LOCAL MATCHING CAN MISS FINITE BEHAVIOR IN MODEL-EDIT INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Local measurements are often used to characterize interventions in a language model. When does agreement on these measurements imply similar effects at a finite intervention scale? We introduce FINITEFIBER, an audit that constructs distinct interventions with matched local profiles and then compares their complete-target log probabilities. The evaluated score is the same score whose derivatives enter the matching constraints. On an independent cohort of 40 Qwen2.5-3B factual edits, a profile matching residual geometry, positionwise derivatives at two anchors, calibration energies, and directional curvature admits controls for 35 facts. With a 1% energy tolerance, 23 facts exhibit score differences of at least .10 nat/token; the prespecified majority-rate test is not significant ($p = .215$). A separate shared eight-dimensional activation family clarifies the role of measurement resolution: two aggregate slopes admit finite-effect differences, whereas retaining the positionwise readings of the same edited-anchor gradient identifies the linear coefficients exactly on all 40 facts. This stronger measurement also rejects controls with small score differences, so it does not establish a selective verification procedure. Local scores nevertheless rank restricted LoRA subedits effectively. These results distinguish local equivalence testing from candidate ranking and show that the implications of local agreement depend on the candidate family and the information retained by the measurement.

1 INTRODUCTION

Interventions on internal representations provide a way to examine how a model edit changes a language model. A proposed component of an edit can be characterized by its reconstruction of the activation update, its magnitude under a calibration metric, and the sensitivity of the target score to that component. These measurements describe different aspects of an intervention. Their agreement raises a further question: does it justify treating two interventions as equivalent at the scale where they will actually be applied?

This question concerns the inference from a measurement to a behavioral claim. Prior work has shown that activation-patching choices can alter localization results (Zhang & Nanda, 2024) and that a successful subspace intervention need not recover the mechanism used by the unmodified model (Makelov et al., 2024). Here we examine the sufficiency of a specified local acceptance rule. The rule combines geometric constraints with derivatives of the same target score used for finite evaluation. Matching this score throughout the protocol separates the limitations of local information from differences between the objectives used for construction and evaluation.

Let H_0 and H_1 denote the activation states before and after an edit at a fixed residual-stream intervention site, and let $\Delta = H_1 - H_0$. Retaining a component C of this update is equivalent to removing the residual $D = \Delta - C$ from H_1, since $H_1 - D = H_0 + C$. For paraphrase p, write $M_p(D)$ for the local measurement profile and $B_p(D) = f_p(H_{1p} - D)$ for the finite target score. Given a reference residual a_p and acceptance tolerances τ, the set of accepted alternatives is

$$\mathcal{F}_{\tau,p}(a_p) = \{D : \mathrm{Accept}_\tau(M_p(D), M_p(a_p)) = 1\}.$$

Our objective is to find accepted alternatives with different values of B_p. Such a pair demonstrates that the specified profile and tolerances do not guarantee finite-score proximity in the declared candidate family.

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.

FINITEFIBER implements this test by constructing a distinct control, verifying its constraints independently, and evaluating its finite effect only after the candidate has been fixed. The audit is one-sided: a verified difference is a counterexample, whereas an unsuccessful search leaves the question unresolved. We apply it first to a broad activation family, then examine a restricted family whose coefficients are shared across paraphrases. Throughout, finite behavior is measured by the teacher-forced mean log probability of the complete target answer.

The paper makes three contributions. First, it provides an executable matched-control audit and independent evidence that a rich local profile can admit substantial finite-score differences. Second, it isolates the role of aggregation in a shared eight-dimensional family: retaining positionwise gradient measurements restores exact linear identifiability that is lost in aggregate slopes. Third, it separates this equivalence question from ranking a restricted set of executable LoRA subedits, where local scores remain useful. Together, the results characterize what a local measurement supports about a specified candidate family. The combined acceptance rule is a controlled test of sufficiency; it is not attributed to a particular published evaluator.

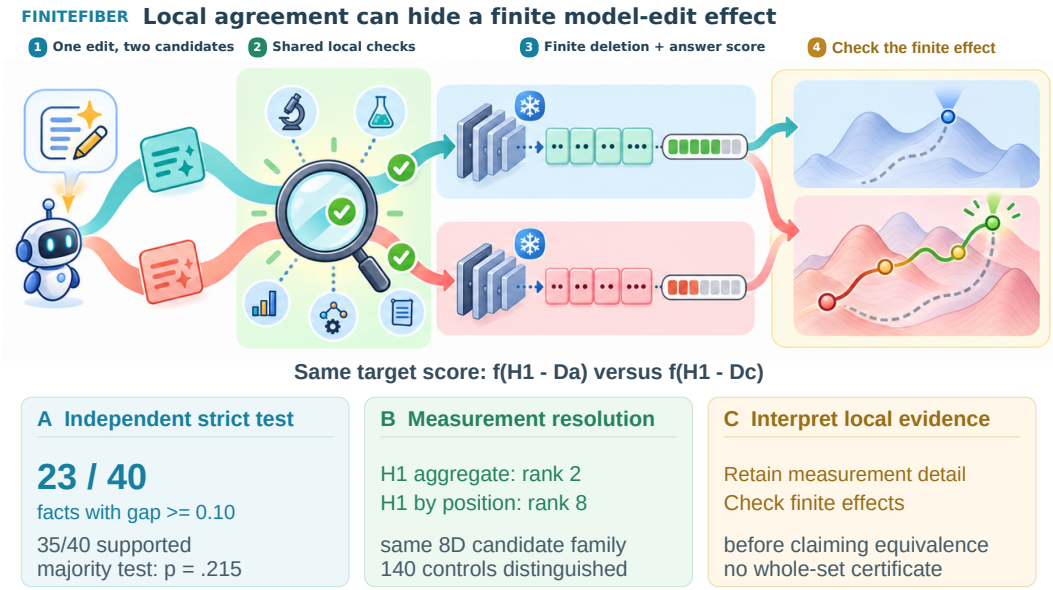

Figure 1: Overview of the audit and its implications. (1–3) Two residual interventions satisfy the specified local constraints but can produce different complete-target scores; paths and score bars are schematic. (A) The independent 1% energy-tolerance cohort contains 23 facts with gaps of at least .10 nat/token among 35 supported facts out of 40; the prespecified majority-rate test is not significant. (B) In a separate shared 8D linear family, retaining positionwise H_1 slopes raises the measurement rank from two to eight. These measurements, and separately the additional aggregate H_0 slopes, distinguish all 140 previously accepted controls. (C) Local agreement should be interpreted relative to its candidate family and checked against finite effects when equivalence is the intended claim.

2 PROBLEM FORMULATION AND AUDIT PROTOCOL

Intervention and outcome. We intervene at the residual state after layer 17 of an edited Qwen2.5-3B model and evaluate the unchanged base-model suffix. For a residual D, the intervention $H_1 - D$ retains the component $C_D = \Delta - D$. On paraphrase p, $f_p(H)$ is the teacher-forced mean log probability over all target-answer tokens obtained by continuing the model from H. The derivatives used for matching and the finite scores are therefore computed from the same function.

Let $\mathcal{M}$ be the valid token positions. We parameterize the residual magnitude by differential fidelity (qDF), a normalized reconstruction score:

$$\mathrm{qDF}(\Delta, D) = 1 - \frac{\|D\|_{\mathcal{M}}^2}{\|\Delta\|_{\mathcal{M}}^2}. \tag{1}$$

The primary setting is $\mathrm{qDF} = .25$, corresponding to $\|D\|_{\mathcal{M}} = \sqrt{.75}\|\Delta\|_{\mathcal{M}}$. Matching the residual radius thus also matches qDF; these are two expressions of the same scalar constraint. They do not determine the direction of the retained component. Table 1 summarizes the common experimental setting.

Table 1: Experimental setting. Each fact is evaluated on its first two CounterFact paraphrases. The independent cohorts test different candidate families and acceptance profiles.

Data	CounterFact (Meng et al., 2022). The original 40-fact cohort is hash-selected from IDs 452–799; the independent 1% cohort from 3200–3599; and the independent shared-family cohort from 800–1199. Calibration (100–111), development (120–139), and cohort-specific auxiliary training facts are disjoint.
Model and edit	Qwen2.5-3B base (Qwen Team, 2025), revision `3aab1f19`; rank-8 LoRA (Hu et al., 2021) on the layer-17 MLP down-projection; ROME-source calibration geometry
Evaluation	Post-layer-17 residual intervention; frozen layers 18–35, final normalization, and LM head. Prompt and target are truncated to 128 tokens in total.
Outcome	Mean complete-target log probability on each paraphrase, followed by a fact-level comparison; primary $\mathrm{qDF} = .25$

Local acceptance profile. For each valid position r, let $g_{kr} = \nabla_{H_r} f_p(H_k)$, where $k \in \{0, 1\}$ denotes the unedited or edited anchor. The base profile contains the residual radius and the two directional derivatives at each position, together with two calibration energies:

$$M_p(D) = \big(\{\|D_r\|, g_{0r}^\top D_r, g_{1r}^\top D_r\}_{r\in\mathcal{M}}, E_0(D), E_1(D)\big).$$

A control c is required to match the reference residual a:

$$g_{0r}^\top c_r = g_{0r}^\top a_r, \quad g_{1r}^\top c_r = g_{1r}^\top a_r, \quad \|c_r\| = \|a_r\|. \tag{2}$$

These equalities specify the matching targets; numerical acceptance uses fixed absolute and relative tolerances. To define calibration energy, let F_k stack normalized rewrite-score gradients from the calibration edits at anchor k, and let $C_k = F_k^\top F_k$. Then

$$E_k(D) = \|DF_k^\top\|_F^2 = \sum_{r\in\mathcal{M}} D_r^\top C_k D_r.$$

Both E_0 and E_1 are matched. The curvature-augmented profile also matches the edited-anchor quadratic $q_{H_1}(D) = \frac{1}{2} D^\top \nabla^2 f_p(H_1) D$ to within .01 nat/token per paraphrase. The independent confirmation uses this augmented profile with a 1% relative energy tolerance. Complete tolerances are given in Appendix 6.

Constructing a distinct control. The reference residual, called the *active* residual in the figures, is obtained by projecting Δ into a calibration-fitted orthonormal span $U \in \mathbf{R}^{d\times 32}$ and rescaling each position to the required radius. The span favors path-gradient drift relative to endpoint-gradient covariance, using only calibration facts. Its generalized eigenproblem and normalization are specified in Appendix 6.

To ensure that the control is distinct, construction additionally requires $U^\top c_r = 0$, while $a_r \in U$. This restriction defines the searched control family; membership in U is not a matched measurement. A minimum-norm affine solution and its nullspace parameterize the positionwise sphere in Eq. 2. The base search uses 64 seeded initializations and at most 750 optimization steps to match the two energies. The curvature extension starts from a dose-matched control and applies 48 projected Gauss–Newton steps. Accepted candidates are checked independently after FP32 serialization. Neither search objective uses the new control's finite score.

Fact-level outcome and experimental status. A fact is supported when both paraphrases have accepted controls. Its signed contrast is

$$\Gamma_i = \tfrac{1}{2} \sum_{p=1}^{2} \left[f_p(H_{1p} - D_{ap}) - f_p(H_{1p} - D_{cp}) \right]. \tag{3}$$

We average signed prompt contrasts before taking $|\Gamma_i|$. The prespecified event is $|\Gamma_i| \geq .10$ nat/token; it need not occur separately on both paraphrases. This threshold corresponds to about a 10.5% ratio of geometric-mean per-token probabilities and is an effect-size criterion, rather than a factual-correctness label.

For N planned facts, K supported facts, and V facts with such an event, we report K/N, V/K, and V/N. Unsupported facts have no finite score. The prespecified one-sided binomial test evaluates whether the full-cohort discovery rate exceeds one half; uncertainty intervals resample whole facts. Independent cohorts were fixed before edit fitting. Analyses that reuse facts with previously observed outcomes are identified as same-cohort follow-ups, even when their newly constructed controls were fixed before evaluation. Table 2 summarizes the audit.

Table 2: FINITEFIBER audit protocol. The acceptance check and finite evaluation are separate stages.

Input	Edited and unedited states, two paraphrases, calibration-fitted U, F_0, F_1, profile M, tolerances τ, residual radius, and search budget.
Construction	Form a reference residual for each paraphrase and search for a distinct control satisfying the specified constraints.
Verification	Recompute the constraints from saved FP32 candidates. If either paraphrase has no accepted control, report construction unresolved.
Evaluation	Fix the accepted candidates, score $H_1 - a$ and $H_1 - c$ on the same complete target, and compute Γ_i.
Interpretation	A gap above δ is a witness against the specified equivalence rule. A smaller gap establishes proximity only for the tested pair. Neither outcome certifies the entire accepted set.

3 FINITE DIFFERENCES UNDER MATCHED LOCAL PROFILES

Table 3 separates the four principal comparisons by candidate family, profile, and experimental status. The broad activation audit constructs one control per paraphrase. The shared-family experiment considers at most four controls per fact, so its discovery rate refers to the existence of a witness within that fixed search budget.

Table 3: Primary evidence. K/N is construction support, V/K is the event fraction among supported facts, and V/N is the discovery rate over all planned facts. Events use $|\Gamma| \geq .10$ nat/token. The first three rows evaluate one pair per fact; the shared-family row counts an event if any of at most four controls exceeds the threshold.

Candidate family	Acceptance profile	Cohort	K/N	V/K	V/N
Promptwise activation	Radius, positionwise H_0/H_1 slopes, energies	Original	40/40	31/40	31/40
Promptwise activation	Base profile + H_1 curvature; 10% energy tolerance	Same-cohort	34/40	25/34	25/40
Promptwise activation	Base profile + H_1 curvature; 1% energy tolerance	Independent	35/40	23/35	23/40
Shared 8D hook $V_p\theta$	qDF + two aggregate H_1 slopes	Independent	35/40	30/35	30/40

3.1 INDEPENDENT CONFIRMATION WITH A STRICT PROFILE

The independent 40-fact cohort was selected and fixed before edit fitting, control construction, and finite evaluation. Under the curvature-augmented profile with 1% relative energy tolerance, both paraphrases admit accepted controls for 35 facts. Of these, 23 have $|\Gamma_i| \geq .10$ nat/token, giving a full-cohort discovery rate of 23/40. The median supported gap is .138 nat/token, with a 95% fact-bootstrap interval of $[.085, .309]$. The five unsupported facts remain unscored. Figure 2 shows the gap distribution and full-cohort outcomes; protocol details are in Appendix E.1.

These observations establish finite-score differences among candidates that pass the stricter local rule on independent facts. The prespecified majority-rate hypothesis, however, is not confirmed: the one-sided test requires at least 26/40 events and yields $p = .215$. The evidence therefore supports the existence and measured magnitude of the differences, while leaving the population majority claim unresolved.

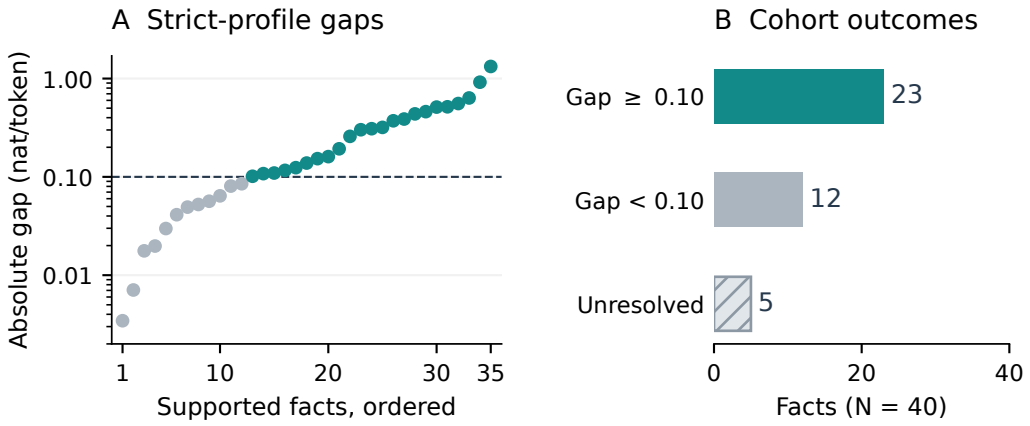

Figure 2: Independent confirmation under the curvature-augmented profile with 1% energy tolerance. (A) Absolute gaps for the 35 supported facts; the dashed line marks .10 nat/token. (B) Outcomes for all 40 planned facts: 23 witnesses, 12 evaluated pairs below the threshold, and five unresolved constructions. Unsupported facts are not assigned zero gaps. Each fact averages signed contrasts across two paraphrases before taking the absolute value. The full-cohort majority test gives $p = .215$.

For comparison, the original independent cohort used the base profile without curvature at the same qDF .25. All 40 facts were supported; 31/40 exceeded the threshold, with median $|\Gamma| = .223$ and 95% interval $[.161, .478]$. Its distribution is reported in Appendix B.1; signed differences occur in both directions. Since both the cohort and the profile change in the strict confirmation, the two rates do not isolate an effect of energy tolerance. Secondary recovery and margin outcomes are reported in Appendix B.1.

Dependence on the reference subspace. The primary candidates are orthogonal and have equal radii, so $\|D_a - D_c\| = 1.225\|\Delta\|$ at qDF .25. The median projected-update fraction is .152, corresponding to a normalization factor of 5.71. On the original cohort, three independent full-space random bases support all 40 facts and yield 27, 28, and 25 threshold-exceeding gaps, compared with 31 for the drift-selected basis under the same base profile and budget. Thus the observed ambiguity is not specific to drift-based subspace selection. This comparison concerns constructed candidates at the same substantial separation (Appendix E).

3.2 Curvature matching and intervention scale

Adding the edited-anchor directional quadratic to the base profile on the original cohort yields controls for 34/40 facts. Among those facts, 25 retain gaps of at least .10 nat/token, with median .238. The absolute edited-anchor first-plus-second-order contrast has median .00129 and maximum .00808 nat/token. Hence the observed low-order contrast is small relative to the finite differences. This comparison does not bound the effect of every permitted geometric or energy mismatch; the Taylor decomposition and numerical checks are given in Appendix B.

Re-optimizing this same cohort with a 1% rather than 10% relative energy tolerance supports 35/40 facts and finds 26 threshold-exceeding gaps among them. This result and the independent 23/40 confirmation show that the observed differences persist under tighter energy matching. The re-optimization is a same-cohort diagnostic because the earlier outcomes had already been evaluated.

Fixed-direction scaling. To isolate scale from direction, we rescale both members of the same 34 curvature-matched pairs by $t = \sqrt{(1 - \mathrm{qDF})/.75}$. This preserves the directions and fact support. Matched slopes scale with t, and calibration energies and directional quadratics with t^2; the scaled candidates pass fresh radius and qDF checks. The number of threshold-exceeding facts decreases from 25/34 at qDF .25 to 15/34 at .75, 3/34 at .90, and 0/34 at .975. The corresponding median gaps

are .238, .058, .030, and .0033 nat/token (Appendix B.3). This paired analysis demonstrates scale dependence along the tested directions. It does not bound the differences obtainable by searching other directions at the smaller scale.

3.3 CONSTRUCTING MATCHED CONTROLS UNDER A FIXED BUDGET

An audit can produce a witness only if its search finds a control that passes the constraints. We compare projected refinement with two sampling baselines initialized from the same dose-matched control. The first evaluates 63 global initializations and the starting control. The second applies seven seeded rotations within a common nullspace, preserving both calibration energies before the independent FP32 curvature check. All methods use the same reference residual, calibration span, positionwise affine sphere, and acceptance tolerances.

Projected refinement supports 34/40 facts and finds 25 witnesses. The stronger rotation baseline supports only 2/40 facts and finds one witness with a similar incremental HVP count (Table 4). This comparison supports the constructor's ability to obtain auditable pairs under the tested budget. It does not establish global search optimality. Wall-clock times exclude shared preprocessing and were measured in separate GPU runs; they are reported descriptively. New controls in all arms were fixed before their finite scores were evaluated.

Table 4: Incremental cost and audit yield from a common dose-matched initialization. Support is reported for prompts and for facts with two accepted prompts. Witnesses are counted over all 40 planned facts. Times exclude calibration, initial-control construction, and endpoint gradients. See Appendix A for the searched subfamilies.

Constructor	HVP	Time (s)	Prompts/80	Facts/40	Witnesses/40
Initial control + 63 global samples	160	118	1	0	0
Initial control + 7 nullspace rotations	720	176	10	2	1
FINITEFIBER projected refinement	692	187	74	34	25

4 CANDIDATE IDENTIFIABILITY AND LOCAL RANKING

A shared candidate family. To examine how candidate restrictions change the meaning of local agreement, we use a separate 40-fact cohort fixed before fitting its LoRA edits. The eight singular-component deletions of each edit supply hook-state directions V_p. A candidate takes the form $D_p = V_p\theta$, with one coefficient vector shared across the two paraphrases. This experiment initially matches qDF and two aggregate edited-anchor slopes, rather than the full positionwise profile used in Section 3.

The aggregate measurement map on the eight-direction span has rank two on all 40 facts. Starting from the same feasible references in a 2D subfamily, expansion to 8D yields accepted controls for 35/40 facts and at least one finite gap of .10 nat/token for 30/40 facts, within a budget of four controls per fact. The prespecified one-sided test gives $p = .00111$. Scaling these directions to qDF .90 leaves 6/40 events; a subsequent search performed directly at that scale finds 13/40. The latter is a same-cohort follow-up, and the paired mean increase on common support is unresolved. These experiments establish finite differences in the shared hook family while separating fixed-direction scaling from renewed search. Figure 3 summarizes candidate expansion and fixed-direction scaling; Appendix D gives the construction protocol.

Separating aggregation from anchor information. For the same 40 facts and eight directions, we compare aggregate and positionwise slope measurements using either H_1 alone or both anchors. All measurements are linear in the shared coefficients: $m(\theta) = A\theta$. For a fixed k-dimensional family,

$$A(\theta_a - \theta_c) = 0, \quad \operatorname{rank}(A) = k \quad \Longrightarrow \quad \theta_a = \theta_c.$$

Full column rank excludes distinct candidates with identical measurements. It does not imply accurate Taylor predictions for different candidates; nonzero tolerances additionally require conditioning and score stability. We compute rank on all 40 facts and re-evaluate acceptance on the previously fixed pool of 140 controls from the 35 supported facts (Table 5).

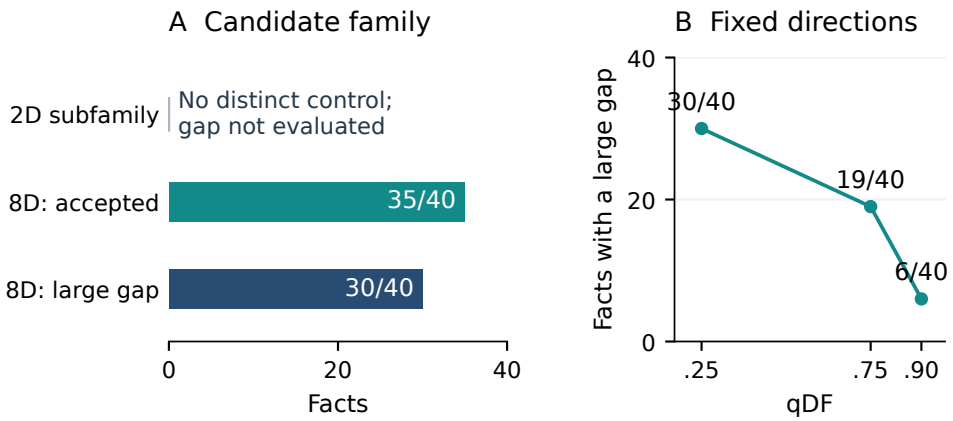

Figure 3: Candidate family and scale under the two-aggregate-slope and qDF rule. (A) Counts are facts: 35/40 admit an 8D control and 30/40 have at least one gap $\geq .10$, with 140 controls in total. The 2D search supplies no distinct control; its finite gap is not evaluated. (B) The same 8D directions are evaluated at the indicated qDF values, placed at their numerical spacing. These counts do not bound a new search at each scale. Both panels use the independent shared-hook cohort.

Retaining the positionwise readings of the H_1 gradient increases the rank from two to eight on every fact. Thus this single anchor already contains enough linear information to identify the coefficients exactly in the tested family. The median number of retained scalars increases from two to 43. In a prospective implementation, saving these readings from the existing gradient computation requires no additional backward evaluation. Our retrospective analysis regenerated the gradients because the positionwise matrices had not been archived.

All three enriched measurements reject all 140 controls, including the 85 with large finite gaps and the 55 with small gaps. Additional aggregate H_0 slopes are therefore sufficient to distinguish this particular pool even though their matrix has rank only four. The pool comparison establishes neither selective retention of behaviorally similar controls nor the absence of other accepted candidates.

Table 5: Aggregation and anchor ablation in the same 8D linear family. Every fact has the indicated numerical rank, with singular-value threshold $10^{-8}\sigma_{\max}$. Reacceptance uses the same 140 controls: aggregate slopes have an absolute tolerance, whereas positionwise slopes have absolute and relative terms. Large gaps are at least .10 nat/token. These are old-pool results under the stated distinct tolerances.

Slope measurement	Rank	Rows†	All accepted	Large accepted	Small accepted
H_1 aggregate	2	2	140/140	85/85	55/55
$H_0 + H_1$ aggregate	4	4	0/140	0/85	0/55
H_1 by position	8	43	0/140	0/85	0/55
$H_0 + H_1$ by position	8	86	0/140	0/85	0/55

†Median number of measurement rows per fact.

We also examine conditioning after normalizing candidate distance by the mean two-prompt hook Gram matrix and measurements by anchor-fixed tolerances and row count. The median smallest singular values are 19.3 for positionwise H_1 and 20.8 for positionwise $H_0 + H_1$. These quantify linear identifiability under the stated normalization. Actual relative acceptance tolerances depend on both candidates, and finite-score guarantees additionally require score stability. Appendix D.3 provides the normalization, acceptance rules, and fact-level results.

Ranking executable subedits on independent facts. On the independent shared-family cohort, we compare three selectors among all eight hook states produced by actual LoRA singular-component

deletions. These candidates have a physical weight-edit origin; scores here use the FP32 frozen suffix. The selectors were fixed before child scores were evaluated. Focal regret is $R_i = \max_j B_i(j) - B_i(\widehat{j}_i)$. Table 6 shows that first- and second-order scores rank this restricted set effectively. The comparison includes all eight children and is distinct from the earlier focal-high-qDF selection and direct BF16 results in Appendix C. Ranking candidates with different measurements does not test their equivalence under a matched profile.

Table 6: Focal selection on the independent 40-fact cohort. All eight physical-child hook states are scored using the FP32 frozen suffix. Regret is in nat/token on the complete-target score.

Selector	Exact optimum	Mean regret	Regret $\geq .10$
Maximum qDF	5/40	.1112	11/40
First order	38/40	.00249	0/40
First order + half-HVP	39/40	.00248	0/40

5 Related work

Activation-patching audits examine how measurement choices affect interpretation. Zhang & Nanda (2024) study the effects of corruption, metrics, and patching procedures on localization; Makelov et al. (2024) demonstrate that successful subspace interventions can admit misleading mechanistic interpretations. The distinction between localization and editing outcomes is studied by Hase et al. (2023). Our audit asks whether candidates accepted by a specified geometric and derivative profile must have similar finite effects on the same target score.

AtP* addresses errors in gradient-based component screening and includes direct activation-patching verification (Kramár et al., 2024). Mechanistic Tomography studies measurement design and calibration for recovering intervention-effect maps (Erramilli, 2026). FiniteFiber instead constructs matched alternatives to measure finite-score variation within a declared acceptance set. It evaluates the sufficiency of that set's defining information, rather than estimating the error rate of a natural editor or replacing direct finite verification. Appendix H gives the broader comparison with editing, attribution, and interpretability methods.

6 Discussion and limitations

Local agreement constrains finite behavior only relative to a candidate family, a measurement profile, and an acceptance tolerance. The independent strict-profile experiment provides 23 verified witnesses among 40 planned facts, while leaving the prespecified majority-rate claim unresolved. The shared-family experiment supplies a complementary result: positionwise measurements from the edited-anchor gradient identify an eight-dimensional linear family that aggregate slopes do not identify. Local target scores can also rank restricted executable subedits effectively. These findings distinguish three questions: whether measurements identify a candidate, whether they justify finite equivalence, and whether they support a useful ranking.

The practical implication is to retain measurement resolution and evaluate the inference it is intended to support. Positionwise readings can preserve information already available in a gradient computation. When finite equivalence is the required conclusion, directly evaluating accepted interventions provides evidence that local agreement alone may omit. The current pool ablation rejects both large- and small-gap controls, so it establishes information loss without providing a selective verification algorithm.

Our conclusions concern specified activation families and target likelihoods. The broad-family controls are constructed separately for each paraphrase; shared hook coefficients do not establish strict BF16 weight matching or transfer to unseen prompts. Full rank identifies exact matches, while nonzero tolerances require additional score-stability assumptions (Appendix F). Free-generation correctness and broader edit quality remain untested. Dose, path, and auxiliary-target analyses are supporting diagnostics rather than independent replications.

AI USE STATEMENT

AI tools were used to assist with experimental design and manuscript editing. The authors have reviewed all content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses publicly available models and datasets, involving no human subjects or personally identifiable information. As our research focuses on diagnostic evaluation and interpretability, we foresee no direct negative societal or ethical impacts.

REPRODUCIBILITY STATEMENT

All mathematical formulations, experimental settings, and hyperparameters are detailed in the main text and appendix. The models and data used are publicly accessible. Complete code and execution scripts will be open-sourced upon publication.

REFERENCES

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster ICA under orthogonal constraint. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018. doi: 10.1109/ICASSP.2018.8461662. URL https://arxiv.org/abs/1711.10873.

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*, 2018. URL https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, 2006. doi: 10.1109/TSP.2006.881199.

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*, 2026. URL https://arxiv.org/abs/2607.08349.

Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*, 2023. URL https://arxiv.org/abs/2307.12976.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023. URL https://arxiv.org/abs/2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://arxiv.org/abs/2104.08164.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. Qwen-Scope: Turning sparse features into development tools for large language models. *arXiv preprint arXiv:2605.11887*, 2026. URL https://arxiv.org/abs/2605.11887.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Vijay Erramilli. Mechanistic tomography: Designed measurement for control-oriented interpretability, 2026. URL https://arxiv.org/abs/2608.19338.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024. URL https://arxiv.org/abs/2406.04093.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

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*, 2025. URL https://arxiv.org/abs/2301.04709. arXiv:2301.04709.

Gene H. Golub and John H. Welsch. Calculation of gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969. doi: 10.1090/S0025-5718-69-99647-1.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2301.04213.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7ab63a5314680e2f083cb288abeaeb8e-Paper-Conference.pdf.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. doi: 10.1109/72.761722.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. URL https://jmlr.org/papers/v22/20-1223.html.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.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.

Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000. doi: 10.1162/089976600300015826.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.03341.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 689–696, 2009. doi: 10.1145/1553374.1553463.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.17030.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.08366.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024. URL https://arxiv.org/abs/2403.19647.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://arxiv.org/abs/2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, 2022b. URL https://arxiv.org/abs/2206.06520.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. doi: 10.1038/381607a0.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. URL https://arxiv.org/abs/2312.06681.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a. URL https://arxiv.org/abs/2404.16014.

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

Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/taking-features-out-of-superposition-with-sparse.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2004.00345.

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

Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023. URL https://arxiv.org/abs/2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.12265.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small. *arXiv preprint arXiv:2211.00593*, 2022. URL https://arxiv.org/abs/2211.00593.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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

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

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024. URL https://arxiv.org/abs/2401.01286.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Data, states, and suffix. The frozen primary dataset is CounterFact (Meng et al., 2022), source file `data/counterfact.json` with SHA-256 beginning `d01705612517`. The 40 fact IDs are selected by SHA-256 ordering from indices 452–799, independently of text and finite candidate scores; LoRA auxiliary facts occupy 14096–14375. Calibration facts 100–111 provide geometry, and development facts 120–139 fix design choices. Each record supplies a requested rewrite and paraphrase prompts; we take the first two paraphrases for evaluation. Prompt plus target is tokenized to a maximum of 128 tokens and the target mask selects only complete target-answer tokens for teacher-forced scoring. The primary Qwen/Qwen2.5-3B *base* checkpoint revision is `3aab1f1954e9cc14eb9509a215f9e5ca08227a9b`; the model index SHA-256 begins `bc8aaa0c87d4`. Its layer-17 MLP down-projection receives the rank-8 LoRA edit. H_0, H_1 are states at the post-layer-17 residual hook under the old and edited forward passes. Every candidate state is read out by the *same frozen base-model suffix*: layers 18–35, final normalization, and LM head. Padded positions have zero intervention. Exact suffix/full-model hook replay is checked before the candidate comparison. The ROME rank-one operator supplies the calibration geometry for the cross-source primary setting; evaluation uses LoRA.

Calibration span and dose. For each of the 12 operator-specific calibration rewrite edits j, compute complete-target score gradients at three states $H_{0j} + \alpha\Delta_j$, $\alpha \in \{0, \frac{1}{2}, 1\}$. Let m_j be valid-token count, $e_j = f_j(H_{1j}) - f_j(H_{0j})$, and $s_j = \max(|e_j|, \eta)$, where the calibration-only effect floor η is $\max(10^{-6}, .01 \times$ pooled median absolute rewrite effect$)$. For $k \in \{0, 1\}$, F_k stacks rows $\nabla_{H_r} f_j(H_{kj})/(s_j\sqrt{12m_j})$ over all valid calibration tokens r. Thus $F_k \in \mathbf{R}^{m_c \times d}$, $C_k = F_k^\top F_k$, and for a target-prompt direction D,

$$E_k(D) = \|DF_k^\top\|_F^2 = \sum_{r \in \mathcal{M}} D_r^\top C_k D_r. \tag{4}$$

The actual drift reference is the *within-prompt, within-position* weighted path-gradient mean, not an endpoint gradient or a population mean. Writing $G_{j\alpha r} = \nabla_{H_r} f_j(H_{0j} + \alpha\Delta_j)/s_j$, $w = (1, 4, 1)/6$, and $\bar{G}_{jr} = \sum_\alpha w_\alpha G_{j\alpha r}$, the implemented covariance is

$$C_{\text{drift}} = \frac{1}{12}\sum_{j=1}^{12} \frac{1}{m_j} \sum_{r \in \mathcal{M}_j} \sum_{\alpha \in \{0,1/2,1\}} w_\alpha (G_{j\alpha r} - \bar{G}_{jr})(G_{j\alpha r} - \bar{G}_{jr})^\top. \tag{5}$$

This direct positive-semidefinite sum, rather than a clipped difference, constructs C_{drift}; the code separately checks its identity with $C_{\text{path}} - C_{\text{secant}}$. Set $\widetilde{C}_d = C_{\text{drift}}/\operatorname{tr} C_{\text{drift}}$ and $\widetilde{C}_e = C_e/\operatorname{tr} C_e$ for $C_e = (C_0 + C_1)/2$. Let S be the top eigenspace of $\widetilde{C}_d$ retaining 95% of its nonnegative eigenvalue mass, with rank clipped to 64–128. The implemented generalized problem *inside that support* is

$$S^\top \widetilde{C}_d S\, v = \lambda \left(S^\top \widetilde{C}_e S + \frac{.001}{d} I \right) v, \qquad U = \mathrm{QR}(S[v_1, \ldots, v_{32}]). \tag{6}$$

The ridge uses the original channel width d, not the support rank. The top generalized vectors are Euclidean-orthonormalized only *after* the solve. This operator-specific fit uses calibration gradients and no held-out finite candidate scores.

Outcome-blind control search. For each paraphrase, project every valid row of Δ into U and rescale to the positionwise radius $\sqrt{1 - .25}\,\|\Delta_r\|$. Given the active row a_r, the control solves the profile constraints in Eq. 2 and the same-radius sphere. The constructor additionally enforces $U^\top c_r = 0$ to keep the witness distinct; this is not a component of the matched profile. The affine minimum-norm solution and its null space parameterize that sphere. A fixed seed supplies 64 starts; up to 750 Adam steps search only the remaining coordinates for the two global $E_k(c) = E_k(a)$ constraints. Selection minimizes the maximum normalized energy residual, with deterministic tie breaking; the pair is accepted only after independent FP32 checks of qDF, anchor gradients, calibration energies, and orthogonality. Padding and zero-update positions remain zero. Failure is recorded without substituting a fact or assigning a finite gap. The enhanced search starts from an accepted control and uses 48 fixed projected Gauss–Newton steps to add $|q_{H_1}(a) - q_{H_1}(c)| \le .01$ nat/token per paraphrase, while retaining all original checks. This is a scalar directional quadratic constraint, not a full-Hessian match. The cross enhancement accepted 74/80 prompt pairs and 34/40 facts with both paraphrases; five failed prompts exhausted the curvature search and one failed stored-control reprojection. Accepted active–control cosines are at most 5.4×10^{-10} in magnitude; the minimum enhanced/original distance ratio is .999999995.

Affine–sphere feasibility and searched subfamily. At a valid position stack the linear constructor constraints as $Ac = b$, including $U^\top c = 0$ and the two target-score slopes, and fix $\|c\| = \rho = \|a\|$. For a consistent affine system let $c_0 = A^\dagger b$ and let N have orthonormal columns spanning $\ker A$. Since $c_0 \perp \ker A$, every full-space solution at the required radius is

$$c = c_0 + \sqrt{\rho^2 - \|c_0\|^2}\, Nv, \qquad \|v\| = 1, \tag{7}$$

provided $\|c_0\| \le \rho$ and the needed null direction exists. If the remaining radius is zero, c_0 is the sole candidate; if it is positive but $\ker A$ is empty, the intersection is infeasible. The implementation avoids a full $d \times d$ null basis: it searches the energy-visible part of $\ker A$ plus one deterministic common-null filler. This is a reduced subfamily of Eq. 7, not a claim to enumerate the whole sphere. The positionwise test guarantees neither the two global energies nor the edited-anchor quadratic. Those quantities are searched and independently re-audited after FP32 serialization; infeasible local geometry and bounded-search non-attainment remain different failure codes.

Table 7: FP32 acceptance thresholds. The original cohort uses the 10% energy row; the strict independent confirmation uses the 1% row and the same remaining constraints, including the half-HVP. For gradients, radii, and energies, absolute and relative terms are added; the relative scale is the maximum absolute active/control value. Global and positionwise qDF use absolute thresholds. All accepted controls are rechecked after serialization.

Quantity	Acceptance tolerance
Global and positionwise qDF	10^{-6} each
Positionwise residual radius	$10^{-6} + 10^{-6} \max(\|a_r\|, \|c_r\|)$
Endpoint target-score slope ($k = 0, 1$)	$10^{-6} + .05 \max(\|g_{kr}^\top a_r\|, \|g_{kr}^\top c_r\|)$
Calibration energy, original ($k = 0, 1$)	10^{-8} absolute, .10 relative
Calibration energy, strict ($k = 0, 1$)	10^{-8} absolute, .01 relative
Added edited-anchor half-HVP	.01 nat/token per paraphrase

Table 8: Frozen focal-plus-auxiliary LoRA fit used by the Qwen2.5 confirmation and the later SVD subedit study. These values are from the archived Model-B YAML, not chosen after the candidate outcomes.

Quantity	Value
Changed module	layer-17 MLP down-projection only
Rank / alpha / training bundle	8 / 8 / focal fact plus seven auxiliary facts
Optimizer	100 steps, learning rate .01, weight decay 0
Locality term	KL coefficient 1.0
Strength selection	grid $\{0, .125, \ldots, 1\}$; closest to $f_0 + .95(0 - f_0)$
Tie break / evaluation	smaller strength / frozen two paraphrases

Phases and estimands. In Table 8, f_0 denotes the unedited focal rewrite's complete-target log probability; the grid minimizes distance to the stated target score, with smaller strength breaking ties. The original setting was chosen on an older, opened 60-fact cohort; the new 40 IDs, deletion branch, $\mathrm{qDF} = .25$, and .10 practical threshold were fixed before their outcomes. All 40 original constructions succeeded. Original confirmation and curvature enhancement have separate direction seals and finite score receipts, but the original finite outcomes were already open when enhancement began. The new control's score was not queried by its constructor. Enhancement reuses the same 40 facts and unchanged active directions; its support is conditional on both paraphrases succeeding. Fact-level inference uses signed prompt averages as in Eq. 3; promptwise threshold events are not the primary estimand. The original one-sided exact test has $H_0 : \Pr(|\Gamma_i| \ge .10) \le .5$, with 26/40 events required at $\alpha = .05$. Wilson intervals describe event rates; bootstrap intervals resample whole facts 10,000 times. A construction failure is a primary non-event in the original 40-fact design, and is unscored in the later conditional enhancement. This distinction preserves each frozen decision rule.

Figure 4 summarizes the independent strict-profile confirmation. Same-cohort extensions retain their separately stated chronology. Candidate-score blindness excludes new intervention endpoints, not the target scores required to fit the original edit.

Table 9: Cohorts and roles. Identifiers index the frozen CounterFact source; rows with different estimands are never pooled.

Setting	Fact IDs	Role
Qwen2.5-3B	452–799, 40 hash-selected	Original same-readout confirmation; enhanced follow-ups; auxiliary facts 14096–14375
Qwen2.5-3B	3200–3599, 40 hash-selected	Independent curvature-augmented 1% confirmation; auxiliary facts 16000–16279
Qwen2.5-3B	800–1199, 40 hash-selected	Independent shared-hook confirmation and child ranking; later same-cohort 2×2 audit; auxiliary facts 15000–15279
Qwen3-4B	220–239 / 240–259	Disjoint primary/replication cohorts for the static-width/path conjunction
Qwen3-4B	332–371	Prospective tail-risk ranking of numerical path error
Qwen3.5-4B	372–411	Absolute-cutpoint transport stopped before finite outcomes
Qwen3.5-4B	412–451	Separate local-scale ranking; same-cohort selective width-error reporting
Qwen3-8B / Qwen3.5-2B	260–279 / 312–331	Held-out official-SAE common-support qualification, with no finite behavioral scoring

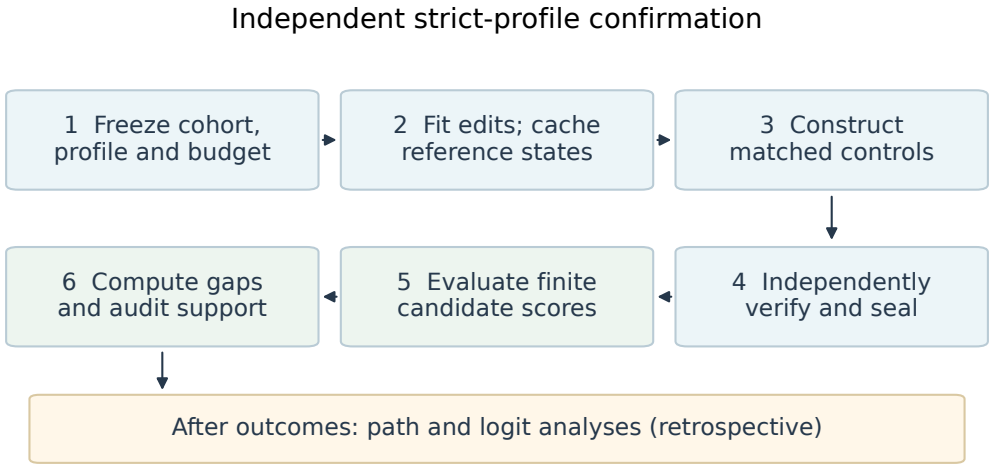

Figure 4: Protocol for the independent strict-profile confirmation. Candidates are constructed, independently checked, and fixed before their finite target scores are evaluated. Path and logit decompositions are subsequent analyses. The earlier width/path experiments use their own prediction protocols (Appendix G).

A MATCHED-CONTROL CONSTRUCTOR BASELINES

Both comparisons reuse the existing 40 Qwen2.5 facts, 80 paraphrases, source ROME span, accepted original control, edited-anchor half-HVP, and all original FP32 tolerances. They change only how a new control is sought *after* the original matched control is available. All 80 random construction records and their accepted directions were sealed before any new random-control finite score. The old active and optimized-control outcomes were already open; hence this is an algorithm comparison on the same cohort, not another independent fact confirmation.

The first baseline uses the existing affine-sphere constructor's seeded initial pool without gradient optimization: its old accepted control plus 63 global Rademacher starts per prompt, each with the source constructor's fixed start radius. All 5,120 attempts obey the local affine-sphere parameterization before serialization. Only the 80 warm controls pass both global dose checks; one also passes the new curvature and FP32 gate. No fact has two supported paraphrases, so this baseline has no conditional finite-gap distribution. Its zero discovered witnesses are a *bounded construction result*, not evidence of equal behavior.

The stronger baseline retains each old control's energy-visible null coordinates and rotates only its common-null component. At a position, write the old control as $c_0 = m + \rho(v + w)$, where m is the minimum-norm affine solution, v lies in the energy-visible affine nullspace, and w lies in the joint nullspace of the affine rows and both dose factors. For unit ξ in that joint nullspace with $\xi \perp w$, sample

$$c(\theta) = m + \rho\,[v + \|w\|(\cos\theta\,\widehat{w} + \sin\theta\,\xi)]\,. \tag{8}$$

Thus local affine conditions, radius, and both E_k are unchanged in exact arithmetic; independent FP32 checks still decide acceptance. Rademacher vectors are projected twice from the affine row space, energy-visible basis, and $\widehat{w}$. Seven fixed van-der-Corput angles $\pi \times (.5, .25, .75, .125, .625, .375, .875)$ plus the old control give eight attempts per prompt, without any optimization or finite-score input. All 640 attempts pass the dose gate; 15 pass curvature and FP32 checks, occurring in 10 distinct prompts. Only facts 463 and 696 support both paraphrases; their absolute gaps are .02044 and .15634 nat/token. Table 4 contrasts these with the unchanged optimized 74/80 prompts, 34/40 facts, and 25/40 discovered practical witnesses. The common-null sampler consumes 720 HVP calls and 176 seconds of incremental search, versus 692 and 187 for projected refinement; these exclude calibration, original-control construction, and endpoint-gradient computation. The wall times were recorded in separate GPU runs and are descriptive, not a matched-load benchmark. It explores a broader common-null set than the optimizer's fixed filler, so it is a strong simple-search sensitivity baseline rather than byte-identical parameter-count comparison. Neither baseline establishes global optimality of the optimizer.

B SAME-READOUT EVIDENCE AND NUMERICAL RECONCILIATION

B.1 ORIGINAL CONFIRMATION AND SECONDARY READOUTS

The independent validator recomputes 31/40 events, exact one-sided $p = 0.0003397741$, Wilson interval $[.624969, .876839]$, mean absolute gap .698414, and median .222780 from all 40 saved fact rows. The complete-target log probability is the primary outcome. Recovery relative to the full edit has a stable denominator on 39 facts and median absolute active–control difference .16122; five facts have opposite mean first-token target–old margins. Neither diagnostic establishes free response correctness. Earlier Gemma2-2B and Qwen2.5-3B settings show exploratory same-readout recovery gaps across operator cells and scales, but use previously opened directions and cannot replace this new-cohort confirmation. In the Gemma confirmation cohort, local second-order MAE ranges from .00151 to .15111 across four cells, while registered GJ4 MAE ranges from .000008 to .01698; those Hessians were measured, not matched as in the new Qwen2.5 enhancement. Figure 5 is a post-outcome Gemma witness of why the matched target-loglikelihood and an unmatched first-token margin must not be treated as the same behavioral readout.

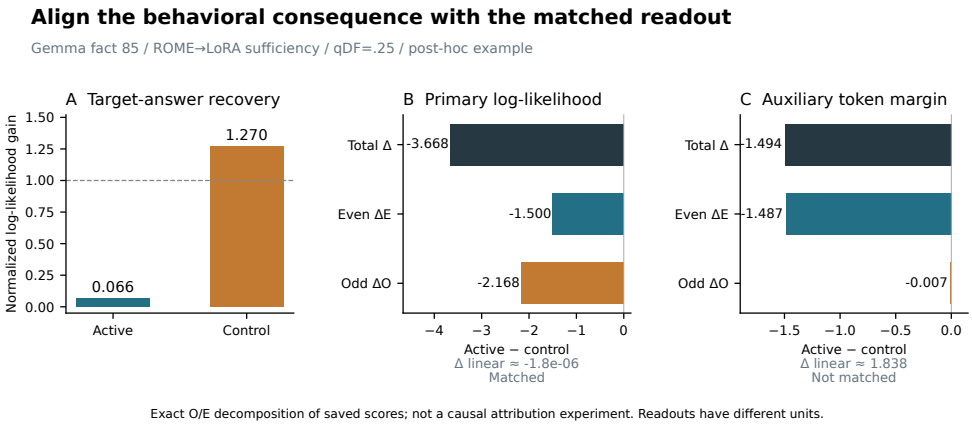

Figure 5: One post-hoc witness illustrates the readout boundary. The original qJ constraint matches the complete target-loglikelihood derivative (panel B), not the first-token target–old margin derivative (panel C). The margin flip is almost entirely even in this case, while the target-loglikelihood difference is not. Values are exact decompositions of stored scores.

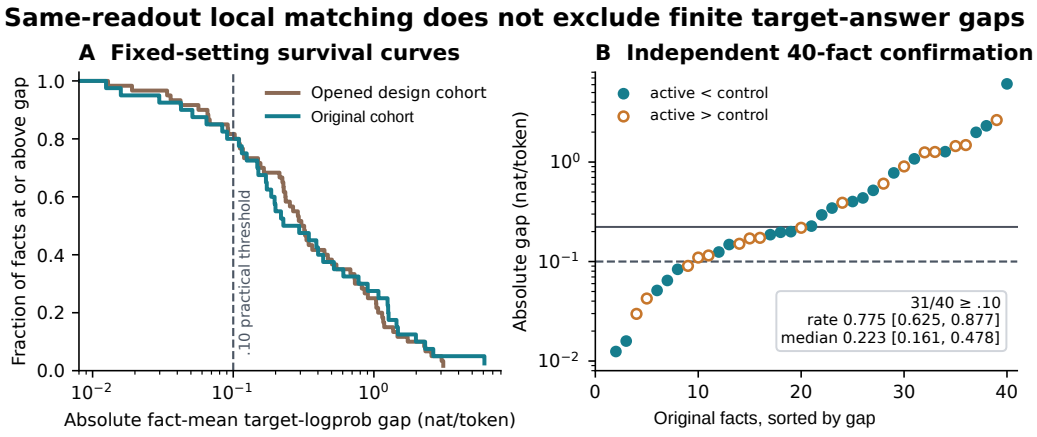

Figure 6: Original base-profile distribution before curvature matching. This earlier confirmation has 31/40 events. It is distinct from the independent 1% curvature-augmented cohort in Figure 2.

B.2 Enhanced support, edited-anchor contrast, and tail

For a fixed fact, define the promptwise contrast and its average by

$$h_p(t) = f_p(H_{1p} - tD_{ap}) - f_p(H_{1p} - tD_{cp}), \qquad h(t) = \frac{1}{2}\sum_{p=1}^{2} h_p(t).$$

Then $h(0) = 0$ and $h(1) = \Gamma$. Taylor's integral identity gives

$$\Gamma = \underbrace{h'(0) + \tfrac{1}{2}h''(0)}_{\ell} + \underbrace{\int_0^1 (1-t)[h''(t) - h''(0)]\,dt}_{V}, \qquad |V| \geq (|\Gamma| - \epsilon)_+. \tag{9}$$

Here ϵ bounds the measured edited-anchor contrast for each accepted pair; it is not a bound on propagated profile-tolerance error.

Table 10: Enhanced support and fact-level threshold transitions. Original →enhanced transitions use the same supported facts and different controls; failures receive no finite score.

Source	Both	One	Neither	Enhanced events
ROME→LoRA	34	6	0	25/34
LoRA→LoRA	12	22	6	9/12

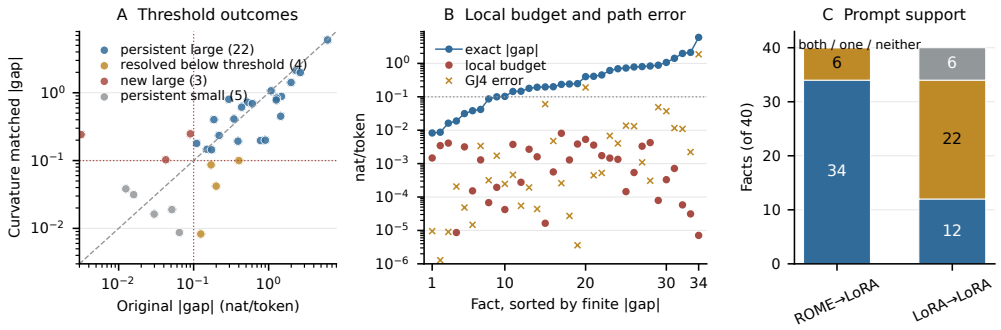

Figure 7: Original 40-fact same-cohort curvature follow-up: (A) paired threshold transitions on 34 supported facts; (B) edited-anchor Taylor budgets, finite gaps, and post-outcome path errors; (C) support under cross-source and native calibration geometry. Unsupported facts have no assigned finite outcome.

Within the 34 cross-supported facts, the original-to-enhanced threshold transitions are 22 large-to-large, four large-to-small, three small-to-large, and five small-to-small. The cross enhancement's paired mean change in absolute gap is $-.131$ [95% fact-bootstrap $-.240, -.025$], while the median of paired changes is $-.030$ $[-.158, .016]$. The difference between marginal medians is $-.08138$, a different statistic. All 74 cross and 46 native accepted prompt pairs pass the independent FP32 audit: maximum positionwise qDF error 7.46×10^{-8}, maximum ratio to allowed signed-gradient error .000484, and maximum dose-energy tolerance ratio .99022. The latter is close to its registered limit and is not assumed away.

For the promptwise curves $h_p(t)$ defined above Eq. 9, the per-fact conservative budget is $\epsilon_i = \frac{1}{2}\sum_p [t|h_p'(0)| + \frac{1}{2}t^2|h_p''(0)|]_{t=1}$. Its maximum is .0080752 and median .00129 nat/token; the maximum prompt-level budget is .00983. Exact saved-value subtraction gives a median lower bound .23347 for $|V|$, and the minimum lower bound among the 25 practical-gap facts is .10268; their median lower bound is .44917. The .23347 median uses all 34 supported facts. No formal floating-point error certificate is claimed, although old-score replay differs by at most 2.91×10^{-5} nat/token. In native geometry, promptwise success is 24/40 and 22/40, with 12 both-prompt facts;

the 22 one-prompt cases are mostly fixed-budget search failures. These counts alone do not identify a cross-paraphrase conflict mechanism. The budget bounds only the observed pair's edited-anchor first- and second-order score contrast. It does not propagate the separate calibration-energy or radius acceptance tolerances through the full model.

For physical deletion, the one-sided path identity is

$$f_p(H_1 - D) - f_p(H_1) = -\nabla f_p(H_1)^\top D + \int_0^1 (1-t) D^\top \nabla^2 f_p(H_1 - tD) D \, dt. \tag{10}$$

The GJ4 rule uses fixed nodes for its $(1-t)$-weighted integral. On 34 cross facts, its signed-gap MAE is .067312 versus .630993 for the local quadratic; the paired mean absolute-error change is $-.564$ [95% bootstrap $-.863, -.323$]. The largest error, fact 760, contributes 81.9% of total GJ4 absolute error. Table 11 retains both errors above .10 and their post-outcome fixed-node replay. Direct finite scores replay within 2.86×10^{-5} nat/token and original HVP nodes replay exactly. On five selected facts, 16- and 32-node weighted Gauss–Legendre rules agree within .00010, and 32-node individual-branch error is at most 4.13×10^{-5}. This diagnoses under-resolution of two selected paths; it does not alter the full-cohort GJ4 estimate.

Table 11: Two selected GJ4 failures in signed nat/token. Denser quadrature was run after outcome access to diagnose the tail; the fixed GJ4 estimates remain in every full-cohort summary.

Fact	Direct	GJ4	GL8	GL16	GL32
760	-6.01250	-4.13729	-5.91245	-6.01246	-6.01249
577	$-.40169$	$-.59220$	$-.37987$	$-.40176$	$-.40167$

B.3 Fixed-direction dose diagnostic

On the 34 cross-geometry facts supported on both paraphrases, the new versioned analysis scores the already sealed enhanced active and control directions at $H_1 - tD$, with $t = \sqrt{(1 - \mathrm{qDF})/.75}$ and qDF targets $.25, .50, .75, .90, .95, .975$. Neither the directions nor the two-prompt fact population are reselected. The local target slopes scale with t; the two calibration energies and edited-anchor directional quadratic scale with t^2. Thus the original active–control profile differences do not grow under the frozen mixed tolerances. Direct checks of every scaled state put the largest global qDF error at 4.89×10^{-8}; the largest positionwise radius error is 3.77×10^{-6} and passes its relative tolerance. New $t = 1$ signed fact gaps replay the original enhanced gaps within 1.77×10^{-5} nat/token. Gradients, energies, and HVPs were not separately recomputed at every scale.

Table 12: Same 34 constructed pairs and complete-target FP32 suffix at each dose. Intervals bootstrap whole facts; $|\Gamma| \geq .10$ retains its original threshold. This same-cohort sweep is descriptive.

qDF	t	Events/34	Median $\lvert\Gamma\rvert$ [95% CI]	Mean $\lvert\Gamma\rvert$
.25	1.0000	25	.2382 [.1468, .6135]	.6310
.50	.8165	21	.1461 [.0689, .3533]	.3584
.75	.5774	15	.0581 [.0430, .1451]	.1389
.90	.3651	3	.0296 [.0144, .0490]	.0462
.95	.2582	1	.0097 [.0050, .0174]	.0172
.975	.1826	0	.0033 [.0017, .0064]	.0056

At qDF .90, facts 497, 690, and 760 exceed .10; at .95, only 690 does. The aggregate decline is not pointwise monotone for every fact. Since the $t = 1$ outcomes were open before this sweep, these numbers locate a scale boundary for the constructed pairs and do not replicate the original confirmation on new facts.

B.4 Anchor-logits proxy and component cancellation

For each sealed direction, one JVP obtains all target-position logits slopes $J_z(H_1)D$. Scoring $z(H_1) - J_z(H_1)D$ with exact log-softmax gives a specified first-order proxy, with no fitted coefficient

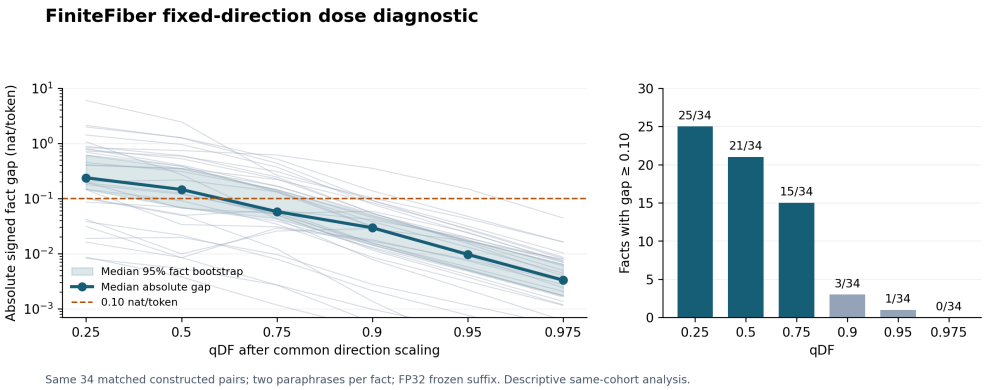

Figure 8: Individual fact trajectories (thin), median and whole-fact bootstrap band, and the unchanged .10 threshold for the same 34 enhanced matched pairs. The right panel keeps the common denominator.

or path query. An independent 136-direction JVP replay agrees with saved scalar slopes to 3.90×10^{-5} nat/token and proxy scores to 9.54×10^{-6}. Table 13 reports its errors on the identical cross/native candidate pairs. The cross proxy-minus-GJ4 paired mean absolute-error difference is .75838 [95% fact-bootstrap .43484, 1.13741]; on the selected native 12 it is .27358 [.12778, .44905]. Native and cross support overlap on 11 facts, so neither row is an independent behavioral replication.

The anchor logits proxy is not nested above the scalar quadratic. Write $f_D(t) = f(H_1 - tD)$ and let $\widehat{f}_D^{\log}(t)$ score the linearized logits $z(H_1) - tJ_z(H_1)D$ with exact log-softmax. From the saved per-role total HVP and readout-curvature records, a post-outcome check forms

$$\widehat{f}_D^{\mathrm{corr}}(1) = \widehat{f}_D^{\log}(1) + \tfrac{1}{2}\big[f_D''(0) - (\widehat{f}_D^{\log})''(0)\big]. \tag{11}$$

The bracket is the inferred anchor logits-curvature term. The stored total HVP is $f_D''(0)$, whereas the constructor's q_{H_1} is already $\frac{1}{2}f_D''(0)$; Eq. 11 applies the half factor exactly once. Recomputing every prompt/role from the saved records yields cross MAE .76958, median .29219, p90 1.65452 (34 facts), and native MAE .31026, median .16003, p90 .69901 (12 selected facts). The original logits proxy gives .82569 and .28260 on those same supports. This saved-value comparison fits no coefficient, makes no model call, and does not constitute a prospectively selected predictor. Its modest cross improvement and native deterioration show that repairing the omitted anchor quadratic alone does not reliably account for the finite response.

At the edited anchor, exact readout-curvature candidate contrast has median absolute magnitude .18761 in cross geometry; 24/34 exceed .10. The matched total HVP contrast has median absolute magnitude .00201. The residual logits-curvature contrast is *inferred* by subtracting readout from total, rather than measured as a separate full-logits Hessian; it has opposite sign in all 34 facts. Matching the total partly enforces this opposition, so the finding is component magnitude and cancellation, not an independently identified circuit. Two post-outcome selected cases show distinct proxy errors. For fact 690 the observed signed gap is $-.836$, the logits proxy gives -4.973, and GJ4 gives $-.836$; readout and inferred logits-curvature contrasts at the anchor are -4.154 and $+4.162$. For fact 760 the proxy gives $-.104$ against an observed -6.013, and GJ4 underestimates at -4.137.

C Executable LoRA subedit selection

Physical candidates and outcome firewall. The original confirmation cache did not retain LoRA factors. We refit the fixed focal-plus-seven-auxiliary bundle with the original code, seed, checkpoint, and strength rule for each of the same 40 facts. Every refit has rank 8 and reproduces the old H_0, H_1 on all valid tokens exactly. For the single modified down-projection, write its effective weight update as

$$\Delta W = \frac{\alpha}{8} sBA = \sum_{j=1}^{8} \sigma_j u_j v_j^\top, \qquad D_{pj}^{\mathrm{ideal}} = \sigma_j (X_p v_j) u_j^\top. \tag{12}$$

Table 13: Signed fact-level gap estimation on identical supported directions. Errors are nat/token. The native subset overlaps cross support on 11 facts; rows compare information sources, not equal computational budgets. The curvature-corrected proxy is a post-outcome, saved-value analysis.

Geometry	Estimator	MAE	Median	p90	Error $\geq .10$	Sign
Cross (34)	Local quadratic	.631	.237	1.313	25	19/34
	Anchor logits proxy	.826	.309	2.021	27	15/34
	Curvature-corrected logits	.770	.292	1.655	28	17/34
	Fixed GJ4 path	.067	.00049	.0453	2	34/34
Native (12)	Local quadratic	.300	.200	.673	9	8/12
	Anchor logits proxy	.283	.189	.723	8	8/12
	Curvature-corrected logits	.310	.160	.699	7	6/12
	Fixed GJ4 path	.009	.0050	.0258	0	12/12

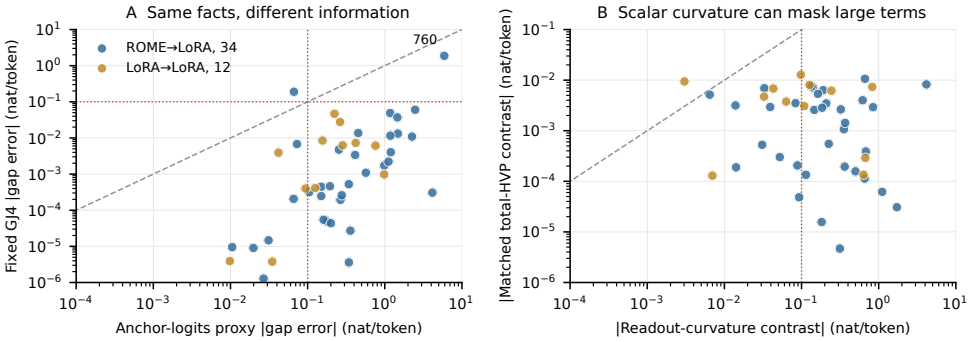

Figure 9: On the same sealed pairs, GJ4 errors are usually below the anchor-logits proxy's errors (A). At the anchor, a small matched total-HVP contrast can conceal much larger opposing readout- and logits-curvature terms (B). Fact 760 and native-support selection are retained.

Here X_p is the actual input to the down-projection on prompt p; thin QR of B and $A^\top$ reduces the decomposition to an 8×8 SVD. Candidate j uses the real weight update $\Delta W - \sigma_j u_j v_j^\top$ on both paraphrases. We capture its post-layer-17 state H_{pj}^{child} with an early-stop hook before the downstream suffix, so no candidate logits or target scores are computed during construction. The local direction is the *actual* $D_{pj} = H_{1p} - H_{pj}^{\text{child}}$, not D_{pj}^{ideal}. Ideal components obey $\sum_j \|D_{pj}^{\text{ideal}}\|^2 = \|\sum_j D_{pj}^{\text{ideal}}\|^2$ because the u_j are orthogonal; thus their average ideal qDF is $1 - 1/8$. BF16 casts make this an inexact identity for captured states: across facts the median average *actual* qDF is $.874924$, and the maximum relative L2 difference between summed ideal responses and the actual hook update is 2.84%. We report each candidate's actual qDF rather than substituting the ideal value.

After all 320 states were saved, three selectors were sealed for all 40 facts: largest two-paraphrase mean qDF; largest mean $f_p(H_{1p}) - \nabla f_p(H_{1p})^\top D_{pj}$; and that first-order score plus $\frac{1}{2} D_{pj}^\top \nabla^2 f_p(H_{1p}) D_{pj}$. Ties choose the lower singular index. The same suffix and complete-target teacher-forced metric as the main experiment supply these local scores. The unrestricted choice uses all eight candidates. A separately frozen high-fidelity choice uses only candidates with actual qDF $\geq .8$ on *both* paraphrases, requiring at least two; every fact has six or seven. Candidate scores were opened only after the prediction seal. For endpoint $E \in \{S, P\}$,

$$R_i^E = \max_{j \in \mathcal{J}_i} B_i^E(j) - B_i^E(\widehat{j}_i), \tag{13}$$

where S replays the actual child state through the paper's FP32 suffix and P runs the direct BF16 full-model weight subedit. The physical score is a separate external endpoint, not silently substituted for the paper's readout. All 320 physical scoring runs recaptured their sealed child hook state exactly. The 320 corresponding FP32/physical mean candidate-score differences have mean .0257, p90 .0542, and maximum .1121 nat/token.

Table 14: Mean *focal* regret (nat/token) for one executable singular-component deletion on 40 facts. The high-qDF family requires both-prompt qDF$\geq .8$. Random is exact uniform-choice expectation and minimum weight singular value is post-outcome descriptive; the other three rules were sealed before child outcomes.

Candidates	Selector	FP32 suffix	BF16 physical
All eight	Random (post-hoc)	.489	.495
	Min weight σ (post-hoc)	.182	.193
	Max qDF	.179	.192
	First order	.070	.067
	First + half HVP	.024	.024
Both-prompt qDF$\geq .8$	Random (post-hoc)	.239	.248
	Min weight σ (post-hoc)	.131	.147
	Max qDF	.129	.146
	First order	.014	.013
	First + half HVP	.014	.018

Results and interpretation. Table 14 reports the complete 40-fact outcomes, with 640 paraphrase-level child scores per endpoint. The second-order rule's unrestricted mean regret is .02373 [95% whole-fact bootstrap $.00075, .05774$] under the FP32 suffix and .02412 [$.00143, .06088$] under direct BF16 scoring. Its mean paired regret change from first order is $-.04672$ [$-.15093, .02761$] and $-.04290$ [$-.14126, .03162$], respectively; both intervals include zero. On the high-qDF family, second-order regret is .01387 [$.00000, .04162$] for the suffix and .01816 [$.00135, .05002$] physically. Only one fact exceeds .10 under either endpoint. The suffix selector reaches the exact finite optimum on 39/40 high-qDF facts; the physical selector reaches it on 29/40 but most remaining losses are small. These observations do not establish a reliable second-order failure rate in this restricted family.

Fact 592 is the sole high-qDF large-regret case: second order selects component 1 with qDF $.911/.833$, while the finite optimum is component 3. Its suffix/physical regrets are $.555/.619$ nat/token. The first-order rule selects component 3. In a *post-outcome selected case study*, the paper's unchanged four-node Gauss–Jacobi path rule predicts all eight suffix candidate scores within .0013 nat/token and selects component 3. This diagnoses an under-resolved endpoint quadratic on one case, not a prospective population-level path-selector result.

Post-outcome baselines, parent scores, and ranking error. On the same saved candidate sets, the exact expected regret of uniform random selection is the best candidate score minus the mean candidate score; no repeated random trials are needed. Deleting the smallest *weight* singular value is another simple baseline. It agrees with max-qDF on 33/40 facts and has similar but nonidentical regret (Table 14). Both were added after candidate outcomes opened and are descriptive, not part of the sealed choice protocol. They show that the low regret of target-aware rules is not obtained by choosing an arbitrary component.

Regret refers to maximizing the *focal* target, not preservation of the whole parent edit. With full-parent score B_i^{par},

$$R_i = \big(\max_j B_i(j) - B_i^{\mathrm{par}}\big) + \big(B_i^{\mathrm{par}} - B_i(\widehat{j}_i)\big). \tag{14}$$

The first term is the best child's gain relative to the parent; the second is the chosen child's score decrease relative to the parent, which may be negative if the deletion improves the target. On all-eight choices, 36/40 FP32 facts and 35/40 direct-BF16 facts have a child above the parent, with mean best-child gains .1645 and .1711 nat/token. Max-qDF's mean chosen-minus-parent change is only $-.0148$ FP32 and $-.0206$ physical, whereas its mean regrets are .1793 and .1917. Its mean *positive* target damage $\max(0, B_i^{\mathrm{par}} - B_i(\widehat{j}_i))$ is .0167 FP32 and .0309 physical. Much of its regret is a missed improvement. The seven auxiliary training facts are a separate outcome measured below, not part of this focal regret estimand.

For any predicted candidate scores $\widehat{B}_j$, write $e_j = B_j - \widehat{B}_j$ and $\widehat{j} = \arg\max_j \widehat{B}_j$. If j^* maximizes true score, $R = (\widehat{B}_{j^*} - \widehat{B}_{\widehat{j}}) + (e_{j^*} - e_{\widehat{j}}) \le \max_j e_j - \min_j e_j$. This inequality was checked on every saved fact and candidate family; the error range requires true outcomes and is not an online certificate. On high-qDF FP32 candidates, mean per-fact candidate MAE is .0385 for first order and .0248 for second; median error ranges are .1107 and .0326, compared with predicted top-two margins .0333 and .0339. The sufficient margin-greater-than-error-range condition holds for 16/40 and 21/40 facts, while actual oracle choices occur on 34/40 and 39/40. The natural family's success is consistent with improved relative accuracy and some ranking robustness. Because the controlled and SVD studies differ in family, scale, and estimand, these observations do not identify which change caused their contrasting outcomes.

The SVD candidates do not generally satisfy the full matched profile of Section 2. The experiment measures local *ranking* among real weight subedits, not the natural frequency of two candidates passing the original acceptance rule. Reusing the original fact IDs and opened full-edit results makes it a same-cohort extension, not a fresh fact confirmation.

Physical auxiliary-target follow-up. The old focal selectors and 320 physical children were fixed before the new scoring phase. For each of the same 40 focal edits, we directly scored the base model, complete parent, and eight saved weight subedits on the first two paraphrases of all seven auxiliary training facts; rewrite prompts are retained separately. Thus 280 auxiliary fact groups supply 8,400 prompt-level scores. Every physical focal parent/child score replayed exactly before auxiliary scoring. The independent saved-value validator rechecks all 40 fact records, source hashes, selector joins, and the selected-child summaries below. The seven-target parent gain over base is mean 4.4621 nat/token. In the high-qDF family, the focal second-order selector raises the focal score by mean $.1075$ relative to the full parent but lowers the auxiliary mean by $.3758$; 28/40 facts lose at least $.10$ on that auxiliary mean. Max-qDF changes those scores by $-.0206$ and $-.0286$, respectively, with 3/40 auxiliary losses of that size. First order gives focal $+.1131$, auxiliary $-.4157$, and 28/40 auxiliary losses. On all eight children, the second-order focal/auxiliary changes are $+.1469/-.6291$ and the max-qDF changes are $-.0206/-.0286$.

The high-qDF paired second-order-minus-max-qDF contrast is $+.1281$ [95% source-fact bootstrap $.0654, .1923$] on the focal target, $-.3471$ $[-.4740, -.2342]$ on auxiliary mean, and $-.2877$ $[-.3963, -.1908]$ on the equal-eight-target mean. In 31/40 facts the second-order child improves the focal target while reducing auxiliary mean. This target mismatch is not a numerical error in the focal predictor, and the equal-eight-target mean is not a deployed edit objective. The seven auxiliaries are training facts; locality and free generation remain unmeasured. This same-cohort follow-up therefore qualifies focal decision value without providing a natural-candidate replication of the full matched-profile counterexample.

We also joined each physical high-qDF fact's post-outcome focal-oracle child index to its seven auxiliary score records, using the same two paraphrases per auxiliary fact as the preceding analysis. Let j_T be the sealed second-order choice, j_F the true physical focal optimum, and $A(j)$ the auxiliary average. The exact per-fact identity is $A(j_T)-A(\text{parent}) = [A(j_T)-A(j_F)]+[A(j_F)-A(\text{parent})]$. The two child indices agree in 29/40 facts. The mean first term is $-.00245$ nat/token [95% paired fact bootstrap $-.0595, .0437$], but its mean absolute value is $.0565$ [$.0167, .1112$] and its maximum absolute value is $.867$. The oracle's mean second term is $-.3733$; 27/40 facts lose at least $.10$ on auxiliary average even under this focal oracle. Thus the *mean* auxiliary loss largely persists after replacing the local focal choice by its true optimum, while individual facts can still differ substantially. This oracle uses the same focal candidate set and target; it does not optimize auxiliary preservation.

D Controlled shared-family stress test

D.1 Candidate family and construction

The protocol was developed on the original 40 Qwen2.5 facts and then applied to an independently frozen cohort. The candidate definition is shared; cohort-specific experimental status is specified below. For fact i and its two paraphrases, set $V_{ip,j} = H_{1,ip} - H^{\text{child}}_{ip,j}$ using the eight saved physical LoRA singular-child hook states. A shared coefficient vector θ_i defines $D_{ip}(\theta_i) = \sum_{j<k} \theta_{ij} V_{ip,j}$, $k \in \{2, 4, 8\}$; coefficients lie in $[-2, 2]$. This is a precisely replayable hook-state family derived from one edit. Its linear combination need not equal the BF16 hook of a single modified LoRA weight update. For each qDF target $.25, .75, .90$, 10,000 seeded candidate directions supply a qDF-feasible anchor (both prompts within $.005$); 16 seeded nullspace-constrained optimizations seek up to four controls. They must match anchor qDF within 10^{-4} and the two edited-anchor signed complete-target slopes within 10^{-5} nat/token, with normalized hook separation at least $.15$. Failed anchor or control searches remain in the 40-fact denominator. All candidate files were independently checked and sealed before new $t = 1$ scores; the maximum realized qDF pair error was 3.33×10^{-5}, and slope error was below 1.5×10^{-15}. This two-slope profile is *not* the earlier radius–calibration–two-anchor acceptance rule, and its constructed controls are not natural SVD children.

D.2 Independent confirmation and local selection

After the development 40-fact outcomes were known, a new 40-fact cohort was frozen *before* any of its edit fits or continuous-candidate scores. It was selected by SHA-256 order from previously unused CounterFact IDs 800–1199; 280 disjoint auxiliary training facts use IDs 15000–15279. The model revision, layer-17 rank-8 LoRA recipe, complete-target readout, two paraphrases, coefficient bound, qDF target, two-slope match, and at-most-four-control search are unchanged. The fact-use metadata scan found no overlap, although it cannot certify absence from unindexed external work. Five facts have no qDF-feasible 2D anchor; the other 35 all have an 8D control. The largest qDF anchor–control error is 3.23×10^{-10} and the largest two-slope error 8.54×10^{-7} nat/token; the saved HVP action check is below 4.30×10^{-6}. All 40 source fits, local probes, and failures were independently validated and sealed before any new $t = 1$ candidate score.

Table 15: Disjoint fresh 40-fact shared-hook confirmation. All dose rows rescore the *same* outcome-blind candidate direction pool from qDF .25. Support is 35/40 at every scale; an event is a fact with at least one of at most four controls having $|\Gamma| \geq .10$ nat/token. The conditional median uses the per-fact maximum among those controls.

qDF	Matched support/40	Events/40	Median supported max $\lvert\Gamma\rvert$
.25	35	30	.303
.75	35	19	.124
.90	35	6	.050

The prespecified qDF .25 full-cohort event test rejects rate $\leq .5$ with 30/40 events (exact one-sided $p = .0011107$; Wilson 95% event-rate interval $[.598, .858]$). This is an accepted-pair *existence* test within a four-control construction budget, not the failure probability of natural candidates. The nominal binomial p-value assumes independent facts; relation dependence could weaken it. The

paired mean change in per-fact maximum absolute gap from qDF .25 to .90 is $-.469$ [95% whole-fact bootstrap $-.706, -.285$] on the 35 supported facts. All supported directions and controls are fixed across dose; changes cannot be attributed to selecting different directions at the smaller scale.

For the same fresh cohort, the predeclared one-control-per-fact probe comparison gives midpoint-minus-random mean selected gap $-.091$ [95% fact-bootstrap $-.207, -.0001$] on 35 multi-control facts. Midpoint minus old-anchor slope is $+.040$ $[-.141, .258]$, and midpoint minus edited-anchor half-HVP is $+.030$ $[-.140, .246]$. The latter two intervals include zero, so this experiment does not identify a universally best next measurement. Its measured budget was 80 gradient calls, 320 HVP actions, 175 half-path suffix forwards, and 525 $t = 1$ suffix forwards; LoRA fitting summed to 478.5 seconds on the occupied H200, excluding model load. These independent facts reproduce the shared-family finite gaps and fixed-direction dose trend; the preferred additional measurement remains unsettled.

An additional fresh-cohort decision check sealed the first-order, first-plus-half-HVP, and max-qDF selectors before opening the scores of the eight saved physical LoRA-child hook states per fact. The child states derive from one fitted weight edit, while scoring uses the FP32 frozen suffix; this is a focal choice check, not a matched-profile comparison. All 40 facts were scored. Max-qDF selected the true focal best child in 5/40 facts (mean regret .1112; 11/40 regrets $\geq .10$), first order in 38/40 (mean regret .00249; 0/40 $\geq .10$), and second order in 39/40 (mean regret .00248; 0/40 $\geq .10$). This independent cohort supports a useful local rule within a restricted candidate set even as the wider continuous hook family admits matched finite gaps.

Figure 10 summarizes the additional-probe comparison.

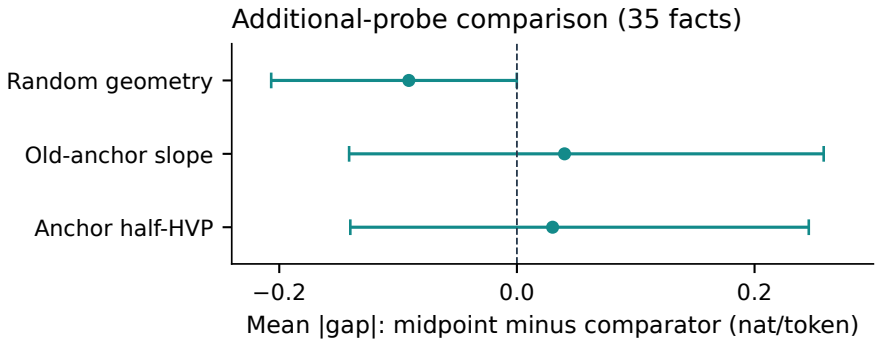

Figure 10: Additional-probe comparison on the 35 independent-cohort facts with multiple controls. A finite midpoint target score selects one control per fact; comparators are a seeded random coefficient projection, the old-anchor slope, and the edited-anchor half-HVP. Points are mean differences in selected absolute finite gap (midpoint minus comparator), in nat/token. Bars are 95% whole-fact bootstrap intervals. Negative values favor the midpoint choice. Costs differ across probes, and the intervals versus the two derivative probes include zero.

D.3 Aggregation and anchor ablation on one sealed pool

We replayed the two endpoint complete-target gradients for the *same* fresh 40 facts and saved eight-direction hook states; no edit was refit and no control was rebuilt. Each row records one candidate coefficient's signed target-score slope at one valid position. Summing rows recovers the previously sealed aggregate slopes with maximum absolute error 3.45×10^{-6}. All 40 measurement matrices were sealed before joining the already opened candidate outcomes. The 35 supported facts contain 140 original qDF .25 controls: 85 have a factwise finite gap $\geq .10$, spanning 30 facts; 55 have smaller gaps, spanning 26 facts. These candidates are repeated observations within facts, not 140 independent facts.

The four cells in Table 5 use either one aggregate slope per prompt or all valid-position slopes, at H_1 alone or at $H_0 + H_1$. The old weak aggregate screen accepts a control when each of its two slope differences is at most 10^{-5} nat/token. For each positionwise slope we use the original full-rule tolerance $10^{-6} + .05 \max(|m_a|, |m_c|)$; the two-anchor aggregate cell applies the same 10^{-5} absolute tolerance at both anchors. All controls retain the already sealed qDF match. Numerical rank uses singular values exceeding 10^{-8} of the largest. The ranks are 2, 4, 8, 8 respectively on *every* fact. The positionwise H_1 map alone therefore identifies an exact coefficient vector in this linear family; adding H_0 is not required for exact full-column rank.

For a candidate-distance scale, define G as the mean of the two full-update-normalized hook Gram matrices. We whiten the coefficient columns by $G^{-1/2}$, scale each measurement row by its *anchor-fixed* tolerance, and divide by the square root of its row count. On the 35 supported facts, the smallest singular value has median 19.32 for H_1 by position and 20.81 for both anchors by position. Their reciprocal candidate-distance sensitivities have medians $.0518$ and $.0481$ in units of the normalized update. This quantifies local conditioning under the stated anchor-fixed scale; the actual acceptance tolerance depends on both candidates and a finite-score bound would require score stability.

The old-pool result is deliberately more limited than a search result: either adding the two aggregated H_0 slopes (rank four) or retaining the H_1 positions (rank eight) rejects all 140 old controls, including all 85 large-gap and all 55 small-gap controls. Their median maximum error-to-tolerance ratios are respectively $37{,}791/11{,}851$ for large/small gaps under both-anchor aggregation, and $34.37/34.41$ under H_1 positionwise measurements. The latter is not selective between large- and small-gap old controls. In particular, rank four does not make the full set identifiable; it merely distinguishes this pool. Without a fresh search under each enriched rule, no zero-width or safety claim follows. Keeping the already computed gradient's position detail is an exact-identification option for this linear family, while finite replay remains necessary before asserting behavioral equivalence for a newly accepted, untested candidate.

D.4 Development-cohort analyses

Hold the anchor, widen the family. The two-slope measurement matrix has rank two on all 40 saved eight-direction families, leaving a six-dimensional linear nullspace at $k = 8$. Identification of a designated 2D subfamily requires its own two-column restriction to have rank two; this does not follow from the rank of the full eight-column matrix alone. Table 16 fixes each 2D or 4D anchor and searches again after padding it with zeros into the eight-direction family. The initial full-grid anchors and the later nested-anchor construction use new outcome-blind seals, but the latter was designed after seeing the first shared-family endpoint results. These rows therefore test an exact anchor confound, not an independent confirmatory hypothesis. Under regularity and interior support,

Table 16: Same-anchor dimension expansion on 40 facts per row. "Anchor" is qDF-feasible source-family support; "source/8D" is distinct slope-and-qDF-matched control support. Events use the complete-target fact-mean gap $\geq .10$ on the expanded support. No unsupported fact is assigned a finite gap.

Source k	qDF	Anchor/40	Source support	8D support	8D events
2	.25	34	0	34	30/34
2	.75	37	0	35	21/35
2	.90	39	0	32	15/32
4	.25	40	14	39	27/39
4	.75	40	21	39	24/39
4	.90	40	17	35	15/35

two independent qDF equalities plus a rank-r slope map leave local feasible dimension $k - r - 2$ by the implicit-function theorem. Here $r = 2$: $k = 2$ is identified when its restricted map has rank two; $k = 4$ can have isolated alternatives, and $k = 8$ leaves four local coordinates. This counts *possible* matches, not finite gaps; bounds, failed construction, or behaviorally flat null directions still require the support and scored-outcome checks in Table 16.

Hold directions, shrink the intervention. A second post-outcome addendum includes *all* 155 original 8D controls constructed at qDF .25 (40 facts), without selecting by their known outcome. Both anchor and control coefficients are scaled by $\sqrt{(1-q)/.75}$ and rescored at $q = .75, .90$; qDF and slope equality follow algebraically for each saved direction pair. Practical pair events fall $94/155 \to 50/155 \to 19/155$, while the same-pair median absolute gap falls $.156 \to .060 \to .027$ nat/token. The qDF $.25 \to .90$ mean paired change in absolute gap is $-.177$ [95% fact-bootstrap $-.245, -.116$]; 143/155 pairs shrink. The 16 original 4D controls on 14 facts show $7 \to 3 \to 1$ practical pair events across the same scales. These are finite-scale observations, not an upper bound on untested paths.

Measurement richness versus abstention. On the sealed first-order pool, E adds two old-anchor signed slopes, Q adds two edited-anchor half-HVPs, P adds two half-path target scores, and R adds two seeded random coefficient projections normalized to the E row norms. The qDF-and-slope-matched pool contains 517 control pairs; pre-endpoint costs were 80 gradient calls, 320 HVP actions, and 867 midpoint suffix forwards. With a fixed $.01$ tolerance, the 8D qDF $.25$ cell retains only 1/40 fact under E, 0 under Q, 8 under P, and 0 under R; zero observed events on empty support is abstention. A pre-score amendment therefore compares the probes at equal retention: each chooses exactly one control from each of 131 fact-cells with at least two controls. The midpoint choice has mean paired gap $-.040$ [95% whole-fact bootstrap $-.072, -.011$] versus random, $-.029$ $[-.047, -.013]$ versus E, and $-.012$ $[-.028, .004]$ versus Q. The last interval includes zero; 101/131 midpoint–quadratic choices tie in finite gap. A midpoint score is already a finite intermediate readout, and these probes have sharply different model-call costs; this is not an equal-compute speed claim or a certificate for the full accepted set.

D.5 High-qDF re-search on the evaluated confirmation cohort

The frozen fresh qDF .25 controls were scaled along fixed rays above. A separate post-outcome same-cohort diagnostic instead reran the unchanged 8D anchor and four-control search directly at qDF .75 and .90, without using the old directions as initial controls. Construction records were sealed before their new finite scores. An initial scorer accidentally dropped the first paraphrase and its results were discarded; the corrected two-prompt analysis independently reproduces the old fixed-ray counts. New searches support 37/40 facts with 24/40 events at qDF .75, and 36/40 with 13/40 events at qDF .90, versus fixed-ray 19/40 and 6/40. On facts supported by both at .90, events are 11/31 versus 6/31. The paired mean maximum-gap increase is .0209 with 95% whole-fact interval $[-.0098, .0542]$, so that mean contrast is unresolved. Re-search shows that a fixed-ray decline does not determine the maximum gap of a newly searched high-qDF accepted set; it neither removes the observed dose dependence nor establishes a global supremum.

D.6 Physical-weight transport boundary

After the hook-family results, a separate 34-fact check selected the *first sealed* $k = 2 \to 8$, qDF .25 control per supported fact, then applied one SVD-coefficient vector to the actual rank-8 LoRA down-projection weight update across both paraphrases. Each run reproduced the saved parent and first physical child hook and child score exactly. Direct BF16 physical practical gaps fell $24/34 \to 13/34 \to 6/34$ across the *nominal* qDF targets $.25 \to .75 \to .90$; the fact-paired mean gap change from .25 to .90 was $-.160$ $[-.221, -.104]$. But BF16 physical hooks were not exact linear combinations: median hook closure error relative to the full update was 3.07% at qDF .25, and median actual physical anchor–control qDF and edited-anchor slope mismatches were .00104 and .0141. *Zero of 34* physical pairs passes both the strict 10^{-4} qDF and 10^{-5} slope gates at any tested scale. Therefore this is qualitative transport of the dose trend, *not* a physical matched-rule counterexample or an additional 34 successful witnesses. Exact physical matching remains separate; the independent hook-family confirmation above does not retroactively satisfy the BF16 physical match gate. On the 40 fresh facts, a necessary-condition audit of all 40×28 within-fact pairs among eight fixed physical-child hook states finds 0/1,120 pairs matching both paraphrases' scalar qDF within 10^{-6}; none of the discrete children attains the original .25 target within that tolerance. This rules out a direct discrete-child full-profile witness in this pool, not continuous shared-weight controls.

E SIMPLE-CONTROL AND NUMERICAL-GATE AUDITS

We independently recomputed geometry on all 80 original confirmation prompts from the sealed direction tensors, cache updates, and calibration basis. The constructor seals qDF .50 directions; the scorer recovers full radius and applies its frozen qDF .25 target, multiplying those stored vectors by $\sqrt{.75/.50}$. All figures here use the *scored* vectors. The maximum active–control cosine magnitude is 1.25×10^{-9}, and the maximum deviation from the identity $\|D_a - D_c\|/\|\Delta\| = \sqrt{1.5}$ is 8.81×10^{-9}. Across prompts, $\|P_U\Delta\|/\|\Delta\|$ has median .1517 (10th–90th percentiles $.1336, .1700$); the ratio of scored active radius to projected-update norm has median 5.711 (10th–90th percentiles $5.095, 6.484$). Mean fact-level amplification has an exploratory Spearman correlation $\rho = .147$ with the original absolute finite gap ($p = .367$); this is neither a causal estimate nor a distance-matched control. The pair distance is fixed by construction, so it cannot explain variation *among* accepted pairs.

Replacing the drift-selected basis. We refit no model weights. On the same opened original 40 facts, three independently seeded full-space random orthonormal rank-32 bases replace only the active span U. Each retains the same calibration energy factors, positionwise radius and two-anchor slope constraints, search budget, qDF .25 scorer, and complete-target readout. All candidate directions for each seed were sealed before that seed's finite scores. These are post-outcome same-cohort controls, not new confirmations. Across seeds the mean random event fraction is .667 versus .775

Table 17: Reference-basis comparison on the original cohort. Dashes denote unsupported, unscored facts.

Active basis	Support/40	Events/40	Median supported $\|\Gamma\|$
High-drift calibration	40	31	.223
Endpoint covariance	0	–	–
Full-space random, seed 1	40	27	.179
Full-space random, seed 2	40	28	.215
Full-space random, seed 3	40	25	.140

for drift; the paired whole-fact bootstrap interval for random-minus-drift is $[-.242, .025]$, so an event-rate ordering is unresolved. The mean random-minus-drift gap is $-.417$ $[-.791, -.136]$, but its median fact-averaged difference is only $-.0547$ and a few large drift-basis gaps influence the mean. The random projection ratios have medians .1205–.1247, smaller than the high-drift .1517, so their qDF-normalization factors are *larger* (6.94–7.19 versus 5.71). Thus projection amplification alone does not account for a strictly larger drift-basis event count. The result shows that the same-rule ambiguity is not unique to a drift-informed active basis, without estimating how natural candidates behave. The endpoint-covariance basis comprises the top 32 right singular directions of the same calibration endpoint-gradient factors. Under the same fixed budget, all 80 prompt constructions ended with `energy_search_inconclusive`, giving 0/40 supported facts and *no* finite scores. The dashes in the table mean unscored, not zero-gap. This is a constructibility contrast, not proof that the accepted set is empty for this basis.

For the curvature-enhanced same-cohort pairs, we reapplied a grid of relative energy tolerances to the saved pre-score construction records, keeping the original absolute tolerance. No optimization or endpoint scoring was repeated. A fact remains supported only when both prompts pass both anchor-energy checks. The event threshold remains .10 nat/token; unsupported facts receive no gap. The roughly stable conditional event proportion is a property of this post-hoc retained subset; it does

Table 18: Reacceptance of the same saved curvature-matched pairs under stricter energy tolerances; no reconstruction is performed.

Energy tol.	Support $K/40$	Events V/K	Discovery $V/40$	Median $\|\Gamma\|$
10%	34/40	25/34	25/40	.238
5%	34/40	25/34	25/40	.238
2%	30/40	22/30	22/40	.221
1%	26/40	19/26	19/40	.199

not estimate the yield of a new tighter-tolerance search or prove that energy mismatch is irrelevant.

We separately reran the same 48-step curvature constructor on the opened 40-fact cohort with relative energy tolerance set to 1%; the absolute energy and .01-nat curvature tolerances, model, readout, and qDF .25 were unchanged. All 80 prompt constructions were sealed before the new control scores. The new search supports 35/40 facts and finds 26/35 practical gaps (26/40 discovery). An independent audit recomputed both anchor-energy inequalities on every passing prompt: the largest error-to-allowed-tolerance ratio is .882; the largest active endpoint replay discrepancy is 2.70×10^{-5} nat/token. On the 33 facts supported by both 10% and 1% searches, each has 24 practical gaps and the mean absolute-gap change is $-.000032$ $[-.000350, .000254]$ under a whole-fact bootstrap. Re-optimization improves support relative to post-hoc filtering, while preserving finite ambiguity in this opened cohort; it is not a new-fact 1% confirmation.

As a calibration, define for each accepted enhanced pair an *exact* quadratic surrogate using the real score's gradient and Hessian at H_1, then apply the same two-prompt gap computation to that surrogate. Among the 34 accepted facts, zero surrogates cross .10 nat/token; the maximum absolute quadratic gap is .00532 and the median .00100. This checks the scorer on a zero-remainder environment for already accepted pairs. It is not a fresh synthetic-model constructor run.

E.1 Independent new-fact full-profile 1% test

To test the stricter rule outside the opened original cohort, we froze 40 CounterFact IDs by a specified hash ordering of the metadata-unused 3200–3599 range *before* fitting their edits. The seven auxiliary training facts per edit occupy a disjoint, previously unused 16000–16279 range. This cohort is disjoint from both the original 452–799 and the separate shared-hook 800–1199 cohorts. The fact-use scan covers indexed project metadata, not unknowable external uses. We retained the Qwen2.5-3B model revision, layer-17 rank-8 focal-plus-seven LoRA fit, two paraphrases, complete-target readout, qDF .25 deletion branch, ROME-source calibration geometry, original 64-start/750-step base constructor, 48-step curvature refinement, and .01-nat curvature tolerance. The changed parameter is the predeclared 1% rather than 10% relative calibration-energy tolerance; its absolute tolerance is fixed. The fixed full-cohort statistic is a fact with $|\Gamma| \geq .10$; unsupported facts remain unscored. The one-sided majority-rate test uses $N = 40$ and critical count 26.

All 40 exact-replay edit caches and all 40 original-rule base directions were sealed before enhancement. Enhanced directions on all 80 prompts were then sealed with 35 facts supported on both prompts, before any new candidate finite endpoint was scored. The independent validator checks cache, base, enhanced, and score receipts by hash; recomputes both energy inequalities on every passing prompt; and confirms 210 finite intervention calls only on the 35 supported facts. The largest actual energy-error-to-tolerance ratio is .6283, and the largest edited-anchor curvature error is .00913 nat/token.

Table 19: Construction support and finite-gap events across separately defined profile/cohort combinations.

Profile and cohort	Support K/N	Events V/K	Discovery V/N
Original rule, first cohort	40/40	31/40	31/40
Curvature, same cohort, 10% energy	34/40	25/34	25/40
Curvature, same cohort, 1% re-search	35/40	26/35	26/40
Curvature, new cohort, 1% energy	35/40	23/35	23/40

The new cohort's supported median $|\Gamma|$ is .138 [95% whole-fact bootstrap $.085, .309$]. Its 23/40 full-cohort discovery rate is .575 (Wilson 95% interval $[.422, .715]$); the prespecified exact one-sided test against rate $\leq .5$ gives $p = .2148$; 23 events fall below the critical count of 26. Thus many accepted-pair counterexamples recur under 1% matching, but the original majority-rate claim is *not* independently confirmed. The five unsupported facts have no assigned finite gap; the result does not identify which change of cohort versus tolerance accounts for the frequency difference.

F Two conditional sufficiency statements

For a fixed candidate family $D = V\theta$ with a known vectorized basis V, stack the signed anchor-gradient measurements into $m(D) = A\theta$. If A has full column rank, equal measurements imply

equal candidates and therefore equal finite behavior, irrespective of downstream nonlinearity. More generally, using Euclidean norms,

$$\|D_a - D_c\| \le \frac{\|V\|_{\mathrm{op}}}{\sigma_{\min}(A)} \|m(D_a) - m(D_c)\|. \tag{15}$$

This is the minimum-singular-value inequality, a candidate-distance bound; a finite-score bound additionally needs score stability on the intervened states. Rank deficiency alone does not prove different behavior, because other profile constraints or a behaviorally flat nullspace may suffice. For the fresh eight-direction *linear hook span*, the original positionwise two-anchor slope matrix has full rank on all 40 facts as measured above. This is a conditional exact-identification result for that span, not a full-profile certificate for nonlinear physical SVD weight edits or for a nonzero acceptance tolerance.

Conversely, for the fact-average contrast $h(t)$ of Eq. 9, let $a = \frac{1}{2}\sum_p |h'_p(0)|$ and $b = \frac{1}{2}\sum_p |h''_p(0)|$. If the actual path satisfies $|h''(s) - h''(0)| \le Ls$ for $0 \le s \le t$, Taylor's integral formula implies

$$|h(t)| \le at + \frac{b}{2}t^2 + \frac{L}{6}t^3, \quad L \ge \frac{6\big(|h(t)| - at - \frac{b}{2}t^2\big)_+}{t^3}. \tag{16}$$

The second inequality is only a *necessary lower bound* on any valid path-variation constant after the endpoint is known. Applied at $t = 1$ to the 34 saved enhanced pairs, its median is 1.4008 nat/token; among the 25 practical-gap facts, the minimum is $.6161$ and median 2.6950. Neither these saved-value lower bounds nor four quadrature nodes certify an upper bound on L for unseen edits. The exact-identification premise *is* observed for the restricted 8D linear hook family when the already computed H_1 gradients are retained by position (Appendix D.3); the required finite-tolerance score-stability bound and the path-variation upper bound are not verified. The broad accepted-pair witnesses establish that the paper's specified profile and tolerances do not guarantee finite-score proximity at the tested scale.

G Other experiments and exact-scope boundaries

The following results concern either a different finite-fiber estimand or a qualification gate. They test how far the audit can be extended and are not pooled with the Qwen2.5 same-readout confirmation.

Table 20: Scope of the earlier experiments. These studies use different estimands from the primary same-target matched-pair audit. Detailed numerical results and negative outcomes are retained below.

Study	Quantity assessed	Scope of the result
Gemma/Qwen2.5	Recovery and raw target-score differences over dose	Exploratory cohorts; curvature measured, not matched
Qwen3-4B	Static profile width and path-estimation error	Registered four-cell conjunction unconfirmed in both cohorts
Qwen3.5-4B	Path-error ranking and width-error deferral	Separate endpoints on one cohort; no runtime-gain claim
Official SAE/ICA	Shared update-fidelity support	264/264 comparison cells unsupported per model; no behavioral ranking

G.1 Earlier recovery endpoints

Opened Gemma2-2B and Qwen2.5-3B cohorts showed differences in normalized target-answer recovery across editor cells and intervention scales. These analyses use a full-edit gain denominator. Their unchanged scored rows also retain the active and control *complete-target mean log probabilities* for each paraphrase. We subtract those raw scores promptwise and then average the two signed gaps within each fact, exactly as in Eq. 3; no product of averaged recovery ratios and gains is used. Table 21 therefore gives a common nat/token unit for the original matcher across six radii and two model families. These cohorts were opened during method development and do not match the added curvature term; they are descriptive dose context, not independent curvature-enhanced confirmation.

Table 21: Older, opened cohorts re-expressed from saved per-prompt complete-target scores. Each cell counts facts with the absolute signed two-paraphrase gap $\geq .10$ nat/token. R→L uses ROME calibration geometry with LoRA evaluation; L→R reverses that pairing. The original rule had no directional-curvature match.

qDF	Gemma R→L (20)	Gemma L→R (20)	Qwen R→L (60)	Qwen L→R (60)
.25	18/20	9/20	48/60	10/60
.50	19/20	9/20	41/60	8/60
.75	16/20	6/20	33/60	3/60
.90	9/20	2/20	23/60	2/60
.95	6/20	0/20	11/60	0/60
.975	5/20	0/20	7/60	0/60

The earlier Gemma scale diagnostic measured directional Hessians without matching them. Its original figure is retained among the historical source assets rather than pooled with the new Qwen2.5 sweep.

G.2 QWEN3-4B STATIC WIDTH AND PATH ACCOUNTING

For each fact/cell, nine rank-r candidate profiles are scored over six frozen qDF targets; d_r is their largest mean pairwise finite-response distance. The static width statistic is $W_{32} = \mathrm{median}(d_{32})/\,\mathrm{median}(d_0)$. P_3 is mean absolute three-node path residual normalized by $\mathrm{median}(d_0)$. The registered positive law requires a 90% bootstrap lower bound on $W_{32} > .5$ and an upper bound on $P_3 < .5$ in *all four* ROME/LoRA by insertion/deletion cells. Two disjoint 20-fact Qwen3-4B cohorts both fail this conjunction (Table 22). Accurate individual ROME path cells cannot replace the registered four-cell decision. Failure to establish the conjunction is not a statistical proof of the opposite universal law. This experiment asks about a nine-profile diameter and normalized path error, not the signed pair gap Γ of Eq. 3.

Table 22: Separate Qwen3-4B static-width/path stress test. Estimate [90% whole-fact interval]. The registered joint decision requires the width lower bound $> .5$ and GJ3 error upper bound $< .5$ in all four cells; neither cohort establishes the conjunction.

Cohort	Editor/branch	W_{32}	P_3
Primary	LoRA/insert	.516 [.373,.885]	1.401 [.072,3.402]
	LoRA/delete	.490 [.340,1.154]	.292 [.044,.696]
	ROME/insert	.522 [.351,.908]	.011 [.006,.021]
	ROME/delete	.727 [.438,1.569]	.010 [.005,.031]
Replication	LoRA/insert	.966 [.505,1.622]	.855 [.202,1.569]
	LoRA/delete	.713 [.423,1.996]	.333 [.084,1.068]
	ROME/insert	.516 [.316,.936]	.632 [.011,1.947]
	ROME/delete	.682 [.265,.996]	.174 [.006,.491]

The original audit checked matching across 3,840 eligible Qwen3-4B candidates. Figure 11 shows the registered gate clearances for this separate negative conjunction; it is not pooled with the Qwen2.5 pair result.

G.3 PATH-RISK RANKING AND SELECTIVE NUMERICAL ESTIMATION

In a separate prospective Qwen3-4B tail test (40 facts), sealed GJ4–GJ3 prediction disagreement separated high/lower numerical path-error strata in both primary LoRA branches: sample superiority 1.000, Holm $p \leq 7.67 \times 10^{-7}$. This result predicts numerical approximation risk, not semantic correctness or a universal path law. An absolute Qwen3 threshold failed pre-outcome stratification on Qwen3.5-4B (3/37 high/lower facts), so those outcomes remained unopened. A subsequent local-scale Qwen3.5-4B cohort of 40 *new* facts (412–451) used a separately frozen continuous-ranking rule: LoRA insertion/deletion Spearman correlations between the selected path-disagreement score and *normalized GJ3 path-estimation MAE* are .979 [90% CI .948, .989] and .968 [.933, .980]. On

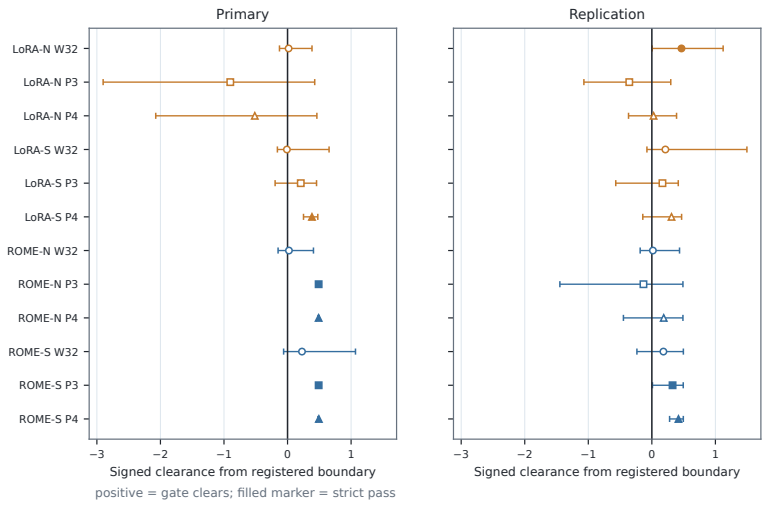

Figure 11: Clearance from the .5 boundary. Positive values clear a width/path gate; filled markers pass its strict confidence bound. Tranches remain unpooled.

these *same* 40 facts, a pre-outcome supplemental 75%-coverage deferral policy gives retained/all width-MAE ratios .097 [.019, .719] and .332 [.147, .663]. It is not a second cohort. A post-hoc cheap predicted-width policy gives .058 and .522 on the same branches: neither score dominates both. None of these results establishes a runtime gain over direct endpoint scoring or repairs the failed Qwen3 joint law.

Table 23: Qwen3.5-4B local-scale risk test on the same 40 facts per branch. Spearman refers to normalized GJ3 path MAE, whereas the two retained/all columns refer to width-estimation MAE. Ranking is prospective; width deferral was frozen before outcomes as a supplement; the width-magnitude comparison is post-hoc.

LoRA branch	Path-error Spearman [90% CI]	Disagreement retained/all	Cheap width rule
Insertion	.979 [.948,.989]	.097 [.019,.719]	.058
Deletion	.968 [.933,.980]	.332 [.147,.663]	.522

The path-risk and width-reporting results use the same 40-fact cohort but different endpoints. Their original visualization is retained in the source assets; Table 23 gives the numerical summary.

G.4 Official SAE/ICA common-support qualification

A causal ranking between sparse dictionaries requires both interventions to attain the registered update-direction qDF. Checkpoint- and hook-matched official Qwen-Scope TopK100 SAEs (Deng et al., 2026) for Qwen3-8B (post-layer 17) and Qwen3.5-2B (post-layer 11) were compared with coordinate-sparse ICA/PCA controls. Dense ICA/PCA rotations yield the same projection up to numerical precision and cannot define a distinct causal lens. For a scaled component C, the attainment curve $\mathrm{qDF}(s) = As - Bs^2$ has a continuous ceiling. Neither model attains exact common support at the lowest registered target .25: *each model's* 264/264 operator–coordinate–target held-out comparison cells return `ICA-SAE-NO-COMMON-SUPPORT`, with zero finite behavioral calls. The newer Qwen3.5-2B official SAE reconstructs centered activations at $.7996/.7901$ on calibration/development, yet its sparse controls fail the registered update-direction target. Lower *diagnostic* targets allow some strength selections, but cannot retroactively authorize the frozen test. This is a replicated support mismatch across two checkpoints, not a behavioral tie or ranking between ICA and SAE.

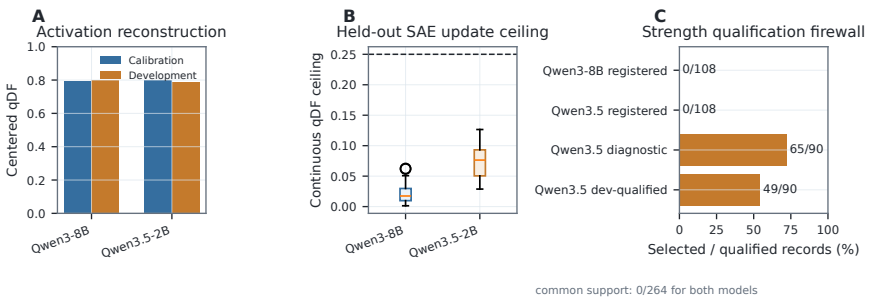

Figure 12: A support gate can stop a causal ranking before behavioral scoring. Official SAE and coordinate controls reconstruct activations, but no pair attains the registered common update-fidelity target in 264 held-out comparison cells *per model* for Qwen3-8B and Qwen3.5-2B. These cells are not independent facts. Lower targets are diagnostic and do not replace the frozen comparison.

H Related-method map and discriminating controls

This map records what prior methods establish and which implication our audit tests. The cited literatures are not interchangeable baselines.

Sparse representations and their validation. Sparse coding and overcomplete dictionaries establish representation objectives (Olshausen & Field, 1996; Lewicki & Sejnowski, 2000; Aharon et al., 2006; Mairal et al., 2009); toy superposition and early SAE proposals motivated their use for model internals (Elhage et al., 2022; Sharkey et al., 2022). Language-model SAEs progressed from interpretable features and monosemanticity demonstrations to scaling studies (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024). Gated and JumpReLU SAEs alter the sparsity–reconstruction tradeoff; Gemma Scope and Qwen-Scope supply open model-specific features (Rajamanoharan et al., 2024a;b; Lieberum et al., 2024; Deng et al., 2026). Task-grounded control and circuit studies test more than reconstruction (Makelov et al., 2025; Marks et al., 2024; Karvonen et al., 2025). Randomization checks show why an apparently interpretable feature may reflect input statistics or model architecture (Adebayo et al., 2018; Heap et al., 2026). None establishes that our registered SAE/ICA comparison has common qDF support; Section G.4 measures that condition directly.

Intervention semantics and competing explanations. Causal mediation and abstraction formalize intervention claims (Pearl, 2009; Vig et al., 2020; Geiger et al., 2021; 2025). IOI circuits and automated circuit discovery instantiate component-level analyses (Wang et al., 2022; Conmy et al., 2023). Patching choices, subspace illusions, attribution-patching false negatives, and second-order corrections identify distinct ways a local readout can mislead (Zhang & Nanda, 2024; Makelov et al., 2024; Kramár et al., 2024; Zhang & Wang, 2026). Erramilli (2026) organizes basis choice, designed measurements, finite calibration, and held-out validation around recovering an intervention-effect map. Our narrower estimand is the finite behavioral spread *among candidates that pass one specified local profile*, exposed by outcome-blind construction; we do not claim the three-factor measurement framing itself as novel. Steering and function-vector interventions form natural candidate families but do not meet our frozen matched-profile rule by definition (Turner et al., 2023; Panickssery et al., 2023; Zou et al., 2023; Li et al., 2023; Todd et al., 2023). Steering-vector non-identifiability runs from similar behavior to uncertain directions, the reverse implication to our local-profile test (Venkatesh & Kurapath, 2026). Distributional statistical certification addresses a different quantifier from a pairwise witness (Asiaee, 2026). The attention debate makes the realizability issue concrete: freely varied per-instance alternatives need not be a single trained intervention policy (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). This is a boundary on our constructed witnesses, not a disproof of their finite score differences.

Editing methods and outcome scope. Editable training, hypernetwork edits, external memory, and direct weight rewrites create different candidate populations (Sinitsin et al., 2020; De Cao et al., 2021; Mitchell et al., 2022a;b; Meng et al., 2022; 2023). Localization need not predict which edit

layer works, while ripple and benchmark studies broaden the tested outcomes (Hase et al., 2023; Cohen et al., 2023; Zhang et al., 2024). Our physical LoRA update follows a low-rank adaptation family (Hu et al., 2021); its focal-versus-auxiliary result does not measure those papers' locality or related-fact criteria. The primary Qwen2.5 model and secondary Qwen3/Qwen3.5 checkpoints have separate provenance (Qwen Team, 2025; Yang et al., 2025; Qwen Team, 2026). Their cohorts are not pooled as replications of one matched-pair estimand.

Coordinates, numerical accounting, and uncertainty. Hessian-vector products and curvature approximations predate this audit (Pearlmutter, 1994; Martens & Grosse, 2015); integrated gradients, interaction attributions, and Gauss quadrature motivate path accounting without certifying a finite endpoint from local data (Sundararajan et al., 2017; Janizek et al., 2021; Golub & Welsch, 1969). ICA methods and principal-component controls identify distinct coordinate objectives (Comon, 1994; Bell & Sejnowski, 1995; Hyvärinen, 1999; Hyvärinen & Oja, 2000; Ablin et al., 2018; Pearson, 1901; Hotelling, 1933). Unsupervised disentanglement limits are a broader identifiability analogy, not a proof about our fixed residual hook (Locatello et al., 2019). Whole-fact bootstrap follows resampling at the independent fact unit (Efron & Tibshirani, 1993).

I Evidence records and experimental chronology

The manuscript source archive contains LaTeX and all figure assets, including historical plots omitted from the streamlined appendix. A separate verification companion contains compact records and standard-library Python checks for the principal tables. The checks validate stored summaries and hashes; they do not replay model fitting or hidden-state generation. Model weights and activation caches are excluded. The original freeze record contains machine-specific metadata and is retained outside the anonymous companion; its digest and the public protocol are included. The distributed checks cannot rehash that withheld record.

Table 24: Contents of the compact verification companion. Filenames are relative to their named companion subdirectory.

Result	Record	Verification entry
Independent strict profile	`fresh_full1_summary.json`	`followup/verify_companion.py`
Main evidence table	Original, enhanced, strict, and shared summaries	`aggregation_anchor_2x2/verify_submission_tables.py`
Aggregation/anchor table	`aggregation_anchor_2x2/SUMMARY.json`	`aggregation_anchor_2x2/verify_summary.py`
Basis, tolerance, and dose controls	Summaries in `followup/`	`followup/verify_companion.py`
Independent child ranking	`fresh_child_summary.json`	Stored validated selector summary; no model replay

Experimental chronology. The original directions were fixed before their confirmation scores. Curvature, constructor, dose, and auxiliary-target follow-ups reused facts whose earlier outcomes were known, while fixing new controls or evaluation choices before their respective new scores. The strict 1% and shared-family confirmation cohorts were each fixed before edit fitting. Their candidate directions and local measurements were sealed before finite evaluation. The aggregation/anchor ablation reused the evaluated shared cohort, fixed the replayed matrices before joining candidate scores, and generated no new candidates.

LoRA child selectors were fixed before child scoring; FP32 suffix and BF16 physical evaluations were retained separately. Path/logit decompositions and the selected fact-592 replay followed outcome access. These distinctions are maintained in the cohort tables and verification records; same-cohort follow-ups are not pooled as new confirmations.

Inference boundary. An exactly equal measurement profile would force any point predictor using only that profile to give the same estimate for two candidates; if their true scores differ by $|\Gamma|$, one error is at least $|\Gamma|/2$. Our empirical matches have tolerances, so the observed gap is a witness to behavioral spread among *accepted constructed pairs*, not a certified bound on arbitrary predictors or the full feasible-set diameter. The anchor-logits analysis tests its specified proxy and a later curvature-corrected saved-value variant; neither exhausts possible uses of local output information. GJ4 is

explanatory on opened outcomes and has a serious tail. Native support is selected; Qwen3 and SAE/ICA results concern separate estimands. Transport to physical weight interventions and unseen prompts remains outside the established scope.