SEMICLASSICAL GEOMETRY OF ATTENTION SWITCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

When one attention head is pushed toward a competing evidence span, do baseline quantities identify useful directions for controlling the answer margin? For a fixed-key, two-branch logit shift, the head's output change splits exactly into a branch-switching term and a total-evidence-mass term (a specialization of softmax differentiation, used as coordinates). *Switching geometry identifies a task-enriched control subspace.* Across eight Qwen3-8B examples (archived custom path), the three-dimensional switching normal subspace retains a mean 26.6% of squared task-gradient norm versus 0.60% for the median matched-random subspace (paired difference 0.260, 95% interval $[0.141, 0.378]$; uses each example's gradient). *Baseline-only prediction holds.* On 50 ConflictQA examples and eight Gemma heads, finite attention recomputation from the baseline gives pooled $R^2 = 0.853$ versus 0.672 for local linearization. In an exploratory native Qwen3-8B cohort (a 32-fact prefix of a 200-fact ConflictQA test set, path gate bypassed), a predictor frozen from baseline forward and backward passes achieves small median local-regime error ($|\alpha| \leq .05$) for all three steering arms—switching, fixed PCA basis, and static mean direction: geometrically grounded directions are collectively useful. *A switching-specific advantage is not established.* On those 32 facts, switching minus PCA at $\alpha = .20$ is -0.020 $[-0.068, 0.014]$ and switching minus static is -0.020 $[-0.085, 0.035]$ (not a test of equivalence); the 200-fact cohort favors switching over the fixed basis only. Switching geometry thus locates useful control directions but is not a privileged axis; path and attribution limits are in Section 5.

1 INTRODUCTION

Scaled dot-product attention is smooth, but a small change in its logits can exchange which source dominates the information a query receives (Vaswani et al., 2017). When two documents support contradictory answers, this exchange has a concrete meaning: the model changes which evidence it reads. ConflictQA (Xie et al., 2024) stages a contextual-versus-parametric knowledge conflict; our narrower intervention contrasts two marked evidence spans at one head and measures which candidate answer gains score. We treat the switching wall $\Sigma = \{G = 0\}$ (Equation (1)) as a one-attention-head, two-evidence-branch realization of a *finite-model commitment-sensitive transition window*, not as a thermodynamic singularity or a claim about an entire generated sequence. The switching susceptibility factor $2p(1-p)$ is maximal at this wall (Equation (5)); the full answer-margin response need not be. We ask whether the local geometry of this competition predicts the direction and magnitude of the resulting behavioral response, and whether a query intervention that moves across the wall can control evidence selection.

Routing geometry alone does not answer this, and the decomposition below is not by itself new machinery: on its own, the finite-key decomposition specializes standard softmax differentiation, and the frozen-gradient predictor it feeds shares its computational principle with attention-corrected approximations that explain how an output changes without saying what that change predicts or enables at the network level. A branch must carry appreciable mass, its value vector must differ from that of its competitor, and the difference must align with the downstream map to the answer. Attention weights are therefore not explanations by themselves (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), and wall proximity in logit space does not identify causal importance. A Laplace expansion of Gibbs attention motivates the local picture: at a transverse two-branch tie the responsibility is logistic

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

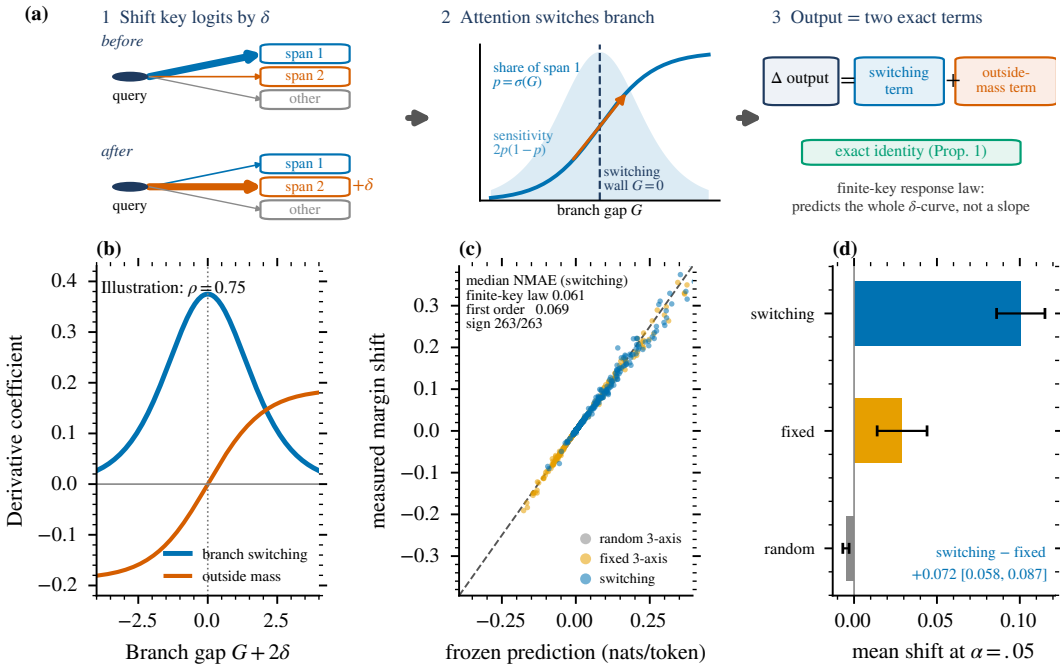


Figure 1: **Finite-key response law.** (a) A key-logit shift δ switches the attention branch; the output change splits into a switching term and an outside-mass term (Prop. 1). (b) Identity coefficients (algebra only; $\rho = 0.75$ is illustrative). (c,d) CX-ID run (Qwen3-8B native, 200 test facts, 1 seed). (d) compares switching with a fixed three-axis basis and a random basis only; no static arm was run on these 200 facts. It is a separate comparison from Fig. 4, which adds a static arm on a 32-fact subset and finds no significant switching advantage; the two cohorts are not pooled. This is an exploratory result: it rests on an in-run zero-hook identity check and frozen predictions; the full path-level validation check was not run.

in the branch action gap, and the output Jacobian has a singular part normal to the tie wall. We use this only as motivation for the coordinates; the sigmoid and the chain rule are not new.

Our claim, in one sentence: baseline switching geometry locates useful, task-enriched control directions and a baseline-only finite recomputation predicts the margin response, but switching is not shown to steer better than PCA or static directions. We relate the normal form to measurements in pretrained language models (Table 1); the contributions are:

1. *Useful control directions (Exp. 4).* The switching normal subspace retains 26.6% of squared task-gradient norm vs. 0.60% for matched-random subspaces (paired difference 0.260, $[0.141, 0.378]$; Section 4.1).
2. *Coordinates (algebra, not a new mechanism).* Under fixed-key competing-evidence interventions, the head's attention-output derivative separates exactly into branch-switching and total-evidence-mass terms (Proposition 1, Equation (5)). The conditional remainder in Appendix A.6 is not a measured error certificate.
3. *Prediction holds (D1).* Finite attention recomputation from the baseline predicts the finite-strength margin response better than same-path local linearization: $R^2 = 0.853$ versus 0.672, median per-curve $R^2 = 0.603$ (Experiment 1, Figure 2; archived custom path). In an exploratory native Qwen3-8B 32-fact cohort, a predictor frozen from baseline forward and backward passes has small median local-regime error ($|\alpha| \leq .05$) for every steering arm, not only switching (Figure 4). Operating-point finite-difference checks in Gemma and Qwen are consistent with the decomposition (Experiments 2–3).
4. *A switching-specific advantage is not confirmed (D2).* On a 32-fact prefix of the 200-fact CX-ID test set, no paired difference between switching and a static mean direction or a fixed PCA basis is significant (every paired interval contains 0; not a test of equivalence; Figure 4), and

Table 1: Correspondence between contributions and the paper. Every claim in the introduction is supported by the listed theory, experiment, and figure.

Contribution	Theory / experiment	Figure
C0 useful control directions	Exp. 4 (§4.1); 4-arm cohort	Fig. 4
C1 finite-key decomposition (algebra)	Prop. 1, Eq. (5)	Fig. 1
C2 prediction holds (D1)	Exp. 1–3 (§3.1); 4-arm cohort	Figs. 2, 4; Tab. 18
C3 no confirmed switching advantage (D2)	Exp. 4–5 (§4.2); 4-arm cohort	Figs. 5, 4
CX-3 strength summaries (exploratory)	CX-3 run (§3)	Fig. 21

a coefficient-held-out controller does not establish an advantage over the static tune-gradient baseline (Figure 5).

Experiments 1–3 test D1 and Experiments 4–5 test D2 on ConflictQA (Xie et al., 2024) with Gemma 4 E4B-it and Qwen3-8B (Gemma Team, 2026; Yang et al., 2025); Experiments 1–3 edit one layer–head site at a time, and Experiments 4–5 jointly edit three frozen Qwen sites. Experiments 1–5 use an archived custom execution path whose checkpoint correspondence was not fully audited; the CX-ID run on the native Qwen3-8B checkpoint (Section 3.1) recorded a zero-strength identity check and froze predictions before measuring (its path-gate receipt was bypassed). Two results bound what the claims mean. First, replacing the example-specific downstream covector by a cross-example mean lowers pooled R^2 from 0.853 to 0.121 (leave-one-example-out: 0.105; Appendix M.2), so the predictive value rests on example-specific downstream sensitivity, not on switching geometry itself. Second, the only positive switching contrast—$+0.072$ nats/token (95% interval $[0.058, 0.087]$) against a fixed three-axis basis on the 200-fact CX-ID cohort—is not established on the separate 32-fact prefix that adds a static arm (Fig. 4); the cohorts differ, so this is not a within-cohort reversal, and the archived controller's switching-minus-static difference is 0.0010 ($[-0.0235, 0.0260]$). Recapture deviations, the auxiliary runners, and the post-hoc Mistral replication are reported in Appendix D.

2 Finite-key steering: two mechanisms

A multi-branch Laplace expansion (Appendix A.1) motivates the switching-wall geometry: near a transverse two-branch tie, the leading singular output derivative has rank at most one and is normal to the tie wall. For more than two branches, the leading singular term is $D_x U_\epsilon = \mathrm{Cov}_\pi(V_a, \nabla\Phi_a)/\epsilon + \mathcal{O}(1)$, and a displacement preserving every relative branch action to first order annihilates this term. We do not treat this asymptotic normal form as an empirical claim; it only motivates the shape of the switching transition, and all finite-key derivations use exact softmax identities.

Fix an attention head and a query position, and let $\mathcal{J}$ be the set of key positions accessible from that query position. Write $q \in \mathbb{R}^{d_h}$ for the query vector and $k_j \in \mathbb{R}^{d_h}$ for the key vector at position j, where d_h is the dimension of these vectors. For standard scaled dot-product attention, the logits are $s_j(q) = \langle q, k_j\rangle/\sqrt{d_h}$. Let v_j denote the value vector associated with key position j, as used in the attention-weighted sum before the attention output projection.

Consider two nonempty, disjoint subsets $B_0, B_1 \subset \mathcal{J}$ consisting of the token positions in two competing evidence passages. Holding the keys and these subsets fixed, define

$$L_a(q) = \log \sum_{j\in B_a} e^{s_j(q)}, \qquad G(q) = L_1(q) - L_0(q), \qquad \Sigma = \{q \in \mathbb{R}^{d_h} : G(q) = 0\}. \quad (1)$$

Here L_a is the log-partition of branch $a \in \{0, 1\}$. Because all weights share one softmax denominator, G is the log ratio of total attention mass on B_1 to that on B_0: $G > 0$ favors branch 1, $G < 0$ branch 0, and $G = 0$ equal mass. The set Σ is the *branch-LSE switching wall*; it is a smooth hypersurface wherever $\nabla_q G \neq 0$.

We intervene on the logits by adding $+\delta$ to every logit in B_1 and $-\delta$ to every logit in B_0, leaving all other logits, the query, and the keys fixed. The uniform shifts give $L_0(q;\delta) = L_0(q) - \delta$ and $L_1(q;\delta) = L_1(q) + \delta$, hence the exact identity $G(q;\delta) = G(q) + 2\delta$.

The fraction of attention mass assigned to B_1, conditional on attending to $B_0 \cup B_1$, is $p_\delta := \mathrm{sigmoid}(G + 2\delta)$.

Let $R = \mathcal{J} \setminus (B_0 \cup B_1)$ be the remaining keys, $Z_a = e^{L_a}$, and $Z_R = \sum_{j \in R} e^{s_j}$. The intervention is

$$s_j(\delta) = s_j + \delta\big(\mathbf{1}_{j \in B_1} - \mathbf{1}_{j \in B_0}\big). \tag{2}$$

Define the conditional branch values $\overline{v}_a = Z_a^{-1} \sum_{j \in B_a} e^{s_j} v_j$ and analogously $\overline{v}_R$. Let $\rho_\delta := \frac{e^{-\delta} Z_0 + e^{\delta} Z_1}{Z_R + e^{-\delta} Z_0 + e^{\delta} Z_1}$, $o_\delta := \sum_j \alpha_j(\delta) v_j$, where $\alpha_j(\delta) = e^{s_j(\delta)} / \sum_k e^{s_k(\delta)}$. ρ_δ is the total attention mass on $B_0 \cup B_1$.

Proposition 1 (Exact routing and output law)**.** *With $v_{B,\delta} = (1 - p_\delta)\overline{v}_0 + p_\delta \overline{v}_1$, one has*

$$o_\delta = \rho_\delta v_{B,\delta} + (1 - \rho_\delta)\overline{v}_R, \tag{3}$$

$$\frac{\mathrm{d}o_\delta}{\mathrm{d}\delta} = \underbrace{2\rho_\delta p_\delta (1 - p_\delta)(\overline{v}_1 - \overline{v}_0)}_{\textit{switching term}} + \underbrace{\rho_\delta (1 - \rho_\delta)(2p_\delta - 1)(v_{B,\delta} - \overline{v}_R)}_{\textit{outside-mass term}}. \tag{4}$$

The outside term vanishes at the branch tie $p_\delta = 1/2$ or when $\rho_\delta = 1$, but cannot generally be discarded away from the tie.

For a scalar downstream margin $M(o)$, let $\lambda_\delta = \partial M / \partial o$ at the intervened output. Then the exact local causal response is

$$\kappa_\delta := \frac{\partial M}{\partial \delta} = 2\rho_\delta p_\delta (1 - p_\delta) \left\langle \lambda_\delta, \overline{v}_1 - \overline{v}_0 \right\rangle + \rho_\delta (1 - \rho_\delta)(2p_\delta - 1) \left\langle \lambda_\delta, v_{B,\delta} - \overline{v}_R \right\rangle. \tag{5}$$

Only the mixing factor $2p(1-p) = [2\cosh^2(G/2)]^{-1}$ is universally maximized at $G = 0$; neither the entire κ_δ nor its magnitude must peak there.

From identity to network-level test. The algebra above is exact softmax differentiation for fixed keys, values, and masks. Its network-level content is the falsifiable composition with the downstream margin: the two separately controllable group log-partitions change a selected head output, and the frozen network maps that output to the teacher-forced score in Equation (7).

For standard dot-product attention with fixed keys, the exact finite-key wall normal is

$$\nabla_q G = \frac{\overline{k}_1 - \overline{k}_0}{\sqrt{d_h}}, \qquad \overline{k}_a = \sum_{j \in B_a} \mathrm{softmax}_{B_a}(s_j) k_j. \tag{6}$$

Thus a tangent $t \perp \nabla_q G$ has $D_t G = D_t p = 0$ to first order; query perturbations can still change within-branch values and outside mass (Appendix A.5).

3 Does the decomposition predict measured responses?

This section tests the paper's two claims and states the answer to each up front. **D1 (prediction):** does a baseline-only finite recomputation of the edited head, passed through the output law of Equation (3), predict the finite-strength margin response? It does—better than local linearization in Experiment 1, and for every steering arm in the matched one-cohort comparison (Fig. 4). **D2 (switching-specific control):** does steering along the switching geometry (Equations (9) and (10)) move the margin more than a static or fixed-basis direction? This is *not* confirmed: switching exceeds a random fixed basis on the 200-fact CX-ID cohort, but not a static mean direction or the PCA basis on matched facts, nor on the matched-norm oracle cohort. Operating-point derivative consistency (Experiment 2) and query-direction sensitivity (Experiment 3) are in Appendix M.2.1 and M.2.3.

Shared setup. We use the GPT-4 PopQA subset of ConflictQA, with tokenizer offsets mapping the two conflicting evidence spans to B_0, B_1. For candidate $y^{(c)}$ with T_c tokens, define

$$S_\delta(y^{(c)} \mid x) = \frac{1}{T_c} \sum_{t=1}^{T_c} \log P_\delta(y_t^{(c)} \mid x, y_{<t}^{(c)}), \qquad M(\delta) = S_\delta(y^{(1)} \mid x) - S_\delta(y^{(0)} \mid x). \tag{7}$$

The outcome is a teacher-forced sequence-score margin in nats per token, not free-generation accuracy. An attention head is indexed by $u = (\ell, h)$; Experiment 1 edits one head at a time with the branch-logit shift of Equation (2). The eight Gemma heads were transferred from an independent synthetic evidence-routing experiment and never selected on ConflictQA (Appendix C). Intervals are example-cluster percentile bootstrap intervals (2,000 resamples).

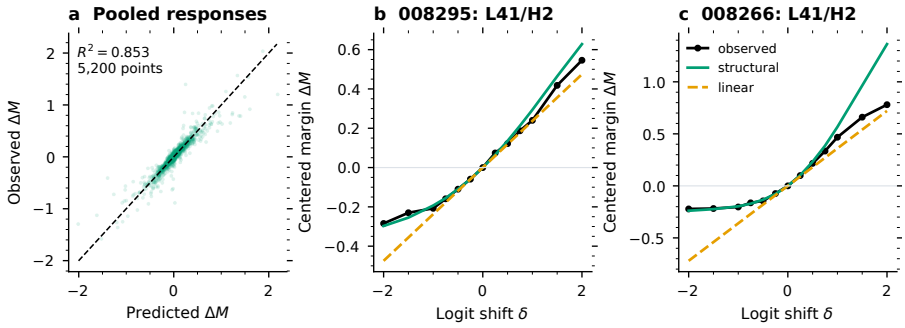


Figure 2: **Finite-strength response prediction on Gemma-4-E4B-it (Experiment 1).** (a) Pooled parity of observed versus structurally predicted centered margin ΔM over all 5,200 centered points (8 frozen heads $\times$ 50 ConflictQA examples $\times$ 13 logit shifts); $R^2 = 0.853$; the dashed line is prediction = observation. (b,c) Two illustrative example–head response curves (examples 008295 and 008266, head L41/H2): observed centered margin (markers), structural prediction (solid), and linear comparator (dashed) against the logit shift δ. The illustrative curves are not independent replications and do not establish uniform per-curve accuracy; the paired error improvement is reported in the text.

Provenance. Experiments 1–5 were measured on a custom attention path whose correspondence to the native checkpoint has not been fully audited. The CX-ID run (Section 3.1) is the one cohort whose zero-strength hook identity and derivative checks were recorded and whose predictions were frozen before outcomes were read (its separate path-gate receipt was bypassed).

3.1 Experiment 1: Finite-strength prediction from baseline

Question and prediction. Can the unmodified computation predict a head's response over a finite range of branch-logit shifts, and does retaining the nonlinear local attention response improve on a linear predictor? We test all eight fixed Gemma heads on 50 deterministically shuffled eligible ConflictQA examples, without outcome filtering; each head is edited separately at 13 strengths in $[-2, 2]$ (400 curves, 5,200 points). From baseline gradients and Equation (3), the structural predictor $\widehat{M}_u(\delta) = M(0) + \sum_c (2c-1) \sum_t \langle \lambda_{c,u,t}(0), \widetilde{o}_{c,u,t}(\delta) - o_{c,u,t}(0) \rangle$ truncates the downstream map at first order but retains the nonlinear local softmax (Appendix E.4). The linear comparator is the zero-slope of this predicted curve. Both use baseline information only.

Results. Across all 5,200 centered points, structural prediction achieves $R^2 = 0.853$ versus 0.672 for the linear comparator (Figure 2). MAE decreases from 0.0412 to 0.0279 nats per token (32.3% reduction); the paired mean gain is 0.0133 (95% interval $[0.0101, 0.0168]$). Mean within-example Spearman correlation is 0.899, top-1 accuracy is 82%, and top-three coverage is 98%. Median per-curve R^2 is 0.603 overall, 0.935 in the largest observed effect-size quartile, and -0.231 in the smallest. In the secondary analysis, 38 of the 47 raw-margin curves that cross zero are also predicted to cross (precision 90.5%, recall 80.9%; Appendix F.4).

Figure 3 shows the Experiment 1 sweep head by head, in the head-evolution view of Sagitova et al. (2026): for Gemma-4 three heads (L29H4, L41H2, L29H2) sweep long arcs along the prediction = measurement diagonal while the remaining five stay near the origin, so the finite-strength effect is carried by a minority of heads and the structural predictor tracks exactly those; for Qwen3-8B the

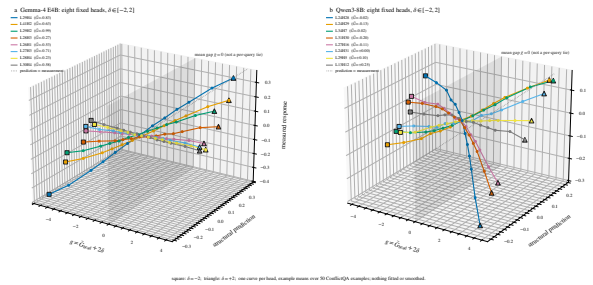


Figure 3: **Head-by-head evolution across the zero of the mean branch gap.** Each curve is one of the eight fixed heads of (a) Gemma-4-E4B-it and (b) Qwen3-8B, swept over $\delta \in [-2, 2]$ (square: $\delta = -2$; triangle: $\delta = +2$; 13 shifts) and plotted in $(g,\ \text{structural prediction},\ \text{measured response})$, where $g = \bar{G}_{\text{head}} + 2\delta$ uses the head's mean baseline branch gap over the 50 examples; prediction and measurement are example means from the archived Experiment 1 curves. The translucent sheet (labelled "mean gap $\bar{g} = 0$") marks $g = 0$ of this head-averaged mean branch gap, not a verified tie of every contributing attention distribution (averaging gaps across examples does not preserve the per-query routing tie); the dotted line is prediction = measurement. Heads are ordered in the legend by swept arc length. Nothing is fitted or smoothed.

arcs split in sign (three heads respond negatively to positive δ) and one head (L13H12) moves off the diagonal, which is where the frozen-covector prediction is zero but the measurement is not.

Interpretation. Retaining nonlinear attention improves finite-strength prediction, especially for larger effects, within the independently selected eight-head pool; it does not imply accurate weak-head predictions or discovery among all heads. A post-hoc fixed-covector ablation (Appendix M.2) shows that example-specific downstream sensitivity drives the gain over a fixed mean covector. Operating-point derivative consistency across branch gaps is in Appendix M.2.1.

A validated measurement path (exploratory CX-ID run). Experiments 1–5 used a custom attention path whose native-checkpoint correspondence has not been audited. We therefore report a separate cohort with in-run measurement checks recorded before any intervention outcome was read. On the native `Qwen/Qwen3-8B` checkpoint (revision `b968826d`) with three frozen CX-3 heads, ConflictQA facts were split into discovery 100, controller 200, validation 50, and test 200. Three in-run checks (exact zero-strength hook agreement, autodiff–finite-difference agreement within 0.3%, and baseline predictions committed before the first test forward) passed; the separate path-gate receipt was bypassed, so they do not certify the intervention map beyond zero strength (details in Appendix F.2). For each test fact we score a predictor in the pre-registered local regime $|\alpha| \le 0.05$, i.e. $\alpha \in \{.005, .05\}$ (the $\alpha = .20$ interventions are measured but excluded from this endpoint):

$$\mathrm{NMAE}_f \;=\; \frac{\mathrm{mean}_\alpha\,|\hat{\Delta}_{f,\alpha} - \Delta_{f,\alpha}|}{\max\!\big(\mathrm{mean}_\alpha|\Delta_{f,\alpha}|,\ \tau\big)}, \qquad \tau = 0.0086 \text{ nats/token},$$

with τ fixed on validation before any test forward. The endpoint is the median of NMAE_f over the 200 test facts: 0.061 for the exact-local predictor and 0.069 for the first-order predictor (switching arm). Sign accuracy is computed separately on the individual (f, α) rows in the same local regime whose measured shift satisfies $|\Delta_{f,\alpha}| \ge \tau$; 263 of the 400 such rows qualify, and the exact-local predictor recovers the sign on all 263 (1.00). The paired contrast between the switching arm and a genuine fixed three-axis query basis is $+0.072$ nats/token at $\alpha = .05$, with a paired fact-bootstrap 95% interval $[0.058, 0.087]$ (Holm-adjusted $p < 0.001$ against a pre-registered practical margin of 0.02). These are the only numbers for which gradient-to-finite-response agreement, native-versus-hooked agreement, and prediction freezing were verified on the measured cohort; they remain exploratory (one model, one seed, 200 test facts, one in-distribution environment). Figure 11 and Appendix F.2 give all strengths, side effects, and artifact hashes. What the cell establishes is therefore narrow: on this cohort the frozen switching predictor tracks the measured response, and the measured response under the switching direction exceeds that under the matched fixed basis by more than the pre-registered margin. It does not reconcile the matched-norm oracle comparison below, on which

switching has the lowest mean shift of the three directions (no interval reported), and it does not supply the separate native-path gate receipt that the full design requires (Appendix F.2).

One-cohort matched comparison including a static direction (exploratory). To test the switching contrast against a static direction on identical facts, the CX-ID protocol was re-run with four arms—no-op, switching, the fixed three-axis PCA basis, and a static mean direction—on a deterministic 32-fact prefix of the 200 test facts, on the same path, with the same model revision and seed, and with the path gate again bypassed (Fig. 4). On the 192 shared (fact, arm, α) keys, the switching and PCA shifts and the baseline margins reproduce the 200-fact run exactly. On this subset no advantage of the switching arm over either comparator is established. The response is the signed margin shift $\Delta_{f,\alpha}$ defined above (intervened minus baseline margin); the one-sided aggregate of Experiments 4–5 was not computed on this cohort. Paired fact-bootstrap 95% intervals for switching minus static are -0.013 $[-0.030, 0.002]$ at $\alpha = .05$ and -0.020 $[-0.085, 0.035]$ at $\alpha = .20$; for switching minus PCA they are -0.003 $[-0.016, 0.005]$ and -0.020 $[-0.068, 0.014]$. The frozen baseline-only predictor has small median local NMAE for every arm (0.011, 0.002, 0.037 for switching, PCA, and static, with correct signs on all qualifying rows); this is a median over facts in the local regime ($\alpha = .20$ excluded), not a per-fact or per-strength guarantee. Two conclusions follow. The local prediction result (D1) holds for each arm in this sense. A switching-specific advantage (D2) is *not* confirmed in the matched comparison that includes a static direction. The $+0.072$ contrast above is specific to the 200-fact cohort against the fixed basis. A concentration of that contrast in a subset of facts is a hypothesis we have not tested. Natural-state retention was not measured on this cohort; the CX-3 examples share no facts with it.

Matched-norm oracle comparison (exploratory). On the 32 frozen-test CX-3 examples at identical injection norm and $\alpha = .05$, the mean margin shift was .077 (switching), .090 (fixed3), .089 (static), and .171 (full-gradient oracle); the switching subspace retains 0.23 of the gradient. Switching thus has the lowest point estimate of the three directions on this cohort (no interval is reported), unreconciled with the CX-ID contrast above. Executed sources match by SHA-256; native-versus-hooked agreement was not measured (Appendix O.1).

4 Can switching geometry control held-out behavior?

We test two stronger statements: that the switching wall normal contains a useful fraction of the downstream task gradient, and that a controller fitted only on switching normals transfers to held-out examples. Both stages use the frozen strong Qwen heads

$$\mathcal{H}_{\text{ctrl}} = \{(24, 28), (34, 7), (24, 29)\}, \tag{8}$$

whose post-RoPE query coordinates form the joint space $\mathcal{U} = \bigoplus_{h \in \mathcal{H}_{\text{ctrl}}} \mathbb{R}^{128} \simeq \mathbb{R}^{384}$.

4.1 Experiment 4: a task-enriched switching subspace

Setup. For input x, let $n_h(x) = \nabla_{q_h} G_h / \|\nabla_{q_h} G_h\|$ and define

$$S_{\text{sw}}(x) = \text{span}\{e_h \otimes n_h(x) : h \in \mathcal{H}_{\text{ctrl}}\} \subset \mathcal{U}. \tag{9}$$

The retained gradient fraction is $R_x(S) = \|P_S g_x\|^2 / \|g_x\|^2$ where $g_x = \nabla_z F_x(0)$ and $F_x(z)$ is the teacher-forced margin under additive joint-query intervention. We compare S_{sw} with 64 block-matched random subspaces per example and a tangent comparator (Appendix M.3).

Results. The three-dimensional switching subspace retains a mean 26.6% of squared task-gradient norm (95% interval $[14.7, 38.3]\%$), compared with 0.60% for the median matched-random subspace and 0.47% for the tangent subspace. The paired switching-minus-random difference is 0.260 ($[0.141, 0.378]$); none of the 64 random draws reaches the switching value in any of the eight examples. At matched finite control norm ($\alpha = .05$), mean causal efficiency is 0.1597 ($[0.0570, 0.2761]$) for the switching projection vs. 0.0173 for median random and 0.0280 for tangent; the full-gradient oracle is 0.2566 ($[0.1235, 0.4169]$). The local normals thus identify a strongly task-enriched search subspace, but this diagnostic uses g_x on every evaluated example.

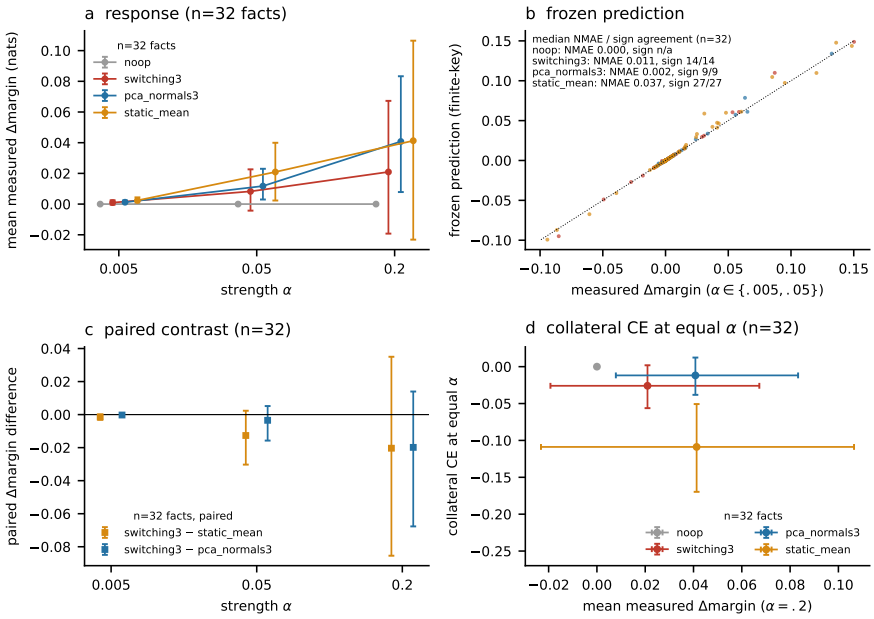


Figure 4: **One cohort, four arms: local prediction holds per arm; no significant switching advantage over static or PCA.** Native Qwen3-8B, ConflictQA, seed 11, $n = 32$ test facts: a separate comparison from Fig. 1(d), whose $+0.072$ switching-minus-fixed contrast is on all 200 CX-ID facts without a static arm and is not re-estimated here. The 32 facts are a subset of those 200, on the same path, with the path gate bypassed; exploratory. All panels use the same facts and 95% paired fact-bootstrap intervals (10k resamples). (a) Mean signed margin shift $\Delta_{f,\alpha}$ (intervened minus baseline) versus α for no-op, switching, the fixed PCA basis, and a static mean direction. (b) Frozen baseline-only predictions versus measured shifts at $\alpha \in \{.005, .05\}$. (c) Paired differences between switching and static, and between switching and PCA; every interval contains 0 (e.g. switching$-$PCA at $\alpha = .20$: -0.020 $[-0.068, 0.014]$). (d) Change in mean teacher-forced candidate cross-entropy versus margin shift at equal $\alpha = .20$. This is a side-effect measurement, not a utility (D3) evaluation, which was not run. Natural-state retention and the Experiments 4–5 aggregate were not computed on this cohort. Source files and hashes are listed in the supplementary material.

4.2 Experiment 5: A frozen three-coefficient controller

Protocol. We remove the test-example task gradient and the artificial tie preload. For each of 16 tune examples, let Q_e contain the three block-supported unit switching normals from zero-increment calibration, and let $b_e = Q_e^\top \nabla_z F_e(0)$. The frozen controller is

$$\widehat{a}_{\mathrm{sw}} = \frac{\bar{b}}{\|\bar{b}\|}, \qquad \bar{b} = \frac{1}{16}\sum_e b_e; \qquad u_e = Q_e \widehat{a}_{\mathrm{sw}}. \tag{10}$$

No test gradient, strength selection, gap preload, or head selection is used on the 32 test examples. Controls: seeded tangent bases, adaptive-random bases, cyclically shuffled switching normals, equal positive normal coefficients, and the normalized 384-dimensional mean tune task gradient.

Results. At norm fraction .05, the frozen switching controller has mean one-sided response .0831 ($[.0523, .1136]$) and improves 24/32 test examples. The static mean tune gradient achieves .0821 ($[.0551, .1088]$), improving 26/32. The paired switching-minus-static difference is 0.0010 ($[-0.0235, 0.0260]$), an interval spanning zero; the switching–static difference is not significant, and the geometry-informed controller does not surpass the static baseline (this is not a test of equivalence). Geometry-disrupting controls (tangent, adaptive-random) have intervals that exclude positive values. Full results are in Table 19 (Appendix H).

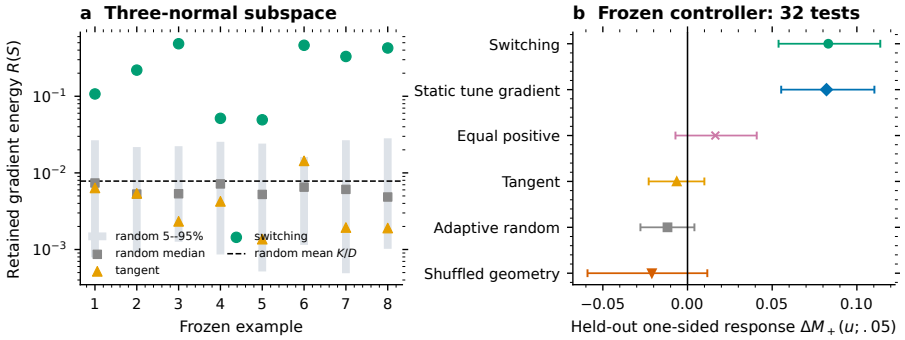


Figure 5: **Held-out control on Qwen3-8B: no significant switching–static difference (paired 95% interval spans zero).** (a) Improved/worsened/zero one-sided margin changes per frozen controller (32 held-out examples, $\alpha = .05$). (b) Post-hoc means with example-bootstrap 95% intervals. The paired switching-minus-static mean difference is 0.0010 ($[-0.0235, 0.0260]$).

5 Related Work

Log-partition gradients and covariances are standard exponential-family identities (Wainwright & Jordan, 2008), and the multi-branch expansion applies Laplace's method (Wong, 2001). Circuit analyses separate query–key routing from value/output information (Elhage et al., 2021); AtP* already recomputes the attention softmax for query/key interventions before applying a downstream gradient (Kramár et al., 2024), so this step is not our contribution. Attention reweighting, dynamic steering, and activation-direction steering (Zhang et al., 2024; Wang et al., 2025; Su & Liu, 2026; Rimsky et al., 2024; Zhao et al., 2025) are related but not novel here; we test a branch-logit response law, not superiority over these controllers. Causal tracing, activation patching, and circuit discovery (Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Zhang & Nanda, 2024) localize components; we parameterize within localized heads. We test whether crossing a competing-evidence wall changes the teacher-forced candidate-score preference. Extended related work is in Appendix L.

The wall $\Sigma = \{G = 0\}$ is the one-head competing-evidence slice of the finite-model commitment-sensitive transition-window framing, rather than evidence for a sharp phase transition in whole-model generation. The switching susceptibility factor $2p(1-p)$ peaks at the tie, but outside mass, value contrast, and the downstream map can prevent a peak in the answer margin (Equation (5)); whether control is stronger near the wall remains open (limits: Appendix M.4).

6 Conclusion

The finite-key decomposition is an exact identity for the stated local logit intervention; finite-strength prediction freezes the downstream covector and is approximate. Archived custom-attention measurements show improved centered-response prediction and wall-normal sensitivity versus sampled tangent controls. The held-out three-coefficient experiment is exploratory; its post-hoc one-sided comparison does not establish superiority, equivalence, or non-inferiority relative to the static tune-gradient baseline. Checkpoint-matched fidelity and joint-query derivative consistency remain unresolved for Experiments 1–5. The native-checkpoint run with in-run checks and pre-frozen predictions (Section 3.1) favors switching over a fixed query basis on one model, seed, and comparator; the hooked matched-norm CX-3 comparison does not. A pre-registered multi-model, multi-basis test with a utility endpoint remains open.

AI USE STATEMENT

During this revision, ChatGPT/Codex assisted with mathematical exposition and proof revision, inspection of the experiment design and notebooks, reanalysis and interpretation of recorded results, manuscript editing, figure-generation and auditing code, and reference checks. The reported numerical measurements come from model-evaluation result files; the figures and tables are generated from those measurements, not from generative image models. The accompanying analysis script checks the recorded identities, selections, sample counts, and summary statistics. The authors are responsible for the final scientific claims and artifacts.

ETHICS STATEMENT

The experiments use public models and a public knowledge-conflict benchmark, with no new human-subject data collection. Routing evidence can amplify either accurate or inaccurate sources; a positive candidate margin is not a truthfulness or safety guarantee. We study controlled internal interventions, not a deployed decision system.

REPRODUCIBILITY STATEMENT

Appendix A states assumptions and proves the mathematical claims. Appendix C documents the original Gemma and Qwen head-selection records, including the selected Gemma scores. Appendices D–G specify preprocessing, intervention positions, finite differences, metrics, and numerical limitations. The source package includes the experiment notebooks, discovery tables, evaluation CSVs, controller splits, and a CPU script that regenerates all figures and audits the principal reported statistics. The exploratory Qwen two-coordinate audit (Appendix M.2.3) had an invalid runtime coordinate check; corrected attempts produced no verifier-passing dataset, and its response table is retracted.

REFERENCES

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2304.14997.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2402.03902.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. arXiv:2309.08600.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Nelson Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. arXiv:2106.02997.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021. arXiv:2012.14913.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*, 2023. arXiv:2304.14767.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2305.00586.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *Proceedings of AISTATS*, 2019. arXiv:1806.01316.

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

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, 2022. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. arXiv:2210.07229.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1905.10650.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024. arXiv:2311.03658.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

M. Sagitova, O. Duranthon, and L. Zdeborová. Specialization of softmax attention heads: insights from the high-dimensional single-location model. *arXiv preprint arXiv:2603.03993*, 2026.

Ye Su and Yong Liu. Geometric capacity of transformers: A tropical geometry perspective. *arXiv preprint arXiv:2604.14727v2*, 2026. URL https://arxiv.org/abs/2604.14727v2.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://functions.baulab.info/.

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. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2004.12265.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 2019. arXiv:1905.09418.

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. URL https://doi.org/10.1561/2200000001.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. arXiv:2211.00593.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*, 2025. URL https://arxiv.org/abs/2410.12299.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

R. Wong. *Asymptotic Approximations of Integrals*, volume 34 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2001. doi: 10.1137/1.9780898719260. URL https://epubs.siam.org/doi/book/10.1137/1.9780898719260.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. arXiv:2309.16042.

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. *arXiv preprint arXiv:2311.02262*, 2024. URL https://arxiv.org/abs/2311.02262.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of NAACL-HLT*, 2025. URL https://aclanthology.org/2025.naacl-long.264/.

A Assumptions, proofs, and geometric interpretation

A.1 Formal statement: multi-branch normal form

Let $x \in X \subset \mathbb{R}^p$ be a query-state coordinate, and let Y be an r-dimensional key manifold with smooth positive density $\mathrm{d}\mu$. For compatibility $A(x,y)$, positive amplitude a, and value $V \in \mathbb{R}^m$, consider

$$Z_\varepsilon(x) = \int_Y a(x,y) e^{A(x,y)/\varepsilon}\,\mathrm{d}\mu(y), \qquad U_\varepsilon(x) = \frac{\int_Y V(x,y) a(x,y) e^{A(x,y)/\varepsilon}\,\mathrm{d}\mu(y)}{Z_\varepsilon(x)}. \tag{11}$$

Suppose finitely many separated, nondegenerate maximizing branches $y_a(x)$ coexist near x_0. Write $\Phi_a(x) = A(x, y_a(x))$, $H_a = -D_y^2 A(x, y_a) > 0$, and $V_a = V(x, y_a)$. If $\mathrm{d}\mu = \rho_\mu(y)\,\mathrm{d}y$, their Laplace prefactors are $B_a = a(x, y_a)\rho_\mu(y_a)/\sqrt{\det H_a}$.

Theorem 1 (Multi-branch normal form)**.** *Under the uniform smoothness, localization, and separation assumptions below, locally uniformly in x,*

$$U_\varepsilon(x) = \sum_{a=1}^{K} \pi_a^\varepsilon(x) V_a(x) + \mathcal{O}(\varepsilon), \qquad \pi_a^\varepsilon = \frac{B_a e^{\Phi_a/\varepsilon}}{\sum_b B_b e^{\Phi_b/\varepsilon}}. \tag{12}$$

For two branches, put $g = \Phi_1 - \Phi_0$ and $\beta = \log(B_1/B_0)$. If $g(x_0) = 0$ and $\mathrm{d}g(x_0) \neq 0$, then

$$\pi_1^\varepsilon = \mathrm{sigmoid}(g/\varepsilon + \beta), \qquad D_x U_\varepsilon = \frac{\pi_0^\varepsilon \pi_1^\varepsilon}{\varepsilon}(V_1 - V_0)(\nabla_x g)^\top + \mathcal{O}(1). \tag{13}$$

The leading derivative is an outer product: every input displacement is read through the scalar $\mathrm{d}g$, and the singular output response lies along $V_1 - V_0$. Its rank is at most one; along directions tangent to $g = 0$ the singular $\mathcal{O}(\varepsilon^{-1})$ term vanishes, so $D_x U_\varepsilon = \mathcal{O}(1)$ there rather than identically zero. The transition has width $\mathcal{O}(\varepsilon)$, and its equal-responsibility center is $g + \varepsilon\beta = 0$, not necessarily $g = 0$ at finite ε. The $\mathcal{O}(1)$ remainder matters: differentiating the uniform $\mathcal{O}(\varepsilon)$ output error can cost a factor ε^{-1}. The proof is below.

Relation to standard attention. For logits $s_j = \langle q, k_j\rangle / \sqrt{d_h}$, the choices $A(q,j) = \langle q, k_j\rangle / d_h$ and $\varepsilon = d_h^{-1/2}$ give $s_j = A/\varepsilon$ exactly. This is an algebraic identification, not evidence that increasing dimension automatically hardens attention: normalized action gaps may also shrink as $d_h^{-1/2}$. For finite keys no integral approximation is needed. Individual maximum-logit ties define a polyhedral routing skeleton; a semantic branch-LSE wall is generally curved and need not coincide with any single-key tie (Su & Liu, 2026). The theorem describes switching between separated, nondegenerate maxima as their values cross. Degenerate critical points and optical caustics are outside its scope. The main text uses this asymptotic normal form only as motivation for the local geometry: the finite-key derivation and all measurements there instead use the exact softmax identities of Section 2.

A.2 Uniform multi-branch Laplace expansion

The following sufficient hypotheses make Theorem 1 precise. Take a relatively compact neighborhood U of x_0 in X. Assume:

(A1) Y is a compact smooth manifold without boundary, with a smooth positive density. The functions A, a, V are smooth on a neighborhood of $\overline{U} \times Y$, and $a > 0$. Compactness can be replaced by uniform exponentially small tail bounds for both the integrals and their first parameter derivatives.

(A2) At x_0, $A(x_0, \cdot)$ has exactly K global maximizers $y_1^0, \ldots, y_K^0$. They lie in disjoint coordinate neighborhoods and have negative-definite Hessians.

(A3) After shrinking U, the critical branches $y_a(x)$ remain in those neighborhoods, their matrices $H_a = -D_y^2 A(x, y_a(x))$ are uniformly positive definite, and outside the neighborhoods

$$A(x,y) \le \max_a \Phi_a(x) - \gamma$$

for a uniform $\gamma > 0$.

The branches need not all remain global maximizers away from the tie; retaining the subdominant local maxima is essential for a uniform formula across it. Smoothness is a convenient sufficient assumption, not an optimal regularity claim.

Choose smooth cutoffs χ_a equal to one near each branch and with disjoint supports. In coordinates centered at $y_a(x)$,

$$A(x, y_a(x) + u) = \Phi_a(x) - \tfrac{1}{2} u^\top H_a(x) u + \mathcal{O}(\|u\|^3). \tag{14}$$

Set $u = \sqrt{\varepsilon}\, z$. Taylor expansion of the phase and amplitude, integration of Gaussian moments, and a uniform localization estimate give

$$I_a(x, \varepsilon) := \int \chi_a a e^{A/\varepsilon} \, \mathrm{d}\mu = (2\pi\varepsilon)^{r/2} e^{\Phi_a/\varepsilon} \left(B_a + \varepsilon C_a + \varepsilon^2 R_a \right), \tag{15}$$

$$J_a(x, \varepsilon) := \int \chi_a V a e^{A/\varepsilon} \, \mathrm{d}\mu = (2\pi\varepsilon)^{r/2} e^{\Phi_a/\varepsilon} \left(B_a V_a + \varepsilon D_a + \varepsilon^2 S_a \right). \tag{16}$$

Here $B_a = a(x, y_a)\rho_\mu(y_a)/\sqrt{\det H_a}$, the coefficients are smooth, and the displayed remainders and their first x derivatives are bounded uniformly *after extracting the exponential*. One obtains these derivative bounds by applying the same localized expansion to the parameter-differentiated integrals. Compactness and the action gap make the complementary integrals exponentially small relative to the retained sum, including any polynomial ε^{-1} factors introduced by differentiation.

Let $W = \sum_a B_a e^{\Phi_a/\varepsilon}$. Since B_a are uniformly positive,

$$Z_\varepsilon = (2\pi\varepsilon)^{r/2} W \big(1 + \mathcal{O}(\varepsilon)\big), \tag{17}$$

$$\varepsilon \log Z_\varepsilon = \tfrac{r}{2} \varepsilon \log(2\pi\varepsilon) + \varepsilon \log W + \mathcal{O}(\varepsilon^2). \tag{18}$$

Dividing the expanded numerator by the denominator gives

$$U_\varepsilon = \sum_a \pi_a^\varepsilon V_a + \varepsilon E_\varepsilon, \qquad \|E_\varepsilon\| \le C, \quad \|D_x E_\varepsilon\| \le C/\varepsilon. \tag{19}$$

The derivative bound follows because E_ε is a quotient of uniformly bounded smooth coefficients and their averages under π^ε, and $D_x \pi_a^\varepsilon = \mathcal{O}(\varepsilon^{-1})$. Consequently the derivative of the output remainder is $\mathcal{O}(1)$, not automatically $\mathcal{O}(\varepsilon)$. This proves the uniform output statement and justifies differentiating its leading term across a switching layer.

A.3 Rank-one singular susceptibility and the shifted center

For $K = 2$,

$$\pi_1^\varepsilon = \frac{B_1 e^{\Phi_1/\varepsilon}}{B_0 e^{\Phi_0/\varepsilon} + B_1 e^{\Phi_1/\varepsilon}} = \mathrm{sigmoid}(g/\varepsilon + \beta), \quad g = \Phi_1 - \Phi_0, \quad \beta = \log(B_1/B_0). \tag{20}$$

Differentiating the leading barycenter gives

$$\begin{aligned} D_x\big((1 - \pi_1^\varepsilon) V_0 + \pi_1^\varepsilon V_1\big) = (1 - \pi_1^\varepsilon) D_x V_0 + \pi_1^\varepsilon D_x V_1 \\ + \pi_0^\varepsilon \pi_1^\varepsilon (V_1 - V_0) \left(\frac{\nabla g}{\varepsilon} + \nabla \beta \right)^\top. \end{aligned} \tag{21}$$

The first line and the prefactor-gradient term are uniformly bounded. Combining with Equation (19) proves Equation (13). For a perturbation $v \in \mathbb{R}^p$, the singular linear map acts as

$$v \longmapsto \frac{\pi_0^\varepsilon \pi_1^\varepsilon}{\varepsilon} \underbrace{\mathrm{d}g[v]}_{\text{one input scalar}} \underbrace{(V_1 - V_0)}_{\text{one output direction}}. \tag{22}$$

It has rank one when its two factors are nonzero, and rank zero otherwise. It does not claim that the whole attention layer or transformer has rank one.

If $\mathrm{d}g(x_0)[v] = 1$, then along $x_\varepsilon(u) = x_0 + \varepsilon u v$,

$$\pi_1^\varepsilon(x_\varepsilon(u)) = \mathrm{sigmoid}\big(u + \beta(x_0)\big) + \mathcal{O}(\varepsilon) \tag{23}$$

for bounded u. The leading equal-responsibility center is $g = -\varepsilon\beta$. For the normalized logistic mixing profile $\operatorname{sech}^2(G/2)$, the half-maximum points are $G = \pm 2 \operatorname{arcosh} \sqrt{2}$. Its full width at half maximum in G is therefore $4 \operatorname{arcosh} \sqrt{2} \simeq 3.5255$. In a local unit-normal spatial coordinate, the leading width is $4\varepsilon \operatorname{arcosh} \sqrt{2} / \|\nabla g(x_0)\|$. In the symmetric logit-shift coordinate $G + 2\delta$, the width is half the G width. This peak height and FWHM were derived independently above; they coincide numerically with Sakamoto's Corollary 3.2 ([F, §3]).

For $K > 2$, the singular term is $\varepsilon^{-1} \operatorname{Cov}_{\pi^\varepsilon}(V_a, \nabla\Phi_a)$; its rank is at most $K - 1$ and may be smaller. Different-dimensional competing strata carry different powers of ε in their prefactors and can shift balance by $\mathcal{O}(\varepsilon \log \varepsilon)$. Degenerate critical points require a different asymptotic analysis; they are not covered by these assumptions.

A.4 Proof of the exact finite-key response law

Let $T_\delta = e^{-\delta} Z_0 + e^{\delta} Z_1$. The conditional responsibility and union mass are $p_\delta = e^{\delta} Z_1 / T_\delta$ and $\rho_\delta = T_\delta / (T_\delta + Z_R)$. Direct differentiation gives

$$p'_\delta = 2p_\delta(1 - p_\delta), \qquad \frac{T'_\delta}{T_\delta} = 2p_\delta - 1, \qquad \rho'_\delta = \rho_\delta(1 - \rho_\delta)(2p_\delta - 1). \tag{24}$$

Uniform shifts cancel inside each conditional softmax, so $\overline{v}_0, \overline{v}_1, \overline{v}_R$ remain constant. Grouping the full attention output gives Equation (3); differentiating it gives Equation (4). When $Z_R = 0$, take $\rho = 1$ and simply omit terms involving $\overline{v}_R$. For multiple independently edited sites, first-order effects add by the chain rule; finite-strength effects across layers need not add because the upstream states change.

A.5 The full query Jacobian need not be rank one

Let $s_j(q) = c_{\mathrm{att}} \langle q, k_j \rangle + b_j$, with keys, values, biases, and masks fixed. The exact log-partition gradient is

$$\nabla_q L_a = c_{\mathrm{att}} \overline{k}_a, \qquad \nabla_q G = c_{\mathrm{att}} (\overline{k}_1 - \overline{k}_0). \tag{25}$$

The general output Jacobian is

$$D_q o = c_{\mathrm{att}} \left(\sum_j \alpha_j v_j k_j^\top - \Big(\sum_j \alpha_j v_j \Big) \Big(\sum_j \alpha_j k_j \Big)^\top \right). \tag{26}$$

Equivalently, with $v_B = (1 - p)\overline{v}_0 + p\overline{v}_1$,

$$\begin{aligned} D_q o = {} & \rho p (1 - p)(\overline{v}_1 - \overline{v}_0)(\nabla_q G)^\top + (v_B - \overline{v}_R)(\nabla_q \rho)^\top \\ & + \rho \{ (1 - p) D_q \overline{v}_0 + p D_q \overline{v}_1 \} + (1 - \rho) D_q \overline{v}_R. \end{aligned} \tag{27}$$

Here $D_q \overline{v}_a = c_{\mathrm{att}} \operatorname{Cov}_a(v, k)$, and

$$\nabla_q \rho = \rho(1 - \rho)(\nabla_q \log(Z_0 + Z_1) - \nabla_q \log Z_R) \tag{28}$$

when $Z_R > 0$. A tangent kills the first, between-branch switching term, but not the other terms. Even at $p = 1/2$, a *query* perturbation may change outside mass; its vanishing at a tie is specific to the symmetric *branch-logit* intervention. Thus normal/tangent output and margin comparisons are measurements, not algebraic consequences of choosing the gap normal. Moreover, among unit vectors the normal maximizes $|D_d G|$ by Cauchy–Schwarz, but the direction maximizing $|D_d M|$ is generally $\nabla_q M / \|\nabla_q M\|$, which can differ from n.

A.6 Meaning and error of the finite-strength predictor

For candidate c, collect all edited query outputs at site u into O_c and hold the other outputs of that layer fixed. Let $F_c(O_c)$ be the remaining network's candidate score. The one-site intervention leaves the input scores and values at that layer fixed, so $\widetilde{O}_c(\delta)$ is exactly determined by the local softmax law in real arithmetic. Taylor's theorem gives

$$F_c(\widetilde{O}_c(\delta)) = F_c(O_c(0)) + \Big\langle DF_c(O_c(0)), \widetilde{O}_c(\delta) - O_c(0) \Big\rangle + R_c(\delta). \tag{29}$$

If the downstream Hessian norm is bounded by L_c on the segment,

$$|R_c(\delta)| \le \tfrac{1}{2} L_c \left\| \widetilde{O}_c(\delta) - O_c(0) \right\|^2 . \tag{30}$$

Subtracting the candidate scores yields Equation (36) and a remainder bounded by the sum of these two terms. This is an interpretation of the approximation, not an empirically estimated uniform error certificate: L_c is not measured here. The exact theoretical identities refer to a real-arithmetic network; BF16 rounding and its interaction with automatic differentiation are addressed separately below.

For fixed values, the prediction can also be written $\langle \lambda, \Delta o \rangle = \sum_j \langle \lambda, v_j \rangle \, \Delta\alpha_j$. Thus exact local softmax recomputation followed by a downstream gradient is the same principle used by the attention correction in AtP* (Kramár et al., 2024). Here the controlled branch-logit coordinate additionally exposes the conditional mixing and outside-mass terms, which are tested separately across operating points.

A.7 Proofs for the Sakamoto extensions

This subsection proves the propositions stated in Appendix K, which were first stated (with proof) in independent research notes by Ryosuke Sakamoto; each proof below is our own derivation in this manuscript's notation, cross-referenced to the corresponding source section.

Proof of Proposition 2 (cf. [F, §3, proof of Theorem 3.1]). Write $U_\varepsilon = T_\varepsilon + r_\varepsilon$. The established normal-form bounds give $\|r_\varepsilon\| = \mathcal{O}(\varepsilon)$, $\|D_x r_\varepsilon\| = \mathcal{O}(1)$, and $\|D_x T_\varepsilon\| = \mathcal{O}(\varepsilon^{-1})$. The chain rule gives

$$D_x W_\varepsilon = D\Psi(T_\varepsilon) D_x T_\varepsilon + [D\Psi(U_\varepsilon) - D\Psi(T_\varepsilon)] D_x T_\varepsilon + D\Psi(U_\varepsilon) D_x r_\varepsilon .$$

The last two terms are bounded. This proves the displayed expansion without differentiating an uncontrolled output-error term. ■

Proof of Proposition 3 (cf. [S, §3.1, Proposition 3.1, proof in [S] Appendix A.2]). Write ρ for the attention mass on the selected pair and $\omega = 1 - \rho$, so that $U_\varepsilon = \rho T_\varepsilon + (1-\rho) R_\varepsilon$ with R_ε the conditional output on the omitted keys (Equation (3) with $\rho_\delta \to \rho$, $v_{B,\delta} \to T_\varepsilon$, $\overline{v}_R \to R_\varepsilon$, in the frozen-value case). Then $U_\varepsilon - T_\varepsilon = \omega(R_\varepsilon - T_\varepsilon)$; both T_ε and R_ε lie in the convex hull of $\{v_j\}$, so $\|R_\varepsilon - T_\varepsilon\| \le D$, giving $\|U_\varepsilon - T_\varepsilon\| \le D\omega$, and the mean value theorem with Lipschitz bound L_1 on F gives $|F(U_\varepsilon) - F(T_\varepsilon)| \le L_1 D\omega$. Let μ_S, μ_R denote the conditional mean action slopes on the selected and omitted key groups respectively; both lie in $[\min_j b_j, \max_j b_j]$, so $|\mu_R - \mu_S| \le B$. Since ω is itself the omitted-group softmax mass, differentiating it directly gives the exact identity $\omega' = \frac{\omega(1-\omega)}{\varepsilon}(\mu_R - \mu_S)$ (the same two-group mixing-derivative identity used to obtain Equation (4) in Appendix A.4, specialised to the selected/omitted partition), hence $|\omega'| \le B\omega/\varepsilon$ — not $\varepsilon^{-1}\omega \cdot D \cdot 1$, which conflates the value diameter D with the action-slope spread B. Differentiating $U_\varepsilon = \rho T_\varepsilon + (1-\rho) R_\varepsilon$ by the same product rule used to obtain Equation (4) (Appendix A.4) gives $U'_\varepsilon - T'_\varepsilon = \omega'(R_\varepsilon - T_\varepsilon) + \omega(R'_\varepsilon - T'_\varepsilon)$; each of $R'_\varepsilon, T'_\varepsilon$ is bounded by the value diameter D times the all-key affine action-slope spread B, i.e. $\|R'_\varepsilon\|, \|T'_\varepsilon\| \le DB/\varepsilon$. Combining these bounds gives $\|U'_\varepsilon - T'_\varepsilon\| \le 3DB\omega/\varepsilon$, and the analogous chain-rule estimate with L_1, L_2 gives the stated bound on the difference of the first derivatives of the two composed readouts in Equation (49). When both selected actions exceed every unselected action by $\eta > 0$, the standard Laplace tail estimate (Assumption (A3) in Appendix A.1) bounds each omitted key's relative mass by $\mathcal{O}(e^{-\eta/\varepsilon})$, summing to $\omega \le (N-2)e^{-\eta/\varepsilon}/2$; feeding this into the previous display recovers the exponential leakage bound stated in the proposition. ■

Proof of Proposition 4 (two-key affine susceptibility; cf. [S, §2.3, Theorem 2.2 and Corollary 2.3, proof in [S] Appendix A.2]). With exactly two keys, unit prefactors, and $A_1(t) - A_0(t) = t$, the responsibility of key 1 is $\mathrm{sigmoid}(t/\varepsilon)$, so $M_\varepsilon(t) = H(\mathrm{sigmoid}(t/\varepsilon))$ exactly, with no omitted-mass term. The chain rule and $\mathrm{sigmoid}'(0) = 1/4$ give $\varepsilon M'_\varepsilon(0) = H'(1/2)/4$, which is Equation (50) for every $\varepsilon > 0$. Substituting $t = \varepsilon h$ and $t = -\varepsilon h$ gives $D_\varepsilon(h) = \big(H(\mathrm{sigmoid}(h)) - H(\mathrm{sigmoid}(-h))\big)/(2\varepsilon h) = C_h/\varepsilon$, hence $D_{\varepsilon/2}(h)/D_\varepsilon(h) = 2$ whenever $C_h \ne 0$. If F is three times continuously differentiable, a symmetric Taylor expansion of $H \circ \mathrm{sigmoid}$ about 0 gives $C_h = H'(1/2)/4 + \mathcal{O}(h^2)$ as $h \to 0$. No dominance or leakage estimate is used: the statement is an identity for two keys and is not asserted for configurations with additional keys. ■

Convex construction of a dominant tie. Proposition 5 is a projected quadratic-program duality statement (Slater-type feasibility criterion plus strong duality for a convex quadratic program with linear constraints); we do not reproduce the full derivation here and refer to [S, §3.2, Proposition 3.2] for the complete proof. The only fact used elsewhere in this manuscript is the existence, when feasible, of a well-defined minimum query-space cost $\frac{1}{2}\|d^\star\|^2$ for constructing a dominant tie; no numerical result in this manuscript currently relies on this construction.

B MOVED REMARK: WHY THE OUTSIDE-MASS TERM IS NECESSARY

Remark 1 (Why the outside-mass term is necessary)**.** *Let $Z_0 = 1$, $Z_1 = 3$, $Z_R = 4$, $\overline{v}_0 = \overline{v}_1 = 1$, $\overline{v}_R = 0$, and $M(o) = o$. At $\delta = 0$, the switching term is zero, whereas the exact response is $\kappa_0 = 1/8$. Thus a switching-only law is not exact away from its stated special cases.*

C DISCOVERY PROVENANCE AND FROZEN HEAD SETS

C.1 ORIGINAL GEMMA SYNTHETIC EXPERIMENT

The source record is the Gemma 4 synthetic routing run, not the earlier Qwen3-4B steering experiment. The archived Gemma files `config_and_audit.json`, `all_head_scores.csv`, `selected_heads.csv`, `split.csv`, and `calibration_runs.csv` agree with the discovery notebook and with the frozen head lists in every later Gemma evaluation archive. The companion notebook is included as `code/gemma_synthetic_discovery.ipynb`.

Formal setup. For synthetic prompt x_i, let a_0, a_1 be its two single-token answer labels, and let $z_i(a; \boldsymbol{\delta})$ denote the model's final vocabulary logit for token a at the last prompt position under head-specific branch shifts $\boldsymbol{\delta} = (\delta_u)_u$. These are logits before the output softmax, not attention logits. Write $z_i(a) = z_i(a; \mathbf{0})$ for the unmodified model and define its signed output margin as

$$m_i := z_i(a_1) - z_i(a_0) = \log \frac{P(a_1 \mid x_i)}{P(a_0 \mid x_i)}. \tag{31}$$

A smaller $|m_i|$ means more similar probabilities for the two labels and thus proximity to their output decision boundary. From the 36 prompts, we choose the eight with the smallest $|m_i|$ as the calibration set $\mathcal{C}$ (a set of prompt indices). This selects near an *output* decision boundary, not an attention tie.

We introduce an independent zero-initialized shift δ_u for each of the $42 \times 8 = 336$ heads at the final prompt query. One backward pass per calibration prompt gives all partial derivatives. We score each head by its mean signed derivative,

$$c_u = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \frac{\partial [z_i(a_1; \boldsymbol{\delta}) - z_i(a_0; \boldsymbol{\delta})]}{\partial \delta_u}\bigg|_{\boldsymbol{\delta} = \mathbf{0}}, \qquad |\mathcal{C}| = 8, \tag{32}$$

and retain the eight largest scores, in order:

$$\mathcal{H}_{\text{fixed}} = \{(29, 2), (28, 3), (29, 4), (41, 2), (27, 3), (28, 4), (30, 4), (26, 1)\}. \tag{33}$$

Indices are zero based. Eight is a preset sparsity budget, not an inferred circuit size or a ConflictQA-tuned optimum. Scores range from 3.8408 to 0.6673; the set contains three full-attention and five sliding-attention heads (2.38% of all heads).

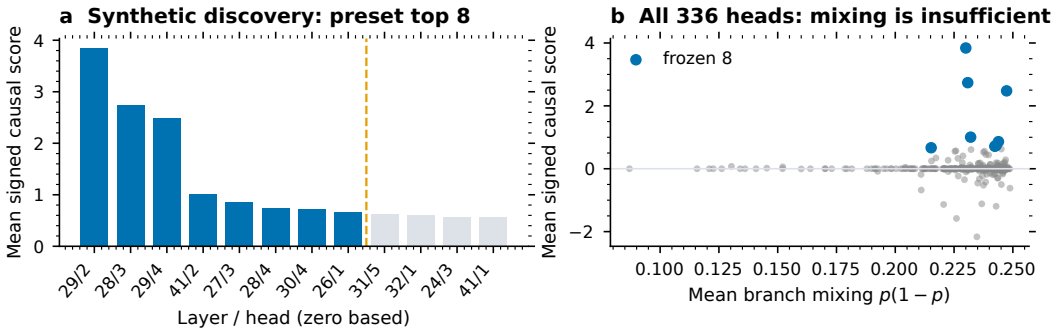


Figure 6: **Independent Gemma head selection.** (a) Original mean signed gradients, with the preset top eight highlighted. (b) All 336 heads: branch mixing alone does not recover the causal ranking (Pearson $r = 0.017$). All scores use synthetic calibration prompts only.

Prompt construction and partition. The seed is 20260728 and the preset budget is `top_k=8`. There are 12 prompts for each of three discovery label pairs (`blue`, `red`), (`green`, `yellow`), (`copper`, `silver`), and 12 for each of three additional transfer pairs (`north`, `south`), (`winter`, `summer`), (`circle`, `square`). Thus the full synthetic pool has 72 prompts, but only 36 are eligible for calibration. Every label is checked to tokenize to one token.

Every prompt contains four reports supporting a_0 and four supporting a_1. Their order is randomly shuffled with seed $20260728 + 10000\, i_{\text{pair}} + i_{\text{prompt}}$. Report wording cycles through six fixed templates, such as "The recorded outcome is {label}," and the question uses one of four templates. The full report lines, rather than only the label tokens, define the branch masks. Equal report counts and equal declared reliability are built into the task; the protocol does not assert equality of token counts or eliminate learned label preferences.

After an unmodified forward scan, the discovery prompts are sorted by $|m_i|$. The first eight form calibration, and the next 18 are the seen-label evaluation set. Six prompts from the additional pairs, also chosen by baseline $|m_i|$, form the synthetic transfer evaluation set. None of these synthetic evaluation outcomes determines the heads. The use of output-boundary selection belongs to the synthetic experiment; the subsequent 50-example ConflictQA cohort is not selected this way.

Table 2: The original eight Gemma calibration prompts. Signed margins are output-logit differences $z(a_1) - z(a_0)$, not attention gaps. All are drawn from the 36 discovery-pair prompts.

Prompt ID	Label pair	Baseline m_i
`seen_blue_red_05`	blue/red	0.500
`seen_green_yellow_04`	green/yellow	0.500
`seen_blue_red_02`	blue/red	-0.750
`seen_blue_red_04`	blue/red	-0.875
`seen_copper_silver_09`	copper/silver	-0.875
`seen_blue_red_00`	blue/red	-1.375
`seen_green_yellow_06`	green/yellow	1.375
`seen_green_yellow_03`	green/yellow	2.000

Selection algorithm. For each calibration prompt, create a zero matrix $\boldsymbol{\delta} \in \mathbb{R}^{42\times 8}$ with gradients enabled. At each layer–head site, add the corresponding signed branch bias at the final prompt query only. Compute $m_i = z_i(a_1) - z_i(a_0)$ and differentiate with respect to the whole matrix. Average the resulting matrices over the eight prompts, then flatten and take `topk(mean_causal, 8)`. This is signed averaging followed by descending ranking: neither absolute gradients nor $p(1-p)$ determine the selected set. All eight retained mean scores happen to be positive. The calibration size eight and the intervention budget eight are separate preset parameters.

Table 3: Exact provenance of the eight frozen Gemma sites, in original rank order. The causal score is Equation (32); the last column is the mean conditional mixing factor on the same eight calibration prompts.

Rank	Layer	Head	Attention type	Signed causal score	Mean $p(1-p)$
1	29	2	Full	3.840800	.229998
2	28	3	Sliding	2.737981	.230814
3	29	4	Full	2.477038	.247348
4	41	2	Full	1.005221	.232054
5	27	3	Sliding	.861593	.243881
6	28	4	Sliding	.740906	.242893
7	30	4	Sliding	.714581	.242331
8	26	1	Sliding	.667340	.215253

Why this is a causal, not a mixing-only, ranking. The eight sites constitute 2.38% of all 336 sites but carry 56.10% of the sum of *positive mean causal scores*. This percentage is a statistic of the discovery gradients, not a fraction of all possible behavior or of finite-strength steering performance. Across all sites, the Pearson correlation between the signed mean causal score and mean $p(1-p)$ is 0.0173. For example, the largest mixing-score site $(30, 5)$ has a negative mean causal score, -0.61794, and is not selected by the signed rule. This difference is consistent with the need for downstream alignment in Equation (5).

C.2 Gemma operating-point subsets are not a new eight-head discovery

For each of ten ConflictQA examples, the operating-point notebook computes, for every $u \in \mathcal{H}_{\text{fixed}}$,

$$s_u^{\text{tie}}(x) = \left| \sum_{c=0}^{1} (2c-1) \left\langle \lambda_{c,u}(0),\ \frac{\mathrm{d}\widetilde{o}_{c,u}(\delta)}{\mathrm{d}\delta}\Big|_{\delta=-\overline{G}_{0,u}/2} \right\rangle \right|, \qquad \overline{G}_{0,u} = \tfrac{1}{2}(G_{0,u}^{(0)} + G_{0,u}^{(1)}). \quad (34)$$

The three largest values and the smallest value are selected for that example. The baseline covectors are used only for this prospective selection, while actual operating-point covectors are recomputed for the derivative test. The eight externally frozen sites therefore remain the candidate pool, even though only four are tested per example in this protocol.

C.3 Independent Qwen discovery and reuse in geometry

The Qwen3-8B discovery notebook scans all $36 \times 32 = 1152$ layer–query-head sites on six ConflictQA examples. It computes Equation (34) and ranks sites by the median across examples of the absolute predicted tie derivative, breaking ties by the mean absolute predicted tie derivative. The top three and the bottom remaining site are frozen, as shown in Table 4. No operating-point finite difference is used in discovery. Subsequent numerical reruns and the geometry experiment retain both these sites and the eight evaluation IDs.

The six discovery IDs are `004824, 003554, 005317, 009442, 000409, 001452`. The eight evaluation IDs are `004712, 004188, 005291, 000629, 003512, 003595, 009152, 008865`; all carry the prefix `conflictqa_`. The two sets are disjoint. Geometry is a different intervention on the same evaluation cohort, not a third independently sampled replication. The initial discovery/selection files are retained separately from the final wide-step evaluation files so that the numerical rerun cannot silently redefine the selected circuit.

Table 4: Qwen3 discovery scores and frozen roles. Values are medians over six discovery examples; the weak site's tiny union mass helps explain its very small analytic sensitivity.

Layer	Head	Role	Median predicted tie strength	Median baseline ρ
24	28	Strong	.136006	.406732
34	7	Strong	.105920	.450975
24	29	Strong	.105363	.386443
13	12	Weak control	6.16×10^{-8}	1.3×10^{-5}

D Implementation, evaluation units, and numerical controls

D.1 Dataset and prompt construction

We use the GPT-4 PopQA file from the repository `osunlp/ConflictQA`. Its filename is `conflictQA-popQA-gpt4.json`; the recorded SHA-256 is

`c9921b733f45c5f473d2ef9ed242b5a04`
`c195d0cc1ae002ba014cf351b847aca`

Candidate 0 comes from `memory_answer` and candidate 1 from `counter_answer`. The evidence fields are `parametric_memory_aligned_evidence` and `counter_memory_aligned_evidence`, with the corresponding shorter memory fields used as fallbacks. Whitespace is normalized, incomplete records and identical normalized candidate/evidence pairs are removed, and duplicate normalized questions are discarded. Both candidate strings must have 2–28 tokens under the model tokenizer before appending. Rendered prompts must fit within 460 tokens.

The system message requests concise answers from supplied sources without assuming agreement. Each evidence string is enclosed in a Source A or Source B delimiter, followed by the question. Source order is determined by the first byte of $\mathrm{SHA256}(\texttt{str(seed)} + \texttt{example_id})$ modulo two, not by intervention outcomes. Branch labels continue to mean benchmark candidate 0/1 regardless of source order. The model chat template is used with a generation prompt; thinking mode is disabled when the template supports that option. Character offsets map the evidence spans to tokens. Candidate appending verifies that the prompt token prefix is unchanged, retrying with an intervening space if necessary.

Eligible examples are shuffled with Python's `random.Random(20260728)`. The first 50 completed Gemma examples form the finite-strength cohort. The ten operating-point examples are its first ten IDs: `008295`, `008266`, `005355`, `007229`, `000467`, `002835`, `000882`, `007915`, `002916`, `004397`, all prefixed by `conflictqa_`. These cohorts are held out from synthetic head discovery, not mutually disjoint. The Qwen discovery and evaluation sets use the same preprocessing rule with its own tokenizer, and have the explicit IDs in Appendix C. All intended final evaluation examples are present in the supplied result tables. The 50-example and Qwen evaluation cohorts have no shared IDs.

D.2 Attention intervention and downstream covectors

The custom eager attention function is intended to preserve the checkpoint's attention scale, masks, key/value head repetition, score transformations, and other upstream computation. Checkpoint-matched verification of this preservation was not performed. The archived measurements characterize the recorded custom execution path and do not by themselves establish equivalence to the native checkpoint implementation. Logit biases are applied before the softmax; masked-out keys remain masked. Softmax is computed in FP32 and attention probabilities are cast back to the model dtype for the value product. Parameters are frozen and dropout is disabled by evaluation mode. All comparisons use the same candidate, branch masks, and chosen layer–head site.

The finite-strength notebook captures every query row predicting a scored answer token, from the final prompt row through the penultimate input row. For the mechanistic experiments, only the final prompt row is selected. The head output is captured before output projection, and `torch.autograd.grad` differentiates each mean candidate score with respect to the captured

layer output tensor. The selected head/query slice is the covector in the theory. The computation therefore includes the same layer's subsequent projection, normalization, residual and feed-forward paths, later transformer layers, and final language-model readout. It is neither an attention-weight gradient nor a coefficient fitted to observed intervention curves.

For finite-strength prediction the covector remains at baseline. For operating-point parity it is recomputed after the actual preload. For prospective tie ranking only, the baseline covector is paired with the exact local derivative evaluated at a hypothetical tie. These three uses of λ should not be conflated.

Table 5: Numerical protocol details. All interventions are evaluated one site at a time. The half-width is in symmetric branch-logit-shift units, so a finite-difference probe changes the branch gap by $\pm 2h$.

Setting	Gemma finite strength	Gemma operating points	Qwen operating points
Evaluation examples	50	10, subset of 50	8, disjoint from discovery
Heads per example	8 frozen	3 strong + 1 weak from 8	4 frozen after 6-example discovery
Strength/gap grid	13 shifts in $[-2, 2]$	9 gaps in $[-4, 4]$	7 gaps in $[-3, 3]$
Primary probe h	Not applicable	.10	.50
Secondary probe h	Structural slope: .05	.05	.25
Inference microbatch	4 strengths	4 probe strengths	1 strength
Candidate passes	Separate	Separate	Separate singleton passes
Covector	Baseline only	Recomputed at each gap	Recomputed at each gap

D.3 Operating-gap calibration

At the final prompt position, candidate 0 and candidate 1 have identical causal prefixes. Their local states therefore coincide in exact arithmetic, even though their downstream covectors differ. In BF16 execution, different full sequence lengths and computation paths can produce small numerical differences in the captured local states. The Gemma preload uses the mean of the two baseline gaps. The final Qwen wide-step rerun and the geometry experiment use candidate 0's no-gradient baseline gap. Candidate-specific local quantities are retained in the derivative calculation; geometry captures local quantities on candidate 0 and computes the margin from both candidates. Maximum recorded imposed-gap errors are 0.11875 in Gemma and 0.07552 in Qwen. The geometry experiment's constant branch preload is held fixed during the $+\eta d$ and $-\eta d$ query perturbations.

D.4 Metric definitions and resampling

For a response curve, let $y_i = M(\delta_i) - M(0)$ and $\widehat{y}_i = \widehat{M}(\delta_i) - \widehat{M}(0)$, using each curve's own recorded zero point. We report

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}, \qquad \mathrm{MAE} = \frac{1}{N} \sum_i |y_i - \widehat{y}_i|. \tag{35}$$

Pooled R^2 uses all eligible points; curve-level R^2 is computed independently for each example–head pair. No affine recalibration or regression fit precedes this calculation. Negative R^2 means the predictor is worse than predicting that curve's observed mean. Pearson correlation is a separate statistic and does not replace R^2. Derivative parity uses $y_i = \kappa_{\mathrm{FD}}$ and $\widehat{y}_i = \kappa_{\mathrm{an}}$ in the same definition.

Head effect size is the range $\max_\delta M_u(\delta) - \min_\delta M_u(\delta)$. Predicted and observed ranges are ranked within each example's eight-head pool. Top-1 accuracy tests whether the largest predicted range identifies the largest observed range; top-three coverage asks whether the observed best head is among the three predicted best. The slope-sign diagnostic uses the nearest grid points on either side of zero, $\delta = \pm 0.25$, for both predicted and observed curves.

For paired gains, we first average over the relevant heads within each example, then bootstrap the resulting example means with replacement (2,000 draws, percentile 95% intervals). This avoids

treating correlated heads or operating points as independent subjects. The primary finite-strength unit is one of 50 examples; the mechanistic units are ten Gemma or eight Qwen examples. Geometry intervals preserve the tangent-median comparison within each condition before averaging over heads. Intervals are descriptive and not adjusted for multiple comparisons. The analysis seed is 20260728; the gap, responsibility, head-output, and margin gain intervals use offsets 0, 1, 2, and 3, respectively.

D.5 Finite precision and audit limits

The notebooks pin Transformers 5.14.1, use eager attention with BF16 model weights, disable TF32, use explicit position IDs, and turn off the generation cache. The original Gemma discovery archive records PyTorch 2.11.0+cu128, CUDA 12.8, and an NVIDIA A100-SXM4-40GB. The later archives do not all record complete package versions, GPU metadata, or immutable model/tokenizer revision hashes. The saved notebook version and dataset hash support reconstruction, but do not provide a bitwise-reproducible snapshot of every execution environment.

An ideal smooth central difference has $\mathcal{O}(h^2)$ truncation error. If each measured margin also has error bounded by ν, the additional difference error can be as large as ν/h. Widening the probe can therefore improve measured parity while moving farther from the infinitesimal regime. For Qwen, $h = .5$ spans gap values $G_\star \pm 1$, a substantial interval inside the logistic layer. We report this as a finite-difference consistency test and do not claim an established $h \to 0$ convergence result.

Table 6: Probe-width sensitivity. "All" includes the weak control; strong-point metrics can be much better because weak responses are near numerical resolution.

Model	h	All R^2	Strong R^2	All MAE	Strong MAE
Gemma	.05	.561	.649	.0755	.0721
Gemma (primary)	.10	.794	.845	.0427	.0402
Qwen	.25	.748	.833	.0614	.0565
Qwen (primary)	.50	.909	.948	.0338	.0290

The Gemma notebook initially designated $h = .05$ primary; the archived reaggregation promotes $h = .10$ after the observed sensitivity analysis. Its generic `kappa_fd_primary` column still refers to the old probe in some files, so this manuscript explicitly reads `kappa_fd_h0p1`. Qwen's final rerun uses singleton inference and $h = .25, .50$, with the previously selected heads and evaluation IDs frozen. The numerical control is adaptive to earlier diagnostics, not a new preregistered replication cohort.

At Qwen's 160 interior example–head–gap points, the $h = .5$ probe endpoints coincide with adjacent integer gap-grid operating points. Separately recorded margins at these same intervention values differ by mean absolute errors 5.86×10^{-8} (minus) and 5.26×10^{-8} (plus). The finite difference and adjacent-grid secant differ by 7.14×10^{-8}. This rules out a substantial mismatch between these recorded evaluations; it does not remove BF16 rounding, prove derivative accuracy, or validate the weak-head effects.

E Shared setup, provenance, and experiment protocol details (moved from the main text)

This appendix contains material moved from the main text without changes to any reported number: the provenance and audit-status statements, the auxiliary intervention primitives, the evaluation design table, the independent Gemma head selection, and the per-experiment intervention and metric details. The main text keeps a condensed statement of each item.

E.1 Provenance of the analyses and status of auxiliary results

then presents its question, intervention and prediction, evaluation metrics, results, and interpretation. The accompanying results ledger distinguishes a smoke run from a subsequent expanded Qwen run; neither implements a fixed query-space basis or a mean task-gradient controller despite their condition names (see the audit-status paragraph below). Neither is incorporated as empirical evidence here

because the measurement function, fitting rules, realized intervention budgets, and result-generation code have not been linked to the archived controller protocol in this manuscript. The expanded comparison must not be described merely as a smoke run, nor treated as validation of the archived controller measurements. Any use of these auxiliary results must identify its own outcome, execution path, analysis chronology, and inferential scope.

For the archived held-out controller cohort, symmetric intervention efficiency was the originally specified endpoint; the additional one-sided analysis is post-hoc and descriptive. Fresh Qwen Experiment 3 results were collected before the comparison convention was adopted, so that amendment is not confirmatory. Multiplicity correction, including Holm correction, does not convert an outcome-informed endpoint or comparison into a prospectively specified test. Any confirmatory fresh evaluation must identify its frozen analysis, outcome-unseen cohort, and corresponding result artifact separately from the archived analyses.

E.2 Shared task, outcome, and intervention positions

The outcome is a teacher-forced sequence-score margin in nats per token, not free-generation accuracy. An attention head is indexed by $u = (\ell, h)$, where ℓ is the layer and h the query-head index; each intervention edits one head at a time. The branch-logit intervention in Equation (2) adds $+\delta$ to B_1 and $-\delta$ to B_0, leaving the other logits unedited. At each query, the branch gap is $G = L_1 - L_0$, with $L_a = \log \sum_{j \in B_a} e^{s_j}$; hence the shift changes G by 2δ.

Separate logit-space intervention primitives. The auxiliary implementation provides legacy symmetric shifts, balanced switching, and evidence allocation. These are single-site logit interventions, not the joint-query intervention used in Experiments 4–5. Write $T_\delta = e^{-\delta} Z_0 + e^{\delta} Z_1$, $C(\delta) = \log(T_\delta / T_0)$, and $\sigma_j = \mathbf{1}_{j \in B_1} - \mathbf{1}_{j \in B_0}$. Let b_j indicate membership in the evidence union. Balanced switching uses

$$s_j(\delta) = s_j + \delta \sigma_j - C(\delta) b_j.$$

It preserves union mass and changes the conditional branch odds by 2δ. With fixed keys, values, and masks, its output derivative is

$$o'_\delta = 2\rho_0 p_\delta (1 - p_\delta)(\overline{v}_1 - \overline{v}_0).$$

Evidence allocation instead uses $s_j(\gamma) = s_j + \gamma b_j$. It preserves the conditional branch weights and satisfies

$$o'_\gamma = \rho_\gamma (1 - \rho_\gamma)(v_B - \overline{v}_R).$$

The downstream margin derivative is the inner product of the corresponding output derivative with the downstream covector. Equation (5) applies to the legacy symmetric intervention, not unchanged to all three variants.

The supplied auxiliary code contains two different scoring paths. Its negative-control routines use an answer0-minus-answer1 raw-logit probe at the first divergent token. Its current production helper instead averages log probabilities over each candidate's post-common-prefix tail and subtracts the answer1 score from the answer0 score. The latter is neither the candidate orientation nor the full-answer normalization used in Equation (7). Removing a shared prefix does not preserve a length-normalized candidate margin in general.

The default intervention position is the final position of the prompt plus the shared answer prefix, rather than necessarily the final prompt position. The default evidence masks follow Source A/B positions rather than the semantic candidate labels. Consequently, neither helper certifies the archived sequence-margin measurements or the intervention positions used in Experiments 1–5.

Scope of the auxiliary numerical audits. The auxiliary zero-hook audit uses a two-layer CPU decoder configured with Gemma-2 attention constants, rather than the Gemma 4 E4B-it checkpoint used in Experiment 1. It therefore does not identify the cause of the archived Gemma 4 offsets or establish fidelity of the Experiment 1 execution path. Agreement between gradient-enabled and inference execution, or between batch layouts, in this small decoder does not rule out discrepancies in the full checkpoint.

Centering removes a constant difference between recorded zero points. It does not remove intervention-dependent differences caused by an omitted score transformation, a changed attention implementation, or finite precision. The archived centered-response statistics remain descriptive

measurements of their recorded execution paths; their correspondence to the intended native checkpoint requires a checkpoint-matched audit.

The auxiliary primitive audit uses synthetic query, key, and value tensors. Its analytic-versus-finite-difference comparisons concern scalar logit-space interventions only. Value-erasure invariance and derivative agreement under scrambled masks are calculus sanity checks, not evidence that semantic masks are useful or that the two mechanisms have been independently identified in the real model. End-to-end validation of the Qwen joint-query intervention used in Experiments 4–5 was not performed.

Checkpoint-matched audit status. As of 2026-09-19, the joint-query intervention used in Experiments 4–5 has been located precisely: it is implemented in `qwen_switching_control_subspace.ipynb` (cell 5, `controlled_eager_attention_forward` with a global `MultiGeomController`), which adds a per-(layer, head) delta directly to the post-RoPE query activation (not the pre-softmax logit operation used by `caustics_model_hooks.py`), and supports simultaneous perturbation of up to three (layer, head) targets in one forward pass. Gradients are obtained via `candidate_delta_gradient()` (`torch.autograd.grad` over all three head deltas jointly) and finite differences via `finite_control()` (signed two-sided steps in joint-query-norm units). The accompanying results ledger reports structural mismatches in auxiliary Qwen3-8B and Gemma-2-9b hook audits. These reports do not establish validation of the archived measurements. In particular, Gemma-2-9b is not the Gemma 4 E4B-it checkpoint used in Experiment 1. The correspondence between each audit and the original measurement function must be documented before its outcome can validate or invalidate an archived result.

For clarity, the intended joint-query intervention is an additive map, not an overwrite of baseline activations. Let z concatenate the three injected query increments. At each selected site, the increment z_h is added after query normalization and RoPE to the query produced by the current forward pass. Thus earlier interventions may change the pre-injection state at a later site. Evidence masks, logit preloads, and direction vectors are fixed from the zero-increment calibration and are not recomputed during a directional probe.

Let $F_x(z)$ be candidate 1's full teacher-forced mean log probability minus candidate 0's under this intervention. The intended gradient and finite efficiency are

$$g_x = \nabla_z F_x(0), \qquad s_x = \alpha \|q_x\|, \qquad \mathcal{E}_x(u;\alpha) = (F_x(s_x u) - F_x(-s_x u))/(2s_x).$$

Here q_x concatenates the zero-increment query blocks. The budget is the norm of the injected increments; it is not a claim that downstream query states differ from their baseline values by exactly that norm. These definitions specify the map that the archived implementation must match. Equality with the archived gradient and finite-response routines remains unverified.

Because this correspondence was not established, we make no native-checkpoint fidelity or joint-query derivative-consistency claim for these archived measurements. The auxiliary results include both a smoke run and an expanded Qwen run. Neither implements the fixed-covector comparison in Section M.2; that separate Gemma ablation changes only the downstream covector and remains subject to the archived-correspondence limitation above. In the supplied runner, the condition named `fixed3` uses fixed relative token-span masks for a balanced logit intervention, rather than three fixed query-space basis vectors with fitted coefficients. The condition named `static_mean_grad` applies a constant evidence-allocation shift and does not compute a mean task gradient. These condition names therefore cannot support claims about fixed three-dimensional query bases or static gradient controllers. We exclude both runs from the empirical evidence for Experiments 4–5. In particular, a state-DP or independent-enumeration quantity is not treated as evidence for the Experiments 4–5 intervention claims.

Correspondence of the auxiliary audits to the original measurement functions. Table 7 records, for each auxiliary audit stored in the results ledger (four ConflictQA calibration examples each; existing artifacts only, no new runs), the checkpoint, outcome, time, hook, dtype, native-versus-zero-hook result, and analytic-versus-finite-difference (AD/FD) result, next to the original measurement function it would have to match. It documents correspondence and mismatch; it is not a fidelity result. The Gemma-2-9b audit fails the native-versus-zero-hook condition in 4 of 4 examples, and both audits fail singleton/micro-batch agreement in 4 of 4; the Qwen3-8B audit's native/zero-hook

Table 7: Audit-to-measurement-function correspondence (all stored conditions; $n = 4$ examples per audit; rows sha256 `432d5f52...` for Qwen3-8B and `226a5b6c...` for Gemma-2-9b). Native/zero-hook and singleton/micro-batch differences are absolute differences of the recorded outcome; tolerances are those stored in each `summary.json`.

Audit	Checkpoint (audited vs. original path)	Outcome	Time (2026-09-19)	Hook	dtype	Native vs. zero-hook	AD vs. FD
Qwen3-8B	Qwen3-8B; same checkpoint as the archived joint-query path, but head selection and route equality with the archived map are not verified	teacher-forced candidate score, 4 examples	05:13:31	joint-query delta on heads (24,28), (34,7), (24,29)	bf16, eager attention	$\lvert\Delta\rvert = 0$ in 4/4 (tol. 0.02); grad on/off 0 in 4/4; singleton vs. micro-batch 0.031–0.101, above tol. 0.001 in 4/4	1/12 head-cases PASS (rel. tol. 0.1); sign agrees in 10/12; overall verdict: structural mismatch
Gemma-2-9b	Gemma-2-9b; *not* the Gemma 4 E4B-it logit path used in Experiment 1	teacher-forced candidate score, 4 examples	05:35:11	layer 29 head 2, query positions 258–467; legacy, balanced, evidence-allocation kinds	not recorded in the stored files	$\lvert\Delta\rvert = 0.125, 0.125, 0.539, 0.141$: 0/4 within tol. 0.05; grad on/off 0 in 4/4; singleton vs. micro-batch 0.125–0.453, 0/4 within 0.001	6/12 head-cases PASS (rel. tol. 0.2), but every PASS has rel. err $= 1.0$ with a finite-difference of 0 at the smallest width, so it does not indicate agreement; overall verdict: structural mismatch

and gradient on/off conditions hold, but its AD/FD comparison fails in 11 of 12 cases. Gemma-2-9b results do not transfer to the Gemma 4 E4B-it path.

Original-measurement-path audit (C5/C6; existing evidence). We first identify each archived checkpoint, scored outcome, and injection map before interpreting an audit result: the Gemma 4 branch-logit forward and batch-one capture, and the Qwen3 joint post-RoPE query hook, candidate-score, and gradient routines are enumerated below. We then check the stored audit rows against native/zero-hook and grad-on/off equality, singleton/microbatch agreement, and directional analytic/central-finite-difference agreement using the recorded tolerances and row hashes in `EXPERIMENT_AUDIT_20260925.md` and Table 7. The four-example Qwen auxiliary run has native/zero-hook equality but fails batch-layout agreement in 4/4 examples and directional AD/FD in 11/12 example–head cases; the four-example Gemma-2-9b auxiliary run fails native/zero-hook and batch-layout agreement in 4/4, and it is not the archived Gemma 4 checkpoint. These are failed or non-transferable auxiliary checks, not a completed original-path C5/C6 audit. The inventory finds no condition-complete original-path matrix or passing decision code, and no fixed-example, outcome-unseen audit with pre-set tolerances and three central-difference widths was performed. Consequently, we make no native-checkpoint fidelity or joint-query derivative-consistency claim for the affected archived measurements.

Original-path E1 condition inventory (not a passing audit). The original Gemma 4 E4B-it branch-logit path is the archived `gemma4_50example_8heads_finite_strength_prediction.py`, whose `controlled_eager_attention_forward` starts at line 524; its baseline capture explicitly requires batch size one (lines 552–554). The original Qwen3-8B joint-query path is the archived `qwen3_switching_control_subspace_test.py` (hook lines 611–672, candidate score lines 684–748, gradient lines 750–774) and the held-out counterpart (hook lines 555–616, score lines 638–702, gradient lines 704–729). In both Qwen notebooks, capture rejects a non-singleton batch, so those capture routines alone do not establish singleton/microbatch equivalence. The stored Qwen auxiliary audit compares native versus zero-hook and grad on/off for four examples, but fails the microbatch comparison in all four and the directional AD/FD comparison in 11 of 12 head-cases (Table 7). The stored Gemma-2-9b auxiliary audit is a different checkpoint and cannot certify the Gemma 4 original path. No complete condition-level matrix for the two original paths establishes native/zero-hook, grad on/off, batch-layout, and AD/FD agreement together. In particular, absence of Gemma 4 rows is not a pass, and Qwen zero-hook agreement is not a derivative pass. The only path verified end to end before outcomes were read is the CX-ID cohort (Section 3.1).

Experiment 1 edits every query position predicting a scored answer token. Experiments 2 and 3 edit *only the final prompt query*, which predicts the first answer token; their outcome still scores the entire candidate sequence. Experiments 4–5 also edit only the final prompt query, but jointly across three heads. Table 8 summarizes the cohorts and positions. Experiments 1–3 use 2,000

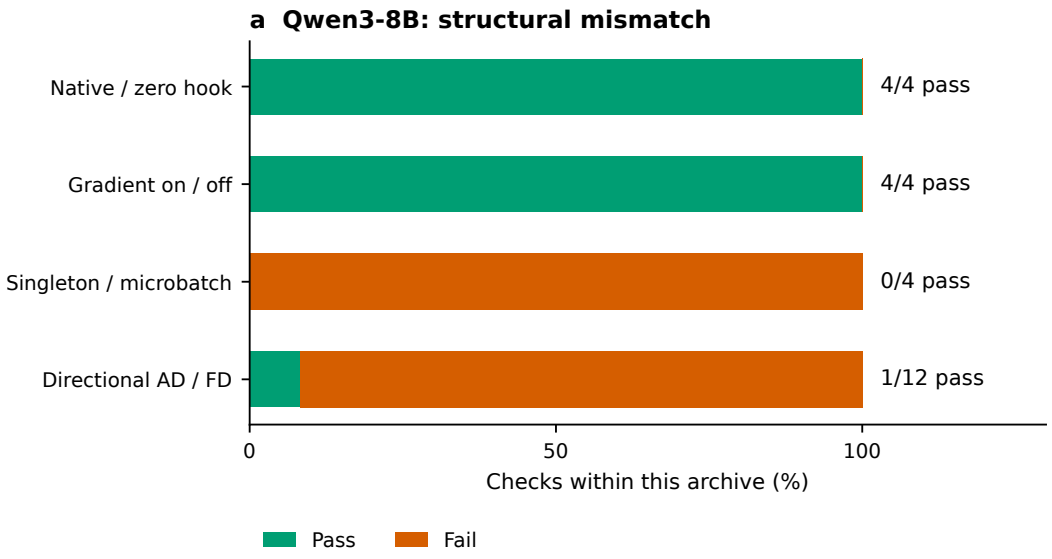


Figure 7: **Archived Qwen hook audit: structural-consistency failure.** Condition checks each contain four examples; directional AD/FD checks contain twelve example–head cases, not twelve independent examples. Counts are descriptive. Native/zero-hook tolerance is 0.02, gradient and batch tolerance 0.001, and directional tolerance 0.1 relative with 0.01 absolute floor at the smallest recorded width. Zero-hook agreement alone does not validate derivatives. Gemma local raw rows were not independently verified for this plot. Source: hash-checked `results/e1_qwen_20260919T051331/rows.jsonl` and summary.

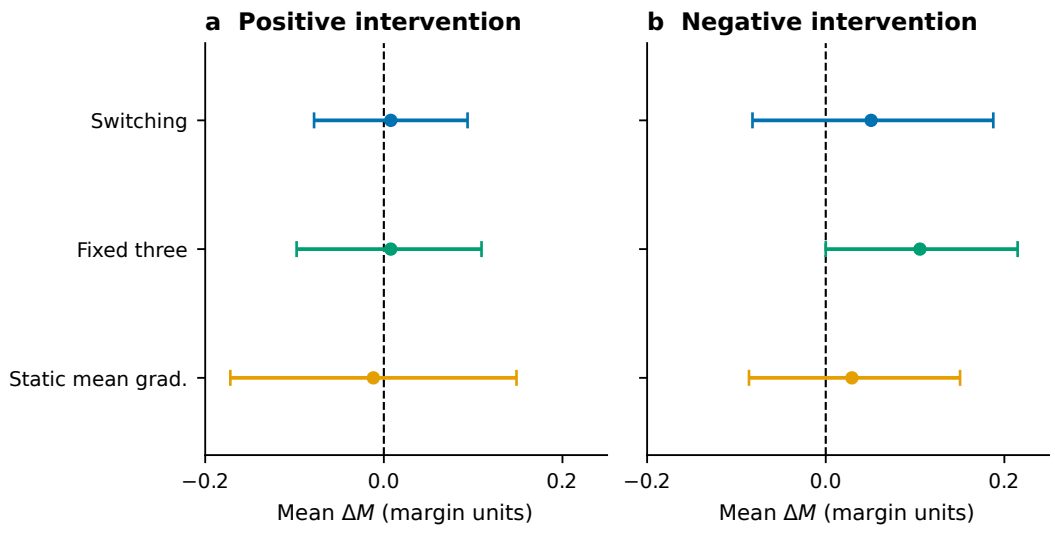


Figure 8: **Recovered E2 token-span cohort, not a query-controller comparison.** Means of $\Delta M_+ = M(+1) - M(0)$ and $\Delta M_- = M(-1) - M(0)$ on the same 32 Qwen3-8B examples; bars are 95% percentile intervals from 10,000 example-level bootstrap resamples (seed 20260919), not seed-level intervals or paired-superiority tests. The named fixed condition is an evidence-allocation shift, not a fixed query basis. Outcome and budget differ from the controller protocol. Source and recomputation: `code/make_audit_figures.py`.

and Experiments 4–5 use 4,000 example-cluster bootstrap resamples for 95% percentile intervals, preserving all conditions within each sampled example.

Table 8: Evaluation design. A fixed head pool is selected before the corresponding evaluation. The Gemma operating-point cohort is a subset of the finite-strength cohort. Experiments 3–4 reuse the Qwen operating-point cohort; Experiment 5 uses new tune and test examples.

Experiment / model	Examples	Heads per example	Edited queries
1: Gemma finite strength	50	All 8 fixed heads	All scored-answer queries
2: Gemma operating points	10	3 strong + 1 weak from 8	Final prompt query
2: Qwen operating points	8	3 strong + 1 weak, fixed	Final prompt query
3: Qwen normal/tangent	Same 8	Same 4 fixed heads	Final prompt query
4: Qwen control subspace	Same 8	3 strong, jointly	Final prompt query
5: Qwen frozen controller	16 tune + 32 test	Same 3, jointly	Final prompt query

E.3 Independent selection of the eight Gemma heads

The eight Gemma heads were transferred from an independent synthetic evidence-routing experiment, not selected on ConflictQA: each of 336 layer–head sites received a zero-initialized branch-logit shift, and the eight sites with the largest mean signed causal gradient toward the correct label, scored on eight calibration prompts chosen near an output-boundary tie, were retained as the fixed pool $\mathcal{H}_{\text{fixed}}$ (Equation (33), Appendix C). No ConflictQA prompt or intervention outcome enters this discovery. Appendix C gives the full synthetic prompt pool, the calibration records, the exact scoring rule (Equation (32)), and the complete 336-head ranking (Figure 6, Table 3).

E.4 Experiment 1: intervention, predictor, and metric details

Intervention and prediction. We test all eight fixed Gemma heads on 50 deterministically shuffled eligible ConflictQA examples, without outcome filtering. Each head is edited separately at 13 strengths in $[-2, 2]$, giving 400 curves and 5,200 points. For a prompt of length n and candidate of length T_c, the edited query positions are $Q_c = \{n-1, \ldots, n+T_c-2\}$ with zero-based indexing; $n-1$ predicts the first answer token. Let $o_{c,u,t}$ be head u's output vector at query t in candidate c's forward pass, before the attention output projection.

Baseline forward/backward passes record attention scores, values, outputs $o_{c,u,t}(0)$, and gradients $\lambda_{c,u,t}(0) = \partial S_0(y^{(c)} \mid x)/\partial o_{c,u,t}$. Automatic differentiation includes the output projection, feed-forward and residual computations, subsequent layers, and final readout. For each hypothetical δ, Equation (3) recomputes the local output $\widetilde{o}_{c,u,t}(\delta)$ while holding its upstream state fixed. A first-order Taylor approximation in the tensor of head outputs gives

$$\widehat{M}_u(\delta) = M(0) + \sum_{c=0}^{1}(2c-1)\sum_{t\in Q_c}\langle \lambda_{c,u,t}(0), \widetilde{o}_{c,u,t}(\delta) - o_{c,u,t}(0)\rangle\,. \tag{36}$$

This truncates higher-order terms in the downstream map from head outputs to candidate scores; it does not linearize the local softmax in δ. The gradients remain fixed at baseline, and prediction uses neither a fitted response curve nor any nonzero full-model intervention outcome. The linear comparator is

$$\widehat{M}_u^{\mathrm{lin}}(\delta) = M(0) + \widehat{\kappa}_u(0)\delta, \qquad \widehat{\kappa}_u(0) := \frac{\widehat{M}_u(h_{\mathrm{lin}}) - \widehat{M}_u(-h_{\mathrm{lin}})}{2h_{\mathrm{lin}}}, \quad h_{\mathrm{lin}} = 0.05. \tag{37}$$

Thus $\widehat{\kappa}_u(0)$ approximates the derivative of the *predicted* curve at zero, not a derivative estimated from observed full-model responses. Observed curves are obtained separately by editing the logits and rerunning the full model.

Evaluation metrics. We center each observed and predicted curve at its own recorded zero-intervention value and compare these responses using pooled R^2 and MAE, without affine recalibration (Equation (35)). We also report per-curve R^2 and descriptive quartiles of the observed response range. Within each example, heads are ranked by their response range $\max_\delta M_u(\delta) - \min_\delta M_u(\delta)$: Spearman correlation measures rank agreement, top-1 accuracy identifies the observed most responsive head, and top-three coverage asks whether that head is among the three highest predicted ranges. The paired MAE gain compares linear and structural errors within examples. A secondary analysis measures raw-margin zero-crossing precision, recall, and crossing-strength MAE, using the interpolation rule in Appendix F.4.

E.5 Experiment 2: intervention and metric details

Intervention and prediction. Strong (top-three) and weak-control heads are selected per example from the fixed pools by ranked absolute predicted tie derivative; the Gemma and Qwen selection procedures, including the frozen Qwen sites, are detailed in Appendix C.

Only the final prompt query is edited. A common preload $\delta_0 = (G_\star - G_0)/2$ targets an operating gap $G_\star$, using the mean of the two candidate baseline gaps for G_0 in Gemma and candidate 0's gap in Qwen. Gemma uses $G_\star \in \{-4, \ldots, 4\}$; Qwen uses $\{-3, \ldots, 3\}$. At each actual preload, we recompute both candidate gradients and combine their signed contributions from Equation (5) to obtain the analytic margin derivative $\kappa_{\mathrm{an}}(G_\star)$. Unlike Experiment 1, this is a local derivative test with a new downstream gradient at each operating point, not a baseline-only prediction of the entire curve. The measured comparator is

$$\kappa_{\mathrm{FD}}(G_\star; h) = \frac{M(\delta_0 + h) - M(\delta_0 - h)}{2h}. \tag{38}$$

The primary half-widths are $h = 0.10$ for Gemma and $h = 0.50$ for Qwen, selected after BF16 sensitivity analyses; the narrower probes remain in Appendix D.5.

Evaluation metrics. We compare κ_{an} directly with κ_{FD} using pooled R^2 and MAE, separately for strong heads and weak controls, with no fitted regression. We also report median per-curve R^2 across gaps for each example–head pair. A mixing-only comparator drops the outside-mass term in Equation (4) while retaining the other local factors. To test profile shape separately, we correlate the mean absolute measured derivative over strong heads and examples with $2p(1-p)$ across operating gaps.

E.6 Experiment 3: intervention and metric details

Intervention and prediction. We reuse the four Qwen heads and eight evaluation examples from Experiment 2. We edit only the final prompt query, after normalization and RoPE and before the attention score product; keys and values are fixed. Equation (6) gives $n = \nabla_q G / \|\nabla_q G\|$. Gram–Schmidt gives three seeded, mutually orthonormal random tangents $t_r \perp n$. Branch preloads target $G_\star \in \{-1, 0, 1\}$ as in Experiment 2, preserving within-branch key averages and thus the normal. Directions are fixed across gaps, and each preload is fixed during the query perturbations. All directions use the same nominal step,

$$\eta = \min\left\{\frac{0.5}{\|\nabla_q G\|},\ 0.2\,\|q\|\right\}, \qquad \mathcal{D}_d F = \frac{F(q+\eta d) - F(q-\eta d)}{2}. \tag{39}$$

These finite odd responses are not divided by η. The norm cap is inactive, giving the linear gap prediction $\mathcal{D}_n G \simeq 0.5$ and zero first-order tangent response. Sigmoid composition predicts $\mathcal{D}_n p \simeq [\mathrm{sigmoid}(G_\star + 0.5) - \mathrm{sigmoid}(G_\star - 0.5)]/2$ without fitting an amplitude. Full-model runs measure all $8 \times 4 \times 3 \times 4 = 384$ direction conditions.

Evaluation metrics. We measure $|\mathcal{D}_d F|$ for $F = G, p, M$ and $\|\mathcal{D}_d o\|_2$ for head output. Each condition compares the normal response with the median of its three matched tangents. We report the median normal/tangent ratio over the 24 strong conditions at the tie, not a ratio of means (Equation (42)), and paired normal-minus-tangent gains with example-cluster intervals. We also measure gap-prediction MAE over all 96 normal conditions and plot responses across the three gaps.

E.7 Operating-point and normal/tangent figures

Experiment 3 calibration note. Normal-gap predictions are essentially constant by calibration, making their correlation with observed shifts uninformative; the archived correlation-based pass flag is not counted as evidence.

F Additional finite-strength and operating-point results

F.1 Additional recovered families and release boundaries

The following inventory separates recovered measurements from claims that passed their scientific or measurement gates. It does not pool archived and fresh cohorts, or treat a failed audit as evidence of native-checkpoint fidelity.

F.2 CX-ID minimal run: adaptive versus fixed-basis comparison and audit receipt

Measurement checks and side effects (moved from Section 3.1). Three in-run checks were passed: (i) native forward and zero-strength hook agree exactly (maximum deviation 0) on all 250 validation and test examples; (ii) on three validation facts the autodiff derivative agrees with finite differences at two widths to within 0.3%; (iii) baseline-only predictions for all 200 test facts were committed (SHA `ca9b530e...`) before the first test-set forward pass. The separate pre-measurement path-gate receipt (grad on/off, batch layout, and native-path equality of the injected map) was bypassed for this run and is recorded as such in its result file; a re-run of that gate stopped on a prefix/query mismatch. The checks above therefore do not certify the intervention map beyond zero strength. Figure 11 draws this contrast at all three strengths together with the per-arm responses and the recorded side effect (the change in teacher-forced candidate cross-entropy, -0.050 for switching versus -0.014 for the

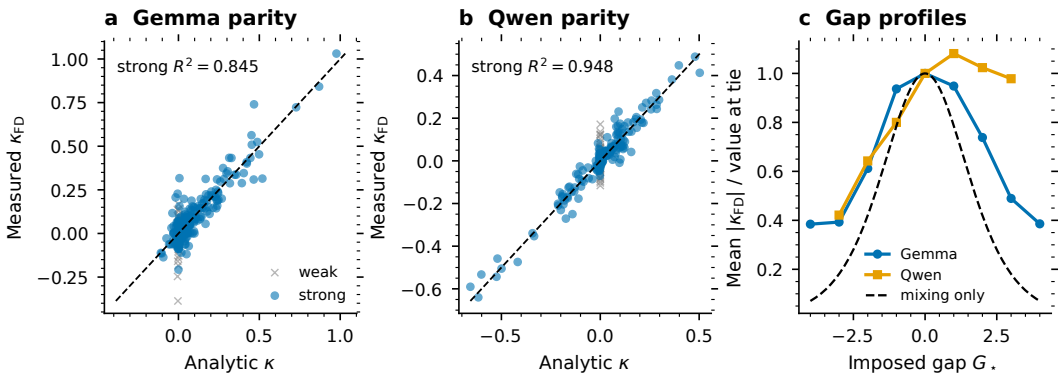


Figure 9: **Experiment 2: derivative agreement and switching profiles.** (a,b) Full analytic derivatives versus measured finite differences, including weak controls. (c) Mean absolute measured susceptibility in strong heads, normalized at the tie. The dashed curve is the normalized mixing factor; it does not assume that the remaining causal factors are constant.

Table 9: Previously unintegrated Caustics families. Entries describe provenance and disposition, not confirmatory effect estimates.

Family	Archived evidence and bounded interpretation
Fresh W3 geometry	A separate frozen 32-example Qwen cohort completed; its geometry-pair rows remain distinct from the archived eight-example geometry in Figure 23. Completion is not an independently validated replacement effect estimate.
Fig. 5 response surface	The receipt-backed rerun records a selected head and single tangent on eight frozen examples. It is a descriptive surface, not general evidence for the proposed mechanism; the earlier surface is superseded.
Fixed-query-basis expansion	The new-example-only summary records 150 additional completed examples. It is not pooled with the original 50-example cohort here, and no new interval or native-fidelity inference is assigned.
E0 measurement retry	Both stored model summaries have `overall_pass=false`. They document a failed measurement audit, not a verified quantitative network result.
W6 variant grid	Archived Qwen and Gemma-2-9b hook-grid rows are descriptive only. The original grid bypassed its negative-control gate, and the recovered Gemma FP32 follow-up reports `AUDIT_FAIL`. Gemma-2-9b is not the Gemma-4 checkpoint in the main experiments.
CX-3 natural-state retention	The reused evaluation cohort has a passing measurement gate but no scientific-pass verdict. Its per-example rows, calibrations, and summary are archived under `results/cx3_source_summary_reconcile_20260926/recovered/` with SHA-256 sums. Figure 10 is exploratory and cannot establish comparative superiority or native fidelity.

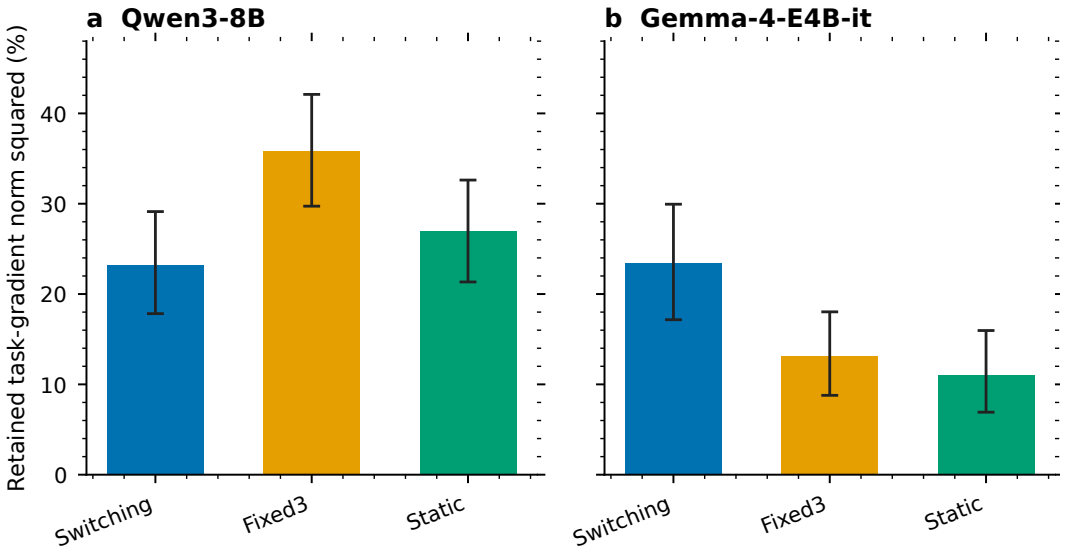


Figure 10: Exploratory natural-state retention differs by controller in the reused evaluation cohort. Bars and example-bootstrap intervals are sourced from the archived CX-3 summary; row-level regeneration is unavailable here. Its measurement gate passed, but `scientific_pass` is unset; the display is not a confirmatory superiority or native-fidelity result.

fixed basis at $\alpha = .05$; the KL divergence of the full next-token distribution was not measured), and Appendix F.2 gives the artifact hashes, commit-pinned paths, and code lines: every statistic in the figure is read from `RESULT.json` or recomputed from the 1 800 measured rows and asserted equal to it.

Against a random fixed three-axis basis the paired contrast is $+0.105$ $[0.090, 0.121]$ at $\alpha = .05$; no static-mean arm was run in this cell. The pre-registered full design is in `prereg/cx_id_full_20260926.json`.

This subsection gives the primary artifact, code lines, and measurement rows behind the CX-ID numbers quoted in Section 3.1, so that the adaptive-versus-fixed comparison can be audited rather than taken on description. Figure 11 is generated by `figures/make_cxid_min_kill_arg.py` from the per-row artifact; every statistic it draws is either read from `RESULT.json` or recomputed from `test_rows.jsonl` and asserted equal to it.

Artifact receipt. The run is `caustics-cx-id-min-20260926` (pre-registration `prereg/cx_id_min_20260926.json`, SHA-256 `c7f15237...`), executed on a single rented NVIDIA RTX A6000 GPU (49 140 MiB, `torch 2.5.1+cu124`, `transformers 5.14.1`); the runner reported 2741.3 s of compute (2911 s including model download; estimated cost $0.32). Result files were committed at `accc436`; from `f3d8d31` on, `results/` is excluded by `.gitignore`, so the tracked copies are read as `git show accc436:results/caustics_cx_id_min_20260926/qwen3_8b/conflictqa/seed11/`*file*. SHA-256 sums of the tracked files (`test_rows.jsonl` has 1 800 rows = 200 facts × 3 arms × 3 strengths):

```
RESULT.json              2301b34d6e0d30ab7a078ccd3655e012feb75eec17992e433d4738434ce2390a
test_rows.jsonl          636d7be19a8e671fa8651450605420adbcaf164d75672a2f35bc8ad75ce0cf92
predictions_frozen.csv   ca9b530eb6f20d387446f0b468fe6661f0d3487c717b005492984b2ddbfdb886
```

The split manifest hashes to `f4b8cfaa...` and the dataset to `c9921b73...`. The runner is `git show 155b696:experiments/cx_id_decision/cxid_run.py` (local regime `LOCAL_MAX=0.05`, line 62; per-fact endpoint and sign-accuracy denominator in `analyze`, lines 361–379) with `cxid_core.py` lines 234–237 (`normalized_mae`). The model is

Qwen/Qwen3-8B at revision b968826d, float32, seed 11, heads $(24, 28)$, $(34, 7)$, $(24, 29)$. In-run gates: audit.jsonl has 250 rows with native-versus-hooked log-probability difference and zero-hook margin difference both exactly 0.0; fd_gate.json passes on all three probe facts. The separate path-gate receipt required by the full design was bypassed on instruction (path_gate=BYPASSED_CXID_PATH_SKIP); the cell therefore does not satisfy the full pre-registration's D1 clause and D3 utility was not evaluated. A separate native-path gate run for this checkpoint (2026-09-26, Modal L40S, float32, torch 2.6.0, transformers 5.3.0, same splits and runner) terminated before writing a receipt: the runner's consistency check that the post-RoPE query at the prompt position is identical under both candidate continuations raised candidate prefix/query mismatch for one of the 303 discovery, controller, and validation facts (conflictqa_003325, head $(24, 29)$, $\max|\Delta q| = 1.55 \times 10^{-5}$ against atol 10^{-5}, prefix tokens identical; 1 of 909 fact–head rows in a full scan, commit 01ff3d5), whereas the same check passed for all 550 facts of the minimal run on the RTX A6000 environment above. The minimal-run statistics are unaffected; the query tolerance was subsequently widened to 10^{-4} for later runs (native, zero-hook, and finite-difference tolerances unchanged), the receipt remains unobtained, and D1 remains unsatisfied. An earlier execution of the same design (155b696:experiments/cx_id_decision/results/cxid_min_run1/, frozen 2026-09-25 12:16 UTC, predictions SHA 129c7815...) yields the same RESULT statistics to fifteen significant digits; the two prediction files differ in hash because their split manifests differ (aa82bbff... versus f4b8cfaa...).

What was and was not measured. (1) One model only (Qwen3-8B), one seed, one in-distribution environment. (2) There is no static_mean arm in this cell; the non-adaptive comparators are the matched fixed three-axis query basis pca_normals3 (primary D2 contrast) and random_fixed3 (secondary). The runner implements static_mean but the minimal specification did not schedule it. (3) The side effect recorded per row is the change in teacher-forced candidate cross-entropy (ce_delta); the KL divergence of the full next-token distribution is not recorded by this runner, and no KL value exists for this run. (4) The endpoint is the per-fact normalized MAE over the local strengths $\alpha \in \{.005, .05\}$ with the validation-fixed floor $\tau = 0.008611986957268529$ nats/token, summarized by the median over 200 test facts; $\alpha = .20$ is measured but excluded. Sign accuracy uses the individual local rows with $|\Delta_{f,\alpha}| \geq \tau$, 263 of 400 for the switching arm (229 and 69 for pca_normals3 and random_fixed3). (5) The measurement path is the native checkpoint with in-run audit, not the custom attention path of Experiments 1–5; the CX-3 companion in panel (d) uses that custom path, a one-sided response ΔM_+, 32 reused examples per model and a different strength grid, and shares only $\alpha = .05$ with the CX-ID grid. The two comparisons are shown side by side and are not pooled.

Values read from RESULT.json. Switching arm: median NMAE 0.06057 (exact-local) and 0.06917 (first-order), sign accuracy $263/263$; mean measured response 0.01066, 0.10067, 0.31279 nats/token at $\alpha = .005, .05, .20$. Fixed basis: median NMAE 0.04099/0.04410, mean response 0.00298, 0.02853, 0.07932. Random fixed: 0.00324/0.00747, mean response -0.00047, -0.00478, -0.02088. Paired contrast switching minus fixed basis: $+0.00767$ $[0.00623, 0.00915]$ at $\alpha = .005$ (Holm $p = 1.0$ against the 0.02 margin), $+0.07213$ $[0.05795, 0.08661]$ at $\alpha = .05$ (simultaneous $[0.05698, 0.08719]$, Holm $p = 0.0$), $+0.23347$ $[0.17985, 0.28611]$ at $\alpha = .20$; minus random fixed: $+0.01112$, $+0.10544$ $[0.09042, 0.12106]$, $+0.33367$. Mean ce_delta (recomputed from the rows, fact bootstrap $B = 2000$): switching -0.0054, -0.0502, -0.1501; fixed basis -0.0015, -0.0141, -0.0368; random fixed $+0.0002$, $+0.0023$, $+0.0104$. Values are truncated for display; the artifact holds full precision.

F.3 Figure-panel source map

Table 10 maps every panel in the submitted figures to its archived row-level input and plotting entry point. Paths are relative to the repository root. In the table, FS means data/finite_strength/, FQ means data/finite_strength_qwen3/, GF means data/gemma_factorization/, QF means data/qwen_factorization/, GE means data/geometry/, HC means data/heldout_controller/, and GD means data/gemma_discovery/. A summary file is named only when the generator uses it; it is

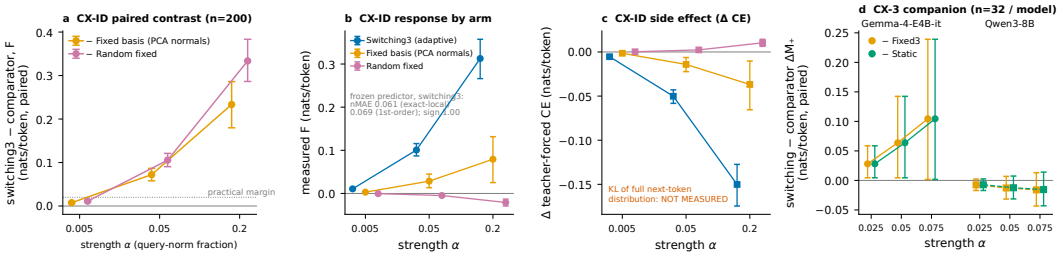


Figure 11: CX-ID minimal run, adaptive versus fixed-basis comparators, with the CX-3 companion. (a) Paired per-fact contrast of the measured response, switching minus the fixed PCA-normal basis and minus a random fixed basis, with the runner's paired fact-bootstrap 95% intervals ($B = 2000$) and the pre-registered 0.02 nats/token margin. (b) Mean measured response per arm with fact-bootstrap intervals recomputed from the rows; the inset gives the frozen switching-arm predictor's median NMAE and sign accuracy. (c) Side effect: change in teacher-forced candidate cross-entropy per arm. KL of the full next-token distribution was not measured and is not drawn. (d) CX-3 natural-state companion from the committed summary (32 reused examples per model, one-sided ΔM_+, custom attention path); different estimand and grid from (a)–(c), shown for side-by-side reading only. All CX-ID values are from `accc436:results/caustics_cx_id_min_20260926/`; see Appendix F.2 for the receipt.

not a substitute for the indicated measurement rows. "Analytic" denotes a schematic or equation-derived panel with no empirical rows, not a missing experimental archive.

The row-level data for the illustrative softmax terms in Figure 1b do not exist by design. The Gemma-2-9b E1 audit archive is now locally present at `results/e1_gemma_20260919T053511/rows.jsonl`; its summary reports `STRUCTURAL_MISMATCH`. Figure 7 plots Qwen only and must not be read as a Gemma pass-count panel. This map records source correspondence, not independent validation of native-checkpoint fidelity or of the archived cohort's scientific gate.

F.4 Finite-strength grid and per-head behavior

The 13 logit shifts are

$$\delta \in \{-2, -1.5, -1, -.75, -.5, -.25, 0, .25, .5, .75, 1, 1.5, 2\}. \tag{40}$$

All 50 examples have complete curves for all eight frozen heads. Table 11 reports equal-weight summaries over their 50 example-level curves. Figure 12 displays errors by head and accuracy by observed effect size. These summaries show both useful nonlinear prediction in the larger-response circuits and the limits of interpreting small-response R^2.

Numerical baseline offsets and centering. The logged baseline from the forward/backward capture and the zero point in the actual intervention sweep are not numerically identical. Across 400 curves, the absolute offset has mean 0.03187, median 0.01953, and maximum 0.17871 nats per token. These are different execution paths, including gradient-enabled capture and intervention microbatching; the archived 400-curve data do not isolate a single source of the discrepancy. The auxiliary CPU audit (`audit/w4_zero_hook_offset_audit.py`, described in Section 3) uses a two-layer decoder configured with Gemma-2 attention constants, not the Gemma 4 E4B-it checkpoint used in Experiment 1. It therefore cannot rule out gradient-mode or batch-layout effects in the archived experiment, identify the cause of its baseline offsets, or establish fidelity to the native checkpoint. The softcap-related behavior reported for the auxiliary decoder is a separate diagnostic, not an explanation of the archived offsets. No checkpoint-matched comparison was performed for this archive. Centering each curve at its own recorded zero removes the constant offset and avoids inflating R^2 with between-example baseline-margin variation. It does not remove δ-dependent numerical error or demonstrate equivalently accurate raw-margin prediction. The figures and primary response metrics consistently use this centered definition.

Exploratory answer-margin crossings. For each raw observed or predicted curve, locate every adjacent-grid sign change and linearly interpolate its zero. Exact zeros are also counted. If multiple

Table 10: Panel-to-archived-row and generator map for every figure included by `main.tex`. Panel letters follow the respective captions.

Figure / panels	Raw rows or analytic source	Figure generator
1 a	schematic, no data	`scripts/fig1_v2_final6h.py`
1 b	analytic identity coefficients ($\rho = .75$, illustrative), no data	`scripts/fig1_v2_final6h.py` (`panel_b` from `scripts/fig1_three_stage.py`)
1 c,d	`results/caustics_cx_id_min_20260926/qwen3_8b/conflictqa/seed11/`: `test_rows.jsonl`, `predictions_frozen.csv`, `RESULT.json`	`scripts/fig1_v2_final6h.py`
3 a,b	`FS,FQ/all_curve_results.csv` joined to `all_baseline_head_stats.csv` by example/layer/head	`analysis/fig_head_trajectory_3d.py`
22 a–d	`FS,FQ/all_curve_results.csv` joined to `all_baseline_head_stats.csv` (branch gap `mean_G_candidate0`); c,d: Gemma only	`analysis/reproduce.py` (`fig2_cui_transition`)
2 a–c	`FS/all_curve_results.csv`, `example_head_metrics.csv`; pooled parity (a), two example curves (b,c)	`analysis/reproduce.py`
23 a–d	`GE/all_normal_vs_tangent_results.csv`, `condition_anisotropy_metrics.csv`	`analysis/reproduce.py`
5 a,b	`data/control_subspace/projection_metrics.csv`; `HC/heldout_causal_results.csv`	`analysis/reproduce.py`
10 a,b	`experiments/cx3_natural_state/results/real_20260925T055152Z/summary.json` (exploratory, reused cohort)	`analysis/plot_cx3_retention_summary.py`
11 a–c; d	`accc436`: `results/caustics_cx_id_min_20260926/qwen3_8b/conflictqa/seed11/RESULT.json`, `test_rows.jsonl` (row recomputation asserted against the summary); d: `155b696`: `experiments/cx3_natural_state/results/real_20260925T055152Z/summary.json`	`figures/make_cxid_min_kill_arg.py`
4 a–d	`results/cxid_e2static_exploratory_qwen3_8b_20260926/`: `test_rows.jsonl`, `predictions_frozen.csv`, `freeze.json` (32 facts; input and output sha256 in `results/ROOT_FIX_manifest.json`)	`figures/make_rootfix_one_cohort.py`
7 a	`results/e1_qwen_20260919T051331/rows.jsonl` (summary hash check)	`code/make_audit_figures.py`
8 a,b	`results/e2_fixed3/e2_expanded_qwen3_8b_n32_20260919/rows.jsonl` (paired-bootstrap hash check)	`code/make_audit_figures.py`
6 a,b	`GD/all_head_scores.csv`, `selected_heads.csv`	`analysis/reproduce.py`
9 a–c	`GF/all_operating_point_results_h0p10_primary.csv`; `QF/all_test_operating_point_results.csv`	`analysis/reproduce.py`
12 a,b	`FS/per_head_summary.csv`, `effect_size_summary.csv`; derived from `FS/all_curve_results.csv`	`analysis/reproduce.py`
13 a,b	`GF/all_operating_point_results_h0p10_primary.csv`; `QF/all_test_operating_point_results.csv`	`analysis/reproduce.py`
14 a–c	`GE/all_normal_vs_tangent_results.csv`, `condition_anisotropy_metrics.csv`	`analysis/reproduce.py`
15 a–c	`HC/frozen_controller.json` (a, tune coefficients); `HC/heldout_causal_results.csv` (b,c)	`analysis/reproduce.py`

Table 11: Gemma finite-strength results by frozen head, with 50 curves per row. All R^2 and MAE values are computed on centered responses.

Layer	Head	Median curve R^2		Mean MAE		Sign accuracy
		Structural	Linear	Structural	Linear	
26	1	.317	.133	.0191	.0204	.66
27	3	.264	-.003	.0190	.0222	.72
28	3	.758	.372	.0221	.0298	.74
28	4	-.182	-.233	.0190	.0202	.66
29	2	.846	.521	.0244	.0483	.86
29	4	.873	.475	.0549	.1033	.88
30	4	.475	.153	.0196	.0231	.72
41	2	.800	.685	.0450	.0622	.98

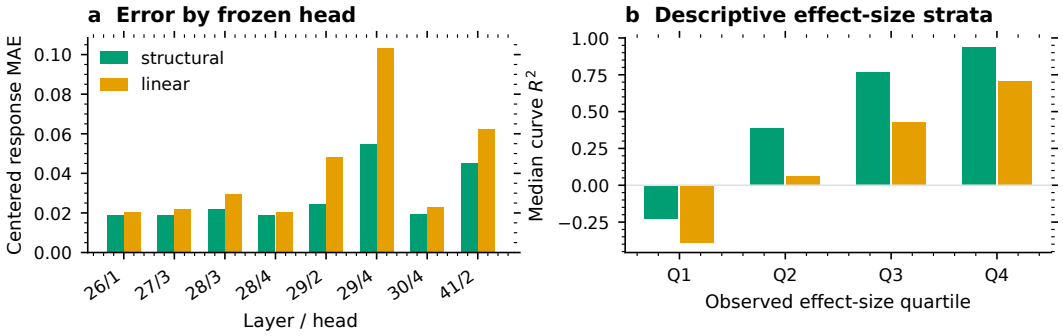


Figure 12: Additional Gemma response summaries. (a) Both predictors have larger absolute errors on some high-effect heads, but structural prediction lowers MAE for every frozen head. (b) Per-curve R^2 improves with effect size. The quartiles are defined by the observed response range and are descriptive strata, not a selection or tuning rule.

Table 12: Descriptive effect-size strata, 100 curves per quartile. Q1 has the smallest observed response ranges. The two R^2 columns are medians across curves, not pooled statistics.

Quartile	Median range	Structural R^2	Linear R^2	Structural MAE	Linear MAE
Q1	.0609	-.231	-.390	.0158	.0164
Q2	.1189	.390	.059	.0200	.0232
Q3	.2174	.770	.426	.0240	.0349
Q4	.6554	.935	.708	.0518	.0902

roots occur, retain the one with smallest $|\delta|$; absence of a root in $[-2, 2]$ means no detected crossing. Table 13 gives the crossing confusion matrix. On the 38 jointly detected curves, the structural crossing MAE is 0.4093 and the predicted sign of the crossing strength is correct in 81.6% of cases. This secondary statistic uses raw margins and is therefore more sensitive to the baseline offsets above than the centered response analysis.

Table 13: Candidate-margin crossing detection over the 400 Gemma curves. This is a crossing of teacher-forced candidate scores, not a free-generation event.

	Predicted crossing	No predicted crossing
Observed crossing	38	9
No observed crossing	4	349

F.5 Full factorization, tie enhancement, and profile differences

Let $\kappa = \kappa_{\text{mix}} + \kappa_{\text{outside}}$ as in Equation (5). The absolute component profiles in Figure 13 separate their effects. Because the two contributions can have opposite signs, $|\kappa|$ is not the sum of their absolute values. Even the mixing contribution need not be proportional to $2p(1-p)$ when ρ and $\langle \lambda, \overline{v}_1 - \overline{v}_0 \rangle$ change with the operating point.

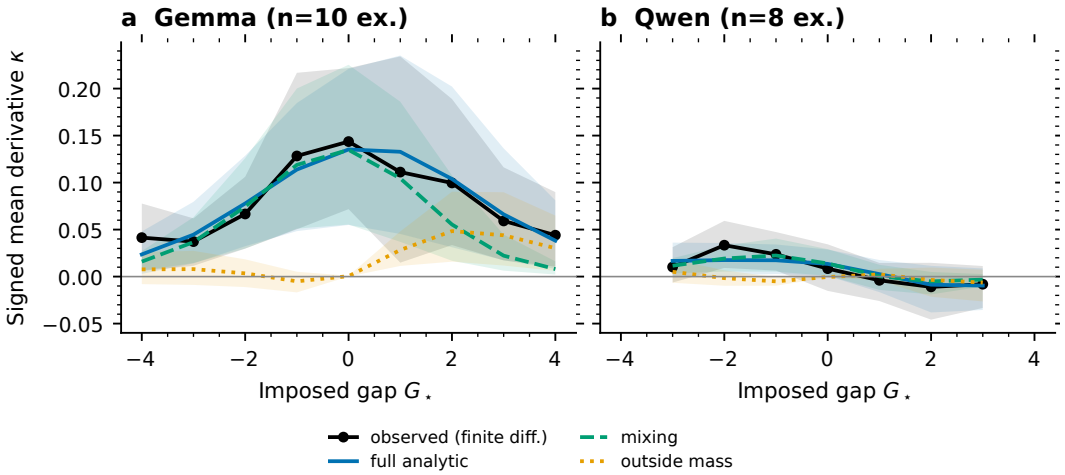


Figure 13: Mean absolute derivative components in the selected strong sites. Outside mass contributes substantially away from the tie, especially in Qwen. Absolute contributions are shown separately and are not additive because of possible cancellation.

For each example–head pair, define tie enhancement as the absolute measured derivative at $G_\star = 0$ minus the average absolute derivative at far operating points. Far means $|G_\star| \geq 3$ in Gemma and $|G_\star| \geq 2$ in Qwen. We then average over heads within each example and bootstrap examples. The recomputed strong-site enhancements are 0.09079 in Gemma (95% interval $[0.03896, 0.14319]$) and 0.03229 in Qwen ($[0.00281, 0.07022]$). The corresponding weak-control intervals contain zero. These are average contrasts, not a claim that every circuit peaks at the tie: only 16/30 strong Gemma curves and 11/24 strong Qwen curves attain their largest measured absolute derivative within $|G_\star| \leq 1$ on the evaluated grids.

The normalized Gemma average profile has Pearson correlation 0.9785 with the logistic mixing factor. Normalizing each pair before averaging instead gives 0.9187 in the archived secondary

analysis. For Qwen's wide-step probe, the unnormalized mean absolute profile has correlation 0.4831. The quantitative support is therefore strongest for the full derivative law, while mixing-only profile universality is model- and circuit-dependent. All intervals in this section are descriptive; the small number of example clusters limits precision and sensitivity to heterogeneous circuits.

G Detailed normal-versus-tangent experiment

G.1 Coordinates, matched directions, and measured quantities

The experiment intervenes on the actual normalized, post-RoPE query entering Qwen's attention score product. It does not perturb the residual stream, token embedding, pre-normalization query, or the key/value vectors. Its Euclidean norm comparison is therefore specific to this internal coordinate system; it need not be preserved by pulling the perturbation back through query normalization. The full model is rerun after each intervention to measure candidate scores.

For each example–head pair, the baseline capture determines $g_q = c_{\mathrm{att}}(\overline{k}_1 - \overline{k}_0)$ and $n = g_q / \|g_q\|$. Seeded Gaussian vectors are orthogonalized against n and previously accepted tangents:

$$\widetilde{t}_r = u_r - \langle u_r, n\rangle\, n - \sum_{s<r} \langle u_r, t_s\rangle\, t_s, \qquad t_r = \widetilde{t}_r / \left\|\widetilde{t}_r\right\|. \tag{41}$$

The seed is derived deterministically from the experiment seed, example ID, layer, and head by SHA-256, as implemented in the supplied notebook. Three accepted tangents are used. Degenerate normals would be rejected, but all 32 intended example–head pairs complete in the supplied data.

The normal and tangents receive the same FP32 step norm η in Equation (39). Steps are cast to BF16 when applied to the model, so exact equality refers to the designed perturbations, not a guarantee of identical realized displacements after rounding. Across all pairs, $\eta / \|q\|$ ranges from 2.67% to 7.52%; no 20% cap is active. The predicted normal gap shift is therefore 0.5 for every pair to FP32 precision. The maximum absolute predicted tangent gap shift is 1.86×10^{-8}. The same four directions and step norm are reused at gaps $-1, 0, 1$.

For condition i and scalar quantity F, define

$$N_i^F = |\mathcal{D}_n F|, \qquad T_i^F = \mathrm{median}_{r=1,2,3}\, |\mathcal{D}_{t_r} F|, \qquad A_i^F = \frac{N_i^F}{T_i^F + 10^{-12}}. \tag{42}$$

For $F = o$, replace the absolute value by the Euclidean norm. Reported anisotropy is the median of A_i^F across the 24 strong conditions at the tie, not the ratio of the mean normal and tangent responses. For the paired absolute gain, use $N_i^F - T_i^F$, average over the three strong heads within each of eight examples, and bootstrap those eight means. The numerical denominator floor is only a division safeguard; it does not make ratios with near-zero tangent responses statistically meaningful.

G.2 Quantitative results at the tie

Table 14: Normal/tangent responses at $G_\star = 0$ in the three strong Qwen heads. The first two numeric columns are means over 24 example–head conditions; intervals on the paired gain resample the eight examples. Margin units are nats per token.

Quantity	Mean normal	Mean tangent	Median ratio	Mean gain	95% gain interval
Gap G	.49912	.00574	121.10	.49337	[.49023, .49626]
Responsibility p	.12224	.00144	118.69	.12081	[.12003, .12148]
Head output o	2.10167	.52464	3.916	1.57703	[1.20688, 2.00077]
Margin M	.04483	.01400	2.971	.03083	[.00894, .05680]

The gap and responsibility comparisons validate the implemented coordinate construction and its finite-step behavior. The output and margin comparisons establish propagation through value aggregation and downstream computation relative to the sampled controls. No singular-value decomposition of the full query-to-output or query-to-margin Jacobian is performed, so these data should not be described as a measured rank-one Jacobian.

Finite-step logistic comparison. With approximately linear gap changes $G(q \pm \eta n) \simeq G_\star \pm 0.5$, the predicted odd responsibility response is

$$\mathcal{D}_n p \simeq \frac{\mathrm{sigmoid}(G_\star + 0.5) - \mathrm{sigmoid}(G_\star - 0.5)}{2}. \tag{43}$$

Figure 23 uses this finite-step expression, without fitting its amplitude. The infinitesimal alternative is $0.5\, p(G_\star)(1 - p(G_\star))$. At the tie the two predict approximately 0.12246 and 0.125, respectively. Only three gap values are available, so a near-perfect three-point profile correlation cannot establish a broadly validated functional shape.

G.3 Normal-gap calibration and the failed correlation flag

The notebook included a check requiring Pearson correlation at least 0.95 between predicted and observed normal gap shifts. Its recorded value is -0.1929, so the archived check fails and is *not* reclassified as a pass. However, the predicted values range only from 0.4999999404 to 0.5000001192: calibration deliberately makes them constant. Correlation with this numerically tiny variation is not a meaningful assessment of displacement accuracy. As a diagnostic, the mean absolute prediction error across all 96 normal conditions is 0.00603, or 1.21% of the nominal shift (Figure 14a). This error analysis explains the inappropriate correlation criterion; it is not a replacement claim that the original threshold was met.

G.4 Weak controls and limits of the anisotropy claim

The frozen weak head $(13, 12)$ has median discovery union mass about 1.3×10^{-5}. At the tie, its mean odd output norms are 1.44×10^{-5} for the normal and 1.06×10^{-5} for the tangent median, compared with 2.10 and 0.525 in strong heads. Yet its mean absolute margin responses are 0.02555 and 0.01032, and the median margin ratio is 267.8 because some tangent denominators are extremely small. The example-bootstrap interval for the weak paired margin gain includes zero ($[-0.00152, 0.02802]$ in the accompanying reanalysis). These observations do not support treating that large ratio as a reliable causal effect. The weak site's poor operating-point parity and very small local displacement require particular caution about BF16 downstream score differences.

The evidence is consequently specific: nominally matched perturbations along the branch normal produce larger output and margin responses than three sampled orthogonal controls in the frozen strong circuits and this evaluation cohort. It does not establish optimality over all query directions, an intrinsic rank of the transformer, or localization of the unmodified model near the wall.

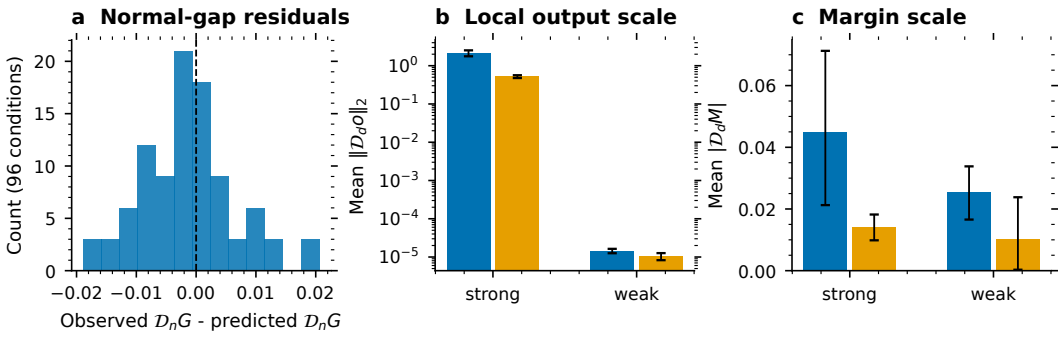


Figure 14: Geometry diagnostics. (a) Normal-gap prediction residuals across all 96 conditions. (b,c) Mean normal (blue) and tangent (orange) responses at the tie, with example-cluster intervals. The weak control has a very small local output change (log scale in b), while downstream margin differences remain comparatively large. Ratios alone would conceal this scale discrepancy.

H CONTROL-SUBSPACE AND HELD-OUT-CONTROLLER DETAILS

H.1 FROZEN COHORTS AND INTERVENTION COORDINATE

Experiments 4–5 use only the three strong Qwen heads in Equation (8); the weak head is not part of the controller. Experiment 4 reuses the eight frozen Qwen evaluation IDs listed in Appendix C. It performs no new discovery, but it does use each evaluation example's task gradient diagnostically. Each of the three heads contributes one 128-dimensional post-normalization, post-RoPE query block, so the joint intervention space has dimension 384 and the switching subspace has dimension three.

Experiment 5 excludes those eight examples, then fixes 16 tune and 32 test examples by the same seeded ConflictQA preprocessing. The tune, test, and earlier eight-example cohorts are pairwise disjoint. The full lists are in `data/heldout_controller/split_manifest.csv`. All 16 tune gradients are computed before the test-gradient firewall is activated. On test, the code permits baseline geometry capture and forward interventions but raises an error on any call to the task-gradient routine. All 32 intended test examples and all six controllers at three strengths complete; both runtime-failure files are empty.

Both experiments concatenate the three baseline query vectors and use $s = \alpha \|q\|$ with $\alpha \in \{.025, .05, .075\}$. The same joint norm is used for every direction on a given example. Responses use symmetric $\pm s$ interventions, so the odd part in Equation (45) cancels the unmodified margin and much of the even finite-step response. Model weights remain BF16, while direction construction and norm bookkeeping use FP32. As in Experiment 3, Euclidean energy is specific to this internal query coordinate.

H.2 Projection and finite-control construction

Because the three normal basis vectors occupy disjoint head blocks, they are mutually orthonormal after within-head normalization. For any orthonormal basis matrix Q_S,

$$P_S g = Q_S Q_S^\top g, \qquad R(S) = \frac{\left\|Q_S^\top g\right\|^2}{\|g\|^2}, \qquad u_S = \frac{Q_S Q_S^\top g}{\left\|Q_S Q_S^\top g\right\|}. \tag{44}$$

Each random comparator also contains exactly one normalized Gaussian direction in each of the three head blocks. Hence it preserves both dimension and head allocation; its isotropic expectation is $\dim S/\dim \mathcal{U} = 3/384 = 0.0078125$. Sixty-four independent seeded subspaces are drawn per example. The causal random comparator uses the draw whose retained fraction is nearest the within-example median, rather than selecting a particularly good or bad random draw. The tangent basis similarly contains one seeded direction in each block, orthogonalized against that block's switching normal.

The tie preload in Experiment 4 is calibrated layer by layer because an earlier-layer preload can change the query and keys at a later layer. Candidate 0 determines the local geometry, while the final margin uses both candidate sequences with their shared causal prefix. The recorded maximum absolute final gap is 1.91×10^{-6}.

Archived scalar-map extraction for Experiments 4–5. The following is a source-level map, not a retrospective fidelity pass. Line numbers refer to the archived Python exports of the original notebooks, included in the supplementary code.

```
F_x(z; p) = score(candidate1, z, p) - score(candidate0, z, p)
score(c, z, p): configure masks/preloads/deltas for c;
                forward c with cache off; mean scored-token log p
Q_x: capture candidate0 at z=0, then form three block normals
g_x: grad_z score(candidate1, 0, p)
     - grad_z score(candidate0, 0, p)
finite(u): [F_x(+s*u; p) - F_x(-s*u; p)]/(2*s)
```

For Experiment 4, `candidate_logp` is at lines 684–690, `_forward_candidate` at 692–710, and `run_margin` at 739–748 in `qwen3_switching_control_subspace_test.py`. The candidate 0 capture for preload calibration is at 835–843, and the condition-level zero-increment capture for geometry is at 1008–1032; the basis construction at 856–866, candidate-wise gradients at 750–774, and signed finite interventions begin at 959. The query increment is added to the live query inside the hook at 611–634, before the attention scores are formed. For Experiment 5, the corresponding source is `qwen3_switching_heldout_controller_test.py`: candidate score 638–665, two-candidate margin 693–702, candidate-wise gradients 704–729, geometry capture at 837–847, tune gradient and $Q_x^\top g_x$ storage at 873–889, and signed finite interventions beginning at 1074. Thus the code has a candidate-wise forward and a candidate 0 geometry capture, rather than an independent downstream query-state overwrite. Whether the archived values were produced by these exact source revisions and whether AD agrees with finite differences remain unresolved by E1.

H.3 Tune-only coefficient fitting and matched controls

The fitted switching vector in Equation (10) is a unit vector in coefficient space. Bootstrap resampling the 16 tune examples gives stable signs and a median cosine 0.994 with the full-tune coefficient vector (95% interval $[0.941, 1.000]$):

The controls separate different possible explanations:

- *Tangent learned*: one deterministic tangent per head and example, with three coefficients fitted from tune gradients by the same normalized-mean rule.
- *Adaptive random learned*: one deterministic isotropic direction per head and example, again with only three tune-fitted coefficients.
- *Shuffled switching*: the learned switching coefficients are retained, but each test example receives the three normals of the next test example in a cyclic permutation. This preserves the marginal distribution of normal vectors while breaking input alignment.

Table 15: Frozen switching coefficients. The intervals resample the 16 tune examples and refit the normalized three-vector. Head order matches Equation (8).

Layer / head	Coefficient	95% interval
24 / 28	-.723	$[-.835, -.601]$
34 / 7	.354	$[.076, .542]$
24 / 29	.593	$[.431, .737]$

- *Equal positive*: the input's correct normals are retained with coefficients $(1, 1, 1)/\sqrt{3}$, testing whether fitted signs and relative weights matter.
- *Static tune gradient*: the normalized mean of the full 384-dimensional tune task gradients. This baseline has no test gradient and uses the same intervention norm, but learns an unconstrained static direction rather than three coefficients in an input-adaptive geometric basis.

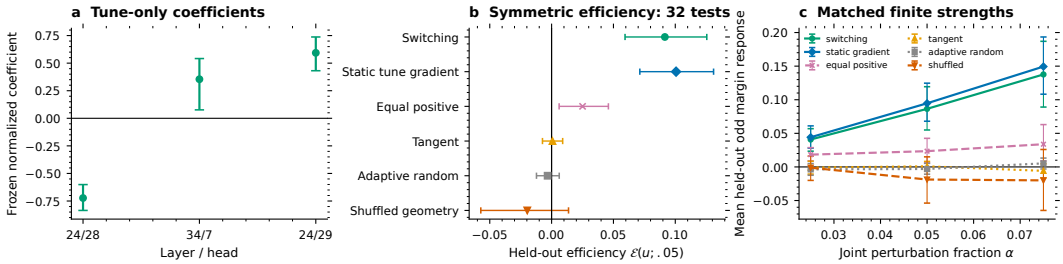


Figure 15: Additional controller diagnostics. (a) Tune-fitted switching coefficients with 95% bootstrap intervals resampling tune examples. (b) Symmetric intervention efficiency $\mathcal{E}(u; .05)$ (Equation (45)), the endpoint originally specified for this archived cohort. (c) Mean held-out odd margin response across the three tested norm fractions. Panels (b,c) use the same 32 held-out examples as Figure 5b and are not independent replications. Example-bootstrap intervals, where shown, apply only to controllers without cyclic test-example donor dependence; any displayed interval for the cyclic shuffled-normal controller is not a valid independent-example population interval and is not used for inference. These symmetric diagnostics establish neither one-sided superiority nor equivalence to the static controller.

H.4 Inference scope

The originally specified endpoint for this archived 32-example cohort was symmetric intervention efficiency at norm fraction $\alpha = .05$. The additional one-sided analysis is post-hoc and descriptive. Its static comparator is the normalized mean of the 384-dimensional tune gradients; it is not a fixed three-dimensional basis with three fitted coefficients. No matched fixed-basis result is reported for this cohort.

The marginal intervals in Table 19 are not intervals for paired between-controller differences; the paired switching-minus-static per-example difference is reported separately in Table 16, and its

95% interval also spans zero. We do not infer one-sided superiority from those marginal intervals or transfer inferential claims from the symmetric endpoint to the one-sided endpoint. The cyclic shuffled-normal control introduces dependence between test examples, so an independent-example bootstrap does not by itself justify population-level inference involving that control. No equivalence or non-inferiority claim is made. Multiplicity correction does not convert an outcome-informed endpoint choice into a confirmatory analysis; confirmation requires a separately frozen analysis on outcome-unseen data.

The controller is held out only with respect to coefficient fitting and per-test-example task gradients. It still receives the evidence-token branch masks and recomputes local normals from each test input. Its three heads were selected by the earlier six-example Qwen causal scan. Consequently, the experiment supports transfer of a signed combination within a fixed mechanism-defined basis; it does not establish unsupervised branch discovery, a model-wide control atlas, cross-task transfer, or reliable changes in sampled generations.

I HELD-OUT CONTROLLER DETAILS MOVED FROM THE MAIN TEXT

The following material was moved from the main text without change to any reported number.

I.1 EXPERIMENT 4 PRELOAD AND EFFICIENCY DEFINITION

The eight examples and three heads are frozen from Experiment 2. We jointly preload the three heads to $G_h = 0$ (maximum recorded error 1.91×10^{-6}), then apply matched joint-query perturbations. Let $F_x(z)$ be the additive joint-query map specified in Appendix E, with masks, preloads, and directions fixed from zero-increment calibration. Let q_x concatenate the zero-increment query blocks and set $s_x = \alpha \, \|q_x\|$. For unit injection direction u and $s_x > 0$, define

$$\mathcal{E}_x(u; \alpha) = \frac{F_x(s_x u) - F_x(-s_x u)}{2 s_x}. \tag{45}$$

The tested norm fractions are .025, .05, and .075, with .05 the originally specified primary strength. The one-sided response is $\Delta M_+ = F_x(s_x u) - F_x(0)$. The budget constrains injected increments, not the total change of downstream query states relative to baseline. These definitions do not establish correspondence with the archived finite-response routine.

I.2 COEFFICIENT OBJECTIVE AND SHUFFLED-CONTROL HANDLING

The normalized mean coefficient rule maximizes the average local directional derivative per unit query displacement. It need not maximize the first-order mean one-sided change when step sizes vary across examples. Writing $s_e = \alpha \|q_e\|$, that objective is

$$J(a) = N^{-1} \sum_e s_e b_e^\top a, \qquad \|a\| = 1.$$

If $\widetilde{b} = N^{-1} \sum_e s_e b_e$ is nonzero, the maximizing coefficient vector is $\widehat{a} = \widetilde{b} / \|\widetilde{b}\|$. The archived controller was not refitted to this objective. We retain its original coefficients and interpret the one-sided results accordingly.

For the shuffled controller, each test example receives normals from another member of the same test cohort. We report its observed mean and sign counts descriptively, but omit an independent-example bootstrap interval because that resampling does not preserve the donor dependence. These summaries do not establish population-level superiority over the shuffled controller.

I.3 PAIRED SWITCHING-MINUS-STATIC DIFFERENCE

J NATIVE QWEN3-8B GEOMETRIC-CONTROL COMPARISON

This experiment is distinct from archived Experiments 1–5, the earlier CX-ID run, and the CX-3 oracle comparison. It measures whether an input-dependent switching-normal controller can increase

Table 16: Paired one-sided response difference (switching minus static mean tune gradient), computed per example before aggregation, on the same 32 unseen examples as Table 19. Intervals are 4,000 example-bootstrap percentile intervals over the paired per-example differences (not a difference of the two marginal intervals above).

Comparison	Mean paired diff.	95% interval	diff. > 0
Switching − static, learned coefficients	0.0010	$[-0.0235, 0.0260]$	15/32

a specified candidate margin, with a static mean-gradient and donor-normal comparison at the same injection norm. Its recorded run is `qwen3_geometry_main_56d0cefebeae`, code version 1.2.0. All files below are under `results/native_control_qwen3_20260926/`.

J.1 Data, scope of held-out evaluation, and frozen design

The official ConflictQA GPT-4 PopQA source contains 9,544 records, reduced to 8,821 normalized unique questions before splitting; normalized duplicate evidence passages are also checked. The SHA-256-pinned source and download URLs are recorded in `manifest.json` and `data_loader.py`. Branch 0 is the source's `parametric_memory` span and branch 1 its `counter_memory` span; their candidate answers are `memory_answer` and `counter_answer`. Branch 1 is a specified steering target, not a factual-correctness label. Document order is assigned deterministically from the question ID. Tokenizer offsets identify the evidence spans. Eligibility uses token lengths (at most 768 total tokens and 24 tokens per candidate), not intervention outcomes. EOS is not added to the score.

The seed is 20260926. The tuning/audit/test split contains 64/12/200 mutually disjoint normalized questions; a preceding 4/2/4 smoke split is disjoint from these and is excluded from scientific inference. Full token sequences, candidate tokens, branch masks, and order are preserved in `encoded_inputs.json`. The three head sites and norm fraction $\alpha = .05$ are fixed before the main outcomes. All test directions are saved before any intervened test score is measured; `prediction_commitment.json` commits the manifest, controller, and ordered plans. This is a local reproducibility commitment, not public preregistration or a guarantee that the questions have never appeared in prior research.

Retrospective archival-overlap audit. The frozen run originally recorded historical overlap as unaudited. After receiving the manuscript source package, we mapped archived source-row IDs and question strings to the same checksum-verified source and normalized-question keys. Across 32 cohort groups in 14 listed archival manifest files, 22 of the 200 test questions and 5 of the 64 tuning questions occur in their union; none of the 12 audit questions does. Fifteen of the test questions occur in the earlier CX-ID discovery/controller/validation/test sets (3/7/3/2, respectively). These counts describe manifest membership, which alone does not prove that every historical row was executed. Broad eligibility pools and later multi-seed planned manifests are excluded from that exposure count. Semantic paraphrase and entity overlap are not tested. Thus "held out" refers to the current coefficient fit, and this cohort is not entirely previously unexamined. The audit changes no frozen inputs or outcomes. Its file list, hashes, matched IDs, and scope are in `results/native_control_cohort_overlap.json`; the reproducible script is `scripts/audit_native_cohort_overlap.py`.

J.2 Native intervention coordinate and controller families

The checkpoint is `Qwen/Qwen3-8B`; its full resolved revision is

`b968826d9c46dd6066d109eabc6255188de91218`

The run uses Transformers 4.51.3, PyTorch 2.11.0+cu128, an A100 40GB, FP32 weights, no quantization, and disabled TF32. The intervention wraps the native SDPA callable and adds to the realized, post-QK-normalization, post-RoPE query at the final prompt position. It preserves the original grouped-query head mapping and mask. A later-layer query is not replaced by a frozen baseline: upstream-to-downstream gradient paths remain present. Only these three additive query blocks are intervention coordinates. Each candidate is teacher-forced separately, with mean token log probability

and candidate difference as in Equation (7). Score reductions and candidate-gradient accumulation use FP64 while the native forward/backward computation remains FP32.

Let g_e be the derivative with respect to the additive intervention coordinate and r_e the baseline joint-query norm. The switching and static coefficients are fitted on the same 64 tuning examples:

$$\widehat{a} = \frac{b}{\|b\|}, \quad b = \frac{1}{64}\sum_e r_e Q_e^\top g_e; \qquad u_{\text{static}} = \frac{v}{\|v\|}, \quad v = \frac{1}{64}\sum_e r_e g_e. \tag{46}$$

The radius weights match the intervention budget. Because each normal occupies a distinct block, $Q_x^\top Q_x = I_3$ and $\|Q_x \widehat{a}\| = 1$. The static-block comparator normalizes each head block of u_{static} and scales it by $|\widehat{a}_h|$. It therefore matches the switching controller's headwise and total injection norms while retaining fixed within-head directions.

For the donor control, a hash-based assignment chooses three distinct tuning examples for each evaluated input, excluding itself during tuning. Each supplies three unit normals Q_j. The donor coefficient vector is independently fitted by normalizing the mean $r_e Q_j^\top g_e$ over these tuning assignments. On each test question, all three donor interventions are measured and their effects averaged. This gives the donor controller the same tuning gradients and its own three coefficients. It does not use other test questions as donors. Every arm injects αr_x in the joint space; equality of directly injected norms does not imply equality of the total realized query displacement after upstream effects.

J.3 Numerical gates and pre-outcome amendments

Before coefficient fitting, each of the 12 audit examples is checked for native-versus-zero intervention agreement, candidate-prefix query/key agreement, local-normal differentiation, and joint intervention derivatives along normal, random, and task-gradient directions. Output/logit identity tolerance is 10^{-5}; shared-prefix query/key tolerance is 10^{-4}. Directional finite differences use

$$D(h) = \frac{M(h) - M(-h)}{2h}, \qquad R(h) = \frac{4D(h) - D(2h)}{3}, \qquad h \in \{.01, .02\},$$

where h scales a unit joint direction by r_x. Both estimates must individually satisfy $|R(h) - A| \le .0005 + .03\max(|R(h)|, |A|)$ against the autograd derivative A. The difference between the two estimates is also checked against the sum of their two error budgets; this is a redundant consistency condition, not an independent convergence certificate. All raw margins are retained. After fitting, all learned controller directions, including the three donor replicates, pass the same derivative check on the first audit example.

Versions 1.0 and 1.1 stopped before any main tuning or test outcomes. The first amendment used FP64 score/gradient accumulation and fixed Richardson estimates to address weak-direction finite differences. The second corrected a supplementary pairwise check that had applied only one error budget to the difference of two estimates. Individual derivative, identity, and prefix tolerances were unchanged; exact encoded cohorts, head sites, intervention strength, and scientific thresholds were retained. The successful version reran the audits and fitting under a new code hash. These are disclosed numerical revisions made using audit data, not post-outcome selection of a successful scientific result.

All 12 audit examples and the learned-controller audit pass. Including the repeated fresh runtime audit, 90 directional derivative comparisons satisfy the fixed criteria; the maximum error-to-bound ratio is 0.7791. The 1,200 measured interventions have equal prescribed injection norms to numerical precision (maximum absolute bookkeeping discrepancy 2.23×10^{-16}). These are checks of the specified coordinate and numerical path, not a claim that identity and derivatives were separately tested on all 200 outcome examples.

J.4 Results and interpretation

The endpoint is $\Delta M_x = M_x(\Delta q_x) - M_x(0)$, with larger values favoring the designated counter-memory candidate. Table 17 gives the recorded arm summaries. Confidence intervals resample whole questions 10,000 times; donor replicates are averaged within question before resampling. Paired contrasts preserve the same test questions in both arms. The two co-primary comparisons use

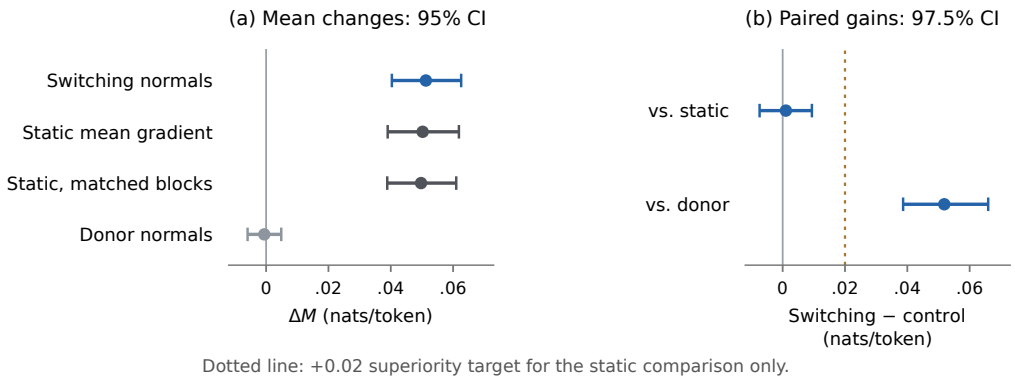


Figure 16: Native Qwen3-8B control comparison (appendix-only; not a main-text claim). 200 test examples held out from the coefficient fit, equal injected query norm ($\alpha = .05$). (a) Mean candidate-margin change with 95% fact-bootstrap intervals. (b) Paired contrasts with 97.5% intervals. The switching-minus-static interval spans zero and the prespecified superiority target of .02 is not met (`TARGET_GAIN_NOT_SUPPORTED`).

Table 17: Native Qwen3-8B control at $\alpha = .05$ on 200 test questions. Arm intervals are 95%; contrast intervals are 97.5%. All values are nats/token.

Controller	Mean ΔM	Improved	Interval
Switching normals	.051265	152/200	[.040346, .062590]
Static mean gradient	.050231	161/200	[.039008, .061864]
Static, matched head norms	.049718	157/200	[.038869, .061005]
Donor normals (three-donor mean)	-.000592	95/200	[−.005949, .004859]
Switching minus static	.001034	—	[−.007403, .009380]
Switching minus donor	.051857	—	[.038652, .065948]

97.5% two-sided intervals for conservative simultaneous coverage of at least 95% under the bootstrap approximation.

The prespecified success rule required a switching-minus-static advantage greater than .02 and a switching-minus-donor advantage greater than zero. Only the donor requirement is met. The result therefore retains its recorded verdict `TARGET_GAIN_NOT_SUPPORTED`; the upper endpoint of even the 97.5% static contrast is below .02. The positive interpretation concerns the demonstrated usefulness of the three-normal controller and its input-matched advantage against the donor control. It does not replace the failed superiority criterion, establish equivalence to the static gradient, or measure full-gradient recovery. The old gradient-retention percentages from the eight-example or CX-3 cohorts are not measurements on these 200 examples.

Reproduction. `scripts/audit_native_control.py` verifies stored hashes, split membership, numerical criteria, injection norms, and arm/paired-bootstrap summaries and regenerates `analysis/native_numbers.tex`. `scripts/plot_native_control.py` regenerates Figure 16. Both operate on recorded data without a model forward pass. The preserved run source is `results/native_control_qwen3_20260926/experiment_core.py`; its SHA-256 is recorded in `settings.json`. Numerical amendment records are retained separately in `results/native_control_numerical_amendments/`.

K External semiclassical extensions (non-load-bearing)

Independent research notes by Ryosuke Sakamoto (Kyoto University, Sept. 2026) extend the asymptotic normal form of Theorem 1 and the exact finite-key law of Proposition 1 in directions this manuscript had not separately stated. We record the statements here for completeness. Proofs of the transport, leakage, and two-key susceptibility statements are given in Appendix A, each with an explicit pointer to the source note's section; the convex-construction statement is a cited summary whose full proof is not reproduced here. None of the experiments in Section 3 or the conclusions in Section 5 depend on these additions.

Transport through a fixed smooth downstream map. Equation (13) describes the derivative of the immediate softmax output U_ε only. For a fixed smooth downstream map, the same singular shape persists, with rank at most one:

Proposition 2 (Transport through a fixed smooth downstream map, cf. Sakamoto *Semiclassical Attention Fronts* Theorem 3.1)**.** *Assume the two-branch case of Theorem 1. Let Ψ be a fixed, ε-independent C^2 map defined on a neighborhood of the relevant output profiles, with bounded first and second derivatives there. Define*

$$T_\varepsilon = (1-\pi_1^\varepsilon)V_0 + \pi_1^\varepsilon V_1, \qquad W_\varepsilon = \Psi(U_\varepsilon).$$

Then

$$D_x W_\varepsilon = (\pi_0^\varepsilon \pi_1^\varepsilon/\varepsilon)\, D\Psi(T_\varepsilon)(V_1 - V_0)(\nabla_x g)^\top + \mathcal{O}(1). \tag{47}$$

Thus the leading singular term has rank at most one. For a scalar readout the same formula gives its differential.

A fixed finite composition of smooth blocks is covered only when its state variables and derivative bounds satisfy these assumptions; the statement is not a rank-one claim for a full Transformer Jacobian.

Corollary 3.2 constant (already independently derived). The peak height $1/(4\varepsilon)$ and full width at half maximum $4\varepsilon \operatorname{arcosh}\sqrt{2} \simeq 3.5255\,\varepsilon$ recorded in Appendix A (Rank-one singular susceptibility) are numerically identical to Sakamoto's Corollary 3.2; we derived them independently and only add the cross-reference.

Omitted-key leakage bounds. The outside-mass term in Equation (4) is an exact identity for the general routing law of Proposition 1. A complementary, explicitly quantitative bound isolating the omitted-key contribution alone holds in the frozen-value, affine-action special case:

Proposition 3 (Leakage bounds, cf. Sakamoto *Low-Temperature Attention Switching* Proposition 3.1)**.** *With unit prefactors, fixed values v_j, and affine actions $A_j(t) = a_j + b_j t$, write $\omega = 1 - \rho_\varepsilon$ for the mass omitted from the selected pair, $D = \max_{j,k} \|v_j - v_k\|$, and $B = \max_j b_j - \min_j b_j$ for the spread of* all *keys' action slopes (selected and omitted alike; the omitted-keys-only spread is insufficient, since it fails to bound the rate at which mass moves between the selected and omitted groups). Let L_1, L_2 bound $\|dF\|$ and $\|d^2F\|$ on the value convex hull. Then, without a dominance assumption on the selected pair,*

$$\|U_\varepsilon - T_\varepsilon\| \le D\omega, \qquad |F(U_\varepsilon) - F(T_\varepsilon)| \le L_1 D\omega, \tag{48}$$

$$\|U'_\varepsilon - T'_\varepsilon\| \le 3DB\omega/\varepsilon, \qquad \big|(F \circ U_\varepsilon)' - (F \circ T_\varepsilon)'\big| \le \frac{\omega}{\varepsilon}\big(3L_1 DB + L_2 D^2 B\big). \tag{49}$$

If every unselected key's action is exceeded by both selected actions by $\eta > 0$, then $\omega \le (N-2)e^{-\eta/\varepsilon}/2$, recovering the exponential leakage bound $\|U_\varepsilon - T_\varepsilon\| \le Ce^{-\eta/\varepsilon}$, $\|dU_\varepsilon - dT_\varepsilon\| \le C\varepsilon^{-1}e^{-\eta/\varepsilon}$ used implicitly wherever this manuscript treats $B_0 \cup B_1$ as isolated.

Equations (48)–(49) sharpen, for the affine/frozen-value case, the qualitative outside-mass term already present in Equation (4); the two are related but not the same statement, since Equation (4) is exact and holds without the affine/frozen-value restriction, while Equations (48)–(49) are asymptotic bounds isolating ω alone.

Two-key affine frozen-value susceptibility constant. In the exactly-two-key, frozen-value, affine-action special case of Proposition 1, Sakamoto's finite-step susceptibility law gives a closed-form constant for the margin derivative at the tie. The statement below is restricted to exactly two keys; it does not assert anything about configurations with additional keys.

Proposition 4 (Two-key affine frozen-value susceptibility, cf. Sakamoto *Low-Temperature Attention Switching* Theorem 2.2 and Corollary 2.3)**.** *Assume exactly two keys with unit prefactors and fixed values v_0, v_1. Their actions are affine, independent of ε, and satisfy $A_1(t) - A_0(t) = t$. Let F be continuously differentiable on a neighborhood of the segment joining v_0 and v_1, and define*

$$p_\varepsilon(t) = \mathrm{sigmoid}(t/\varepsilon), \qquad U_\varepsilon(t) = (1 - p_\varepsilon(t))v_0 + p_\varepsilon(t)v_1,$$
$$M_\varepsilon(t) = F(U_\varepsilon(t)), \qquad H(p) = F((1-p)v_0 + pv_1).$$

For every $\varepsilon > 0$,

$$\varepsilon M'_\varepsilon(0) = H'(1/2)/4. \tag{50}$$

For fixed $h > 0$, define

$$D_\varepsilon(h) = \big(M_\varepsilon(\varepsilon h) - M_\varepsilon(-\varepsilon h)\big)/(2\varepsilon h), \qquad C_h = \big(H(\mathrm{sigmoid}(h)) - H(\mathrm{sigmoid}(-h))\big)/(2h).$$

Then $D_\varepsilon(h) = C_h/\varepsilon$. If $C_h \ne 0$ and $\tau = \varepsilon/2$, then $D_\tau(h)/D_\varepsilon(h) = 2$. If additionally F is three times continuously differentiable, then $C_h = H'(1/2)/4 + \mathcal{O}(h^2)$ as $h \to 0$.

The coefficient $1/4$ is the logistic derivative at equal responsibility when the action gap has unit slope in t. The symmetric branch-logit intervention of Equation (2) changes the logit gap at rate 2 with respect to δ, so these coefficients refer to different intervention coordinates. This proposition makes no claim about additional keys. An extension using the leakage bound of Proposition 3 requires explicit dominance of the selected pair: the hypothesis that unselected actions are merely bounded away from the selected ones does not exclude an unselected key whose action exceeds both selected actions, in which case the selected pair does not carry the mass and the two-key identity does not apply. This manuscript's Experiment 2 tests the general finite-key factorization (5), not this two-key temperature-scaling identity; the two are complementary, not duplicate, claims.

Convex construction of a dominant tie (summary). Sakamoto's construction produces, from queries and keys alone, a minimum-norm query displacement that makes a chosen pair of keys the dominant tie:

Proposition 5 (Minimum-norm construction of an individual-key tie, summary, cf. Sakamoto *Low-Temperature Attention Switching* §3.2 Proposition 3.2)**.** *Assume fixed keys and affine query scores with scale c. Choose two keys i_0, i_1 and a fixed preload that ties their scores at the reference query. Assume $n = c(k_{i_1} - k_{i_0}) \ne 0$. For each remaining key, define*

$$A_j = c(k_{i_0} - k_j)^\top, \qquad b_j = a_j^{\mathrm{pre}} - a_{i_0}^{\mathrm{pre}} + \eta_0, \qquad \Pi = I - nn^\top/\|n\|^2.$$

Consider minimizing $\|d\|^2/2$ *subject to* $n^\top d = 0$ *and* $Ad \geq b$.

$$\min_d \ \tfrac{1}{2}\|d\|^2 \quad \textit{s.t.} \quad \langle n, d\rangle = 0,\ Ad \geq b. \tag{51}$$

By the Farkas alternative, this problem is feasible exactly when there is no vector $\lambda \geq 0$ *satisfying* $\Pi A^\top \lambda = 0$ *and* $b^\top \lambda > 0$. *If feasible, the primal minimizer is unique. The dual objective is*

$$D(\lambda) = b^\top \lambda - \|\Pi A^\top \lambda\|^2/2, \qquad \lambda \geq 0.$$

For any maximizing dual multiplier $\lambda^\star$, *the primal solution is* $d^\star = \Pi A^\top \lambda^\star$. *The dual multiplier need not be unique.*

This construction ties two individual keys and enforces their dominance margin over the other keys. It does not in general produce an exact tie between the total masses of two semantic token groups. Consequently it is not identical to the branch-LSE operating-point construction used in the experiments. This constructive method has no counterpart elsewhere in this manuscript, which selects operating points by preloading the branch-logit coordinate δ directly (Equation (2)) rather than by a query-space displacement; we state it here because it gives an alternative, query-space route to the same class of operating points used in Experiments 2–3, and refer to the source note for the full proof.

L Extended related work (moved from the main text)

Log-partition gradients and covariances are standard exponential-family identities (Wainwright & Jordan, 2008). The multi-branch expansion applies Laplace's method (Wong, 2001); the sigmoid routing law and chain-rule factorization are not new mathematical identities. Transformer-circuit analyses distinguish routing through query–key interactions from information carried through value/output maps (Elhage et al., 2021). Causal head localization and function-vector transfer already demonstrate that small head sets can carry behaviorally relevant computations (Todd et al., 2024). AtP* already recomputes the attention softmax for query/key interventions before applying a downstream gradient (Kramár et al., 2024). Our finite-strength predictor uses the same computational principle for a controlled semantic branch-logit coordinate. The contribution here is its branch-mass decomposition, operating-gap tests, normal/tangent measurements, and the held-out test of an input-adaptive switching basis, not introducing nonlinear attention followed by a frozen downstream gradient. PASTA performs post-hoc attention steering by reweighting a small set of heads toward user-emphasized text (Zhang et al., 2024). Our contribution is not attention reweighting itself, but evidence-pair coordinates that separate relative allocation from union mass, expose when coefficients transfer, and permit finite-strength response prediction. SADI constructs input-adaptive dynamic steering vectors from contrastive activation differences (Wang et al., 2025); our basis is instead constructed from gradients of explicit evidence-allocation coordinates, so dynamic adaptation alone is not novel here. Tropical analyses characterize routing partitions and finite-temperature behavior away from boundaries (Su & Liu, 2026); they establish neither that routing boundaries are novel nor that wall proximity alone predicts control. Our tests address finite-width, boundary-near interventions and their downstream response. CAA and SAE-based knowledge steering intervene on activation directions (Rimsky et al., 2024; Zhao et al., 2025); we test a branch-logit response law, not superiority over these controllers.

Causal localization, patching, and circuit discovery. Causal mediation analysis attributes model behavior to individual neurons and attention heads (Vig et al., 2020), and interchange interventions test whether internal representations realize the variables of a higher-level causal model (Geiger et al., 2021). Causal tracing localizes factual recall to mid-layer computations and motivates rank-one and mass weight edits (Meng et al., 2022; 2023); complementary analyses describe how subject and relation information is routed to the prediction by attention and MLP sublayers (Geva et al., 2021; 2023). Circuit studies identify head-level mechanisms for indirect-object identification, greater-than comparison, and induction (Wang et al., 2023; Hanna et al., 2023; Olsson et al., 2022), and automated circuit discovery and attribution patching scale this search (Conmy et al., 2023; Syed et al., 2023). Activation-patching results are known to depend on the metric and corruption method (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Our measurements inherit this dependence: we report a teacher-forced candidate-score margin for heads fixed by earlier causal analysis, and we do not claim a new localization or circuit-discovery method. The branch-logit coordinate is instead a within-head parameterization of what those methods already localize.

Attention heads and representation geometry. Head-importance and pruning studies show that a few specialized heads carry much of the behaviorally relevant computation (Voita et al., 2019; Michel et al., 2019), consistent with our restriction to a small, pre-selected head set. Work on superposition and sparse dictionary features studies how many features share residual-stream directions (Elhage et al., 2022; Cunningham et al., 2024), and layer-wise decoding tracks how residual-stream predictions are refined across depth (Belrose et al., 2023). The linear-representation view identifies a non-Euclidean causal inner product in which probing and steering directions are related (Park et al., 2024); in the same spirit, our wall-normal versus tangent comparison depends on the chosen coordinates and is not a metric-free statement. Information geometry supplies the Fisher metric on exponential-family parameters (Amari, 2016), whose spectrum in deep networks is highly anisotropic (Karakida et al., 2019). The softmax branch mass is an exponential-family mean parameter, and the switching factor $2p(1-p)$ is proportional to the Bernoulli Fisher information $p(1-p)$ of the branch logit. We use this only as a local sensitivity identity, not as evidence that attention learns an information-geometric representation.

M Main-text material moved to the appendix (9-page compression, 2026-09-26)

This appendix collects, verbatim, the paragraphs, figures, and tables that were moved out of the nine-page main text. Cross-references from the main text resolve here; no results were altered.

M.1 Theory material moved from Sections 2–3

M.1.1 What does the local softmax law predict?

A multi-branch Laplace expansion of Gibbs attention motivates the local geometry we use below (Eq. (11), Appendix A.1): each branch's responsibility is a softmax in the branch actions; at a transverse two-branch tie it obeys a logistic law in the action gap, and the leading singular part of the output derivative is a rank-one outer product normal to the tie wall (Theorem 1, Appendix A). For more than two branches, the leading singular term is

$$D_x U_\epsilon = \mathrm{Cov}_\pi(V_a, \nabla\Phi_a)/\epsilon + \mathcal{O}(1),$$

with the branch weights defined in Equation (12), and its rank is at most $K-1$. A displacement v annihilates this term if it preserves every relative branch action to first order:

$$v^\top \nabla(\Phi_a - \Phi_1) = 0, \qquad a = 2, \dots, K.$$

Tangency to only one pairwise tie wall is not sufficient in general. For two branches this condition reduces to tangency to their action-gap wall. In either case, bounded within-branch and prefactor contributions may remain. We do not treat this asymptotic normal form as an empirical temperature-scaling claim for finite-key attention; it only motivates the shape of the switching transition, and the finite-key derivation and all measurements use the exact softmax identities that follow.

M.1.2 How does finite-key steering split into two mechanisms?

An explicit example in which the switching term vanishes but the exact response does not is in Remark 1 (Appendix A).

From a softmax identity to a network-level test. The algebra above is not, by itself, a new property of a transformer: it is exact softmax differentiation for fixed keys, values, and masks. Its network-level content is the falsifiable composition with the downstream margin: the two separately controllable group log-partitions change a selected head output, and the rest of the frozen network maps that output to the teacher-forced score in Equation (7). In addition to the one-parameter derivative agreement in Table 18, the two-coordinate construction proposes a test of margin changes when branch switching and total evidence mass are manipulated separately. Its archived response table is retracted because intervention verification failed; no validated network response is established. The matched nonsemantic-span control is necessary: a large response to an arbitrary set of keys would show attention leverage without evidence-specific routing. The present audit does not establish

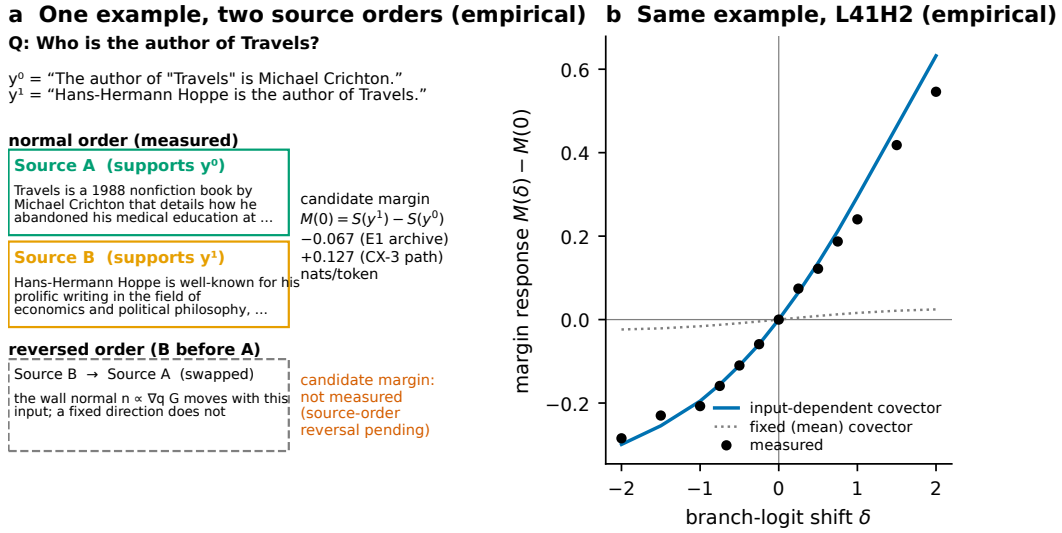


Figure 17: **(a) The problem: the same two sources in two orders, and a direction that must move with the input.** Left (empirical text): ConflictQA example `conflictqa_008295` (GPT-4 PopQA subset; source text verbatim from the frozen pool manifest, truncated for display) with its two candidate answers y^0, y^1; no generated answer is shown. The normal-order candidate margin $M(0)$ is the archived Gemma-4-E4B-it Experiment 1 baseline (-0.067 nats/token); the CX-3 natural-state path records $+0.127$ for the same example, a measurement-path discrepancy we report rather than reconcile. The reversed-order margin has not been measured in this repository and is marked as such. Right (empirical): for the same example and head L41H2, the measured margin response over the 13 archived branch-logit shifts (markers) against the example-specific downstream-covector prediction (solid) and the fixed mean-covector prediction (dotted) from the post-hoc ablation of Section M.2; the fixed direction misses this example's response. This contrasts downstream covectors, not query normals; the fixed query-axis comparison is Figure 19.

that specificity, so we do not claim a distinctive network-level advantage over generic attention reweighting from these measurements.

The key-average expression for the wall normal assumes scores that are affine in the query. Changing the attention scale changes its scalar prefactor. If a checkpoint applies a nonlinear score transformation, the wall normal must differentiate the implemented transformed scores; replacing the scale alone is insufficient. Thus a tangent $t \perp \nabla_q G$ has $D_t G = D_t p = 0$ to first order. The leading singular contribution vanishes along directions tangent to the tie wall; bounded within-branch and prefactor contributions may remain. Query perturbations can still change within-branch values and outside mass, so this does not assert a rank-one *full* finite-model Jacobian

M.2 Experiment details moved from Section 3

For the archived held-out controller cohort, symmetric intervention efficiency was the originally specified endpoint; the one-sided analysis is post-hoc and descriptive, and the Qwen Experiment 3 comparison convention was adopted after those results, so it is not confirmatory. Auxiliary smoke and numerical audits are not evidence here and support no native-checkpoint fidelity claim (Appendix E).

cross (precision 90.5%, recall 80.9%, crossing-strength MAE 0.409 logit-shift units; Appendix F.4). Figure 22 re-plots the same archived points on the branch-gap coordinate $g = G + 2\delta$ so that theory and measurement are compared across the wall $g = 0$ of Equation (3), following the presentation of Cui et al. (2024). For Gemma-4 the measured means cross the wall smoothly and follow the structural prediction (solid) within the example-clustered intervals; on this binned view the linear comparator (dashed) is visually indistinguishable from it, so the pooled R^2 gap reported below (0.853 vs. 0.672) is carried by per-point residuals rather than by the binned response. For Qwen3-8B the response is roughly five times smaller and the measured means sit systematically below both predictors for

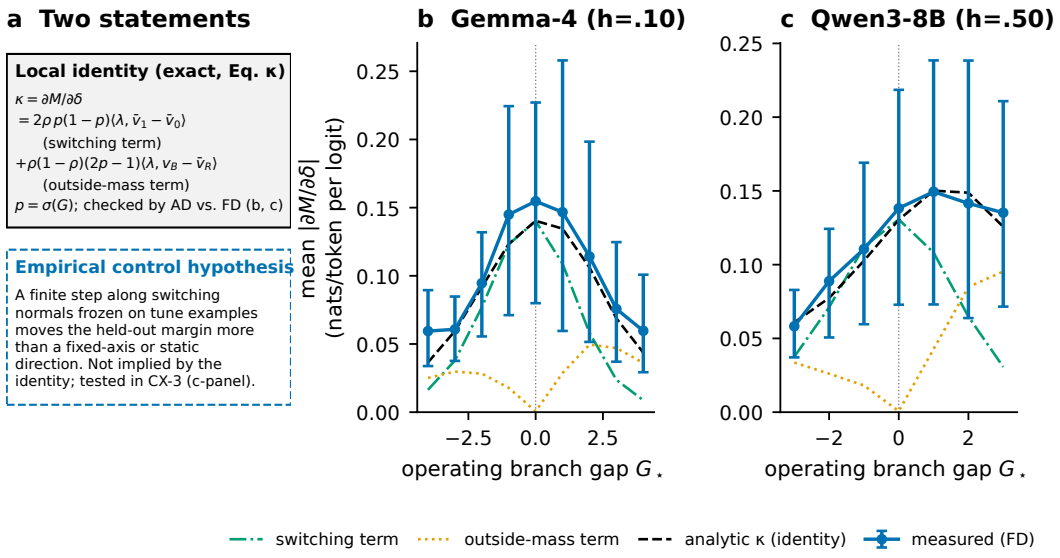


Figure 18: **(b) The lens: local response from G, ρ and value/readout alignment, with the identity kept apart from the control hypothesis.** (a) The exact local identity of Equation (5) (grey box) and the empirical control hypothesis it does *not* imply (dashed box). (b, c) Empirical: mean absolute local response $|\partial M/\partial\delta|$ of strong heads against operating branch gap $G_\star$ for Gemma-4-E4B-it (10 examples $\times$ 3 heads, $h = .10$) and Qwen3-8B (8 examples $\times$ 3 heads, $h = .50$), from the archived Experiment 2 operating-point tables. Blue: central finite difference with example-cluster percentile bootstrap 95% intervals (2,000 resamples); dashed black: analytic κ; green and orange: the switching and outside-mass terms of the identity. The outside-mass term vanishes only at the tie $G_\star = 0$; the measured maximum sits at $G_\star = 0$ for Gemma-4 but at $G_\star = +1$ for Qwen3-8B, so the tie alone does not locate the largest margin response.

$g < 0$, i.e. the replication is weaker and the prediction is biased there. The error maps (c,d) show where the two predictors separate on Gemma-4: near $\delta = 0$ they are equal (mean absolute error 0.015 vs. 0.016 nats/token for $|\delta| \leq 0.5$), while for $|\delta| \geq 1.5$ the linear error is 0.081 against 0.048 for the frozen-covector prediction. The wall itself is not resolved as an error boundary in either map.

or discovery among all heads. Raw-margin crossings are exploratory candidate-score events, not generated-answer success rates (Appendix F.4).

Post-hoc fixed-covector ablation. To ask whether the example-specific downstream covector matters beyond a fixed sensitivity direction, we recaptured baseline gradients for the same 50 examples and eight heads. For each head and candidate slot, the fixed comparator averages the within-example mean gradient across the 50 examples and broadcasts it over that example's scored-answer query positions. A leave-one-example-out version excludes the evaluated example from this average. Both comparators retain the original example-specific attention scores, values, masks, softmax reweighting, and recorded finite-response target. This is an ablation of the downstream covector, *not* of the input-dependent attention geometry. On 5,200 centered points, the recaptured example-specific predictor has pooled $R^2 = 0.853$; the fixed and leave-one-example-out fixed covectors obtain $R^2 = 0.121$ and 0.105, respectively. The paired example-cluster bootstrap interval for the full-mean R^2 difference is $[0.663, 0.793]$ (10,000 resamples), but this post-hoc cohort is not independent of the choice of comparison. A paired sign-flip test detects an overall gain; none of the question-category comparisons survives Holm correction. Moreover, recaptured baseline margins differ from the archive by up to 0.0949 nats/token, and recaptured example-specific predictions by up to 0.0695. The matching pooled R^2 does not erase this pointwise discrepancy. Consequently, these numbers show a paired comparison against archived response targets favoring example-specific downstream sensitivity, not validated correspondence to the archived joint-query gradients, not an isolation of the switching coordinates from generic softmax attribution, and not evidence of held-out task transfer. The runner and paired rows are included in the supplementary material. In particular, the

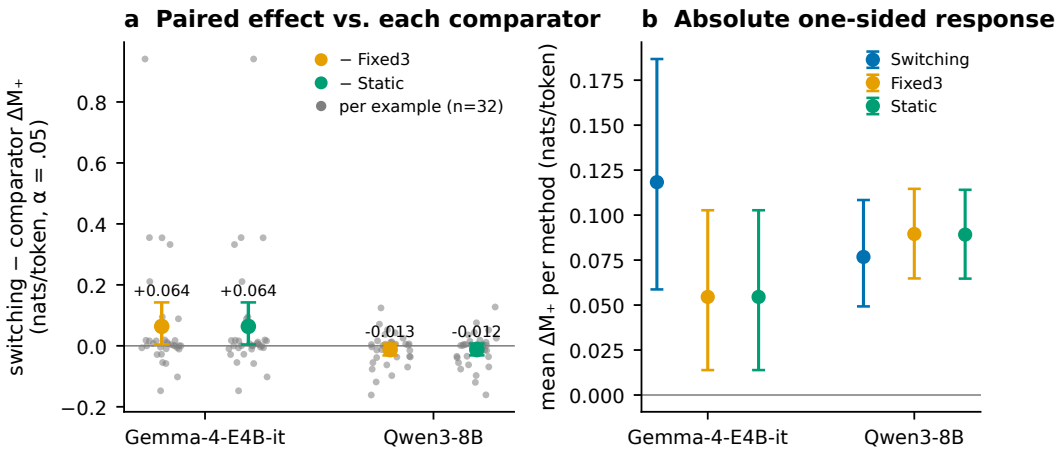


Figure 19: **(c) Main result: at $\alpha = .05$ the switching controller moves the held-out margin more than fixed and static directions on Gemma-4, but not on Qwen3-8B.** Empirical, exploratory CX-3 natural-state run on 32 reused archived held-out examples per model (controllers frozen on 16 tune examples; no test gradient or test margin selects a controller; Figure 10). (a) Paired per-example difference in one-sided response $\Delta M_+ = M(s_x u) - M(0)$, $s_x = \alpha \|q_x\|$, switching minus a genuine fixed three-axis basis (orange) and minus the static tune-gradient direction (green): mean with example-level percentile bootstrap 95% interval (2,000 resamples) from the committed summary; grey dots are the 32 paired examples. Gemma-4-E4B-it: $+0.064$ $[0.004, 0.142]$ against both comparators; Qwen3-8B: -0.013 $[-0.032, 0.007]$ and -0.012 $[-0.031, 0.007]$, an interval that includes zero with a point estimate favouring the comparators. (b) Mean ΔM_+ per method with the same intervals. For Gemma-4 the fixed3 and static arms coincide to numerical precision in mean ΔM_+ and its interval (per-example differences $< 5 \times 10^{-6}$); they are not identical in every statistic, since their mean retained fraction differs by 0.021 at every α. Measurement gates passed; the saved Gemma revision is unresolved and no scientific-pass review exists, so this is not a confirmatory result. Source-order check (not plotted): re-running with the two sources reversed (same cohorts and seed, controllers re-fitted on the reversed tune examples, 2,048 rows, 2,000 resamples) gives switching minus fixed3 at $\alpha = .05$ of $+0.061$ $[0.010, 0.126]$ (Gemma-4) and -0.003 $[-0.021, 0.016]$ (Qwen3-8B); at every $\alpha \in \{.025, .05, .075\}$ the Gemma interval excludes zero and the Qwen interval includes it in both orders, so the model-dependent contrast is not an artefact of source order. This does not supply per-example gate values or an executed-source audit.

R^2 gap cannot be credited to the proposed switching coordinates: an attention-corrected attribution method can use the same example-dependent softmax response in both arms. The coordinates instead specify an intervention contrast to test separately—change the branch split while holding evidence-versus-outside mass fixed, or change that mass while holding the branch split fixed. The fixed-covector ablation measures neither contrast.

The archived Experiments 1–5 above were measured on a custom attention path whose correspondence to the native checkpoint has not been audited. We therefore report separately the one cohort in this paper for which the whole measurement path was checked on the same examples before any intervention outcome was read. On the native `Qwen/Qwen3-8B` checkpoint (revision `b968826d`, eager attention, float32) with the three heads frozen from the CX-3 cohort (layer 24 heads 28 and 29, layer 34 head 7), ConflictQA facts were split disjointly into discovery 100, controller 200, validation 50, and test 200 (in-distribution only). The estimand is the joint post-RoPE additive query intervention at the final prompt position with the natural baseline and no tie preload, $F_x(\delta) = M_\delta(x) - M_0(x)$ in nats per token. Three gates were applied in order: (i) for every audited example the native forward and the hook at zero strength agreed exactly (maximum absolute log-probability and margin deviation 0 over 250 validation and test examples); (ii) on three validation facts the automatic-differentiation derivative $\partial F/\partial \alpha$ along the switching direction agreed with central finite differences at two widths to within 0.3% (for example 0.7869 versus 0.7866 and 0.7856); (iii) the baseline-only predictions of all arms for the 200 test facts were written to a file whose SHA-256 (`ca9b530e...`) was recorded before the first post-intervention test forward and re-verified after

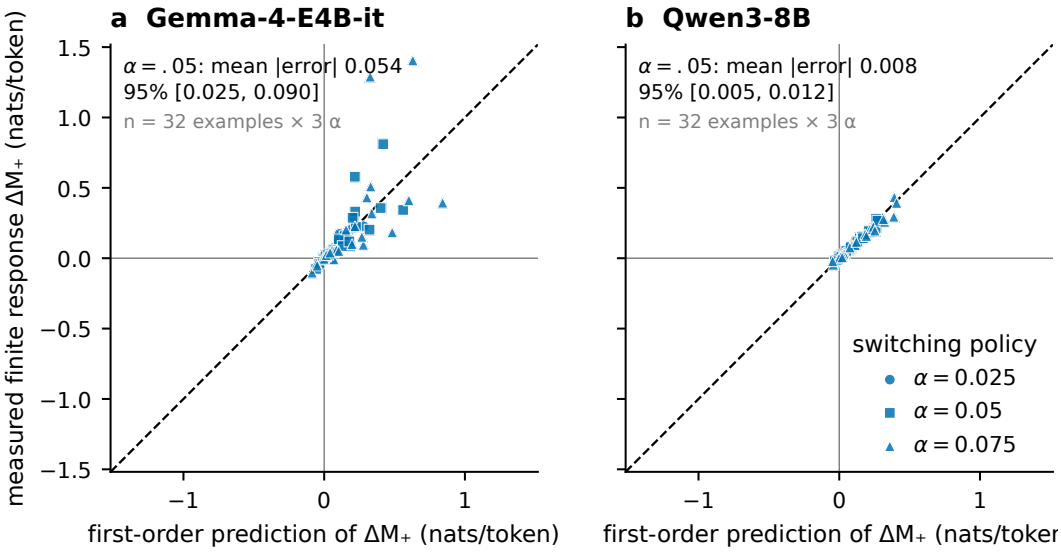


Figure 20: **Baseline directional prediction against finite response on the same examples under the same policy.** Empirical: CX-3 natural-state run, switching policy, one point per example and strength $\alpha \in \{.025, .05, .075\}$ (32 examples per model). Horizontal: first-order prediction of ΔM_+ from the baseline task gradient along the frozen direction; vertical: measured ΔM_+; dashed: identity. Text: mean absolute prediction error at $\alpha = .05$ with example-level percentile bootstrap 95% interval (2,000 resamples) from the committed summary. For Gemma-4-E4B-it the finite response of the largest movers exceeds the first-order prediction; for Qwen3-8B the points stay near the diagonal. Diagnostic only: the test gradients enter after the controllers were frozen and select no policy.

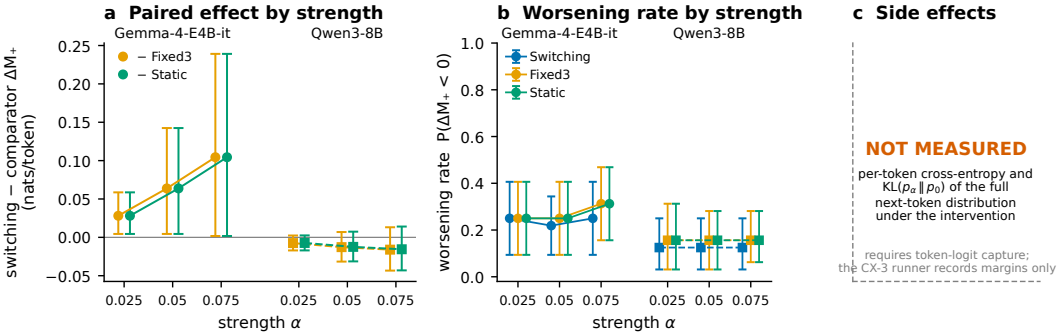


Figure 21: **Exploratory CX-3 strength summaries and unmeasured side effects.** 32 reused archived evaluation examples per model; controllers fitted on 16 tune examples. (a) Paired one-sided response difference ΔM_+, switching minus the fixed three-axis basis (orange) and minus the static tune-gradient direction (green), at $\alpha \in \{.025, .05, .075\}$. (b) Fraction of examples with negative ΔM_+. Intervals are descriptive example-level bootstrap 95% intervals (2,000 resamples); the multiplicity of models, comparators, and strengths is not resolved by whether individual intervals exclude zero. (c) Intentionally empty: per-token cross-entropy and next-token KL divergence were not measured. No `scientific_pass` verdict and an unresolved Gemma revision: these summaries do not establish generation utility or native-path fidelity.

the last. Only then were finite interventions measured at norm fractions $\alpha \in \{.005, .05, .20\}$. The exact-local predictor for the switching arm has median normalized absolute error 0.061 (0.069 for the first-order predictor) with sign accuracy 1.00 on the 263 test rows above the validation-fixed floor of 0.0086 nats/token. The paired measured contrast between the three-coefficient switching arm and a genuine fixed three-axis query basis fitted on discovery only (first principal component of the discovery switching normals per head) is $+0.072$ nats/token at $\alpha = .05$, with a paired fact-bootstrap

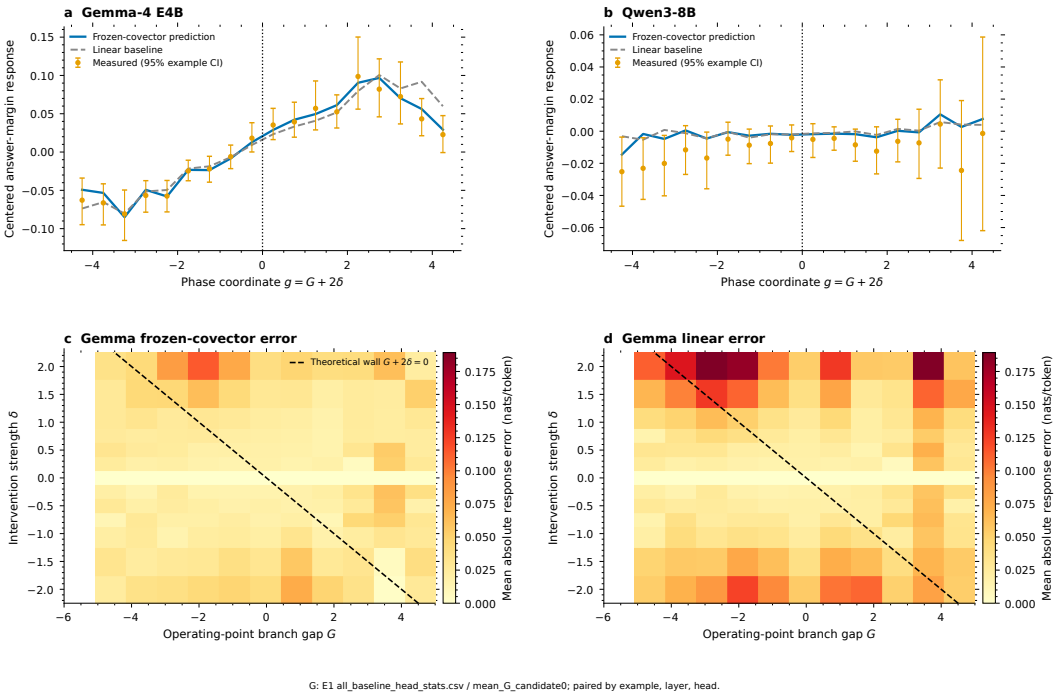


Figure 22: **Response curves and prediction-error maps across the branch-gap wall. Top:** 5,200 archived Experiment 1 points for (a) Gemma-4-E4B-it and (b) Qwen3-8B (50 examples, eight heads, 13 shifts each), binned on the phase coordinate $g = G + 2\delta$, where G is the operating-point branch gap of the same (example, layer, head) taken from the Experiment 1 baseline-head statistics (`mean_G_candidate0`) and δ is the shift strength. Solid: frozen-covector prediction; dashed: linear baseline; markers: measured means with example-clustered bootstrap 95% intervals (not joined). Dotted vertical line: wall $g = 0$. **Bottom:** Gemma-4-E4B-it mean absolute response error on the (G, δ) plane (12 bins in G, one row per δ, shared colour scale) for (c) the frozen-covector prediction and (d) the linear baseline. The dashed line is the theoretical wall $G + 2\delta = 0$. The two maps agree near $\delta = 0$ and separate at $|\delta| \geq 1.5$, where the linear error roughly doubles; the wall is a theoretical overlay and is not itself resolved as an error boundary at this sampling. The $\delta = 0$ row is identically zero by centering.

95% interval $[0.058, 0.087]$ (simultaneous $[0.057, 0.087]$; Holm-adjusted $p < 0.001$ against the pre-registered practical margin of 0.02; 2,000 resamples). These are the only numbers in this paper for which gradient-to-finite-response agreement, native-versus-hooked agreement, and prediction freezing were verified on the measured cohort. They remain exploratory: one model, one seed, one in-distribution environment, 200 test facts, and a single fixed-basis comparator, so "strongest fixed basis" is not validation-selected, and no supervised, gradient-based, or position-aligned fixed basis and no generation-utility or neutral-accuracy endpoint were measured. They do not validate the archived measurements of Experiments 1–5. The pre-registered full design (six checkpoints, three task families, ten seeds, twelve matched query-space arms, frozen predictions before measurement) is fixed in `prereg/cx_id_full_20260926.json`; its launcher refuses to run any subset until every capability, checkpoint revision, dataset, and path-gate receipt it requires is present. Run artifacts: `experiments/cx_id_decision/results/cxid_min_run1/`.

M.2.1 Experiment 2: A finite-difference consistency check across operating points

Question and protocol. Does the full local derivative in Equation (5) agree with measured responses at different branch gaps in both Gemma and Qwen, and is the mixing factor $2p(1-p)$ alone sufficient for the average susceptibility profile? Strong (top-three) and weak-control heads are selected per example from the fixed pools (Appendix C). Only the final prompt query is edited; a common

Table 18: **Experiment 2: operating-point derivative agreement.** R^2 compares the analytic derivative directly with finite differences. "Mixing only" omits the outside-mass term. MAE is in nats per token per unit logit shift.

Model	Heads	h	Points	Full R^2	Mixing-only R^2	Full MAE
Gemma	Strong	.10	270	.845	.733	.0402
Gemma	Weak	.10	90	.011	.000	.0500
Qwen	Strong	.50	168	.948	.748	.0290
Qwen	Weak	.50	56	-.015	-.015	.0481
Mistral (post-hoc, 64 fresh)	Strong	.50	1,344	0.941	0.870	0.0113
Mistral (post-hoc, 64 fresh)	Weak	.50	448	-0.000	-0.000	0.0083

preload $\delta_0 = (G_\star - G_0)/2$ targets an operating gap $G_\star$ ($\{-4, \dots, 4\}$ for Gemma, $\{-3, \dots, 3\}$ for Qwen). At each actual preload we recompute both candidate gradients to obtain the analytic derivative $\kappa_{\mathrm{an}}(G_\star)$ and compare it with the central difference $\kappa_{\mathrm{FD}}(G_\star; h) = [M(\delta_0 + h) - M(\delta_0 - h)]/(2h)$, with $h = 0.10$ for Gemma and $h = 0.50$ for Qwen, chosen after BF16 sensitivity analyses (Appendix D.5). This is a local derivative test with a new downstream gradient at each operating point, not a baseline-only prediction of a curve. Pooled R^2 and MAE are computed without fitted regression, separately for strong heads and weak controls; a mixing-only comparator drops the outside-mass term (Appendix E.5).

Results. The full analytic derivative agrees with measured responses in the strong heads (Table 18, Figure 9) at the pooled example–head level. Including outside mass improves pooled R^2 from 0.733 to 0.845 in Gemma and from 0.748 to 0.948 in Qwen. On the 64 fresh Mistral examples, the post-hoc replication gives $R^2 = 0.941$ for strong heads, while its weak control has $R^2 = -0.000$. The near-zero predictive R^2 neither establishes a null intervention effect nor characterises the shape of the operating-gap dependence: the weak control's absolute mean prediction error is 0.0083 nats/token, comparable in scale to the strong-head error of 0.0113. This audit is descriptive rather than preregistered confirmatory evidence, and we do not yet report a bootstrap confidence interval for this post-hoc MAE. Weak controls have nearly zero analytic sensitivity and poor finite-difference agreement. Median strong per-curve R^2 values are only 0.096 and 0.277, respectively; thus the pooled values primarily capture large between-curve sensitivity differences rather than uniformly explaining each head's operating-gap dependence. The Gemma mean absolute susceptibility profile correlates with $2p(1-p)$ at $r = 0.979$; the corresponding Qwen correlation is 0.483. An adjacent-grid audit agrees with Qwen's $h = 0.5$ secants to MAE 7.14×10^{-8}.

Interpretation. The two-model comparison supports the full local factorization (5), not a universal mixing-only curve: outside mass and downstream alignment change with the operating point. The pooled-versus-per-curve gap limits claims about individual profiles, and the adjacent-grid audit checks evaluation consistency, not convergence as $h \to 0$.

M.2.2 Exploratory two-coordinate intervention audit

Protocol and estimand. For Qwen3-8B we tested the frozen eight ConflictQA examples and four heads (three previously designated strong heads and one weak control) with a second intervention that controls the two group log-odds independently. Writing $H = \log(Z_0 + Z_1) - \log Z_R$, $p = \mathrm{sigmoid}(G)$ and $p' = \mathrm{sigmoid}(G + u)$, the per-key logit shifts are $\beta_0 = v + \log[(1-p')/(1-p)]$, $\beta_1 = v + \log(p'/p)$, and $\beta_R = 0$. Algebraically these give $\Delta G = u$ and $\Delta H = v$, so the u axis varies branch switching at fixed evidence-versus-outside mass while the v axis varies outside-mass allocation at fixed branch split. We measured the change in teacher-forced candidate margin $M(u, v) - M(0, 0)$ for each axis ($u, v \in \{-2, -1.5, -1, -0.5, -0.25, 0, 0.25, 0.5, 1, 1.5, 2\}$), plus a 5×5 joint grid and length/position/approximate-attention-mass-matched nonsemantic spans. We retract the archived response table because its runtime coordinate capture preceded the intervention, and the corrected attempts did not produce a verifier-passing matched-control dataset. The recorded response magnitudes therefore are not used as evidence for a network-level mechanism.

Interpretation and validity gate. The original capture point preceded the intervention; its failure is in the verifier, not an affirmative validation of the intervention path. A rerun with post-bias score capture and all coordinate checks passing is required before promoting a network-level result. A later archived E1 attempt (23:02) reports `verify_all_pass=false` and 216 failed checks, all in the nonsemantic control. A subsequent run with the disjoint-mask fix at commit `a294255` failed before producing a new result: `conflictqa_004712` L24H28 b1 required 141 keys but had only 3 eligible disjoint keys. No valid post-fix verifier-passing dataset exists. In a subsequent real-GPU four-example smoke with the corrected code, all four frozen examples were excluded because no equal-length disjoint nonsemantic control could be formed (required/eligible key counts: 141/3, 149/97, 150/100, and 176/99). The run produced zero valid rows and `verify_all_pass=false`; its zero recorded verifier failures reflect the absence of evaluable rows, not a passing intervention. Thus the archived attempts do not establish semantic specificity or a network-level switching mechanism. A new control definition or cohort would require an explicit protocol decision before further confirmatory testing. Under the pre-registered retry3 deviation, the falsification comparator is the longest available disjoint nonsemantic control with at least half of each evidence-branch length, not an equal-length span; any positive contrast remains length-confounded and cannot alone establish semantic specificity. Thus the two-coordinate construction currently *enables a falsifiable measurement design*, not an observed advantage over generic attention reweighting: a decisive network-level test would require post-intervention verification of both coordinates and a feasible matched nonsemantic control, then compare their teacher-forced margin responses on the same examples and heads. No such verifier-passing comparison is available here.

M.2.3 Experiment 3: normal versus tangent query perturbations

Question and protocol. Do nominally equal-norm query perturbations along the branch normal produce larger head-output and answer-margin responses than sampled tangent perturbations? We reuse the four Qwen heads and eight examples of Experiment 2 and edit only the final prompt query (after normalization and RoPE, keys and values fixed). The normal is $n = \nabla_q G / \|\nabla_q G\|$ (Equation (6)); three seeded orthonormal random tangents $t_r \perp n$ are the controls, at branch preloads $G_\star \in \{-1, 0, 1\}$. Every direction uses the same nominal step $\eta = \min\{0.5/\|\nabla_q G\|, 0.2\|q\|\}$ (norm cap inactive), so the predicted normal gap response is 0.5 and the first-order tangent response is zero; 384 direction conditions are measured. We compare $|\mathcal{D}_d F|$ for $F = G, p, M$ and the head-output norm, using the median of three matched tangents and the median normal/tangent ratio over the 24 strong tie conditions (Appendix E.6).

Results. At the tie, median normal/tangent ratios are 121.1 for gap, 118.7 for responsibility, 3.92 for head output, and 2.97 for margin (Figure 23). The mean paired margin gain is 0.0308 nats per token (95% interval $[0.0089, 0.0568]$); the output gain is 1.577 ($[1.207, 2.001]$). Normal-gap MAE is 0.00603 (1.21% of the nominal shift). The three-gap responsibility profile is consistent with the finite-step prediction. Weak-control output changes are of order 10^{-5}, with unstable margin responses (Appendix G).

Interpretation. The output and margin comparisons support propagation of wall-normal sensitivity relative to the sampled tangents. They do not identify the normal as the optimal downstream direction or a rank-one Jacobian, and large weak-control ratios are not evidence of reliable causal amplification.

M.3 Control-experiment details moved from Section 4

Let $F_x(z)$ denote the full teacher-forced candidate-1-minus-candidate-0 mean log-probability margin under the additive joint-query intervention specified in Appendix E. The vector $z \in \mathcal{U}$ contains the three injected post-RoPE query increments. Each increment is added to the live query produced by the current forward pass; masks, preloads, and probe directions remain fixed from zero-increment calibration. The intended task gradient is $g_x = \nabla_z F_x(0)$, and the retained fraction

$$R_x(S) = \frac{\|P_S g_x\|^2}{\|g_x\|^2} \tag{52}$$

is a property of this map on its nondegenerate domain; it measures the squared local task sensitivity available inside S. Correspondence between this definition and the archived gradient routine was

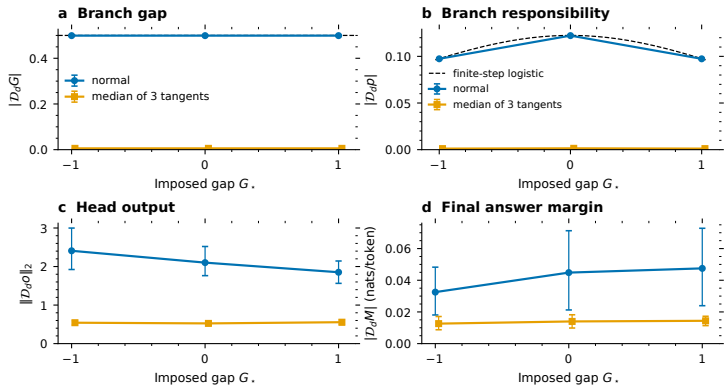


Figure 23: **Experiment 3: normal versus tangent perturbations in Qwen.** Mean finite odd responses across three strong heads and eight examples, with 95% example-cluster intervals. Tangents use the median of three nominally equal-norm controls. Dashed gap/responsibility predictions use step 0.5 without amplitude fitting; output and margin test downstream propagation.

not established (Appendix D) and is excluded from the frozen result set. We compare S_{sw} with 64 block-matched random subspaces per example, each containing one isotropic direction in each of the same three heads, and with a tangent comparator. The tangent comparator is also three-dimensional: it contains one seeded unit direction in each selected head block, orthogonal to that block's switching normal. It is not the full orthogonal complement of the switching subspace. The full task gradient provides an oracle direction for maximizing the local first-order response per unit displacement, not a guaranteed ceiling on finite-strength response. Within any tested subspace, the locally best unit direction is $u_S = P_S g_x / \|P_S g_x\|$. Every normalized direction is defined only when its denominator is nonzero. A zero task gradient makes the retained fraction undefined. A zero projected gradient means that the subspace has no first-order task response and does not determine a unique normalized oracle direction. A zero mean tune coefficient vector likewise does not define the normalized-mean controller. The archived handling of these degenerate cases has not been documented here; the displayed formulas apply only on their nondegenerate domains.

The eight examples and three heads are frozen from Experiment 2. We jointly preload the three heads to $G_h = 0$ (maximum recorded error 1.91×10^{-6}), then apply matched joint-query perturbations at norm fraction $\alpha = .05$ (the primary strength); the finite causal efficiency $\mathcal{E}_x(u;\alpha)$ is defined in Appendix I, Equation (45).

tangent subspace. The paired switching-minus-random difference is 0.260 ($[0.141, 0.378]$), and the median per-example switching/random ratio is 47.8. None of the 64 random draws reaches the switching value in any of the eight examples.

At matched finite control norm, mean efficiency at $\alpha = .05$ is 0.1597 ($[0.0570, 0.2761]$) for the switching projection, 0.0173 ($[-0.0037, 0.0411]$) for the median random projection, 0.0280 ($[0.0050, 0.0589]$) for the tangent projection, and 0.2566 ($[0.1235, 0.4169]$) for the full-gradient oracle. The switching advantages over random and tangent are respectively 0.1424 ($[0.0444, 0.2445]$) and 0.1318 ($[0.0373, 0.2379]$). Thus the local normals identify a strongly task-enriched search subspace, but this diagnostic uses g_x on every evaluated example to choose a direction inside that subspace; it is not yet a held-out controller.

We next remove the test-example task gradient and the artificial tie. For each of the 16 tune examples, let $F_e(z)$ be the additive joint-query map used for finite interventions, without a tie preload. Let Q_e contain the three block-supported unit switching normals obtained from zero-increment calibration. The intended fitting rule is

$$g_e = \nabla_z F_e(0), \qquad b_e = Q_e^\top g_e, \qquad \bar{b} = \frac{1}{16}\sum_e b_e, \qquad \widehat{a}_{\mathrm{sw}} = \frac{\bar{b}}{\|\bar{b}\|}. \tag{53}$$

Table 19: Held-out one-sided margin response $\Delta M_+ = F_x(s_x u) - F_x(0)$ at norm fraction $\alpha = .05$, with $s_x = \alpha \|q_x\|$, on 32 coefficient-held-out examples. Intervals are 4,000 example-bootstrap percentile intervals; the worsening tail counts $\Delta M_+ < 0$. The symmetric odd component is retained separately as a mechanism diagnostic. This one-sided endpoint is a post-hoc descriptive reanalysis; the symmetric endpoint was originally specified for this cohort.

Controller	Mean ΔM_+	95% interval	$\Delta M_+ > 0$	Worsened
Switching, learned coefficients	.0831	$[.0523, .1136]$	24/32	8/32
Static mean tune gradient	.0821	$[.0551, .1088]$	26/32	6/32
Switching, equal positive	.0164	$[-.0072, .0404]$	18/32	14/32
Tangent, learned coefficients	-.0064	$[-.0231, .0104]$	16/32	16/32
Adaptive random, learned coefficients	-.0118	$[-.0270, .0041]$	12/32	20/32
Shuffled switching geometry	-.0211	not reported	17/32	15/32

The normalization requires $\|\bar{b}\| > 0$. Derivatives are with respect to injected increments, not independently overwritten downstream query states. Correspondence with the archived coefficient-generation routine remains unverified. The archived coefficients are $(-0.723, 0.354, 0.593)$ in the head order of Equation (8) and are retained without refitting. For each of 32 disjoint test examples, we compute only its three switching normals and apply the intended direction

No test gradient, coefficient fitting, strength selection, gap preload, or head selection is used. Evaluation remains the teacher-forced margin in Equation (7).

Controls use the same tune-only fitting rule with seeded tangent or adaptive-random bases, cyclically shuffled switching normals from another test example, equal positive normal coefficients, and the normalized 384-dimensional mean tune task gradient. The last is a strong unconstrained baseline rather than a geometry control. The originally specified endpoint for this archived cohort was symmetric finite intervention efficiency at norm fraction $\alpha = .05$ (Equation (45)). The one-sided margin changes below are a post-hoc descriptive analysis, not a prospectively specified endpoint.

The normalized mean coefficient rule maximizes the average local directional derivative per unit query displacement, not necessarily the first-order mean one-sided change when step sizes vary across examples; the archived controller was not refitted, and we retain its original coefficients (Appendix I).

The shuffled controller (cyclic donor normals from another test example) is reported descriptively without an independent-example interval, since resampling does not preserve the donor dependence; these summaries do not establish population-level superiority over it. The paired switching-minus-static difference is given in the text below and in Appendix I.

Results and interpretation. At norm fraction .05, the frozen switching controller has mean one-sided response .0831 (95% interval $[.0523, .1136]$) and improves 24/32 test examples; 8/32 are worsened. The static mean tune gradient has nearly the same mean response (.0821, $[.0551, .1088]$) and improves 26/32. The paired per-example difference (Table 16) is 0.0010 (95% interval $[-0.0235, 0.0260]$), an interval that spans zero, so the paired analysis does not establish superiority over the static baseline either. The symmetric odd component is positive on 27/32 switching examples, but it is not a control-effect endpoint because it can be positive when ΔM_+ is negative. Comparisons to geometry-disrupting controls remain descriptive under this revised endpoint.

The test examples are not artificially placed at a wall: their median $|G_h|$ is 0.603 and mean branch-union mass is 0.410. The result therefore supports an input-adaptive, three-coefficient controller at natural operating points. It does not establish control without known evidence masks, transfer beyond this model and task, or free-generation reliability. Moreover, the heads and branches were fixed by earlier causal analysis; the geometry reduces the within-head control search, not the full circuit-discovery problem. Additional split, coefficient, and baseline details are in Appendix H.

M.4 Extended related-work discussion and measurement limits moved from Section 5

followed by a frozen downstream gradient is not itself our contribution. Attention reweighting, dynamic steering, and activation-direction steering (Zhang et al., 2024; Wang et al., 2025; Su

& Liu, 2026; Rimsky et al., 2024; Zhao et al., 2025) are not novel here; we test a branch-logit response law, not superiority over these controllers (Appendix L). Contrastive activation addition steers behavior through activation directions (Rimsky et al., 2024), while SAE-based representation engineering explicitly targets knowledge-selection behavior (Zhao et al., 2025). Post-hoc attention steering instead reweights where a model attends (Zhang et al., 2024). Our query intervention is a narrower mechanistic test: can movement across a competing-evidence wall change the teacher-forced candidate-score preference in a context–parametric knowledge conflict? This does not directly identify an internal parametric-memory mechanism. It is neither a weight edit nor an established improvement over those methods.

the answer margin (Equation (5)). Query steering is the proposed operation of moving across this wall to change the measured answer preference; Experiments 1–5 measure prediction, local consistency, direction sensitivity, subspace enrichment, and held-out control, respectively, without establishing a native-checkpoint control law. Whether control is actually stronger near the wall than away from it remains an empirical question not answered by these results.

Limits of the evidence. The finite-strength predictor is algebraically identical to attention-corrected attribution when both use the same exact attention recomputation and frozen downstream covector; Experiment 1 compares it only with a weaker local linearization. The conditional Taylor bound is not a measured uniform error certificate because its downstream Hessian constants were not estimated. The archived pooled R^2 gain therefore does not establish added predictive value beyond attention-corrected attribution. Preloading a head to a tie establishes controlled behavior near a wall; it does not show that influential heads naturally lie near $G = 0$. Head selection uses causal gradients, so it is not geometry-only. Equal-norm comparisons establish anisotropy relative to three sampled tangents, not that the wall normal maximizes the full downstream response over all query directions or that the entire Jacobian has rank one. The control-subspace diagnostic uses the evaluated example's task gradient to select the best direction within each subspace; only the subsequent three-coefficient experiment removes test gradients. That controller still assumes known evidence masks and previously selected heads. Its observed one-sided mean response exceeds that of the geometry-disrupting controls; the advantage over the unconstrained static mean tune-gradient baseline is not supported (paired interval spans zero), and no equivalence or non-inferiority claim is made. The post-hoc fixed-covector comparison in Section M.2 reports better archived-target prediction for an example-specific covector in the recaptured cohort. The arms differ in magnitude and query-position structure as well as direction; because both share the same input-dependent softmax response, this is not an ablation of switching geometry against attention-corrected attribution. The pointwise drift from archived margins and predictions further prevents a native-checkpoint fidelity claim; the original joint-query-gradient-to-finite-response correspondence remains unverified for Experiments 1–5. The separate native-checkpoint run of Section 3.1 is the only cohort in which that correspondence, native-versus-hooked agreement, and prediction freezing were verified before outcomes were read; its positive switching-minus-fixed-basis contrast is a single-model, single-seed, single-comparator result without a utility endpoint and is not the pre-registered full design. The additional role of the proposed coordinates is operational rather than an extra softmax identity: they define two independently manipulable axes, branch split and evidence-versus-outside mass, on which to ask whether the network's answer-margin response is specific to competing evidence rather than to arbitrary attention redistribution. The archived fixed-covector comparison does not answer that question because it changes only the downstream covector, and the two-coordinate audit cannot answer it because its runtime verification and matched nonsemantic controls did not pass. We therefore state this as a testable use of the coordinates, not as an established network-level contribution. The study uses a single benchmark subset, two checkpoints, and small mechanistic cohorts; the Qwen geometry cohort is reused from the operating-point experiment, and finite-difference widths were adjusted after numerical diagnostics. BF16 error is especially consequential for weak responses (Appendix D.5). Teacher-forced margin changes do not by themselves establish free-generation reliability, and we make no claim about fluency or factual accuracy under open-ended decoding. Experiments 1–3 are evaluated at explicitly constructed or selected operating points near a wall, not at arbitrary natural operating points.

Recovered hook diagnostics. Appendix Figures 7 and 8 expose the recovered audit results rather than treating a completed run as a passed measurement gate. The E1 Qwen archive fails batch consistency and most directional derivative checks. The E2 expanded cohort is a descriptive comparison

among the implemented token-span logit interventions only. Its fixed condition uses a constant evidence-allocation shift rather than a fixed query basis or static task-gradient controller; its outcome and intervention budget do not match the controller comparison in this paper. Thus neither this plot nor its paired intervals test the additional value of switching geometry. A separate E0 measurement-audit retry reports `overall_pass=false` for both Qwen and Gemma; it is not positive support for the primary measurements.

N FIGURE PROVENANCE INVENTORY

Tables 20 and 22 identify the archived input and in-repository generator for each manuscript figure. A `data/` entry means no corresponding raw `results/` directory exists; a diagnostic inventory is not a substitute for model rows. This is a traceability inventory, not a fresh verification of model fidelity.

O POST-ARCHIVE RESULT INTEGRATION AND AUDIT LIMITS

Table 21 records seven result families absent from the original manuscript mapping in `INTEGRATION_GAP_20260924.md`. These entries are provenance and limitations, not additional positive evidence for native-checkpoint fidelity or a general two-coordinate mechanism.

Retracted verifier status. The two-coordinate eight-row table in Section 3 is *diagnostic, failed verifier*; it must not be read as a confirmed quantitative intervention effect. The recorded `results/e1_two_coordinate_20260923/summary.json` has `verify_all_pass=false`, and the corrected smoke yielded no valid rows.

O.1 MATCHED-NORM ORACLE COMPARISON ON THE CX-3 COHORT

On the 32 frozen-test CX-3 examples (same checkpoint, revision, and three heads; hooked path), switching, fixed three-axis, and static query interventions were applied at identical norms (maximum mismatch $< 5 \times 10^{-16}$), alongside full-gradient and complement oracles that use the evaluation gradient and are upper-bound references, not controllers. At $\alpha = .05$ the mean measured margin shift was .077 (switching), .090 (fixed3), .089 (static), .142 (complement oracle), and .171 (full oracle); the mean retained gradient fraction was 0.23 for the switching subspace and 0.77 for its complement. Paired against static, switching was lower in 15 of 32 examples (mean difference $-.012$). On this cohort, therefore, switching does not exceed the fixed or static directions at matched norm, and most of the gradient lies outside the switching subspace; this cohort and the CX-ID contrast above are not reconciled here. The executed runner and entry script match the archived sources by SHA-256 (`run_cx3.py` `a48cc7f3...`, `e2_oracle_audit.py` `655c5401...`). Native-versus-hooked agreement was not measured in this run, so it does not establish a native-path controller comparison.

Rows and summary: `results/cx3_e2_oracle_frozen_test_20260926/` ($n = 32,480$ rows); the tune split ($n = 16$, in-sample for the calibration) is in `results/cx3_e2_oracle_tune_20260926/`. Paired against static, switching was lower in 17/32, 15/32, and 15/32 examples at $\alpha = .025, .05, .075$.

Table 20: Figure to archived input to generator (part 1). Paths are relative to the repository root.

Figure PDF	Input directory	Generator
fig1_v2_final6h.pdf	results/caustics_cx_id_min_20260926/ (panels c,d); panels a,b: no data	scripts/fig1_v2_final6h.py
fig_rootfix_one_cohort.pdf	results/cxid_e2static_exploratory_qwen3_8b_20260926/ (32 facts; hashes in results/ROOT_FIX_manifest.json)	figures/make_rootfix_one_cohort.py
fig2_cui_transition.pdf	data/finite_strength/; no raw results/ dir	analysis/reproduce.py (also analysis/fig2_cui_transition.py)
fig_head_trajectory_3d.pdf	data/ archived CSVs; no raw results/ dir	analysis/fig_head_trajectory_3d.py
finite_strength.pdf	data/finite_strength/; no raw results/ dir	analysis/reproduce.py
finite_strength_details.pdf	data/finite_strength/; no raw results/ dir	analysis/reproduce.py
operating_points.pdf	data/gemma_factorization/, data/qwen_factorization/; no raw results/ dir	analysis/reproduce.py
geometry.pdf	data/geometry/; no raw results/ dir	analysis/reproduce.py
geometry_schematic.pdf	data/geometry/; no raw results/ dir	analysis/reproduce.py
geometry_controls.pdf	data/geometry/; no raw results/ dir	analysis/reproduce.py
discovery.pdf	data/; no raw results/ dir	analysis/reproduce.py

Table 21: Integration of seven previously unmapped result families. A completed run is distinguished from a passed scientific or verifier gate.

Family	Local provenance	Permitted interpretation
Fresh W3 geometry	results/w3_fresh_20260918/; RUNS.md	Fresh cohort is separate from the archived eight-example geometry result; no pooled estimate is reported here.
Fig. 5 $q(a, b)$	surface_all.csv in the Fig. 5 rerun receipt (full path in INTEGRATION_GAP_20260924.md)	Collected surface is selected-head and single-tangent only; the superseded original surface is not used as general mechanism evidence.
W6 grids / controls	results/w6_qwen_20260918T061141/; results/w6_gemma_20260918T061221/	Grid rows bypassed the negative-control gate; the Gemma fp32 audit was recorded as failing and its cited raw directory is unavailable locally. No native-fidelity claim follows.
E1 Qwen hook audit	results/e1_qwen_20260919T051331/summary.json	Structural mismatch and directional AD–FD failure are audit failures, not a successful equivalence measurement. The Gemma companion raw directory is unavailable locally.
E2 Qwen expanded $n = 32$	results/e2_fixed3/e2_expanded_qwen3_8b_n32_20260919/	Within-hook comparisons only; neither native equivalence nor fixed3 superiority is established.
Fixed-query-basis extension	results/e1_fqb_conflictqa200_20260923/summary_extension.json	The 150 new examples form an extension cohort. Archive drift and cohort separation require audit before combination with the original 50.
E0 measurement retry1	results/e0_measurement_audit_20260924_retry1/	Qwen and Gemma overall audit gates failed; neither is presented as supporting fidelity.

Table 22: Figure to archived input to generator (continued).

Figure PDF	Input directory	Generator
control_validation.pdf	data/control_subspace/, data/heldout_controller/; no raw results/ dir	analysis/reproduce.py
control_details.pdf	data/heldout_controller/; no raw results/ dir	analysis/reproduce.py
factor_components.pdf	data/; no raw results/ dir	code/make_fig2.py (also analysis/reproduce.py)
e1_hook_gradient_audit.pdf	results/e1_qwen_20260919T051331/	code/make_audit_figures.py
e2_expanded_within_hook.pdf	results/e2_fixed3/e2_expanded_qwen3_8b_n32_20260919/	code/make_audit_figures.py
w6_grid.pdf	results/w6_qwen_20260918T061141/, results/w6_gemma_20260918T061221/	code/make_w6_figure.py
wall_proximity.pdf	results/wall_proximity_20260924/; data/finite_strength/, data/heldout_controller/	code/make_fig_wall_proximity.py
fig1_right_finite_strength.pdf	data/finite_strength/; no raw results/ dir	code/make_fig1r.py
fig3_geometry_to_behavior.pdf	data/geometry/; no raw results/ dir	code/make_fig3.py
fig4_paired_heldout.pdf	data/heldout_controller/; no raw results/ dir	code/make_fig4.py