WHEN DO EXPLANATIONS PREDICT MODEL CHANGES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can an explanation predict a model's response to an internal change it has not observed? Matching the original outputs is insufficient: two models can agree on every input yet react differently to the same internal change. We study what finite measurements must retain to predict specified changes, and what to improve when those predictions fail. In a linear response space, a necessary and sufficient condition distinguishes inaccessible information from information discarded by compression; a stability bound separates both from unreliable recovery. Controlled experiments make this distinction actionable: ordinary least squares recovers local responses, whereas principal component analysis (PCA) can lose the required differences despite preserving almost all variance. Retaining those differences repairs prediction. Larger MLPs and pretrained GPT-2 show why local recovery is only part of the problem: new inputs and finite changes introduce different errors. In GPT-2, measuring more inputs improves prediction when local fitting is already accurate. Finite attention-head replacements expose a sharper trade-off: interaction measurements greatly improve predictions of unqueried combinations on the fitting inputs, yet can lose to simpler measurements on more inputs at the same budget. Together, the theory and experiments distinguish when to change the measurements, the representation, or the inputs used to predict model edits.

1 INTRODUCTION

Consider two models with named branches: one computes $x + x$, the other $3x - x$. Both return $2x$ on every input. Removing the first branch, however, leaves x in one model and $-x$ in the other. Perfect output mimicry cannot resolve this difference. The missing information concerns the response to a specified internal change, not ordinary prediction accuracy.

Internal changes are practical tools for understanding and modifying trained networks. Activation patching replaces component activations to locate a behavior, while ROME updates weights to change a factual association (Zhang & Nanda, 2024; Meng et al., 2022). These uses make a further question important: can measured responses predict an edit that has not yet been tried? For example, separate attention-head replacements may inform a prediction for replacing several heads together on new prompts. Which component effects must be measured and retained, and how many inputs should supply them? Agreement with the unchanged model does not answer these questions.

Figure 1 gives an analytic illustration. The fixed network $f_\eta(x) = \tanh(x + \eta_1) + \frac{1}{2}\tanh(2x + \eta_2)$ starts at $\eta = 0$. We record outputs at $x = -1, 0, 1$ after shifting each bias separately by ± 0.01. The two matrix rows estimate their sensitivities. For the unqueried joint edit $\eta = (0.2, -0.2)/\sqrt{2}$, their scaled difference predicts the output change. Keeping the sum of the rows would erase the answer to this question. Even with both rows retained, the prediction is not exact: the sensitivities describe local changes, whereas the edit is finite. This one example already separates lost information from an imperfect approximation.

Response-based evaluation already goes beyond output mimicry (Yeh et al., 2019). Activation-patching conclusions can depend on the response metric and corrupted inputs (Zhang & Nanda, 2024); AtP* identifies and repairs failures of gradient-based approximations (Kramár et al., 2024). We ask a complementary question: *given finite model access, which distinctions must a reusable state retain to predict a specified family of future operations?* We address access, compression, and

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.

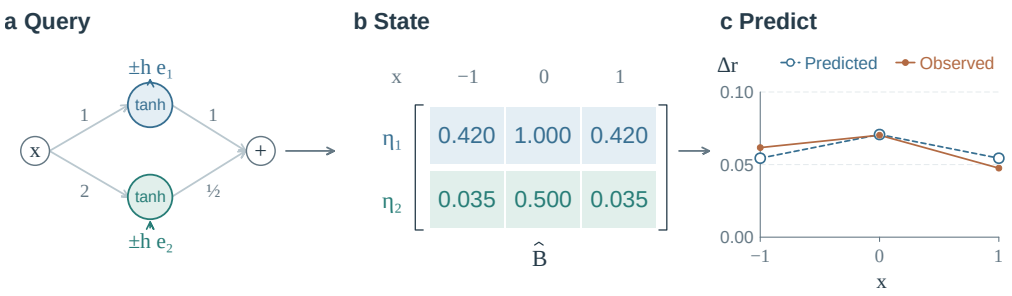

Figure 1: An analytic example: separate bias queries predict a joint edit.

prediction on new inputs together. A *state* here is reusable response information, such as the matrix in the example, not a new hidden layer. Measurement sufficiency and error analysis distinguish failures with different remedies; matched experiments test those remedies on unqueried operations and new inputs.

Our contributions have three connected parts:

1. **Conditions for retaining the answer.** We connect accessible measurements, compression and requested responses through a necessary and sufficient condition, then distinguish information loss from unstable recovery.
2. **A test of where prediction fails.** Matched-query experiments show that ordinary least squares (OLS) is sufficient for local recovery, while principal component analysis (PCA) can discard the response difference needed for prediction. Preserving that difference repairs the failure.
3. **Consequences beyond local fitting.** An exact error decomposition connects larger-model predictions to input sampling and finite-change approximation. GPT-2 experiments test two resulting choices: more input coverage after accurate local fitting, and simpler versus interaction-aware measurements under a fixed budget.

Small networks let us check exact derivatives, complete query geometry, and independent input sets. Larger MLPs and a pretrained Transformer test whether the same reasoning remains useful at scale. All targets are trained before state construction. Our operations modify these frozen models.

2 From a Model Change to a Response Matrix

Let $f_0 : \mathcal{X} \to \mathcal{Y}$ be a trained, frozen model. We choose an internal location and operation coordinates $\eta \in \mathbb{R}^d$; f_η denotes the function after that operation. For example, η can shift a hidden-layer bias or a specified attention output. A vector $r_S(f_\eta) \in \mathbb{R}^p$ records the response statistics of interest on a finite collection of inputs S, called an *input set*. The finite change is $\Delta r_S(\delta, a) = r_S(f_{\delta a}) - r_S(f_0)$. The direction a, magnitude δ, input set, and response definition together specify a prediction problem. In Figure 1, $S = \{-1, 0, 1\}$ and $r_S(f_\eta)$ simply lists the three outputs. In the language experiment, it instead lists four average logit margins.

Writing response vectors as rows, define $B_S \in \mathbb{R}^{d \times p}$ by

$$a^\top B_S = \frac{d}{dt} r_S(f_{ta})\Big|_{t=0}. \tag{1}$$

Thus B_S is the transpose of the response Jacobian with respect to η; $a^\top B_S$ is a directional derivative. Its rows retain named operation coordinates and its columns retain named response statistics. For $d = p = 8$, the state stores 64 coefficients, and one direction decodes eight values. Responses to two bias coordinates yield rows $b_1^\top, b_2^\top$; the unqueried direction $(e_1 - e_2)/\sqrt{2}$ has tangent $(b_1 - b_2)^\top/\sqrt{2}$. Multiplying by δ predicts a small finite change; the accuracy of that additional approximation is tested separately.

To construct the matrix without differentiating the model, query positive and negative edits. Fix the directions, scales, input splits and steps before evaluation. With directions in the rows of X and positive steps h_j,

$$Y_{j,:} = \frac{r_S(f_{h_j a_j}) - r_S(f_{-h_j a_j})}{2h_j}, \qquad \widehat{B}_S = X^+ Y, \tag{2}$$

where X^+ is the Moore–Penrose pseudoinverse. The prediction for an unqueried direction is $a^\top \widehat{B}_S$. Full-column-rank X identifies all declared directions in the exact linear model; otherwise it identifies their projection onto the row span of X. Each row of Y costs two response evaluations; one response may aggregate many network inputs. Small steps amplify value noise and large steps bias the derivative estimate (Appendix A.6). Direct derivatives, when available, provide a separately costed reference for checking this estimate.

The response definition expresses what the explanation is asked to predict. Our controlled networks measure how edits change unit-removal effects; larger MLPs use output moments, and GPT-2 uses template-specific logit margins. Appendix E gives their exact formulas and the weighted version of Equation 2. The rest of the procedure does not depend on selecting one of these particular responses.

Models are compared on declared operations and responses. Matching local matrices establishes local response agreement; finite changes are tested separately. Appendix A formalizes this equivalence.

3 When Is the Retained Information Sufficient?

3.1 Access and compression together

Vectorize the complete response object as $b \in \mathbb{R}^n$. In the linear measurement model, allowed measurements give $y = Mb$, a linear compression stores $s = Ty$, and the requested responses are Qb. Vector norms are Euclidean and matrix operator norms are induced by them. The maps and coordinate ordering are fixed independently of the object being evaluated. This is a finite-dimensional linear model, not an assertion that every model edit has a globally linear effect.

Theorem 1 (Query sufficiency after measurement and compression)**.** *For all $b \in \mathbb{R}^n$, an exact decoder from TMb to Qb exists if and only if*

$$\ker(TM) \subseteq \ker(Q). \tag{3}$$

Whenever such a decoder exists, a linear one exists. If the queries are identifiable from M, the smallest exact linear state dimension is $\operatorname{rank}(Q)$.

The proof uses standard linear factorization: two objects with the same state must have the same answer. The criterion distinguishes what to change when prediction fails. If $\ker M \not\subseteq \ker Q$, the measurements cannot distinguish all requested answers. If measurements suffice but $\ker(TM) \not\subseteq \ker Q$, compression has discarded a needed distinction. The former requires new information; the latter requires a different representation of information already measured. Full reconstruction is unnecessary when omitted directions do not affect the requested answers. Complete proofs appear in Appendix A.

Exact identifiability is not numerical reliability. Set $A = TM$, observe $s = Ab + e$, and use linear decoder D. If $\|b\| \le R$ and $\|e\| \le \epsilon$, then

$$\|Qb - Ds\| \le R\|Q - DA\|_{\mathrm{op}} + \epsilon\|D\|_{\mathrm{op}}. \tag{4}$$

The first term is the chosen decoder's approximation error, not necessarily information loss: with $A = Q = I$, a zero decoder fails although all answers are available. The part that no decoder can recover is $Q(I - A^+A)b$. Indeed, the smallest worst-case linear-decoder residual is $\inf_D \|Q - DA\|_{\mathrm{op}} = \|Q(I - A^+A)\|_{\mathrm{op}}$ (Appendix A.3). The second term measures noise amplification. A small positive singular value can preserve the answer yet make recovery unstable. We report residuals and decoder norms separately, with an explicit numerical rank tolerance.

3.2 Which compression is sufficient?

Return to the two-coordinate example. Keeping only $b_1 + b_2$ preserves the response to $e_1 + e_2$, but erases the response to $e_1 - e_2$. Keeping more variance cannot certify that the missing difference is small. Theorem 1 gives two useful consequences.

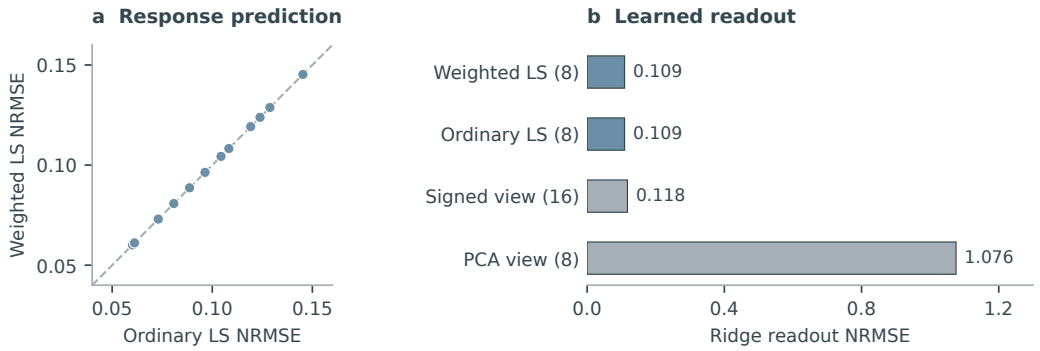

Figure 2: Identical paired queries: response prediction and compressed-view readout.

Corollary 1 (Storage for a specified query family)**.** *For a matrix response object* $\mathbf{F}$ *with Frobenius inner product, store* $s_{\mathcal{Q}}(\mathbf{F}) = (\langle \mathbf{F}, Q_j \rangle)_j$ *for a fixed family of query matrices. Every query* $Q = \sum_j \alpha_j Q_j$ *is decoded exactly as* $\langle \mathbf{F}, Q \rangle = \alpha^\top s_{\mathcal{Q}}(\mathbf{F})$*; the minimum exact linear storage equals the query-span rank. In contrast, if all bilinear queries are allowed on the full matrix space, every fixed rank-deficient row projection loses at least one answer.*

For the second claim, choose a discarded row direction $h \neq 0$ and a nonzero column direction c. Adding $hc^\top$ leaves the compressed state unchanged but changes the answer to query (h, c) by $\|h\|^2\|c\|^2$. The full proofs, including the block-structured response space, remain in Appendix A. A finite state is therefore sufficient for a declared span; larger families require checking whether their new questions introduce additional directions.

3.3 Test the retained information, not just the estimator

The two-row example suggests a concrete test: give estimators identical measurements, then vary what the representation keeps. We use frozen tanh and residual-SiLU networks with eight hidden units. Bias and output-weight edits change unit-removal means and variances; their fixed projections form the eight response coordinates (Appendix E). The edit is the change to predict, while unit removal defines what we measure about the edited model. All input roles, transformations and readout choices are fixed before testing. Reported NRMSE divides each response error by its coordinate's training standard deviation before taking the root mean square.

A preregistered comparison supplies the same paired queries, inputs, edits, splits and readout budget to every route. Across 24 fresh training seeds and 96 model-operation cases, a weighted fit and ordinary least squares reach direct NRMSE 0.1025590 and 0.1025592. The weighted estimator adds no recovery advantage here. This is important: it isolates what happens to the same information when the representation changes.

The eight-value decoded response gives ridge-readout NRMSE 0.109163, compared with 0.117591 for the sixteen-value paired-response view $v_\pm(a)$ (Appendix E.2). PCA compresses this view to eight values before differencing; these are per-direction readout dimensions. This rank-matched representation has NRMSE 1.075509 despite preserving 99.9647% of training variance (Figure 2). All methods see the same measurements. Thus the failure cannot be attributed to withholding response queries from PCA.

Why does nearly complete variance retention fail? Before differencing, the step-scaled paired responses retain a common baseline whose variance can grow as h^{-2}, while the informative difference has bounded variance. PCA can retain the former and discard the latter. Retaining the difference directly preserves the query in Equation 2.

A 24-seed development diagnostic and three newly registered 24-seed cohorts test this explanation. With transformations and readouts frozen before test generation, standardized PCA has NRMSE 1.0170 ± 0.0225, whereas retaining the eight paired differences gives 0.1231 ± 0.0184 (cohort mean and sample SD). SVD and normal-equation readouts agree within 10^{-12}. Changing the solver does

not repair the loss; changing the retained information does. Appendix B gives the derivation and full matched controls. Figure 2 shows the original matched-query study, not these later repair cohorts.

Coordinate tests establish a separate requirement: measurements and queries must refer to the same named coordinates. Relabeling both consistently leaves the aligned predictions unchanged; unlabeled coordinate destruction increases direct error by 15.35 times in the R36 matched-query study. Appendix E.5 retains the earlier comparisons with parameter, activation and output summaries. Those summaries lack the additional response queries, so their larger gaps measure the value of access, not an advantage over an estimator given the same information.

4 Predicting Finite Changes on New Inputs

Retaining the right local information solves only part of the problem. In Figure 1, both response rows are available, yet a finite edit is not predicted exactly. A second source of disagreement appears when the inputs used to estimate the matrix differ from those on which the edit will be evaluated. These two changes require different checks.

4.1 Separate the errors before choosing a remedy

If the response is twice continuously differentiable along the edit segment, with $\|d^2 r_S(f_{ta})/dt^2\| \le L_{S,a}$ throughout it, Taylor's theorem gives

$$\|\Delta r_S(\delta, a) - \delta a^\top \widehat{B}_S\| \le |\delta| \|a^\top (B_S - \widehat{B}_S)\| + \tfrac{1}{2} L_{S,a} \delta^2. \tag{5}$$

Even an exact matrix leaves a finite-change approximation error. Scanning edit magnitudes tests its useful range; a sampled curvature maximum is not a certificate of a uniform bound.

Now estimate on input set S and predict on a new set E. The residual is exactly

$$\begin{aligned} \delta a^\top \widehat{B}_S - \Delta r_E(\delta, a) = {} & \underbrace{\delta a^\top (\widehat{B}_S - B_S)}_{\text{estimation}} + \underbrace{\delta a^\top (B_S - B_E)}_{\text{input sampling}} \\ & + \underbrace{\delta a^\top B_E - \Delta r_E(\delta, a)}_{\text{linearization}}. \end{aligned} \tag{6}$$

The first term calls for better matrix estimation. The second can remain after perfect recovery on S, even when both sets come from the same distribution. The third remains after replacing the estimate by the exact matrix on E. Their vectors add, but their norms need not: errors can cancel. Direct derivatives on both sets provide separately recorded validation access for measuring these components.

This separation also clarifies uncertainty. For a fixed input distribution and set size, average aligned matrices over estimation sets to obtain a fixed prediction center; use separate calibration sets to assess variation on a new set. In the controlled 22-target study, local-operator prediction sets cover 5087 of 5632 evaluation input sets (90.32%; simultaneous pooled lower bound 88.51%). This is marginal coverage of B_E under exchangeable input-set sampling; finite-response prediction additionally involves linearization. Appendix E.6 gives the center, score, coordinate alignment and calibration access; Table 10 reports the two construction routes. The result quantifies finite-input variability rather than assuming it disappears once the matrix has been estimated accurately.

4.2 Accurate local recovery does not remove input dependence

We next keep the same prediction question while increasing model size. California housing MLPs range from 8,961 to 1,318,401 parameters, including a 512-wide, six-layer residual model. Nine prospectively evaluated checkpoints have test R^2 of 0.777–0.794. Across 108 early/middle-layer cases, local tangent NRMSE is below 1.5×10^{-5}, yet input-sampling RMS exceeds estimation RMS throughout the finite-edit grid. Recovering the local matrix more precisely cannot remove a discrepancy between the matrices for two different input sets.

GPT-2 (124M) supplies an interpretable response: the next-token logit margin between the recipient and subject in a name-completion prompt. For example, the model should complete a gift-giving

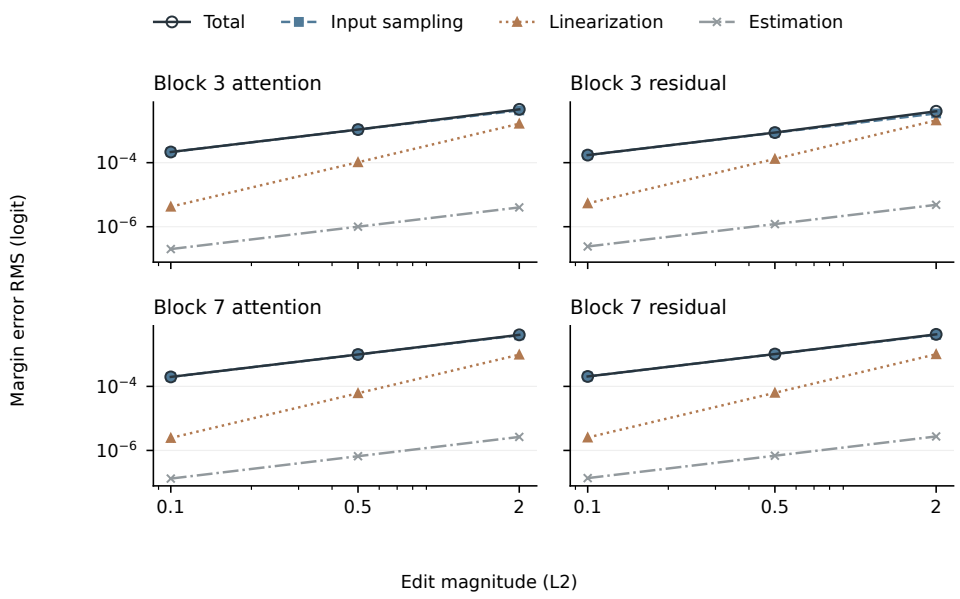

Figure 3: GPT-2 development: RMS of Equation 6's estimation (gray), input-sampling (blue), and linearization (brown) terms; the black circles show total prediction error.

sentence with the recipient rather than repeat the giver. Four template averages form the response vector. The model is qualified on separate prompts before edits are evaluated.

Figure 3 shows the development study's four attention/residual sites in 32-coordinate operation spaces. Input-sampling error remains much larger than estimation error as edit magnitude grows, while the linearization component increases. These are component RMS values, not additive shares of total error. The visual diagnosis motivates the prospective checks rather than counting as an additional independent test.

Three fresh name-pair cohorts, four sites and nested 8-/32-coordinate spaces then give 24 prospectively fixed cases. Local tangent NRMSE is below 0.00030, whereas magnitude-2 edits on new input sets have NRMSE 0.085–0.355. Input-sampling RMS exceeds estimation RMS in every case. Here finite-edit NRMSE divides error RMS by actual effect RMS, floored at 0.001; it is not the training-scaled NRMSE used in Section 3.3. Appendices D and C report the response definitions, weak effects, floors, splits and costs. Same-query OLS remains numerically equivalent in both model families.

These observations suggest spending additional measurements on inputs rather than improving an already accurate local fit. A registered 192-case GPT-2 study compares four complete direction matrices on 64 prompts with one on 256 distinct prompts at equal input/edit evaluation cost. Each matrix already spans the local space. At magnitude 2, more inputs reduce raw MSE from 4.0484×10^{-5} to 2.3260×10^{-5}, or 42.5%. Of 24 fresh name-pair groups, 21 improve and three worsen; pooled-input OLS achieves the same gain (Appendix F).

This control isolates input coverage after local identification is accurate. We next test the harder choice: measuring genuinely new interactions.

5 Response Detail Versus Input Coverage

A more detailed response approximation can genuinely improve prediction and still be the less effective use of a fixed measurement budget. To test this possibility, we replace named attention heads in GPT-2 layer 8 with their activations from paired clean prompts. The target is the change in four mean logit margins after replacing three or four heads together, not their local derivative. The clean donor, recipient and subject identities are fixed before measuring the response.

Let $F_S(v)$ be the finite response change for a binary head mask $v \in \{0,1\}^{12}$. Single-head measurements store $\alpha_j = F_S(e_j)$; pairwise measurements additionally store $\beta_{jk} = F_S(e_j + e_k) - \alpha_j - \alpha_k$. The two predictions are

$$\widehat{F}_{\text{single}}(v) = \sum_j v_j \alpha_j, \qquad \widehat{F}_{\text{pair}}(v) = \sum_j v_j \alpha_j + \sum_{j<k} v_j v_k \beta_{jk}. \tag{7}$$

These are finite-effect measurements, not gradients. They enlarge the declared response coefficient space from single-head to pairwise effects. Theorem 1 applies to measurement and compression within that space; unmeasured higher-order interactions remain approximation error. Appendix G specifies the activation operation and the connection to the linear coefficient model.

Table 1: Finite-head response prediction: MSE (10^{-3}).

	Central difference	Single-head	Pairwise
Fitting inputs; unqueried head combinations			
12 input pairs	10.704	0.873	0.01088

Forward budget	Central difference	Single-head	Pairwise
New inputs; fixed measurement budget			
5,120	21.465	7.699	10.020
15,360	27.306	4.941	6.445

The pairwise measurements are valuable. On the same 12 fitting input pairs, they reduce MSE for *unqueried* three- and four-head combinations from 0.873×10^{-3} to 0.01088×10^{-3}, about eighty-fold (Table 1, upper panel). This is not their guaranteed interpolation of measured single-head and two-head responses.

Their cost changes the comparison on new inputs. Measuring 12 single-head effects, or those effects plus all 66 pairwise interactions, costs 224 or 1,280 individual-prompt forward evaluations per input pair, including clean donors. At a budget cap of 15,360, the single-head method can measure 68 pairs and the pairwise method 12. They use 15,232 and 15,360 evaluations, respectively; neither has enough remaining budget for another complete pair.

On 24 unqueried masks and 16 new input sets, single-head measurements have 23.3% lower mean MSE, improving 15 of 16 sets (lower panel). Actual effects are 1.05–1.25 logit RMS, so the comparison concerns appreciable finite replacements. It combines input coverage with differences in cross-input approximation: even at the same 12 fitting pairs, single-head prediction has slightly lower new-input MSE (5.999 versus 6.445, in units of 10^{-3}). The benefit combines broader input coverage with the cross-input behavior of the retained interaction coefficients.

At budget 5,120, eight sets favor each method and the paired 95% bootstrap interval includes zero; the clear advantage occurs at the larger budget. Intervals and these counts are conditional on the fixed fitting inputs, checkpoint, operation site and templates. Head-space integrated gradients (Sundararajan et al., 2017) improve over local derivatives, but their backward access is costed separately. All six methods, fitting-input curves, intervals and selection costs are retained in Appendix G. The result identifies a practical measurement choice: evaluate gains in within-input approximation together with measurement cost and prediction on the intended new inputs.

6 Related Work

Local surrogates and additive attributions explain predictions through local models or feature credit (Ribeiro et al., 2016; Lundberg & Lee, 2017). Infidelity measures perturbation-weighted response error (Yeh et al., 2019); randomization tests and human simulation examine other aspects of explanation quality (Adebayo et al., 2018; Hase & Bansal, 2020). Rule extraction and distillation establish output approximation (Craven & Shavlik, 1995; Bucilă et al., 2006), while CKA and generalized shape metrics compare representations (Kornblith et al., 2019; Williams et al., 2021). Our question is narrower than explanation quality in general: what information must measurements and their representation retain for specified internal-edit predictions, and what should change when those predictions fail?

Internal interventions already have a substantial methodological literature. Causal abstraction studies faithful simplification under mechanism transformations (Geiger et al., 2025); ROME edits learned associations (Meng et al., 2022). Activation-patching results depend on metric and corruption choices (Zhang & Nanda, 2024). EAP uses linear approximation (Syed et al., 2024), AtP* diagnoses and repairs attribution-patching failures (Kramár et al., 2024), and EAP-IG improves circuit faithfulness (Hanna et al., 2024). First-order relation decoding (Hernandez et al., 2024) and averaged Jacobian readouts (Gurnee et al., 2026; Yan et al., 2026) are also close precedents. Building on this work, our analysis connects measurement sufficiency, compression loss and cross-input error to tested changes in representation and measurement use. The head-space controls are not full implementations of these editing or circuit-discovery pipelines.

Goal-oriented inverse methods already recover requested quantities rather than complete objects (Spantini et al., 2017); our linear criterion separates access and compression without a Bayesian linear-Gaussian assumption. Predictive state representations describe action-conditioned futures (Littman et al., 2001; Singh et al., 2004); here the operations act on a frozen model. Split conformal inference (Lei et al., 2018; Angelopoulos & Bates, 2023) supplies our input-set coverage tool. Certified Interventional Fidelity estimates a bounded evaluation score for a supplied abstraction or component set (Asiaee, 2026). We distinguish identifying the requested responses from quantifying uncertainty over new finite input sets; the latter is not a confidence interval for a population mean score.

7 Discussion and Conclusion

The framework connects prediction failures to distinct remedies: new measurements for inaccessible answers, representation repair for discarded differences, and improved query design or regularization for unstable recovery. The kernel condition and stability bound separate these cases; the matched-query and PCA experiments show why the distinction matters in practice.

Once local recovery is accurate, input sampling and finite-change approximation become central. Larger MLPs and GPT-2 demonstrate the former; finite-head replacements show why refining the latter has a cost. The pairwise predictor's strong same-input gain and the single-head predictor's larger-budget new-input advantage demonstrate why response detail, input coverage and measurement cost must be evaluated together.

Application to another model starts with its edit coordinates and response statistics. The linear criterion depends on this declared response space, not a particular network architecture. Computational cost still matters: a local d-by-p matrix stores dp coefficients, and each paired direction requires two response evaluations. The cost of a response depends on the network, input count and operations needed to compute it. Broader operation families can increase both measurement and storage requirements.

The prediction-set guarantee is marginal over exchangeable input sets in the controlled eight-unit protocol, not a larger-model guarantee. The larger-model evidence covers additive internal edits and clean-donor head replacements in fixed operation families, including one GPT-2 checkpoint and controlled templates. Free-form language, weight editing such as ROME, nonsmooth systems and human usefulness require their own evaluation.

The common lesson is constructive: preserve the distinctions needed by the question, then test what limits their use on new inputs. An explanation's predictive value depends not on the sophistication of its estimator alone, but on the information measured, retained and used under the intended prediction conditions.

AI USE STATEMENT

AI tools supported translation, writing and revision, literature retrieval, and reference checking. The authors reviewed the resulting content and retain full responsibility for the research, its claims, and the final manuscript.

ETHICS STATEMENT

The experiments use synthetic inputs, public California housing data, controlled language templates and frozen learned systems, including a publicly available pretrained language model. They involve no human participants, private data or deployment decisions. The work does not claim that operations on a learned model identify causal relations in the external world. Its main foreseeable risk is over-interpreting a bounded operational state as a universal or causal mechanism; the protocol, proxy, uncertainty checks, and explicit claim boundaries are designed to prevent that use.

REPRODUCIBILITY STATEMENT

The reproducibility materials contain standalone PyTorch methods, per-study replay entry points, immutable reports, seed and split registries, and source and data hashes. The selected methods include 137 tests of response estimation, coordinate alignment, readout, input-set uncertainty and numerical checks, plus 24 tests in the isolated head-replacement component. The larger-model, prospective compression, input-allocation and head-replacement records support reconstruction of training or a pinned checkpoint, paired queries, direct derivatives and the applicable finite edits. Historical report verification and model reconstruction are recorded separately; failed reproduction attempts remain traceable rather than replacing a frozen result. Reproduction does not reopen a one-use formal registry or create additional scientific samples. Appendices give the proofs, operation definitions, fitting roles, error targets and runtime scope. Coverage certification applies to the original controlled-network protocol.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://www.nowpublishers.com/article/Details/MAL-101.

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.

Cristian Bucilă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006. doi: 10.1145/1150402.1150464. URL https://dl.acm.org/doi/10.1145/1150402.1150464.

Mark Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, volume 8, 1995. URL https://proceedings.neurips.cc/paper/1995/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. arXiv preprint arXiv:2607.15495, 2026. URL https://arxiv.org/abs/2607.15495.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. arXiv preprint arXiv:2403.17806, 2024. URL https://arxiv.org/abs/2403.17806.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, 2020. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491/.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2b6d7285306fe896fca9b2e9a02ff6ec-Abstract-Conference.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1307116.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.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, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004. URL https://web.eecs.umich.edu/~baveja/uaipapers.html.

Alessio Spantini, Tiangang Cui, Karen Willcox, Luis Tenorio, and Youssef Marzouk. Goal-oriented optimal approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 39(5):S167–S196, 2017. doi: 10.1137/16M1082123. URL https://epubs.siam.org/doi/10.1137/16M1082123.

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/.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Shi-Qi Yan, Kai-Xuan Ding, Chao-Hong Tan, Qian Chen, Wen Wang, Xiangang Li, and Zhen-Hua Ling. Short horizons and sparse concepts: a mathematical view of the readout in the J-lens. arXiv preprint arXiv:2608.25347, 2026. URL https://arxiv.org/abs/2608.25347.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

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

NOTATION GUIDE

The main text uses the following symbols. A local response matrix and a finite response change are different objects: the former predicts the latter only through a separately checked approximation.

Symbol	Meaning
f_0, f_η	Frozen model and model after the specified internal operation.
S, E	Input sets used for measurement and new-input evaluation.
$r_S(f_\eta) \in \mathbb{R}^p$	Response statistics on input set S.
$a \in \mathbb{R}^d, \delta$	Operation direction and its magnitude.
$\Delta r_S(\delta, a)$	Finite response change from f_0 to $f_{\delta a}$.
$B_S, \widehat{B}_S$	Exact and estimated d-by-p local response matrices.
X, Y	Rows of measured directions and their central-difference response estimates.
b, M, T, Q	Complete response coefficients, measurement map, compression map and requested-response map.
s, A, D	Stored state $s = TMb$, combined map $A = TM$, and linear decoder.
$\mathbf{F}, Q_j$	Response array and question matrices in Frobenius inner products.
$F_S(v), \alpha_j, \beta_{jk}$	Finite head-replacement response, single-head effects and pairwise interactions.

The map Q in Theorem 1 returns the requested responses; the matrices Q_j in Corollary 1 express individual questions as inner products. The theory appendix gives their domains and assumptions.

A Complete Theory Supplement

This appendix gives the full proofs and finite-sample statements behind the main results. The objects are the model responses, allowed measurements, stored coordinates, and input sampling rules introduced in the main text. The first results concern information retained by measurements; the later results concern numerical error and uncertainty from finite input samples.

A.1 Standing assumptions and notation

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a learned system and let

$$\Pi = (\mathcal{X}_\Pi, \mathcal{A}_\Pi, \mathcal{O}_\Pi)$$

be a declared access protocol on a common class of systems. Each operation has a declared realization on every compared system. The observation map takes values in a normed vector space $\mathcal{V}$; its response profile is

$$R_\Pi(f_\theta)(x, a) = \mathcal{O}_\Pi(a \cdot f_\theta, x) - \mathcal{O}_\Pi(f_\theta, x).$$

When the observation is random, R_Π denotes its probability law. We declare the joint sampling rule for paired differences, using the same realization on both sides for the identity operation. The identity operation has zero difference response, so baseline outputs, when needed, are recorded as a separate observation channel. We assume that the response distance is finite on the systems being compared; without this condition, the same construction yields an extended pseudometric.

To cover several response matrices at once, collect them in a finite-dimensional Hilbert space

$$\mathcal{H} = \bigoplus_{b=1}^{B} \mathbb{R}^{L \times D_b}$$

with the direct-sum Frobenius inner product: sum the entrywise inner products over blocks. Each block has L operation coordinates and D_b response coordinates. The exact response array is $\mathbf{F}$, and its estimate is $\widehat{\mathbf{F}}$. For one block these are simply B_S and $\widehat{B}_S$. A fixed collection $\mathcal{Q} = (Q_1, \ldots, Q_m) \subset \mathcal{H}$ specifies linear questions $\langle \mathbf{F}, Q_j \rangle_{\mathcal{H}}$. For example, $Q_j = ac^\top$ asks for $a^\top B_S c$. These question matrices are distinct from the operation directions used to acquire measurements.

A.2 Operational quotient

Proposition S1 (Response distance is a pseudometric)**.** *Let $d_{\mathcal{O}}$ be a metric on $\mathcal{V}$ in the deterministic case, or on the declared class of response probability laws in the random case, and define*

$$d_\Pi(f, g) = \sup_{(x,a) \in \mathcal{X}_\Pi \times \mathcal{A}_\Pi} d_{\mathcal{O}}(R_\Pi(f)(x, a), R_\Pi(g)(x, a)) .$$

Then d_Π is a pseudometric.

Proof. Non-negativity and symmetry hold pointwise and are preserved by the supremum. For any f, g, h, the observation-space triangle inequality gives

$$d_{\mathcal{O}}(R_\Pi(f), R_\Pi(h)) \le d_{\mathcal{O}}(R_\Pi(f), R_\Pi(g)) + d_{\mathcal{O}}(R_\Pi(g), R_\Pi(h))$$

at every authorized (x, a). Taking the supremum and using $\sup(u + v) \le \sup u + \sup v$ proves the triangle inequality. Distinct systems may have zero distance when the protocol cannot observe their internal difference, so identity of indiscernibles need not hold. □

Corollary S1 (Exact operational equivalence)**.** *The relation*

$$f \sim_\Pi g \quad \Longleftrightarrow \quad d_\Pi(f, g) = 0$$

is an equivalence relation, and the quotient class $[f]_\Pi$ *is well-defined.*

Proof. Reflexivity and symmetry follow from the pseudometric axioms. If $d_\Pi(f, g) = d_\Pi(g, h) = 0$, the triangle inequality gives $d_\Pi(f, h) = 0$. A positive-tolerance relation need not be transitive and is not asserted to define a quotient. □

Lemma S1 (Access refinement)**.** *Let* $\Pi_1 \preceq \Pi_2$ *mean that every comparison available under* Π_1 *is also available under* Π_2*, with the same metric on the shared observation domain. Then*

$$d_{\Pi_1}(f, g) \le d_{\Pi_2}(f, g).$$

Consequently, $f \sim_{\Pi_2} g$ *implies* $f \sim_{\Pi_1} g$*.*

Proof. The supremum for Π_1 is taken over a subset of the comparisons in Π_2. If the richer distance is zero, every shared comparison is zero. □

Proposition S2 (Binary non-identifiability under equal accessible laws)**.** *Let* $M \in \{0, 1\}$ *be a latent mechanism label with* $\Pr(M = 1) = \pi$*, and let* O_Π *collect every observation available under* Π*. If*

$$\mathrm{Law}(O_\Pi \mid M = 0) = \mathrm{Law}(O_\Pi \mid M = 1),$$

then every estimator $\widehat{M}$ *measurable with respect to* O_Π *satisfies*

$$\Pr(\widehat{M} = M) \le \max\{\pi, 1 - \pi\}.$$

Under equal priors, the upper bound is $1/2$*.*

Proof. Equality of the conditional laws implies independence of O_Π and M. Thus $\Pr(M = 1 \mid O_\Pi) = \pi$ almost surely. For every observation, the Bayes rule selects the more probable prior class and succeeds with probability $\max\{\pi, 1 - \pi\}$. No measurable estimator can outperform it. This is a finite binary statement, not a universal minimax theorem. □

A.3 Access, compression, and stable query recovery

Let $b \in \mathbb{R}^n$, $M \in \mathbb{R}^{m \times n}$, $T \in \mathbb{R}^{k \times m}$, and $Q \in \mathbb{R}^{p \times n}$. Set $A = TM$. All maps are fixed; the result concerns the full declared linear object space. A smaller admissible object set may admit recovery under weaker conditions and must be analyzed separately.

Theorem S1 (Exact query recovery)**.** *There exists a decoder* g*, not assumed linear, satisfying* $g(Ab) = Qb$ *for every* $b \in \mathbb{R}^n$ *if and only if* $\ker A \subseteq \ker Q$*. When this holds, there is a linear decoder* D *with* $DA = Q$*.*

Proof. Necessity follows by taking $h \in \ker A$: the states of h and zero coincide, so $Qh = g(0) = Q0 = 0$. Conversely, define a map on $\mathrm{im}\, A$ by $D(Ab) = Qb$. If $Ab_1 = Ab_2$, the kernel condition implies $Qb_1 = Qb_2$, so this definition is well-defined. It is linear by construction and extends linearly to the entire state space. Equivalently, $D = QA^+$ works because $Q = QA^+A$, where A^+ is the Moore–Penrose inverse. □

Corollary S2 (Minimum exact linear storage)**.** *If* $\ker M \subseteq \ker Q$*, then among linear states formed from* Mb*, the minimum dimension that answers all components of* Qb *exactly is* $\mathrm{rank}\, Q$*.*

Proof. For any sufficient linear state, the previous theorem yields $Q = DTM$, so $\operatorname{rank} Q \leq \operatorname{rank}(TM) \leq k$. For attainability, choose a row basis W of Q. Identifiability implies $\operatorname{row} Q \subseteq \operatorname{row} M$, hence $W = CM$ for some C. The state CMb has dimension $\operatorname{rank} Q$ and linearly reconstructs Qb. This dimension claim is about linear storage, not arbitrary discontinuous encodings into real numbers. □

Proposition S3 (Approximation and noise amplification)**.** *For $s = Ab + e$, $\|b\|_2 \leq R$, $\|e\|_2 \leq \epsilon$, and any linear decoder D,*

$$\|Qb - Ds\|_2 \leq R\|Q - DA\|_{\mathrm{op}} + \epsilon\|D\|_{\mathrm{op}}.$$

Proof. Write $Qb - Ds = (Q - DA)b - De$, then apply the triangle inequality and the definition of the induced operator norm. □

Information loss versus an imperfect decoder. The residual above depends on the chosen decoder. It can be nonzero even when the measurements retain every answer: take $A = Q = I$ and $D = 0$. For the full linear object space, the exact distinction uses the orthogonal projector $P = A^{+}A$ onto the row space of A:

$$Q - DA = Q(I - P) + (QP - DA).$$

The first part acts on directions invisible to the measurements. The second acts on the measured subspace and can arise from decoder choice, including regularization. These matrix terms are not additive in operator norm.

Proposition S4 (Smallest decoder residual)**.** *For fixed finite real matrices A, Q with the same number of columns,*

$$\inf_{D} \|Q - DA\|_{\mathrm{op}} = \|Q(I - A^{+}A)\|_{\mathrm{op}}.$$

The infimum is attained by $D = QA^{+}$.

Proof. Since $A(I - P) = 0$, every decoder satisfies $(Q - DA)(I - P) = Q(I - P)$. Right multiplication by an orthogonal projector cannot increase the operator norm, which gives the lower bound. For $D = QA^{+}$, the residual is exactly $Q(I - P)$, proving attainment. Equivalently, over $\|b\| \leq R$, the smallest worst-case noiseless linear-decoder error is $R\|Q(I - P)\|_{\mathrm{op}}$. □

This is a standard projection identity, not a new estimation algorithm. It distinguishes a loss intrinsic to the retained measurements from bias introduced by a particular decoder. The exact inverse attaining the noiseless bound need not be preferable with measurement noise: its norm can be large. Choosing regularization therefore requires considering both terms of Equation 4, not only the noiseless residual.

These are standard finite-dimensional factorization and perturbation arguments. Their role here is to separate measurement access, compression loss, and numerical reliability in the same operational prediction problem. If $A = U\Sigma V^{\top}$, the implementation forms a truncated inverse using a declared threshold. It reports $\|Q - DA\|_{\mathrm{op}}$ and $\|D\|_{\mathrm{op}}$ separately. For a witness direction h, both $\|Ah\|$ and $\|Qh\|$ are reported. A direction suppressed by a numerical threshold is not silently called an exact null-space direction. The implementation uses a thin SVD and never requires a dense $n \times n$ null-space projector.

A.4 Compression for a specified set of responses

Theorem S2 (Operation-axis compression can lose an answer)**.** *Let $P \in \mathbb{R}^{r \times L}$ have rank $r < L$. Suppose the legal query family contains all bilinear functionals*

$$R_{a,c}(B) = a^{\top}Bc, \qquad a \in \mathbb{R}^{L}, \quad c \in \mathbb{R}^{D}.$$

Then PB is not sufficient for all legal queries.

Proof. Choose nonzero $h \in \ker P$ and nonzero $c \in \mathbb{R}^{D}$. For an arbitrary response matrix B_1, define $B_2 = B_1 + hc^{\top}$. Then

$$PB_2 = PB_1 + (Ph)c^{\top} = PB_1.$$

For the legal query $(a, c) = (h, c)$,

$$R_{h,c}(B_2) - R_{h,c}(B_1) = h^\top h\, c^\top c = \|h\|_2^2 \|c\|_2^2 > 0.$$

Thus two matrices with the same compressed representation have different allowed responses, so no decoder receiving only PB can recover every query. For a multi-block array, place $hc^\top$ in one selected block and set the other blocks to zero. □

Theorem S3 (Exact dual coordinates and minimum rank)**.** *For a fixed query family $\mathcal{Q} = (Q_1, \ldots, Q_m) \subset \mathcal{H}$, define*

$$s_{\mathcal{Q}}(\mathbf{F}) = \left(\langle \mathbf{F}, Q_1\rangle_{\mathcal{H}}, \ldots, \langle \mathbf{F}, Q_m\rangle_{\mathcal{H}}\right).$$

For every $Q = \sum_{j=1}^m \alpha_j Q_j$,

$$\langle \mathbf{F}, Q\rangle_{\mathcal{H}} = \alpha^\top s_{\mathcal{Q}}(\mathbf{F}).$$

If $q = \dim \operatorname{span}(\mathcal{Q})$, every exact linear state map with linear query decoders has rank at least q.

Proof. Exactness follows from linearity. For minimality, let $T : \mathcal{H} \to \mathbb{R}^r$ be an exact linear state map. Exact linear decodability means that for every Q in the declared span there is a vector b_Q such that, for every $\mathbf{F}$,

$$\langle \mathbf{F}, Q\rangle_{\mathcal{H}} = \langle T\mathbf{F}, b_Q\rangle = \langle \mathbf{F}, T^* b_Q\rangle_{\mathcal{H}}.$$

Hence $Q = T^* b_Q$, so

$$\operatorname{span}(\mathcal{Q}) \subseteq \operatorname{range}(T^*), \qquad q \le \operatorname{rank}(T^*) = \operatorname{rank}(T) \le r.$$

The result is conditional on the declared query span; it does not make the frame complete, causal, identifiable from finite observations, or sufficient for an enlarged operation family. □

A.5 Estimation, coordinate, and decoding errors

Proposition S5 (Vector error with a Lipschitz decoder)**.** *Let $\mathcal{T}, \widehat{\mathcal{T}} : \mathcal{H} \to \mathbb{R}^m$ be linear frame maps, with $s = \mathcal{T}\mathbf{F}$ and $\widehat{s} = \widehat{\mathcal{T}}\widehat{\mathbf{F}} + e$. Let $D : \mathbb{R}^m \to \mathbb{R}^p$ be L_D-Lipschitz on a domain containing $s, \widehat{s}$, and let $\widehat{D}$ be any realized decoder. Use Euclidean coordinate norms, the Hilbert array norm and the induced operator norms. With $\epsilon_{\mathrm{dec}} = \|\widehat{D}(\widehat{s}) - D(\widehat{s})\|_2$,*

$$\begin{aligned}\|\widehat{D}(\widehat{s}) - D(s)\|_2 \le \epsilon_{\mathrm{dec}} + L_D\big(&\|e\|_2 + \|\mathcal{T}\|_{\mathrm{op}}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathcal{H}} \\ &+ \|\widehat{\mathcal{T}} - \mathcal{T}\|_{\mathrm{op}}\|\widehat{\mathbf{F}}\|_{\mathcal{H}}\big).\end{aligned}$$

Proof. The exact state difference is $\widehat{s} - s = e + \mathcal{T}(\widehat{\mathbf{F}} - \mathbf{F}) + (\widehat{\mathcal{T}} - \mathcal{T})\widehat{\mathbf{F}}$. Bound its norm term by term, then use $\|\widehat{D}(\widehat{s}) - D(s)\|_2 \le \epsilon_{\mathrm{dec}} + L_D\|\widehat{s} - s\|_2$. For fixed linear prediction $D(s) = a^\top s$, $L_D = \|a\|_2$. An arbitrary learned head requires its own Lipschitz bound; downstream fit alone does not provide one. The following scalar version uses the requested query directly and can be sharper. □

Let

$$Q_a = \sum_j a_j Q_j, \qquad \widehat{Q}_a = \sum_j a_j \widehat{Q}_j,$$

and suppose the realized state coordinates are

$$\widehat{s}_j = \langle \widehat{\mathbf{F}}, \widehat{Q}_j\rangle_{\mathcal{H}} + \delta_j.$$

For a readout $\widehat{g}$, define

$$\epsilon_{\mathrm{dec}} = \left|\widehat{g}(\widehat{s}, a) - a^\top \widehat{s}\right|.$$

Theorem S4 (Error bound for a requested response)**.** *For any norm on the state coordinates and its dual norm,*

$$\begin{aligned}|\widehat{g}(\widehat{s}, a) - \langle \mathbf{F}, Q_a\rangle_{\mathcal{H}}| \le \epsilon_{\mathrm{dec}} &+ \|a\|_*\|\delta\| \\ &+ \|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathcal{H}}\|Q_a\|_{\mathcal{H}} \\ &+ \|\widehat{\mathbf{F}}\|_{\mathcal{H}}\|\widehat{Q}_a - Q_a\|_{\mathcal{H}}.\end{aligned}$$

Proof. Add and subtract $a^\top \widehat{s}$, $\langle \widehat{\mathbf{F}}, \widehat{Q}_a \rangle_{\mathcal{H}}$, and $\langle \widehat{\mathbf{F}}, Q_a \rangle_{\mathcal{H}}$. The absolute error is bounded by

$$\begin{aligned} &\left|\widehat{g}(\widehat{s}, a) - a^\top \widehat{s}\right| + \left|a^\top \delta\right| \\ &\quad + \left|\langle \widehat{\mathbf{F}} - \mathbf{F}, Q_a \rangle_{\mathcal{H}}\right| + \left|\langle \widehat{\mathbf{F}}, \widehat{Q}_a - Q_a \rangle_{\mathcal{H}}\right|. \end{aligned}$$

Apply Hölder's inequality to the state-realization term and Cauchy–Schwarz to the two array terms. The four terms measure the chosen decoder's departure from linear prediction, error in stored coordinates, response-array estimation error, and error in the question matrices. Each requires its own estimate or bound. □

A.6 Value fidelity and response differentiation

Even uniform value fidelity does not bound a response derivative without additional regularity. On $[-1, 1]$, take true response path $h(t) = 0$ and proxy $\widetilde{h}_n(t) = \sin(n^2 t)/n$. Then $\sup_t |\widetilde{h}_n(t) - h(t)| \leq 1/n \to 0$, while $|\widetilde{h}_n'(0) - h'(0)| = n$. Here t parameterizes a model edit, not an ordinary input. This counterexample complements the two-branch example: output agreement and operation-response fidelity are different requirements even before compression.

Proposition S6 (Central-difference error under smoothness assumptions)**.** *For twice continuously differentiable $h : [-\rho, \rho] \to \mathbb{R}^p$, suppose $\|h''(t)\|_2 \leq L$ and observe $\widetilde{h}_\pm = h(\pm\rho) + e_\pm$, where $\|e_\pm\|_2 \leq \varepsilon$ and $\rho > 0$. The central secant $\widehat{v}_\rho = (\widetilde{h}_+ - \widetilde{h}_-)/(2\rho)$ satisfies*

$$\|\widehat{v}_\rho - h'(0)\|_2 \leq \varepsilon/\rho + L\rho/2.$$

If instead h is three times continuously differentiable and $\|h'''(t)\|_2 \leq M$, the approximation term improves to $M\rho^2/6$.

Proof. The noiseless secant is $(2\rho)^{-1} \int_{-\rho}^{\rho} h'(t)\, dt$. The second-derivative bound gives $\|h'(t) - h'(0)\|_2 \leq L|t|$; integrating gives $L\rho/2$. The noise term is bounded by $(\|e_+\|_2 + \|e_-\|_2)/(2\rho) \leq \varepsilon/\rho$. Under the third-derivative assumption, Taylor expansion of $h'(t)$ has linear term $th''(0)$, whose symmetric integral vanishes, and remainder bounded by $Mt^2/2$. Its average is $M\rho^2/6$. □

These standard numerical-differentiation bounds explain the step-size tradeoff: smaller steps reduce approximation bias but amplify measurement error. The experiments therefore compare against directly computed derivatives and check query consistency separately. A finite amplitude grid does not certify uniform derivative bounds L or M.

A.7 Finite changes and local response error

Fix the input set S and edit direction a, and define the vector-valued path $h(t) = r_S(f_{ta})$. Suppose h is twice continuously differentiable on the segment between zero and δ, with $\|h''(t)\|_2 \leq L_{S,a}$ throughout. By definition, $h'(0) = a^\top B_S$.

Proposition S7 (Finite-edit error decomposition)**.** *For every estimated local operator $\widehat{B}$,*

$$\|h(\delta) - h(0) - \delta a^\top \widehat{B}\|_2 \leq |\delta| \|a^\top (B_S - \widehat{B})\|_2 + \tfrac{1}{2} L_{S,a} \delta^2.$$

Proof. For either sign of δ, the integral remainder is

$$h(\delta) - h(0) - \delta h'(0) = \delta^2 \int_0^1 (1 - v) h''(v\delta)\, dv.$$

Its norm is at most $L_{S,a}\delta^2/2$. Add and subtract $\delta h'(0) = \delta a^\top B_S$ and apply the triangle inequality. □

The tangent error may include input-sampling and proxy error. This result does not estimate the curvature constant from data, nor does it assert a uniform guarantee beyond the declared segment. Finite grid checks remain empirical diagnostics unless supplemented by a valid uniform bound. For a nonsmooth operation path the differentiability assumptions must be replaced by an appropriate local regularity analysis.

A.8 ESTIMATION ACROSS INPUT SETS

Let μ be a declared probability law on inputs, distinct from the access protocol Π. Let B_S be a measurable random response operator for $S \sim \mu^{\otimes n}$. Define

$$B_{\mu,n} = \mathbb{E}[B_S]$$

with covariance operator Σ_B of finite trace. For independent input sets $S_1, \ldots, S_{N_{\text{est}}}$, let

$$\widehat{B}_{\mu,n} = \frac{1}{N_{\text{est}}} \sum_{j=1}^{N_{\text{est}}} B_{S_j}.$$

Proposition S8 (Error of the mean response matrix)**.**

$$\mathbb{E}\Big[\|\widehat{B}_{\mu,n} - B_{\mu,n}\|_F^2\Big] = \frac{\operatorname{tr}(\Sigma_B)}{N_{\text{est}}}.$$

For a fixed linear query $L_a(B) = a^\top B$,

$$\mathbb{E}\Big[\|L_a(\widehat{B}_{\mu,n} - B_{\mu,n})\|_2^2\Big] = \frac{1}{N_{\text{est}}} \mathbb{E}\big[\|L_a(B_S - B_{\mu,n})\|_2^2\big].$$

Proof. Vectorize the response operator. The estimation error is the mean of N_{est} independent centered random vectors, so its covariance is $\Sigma_B / N_{\text{est}}$. Taking the trace gives the first identity. Applying the fixed linear map L_a before the same covariance calculation gives the second. The $1/N_{\text{est}}$ identity is an expectation statement; it does not imply that every finite collection of estimation sets improves prediction on every future set. □

For prediction on a new input set, averaging reduces only one part of the sampling error. Let S_* be independent of the N_{est} estimation sets and follow the same sampling rule. Write $\sigma_a^2 = \mathbb{E}\|a^\top(B_S - B_{\mu,n})\|_2^2$. Then

$$\mathbb{E}\|a^\top(\widehat{B}_{\mu,n} - B_{S_*})\|_2^2 = \left(1 + \frac{1}{N_{\text{est}}}\right)\sigma_a^2.$$

To see this, subtract the common mean from the estimated center and the future matrix. Independence makes their cross term zero; their variances are $\sigma_a^2/N_{\text{est}}$ and σ_a^2. Thus more estimation inputs can improve the center without removing variation in the future inputs. This identity concerns exact local derivatives under the stated independent sampling model. Biased fits, finite edits and sampling without replacement require their corresponding error or dependence terms.

Proposition S9 (Coordinate relabeling invariance)**.** *Let $B, C \in \mathbb{R}^{d_{\text{op}} \times d_{\text{resp}}}$ share named operation and response axes, and let $X_{\text{eval}} \in \mathbb{R}^{J \times d_{\text{op}}}$ contain the evaluation directions. For permutation matrices $P_{\text{op}}, P_{\text{resp}}$, define $B' = P_{\text{op}} B P_{\text{resp}}^\top$, $C' = P_{\text{op}} C P_{\text{resp}}^\top$, and $X'_{\text{eval}} = X_{\text{eval}} P_{\text{op}}^\top$. For the ratio below, assume $\|X_{\text{eval}} B\|_F > 0$. Then*

$$X'_{\text{eval}} B' P_{\text{resp}} = X_{\text{eval}} B, \qquad \frac{\|X'_{\text{eval}}(C' - B')\|_F}{\|X'_{\text{eval}} B'\|_F} = \frac{\|X_{\text{eval}}(C - B)\|_F}{\|X_{\text{eval}} B\|_F}.$$

Flooring both denominators at the same $\epsilon > 0$ also gives invariance at zero response. The coordinate array of the mean matrix transforms with the relabeling; after restoring the coordinate order, its decoded predictions and score are unchanged. Values and their coordinate IDs must move together.

Proof. Orthogonality gives $P_{\text{op}}^\top P_{\text{op}} = I$ and preserves the Frobenius norm. Hence $X'_{\text{eval}} B' = X_{\text{eval}} B P_{\text{resp}}^\top$, and similarly $X'_{\text{eval}}(C' - B') = X_{\text{eval}}(C - B) P_{\text{resp}}^\top$. Right multiplication by $P_{\text{resp}}^\top$ preserves both norms. A value permutation that leaves IDs fixed is not a relabeling and is not covered by this invariance result. □

A.9 PREDICTION SETS FOR NEW INPUT SAMPLES

For a frozen center C, evaluation directions X_{eval}, and denominator floor $\epsilon > 0$, the implemented score is Equation 12. Its prediction set is $\mathcal{C}_r = \{B : e_{X_{\mathrm{eval}}}(C, B) \le r\}$. Because normalization uses the candidate B, r is dimensionless, not an absolute output radius. The set constrains the response $X_{\mathrm{eval}}B$; it need not constrain every operator entry if X_{eval} is rank deficient. E023 uses directly computed matrices as calibration and evaluation labels for both construction methods. Its paired-query estimator therefore uses derivative access during calibration; that additional access is included in the accounting.

Let $E_1, \dots, E_m, E_*$ be exchangeable nonconformity scores, where the first m scores are computed on calibration input sets and E_* is the score for one future set. The score rule, center, directions and floor are fixed independently of these sets. Exchangeability of the sets then gives exchangeability of their scores; refitting on the evaluation set would not satisfy this argument. Let

$$k = \lceil (m+1)(1-\alpha) \rceil ,$$

and, when $k \le m$, let $r = E_{(k)}$ be the k-th calibration order statistic. For $k = m+1$, set $r = +\infty$. Assume $0 < \alpha < 1$.

Proposition S10 (Split-conformal finite marginal coverage)**.**

$$\Pr(E_* \le r) \ge \frac{k}{m+1} \ge 1 - \alpha.$$

Proof. Under exchangeability and randomized tie breaking, the rank of E_* among the $m+1$ scores is uniform on $\{1, \dots, m+1\}$. Hence the probability that its rank is at most k equals $k/(m+1)$. Without randomized tie breaking, the weak inequality and calibration order statistic are conservative. In the frozen evaluation, $m = 96$, $1 - \alpha = 0.90$, and $k = 88$, giving $88/97 \approx 0.907216$. This is marginal coverage over exchangeable input sets, not conditional coverage for every model, operation, or input configuration. □

For validation of frozen states, the next result is applied conditionally on the trained targets, query designs, and estimation/calibration data. The remaining randomness comes from independent, previously unopened evaluation sets; the expectations in the bound use this same conditioning. This differs from the marginal conformal guarantee above, which also averages over calibration sets. The finite-radius implementation abstains if $k > m$ rather than replacing $+\infty$ by a finite empirical quantile; E023's $k = 88 \le 96$ does not encounter that boundary.

Proposition S11 (Simultaneous finite-registry lower bounds)**.** *For comparison $j \in \{1, \dots, J\}$, let $X_{j1}, \dots, X_{jN_j} \in [0,1]$ be independent coverage indicators, not necessarily identically distributed. Write*

$$\widehat{p}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} X_{ji}, \qquad \bar{p}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbb{E}[X_{ji}].$$

Then, with probability at least $1 - \delta$, simultaneously for every j,

$$\bar{p}_j \ge \widehat{p}_j - \sqrt{\frac{\log(J/\delta)}{2N_j}}.$$

Proof. Hoeffding's inequality gives

$$\Pr\left(\bar{p}_j < \widehat{p}_j - \sqrt{\frac{\log(J/\delta)}{2N_j}} \right) \le \frac{\delta}{J}.$$

A union bound over the J comparisons proves the simultaneous statement. The frozen evaluation applies it prospectively to derivative/paired-query pooled coverage and to architecture-operation strata. With $(J, N, \delta) = (2, 5632, 0.05)$, the pooled penalty is 0.0180968; with $(J, N, \delta) = (8, 1408, 0.05)$, the stratified penalty is 0.0424531, matching the frozen report. □

A.10 Evaluation noise on independent input sets

Let C be a system-operation context and $r(C) \in \mathbb{R}^q$ its mean response under the specified input sampling rule. Let

$$Z = r(C) + \epsilon_O, \qquad Y = r(C) + \epsilon_Y,$$

where Z is a state-side estimate and Y is an independently estimated evaluation outcome.

Proposition S12 (Predictive-risk decomposition)**.** *Suppose $r(C)$, ϵ_Y, and $h(Z)$ are square-integrable, where h is a fixed measurable readout. If $\mathbb{E}[\epsilon_Y \mid C, Z] = 0$, then*

$$\mathbb{E}\|h(Z) - Y\|_2^2 = \mathbb{E}\|h(Z) - r(C)\|_2^2 + \mathbb{E}\|\epsilon_Y\|_2^2.$$

Proof. Expanding the square gives

$$\begin{aligned}\|h(Z) - Y\|_2^2 = \|h(Z) - r(C)\|_2^2 + \|\epsilon_Y\|_2^2 \\ - 2\langle h(Z) - r(C), \epsilon_Y \rangle.\end{aligned}$$

Conditioning on (C, Z) makes the cross term zero. The evaluation-sampling term is a common evaluation noise floor; it cannot be subtracted from only one representation, and the decomposition does not imply that Z is unbiased, sufficient, causal, or optimal. □

A.11 Relation to the frozen evaluation protocol

The experiment instantiates the theory as follows. The response operator $\widehat{B}_{\mu,n,u}$, estimated from 16 independent input sets for case u, is the stored mean response matrix. A held-out direction a is decoded as $a^\top \widehat{B}_{\mu,n,u}$, so decoder departure is zero for the primary direct readout. The derivative and paired-query methods estimate the same named response coordinates while retaining separate fidelity checks. Ninety-six disjoint calibration input sets define the predictive radius, and 128 untouched evaluation sets per case test point prediction and marginal coverage. Shuffled and coordinate-destroyed states preserve tensor size while breaking state-query alignment. Parameter, activation, gradient, output, strong concatenated, and compact projected descriptions face matched ridge and nonlinear readout evaluations.

Near-equality of derivative-based and paired-query prediction errors shows that restricted queries suffice in these cases. Estimation accuracy, variation across input sets, and prediction-set coverage remain separate questions with separate measurements.

A.12 Claim boundary

Together, the results specify what the permitted observations can distinguish, what linear compression must retain, and how estimation and input sampling affect response predictions. Each result states the relevant domain, regularity or exchangeability assumptions. The experiments test specified model operations and response coordinates. They do not identify unique physical mechanisms or external-world causes, and marginal coverage does not imply a guarantee at every context.

B Why Compression Loses the Response Difference

The earlier access-matched result leaves an important question: does PCA lose the response information, or does its readout merely fail to recover it? We distinguish these possibilities in a prospectively specified development diagnostic on 24 new training-seed groups. The 8/4/12 train/validation/test split yields 48 trained targets and 96 architecture–operation combinations. Every combination uses 16 estimation input sets of size 128 and 16 held-out directions as the earlier protocol. One independent evaluation input set supplies the directly computed derivative targets. This is a new attribution study, not a rerun or replacement of the frozen formal coverage experiment.

Separate the common response from the difference. Let $r_+, r_- \in \mathbb{R}^p$ be the coordinate-aligned positive and negative response coordinates, each already divided by its query's $2h$. Define

$$c = \frac{r_+ + r_-}{\sqrt{2}}, \qquad o = \frac{r_+ - r_-}{\sqrt{2}}.$$

Here c is the common (even) component, and o the difference (odd) component under exchanging positive and negative queries. This is an orthogonal coordinate change: centered paired-response energy equals the sum of centered common and odd energies. The decoded central secant is $r_+ - r_- = \sqrt{2}\, o$, so an odd-only view preserves that query exactly without using future reference responses or task labels.

The scale imbalance has a simple local explanation. For a fixed scalar response path $y(h)$, Taylor expansion gives, with bounded local third derivative,

$$\frac{y(h) + y(-h)}{2\sqrt{2}\, h} = \frac{y(0)}{\sqrt{2}\, h} + O(h), \qquad \frac{y(h) - y(-h)}{2\sqrt{2}\, h} = \frac{y'(0)}{\sqrt{2}} + O(h^2).$$

Thus small signed steps can enlarge the common response relative to the derivative signal. Across paths with square-integrable baseline values and uniform local derivative bounds, and nonzero baseline variance, the leading common variance scales as h^{-2}, while the odd variance remains bounded. This explains a possible variance imbalance; it does not assert that every dataset or principal projection must discard odd information. The following measurements test this mechanism directly.

Matched views and numerical controls. We compare the full signed view (16 coordinates) with five eight-coordinate views: raw centered PCA, feature-standardized PCA, common/odd rotation followed by separate standardization and PCA, fixed common-only, and fixed odd-only. Ordinary LS and the weighted state are additional eight-coordinate controls. Only training data fit means, scales, and projections. The paired ridge readouts share the objective $\|Y - XW - b\|_F^2 + \lambda\|W\|_F^2$, with an unpenalized intercept and $\lambda \in \{0, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ selected on validation data. One solver uses the original normal equations; the other directly decomposes the design matrix by SVD. Both use the same feature and target standardization. We report all routes rather than selecting a winner on test data. In Table 2, secant error measures affine reconstruction of the available central secant; tangent error measures prediction on an independent evaluation input set. Both are pooled NRMSE on twelve test-seed groups using direct-SVD ridge. Size counts decoded coordinates, not full-operator storage.

Table 2: Compression attribution: secant reconstruction and tangent prediction.

View	Size	Secant error	Tangent error
Raw signed	16	$< 10^{-10}$	0.1121
Raw PCA	8	0.9998	1.0239
Standardized PCA	8	0.9998	1.0239
Common/odd PCA	8	0.5946	0.6186
Common only	8	1.0001	1.0242
Odd / central secant	8	$< 10^{-10}$	0.1109
Ordinary LS	8	$< 10^{-10}$	0.1109
Weighted state	8	0.00004	0.1109

What the measurements establish. The common part contains 99.9684% of raw centered training energy. Raw PCA retains 99.9714% of raw variance, yet its validation-selected affine secant reconstruction has test NRMSE 0.9998. Standardized PCA also fails; its explained variance is computed in a different, standardized metric and must not be read as retained raw energy. Balancing common and odd coordinates before PCA partially improves both reconstruction and prediction, but still trails the exact odd-only view. The fixed common-only control behaves like raw PCA, consistent with removal of the relevant odd information.

Normal-equation and direct-SVD test NRMSE differ by less than 10^{-10} for every route in this study. Hence changing the numerical solver does not repair the compression failure. The uncompressed signed view reconstructs the available secant to numerical precision, distinguishing lost information from extra statistical work imposed on a learned readout. Ordinary LS, the exact odd view, and the weighted state agree in predictive quality. A query-aware change of coordinates suffices here; estimator complexity is not the source of the result.

For each fitted affine compression we also evaluate $\|Q - QA^+A\|_{\mathrm{op}}/\|Q\|_{\mathrm{op}}$, where $Q = [I_p, -I_p]$ and A is the compression's linear part in the original signed coordinates. This quantity is zero up to

floating-point tolerance for the full and odd views, and approximately one for raw and standardized PCA. It diagnoses failure of exact recovery on the entire declared signed vector space. It is not a lower bound for predictors restricted to a particular empirical data manifold.

All targets meet the unchanged quality rule; the largest validation relative MSE is 0.0563. No model, direction, or input set is removed. Seed-group errors, paired response tensors, transforms, coefficients, and named coordinates are retained. The study diagnoses this linear response-view compression; it does not claim that PCA fails generally, that nonlinear decoders cannot exploit additional correlations, or that weighting outperforms same-access ordinary LS.

B.1 Prospective three-cohort replication

We then registered three new cohorts with the same 8/4/12 allocation to train, validation and test, and identical views, queries, ridge grid and qualification rules. The 72 training-seed groups produce 144 trained targets and 288 architecture–operation cases. All targets qualify; the largest validation relative MSE is 0.1094 under the fixed 0.2 rule. No seed group or outcome was replaced. For each cohort, training/validation data determine every projection, scale, readout coefficient and selected regularization value. A hash-linked fit snapshot is saved before generating test systems. The verification regenerates all original models, paired queries and outcome tensors, replays the train/validation fitting, and reconstructs each test prediction from the original frozen parameters. Table 3 reports NRMSE means, sample standard deviations and ranges across the three independently fitted cohorts, using direct-SVD ridge and frozen test-time maps. These summaries describe cohort variability rather than confidence intervals.

Table 3: Compression replication across three independently fitted cohorts.

View	Size	Secant error	Tangent error	Tangent range
Raw signed	16	$< 10^{-10}$	0.1325 ± 0.0166	$[0.1179, 0.1505]$
Raw PCA	8	1.0036 ± 0.0177	1.0170 ± 0.0225	$[0.9971, 1.0413]$
Standardized PCA	8	1.0036 ± 0.0177	1.0170 ± 0.0225	$[0.9971, 1.0413]$
Common/odd PCA	8	0.6682 ± 0.0396	0.6861 ± 0.0310	$[0.6505, 0.7069]$
Common only	8	1.0032 ± 0.0183	1.0166 ± 0.0228	$[0.9960, 1.0411]$
Odd / central secant	8	$< 10^{-10}$	0.1231 ± 0.0184	$[0.1091, 0.1440]$
Ordinary LS	8	$< 10^{-10}$	0.1231 ± 0.0184	$[0.1091, 0.1440]$
Weighted state	8	0.000045 ± 0.000010	0.1231 ± 0.0184	$[0.1091, 0.1440]$

Table 3 confirms the earlier attribution without reusing its training seeds. Common response accounts for 99.9600–99.9674% of centered training energy. Raw PCA retains 99.9632–99.9695% of raw variance, yet its fitted reconstruction of the available secant has NRMSE 0.9929–1.0240. Standardizing before PCA leaves essentially the same error. Common/odd balancing helps but does not guarantee preservation of the query: its normalized full-space query residual ranges from 0.4832 to 0.9869. The corresponding residual is below 10^{-12} for full and odd-only views and approximately one for raw and standardized PCA. This worst-case linear-space audit and empirical prediction error answer different questions; their rankings need not coincide on a restricted data distribution.

Across all eight routes and three cohorts, normal-equation and direct-SVD tangent NRMSE differ by at most 8.74×10^{-14}. The direct fixed decoder gives NRMSE 0.11408 ± 0.00931 for both ordinary LS and the weighted state; these are distinct from the learned-readout entries above. The weighted state has a small nonzero secant reconstruction error, retained rather than called exact. The full signed control preserves the secant but has slightly higher new-input readout error than the explicit difference view. Thus algebraic preservation and finite-data readout estimation remain separate.

Each cohort shares one fitted readout across its twelve test-seed groups. We therefore summarize three cohort-level errors rather than treat 36 test-seed groups or thousands of response rows as independent fitted replications. These are controlled-model confirmations, not additional architecture-scale evidence or a population coverage certificate. Generation, fitting and evaluation take 42.3–43.8 seconds per cohort on one CPU thread, with peak process memory below 484 MiB; separate model replay is not included in those timings.

C Internal Transformer Edits and Prospective Replication

Task and qualification. In a name-completion prompt, two people visit a location and one gives an object to the other. The next name should identify the recipient rather than repeat the subject. Indirect object identification is an established task for studying GPT-2 internals (Wang et al., 2023); we author an IOI-style fixture, not a reproduction of the original data generator or circuit experiment. The target is frozen pretrained GPT-2 small with 124,439,808 parameters.[1] It is never trained or coupled to our estimator.

Sixteen single-token names and four templates are fixed before task scores. Each unordered name pair contributes both mention orders and both subject choices in every template. Twelve qualification pairs, eight estimation pairs and eight future pairs produce 192, 128 and 128 prompts. Names and templates may recur; pairs and prompts never cross roles. Correlated variants are not treated as independent tokens. No incorrectly answered prompt is removed. Qualification requires pair accuracy at least 0.90, at least 0.80 per template, and positive mean logit margin. The development cohort passes at 0.9896 pair accuracy, 0.9010 full-vocabulary top-1, and mean margin 2.6576. Estimation/evaluation pair accuracies are 1.0, with full-vocabulary accuracies 0.9375/0.9297; these scores do not select operations or samples.

Operations and response coordinates. Four sites are fixed: block 3 attention output and completed residual at the second subject token; block 7 attention output and completed residual at the final prediction token (zero-based indices). Earlier-token edits propagate through downstream attention and residual computations. None edits only the last linear readout. A functional additive shift occurs once at its declared site; parameters remain bitwise unchanged.

Each 768-channel site has response-independent nested orthonormal subspaces of dimension $d = 8$ and $d = 32$, with response dimension $p = 4$ as in the main notation. Paired positive/negative queries at step 0.1 build the matrix; eight normalized combinations not in the query table test it. Four response coordinates are template-specific mean recipient-minus-subject logit margins. A 32-coordinate state stores 128 coefficients and decodes to four responses. This is a declared subspace, not whole-model mechanism coverage. Independent OLS receives the identical query table and agrees exactly with the state. Direct Jacobians are separately accessed for diagnosis, never estimation. Evaluation on new input sets uses both signs at magnitudes 0.1, 0.5 and 2.0; small or null direction effects are retained.

Three errors, not one. Equation 6 separates matrix estimation, input-set variation and finite-edit linearization. An independent reader reconstructs the exact vector identity and each metric from tensors. Component norms need not add: errors can cancel. Table 4 compares local derivative NRMSE on estimation inputs with finite-edit NRMSE on new inputs at magnitude 2, pooling both signs, eight directions and four template responses.

Table 4: GPT-2 development: local tangents and finite edits at magnitude 2.

Site	d	Local NRMSE	Finite NRMSE
Block 3 attention, subject	8	0.000065	0.1477
Block 3 attention, subject	32	0.000132	0.1689
Block 3 residual, subject	8	0.000169	0.1351
Block 3 residual, subject	32	0.000176	0.1631
Block 7 attention, final	8	0.000233	0.3616
Block 7 attention, final	32	0.000149	0.2484
Block 7 residual, final	8	0.000142	0.3241
Block 7 residual, final	32	0.000102	0.1741

Input-set variation is much larger than estimation error at every site, while linearization grows with magnitude (Figure 3). Excellent local estimation therefore does not make the fitted matrix independent of its estimation inputs. At magnitude 2, measured effects range from 0.0079 to 0.0472 logit RMS: real but modest changes relative to the unedited margins, not evidence of frequent decision reversals.

[1] `https://huggingface.co/openai-community/gpt2`, revision `607a30d783df` (full revision and weight hash in the artifact manifest).

A declared 0.001-logit denominator floor applies to smaller effects; none of the magnitude-2 table entries activates it. The 8/32 comparison changes held-out mixtures, so it is not a matched proof that increasing dimension improves prediction.

Cost and scope. Timed model loading, qualification and all eight audits take 42.42 seconds on one RTX 4090, with 1.52 GiB peak allocated memory. Downloads, preflight hashing and independent readback are excluded. Repeated queries contract final hidden states with two output rows rather than materialize full-vocabulary logits; parity is checked against the full computation. Target arithmetic is FP32 and small operator algebra FP64. Forward batches are bounded at 16 sequences, but direct-Jacobian computation retains the graph for an input set. This development cohort demonstrates internal attention/residual applicability, not complete circuit recovery, arbitrary language-task transfer, long-context scalability or new population coverage.

Prospective disjoint-cohort replication. Before opening further pretrained scores or edit responses, we commit a three-cohort protocol that excludes all 28 name pairs used anywhere in development. A fixed permutation partitions 84 of the remaining 92 pairs into three groups of 28. Each group retains the 12/8/8 qualification/estimation/evaluation roles and all template/order/subject variants. No pair or prompt recurs across these cohorts or development; names and templates still recur. Fresh nested operation frames and held-out directions are fixed for each cohort. The checkpoint, sites, qualification rule, dimensions, query step and edit magnitudes are unchanged. All three cohorts are mandatory, including failures; none may be replaced after inspecting its outcome.

All three qualify, with pair accuracy 1.0 in every cohort and full-vocabulary top-1 mean 0.9063 (sample standard deviation 0.0188, range 0.8906–0.9271). All 24 site/dimension cases pass independent tensor, role and hash readback. The maximum estimation-set derivative NRMSE is 0.000291. Same-access OLS remains equivalent, and input-set variation RMS exceeds estimation RMS in all 72 case/magnitude summaries. The raw prediction RMS is below the zero-change baseline in each summary; this comparison is not inferred from floored ratios. The denominator floor is active in 10 summaries at smaller magnitudes, but never in Table 5. Weak directions remain included.

Table 5: GPT-2 replication: new-input NRMSE at magnitude 2.

Site	d	Mean	Std	Min	Max
Block 3 attention, subject	8	0.1700	0.0095	0.1605	0.1795
Block 3 attention, subject	32	0.1785	0.0908	0.1141	0.2823
Block 3 residual, subject	8	0.1756	0.0321	0.1490	0.2112
Block 3 residual, subject	32	0.1758	0.0392	0.1434	0.2194
Block 7 attention, final	8	0.2076	0.1283	0.1201	0.3549
Block 7 attention, final	32	0.1939	0.0866	0.0963	0.2613
Block 7 residual, final	8	0.2031	0.0370	0.1737	0.2447
Block 7 residual, final	32	0.1932	0.0954	0.0850	0.2651

At magnitude 2, actual effects span 0.0099–0.0417 logit RMS across all cases. Measured cohort times are 42.52–42.66 seconds, with the same 1.52 GiB peak allocation and timing exclusions as development. These repetitions strengthen the distinction between local recovery and prediction on new inputs, not a claim of frequent decision reversals. Three fixture cohorts on one checkpoint are not three trained-model replicates or a population confidence interval. Their variability combines input-pair and operation-basis changes; it does not isolate either source. The input-set coverage guarantee of E023 is not transferred to this experiment.

D Larger MLPs: Task-Qualified Finite-Edit Validation

Data and targets. The California housing training partition contains 16,512 rows. We divide these into target fitting, checkpoint selection, and operation development (70/15/15%, split seed 93701). Inherited feature normalization is inverted and refitted on target-fitting rows only; target and response scales are also train-fitted. The original 4,128 test rows are not loaded by this development experiment. Rows are sampled randomly; the sets are not spatially separated.

Three SiLU networks use widths/depths 64/3, 256/4, and 512/6; the last uses residual hidden layers. Training uses seed 78311, AdamW at learning rate 10^{-3} and weight decay 10^{-4}, batch 256, and 150 epochs. Checkpoint selection minimizes selection MSE. The predeclared qualification rule requires selection $R^2 \geq 0.65$ and at least 0.02 above a train-fitted linear baseline. All three pass; the linear baseline has selection/development R^2 0.5826/0.6178. Table 6 reports model quality separately from response prediction. These task scores qualify the studied target, not the explanation estimator. Time includes training and the registered audits; memory is peak allocated CUDA memory, not total device usage.

Table 6: Larger-MLP development: target quality and computation cost.

Target	Parameters	Select R^2	Dev R^2	Seconds	MiB
MLP 64/3	8,961	0.7892	0.7795	8.56	84.51
MLP 256/4	199,937	0.7949	0.7879	10.42	136.76
Residual 512/6	1,318,401	0.7972	0.7828	16.52	306.57

Operation and response identities. Frozen models receive functional additive bias shifts at the first and middle hidden layers. A response-independent orthonormal bank supplies nested 8/32 coordinate subspaces; each coordinate ID includes checkpoint, layer and bank. The response vector contains output mean, second moment, and six context-weighted moments, with scales fitted from train predictions and no operation-evaluation labels. A 32×8 operator stores 256 coefficients; one direction decodes eight response components.

Each layer/subspace uses three disjoint 256-row operation-development input sets and 16 held-out unit-direction combinations. A rotated orthonormal direction matrix at positive/negative step 0.01 supplies the state and independent OLS with the same table. Actual held-out edits use both signs at magnitudes 0.025, 0.1 and 0.4. Each audit builds and tests on the same named input set; it is not the different-input-set experiment reported for GPT-2. Separate direct Jacobians isolate local estimation from finite-edit linearization. Parameters stay unchanged.

What the experiment establishes. All 36 target/layer/subspace/input-set cases complete. State and OLS decoded predictions differ by less than 10^{-15}. Maximum local tangent NRMSE under the declared floor is approximately 1.04×10^{-5}. Yet early-layer 32-coordinate edits in the 64-wide model incur about 1% error relative to effect at magnitude 0.025, 4% at 0.1, and 15–17% at 0.4. Thus accurate local response estimation does not remove the approximation term in Equation 5. Raw effects, predictions and reference residuals are retained for every direction, including weak effects.

These initial results are one-training-seed development results, not population coverage or independent repeated-model evidence. Larger targets sometimes produce smaller effects under the same normalized edit magnitude, so lower absolute error is not evidence that size alone improves interpretability. Training uses FP32; frozen MLP queries use FP64. Timings are for an RTX 4090. A coordinate-bank correction was replayed before this table: the 8-D bank is now the actual prefix of the 32-D bank. Historical non-nested results remain separately recorded rather than being relabeled retrospectively.

D.1 Prospective Replication on New Models and Inputs

Freeze before test access. Following development, a separate registration fixes three new training seeds (94111–94113), all three architectures, the existing fit/selection partition, 150-epoch optimization budget and qualification rule. All nine checkpoint and selection records are frozen before any original test array is loaded. All nine pass qualification; none is replaced. Table 7 reports direct performance on all 4,128 test rows. The train-fitted linear baseline has test $R^2 = 0.5870$, versus -0.0002 for the constant baseline. These task metrics never select an operation, seed or checkpoint.

Table 7: Prospective MLP test R^2 across three training seeds.

Target	Mean	Std	Min	Max
MLP 64/3	0.7812	0.0037	0.7769	0.7837
MLP 256/4	0.7888	0.0043	0.7854	0.7937
Residual 512/6	0.7901	0.0035	0.7880	0.7942

Paired operational evaluation. A fixed permutation (seed 94121) forms three pairs of disjoint 256-row estimation/evaluation sets, six non-overlapping blocks in total. All models and training seeds reuse these input sets for paired comparison. Basis, query-direction and held-out-direction seeds are also frozen independently of responses and held fixed across training seeds. The nested 8/32 subspaces, eight train-fitted response coordinates, sixteen directions and signed magnitude grid are the same design as above, but estimation and finite-edit evaluation now use different input sets. No response normalization is refitted on test data. Common channel addresses do not assert neuron-level semantic alignment across trained models; only scalar prediction errors are aggregated across seeds.

The nine models yield 108 layer/subspace/input-pair cases, each evaluated at three absolute magnitudes and both signs. Independent readback reconstructs the paired-query least-squares solution and all error vectors, then replays every stored query, direct Jacobian and finite edit on the frozen network. Maximum local tangent NRMSE is 1.4544×10^{-5}; maximum decoded state/OLS discrepancy is 1.18×10^{-15}. At each of the 324 grouped case/magnitude summaries, input-sampling RMS exceeds estimation RMS, and raw state error is below zero-change prediction error. Input sampling also exceeds linearization in all small/medium-magnitude summaries and 104/108 at magnitude 0.4. This replication distinguishes accurate local recovery from portability across samples, rather than proposing a stronger least-squares estimator.

Table 8: New-input finite-response NRMSE at magnitude 0.4.

Target	Layer	d	Mean	Std	Min	Max
MLP 64/3	Early	8	0.2265	0.0186	0.2139	0.2479
MLP 64/3	Early	32	0.2047	0.0060	0.2006	0.2116
MLP 64/3	Middle	8	0.2362	0.0223	0.2220	0.2620
MLP 64/3	Middle	32	0.2259	0.0197	0.2090	0.2476
MLP 256/4	Early	8	0.2690	0.0328	0.2382	0.3035
MLP 256/4	Early	32	0.2530	0.0333	0.2280	0.2908
MLP 256/4	Middle	8	0.3647	0.1258	0.2800	0.5093
MLP 256/4	Middle	32	0.3782	0.1373	0.2915	0.5365
Residual 512/6	Early	8	0.2057	0.0054	0.2017	0.2119
Residual 512/6	Early	32	0.2104	0.0066	0.2047	0.2177
Residual 512/6	Middle	8	0.1557	0.0227	0.1320	0.1772
Residual 512/6	Middle	32	0.1586	0.0293	0.1336	0.1908

Aggregation, weak effects and cost. Within each trained model, squared errors are pooled over both signs, sixteen directions, eight response coordinates and three input-set pairs before RMS normalization. Table 8 then gives mean/sample-standard-deviation/minimum/maximum across the three training seeds, not a confidence interval or 108 independent replications. The predeclared denominator is the larger of response RMS and 0.001. At magnitudes 0.025, 0.1 and 0.4, this floor is active in 79, 36 and 18 of the 108 individual case summaries, respectively. All raw effects and errors remain available; small floored ratios should not be read as uniformly accurate relative prediction of tiny effects. No case is removed for weak response. At magnitude 0.4 actual response RMS spans 0.000517–0.05967 in train-standardized moment units. The component vectors add exactly, whereas their RMS values need not.

On one RTX 4090, sequential training and qualification of all nine targets takes 88.13 seconds; operational evaluation takes 14.28 seconds. The measured peak allocation is 315.00 MiB, excluding driver reservation and other device usage. Timings exclude preflight file hashing and independent readback. The result is a multi-training-seed, fixed-dataset validation: it does not establish spatially independent population coverage, model-size monotonicity or complete network mechanism coverage.

E RESPONSE DEFINITIONS, CONSTRUCTION AND ERROR UNITS

This appendix makes the common procedure in Section 2 explicit for each experiment. A response statistic is fixed before operations are evaluated. Changing its definition changes the prediction problem, even when two statistics have the same vector dimension.

E.1 THREE RESPONSE OBJECTS

Controlled networks: projected unit-removal moments. Both eight-unit families have a scalar linear output from their final hidden responses. For each named unit, its removal changes the output by $z_u(x) = -w_u h_u(x)$. For the residual-SiLU family, h_u includes its declared residual contribution. At every modified model, recompute these unit-removal responses, then their empirical means and population variances. With empirical average $\mathbb{E}_S$, define

$$m_u = \mathbb{E}_S[z_u], \quad v_u = \mathbb{E}_S[(z_u - m_u)^2], \qquad r_S(f) = \big((Lm)_\ell/\sigma_r, (Lv)_\ell/\sigma_r^2\big)_{\ell=1}^4. \tag{8}$$

Here $\sigma_r > 0$ is a frozen response scale. Variances precede projection; they are not variances of projected responses. Thus the edit is the change to predict, while unit removal defines what we measure about the edited model. The fixed projection is

$$L = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}.$$

The coordinate order is mean/variance for row 1, mean/variance for row 2, and so on. The scale σ_r is fixed by the independent effect-calibration stage and reused for both access routes. A signed projection of variances can be negative; it remains a variance contrast, not itself a variance. There are no cross-unit covariance terms in this response definition. For direction a, direct differentiation gives

$$\dot{m}_u = \mathbb{E}_S[\dot{z}_u], \qquad \dot{v}_u = 2\mathbb{E}_S[(z_u - m_u)(\dot{z}_u - \dot{m}_u)],$$

followed by the same L and scale. Thus the experiment predicts how *unit-removal response statistics* change under a weight or bias shift; it does not replace this target with the ordinary output mean.

Larger MLPs: output and context-weighted moments. Inputs and target predictions use target-training scales. For a normalized prediction $y = f(x)$ and the first six declared input fields, define

$$\phi(x, y) = (y, y^2, y\tanh x_1, \ldots, y\tanh x_6), \qquad r_S(f)_j = \frac{\mathbb{E}_S[\phi_j(x, f(x))]}{\sigma_j}.$$

Each σ_j is the population standard deviation of ϕ_j on target-fitting rows, floored at 10^{-6} and frozen. The second coordinate is a second moment, not a variance. The context fields and their ordering are retained with the operator; no outcome label defines a response coordinate. Additive early/middle-layer bias edits are represented in nested response-independent orthonormal subspaces of dimension 8 or 32. In either case the response has eight coordinates.

GPT-2: grouped logit margins. For prompt x, let $\ell_{\text{recipient}}(x)$ and $\ell_{\text{subject}}(x)$ be the two next-token logits. Each of the four template groups S_t defines one response coordinate,

$$r_S(f)_t = \frac{1}{|S_t|} \sum_{x \in S_t} \big(\ell_{\text{recipient}}(x) - \ell_{\text{subject}}(x)\big).$$

These coordinates are in logit units. Name pairs, mention orders and subject choices are retained as grouped samples; template means are not independent token observations. Operations add an activation shift at a named block/site/token position in an 8- or 32-dimensional subspace. The four template identities remain fixed between estimation and evaluation input sets.

E.2 ESTIMATION FROM PAIRED QUERIES AND ITS COST

For query directions in the rows of X, the measurement table contains $r_S(f_{h_j a_j})$ and $r_S(f_{-h_j a_j})$. The central finite-difference matrix Y contains these paired differences divided by $2h_j$. With positive diagonal precision matrix W, the declared weighted estimate is

$$\widehat{B}_S = (W^{1/2}X)^+ W^{1/2} Y.$$

The unweighted comparison sets $W = I$. The original controlled experiments freeze direction-specific steps through independent effect calibration and check a second, smaller response scale. The larger MLP and GPT-2 experiments use common steps 0.01 and 0.1, respectively, with square rotated orthonormal direction matrices. In this full-rank setting, either route solves the same paired-difference system; weighting cannot create missing information.

A table with m query directions costs $2m$ evaluations of r_S. One such evaluation aggregates all samples in its input set: it is not one scalar network forward pass. The original controlled study (archive ID E023) also uses separately recorded effect-fit, scale-consistency and derivative-validation queries. We report construction, validation and calibration access separately. Direct derivatives are a reference for the paired-query method; both controlled-study prediction sets use these derivatives as calibration labels. The larger-model finite-edit experiments do not inherit the controlled study's calibrated coverage guarantee.

In the controlled study, the query/direct prediction-error ratio is 1.000017. Independent checks give maximum direct-derivative error 1.56×10^{-9}, matrix-fit residual 4.88×10^{-4} and center-direction error 5.44×10^{-4}. These checks evaluate the construction, separately from the learned readout and input-set coverage.

Fix response normalization and query design before evaluation, solve on the estimation input sets, average only coordinate-aligned matrices across sets, and freeze that center. At evaluation, a new direction multiplies this fixed matrix. The original study uses sixteen estimation sets per case; the prospective larger-model experiments pair each estimation set with a different future set. Future responses and their direct derivatives are used to measure error, never to refit the center, choose an edit or discard a weak direction.

From the measurement table to readout views. The matched-query study first averages coordinate-aligned response tables over its estimation input sets $S_1, \dots, S_n$, keeping X and h_j fixed. Write

$$\overline{R}^{\pm}_{j,:} = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{S_i}(f_{\pm h_j a_j})}{2h_j}.$$

For an unqueried direction a, the positive/negative view is

$$v_{\pm}(a) = \left[a^\top X^+ \overline{R}^+,\ a^\top X^+ \overline{R}^- \right] \in \mathbb{R}^{2p}.$$

This uses only the stored measurements and a, not new responses at $\pm a$. Subtracting its two halves gives $a^\top \widehat{B}_{\mathrm{OLS}}$, where $\widehat{B}_{\mathrm{OLS}} = X^+(\overline{R}^+ - \overline{R}^-)$. The weighted route instead decodes its averaged weighted matrix. With $d = p = 8$, these matrices each store 64 coefficients; their decoded views have eight values, while $v_{\pm}$ has sixteen. PCA is fitted to the latter on training groups before the shared readout comparison.

E.3 Conventional descriptions

The output summary in the controlled-network comparison uses scalar predictions $y = f(x)$. Its fixed-frequency coordinates are ordinary sine and cosine averages:

$$v_{\mathrm{out}}(S) = (\mathbb{E}_S[y], \mathrm{Var}_S(y), (\mathbb{E}_S[\cos(\omega y)])_{\omega \in \Omega}, (\mathbb{E}_S[\sin(\omega y)])_{\omega \in \Omega}), \qquad \Omega = (0.5, 1, 2).$$

Variance uses the population convention on the finite input set. Frequencies are fixed by the original protocol, not tuned against evaluation responses. The output-based description also includes input-gradient statistics and the held-out operation direction. Activation descriptions contain moments of hidden responses and unit-removal responses; gradient descriptions contain input-gradient moments. The parameter description includes the operation direction and model parameters. Their learned readouts are fit on training groups and selected on validation groups.

These summaries and readouts are the tested baselines, not a claim that full model parameters with unrestricted computation cannot answer an edit. The paired-query comparison separately tests standard estimators with the same response measurements.

E.4 METRIC DEFINITIONS AND AGGREGATION

Let $y_i, \widehat{y}_i \in \mathbb{R}^p$ be a true and predicted decoded response, with i indexing the evaluation examples for the specified study. Two normalizations answer different questions:

$$\mathrm{NRMSE}_{\mathrm{train}} = \sqrt{\frac{1}{Np}\sum_{i,j}\left(\frac{\widehat{y}_{ij} - y_{ij}}{\widehat{\sigma}_j^{\mathrm{train}}}\right)^2}, \tag{9}$$

$$\mathrm{NRMSE}_{\mathrm{effect}} = \frac{\mathrm{RMS}(\widehat{y} - y)}{\max\{\mathrm{RMS}(y), 10^{-3}\}}. \tag{10}$$

The first fixes the evaluation scale from training-target coordinate variation; subtracting the fitted mean cancels in the prediction residual. The second scales by the actual evaluated effect and records whenever its floor is active. Neither is an accuracy or a universal percentage of error.

Table 9: Error targets and aggregation units.

Reported quantity	Normalization	Evaluation unit
Controlled direct/readout	Train-coordinate SD	Test-seed response rows
Controlled input-set error	Equation 12	New input-set matrix
Prediction radius/coverage	Same relative score	New input-set matrix
Larger-model local error	Effect RMS, floor 10^{-3}	Estimation-set derivatives
Larger-model finite error	Effect RMS, floor 10^{-3}	Future finite edits
Component RMS	None	Same finite-edit residual coordinates

The original direct/readout metrics use the 4/3/4 training-seed split. Four cases per seed group combine two independently trained architectures and two operation families. Direct test prediction is scored on held-out directions and future input sets of the four test groups. In contrast, the prediction-set analysis evaluates all 44 case-specific centers, each with 96 calibration and 128 evaluation input sets: 5632 coverage events per route. The mean input-set score averages set-specific relative Frobenius errors, not the train-standardized pooled RMS of Equation 9. The prediction radius is calibrated using that same set-specific score; it is a set radius for a future operator, not a confidence interval for the mean operator or an output-unit error bar.

For the prospective MLP study, pool squared errors over signs, sixteen directions, eight coordinates and the three paired input sets within each model/layer/subspace before normalization. The reported mean, sample SD, minimum and maximum then summarize the three independently trained models. GPT-2 pools signs, eight directions and four template responses within each site/subspace/cohort, then summarizes three disjoint pair cohorts of the same checkpoint. Those cohorts vary both input pairs and operation frames; their SD is not a model-population confidence interval. The compression study similarly reports three cohort-level metrics, retaining paired training-seed analyses separately. No descriptive SD is relabeled as a confidence bound.

Raw estimation, input-sampling and linearization vectors satisfy Equation 6. Component RMS plots expose their magnitudes, not an additive variance decomposition. A zero-change baseline predicts $\Delta r = 0$ and has raw RMS error equal to actual effect RMS. Comparisons against it are checked in raw units, including directions whose normalized scores activate the denominator floor.

Additional readout precision. In the original controlled comparison, the decoded response has shared ridge NRMSE 0.11285 and matched nonlinear NRMSE 0.14802. These are readout scores, not the direct matrix-prediction error. In the separate access-matched study, the weighted-fit/OLS paired seed-level direct-error ratio has 95% bootstrap interval $[0.9999900, 1.0000040]$, consistent with numerical agreement. The finite-head study instead reports raw MSE over held-out subsets and template responses; its conditional input-set resampling and error/effect normalization are specified in Appendix G.

E.5 ORIGINAL CONTROLLED EVALUATION

Frozen learned systems and operations. The controlled evaluation uses 11 untouched training-seed groups. Each independently trains one tanh MLP and one residual-SiLU MLP before any mechanism object is constructed, yielding 22 frozen targets. The evaluation protocol creates two operation-family cases per target (44 total): output-weight shifts and hidden-bias shifts. Target qualification, all thresholds, input-set roles, operation directions, proxy rules, baselines, and claims are frozen before opening evaluation input sets.

Each case uses 16 estimation input sets of size 128 to form the center, 96 disjoint calibration sets, and 128 untouched evaluation sets. Evaluation draws use distinct seeds indexed by target, architecture, operation family and draw: 5632 distinct draws across 44 cases. The two construction routes share these draws for paired comparison and are assessed separately. Direct-derivative and restricted paired-query routes are evaluated separately. All cases, including failures, are retained; Appendix E.4 and the reproducibility materials specify integrity and split checks.

Prediction tests and controls. The seed groups are split 4/3/4 for readout training, validation and testing. Direct prediction evaluates Equation 11 on held-out operation directions. We also fit ridge readouts, either shared across both network families or fitted separately to each. They use training-seed groups, choose regularization on validation groups, and open test groups once. A capacity-matched nonlinear readout has the same 4096-parameter budget and 250-step validation procedure for the state and strongest conventional view. These learned readouts receive the decoded response view $r_{\widehat{B}}(a) = a^\top \widehat{B}_{\mu,n}$, whereas the state itself is the mean response matrix $\widehat{B}_{\mu,n}$. The primary state claim uses the fixed bilinear decoder and does not depend on a learned head.

Conventional descriptions include parameters and operation direction; activation and primitive-response moments; input-gradient moments; output moments, fixed-frequency output sine/cosine averages, and output gradients; their concatenation; and a train-fitted eight-dimensional principal projection (Appendix E.3). They never observe the held-out operation response. They receive the held-out direction but not the paired responses used to estimate the response matrix. This comparison matches readout selection and capacity, but gives the state additional response-query access. Section 3.3 also matches response-query access. Row shuffling and independent coordinate permutation/sign flips preserve tensor size while breaking state–query alignment.

Validation criteria and retained failures. Predeclared criteria test prediction, coordinate alignment, agreement of derivatives and paired queries, and prediction-set coverage. Three earlier failures remain: a matrix did not transfer unchanged to new inputs; a smaller evaluation did not establish simultaneous coverage bounds; and a fixed-radius policy failed development criteria.

Read each error in its own units. Controlled direct and learned-readout errors use normalized root-mean-square error (NRMSE): divide each residual by its coordinate's training standard deviation, then take the root mean square (RMS). The input-set score in Equation 12 instead normalizes each future operator by its own response norm. Thus direct error on test-seed groups and input-set error over all 44 cases measure different quantities. The larger-model finite-edit score divides error RMS by the larger of actual effect RMS and 0.001. Raw effects and floor counts accompany it. Appendix E.4 specifies populations, denominators and aggregation order; these scores are not interchangeable percentages.

The worst target validation relative MSE is 0.017981.

Table 10: Controlled response prediction and input-set coverage.

Quantity	Derivatives	Paired queries
Direct state NRMSE	0.095197	0.095199
Mean relative input-set error	0.097510	0.097511
Predictive radius mean	0.148075	0.148076
Coverage	0.903232	0.903232
Simultaneous pooled lower bound	0.885135	0.885135
Shuffled error factor	–	15.279
Destroyed error factor	–	16.850

Response accuracy and coordinate alignment. Direct-derivative and query-estimated errors agree to five decimal places; the maximum matrix-fit residual is 4.88×10^{-4}. Appendix E gives the independent numerical checks. Shuffling operation rows raises proxy error by 15.279 times; independently permuting and sign-flipping state coordinates raises it by 16.850 times. The state therefore depends on declared query alignment, not merely dimension or marginal scale.

What additional response queries provide. The decoded response reduces within-architecture error relative to the best conventional description by 87.35% under ridge and 84.57% under the equal-budget nonlinear comparison. These descriptions combine parameter, activation, gradient, output, direction and interaction evidence, but lack paired response queries. Section 3.3 tests the same-query comparison separately.

E.6 Input-set prediction and calibration

A matrix estimated from one finite input set need not predict another set equally well. Fix an input distribution μ and set size n. Let $B_{S,u}$ be the response matrix for $S \sim \mu^{\otimes n}$, where u identifies a model and operation family. For exactly measured matrices, form a center by averaging across independent estimation input sets:

$$\widehat{B}_{\mu,n,u} = \frac{1}{N_{\mathrm{est}}} \sum_{j=1}^{N_{\mathrm{est}}} B_{S_j,u}, \qquad \widehat{y}_u(a) = a^\top \widehat{B}_{\mu,n,u}. \tag{11}$$

Here N_{est} is the number of estimation input sets. For paired-query construction, replace each $B_{S_j,u}$ by its estimated matrix $\widehat{B}_{S_j,u}$; its estimation error is evaluated separately. Every operator in this average shares named operation and response-coordinate IDs. Stack the fixed evaluation directions in X_{eval}. A legal coordinate relabeling is aligned before averaging or scoring: if permutations $P_{\mathrm{op}}, P_{\mathrm{resp}}$ induce $B' = P_{\mathrm{op}} B P_{\mathrm{resp}}^\top$ and $X'_{\mathrm{eval}} = X_{\mathrm{eval}} P_{\mathrm{op}}^\top$, then $X'_{\mathrm{eval}} B' P_{\mathrm{resp}} = X_{\mathrm{eval}} B$. Thus equivalent relabelings contribute zero uncertainty; permuting values without the corresponding IDs breaks the state–query meaning. For center C, calibration uses the score

$$e_{X_{\mathrm{eval}}}(C, B) = \frac{\|X_{\mathrm{eval}}(C - B)\|_F}{\max\{\|X_{\mathrm{eval}} B\|_F, \epsilon\}}, \qquad \epsilon > 0. \tag{12}$$

A calibrated radius defines a set of candidate future operators through $e_{X_{\mathrm{eval}}}(C, B) \le r$, not a confidence interval for the mean operator. Its coverage concerns this fixed bank of directions and exchangeable input-set draws. In the controlled study, both routes calibrate against directly computed derivatives; constructing a matrix without derivatives does not make its calibration derivative-free. Evaluation input sets are opened once.

Appendix A bounds proxy, frame and decoder perturbations. Section 4 relates these errors to finite edits.

Uncertainty on new inputs. Predictive sets cover 5087 of 5632 evaluation input sets (90.32%) in both access routes. The simultaneous pooled lower bound is 88.51%; every architecture-operation stratum passes its observed-coverage and lower-bound gate. Radius correlates 0.870 with point error. The previously rejected fixed radius threshold remains report-only even though it appears favorable on this registry; selective utility requires its own prospective evaluation.

F From Error Diagnosis to Input Allocation

Question and prospective design. The preceding MLP and GPT-2 studies find that differences between estimation and new input sets dominate local matrix-fitting error. We test the resulting allocation choice on new inputs: spend a fixed query budget on more directions at the same inputs, or on more distinct inputs. A development group motivates the comparison; all subsequent choices and the 192-case roster are committed before confirmation responses are observed. This is an allocation experiment, not a comparison of all tuned adaptive or nonlinear estimators.

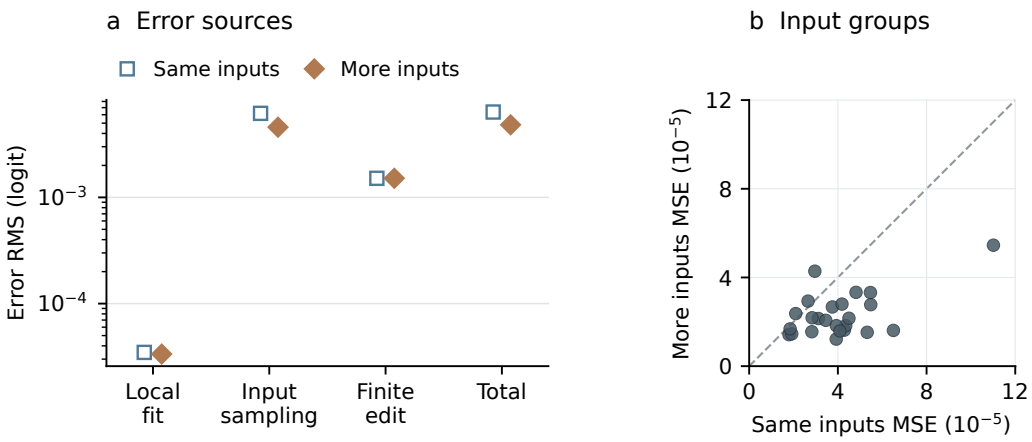

Figure 4: GPT-2: input-sampling error and a fixed-budget remedy.

The original GPT-2 checkpoint, four internal sites and four template-specific mean logit margins remain as defined in Appendix C. We add names that satisfy a response-independent single-token rule. All pairs from the original name bank are excluded, as are the new development and qualification pairs. Twenty-four disjoint groups each contain 20 pairs: four estimation sets and one evaluation set, each with four pairs. Every pair retains all 16 template, mention-order and subject variants. Thus each set has 64 prompts. Names and templates may recur, but no pair or prompt crosses these roles. A separate 16-pair qualification set passes the unchanged criteria: pair accuracy 1.0, vocabulary top-1 accuracy 0.9414, and mean margin 3.0092. No confirmation group is screened by its response or score.

Response-independent operation bases, four distinct orthonormal direction matrices and eight held-out directions are fixed across all groups. Both $d = 8$ and $d = 32$ are mandatory; their direction mixtures differ, so this is not a matched capacity comparison. Query steps are 0.1 and tested edit magnitudes are $\pm 0.1, \pm 0.5, \pm 2$. No weak direction or adverse case is removed.

Compared allocations and query cost. Let $S_0, \ldots, S_3$ be the four estimation sets, and let X_i be the ith direction matrix. Write $Y_{j,i}$ for the positive/negative response differences on S_j, divided by twice the query step. The same-input fit and expanded-input fit are

$$\widehat{B}_{\text{same}} = \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix}^{+} \begin{bmatrix} Y_{0,0} \\ Y_{0,1} \\ Y_{0,2} \\ Y_{0,3} \end{bmatrix}, \qquad \widehat{B}_{\text{more}} = \frac{1}{4} \sum_{j=0}^{3} X_0^{+} Y_{j,0}.$$

The first uses distinct directions, not repeated identical queries. Both use $2 \times 4 \times d \times 64$ input-endpoint evaluations: 4,096 at $d = 8$ and 16,384 at $d = 32$. An endpoint means one edited model evaluation for one prompt. The single-set control $X_0^{+} Y_{0,0}$ uses one quarter of that budget. Ordinary OLS fitted on the pooled 256 prompts is a second mandatory control. For this equally weighted response mean and common frame, $X_0^{+}(\frac{1}{4} \sum_j Y_{j,0}) = \widehat{B}_{\text{more}}$. Its numerical equality is expected, not evidence for a new LS estimator.

All matrices are estimated before evaluation edits or reference derivatives are accessed. The reference matrices on S_0, the pooled inputs and the evaluation set use 12 separately counted backward sweeps per case. They identify the terms in Equation 6; they are not free estimator inputs. Actual execution counts additionally include shared checks and repeated controls. Equal acquisition budgets do not imply equal total time or unique inputs.

What is averaged and what improves. The registered primary score is raw finite-edit MSE at absolute magnitude 2, averaged over both signs, eight directions and four response coordinates. The eight site/dimension cases are averaged within each group, and the 24 groups receive equal weight.

This is a ratio of mean MSEs, not a mean of direction-wise percentage gains. Table 11 retains all four methods. Figure 4 uses the same grouping.

Table 11: Fixed-budget allocation: primary raw prediction error.

Method	Prompts	Relative budget	MSE (10^{-5})
Single input set	64	1	4.0499
More directions, same inputs	64	4	4.0484
More distinct inputs	256	4	2.3260
Pooled-input OLS	256	4	2.3260

The realized mean reduction is 42.5447%. Twenty-one group means improve; groups 0, 9 and 17 worsen. Every site/dimension/magnitude mean improves, as shown in Table 12; these are not 24 independent significance claims. Actual magnitude-2 effects range from 0.00818 to 0.03579 logit RMS. Across all 1,152 case/magnitude evaluations, 131 activate the 0.001-logit floor used only for secondary normalized scores. Primary raw MSE does not divide by this floor, and every case remains in the score.

Table 12: Mean-MSE reduction by site, dimension and edit magnitude (%).

Site	d	0.1	0.5	2.0
Block 3 attention	8	43.71	43.52	40.90
Block 3 attention	32	49.09	48.86	45.87
Block 3 residual	8	38.96	38.88	36.87
Block 3 residual	32	43.87	43.65	40.56
Block 7 attention	8	46.06	45.96	43.69
Block 7 attention	32	49.37	49.30	47.48
Block 7 residual	8	44.42	44.34	42.30
Block 7 residual	32	45.25	45.22	44.13

Conditional uncertainty. Let W_c indicate whether expansion improves primary MSE in group c. Condition on the frozen checkpoint, operation geometry and each group's unordered 20-pair pool. The uniform pair permutation assigns roles uniformly within these pools. These within-pool assignments are conditionally independent: every combination has the same number of compatible global permutations. Each fitted predictor depends only on its own group. Hoeffding's inequality (Hoeffding, 1963), for independent but not necessarily identically distributed indicators, gives the predeclared one-sided 95% lower bound

$$\frac{1}{24}\sum_{c=0}^{23}\mathbb{E}[W_c \mid \text{fixed pools}] \;\geq\; \frac{21}{24} - \sqrt{\frac{\log(1/0.05)}{2\cdot 24}} = 0.62518.$$

This bounds the average probability of improvement under new role assignments within the fixed pools. It is not a confidence interval for the 42.5% mean reduction, future checkpoints or a language population. Sites, directions, tokens and templates are not additional independent replicates. The three adverse groups and all descriptive strata remain part of the result.

Why averaging helps, and what it leaves. Under the independent, identically distributed input-set model of Appendix A, define $\sigma_a^2 = \mathbb{E}\|a^\top(B_S - \mathbb{E}B_S)\|_2^2$. Averaging M exact matrices gives expected squared error $(1+1/M)\sigma_a^2$ against an independent future matrix: only the center-estimation part shrinks with M. This explains the direction of the remedy, not the exact numerical gain under our without-replacement sampling, estimated matrices and finite edits. The error-component vectors and all their cross terms are retained. Their RMS values in the figure are separate scales, not additive percentages of total error.

Execution and reproduction. The experiment uses full original-model logits for every method. Model arithmetic is float32 and response and operator arithmetic float64, with batch size 16 and TF32 disabled. The 192 timed cases take 2,866.1 seconds in total on one RTX 4090, with peak

allocated memory 2.56 GiB. This includes the compared methods and validation within each case, not loading, qualification, downloads or independent replay. It is not a per-method throughput benchmark. Every paired query, finite edit and reference derivative was independently recomputed through the unchanged checkpoint, with zero maximum discrepancy. The selected reproduction component retains all raw records, input roles, geometry and source identities, with separate CPU arithmetic and original-model entry points. This confirmation supports the tested allocation choice on one checkpoint, not an untested adaptive policy or a claim about weight-editing methods.

G Predicting Finite Attention-Head Replacements

From a local matrix to a component-level prediction. The previous allocation experiment holds the local approximation fixed. Here we ask whether fitting finite interactions more precisely is a better use of measurements than observing more inputs. We compare standard predictors of actual component replacements. The linear sufficiency theorem concerns a declared linear response space; it does not assert that finite head interactions are exactly linear.

Operation and measured response. We use the same pinned, frozen GPT-2 checkpoint and four language templates as Appendix C. For each unordered name pair, both mention orders and both subject choices give 16 prompts. A corrupted prompt changes only the repeated subject name to the other name. For example, the ending of "When Alice and Bob went to the store, Bob gave a book to" becomes "Alice gave a book to." The clean recipient is Alice and the clean subject is Bob; their logit identities stay fixed even on the corrupted prompt. This example illustrates the operation, not an evaluation pair. Clean and corrupted sequences have aligned token positions. At the prediction token of layer 8 (zero-based), let $h_j^c(x)$ and $h_j^d(x)$ be head j's corrupted and clean activations, respectively, before the attention output projection. For $j = 1, \dots, 12$, set

$$h_j(x;\eta) = h_j^c(x) + \eta_j\big(h_j^d(x) - h_j^c(x)\big). \tag{13}$$

The rest of the corrupted forward computation continues normally. Thus $\eta_j = 1$ replaces that head with its clean donor, while $\eta = 0$ leaves the corrupted computation unchanged. Replacing all heads at this one site does not replace the entire model computation with the clean one.

Write $r_S(\eta) \in \mathbb{R}^4$ for the four template-specific mean logit margins on input set S. Each margin is the clean recipient's logit minus the clean subject's logit. These two token identities remain fixed after corruption and replacement. A binary mask v specifies three or four heads; the prediction target is the finite change $F_S(v) = r_S(v) - r_S(0)$. This is not a derivative, a discovered circuit, or the official IOI benchmark score. The model's mean pair accuracy is 1.0 on fitting and validation pairs and 0.9929 on the 247 distinct evaluation pairs; the least accurate evaluation pair has accuracy 0.8125 and is retained.

Predictors and their information. All fitting statistics are unweighted averages over the specified input pairs. Let e_j be the jth coordinate vector, $\alpha_j = F_S(e_j)$ and $\beta_{jk} = F_S(e_j + e_k) - \alpha_j - \alpha_k$. The finite predictors are

$$\widehat{F}_{\text{single}}(v) = \sum_j v_j\alpha_j, \qquad \widehat{F}_{\text{pair}}(v) = \sum_j v_j\alpha_j + \sum_{j<k} v_j v_k \beta_{jk}. \tag{14}$$

The second interpolates queried singletons and pairs exactly; its prediction on larger subsets can omit higher-order interactions. Within this degree-two approximation class, the query features are $(v_1, \dots, v_{12}, (v_j v_k)_{j<k})$ and the stored coefficients are (α_j, β_{jk}). Theorem 1 applies to measurements and compression of that coefficient space; omitted higher-order effects remain approximation error. With four response components, the singleton and pairwise representations store 48 and 312 coefficients, respectively. The local predictor uses $\sum_j v_j \partial_j r_S(0)$, and the forward-only central-difference predictor substitutes $[r_S(0.01e_j) - r_S(-0.01e_j)]/0.02$ for each partial derivative.

Integrated gradients (Sundararajan et al., 2017) provide a stronger standard gradient control. With $\mathbf{1} \in \mathbb{R}^{12}$ the all-ones vector, we average $\partial_j r_S(t\mathbf{1})$ at the 16 midpoint nodes $t = (k + 1/2)/16$, $k = 0, \dots, 15$, then weight the result by v_j. An additional control averages derivatives at $\eta = 0$ and $\eta = \mathbf{1}$. The integrated path runs through *all* head coefficients; it is fixed before a held-out subset is given. Its completeness identity concerns $F_S(\mathbf{1})$, not every subset v. Integrating a separate

path to each evaluation subset would require new operation-specific queries and is not this reusable predictor. Subset error therefore includes the fixed path's approximation to subset-specific sensitivities, separately from quadrature error. A fixed 32-node check on the first four fitting pairs assesses numerical sensitivity without changing the primary rule. These are head-space controls, not complete EAP-IG, AtP* or ROME procedures.

Prospective inputs and operations. Development includes finite-head, input-transfer, input-selection and path-gradient studies. Their failures and results remain in the reproducibility record; the following confirmation uses none of their name pairs or subset masks. We first exclude the original 120-pair name universe and all 516 pairs exposed in the earlier expanded-name studies from the 48-name single-token bank. The remaining 492 pairs are shuffled with seed 94601: 68 for fitting, 48 for validation, and 376 for the sealed evaluation population. Names and templates can recur across roles, but pairs and complete prompts do not.

With seed 94602 we independently draw 16 evaluation sets of 24 pairs *with replacement* from that sealed population. There are 384 draws and 247 distinct pairs; repeated draws keep their multiplicities. Evaluation sets can overlap while their draws remain independent conditional on the finite population. All predictors and validation choices are fixed before their responses are queried. With seed 94603, 24 new validation masks and 24 disjoint evaluation masks each contain 12 three-head and 12 four-head subsets. The same evaluation masks are used in all 16 input sets; masks and template components are not treated as independent statistical replicates.

A budget must count the information actually acquired. The three forward-only methods require, per prompt, 24 queries for central differences, 13 for singleton effects (including the unchanged response), or 79 for pairwise effects (including 66 pairs). Each also needs one clean donor evaluation. For 16 prompts per name pair the costs are 400, 224 and 1280 forward-input evaluations. A forward-input evaluation is one model evaluation of one prompt, not a batched launch. Table 13 gives the fixed allocations. Cheaper methods receive more inputs, rather than being forced to use the expensive method's smaller fitting set.

Table 13: Predeclared fitting inputs and forward-query costs.

Budget	Predictor	Name pairs	Forward-input evaluations
5,120	Central difference	12	4,800
	Single-head effects	22	4,928
	Pairwise interactions	4	5,120
15,360	Central difference	38	15,200
	Single-head effects	68	15,232
	Pairwise interactions	12	15,360

Gradient access is counted separately. Per prompt, the local, integrated-gradient and endpoint-average methods require respectively 2, 17 and 3 forwards (including the donor), and 4, 64 and 8 backward response sweeps. Each Jacobian has four output components. These controls do not have the same access or cost as a forward-only candidate. Recorded wall times additionally reflect shared caches, batching and verification; they are not isolated latency comparisons between methods.

For each budget, validation selects the lowest-MSE forward-only candidate on its 48 pairs and 24 masks. It chooses the singleton predictor at both budgets. Acquiring all candidates and validation responses costs 33,728 and 62,464 forward-input evaluations, including 19,968 for validation in each case. Therefore this selector is not a same-total-budget improvement over choosing singleton effects in advance. The fixed-method results stand independently of that selection experiment.

Results and variation across inputs. For each evaluation set we take MSE over 24 masks and four response components, then average the 16 set-level scores. Tables 14 and 15 retain all nested fitting counts, not only the best prefix. Effects range from 1.0509 to 1.2510 logit RMS. The zero-prediction MSE is 1.360312; the secondary error/effect normalization never activates its 0.001 floor. No effect-size filtering is applied.

Table 14: New-input MSE across all fitting sizes (10^{-3}).

Pairs	Local	Central	Single-head	Pairwise	IG16	Endpoints
4	13.558	13.552	11.980	10.020	9.117	23.882
12	21.469	21.465	5.999	6.445	8.239	59.760
22	26.318	26.302	7.699	9.745	15.393	77.816
38	27.327	27.306	9.975	12.431	19.175	77.625
68	17.941	17.925	4.941	5.836	8.568	54.139

Table 15: Held-out-subset MSE on fitting inputs (10^{-3}).

Pairs	Local	Central	Single-head	Pairwise	IG16	Endpoints
4	2.260	2.258	1.679	0.00927	3.666	21.327
12	10.703	10.704	0.873	0.01088	1.633	41.741
22	7.028	7.024	0.983	0.00799	1.938	39.212
38	5.269	5.263	0.812	0.00498	1.630	31.905
68	7.128	7.118	0.575	0.00575	1.056	32.182

Pairwise effects give the best within-input approximation at every measured fitting size. At budget 15,360, however, singleton effects on 68 pairs reduce new-input mean MSE by 23.34% relative to pairwise effects on 12 pairs. Fifteen evaluation sets improve and one worsens. At budget 5,120, the mean also favors singleton effects, but eight sets favor each predictor and the interval includes zero. More inputs are not a monotone improvement along this one fitted input ordering. IG16 improves over local derivatives on all 16 evaluation sets at 68 fitting pairs; endpoint averaging does not.

Five comparisons are specified before evaluation. For each, we bootstrap the paired set-level differences 20,000 times (seed 94604). The percentile intervals in Table 16 are approximate, marginal 95% intervals conditional on this checkpoint, fitted predictors, validation choices, fixed operations and finite evaluation population. They are not simultaneous guarantees, do not include fitting/selection variability, and do not establish transfer to new names, models or free-form language.

Table 16: Predeclared paired MSE contrasts (10^{-3}).

Contrast	Mean	95% interval	Favor first
Single-head − pairwise, 5,120	-2.321	$[-6.638, 2.030]$	8/16
Single-head − pairwise, 15,360	-1.504	$[-2.028, -1.006]$	15/16
Selected − single-head, 5,120	0	$[0, 0]$	Tied
Selected − single-head, 15,360	0	$[0, 0]$	Tied
IG16 − local, 68 pairs	-9.373	$[-11.756, -7.039]$	16/16

What this adds to the diagnosis. A model of finite interactions can be excellent on the inputs used to build it yet be a worse budget choice for predicting new inputs. The relevant comparison is therefore not simply which approximation has the smallest within-input residual. It must also account for how many inputs the required measurements afford. This result complements the exact three-error decomposition for additive edits. The present study directly checks finite predictions and within-input approximations; it does not compute a complete three-source decomposition for these new head replacements.

Numerical verification and runtime. Source, input populations and operation masks are committed before model queries; validation choices are sealed before evaluation. All 480 set-level MSE entries are independently rebuilt from saved responses and coefficients. A separate hook and metric implementation replays eight fixed finite effects and one derivative table directly through the frozen model: maximum absolute differences are zero and 1.20×10^{-7}, respectively. This is an independent implementation check, not a second full confirmation run. An isolated reproduction component subsequently reexecutes all 68 fitted coefficient sets, 48 validation pairs, 247 queried evaluation pairs

and 68 fitting-input diagnostics. Their recorded numerical values reproduce exactly; 24 component tests also pass. This replay adds no statistical samples. The full study takes 152.52 seconds with peak allocated CUDA memory 652.52 MiB on one RTX 4090, using FP32 model computation, FP64 response arithmetic and 16-prompt batches. Parameters are unchanged and hooks removed. Evaluation and post-evaluation diagnostic accesses are separately recorded; the latter never choose predictors. The results concern this fixed checkpoint, site, prompt family and finite operation suite.

H Additional Reproducibility Checks

The 11 formal training-seed groups and all derived seeds were absent from historical, development, diagnostic, and previous formal registries. Each target was trained and qualified before proxy or state construction. The estimation, calibration, and evaluation input sets are disjoint. The derivative and restricted-query methods share directions and response units but retain separate records. All 44 model-operation cases are required, including failed cases.

The frozen evaluation also checks source cleanliness, configuration and report hashes, deterministic replay, target immutability, split provenance, device and dtype consistency, and zero CUDA allocation. These integrity gates are conjunctive with the scientific claims rather than post-hoc documentation.

Reproduction checks retain fitted transformations with their readouts. PCA signs can vary across backends and affect a fixed nonlinear readout if left unaligned. Coordinate mismatches are distinguished from changes in subspace or predictive information.