WHAT CAN ACTIVATION-PATCHING EVIDENCE ESTABLISH?

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful activation patching does not necessarily establish the functional claim it is used to support. Even when expected outputs and measured responses are restored, unmeasured claim-changing alternatives can remain indistinguishable. We formalize this gap as *target-relative observability*: relative to a declared class of admissible responses, evidence is sufficient only when every response consistent with the observations agrees on the target. On finite-dimensional linear response spaces, we characterize claim-changing blind directions, minimum target-specific measurement completion, and sharp error amplification; an activation-specific corollary shows when a finite probe panel cannot certify unrestricted local response preservation. Prospectively, among center-restoring GPT-2 patches, full-gradient discrepancy predicts withheld finite-response failures better than public-direction screening by 0.080/0.055 within-layer AUC on IOI/GT, with recurring but model-dependent gains across six external model families. Controlled studies show that variance, generic geometric matching, and span coverage can each misrank evidential value, while reference queries can transfer shared nonlinear structure. Reliable validation need not recover the whole response; it must rule out the admissible alternatives that can change the claim.

1 INTRODUCTION

Activation interventions are used to infer how neural networks compute, from tracing circuits to locating factual associations (Vig et al., 2020; Wang et al., 2023; Meng et al., 2022). Yet an intervention result is evidence for a functional claim only insofar as the collected observations exclude alternatives that would change that claim. We study this structural sufficiency problem. *Restoring what is observed need not restore what a functional claim requires.*

Suppose an activation correction restores the desired answer and agrees with a finite set of response probes. A broader claim—for example, preservation across alternative formulations or activation changes—requires the evidence to exclude every admissible response that matches those observations but changes the claim. If two such responses remain indistinguishable, measuring the same quantities more precisely cannot decide between them.

Our organizing principle is *target-relative observability* (Figure 1). For an admissible response class F, observation map A, and target H, evidence is sufficient when agreement under A forces agreement under H. Sufficiency is therefore relative both to *what the claim asks* and to *which alternatives are admitted*. This separates restoration, target identification, and stable reconstruction, and turns "what else should we measure?" into a target-specific question.

Our contributions are:

1. **A claim-level criterion.** Using classical linear-identifiability tools, we derive target-specific completion, stability, and an activation-specific finite-panel insufficiency result (Section 3).
2. **A prospective falsification test.** Frozen IOI/GT evaluations show that center restoration and a finite public panel leave withheld failures predictable from broader response information; a six-family panel tests recurrence without model-specific fitting (Section 4).

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.

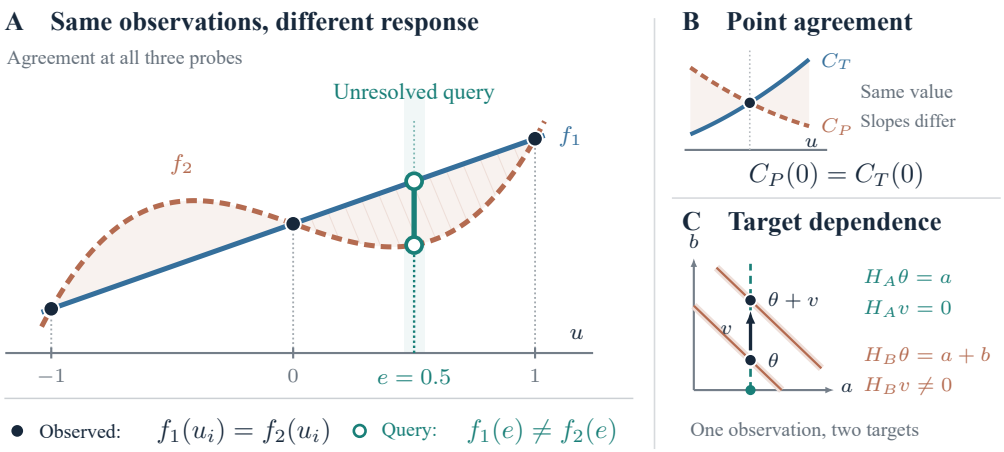


Figure 1: **The claim determines which indistinguishable alternatives must be ruled out. A**, $f_1(u) = u/2$ and $f_2(u) = u/2 + u(u^2 - 1)$ agree at public probes $u_i \in \{-1, 0, 1\}$ but differ at $e = 0.5$. **B**, Patched C_P and clean-target C_T computations can agree at the center while differing under activation changes. **C**, observing a in $\theta = (a, b)$ leaves b unresolved: this blind direction is irrelevant to $H_A\theta = a$ but changes $H_B\theta = a + b$. All panels are schematic.

3. **Controlled measurement-design reversals.** Variance, generic geometric matching, and coverage can misrank evidential value; reference transfer identifies shared nonlinearity as a reusable source of finite-response information (Section 5).

2 Related Work

Intervention validity and mechanistic explanation. Causal abstraction and alignment (Geiger et al., 2021; 2024; Wu et al., 2023), circuit discovery (Conmy et al., 2023; Marks et al., 2025), and faithfulness criteria (Hanna et al., 2024; Mueller et al., 2025; Palumbo et al., 2025; Hadad et al., 2026) formalize what interventions can reveal about computation. Corruption choices, mediator interactions, dormant pathways, and representational divergence expose complementary validity failures (Zhang & Nanda, 2024; Makelov et al., 2024; Vaidyanathan et al., 2026; Grant et al., 2026). These works motivate careful intervention design; our question is whether a fixed collection of observations structurally determines the particular target being claimed.

Explanation identifiability, statistical certification, and claim identifiability. Méloux et al. (2025) ask whether an entire mechanistic explanation is unique under fixed validity criteria and find extensive explanation-level non-identifiability. Certified Interventional Fidelity instead treats a declared interventional quantity as a causal estimand and supplies finite-sample and anytime-valid statistical guarantees for estimating it (Asiaee, 2026). We ask a different question: given observations $A\theta$, do they identify a declared target $H\theta$ over the admissible alternatives F? A full explanation may remain non-unique while a target is already determined; conversely, arbitrarily precise estimation of $A\theta$ cannot recover $H\theta$ when a claim-changing blind direction remains. We therefore characterize target-specific completion and stability after the target and alternative class are declared.

Classical tools, new evidential role. Kernel inclusion, pseudoinverse reconstruction, and rank arguments are classical linear-algebraic tools (Penrose, 1955); optimal design studies which measurements inform specified quantities (Elfving, 1952). We do *not* claim these identities as new mathematics. Our contribution is their claim-level formulation for activation-patching evidence, intervention-specific consequences, and prospective tests. Attribution, path integration, and higher-order methods approximate responses (Syed et al., 2024; Sundararajan et al., 2017; Zhang & Wang, 2026); we instead ask which response information a particular claim requires.

3 WHAT CAN THE OBSERVATIONS ESTABLISH?

We characterize when observations determine a response target, how much target-relevant information remains missing, and how observation error affects the conclusion.

3.1 WHAT IS THE RESPONSE-PRESERVATION TARGET?

At intervention site s, let C denote the scalar task contrast. For clean, corrupt and patched runs, restoration is

$$O_s = \frac{C_{\text{patch}} - C_{\text{corrupt}}}{C_{\text{clean}} - C_{\text{corrupt}}}. \tag{1}$$

With a nondegenerate denominator, unclipped O_s measures center restoration; it leaves responses to activation changes unresolved.

Two full responses define the target. Let $u \in \mathbb{R}^{d_h}$ displace the clean activation, with model width d_h. Patched-recipient and clean-target branches $C^{PJ}(u), C^{TJ}(u)$ share this activation and displacement, retaining their own non-patch contexts and live downstream computation. Write $p = \nabla C^{PJ}(0)$ and $q = \nabla C^{TJ}(0)$. The centered response discrepancy is $\delta(u) = [C^{PJ}(u) - C^{PJ}(0)] - [C^{TJ}(u) - C^{TJ}(0)]$, with $\nabla\delta(0) = p - q$. The target is agreement of changes relative to each branch's center, separating response preservation from restoration.

Finite response disagreement. On M hidden displacements e_m reserved for validation, let $P_m^{\text{end}} = C^{PJ}(e_m) - C^{PJ}(0)$ and $T_m = C^{TJ}(e_m) - C^{TJ}(0)$. Using root mean square (RMS) across displacements, the endpoint is

$$E_s = \frac{\text{RMS}_m(P_m^{\text{end}} - T_m)}{\max\{\text{RMS}_m(P_m^{\text{end}}), \text{RMS}_m(T_m), \epsilon\}}, \qquad \epsilon = 10^{-12}. \tag{2}$$

Here $E_s \in [0, 2]$, with zero indicating identical increments; changing the displacement set or readout changes the target (Section 6).

3.2 WHEN IS THE TARGET OBSERVABLE?

Target-relative observability requires admissible responses agreeing on the observations to agree on the target. Both pieces matter: H states what the claim asks for, while the declared response space F states which alternatives the experiment is meant to distinguish. For linear measurements $A\theta$ and target $H\theta$ on F, sufficiency means $A\theta_1 = A\theta_2 \Rightarrow H\theta_1 = H\theta_2$ for all $\theta_1, \theta_2 \in F$.

Theorem 1 (Target-relative linear observability)**.** *Let F be a finite-dimensional inner-product space of admissible local parameters, fixed independently of tested outcomes. For linear operators $A : F \to \mathbb{R}^r$ and $H : F \to \mathbb{R}^s$, a finite nonnegative κ with $\|H\theta\|_2 \le \kappa\|A\theta\|_2$ for every $\theta \in F$ exists exactly when* $\ker(A|_F) \subseteq \ker(H|_F)$. *In orthonormal coordinates on F, with $\dagger$ denoting the Moore–Penrose pseudoinverse, the smallest constant is $\kappa_* = \|H_F A_F^\dagger\|_{2\to 2}$ (zero when $H_F = 0$).*

The kernel condition is the familiar linear-identifiability criterion; its role here is evidential. A blind direction prevents target identification only if it changes the target value. If the condition fails, no estimator using only $A\theta$ identifies the target over F, regardless of measurement precision; target-irrelevant variation may remain unobserved without harming the claim.

The missing information is target-specific. Unresolved target effects determine what to measure next.

Corollary 1 (Minimal unrestricted measurement completion)**.** *Write A, H in orthonormal coordinates on F, and let $N = I - A^\dagger A$ project onto the unobserved subspace* $\ker A$. *The minimum number of additional arbitrary linear scalar measurements needed to identify $H\theta$ is $k = \text{rank}(HN)$. If the rows of B form an orthonormal basis of* $\text{row}(HN)$, *then*

$$H\theta = HA^\dagger(A\theta) + HNB^\top(B\theta). \tag{3}$$

Thus the minimum completion is set by the *claim-changing part of the blind subspace*, not by the dimension of the full response. For a concrete patching instance, let $x = p - q$ and suppose public probes reveal Ux while the claim asks only for first-order response along a declared displacement d_*, $Hx = d_*^\top x$. Writing $d_* = d_\parallel + d_\perp$ with $d_\parallel \in \mathrm{row}(U)$ and $d_\perp \perp \mathrm{row}(U)$, the public panel already determines $d_\parallel^\top x$; if $d_\perp \neq 0$, one additional scalar measurement $d_\perp^\top x$ completes this target without recovering the full gradient. With measurements restricted to a realizable dictionary D, a query subset S suffices exactly when $\ker[A; D_S] \subseteq \ker H$, characterizing feasibility and minimum dictionary-query count (Appendix B.2). Exact target identification is stronger than some decision claims require: if all admissible target values consistent with the observations lie on the same side of a threshold, the threshold claim is already determined.

Identification and precision are distinct. When the kernel condition holds, reconstruction with $H_F A_F^\dagger$ incurs at most $\kappa_* \|\eta\|_2$ target error under additive measurement error η. In the two-dimensional example, measuring b completes the target $a + b$. Measuring εb instead ($\varepsilon > 0$) also identifies it, but amplifies errors in that measurement by $1/\varepsilon$. Completion changes what is knowable; observation geometry changes how accurately it can be recovered.

3.3 From partial observations to finite responses

For patching, let $d_1, \ldots, d_K$ be public activation displacements and $e_1, \ldots, e_M$ hidden displacements reserved for evaluation. The vectors include their magnitudes. Define $U = K^{-1/2}[d_1, \ldots, d_K]^\top$ and $V = M^{-1/2}[e_1, \ldots, e_M]^\top$, so Euclidean norms of their linear responses equal panel RMS values.

Separate full response from selected-path evidence. $C^{PB}(u)$ and $C^{TB}(u)$ freeze selected downstream gates at values captured separately in the patched-recipient and clean-target runs at $u = 0$; other downstream computation remains live. The selected contribution is

$$\Psi(u) = C^{PJ}(u) - C^{PB}(u) - C^{TJ}(u) + C^{TB}(u). \tag{4}$$

Set $g(u) = \Psi(u) - \Psi(0)$, $c = \nabla\Psi(0)$ and $b(u) = C^{PB}(u) - C^{PB}(0) - C^{TB}(u) + C^{TB}(0)$, with $v = \nabla b(0)$. Then $\delta(u) = g(u) + b(u)$ and $p - q = c + v$: selected-path agreement can leave bypass response unresolved.

For all four branch gradients, let $\theta = (p, p_B, q, q_B) \in F \subseteq \mathbb{R}^{4d_h}$, with $p_B = \nabla C^{PB}(0)$ and $q_B = \nabla C^{TB}(0)$. The public observation and hidden target maps are $A = \mathrm{diag}(U, U, U, U)$ and $H = (V, 0, -V, 0)$. This gives an intervention-specific consequence of Theorem 1.

Corollary 2 (Finite-panel insufficiency for unrestricted local response)**.** *For $F = \mathbb{R}^{4d_h}$, if $\mathrm{row}(V) \nsubseteq \mathrm{row}(U)$, then the exact public directional derivatives do not identify the hidden patched-minus-target linear response $V(p - q)$. In particular, a hidden displacement with a component outside the public span creates a claim-changing blind direction.*

Repeatedly measuring the same public span more precisely cannot repair this ambiguity; new target-relevant directions are required. Retaining only the selected contrast, $A_g = (U, -U, -U, U)$, can lose additional bypass information, whereas the full public records recover $Uv = Up - Uq - Uc$. Acquisition and retention in a decision statistic are therefore separate design choices.

From derivatives to finite responses. Linear observability alone does not certify a finite intervention response. Let $D_{\mathrm{pub}} = \max(\|Up\|_2, \|Uq\|_2, \epsilon)$ and $R = \|Uc\|_2 / D_{\mathrm{pub}}$. Suppose c lies in a declared subspace S on which U observes V, with $\kappa = \|V_S U_S^\dagger\|_{2\to 2}$; suppose also $\|Vv\|_2 \leq \beta$. For $r_m = \delta(e_m) - (c + v)^\top e_m$, assume $\mathrm{RMS}_m(r_m) \leq \gamma$ on the actual hidden segments and endpoint denominator $D_{\mathrm{end}} \geq a > 0$. Then

$$E_s \leq \min\{2, (\kappa R D_{\mathrm{pub}} + \beta + \gamma)/a\}. \tag{5}$$

Thus finite-response control has four distinct requirements: observation geometry κ, bypass response β, nonlinear remainder γ, and endpoint scale a (Corollary 3). The predictive tests examine whether broader response information predicts withheld finite-response failures; they do not identify a unique cause for each failure. The reconstruction studies ask which additional observations are useful once acquired. Appendix B gives proofs and exact counterexamples.

3.4 FROM UNRESOLVED INFORMATION TO A TESTABLE PREDICTION

Motivated by the blind-direction analysis, we test an empirical consequence: whether broader response information adds predictive value for finite-response failure, $E_s > 0.20$, after center restoration. With $K = M = 8$, we compare public-direction and full-gradient discrepancies. Define $D_{\text{full}} = \max(\sigma\|p\|_2, \sigma\|q\|_2, \epsilon)$, with the executed scale $\sigma = 10^{-3}/\sqrt{768}$, also retained externally. Our primary comparison is

$$S_{\text{DD}} = \frac{\|U(p-q)\|_2}{D_{\text{pub}}}, \qquad S_{\text{GG}} = \frac{\sigma\|p-q\|_2}{D_{\text{full}}}. \tag{6}$$

DD measures public-direction mismatch; GG measures full first-order mismatch, both including bypass response through $p-q = c+v$. These magnitude summaries rank failures; identifying arbitrary signed responses requires the directional information they discard (Appendix E.1). Under a mean-zero displacement law with covariance $\sigma^2 I$, $\sigma\|p\|_2$ is population RMS linear response. The fixed σ cancels from GG when its denominator floor is inactive, including at external widths. Derivatives use float64, without fitted score weights or layer thresholds; Section 4.3 controls normalization (protocol: Appendix C).

4 OBSERVED SUCCESS CAN CONCEAL PREDICTABLE RESPONSE FAILURE

We test whether response information not retained by the public panel adds prospective predictive value for failure under withheld changes among center-restoring patches. Scores are frozen before hidden outcomes.

4.1 EXPERIMENTAL SETUP AND FROZEN EVALUATION

GPT-2 Small residual-stream post sites at layers 0–8 use IOI name-logit contrasts (Wang et al., 2023) and GT balanced suffix-logit contrasts (Hanna et al., 2023). Development used 3,744 sites; the disjoint prospective cohort contains 4,608 sites from 128 IOI/384 GT prompt clusters. Eligibility requires finite endpoints and $O_s \geq 0.8$; GT also requires higher clean probability mass above the year boundary than at/below it. Appendix C gives the selected-path instrument and split checks.

Primary estimand and uncertainty. We compare GG and DD by within-layer AUC for $E_s > 0.20$, weighting layers by their positive–negative site pairs:

$$\text{AUC}_{\text{within}} = \frac{\sum_\ell n_{\ell,+} n_{\ell,-} \, \text{AUC}_\ell}{\sum_\ell n_{\ell,+} n_{\ell,-}}. \tag{7}$$

Single-class layers contribute no pairs. The primary criterion requires positive lower bounds for both approximate 97.5% paired GG-minus-DD intervals, using 2,000 stratified prompt-cluster resamples before eligibility filtering and recomputing pair weights.

4.2 A RESTORED CENTER LEAVES PREDICTABLE RESPONSE DIFFERENCES

Of 1,152/2,420 eligible IOI/GT sites, 869/229 fail the hidden-response test despite passing center restoration. More stringently, 117 eligible IOI sites pass the exact public finite-response test ($E_{\text{pub}} \leq 0.20$) yet fail the withheld test; Figure 2A shows the member of this set with center restoration closest to one ($O_s = 1.0006$, $E_{\text{pub}} = 0.1133$, $E_s = 0.3080$). GG improves within-layer AUC over DD by 0.0802/0.0549, with both primary intervals above zero (Table 1a). Thus observed success can leave response distinctions for which broader information has prospective predictive value; the experiment does not attribute every failure to a particular orthogonal blind-space component.

4.3 PREDICTIVE GAINS PERSIST UNDER MEASUREMENT CONTROLS

Restoration error $|O_s - 1|$ gives IOI/GT AUCs of 0.5430/0.5726, whereas GT layer identity alone gives pooled AUC 0.9534. GG gains persist under equal-layer weighting, with a negative GT L7 contrast (Appendix C.2). At a common denominator, DG uses DD's public numerator with GG's normalization. GG−DG gains at the predeclared factor-100 endpoint are 0.0649/0.0344,

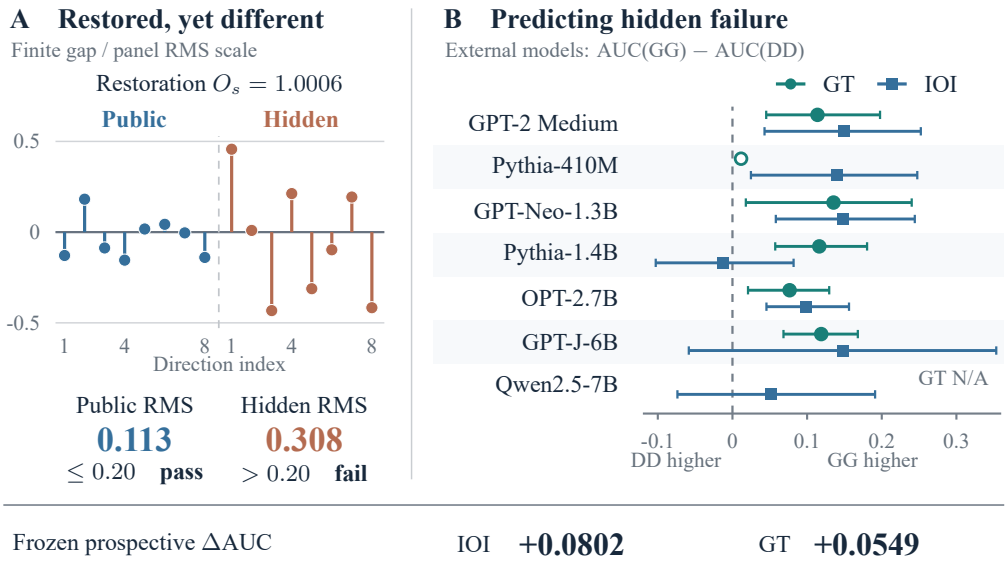


Figure 2: **Observed success can conceal hidden response failure. A**, Finite patched-minus-target responses at an IOI layer-5 site with nearly exact center restoration. Each panel uses its own RMS denominator (Eq. 2); the 0.20 threshold applies to aggregate error. Among 117 eligible IOI sites that pass the exact public test but fail the hidden test, this post-hoc illustration minimizes $|O_s - 1|$ (Appendix C.5). **B**, All executable external within-layer GG−DD AUC contrasts, with 97.5% intervals and no family-wise correction. The hollow GT interval fails the all-draw-validity rule; Qwen GT is unsupported by the readout. The bottom row and Table 1a give the frozen prospective comparison.

with both supporting 95% intervals above zero; GG also outperforms exact finite public responses (Appendix C.3). At 50% acceptance, IOI failure risk falls from 59.38% to 56.77%, while neither score accepts a GT failure (Appendix C.4).

4.4 Response information has recurring, task-dependent value

The frozen scores and factor-100 protocol, without model-specific fitting, yield positive GG−DD gains in twelve of thirteen executable conditions across seven model instances from six external families (Figure 2B). All six GT contrasts and six of seven IOI contrasts are positive; five GT and four IOI 97.5% intervals lie above zero. Pythia-1.4B IOI reverses, while the GPT-J and Qwen IOI intervals cross zero. This exploratory panel establishes recurrence with task- and model-dependent uncertainty; Appendix D details the interval-validity and readout conditions.

5 Evidence quality is target- and use-relative

The observability criterion is not a prescription to collect the largest possible response summary. It asks whether the observations remove the distinctions that matter for the declared target, and whether those distinctions can be used stably. We therefore test three common surrogates for evidential strength—variance, generic geometric matching, and span coverage—and construct settings in which each misranks predictive value (Figure 3). Separately, reference queries test when nonlinear response structure can be reused rather than remeasured. We reconstruct correct-answer-versus-competitor logit-margin vectors, distinct from the scalar failure-screening task above. Calibration uses both branches and reference queries at evaluation displacements; patched endpoints remain withheld. We report mean absolute/squared error (MAE/MSE); PC1 and OLS denote the first principal component and ordinary least squares.

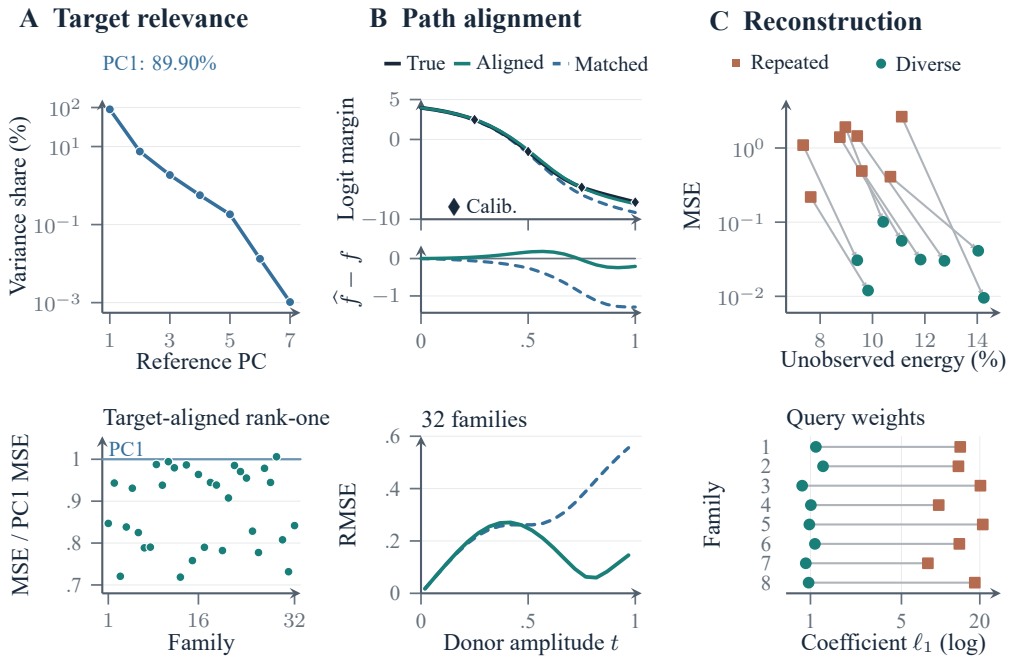


Figure 3: **Three "more information" heuristics can misrank evidential value. A**, *High variance $\not\Rightarrow$ high target value*: the normalized reference-remainder spectrum is compared with held-out rank-one/PC1 prediction error at fixed acquisition. **B**, *Matched geometry $\not\Rightarrow$ intervention relevance*: donor-path calibration (diamonds) is compared with moment-matched independent directions on fresh amplitudes; lower curves aggregate RMSE over 32 families. **C**, *More coverage $\not\Rightarrow$ more stable reconstruction*: repeated and diverse four-observation OLS allocations trade unobserved response energy against query-coefficient norm and MSE. Axes and constructions match the original protocols; details and stronger controls are in Appendices G–I.

5.1 Finite transfer depends on relative nonlinearity

An observed reference transfers finite-response information after correcting center and signed derivative differences. Here K counts competitors, not public directions. Let $f_b(v) \in \mathbb{R}^K$ be the vector of correct-answer versus competitor logit margins for branch b, with public center m_b, Jacobian J_b, and remainder $\mathcal{R}_b(v) = f_b(v) - m_b - J_b v$. Reference T and patched branch P share the injected activation and displacement, retaining their respective non-patch contexts. When $f_T(v)$ is available, signed relative derivatives define

$$\widehat{f}_P(v) = f_T(v) + m_P - m_T + (J_P - J_T)v = m_P + J_P v + \mathcal{R}_T(v). \tag{8}$$

Its error is exactly $\mathcal{R}_P(v) - \mathcal{R}_T(v)$: nonlinear branches can still transfer accurately when their remainders are close, even if individually large. With a measured-direction projector Π and $\Delta = J_P - J_T$, define $\widehat{f}_{P,\Pi}(v) = f_T(v) + m_P - m_T + \Delta\Pi v$. Its error is

$$f_P(v) - \widehat{f}_{P,\Pi}(v) = \underbrace{\Delta(I - \Pi)v}_{\text{unobserved linear response}} + \underbrace{\mathcal{R}_P(v) - \mathcal{R}_T(v)}_{\text{relative nonlinear remainder}}. \tag{9}$$

Appendix F.1 gives the derivation, relative-Hessian bound and squared-error cross term. The signed terms can reinforce or cancel, linking derivative coverage to the nonlinear structure supplied by reference queries.

A mechanism for reuse: shared nonlinearity. Reference-transfer error depends on the *difference* between branch remainders (Eq. 9), not on either branch being individually linear. Of 288 sites in 32 new IOI families, 201 from 29 families pass public center-correctness eligibility for both branches. At the primary relative radius 0.10, aligned reference transfer reduces MAE by 46.28% relative to

radial reconstruction; public scalar calibration improves it further, while target-row-space MAE is 0.04142 (Table 1b). All three fixed comparisons satisfy the prespecified absolute and relative gain criteria under six simultaneous intervals. Thus shared nonlinear structure can become useful evidence when transfer is measured rather than assumed (Appendix F.2).

Context can create or destroy that reuse. Within a separate 32-family cohort (96 L0/L4/L8 sites), H/N contexts match the injected activation, final-token row, and target distance while differing internally. Their remainders differ by 0.003983; all five fixed cohort contrasts are supported. At L8, a factorial contrast shows that the internal/final-row interaction cancels **67.9%** of additive reference error, with positive cancellation in all 32 families. Reference quality is therefore a property of the joint computation, not of the injected activation alone (Appendix F.3).

5.2 Three controlled reversals

High variance need not mean high target value. At equal information cost on 32 known families/96 sites, public reference/recipient remainder pairs define correction subspaces. Four fresh 64-direction public panels and one independent hidden panel test new directions within these families. PC1 captures 89.90% of reference variance, yet target-aligned rank-one correction reduces MSE by 16.99% and retains 97.68% of full correction's incremental gain over the mean correction (Figure 3A; Table 1c). Retention is $(L_{\text{mean}} - L_1)/(L_{\text{mean}} - L_f)$, where L_1 and L_f are rank-one and full-correction losses. All six fixed comparisons support target-aligned compression, although L8 contributes 99.67% of the absolute gain. The result is deliberately about *retention after acquisition*, complementing Corollary 1's measurement count for exact identification: variance describes what changes most, not what the target needs most (Appendix G).

Matched geometry need not mean intervention relevance. A rank-one correction calibrated on random directions has 2.1406 times the reference baseline's MSE on donor displacements. Four donor-aligned observations at amplitudes $0.25, 0.5, 0.75, 1$ outperform norm-, moment-, and layout-matched controls on 32 known families, three layers, five contexts, and 20 fresh query amplitudes. Donor calibration gives MSE 0.0310 versus 0.0985/0.1512 for matched independent directions/rays; both fixed contrasts and all 480 context-cell means favor alignment (Table 1d). The gap widens at

Table 1: **Prospective screening and controlled measurement-design results.** Bold marks the higher AUC within each task in a and the lowest displayed error within each of b–e; comparisons are meaningful within, not across, blocks.

a Prospective AUC ↑

Task	Public DD	Full-gradient GG	GG−DD	97.5% interval
IOI	0.7352	**0.8154**	+0.0802	[0.0470, 0.1152]
GT	0.8342	**0.8890**	+0.0549	[0.0175, 0.0968]

b Reference transfer

Method	MAE (10^{-3}) ↓
Target row space	41.424
Radial quadratic	12.180
Aligned reference	6.543
Public scalar calibration	**5.705**

c Target-aligned correction

Method	MSE (10^{-5}) ↓
Mean	5.1431
PC1	4.4469
Target rank-one	3.6916
Full	**3.6572**

d Direction alignment

Method	MSE ↓
Moment-matched indep.	0.098534
Moment-matched ray	0.151223
Donor affine	0.031018
Direct cubic	**0.000007517**

e Four-point allocation

Method	MSE ↓
Reference shift	0.036118
Repeated OLS	1.202821
Diverse OLS	0.038953
Diverse ridge	0.015364
Diverse scalar	**0.012509**

Compare values within blocks. Gains use unrounded AUCs. Rank retention uses Mean; acquisition is unchanged. Direct cubic matches total suffix-query count. Scope and uncertainty: Appendix A.

larger amplitudes (Figure 3B), and direct interpolation is the strongest tested predictor at the same total suffix-query budget. Matching generic geometry is therefore not equivalent to interrogating the change invoked by the claim (Appendix H).

More coverage need not mean more stable reconstruction. At a fixed four-observation budget, diverse allocation *increases* unobserved response energy from 9.19% to 11.71% on eight held-out families from the known pool, yet lowers the query-coefficient ℓ_1 norm from 14.88 to 1.02. OLS MSE simultaneously falls by 96.76%, improving all 120 family–layer–context means; ridge and scalar calibration improve further (Table 1e). The reversal holds in all eight family averages (Figure 3C). Coverage and numerical usability are therefore distinct: an observation set can span less of the response while producing a far more stable reconstruction. Because the allocation also changes finite-query locations and nonlinear remainders, the reduced coefficient norm is a concrete stability mechanism consistent with—not a sole causal attribution of—the MSE gain (Appendix I).

Taken together, these results sharpen the completion view: the goal is not maximal generic information, but target-relevant distinctions acquired and used with adequate stability.

6 Discussion

Three distinctions organize the evidence. The results separate *restoration* from *identification*, *identification* from *stable reconstruction*, and generic information volume from *evidential value*. A restored center can leave claim-changing directions unresolved; an identified target can still amplify measurement error; and higher variance or broader coverage can reconstruct less reliably when measurements are poorly aligned. These are different questions, not competing metrics of one notion of success.

The claim includes its admissible alternatives. The response class F is part of the scientific statement: narrowing it rules out alternatives by assumption, while broadening it can expose new blind directions. Changing the target or assessed displacements likewise changes what counts as sufficient evidence. Across four repeated direction panels, mean GG$-$DD gains remain positive, yet 37.0%/10.9% of eligible IOI/GT sites change failure labels; a donor study likewise shows that response preservation and answer correctness can favor different screens (Appendices E.1 and E.2). The assessment target is therefore part of the result, not merely an evaluation detail.

Evidence can be reused through relational structure. Reference transfer yields a second conclusion beyond missing-information diagnosis. Finite-response reuse does not require either branch to be nearly linear: Eq. 9 shows that transfer error depends on the *difference* between their nonlinear remainders. The context study then shows that this agreement is not a property of the injected activation alone; surrounding computation can create or destroy reuse. Thus reusable evidence is relational, and its transfer must itself be tested.

Validity assumptions and scope. Other validity analyses can restrict the admissible response class F; such restrictions may remove blind directions, but they are scientific assumptions rather than substitutes for checking whether the resulting observations identify H. Our exact results hold on a declared finite-dimensional linear response space and do not select the scientifically correct F or H. Finite-response control additionally requires bypass, nonlinearity, and endpoint-scale assumptions. The primary tests use local activation changes on canonical GPT-2 Small IOI/GT tasks; the external panel is exploratory, and random hidden directions are a controlled target rather than a complete proxy for semantic behavior. Full gradients are diagnostic evidence for the predicted information gap, not a certificate of arbitrary finite-response preservation. A stronger next step is to predeclare semantic or natural input-change targets and withhold their activation trajectories until evaluation.

7 Conclusion

Activation-patching evidence supports a functional claim only when its observations rule out admissible alternatives that would change it. Target-relative observability separates missing information from precision; prospective failures and controlled reversals show why restoration, variance, and coverage are insufficient proxies for evidential strength.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the response targets, frozen statistical procedure and reconstruction access. Appendices provide proofs, all-layer and coverage summaries, measurement controls, and protocols for reconstruction and computational-use studies. A local analysis-only entry regenerates reference, context, budget, workload-cost, rank, direction and allocation assets from saved analyses and compiles without model calls. Separate reconstruction scripts recompute primary AUC estimates and the new family-level transfer/context statistics from retained records; the display builder preserves stored intervals.

AI USE STATEMENT

The core research ideas originated with the authors, who led the reasoning, study design, and manuscript writing. Under author direction, generative AI tools assisted with literature search, method implementation and debugging, figure preparation, and text drafting and revision. They also supported refinement of theoretical framing and mathematical claims, proof drafting and review, feedback on experimental methodology, and interpretation of results. The authors retain full responsibility for the final scientific claims, proofs, experiments, citations, code, and presentation.

REFERENCES

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 210–227. PMLR, 2026. URL https://proceedings.mlr.press/v337/asiaee26d.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. doi: 10.1214/aoms/1177729442. URL https://doi.org/10.1214/aoms/1177729442.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_file

s/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b5afe13494c825089b1e3944fdaba212-Abstract-Conference.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76033–76060. Curran Associates, Inc., 2023. doi: 10.52202/075280-3322. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51157713f1f1f572a05647e1aef1b1ed-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S. Pasareanu, and Somesh Jha. Validating mechanistic interpretations: An axiomatic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47509–47544. PMLR, 2025. URL https://proceedings.mlr.press/v267/palumbo25a.html.

R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955. doi: 10.1017/S0305004100030401. URL https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0305004100030401.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

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*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026. doi: 10.48550/arXiv.2606.27510. URL https://arxiv.org/abs/2606.27510.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

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*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78205–78226. Curran Associates, Inc., 2023. doi: 10.52202/075280-3418. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction. *arXiv preprint arXiv:2606.09899*, 2026. doi: 10.48550/arXiv.2606.09899. URL https://arxiv.org/abs/2606.09899.

A Study design and comparison guide

The supplement separates mathematical sufficiency from empirical scope. Appendix B proves the target-relative identification, finite-panel, completion and stability statements, including realizable-query restrictions and finite-response counterexamples. Appendices C and D document the frozen prospective and exploratory external predictive evaluations. Appendix E.1 and Appendix E.2 show how changing assessed directions or behavioral targets changes the scientific estimand. The remaining sections support reference transfer, context interaction, target-aligned retention, intended-path calibration, allocation stability and computational-use studies.

Which comparisons share a target? Table 1a measures prediction of hidden response failures, with approximate two-task 97.5% intervals; the exploratory external panel is not a new simultaneous confirmatory family. Blocks b–e concern competition-margin reconstruction under different protocols, so losses are compared within blocks. Reference transfer uses new lexical families and allowed target-branch queries at evaluation displacements. The rank study retains full acquisition and tests compression on fresh directions of known families. The directional study compares intended and matched control directions, with direct interpolation matched in total suffix-query count. Allocation holds the four-observation budget fixed on held-out families within a known pool; its loss comparisons

are descriptive. The independently timed random-direction workload evaluates the cost and finite-workload guarantee of its own sharing rule.

B PROOFS AND SCOPE OF TARGET-RELATIVE OBSERVABILITY

Throughout the linear results, F is the finite-dimensional admissible parameter space fixed independently of tested outcomes. We use orthonormal coordinates on F and omit restriction subscripts when unambiguous; vector norms are Euclidean and matrix norms are induced $2 \to 2$ norms. The target map H is specified, whereas its value $H\theta$ is unknown.

B.1 EXACT IDENTIFICATION, MINIMAL COMPLETION AND STABILITY

Using the generalized-inverse identities of Penrose (1955), a feasible observation $y = A\theta$ leaves the affine set $A^\dagger y + \ker A$ of indistinguishable parameters. The target is constant on every such set exactly when $\ker A \subseteq \ker H$. This also proves the pairwise formulation in Section 3.2, since differences of admissible parameters range over F.

Proof of Theorem 1. If $Az = 0$ but $Hz \neq 0$, no finite constant can satisfy the asserted inequality. Conversely, kernel inclusion implies $H = HA^\dagger A$: the complementary projector $I - A^\dagger A$ has range $\ker A$. Consequently $\|H\theta\| \leq \|HA^\dagger\| \, \|A\theta\|$. The operator $HA^\dagger$ vanishes on $(\operatorname{range} A)^\perp$. If it is nonzero, a unit right singular vector attaining its norm can therefore be chosen in $\operatorname{range} A$ and written as $A\theta$. The inequality on that parameter requires $\kappa \geq \|HA^\dagger\|$, proving sharpness. If $HA^\dagger = 0$, the factorization gives $H = 0$ and the smallest nonnegative constant is zero. □

Proof of Corollary 1. Let $N = I - A^\dagger A$, the orthogonal projector onto $K = \ker A$. Any completing matrix B' with m rows must satisfy $\ker(B'|_K) \subseteq \ker(H|_K)$. Rank–nullity on K gives

$$m \geq \operatorname{rank}(B'|_K) \geq \operatorname{rank}(H|_K) = \operatorname{rank}(HN).$$

Conversely, if the $k = \operatorname{rank}(HN)$ rows of B form an orthonormal basis of $\operatorname{row}(HN)$, then $B^\top B$ projects onto that row space and $HNB^\top B = HN$. Thus $H = HA^\dagger A + HNB^\top B$, yielding Eq. 3 with exactly k additional scalar measurements. The zero-rank case needs no additional measurement. □

Proof of Corollary 2. For $A = \operatorname{diag}(U, U, U, U)$ and $H = (V, 0, -V, 0)$ on unrestricted F, the kernel condition is equivalent to $\ker U \subseteq \ker V$ for the patched and target gradient coordinates. For finite-dimensional Euclidean spaces, $\ker U \subseteq \ker V$ holds exactly when $\operatorname{row}(V) \subseteq \operatorname{row}(U)$. If the row inclusion fails, choose $z \in \ker U$ with $Vz \neq 0$ and set $\theta = (z, 0, 0, 0)$. Then $A\theta = 0$ but $H\theta = Vz \neq 0$, violating Theorem 1. □

Observation error does not change the identification criterion. Under kernel inclusion, measurements $y = A\theta + \eta$ and reconstruction $\hat{t} = HA^\dagger y$ obey $\hat{t} - H\theta = HA^\dagger \eta$, hence error at most $\kappa_* \|\eta\|$. For $A = \operatorname{diag}(1, \varepsilon)$ and $H = (1, 1)$, reconstruction uses $(1, 1/\varepsilon)$: any $\varepsilon > 0$ identifies the target, but error confined to the second measurement is amplified by $1/\varepsilon$. Improving precision can reduce this error; it cannot remove a direction in $\ker A$ that changes the target.

B.2 COMPLETION WITH REALIZABLE QUERIES

Let the rows of a finite dictionary D be the allowed scalar measurements, and set $M_S = [A; D_S]$ for a subset of row indices S. Define

$$k_D = \min\{|S| : \ker M_S \subseteq \ker H\}, \qquad \min \varnothing = \infty.$$

Proposition 1 (Restricted completion and residual uncertainty)**.** *Completion is possible exactly when* $\ker[A; D] \subseteq \ker H$. *When possible,* $k_D \geq \operatorname{rank}(HN)$*, and the inequality can be strict. For any subset, noisy observations* $y = M_S\theta + \eta$ *and estimate* $\hat{t} = HM_S^\dagger y$ *satisfy*

$$\|H\theta - \hat{t}\| \leq \|H(I - M_S^\dagger M_S)\| \, \|\theta\| + \|HM_S^\dagger\| \, \|\eta\|. \tag{10}$$

Proof. Apply Theorem 1 to M_S. If the full dictionary fails the kernel condition, every subset fails; if it passes, taking all rows suffices. The rank lower bound follows from unrestricted minimality. For strict inequality, take A with no rows, $H = (1, 0)$ and dictionary rows $(1, 1)$ and $(1, -1)$. An unrestricted query needs one row; neither dictionary row alone suffices, while their pair does. Finally, $H\theta - \hat{t} = H(I - M_S^\dagger M_S)\theta - HM_S^\dagger\eta$; the triangle inequality proves the bound. □

The two terms separate unresolved target variation from measurement-error amplification. An operational bound requires justified bounds on both $\|\theta\|$ and $\|\eta\|$. The criterion does not supply a target-optimal selector without knowledge of the target, and scalar-row count is not a forward-call count when a query returns multiple coordinates.

B.3 From linear information to finite response control

Retain the main-text definitions $\delta(u) = g(u) + b(u)$, $c = \nabla g(0)$, $v = \nabla b(0)$ and D_{pub}, with displacement magnitudes included in U, V.

Corollary 3 (Conditional finite-response bridge)**.** *Suppose $c \in S$, $\ker(U|_S) \subseteq \ker(V|_S)$ and $\kappa = \|V_S U_S^\dagger\|$. If $\|Vv\| \le \beta$, the remainders $r_m = \delta(e_m) - (c + v)^\top e_m$ satisfy $\mathrm{RMS}_m(r_m) \le \gamma$, and $D_{\mathrm{end}} \ge a > 0$, then*

$$E_s \le \min\{2, (\kappa R D_{\mathrm{pub}} + \beta + \gamma)/a\}.$$

Proof. Because V includes the factor $M^{-1/2}$, $\mathrm{RMS}_m \delta(e_m) \le \|Vc\| + \|Vv\| + \mathrm{RMS}_m(r_m)$. Observability on S bounds $\|Vc\|$ by $\kappa\|Uc\| = \kappa R D_{\mathrm{pub}}$. Divide by the endpoint denominator. Independently, the RMS triangle inequality bounds the discrepancy numerator by the sum of the two branch RMS responses, at most twice their maximum, so $E_s \le 2$ even with an active floor. □

One sufficient remainder condition is twice continuous differentiability on the actual hidden segments, with $|e_m^\top \nabla^2 \delta(te_m) e_m| \le L_m \|e_m\|^2$ for $0 \le t \le 1$ and $L_m \ge 0$. Taylor's integral formula gives

$$r_m = \int_0^1 (1 - t) e_m^\top \nabla^2 \delta(te_m) e_m \, dt, \qquad \gamma = \mathrm{RMS}_m(L_m \|e_m\|^2/2).$$

These are separate conditions: public-ray curvature does not bound hidden-segment curvature, and a small selected-path statistic alone does not bound bypass response or endpoint scale.

Exact counterexamples. These are smooth response-function constructions, not prevalence claims about a model. For a scalar displacement t, write a tuple in the order $(C^{PJ}, C^{PB}, C^{TJ}, C^{TB})$ and assess the derivative at zero and finite response at $t = 1$, with inactive floors. The tuple $(t, 0, -t, -2t)$ has $R = 0$ but $E_s = 2$ because of bypass mismatch. The tuple $(t, -t, t, 0)$ has $R = 1$ but $E_s = 0$ because selected and bypass discrepancies cancel. The tuple $(t - 2t^2, t, t, t)$ has zero selected derivative and bypass difference but $E_s = 2$ from curvature. All four branches agree at the center in each example.

For directional ambiguity, choose $w \ne 0$ orthogonal to all public displacements and a with $a^\top w \ne 0$. Start with $C^{PJ}(u) = C^{TJ}(u) = a^\top u$ and $C^{PB}(u) = C^{TB}(u) = 0$. Add $-2(a^\top w) w^\top u/\|w\|^2$ to both C^{PJ} and C^{PB}. Every public-ray branch value and the selected contrast remain unchanged, but the patched full response along w reverses sign, changing its hidden error from zero to two when the floor is inactive. Repeating or refining measurements on those same rays cannot distinguish the constructions; measuring a new direction can.

B.4 What scalar scores retain and discard

Simultaneously replacing (p, q) by $(-p, -q)$ leaves DD and GG unchanged while reversing every signed linear response discrepancy. Thus these screening summaries cannot identify arbitrary signed responses. Nor does GG alone determine the distribution of a finite-panel normalized error.

Proposition 2 (Geometry of the finite-panel response law)**.** *For nonzero p, q, inactive floors, $M \ge 1$ and independent uniform spherical directions in dimension $d \ge 2$, the linearized endpoint law depends on (r, ρ, d, M), where $r = \min(\|p\|, \|q\|)/\max(\|p\|, \|q\|)$ and $\rho = p^\top q/(\|p\|\|q\|)$. It is not determined by $S_{\mathrm{GG}} = \sqrt{1 + r^2 - 2r\rho}$ alone.*

Proof. Branch exchange, common scaling and rotation allow $p = e_1$ and $q = r(\rho e_1 + \sqrt{1-\rho^2}e_2)$. Let (X_i, Y_i) be the first two coordinates of independent uniform unit vectors in $\mathbb{R}^d$ and put $Q_i = \rho X_i + \sqrt{1-\rho^2}Y_i$. Then

$$E_{\mathrm{lin}} \overset{d}{=} \frac{\mathrm{RMS}_i(X_i - rQ_i)}{\max\{\mathrm{RMS}_i\, X_i,\ r\,\mathrm{RMS}_i\, Q_i\}}. \tag{11}$$

For $q = 0.8p$, both GG and E_{lin} equal 0.2 almost surely. For unit p, q with $p^\top q = 0.98$, GG again equals 0.2, but panels near the direction $p + q$ give errors near zero and panels near $p - q$ give errors near two. Both neighborhoods have positive probability for every finite M, so the laws differ. □

For unit directions $u_1, \ldots, u_M$, if each nonzero $z \in \{p, q, p-q\}$ satisfies $M^{-1}\sum_i (z^\top u_i)^2 \in [(1-\eta), (1+\eta)]\|z\|^2/d$ for $0 < \eta < 1$, taking square roots in the numerator and denominator yields

$$S_{\mathrm{GG}}\sqrt{\frac{1-\eta}{1+\eta}} \le E_{\mathrm{lin}} \le S_{\mathrm{GG}}\sqrt{\frac{1+\eta}{1-\eta}}.$$

This is conditional, not a claimed probability guarantee for eight directions. The spherical law is an ideal sampling model: actual payloads are quantized to float32 and evaluated in float64. Active floors and finite nonlinear remainders require separate treatment; repeated-panel evidence is reported in Appendix E.1.

B.5 Task directions and natural-context changes

For a donor displacement d, define centered changes $P(t) = C^{PJ}(td) - C^{PJ}(0)$ and $T(t) = C^{TJ}(td) - C^{TJ}(0)$. For continuously differentiable branches, the fundamental theorem of calculus gives

$$P(1) - T(1) = \int_0^1 [\nabla C^{PJ}(td) - \nabla C^{TJ}(td)]^\top d\, dt. \tag{12}$$

Its normalization depends on response scale along this particular direction. For example, $C^{PJ}(x, y) = 1 + x + \eta y$ and $C^{TJ}(x, y) = 1 + x$ have GG tending to zero as $\eta \to 0$, while a y-direction change has normalized discrepancy one whenever the floor is inactive.

Here let N denote the same fixed contrast's centered change under the complete natural donor input, rather than under an activation transplant in the clean target context. Adding and subtracting $T(1)$ gives

$$P(1) - N = \underbrace{P(1) - T(1)}_{\text{local response gap}} + \underbrace{T(1) - N}_{\text{context gap}}. \tag{13}$$

The signed terms can reinforce or cancel. Local response agreement therefore does not by itself bound the effect of the remaining context changes induced by a natural input change.

C Predictive evaluation: protocol and supporting comparisons

C.1 Tasks, measurements and frozen evaluation

The primary study uses GPT-2 Small with fixed weights and tokenizer, residual-stream post sites at layers 0–8, and ten fixed final-position layer-10 MLP neurons for the selected-path instrument. IOI reads the indirect-object name logit minus the subject-name logit. For suffix logits L_j and clean year boundary y, GT reads $\frac{1}{99-y}\sum_{j=y+1}^{99} L_j - \frac{1}{y+1}\sum_{j=0}^{y} L_j$. Its separate clean-validity rule compares summed full-softmax probabilities of suffixes above and at/below y; it does not require a correct argmax answer. Restoration uses Eq. 1, with clean-minus-corrupt denominators of absolute value below 10^{-12} excluded. Patched and target contexts retain their own bypass center anchors.

Each site has eight public and eight hidden directions of width 768 and nominal norm 10^{-3}. Frozen float32 direction payloads are cast into float64 model evaluation; the score convention is $\sigma = 10^{-3}/\sqrt{768}$. Public and hidden directions retain their task/site bindings and are selected without endpoint outcomes. Prediction processes do not receive the hidden payloads. All 4,608 prediction

records and 384 GT validity records were committed before separate processes collected hidden-direction endpoints. The prospective split shares no cluster keys, exact clean-token sequences or reversed pairs with the 3,744-site development collection.

All 1,152 IOI sites are eligible, with 869 hidden-response failures. GT retains 2,420 of 3,456 sites after restoration conditioning, with 229 failures; all 384 GT prompts pass the clean-validity gate. The other 1,036 GT sites are outside the restoration-conditioned estimand. All predictions, endpoints and validity records were available; no numerical-invalid record or normalization-floor hit occurred. The primary comparison and intervals are in Table 1a.

Statistical units and uncertainty. The within-layer AUC in Eq. 7 weights each layer by its number of positive–negative site pairs; score ties receive half credit. Following the bootstrap principle of estimating a sampling distribution by resampling (Efron, 1979), our 2,000 paired draws resample prompt clusters within original strata before applying eligibility and recomputing pair weights (seed 20260908). IOI has two strata of 64 clusters; GT has 48 noun/century/distance strata of eight clusters. All layers of a prompt share its resampling weight. The two-task primary criterion uses percentile quantiles $(0.0125, 0.9875)$ for each GG$-$DD contrast; supporting comparisons use $(0.025, 0.975)$. A single-class layer has no AUC, and a full layer interval is withheld if any draw lacks both classes. Both task-level primary contrasts have 2,000 valid draws.

Implementation checks. Saved records retain raw public/hidden effects, gradients, scores, restoration and validity. A separate NumPy calculation reproduces DD and endpoint error exactly and GG to 2.22×10^{-16}; an alternate weighted-AUC calculation reproduces the primary estimates. These are arithmetic consistency checks of the float64 empirical instrument, rather than formal numerical enclosures.

C.2 Layer heterogeneity and restoration controls

Table 2 reports every layer, including negative contrasts and unavailable intervals. GT L5–L6 contribute 71.53% of primary pair weight; the L7 contrast is negative. To separate directional information from restoration and layer weighting, Table 3 compares $|O - 1|$, $-O$, and equal-valid-layer means on the same eligible population. These post-hoc checks retain positive mean GG$-$DD differences with equal layer weights (0.0820 IOI, 0.0303 GT). The selected-path controls use $\|Uc\|_2/D_{\mathrm{pub}}$ and $\sigma\|c\|_2/D_{\mathrm{full}}$; their prospective AUCs are 0.5330/0.7806 and 0.5196/0.8639 (IOI/GT). Exact finite public-response AUC equals DD to the displayed precision at the primary radius.

Table 2: **All-layer diagnostics.** GG−DD AUC differences and supporting 95% intervals. Intervals are not adjusted for layer-wise multiplicity; – denotes unavailable, not zero.

		Eligible population		GG−DD contrast		
Task	Layer	Eligible	Failures	Difference	95% interval	Valid draws
IOI	0	128	31	+0.0918	[+0.0341, +0.1586]	2000/2000
IOI	1	128	95	+0.0839	[+0.0289, +0.1432]	2000/2000
IOI	2	128	98	+0.0065	[−0.0776, +0.0923]	2000/2000
IOI	3	128	105	+0.1925	[+0.0936, +0.2986]	2000/2000
IOI	4	128	119	+0.1008	–	1999/2000
IOI	5	128	111	+0.0567	[−0.0466, +0.1725]	2000/2000
IOI	6	128	109	+0.1028	[−0.0119, +0.2121]	2000/2000
IOI	7	128	108	+0.0255	[−0.0665, +0.1207]	2000/2000
IOI	8	128	93	+0.0777	[−0.0231, +0.1735]	2000/2000
GT	0	378	0	–	–	0/2000
GT	1	371	1	+0.0054	–	1313/2000
GT	2	370	1	+0.0054	–	1303/2000
GT	3	370	1	+0.0217	–	1337/2000
GT	4	362	9	+0.0264	–	1999/2000
GT	5	287	41	+0.0776	[+0.0187, +0.1416]	2000/2000
GT	6	138	59	+0.0468	[−0.0160, +0.1117]	2000/2000
GT	7	75	60	−0.0300	[−0.1684, +0.1139]	2000/2000
GT	8	69	57	+0.0892	[−0.0167, +0.2037]	2000/2000

Table 3: Post-hoc restoration and layer-weighting checks on the observed prospective population. No new confirmatory intervals are assigned. Equal-layer means use only layers containing both endpoint classes (nine IOI, eight GT); GT L0 remains unavailable.

	Pair-weighted AUC ↑		Equal-layer AUC ↑	
Score	IOI	GT	IOI	GT
GG	0.8154	0.8890	0.8048	0.9059
DD / finite public	0.7352	0.8342	0.7228	0.8756
Selected-path, full-space	0.5196	0.8639	0.5251	0.8756
$\lvert O-1\rvert$	0.5430	0.5726	0.5212	0.6941
$-O$	0.4770	0.5569	0.5014	0.3649

C.3 Scale, normalization and finite public-response controls

The scale extension reuses the prospective sites, public/hidden orientations, scores and center-based eligibility, multiplying displacements by $1, 10, 100, 1000$. Factor 100 is the predeclared primary extension; other factors are supporting checks. DG combines DD's public numerator with GG's denominator, isolating their directional-information difference at a common scale. The finite-public comparator evaluates exact response increments on the public directions at the tested scale. Table 4 gives the complete fixed factor-100 contrasts, using the original paired cluster-bootstrap procedure.

At factor 100, median displacement relative to activation norm is 0.1151% on IOI and 0.1563% on GT. Only two IOI labels change and no GT label changes, explaining the identical GT AUC across these radii. At factor 1000, the relative medians are 1.1506%/1.5628%, with eight/three label changes. Writing r^P, r^T for finite-minus-linear branch increments at this radius, define $Q_b = \max\{\mathrm{RMS}(r^P), \mathrm{RMS}(r^T)\}/D_{\mathrm{end}}$ and $Q_\delta = \mathrm{RMS}(r^P - r^T)/D_{\mathrm{end}}$, using the same finite-response denominator. At factor 1000, median/maximum Q_b is 0.021932/0.084930 for IOI and 0.026987/0.171861 for GT; Q_δ medians are 0.004593/0.001129. These measured endpoint remainders characterize the tested local range; they are not the uniform curvature bounds assumed in Corollary 3.

Table 4: Factor-100 within-layer AUC contrasts. The primary DD contrasts use 97.5% intervals; DG and finite-public contrasts use supporting 95% intervals. All have 2,000 valid draws.

Task	GG minus comparator	AUC gain	Interval
IOI	DD	0.080628	[0.047344, 0.116063]
IOI	DG (common scale)	0.064925	[0.037622, 0.093009]
IOI	Finite public response	0.080720	[0.050318, 0.111712]
GT	DD	0.054860	[0.017487, 0.096787]
GT	DG (common scale)	0.034439	[0.011859, 0.060210]
GT	Finite public response	0.054666	[0.021991, 0.091498]

C.4 Screening at fixed acceptance budgets

At coverage a, each score accepts its $\lceil aN \rceil$ lowest-scoring eligible sites, with stable site identity breaking ties. Table 5 reports every frozen budget; 50% is the fixed supporting operating point. The reported risks are observed score-quantile selection risks. At 50% coverage GG accepts 327 failures among 576 IOI sites, versus DD's 342; both accept zero GT failures among 1,210 sites. These finite-sample operating points do not supply calibrated deployment thresholds.

Table 5: **Frozen descriptive acceptance budgets.** Failure risk (a fraction) and mean endpoint error are computed on each method's lowest-scoring eligible sites; lower is better.

		Failure risk		Endpoint error	
Task	Coverage	DD	GG	DD	GG
IOI	25%	0.458333	0.354167	0.212239	0.181247
IOI	50%	0.593750	0.567708	0.251484	0.234822
IOI	75%	0.696759	0.685185	0.283857	0.271580
IOI	100%	0.754340	0.754340	0.313138	0.313138
GT	25%	0.000000	0.000000	0.017163	0.014943
GT	50%	0.000000	0.000000	0.025181	0.023104
GT	75%	0.001653	0.000551	0.037850	0.035780
GT	100%	0.094628	0.094628	0.077202	0.077202

C.5 Selection and normalization of the Figure 2 example

Among the 1,152 eligible IOI sites, 117 pass the exact public-response test but fail the hidden test. Figure 2A selects, post hoc, the site minimizing $|O_s - 1|$ in this subset, breaking ties by saved site identifier. The selected layer-5 site has $O_s = 1.0006481$, $E_{\mathrm{pub}} = 0.1132501$ and $E_s = 0.3079897$. Each panel shows all eight saved finite patched-minus-target response increments in their original order, divided by its own denominator from Eq. 2; the plotted RMS therefore equals that panel's endpoint error. This selected example illustrates the mismatch; the prospective population comparison establishes predictive value.

D External models and endpoint applicability

The exploratory external panel uses the same prediction-before-endpoint contract, float64 evaluation, layers 0–8, restoration eligibility, $E_s > 0.20$ failure rule and within-layer pair weighting. Table 6 reports every model–task condition, eligible denominator and interval. The panel includes two sizes from the Pythia model suite (Biderman et al., 2023). Fixed-stratum prompt-cluster draws recompute eligibility and pair weights; a 97.5% interval is reported only if all 2,000 draws are valid. There is no family-wise correction across this post-hoc panel.

Width, implementation and readout. Activations and probes use each model's actual width d_h. The executed score retains $\sigma = 10^{-3}/\sqrt{768}$; although same-radius spherical probes have coordinate

Table 6: **Conditional cross-model transfer under the frozen external panel.** Entries are within-layer GG−DD contrasts. The panel spans six model families and 14 model–task entries: 13 executable cells plus one tokenizer-compatibility audit. Intervals are fixed-stratum cluster-bootstrap intervals when all 2,000 draws were valid.

		Sites		GG−DD contrast	
Task	Model	Raw/elig.	Fail.	Estimate	97.5% interval
GT	GPT-2 Medium	1152/1031	20	0.1139	[0.0452, 0.1977]
GT	Pythia-410M	1152/1106	3	0.0115	unavailable[†]
GT	GPT-Neo-1.3B	1152/1082	19	0.1354	[0.0179, 0.2401]
GT	Pythia-1.4B	1152/896	98	0.1163	[0.0573, 0.1803]
GT	OPT-2.7B	1152/1097	74	0.0766	[0.0208, 0.1296]
GT	GPT-J-6B	1152/914	210	0.1189	[0.0683, 0.1679]
GT	Qwen2.5-7B	–	–	–	N/A[‡]
IOI	GPT-2 Medium	288/288	163	0.1493	[0.0430, 0.2523]
IOI	Pythia-410M	288/288	116	0.1401	[0.0247, 0.2476]
IOI	GPT-Neo-1.3B	288/285	117	0.1477	[0.0581, 0.2443]
IOI	Pythia-1.4B	288/280	233	−0.0122	[–0.1028, 0.0819]
IOI	OPT-2.7B	288/288	149	0.0986	[0.0455, 0.1561]
IOI	GPT-J-6B	288/278	242	0.1480	[–0.0585, 0.3534]
IOI	Qwen2.5-7B	288/211	179	0.0510	[–0.0737, 0.1911]

[†]Formal interval withheld under the predeclared all-draw-validity rule; 1,916/2,000 draws were valid and the partial interval [−0.1251, 0.2018] is descriptive only. [‡]Qwen GT is outside the frozen readout contract: 0/100 exact two-digit suffixes are single tokenizer tokens. All executable cells use independent factor-100 endpoint directions and sealed outcome-blind strata. The expanded results are post-hoc exploratory and have no family-wise multiple-test correction. Cross-zero IOI intervals coincide with sparse non-failure classes in the GPT-J, Qwen and Pythia-1.4B cells; they mark uncertainty/boundaries rather than proof of no effect.

RMS $10^{-3}/\sqrt{d_h}$, this common positive scale cancels from GG whenever its denominator floor is inactive. Implementation checks cover hook shape, no-op and center-anchor identity, projection semantics, finite gradients and intervention sensitivity. The finite-difference engineering gate was amended after initial qualification attempts; these checks support the executed exploratory protocol. Qwen GT is inapplicable under the frozen readout: none of the exact suffix strings $00, \ldots, 99$ is a single tokenizer token. Sequence scoring would define a different task/readout.

Induction applicability check. The same 32 fixed induction rows and layers 0–8 were evaluated with direct two-branch float64 responses at factor 100. Failures were 286/288 (GPT-2 Medium), 288/288 (Pythia-410M), 286/288 (GPT-Neo-1.3B) and 288/288 (Pythia-1.4B). With only two non-failures or none, these cells do not support informative failure-ranking comparisons. They delimit the current IOI/GT assessment protocol; thresholds and rows were not retuned after this outcome.

E How the assessed response changes the useful evidence

E.1 Repeated direction panels

Proposition 2 separates the scalar GG score from the distribution of a finite-panel response test. Four newly drawn hidden panels, each containing eight nominal-norm-10^{-3} directions per site, test repeatability on the same 4,608 collected sites. All original scores, eligibility and model choices remain fixed. The primary estimand averages the four within-layer AUC contrasts; it does not pool all 32 responses before thresholding. The 2,000 prompt-cluster draws are shared across panels and scores and condition on these four panels. Table 7 reports all panel results and mean contrasts.

The histograms of failures across zero through four panels are $(82, 79, 115, 232, 644)$ for IOI and $(2054, 118, 75, 70, 103)$ for GT. Thus 37.0%/10.9% of eligible IOI/GT sites change status with the panel even though both mean ranking gains remain positive. Only one IOI and two GT site–panel labels differ between finite and linearized endpoints at this scale.

Table 7: Four newly generated hidden panels on the same prospective sites. Mean contrasts use approximate 97.5% primary intervals; shared-scale and finite-public controls use supporting 95% intervals.

Task	Contrast / panel	AUC gain	Interval
IOI	Panel 1, GG−DD	0.0993	descriptive
IOI	Panel 2, GG−DD	0.0553	descriptive
IOI	Panel 3, GG−DD	0.1043	descriptive
IOI	Panel 4, GG−DD	0.0965	descriptive
IOI	Mean GG−DD	0.0888	[0.0690, 0.1113]
IOI	Mean GG−DG	0.0726	[0.0564, 0.0903]
IOI	Mean GG−finite	0.0888	[0.0712, 0.1084]
GT	Panel 1, GG−DD	0.0941	descriptive
GT	Panel 2, GG−DD	0.1233	descriptive
GT	Panel 3, GG−DD	0.1179	descriptive
GT	Panel 4, GG−DD	0.0909	descriptive
GT	Mean GG−DD	0.1066	[0.0783, 0.1388]
GT	Mean GG−DG	0.0668	[0.0473, 0.0883]
GT	Mean GG−finite	0.1067	[0.0827, 0.1342]

E.2 Donor screening: response discrepancy and answer error

Eight new prompt families per task supply nine sites per family, four public donors and four held-out donors. There are 67/72 eligible IOI sites and 50/72 eligible GT sites. Screening compares families within each layer, with at least two eligible families required: all nine IOI layers and seven GT layers qualify, covering 67 and 49 sites respectively. This separates the comparison from selecting earlier versus later layers within a prompt. It is a descriptive qualification study; center-margin controls were added post hoc.

For natural and patched full-vocabulary probability vectors p, b and their centers p_0, b_0, define $R_{\text{full}} = \|(b - b_0) - (p - p_0)\|_1/4$ and $\mathrm{TV}(b, p) = \|b - p\|_1/2$. For valid token set G, the margin is $m_G(p) = \max_{i \in G} p_i - \max_{j \notin G} p_j$. Added error is $\mathbf{1}\{\text{natural valid, patched invalid}\}$; net loss subtracts corrections from added errors. IOI validity means the exact donor name is the top-1 token; GT validity means top-1 belongs to the donor-valid suffix set. Exact-name mismatches do not by themselves determine the semantic validity of a continued response.

Scores and acceptance. Lower scores are accepted first. Public full response averages public-donor R_{full}, public error averages added-error indicators, and the center patched-margin score is $-m_G(b_0)$ for the original answer set. Center probability uses $\mathrm{TV}(Qb_0, Qp_0)$, where Q retains the task tokens and an OTHER bin. GG and DD retain their original definitions. For n_ℓ eligible families in layer ℓ, let b_i count strictly lower scores and t_i count tied scores. At expected coverage a, acceptance weight is

$$w_i(a) = \min\{1, \max\{0, (an_\ell - b_i)/t_i\}\}.$$

Loss L_i first averages hidden donors within a site; each layer contributes $\sum_i w_i(a) L_i/(an_\ell)$, and valid layers are averaged equally. The fixed descriptive point is $a = 0.5$; 25% and 75% are sensitivity checks.

Center patched-margin screening has fewer added IOI errors than public full-response screening (0.145833 versus 0.208333), but greater response discrepancy (0.080551 versus 0.055588). Public-error screening gives the lowest displayed non-oracle net loss (0.117130), because net loss credits corrections; all five observed IOI corrections belong to one family. GT favors center probability for response discrepancy, while the eligible GT sample contains no added top-1 errors to discriminate. The small IOI response-loss advantage of public full response over center probability is family-sensitive: its leave-one-family-out range at 50% coverage is $[-0.003629, 0.000168]$. Including restoration-ineligible sites also changes some rankings, so the comparison concerns the declared eligible population. Center rules use the common 11-forward center acquisition; acquiring four public donors across nine layers adds 40 forwards for public full response. The results identify target-dependent evidence value, without establishing a uniformly preferable screening rule.

Table 8: Response and task outcomes at 50% expected within-layer coverage. Lower is better. IOI added error and net loss are fractions. All eligible GT added errors and net losses are zero. Each oracle column separately orders sites by that column's hidden loss; these unavailable selectors are not one common policy.

	IOI			GT	
Rule	R_{full}	Added error	Net loss	R_{full}	TV
Center probability (Q)	0.0569	0.2282	0.2123	0.0185	0.0360
GG	0.0600	0.2490	0.2143	0.0210	0.0424
DD	0.0609	0.2371	0.2063	0.0205	0.0410
Public full response	0.0556	0.2083	0.1696	0.0202	0.0402
Public added-error rate	0.0645	0.1544	0.1171	0.0213	0.0412
Center patched margin	0.0806	0.1458	0.1300	0.0215	0.0386
Uniform	0.0778	0.2336	0.2143	0.0213	0.0412
Hidden oracle	0.0547	0.0754	0.0367	0.0179	0.0332

F FINITE REFERENCE TRANSFER AND CONTEXTUAL RESPONSE

F.1 ERROR DECOMPOSITION AND RELATIVE CURVATURE

For a fixed public competitor bank, let $f_b(v) \in \mathbb{R}^K$ denote answer-minus-competitor logits with the patch activation displaced by $v \in \mathbb{R}^d$ and the non-patch context held fixed. Write $m_b = f_b(0)$, $J_b = D_v f_b(0)$, $\mathcal{R}_b(v) = f_b(v) - m_b - J_b v$ and $\Delta = J_P - J_T$. Substituting these definitions into Eq. 8 gives $\widehat{f}_P = m_P + J_P v + \mathcal{R}_T$, hence $f_P - \widehat{f}_P = \mathcal{R}_P - \mathcal{R}_T$. Replacing Δv by $\Delta\Pi v$ for an orthogonal measured-direction projector Π gives Eq. 9 by the same expansion. For $a(v) = \Delta(I-\Pi)v$ and $b(v) = \mathcal{R}_P(v) - \mathcal{R}_T(v)$,

$$\|f_P(v) - \widehat{f}_{P,\Pi}(v)\|_2^2 = \|a(v)\|_2^2 + \|b(v)\|_2^2 + 2\langle a(v), b(v)\rangle.$$

Thus omitted linear response and relative nonlinearity can reinforce or cancel. For $v = \rho u$ with u uniform on the unit sphere, $\mathbb{E}uu^\top = I/d$ additionally gives $K^{-1}\mathbb{E}\|a(v)\|_2^2 = \rho^2\|\Delta(I-\Pi)\|_F^2/(dK)$.

If each margin is twice continuously differentiable on the actual displacement segment, Taylor's integral remainder gives

$$f_{P,j}(v) - \widehat{f}_{P,j}(v) = \int_0^1 (1-t)v^\top \left[D^2 f_{P,j}(tv) - D^2 f_{T,j}(tv)\right] v\, dt. \tag{14}$$

A segmentwise operator-norm bound L_j on the Hessian difference therefore bounds the error by $L_j\|v\|_2^2/2$. The experiments measure endpoint remainders, rather than a certified supremum L_j.

F.2 REFERENCE-TRANSFER PROTOCOL AND FIXED COMPARISONS

The study used 32 IOI lexical families balanced over four grammar strata, with names disjoint from the preceding 16-family competitor-response development cohort. Grammar templates and GPT-2 Small remained shared. All 288 sites at layers 0–8 were collected at relative radii $0.03/0.10$, with 0.10 primary. Public eligibility required correct exact-name answers at both target and patched centers, retaining 201 sites from 29 families. Each site used eight public and sixteen hidden normalized Gaussian directions, scaled by the specified fraction of the patch-center norm, and a fixed bank of 4–6 competitors from public centers. Directions were reused across radii within each phase. Banks and fitted coefficients were completed before the hidden phase; hidden target responses were allowed reference queries, whereas hidden patched responses were reserved for evaluation.

Table 9 retains all eight methods. The radial control adds the public mean remainder at the corresponding radius. The target-row-space control uses the target Jacobian's actual rank of 4–6. Public scalar calibration replaces $\mathcal{R}_T(v)$ by $\beta\mathcal{R}_T(v)$, where $\beta = \langle \mathcal{R}_P, \mathcal{R}_T\rangle_{\text{public}} / \max(\|\mathcal{R}_T\|^2_{\text{public}}, 10^{-24})$. No coefficient or bank is chosen using hidden patched responses.

Absolute errors are averaged over competitors and sixteen directions, then over eligible families within layers, and finally equally over nine layers. Each of the three prespecified control-minus-reference comparisons at $r = 0.10$ required an absolute gain above 0.001 and a relative gain above 20%. The six approximate 99.1667% Bonferroni intervals use 2,000 grammar-stratified family resamples, preserving layers, methods and radii within families; all draws were valid (Table 10). The smaller radius is supportive, not an independent primary trial.

Table 9: Competition-margin prediction on 32 new lexical families. Per-site direction/bank MAE is averaged within layers and then across nine layers; 201 eligible sites from 29 families. Smaller is better. Reference access means the target branch is evaluated at the query displacement.

		Margin MAE (↓)	
Method	Query reference	$r = 0.03$	$r = 0.10$
Patch linear	No	0.003628	0.040107
Radial quadratic	No	0.001091	0.012180
Reference center	Yes	0.013714	0.045642
Random eight directions	Yes	0.013620	0.045320
Eight finite observations	Yes	0.013620	0.045325
Target row space	Yes	0.012404	0.041424
Aligned reference	Yes	0.000591	0.006543
Public scalar calibration	Yes	0.000511	0.005705

Table 10: Prespecified comparisons at $r = 0.10$. All six approximate 99.1667% intervals use 2,000 grammar-stratified family draws. e_C and e_F denote control and aligned-reference MAE.

		Adjusted MAE contrast intervals	
Control C	Reduction	$e_C - e_F$	$0.8e_C - e_F$
Radial quadratic	46.28%	[0.004211, 0.006766]	[0.001849, 0.004276]
Eight finite observations	85.56%	[0.033185, 0.044943]	[0.025335, 0.034531]
Target row space	84.20%	[0.029820, 0.039912]	[0.022620, 0.030585]

Aligned reference transfer reduces MAE by 46.28% against radial reconstruction; public scalar calibration is stronger still. The radial comparison favors aligned reference in 186/201 sites and all 29 eligible family averages; including all 288 sites gives reference/radial MAE $0.006781/0.011923$. The fixed direction-mismatch control cycles the target remainder by one hidden-direction index: mismatch-minus-aligned MAE is 0.010249 with a supporting 95% family interval $[0.009380, 0.011069]$. Lower margin error does not imply uniform answer-screening superiority: at $r = 0.10$, reference and radial prediction make eight and ten classification mistakes, respectively, but their post-hoc paired improvement interval includes zero. Direct patched evaluation returns full-vocabulary outputs; the estimator predicts the declared margin bank using paid reference queries.

F.3 Independent lexical confirmation of context effects

The context confirmation used 32 new families and 64 distinct single-token names absent from the reference-transfer cohort and four preceding development cohorts. Names were selected by a seeded permutation after tokenizer-only filtering, without model-outcome screening; the four grammar templates, GPT-2 Small checkpoint and float64 route remained shared. Layers 0/4/8, radii $0.03/0.10$, eight public/sixteen hidden directions and seed 2026091391 were fixed; 0.10 was the sole primary scale. The six states shared each site's public-center bank of 4–7 competitors, directions and displacement norm.

Matched contexts and factorial response. Let T, C be clean/corrupt residual matrices, p the independently injected patch row, and j the final-token row. Set $D = C - T$ with $D_p = 0$. H has non-patch rows $T + 0.5D$; P cycles $D_{p+1:j}$ by one row before half-scaling. I equals P with final row j restored to H, and O equals H with row j replaced by P. N cycles only $D_{p+1:j-1}$, keeping row j fixed. T is the reference state, and all states use the same injected activation $h_0 + v$. H/N

consequently match the injected activation, injection-layer final row and global target distance; I/O are not distance-matched. Although $P = I + O - H$ as state matrices, their remainders can interact:

$$\mathcal{Q} = \mathcal{R}_P - \mathcal{R}_I - \mathcal{R}_O + \mathcal{R}_H, \qquad \mathcal{R}_{\text{add}} = \mathcal{R}_I + \mathcal{R}_O - \mathcal{R}_H.$$

Writing $e_b = \langle|\mathcal{R}_b - \mathcal{R}_T|\rangle$ and $e_{\text{add}} = \langle|\mathcal{R}_{\text{add}} - \mathcal{R}_T|\rangle$, define cancellation $C = e_{\text{add}} - e_P$. Angle brackets average measured directions and competitors. The additive response is an algebraic reconstruction of measured outputs, not another executed intervention.

Population, inference and interpretation. All 96 structural sites enter the primary analysis regardless of correctness. Internal contrasts average families within each layer and then the three layers equally; L8 contrasts use all 32 L8 families. For $D_{NH} = \langle|\mathcal{R}_N - \mathcal{R}_H|\rangle$ and $q = \langle|\mathcal{Q}|\rangle$, the five prespecified effect-size contrasts appear in Table 11. The same 20,000 grammar-stratified family draws (seed 2026091392) preserve layers, scales and states; all draws were valid. The five 99% marginal intervals target approximate 95% family-wise coverage. Thresholds were set from development for this confirmation, not as universal constants.

Table 11: Five prespecified new-lexical context contrasts, $r = 0.10$. Internal contrasts use all 96 sites; L8 contrasts use all 32 families. Bonferroni intervals (99% each; approximate 95% family-wise) use 20,000 grammar-stratified family draws.

Contrast	Estimate	Adjusted interval
$D_{NH} - 0.0005$	0.003483	[0.002721, 0.004300]
$D_{NH} - 0.2e_H$	0.003232	[0.002520, 0.003998]
$\langle|\mathcal{Q}|\rangle - 0.001$ (L8)	0.007392	[0.004480, 0.010602]
$C - 0.001$ (L8)	0.005533	[0.003258, 0.007965]
$C - 0.2e_{\text{add}}$ (L8)	0.004608	[0.002885, 0.006474]

All five adjusted intervals have positive lower endpoints. The matched H/N remainder difference is 0.003983; at L8, $C/e_{\text{add}} = 0.678585$, with post-hoc 95% family interval $[0.611750, 0.739213]$, and all 32 families have positive cancellation. N lowers reference MAE in 28/32 and 30/32 families at L0/L4, but only 3/32 at L8: context influence is not uniformly beneficial. Public scalar calibration is the strongest layer-macro MAE estimator in every state, and its N-minus-H changes retain the layerwise sign reversal.

The matched-summary ambiguity has a direct bound. For any shared correction $q_0(v)$ depending only on the matched summaries and v, the coordinatewise triangle inequality gives

$$\frac{\langle|\mathcal{R}_H - q_0|\rangle + \langle|\mathcal{R}_N - q_0|\rangle}{2} \geq \frac{D_{NH}}{2}. \tag{15}$$

The confirmation bound is 0.001991 with transformed interval $[0.001611, 0.002400]$ on this finite panel. H/N have different Jacobians, so the result does not exclude prediction from richer context or derivative information. The measured continuous-response interaction does not establish answer repair.

G TARGET-ALIGNED LOW-RANK RESPONSE CORRECTION

Protocol and information access. We use the 32 known IOI families, four grammar strata, all 96 L0/L4/L8 sites, and five recipient contexts H/P/I/O/N from Appendix F.3. The checkpoint is GPT-2 Small and arithmetic is float64. At each site, four independent 64-direction public panels precede one independent 64-direction hidden panel shared by all fits. Normalized Gaussian directions in 768 dimensions use PCG64DXSM keys $[2026091467, \text{ordinal}, \text{layer}, \text{hiddenFlag}, \text{replicate}]$; the hidden replicate is zero. Radii $0.03/0.10$ share unit directions. The fixed competitor bank contains the incorrect subject name and the union of the branches' top-five center tokens, excluding the correct answer. Public calibration observes both reference and recipient responses; at each hidden displacement only the reference is available to the predictor. All public fits are saved before collection of hidden recipient labels.

Estimators and reduced-rank solution. For one recipient and radius, let $X, Y \in \mathbb{R}^{N\times K}$ contain public reference and recipient remainders as rows, with $D = Y - X$. Define column means $\mu = X^\top \mathbf{1}/N$, $a = D^\top \mathbf{1}/N$, and centered matrices $X_c = X - \mathbf{1}\mu^\top$, $D_c = D - \mathbf{1}a^\top$. With $\lambda = \max\{\|X_c\|_F^2/K, 10^{-24}\}$ and $A = X_c^\top X_c + \lambda I$, define

$$W_f = A^{-1}X_c^\top D_c, \qquad W_1 = W_f v v^\top, \qquad v \in \arg\max_{\|w\|=1} w^\top W_f^\top A W_f w. \tag{16}$$

For a column-vector query remainder x, the prediction is $x + a + W_1^\top(x - \mu)$, followed by the recipient's own center and full Jacobian response. Only the relative correction is rank constrained. Indeed, the penalized objective equals $C + \|A^{1/2}(W - W_f)\|_F^2$, with C independent of W. Since A is positive definite, multiplication by $A^{1/2}$ preserves rank; its leading singular component gives the rank-at-most-one minimizer above. The sign of v cancels and ties use the public solver output.

All controls use the same pairs and full-input penalty. Mean sets $W = 0$. For an orthogonal input projector S, set $Z = X_c S$ and $W_S = S(Z^\top Z + \lambda I)^{-1} Z^\top D_c$. Shared uses $S = \mathbf{1}\mathbf{1}^\top/K$, contrast uses $I - S_{\text{shared}}$, PC1 uses the leading public right-singular input projector, and full uses I. Shared and PC1 both fit vector outputs. Diagonal fits each coordinate with denominator $\|X_{c,\cdot j}\|^2 + \lambda$; reference uses x without correction or bias. Each ensemble averages two full-strength half-panel fits, recomputing means, projectors and penalties from its own 32 observations. The rank-one ensemble need not itself have rank one.

Inference and complete fixed comparisons. Losses first average hidden directions and bank coordinates, then weight the five contexts, three layers, four public fits and 32 families equally. The six predeclared intervals in Table 12 use 20,000 family bootstrap draws, sampling eight families within each grammar stratum (PCG64DXSM seed 2026091468); each family's repeated observations stay together. The Bonferroni percentile coverage is nominally simultaneous 95% and approximate. Every public fit improves macro MSE over PC1/shared, mean-to-full improvement is positive, and rank-one/full MSE and MAE ratios are both below 1.05.

Table 12: The six predeclared family-level MSE contrasts at $r = 0.10$. Positive favors the stated effect threshold. L_1, L_f denote rank-one and full losses; subscripts 0/4 select a layer. Percentile intervals use 20,000 grammar-stratified family draws, nominal 99.1667% individually and approximate joint 95% Bonferroni coverage. All lower endpoints are positive.

Contrast	Unit	Estimate	Adjusted interval	+/32
$0.95L_{\text{PC1}} - L_1$	10^{-6}	5.330	[2.187, 8.991]	22
$0.95L_{\text{shared}} - L_1$	10^{-6}	6.453	[2.644, 10.812]	22
$0.95L_{\text{PC1,ens}} - L_1$	10^{-6}	5.262	[2.097, 8.897]	21
$(L_{\text{mean}} - L_1) - 0.90(L_{\text{mean}} - L_f)$	10^{-6}	1.141	[0.342, 2.138]	25
$1.05L_{\text{mean},0} - L_{1,0}$	10^{-8}	2.442	[1.471, 3.711]	32
$1.05L_{\text{mean},4} - L_{1,4}$	10^{-7}	3.655	[2.628, 4.928]	32

Table 13: All twelve fixed methods on the 32 known families. Each method has four 64-observation public fits and shares the hidden panel. MSE and MAE are lower-is-better macro means; the primary radius is 0.10. Every method retains the full reference remainder.

	$r = 0.10$		$r = 0.03$	
Method	MSE (10^{-5})	MAE (10^{-3})	MSE (10^{-7})	MAE (10^{-4})
Reference	8.1901	3.4980	7.1622	3.2062
Mean	5.1431	2.5986	4.4288	2.3760
Shared	4.5652	2.4936	3.8142	2.2659
Contrast	3.8681	2.3871	3.2285	2.1653
PC1	4.4469	2.4775	3.7059	2.2502
Full	3.6572	2.3375	2.9995	2.1148
Diagonal	4.3819	2.4792	3.7195	2.2581
Shared ensemble	4.5666	2.4891	3.8255	2.2623
PC1 ensemble	4.4398	2.4706	3.7111	2.2445
Full ensemble	3.6725	2.3345	3.0263	2.1128
Target rank-one	3.6916	2.3454	3.0197	2.1213
Rank-one ensemble	3.7045	2.3410	3.0454	2.1184

PC1 retains 89.90% of reference variance, but target rank-one lowers its MSE 16.99% and retains 97.68% of full's incremental improvement over mean: $(L_{\text{mean}} - L_1)/(L_{\text{mean}} - L_f)$. All fifteen layer–context means favor rank-one over mean, shared and PC1; nevertheless, L8 accounts for 99.67% of the absolute gain over mean. Across the four public fits, layer gains over mean range from 0.9–2.0% at L0, 2.5–3.4% at L4 and 28.2–29.5% at L8. Only L0/L4 non-inferiority was predeclared; their positive means are descriptive. Full has the lowest MSE and its ensemble the lowest MAE. At $r = 0.10$, bank-minimum sign disagreements are 57 for rank-one, 56 for PC1 and 57 for full across 122,880 reused predictor–endpoint evaluations. Thus the squared-error advantage does not establish a discrete decision gain. The smaller radius reuses unit directions and is a scale check.

The executed rank solver first fits the full ridge map and then truncates it; full reference acquisition remains available. One calibration shared by five recipients plus Q reference queries costs $6N + Q$ suffix evaluations: 448 at $N = Q = 64$, versus 320 direct recipient evaluations, before center/Jacobian acquisition and algebra. This is a matched-acquisition test of correction structure, not evidence of cheaper acquisition or new-family/model transfer.

Figure construction. Figure 3A averages public reference-remainder spectra after normalizing each public spectrum separately. The lower panel reports held-out target-aligned rank-one MSE relative to PC1 for each of the 32 families, using the loss aggregation above. The displayed variance spectrum and prediction losses therefore share the fixed acquisition protocol but measure different quantities.

H Intervention-aligned observations and matched controls

Transfer failure motivating the comparison. A separate random-calibration study on the same 32 families retained the derivative-subtracted remainder predictors, four Gaussian public panels and 64 observations per fit. At the norm-matched donor query $h + 0.10\|h\|d/\|d\|$, where d is the corrupt-minus-clean activation, rank-one/reference MSE is 2.1406; full also deteriorates and reference is the best of the fifteen fixed predictors in this query group. The failure therefore does not arise solely from rank truncation. These results concern different query directions from the preceding rank study.

Public observations, estimand and exact geometry. The directional confirmation retains 32 families, 96 sites and five contexts, with four nonzero public amplitudes $t = 0.25, 0.5, 0.75, 1$ on $h + td$ and a new 20-query grid $t = (j + 0.37)/20$, $j = 0, \ldots, 19$. All queries differ from the public nodes. This is a new-grid/control comparison on known paths, not a new-family test. Unlike remainder regression above, here the rows $X_{j,:} = (f_T(v_j) - f_T(0))^\top$ and $Y_{j,:} = (f_b(v_j) - f_b(0))^\top$

contain full margin increments. With the same ridge scale $\lambda = \max\{\|X\|_F^2/K, 10^{-24}\}$, fit

$$W = (X^\top X + \lambda I)^{-1} X^\top (Y - X), \qquad \widehat{f}_b(v) = f_b(0) + x(v) + W^\top x(v), \quad x(v) = f_T(v) - f_T(0). \tag{17}$$

The origin anchors each fit; the reference increment at a query is available and its recipient endpoint is withheld until predictions are frozen.

Let $e_0 = \mathbf{1}/\sqrt{768}$ and $e_h = (h - \bar{h}\mathbf{1})/\|h - \bar{h}\mathbf{1}\|$. Write $d = ae_0 + be_h + w$ and replace it by $d^* = ae_0 + be_h + \|w\|q$, where q is a random unit vector orthogonal to e_0, e_h. Then displacement norm, clean inner product, and the mean, variance and norm of $h + td$ agree for every t. These scalar moments do not equate coordinatewise normalized activations or later nonlinear computations. Controls cross norm-only versus moment-matched directions with four independent directions versus one repeated ray across amplitudes. PCG64DXSM keys $[2026091525, \text{ordinal}, \text{layer}, 1, \text{replicate}]$ supply paired Gaussian vectors; matching projects these vectors onto the orthogonal complement and ray controls reuse the first vector. Each random condition has four separate four-observation fits. Their losses are averaged, not their predictions; no method accesses sixteen observations. The competitor bank, public-fitting barrier and complete-context aggregation are fixed before query evaluation.

Fixed comparisons and stronger direct predictors. Losses average queries and competitors within each cell, then equally weight families, layers, contexts and public fits. The seven contrasts in Table 14 use 20,000 grammar-stratified family bootstrap draws (seed 2026091526), preserving all within-family dependence. Directional specificity also requires donor MAE at most 1.05 times the matched control, no greater competitor-contrast MSE and each layer's MSE at most 1.05 times that control. The donor contrasts meet these conditions; repeating a random ray by itself does not improve reconstruction.

Table 14: Seven predeclared contrasts. D is donor loss; I, R are norm-only independent/ray losses and M, C their moment-matched counterparts. Random losses average four separate fits. Intervals use 20,000 grammar-stratified family draws and quantiles $0.05/14, 1 - 0.05/14$ (approximate nominal joint 95% coverage). Positive lower bounds pass the 5% effect criterion.

Contrast	Estimate	Adjusted interval	Positive/32	Pass
$0.95I - D$	0.078492	[0.054730, 0.110190]	32	Yes
$0.95M - D$	0.062590	[0.045182, 0.083473]	32	Yes
$0.95C - D$	0.112644	[0.072194, 0.166832]	32	Yes
$0.95I - M$	0.010975	[0.000544, 0.024160]	21	Yes
$0.95R - C$	0.032443	[0.010303, 0.059017]	26	Yes
$0.95I - R$	–0.083823	[–0.158508, –0.031122]	7	No
$0.95M - C$	–0.057616	[–0.114627, –0.014731]	10	No

Table 15: All frozen directional predictor groups. Each random row averages four separate public fits, representing all 21 original predictors. Full-vocabulary errors and false positives (FP) cover the same 1,920 half-context endpoints per fit; fractional counts are panel means.

Method	MSE	Full errors	Full FP
Reference shift	0.060393	21	18
Reference + Jacobian	0.275452	7	1
Donor affine	0.031018	26	25
Norm-only independent	0.115273	21.25	18.75
Norm-only ray	0.193333	25	20
Moment-matched independent	0.098534	21.25	18.75
Moment-matched ray	0.151223	24	20.25
Direct PCHIP (10 nodes)	0.000031056	1	0
Direct cubic (10 nodes)	0.000007517	0	0

Donor MSE is 0.031018 versus 0.098534/0.151223 for moment-matched independent/ray controls, reductions of 68.52%/79.49%. All 32 family means and all 480 family–layer–context means favor

donor over each group's four-fit average. Every individual random panel also has higher family-mean MSE in every family. Moment matching reduces independent/ray MSE by 14.52%/21.78%, while repeating a random ray increases error by 67.72%/53.47% without/with matching. Thus scalar geometry and ray layout do not account for the donor advantage.

Direct piecewise cubic Hermite interpolation (PCHIP) and cubic interpolation use ten public nodes $t = j/9$. For five recipients and twenty queries, their $5 \times 10 = 50$ suffix evaluations match the affine predictor's $(1 + 5) \times 5 + 20 = 50$, including centers and query references. Exact recipient querying costs 100. This is total suffix-query accounting, not measured end-to-end speedup. Direct cubic is substantially more accurate, and donor affine has more full-vocabulary sign errors than reference shift (26 versus 21). The bank truth agrees with full-vocabulary truth at these 1,920 endpoints (478 correct, 1,442 incorrect), without implying general bank sufficiency.

Illustrative path and cohort curves. Figure 3B shows true logit margins, the affine predictions fitted to full margin increments, and prediction minus truth. Diamonds mark the four donor-aligned calibration values. The L0 half-context illustration is selected at median difficulty among 96 paths using the two methods' mean bank MSE; the original-subject readout and matched panel p0 are fixed. The lower RMSE curves aggregate 32 families, three layers and five contexts; the matched control averages losses across four fits, rather than averaging their predictions. All B axes use donor amplitude t.

I Measurement allocation and stable reconstruction

Held-out allocation roles and balanced access. The allocation comparison uses the next eight families after an eight-family development study, within the same known 32-family pool; only their role in this question is held out. It retains L0/L4/L8 and H/P/I/O/N. At the first-name patch site, eight token substitutions produce donor activations: four supply calibration labels and four supply query inputs. The first public donor is the original subject; the others follow a fixed cyclic name-pool rule, with an eight-family shift for this replication. Names may recur across families. The original clean answer remains the assessment target. Each query donor has amplitudes $0.25, 0.5, 0.75, 1$, giving 1,920 recipient endpoints and 384 complete-vocabulary H endpoints. The fixed bank includes the original subject, all donor names and the union of six centers' top-five tokens, excluding the original answer. All public records precede a global fit/prediction barrier, followed by hidden recipient labels. Query reference responses remain available.

For each of the six unordered pairs (i, j) of public donors, repeated allocation observes both at $0.5, 1$; diverse allocation observes i, j at 1 and the other two at 0.5. Both use four labels per response branch. Across the six designs, each donor occurs three times at each scale; losses, not predictions, are averaged. The collection union contains thirteen public nodes for the fixed controls, but no four-point fit accesses that union. Families are the independent descriptive groups; repeated contexts and allocations do not increase their number.

Estimators and error decomposition. Let X, Y be reference/recipient public increments and Z query reference increments. Shift predicts $y_0 + Z$; OLS predicts $y_0 + Z + ZX^\dagger(Y - X)$ with relative cutoff 10^{-12}; ridge uses Eq. 17. Scalar predicts $y_0 + \beta Z$, where $\beta = \langle X, Y\rangle_F / \max\{\|X\|_F^2, 10^{-24}\}$. Direct linear fits recipient increments to donor activation displacements. Every four-point estimator uses only its own allocated labels.

For any declared finite-response map B, write public corrections $E = XB + R_{\mathrm{pub}}$ and query corrections $D = ZB + R_{\mathrm{query}}$. With $C = ZX^\dagger$, substitution gives

$$\begin{aligned} CE - D &= -(Z - CX)B + CR_{\mathrm{pub}} - R_{\mathrm{query}}, \\ \|CE - D\|_F &\le \|Z - CX\|_2 \|B\|_F + \|C\|_2 \|R_{\mathrm{pub}}\|_F + \|R_{\mathrm{query}}\|_F. \end{aligned} \tag{18}$$

Public interpolation alone controls neither the missing-feature term nor amplified public and query remainders. Applying the bound requires justified restrictions on B and both remainders. Here X is a finite-response feature matrix, not the gradient operator in Theorem 1.

Consistency, stronger controls and scope. In Figure 3C, each upper link joins repeated and diverse four-observation OLS within one of the eight families. Unobserved energy is reference-response

Table 16: All four-point allocation methods and the shifted-reference control. Four-point rows average six fixed losses on the same eight replication families; full-vocabulary events cover 384 H-context endpoints per allocation.

Method	MSE	MAE	Full errors	Full FP
Reference shift	0.036118	0.104992	4.000	2.000
Repeated OLS	1.202821	0.472765	11.500	10.000
Repeated ridge	0.015758	0.067772	3.500	2.000
Repeated scalar	0.012621	0.061166	3.833	3.000
Repeated direct linear	8.316861	1.953090	34.667	2.167
Diverse OLS	0.038953	0.095596	2.833	1.833
Diverse ridge	0.015364	0.066175	3.167	2.000
Diverse scalar	0.012509	0.060848	4.000	3.000
Diverse direct linear	7.501145	1.821300	34.333	0.833

Table 17: Family-mean MSE in the held-out allocation replication; layers, contexts and six configurations are equally weighted.

Family	**Repeated OLS**	**Diverse OLS**	**Diverse ridge**	**Diverse scalar**
Family 1	0.410377	0.041115	0.010318	0.008356
Family 2	1.400886	0.056003	0.014542	0.004297
Family 3	2.644099	0.009540	0.006231	0.004410
Family 4	0.218386	0.011987	0.008259	0.008817
Family 5	1.915393	0.100911	0.038250	0.043324
Family 6	1.094034	0.030637	0.012946	0.013746
Family 7	0.489178	0.031303	0.018150	0.011309
Family 8	1.450218	0.030131	0.014218	0.005810

energy outside the observed span; the lower rows show the corresponding query-coefficient ℓ_1 norms. Family losses average contexts, layers and fixed allocations, whereas geometry averages layers and allocations. The variance axis in A and the MSE and coefficient axes in C are logarithmic.

Diverse OLS lowers MSE 96.76% relative to repeated OLS, although unobserved reference-response energy rises from 9.19% to 11.71%. The mean coefficient ℓ_1 norm falls from 14.88 to 1.02 and condition number from 172.98 to 12.45. All 144 paired site/allocation comparisons improve these two stability quantities, whereas unobserved response energy rises in 142/144 pairs. Both public fits have MSE below 1.5×10^{-32}. Allocation changes feature geometry and sampled remainders jointly, so the comparison does not identify conditioning as the sole causal mediator.

All eight family means (Table 17), 24 site means and 120 family–layer–context means improve after averaging the six allocation losses. Family improvements span 89.98–99.64%; leave-one-family-out gains span 95.67–97.27%, a sensitivity range rather than a confidence interval. At the individual allocation/cell level, 681/720 improve and 39 do not. These comparisons show the value of stable allocation at fixed OLS estimation. For absolute reconstruction accuracy, diverse ridge and scalar are stronger, and shifted reference also has slightly lower MSE than diverse OLS.

Of the 384 H endpoints, 77 retain the original answer and 307 do not; bank and full-vocabulary truth signs agree on this sample. Diverse OLS averages 2.833 errors/1.833 false positives versus shift's 4/2, a false-positive difference of 0.167 events; lower reconstruction MSE does not by itself guarantee answer-status accuracy. A four-point shared fit costs 6 centers, 24 calibration branch values and 16 query references, totaling 46 suffix evaluations per family/layer across five recipients, versus 22 for shift and 86 for complete direct querying. These counts exclude native capture and wall time; the separate workload experiment supplies the measured deployment-cost claim.

J FINITE-WORKLOAD AUDITING AND INDEPENDENTLY MEASURED COST

Workload and information access. The independently timed study reuses four old development families with a new random-direction stream, layers 0/4/8, five recipient contexts, eight public-fit directions and 64 query directions at 0.10 times the patch-center norm. All structural sites are included, giving $N = 4 \cdot 5 \cdot 64 = 1280$ vector endpoints per layer and 3,840 overall. A site shares its fixed competitor bank and queried reference responses across recipients. Recipient query outputs are unavailable to prediction and routing except for sampled audit labels; full direct outputs are retained subsequently for evaluation. Model operations use float64 and batches of at most eight within a structural site.

For a predicted bank vector $\widehat{y}_i$ and its direct numerical output y_i, define combined failure by

$$B_i = \mathbf{1}\left\{ \|\widehat{y}_i - y_i\|_\infty > 0.01 \text{ or } \mathbf{1}\{\min_j \widehat{y}_{ij} \geq 0\} \neq \mathbf{1}\{\min_j y_{ij} \geq 0\} \right\}.$$

This evaluates numerical error and bank-minimum sign, rather than full-vocabulary correctness.

Executed audit and finite-population guarantee. Fix predictions and positive scales g_i before audit labels. Sample $n = 114$ endpoints uniformly without replacement and set $q = \max_{i \in S} \|\widehat{y}_i - y_i\|_\infty / g_i$. An unobserved prediction is released only if $qg_i \leq 0.01$ and its predicted bank minimum is strictly outside $[-qg_i, qg_i]$; sampled or rejected endpoints return direct outputs. With $e_i = \|\widehat{y}_i - y_i\|_\infty$,

$$\Pr\{\#\{i : e_i/g_i > q\} > 64\} \leq \frac{\binom{1215}{114}}{\binom{1280}{114}} \approx 0.0019734521. \tag{19}$$

The event requires missing a fixed set of 65 highest-scoring endpoints; ties can only reduce strict exceedances. Every delivered failure exceeds q, while audited and fallback outputs contribute zero failures relative to the direct numerical oracle. The primary uses this rule with constant scales at L0/L4 and evaluates L8 directly. A union bound over the two audited early layers is approximately $2(0.0019734521) = 0.0039469042 < 0.05$; L8 is direct. Thus each delivered layer has at most 5% combined failures with at least 95% simultaneous confidence over audit randomization, without assuming independence across endpoints or methods. The guarantee concerns this finite workload and numerical oracle, not new prompts, individual predictions or the conditional risk among accepted predictions.

Independent timing and strong controls. Each method recomputed its own clean/corrupt captures, bank, required gradients, reference queries, public fit, audits and fallback; none reused another method's intermediate results. One synchronized clock covered fresh setup through all delivered arrays, including Python routing, with resident model weights. Model loading, subsequent serialization, scientific archive I/O and post-delivery validation were excluded. Execution order was cyclically rotated over six repeats, reusing the same query and audit draws as timing repetitions.

Table 18: Independent setup-to-delivery timing and combined bank-vector numerical/sign failures. Times are six-repeat medians with resident model weights; speedup is relative to Direct. CV: sample standard deviation divided by mean time (%). G: finite-workload guarantee.

	Resident execution			Combined failures			
Method	Time (s)	Speedup (×)	CV (%)	L0	L4	L8	G
Direct	10.352	1.000	3.24	0	0	0	yes
Shared reference	6.242	1.658	7.08	0	0	191	–
L8-direct	7.163	1.445	7.30	0	0	0	–
Audited L8-direct	7.889	1.312	2.24	0	0	0	yes

Audited L8-direct saves 23.79% with zero observed failures at every layer (Table 18). It predicts 2,332 endpoints, returns 228 audit answers and evaluates 1,280 L8 endpoints directly. The unaudited L8-direct control also has zero observed failures; auditing adds 10.13% time for the finite-workload guarantee. Raw sharing incurs 191 L8 failures (14.92% of that layer); pooling its 4.97% overall rate conceals failure of the per-layer requirement. This workload is separate from the donor-path and rank-compression studies.