WHEN ACTIVATION INTERVENTIONS FAIL TO IDENTIFY COMPONENT RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability often treats a successful activation intervention as evidence that a component's causal role has been identified. We show this inference can fail for the reported ranking itself: for a fixed finite intervention family, a second decomposition can match every observed outcome exactly yet reverse the ordering of two components on a held-out contrast, while leaving every residual-stream activation and output unchanged. The ambiguity is governed by coverage, the portion of the reported contrast outside the span of the probed directions, measurable on the original model before any alternative is built. On GPT-2 Small, a construction reverses the IOI name-mover ordering that single-position interchange measures on all 24 held-out pairs while agreeing with a 512-prompt probe family, and generalizes to unseen templates. On Gemma-2 2B, an edited same-architecture JumpReLU SAE matches reconstruction and every probe outcome where the intervened feature is inactive, while reversing the feature-pair ordering on most active prompts. Coverage separates evidence styles: single-position families can leave the ordering steerable until probe count nears the model width (768 on IOI here), while position-resolved families and the per-context trace of a discovery sweep drive the uncovered portion to or near zero at this width. Successful interventions alone do not establish identification; whether the evidence covers the claim does.

1 INTRODUCTION

Mechanistic interpretability infers causal structure by intervening on internal activations (Vig et al., 2020; Wang et al., 2023; Geiger et al., 2024), within the circuits program (Olah et al., 2020; Elhage et al., 2021). Yet different operators can rank the same components differently: critiques document circuit-level disagreement (Makelov et al., 2024; McGrath et al., 2023; Zhang & Nanda, 2024; Hanna et al., 2024; Sharkey et al., 2025), and MIB (Mueller et al., 2025) reports sparse autoencoder features (Huben et al., 2024; Bricken et al., 2023) losing to alignment search (Geiger et al., 2024). We ask: *even if every outcome of a chosen intervention family were known exactly, would those observations uniquely identify the component ranking being reported?* Not necessarily; enlarging the family helps only insofar as it covers the reported contrast (Fig. 1).

Contributions.

1. **Finite activation interventions need not identify a component ranking** (Thm. 1). We construct two behaviorally identical decompositions agreeing on every observed outcome yet reversing a held-out ranking, and characterize which interventions distinguish them (Prop. 1).
2. **The ambiguity occurs at realistic scale and within a same-architecture SAE** (§4). Per-contrast, fixed, and composed witnesses reverse IOI name-mover orderings out of sample on GPT-2 Small, and the edited same-width JumpReLU SAE extends the ambiguity to dictionary granularity.
3. **Claim-specific coverage, not probe count alone, determines exposure** (Fig. 2). Position-resolved patching drives it to zero at $d = 768$; the rank law replicates at $d = 2048$.
4. **Claim-specific coverage is weaker than global spanning, and directly measurable** (§3). The witness dies once the reported contrast enters the probe span, even when a large annihilator remains (Cor. 1), so we report the uncovered mass and the exposure ratio R.

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.

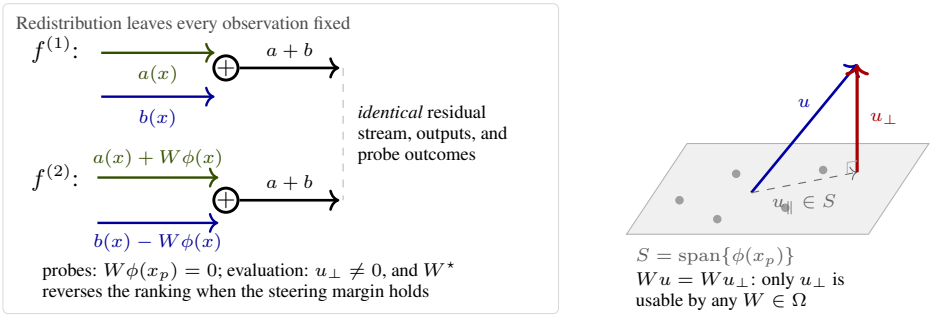

Figure 1: **The witness of Theorem 1.** *Left:* redistributing $\pm W\phi(x)$ between two parallel writes leaves the stream, outputs and probe outcomes unchanged. *Right:* the probe features span S; an evaluation contrast decomposes as $u = u_\parallel + u_\perp$, and $Wu = Wu_\perp$ is all the witness can move, so the ranking reverses when that reach exceeds the baseline gap (H3, §3) and dies when $u_\perp = 0$.

2 Setup

Let $f = f_{\geq \ell} \circ f_{<\ell}$ be a trained network of residual width d, let $h(x) \in \mathbb{R}^d$ be the activation on input $x \sim \mathcal{D}$, and $g := f_{\geq \ell}$ the downstream readout, composed with a measurement m (a logit difference, say) into $\Psi := m \circ g$. A *site* is a layer-token pair (ℓ, t); everything below acts at one site: displacements are vectors, not sequences. At a *parallel-write* site the stream decomposes as $h = c + a + b$, with c the incoming stream and a, b the two component writes, neither depending on the other.

Notation. A feature map $\phi : \mathcal{X} \to \mathbb{R}^q$, measurable with respect to the stream entering the site, is evaluated on P probe inputs $\{x_p\}$. Write

$$\Phi := [\,\phi(x_1)\ \cdots\ \phi(x_P)\,], \qquad S := \operatorname{col}\Phi, \qquad \Omega := \{W \in \mathbb{R}^{d\times q} : W\Phi = 0\},$$

the last being the *annihilator*; each $W \in \Omega$ has all its rows in $S^\perp$, so we report its size as $\dim S^\perp$ throughout ($\dim\Omega = d\,(q - \operatorname{rank}\Phi)$). For an evaluation pair (x', x) outside the probe set, $u := \phi(x') - \phi(x)$ with transversal part $u_\perp := \mathsf{P}_{S^\perp} u$, the operative quantity below. The *redistributed* network f_W replaces the two writes by $a_W := a + W\phi$ and $b_W := b - W\phi$, leaving $h = c + a_W + b_W$ unchanged pointwise for every W. Interchanging write $w \in \{a, b\}$ from x' into x gives the state $s_w(x' \to x)$, hence the implemented effect $\theta_W(w) := \Psi(s_w(x' \to x)) - \Psi(h(x))$ and the *designated gap* $\Delta(W) := \theta_W(a) - \theta_W(b)$, both under f_W, so $\Delta(0)$ is the baseline gap under f. Write $\tilde{s}_a := s_a(x' \to x)$ and $\tilde{s}_b := s_b(x' \to x)$ for the two interchange states under f, and $G := \nabla\Psi(\tilde{s}_a) + \nabla\Psi(\tilde{s}_b)$ for the gap gradient there: a sum, not a difference, because the two displacements enter Δ with opposite signs (hypothesis (H2) of Thm. 1).

Definition 1 (Exposure, and the exposure ratio)**.** The *exposure* of a reported ordering is its first-order steerable range $\|G\|\,\|u_\perp\|$; the *first-order exposure ratio* is exposure divided by the gap the ordering reports,

$$R \;:=\; \frac{\|G\|\,\|u_\perp\|}{|\Delta(0)|}, \qquad \Delta(0) \neq 0. \tag{1}$$

R is computable on f alone from two intervened forward passes and two backward passes beyond the clean runs any analysis already has, before any witness exists; the name is deliberate: R omits the remainder ω_Ψ in the theorem's margin, so $R > 1$ is a first-order threshold, not a certificate.

3 Fixed Finite Probe Families Need Not Identify Component Rankings

Lemma 1 (Exact behavioral invariance)**.** *For every W and every input x, $f_W(x) = f(x)$ exactly: the layer-ℓ activation, all downstream activations, the set of reachable activations (the activation manifold $\mathcal{M}$), and the output distribution are identical.*

Both writes evaluate ϕ on the stream *entering* layer ℓ, so their sum is $a + b$ and f, f_W are indistinguishable on natural behavior for any amount of data. Throughout, an estimator observes inputs, outputs, residual streams and the family $\mathcal{I}$'s outcomes, not the site's two writes separately, which differ by $\pm W\phi(x)$ off the probe span (the $f^{(2)}$ paragraph below). The question is what interventional evidence adds: given every probe outcome exactly, is the held-out ranking determined? It need not be. A ranking asserted only of the probed prompts is covered by construction (Cor. 1); the claim at issue is the ranking read as a statement about the task, and it is sharpest for an evaluator who receives intervention outcomes but not weights or write logs.

Theorem 1 (Non-identification from a fixed finite probe family)**.** *Let the site carry two components admitting the parallel-write decomposition of §2, both reading the same incoming state and writing additively into it, and let $\mathcal{I}$ be a finite probe family of zero ablations, within-$\mathcal{I}$ interchange interventions, and mean ablations whose reference mean satisfies $\overline{\phi}_R \in S$. Let (x', x) be an evaluation pair with $x, x' \notin \{x_p\}$, and let $\tilde{s}_a, \tilde{s}_b$ be the two held-out interchange states under f at that pair. Assume*

(H1) Transversality: $u_\perp \neq 0$.

(H2) Regularity: $\nabla\Psi$ *exists and is $L_{\nabla\Psi}$-Lipschitz on the closed balls $B(\tilde{s}_a, \|u_\perp\|)$ and $B(\tilde{s}_b, \|u_\perp\|)$, and the gap gradient $G := \nabla\Psi(\tilde{s}_a) + \nabla\Psi(\tilde{s}_b)$ is nonzero.*

(H3) Steering margin: $0 < |\Delta(0)| < \|G\|\,\|u_\perp\| - \omega_\Psi(\|u_\perp\|)$, *with the first-order remainder modulus*

$$\omega_\Psi(r) \;:=\; \sum_{k\in\{a,b\}} \sup_{\|v\|\le r} \big|\, \Psi(\tilde{s}_k + v) - \Psi(\tilde{s}_k) - \langle \nabla\Psi(\tilde{s}_k), v\rangle \,\big| \;\le\; L_{\nabla\Psi}\, r^2,$$

the bound following from (H2) by the integral form of the Taylor remainder.

Then there is a $W^\star \in \Omega$ with $\|W^\star\|_{\mathrm{op}} = 1$ such that f and $f_{W^\star}$

1. *agree pointwise on every output and on every residual stream at every layer, although their per-component writes differ;*
2. *return identical estimates for every intervention in $\mathcal{I}$; and*
3. *carry opposite designated gaps,* $\operatorname{sign}\Delta(W^\star) = -\operatorname{sign}\Delta(0)$, *with* $|\Delta(W^\star)| \ge \|G\|\|u_\perp\| - \omega_\Psi(\|u_\perp\|) - |\Delta(0)| > 0$.

Hence no estimator measurable with respect to natural behavior and the outcomes of $\mathcal{I}$ identifies the ordering of a and b.

Clause (2)'s family is stated for concreteness and understates what the construction survives: Lemma 1 is pointwise at every input, not merely at the probes, ruling out every stream-level class below.

Proposition 1 (What the ambiguity survives)**.** *Fix $W \in \Omega$. Then f and f_W return identical outcomes, at* every *input and not only on the probe family, for any intervention (i) upstream of ℓ, (ii) downstream of ℓ, (iii) on a component writing at ℓ other than the designated pair, or (iv) on the pair's* sum *$a + b$. Only an intervention that separately exposes a_W or b_W can distinguish them, and such an intervention agrees with f at the state level exactly when the displacement Lemma 2 assigns it vanishes: $W\phi(x)$ for an ablation, $W(\phi(x') - \phi(x))$ for an interchange.*

Each case is immediate from the pointwise cancellation ($a_W + b_W = a + b$ holds at every argument, and h is unchanged). So the ambiguity is not an artifact of the family we chose: it survives every class above except those that decompose the designated pair, and for those, probe-span directions are guaranteed invisible. Evidence reading one component's write directly (direct logit attribution off the probe span) decomposes the pair and falls under the same exception; mean ablation against an external corpus reference generally violates $\overline{\phi}_R \in S$ and likewise exposes it, an external reference mean being a cheap partial defense (Rem. 8 places stream-level alignment evidence in the invariant class).

Remark 1 (Scope)**.** Transversality forces $S \neq \mathbb{R}^q$; the designated components must be the two parallel writes; the held-out states must come from the same intervention operator as the probe family (Lem. 2(ii)); and the result is existential per-family. The restriction is on the *pair*, not the architecture: sequential pre-norm is out of scope at block granularity (the cancellation fails through the MLP

nonlinearity), but two heads in one attention layer read the same input and write additively in standard multi-head attention, including every model studied here, licensing the GPT-2 Small results of §4.

(H3) constrains two quantities computable on f alone, $\|G\|$ and $|\Delta(0)|$, against one that must be bounded, $L_{\nabla\Psi}$, read at the *realized* $\|u_\perp\|$ (Rem. 6). We cannot supply $L_{\nabla\Psi}$: sampled estimators lower-bound a supremum over a ball (App. D.2); a checkable f-only form of (H3) is therefore not instantiated, nor needed: the realized remainder is measured well below $\widehat{L}\|u_\perp\|^2$ for our sampled-ball estimate $\widehat{L}$.

What kind of object $f^{(2)}$ is, and what is being claimed. Write $f^{(1)} := f$ and $f^{(2)} := f_{W^\star}$ for the pair the theorem produces (Fig. 1). The redistribution changes the two components' *writes*, not the function they compose to, and we do not claim $f^{(2)}$ is reachable by another setting of the same weights: $W\phi$ is a pair of canceling linear side-paths, and weight inspection, or the per-component write logs that running an interchange already produces, distinguishes the pair immediately. The identification target is therefore exact: not whether the components are knowable, but whether the intervention outcomes a study reports identify its ranking, and Theorem 1 says they need not. At head granularity the ambiguity is resolved by data the protocol already collects, so the theorem's content there is that the resolution comes from the architectural decomposition, an identifying assumption that is often left implicit. A learned decomposition, whose feature basis the architecture does not privilege, inherits no such assumption (§4). The usual gauge transformations, relabeling and positive rescaling, are *intervention-invariant* and cannot produce this failure. The witness is non-trivial because it does both at once, annihilating the probe span and steering the uncovered contrast. Operator-norm budgets yield two-sided radii on the attainable gaps in the partial-identification tradition (Manski, 2003) (Cor. 3; Cor. 2 as the two-point form).

Coverage removes the witness. If $u \in S$, no redistribution in Ω moves the designated effects (Cor. 1); conversely $\|u_\perp\| > 0$ alone does not deliver a reversal, which needs (H3). Under *full-output* zero-ablation agreement the surviving set is $\mathcal{N}_0 \supseteq \Omega$, strictly larger under centered LayerNorm; coverage eliminates all of $\mathcal{N}_0$ under a stated invariance condition (Prop. 3, Rems. 2–3), and a study checking agreement only on its reported statistic rules out less.

The canonical feature map. S, and hence $\|u_\perp\|$, depends on ϕ and q, which belong to the *construction* rather than the network; a diagnostic defined against a ϕ the analyst does not control is not a diagnostic, so we fix the canonical choice, ϕ the identity on the layer-ℓ input residual with $q = d$, making transversal mass basis-free.

Proposition 2 (Global coverage requires d independent probe directions)**.** *With $\phi = \mathrm{id}$ and $q = d$, every $W \in \Omega$ has all d of its rows in $S^\perp$, so $\dim \Omega = d\,(d - \operatorname{rank}\Phi) \geq d\,(d - P)$. Hence $\Omega \neq \{0\}$ whenever $P < d$, and $\Omega = \{0\}$ requires at least d probes in general position.*

What a larger family buys is therefore not the elimination of the witness class, which stays high-dimensional over single-position budgets, but a reduction in $\|u_\perp\|$ that the reversal rate is insensitive to until the span nears saturation (§4); removing the witness for a specific claim needs coverage of one direction, not global spanning, at a realizable cost App. A.2 measures.

Corollary 1 (Covering the evaluation contrast eliminates this witness)**.** *If $u \in S$ then $Wu = 0$ for every $W \in \Omega$, so no redistribution in Ω moves the designated effects and the witness is unavailable, however large Ω remains. The transversal mass $\|u_\perp\|$, not $\dim S^\perp$, is the operative necessary quantity, and full feature-span coverage ($S = \mathbb{R}^q$, forcing $\Omega = \{0\}$) is the stronger special case that delivers this for every evaluation pair at once. Proof and converse: App. A.*

The gap between the two conditions is measured in §4: zero exposure with $\dim S^\perp = 176$ on one side, failure to cover at $16\times$ the intrinsic dimension on the other; hence the recommendation: report the first-order exposure ratio R of (1) under the canonical ϕ for each ordering claimed. Probe count alone is not a defense: under the canonical ϕ protection appears only near $P = d$ (§4), and at fixed count exposure returns as soon as a construction enlarges its feature dimension; at this site, the observed outcome is tracked by whether R lies above or below 1. Two caveats: R omits ω_Ψ, so neither $R > 1$ nor $R < 1$ is a certificate, and it is trustworthy only where the remainder is measured small, as here; §4 evaluates the threshold on both sides, including naturally occurring sub-threshold contrasts.

4 EXPERIMENTS

The experiments show five things: a single-position family at published-analysis scale does not cover the ordering it reports; one *fixed* witness steers it out of sample; untargeted growth helps only near the width-scale ceiling, while positions buy coverage by a rank law; covering the reported contrast alone removes the witness; and the observed reversal threshold closely follows $R = 1$, including on a structurally different Greater-Than circuit (protocols: App. B.3; per-cell tables: App. D).

A published circuit's head pair, at the probe scale that analysis collects. Two heads in one attention layer read the same layer-normed input and write additively into the same stream, so Lemma 1 holds exactly even in sequential blocks; we test this on GPT-2 Small, on the name-mover pair L9H9 / L9H6 that Wang et al. (2023) rank, plus two replication pairs. (IOI prompts such as "When Mary and John went to the store, John gave the milk to" are completed with the non-repeated name; name movers' final-token writes promote it.) The probe family is scaled to what such an analysis collects, not to what suits the construction: up to 512 prompts from 15 templates under the canonical $\phi = \mathrm{id}$, $q = d = 768$, so no PCA dimension is free. The witness reverses 24 of 24 held-out contrasts at every probe count through $P = 512$ (Tab. 1), where $\operatorname{rank}\Phi = 512$ still leaves $\dim S^{\perp} = 256$; pushed toward the ceiling, reversal turns over sharply ($22/24$ at $P = 640$, $1/24$ at 768); the $P = 768$ cell annihilates only to a structural 1.5×10^{-3}, a transversal-mass collapse rather than a clause-(2) instance, so prompt growth is empirically protective only near width-scale saturation here, and nowhere at $d = 2048$, even position-resolved (Tab. 3); the exact result remains the covered-contrast arm. The reversals need no oversized side path: at $P = 128$ all $24/24$ reverse with the added write capped at a quarter of the smaller head's own write, jointly inside unit operator norm; at $P = 512$ unit norm realizes 0.17–$0.55\times$ component scale (App. B.3). Three scope notes: the family matches the published analysis in scale and template diversity, not in protocol (Wang et al. (2023) use path patching we do not reproduce), so the claim is about coverage, not published numbers, and the baseline ordering our single-position interchange measures, which the witness reverses, need not match and here does not match the published path-patching ordering; the heads' output row spaces intersect trivially, so the added write is a re-decomposition, not a disguised weight edit; and annihilation residuals are reported against a double-precision nullspace basis (App. B.3).

One fixed witness reverses that ordering aggregated over prompts, out of sample. The witness above is built per contrast, whereas a circuit analysis reports an ordering aggregated over a prompt distribution, and reversing that is strictly harder: a single W at unit operator norm must steer k contrasts whose $u_{\perp}$ point in different directions. It is available in closed form, a nuclear-norm dual solved by one SVD, with no optimizer and no tuned hyperparameter (App. B.3). Writing $\Delta(\alpha W)$ for the gap under the witness scaled by α: at the same head pair with a 128-prompt family, one W fitted on 24 contrasts flips their aggregate gap from -1.081 to $+1.082$ at $\alpha = 0.25$, and on a

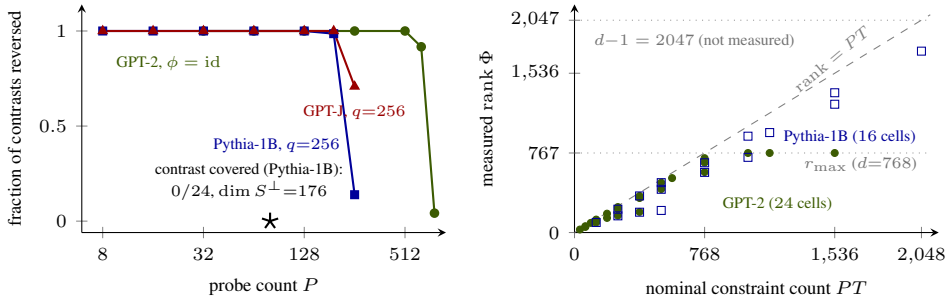

Figure 2: **Claim-specific coverage, not probe count alone, controls exposure to the ambiguity.** *Left:* reversal rate stays at essentially 100% as single-position families grow (models and feature maps as labeled), turning over only near effective saturation (GPT-2 at $P \in \{640, 768\}$, Tab. 1; Tab. 6); the starred point is the covered-contrast arm: exposure dies at *coverage*, not at exhaustion. *Right:* every position-resolved cell at two widths obeys Eq. (2); the nominal count PT (dashed) overstates the rank by the first-position deficit, so cells at $PT = d$ still reverse (40–91 of 96) and only cells reaching $r_{\max}$ (dotted) collapse. Per-cell values: Tabs. 2 and 3.

disjoint 24 the held-out aggregate flips at the same scale, from -0.777 to $+0.541$, with 23 of 24 contrasts individually reversing at $\alpha = 0.5$; the sign at unit norm is checked directly ($+7.33$ in sample, $+4.39$ held out) and holds at every scale on the frozen grid through $\alpha = 5$. A single fixed redistribution therefore misleads contrasts it was never fitted to, the form the reported ordering takes. The construction also survives *template* shift: a witness whose probes and fitting contrasts use only ten of the fifteen templates flips the aggregate on 24 contrasts from the five unseen templates at $\alpha = 0.25$ ($-0.468 \to +0.586$; $+3.70$ at unit norm), with 23 of 24 individually reversing from $\alpha = 0.5$ (App. B.3). W_{joint} meets the 10^{-4} annihilation criterion at unit norm (2.0×10^{-12} in double precision; the template-shift witness at 2.3×10^{-12}).

A composed $k = 3$ sibling chain extends the reversal to a three-head ordering, exactly inverted out of sample at budgets 2 and 4; the $k = 4$ chain fails the pre-registered annihilation criterion, so no $k = 4$ instance is reported (protocol and results: App. B.3).

Two JumpReLU SAEs, one audited reconstruction, opposite orderings. Beyond architectural components, where §3 concedes weight inspection adjudicates, the dictionary analogue is exact: repurposing a corpus-dead unit of the canonical 16k Gemma Scope JumpReLU SAE (Lieberum et al., 2024; Rajamanoharan et al., 2024) on Gemma-2 2B (Gemma Team et al., 2024) (one unit's encoder parameters and two decoder rows, App. D.1) yields a second JumpReLU SAE of identical architecture and width, pointwise identical reconstruction on both audited corpora (instantiated and verified), and probe outcomes agreeing to 0.0 exactly on all 64 probes per pair, each drawn where the intervened features are inactive, while the reported importance ordering reverses on a median 20.5 and 20.0 of 24 held-out active prompts across 12 of 12 screened pairs on both corpus seeds (pre-committed threshold: 3 of 12 at 12 of 24, one seed). The reversal holds at the within-norm cap, with aiming gradient, steering sign, and screening statistic from a witness slice disjoint from evaluation, and survives excluding the shared dense partner ($6/6$ fresh pairs, both seeds). The alternative is edited, not trained, L_0 rising by one where the sheared feature fires: this intervention evidence does not separate the two SAEs; the training objective, weight inspection, and co-activation do, exactly as for $f^{(2)}$. Two precursors, an unaimed variant at 1 to 3 per cell and a pilot shear whose pre-committed unsigned statistic returned 0 of 24, are in App. D.1.

The witness at 1B scale. On Pythia-1B at layer 8, site-local at the final token, we parameterize $W = AN^\top$ with N an orthonormal basis of $S^\perp$, so $W\Phi = 0$ holds structurally; evaluation contrasts are held out from the probe family and from a pool slice never examined during development, every candidate reported rather than selected (App. B.3). Four arms share the site, family, implementation and contrasts (Tab. 5): the theorem's analytic $W^\star = \hat{e}\, u_\perp^\top / \|u_\perp\|$; a non-annihilating W_{bad} as positive control; a random direction inside Ω at the same norm; and an optimized direction inside Ω. The primary statistic is $\widehat{\alpha}_{\min}$, the smallest tested scale on a frozen grid at which the gap changes sign, with the criterion a reversed sign at $\alpha = 1$ checked directly.

Table 1: **A single-position family the size a circuit analysis collects does not cover the contrast it reports.** IOI name-mover ordering on GPT-2 Small, $\phi = \text{id}$, $q = 768$, nested families over 15 templates; every cell through $P = 512$ reverses all 24 held-out contrasts and meets $\|W\Phi\|_F \le 10^{-4}$. For this *clean-prompt* family $\text{rank}\,\Phi = P$ exactly, and the full 1,576-prompt task pool measures rank 767 of 768: approaching that ceiling, the $P \in \{640, 768\}$ extension (an 842-prompt pool; rank 765 at $P = 768$) turns over at $22/24$ ($\|W\Phi\|_F = 4.6 \times 10^{-5}$; 4.0×10^{-13} in double) and collapses to $1/24$, the $P = 768$ cell annihilating only to a structural 1.5×10^{-3}, unchanged in double precision, the $\sigma_{\min}$ mechanism of Tab. 2: a numerical turnover, not a certified probe-preserving witness. Bold marks extremes; n/a cells are replication cells not run, not failures.

probe count P	32	64	128	256	512	640	768
dim $S^\perp$	736	704	640	512	256	128	3
L9H9 / L9H6 (name movers)	**24/24**	**24/24**	**24/24**	**24/24**	**24/24**	22/24	**1/24**
L9H9 / L9H0	24/24	n/a	24/24	n/a	24/24	n/a	n/a
L10H10 / L10H7	24/24	24/24	24/24	24/24	24/24	n/a	n/a

$W^\star$ gives $\Delta = +2.61$ at $\alpha = 1$ and first changes sign at $\alpha = 0.25$, so *the theorem's witness reverses the ranking at a quarter of unit operator norm*; across the 24 fresh contrasts all 24 reverse at or below unit norm, at a median $\widehat{\alpha}_{\min}$ of 0.5.

$W^\star$ annihilates the probe span $40\times$ inside the criterion, while the exposed control registers 2.6×10^6 times the calibrated ceiling. Direction, not magnitude, is what reverses: in a matched null of 400 random directions inside Ω at identical operator norm, 1 reverses at $\alpha = 1$ and *none* at or below the witness's $\alpha = 0.25$ (App. B.3). Three caveats: W_{opt} is a post-hoc-normalized search, reported as supplementary; the tabled contrast sits at the 14th percentile of the baseline-gap distribution, a favorable case; and the feature map is a PCA map at $q = 64$ against $d = 2048$, as is the sweep below at $q = 256$, a restriction conservative for the exposure claim ($q \ll d$ *understates* it, App. D) but one that makes the turnover cells properties of the reduced map. The width replication below supplies a canonical-ϕ head-level measurement at 1B, reversing $24/24$ in 14 of 16 cells; the four-arm protocol itself is not rerun at $q = d$.

Enlarging the probe family reduces exposure only near the effective ceiling. The insensitivity anticipated after Prop. 2 holds at 1B and 6B scale and under a learned feature map: a *nested* family at fixed $q = 256$ leaves reversal at 98.6–100% from $P = 8$ through 192 on Pythia-1B and GPT-J-6B, and sweeping q up to 1024 moves nothing either, $216/216$ reversing (Fig. 2, left; Tab. 6). In these *untargeted* sweeps exposure approaches zero only near saturation, a property of untargeted growth rather than of Cor. 1: the targeted arm below reaches zero exposure with $\dim S^\perp = 176$ intact.

Positions buy coverage where prompts do not. The families above expose ϕ at one token position; a protocol patching at T positions contributes T constraints per prompt, since an upstream patch changes the site's feature and probe agreement then requires $W\phi(x_p, t) = 0$ for every (p, t) (the redistribution stays at the final-token site; the family probes the designated pair within each patched context, exposing that context's feature, Lem. 2(i)). Measured on the same head pair over four pool seeds, those constraints are not collinear: each position beyond the first adds exactly P fresh directions, while the first contributes only a sublinear $r_1(P)$, falling from $0.91P$ to $0.52P$ because patching the first differing position merges prompts sharing a template and a name pair. The empirical law at this site is therefore

$$\operatorname{rank} \Phi(P, T) \;=\; \min\{\, r_1(P) + (T-1)P,\; r_{\max} \,\}, \tag{2}$$

with $r_{\max} = 767$ the effective rank at relative SVD tolerance 10^{-6}, measured rather than derived (ϕ reads the pre-normalization residual, so no LayerNorm constraint applies). The law reproduces all 24 GPT-2 cells and, on Pythia-1B at $d = 2048$ under the same canonical ϕ, all 16 cells, with near-identical r_1, suggesting r_1 depends primarily on the patched prompt distribution, not model width (Fig. 2, right; Tab. 2). The clean-prompt family of Tab. 1 instead has $\operatorname{rank} \Phi = P$ exactly and approaches the ceiling near $P = d$: positions supply directions the prompt distribution does not.

In our evaluation sets, exposure collapses exactly when Eq. (2) reaches $r_{\max}$, and the nominal product PT overstates that point by the first-position deficit: every starred $PT = d$ cell of Tab. 2 still reverses, the smallest collapsing cell is $(256, 4)$ at $PT \approx 1.33d$, and at $d = 2048$ no reachable budget collapses (Tab. 3). These cells measure the collapse of the empirical transversal mass, not theorem-certified indistinguishability; the theorem-faithful zero-exposure test is the covered-contrast arm below. What positions add is mostly, but not only, constraint count: at matched rank, $T = 4$ still leaves less transversal mass than $T = 1$ (Tab. 4). In short, probe count, probe rank and contrast coverage are distinct, and only the last directly controls $\|u_\perp\|$; collapsing it costs 1024 prompt-positions at $d = 768$, with a rank requirement that scales with width (§6). Note what does *not* happen: $\dim S^\perp = 1$ at the collapsing cells, so $\dim \Omega = d$ and the algebraic freedom survives; what is exhausted is the steering mass. Held there, three models separate by transversal mass alone (GPT-J-6B at $\|u_\perp\| = 0.44$ reverses 71%, Pythia-1B at 0.033 reverses 13.9%, GPT-2 at 0.000 reverses nothing), the most direct indication in our data that $\|u_\perp\|$ and not $\dim S^\perp$ is operative. One caveat, stated rather than found: at $\dim S^\perp = 1$ the annihilation residual is $\sigma_{\min}(\Phi)$ by construction and exceeds the pre-fixed 10^{-4} tolerance in three of the four seed-cells (1.1 to 2.8×10^{-4}), so those cells do not satisfy clause (2) of Thm. 1 on its own terms and the comparison is a consistent cross-model pattern rather than three certified instances (App. D).

Covering the reported contrast removes the witness, with the annihilator still large. Cor. 1 predicts that covering the *evaluation contrast* suffices however much freedom remains, and we test

it on Pythia-1B by placing both endpoints of every evaluation pair inside the probe family. With 80 probes at $q = 256$ the annihilator retains $\dim S^{\perp} = 176$, yet the transversal mass collapses to a maximum 5.5×10^{-6} and no contrast reverses at any scale up to $\alpha = 10$, the gap moving by at most 1.3×10^{-3} on gaps of order one: the only zero-exposure point we measure that is not span saturation, and the one the recommendation rests on. That coverage is also what a study cannot easily buy: orthogonal matching pursuit against a *single* contrast reduces $\|u_{\perp}\|$ by a median $7.8\times$ at 400 probes, $16\times$ the intrinsic dimension, yet leaves 23 of 24 reversing; the budget at which targeted coverage becomes affordable is not something we can bound (App. A.2).

The observed reversal threshold closely follows $R = 1$, and R is measurable before any witness exists. Across 384 cases (360 held-out contrast–q cases from five probe families, plus 24 from probe selection), **382** of 384 reverse at unit operator norm, and the two exceptions are exactly the two cases with $R < 1$; in the five-family held-out grid alone, 359 of 360 reverse, at baseline gaps up to 1.35 logits. This is one measurement reported 384 times, and we say so: with the realized remainder at median 0.006 (arm-run site), $R > 1$ and reversal coincide *by arithmetic*: Ψ is close to affine over the realized displacement. The dose–response sweep locates the crossing at median $\widehat{\alpha}_{\min} \cdot R = 1.028$, 92.2% of contrasts in $[0.95, 1.15]$ and none below 0.95 (App. D); it scales one witness, so it does not rule out other constructions below threshold. R's input is measurable in advance: a 32-prompt Pythia-1B family leaves 56–81% of held-out contrast mass uncovered. Pipeline-shaped evidence inverts the comparison: an ACDC-shaped family (50 clean/corrupted IOI pairs, every upstream node patched at all positions) reaches the rank-767 ceiling by *ten* pairs, median uncovered fraction 1.7×10^{-6}, median $R = 3.8 \times 10^{-5}$, 0 of 24 above threshold (App. D): the diagnostic separates evidence *shapes*, single-position interchange tables leaving the reported contrast steerable while the trace of a diverse-node patching sweep does not. Both shapes replicate with prompts sampled from the verbatim released IOI assets of Wang et al. (2023): single-position families leave median R between 9.6 and 1.7 for $P = 32$ to 512 ($R > 1$ on 22 to 24 of 24), while the ACDC-shaped family reaches rank 767 by ten pairs, median $R \leq 2.9 \times 10^{-5}$ under both operators (App. D).

The geometry generalizes beyond IOI. The same protocol, its outcome reading pre-committed in the producing script, runs on the Greater-Than task of Hanna et al. (2023) on two sibling head pairs, L8H11/L8H8 and L6H9/L6H1 (task format and measurement: Tab. 7). Both pairs reverse 24 of 24 held-out contrasts at every $P \in \{32, 64, 128\}$; unlike IOI, reversal then *falls* as the family grows, $24/24$ to $19/24$ to $12/24$ at $P = 128, 256, 512$ on the layer-8 pair, median R falling alongside it to 1.2 at $P = 512$. Across both pairs and all five probe counts, $R > 1$ agrees with reversal on 238 of 240 contrasts; both exceptions sit just above the boundary ($R = 1.11$ and 1.03, not reversing), and none of the 28 naturally occurring $R < 1$ contrasts reverses (Tab. 7). A task with different structure and an earlier probe-count turnover shows the same exposure relationship: R, not probe count alone, tracks reversal across tasks.

5 Related Work

Estimation pathologies and attribution impossibility. Off-distribution representations (Grant et al., 2026), geometric steering criteria (Wurgaft et al., 2026) and sampling uncertainty (Asiaee, 2026) are estimation pathologies under a fixed network; ours persists with exact probe outcomes, every natural activation unchanged. Bilodeau et al. (2024) prove an impossibility for complete-and-linear attribution *methods* given a model; ours is over *models* given the evidence; the two compose.

Behavior-preserving reparameterization. Permutation symmetries (Entezari et al., 2022; Ainsworth et al., 2023) give behaviorally identical networks with different decompositions, but they are *intervention-invariant*, so they cannot produce this failure (§3).

Identifiability under interventions. The argument shape is old and we do not claim it: Hyvärinen & Pajunen (1999) for nonlinear ICA, and Locatello et al. (2019) as this paper's structural twin, preserving all *observational* evidence while destroying the target. Interventions are the canonical escape, and a substantial line establishes when they, or auxiliary structure, suffice (Khemakhem et al., 2020; von Kügelgen et al., 2021; Ahuja et al., 2023; Squires et al., 2023). That line asks whether interventions identify latents of an *unknown* generative process with intervention *targets* as the lever; we ask whether they identify the ordering of named components of a *known* network with probe *inputs* as the lever.

Sparse representations. Dictionary learning admits identifiability up to permutation and scaling under sparsity conditions (Spielman et al., 2012); SAE non-uniqueness is a stated open problem (Sharkey et al., 2025). Ours asks instead whether a finite feature-intervention family identifies the feature ranking with architecture and reconstruction fixed on the audited distribution; the dead-unit construction answers no, claiming no loss-equivalence.

Intervention success versus localization. Hase et al. (2023) show a successful edit does not license the localization claim motivating it; Miller et al. (2024) document circuit-faithfulness conclusions shifting with the ablation (practice guidance: Heimersheim & Nanda, 2024). Automated discovery selects components by this same evidence (ACDC; Conmy et al., 2023), lately via gradient approximations (Syed et al., 2024; Kramár et al., 2024), whose per-write Jacobians are exposing (Rem. 8), and over learned features (Marks et al., 2025); whether its family covers the reported contrasts is checkable; §4 measures an ACDC-shaped instance.

Prior non-identifiability results in interpretability. Méloux et al. (2025) is the closest antecedent: enumerating explanations for Boolean tasks on small MLPs, they show many circuits replicate a behavior. Their evidence is enumerative, over models small enough to search completely; ours is a constructive witness with stated conditions at transformer scale. Sutter et al. (2025) show unconstrained alignment maps make causal abstraction vacuous, a degeneracy in *explanation* space; ours is in the space of *networks* consistent with a fixed evidence base, alignment map held fixed. Bolukbasi et al. (2021) and Friedman et al. (2024) are representation-level antecedents.

Causal-inference identifiability. Causal scrubbing (Chan et al., 2022) resamples from hypothesis-consistent partitions; causal abstraction (Geiger et al., 2021; 2025) asks whether a network realizes a chosen high-level model; we fix the question and ask whether the probe family identifies the answer.

6 Limitations

The result has four scope boundaries. **Granularity, not architecture:** the two components must be siblings at a shared write point (Rem. 1); for the three IOI head pairs whose output spaces we measure, the witness is not weight-realizable, since each head's write lies in its rank-64 output-projection column space and the measured trivial intersection of those spaces leaves no nonzero witness (App. B.3); dictionary granularity, where no decomposition is architecturally privileged, admits an exact within-architecture instance on the audited corpus (§4); the edited SAE is weight-visible and not loss-equivalent (its L_0 rises where the repurposed unit fires), so the reversal within a *trained*, loss-matched family remains open. **Probe family:** the construction is separable only by interventions that decompose the designated pair (Prop. 1); covering the reported contrast eliminates the directions the witness needs. **Position and budget:** uniform protection over arbitrary feature-space contrasts requires the full probe span, claim-specific protection only coverage of the reported contrast (Cor. 1); the rank law is measured at $d = 768$ and 2048, exposure collapse only at 768, so we do *not* establish that position-resolved protocols confer identification at scale. A protocol that exactly covers the reported contrast identifies the ordering against this witness class; the ACDC-shaped audit is empirical near-coverage, not a certificate, and matches that pipeline's shape: a replication samples its prompts from the released IOI assets (App. D), but the operators are ours, so we make no claim about the path patching of Wang et al. (2023) itself. **Prevalence:** the theorem is existential: the witness reverses 359 of 360 held-out contrasts on the main grid, but at one Pythia-1B site and five GPT-2 head pairs over two task families (IOI and Greater-Than); a rate for published orderings requires reconstructing their probe families. Off-manifold bias is an orthogonal failure mode (App. C): the redistribution leaves natural activations and the manifold unchanged, though interchange can still generate off-manifold states.

7 Conclusion

Natural behavior and every supplied probe outcome can remain exactly unchanged while the interchange ordering of two components reverses on a held-out contrast (Thm. 1), structurally (Cor. 2), not as a family artifact (Prop. 1). The implication is a stricter, coverage-based standard, with first-order exposure measurable through R: judge an intervention by whether its evidence covers the quantity reported, not only by whether it moves behavior; successful interventions are not, by themselves, evidence that the reported ranking has been causally identified.

REPRODUCIBILITY STATEMENT

Compute and artifacts. Each experiment runs on a single device, on one of three platforms: one NVIDIA GB10 Grace-Blackwell device with 128 GB unified memory (Pythia-1B, GPT-J-6B, the precursor Gemma-2 2B runs, the k-NN projection runs of App. C in fp16, and most GPT-2 Small runs, including the cells of Tabs. 1 through 4 apart from Tab. 1's $P \in \{640, 768\}$ extension, the joint witness and the composed chain, the Greater-Than replication of Tab. 7, and the covering experiment of App. A.2); Apple-silicon MPS with 128 GB unified memory (Pythia-160M, the primary Gemma-2 2B dead-unit runs, and the GPT-2 Small coverage audits, component-scale audit and template-shift witness); or CPU (the $P \in \{640, 768\}$ extension cells of Tab. 1).

Device, dtype and seed are recorded in every archived result file and stated in the appendix for each experiment. The unified-memory platform is worth naming because it fixes the host-device transfer behavior and the available reference-bank size, both of which bound the k-NN estimator of App. C.

An anonymous artifact bundle with the pre-execution protocol code, the analysis scripts for every experiment reported here, and every archived result file referenced in the text (`results/*.json`, `results/*.csv`) accompanies the submission, with a public release at camera-ready; it additionally contains exploratory analyses not reported here, the pre-registration of the circuit-scale experiment, and the pre-correction artifacts named below. Pre-committed outcome readings are recorded in the protocol scripts' own headers, written before the producing runs.

Pre-specification. Success criteria were fixed in protocol code before each producing experiment (Apps. B.2, D). The four-arm dose grid was fixed before the arm run; after that run revealed censoring at the lower end of the grid, the higher-resolution sweep grid was fixed before any sweep outcome was generated. Seeds and per-experiment configurations are given with each experiment in the appendix.

Artifact audit. Three implementation defects, all in activation-site semantics (where the redistribution was applied, or where activations were read), were found during development; the affected pipeline's results were withdrawn in full (internally E3, App. B.2) and every downstream measurement re-run. The third defect, found after the replacement protocol had produced results, had *understated* the headline, reporting 6 of 24 reversals where the corrected figure, reproduced through an independent hook path sharing no code with the reported one, is 24 of 24. App. B.2 records all three defects, the checks that now guard the interchange semantics, and that recomputation.

AI USE STATEMENT

Generative AI tools were used in this work for code assistance (drafting and refactoring experiment scripts and analysis utilities), for editorial assistance (copy-editing prose and LaTeX), and for research discussion, including critique of experimental design, theoretical exposition, and manuscript framing. The research questions, the construction, the theoretical statements and proofs, and the experimental protocols are the authors'; all research decisions, analyses, and numerical results were independently checked and validated by the authors. AI-assisted code was read line by line and its outputs cross-checked against independently archived result files, and every numerical claim in the paper was traced to its producing artifact. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work analyzes the validity of interpretability interventions on publicly released pretrained models and public datasets; no human subjects or personal data are involved. Its intended effect is to reduce over-claiming of causal structure in model audits by making intervention evidence auditable; the constructive impossibility pair is a diagnostic for evaluators, not a deployment technique.

The construction requires white-box weight access, which in an audit the developer holds and the evaluator may not, so the practical reading is a condition on what an audit must collect rather than a technique: where an evaluator receives intervention outcomes but not weights or per-component write logs, identification of a component ranking is conditional on trusting the reported decomposition. The separators are weight inspection, component-write access, and, for learned dictionaries, the training objective and co-activation structure.

The dictionary case is the one where an alternative is realizable within the architecture (App. D.1); at head granularity the alternative is not itself a pair of attention heads (App. B.3), so there the assumption being trusted is the architectural decomposition itself.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 372–407, 2023.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 337 of *PMLR*, pp. 210–227, 2026. arXiv:2607.08349.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. *Alignment Forum*, 2022.

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 (NeurIPS)*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 14035–14059, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning (CLeaR)*, volume 236 of *PMLR*, pp. 160–187, 2024.

Atticus Geiger, Duligur Ibeling, Amir Zur, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. Oral presentation. arXiv:2511.04638.

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 (NeurIPS)*, 2023.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling (COLM)*, 2024.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

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

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Ilyes Khemakhem, Diederik P Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *PMLR*, pp. 2207–2217, 2020.

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.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 4114–4124, 2019.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.17030.

Charles F Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.20914.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit evaluation metrics are not robust. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2407.08734.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 45069–45108, 2025.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2501.16496.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.13032.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *PMLR*, pp. 37.1–37.18, 2012. arXiv:1206.5882.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 32540–32560, 2023.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2507.08802.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024. arXiv:2310.10348.

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 (NeurIPS)*, 2020.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

Kevin 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 (ICLR)*, 2023.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026.

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

A FULL PROOFS

Scope remarks on Theorem 1. The strict baseline gap $\Delta(0) \neq 0$ is needed for $\operatorname{sign}(\Delta(0))$ to be defined; a baseline tie would require a symmetric $\{f_{+W}, f_{-W}\}$ pair, which we do not pursue. No labeling convention on the sign of $\Delta(0)$ is needed, since the witness carries $\operatorname{sign}(\Delta(0))$ explicitly through $\hat{e}$. The hypothesis list is not minimal: since (H3) forces $\|G\|\|u_\perp\| > \omega_\Psi(\|u_\perp\|) \geq 0$, it already implies both the transversality of (H1) and the nonvanishing of G in (H2), and the Lipschitz clause of (H2) is used only for the displayed bound $\omega_\Psi(r) \leq L_{\nabla\Psi} r^2$, whose constant App. D.2 reports we cannot obtain; the expansion itself needs only differentiability at $\tilde{s}_a, \tilde{s}_b$ and finiteness of ω_Ψ. We keep the three labels because each names a distinct role.

The result is existential per-family: we prove no universal lower bound over arbitrary intervention families. And the evaluation pair must lie outside the probe set, since if $x, x' \in \{x_p\}$ then $u \in S$ and Corollary 1 removes the witness, so the theorem concerns an intervention outcome at a contrast the analyst did *not* probe, which is the held-out generalization regime the experiments study.

Corollary 2 (Two-point estimation lower bound)**.** *Assume (H1) and (H2) of Thm. 1; the steering margin (H3) is not needed for the estimation bound and enters only where noted. Write $\Delta_\theta(f) := \theta_f(a; x', x) - \theta_f(b; x', x)$ for the designated implemented effect gap, set $\hat{g} := G/\|G\|$ (unsigned, unlike the witness direction $\hat{e}$ used in Thm. 1) and $W_\pm := \pm\, \hat{g}\, u_\perp^\top / \|u_\perp\|$. Both lie in Ω with unit operator norm, so by Lemma 1 and clause (2) the networks f_{W_+} and f_{W_-} induce identical natural behavior and identical outcomes on every probe in $\mathcal{I}$; any estimator $\widehat{\Delta}$ measurable with respect to that evidence therefore returns the same value on both, and*

$$\inf_{\widehat{\Delta}} \max_{\pm} \left|\widehat{\Delta} - \Delta_\theta(f_{W_\pm})\right| \;\geq\; \tfrac{1}{2}\left|\Delta_\theta(f_{W_+}) - \Delta_\theta(f_{W_-})\right| \;\geq\; \|G\|\|u_\perp\| - \omega_\Psi(\|u_\perp\|),$$

since $W_\pm u = \pm\|u_\perp\|\hat{g}$ gives $\Delta_\theta(f_{W_\pm}) = \Delta(0) \pm \|G\|\|u_\perp\| + \varrho_\pm$ with $|\varrho_\pm| \leq \omega_\Psi(\|u_\perp\|)$, so the baseline gap cancels in the difference.

Two consequences worth noting. The bound is twice what the asymmetric pair $\{f, f_{W^\star}\}$ yields, because anchoring one endpoint at f wastes half the attainable spread. And this estimation *bound requires no steering margin, since $\Delta(0)$ drops out; it is however vacuous unless $\|G\|\|u_\perp\| > \omega_\Psi(\|u_\perp\|)$, which (H1) and (H2) alone do not supply. Under (H3) it is automatically non-vacuous and in fact exceeds $|\Delta(0)|$, since (H3) reads $|\Delta(0)| < \|G\|\|u_\perp\| - \omega_\Psi(\|u_\perp\|)$. The infimum is over estimators measurable with respect to natural behavior and the outcomes of $\mathcal{I}$; without that restriction the claim is false, since an estimator with weight access returns the truth.*

The ranking *statement is separate and does need the margin. Under an equal prior over the pair, the* best *(possibly randomized) ranking rule errs with probability $1/2$ precisely when $\operatorname{sign}\Delta_\theta(f_{W_+}) \neq \operatorname{sign}\Delta_\theta(f_{W_-})$ (the evidence being identical, no rule can do better than chance; when the signs agree the constant rule errs never), for which it* suffices *that $|\Delta(0)| < \|G\|\|u_\perp\| - \max(|\varrho_+|, |\varrho_-|)$, essentially the steering margin of Theorem 1. The exact condition for this orientation ($\Delta_\theta(f_{W_+}) > 0 > \Delta_\theta(f_{W_-})$) is two-sided in the realized remainders, $-A - \varrho_+ < \Delta(0) < A - \varrho_-$ with $A = \|G\|\|u_\perp\|$, and is implied by but not equivalent to the displayed inequality. Absent it both members can carry the same sign, and the constant rule returning that shared sign is then correct on both.*

Proof of Cor. 1. Write $S := \operatorname{span}\{\phi(x_p)\}_{p=1}^P$ and $u := \phi(x') - \phi(x)$. Every $W \in \Omega$ annihilates S row-wise, so $u \in S$ gives $Wu = 0$, and by Lem. 2 the interchange states under f_W are $\tilde{s}_a + Wu$ and $\tilde{s}_b - Wu$, i.e. exactly the states under f. Hence $\theta_W(a) = \theta_0(a)$ and $\theta_W(b) = \theta_0(b)$, so $\Delta(W) = \Delta(0)$ for every such W and no reversal is available at any operator norm. Nothing in the argument bounds $\dim\Omega$, which is why coverage and exhaustion are independent conditions.

Two scope notes. The quantifier is over Ω rather than over all probe-preserving redistributions because annihilation is sufficient but not necessary for probe agreement (App. B): a nonzero $W\phi(x_p)$ could in principle lie in a null direction of the readout, and Prop. 3 supplies the converse under a strengthened notion of agreement. The coverage implication itself has a one-line algebraic converse: if $u \notin S$ then $u_\perp \neq 0$, and the rank-one lift $W_e := e\, u_\perp^\top / \|u_\perp\| \in \Omega$ of Rem. 6 gives $W_e u = \|u_\perp\|\, e \neq 0$, so some witness displaces the evaluation states. And $\|\mathsf{P}_{S^\perp} u\| > 0$ does not by itself deliver a reversal, which additionally requires the steering margin (H3); nonzero transversal mass is the necessary condition on the evaluation directions, not a sufficient one for a reversal. □

Proposition 3 (Probe-preserving redistributions, exactly)**.** *Let $\mathcal{I}_0 \subseteq \mathcal{I}$ be the zero-ablation subfamily, containing the zero ablation of one designated component at each probe x_p; write $z_p := h(x_p) - a(x_p)$ for the exposed state, and put*

$$K_p := \{v \in \mathbb{R}^d : g(z_p - v) = g(z_p)\}, \qquad \mathcal{N}_0 := \{W : W\phi(x_p) \in K_p \ \forall p\},$$

$$K_{\mathrm{probe}} := \mathrm{span}\Big(\bigcup\nolimits_p K_p\Big),$$

the displacements at z_p that the full output cannot see, the redistributions confined to them, and the span of all of them. Then:

1. *f_W agrees with f on the* full output *$g(\cdot)$, not merely on a scalar measurement $m \circ g$, for every intervention in $\mathcal{I}_0$* if and only if *$W \in \mathcal{N}_0$. Agreement on the larger family $\mathcal{I}$ therefore implies $W \in \mathcal{N}_0$; the converse is not claimed, since interchange and mean ablations expose states other than z_p and invisibility there is a separate condition.*
2. *$\Omega \subseteq \mathcal{N}_0$. The inclusion can be strict, and can also be equality even when some K_p is nontrivial: if two probes share a feature vector and one of their kernels is trivial, membership forces the displacement to zero regardless of the other. We therefore claim the inclusion only, and exhibit strictness by example rather than characterizing it.*
3. *If $\Psi = m \circ g$ is translation-invariant along K_{probe} at the two held-out states, then covering the evaluation contrast, $u \in S$, leaves the designated gap unchanged for every $W \in \mathcal{N}_0$, so the two-sided form of Cor. 1 holds over $\mathcal{N}_0$ and not merely over Ω.*

Proof. Zeroing the designated component a at x_p exposes z_p under f and, since $a_W = a + W\phi$ and h is invariant (Lem. 1), $z_p - W\phi(x_p)$ under f_W. Full-output agreement at x_p is therefore precisely $g(z_p - W\phi(x_p)) = g(z_p)$, i.e. $W\phi(x_p) \in K_p$; ranging over p gives clause (1), the converse implication being immediate, and agreement on $\mathcal{I} \supseteq \mathcal{I}_0$ implies agreement on $\mathcal{I}_0$. Clause (2) holds because $0 \in K_p$ for every p. For clause (3), $u \in S$ means $u = \sum_p c_p \phi(x_p)$, so $Wu = \sum_p c_p W\phi(x_p)$ with each $W\phi(x_p) \in K_p \subseteq K_{\mathrm{probe}}$, hence $Wu \in K_{\mathrm{probe}}$, which as a subspace absorbs the scalings and the sum. The displaced held-out states are $\tilde{s}_a + Wu$ and $\tilde{s}_b - Wu$, and Ψ is invariant along K_{probe} at those states, so both carry their undisplaced values and $\Delta(W) = \Delta(0)$. □

Remark 2 (The invariance hypothesis is substantive)**.** Clause (3) needs Ψ invariant along the *span* of the kernels, not along a subspace they share, and the former does not follow from the latter: $K \subseteq K_p$ for every p says nothing about where an individual $W\phi(x_p) \in K_p$ lands. The distinction matters because K_{probe} can be large, as g fails injectivity for reasons beyond any single architectural symmetry, so the hypothesis is a real restriction and not a formality. It is satisfied when the kernels reduce to a common direction along which the readout is constant, which is the case Remark 3 exhibits.

We should be plain about how much clause (3) then buys, because in that instance the answer is little. Under centered LayerNorm the explicit redistributions we exhibit in $\mathcal{N}_0 \setminus \Omega$ are $\mathbf{1}$-displacements, and those are invisible to Ψ as well as to g, so they could not have steered Δ whether or not the contrast is covered. Clause (3) is therefore true and not load-bearing there: it neutralizes a class that was never a threat. It would do real work at a site where K_{probe} is strictly larger than $\mathrm{span}\{\mathbf{1}\}$ and Ψ is still invariant along it, and we exhibit no such site. Whether one exists is open, and until it is settled the operative guarantee remains Cor. 1 over Ω.

Remark 3 ($\mathcal{N}_0$ is strictly larger than Ω under centered LayerNorm)**.** The inclusion $\Omega \subseteq \mathcal{N}_0$ is proper for the models we study, and not for exotic reasons. In GPT-2 and Pythia every read of the residual stream is preceded by a *centered* LayerNorm, which subtracts the mean of its input, and the unembedding is preceded by a final one; hence $g(z + c\mathbf{1}) = g(z)$ identically, for every z and every $c \in \mathbb{R}$. So $\mathbf{1} \in K_p$ for all p, and $W = \mathbf{1}v^\top$ with $v^\top\phi(x_p) \neq 0$ lies in $\mathcal{N}_0 \setminus \Omega$: it preserves every probe outcome at the level of the *full logit vector*, not merely a scalar summary, while $W\Phi \neq 0$. Any statement of the form "full-output agreement forces $W \in \Omega$" is therefore false for Pythia-1B and GPT-2 Small alike, which is why we state the quotient version. The direction that carries the paper survives intact, and for the same reason the counterexample exists: $\mathbf{1}$-displacements are invisible to Ψ as well as to g, so they cannot steer Δ, and if the kernels reduce to $\mathrm{span}\{\mathbf{1}\}$ then K_{probe} does too and clause (3) applies. The argument is specific to *centered* normalization and does not transfer to RMSNorm architectures, which rescale without subtracting the mean, so $g(z + c\mathbf{1}) \neq g(z)$ there in

general; Gemma-2, used in §4 and App. C, is such a model. What the correction costs is the clean identification $\mathcal{N}_0 = \Omega$; what it buys is a statement true of the models it is asserted about.

Remark 4 (Why the strengthening is needed, and what it recommends)**.** The escape Cor. 1 guards against is a nonzero $W\phi(x_p)$ lying in a null direction of the readout. That escape is available only because m collapses g to a scalar: a displacement invisible to $m \circ g$ need not be invisible to g. Proposition 3 closes it by asking for agreement one level up. Full column rank of J_g at z_p is a convenient *sufficient* condition for the injectivity hypothesis: g maps into a higher-dimensional space, so full column rank makes g an immersion at z_p, injective on a radius the immersion theorem supplies. Under centered LayerNorm the condition must be read on the quotient modulo $\mathrm{span}\{\mathbf{1}\}$, where the reduced Jacobian can attain full rank: the shift invariance of Rem. 3 forces $J_g\mathbf{1} = 0$ on the ambient space, so ambient full column rank never holds there. It is not a characterization either way, since a map can be injective where its Jacobian is rank-deficient. The practical reading is a reporting recommendation rather than a theoretical nicety: a study that verifies probe agreement on its *summary statistic* has ruled out less than one that verifies it on the logit vector, and only the latter licenses the two-sided form.

A.1 Buying coverage: a probe-family design procedure

Proposition 2 and Proposition 3 pull in opposite directions, and the tension is the useful part. Full feature-space coverage is expensive for ordinary single-position probe families: under the canonical ϕ the annihilator has $\dim \Omega \geq d(d-P)$, so fewer than d independent probe directions cannot empty it, and a prompt family supplies rank one per prompt (a diverse-node patching family does reach the effective ceiling, App. D). The measured ceiling itself, $767 < d$ at tolerance 10^{-6} (§4), means natural probe features never spanned all of $\mathbb{R}^d$ at that tolerance in our runs, so emptying Ω outright was unattainable from natural inputs there. But coverage is not needed in the whole feature space. It is needed only in the directions the study's own evaluation contrasts occupy, and that subspace is small.

Proposition 4 (Targeted coverage, and what it costs)**.** *Let $U := \mathrm{span}\{u_1, \ldots, u_k\}$ be the span of the contrasts a study reports, with $\dim U = r$, and let probes be drawn from a candidate pool $\mathcal{X}$. Write*

$$P^\star(U, \mathcal{X}) \;:=\; \min\big\{|A| \;:\; A \subseteq \mathcal{X},\;\; U \subseteq \mathrm{span}\{\phi(x) : x \in A\}\big\}$$

for the smallest probe set that covers U, with $P^\star = \infty$ if no subset does. The pool may contain the evaluation endpoints themselves, in which case $P^\star \leq 2k$ trivially: that is the covered-contrast arm of §4, and the measured pools of App. A.2 exclude the evaluation pairs, which is what makes the budgets there nontrivial. Whenever $P^\star$ is finite and the probe set attains it, every reported contrast has $u_\perp = 0$, so by Cor. 1 no $W \in \Omega$ moves the designated effect, and under the hypotheses of Prop. 3 no redistribution in the family preserves the probe outcomes while moving it. The cost obeys

$$P^\star(U, \mathcal{X}) \;\geq\; r,$$

since $\dim S \leq |A|$, and the inequality can be strict.

Proof. Necessity is immediate from $\dim S \leq |A|$. For strictness, take $U = \mathrm{span}\{e_1\}$ with $r = 1$ and a pool whose features are $e_1 + e_2$ and $e_1 - e_2$: their span contains U, but neither alone does, so $P^\star = 2 > r$. □

Remark 5 (Why r is a lower bound and not the price)**.** The gap between r and $P^\star$ is a property of the candidate pool, not of the network: covering U requires probe features that span it, and a pool can contain U in its span while every individual feature lies off U. Two consequences follow. Greedy selection on $\|\mathsf{P}_{U \cap S_p^\perp}\phi(x)\|$ is *not* guaranteed to terminate at $U \subseteq S$, because $U \cap S_p^\perp$ can be trivial while $U \not\subseteq S_p$, as with $S_p = \mathrm{span}\{e_1 + e_2\}$ against $U = \mathrm{span}\{e_1\}$, so a covering set must be sought by a criterion that tracks $\|\mathsf{P}_{S_p^\perp} u_i\|$ for the reported contrasts directly rather than by projecting onto $U \cap S_p^\perp$. And $P^\star = r$ holds only under an additional condition on the pool, for instance that it contains r features already spanning U. We therefore report r as the intrinsic lower bound and $P^\star$ as the realizable cost, and do not claim the two coincide.

Three consequences are worth stating separately, because they are what a practitioner acts on.

The evaluation pairs need not be probed. Theorem 1 requires the evaluation pair to be held out, $x, x' \notin \{x_p\}$, which appears to make covering its contrast circular. It does not: covering

$u = \phi(x') - \phi(x)$ requires only that the probe features *span* that direction, not that they include either endpoint. A study can hold its evaluation pairs out and still cover the direction they contrast, using different prompts from the same family.

The intrinsic lower bound is the contrast-space dimension; the realizable cost depends on the probe pool. At least $r = \dim U$ probes are necessary, and for a single task family r is typically far below both d and the number of prompts available, but the attainable minimum $P^\star$ can exceed it and does so whenever no r pool features already span U. This is the sense in which "run more probes" is the wrong prescription and "cover the directions you report" is the right one: the first scales with a quantity that does not control exposure, the second with one that does, even though the second's price is set by the pool rather than by r alone.

The residual is measurable before any witness exists. At a budget short of covering U, $\max_i \|\mathsf{P}_{S^\perp} u_i\|$ is the reportable exposure of the uncovered remainder, computable on f alone and requiring no witness to be built.

A.2 Empirical difficulty of buying targeted coverage

We report what a natural targeted-selection heuristic achieves, on the GPT-2 name-mover pair with $\phi = \mathrm{id}$, against a pool of 1,076 candidates (676 IOI prompts plus 400 heterogeneous ones). The reported contrasts span $r = \dim U = 24$, and U lies *entirely* inside the pool's span: $\mathrm{rank}(\Phi_{\mathrm{pool}}) = \mathrm{rank}([\Phi_{\mathrm{pool}} \,|\, U]) = 767$, energy fraction 1.0000. So a covering set exists, and the question is its size.

It is far larger than r. Orthogonal matching pursuit against a single reported contrast, run to a budget of 400 probes, reduces $\|\mathsf{P}_{S^\perp} u\|$ by a median $7.8\times$, taking R from a median 16.3 to 2.07; 0 of 24 contrasts reach a residual of 2% of $\|u\|$, and 23 of 24 still reverse. The one contrast whose R crosses below 1 (at 0.94) is the one that stops. Covering the 24 contrasts jointly is harder still: 160 probes buy a $3.1\times$ reduction in $\max_i \|\mathsf{P}_{S^\perp} u_i\|$ and leave the reversal rate at $24/24$ throughout.

We are careful about what this does and does not establish. It does *not* measure $P^\star$: orthogonal matching pursuit failing to cover within 400 probes leaves open that some other subset of the pool, possibly a much smaller one, covers. Nor can a single experiment at $d = 768$ establish how the cost scales with d. What it establishes is that a natural targeted selection, run on the quantity the paper asks studies to report, remains far from coverage at budgets an order of magnitude above the intrinsic dimension $r = 24$. Task-matching is not the obstruction: adding 400 heterogeneous prompts to a 676-prompt IOI pool changed the achievable reduction by under 2%.

These results suggest that realizable coverage can be substantially more expensive than its intrinsic lower bound r; they do not identify the optimum $P^\star$. Coverage remains the correct criterion, and at this site we could not buy it. That is the mechanism behind Tab. 1's flat rows: at this site R begins around 16 and nothing we achieved moves the median below 1.

Remark 6 (Steering capacity is exactly characterized within Ω)**.** Every $W \in \Omega$ annihilates $\mathrm{span}\{\phi(x_p)\}$ row-wise, so $Wu = Wu_\perp$, and for any unit $e \in \mathbb{R}^d$ the rank-one lift $W_e := e\, u_\perp^\top / \|u_\perp\|$ lies in Ω with $\|W_e\|_{\mathrm{op}} = 1$ and $\langle W_e u, e\rangle = \|u_\perp\|$; by Cauchy–Schwarz this is extremal, so $\sup_{W \in \Omega,\, \|W\|_{\mathrm{op}} \le 1} \langle Wu, e\rangle = \|u_\perp\|$ *exactly, for every direction* e.

For Theorem 1 the leading scale is $\|G\|\|u_\perp\|$, determined by two measurable quantities: the transversal mass $\|u_\perp\|$ and the gap-gradient norm $\|G\|$, the latter obtained by autodifferentiation or finite differences through $\Psi = m \circ g$ directly, *not* through the projector. Whether the margin holds additionally depends on the finite-displacement remainder $\omega_\Psi(\|u_\perp\|)$, which $\|u_\perp\|$ and $\|G\|$ do not determine. Increasing q can enlarge $S^\perp$, but it need not increase the transversal mass of any particular evaluation contrast; that has to be measured, and App. D reports a configuration in which large algebraic slack coexists with transversal mass at essentially zero.

Note also that the condition is a *near-tie* requirement: the witness reverses pairs whose baseline gap falls below the steerable range $\|G\|\|u_\perp\|$. How far that reaches is measured in §4: over 360 held-out contrast–q cases the witness reverses 359, at baseline gaps up to 1.35 logits, and the sole exception is the only case with $R < 1$. At this site the remainder is small enough that $R > 1$ perfectly predicts reversal on that IOI grid, so the near-tie form of the condition is not the operative restriction it appears

to be; the Greater-Than replication of §4 tests both sides of the boundary, with 28 sub-threshold contrasts (none reversing) and two exceptions just above it.

The condition $|\Delta(0)| < \|G\|\|u_\perp\| - L_{\nabla\Psi}\|u_\perp\|^2$ is a sufficient f-only certificate in principle, provided $L_{\nabla\Psi}$ can be upper-bounded, which App. D.2 reports we cannot do by sampling. It must be evaluated at the realized displacement: maximizing $\|G\|r - L_{\nabla\Psi}r^2$ over a free radius r returns its value at $r^\dagger = \|G\|/2L_{\nabla\Psi}$. When $r^\dagger < \|u_\perp\|$ that displacement is realizable inside the unit ball, by $\alpha W^\star$ with $\alpha = r^\dagger/\|u_\perp\| < 1$, and evaluating at $\|u_\perp\|$ understates the certifiable margin; when $r^\dagger > \|u_\perp\|$ it certifies a margin at a displacement no witness with $\|W\|_{\mathrm{op}} \le 1$ realizes, since that bound caps the attainable displacement at $\|u_\perp\|$.

Corollary 3 (Attainable-set radii for the designated gap under a budgeted witness class)**.** *Fix the probe family, the evaluation pair (x', x) and a budget $\eta > 0$, write $\mathcal{I}_\eta := \{\Delta_\theta(f_W) : W \in \Omega,\ \|W\|_{\mathrm{op}} \le \eta\}$ for the attainable set of designated gaps, and put*

$$A_\eta := \eta\|G\|\|u_\perp\|, \qquad R_\eta := \omega_\Psi(\eta\|u_\perp\|), \qquad c_- := \big(A_\eta - R_\eta\big)_+, \qquad c_+ := A_\eta + R_\eta.$$

Under the hypotheses of Theorem 1 except the steering-margin inequality, with (H2) assumed on balls of radius $\eta\|u_\perp\|$ and Ψ continuous on the corresponding segments when $\eta > 1$,

$$\big[\,\Delta(0) - c_-,\ \Delta(0) + c_-\,\big] \ \subseteq\ \mathcal{I}_\eta\ \subseteq\ \big[\,\Delta(0) - c_+,\ \Delta(0) + c_+\,\big].$$

Consequently the designated ranking is not *identified when $|\Delta(0)| < c_-$, since a sign reversal is then attainable (at $|\Delta(0)| = c_-$ only a tie is guaranteed); when $|\Delta(0)| > c_+$ no admissible redistribution in Ω reaches zero, which certifies robustness* against this witness class *and not identification, since Ω-membership is sufficient but not necessary for probe agreement (proof of Thm. 1) and η truncates the class further; and for $c_- \le |\Delta(0)| \le c_+$ these bounds decide nothing. Note $\mathcal{I}_\eta$ ranges over Ω truncated at η, while by Prop. 3 the probe-consistent redistributions are $\mathcal{N}_0 \supseteq \Omega$, untruncated; so c_+ bounds a* subset *of the evidence-consistent set rather than containing the identified set, and is not an outer bound in Manski's sense.*

Proof of Corollary 3. Outer: for $W \in \Omega$ with $\|W\|_{\mathrm{op}} \le \eta$, $Wu = Wu_\perp$ and $|\langle G, Wu\rangle| \le \|G\|\|Wu_\perp\| \le A_\eta$, while the remainder is at most R_η since ω_Ψ is nondecreasing. Inner: if $A_\eta \le R_\eta$ then $c_- = 0$ and the inclusion is trivial, since $W = 0$ attains $\Delta(0)$. Otherwise put $\hat{g} := G/\|G\|$ and $W_t := t\,\eta\,\hat{g}\,u_\perp^\top/\|u_\perp\|$ for $t \in [-1, 1]$; each W_t lies in Ω with $\|W_t\|_{\mathrm{op}} \le \eta$, and $W_{\pm1}u = \pm\eta\|u_\perp\|\hat{g}$ gives $\Delta(W_{+1}) \ge \Delta(0) + c_-$ and $\Delta(W_{-1}) \le \Delta(0) - c_-$. Since $t \mapsto \Delta(W_t)$ is continuous, the intermediate value theorem attains every value between. □

B The Impossibility Construction and the E3 Protocol

B.1 Component redistribution

Fix a site ℓ at which two components write additively into the residual stream: $h(x) = c(x) + a(x) + b(x)$, where c is the incoming stream and a, b are the component writes (in a transformer, the attention and MLP block outputs at layer ℓ; the construction generalizes to any additive decomposition).

Architectural scope. The construction is stated for *parallel-write pre-norm* transformer blocks (PaLM, GPT-J, GPT-NeoX convention), in which attention and MLP both read from the residual stream entering the block and write back into it additively in parallel: $h_{\mathrm{out}} = h_{\mathrm{in}} + a(h_{\mathrm{in}}) + b(h_{\mathrm{in}})$. In this setting the cancellation in Lemma 1 is exact: $(a + W\phi) + (b - W\phi) = a + b$ pointwise in $\mathbb{R}^d$, because the MLP's input does not depend on the attention output.

The construction does *not* apply verbatim to *sequential* pre-norm blocks (the original GPT-2 / Vaswani convention, where MLP reads $h_{\mathrm{in}} + a$): there, the additional $+W\phi$ in the attention write modifies the MLP's input, and the MLP's nonlinearity prevents the subsequent $-W\phi$ from restoring identity at the output. Extending the *block-level* construction to sequential blocks requires a more elaborate redistribution that compensates for the MLP's nonlinear response, which we leave to future work.

This restriction is on granularity, not on architecture: at *head* granularity two heads in one attention layer read the same layer-normed input and write additively into the same stream in standard multi-head attention, including every model studied here, so the construction applies unchanged to sequential pre-norm models, which is what licenses the GPT-2 Small experiments of §4 (Rem. 1).

Definition 2 (Redistributed network)**.** Let $\phi : \mathcal{X} \to \mathbb{R}^q$ be a feature map computable from information available upstream of both components (formally, a measurable function of the stream entering layer ℓ), and let $W \in \mathbb{R}^{d\times q}$. The redistributed network f_W has writes $a_W(x) = a(x) + W\phi(x)$ and $b_W(x) = b(x) - W\phi(x)$, all other parameters unchanged.

Remark 7 (The redistributions compose across more than two components)**.** Stated for a pair, the construction extends to any k siblings at a shared write point by telescoping. Give C_1 a displacement $+W_{12}\phi$, give C_2 the pair $-W_{12}\phi + W_{23}\phi$, and so on to C_k with $-W_{k-1,k}\phi$; every interior term cancels against its neighbor and the sum of the writes is unchanged pointwise, exactly as in Lem. 1. Each $W_{i,i+1}$ may be chosen independently in Ω, so the gauge group over k siblings has dimension $(k-1)\dim\Omega$ rather than $\dim\Omega$. Probe invisibility is inherited, since each summand annihilates Φ, and the freedom available to permute a *ranking* over k components, which is the object a circuit analysis reports, is correspondingly larger than the freedom to reverse one designated pair. The freedom is not vacuous in practice: §4 reports a measured $k = 3$ instance whose chain, at annihilation residuals inside the pairwise criterion, exactly inverts a held-out three-head ordering at twice unit operator norm (protocol in App. B.3); the same chain fails the annihilation criterion at $k = 4$, so behavior at larger k remains open.

Proof of Lemma 1 (stated in §3).

Proof. $c(x) + a_W(x) + b_W(x) = c(x) + a(x) + b(x) = h(x)$ pointwise, where the cancellation $+W\phi(x) - W\phi(x) = 0$ uses that ϕ is measurable with respect to the stream entering layer ℓ (Definition 2), so both writes evaluate ϕ on the same argument x via the same upstream activations; if ϕ were a function of the post-block residual, the two evaluations would differ and the cancellation would fail. Everything downstream is then a function of $h(x)$. □

Lemma 2 (What interventions expose)**.** *(i) Zero-ablating component a on input x yields downstream input $h(x) - a(x) - W\phi(x)$ in f_W versus $h(x) - a(x)$ in f: the probe responses differ by the action of g along $-W\phi(x)$ at the probe state. (ii) Interchanging component a from source x' into base x yields $h(x) - a(x) + a(x') + W(\phi(x') - \phi(x))$ in f_W: component-level counterfactual states differ by $W(\phi(x') - \phi(x))$, so the resulting effect estimates differ too. Mean ablation replaces the write $a(x) + W\phi(x)$ by its reference mean, exposing $W(\overline{\phi}_R - \phi(x))$ rather than case (i)'s $-W\phi(x)$.*

Proof. Direct computation from Definition 2. □

Remark 8 (Stream-level claims are gauge-invariant; stream-level evidence buys no coverage)**.** Three consequences of Lemma 1 and Prop. 1 place the evidence styles that operate on the summed stream. (i) Any claim measurable with respect to the residual stream and the readout, an alignment subspace of h and its interchange-intervention accuracy included, takes identical values under f and f_W for every $W \in \Omega$ at every input; the authority is Lemma 1, which makes h identical pointwise, with case (iv) of Prop. 1 covering interventions on the summed stream. Such claims are *identified relative to this witness class*, and counterfactual-label supervision changes nothing, the labels being themselves functions of those outcomes. Relative is the operative word: a stream-level claim is untouched by *this* ambiguity while remaining subject to the separate freedom in the alignment map itself (Sutter et al., 2025), so invariance here is not identification. (ii) For the same reason, stream-level interventions contribute no columns to Φ: they never decompose the designated pair, so they add no coverage to any component-ranking claim riding on them, however accurate the alignment. (iii) An alignment map trained on one component's write decomposes the pair and joins direct logit attribution in the exposing class; so does a resampling scheme that gives the two writes different inputs, treeified scrubbing included, by the edge-patch mechanism of App. D.

Gradients require the same distinction. A gradient of the readout is unchanged: g is untouched and every state is identical under f and f_W (Lemma 1), so $\nabla\Psi$ agrees wherever it is evaluated, and total derivatives of f_W agree with those of f because the two functions coincide. A gradient taken *through one pair member's write* does not: since $a_W = a + W\phi$, its Jacobian is $\partial a/\partial z + W\,\partial\phi/\partial z$ against $\partial a/\partial z$ under f, and the two members' Jacobians differ individually while cancelling in sum. Per-component or edge-level gradient attribution therefore separately exposes a pair member and belongs in the exposing class, and the constraint it imposes, a two-sided contraction $\langle\nabla\Psi, W\delta\rangle = 0$ against an observed upstream displacement δ, is not of the row form $W\Phi = 0$ that defines Ω. The

unit of analysis, stream variable or component write, thus determines whether this ambiguity cannot be constructed or is simply not probed.

Proof of Theorem 1. Take $f^{(1)} := f$ and $f^{(2)} := f_W$ for $W \in \mathbb{R}^{d\times q}$ to be chosen. Clause (1) of the theorem holds for every W by Lemma 1. For clause (2), Lemma 2 gives that a *sufficient* condition for exact probe agreement is $W\phi(x_p) = 0$ for every p: under this condition zero ablation exposes no redistribution term, and within-$\mathcal{I}$ interchange exposes $W(\phi(x_{p'}) - \phi(x_p)) = 0$ as a consequence of the zero condition on both terms. We do not claim necessity, and none is needed: a nonzero $W\phi(x_p)$ could in principle lie in a null direction of the downstream readout and still produce equal measured responses. Mean ablation against a reference set R exposes $W\overline{\phi}_R - W\phi(x_p)$, matching the sign convention of Def. 2; the second term vanishes by the row constraint. Since $W \in \Omega$ annihilates $\mathrm{span}\{\phi(x_p)\}_{p=1}^P$ row-wise (each row of W lies in the orthogonal complement of that span), we have $W\overline{\phi}_R = 0$ for any $\overline{\phi}_R \in \mathrm{span}\{\phi(x_p)\}_{p=1}^P$ (linear-combination membership suffices, not just convex combination). The condition holds in particular when $R = \{x_p\}$ or $R \subseteq \{x_p\}$, which is the operative scoping in the empirical instantiation (Section 4). For external reference sets violating this condition, mean ablation must be excluded from $\mathcal{I}$. The system

$$W\phi(x_p) = 0 \quad \forall p \in \{1, \dots, P\}$$

is linear in W and decomposes row-wise: each row w_i of W must lie in $S^\perp \subset \mathbb{R}^q$, where $S := \mathrm{span}\{\phi(x_p)\}_{p=1}^P$. By transversality, S is a proper subspace of $\mathbb{R}^q$, so $S^\perp$ is nontrivial and the constraint set $\Omega := \{W : W\phi(x_p) = 0\ \forall p\}$ is a linear subspace of $\mathbb{R}^{d\times q}$ of positive dimension.

For clause (3), fix the transversal evaluation pair (x', x) with $u := \phi(x') - \phi(x)$ and $u_\perp \neq 0$ its component orthogonal to $\mathrm{span}\{\phi(x_p)\}$. By Lemma 2(ii), the held-out interchange on component a at (x', x) produces evaluation state $\tilde{s}_a$ in f and $\tilde{s}_a + Wu$ in f_W; the same interchange on b produces $\tilde{s}_b$ and $\tilde{s}_b - Wu$ (the redistributed write enters the two components with opposite signs, Definition 2). With $\Psi := m \circ g$ the readout, literally the same function for f and f_W since the redistribution leaves g untouched (Lemma 1), the implemented interchange effects are $\theta_W(a) = \Psi(\tilde{s}_a + Wu) - \Psi(h(x))$ and $\theta_W(b) = \Psi(\tilde{s}_b - Wu) - \Psi(h(x))$, so the designated gap is

$$\Delta(W) := \theta_W(a) - \theta_W(b) = \Psi(\tilde{s}_a + Wu) - \Psi(\tilde{s}_b - Wu), \qquad \Delta(0) = \theta_f(a) - \theta_f(b).$$

Choice of $W^\star$. Set $\hat{e} := -\mathrm{sign}(\Delta(0))\, G/\|G\|$ with $G = \nabla\Psi(\tilde{s}_a) + \nabla\Psi(\tilde{s}_b)$ as in the hypothesis, and $W^\star := \hat{e}\, u_\perp^\top / \|u_\perp\|$. Then $W^\star \in \Omega$ (each row is proportional to $u_\perp$, which is orthogonal to every $\phi(x_p)$), $\|W^\star\|_{\mathrm{op}} = 1$, and $W^\star u = W^\star u_\perp = \|u_\perp\|\, \hat{e}$; by Remark 6 this choice is extremal over Ω. The expansion below therefore needs only differentiability of Ψ at $\tilde{s}_a, \tilde{s}_b$ with locally Lipschitz derivative, which the hypothesis supplies; no reach, tube containment, or projector regularity is used. First-order expansion of Ψ at $\tilde{s}_a$ and $\tilde{s}_b$ then gives

$$\begin{aligned}\Delta(W^\star) &= \Delta(0) + \|u_\perp\| \left\langle \nabla\Psi(\tilde{s}_a) + \nabla\Psi(\tilde{s}_b),\, \hat{e} \right\rangle \pm \omega_\Psi(\|u_\perp\|) \\ &= \Delta(0) - \mathrm{sign}(\Delta(0))\, \|G\|\, \|u_\perp\| \pm \omega_\Psi(\|u_\perp\|).\end{aligned}$$

The steering-margin hypothesis $|\Delta(0)| < \|G\|\|u_\perp\| - \omega_\Psi(\|u_\perp\|)$ forces $\mathrm{sign}(\Delta(W^\star)) = -\mathrm{sign}(\Delta(0))$: the designated ranking reverses, with post-flip gap at least $\|G\|\|u_\perp\| - \omega_\Psi(\|u_\perp\|) - |\Delta(0)| > 0$. Clauses (1)–(2) are exact for every $W \in \Omega$ by Lemma 1; clause (3) holds with the explicit post-flip bound displayed above, which is strictly positive but need not be a constant fraction of $\|G\|\|u_\perp\|$, since the margin hypothesis requires only strict inequality. The sibling-write condition this argument uses is exactly the scope stated in Rem. 1 and the architectural-scope discussion above: parallel writes at a shared point, which head pairs satisfy in any attention layer and attention-versus-MLP pairs satisfy only in parallel-write blocks. □

B.2 Invalidated E3 pipeline and artifact audit

The E3 pipeline used during development is invalidated in full and none of its numerical results are evidence for any claim in this paper. We record the defects and the diagnostics that establish them; all downstream values from those pipelines are retained in the artifact bundle as archival material only. Two defects invalidated the original pipeline. A third, of the same kind, was found later in the replacement and is recorded here with the others rather than separately: all three are activation-site errors, two reads placed where the construction is invisible and one redistribution applied off its

specified single-site support, which is the failure mode this construction is most exposed to, and we think the pattern is more informative to a reader than any one instance.

Defect 1: the probe instrumentation, and hence the training objective, was blind to the construction. The helper computing the S2 probe-match measurement evaluated the first three entries of the probe list, `blocks.{0,1,2}.hook_attn_out`, which sit upstream of the layer-8 redistribution site. Execution has not yet reached the redistribution site when those activations are computed, so they are identical for every W and the test could not fail. The signature is visible in the archive: the recorded probe "KL" medians are *negative* (e.g. -8.65×10^{-9}), which a divergence cannot be, and are floating-point reduction-order noise between two numerically identical passes. The same helper supplied the probe term of the training objective, so $\mathcal{L}_{\mathrm{probe}}$ carried no gradient: the reported objective $20\,\mathcal{L}_{\mathrm{probe}} + \mathcal{L}_{\mathrm{rank}}$ was in effect $\mathcal{L}_{\mathrm{rank}}$ alone, and the trained W was never pressured toward annihilation. This accounts for the recorded annihilation residual $\|W_{\mathrm{opt}}\Phi\|_F = 889.9$ against a pseudoinverse solution's 2.1×10^{-3}. It also voids that pipeline's scaling criterion (internally S2b), which regressed that same noise against $\|W\phi\|$ and therefore carries no information about the scaling it was written to check.

Defect 2: the redistribution support did not match the protocol. The implementation applied $W\phi(h_t)$ at every token position, whereas the protocol specifies a single final-token site. The two are not interchangeable: a position-wise probe family exposes ϕ at every position, and the resulting feature matrix saturates the feature space, leaving $\Omega = \{0\}$ and no non-trivial witness at all.

Defect 3, found after the replacement protocol had produced results: the interchange read the source component from the wrong network. An interchange performed on f_W must install the source component as f_W computes it, $a_W(x') = a(x') + W\phi(x')$. Both the block-level gap routine and its head-level counterpart cached that component *before* installing the redistribution hooks, so they interchanged $a(x')$ into a base run that did carry the redistribution. The realized held-out states were then $\tilde{s}_a - W\phi(x)$ and $\tilde{s}_b + W\phi(x)$ rather than the theorem's $\tilde{s}_a + Wu$ and $\tilde{s}_b - Wu$, a discrepancy of exactly $\mp W\phi(x')$.

The defect is invisible on probes, since $W\phi(x_p) = 0$ by construction, so probe agreement, the annihilation residuals, the calibrated divergence ceiling and the positive control were unaffected and are not revised. On held-out pairs it is an $O(1)$ error: the median discrepancy in $\Delta(W)$ was 1.04 on gaps of order one, and the sign differed on 20 of 24 contrasts. Because the displacement it produced was still probe-invisible, reversals still occurred and the measurement did not look broken; it understated the witness, reporting 6 of 24 reversals where the correct figure is 24 of 24.

It was found by a self-test rather than by inspection. The decomposition $\Delta(W^\star) = \Delta(0) - \mathrm{sign}(\Delta(0))\|G\|\|u_\perp\| + (\rho_a - \rho_b)$ is an algebraic identity under Lemma 1, so evaluating Ψ at the displaced states and comparing against the gap routine tests the routine rather than the theory; the residual came back at 1.04 rather than at floating-point noise. Every measurement downstream of the gap routine was re-run after the fix, and all figures and tables in this paper report post-fix values; the pre-fix outputs are retained in the artifact bundle under their original filenames alongside the corrected ones. Three checks now guard the semantics: a regression test asserting that the gap routine reproduces $\Psi(\tilde{s}_a + Wu) - \Psi(\tilde{s}_b - Wu)$ and that a stale-source variant does not, a comparison against states written directly at the post-site residual (agreement 3.5×10^{-5}), and a recomputation of $\Delta(0)$ and $\Delta(W^\star)$ through PyTorch module hooks sharing no code with the library hook path used to produce the reported numbers (agreement exact on all 24 contrasts, flip decision identical on all 24).

B.3 Replacement E3 producing protocol

The replacement builds annihilation into the parameterization rather than into a penalty, so no implementation error can silently relax it.

Site. Layer 8 of 16 on Pythia-1B (fp32), final token position; the site is the pair (ℓ, t) of Section 2. Zero, mean, and interchange probe operators all act site-locally on the layer-8 attention and MLP writes at that position.

Feature map and probe span. ϕ is the projection of the layer-8 input residual at the final token onto the top q PCA directions, fitted once on separately collected references and frozen. $\Phi = [\phi(x_1) \cdots \phi(x_P)] \in \mathbb{R}^{q \times P}$ over the P training probes; $S = \mathrm{col}(\Phi)$; $\mathrm{rank}(\Phi)$ is taken at relative SVD tolerance 10^{-6}; N is an orthonormal basis of $S^\perp$ from the same decomposition.

Parameterization. $W = AN^\top$ with $A \in \mathbb{R}^{d \times \dim S^\perp}$, so $W\Phi = AN^\top\Phi = 0$ identically for every A. Numerical annihilation tolerance: $\|W\Phi\|_F \le 10^{-4}$.

Arms, sharing site, probe family, intervention implementation and evaluation pairs: (i) W_{bad}, deliberately outside Ω, as a positive control that the corrected probe-agreement outcome S2 below detects exposed redistribution; (ii) the theorem's rank-one witness $W^\star = \hat{e}\, u_\perp^\top / \|u_\perp\|$ at unit operator norm, with G obtained by autodifferentiation through $\Psi = m \circ g$; (iii) A optimized for ranking reversal within Ω, by Adam ascent on the signed designated interchange gap driven against $\mathrm{sign}(\Delta(0))$ (learning rate 5×10^{-3}, 300 steps, the protocol script's defaults), with only the *direction* optimized and the result renormalized to $\|W\|_{\mathrm{op}} = 1$, since $W = AN^\top$ annihilates Φ for arbitrarily large A and an unconstrained optimum would reverse rankings by magnitude alone; (iv) $W_{\mathrm{rand}} = A_r N^\top$ with A_r random at the same operator norm, as the null distinguishing direction from magnitude.

Outcomes. S1, residual-stream identity, reported as an implementation check rather than a threshold test, since Lemma 1 makes it exact for every W. S2, probe agreement, measured on the full probe family and separately on the two site-crossing probes, the zero ablations of the attention and MLP writes at the redistribution site itself. S3, the primary outcome, is the *raw held-out interchange ranking* of Theorem 1, with no projection. The primary statistic is $\widehat{\alpha}_{\min} = \min\{\alpha \in \mathcal{A} : \mathrm{sign}\,\Delta(\alpha W) \neq \mathrm{sign}\,\Delta(0)\}$, a smallest *tested* grid value rather than a continuous infimum. The grid $\mathcal{A}$ is $\{0.25, 0.5, 0.75, 1, 1.5, 2, 3, 5, 10, 30, 100\}$ for the four arms. It was fixed before the arms were run and was not revised afterwards, so the arm table is a grid-frozen measurement. The sweep grid is $\{0.05, 0.1, 0.15, 0.25, 0.4, 0.6, 0.8, 1, 1.5, 2, 3, 5, 10, 30\}$ with bisection to ± 0.02, and we state its provenance plainly: it was chosen *after* the arm run, because two of the six reversing contrasts there sat exactly at the grid floor of 0.25 and were therefore censored from below, which attenuates any correlation computed from them. That count of six is the pre-correction figure withdrawn in App. B.2; the censoring it revealed is a property of the grid rather than of the defect, so the motivation survives the correction, but we flag the provenance. The corrected $W^\star$ arm likewise records $\widehat{\alpha}_{\min}$ at that floor, so its 0.25 is an upper bound on the crossing rather than a located value. The extension is a resolution change to the same estimator and moves $\widehat{\alpha}_{\min}$ only downward; it was not selected by inspecting sweep outcomes, which had not been run. The unit-norm criterion, which is what the theorem's claim rests on, is unaffected by grid resolution. Because the response need not be monotone in α, the sign at $\alpha = 1$ is checked directly rather than inferred from $\widehat{\alpha}_{\min} \le 1$; the remainder of the grid is secondary dose-response evidence, since a reversal at large α would not test a unit-norm construction.

Two controls on the witness arm. The arm table reports one random draw, which cannot separate direction from magnitude, and the coverage prediction of Cor. 1 needs a direct test. Both controls run on Pythia-1B layer 8, $q = 256$, fp32, seed 42, and are archived in `results/missing_controls_q256.json`. *Covered contrast.* The probe family is the standard 32 probes plus both endpoints of all 24 evaluation pairs, 80 in total, so that $u = \phi(x') - \phi(x)$ lies in S by construction; this deliberately violates the theorem's held-out requirement, which is the point. Because $q = 256$ the family does not saturate the feature space, and $\dim S^\perp = 176$ remains: the test is coverage with the annihilator intact, not coverage by exhaustion. The grid runs to $\alpha = 10$, ten times the norm at which the claim is stated. *Random null.* For each of 8 held-out contrasts we draw 50 matrices $W_r = A_r N^\top$ with A_r standard normal, renormalized to the witness's operator norm, and record the smallest grid α at which each reverses. Both controls were specified before either was run, and the covered-contrast control is a falsification test of Cor. 1: any reversal there would contradict it.

Component-scale normalization of the budget. Operator norm does not by itself locate the witness on the scale of the writes it perturbs, so we measure it on the head-level pair at $P \in \{128, 512\}$. The added write $\|W^\star\phi(z)\|$ at the two evaluation endpoints is a median 0.02 and $0.79\times$ ($P = 128$), and 0.003 and $0.26\times$ ($P = 512$), the smaller of the two heads' own write norms, and the witness's held-out displacement $\|u_\perp\|$ is a median 0.12 to $0.66\times$ the head-write difference the interchange itself installs; on the probes the added write is 10^{-14} (algebraic). Rerunning the held-out test with the budget capped in component terms, $\|\alpha W^\star\phi(z)\| \le \beta \min(\|a(z)\|, \|b(z)\|)$ at both endpoints, all 24 of 24 contrasts reverse at every $\beta \in \{0.25, 0.5, 1\}$ at both probe counts (pre-committed reading, written before the run: ≥ 12 of 24 at $\beta = 0.5$, $P = 512$). The cap fixes the component-relative size, not α: at $P = 128$, $\beta = 0.25$ the realized α is 0.19–0.51, so all 24 reversals satisfy the component cap and unit operator norm *jointly*; at $P = 512$, $\beta = 0.25$ the realized α is 0.46–1.44 (16 of 24 inside $\alpha \le 1$, all 16 reversing), and equivalently the unit-norm reversals of Tab. 1 realize added writes of

0.17–0.55× component scale (per contrast, $\max_z \|W^\star\phi(z)\| / \min_z \min(\|a(z)\|, \|b(z)\|)$ over the two evaluation endpoints $z \in \{x, x'\}$) with 24 of 24 reversing; the $\beta \in \{0.5, 1\}$ cells at $P = 512$ require α up to 2.9 and 5.8 and sit above the theorem's unit-norm budget, reported as scale statements only. At $P = 512$ the unit-norm reversals therefore operate at a fraction of the components' own scale, not by outsized displacement. At $P = 128$ they do not: the same quantity runs 0.49–1.31× and the contrast-relative ratios reach 1.61×, so at the smaller probe count the added write matches or exceeds the smaller head's own write on some contrasts. We report the medians and the maxima rather than describing either count as typical. GPT-2 Small, $\phi = \mathrm{id}$, $q = 768$, MPS fp32, seed 42; archived in `results/witness_scale_audit_L9.json`.

Second architecture. The probe-family sweep is repeated on GPT-J-6B (fp32) at layer 14 of 28, final token, with the same feature map, the same nested-family design and the same annihilation tolerance; $q = 256$, pool seed 1, 24 held-out contrasts. GPT-J is a parallel-write pre-norm model, so the block-level construction applies without the head-level restriction of Section 4. Only the sweep is run there, not the four-arm grid: it is a replication of the exposure-versus-coverage shape on a second architecture at three times the width, not an independent producing experiment. It is also the cleaner dose-response of the two, because its transversal mass is larger at every probe budget: as $\dim S^\perp$ falls from 248 to 1, median $\|u_\perp\|$ falls from 7.55 to 0.44 and median $\widehat{\alpha}_{\min}$ rises monotonically from 0.05 to 0.575, still inside unit norm at every point. Pythia-1B enters the same regime with $\|u_\perp\|$ an order of magnitude smaller and therefore crosses out of unit norm, which is why the two models separate only at the last cell.

Evaluation pairs. Held out in two senses. They lie outside the probe set, as the theorem requires, *and* they are drawn from a slice of the prompt pool never examined under the invalidated pipeline, so no outcome seen during development informed their selection. Every fresh candidate is evaluated; we do not select one. The pre-registered ordering is $R = \|G\|\|u_\perp\|/|\Delta(0)|$, computed on f alone and never involving f_W, and the theorem's prediction is that $R > 1$ is the favorable regime. Reporting a single $\arg\max R$ pair would systematically select the nearest tie, since $|\Delta(0)|$ sits in the denominator.

Probe-count sweep. Model, layer, feature dimension ($q = 256$) and evaluation contrasts are held fixed; only the probe family grows. Families are *nested*, so S is monotone in P by construction and the curve cannot be an artifact of redrawing probes at each P. The 24 evaluation contrasts are taken from the pool *after* the largest probe family, so the held-out property holds simultaneously at every P rather than being re-established per point. Three pool seeds permute the deduplicated pool before splitting, redrawing probe family and contrasts together; a probe/evaluation overlap assertion runs at every seed. Each cell recomputes the theorem's rank-one witness from the nullspace of that family's Φ, so the sweep compares witnesses at matched operator norm rather than rescaling a single one. Reported per row: the pooled reversible count over the three seeds, the pooled median $\widehat{\alpha}_{\min}$ over contrasts with a finite value, and the worst annihilation residual over all cells in the row.

A criterion change we record explicitly. The pre-specified rule excluded the $P = 256$ row, on the grounds that $\mathrm{rank}(\Phi) = 255$ leaves a one-dimensional $S^\perp$ whose residual reaches 2.8×10^{-4}, above the algebraic criterion, making the witness unverifiable rather than ineffective. We later determined that the rule as applied conflated two different things. Recomputing the contrast projection in double precision separates them: at every cell with $\dim S^\perp \geq 64$ the residual falls by one to eight orders of magnitude, so there it was floating-point accumulation and the cells pass with room to spare; at $\dim S^\perp = 1$ it does not move at all, because the single basis direction is the last left-singular vector of Φ, whose singular value lies below the 10^{-6} relative rank tolerance but is not zero, so $\|W\Phi\|_F$ equals that singular value and the direction annihilates the probe span only approximately. The excluded row was therefore not a numerical artifact to be discarded but a measurement of how nearly the probe span covers the feature space. We report it, marked separately in Tab. 6 and never pooled with the exactly-annihilating cells, and we flag the change rather than presenting the revised rule as the original one. The mechanism claim of §4 draws on these two cells together with the GPT-2 cells of Tab. 2 at $\dim S^\perp = 1$, giving three points across three models with $\|u_\perp\|$ spanning 0.44, 0.033 and 0.000 and exposure monotone in it. The three differ in model, feature map and probe construction, so we read it as a consistent cross-model pattern rather than a controlled comparison; a controlled version would hold everything but $\|u_\perp\|$ fixed, which we did not run.

Head-level protocol (GPT-2 Small). The head-level runs of Tab. 1, Tab. 2 and the joint witness redistribute $\pm W\phi$ between two heads of one attention layer, both reading the same layer-normed input, so Lemma 1 applies without the parallel-write restriction. Feature map: canonical $\phi = \mathrm{id}$

on the layer-9 `resid_pre` at the final token, $q = d = 768$. Probe families are nested over a 586-prompt IOI pool (15 templates); evaluation contrasts are 24 held-out pairs disjoint from the largest family. Four head-level pools are distinct: this 586-prompt nesting pool, the 842-prompt extension pool for the $P \in \{640, 768\}$ cells of Tab. 1 (the name-pair grid exhausts near that size), the full 1,576-prompt task pool used only for the rank-ceiling measurement, and the mixed 1,076-prompt candidate pool of the covering experiment (App. A.2). The Greater-Than replication (Tab. 7) runs this same head-level protocol on its own 586-prompt Greater-Than pool, 20 nouns and start years 12–88; the equality with the IOI nesting-pool size is coincidental, both builders requesting $P_{\max} + 24 + 50$ base prompts (the largest probe family, the 24 evaluation bases, whose counterfactual sources are generated per prompt rather than drawn from the pool, and 50 unused reserve). The nullspace basis N is computed in double precision, which is why these runs annihilate to $\leq 4.7 \times 10^{-13}$ (Tab. 6 caption); the same grid and unit-norm criterion as above apply.

Write-space realizability, measured. A head's write is confined to the column space of its output matrix, of dimension $d_{\text{head}} = 64$ against $d = 768$, so a displacement absorbable into *both* heads' own weights would have to lie in the intersection of two 64-dimensional subspaces. We measure it rather than assert it: the intersection is trivial for all three reported pairs; for L9H9/L9H6 the minimum principal angle is $49.5°$ (largest principal-angle cosines 0.650, 0.703, and 0.903 across the three pairs), archived in `results/head_level_witness_gpt2_analytic.json`. Two consequences, in opposite directions. The added write is a genuine re-decomposition rather than a disguised weight edit, which is what makes the witness non-trivial. But it also means a_W and b_W are not themselves attention-head outputs, so at head granularity an analyst who requires each write to be realizable by its own head faces no admissible nonzero redistribution at all, and the singular spectrum of one head's write over a prompt pool separates the two decompositions without consulting any weights (rank 64 under f, one more direction under f_W). This is the concrete form of the concession in §3: the head-level ambiguity is resolved by the architectural decomposition, and it is the dictionary granularity of App. D.1, where no basis is privileged, that carries the identification argument.

Joint witness, closed form. For contrasts $u_1, \ldots, u_k$ with gap gradients $G_1, \ldots, G_k$, the aggregate gap is linear in W to first order with coefficient $M := \sum_i G_i u_{i\perp}^\top$. Maximizing $\langle W, M \rangle$ over $\{W = AN^\top : \|W\|_{\text{op}} \leq 1\}$ is the nuclear-norm dual: with the SVD $MN = U\Sigma V^\top$ taken in its thin form (the full-form factors do not compose to $UV^\top$ when $\dim S^\perp < d$), the optimizer is $W_{\text{joint}} = -\text{sign}(\overline{\Delta(0)})\, UV^\top N^\top$, where $\overline{\Delta(0)}$ is the mean baseline gap over the fitting contrasts, required nonzero; the rows of W_{joint} lie in $S^\perp$ so $W_{\text{joint}} \in \Omega$ automatically. No optimizer and no tuned hyperparameter: one SVD. Fit on 24 contrasts at $P = 128$; evaluated on those and on a disjoint 24 from the same pool; frozen grid $\alpha \in \{0.25, 0.5, 1, 1.5, 2, 3, 5\}$ with the sign at $\alpha = 1$ checked directly. The template-shift arm repeats the construction with probes ($P = 128$) and 24 fitting contrasts drawn only from templates 1–10 and 24 held-out contrasts drawn only from templates 11–15, which appear nowhere in the probe family or the fit; same grid and criteria, archived in `results/joint_witness_template_shift_L9_h9h6.json`. One provenance note, since the archived payloads record it and a reader checking them will meet it first: for both arms the residual evaluated in single precision is above the criterion (2.4×10^{-4} for the fit arm, 2.9×10^{-4} for the template-shift arm). As for every head-level run of §4, the criterion is assessed against the double-precision nullspace basis, where the residuals are 2.0×10^{-12} and 2.3×10^{-12}; the single-precision figure is floating-point accumulation in re-promoting a basis stored in `fp32`, not structural leakage, the same mechanism the † cells of Tab. 6 document. The template-shift payload archives its double-precision residual; the fit arm's was recorded in the run log rather than the payload.

Composed (k-sibling) witness. The $k = 3$ chain of §4 assigns head $m \in \{1, 2, 3\}$ the displacement $-W_{m-1}\phi + W_m\phi$ with $W_0 = W_3 = 0$ (Rem. 7); heads (L9H9, L9H6, L9H0), canonical $\phi = \text{id}$ on the layer-9 `resid_pre` at the final token, $q = d = 768$. Probes are the first 128 pool prompts ($\text{rank}\,\Phi = 128$, $\dim S^\perp = 640$), disjoint from the 24 fitting and 24 held-out contrasts (asserted programmatically). The per-head score is the head's interchange effect on the signed logit difference at the final token, averaged over contrasts. The chain is fitted on the fitting contrasts only, by Adam ascent (300 steps, learning rate 0.05) on a soft minimum of the adjacent-pair margins of the target ordering, $-\frac{1}{\tau}\log\sum_j \exp(-\tau\,\text{margin}_j)$ at temperature $\tau = 8$, computed on the first-order predicted scores; each step projects every W_m onto Ω (in double precision) and onto the operator-norm ball of radius α; the chain is refitted at each $\alpha \in \{0.5, 1, 2, 4\}$, so α is the per-pair operator-norm budget rather than a scaling of one fixed witness, and an interior head's net displacement

operator $-W_{m-1} + W_m$ can reach operator norm 2α. Residuals $\max \|W\Phi\|_F$: 6.7–7.8 $\times 10^{-5}$ in fp32, 7.5–8.7 $\times 10^{-6}$ rechecked in double, inside the 10^{-4} criterion at every scale. Orderings by descending score, baseline H6 > H0 > H9 on both splits: at $\alpha = 0.5$, fit and held-out give H6 > H9 > H0; at $\alpha = 1$, fit H6 > H9 > H0 and held-out H9 > H6 > H0; at $\alpha \in \{2, 4\}$ both splits give the exact inversion H9 > H0 > H6. The $k = 4$ chain (adding L9H7) reaches only $\|W\Phi\|_F = 3.1$–3.5×10^{-3}, above the 10^{-4} criterion, so no $k = 4$ instance is claimed; the failure is the composed projection's residual, not a measured ordering. GPT-2 Small, fp32, seed 42; archived in `results/multicomponent_*.json`.

Pass criteria, fixed before the producing runs. The annihilation criterion is algebraic: $\|W\Phi\|_F \leq 10^{-4}$, and we additionally report $\max_p \|W\phi(x_p)\|_2$. This, not any behavioral quantity, is what the theorem's clause (2) requires. Probe divergence is reported as an *instrumentation check* rather than as a second proof of annihilation: we record, for each probe operator, the 95th-percentile KL over the complete probe family and report the worst across operators, calibrate the scale the same way by constructing W whose annihilation residual sits exactly at 10^{-4} in random directions, and use W_{bad} to establish that the corrected measurement has the dynamic range to detect exposure. We do not claim this calibration upper-bounds the divergence of every W passing the algebraic check: matrices with equal $\|W\Phi\|_F$ can point in different residual directions and be amplified differently downstream. Finally, the held-out interchange gap must carry the reversed sign at $\alpha = 1$, evaluated directly rather than inferred from $\widehat{\alpha}_{\min} \leq 1$, since the response need not be monotone in α.

C A Negative Result on k-NN Projection Repair

Off-manifold displacement is an intervention pathology orthogonal to the non-identification of the main text: an intervened state can sit far from any activation the network produces naturally, and projecting it back onto an estimated manifold is a proposed repair. A repair claim has two parts. Geometrically, the projected state should return to the natural range; semantically, the intervention's effect should survive, since the practitioner's target is a semantic counterfactual rather than the nearest point on a manifold, and we call the gap between the two the semantic residual, which we do not estimate. The experiment below tests a k-NN projection estimator and reports a negative result with two faces: the projector destroys the effect of a reference intervention known to carry one, so post-projection effects carry little information about pre-projection ones, and under a coarse reference bank it manufactures an apparent effect by collapsing every input to the same point. Two reporting items follow for any study that projects: the cross-fit manifold deviation of intervened states before and after projection, and a preservation control, verifying that the projector preserves an intervention known to carry an effect before trusting any post-projection number.

The test is on Gemma-2 2B (Gemma Team et al., 2024) with the canonical Gemma Scope 16k SAE (Lieberum et al., 2024) at layer 19, n=50 IOI pairs, reporting the *normalized interchange effect* $\text{NIE}(\tilde{h}) = (m(\tilde{h}) - m(h))/(m(h_{\text{patch}}) - m(h))$ on the logit difference (0 = no effect, 1 = full patching), where h_{patch} is the *whole-sequence* layer-19 residual patch from the counterfactual prompt; the table's activation-patch preservation control is the *final-token-only* patch, the like-for-like comparison for the single-position swap, which is why its unprojected NIE is 0.283 rather than 1. Alongside it we report the manifold deviation $D(\tilde{h}, \mathcal{M}) := \|\tilde{h} - \Pi_{\mathcal{M}}(\tilde{h})\|_2$, the distance to the k-NN mean projection $\Pi_{\mathcal{M}}$ computed against the ruler half of the bank. NIE is the mean of the 50 per-pair ratios, not a ratio of means; a pair is dropped if $|m(h_{\text{patch}}) - m(h)| < 10^{-6}$, which never fired here ($n = 50$ of 50 in every cell). Every D is *cross-fit*: the reference bank is split, the first half supplies the projector and the second is the ruler, so natural and projected states are measured against a bank that did not produce them.

Measuring a projected state against its own bank would evaluate in-sample, shrinking D through the shared neighborhood, and would falsely read as the projected state being more on-manifold than real data. A repair claim needs the geometric term to shrink *and* the semantic effect to survive. Two banks estimate $\mathcal{M}$: a task-matched bank of 1,948 held-out IOI activations at k=4, and, as a cautionary comparison, the 500-sample wikitext bank at k=10 used in a pilot run of this experiment. Each bank is split in half by the cross-fit, so 974 and 250 references supply the projector and as many again the ruler.

	task-matched bank		wikitext bank	
state	NIE	$D(\cdot, \mathcal{M})$	NIE	$D(\cdot, \mathcal{M})$
held-out natural	n/a	21.6 ± 10.2	n/a	223.3 ± 10.9
clean base	0	19.4	0	222.9
activation patch (preservation control)	0.283	59.7	n/a	n/a
+ projection	**0.044**	17.0	n/a	n/a
SAE swap	**0.371**	**125.7**	**0.371**	182.4
+ projection	**0.028**	**15.2**	0.557†	46.7†
clean state projected (control)	−0.002	12.4	0.557†	46.7†
matched-norm random, projected (control)	−0.001	12.5	0.556†	46.8†

† Constant-output artifact: under the wikitext bank every projected state returns the same value, so these are not repairs. Not bolded for that reason.

Only the task-matched bank has resolution, and under it projection erases the intervention. Held-out natural activations sit at cross-fit $D = 21.6$ against the swap's 125.7, a $5.8\times$ separation; under the wikitext bank the same quantities are 223.3 and 182.4, so there the intervened state is *closer* to the estimated manifold than a genuine one and no geometric claim it supports is meaningful. With a bank that separates them, the swap attains 37% of the full patching effect and projecting it drops that to 2.8% (paired bootstrap over pairs, $\Delta\text{NIE} = 0.343$, 95% CI $[0.326, 0.361]$), while both controls sit at ≈ 0. The geometric term does reach the natural range: cross-fit D falls from 125.7 to 15.2 against a held-out natural mean of 21.6 ± 10.2, a paired-mean drop of 110.5 $[108.3, 112.7]$. The preservation control, however, shows this cannot be read as a semantic residual. A reference activation patch, carrying a large known pre-projection effect, goes from $\text{NIE} = 0.283$ at $D = 59.7$ to 0.044 at $D = 17.0$ under the identical projector: an 84% loss. Both states are sent to the same place irrespective of provenance, so the post-projection effect carries little information about the pre-projection one, and the swap's collapse bounds the estimator rather than the semantic residual. We therefore report this as a constraint on *this projection estimator*, not on metric projection in general and not as a measurement of the semantic residual.

Under a coarse bank the apparent repair is a constant-output artifact. There the projected swap, projected clean state and projected random displacement give $\text{NIE} = 0.557$, 0.557 and 0.556, and the paired difference between the projected swap and the projected clean state is 0.0000 with a bootstrap CI of $[0.0000, 0.0000]$: the projector returns the same behavior whatever it is given. The cross-fit deviations say the same thing more directly. Within each condition the 50 projected states sit at a single deviation to two decimal places ($D = 46.7$ for the swap, standard deviation 0.00), against 223.3 for held-out natural activations, so the projector collapses every input to a single point nearly five times closer to its own bank than real data ever gets. It also inverts the direction, appearing to *raise* the effect from 0.371 to 0.557, which is how the wikitext-bank pilot initially read as evidence for repair.

Robustness checks. NIE for the unprojected swap is identically 0.371 under both banks, as it must be, since the swap never touches $\mathcal{M}$; this validates the comparison across banks. Sweeping the number of swapped features over $\{50, 200, 1000, \text{all differing}\}$ changes nothing to three decimals, so the top-50 features carry the entire effect. We also note that a binary interchange-accuracy metric scores this same swap at 2%: because that metric asks only whether the source's answer now wins, a state whose task content has been destroyed scores near chance, and it cannot separate installing the source's answer from erasing the base's, which is why the normalized effect is reported. GB10 CUDA fp16, seed 0; archived in `results/e5_repair_v3_layer19.json`.

D Additional Experimental Results

Coverage audit of an ACDC-shaped evidence family. The audit of §4 builds the evidence family automated circuit discovery uses on IOI in shape, not from a released prompt set: 50 clean/corrupted pairs (ABC corruption, the repeated name's second occurrence replaced by a length-matched third name), and for each pair every upstream node (attention heads and MLPs, layers 0–8, 117 nodes) patched one at a time with the corrupted run's activation at all positions. Each patched run realizes one feature column (site `resid_pre`, layer 9, final token); Φ additionally carries the clean and

corrupted states. The family probes the designated pair within each realized context, as a discovery sweep does in order to rank those heads; by Lem. 2(i) each such probe exposes W along that context's site feature, so every column is a constraint (the most conservative operator reading, constraints as differences against each pair's corrupt reference, is computed alongside and reported below). An edge-level pipeline is exposed by a second mechanism: an edge patch into one head of the pair gives the two writes different inputs, breaking the cancellation of Lem. 1 outright and exposing W along the two contexts' feature difference, whether or not the pair is separately probed.

Two evidence conventions should be kept distinct: these columns model *trace-level* evidence, the per-context outcomes an edge-realizing sweep computes as it runs; a node-level sweep that scores each upstream node by an output divergence probes the pair in few contexts and reports threshold-compared aggregates, and its *reported* evidence sits nearer the single-position regime of Tab. 1 than to this family's rank ceiling. Rank and nullspace in double precision at 10^{-6} relative tolerance; evaluation on 24 held-out swapped-name contrasts disjoint from the audit pairs. The reading was committed in the producing script's header before the run: rank at the ceiling with median $R < 1$ reads as empirical near-coverage (exact coverage in Cor. 1 requires $u \in S$); median $R > 1$ reads as exposure surviving pipeline-shaped evidence (the conservative-operator recomputation was added after the first run and changed no reading).

Measured: rank 767 of 768 already at 10 pairs (1,190 columns), with median uncovered fraction 1.7×10^{-6} and median $R = 3.8 \times 10^{-5}$ (maximum 8.7×10^{-4}); at 50 pairs (5,950 columns) the medians fall to 1.2×10^{-7} and 3.3×10^{-6}. Ten pairs is the smallest budget run, so the ceiling is reached at or below it rather than located there. We note which protective mechanism this is: $\dim S^{\perp} = 1$ at every budget under both operators, so the family has effectively saturated the feature space at this width, and the near-coverage is of the span-exhaustion kind rather than the claim-specific kind that Cor. 1 describes and the covered-contrast arm exhibits with a 176-dimensional annihilator intact. Under the conservative difference reading the numbers are indistinguishable: rank 767 at every budget, median uncovered fraction 2.1×10^{-6} and median $R = 4.5 \times 10^{-5}$ at 10 pairs, 1.6×10^{-7} and 3.9×10^{-6} at 50. No contrast reaches $R = 1$ at any budget under either operator. GPT-2 Small, MPS fp32, seed 42; archived in `results/acdc_family_audit_L9.json`. Statistical tests of whether a claimed circuit satisfies behavior-preservation, localization, and minimality criteria (Shi et al., 2024) are complementary to this audit: they assess a circuit against those criteria taking the evidence family as informative about it, while the audit asks whether the family covers the reported contrast at all.

Released-distribution replication. The audit above matches the ACDC pipeline in shape with constructed prompts; a second run samples prompts from the verbatim released IOI assets of Wang et al. (2023) (99 names single-token under the GPT-2 tokenizer, all 30 BABA/ABBA templates, released places and objects, vendored with provenance in the bundled `vendored_ioi_assets.py`; the prompts themselves are freshly sampled, seed 42, and the ABC corruption is ours), 636 distinct prompts, same site and heads. The reading was committed in the script's header before the run, two-sided: single-position median $R > 1$ at every P and ACDC-shaped median $R < 1$ confirm the evidence-shapes claim on the released distribution; a single-position median $R < 1$ would contradict the exposure claim at published scale and be reported prominently. Measured: nested single-position families give median uncovered fraction 0.75, 0.60, 0.46, 0.29, 0.14 and median $R = 9.57$, 8.35, 6.26, 3.76, 1.72 at $P = 32$, 64, 128, 256, 512 ($R > 1$: 24, 24, 24, 24, 22 of 24); the ACDC-shaped family (50 clean/ABC pairs, 117 upstream nodes, all positions) reaches rank 767 at ten pairs, median $R = 2.9 \times 10^{-5}$ raw and 2.7×10^{-5} centered, 0 of 24 above threshold at every budget under either operator. No witness is run here; this experiment measures coverage and R only. GPT-2 Small, MPS fp32, seed 42; archived in `results/verbatim_ioi_audit.json`.

The first-order margin is accurate at this site. Over all 360 contrast–q cases, none censored, the scale at which the gap changes sign is a median $1.03\times$ the first-order prediction $|\Delta(0)|/(\|G\|\|u_{\perp}\|)$, with interquartile range $[1.01, 1.06]$. Measuring the remainder directly rather than through the required scale agrees: writing the two held-out states at their realized displacements $\pm\|u_{\perp}\|\hat{e}$ gives $|\rho_a - \rho_b|$ at a median of 0.006 against a first-order term of order one, so the identity $\Delta(W^{\star}) = \Delta(0) - \mathrm{sign}(\Delta(0))\|G\|\|u_{\perp}\| + (\rho_a - \rho_b)$ reproduces the measured gap to within 0.005% (median 0.001%; the residual is a self-test of the decomposition). The implied curvature $|\rho_a - \rho_b|/\|u_{\perp}\|^2$ has median 0.02 (GB10 CUDA fp32, seed 42, $q = 64$, $P = 32$; archived

Table 2: **Exposure collapse tracks the effective probe rank and the transversal mass, which the nominal product PT overstates** (plotted in Fig. 2, right). IOI name-mover ordering on GPT-2 Small, $\phi = \mathrm{id}$, $q = d = 768$; P prompts patched at T positions, the designated pair probed at the site within each patched context (each probe exposing that context's feature, Lem. 2(i)), four pool seeds, 24 held-out contrasts each. The T positions are, in order, the token positions at which base and counterfactual prompt differ (the name positions) and the final token; each patch writes the counterfactual prompt's own activations into `resid_pre` at an upstream layer. Geometry columns are medians over seeds. Bold marks the extremes; starred cells sit at the nominal threshold $PT = d$ and none collapses (Eq. (2)). Cells at $\dim S^{\perp} = 1$ annihilate only to $\sigma_{\min}$; there $\|u_{\perp}\|$ is numerically zero, so the witness is unavailable rather than weak. Twelve archived cells at $P \in \{32, 64, 128\}$ obey the same law; Tab. 3 gives the width replication per cell. Since $r_1(P)$ is read off each $T = 1$ cell, the law's predictive content is the $T \geq 2$ cells, 18 of the 24 here and 12 of the 16 at $d = 2048$, every one of which it reproduces exactly.

P	T	PT/d	$\mathrm{rank}\,\Phi$	$\dim S^{\perp}$	median $\|u_{\perp}\|$	reversed
192	1	0.25	141	627	17.61	**96/96**
192	2	0.50	333	435	8.00	**96/96**
192	3	0.75	525	243	3.56	89/96
192	4	1.00*	717	51	1.14	40/96
256	1	0.33	163	605	16.33	**96/96**
256	2	0.67	419	349	6.56	**96/96**
256	3	1.00*	675	93	1.97	58/96
256	4	1.33	767	1	0.00	**0/96**
384	1	0.50	199	569	14.95	**96/96**
384	2	1.00*	583	185	3.85	91/96
384	3	1.50	767	1	0.00	**0/96**
384	4	2.00	767	1	0.00	**0/96**

Table 3: **At $d = 2048$, no reachable position-resolved budget collapses exposure**: the width replication, per cell (plotted in Fig. 2, right). Pythia-1B, layer 12, heads L12H0 / L12H1, canonical $\phi = \mathrm{id}$, $q = d = 2048$, fp32, seed 42, 24 held-out contrasts per cell; same protocol as Tab. 2. Eq. (2) reproduces every rank; no cell reaches an effective ceiling, and no reachable budget collapses exposure at this width.

P	T	PT	$\mathrm{rank}\,\Phi$	$\dim S^{\perp}$	median $\|u_{\perp}\|$	reversed
128	1–4	128–512	97, 225, 353, 481	1951–1567	14.0–6.8	96/96
256	1–4	256–1024	161, 417, 673, 929	1887–1119	13.3–4.5	96/96
384	1–4	384–1536	196, 580, 964, 1348	1852–700	13.0–3.2	93/96
512	1–4	512–2048	213, 725, 1237, 1749	1835–299	12.8–1.7	80/96

Per-T reversals: $24/24$ everywhere except $(384, 4)$ at $21/24$ and $(512, 4)$ at $8/24$, the two cells with the smallest remaining $\|u_{\perp}\|$.

in `results/realised_remainder_fixed_pythia-1b_q64_seed42.json`, this measurement rerun on the corrected interchange path of App. B.2; the pre-correction run is retained in the bundle, and its failed identity check, median residual 1.04, is the defective path's signature). We therefore do *not* treat ω_{Ψ} as the binding term in (H3) at this site, and we do not certify it: two sampling estimators of the gradient-Lipschitz constant C (chord quotients along the steered segment, and Hessian power iteration over a sample of the ball) are reported in App. D.2 for completeness, but with the remainder measured this small the certification question does not arise here. Whether it arises at sites with larger curvature is open.

Baseline-gap stratification. The 360-case pool reported in §4 pools the five probe families of `results/q_sweep_fixed_seed[1-5].json` (the corrected runs; the pre-fix files of App. B.2 are retained beside them under their original names): 5 pool seeds $\times$ 24 held-out contrasts $\times\ q \in \{64, 128, 256\}$ gives 360 cases, all annihilating ($\max \|W\Phi\|_F = 1.25 \times 10^{-5}$). Quintile edges of $|\Delta(0)|$ are 0.177, 0.408, 0.532 and 0.743 logits; since $|\Delta(0)|$ is invariant in q the cases

Table 4: **The matched-rank arm: positions buy modestly more than constraint count.** IOI name-mover ordering on GPT-2 Small, $\phi = \mathrm{id}$, $q = d = 768$, fp32, pool seed 42, 24 held-out contrasts; at each target rank a $T = 1$ family of P prompts is compared with a $T = 4$ family of roughly $P/4$ prompts patched at four positions, matched on measured $\mathrm{rank}\,\Phi$ (within 3). Transversal mass is 1.24–1.50× smaller under $T = 4$ throughout, and reversal rates separate only at rank 640.

	$T = 1$		$T = 4$	
target rank	median $\|u_\perp\|$	reversed	median $\|u_\perp\|$	reversed
128	13.30	24/24	10.73	24/24
256	9.24	24/24	6.28	24/24
384	6.64	24/24	4.41	24/24
512	4.57	24/24	3.13	24/24
640	2.72	23/24	1.85	17/24

Table 5: **The theorem's own witness reverses the held-out ranking at a quarter of unit operator norm while the probes see nothing** (the four-arm run of §4). Pass criteria, fixed before the run: $\|W\Phi\|_F \le 10^{-4}$ and a reversed sign at $\alpha = 1$; probe KL is the worst across probe operators of the 95th percentile over the probe family, an instrumentation check against a ceiling calibrated the same way at 2.3×10^{-7}, not a criterion; for the two control arms the desired outcome at $\alpha = 1$ is no reversal. $\widehat{\alpha}_{\min} = 0.25$ is the grid floor, hence an upper bound on the crossing. Pythia-1B, layer 8, final-token site, $q = 64$, fp32, seed 42; the tabled contrast has $\Delta(0) = -0.161$, and $\Delta(0)$, $\Delta(1)$ are reproduced through an independent hook implementation.

arm	$\|W\Phi\|_F$	probe KL (p95)	$\widehat{\alpha}_{\min}$	$\Delta(1)$	reverses at $\alpha = 1$
$W^\star$ (theorem's witness)	2.4×10^{-6}	1.9×10^{-7}	**0.25**	$\mathbf{+2.61}$	✓
W_{opt} (supplementary search)	3.4×10^{-6}	1.8×10^{-7}	0.5	$\mathbf{+0.227}$	✓
W_{rand} (random in Ω)	1.9×10^{-5}	1.8×10^{-7}	1.5	-0.048	×
W_{bad} (positive control)	1.7×10^{2}	6.1×10^{-1}	2.0	-0.069	×

arrive in triplets and the strata are 72, 72, 69, 75 and 72 rather than exactly one fifth each. Counts share contrasts across q, so we report them raw, and binomial intervals would be mildly optimistic. The median gap is 0.458. Reversal is 359/360 pooled, and by stratum from the smallest gaps up 72/72, 72/72, 69/69, 75/75 and 71/72: the single exception sits in the top stratum, not among the near-ties. The R-calibration bins are the same cases partitioned by the pre-registered ordering statistic rather than by $|\Delta(0)|$; within this 360-case pool exactly one case has $R < 1$ (the second $R < 1$ case of §4 is the probe-selection contrast of App. A.2), so these data cannot test the first-order predicate's soundness at the boundary, and we do not claim they do; the Greater-Than cells of Tab. 7, with 28 sub-threshold contrasts, do.

The q-confound families (model, layer and contrasts fixed, nested $P \in \{64, 128, 192\}$ at $q \in \{256, 512, 1024\}$, pool seed 1) reverse 216 of 216 with worst annihilation residual 3.5×10^{-5}, which is the support for the fixed-P q-sweep claim of §4.

The measurable-in-advance figures of §4 (a 32-probe Pythia-1B family at $q \in \{64, 128, 256\}$ leaving median 56–81% of held-out contrast mass uncovered, cross-task contrasts at 54–79%) are archived in `results/transversality_gate_dedup.json`.

Why a PCA feature map with $q \ll d$ understates exposure. With $q \ll d$, a few hundred diverse inputs can saturate the reduced feature space, killing the witness by capacity; under the canonical $\phi = \mathrm{id}$ the prompt axis buys rank one per clean prompt, and Tab. 1 measures $\mathrm{rank}\,\Phi = P$ exactly with 24/24 reversals through $P = 512$ at $q = d$. Reduced PCA feature spaces can saturate after only a few hundred probes, substantially earlier than the canonical $q = d$ space; this is why the reporting recommendation fixes ϕ. A precursor diagnostic (Pythia-160M, PCA maps at $q \le 512$ over a 500-input diverse sweep, archived in the bundle) found the witness unavailable at every q, by capacity where the family saturated the reduced space and by transversality where rank slack remained but the evaluation contrasts sat at median angle 0.01° to the probe span; the operative

Table 6: **Reversal is insensitive to probe count over the whole practical range**: the nested probe-family sweep, per cell. The single-position sweep of §4, with the quantity the witness consumes, $\|u_\perp\|$, alongside the outcome. Pythia-1B pools three pool seeds (72 contrasts per cell; geometry entries are medians of per-seed medians, reversal counts pooled); GPT-J-6B is one seed (24). Reversal is insensitive to probe count over the whole practical range even as $\|u_\perp\|$ falls by a factor of 3, and $\widehat{\alpha}_{\min}$ rises only from 0.11 to 0.36: the family is buying transversal mass slowly and exposure not at all. Both models turn over only at $\dim S^\perp = 1$, and they turn over differently, because GPT-J retains $13\times$ the transversal mass there. Cells marked † annihilate only to $\sigma_{\min}(\Phi)$ rather than exactly, for the structural reason given in §4; residuals are `fp32`, and recomputing the contrast projection in double drops every unmarked cell to $\leq 2.7 \times 10^{-6}$ while leaving the marked cells where they are, which is what identifies their residual as structural. The 10^{-6} floor is itself an artifact of re-promoting a basis stored in `fp32`; carrying the basis in double throughout, as the head-level runs of Tab. 1 do, reaches $\leq 4.7 \times 10^{-13}$. Layer 8 of 16 (Pythia-1B) and 14 of 28 (GPT-J-6B), $q = 256$, fp32, seed 42.

		measured			
P	$\dim S^\perp$	median $\|u_\perp\|$	reversed at $\alpha = 1$	median $\widehat{\alpha}_{\min}$	max $\|W\Phi\|_F$
Pythia-1B, three pool seeds					
8	248	1.295	72/72	0.113	4.1×10^{-6}
16	240	1.247	72/72	0.113	9.6×10^{-6}
32	224	1.133	72/72	0.125	1.3×10^{-5}
64	192	0.946	72/72	0.150	1.0×10^{-5}
128	128	0.707	72/72	0.213	2.2×10^{-5}
192	64	0.433	71/72	0.363	3.8×10^{-5}
256	1	0.033	10/72	4.18	2.8×10^{-4} †
GPT-J-6B, one pool seed					
8	248	7.551	24/24	0.050	7.4×10^{-6}
16	240	6.758	24/24	0.050	1.2×10^{-5}
32	224	5.945	24/24	0.050	1.5×10^{-5}
64	192	5.069	24/24	0.056	1.8×10^{-5}
128	128	3.600	24/24	0.081	1.9×10^{-5}
192	64	2.379	24/24	0.113	3.9×10^{-5}
256	1	0.442	17/24	0.575	1.1×10^{-4} †

diagnostic there, as in the main text, is the transversal mass of the evaluation contrast, not rank slack against q.

D.1 The dictionary constructions in full

The dead-unit construction: a second JumpReLU SAE, exactly. An activation-space shear (the precursor below) edits activations, which no JumpReLU encoder (Rajamanoharan et al., 2024) need realize. Editing one unit and two decoder rows realizes it exactly within the architecture: pick a unit k dead on the corpus and any decoder direction d_m, $m \notin \{i, j, k\}$, and set $w_k = c\,w_i$, $b_k = c\,b_i$, $\theta_k = c\,\theta_i$ (the unit's affine encoder parameters and threshold scaled together; JumpReLU commutes with positive scaling, so $a_k = c\,a_i$ exactly and stays nonnegative), $d_k = -s\,d_m$, $d_i' = d_i + s\,c\,d_m$, $s \in \{\pm 1\}$.

The reconstruction is pointwise identical ($a_i\,s\,c\,d_m$ cancels $a_k \cdot (-s\,d_m)$), and on prompts with $a_i = 0$ unit k is silent, so every single-feature ablation agrees exactly under the two factorizations: the result is a valid JumpReLU SAE of identical architecture and width, removing the encoder-realizability caveat the precursor carries (scope note (i) below), though still not a trained Gemma Scope SAE (see disclosure (d)). Exactness holds on every input where the repurposed unit was originally inactive, measured over all 400 prompts of both corpora; we do not certify deadness off-corpus.

The evaluation runs simulate the edited SAE's ablations algebraically; a separate verification instantiates the weight edits in the SAE's parameters for all 36 archived pair configurations and measures, over each full corpus, zero JumpReLU gate flips, maximum $|a_k - c\,a_i|$ of 2.3×10^{-5}, and maximum reconstruction delta 3.1×10^{-5} against a pre-committed 10^{-4} (36 of 36 pass; feature i's *encoder* parameters are untouched and its activations are bit-identical).

Table 7: **Greater-Than replication: a different task, an earlier dose–response turnover, the same exposure relationship.** Greater-Than task of Hanna et al. (2023) ("The *noun* lasted from the year 17YY to the year 17", the model preferring continuations above the span start), GPT-2 Small, canonical $\phi = \mathrm{id}$, $q = d = 768$, single-position interchange at the final token; measurement $\Psi = \mathrm{logit}(\mathrm{YY}{+}10) - \mathrm{logit}(\mathrm{YY}{-}10)$, a single-token analogue of that work's probability-difference metric; 20 nouns, start years 12–88, pool of 586 unique prompts, nested probe families, 24 held-out contrasts per pair; GB10 CUDA fp32, seed 42; outcome reading pre-committed in the producing script before the run, and both head pairs fixed in advance, no pair selected or replaced on outcome (L8H11, L8H8, and L6H9 are heads of that work's published circuit; L6H1 appears in its appendix variant-task analysis). $\mathrm{rank}\,\Phi = P$ in every cell; worst annihilation residual 2.9×10^{-5} (3.5×10^{-13} in double). Across all cells $R > 1$ agrees with reversal on 238 of 240 contrasts, the two exceptions at $R = 1.11$ and 1.03 (above the boundary, not reversing); none of the 28 contrasts with $R < 1$ reverses. Archived in `results/greaterthan_circuit_attack_L8_h11h8.json` and `..._L6_h9h1.json`.

			L8H11 / L8H8		L6H9 / L6H1	
probe count P	$\mathrm{rank}\,\Phi$	$\dim S^{\perp}$	reversed	median R	reversed	median R
32	32	736	24/24	28.0	24/24	35.6
64	64	704	24/24	10.9	24/24	21.8
128	128	640	24/24	6.3	24/24	12.1
256	256	512	19/24	3.5	21/24	5.1
512	512	256	12/24	1.2	14/24	1.5

Only the i-ablation shifts, by $a_i\, s\, c\, d_m$; the j-ablation is unchanged, so its baseline outcome is reused. Jointly ablating the co-activating pair $\{i, k\}$ cancels the edit exactly, $\hat{x} - a_i(d_i + s\, c\, d_m) - c\, a_i \cdot (-s\, d_m) = \hat{x} - a_i d_i$, reverting to the original SAE's i-ablation: the reversal is a property of single-coordinate interventions, one more face of the co-activation separator.

Per pair, the active prompts split into a witness slice (first 8: the aiming gradient G, the mean gap fixing s, and the screening statistic) and an evaluation slice (next 24, touched by nothing in witness construction): $m^{\star} = \arg\max_m |\langle G, d_m\rangle| / \|d_m\|$, s opposing the witness-slice gap sign, and $c = \|d_i\| / \|d_{m^{\star}}\|$ capping the added component at the replaced row's norm (measured $c = 1.00$, Gemma Scope decoder rows being near unit norm; the edited row d_i' has norm near $\sqrt{2}$). Pairs are screened by a selection heuristic on the witness slice (median contribution over baseline gap; not the paper's R); feasibility conditions (activity and probe-support counts) use the full corpus, the screening statistic and the witness only the witness slice.

Pre-committed readings, results, and disclosures. Committed in the producing script's header before the run: exactness (probe-ablation disagreement $\leq 10^{-4}$ over all 64 probes per pair, both features; algebraic, violations are bugs) and, primary, at least 3 of the top 12 screened pairs reversing ≥ 12 of 24 evaluation prompts on seed 0, seed 1 replication reported either way. Measured: probe-ablation disagreement 0.0 exactly, every pair, both seeds; 12 of 12 pairs pass on seed 0 (per-pair 12–24 of 24, median 20.5) and 12 of 12 on seed 1 (13–24, median 20.0); 475 of 576 contrasts reverse in total (not independent trials: prompts within a pair share one witness, pairs share the partner), per-prompt rows archived. The reversal is sign-structured: the single edit flips 475 of 482 evaluation contrasts whose baseline gap sign matches the witness-slice mean and 0 of 94 opposite-sign contrasts, and flipped contrasts have larger baseline gaps (median 0.221 versus 0.072 logits), so the reversals do not live on near-ties. Disclosures: (a) lineage: an unaimed variant steering along d_j (its signed reversal criterion pre-committed before that first dead-unit run) reversed only 1–3 per cell (two pairs, two seeds, one cell $n = 23$); a first aimed run whose witness construction drew on its evaluation prompts (the gradient on 8 of 24, the steering sign on all 24) passed 12 of 12 on both seeds and prompted this stricter protocol (both precursors on GB10 fp32, readings committed before each run, all archived). On the 14 pairs the two aimed runs share, per-pair counts differ by at most 2 (net -4 of 336), so the overlap inflated little; the median drop to 20.5/20.0 is pair-set composition.

(b) All screened pairs share one partner feature, a dense unit active on all 400 corpus prompts (decoder aligned with whitespace and ellipsis tokens); the concentration is structural, the screen taking the densest feasible partner. A pre-committed exclusion arm promotes a second, disjoint dense partner (active on 348 of 400) and 6 of 6 pairs reverse on both seeds (14–23 and 12–22 of 24, medians 20.0

and 18.5; 225 of 288 in total): the reversal does not depend on the particular dense partner; partner density itself is a construction choice, not a tested variable.

(c) The aiming selects one of three decoder directions across all pairs (a punctuation-aligned feature at 187 of 400 fires and two rare features at 22 and 9 of 400, the third on one seed-0 pair only); that semantically arbitrary decoder rows carry the steering power, and that the edited d'_i is visibly contaminated in weight space, is the point of §3 at this granularity. Feature identities for every involved feature are archived with the run; the reversing i-features include kinship, object-pronoun, and shop and staff features alongside uninterpretable ones, screened by activity, not interpretability.

(d) The alternative SAE is edited, not trained: reconstruction and probe outcomes are identical, but L_0 rises by one wherever $a_i > 0$, so a sparsity-penalized objective separates the two factorizations, as do weight inspection and co-activation; reachability by training is not claimed. Gemma-2 2B, canonical Gemma Scope 16k, layer 19, Apple MPS fp32, seeds 0/1, 400-prompt corpora (half IOI, half heterogeneous), 64 probes, 8 witness and 24 evaluation prompts per pair; archived in `results/dictionary_deadunit_aimed_v3_layer19.json`, `..._v3_jdiv.json`, and `..._verify_layer19.json`, precursors in `..._exact_layer19.json` and `..._aimed_layer19.json`.

The precursor shear. A feature's contribution must stay a scalar times a fixed direction or the object stops being a dictionary, so the dictionary-form redistribution is a shear in the pair's own two-dimensional span,

$$d'_i = d_i + c\,d_j, \qquad a'_j = a_j - c\,a_i, \qquad d'_j = d_j, \qquad a'_i = a_i,$$

for which $a'_i d'_i + a'_j d'_j = a_i d_i + a_j d_j$ pointwise: the reconstruction, the residual stream and every downstream computation are unchanged on every input, and the support pattern moves only where $a_i \neq 0$. Ablating either feature differs between the two factorizations by exactly $c\,a_i(x)\,d_j$, so both ablations agree on any input with $a_i(x) = 0$. Sparsity makes those the common case, but a feature cannot be ablated where it is inactive, so these probes are informative about feature j and about reconstruction while being, by construction, the prompts on which the shear is invisible; the evaluation contrasts are the prompts with $a_i \neq 0$, which is where a study establishing feature i's importance would actually look. We do not claim sparse dictionaries are generically non-identifiable: dictionary learning admits identifiability results up to permutation and scaling under standard conditions, and this construction neither exploits nor contradicts them, since the second factorization is not shown to lie in the trained family.

Protocol and result. Gemma-2 2B, canonical 16k Gemma Scope decoder at layer 19, a matched-contribution feature pair screened from 400-prompt corpora (half IOI, half heterogeneous), 64 probe prompts with $a_i = 0$, 24 held-out prompts with $a_i > 0$, shear $c = \|d_i\|/\|d_j\|$, two corpus seeds. Max probe-ablation disagreement between the two factorizations: 0.0 exactly, both seeds. Reported importance ordering of the pair under the signed logit difference: reversed on 22 and 21 of 24 held-out prompts, on 8 and 6 of 24 within the norm window, shears keeping the transferred component no larger than the replaced one ($\|c\,d_j\| \leq \|d_i\|$).

The four scope notes. (i) *Admissibility.* Gemma Scope SAEs are JumpReLU autoencoders with rectified feature magnitudes, so an admissible state has $a'_j \geq 0$. On the probes this holds exactly ($a_i = 0$ there). On the evaluation prompts it does not: at the reported shear $a'_j < 0$ on 5 and 6 of 24; the largest shear keeping every evaluation prompt nonnegative is 0.86 and 0.79 of the reported one on the two corpora, and within that window the reversal rate falls to 3 and 6 of 24; at twice the reported shear a'_j is negative on 23 to 24 of 24 (archived per prompt). Nonnegativity is necessary but not sufficient: we do not show $a_j - c\,a_i$ is realizable by any JumpReLU affine encoder unit, so even the within-window reversals establish an alternative exact factorization of the reconstruction, not a second Gemma Scope SAE; the dead-unit construction above removes this caveat. We attempted both, under pre-committed gates, and both returned negatives, which we report with their residuals as that protocol's header requires. *Encoder realizability:* fitting a single JumpReLU unit to the sheared activation passed the fit gate on four of five attempted pairs (≤ 0.05 relative L_2; best 0.0029) but failed the probe-agreement gate on all five, at 1.06, 8.07, 11.97, 14.03 and 14.39 in logit L_2 against a pre-committed 0.01, so no pair yields an approximate within-family instance. *Pair dependence:* a 40-pair screen, pre-committed to read three or more pairs reversing ≥ 12 of 24 inside their admissible window as evidence that the pilot pair is not unique, returned one pair of 40 (median 3 of 24, twelve pairs at zero), so the pair-dependence concern

of scope note (iii) stands as measured. Both negatives are why the dead-unit construction, which needs no window and no fit, supersedes the shear rather than merely improving it; archived in `results/dictionary_realizability_layer19.json`. (ii) ***Metric.*** The pre-committed statistic was the unsigned magnitude ordering $\|\Delta \text{logits}\|_2$, which reverses 0 of 24; that is structural, since the two magnitudes are $|a_i|\,\|d_i + cd_j\|$ and $|a_j - ca_i|\,\|d_j\|$, whose ratio tends to 1 from below as c grows when a_i dominates, so an unsigned ordering asymptotes to a tie and cannot cross. The signed logit difference, the m used everywhere else in this paper, was adopted after the magnitude arm returned zero; both arms are reported. (iii) ***One pair.*** Both corpora selected the same feature pair, so the two seeds replicate the effect across evaluation prompts, not the pair choice. (iv) ***Scale.*** The within-norm rate of 6 to 8 of 24 is well below the 21 to 22 at larger shear, and the admissibility bound of (i) cuts into exactly that within-norm subset.

D.2 Certifying the remainder modulus

(H3) is the only hypothesis of Theorem 1 that the paper cannot read off f directly, because ω_Ψ is a supremum over a displacement region rather than a quantity two forward passes return. Under (H2) it is controlled by the gradient-Lipschitz constant $L_{\nabla\Psi}$ on the two balls, $\omega_\Psi(r) \le L_{\nabla\Psi} r^2$, so bounding $L_{\nabla\Psi}$ would make the antecedent checkable.

We report two estimators, because the difference between them is itself the finding. The first takes the largest realized difference quotient of $\nabla\Psi$ on a uniform mesh along the two segments each held-out state actually travels, $\tilde{s}_k \pm t\,\|u_\perp\|\,\hat{e}$. That samples two chords through a ball in $d = 2048$ dimensions, so curvature in the remaining directions is never seen. On the 24 held-out contrasts of the arm run it admits all 24, which is uninformative because all 24 also reverse: the predicate excludes nothing, and would have to be built on an $L_{\nabla\Psi}$ a median $7.2\times$ larger (range 2.4–$21.3\times$) before it excluded any contrast at all. A chord estimate is a lower bound on the constant (H2) posits, and at this dimension it is not a close one.

The second targets the quantity (H2) names. The gradient-Lipschitz constant at a point is the operator norm of the Hessian, and that is what power iteration on Hessian-vector products returns, at the cost of one extra backward pass per power-iteration step (8 here) through a graph retained with `create_graph`. We take

$$\widehat{L} \;=\; \max_{k\in\{a,b\}} \; \max_{x\in B_k} \; \left\|\nabla^2\Psi(x)\right\|_{\text{op}},$$

with B_k a sample of the ball $B(\tilde{s}_k, \|u_\perp\|)$ containing both the steered-segment mesh and random directions at radius $\|u_\perp\|$. On Pythia-1B layer 8, $q = 64$, $P = 32$, this returns a median $\widehat{L}$ of 1.0, a factor 1.4–1.8 above the chord estimate's 0.57 (both archived: `results/curvature_certificate_fixed_offset150.json`, `..._hvp_fixed_offset150.json`, reruns on the corrected interchange path): sampling the ball rather than one line through it recovers some of the curvature the chord misses, but not much of it.

Neither estimator supports a certificate, and we report this as a negative result rather than presenting either as one. Under the chord estimate the predicate $|\Delta(0)| < \|G\|\|u_\perp\| - \widehat{L}\|u_\perp\|^2$ admits all 24 held-out contrasts; the Hessian estimate admits them too. Since the corrected measurement reverses all 24, the predicate happens to be right on this set, but it is right vacuously: it would need the chord estimate to be a median $7.2\times$ larger (range 2.4–$21.3\times$), or the Hessian estimate $4.3\times$ (range 1.1–$10.9\times$), before either excluded anything at all. The main text's statement that the realized remainder sits well below $\widehat{L}\|u_\perp\|^2$ corresponds, on the same 24 contrasts, to a per-contrast median factor of 42 for this Hessian estimate against the realized curvature (29 for the chord estimate; as ratios of medians, 40 and 23). And a maximum over finitely many sampled points cannot bound a supremum over an uncountable ball in $d = 2048$ dimensions in any case. The reason this costs the paper nothing is measured rather than assumed: the realized remainder at this site is a median 0.006 (App. D), so ω_Ψ is not the binding term and the certificate is not needed to reach the conclusion. We therefore call both quantities *sampled-ball estimates* of $L_{\nabla\Psi}$, never certificates, and draw no conclusion from a predicate built on them; under a genuine upper bound (H3) would be sound by construction, and the empirical question would become *coverage*, the fraction of the reversing set a conservative predicate still admits. A genuine bound would require interval arithmetic or a layerwise composition rather than sampling, and at 16 layers we expect the composed constant to be vacuous.