129, +0.312], with 0 outcome ties and an identity residual of 2.2e-06. Always steering the evidence scores 0.344 here. The three-model verdict is reported as it stands and is not recomputed over four models.

Models and data. Qwen2-1.5B, SmolLM2-1.7B and OLMo-2-1B-Instruct (`allenai/OLMo-2-0425-1B-Instruct`; Team OLMo et al., 2025), bf16 weights on MPS with every geometric quantity computed in float32. HaluEval-QA recast as balanced A/B, with a tokenisation oracle asserting per example that the teacher-forced gold token and the greedy generated token are the same

token id; 256 fresh examples per model at sequence length at most 256 (seed 20260912), every source row used by any earlier run in this paper excluded through the recorded exclusion summaries, and examples whose spans map to an empty token set dropped and counted (0 on every model). The two candidate spans are fixed by construction, never by outcome: S_e, the supplied knowledge passage, and S_q, the question sentence.

Primitive and readouts. The constant bias b of Appendix F on the span's keys within each row's causal support, at every layer and head (the InstABoost form), at $b = 1.0$ (primary) and $b = 2.0$ (secondary); a single-layer arm at layer $\lfloor L/2 \rfloor$ and $b = 1.0$ is the implementation check. The task margin $m = \text{logit(correct)} - \text{logit(incorrect)}$ is read at the answer position under constrained scoring; the generated answer is greedy over 4 new tokens with a strict first-token parse, invalid counted as wrong. The realised outcome is $\Delta_m(S) = m(\text{steered on } S) - m(\text{clean})$, one forward per span and budget.

Predictors. All are computed from the clean pass alone. At the final query row of layer ℓ and head h, with clean span mass $\psi_{\ell h}$, conditional value centroids μ_S and $\mu_{\bar{S}}$, and $g_{\ell h}$ the gradient of m with respect to the head's weighted value (one backward pass), both taken before the output projection,

$$P_{\text{contrast}}(S) = \sum_{\ell,h} \big(\psi_{\ell h}(b) - \psi_{\ell h}\big)\, g_{\ell h}^{\top} \big(\mu_S - \mu_{\bar{S}}\big)_{\ell h}, \qquad P_{\text{mass}}(S) = \sum_{\ell,h} \big(\psi_{\ell h}(b) - \psi_{\ell h}\big),$$

$$P_{\text{aperture}}(S) = \sum_{\ell,h} \gamma_{\ell h}, \qquad P_{\text{length}}(S) = |S|,$$

with $\psi(b) = \psi e^b/(\psi e^b + 1 - \psi)$ and P_{random} a seeded coin. P_{contrast} is not a payload-only statistic: the contrast is its direction factor and the routing-mass gain and downstream gradient are its scale. It is exact in routing at fixed values and first order downstream, but it is not the whole first-order effect of the intervention that is run: the bias is installed on every causal query row, while the score is read at the final row alone, so the direct contributions of earlier rows through later attention layers are absent from it even to first order. It stands in for the quantity it ranks, and the ranking is what is tested; P_{aperture} is the whole-row Jacobian norm at the final row and does not depend on S, so it ties on every example.

Endpoints and decision rule. Primary: the pairwise ranking accuracy

$$A_P = \frac{1}{n} \sum_{\text{examples}} \mathbf{1}\big[\text{sign}(P(S_e) - P(S_q)) = \text{sign}(\Delta_m(S_e) - \Delta_m(S_q))\big]$$

at $b = 1.0$, excluding and counting examples with $|\Delta_m(S_e) - \Delta_m(S_q)| < 10^{-4}$ (0 on every model); a predictor tie counts as incorrect, and the aperture column reports the half-credit score. Uncertainty: 10000 example-level bootstrap resamples. The protocol states: "Success requires, on at least two of the three models: (1) the 95% interval of A_{contrast} lies above 0.5, and (2) the 95% paired interval of $A_{\text{contrast}} - A_{\text{mass}}$ lies above 0." Both hold on SmolLM2-1.7B and OLMo-2-1B-Instruct; on Qwen2-1.5B (1) holds and (2) fails (Table G.1). The registered secondary endpoints are in Table G.2: the correlation between P_{contrast} and the realised gain is positive on every model, the inertness screen is above its no-information value of 0.5 on every model except Qwen2-1.5B, and at the secondary budget the gap over span mass clears zero on SmolLM2-1.7B and the 3B arm and not on Qwen2-1.5B or OLMo-2-1B-Instruct. Table G.3 gives the generated-answer accuracy and the realised margin gain of the chosen span. Qwen2-1.5B is a base model and emits no valid A/B generation, so only its margin endpoints count.

Integrity checks. In the single-layer arm the identity residual $\|\Delta Y_{\text{row}} - (\psi(b) - \psi)(\mu_S - \mu_{\bar{S}})\|/\|\Delta Y_{\text{row}}\|$ at the final row is at float32 precision on every model when the attention probabilities are recomputed in float32 (maximum 1.5e-05 on Qwen2-1.5B, 1.9e-06 on SmolLM2-1.7B, 2.0e-06 on OLMo-2-1B-Instruct), and the realised span mass matches its closed form to at most 1.1e-06, 5.4e-07 and 4.2e-07. That is a check on the algebra, not on the run: recomputed from the bfloat16 probabilities the models actually used, the same residual reaches 14.8%, 2.8% and 3.3%. The identity holds exactly; the arithmetic the run executed does not, and the ranking endpoint is what the claim rests on.

Model	A_{contrast} [95%]	A_{mass} [95%]	difference [95%]	aperture	length	random	passes
Qwen2-1.5B	0.617 [0.559, 0.676]	0.582 [0.523, 0.641]	+0.035 [-0.055, +0.125]	0.500	0.430	0.500	no
SmolLM2-1.7B	0.648 [0.590, 0.707]	0.402 [0.344, 0.461]	+0.246 [+0.164, +0.324]	0.500	0.418	0.488	yes
OLMo-2-1B-Instruct	0.582 [0.523, 0.641]	0.395 [0.336, 0.457]	+0.188 [+0.094, +0.277]	0.500	0.418	0.531	yes
Qwen2.5-3B-Instruct	0.613 [0.555, 0.672]	0.391 [0.332, 0.449]	+0.223 [+0.129, +0.312]	0.500	0.355	0.496	yes

Table G.1: Span-choice primary endpoint: pairwise ranking accuracy A_P at $b = 1.0$ for every predictor, with bootstrap intervals for the contrast, the mass and their paired difference, and the protocol's pass verdict. The aperture predictor ties on every example and is shown at half credit. The last row is the 3B replication arm, outside the three-model verdict.

Model	Spearman	sign agr.	inertness	A_{contrast}, $b = 2.0$	gap, $b = 2.0$
Qwen2-1.5B	$+0.466$ [$+0.392$, $+0.540$]	0.682	0.442	0.586 [0.527, 0.648]	$+0.055$ [-0.035, $+0.141$]
SmolLM2-1.7B	$+0.575$ [$+0.504$, $+0.636$]	0.715	0.615	0.609 [0.547, 0.668]	$+0.191$ [$+0.109$, $+0.277$]
OLMo-2-1B-Instruct	$+0.317$ [$+0.226$, $+0.395$]	0.600	0.654	0.523 [0.461, 0.586]	$+0.062$ [-0.035, $+0.156$]
Qwen2.5-3B-Instruct	$+0.322$ [$+0.228$, $+0.415$]	0.654	0.577	0.629 [0.570, 0.688]	$+0.227$ [$+0.133$, $+0.316$]

Table G.2: Registered secondary endpoints of the span-choice test, reported and not decisive. The Spearman correlation between P_{contrast} and the realised margin gain Δ_m is over all span-example pairs at $b = 1.0$; sign agreement is the fraction of pairs with nonzero Δ_m whose signs match; the inertness screen is the fraction of the lowest decile of $|P_{\text{contrast}}|$ whose $|\Delta_m|$ lies below its median, 0.5 with no information; the last two columns repeat the primary endpoints A_{contrast} and $A_{\text{contrast}} - A_{\text{mass}}$ at the secondary budget. The last row is the 3B replication arm, outside the three-model verdict.

The constant policies. Neither span is the right one to steer in general, so the baseline that decides whether inspecting an example helps is the better of the two constant policies. Always steering the question ranks at 0.598, 0.602, 0.574 and 0.656 on the four models, always steering the evidence at one minus each. The contrast minus that policy is $+0.020$ [-0.066, $+0.109$], $+0.047$ [-0.039, $+0.133$], $+0.008$ [-0.070, $+0.086$] and -0.043 [-0.117, $+0.027$]: every interval covers zero and the 3B point estimate is negative. This baseline was added after the protocol was frozen; it is a comparison the protocol did not require and does not relax, and it is reported because it is the one that decides what the result licenses.

	generated accuracy			difference [95%]		Δ_m of chosen span	
Model	clean	contrast	mass	contrast − mass	contrast − clean	contrast	mass
Qwen2-1.5B	0.0%	0.0%	0.0%	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]	+0.238	+0.166
SmolLM2-1.7B	60.9%	61.7%	55.9%	+5.9 [+2.3, +9.4]	+0.8 [-3.9, +5.5]	+0.048	-0.047
OLMo-2-1B-Instruct	84.4%	80.1%	79.7%	+0.4 [-1.6, +2.3]	-4.3 [-7.8, -0.8]	-0.326	-0.431
Qwen2.5-3B-Instruct	77.3%	78.5%	74.6%	+3.9 [+0.8, +7.0]	+1.2 [-2.0, +4.7]	+0.538	-1.208

Table G.3: Span-choice task-level endpoints at $b = 1.0$: generated-answer accuracy (greedy, strict first-token parse) under clean inference and under steering of the contrast-chosen and the mass-chosen span, with paired bootstrap intervals in accuracy points, and the mean realised margin gain of the chosen span. Qwen2-1.5B, a base model, emits no valid A/B generation. The last row is the 3B replication arm, outside the three-model verdict.

H Logit-Space Intervention Protocol

All intervention experiments perturb the pre-softmax logits. For each model family (GPT-2, GPT-NeoX, Qwen2, Llama) the eager attention forward is replaced by a copy that exposes three injection points: the scaled and causally masked logits (pre-softmax; every logit-space arm of this paper), the attention weights after the softmax (the post-softmax control), and the head output after the value contraction (the state-interface control). A perturbation $\delta\mathbf{L}$ is defined on each row's causal support and is zero elsewhere; the model's own softmax is then recomputed. For grouped-query models the value tensor is recorded after the key/value expansion, so every query head sees its own group's values.

Arms. At the target layer a clean forward records $\mathbf{A}$, the per-head value-output operator $U^{(h)} = V^{(h)}W_O^{(h)}$, and per row the Jacobian J_i, its norm γ_i, and the support-aware distance d_i^{Σ}. Three logit-space arms are built row by row: the spectral direction $\delta\mathbf{L}_i \propto J_i(u_2(i)\, v_2)$, the first-order maximiser of the change of the second singular value σ_2 of $\mathbf{A}$ with singular vectors u_2, v_2 where σ_2 is simple; the value-aligned direction, the leading right singular vector of $J_iU^{(h)}$ (power iteration, checked against a dense eigendecomposition); and a Gaussian direction with the all-ones component removed. Each arm receives the per-row budget $\varepsilon/\sqrt{\text{rows}}$, so that the per-head budget is $\|\delta\mathbf{L}\|_F = \varepsilon$; a fourth arm concentrates the whole budget on the row maximising $\sigma_1(J_iU^{(h)})$. Budgets $\varepsilon \in \{0.01, 0.03, 0.1, 0.3, 1, 3\}$; seed 1319.

Checks. At the smallest budget the realised attention change is compared with its linearisation $\|\delta\mathbf{A} - J\,\delta\mathbf{L}\|_F / \|J\,\delta\mathbf{L}\|_F$, and the two targeted arms are compared by cosine similarity; both statistics are reported with the results below.

Recorded quantities. Per sample, arm, and budget: the realised $\|\delta\mathbf{A}\|_F$, the row utilisation $u_i = \|J_i\delta\mathbf{L}_i\|_2/(\gamma_i\|\delta\mathbf{L}_i\|_2)$ (median over rows), and $\|\delta Y\|_F$ for the target head; per token the pre-intervention margin m_t, the post-intervention argmax, $\|\delta\boldsymbol{\ell}_t\|_\infty$, the next-token KL divergence, and the sequence perplexity. Models: GPT-2, Pythia-160M (fp32), Qwen2-0.5B, Qwen2-1.5B, SmolLM2-1.7B; data: WikiText-103 validation, HaluEval-QA, TruthfulQA; intervention depths $\ell/L \in \{0.25, 0.5, 0.75\}$.

Per-model checks and utilisation. Table H.1 reports the two protocol checks and the per-row utilisation of each arm. The utilisation ordering (value-aligned > spectral $\gg$ random) is stable across architectures and is the mechanism behind the reach ordering of Table H.2.

Model	n	lin. rel. err.	$\lvert\cos(\text{spec., val.})\rvert$	u (spec.)	u (val.)	u (rand.)	median γ_i
GPT-2	500	6.4e-04	0.054	0.957	0.989	0.124	0.373
Pythia-160M	500	2.4e-03	0.049	0.960	0.989	0.131	0.412
Qwen2-0.5B	200	8.1e-04	0.054	0.947	0.994	0.126	0.380
Qwen2-1.5B	200	9.3e-04	0.053	0.949	0.998	0.121	0.373
SmolLM2-1.7B	200	7.3e-04	0.058	0.951	0.995	0.123	0.383

Table H.1: Protocol checks and per-row utilisation, WikiText-103 at depth $L/2$. The linearisation relative error is measured at the smallest budget; u is the median over rows at $\varepsilon = 1$. Spec. is the spectral arm, val. the value-aligned arm.

Reach at matched budget. Table H.2 reports the realised output change, the next-token KL divergence from the clean prediction, and the argmax-flip rate at $\varepsilon = 1$. All three orderings follow the utilisation, and the output-change ratio is close to constant across budgets, as a first-order quantity should be.

	median $\|\delta Y\|_F$			median KL			argmax-flip rate		
Model	spec.	val.	random	spec.	val.	random	spec.	val.	random
GPT-2	7.92	9.16	2.05	7.64e-05	7.53e-05	2.69e-06	0.77%	0.82%	0.17%
Pythia-160M	1.69	2.01	0.428	7.76e-05	9.50e-05	3.14e-06	0.72%	0.92%	0.15%
Qwen2-0.5B	1.11	1.28	0.3	1.58e-05	1.77e-05	6.29e-07	0.32%	0.41%	0.10%
Qwen2-1.5B	2.23	2.71	0.49	7.68e-06	8.14e-06	2.98e-07	0.23%	0.27%	0.04%
SmolLM2-1.7B	22.3	25.4	5.7	3.83e-05	3.40e-05	1.25e-06	0.50%	0.43%	0.10%

Table H.2: Realised output change, median next-token KL(post $\|$ pre), and argmax-flip rate at $\|\delta\mathbf{L}\|_F = 1$, WikiText-103, depth $L/2$; arms as in Table H.1.

Margin sweep. Pooling the sweep over 5 models, three datasets, three depths, six budgets and four arms gives 11,636,856 token-interventions, with observed flip rate 0.42% against a predicted 9.24%, the share of tokens whose margin a first-order bound on the edit's logit change does not certify, and 0 flips among the tokens it does certify. Flip rates fall with depth (0.54% at $\ell/L = 0.25$, 0.42% at 0.5, 0.29% at 0.75): an earlier intervention has more downstream stack to act through, matching the x_ℓ

ordering of Proposition A.1 for expansive stacks. Runner-up rates stay at or near one for every arm and budget (value-aligned 0.973 at $\varepsilon = 1$, 0.916 at $\varepsilon = 3$), so the flips are the top-1/top-2 swap rather than wholesale redistribution.

I Product Bound: Small-m Verification

Theorem 6 defines $C(m) = \max_d d \min\{\frac{1}{2}, 1 - 1/m - d^2\}$. Both envelopes are met with equality on the two-key family $(p, 1-p, 0, \ldots, 0)$, so the constant is the maximum of $d\,\|J\|_{\text{op}}$ over that family; the machine-checked proof works in d directly, and the family is parametrised here by $u = p(1-p) \in (0, \frac{1}{4}]$: there $\|J\|_{\text{op}} = 2u$ and $d^2 = \|\mathbf{A}_i\|_2^2 - 1/m = 1 - 1/m - 2u$, so

$$C(m) \;=\; \max_{0<u\le 1/4} 2u\,\sqrt{1 - 1/m - 2u}, \tag{28}$$

whose unconstrained maximiser is $u^\star = (1 - 1/m)/3$. It lies in $(0, \frac{1}{4}]$ exactly when $m \le 4$; for $m \ge 4$ the maximum sits at the endpoint $u = \frac{1}{4}$, the balanced two-key row, where the cap $\frac{1}{2}$ and the trace envelope meet. Table I.1 lists the values. Only at $m = 2$ is the extremiser interior to the simplex; for $m \ge 3$ it has zero entries, so over strictly positive (unmasked) rows $C(m)$ is a supremum approached along rows tending to the boundary, while masked rows attain it on their feasible support. All values lie below $\sqrt{2}/4 \approx 0.354$.

m	extremiser	$u^\star$	$C(m)$	attained?
2	$\left(\frac{3+\sqrt{3}}{6}, \frac{3-\sqrt{3}}{6}\right)$, interior	$\frac{1}{6}$	$\sqrt{6}/18 \approx 0.1361$	yes, by a positive row
3	$(\frac{2}{3}, \frac{1}{3}, 0)$	$\frac{2}{9}$	$4\sqrt{2}/27 \approx 0.2095$	supremum; masked rows attain it
4	$(\frac{1}{2}, \frac{1}{2}, 0, 0)$	$\frac{1}{4}$	$\frac{1}{4}$	supremum; masked rows attain it
$m \ge 5$	$(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$	$\frac{1}{4}$ (endpoint)	$\frac{1}{2}\sqrt{(m-2)/(2m)}$	supremum; masked rows attain it

Table I.1: The aperture constant for small supports: the extremising row, its two-key parameter $u^\star = p(1-p)$, the value of $C(m)$, and whether a strictly positive softmax row attains it.

The constant does not depend on the envelope. Sharper bounds on $\|J_i\|_{\text{op}}$ than Lemma B.1 are available, and one might expect a sharper envelope to give a smaller constant. It does not, for a structural reason. Let E be any function on the simplex with $\|J_i\|_{\text{op}} \le E(\mathbf{A}_i) \le \frac{1}{2}$ that agrees with $\frac{1}{2}$ at the balanced two-key row. Replacing the cap by E leaves $C(m)$ unchanged for every $m \ge 4$: it cannot rise, since $E \le \frac{1}{2}$ pointwise and the trace envelope is untouched, and it cannot fall, since the balanced two-key row meets the trace envelope at $d_i = \sqrt{\frac{1}{2} - \frac{1}{m}}$ with $E = \frac{1}{2}$ there. The minimum against the trace envelope is necessary: without it $E \equiv \frac{1}{2}$ is admissible and $\max d_i/2 > C(m)$. Yudin et al. (2025)'s $\|J_i\|_{\text{op}} \le a_{(1)}(1 - a_{(1)} + a_{(2)})$ satisfies the hypothesis, being at most the bound of Lemma B.2 and equal to $\frac{1}{2}$ at the balanced two-key row, so the constant is a property of the geometry at the extremiser rather than an artifact of the envelope used to reach it. Nair (2026) sharpens the same picture across ℓ_p: for $p \in (1, \infty)$ the value $\frac{1}{2}$ is a supremum not attained in the open simplex for $m > 2$, which is the attainment structure Table I.1 records case by case and which causal masking is what lets a real row realise.

J Context Control on Counterfactual Passages

The context-control experiment runs on NQ-Swap (Longpre et al., 2021), on OLMo-2-1B-Instruct and Qwen2.5-3B-Instruct (Qwen Team, 2024), whose passages are edited to contradict the model's parametric answer, so a context-faithful answer and a memorised one are distinguishable. The readout is whether the model selects the option supported by the passage. The edit adds one bias to the passage-span logits at a single layer, chosen on calibration data as the layer of maximum median alignment between the readout gradient and the projected centroid contrast, and the number of heads is fixed on the same calibration split. The test split is read once.

The arms differ only in which heads receive the budget. The spread arm biases every head of that layer; the concentrated arm biases the K heads whose projected centroid contrast aligns with the

readout, with the per-head dose either set by matched transfer, every selected head's largest mass transfer scaled by one shared factor so that the predicted head-output movement equals the spread arm's (or, if lower, the movement of the selected heads all at the cap, their capacity), or solved for the largest predicted gain at no more than the spread arm's predicted movement. The concentrated arm also moves the head outputs less than the spread arm, 7.01 against 9.27 on OLMo-2 and 17.9 against 21.2 on Qwen2.5-3B. The two concentrated variants agree on every example of both models, so the selection rather than the dose rule carries the effect. The concentrated arm against the spread arm was registered as an ablation, whether the concentrated arm under matched transfer beats the spread arm, with the same primary as the selection comparison below and each arm's realised movement printed beside it; it clears zero on both models (Table J.1). A fourth arm reproduces the published all-layer form (Guardieiro et al., 2025), biasing every head of every layer; it costs -6.53 [-15.33, $+2.51$] pp on OLMo-2 1B, an interval covering zero, and -14.32 [-21.11, -7.29] pp on Qwen2.5-3B. The matched-scope selection comparison, PASTA's profiling-based head ranking (Zhang et al., 2024a) restricted to the edited layer at the same head count and budget so that only the selection rule differs, was registered with its readings fixed before the test split was read. Its primary is the paired difference in context-faithful accuracy over all swapped test views, with a 10,000-resample example-level bootstrap interval, under both the solved dose and matched transfer. The contrast selection beating the profiling ranking was to be read as the better selection principle; the reverse, as profiling supplying the better head set; neither interval clearing zero, as comparable rules; and the two comparisons disagreeing in sign, as a dependence on the dose procedure, with no selection claim made. Under the solved dose the matched-scope comparison gives $+1.01$ [$+0.25$, $+2.01$] and $+0.25$ [-0.50, $+1.26$] pp; the Qwen2.5-3B solved dose is the one comparison of the four not separated from zero. On OLMo-2 both comparisons clear zero in the contrast selection's favour, the first reading. On Qwen2.5-3B matched transfer clears zero and the solved dose does not, with the same sign; the four readings do not name that case, so it is reported as it stands, one comparison of four not separated from zero, with the selection claim resting on the other three.

Three points qualify the numbers. The unrestricted comparison against a profiling-based selector is not a selection result: that selector edits 75 heads across the stack against the concentrated arm's 12 at one layer, with 2 and 0 of them at the edited layer on the two models, so scope differs as much as selection, and that outcome splits across models, favouring the contrast selection on OLMo-2 and the profiling selector on Qwen2.5-3B. The matched-scope comparison, which gives that selector its own ranking restricted to the edited layer at the same head count and budget, is the one that isolates selection, and its two arms' realised movements differ by 4.3% at the solved dose and 20.7% under matched transfer on OLMo-2, and by 23.1% and 1.8% on Qwen2.5-3B, each as a fraction of the smaller of the two arms. The contrast arm moved more at the solved dose, 7.3 against 7.01 and 16.2 against 13.1, so part of its gain there may be displacement; under matched transfer it moved less, 7.3 against 8.81 and 16.2 against 16.5, so there it is not. The free-form readout is excluded on Qwen2.5-3B, where only 13.4% of clean generations parse to one of the two options, so that column measures output format rather than faithfulness. And the deficit subpopulation, the examples the clean model answers against the passage, is wrong by construction, so a flip count there is never reported as an accuracy improvement; the numbers quoted are changes over all swapped views.

Comparison	Model	pp	95% interval	records won : lost
Concentrated vs spread, matched transfer	OLMo-2-1B	$+4.02$	[$+2.26$, $+6.03$]	–
	Qwen2.5-3B	$+3.02$	[$+1.51$, $+4.77$]	–
Contrast vs PASTA, solved dose	OLMo-2-1B	$+1.01$	[$+0.25$, $+2.01$]	4 : 0
	Qwen2.5-3B	$+0.25$	[-0.50, $+1.26$]	2 : 1
Contrast vs PASTA, matched transfer	OLMo-2-1B	$+3.52$	[$+1.51$, $+5.78$]	16 : 2
	Qwen2.5-3B	$+2.76$	[$+1.26$, $+4.52$]	11 : 0

Table J.1: NQ-Swap context control at one layer: paired difference in context-faithful accuracy over the 398 swapped test views (397 for the Qwen2.5-3B solved dose), in percentage points, with 10,000-resample example-level bootstrap intervals, and the records on which the two arms' answers differ, won against lost by the first arm.

How many records actually differ. The endpoint is binary and the two arms agree on almost every record, so an interval on the mean can rest on a handful of one-signed pairs. The discordant counts

are, for matched transfer, 16 to 2 on OLMo-2 and 11 to 0 on Qwen2.5-3B, and for the solved dose 4 to 0 and 2 to 1, over 398 paired swapped views, except the Qwen2.5-3B solved dose at 397 after one abstention. Under an exact two-sided paired sign test the two matched-transfer comparisons separate and neither solved dose does, so the OLMo-2 solved dose clears the registered bootstrap on four discordant pairs alone. The registered interval stands as the primary; this count is reported beside it for all four comparisons.

Where the first-order account applies. The ceiling is built on the gradient of the readout at the head output, so it is exact to the extent the map from there to the logits is affine. Five edited layers of OLMo-2 test it, over 255 deficit examples each. The registered endpoint is the gain in ranking from the headroom ratio of Appendix E, here on the test split at each edited layer and $k = 2$, over the deficit alone, the margin the edit must close and the ratio's denominator, and the registered prediction is that the gain declines monotonically with depth. It does not. The ratio ranks at 0.985, 0.985, 0.957, 0.840 and 0.910 against a deficit that alone ranks at 0.945, 0.970, 0.921, 0.864 and 0.892, so the gain runs $+0.040$ [$+0.021$, $+0.063$], $+0.016$ [$+0.002$, $+0.032$], $+0.035$ [$+0.007$, $+0.066$], -0.024 [-0.084, $+0.020$] and $+0.018$ [-0.007, $+0.043$] at zero, one, four, eight and twelve blocks after the edit: clear of zero at the first three depths, not at the last two, and not monotone. The protocol fixed three readings in advance: a monotone decline supports the depth account; no decline withdraws it; and a decline that is not monotone is reported as underpowered or as the wrong variable, not as partial support. This outcome is the third. On Qwen2.5-3B the same endpoint was computed at the edited layer alone, 4 blocks before the readout, over 166 deficit examples of which 25 flip: the ratio ranks at 0.936 against 0.946 for the deficit alone, a gain of -0.010 [-0.039, $+0.016$], and the ceiling alone ranks at 0.544, near chance, as OLMo-2's does four blocks from the readout, where OLMo-2's registered gain still clears zero and Qwen2.5-3B's does not. The account below is OLMo-2's.

What does separate with depth is the ceiling's standalone ranking, which an addendum written after the four-point sweep had been read, and before the one-block layer ran, fixed as a second endpoint with three readings: near its zero-block value, the account survives one block; near chance, one block destroys it; anything between is reported as the measured value with no claim about where between one and four blocks the fall occurs, since two points cannot locate it (the addendum's wording; it fixes no threshold for "near"). It is 0.804 where nothing follows the edit but the final norm and the unembedding, 0.609 with one block after it, and 0.509, 0.497 and 0.491 at four, eight and twelve blocks, the last three being chance. The one-block value is the third case, so the shape of the fall is not claimed. Neither endpoint tracks the head count, which differs by layer and runs 12, 16, 16, 8 and 4 across those depths: the one- and four-block layers share a head count of 16 and sit at 0.609 and 0.509 on the standalone ranking and at $+0.016$ and $+0.035$ on the registered gain. The clean test the registration names, fixing the head count across depths, was not run, and whether the fall survives it is open. The scope of the account is therefore narrow and, in the endpoint that carries it, post hoc: the ceiling's own ranking carries measured predictive content where the edited head is read out within about a block and none at the depths tested beyond that, while the comparison registered in advance did not decline with depth on this model.

Run-to-run variation on this task is zero: two independent passes of the identical configuration agree on all 4740 arm-by-example rows of OLMo-2, on both readouts, so a difference of a few records is not noise. That floor is measured on OLMo-2 at these prompt lengths and is not a property of the platform, since reduction order depends on tensor shapes.

K The Selector Comparison and the Finite-Response Predictor

Two analyses here were registered before they were run, the selector comparison, whose registered primary is the binary readout, and the held-out test of the finite-response predictor; two exploratory analyses, the dose path and the float32 diagnostic, had their readings fixed in advance. The criteria each is judged against are stated with its result. Every figure here is on the swapped primary population unless stated otherwise.

The arm. The matched arm scales each selected head's maximal mass transfer by one shared factor, chosen so that the predicted head-output movement equals that of the spread arm of Appendix J, and inverts per head, so it applies per-head logit biases that reduce to the common cap only at full capacity, which 311 of 398 Qwen2.5-3B and 396 of 398 OLMo-2 primary records reach; the remaining 87

Qwen records are scaled below the cap to that movement. The transfer, not the bias, is what is matched, and the reduction at capacity is mathematical rather than a demonstrated numerical equality of the implemented biases.

Selectors. Four scores rank heads for selection while the dose objective is held fixed: the transfer score $c_h = g_h^\top d_h$ the paper uses, $p_h(1-p_h)c_h$, $[\psi_h(B) - p_h]c_h$, and the autodiff derivative of the margin with respect to the shared bias through every causal query row, the all-row derivative of §5. The three analytic scores share a sign, hence a positive pool, and agree whenever that pool fits inside K, which it does on 398 of 398 Qwen records; agreement among them carries nothing about ranking. The registered compute rule runs the behavioural comparison of two selectors only if they choose different head sets on more than 5% of records; below that it is reported as not evaluated, not as showing no difference, because two interventions that agree on a fraction $1 - f$ of records can still differ in accuracy by up to f. On this rule the comparisons among the analytic scores were not evaluated, and neither were any OLMo-2 pairs. A later addendum, registered before any behavioural outcome of it was read, evaluates a pair anyway: the second intervention is run only on records whose complete plans, the selected head sets with their per-head biases, differ, and every other record, where the two interventions coincide, enters the full-population estimate as a zero paired difference. The OLMo-2 comparison below is that evaluation. Against the autodiff derivative, selected-set disagreement is 97.7% on Qwen2.5-3B (389 of the 389 such records carrying at least one sign reversal, 23.6% of heads on average) and 1.8% on OLMo-2, whose edited layer is the last, so there the derivative equals $p_h(1-p_h)c_h$ exactly and the two agree to bfloat16. Selection follows each score; the solver's variable order follows the transfer score for every selector, since SLSQP is not permutation invariant and head-set order alone changes the intervention.

Accuracy. Transfer minus autodiff on the two-choice readout, paired, 10,000 example-level resamples, the exact paired sign test beside each. Qwen2.5-3B: $+0.25$ $[-0.50, +1.26]$ pp under matched transfer with 2 to 1 discordant records, -0.25 $[-0.76, +0.00]$ at the solved dose with 0 to 1. OLMo-2: no discordant record in 398 under either arm, an observation and not a population equivalence. Neither comparison establishes an accuracy advantage; the Qwen interval permits about one point either way. Three outcomes were fixed in advance: the transfer score beating the derivative; one of the two routing-weighted scores, $p_h(1-p_h)c_h$ and $[\psi_h(B) - p_h]c_h$, beating the transfer score; and no pair separating, to be reported as a null showing that selection is not where the geometry's value lies, not that it has none. The third obtains.

Margins. The binary readout hides continuous effects. On Qwen2.5-3B the two selections move the target margin on 389 of 398 records; the signed paired difference, transfer minus autodiff, is $+0.142$ $[+0.101, +0.183]$ under movement matching and $+0.041$ $[+0.013, +0.069]$ at the solved dose, both excluding zero. These are exploratory: the registered primary endpoint was the binary readout. Under matched transfer the heads the autodiff selects with non-positive final-row gain receive dose on all 343 of 343 records selecting one, at 21.1% of total transfer on average, so the dose objective does not discard the selection there.

Mechanism. Each term of the per-row sum $\sum_{q,h} p_{qh}(1-p_{qh})\langle \nabla_{y_{qh}} m,\, d_{qh}\rangle$, with p_{qh} the base span mass on query row q of head h and $\nabla_{y_{qh}} m$ and d_{qh} the readout gradient and the payload contrast there, both taken before the output projection as in Appendix G, is one query row's routing sensitivity times its payload's alignment with the readout gradient, so the all-row and final-row scores differ exactly by the rows the final-row score omits. On 6 records the autodiff derivative is reconstructed from that sum to scale-relative error at most 8.7e-03, with the final-row term and the earlier-row remainder separated; all 19 of 19 heads whose sign the remainder reverses reconstruct their sign, one of them within about twice the reconstruction error. On OLMo-2 the remainder is exactly 0, as the architecture requires. Those records support the omitted-row explanation there; the full scan establishes the associated disagreement.

Dose path. Each plan's bias vector is frozen and scaled by α on the grid $0, 0.02, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 1$, nothing re-solved. Three curves per plan: the linear prediction $L_b(\alpha) = \alpha\, \nabla_b m(0)^\top b$; the geometry prediction $G_b(\alpha) = \sum_{q,h} \langle \nabla_{y_{qh}} m(0), d_{qh}\rangle [\psi_{qh}(\alpha b_h) - p_{qh}]$ with ψ_{qh} the law (4) on row q of head h from the base mass p_{qh} and the other factors as above, exact in the routing at fixed values and first order downstream; and the measured margin change

M. The realised per-row masses match G's routing to 8.1×10^{-7} (8.3×10^{-7} held out) and the head outputs match $(\Delta p)\, d$ to 9.8×10^{-6} (1.2×10^{-5}) at every α, maxima over all rows. On the development population (matched-transfer arm, 398 records) the autodiff plan is steeper at zero on 382 of them, and on 256 of those transfer is ahead at the endpoint, 250 bracketed by the grid; G orders the endpoint pairs at 83.7% against 67.9% for the constant choice and 32.9% for L. The run is exploratory and uses the matched-transfer arm only, where the selection difference is actually dosed; the solved dose was not run. Its four readings were fixed beforehand: G reversing the ordering and tracking M, so that finite routing saturation explains an important part of the reversal; G keeping the local ordering while M reverses, so that downstream nonlinearity is essential; G predicting part of it with a substantial residual, so that both contribute; and G improving substantially on L without capturing every reversal, predictive value short of an explanation. The fourth applies: G orders most endpoint pairs, not all of them.

Held out. 400 swapped records prepared from source questions used by no earlier stage, 398 eligible after the label-free rule and 2 excluded, no top-up; the edited layer and K pinned to their development values, nothing retuned. Primary, paired-difference error at $\alpha = 1$: L 0.350, G 0.191, improvement +0.160 [+0.130, +0.190] logits of margin, against a mean measured paired difference of +0.073. The outcomes were fixed before the run: an improvement whose interval excludes zero replicates the development result, one whose interval covers zero is a development finding that failed to replicate, and one present but smaller is reported at its held-out size with the development figure named as the optimistic one. The first and third both apply: the development improvement was +0.215 [+0.181, +0.251], and it is the optimistic figure. Secondary, ordering over 384 pairs with 14 near-ties held apart: G 77.1%, the development-fixed constant choice 60.2% (paired +16.9[+11.2, +22.7]), L 39.8%, which misorders 231 of the scored pairs (paired +37.2 [+30.2, +44.0]).

Float32 diagnostic. At the tested edit the bfloat16 instrument has a floor: a bias of 10^{-3} on every head, whose linear prediction is of order 10^{-4}, moves the margin by 0.038 logits at the median over 24 held-out records (maximum 0.256), while the hooked forward at zero bias reproduces the clean margin to 0, which rules out a zero-edit mismatch; the small-dose residuals of the dose path are of the size of this floor. The floor is bfloat16 forward arithmetic, and an edit of 10^{-3} is below one bfloat16 unit in the last place on scores of magnitude above one quarter, so that forward cannot carry it as a small perturbation. With the same bfloat16-rounded weights cast to fp32, the same frozen biases (24 of 24 of the held-out records reproducing the pinned plans, both selectors, 48 curves) and the gradients recomputed in fp32, the same edit moves the margin by 9.0×10^{-5} at the median, 0.97 times its linear prediction (0.74 to 1.34 between the tenth and ninetieth percentiles). The initial slopes of the two precisions agree to 5.6% at the median, 15 of 48 differing by more than ten percent and one by 70.8%, on the smallest slope in the set (0.12 in fp32 against 0.04 in bfloat16, the median slope being 0.51). At $\alpha = 0.02$ the residual $|M - G|$ is 0.059 in bfloat16 and 2.1×10^{-5} in fp32, the fp32 value the smaller in 48 of 48 curves, and the two forwards themselves differ by 0.058 at that dose. Fitted as a power of α over the four smallest doses, the fp32 residual grows with exponent 1.67 (interquartile range 1.34 to 1.98 over the curves; the residual at $\alpha = 0.01$ is 1.1×10^{-5}, near the fp32 resolution of the margin), consistent with the second-order growth of a Taylor remainder, whereas the bfloat16 residual has exponent -0.01: it does not shrink with the dose. The empirical curvature $2|M - G|/\|\delta\|^2$, with δ the head-output displacement, is 1.7×10^{-5} at the median and 2.0×10^{-4} at most in fp32 at $\alpha = 1$, a description of these records. At $\alpha = 1$ the fp32 residual is 0.147 against 0.177 in bfloat16, the smaller in 29 of 48 curves, while the two forwards differ by 0.119 at the median against a median measured change of 1.283: the residual at the operating budget is of the same size in fp32 arithmetic, and the bfloat16 forward there carries an arithmetic error of about a tenth of the measured change. The diagnostic's five readings were fixed before it ran, and these numbers settle them: the zero-bias forward reproduces the clean one exactly in fp32 too (largest deviation 0); at the small dose the fp32 residual is far below the bfloat16 one while the two forwards differ by about the bfloat16 residual, so the small-dose discrepancy is bfloat16 forward arithmetic; the initial slopes differ by 5.6% at the median, under the ten percent that would implicate the bfloat16 gradient; the fp32 exponent of 1.67 is below two (the protocol fixes no threshold for "near"), which is consistent with the smoothness account without establishing a bound; and the tiny edit's fp32 response is 0.97 of its linear prediction, so the floor is bfloat16 arithmetic.

What the held-out result licenses. L and G share their downstream gradients and differ only in linearised against finite routing, so the improvement is the routing correction's within this comparison; it does not make downstream nonlinearity negligible, and a base-point derivative recomputed along the path is a different comparison not run here. One arm, one model, one layer, one budget, one task.

L Related Work in Detail

Sensitivity geometry. Self-attention is not globally Lipschitz (Kim et al., 2021); Nair (2026) proves the sharp $1/2$ softmax cap in every ℓ_p norm and Castin et al. (2024) analyse self-attention smoothness. Emadi (2026) gives an exact $\infty \to 1$ induced norm, $\theta(p)/\tau$ with $\theta(p) = 4 \max_S p(S)(1 - p(S))$, and read it as bisectability rather than concentration; in our Euclidean norm the two agree at the extremiser, the balanced two-key row being both the $\theta = 1$ case and the equality case of (12), and our forbidden region is a statement about the product $d_i \gamma_i$ rather than about either factor. Yudin et al. (2025) give the row-adaptive bound $\|J_i\|_{\text{op}} \le a_{(1)}(1 - a_{(1)} + a_{(2)})$ and already read a Jacobian-norm bound as an exclusion on the top mass at the same extremal row; we instead couple the norm to the distance from uniform, giving a sharp product bound with closed-form constant, read interventionally and propagated through a causal path.

The groupwise routing factor. The routing factor of (2) is the natural-to-mean map of an exponentially tilted categorical family (Amari, 2016; Wainwright & Jordan, 2008). The closed-form bias is in Kang et al. (2026, App. A.1); the $b = 0$ factorisation is the decomposition theorem of Su et al. (2025) and the class-based softmax (Goodman, 2001); and Park et al. (2026) steer the vocabulary softmax in the same geometry, working in the linear-representation framework of Park et al. (2024), with an optimality result independent of our exclusion. What the value map adds is the payload contrast.

Attention identifiability. Jain & Wallace (2019) and Wiegreffe & Pinter (2019) debate what distinct attention distributions with equivalent predictions show, Brunner et al. (2020) prove that the left nullspace of the value map is nontrivial once the sequence outruns the head width, and construct simplex-valid alternative rows in it, together with the effective-attention projection that removes them; Sun & Marasović (2021) compute that projection and show it reads differently from the raw weights; Venkatesh & Kurapath (2026) carry the question to steering vectors; we ask which output directions a prescribed control partition can reach, and answer it with a criterion, not an existence statement.

Control-surface taxonomies. Ostermann et al. (2026) and Miehling et al. (2026) organise steering methods by adaptation type and control surface, and Alpay & Alpay (2025) frame interventions control-theoretically; Table 1 adds the cost axis. Adila et al. (2026) and Sun et al. (2026) map activation steering to post-block weight updates and rank-one edits; both live in the post-softmax route of Theorem 4, without an aperture factor. The same split bears on the weak correlation between causal-tracing localisation and edit success (Hase et al., 2023), as an interpretation rather than a theorem.

Attention-steering primitives. PASTA (Zhang et al., 2024a) rescales post-softmax weights on profiled heads, AutoPASTA (Zhang et al., 2024b) selects the span, SpotLight (Venkateswaran & Contractor, 2026) biases pre-softmax logits, and InstABoost (Guardieiro et al., 2025) adds a constant instruction-key logit bias in every layer and head, proving the tilt and reporting over-boosting. All implement Corollary 2's odds shift; their reported task gains are not adjudicated here, since the budget they spend follows from the corollary either way; we contribute the ex-ante certificate and the bypass control. Billa (2026) and Olivieri & Pérez Rodríguez (2026) predict steering effects from the logit lens and linear response, and Piontkovskaia & Nikolenko (2026) find local adaptation first-order predictable but pairwise fragile; we add a one-sided unreachability guarantee and, at finite budget, a predictor that keeps the routing exact.

M Reproducibility

Data. The structural measurements and the perplexity arm use the WikiText-103 validation split (Merity et al., 2017, CC BY-SA 3.0); the margin sweep adds HaluEval-QA (Li et al., 2023a) and TruthfulQA (Lin et al., 2022); the context-control, selector and finite-response arms use NQ-Swap (Longpre

et al., 2021). No training is performed: every experiment measures properties of frozen, publicly available pretrained models. Sample counts and sequence lengths are documented in §5 and Appendix E.

What was registered in advance. The following were tested against criteria fixed before they were run, and each criterion is stated where its result is reported, with any part of these registrations that was not run stated as an open question, so that none has to be looked up elsewhere: the reachable-set install and its gates (Appendix E), span choice (Appendix G), context control and the depth sweep (Appendix J), and the selector comparison with the held-out test of the finite-response predictor (Appendix K); two exploratory analyses in that appendix, the dose path and the float32 diagnostic, had their readings fixed before they ran and are read against them there. The frozen protocols are released with the code that implements them. Two of the reachable-set gates failed and are reported as failures in Appendix E. Two criteria were added after a first outcome had been read, and are disclosed where they bear: the depth sweep's one-block criterion, fixed once the four-point sweep had been read (Appendix J), and the span-choice baseline that decides what that result licenses (Appendix G). One rule was applied that the frozen protocols do not contain: the context-control arm selected its profiling size by a stability criterion, mean top-k overlap at least twice a random baseline and mean Spearman at least 0.4 at the smallest qualifying size, fixed before any stability number was read but absent from the frozen addendum. It is material, because it changed which head set was used: the 32-example profile that would otherwise have been taken sits at 1.76 times the random baseline, below the threshold. The saturation and utilisation sweeps, the interface comparison and the payload and path-aperture measurements were not pre-registered; they report measured values with the filters and counts their protocol sections give.

Model checkpoints. All models are publicly available on HuggingFace: GPT-2 (MIT licence), the Pythia suite at 160M and 1B (Apache 2.0, revision `step143000`), Qwen2 at 0.5B and 1.5B (Tongyi Qianwen / Apache 2.0), SmolLM2-1.7B (Apache 2.0), Phi-2 (MIT-style), OLMo-1B (Apache 2.0), OLMo-2-1B-Instruct (Apache 2.0, `allenai/OLMo-2-0425-1B-Instruct`) and Qwen2.5-3B-Instruct (Qwen Team, 2024) (Qwen licence, `Qwen/Qwen2.5-3B-Instruct`). Exact checkpoint identifiers are recorded in the experiment configuration files of the anonymised repository.

Compute environment. The measurements run on Apple Silicon (MPS backend). The structural, reachable-set, span-steering, path, logit-space and saturation measurements run in float32; the span-choice, context-control, selector and finite-response arms load bfloat16 weights, with the geometric quantities computed in float32 and the bfloat16 floor measured in Appendix K. The pipeline also runs on CUDA and on CPU for the smaller models. Float32 is required where it is used: the GPT-NeoX implementation produces NaNs in half precision on this backend, and the per-row singular-value computations need the extra precision. Dependencies are pinned via `uv.lock` (Python 3.11, PyTorch 2.10, Transformers 5.3). The seed is 1319 except where an appendix states another (the span-choice draw and its bootstrap use 20260912, Appendix G); power iteration counts, convergence tolerances, and the finite-difference step used to validate the Jacobian-vector products are reported in the relevant appendix sections.

Formal verification scope. The load-bearing statements are formalised in Lean 4 (de Moura & Ullrich, 2021) with Mathlib (Apache 2.0), zero `sorry`, axioms restricted to `propext`, `Classical.choice`, and `Quot.sound`. Table M.1 lists the correspondence; the curated import surface is `ShadeFormal.Papers.SoftmaxAperture.API`. Every theorem in Table M.1 is checked to depend on exactly `propext`, `Classical.choice` and `Quot.sound` by `ShadeFormal.Papers.SoftmaxAperture.AxiomAudit`, which prints the dependencies of every audited declaration rather than asserting them; the table entries that are definitions rather than theorems (`pathAperture`, `LogitRangeBounded`, `LikelihoodRatioReachable`, `responseNorm`, `headOutputPath`, `multiHeadOutput`, `centroidContrast`, `multiHeadReach`, `spanBiasForMass`, `protectedDirections`) name the objects the theorems are about. The aperture bound and the interface theorem are formalised about the Jacobian of an actual softmax row, not about a matrix stipulated to have the required properties: `softmaxJacobianMatrix` is $\mathrm{diag}(\mathbf{A}_i) - \mathbf{A}_i\mathbf{A}_i^\top$ built from `softmaxRow`, and `attentionResponse_row_eq_fderiv` identifies each row of $\mathcal{R}$ with the Fréchet derivative of the softmax certified by `softmaxRow_hasFDerivAt`, so the bounds are first-order statements

about softmax rather than about a named family. What remains a hypothesis is what a formalisation cannot supply: the downstream Jacobian is an arbitrary matrix rather than a particular model's, so its norm enters every downstream bound as a measured input.

The finite-path correspondence is modular, and its two machine-checked steps should be read precisely: `interface_frobenius_le_aperture` supplies the pointwise response bound at each point of the path, and `path_displacement_le_integrated_bound` supplies the calculus step, a Banach-space fundamental theorem of calculus with the norm-of-integral bound. The differentiability of the softmax path and the integrability of its velocity and of $\gamma(\mathbf{A}_t)$ enter `finite_path_aperture_bound` as hypotheses. At the sharp constant, `headOutputPath_hasDerivAt` proves that the velocity of the head output along the affine logit path is the response operator applied to δL, so `finite_path_response_bound` carries only the integrability of $t \mapsto \max_i \sigma_1(J_{i,t}U)$ as a hypothesis; continuity of σ_1 in the matrix, which discharges it, is established in Appendix B but is not in the development, so the Lean statement keeps it as a hypothesis.

The development also formalises the classical likelihood-ratio reachability equivalence to make its downstream use auditable. Novelty is not claimed for that Hilbert-simplex geometry, nor for the exponential-family facts underlying the coarse routing weights: a groupwise-constant logit bias is an exponential tilt by a partition-measurable statistic, so the group indicator is sufficient for the family and the within-group conditionals are unchanged, and the map $b \mapsto w(b)$ is the canonical natural-to-mean parameterisation of the k-atom categorical family with base measure π, a surjection onto the relative interior of the simplex that is injective modulo a common additive constant, with the log-ratio $b_g = \log(q_g/\pi_g)$ as the normaliser-one inverse. Both are formalised here only so their downstream use is auditable. The attention-specific formal content is what those weights act on: the exact span routing–payload identity, the all-budget inertness criterion, and the affine span ceiling. The Lean split places the factor $\psi(1-\psi)$ on the payload rather than the routing side: `spanPayloadContrast` is $\psi(1-\psi)(\mu_S - \mu_{\bar{S}})$ and the routing scalar is $(e^b - 1)/Z$, whose product is the $(\psi(b) - \psi)(\mu_S - \mu_{\bar{S}})$ of (5).

Not formalised. The following statements are proved in the appendix only, or enter the Lean development as hypotheses. The finite-response predictor of Appendix K: its routing factor is the per-row instance of Corollary 2, which is formalised, but its downstream linearisation and the predictor as a whole are not. Proposition A.1, the L-layer composition bound, whose block Lipschitz constants are supplied inputs. The equality case of Lemma B.2; the bound itself is `variance_le_binary_bound` for every p. The attainment clause of Corollary B.3, $2\varepsilon(1-\varepsilon)$ at the row $(1-\varepsilon, \varepsilon, 0, \ldots)$. The integrated form of Proposition B.4; only the pointwise cap $\frac{1}{2}$ is checked. Envelope invariance (Appendix I). The uniqueness clause of Theorem 6 (equality only at balanced two-key rows): the constant is machine-checked as an upper bound that every row of Table I.1 attains, not as attained only there. The equality characterisation in Lemma B.1 (trace equality exactly at rows with at most two nonzero entries). The row-level form of the equality case, checked as saturation of the composite norm, $\|\mathcal{R}\|_{F\to F} = \gamma(\mathbf{A})\|U\|_{\mathrm{op}}$, with the operator norms entering as reals; identifying that with a single row attaining both maxima is the unfolding of $\|\mathcal{R}\|_{F\to F} = \max_i \sigma_1(J_i U)$ and is not itself checked. The multi-head aggregation bound (14): only the scalar triangle inequality behind it, `multihead_triangle_l1`, is checked. Equality in Proposition C.1(ii) and the curvature of the shared-bias set. The convex-hull image step of Proposition C.2. Continuity of $t \mapsto \max_i \sigma_1(J_{i,t}U)$ along the path, proved in Appendix B and entering `finite_path_response_bound` as an integrability hypothesis. Clause (ii) of Theorem 4, the post-softmax half, which is the triangle inequality. The velocity identity in (17), the product rule along a differentiable path, which enters `finite_mixed_path_bound` as a hypothesis. The identification of the reachable set of Theorem 1 with the relative interior of the centroid hull and the equality in (3): the development proves the containment in the affine span and the surjectivity onto the open simplex, and Step 3 of the proof supplies the two-line identification. Remainder bounds for nonlinear downstream readouts.

Code and proof release. A reference implementation of the measurement pipeline (the support-aware saturation sweep, the pre-softmax intervention harness, the downstream-factor estimator, and the analysers that emit every numeric macro used in this paper) together with the Lean 4 formalisation is available for review at `https://anonymous.4open.science/r/`

Paper statement	Lean declaration(s)
Lemma B.1, trace envelope and cap	`variance_le_trace`, `weightedVariance_le_half`
Theorem 6, $d_i\gamma_i \le C(m)$ for all $m \ge 2$	`product_bound`, `product_bound_via_envelope`, `envelope_max_le_aperture_const`
Theorem 6 at the operator norm of an actual softmax row, $d_i\,\sigma_1(J_i) \le C(m)$	`product_bound_normSq`, `softmaxJacobianMatrix_deviation_top_eigenvalue_le`, `softmaxJacobianMatrix_deviation_topSingularValue_le`
Corollary B.3, the aperture bound, the distance floor and the product bound in the concentrated regime; Lemma B.2 for every p	`softmax_barrier`, `variance_le_binary_bound`, `deviation_sq_ge_concentrated`, `product_bound_concentrated`
Table I.1, every row attains $C(m)$: $m \ge 4$, $m = 3$, $m = 2$	`twoKeyRow_attains_aperture_const`, `threeKeyRow_attains_aperture_const`, `twoKeyRowInterior_attains_aperture_const`
Theorem 4, the response $\mathcal{R}$ with $\|\mathcal{R}\|_{F\to F} = \max_i \sigma_1(J_iU)$ in both directions (bound at the named constant; attainment on a maximising row along a unit vector attaining σ_1), and the factorisation through $\gamma(\mathbf{A})\|U\|_{\mathrm{op}}$	`rowwiseMulVec_frobenius_le`, `attentionResponse_row_eq_fderiv`, `responseNorm`, `interface_frobenius_le_responseNorm_sup`, `exists_unit_norm_toEuclideanLin_eq_topSingularValue`, `interface_frobenius_responseNorm_attained`, `responseNorm_isGreatest`, `interface_frobenius_le_aperture`
Theorem 4 as an abstract chain rule over normed groups (the composed constant, with Γ a free real)	`pre_softmax_interface_bound`, `interface_penalty_ratio`
Corollary 5, finite and fully mixed path bounds, and the sharp-constant bound with the velocity identified as $\mathcal{R}_t(\delta L)$	`pathAperture`, `path_displacement_le_integrated_bound`, `finite_path_aperture_bound`, `finite_mixed_path_bound`, `headOutputPath`, `headOutputPath_hasDerivAt`, `finite_path_response_bound`
Eq. (24), classical finite reachability, and Proposition C.2, the polytope correspondence and the two-level optimality certificate	`LogitRangeBounded`, `LikelihoodRatioReachable`, `likelihoodRatioReachable_iff_exists_tilt`, `weightedScore_le_of_twoLevel_certificate`
Theorem 1, factorisation, coefficient reachability, the gauge fibre $b + c\mathbf{1}$ with its normaliser-one representative, and affine-span ceiling	`groupShift_weightedValue_eq_centroid_mixture`, `groupWeight_biasForWeight_eq`, `groupShift_biasForWeight_weightedValue_eq`, `groupShift_weightedValue_sub_centroid`, `groupWeight_add_const`, `exists_const_of_groupWeight_eq`, `groupNormalizer_biasForWeight_eq_one`
Corollary C.3, the budget-free readout ceiling, the gap bound, the exponential budget cost, and readout inertness	`readout_le_readoutCeiling`, `readoutFloor_le_readout`, `readoutCeiling_sub_readout_le`, `one_sub_groupWeight_concentrate_le`, `readout_constant_iff_centroidReadout_constant`, `readout_constant_of_centroidReadout_constant`, `readout_groupShift_eq_weighted_sum`

Table M.1: Paper statements and the Lean declarations certifying their formal content, exported from `ShadeFormal.Papers.SoftmaxAperture.API` (the witnesses are declared in that namespace by `Witnesses`); statements absent from both parts are listed in the paragraph "Not formalised". Continued in the next table.

Paper statement	Lean declaration(s)
Proposition C.1, displacements in the reach submodule and the additive ceiling, independent and shared biases	`multiHeadOutput`, `centroidContrast`, `multiHeadReach`, `multiHeadOutput_sub_mem`, `finrank_multiHeadReach_le`, `multiHeadOutput_shared_sub_mem`, `finrank_multiHeadReach_shared_le`
Proposition C.4, canonical inverse bias, exact interval and box images, positive and signed capacities, signed-box support function	`spanBiasForMass`, `spanMass_spanBiasForMass_eq`, `spanBiasForMass_mem_Icc`, `spanMass_sub_image_pi`, `spanBias_positiveBox_target_feasible_iff`, `spanBias_box_target_feasible_iff`, `signedBox_inner_isGreatest`
Corollary 2, exact response and gain	`span_shift_mass`, `span_shift_mass_gain`, `spanShift_weightedValue_sub`, `norm_spanShift_weightedValue_sub`, `spanShift_weightedValue_eq_of_contrast_eq_zero`, `spanPayloadContrast_eq_zero_of_shift_eq`
Eq. (19), response ratio and its monotonicity in ψ	`span_shift_eq_firstOrder_mul_ratio`, `responseRatio_strictAntiOn_mass`, `responseRatio_symm`
Eq. (21), factored identity and the sign of the error	`responseRatio_sub_one_eq`, `one_lt_responseRatio_of_lt_exactMass`, `responseRatio_lt_one_of_half_le`
Eq. (22), balanced-row bound and validity radius	`responseRatio_half_eq_tanh`, `abs_responseRatio_half_sub_one_le`, `validity_radius_sufficient`
Eq. (20), zero-error mass: value, sandwich, limit, and the sign of the error on either side of it	`exactMass_lt_half`, `responseRatio_lt_one_of_exactMass_lt`, `responseRatio_exactMass`, `exactMass_eq_inv_sub`, `half_sub_le_exactMass`, `exactMass_strictAntiOn`, `tendsto_exactMass_zero`
Eq. (23), the bound does not degrade with the budget	`abs_spanMass_sub_le`
Worked example as witness: twin spans (inert span, nonzero contrast)	`twinSpanA_inert`, `twinSpanB_contrast`

Table M.1: Paper statements and Lean declarations, continued.

`softmax-aperture-6D50`; it will be released on a permanent public repository with a DOI upon acceptance.