INTERVENTIONAL COMMUTATIVITY: EFFECT PRESERVATION AND REFERENCE SELECTION UNDER WEIGHT-ONLY QUANTISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can an intervention discovered before quantisation be reused after deployment? This question has two targets: preserving its original effect and matching an intervention re-discovered on the deployed model. We formalise the latter as *interventional commutativity* and audit both on behaviourally matched inputs. A round-to-nearest construction proves that exact functional equivalence can coexist with complete source-effect loss; an effect-mass bound separates mean sign agreement from magnitude retention. Empirically, a corrected 504-cell FP16/INT8 audit finds median native-relative retention 0.999 across 245 informative INT8 cells, with eight severe gaps in four configurations. However, the 243 cells whose source-effect magnitudes exceed the same descriptive guard have source-fidelity point estimates between 0.792 and 1.186. Crossed execution explains the distinction: on 117 matched template-held-out Qwen prompts, 4.481 logits of a 4.497-logit native-reference gap already exist before quantisation; only 0.016 comes from different responses to deployment. Thus a large native gap can diagnose reference selection while the transported effect's point estimate stays near its source value. Control-calibrated floors, paired gaps and cluster uncertainty make the comparison measurable; sequence checks show that its verdict remains endpoint-specific. The resulting audit quantifies preservation against the chosen reference and clarifies the target of repair.

1 INTRODUCTION

Mechanistic interpretability validates explanations through interventions: an ablation tests a circuit, activation patching tests a causal variable, and a steering direction is useful because it changes behaviour (Olah et al., 2020; Wang et al., 2023; Geiger et al., 2024; Rimsky et al., 2024). These artefacts are commonly discovered on a development checkpoint and then reused on a compressed deployment (Dettmers et al., 2022; Frantar et al., 2023; Lin et al., 2024; Sun et al., 2024). Preserved benchmark accuracy leaves the effect of that reuse untested.

Consider an intervention whose mean effect is 1.016 logits before quantisation and 0.908 afterwards. It retains 89% of its original effect. Yet re-discovery on the quantised model produces an intervention with effect 5.391, making the same transported intervention only 17% as effective as the native reference. Both statements describe one Qwen configuration in our audit. They answer different practical questions: can a previously validated intervention retain its effect, and does reuse match what the discovery procedure now finds?

We call the second property *interventional commutativity*, and distinguish it from *source-effect fidelity*. Both are measured on inputs where the unintervened models agree. Native parity matters when deployment reuses an explanation or control in place of re-running discovery; source fidelity matters when its original effect is the deployment requirement. Neither target implies the other. In particular, improving native parity can move an intervention away from the effect it was originally meant to preserve.

Our audit makes this distinction operational. Execute each frozen artefact on both the development and deployment models. The four effects separate a gap already present between selected artefacts

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

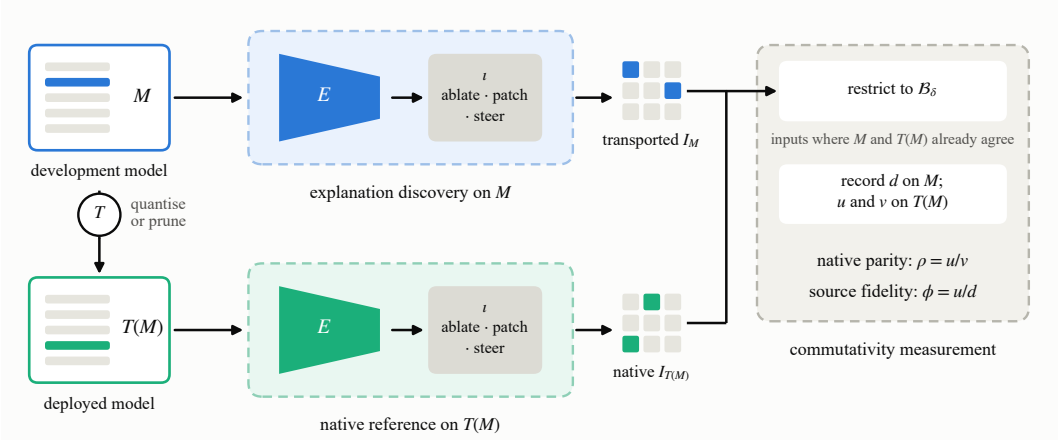


Figure 1: **The intervention audit from discovery to diagnosis.** Discover source and native artefacts, freeze their sites and payloads, and evaluate on the same behaviour-matched inputs. Source fidelity compares u with d; native parity compares u with v. For the additional crossed diagnostic in Section 6.1, execute the native artefact back on development to obtain b. Controls and paired uncertainty resolve the comparisons.

from their different responses to quantisation. This requires one additional execution of the native artefact on development, beyond the three effects used for the two retention ratios. Control-based scale guards and paired uncertainty establish whether a measured gap is resolved. The resulting diagnosis determines whether to investigate execution changes, discovery choices, or the repair objective.

Contributions. **(1) A reference-explicit formulation.** We define native-relative commutativity alongside source fidelity and use crossed execution to distinguish their sources of disagreement. An RTN construction proves that even exact functional equivalence can destroy a transported effect; a sharp effect-mass bound certifies mean sign under partial alignment. **(2) A calibrated audit.** Behavioural matching, FP16 and random-site controls, denominator guards and prompt-cluster intervals separate unresolved measurements from substantial effect gaps. **(3) An empirical diagnosis.** A corrected 504-cell sentinel finds native-relative retention near one in the bulk and eight severe cells in four configurations. Source-fidelity re-analysis and a targeted held-out Qwen study show why this tail cannot be read as a count of destroyed source effects. Discovery, repair and sequence checks establish the consequences of choosing a reference and endpoint.

2 Related Work

Causal explanations and their reliability. Circuit analysis, automated discovery, activation patching, alignment and steering validate explanations through interventions (Wang et al., 2023; Conmy et al., 2023; Syed et al., 2024; Zhang & Nanda, 2024; Geiger et al., 2024; Rimsky et al., 2024); reliability studies vary examples, metrics, or intervention design on a fixed checkpoint (Heimersheim & Nanda, 2024; Tan et al., 2024). We vary deployment precision, condition on behavioural agreement, and distinguish reference selection from preservation of a fixed intervention.

Across-model transport and similarity metrics. CKA and related metrics compare representations (Kornblith et al., 2019), but which measure tracks functional correspondence is contested (Bo et al., 2024). Learned maps move interventions between separately trained models (Rodriguez et al., 2025; Oozeer et al., 2025), and Model Alignment Search fits per-model alignment functions so patched activity can be compared causally (Grant, 2025). Compression supplies shared site names and shapes, but still requires an effect comparison. Our Procrustes/ridge maps fit activations; the targeted gate pilot instead fits intervention effects. Neither implements MAS's designated causal subspaces, and failure of an activation fit establishes no limit on causal repair.

Compression beyond task accuracy. GPTQ, AWQ, NF4, 8-bit inference, and one-shot pruning target deployment efficiency (Frantar et al., 2023; Lin et al., 2024; Dettmers et al., 2023; 2022; Sun et al., 2024; Frantar & Alistarh, 2023). We audit weight-only rounding and pruning plus official GPTQ/AWQ checkpoints. Recent studies analyse internal effects and factual recall under quantisation (Singh & Sajjad, 2025; Wang et al., 2025), while large-scale evaluations chart accuracy and alignment-flavoured effects (Lee et al., 2024). UniComp jointly evaluates performance, reliability, and efficiency across compression families (von Rad et al., 2026). Commutativity adds a distinct paired question: does the *same deployed model* respond differently to a transported artefact and native re-discovery? It complements these broader evaluations. Zhou et al. (2026) distinguish signal degradation from computation collapse under PTQ. Our reference decomposition asks a prior diagnostic question: did the original intervention change, or did the selected reference change? A native-relative gap alone does not establish either network failure mode, especially when the source effect remains intact.

Concurrent work. Rabbi et al. (2026) report that quantisation can change saliency-style explanations while predictions stay fixed—the correlational analogue of our causal finding; Belkhiter et al. (2024) chart how quantisation moves harm-level compliance; Braun (2026) and Raju (2026) study steering unreliability and representational drift within a checkpoint; and Sun & Toneva (2026) estimate when two networks admit equivalent interpretations. To our knowledge, none measures transported-versus-native intervention efficacy on a compressed deployment.

3 Interventional Commutativity

Let a development model M expose activation sites $\mathcal{L}$, and let an architecture-preserving deployment transformation T produce $T(M)$. An explainer E and compiler ι produce an intervention $I = \iota(E(M))$, such as a component ablation, activation patch, or steering vector. We first isolate inputs with matched unintervened behaviour:

$$\mathcal{B}_\delta = \{x \in \mathcal{D} : \arg\max M(x) = \arg\max T(M)(x),\ D_{\mathrm{KL}}(M(x) \,\|\, T(M)(x)) \le \delta\}\,. \quad (1)$$

This conditioning defines the estimand. Insufficient matched support is reported separately from an intervention gap.

Definition 1 (Efficacy and commutativity)**.** For task score m and intervention I applied to M', define

$$\mathrm{Eff}(I, M') = \mathbb{E}_{x \sim \mathcal{B}_\delta}[m(I[M'], x) - m(M', x)]. \quad (2)$$

Let $I_M = \iota(E(M))$ be transported from the development model and $I_{T(M)} = \iota(E(T(M)))$ be re-discovered on the deployment. Their native-relative efficacy retention and commutativity gap are

$$\rho = \frac{\mathrm{Eff}(I_M, T(M))}{\mathrm{Eff}(I_{T(M)}, T(M))}, \qquad \mathrm{CG} = 1 - \rho. \quad (3)$$

$\rho = 1$ denotes equal mean effects, $\rho = 0$ a zero transported mean effect, and $\rho < 0$ opposite transported/native signs. The ratio is unclipped; equality of means alone does not imply per-example correspondence.

Four effects distinguish preservation from native parity. Execute both frozen artefacts on both models, measuring each effect relative to that model's baseline on the same $\mathcal{B}_\delta$:

$\mathrm{Eff}(\text{artefact}, \text{model})$	M	$T(M)$
I_M	d	u
$I_{T(M)}$	b	v

$$\phi = \frac{u}{d}, \quad \rho = \frac{u}{v}. \quad (4)$$

Here ϕ measures *source-effect fidelity*, with its own denominator check. Native re-discovery supplies an explainer-defined reference, not an optimum: $\phi \approx 1$ can coexist with $\rho \ll 1$ when the native artefact is stronger. To diagnose that discrepancy, use the exact decomposition

$$v - u = \underbrace{b - d}_{\text{reference gap on development}} + \underbrace{(v - b) - (u - d)}_{\text{different responses to deployment}}. \quad (5)$$

Crossed execution separates a difference between the selected artefacts from a change in their execution effects. It does not identify the neural mechanism behind selection. A negative ρ likewise need not mean that the source effect reversed across models. Throughout, "severe" refers to native parity.

Transport ships the artefact unchanged. I_M retains its source sites and payload; $I_{T(M)}$ re-estimates them with the same explainer hyperparameters (Figure 1). Discovery data, seeds and tie-breaking are part of the explainer specification; payload fitting is part of the compiler. Each evidence layer states whether discovery and evaluation data are shared or disjoint. Re-fitting a transport map is adaptation. **L1 behavioural agreement** defines $\mathcal{B}_\delta$; **L2 representation or explanation alignment** includes CKA and site overlap; **L3 intervention-effect parity** compares causal effects. L1 and L2 do not certify L3.

A ratio requires a measurement-scale guard. When the native effect is small relative to run-to-run variation, the ratio is poorly resolved. For the FP16 deployment control T_0, group runs by model, task, and explainer and set

$$\tau_{\text{eff}} = 3 \ \text{median}_g \ \text{sd}_s \big[\text{Eff}(I_{T(M)}, T_0(M))\big] \, . \tag{6}$$

These runs combine sampling, discovery and execution variation; the floor sets a descriptive measurement scale rather than a hypothesis-test threshold. The sentinel floor is 0.229 logits. We score ρ only when $|v| \geq \tau_{\text{eff}}$, reporting $R_\alpha = \Pr(\rho < \alpha \mid |v| \geq \tau_{\text{eff}})$ at $\alpha = 0.8$ and 0.5, paired gaps $v - u$, and prompt-cluster intervals. Random-site controls test sensitivity to site choice; FP16-control transport checks discrepancies under a small deployment change (Section 5). Appendix B documents the SD convention, repeat counts, layer-specific floors and held-out sensitivity.

3.1 Exact behaviour does not imply intervention preservation

Proposition 1 (Exact behaviour, total transported-intervention failure)**.** *Fix a quantisation step $\Delta > 0$, an integer $k \geq 2$, and $0 < \epsilon < (k-1)\Delta/[2(k+1)]$. There is a $(k{+}1)$-component linear model f with one leading and k redundant components (weights $\Delta/2 \mp \epsilon$) such that rounding every weight to the nearest multiple of Δ leaves $f(x) = T(f)(x) = \Delta x$ for every input, yet a zero-ablation explainer selects the leading component before quantisation and a backup after it. On positive inputs with finite mean, the transported ablation has zero effect on $T(f)$ while native re-discovery has non-zero effect: $\rho = 0$ despite exact behavioural equivalence.*

Here $\phi = \rho = 0$; source selection is strict and every native backup has the same effect, making native tie-breaking immaterial. Small nonzero weight neighbourhoods preserve zero transport with arbitrarily small output error (Appendix A); this is a possibility result, not an explanation of the empirical tail. The next bound separates sign preservation from magnitude retention.

Proposition 2 (An effect-mass bound allowing sign exceptions)**.** *Let $u(x)$ and $v(x)$ be the integrable transported and native per-example effects in Equation 2. Write $s = \text{sign}\,\mathbb{E}[v]$ and $\mu = \mathbb{E}[sv] > 0$. On a subset $G \subseteq \mathcal{B}_\delta$, assume $uv \geq 0$ and $\lambda|v| \leq |u| \leq \kappa|v|$, where $0 \leq \lambda \leq \kappa < \infty$. Define aligned majority and minority native masses, and the transported exception budget, by*

$$\begin{aligned} A &= \mathbb{E}[|v|\mathbf{1}\{x \in G,\, sv > 0\}], \quad B = \mathbb{E}[|v|\mathbf{1}\{x \in G,\, sv < 0\}], \\ \varepsilon &= \mathbb{E}[|u|\mathbf{1}\{x \notin G\}]. \end{aligned} \tag{7}$$

Then

$$\frac{\lambda A - \kappa B - \varepsilon}{\mu} \ \leq \ \rho \ \leq \ \frac{\kappa A - \lambda B + \varepsilon}{\mu}. \tag{8}$$

The bounds are sharp under this group-mass information. In particular, $\lambda A \geq \kappa B + \varepsilon$ precludes a mean sign reversal.

The majority's guaranteed effect must cover both an amplified native minority and the absolute transported mass outside the aligned subset. No site additivity or linearisation is assumed. With $G = \mathcal{B}_\delta$ and compatible zeros, $\varepsilon = 0$, $\mu = A - B$, and the bound reduces to $[(\lambda - \kappa\pi)/(1-\pi), (\kappa - \lambda\pi)/(1-\pi)]$ for $\pi = B/A$. The exception budget avoids abstaining solely because a few small effects change sign. For example, $v = (1, 1)$ and $u = (0.8, -0.01)$ violate global alignment, yet Equation 8 gives $\rho \in [.395, .405]$: the mean cannot reverse, but retention is still severe. The empirical bound is a post-hoc diagnostic on fixed paired observations, not a predictor for unseen prompts or a neural explanation. Appendix A gives the proof, special case, and sign–rank examples (Figure 4).

For the sequence-level arm, m becomes a refusal-phrase indicator on a deterministic 48-token continuation. We report paired rate gaps and disagreement counts directly; a logit-unit floor cannot be transferred to this discrete score.

Table 1: **Evidence for preservation, reference selection and endpoint dependence.** The sentinel supplies the primary INT8 estimate; extensions are analysed separately.

Evidence	Observation	Scope
Primary sentinel	504 FP16/INT8 cells; 8/245 severe native gaps, 4/83 configurations	Seven base models; both effect references; six severe cells use eight prefix prompts
Reference diagnosis	205 matched new Qwen prompts; 6 GPT-2 splits	Crossed execution on CPU BF16 Qwen; disjoint discovery on CPU FP32 GPT-2
Deployment extensions	Instruction RTN8: 0/17 severe; official GPTQ/AWQ: 0/24	Two Qwen-Instruct models; official $\delta = .2$ results, with 12 valid cells at .05
48-token outcomes	4/12 official gaps survive multiplicity correction	Paired phrase-based rates; separate RTN8 follow-up records exact token IDs
Exploratory checks	858/1,764 breadth cells complete; 30 fitted-map cells; 12-gate pilot	Legacy transforms, activation alignment, and effect-supervised fitting; detailed provenance in appendices

4 Experimental Design

The primary audit measures both source fidelity and native parity; follow-ups diagnose their difference and test other endpoints (Table 1). A *cell* is one model/task/intervention/deployment/seed run; a *configuration* groups its seeds. Table 25 records the grids.

Primary audit and provenance. Seven base checkpoints (GPT-2 small/medium (Radford et al., 2019), Pythia-410M/1.4B (Biderman et al., 2023), Qwen2.5-0.5B/1.5B (Qwen Team, 2024), Llama-3.2-1B (Grattafiori et al., 2024)) cross four tasks (IOI, Greater-Than, factual recall, assistant-prefix contrast), three interventions (attribution-based mean ablation (Syed et al., 2024), localised activation patching (Geiger et al., 2024), difference-in-means steering (Rimsky et al., 2024)), FP16/RTN8, and seeds $\{0, 1, 2\}$: 504 cells. After correcting padding masks and mask-derived positions and adding regression tests, we reran this entire grid; severe RTN8 counts changed from 14 to eight. The exploratory breadth screen crosses the same model/task/intervention grid with seven transforms: FP16, RTN INT8, RTN INT4 (groups 128/64), NF4, Wanda 50%, and magnitude pruning 50% (Appendix C). A separate follow-up corrects partial-minibatch attribution weighting (Appendix L).

Intervention budgets. Ablation and patching select $k = 12$ sites by absolute score. Attribution uses single-site effects for at most 256 candidates and gradient estimates otherwise; patching imports counterfactual activations. Steering selects the strongest absolute effect at depth fractions $\{.4, .5, .6, .7, .8\}$, using an activation-norm-scaled class-mean direction with unit coefficient. Source and native arms share these budgets; Table 26 specifies site families and task scores.

The prefix contrast repeats eight clean prompts on base models; factual recall uses 12, with varying counterfactuals. These are controlled stress tasks. Instruction-tuned cells instead sample 224 prompts per seed from HarmBench (Mazeika et al., 2024) and JailbreakBench (Chao et al., 2024), with benign Alpaca instructions (Taori et al., 2023); their endpoint is the refusal-opener minus compliance-opener logit. Both contrasts are proxy scores.

Matching, execution and uncertainty. Base cells use 256 evaluation rows and instruction/official-checkpoint cells use 224. Defaults are top-1 agreement, $\delta = .05$, and $|\mathcal{B}_\delta| \geq 16$. Simulated low-bit weights are quantise–dequantise values under a shared BF16 forward; grouped INT4 uses RTN. Base-task transformation calibration uses the first 128 clean rows when required; RTN is data-free. Seeds vary task sampling or counterfactual pairing. The sentinel scale guard uses three FP16-control runs in each of 84 model/task/explainer groups; Table 8 records repeat counts and floors for each evidence layer. Primary median intervals resample configurations with their seeds together (10,000 draws); per-cell intervals resample distinct clean-prompt clusters with their repeated rows. Both condition on fixed interventions. Sentinel discovery and evaluation share a prompt pool; GPT-2 32/64-prompt splits and targeted Qwen studies freeze rankings and payloads before held-out evaluation (Appendices I and L). Legacy cell-bootstrap/Wilson intervals describe their grids. CPU follow-ups are reported separately from the GPU sentinel.

Table 2: **INT8 usually preserves native-relative efficacy.** The corrected sentinel uses a 0.229-logit floor; OK means $\rho \geq 0.8$, severe means $\rho < 0.5$. Agreement is median top-1 agreement over completed cells. Silent severe cells also retain at least 90% output agreement. The prefix task repeats eight base-model prompts.

(a) Corrected FP16/INT8 sentinel

Transform	Complete	Info.	Agreement	Median ρ	OK	Severe
FP16 cast	252/252	246	100.0%	1.000	100.0%	0.0%
RTN INT8	252/252	245	94.5%	0.999	95.1%	3.3%

(b) INT8 by task and intervention

Task	Intervention	n	Median ρ	OK	Severe	Silent severe
Factual recall	Circuit ablation	21	1.002	100.0%	0.0%	0/15
	Causal patching	21	0.999	100.0%	0.0%	0/15
	Steering vector	21	0.996	100.0%	0.0%	0/15
Greater-Than	Circuit ablation	21	0.997	95.2%	4.8%	1/6
	Causal patching	17	0.999	100.0%	0.0%	0/5
	Steering vector	21	0.993	95.2%	4.8%	0/6
IOI	Circuit ablation	20	1.000	100.0%	0.0%	0/17
	Causal patching	21	1.002	100.0%	0.0%	0/18
	Steering vector	21	1.002	100.0%	0.0%	0/18
Assistant-prefix	Circuit ablation	21	0.895	57.1%	28.6%	6/9
	Causal patching	19	0.982	94.7%	0.0%	0/9
	Steering vector	21	0.995	100.0%	0.0%	0/9

Table 3: **The same severe native gap can coexist with source fidelity.** All four configurations containing severe sentinel cells, with mean effects in logits. d is the source effect, u its transported effect and v the native reference. Prefix values coincide across three seeds. † marks a source mean below the descriptive guard whose interval crosses zero.

Configuration	Cells	d	u	v	$\rho = u/v$	$\phi = u/d$
Qwen-0.5B / prefix	3	1.016	.908	5.391	.168	.894
Qwen-1.5B / GT ablation	1	−.016	−.037	.725	−.051	†
Qwen-1.5B / prefix	3	.969	.967	5.445	.178	.998
Llama-1B / GT steering	1	.584	.606	−.509	−1.189	1.037

5 Native-Reference Gaps and Source-Effect Preservation

Preservation in the bulk, disagreement in a small tail. The corrected INT8 sentinel has 245 informative cells from 83 configurations, median $\rho = 0.999$ (configuration-bootstrap CI $[0.996, 1.002]$); 233/245 retain at least 80% of the native effect. Eight cells in four configurations fall below 0.5, with each prompt-cluster interval wholly below that threshold. Their median absolute native effect and transported–native gap are 5.39 and 4.48 logits. The gaps are therefore substantial in task units. Seven retain at least 90% output agreement, demonstrating that this behavioural screen does not detect native-reference disagreement.

The same records give a different answer about original-effect preservation. Among 243 cells whose source-effect magnitudes also exceed the .229-logit descriptive guard, median ϕ is .998 and point estimates span $[.792, 1.186]$; none is below .5. Table 3 places the two ratios side by side for all four severe configurations. In the Qwen-0.5B prefix case, all 256 rows retain their top prediction and CKA is .9997; $\rho = .168$ coexists with $\phi = .894$. The Llama steering source effect stays positive ($.584 \to .606$), while the native reference is negative. These are disagreements with the selected reference, not evidence that the original effect changed sign. The source-fidelity points remain uncertain: the Qwen-0.5B interval reaches .487, and the source mean of the Qwen Greater-Than case is unresolved. Absence of severe point estimates is not a confidence guarantee.

Controls establish that the contrast is measurable. The exploratory seven-transform screen has median $\rho \in [.982, 1.008]$ wherever estimable, but coverage contracts: NF4 and grouped INT4 (groups 128/64) yield 154, 54 and 104 informative cells out of 252 each; magnitude pruning yields none (Table 12). These pre-correction results describe surviving matched subsets, not a broader corrected tail estimate. Random-site selections are severe in 48.0% of the exploratory breadth cells, versus 5.3% for transported selections in that layer (Table 12). In the corrected sentinel, FP16-control transport has zero severe cells out of 246, versus eight out of 245 under INT8. These arms check sensitivity to site choice and resolution under a small deployment change. They are descriptive paired-grid controls, not independent samples for a population quantisation test.

Where the tail occurs determines its scope. Six severe cells are Qwen prefix ablations; factual recall and IOI have none. The cell-weighted severe fraction has configuration-cluster interval $[0.4, 7.3]\%$, reflecting concentration in four configurations. Severe cells lie well inside the matching cutoff: their median within-cell KL is .0034, versus .0046 elsewhere; median within-cell 95th-percentile KL is .0069 versus .0195. This task-confounded comparison and the eight-prompt prefix population motivate the targeted held-out test below, rather than extrapolation to a deployment-wide failure rate.

6 Diagnosing the Gap: Selection, Repair and Endpoint

6.1 Crossed execution identifies a reference-selection gap

We test the severe Qwen-0.5B prefix configuration on CPU BF16/RTN8, retaining historical GPU site identities and fitting mean payloads on the original eight prompts before freezing them. Eight authored request domains provide 32 calibration, 96 new-instance and 128 new-template prompts. Behavioural matching retains 29, 88 and 117, respectively. On the two evaluation partitions, $\rho = .097/.101$, with domain-cluster 95% upper endpoints $.172/.193$. The fixed-site reference gap thus extends to new prompts. Source-fidelity point estimates are $.865/.918$; for the new-template partition the source-mean interval crosses zero, so its ratio interval is withheld (Appendix L). This is a CPU refit of historical sites, with held-out authored prompts, rather than a replay of the full GPU artefacts.

Almost all of the reference gap exists before quantisation. On new templates, the source artefact has mean effects $.550$ on development and $.505$ on deployment; the native artefact has effects 5.031 and 5.002. Equation 5 decomposes the 4.497-logit gap into 4.481 $[4.471, 4.490]$ already on development and only $.016$ $[.004, .029]$ from their different responses to deployment (Appendix L). The former accounts for 99.6% of the point-estimated gap. The source artefact's own execution change is $-.045$ $[-.077, -.014]$ logits. Site/payload crossing further assigns symmetric contributions of 4.618 and $-.120$ logits to site choice and mean payload, respectively. These controlled executions locate the discrepancy in the selected artefacts, while Figure 2 shows the paired prompt effects and execution-change distributions. They do not establish the redundant-weight mechanism of Proposition 1.

Controlled site swaps sharpen this diagnosis. On 88 matched new instances, holding native means fixed, replacing source site 22 with 21 raises the effect from $.426$ to 4.388 logits; replacing both differing sites gives 4.979. The joint-swap interaction is $.787$ $[.770, .802]$ logits (Appendix L.2).

6.2 Selection diagnostics connect the theory to the audit

The RTN construction loses the source effect through compensating weight rounding; the observed Qwen gap instead largely reflects reference choice. To distinguish candidate mechanisms, first use crossed execution, then inspect discovery stability on fixed fitting data: selected-site overlap, the score margin at the top-k boundary, and sensitivity to tie-breaking. Small margins motivate controlled site substitutions; they do not certify effect loss. Among 161 informative sentinel cells with complete scores, exact source boundary ties occur in 3/7 severe cells and 32/154 other cells, with median relative margins $.018$ and $.026$. All three severe Qwen-0.5B cells have ties; the three Qwen-1.5B prefix cells do not (Table 22). Thus tie-breaking is a candidate for a subset of selections, rather than an explanation of the whole tail. Exact re-discovery falls from 93.9% in FP16 controls to 67.3%

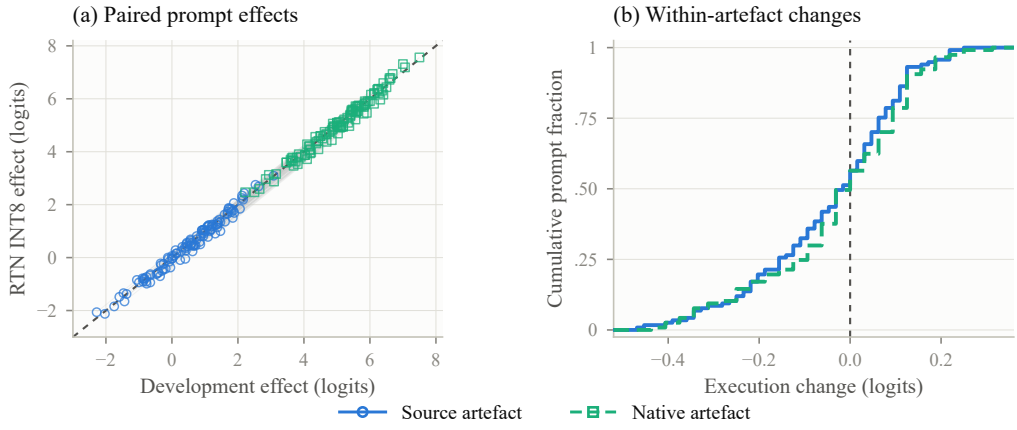


Figure 2: **A large native gap need not be a large execution change.** Qwen-0.5B, CPU BF16/RTN8; all 117 matched template-held-out prompts. (a) Each marker pairs one frozen artefact's effects on the two models; faint segments connect source and native artefacts on the same prompt. The dashed diagonal marks unchanged effects. (b) Empirical cumulative distributions of $u_i - d_i$ and $v_i - b_i$; the dashed vertical marks zero change. The plot shows prompt variation, not uncertainty in a population estimate.

under INT8 ($91.4\% \to 39.8\%$ for ablation), but overlap and retention correlate weakly ($r_s = .13$; Table 20). Turnover supports triage, not a quantitative failure predictor.

Fresh, equal-example Qwen discovery likewise separates selection from fidelity: new-instance/template $\rho = -3.348/-3.314$ while $\phi = 1.000/1.003$. This follow-up corrects partial-batch gradient weighting; it concerns newly selected artefacts, and the historical full-batch prefix selections are unaffected by that defect. Earlier disjoint GPT-2 discovery gives 48 RTN8 measurements with $\rho = .982$–1.006 across site budgets $k = 6, 12, 24, 48$. The Qwen budget sweep is non-monotonic, including FP16-control drift (Appendix I). Across the sentinel, Proposition 2 certifies mean sign in 166/245 cells, including six severe attenuations; it abstains on both native-relative reversal points. Selection stability, sign and magnitude therefore require separate checks. Earlier diagnostics on ten selected configurations give median $\rho = .490$ for identity, .345 for calibration-fitted Procrustes and .349 for ridge transport (Table 18). These maps do not improve median retention in this selected sample; other alignment objectives remain open.

Repairing native parity can sacrifice source fidelity. Twelve source-site gates, $h \mapsto (1-a)h + a\bar{h}$ with $a \in [0, 2]$, are fitted to either target using calibration prompts and 24 fixed updates. Native-target gates reach template $\rho = .917$ and reduce native-target MSE from 20.228 to .186 logits^2, while raising source-target MSE from .027 to 16.349. Source-target gates change their own MSE only from .027 to .026, with an improvement interval crossing zero. The pilot's response structure is simple: 99.99% of the original native-target MSE is squared mean bias. A calibration-only additive prediction has MSE .001840, below the gates' .186. This prediction executes no repair; it shows that the pilot provides no evidence of correspondence recovery beyond an offset. The empirical lesson is to specify the repair target, with richer functional alignment left open (Appendix L.2).

6.3 A VERDICT APPLIES TO THE MEASURED ENDPOINT

Instruction-tuned RTN8 and selected official GPTQ/AWQ checkpoints have no severe next-token cells (Table 15); instruction-tuned RTN8 has median $\rho = .991$ across 17 informative cells. At strict $\delta = .05$, only 12/24 official cells are informative; selected $\delta = .2$ admits 24/24, with deployment medians .834–1.019. This is checkpoint validation with explicit coverage limits, not an in-pipeline matched-everything quantizer comparison. In the 12 official-checkpoint sequence cells (976 paired prompt evaluations), however, four paired refusal-rate gaps survive Holm correction. Seven cells have next-token ρ within 10% of one, and four of these have sequence-rate gaps above .1. These

comparisons span separate outcome evaluations; historical logs lack intervened first-token IDs and do not establish identical prompt membership across experiments.

A targeted RTN8 follow-up supplies exact token records on 32 matched prompts. Comparing already fitted native-target gates with the native intervention gives scalar $\rho = .919$: 27 pairs share the first generated token, 12 continuations differ, and seven share the first token but diverge later. Native replay reproduces all 32 original continuations. This demonstrates why first-token agreement cannot substitute for a continuation check; the phrase-based endpoint and token identity measure different properties (Appendix L.3).

6.4 Robustness of the audit

Holding interventions fixed, twelve floor, matched-set-size, and KL re-scoring settings give median retention 0.999–1.000 and severe fractions 3.3–4.8% (Table 21). These are sensitivity checks on the same observations. They do not estimate failures outside the matched set.

The calibration check also separates eligibility from the observed failures. Leaving the evaluated configuration, model, task, or explainer out of floor estimation yields 241–245 informative cells and retains all eight original severe cells (Table 9). Using sample rather than population SD raises the sentinel floor from 0.229 to 0.281 with the same eight failures. Nested group/run resampling exposes substantial uncertainty in a three-run scale estimate; its 95% percentile range is $[0.081, 0.191]$ logits, while every draw retains the original eight. This is an estimator sensitivity analysis, not an independently calibrated kernel-noise interval.

Ratio-free diagnostics recover the discrepancies: a through-origin slope below 0.5 flags seven severe cells, as does a one-logit mean-effect gap. The union of that gap and a mean sign reversal identifies all eight and no other recorded cells (Table 23). These retrospective criteria corroborate the observed tail; their thresholds have no held-out predictive validation. Algorithm 1 records coverage, unresolved denominators, bulk effects, and the tail separately. The operational report preserves absolute effects and both denominator statuses: an empty matched set is a premise failure, and a sub-floor reference is unresolved; neither is labelled catastrophic transport.

7 Discussion and Conclusion

Intervention portability has two targets: preserving the original effect and matching native re-discovery. Native-relative retention has median .999, yet eight severe reference gaps coexist with source-fidelity point estimates above .79 where both references pass the guard. Crossed Qwen execution identifies a concrete distinction: almost all of the selected artefacts' effect gap already exists on development. Specifying the reference therefore changes both the diagnosis and repair objective.

A three-step audit. **(1)** Declare the required effect reference, endpoint, backend and held-out population; verify behavioural matching and report coverage. **(2)** Measure original and transported effects, adding native re-discovery for native parity. Use documented controls, paired gaps and prompt-cluster intervals; execute the native artefact on development to separate selection gaps from changed execution. With insufficient repeats, report denominator uncertainty and absolute gaps. **(3)** Fit repairs to the declared target on fitting data only and evaluate held-out effects against a predeclared practical threshold. For generation, include paired continuations alongside next-token scores.

Scope. The primary estimate covers small models and weight-only RTN; six severe cells come from an eight-prompt stress task. The crossed study covers one CPU BF16 Qwen configuration and eight authored domains. Source-fidelity point estimates are not equivalence guarantees. Official GPTQ/AWQ results are checkpoint comparisons, not matched in-pipeline quantizer tests; phrase endpoints establish no safety guarantee. Activation quantisation, rotation-based and learned rounding (Chee et al., 2023; Ashkboos et al., 2024; Li & Panda, 2025), broader GPU replication, and sequence-level control calibration are the next tests of this audit.

REPRODUCIBILITY STATEMENT

The supplement contains the original breadth screen, corrected 504-cell sentinel, instruction-tuned and official-checkpoint extensions, sequence records, and all executed CPU follow-ups. Historical raw observations are unchanged. New runs record prompt-content hashes, fitting/evaluation membership, selected sites and payloads, software/backend settings, source hashes and completion manifests. The Qwen study additionally ships both causal training traces, final gate coefficients, all crossed per-prompt effects, exact generated token IDs, and model-weight restoration checks. The separate equal-example discovery correction and frozen-gate sequence follow-up each have their own execution and source manifest. The latter repeats the parent's exact 32 prompts and records its post-parent specification.

The README gives the ordered, model-free regeneration commands. In particular, `analysis_review2.py` and `analysis_review3.py` recover the sentinel uncertainty, effect bounds, matching composition and historical sequence comparisons; `analysis_reference_fidelity.py` adds the source reference. `summarize_prefix_holdout.py` and `summarize_weighted_followup.py`, together with `summarize_prefix_gate_sequence.py`, reject incomplete or hash-inconsistent studies before generating paired statistics and CSV files. Their inference counterparts have matching `run_` entry points. Figure and table generators consume these outputs. CPU regression tests cover masking, matching, payload alignment, quantisation, denominator handling, actual-model split isolation, gate gradients, paired resampling and equal-example attribution. The submission includes the anonymous source tree; code and results will receive a permissive public release at camera-ready.

A historical naming hazard remains documented: the raw `effect_spearman` field is not the per-example rank correlation Ex-r_s in Table 7. Re-analysis recomputes the latter from the paired `diagnostics.effect_values` arrays.

The paper distinguishes (i) the original breadth screen from the corrected confirmatory sentinel, (ii) simulated quantise–dequantise transforms from official low-bit checkpoints, (iii) strict $\delta = .05$ packed support from the common selected $\delta = .2$, and (iv) valid refusal experiments from the base-model assistant-prefix stress test. Interrupted GPU jobs append by cell key and final archives are emitted only after all required stages finish. The per-cell wall-clock times stored in the released records sum to 41.1 GPU-hours on a single RTX 5090 (breadth 39.6, corrected sentinel 0.5, instruction 0.4, packed, sequence and diagnostic layers 0.6); this counts the runs that landed, and excludes discarded attempts, model downloads, and CPU inference extensions. New CPU job timings and exact software versions are stored separately; they are not added to the historical GPU-hour total.

ETHICS STATEMENT

This work audits reliability after model compression. The instruction-tuned extension uses public HarmBench/JailbreakBench behaviours and benign Alpaca instructions. Their historical sequence records contain refusal labels and token hashes. The separate authored-prefix pilot releases generated token IDs for exact paired comparison and makes no semantic safety judgement. The work does not release jailbreak methods or claim that a next-token contrast is a complete safety evaluation. Its recommendation—re-audit causal artefacts on the served precision and backend—is intended to expose, not conceal, deployment-specific failures. No human subjects or personal data are involved.

AI USE STATEMENT

Generative AI tools assisted with prose editing, code drafting, automated experiment execution and analysis, plotting, and candidate literature discovery. Experimental observations come from executed model runs; historical raw records are preserved. Numerical claims are regenerated from the released results, AI-assisted code is covered by the test suite, and the authors remain responsible for reference verification, interpretation, and the submitted artefacts.

REFERENCES

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, 2025.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Yannis Belkhiter, Giulio Zizzo, and Sergio Maffeis. HarmLevelBench: Evaluating harm-level compliance and the impact of quantization on model alignment. *arXiv preprint arXiv:2411.06835*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*, 2024.

Joschka Braun. Understanding unreliability of steering vectors in language models: Geometric predictors and the limits of linear approximations. *arXiv preprint arXiv:2602.17881*, 2026.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 37, 2024.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, 2023.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Satchel Grant. Model alignment search. *arXiv preprint arXiv:2501.06164*, 2025. doi: 10.48550/arXiv.2501.06164.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Jung Hwan Heo, Jeonghoon Kim, Beomseok Kwon, Byeongwook Kim, Se Jung Kwon, and Dongsoo Lee. Rethinking channel dimensions to isolate outliers for low-bit weight quantization of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. Exploring the trade-offs: Quantization methods, task difficulty, and model size in large language models from edge to giant. *arXiv preprint arXiv:2409.11055*, 2024.

Yuhang Li and Priyadarshini Panda. TesseraQ: Ultra low-bit LLM post-training quantization with block reconstruction. In *International Conference on Learning Representations (ICLR)*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224, 2024.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models. In *International Conference on Machine Learning (ICML)*, 2025.

Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. Outliers and calibration sets have diminishing effect on quantization of modern LLMs. *arXiv preprint arXiv:2405.20835*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Kazi Kamruzzaman Rabbi, Md. Zami Al Zunaed Farabe, and M. Sohel Rahman. Same predictions, different reasons: The effect of quantization on model explanations. *arXiv preprint arXiv:2607.22872*, 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Prashant C. Raju. The geometric canary: Predicting steerability and detecting drift via representational stability. *arXiv preprint arXiv:2604.17698*, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations (ICLR)*, 2025.

Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. doi: 10.1214/aos/1176345338.

Manpreet Singh and Hassan Sajjad. Interpreting the effects of quantization on LLMs. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2267–2281, 2025. doi: 10.18653/v1/2025.ijcnlp-long.123.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. *arXiv preprint arXiv:2603.30002*, 2026.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. GitHub repository, 2023. URL `https://github.com/tatsu-lab/stanford_alpaca`.

Jonathan von Rad, Yong Cao, and Andreas Geiger. UniComp: A unified evaluation of large language model compression via pruning, quantization, and distillation. *arXiv preprint arXiv:2602.09130*, 2026. URL `https://arxiv.org/abs/2602.09130`.

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

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv preprint arXiv:2505.13963*, 2025.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, 2022.

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

Manyi Zhang, Ji-Fu Li, Zhongao Sun, Haoli Bai, Hui-Ling Zhen, Zhenhua Dong, and Xianzhi Yu. Benchmarking post-training quantization of large language models under microscaling floating point formats. *arXiv preprint arXiv:2601.09555*, 2026.

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao, and Kang Liu. From signal degradation to computation collapse: Uncovering the two failure modes of LLM quantization. *arXiv preprint arXiv:2604.19884*, 2026. URL `https://arxiv.org/abs/2604.19884`.

A THEORY AND OBSERVABLE EFFECT BOUNDS

Table 4: **Notation and vocabulary in one place.** Every quantitative claim in the paper is tagged with the evidence layer it comes from; the layers use different efficacy floors and different cell populations and are never pooled.

Term	Meaning
$\mathcal{B}_\delta$	Behaviour-matched set: inputs where M and $T(M)$ share the top-1 token and have $D_{\mathrm{KL}} \le \delta$
ρ, CG	Native-relative retention $\mathrm{Eff}(I_M)/\mathrm{Eff}(I_{T(M)})$ on the deployment and gap $1-\rho$
ϕ	Source-effect fidelity: transported effect divided by the original intervention's effect on the development model, with a separate denominator check
d, u, v, b	Mean effects: source artefact on development/deployment (d, u), native artefact on deployment/development (v, b)
τ_{eff}	Efficacy floor: three times the median FP16-control run-to-run standard deviation of the native effect
π	Minority-sign mass of the native per-example effect (Equation 17)
Ex-r_s	Spearman correlation between transported and native *per-example* effects, recomputed from stored arrays
L1 / L2 / L3	Behavioural agreement / representational alignment / transported–native intervention-effect parity
Informative	Cell with $\lvert\mathcal{B}_\delta\rvert \ge 16$ and $\lvert\mathrm{Eff}_{\mathrm{native}}\rvert \ge \tau_{\mathrm{eff}}$; only these enter calibrated grid summaries; targeted raw ratios are labelled
Complete	Cell that ran to termination, whether or not it is informative
Degraded / severe	$\rho < 0.8$ / $\rho < 0.5$
Silent severe	Severe cell that nonetheless keeps $\ge 90\%$ output agreement
Repeated indices	Equal local row numbers across seeds; not necessarily equal task inputs (Appendix B.1)
Breadth screen	1,764 planned cells, seven transforms, floor 0.279; establishes coverage
Corrected sentinel	504 completed FP16/RTN INT8 cells, floor 0.229; the primary INT8 native-relative tail estimate
Instruction layer	72 cells on Qwen-Instruct, floor 0.311
Packed layer	24 cells on official GPTQ/AWQ checkpoints

Figure 1 in the main text gives the full audit workflow.

Round-to-nearest on $\Delta\mathbb{Z}$ maps $w_0 = \Delta/2 - \epsilon$ to zero and every $w_j = \Delta/2 + \epsilon$ to Δ. Before quantisation,

$$f(x) = w_0 x + \sum_{j=1}^{k} w_j \frac{x}{k} \tag{9}$$

$$= \left(\frac{\Delta}{2} - \epsilon + \frac{\Delta}{2} + \epsilon\right) x = \Delta x. \tag{10}$$

After quantisation,

$$T(f)(x) = 0 \cdot x + \sum_{j=1}^{k} \Delta \frac{x}{k} = \Delta x. \tag{11}$$

Thus the models agree for every input, not only in top prediction. The component activations h_j are unchanged, so representation similarity before the scalar readout is also exact.

For $x > 0$, component 0 contributes $(\Delta/2 - \epsilon)x$, while each backup contributes $(\Delta/2 + \epsilon)x/k$. The assumed bound is equivalent to

$$\epsilon < \frac{(k-1)\Delta}{2(k+1)} \quad \Longleftrightarrow \quad \frac{\Delta}{2} - \epsilon > \frac{\Delta/2 + \epsilon}{k}, \tag{12}$$

so component 0 is uniquely largest before quantisation. Afterwards it contributes zero and every backup contributes $\Delta x/k$. Zero-ablating the transported component has efficacy zero; ablating a

natively selected backup changes the output by $-\Delta x/k$. Hence $\rho = 0$ and $\mathrm{CG} = 1$ despite exact behavioural and representational agreement.

Remark 1 (Perturbation robustness and native ties). The source selection is strict. Although the native backups tie, every one has the same ablation effect, so native tie-breaking does not affect the result. Moreover, let $\epsilon_\star = (k-1)\Delta/[2(k+1)]$, choose $0 < \eta < \min\{\epsilon, \epsilon_\star - \epsilon\}$, and perturb every weight to $\widetilde{w}_j = w_j + r_j$ with $|r_j| \le \eta$. The bounds imply $0 < \widetilde{w}_0 < \Delta/2$ and $\Delta/2 < \widetilde{w}_j < 3\Delta/2$ for $j \ge 1$, so all RTN bins are unchanged. The source's leading component also remains uniquely largest: for every backup,

$$\widetilde{w}_0 - \frac{\widetilde{w}_j}{k} \ge \frac{k+1}{k}(\epsilon_\star - \epsilon - \eta) > 0. \tag{13}$$

Writing $\widetilde{f}$ for the perturbed model therefore gives

$$T(\widetilde{f})(x) = \Delta x, \qquad |\widetilde{f}(x) - T(\widetilde{f})(x)| = \left| r_0 + \frac{1}{k}\sum_{j=1}^{k} r_j \right| |x| \le 2\eta|x|. \tag{14}$$

Transport still ablates the zero-weight component, while every native backup has effect $-\Delta x/k$. Thus $\phi = \rho = 0$ throughout this nonzero weight neighbourhood on positive-input distributions with finite mean; mean absolute output error is at most $2\eta\mathbb{E}[X]$ and can be arbitrarily small. Exact cancellation supplies the centre of the construction, rather than a necessary condition for nearby low-error examples. This statement establishes possibility, not the prevalence or neural mechanism of empirical gaps.

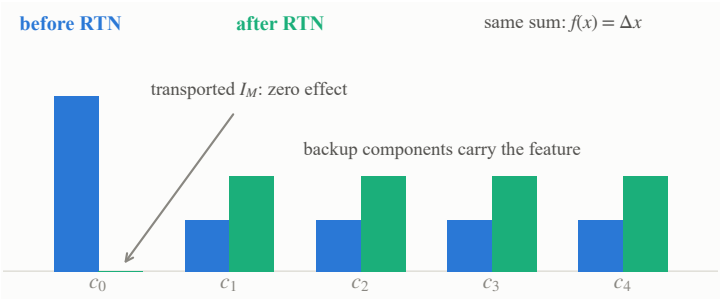


Figure 3: **Exact behaviour can coexist with source-effect loss.** Analytic construction from Proposition 1: RTN zeroes the selected component c_0, while rounding the redundant backups preserves the full function. The transported ablation consequently has zero effect. This illustrates the possibility result, not an observed neural mechanism.

Proof of Proposition 2. Orient both effects by $s = \operatorname{sign}\mathbb{E}[v]$. Their ratio is unchanged and $\mu = \mathbb{E}[sv] > 0$. On G, alignment and the response bounds give

$$\lambda A \le \mathbb{E}[su\mathbf{1}\{x \in G, sv > 0\}] \le \kappa A, \tag{15}$$

$$-\kappa B \le \mathbb{E}[su\mathbf{1}\{x \in G, sv < 0\}] \le -\lambda B. \tag{16}$$

On $G \cap \{v = 0\}$ the upper bound requires $u = 0$. Outside G, the triangle inequality gives $|\mathbb{E}[su\mathbf{1}\{x \notin G\}]| \le \varepsilon$. Adding these inequalities and dividing by μ proves Equation 8. To attain the lower endpoint, set $su = \lambda|v|$ on the aligned majority, $su = -\kappa|v|$ on the aligned minority, and put the entire exceptional transported mass in the negative orientation. Interchange λ, κ and the exceptional orientation for the upper endpoint. These choices preserve the stated group masses. Sharpness refers to these coarse premises; requiring G additionally to be the maximal aligned subset can provide extra information that the bound does not use. If an upper bound on ε is available instead of its exact value, substituting it remains conservative. □

The no-exception special case. When $G = \mathcal{B}_\delta$, zeros are compatible and $\varepsilon = 0$. Define

$$\pi = \frac{\mathbb{E}[|v|\mathbf{1}\{sv < 0\}]}{\mathbb{E}[|v|\mathbf{1}\{sv > 0\}]} = \frac{B}{A}, \qquad 0 \le \pi < 1. \tag{17}$$

Then $\mu = A(1-\pi)$ and

$$\frac{\lambda - \kappa\pi}{1-\pi} \le \rho \le \frac{\kappa - \lambda\pi}{1-\pi}. \tag{18}$$

For $\kappa > 0$, $\pi \le \lambda/\kappa$ precludes sign reversal; for $\lambda = \kappa = 0$ all transported effects vanish. The strict positive-response check below uses $\lambda > 0$ to match the original diagnostic. On two equally weighted inputs, $u = (100, -1)$ and $v = (1, -100)$ agree in sign pointwise but have $\pi = .01 > \lambda/\kappa = .0001$ and $\rho = -1$.

This result concerns measured total effects. A site-additive model with $c_j(x)v(x) \ge 0$ at every transported site is one sufficient way to obtain aggregate alignment, but is unnecessary for the bound. We neither assume such a decomposition for a trained network nor infer it from the certificate.

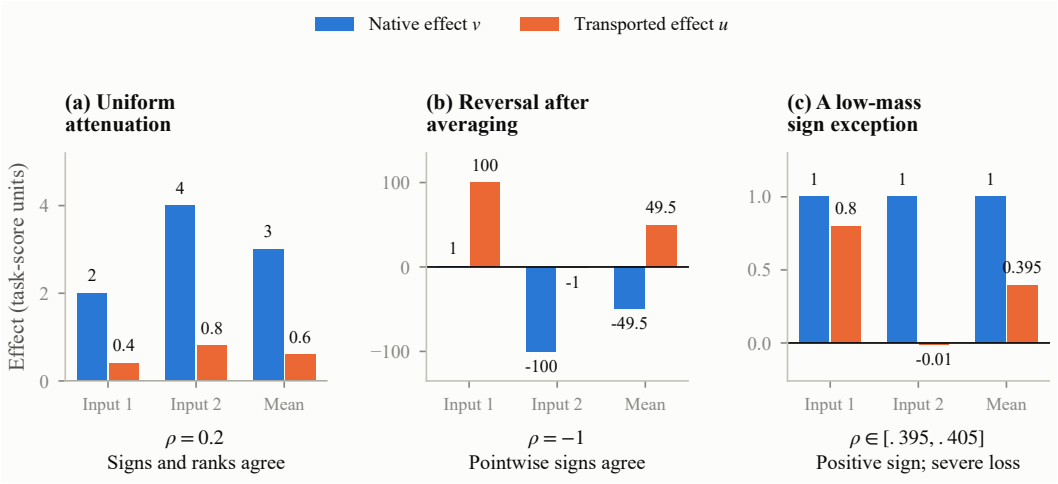


Figure 4: **Why signs and magnitudes must both be audited.** Analytic examples, not model observations. (a) Uniform attenuation preserves ranks and signs but retains only 20% of the effect. (b) Native and transported effects have the same sign on each input, yet their means have opposite signs. The minority native mass is amplified enough to reverse the average. (c) One small sign exception breaks global alignment, but the exception budget certifies a positive mean without certifying retention above 0.5. Hatched bars show means; the panels use different vertical scales.

A complete check of the no-exception special case. For every recorded paired effect array, we check (i) $uv \ge 0$ everywhere and (ii) $u = 0$ on $v = 0$ and $\lambda = \min_{v \ne 0} |u/v| > 0$, with $\kappa = \max_{v \ne 0} |u/v| < \infty$. Zeros are checked exactly, without an epsilon denominator. Only then do we compare π with λ/κ. This is a deterministic certificate on the recorded matched set; it does not extend the bounds to unobserved inputs or quantify discovery uncertainty.

Table 5: **Complete premises of the no-exception special case.** The separate premise counts need not be nested; the last two rows require their conjunction. All certified cells have nonnegative retention.

Check	Cells or configurations
Total-effect sign alignment (i)	182/245
Zero compatibility	179/245
Bounded response (ii), including $\lambda > 0$	160/245
Both premises (i) and (ii)	129/245
Both premises and $\pi \le \lambda/\kappa$	128/245
Distinct configurations with a sign certificate	49/83

The 128 no-exception certificates cover 49 configurations; 117 cells remain uncertified by that special case. An uncertified cell is not a predicted failure. All eight severe cells fail total-effect alignment, including the six with $\pi = 0$: in the two repeated prefix-ablation configurations, 96/256 and 32/256 individual transported effects oppose the native sign. Consequently, minority mass

alone cannot explain these failures. Checking only $\pi \leq \lambda/\kappa$ counts 167 cells, of which 39 fail the complete premises. Likewise, the count 179 for zero compatibility must not be substituted for the full bounded-response premise, which holds in 160. The scripts `analysis_review2.py` and `probe_premise_coverage.py` reproduce the complete check.

Allowing a measured exception budget. For the general bound, set $G = \{x : v(x) \neq 0,\ u(x)v(x) \geq 0\}$, use the minimum and maximum observed $|u/v|$ on G, and charge every other observation, including $v = 0, u \neq 0$, to ε. We allow $\lambda = 0$ and set $\lambda = \kappa = 0$ when G is empty, yielding the trivial absolute-mass bound. This gives nonnegative lower bounds in 166/245 informative INT8 cells, 38 more than the strict no-exception check. Six of the eight severe cells now have positive sign lower bounds (Table 6). In Qwen-0.5B and Qwen-1.5B prefix ablations, respectively, 3/8 and 1/8 distinct prompts violate alignment, but they contribute only 7.96% and 3.04% of transported absolute mass. Their lower bounds are 0.024 and 0.016; observed retention remains 0.168 and 0.177. Thus a positive certificate does not certify retention above 0.5. The two reversal point estimates remain uncertified, with lower bounds -21.565 and -36.912. These calculations diagnose which premise is too brittle for an observed effect array; they neither predict the array nor identify the neural mechanism producing it.

Table 6: **Exception budgets separate sign from attenuation.** Fractions count recorded rows; the first six cells repeat eight prompts. Exceptional mass is $\varepsilon/\mathbb{E}|u|$. All eight cells retain $\rho < 0.5$ regardless of whether the bound certifies nonnegative native-relative retention.

Model / task / seed	ρ	Exceptions	Mass	Lower bound	Sign certified
Qwen-0.5B / Prefix / 0	0.168	96/256	8.0%	0.024	yes
Qwen-0.5B / Prefix / 1	0.168	96/256	8.0%	0.024	yes
Qwen-0.5B / Prefix / 2	0.168	96/256	8.0%	0.024	yes
Qwen-1.5B / GT / 1	-0.051	89/254	20.3%	-21.565	no
Qwen-1.5B / Prefix / 0	0.178	32/256	3.0%	0.016	yes
Qwen-1.5B / Prefix / 1	0.178	32/256	3.0%	0.016	yes
Qwen-1.5B / Prefix / 2	0.178	32/256	3.0%	0.016	yes
Llama-1B / GT / 1	-1.189	120/226	51.3%	-36.912	no

Remark 2 (Sign and rank are independent). A uniform rescaling $u = \alpha v$, $\alpha > 0$, preserves per-example ranks while $\rho = \alpha$: the attenuation branch. But rank agreement does not constrain the sign, since $u = (1, 2)$ and $v = (-2, -1)$ are comonotone ($r_s = 1$) with $\rho < 0$. Sign reversal therefore neither implies nor is implied by rank breakage, and we report their co-occurrence (Section 6) descriptively.

Rank agreement is a separate property. Uniform positive scaling gives $r_s = 1$ but arbitrary positive retention. Conversely, $u = (1, 2)$ and $v = (-2, -1)$ have $r_s = 1$ and $\rho = -1$. The two empirical cells with per-example $r_s < 0.5$ happen to be the two reversal point estimates; the next cell has $r_s = 0.500$ and $\rho = 1.05$. This thresholded description has no implication for a general sign–rank law.

Table 7: **The two severe branches, all 245 informative RTN INT8 cells.** Ex-r_s is the per-example transported/native Spearman, slope a through-origin regression of one on the other, and π the native minority-sign mass of Equation 17. Only reversal carries appreciable π.

Group	n	Median ρ	Ex-r_s	Slope	Jaccard	π	Sign kept
Attenuation (coherent)	6	0.173	1.00	0.20	0.71	0.000	6/6
Reversal (decoherent)	2	-0.620	0.31	0.31	0.36	0.346	0/2
Non-severe	237	1.000	0.99	1.00	1.00	0.001	237/237

B Conditional Estimand and Calibration Details

The audit estimates a conditional deployment quantity. For a task distribution $P_{\mathcal{D}}$ and the event $A_\delta = \{X \in \mathcal{B}_\delta\}$,

$$\mathrm{Eff}(I, M'; \delta) = \mathbb{E}_{P_{\mathcal{D}}}[m(I[M'], X) - m(M', X) \mid A_\delta]. \tag{19}$$

Changing δ can therefore change both support and the population being averaged. The sensitivity analysis in Table 21 avoids re-discovery confounding by holding the intervention fixed and re-filtering the stored per-example effects.

The global guard in Equation 6 uses the median noise scale across groups. The per-configuration alternative is

$$\tau_g = 3\ \mathrm{sd}_s[\mathrm{Eff}(I_{T(M)}, T_0(M)) \mid g], \qquad \rho_g \text{ is scored only if } |\mathrm{Eff}_{g,\mathrm{native}}| \geq \tau_g. \tag{20}$$

Groups without an estimable τ_g fall back to the global floor. The corrected sentinel gives 235 informative real-transform cells under this rule, median $\rho = 0.999$, 3.4% severe risk, seven high-agreement severe cells, and three persistent configurations—the same qualitative result as the global floor.

Repeat counts and what the floor measures. Table 8 gives the exact number of usable FP16 control runs, grouped by model, task, and explainer. Each group uses seeds $\{0, 1, 2\}$, except one instruction group with two completed runs. The SD is $[R^{-1}\sum_s (e_s - \bar{e})^2]^{1/2}$; it is an empirical run-spread scale, not an unbiased estimate of a universal noise variance. The runs can change prompts, counterfactuals, selected sites, and floating-point execution together. No fixed-input repeated-kernel experiment isolates the last contribution.

Table 8: **Exact calibration sample sizes.** Runs are seed-indexed discovery/evaluation executions; repeated rows within a run are not additional calibration repeats.

Layer	Groups	Runs	Repeats per group	τ_{eff}
Breadth	83	249	83 groups × 3	0.279
Corrected sentinel	84	252	84 groups × 3	0.229
Instruction-tuned	6	17	1 groups × 2; 5 groups × 3	0.311

The sentinel's explainer-specific pooled floors are 0.292 for attribution, 0.168 for patching, and 0.304 for steering (28 groups each). Thus the global median is a shared convention, not evidence of equal explainer noise. With sample SD (divisor $R - 1$), the sentinel floor becomes 0.281; 242 cells remain informative and the eight original severe cells remain.

Held-out and nested calibration sensitivity. For each evaluated cell, we refit the floor excluding its configuration, model, task, or explainer from the control groups. These leave-out fits do not reuse that stratum's control effects. They retain all eight original severe cells; eligibility changes by zero, one, two, and four cells respectively. This checks calibration reuse, not out-of-sample intervention discovery.

Table 9: **Calibration changes eligibility more than the observed tail.** Floor ranges span held-out strata. Every row retains the original eight severe cells; all numbers come from the corrected sentinel.

Calibration	Floor	Info.	Median ρ	Severe	Severe %
Three-run global floor	0.229	245	0.999	8	3.3
Sample SD (divisor $R - 1$)	0.281	242	0.999	8	3.3
Leave configuration out	0.224–0.234	245	0.999	8	3.3
Leave model out	0.212–0.264	244	0.999	8	3.3
Leave task out	0.172–0.311	243	0.999	8	3.3
Leave explainer out	0.180–0.304	241	0.999	8	3.3

With 10,000 draws, resampling the 84 whole control groups gives floor percentiles $[0.166, 0.311]$ and informative-count percentiles $[241, 248]$. A nested scheme also resamples the three runs within every sampled group, giving $[0.081, 0.191]$ and $[246, 249]$, respectively. The latter shifts downward because resampling just three empirical values often duplicates a run. Every draw in both schemes retains all eight original severe cells; severe-fraction percentiles are $[3.25, 3.64]\%$. Dropping one seed at a time gives floors 0.114–0.162 and 248–249 informative cells, with nine severe cells. These are sensitivity distributions of the available estimator, not validated confidence intervals for independent kernel noise. Additional independent repeats are needed to identify that quantity.

B.1 PROMPT IDENTITY AND THE RESAMPLING UNIT

The shipped base-task generators expose important repetition. The prefix task cycles through eight clean prompts; factual recall cycles through 12, while varying counterfactuals. IOI and Greater-Than sample new text cases by seed. We reconstruct clean strings from the shipped generator and cluster all rows sharing a clean prompt, keeping their counterfactual variants together. Regression tests compare reconstructed texts and targets with the actual task builders for all four tasks and all three seeds.

Equal local row numbers do not establish cross-seed input identity. Of 168 multi-seed configurations, 96 have identical matched index arrays but only 14 have identical reconstructed task cases relevant to their explainer. For ablation the signature contains clean text and both target tokens; patching and steering also include counterfactual text. Therefore 82 index-identical configurations still vary their relevant cases. A full matched mask says only that each run retained its own pool. It cannot isolate kernel nondeterminism.

Table 10: **Cross-seed identity uses task content, not local row numbers.** Text signatures are reconstructed from the shipped default generators.

Property	Configurations
Multi-seed FP16/RTN INT8 configurations	168
Identical local matched row indices	96
Identical relevant text cases	14
Identical row indices but different relevant cases	82
RTN8 configurations degraded at every seed	3
all three reuse the prefix-ablation task cases	3

Table 11: **Severe-cell uncertainty after clustering repeated prompts.** Each of 10,000 bootstrap draws resamples clean-prompt clusters, retaining all their rows. Ratios are scored only when the resampled native mean clears $\tau_{\text{eff}} = 0.229$; the last column reports resolved draws. All upper endpoints remain below 0.5. GT denotes Greater-Than.

Model/task/seed	Rows	Prompts	ρ	95% interval	Resolved
Qwen-0.5B / Prefix / 0	256	8	0.168	[0.034, 0.277]	10000/10000
Qwen-0.5B / Prefix / 1	256	8	0.168	[0.037, 0.279]	10000/10000
Qwen-0.5B / Prefix / 2	256	8	0.168	[0.035, 0.280]	10000/10000
Qwen-1.5B / GT / 1	254	216	-0.051	[-0.697, 0.328]	9999/10000
Qwen-1.5B / Prefix / 0	256	8	0.178	[0.081, 0.259]	10000/10000
Qwen-1.5B / Prefix / 1	256	8	0.178	[0.080, 0.261]	10000/10000
Qwen-1.5B / Prefix / 2	256	8	0.178	[0.081, 0.260]	10000/10000
Llama-1B / GT / 1	226	196	-1.189	[-2.061, -0.591]	9998/10000

Prompt clustering widens the prefix intervals but does not remove severity. It also distinguishes the Greater-Than sign claims: the Qwen interval crosses zero whereas the Llama interval is entirely negative. Interventions remain fixed during resampling, so these intervals do not account for adaptive discovery or establish generalisation beyond the observed prompt pool.

Why a numerical epsilon is not an efficacy floor. In the breadth screen, 35 completed cells fall below $\tau_{\text{eff}} = 0.279$. The 34 non-zero denominators generate raw ratios from -31.2 to $+20.0$ (Figure 5). Including them moves the overall mean from 0.949 to 0.894 while leaving the median at one. Machine precision prevents division by zero; it does not identify a task-logit effect.

Algorithm 1 Noise-calibrated commutativity audit

Require: development model M, deployment $T(M)$, explainer E, compiler ι, task score m, disjoint fitting/evaluation prompts, seeds S, matched-set thresholds

1: Run documented FP16 controls over S and estimate τ_{eff} by Equation 6; state a separate source-effect guard
2: **for** each model–task–intervention–deployment–seed cell **do**
3: build $\mathcal{B}_\delta$ on evaluation prompts; if below the minimum, record a premise failure and stop the cell
4: derive I_M on M and $I_{T(M)}$ on $T(M)$ using matched fitting prompts only; freeze sites and payloads
5: measure source effect d on M and transported/native effects u, v on $T(M)$, all on the same evaluation $\mathcal{B}_\delta$
6: if $|v| < \tau_{\text{eff}}$, record an unresolved native denominator; otherwise compute $\rho = u/v$
7: report $\phi = u/d$ only with its source-denominator status; always keep the paired gaps $u - v$ and $u - d$
8: if diagnosing a reference gap, measure $b = \text{Eff}(I_{T(M)}, M)$ and decompose $v - u = (b - d) + [(v - b) - (u - d)]$
9: **end for**
10: report planned/completed/informative coverage, agreement, absolute effects, native parity, source fidelity, $R_{0.8}$ and $R_{0.5}$
11: group repeated seeds by configuration, and bootstrap over the examples in $\mathcal{B}_\delta$ by distinct clean-prompt cluster for each cell's interval

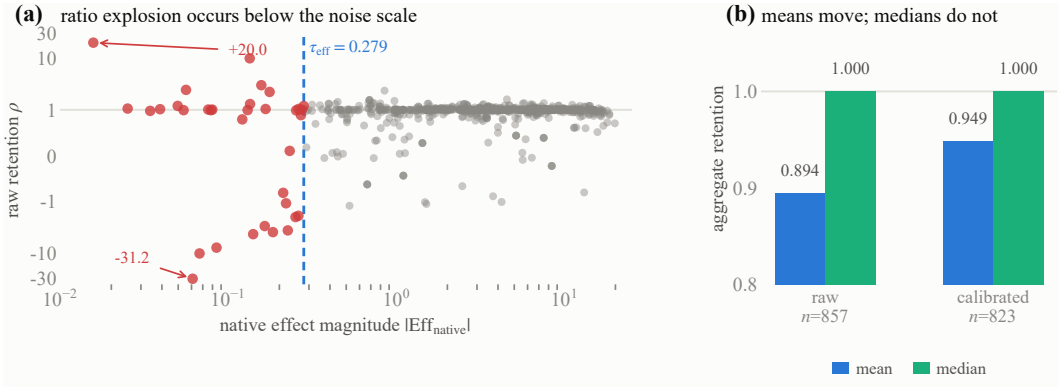


Figure 5: **Weak denominators distort the raw ratio.** *(a)* Breadth-screen retention against native-effect magnitude. *(b)* The median is stable while the mean is pulled by cells below the control-calibrated floor.

C Full Breadth Grid

The breadth-screen completed/informative counts for FP16, RTN8, NF4, RTN4 (g128), RTN4 (g64), Wanda, and magnitude pruning are respectively 249/241, 249/244, 162/154, 63/54, 105/104, 30/26, and 0/0 (Table 13 gives the exact Cartesian product, and Figure 6 the per-transform disposition and global agreement). Of 906 failed cells, 897 have fewer than 16 matched inputs and nine are OOM. These disposition counts are retained for transparency; corrected RTN8 counts are in the main paper.

Table 12: **Exploratory breadth and resolution controls.** Breadth rows predate the mask correction. Informative means at least 16 matched rows and native efficacy above the layer's floor. Cell-bootstrap/Wilson intervals are descriptive grid summaries. Control arms calibrate measurement contrast; their intervals do not provide independent-samples tests.

(a) Breadth screen (simulated transforms)

Transformation	Informative	Agreement	Median ρ [95% CI]	OK	Severe
FP16 cast *(control)*	241/252	1.000	1.000 [1.000,1.000]	100.0%	0.0%
RTN INT8	244/252	0.961	1.001 [1.000,1.003]	93.4%	5.7%
NF4	154/252	0.250	1.002 [0.992,1.021]	88.3%	8.4%
RTN INT4 (g128)	54/252	0.273	1.000 [0.982,1.016]	85.2%	9.3%
RTN INT4 (g64)	104/252	0.266	0.982 [0.913,1.006]	79.8%	7.7%
Wanda 50%	26/252	0.492	1.008 [0.874,1.085]	76.9%	15.4%
Magnitude 50%	0/252	–	–	–	–

(b) Resolution of the retention ratio

Arm	Layer	n	Median ρ	Severe % [95% CI]
Random k sites	Breadth	823	0.556	48.0 [44.6, 51.4]
Transported (identity)	Breadth	823	1.000	5.3 [4.0, 7.1]
Native re-discovery (denominator)	Breadth	823	1.000	0.0 [0.0, 0.5]
Transported, no-op FP16	Sentinel	246	1.000	0.0 [0.0, 1.5]
Transported, RTN INT8	Sentinel	245	0.999	3.3 [1.7, 6.3]

Table 13: Exact breadth-screen Cartesian product.

Axis	Levels
Models	GPT-2 small/medium; Pythia-410M/1.4B; Qwen2.5-0.5B/1.5B; Llama-3.2-1B
Tasks	IOI; Greater-Than; factual recall; assistant-prefix contrast
Explainer → intervention	edge attribution → mean ablation; causal localisation → activation patching; difference in means → additive steering
Transformations	FP16; RTN INT8; RTN INT4 (group 128); RTN INT4 (group 64); NF4; Wanda 50%; magnitude 50%
Matched rule	256 inputs; same top prediction; KL $\leq .05$; at least 16 matched inputs
Seeds	$\{0, 1, 2\}$ for every Cartesian-product cell

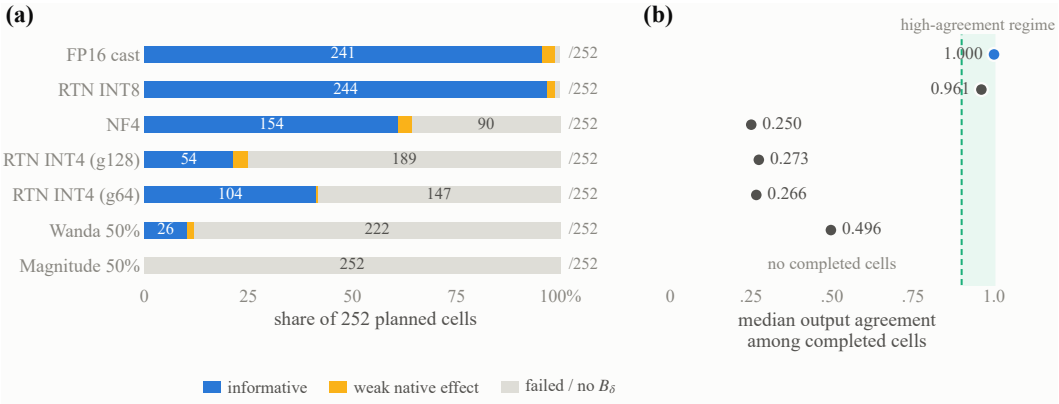


Figure 6: **The matched-set premise is an empirical result.** *(a)* Disposition of all 252 planned cells per breadth-screen transformation. *(b)* Median global top-prediction agreement among completed cells. Four-bit and pruning rows describe selected matched subsets, not globally behaviour-preserving deployments.

D CORRECTED PERSISTENT CONFIGURATIONS

Table 14: All corrected-sentinel configurations with $\rho < 0.8$ at every seed.

Model	Task	Intervention	Transform	ρ_0, ρ_1, ρ_2
Qwen2.5-0.5B	Assistant-prefix	Ablation	RTN INT8	0.168, 0.168, 0.168
Qwen2.5-1.5B	Assistant-prefix	Ablation	RTN INT8	0.178, 0.178, 0.178
Pythia-410M	Assistant-prefix	Ablation	RTN INT8	0.789, 0.789, 0.789

The corrected failure-count histogram is $(75, 2, 0, 3)$ for zero through three degraded seeds. The two sign reversals are Qwen2.5-1.5B Greater-Than attribution at seed 1 ($\rho = -0.0509$, 99.2% agreement) and Llama-3.2-1B Greater-Than steering at seed 1 ($\rho = -1.189$, 88.3% agreement).

E INSTRUCTION AND PACKED-DEPLOYMENT DETAILS

Table 15: **Extensions beyond the base-model grid.** Instruction-tuned Qwen with published harmful prompts (floor 0.311), and official GPTQ/AWQ checkpoints at the strict $\delta = .05$ and the common selected $\delta = .2$. Neither instruction-tuned RTN INT8 nor any official checkpoint produces a severe cell; the low-support Wanda stress test produces one. In (c), the gap is the absolute transported/native refusal-rate effect difference, CI counts 95% paired bootstrap intervals excluding zero, and disagreement is the fraction of paired refusal labels that differ. Detailed intervals and multiplicity correction appear in Table 17.

(a) Instruction-tuned refusal, Qwen2.5-0.5B/1.5B-Instruct

Transform	Complete	Info.	Agreement	Median ρ	OK	Severe
FP16 cast	17/18	17	97.8%	0.999	100.0%	0.0%
RTN INT8	17/18	17	96.0%	0.991	82.4%	0.0%
NF4	9/18	9	8.9%	0.998	100.0%	0.0%
Wanda 50%	6/18	6	7.4%	0.964	83.3%	16.7%

(b) Official GPTQ/AWQ checkpoints

Deployment	Top-1	Match .05/.2	Strict	Selected	Median ρ	OK	Severe
Qwen-0.5B GPTQ	62.1%	2/51	0/6	6/6	1.019	100.0%	0.0%
Qwen-1.5B GPTQ	73.7%	35/114	6/6	6/6	0.981	100.0%	0.0%
Qwen-0.5B AWQ	62.1%	1/45	0/6	6/6	0.834	66.7%	0.0%
Qwen-1.5B AWQ	71.0%	70/133	6/6	6/6	0.988	100.0%	0.0%

(c) 48-token steering, three cells per deployment

Deployment	Prompts/cell	Median |gap|	CI excludes 0	Disagreement
0.5B GPTQ	44–51	0.196	2/3	0.333
1.5B GPTQ	101–114	0.228	3/3	0.320
0.5B AWQ	36–44	0.194	1/3	0.361
1.5B AWQ	129–133	0.000	0/3	0.000

Instruction prompts are normalised from the published `Compumacy/HarmBench-Test-Standard` (Mazeika et al., 2024) and `JailbreakBench/JBB-Behaviors` (Chao et al., 2024) datasets; benign prompts come from `tatsu-lab/alpaca` (Taori et al., 2023). Stable de-duplication is applied before seeded sampling. The exact source revisions, input hashes, and derived prompt hashes are stored in the data manifest. Prompts are formatted using each tokenizer's chat template with an assistant-generation marker.

For packed deployments, let $n_d(\delta)$ be the number of inputs satisfying the matched rule for deployment d. The per-deployment support threshold is

$$\delta_d^* = \min\{\delta \in \{.05, .1, .2, .5, 1, 2\} : n_d(\delta) \geq 16\}, \tag{21}$$

and the common selected threshold is $\delta^* = \max_d \delta_d^* = .2$. We report both $\delta = .05$ and δ^* in Table 15(b).

The first generic GPTQModel AWQ attempt produced NaN logits under the available Torch/SM120 combination for its Marlin, Triton GEMM, and Torch paths. These rows are discarded as invalid runtime results, not interpreted as behavioural collapse. The replacement AutoAWQ 0.2.9 path

produces finite logits and logs `WQLinear_GEMM` modules. Its 0.5B/1.5B global top-1 agreements are 0.621/0.710 and median KL values are 0.271/0.142.

For sequence-level evaluation, with phrase-matching indicator $r(y) \in \{0, 1\}$,

$$\mathrm{Eff}_{\mathrm{seq}}(I, M') = \frac{1}{|\mathcal{B}_\delta|} \sum_{x \in \mathcal{B}_\delta} \left[r(\mathrm{gen}_{48}(I[M'], x)) - r(\mathrm{gen}_{48}(M', x)) \right]. \tag{22}$$

Generation is greedy. Records store labels, lengths, and token hashes. The indicator detects refusal-related phrases in the first 500 decoded characters; it is not a validated semantic refusal classifier. The matched set is selected by the next-token behavioural rule, not by sequence equality.

Table 16: **Sequence-level descriptive effects.** Three seeds per deployment, summarised by median transported/native rate effects, absolute gap, maximum gap, and counts of matching effect signs and exactly equal effects. No deployment has a severe next-token cell; that does not certify a sequence endpoint.

Deployment	n	Transp.	Native	|gap|	Max |gap|	Sign	Exact
Qwen-0.5B GPTQ	3	-0.118	-0.159	0.196	0.205	1/3	0/3
Qwen-1.5B GPTQ	3	-0.178	-0.436	0.228	0.257	2/3	0/3
Qwen-0.5B AWQ	3	0.139	0.244	0.194	0.463	1/3	0/3
Qwen-1.5B AWQ	3	-0.612	-0.612	0.000	0.000	3/3	3/3

Paired uncertainty and discrete resolution. Let b_i, t_i, n_i be the deployed baseline, transported, and native binary labels. The native effect is the mean of $n_i - b_i$; the signed gap is the mean of $g_i = n_i - t_i$, so the shared baseline cancels. We resample paired rows 10,000 times per cell. One prompt changes either mean by $1/N$, here 0.0075–0.0278; four of the 12 native-effect intervals include zero. We therefore avoid sequence retention ratios. These sampling intervals are not a sequence-level no-op floor: the required repeated no-op generations were not recorded, and a logit-scale floor cannot be reused in rate units.

We additionally test paired label symmetry by the exact McNemar test. Conditional on the discordant count D, the null gives a $\mathrm{Binomial}(D, 1/2)$ split between native-only and transported-only refusals. Holm correction across 12 cells permits dependence between tests; its input-wise sampling premise is still conditional on fixed discoveries and the observed matched pool.

Table 17: **Sequence differences with paired uncertainty.** Gap is native minus transported. Brackets give paired 95% bootstrap intervals; disagreement counts pairs with unequal refusal labels. The last column is the exact paired-binary p-value adjusted across all 12 comparisons.

Deployment/seed	N	Native [95% CI]	Gap [95% CI]	Disagree	Holm p
0.5B GPTQ / 0	51	-0.118 [-0.235, -0.020]	-0.196 [-0.333, -0.059]	16/51	0.17
0.5B GPTQ / 1	48	-0.250 [-0.458, -0.042]	+0.167 [+0.021, +0.333]	16/48	0.538
0.5B GPTQ / 2	44	+0.045 [+0.000, +0.114]	+0.205 [-0.023, +0.432]	29/44	0.68
1.5B GPTQ / 0	114	-0.430 [-0.544, -0.316]	+0.228 [+0.149, +0.307]	28/114	2.59e-06
1.5B GPTQ / 1	103	+0.049 [-0.049, +0.146]	+0.223 [+0.126, +0.320]	33/103	0.000662
1.5B GPTQ / 2	101	-0.178 [-0.287, -0.069]	+0.257 [+0.149, +0.366]	36/101	0.000142
0.5B AWQ / 0	44	+0.273 [+0.114, +0.432]	-0.023 [-0.136, +0.091]	7/44	1
0.5B AWQ / 1	41	-0.220 [-0.439, +0.000]	-0.463 [-0.659, -0.268]	25/41	0.00141
0.5B AWQ / 2	36	+0.139 [-0.028, +0.306]	+0.194 [+0.000, +0.389]	13/36	0.554
1.5B AWQ / 0	133	-0.609 [-0.692, -0.526]	+0.000 [+0.000, +0.000]	0/133	1
1.5B AWQ / 1	132	-0.621 [-0.705, -0.538]	+0.000 [+0.000, +0.000]	0/132	1
1.5B AWQ / 2	129	-0.612 [-0.698, -0.527]	+0.000 [+0.000, +0.000]	0/129	1

Six gap intervals exclude zero, but only four cells survive Holm correction at 0.05: three Qwen-1.5B GPTQ and one Qwen-0.5B AWQ. Seeds are not pooled as independent prompt sets. For Qwen-1.5B AWQ, every paired refusal label matches, so a nonparametric bootstrap has the degenerate interval $[0, 0]$. This is not certainty of zero population disagreement: the exact one-sided 95% binomial upper bounds are 2.23–2.30% for the three cell sizes. Moreover, only 5/133, 1/132, and 6/129 stored token hashes match. Equal phrase rates thus leave finer sequence properties unresolved.

Next-token parity and continuation discrepancies. All sequence inputs pass the *unintervened* next-token top-1 matching rule. There are 751 differing dense/deployed baseline continuation hashes

among 976 paired prompt evaluations. Historical hashes include the stored batch-length padding after EOS, so a length difference alone is not proof of differing generated content. Restricting to the 942 equal-length stored pairs still gives 717 hash mismatches. Intervened first-token IDs were not logged, so these records cannot establish per-example parity of the *intervened* next token.

At the deployment/seed level, seven cells have next-token $\rho \in [.9, 1.1]$; four of those have sequence absolute rate gaps above .1. We join by deployment/seed rather than assuming identical prompt membership: for example, one Qwen-0.5B AWQ cell has 45 next-token versus 44 sequence matched prompts. The small set of layer choices and 12 dependent cells does not support a reliable layer-level predictor of these discrepancies. The complete join, sample sizes, layer names, label disagreements, and equal-length hash counts are released in `sequence_crosscheck.csv`.

F Per-Configuration Component Diagnostics

Table 18: **Targeted alignment diagnostics, configuration-level medians.** Ten completed configurations selected by the three-seed degradation rule. Procrustes and linear maps are fitted on calibration activations. The native row is the ratio's own denominator, so its 1.000 is an identity rather than a measured result; it is listed to make the reference point explicit, and the informative comparison is between the fitted maps and identity transport.

Transport	n	**Median ρ**	**OK**	**Severe**
Identity transport	10	0.490	20.0%	50.0%
Orthogonal Procrustes	10	0.345	30.0%	70.0%
Linear ridge map	10	0.349	30.0%	70.0%
Native re-discovery	10	1.000	100.0%	0.0%

Two stored configuration keys retain the legacy `awq4`/`gptq4` names. Both are simulated RTN INT4 (group sizes 64 and 128) and are displayed as such below. This naming is confined to those keys: a regression test rejects any RTN run configuration named GPTQ or AWQ, and no statistic is paired with the wrong transform.

Across the 30 successful diagnostic cells, source/native selections contain 229/232 MLP sites, 39/33 attention heads, and 26/29 residual sites. Source selections are 43% early, 28% middle, and 29% late; native selections are 50%, 26%, and 24%. The result is not a universal late-layer shift.

Table 19: **Per-configuration transport diagnostics.** All ten completed targeted configurations. "Oracle" relocates the source payload to the union/native-selected sites using an oracle reference. This is a diagnostic adaptation, not a guaranteed upper bound or a method without native discovery.

Model	**Task**	**Int.**	**Transform**	**Jacc.**	**Id.**	**Proc.**	**Linear**	**Oracle**
Pythia-1.4B	IOI	Steering	RTN INT4 (g64)	1.000	0.709	0.272	0.317	0.709
Pythia-1.4B	Assistant-prefix	Ablation	NF4	0.714	0.426	-0.744	-0.089	0.535
Pythia-1.4B	Assistant-prefix	Ablation	RTN INT8	0.846	1.160	-0.670	0.380	0.695
Pythia-410M	Assistant-prefix	Ablation	RTN INT8	0.846	0.404	0.418	0.429	0.425
Qwen2.5-0.5B	IOI	Steering	RTN INT4 (g64)	1.000	0.683	0.419	0.093	0.683
GPT-2 medium	IOI	Ablation	RTN INT4 (g128)	0.600	0.073	0.041	0.092	1.134
GPT-2 medium	Assistant-prefix	Ablation	RTN INT4 (g64)	0.500	-3.750	0.938	1.125	0.125
GPT-2 medium	Assistant-prefix	Ablation	NF4	0.412	-0.350	-0.525	-0.450	0.550
GPT-2 medium	Assistant-prefix	Ablation	RTN INT8	1.000	0.909	1.182	1.182	0.909
GPT-2 medium	Assistant-prefix	Patching	RTN INT4 (g64)	0.714	0.554	3.585	4.215	0.554

The instruction-tuned Qwen-0.5B degraded configuration was rerun with a smaller batch to enable component logging; its identity median changed from 0.723 in the grid to 1.045. We treat this as batch-sensitive non-reproduction and exclude it from the component-mechanism claim.

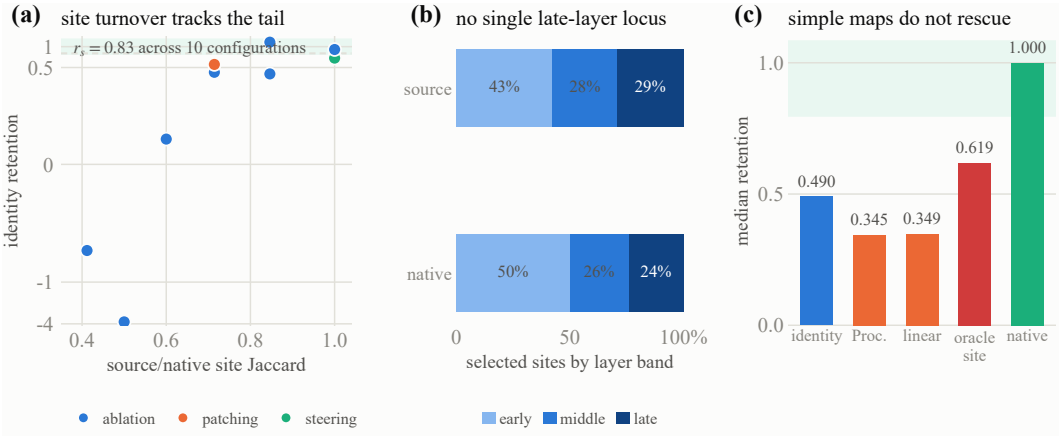


Figure 7: **What targeted diagnostics do and do not explain.** *(a)* Site overlap is associated with identity retention in the selected tail. *(b)* Selections are not confined to late layers. *(c)* Simple activation maps and an oracle site-relocation diagnostic do not recover native efficacy on average.

G Exact Re-Analyses of the Corrected Sentinel

Every analysis in this section is a CPU re-computation over the landed per-example records of the corrected sentinel; no model is re-run. Because a cell's efficacy is stored as the mean of its per-example effects, re-scoring a cell on any stored subset is exact rather than approximate. The re-analysis script first reproduces the published sentinel statistics (245 informative cells, median $\rho = 0.999$, eight severe, seven silent severe among 142, three persistent among 80) and aborts on any discrepancy before computing a new number.

Table 20: **Population-level site turnover in the corrected sentinel.** Exact match means the natively re-discovered site set equals the development set; the FP16 column is the re-discovery noise floor on the no-op cast, and r_s is the Spearman correlation between Jaccard overlap and retention over informative RTN INT8 cells. Discussed in Section 6.

Intervention	n	**FP16 exact**	**RTN8 exact**	**Median Jaccard**	r_s(Jaccard, ρ)
Circuit ablation	83	91.4%	39.8%	0.85	0.11
Causal patching	78	90.1%	65.4%	1.00	0.21
Steering vector	84	100.0%	96.4%	1.00	0.15
All cells	245	93.9%	67.3%	1.00	0.13

Observed turnover rises under RTN INT8: exact re-discovery drops from 93.9% of FP16 cells to 67.3% under RTN INT8, and from 91.4% to 39.8% for circuit ablation. It marks the tail without grading the bulk—every severe cell has Jaccard at most 0.71 while the median informative cell has Jaccard one (rank-sum $p = 1.9 \times 10^{-6}$, permutation $p = 0.002$; the median is exactly one, so more than half the cells tie and the rank-sum statistic is computed with midranks)—yet the population-wide r_s between overlap and retention is only 0.13. Turnover concentrates in late layers (17/30 versus 91/316 overall).

The persistent column below is descriptive: Appendix B.1 shows that the seed is inert in every persistent configuration. Quantile floors are estimated from the transported/native absolute discrepancy in 252 no-op control cells.

The calibration inventory, leave-out fits, nested resampling, and prompt-cluster intervals are reported in Appendix B. They retain the eight original severe cells while making the dependence and uncertainty of the calibration estimator explicit.

The sweep uses the 95 RTN INT8 cells with more than 48 stored attribution scores and selects the top-k sites by absolute score.

Table 21: **Sensitivity to matched-set and efficacy-floor choices.** *(a)* Re-filtering stored examples while holding each intervention fixed. *(b)* Replacing the standard-deviation floor with no-op discrepancy quantiles. Both leave the tail estimate essentially unchanged.

(a) Pre-specified threshold grid

Setting	n	Median ρ	OK	Severe	Silent	Persistent
2σ floor	249	0.999	94.8%	3.6%	7/144	3/83
3σ floor	245	0.999	95.1%	3.3%	7/142	3/80
4σ floor	241	0.999	95.4%	3.3%	7/140	3/78
$\lvert\mathcal{B}_\delta\rvert \geq 8$–$32$	245	0.999	95.1%	3.3%	7/142	3/80
$\lvert\mathcal{B}_\delta\rvert \geq 64$	239	0.999	95.0%	3.3%	7/142	3/77
$\delta = 0.005$	189	1.000	95.2%	4.8%	4/44	2/59
$\delta = 0.01$	218	1.000	94.0%	3.7%	4/62	3/68
$\delta = 0.025$	241	1.000	95.4%	3.3%	7/100	3/79
$\delta = 0.05$	245	0.999	95.1%	3.3%	7/142	3/80
Per-configuration floor	235	0.999	94.9%	3.4%	7/133	3/76
Lowest-KL 50% of $\mathcal{B}_\delta$	241	1.000	96.7%	3.3%	7/142	2/79
Lowest-KL 25% of $\mathcal{B}_\delta$	237	1.000	94.5%	3.8%	7/139	3/75

(b) Floors set from quantiles of the no-op distribution

Floor construction	τ	n	Median ρ	Severe % [95% CI]
90th percentile of no-op $\lvert\Delta\rvert$	0.031	252	0.999	3.6 [1.9, 6.6]
95th percentile of no-op $\lvert\Delta\rvert$	0.055	251	0.999	3.6 [1.9, 6.7]
99th percentile of no-op $\lvert\Delta\rvert$	0.112	249	0.999	3.6 [1.9, 6.7]

Table 22: **Top-k sensitivity for attribution-based cells.** *(a)* Development/native overlap across cuts. *(b)* Relative score gap at the boundary. The severe group has $n = 7$ because the steering reversal has no top-k cut; the groups overlap strongly, and their separation at $k = 12$ does not persist at larger k.

(a) Selection overlap (Jaccard) by cut

Group	n	$k = 6$	$k = 12$	$k = 24$	$k = 48$
Severe	7	0.71	0.71	0.92	0.92
Non-severe	88	1.00	0.85	0.92	0.92

(b) Distance from the boundary

Group	n	Median gap (dev)	Median gap (native)	Tied at boundary
Severe	7	0.018	0.018	3
Non-severe	154	0.026	0.018	32

Site turnover is measured against the same stored explanations. Native re-discovery on the FP16 no-op cast reproduces the development selection exactly in 93.9% of cells, so about 6% of turnover is intrinsic explainer variation; RTN INT8 raises non-exact re-discovery to 32.7%. Turnover sites in severe cells fall in the late layer band in 17 of 30 instances, versus 91 of 316 across all informative cells.

In the cell-level view below, Jacc. is selected-site Jaccard overlap, |gap| is the absolute logit-effect discrepancy, Ex-r_s is the per-example effect Spearman, and slope regresses transported on native effects through the origin. The six assistant-prefix cells are coherent attenuations; the two Greater-Than cells are sign reversals. In part (b), overlap counts flagged cells among these eight.

Which selected sites change. For each severe cell we compare the sites the development selection holds alone with those the native re-discovery holds alone. Qwen2.5-0.5B assistant-prefix ablation trades `mlp_out.22` and `mlp_out.6` for `mlp_out.11` and `mlp_out.21`; Qwen2.5-1.5B trades late sites `head.19.6` and `mlp_out.26` for two layer-0 heads. In the two reversal cells, Qwen2.5-1.5B Greater-Than ablation trades `head.26.6` and `mlp_out.24` for `mlp_out.19` and `mlp_out.26`, while Llama-3.2-1B steering moves `resid.10` to `resid.8`. The two assistant-prefix patterns repeat identically over all three seeds; each reversal pattern is observed at seed 1 only. Across the severe cells, the median layer distance from a dropped site to its nearest replacement

Table 23: **Severe cells and ratio-free tail criteria.** *(a)* All eight severe corrected RTN INT8 cells. *(b)* Ratio-free criteria over the same 245-cell population; their union at a one-logit gap or sign reversal recovers exactly the eight severe cells without additional flags.

(a) Every severe cell

Model	Task	Int.	Seed	ρ	Jacc.	|gap|	Ex-r_s	Slope
Llama-3.2-1B	Greater-Than	Steering	1	-1.189	0.00	1.11	0.15	-0.01
Qwen2.5-1.5B	Greater-Than	Ablation	1	-0.051	0.71	0.76	0.47	0.63
Qwen2.5-0.5B	Assistant-prefix	Ablation	0	0.168	0.71	4.48	1.00	0.20
Qwen2.5-0.5B	Assistant-prefix	Ablation	1	0.168	0.71	4.48	1.00	0.20
Qwen2.5-0.5B	Assistant-prefix	Ablation	2	0.168	0.71	4.48	1.00	0.20
Qwen2.5-1.5B	Assistant-prefix	Ablation	0	0.178	0.71	4.48	1.00	0.20
Qwen2.5-1.5B	Assistant-prefix	Ablation	1	0.178	0.71	4.48	1.00	0.20
Qwen2.5-1.5B	Assistant-prefix	Ablation	2	0.178	0.71	4.48	1.00	0.20

(b) Ratio-free tail criteria

Criterion	n	Flagged	Rate	Overlap
Additive gap $>$ FP16 P95 (0.055 logits)	245	100	40.8%	8/8
Additive gap $>$ 1 logit	245	7	2.9%	7/8
Per-example effect Spearman < 0.5	245	2	0.8%	2/8
Through-origin slope < 0.5	245	7	2.9%	7/8
Sign reversal	245	2	0.8%	2/8

is five and only 26.7% of replacements lie within one layer. This identifies depth-diffuse selection changes, not a demonstrated reallocation of neural computation. The crossed-artefact diagnostic in Appendix K separates a changed reference from changed execution effects.

Table 24: **Original audit registry.** Landed result file, matched rule, efficacy floor, and informative cells. Floors are three control standard deviations except where stated; packed next-token rows use the instruction-derived floor. Sequence effects are in rate units and use paired intervals rather than a logit-unit floor. The new CPU studies have separate records and calibration in Appendix I.

Layer	Result file	Matched rule	Floor	Informative
Breadth screen	`results/cells.jsonl`	δ=.05, top-1, $\geq$16	0.279	823 (241 control)
Corrected sentinel	`sentinel_cells_final.jsonl`	δ=.05, top-1, $\geq$16	0.229	246 FP16, 245 RTN8
Instruction-tuned	`instruction_priority_cells_final.jsonl`	δ=.05, top-1, $\geq$16	0.311	17+17+9+6
Packed strict	`packed_kernel_strict.jsonl` (+AWQ valid)	δ=.05, top-1, $\geq$16	0.311	12 of 24
Packed selected	`packed_kernel_selected.jsonl` (+AWQ valid)	δ=.2, top-1, $\geq$16	0.311	24 of 24
Sequence refusal	`sequence_refusal_*.jsonl`	δ=.2, top-1, $\geq$16	rate effect	12 cells
Targeted diagnostics	`baseline_priority_diagnostics.jsonl`	δ=.05, top-1, $\geq$16	0.28	30 cells, 10 configs
Sentinel re-analysis	`analysis_core/analysis.json`	derived from sentinel	–	re-analysis

H Tasks, Interventions, and Settings at a Glance

The layers are not pooled. Breadth establishes coverage and conditional bulk effects; the corrected sentinel supplies the primary INT8 tail estimate. Instruction, official-checkpoint, and sequence extensions retain their own matched-set coverage and outcome-specific uncertainty. The CPU studies test disjoint discovery and budget sensitivity, without adding a population tail estimate. Table 26 fixes the task scores and interventions; Appendices I and L give the CPU settings. The last two follow-ups reuse the targeted Qwen population. Their measurements are not independent replications or additional sentinel cells. Source fidelity, native parity, scalar score agreement and token identity are reported as separate outcomes.

Shared settings, identical for every cell unless a table states otherwise: 256 evaluation and 128 calibration examples for base models (224 for instruction and packed cells); matched-set rule $\delta = .05$ with top-1 agreement and $|\mathcal{B}_\delta| \geq 16$; seeds $\{0, 1, 2\}$; BF16 CUDA execution with simulated low-bit weights as quantise–dequantise values on a shared forward path. Seeds change task sampling or counterfactual pairing according to the task builder; equal matched row indices do not imply equal inputs. The first 128 clean rows provide transformation calibration when needed; RTN itself is

Table 25: **Evidence ladder.** Planned/completed counts are grid cells or explicitly labelled CPU measurements, not independent datasets. Strict and selected-δ packed evaluations are reported separately rather than pooled.

Layer	Scope	Cells plan/done	Primary purpose
Breadth screen	7 base models, 4 tasks, 3 interventions, 7 simulated transforms, 3 seeds	1,764 / 858	Transformation coverage and bulk median
Corrected sentinel	Same 7 models/tasks/interventions; FP16 and RTN INT8	504 / 504	Confirmed tail, sensitivity, absolute effects
Instruction-tuned	2 Qwen2.5-Instruct models; refusal task; FP16, RTN8, NF4, Wanda	72 / 49	Task validity and instruction tuning
Official checkpoints	GPTQ/AWQ on 2 Qwen-Instruct models; patching and steering	24 / 12 strict; 24 / 24 selected	Deployment-path validation
Sequence + components	12 sequence cells; 33 targeted diagnostic reruns	12 / 12; 33 / 30	Multi-token effects, sites, and alignment maps
Disjoint discovery	GPT-2; 2 tasks, 3 splits, 2 transforms, 4 budgets	96 / 96 measurements	Frozen artefacts on new prompt instances
Prefix budget	Qwen-0.5B; 8 original prompts, 2 transforms, 4 budgets	8 / 8; 1 follow-up	CPU BF16 budget check and historical site sets
Qwen held-out	32 calibration, 96 new-instance and 128 new-template prompts	48 / 48 effects; 64 / 64 sequences	Crossed artefacts and two gate targets
Equal-example discovery	Same Qwen prompts, backend and $k = 12$; corrected gradient accumulation	12 / 12 effects	Partial-batch weighting sensitivity
Frozen-gate sequence	Same 32 fixed prompts; final native gates and native replay	64 / 64 sequences	Exact first-token and continuation comparisons

data-free. The two arms share explainer hyperparameters, and native re-discovery is a reference procedure rather than a guaranteed optimal intervention.

Table 26: **Every task score m and every intervention, in one place.** Each task scores a single next-token logit difference; the sequence-level audit replaces m with a refusal rate over a 48-token greedy continuation. Interventions are the compiled form $\iota(E(\cdot))$ of the explainer to their left.

Task	Score m	Counterfactual
IOI	logit(correct name) − logit(other name)	subject swapped so the answer flips
Greater-Than	logit(plausible later year) − logit(implausible year)	year range altered
Factual recall	logit(true object) − logit(counterfactual object)	subject replaced
Assistant-prefix	logit("I") − logit("Sure"), base models only	benign/stress request pair
Instruction refusal	logit(refusal opener) − logit(compliance opener), chat template applied	harmful vs. benign instruction
Explainer	**Selection rule**	**Compiled intervention**
Component attribution	top-k=12 head/MLP sites by absolute score; single-site ablation scores for at most 256 candidates, otherwise a first-order gradient estimate	mean ablation of the selected set
Causal localisation	top-k=12 residual/MLP-output sites by absolute single-site patch effect	activation patches from the counterfactual run
Difference in means	among residual sites at depth fractions $\{.4, .5, .6, .7, .8\}$, select the one with largest absolute causal effect on m	additive steering along that model's activation-norm-scaled class-mean difference, unit coefficient; the source site/direction are transported unchanged, while the native arm re-estimates both

The result schema stores exact configuration dictionaries, behaviour statistics, transported/native effects, per-example KL/effects where available, selected sites, CKA, top-k overlap, effect-rank correlation, and elapsed time; figures and tables share the same generated summary. Empirical rates should not be extrapolated beyond the model, task, matched-set, and backend regimes reported here.

I Independent Discovery and Executed Selection Budgets

Discovery and evaluation are separated before model scoring. For each of IOI and Greater-Than and seeds $0, 1, 2$, we construct a new task pool, deduplicate unpadded clean token sequences, shuffle the content hashes, and take 32 distinct discovery prompts and 64 disjoint evaluation prompts. Clean-prompt identity, rather than a counterfactual draw or a local row number, defines disjointness. The split precedes behavioural matching and uses no model scores. We fit source and native site rankings and mean-ablation payloads only on the matched discovery partition; all evaluation arms use these frozen artefacts. The two partitions share the original task templates, so this evaluates new prompt instances, not unseen template families.

We execute every $k \in \{6, 12, 24, 48\}$ under FP16 and RTN INT8 for GPT-2. The 156-site candidate pool uses exact single-site ablation scoring. Rankings are reused across k within one model state and matched discovery set, but each joint intervention is executed. The same examples are reused across budgets. Thus the 12 jobs yield 96 measurements over six task/seed splits. Runtime is CPU FP32 with eager attention (Torch 2.13.0+cpu, Transformers 5.14.1), separate from the sentinel's GPU BF16 path. The model's final-logit-only option is checked against ordinary forward output, and cached-baseline effects are checked against the audit's original efficacy routine on an actual intervention arm.

Calibration and uncertainty. At each k, we compute the population SD of the three FP16 native effects in each task/partition group, then multiply the median of these four SDs by three. The resulting floors are .242, .336, .406, and .463 logits. A single floor per k is used for both partitions. These scales include seed-indexed sampling and discovery variation; the four SD estimates are not independent kernel-noise estimates. Ratios are accompanied by 10,000-draw paired-prompt bootstrap intervals for native means and signed gaps. We do not add an epsilon to any denominator. Intervals condition on frozen interventions; they do not include uncertainty from fitting a new explainer on another discovery sample.

Table 27: **Frozen GPT-2 interventions on disjoint evaluation prompts.** RTN INT8 medians over three seeds per row. Each entry is discovery / evaluation. All budgets are executed; 31–32 discovery and 62–64 evaluation prompts survive matching. The final column reports floor eligibility in the two partitions. All per-seed effects, native intervals, gaps, and selections are in the CSV companion; rows are dependent across k.

Task	k	Median ρ	Median absolute gap	Informative
GT	6	1.000 / 1.000	0.000 / 0.000	3/3 / 3/3
GT	12	1.001 / 1.000	0.002 / 0.001	3/3 / 3/3
GT	24	0.999 / 1.000	0.001 / 0.000	3/3 / 3/3
GT	48	0.999 / 0.998	0.002 / 0.004	3/3 / 3/3
IOI	6	0.999 / 1.000	0.002 / 0.002	3/3 / 3/3
IOI	12	0.999 / 1.000	0.008 / 0.012	3/3 / 3/3
IOI	24	0.999 / 0.999	0.005 / 0.005	3/3 / 3/3
IOI	48	0.998 / 0.999	0.008 / 0.012	3/3 / 3/3

All 48 RTN8 measurements are informative. Their minimum absolute native mean is 1.328 logits, and none of the native-effect intervals crosses zero. Retention spans $.982$–1.006, with a maximum absolute gap of $.067$ logits. The discovery and evaluation medians are $.99945$ and $.99959$; the largest change in a paired seed/budget absolute effect gap between partitions is $.0119$ logits. No large native-relative gap emerges on these held-out instances. This does not bound discovery overfitting in the Qwen severe configurations. Across seeds, the RTN8 Greater-Than evaluation partitions have 188 matched prompt evaluations but only 174 distinct prompts; IOI has 191 distinct matched evaluation prompts. We do not treat seed or budget measurements as independent samples. The

shipped manifests record the full grid and source hashes before execution, and a summary is emitted only after all planned jobs and their hashes verify.

I.1 A TARGETED PREFIX SWEEP WITH THE ORIGINAL PROMPT POPULATION

The Qwen2.5-0.5B prefix check executes FP16 and RTN8 at the same four budgets on the eight original distinct prompts, once each, with equal weights. This is a computational deduplication and CPU-backend check, not new held-out data. It retains BF16 parameters and computation, eager attention, and the original explainer's 256-candidate threshold: this 360-site model uses first-order attribution. Batch size is eight rather than the GPU sentinel's 32. FP16 is a round trip of eligible weights through FP16 back to BF16. All eight prompts pass the top-1/KL rule in both arms. We estimate no run-noise floor from a single eight-prompt control and do not add these measurements to the sentinel counts.

The RTN8 $k = 6$ point ratio is negative, but its interval $[-.248, .031]$ includes zero: severity is better resolved than the sign. At $k = 24$ the effect gap falls to $.113$ logits and Jaccard rises to .920, yet a few bootstrap native means are zero, so a finite ratio interval is withheld. At $k = 48$, the raw ratio 35 comes from a $.031$-logit native mean with interval $[-.680, .836]$ and must not be read as effective transport. Increasing the budget changes the native intervention as well as the transported one; improved overlap or a large raw ratio need not mean a better intervention (Figure 8).

Table 28: **Actual Qwen-0.5B budget executions with their controls.** Eight distinct prefix prompts per arm. Brackets are paired-prompt 95% bootstrap intervals, conditional on fixed discoveries. Absolute-gap intervals are mapped from the signed-gap interval. Raw ratios are shown for transparency; † means no finite ratio interval is reported because the native interval crosses zero or a bootstrap denominator is exactly zero. These observations do not establish a calibrated failure rate.

Arm	k	Native effect [CI]	Raw ρ	Absolute gap [CI]
FP16	6	-0.250 [-0.938,0.547]	0.969†	0.008 [0.000,0.059]
FP16	12	1.176 [0.457,2.000]	2.322	1.555 [1.480,1.629]
FP16	24	-0.832 [-1.542,-0.027]	0.999	0.000 [0.000,0.002]
FP16	48	1.281 [0.555,2.078]	0.939	0.078 [0.062,0.102]
RTN8	6	5.809 [5.156,6.559]	-0.101	6.395 [6.258,6.547]
RTN8	12	3.793 [3.168,4.539]	0.514	1.844 [1.734,1.945]
RTN8	24	-0.812 [-1.531,-0.016]	1.139†	0.113 [0.083,0.144]
RTN8	48	0.031 [-0.680,0.836]	35.000†	1.062 [1.031,1.094]

Separating discovery drift from executing a recorded site set. The fresh CPU $k = 12$ ratio, $.514$, differs from the GPU sentinel's $.168$. In a post-hoc follow-up, we freeze the historical GPU source/native site identities and refit their mean payloads on CPU. Their transported and native effects are .916 and 5.406 logits, giving $\rho = .169$ (95% prompt interval $[.040, .275]$). This is not a full replay of the historical payload, but shows that the severe native-reference gap recurs on CPU when the selected sites are held fixed. Fresh discovery is a material source of the discrepancy.

The CPU FP16 control's $k = 12$ gap is itself 1.555 logits. Counting parameter changes shows that the BF16–FP16–BF16 round trip alters 84,593 of 357,826,560 eligible weights, with maximum absolute change 2.98×10^{-8}. It is therefore an approximate control, not an algebraic identity. The two recorded untransformed discoveries have exactly equal score arrays and mean-payload hashes; these records do not support attributing the drift to nondeterminism. Together the control and frozen-site check argue for logging the actual site identities and backend, and for testing controls at the intended budget.

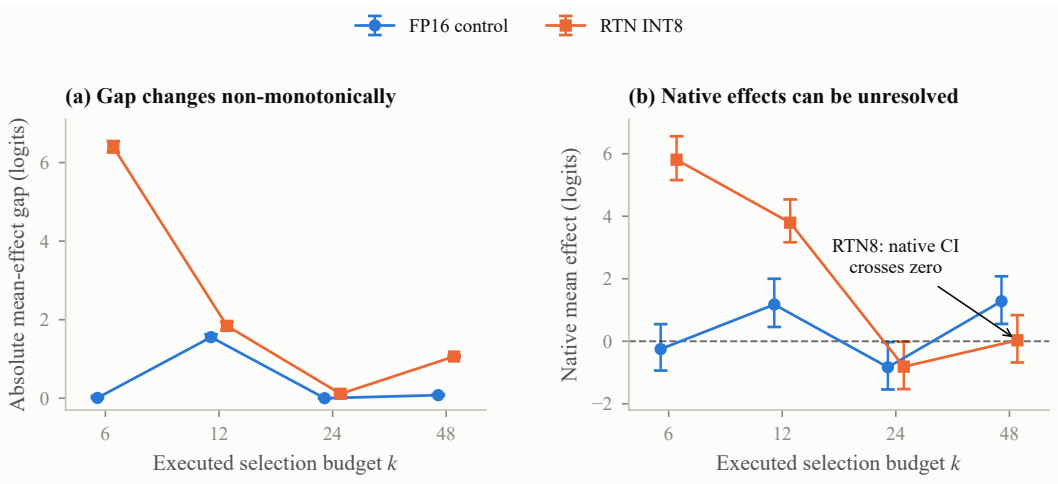


Figure 8: **Selection budget changes both the gap and the denominator.** Executed CPU BF16 measurements on one targeted eight-prompt configuration. Error bars are paired-prompt 95% bootstrap intervals. (a) The gap shrinks at $k = 24$ and grows again at $k = 48$; the FP16 control also drifts at $k = 12$. (b) The RTN8 native mean at $k = 48$ is unresolved, explaining why its raw ratio is misleading. Neither panel estimates population prevalence.

J Matched-Set Composition and Source-Only Screening

We reconstruct $\mathcal{B}_\delta$ from every stored top-1 indicator and KL value and verify exact equality with the recorded matched indices in all 252 completed RTN8 cells. The CSV companion reports total and matched prompt counts, distinct clean prompts, top-1 agreement, coverage, all-input and matched KL quantiles $(0, 25, 50, 75, 90, 95, 99, 100)$, and exclusions due to top-1, KL, or both. Configuration summaries retain the seed grouping.

Table 29: **Severe cells are not concentrated at the KL threshold.** Descriptive medians over the eight severe and 237 other informative cells. For the KL rows, each cell's quantile is computed before the median across cells. Repeated prompts and task composition prevent an independent population comparison. The full cell/configuration distributions are supplied as CSV.

Quantity	Severe (8)	Other informative (237)
Top-1 agreement	1.000	.961
Matched fraction	1.000	.938
Median matched KL	.00344	.00463
95th-percentile matched KL	.00686	.01953
Distinct matched clean prompts	8	45

For an upstream probe, use only source-selected mean relative layer, fraction of MLP sites, and relative score margin at the selection boundary. A fixed $L_2 = 1$ class-balanced logistic model is trained on configuration units, leaving one checkpoint out at a time; all seed cells stay in the same fold. Missing margins are median-imputed from training configurations and accompanied by missingness indicators. The flag threshold is the 90th percentile of training-negative predictions in each fold. Across 83 configurations, the resulting AUROC is .627; nine flags catch one of four severe configurations, with eight false positives. This is a retrospective feature shortlist selected after seeing outcomes, with only four positive configurations. It is a failed screening probe, not a prospective predictor. Native site turnover cannot be an upstream predictor when computing it already requires deployment re-discovery.

Settings when calibration compute is limited. Choose the outcome and practical effect-gap tolerance before fitting or scoring interventions. Select δ using only unintervened behavioural distributions and report a coverage curve, rather than raising it after inspecting retention; .05 is our

reference point, not a universal deployment threshold. Freeze discovery and payload fitting before evaluation. If possible, reuse these frozen artefacts across outcomes and budgets while executing each intervention. Document whether control repeats vary prompts, discovery seeds, or kernels. Two or three mixed repeats cannot isolate kernel nondeterminism or give a precise noise quantile. When a floor cannot be stably estimated, report the paired absolute gap and the native-effect interval as primary outputs, expose ratio eligibility over plausible floors, and mark unstable denominators unresolved. A conservative empirical floor is a reporting convention, not a confidence guarantee; neither the sentinel's .229 logits nor a logit floor in general transfers to a new task, backend, or discrete sequence score.

Machine-readable companions. `results/discovery_holdout/measurements.csv` contains every executed discovery/evaluation arm; `summary.csv` and `partition_changes.csv` give the seed summaries and disjoint-partition contrasts. `results/prefix_topk_cpu/measurements.csv` contains all eight budget/control arms, and `historical_selection_check.json` records the frozen-site follow-up. `results/analysis_review3/` contains `coverage_cells.csv`, `coverage_configurations.csv`, `exception_bounds.csv`, and `sequence_crosscheck.csv`. These provide exact values when PDF table extraction loses minus signs, cells, or mathematical symbols.

K Source Fidelity and Native-Reference Gaps

Re-analysis of the unchanged sentinel. The stored per-example records contain three effects: the source intervention on the development model, that same intervention on the deployment, and native re-discovery on the deployment. We recompute every mean and first recover all published sentinel counts. Native-relative retention ρ and source fidelity ϕ answer different questions, even when evaluated on identical matched prompts. A weak source intervention may transport faithfully while differing greatly from a stronger native selection. The native reference can also have a different sign from the source intervention before transport.

Table 3 in the main text reports all four severe configurations with both references.

Of the 245 native-informative RTN8 cells, 243 also have source-effect magnitude above .229 logits. Their source-fidelity point estimates range from .792 to 1.186 (median .998); none is below .5. This applies the same empirical scale as a descriptive guard, not a new independently calibrated source-noise threshold. The two excluded source means belong to GPT-2 prefix patching seed 1 and Qwen-1.5B Greater-Than ablation seed 1. Source-fidelity uncertainty still matters: the Qwen-0.5B prefix interval is approximately $[.487, .985]$, despite its .894 point estimate, because there are only eight prompt clusters. The Qwen-1.5B prefix and Llama steering intervals are approximately $[.929, 1.250]$ and $[1.018, 1.082]$. These are paired clean-prompt bootstrap intervals with the artefacts fixed. We do not convert the absence of severe source-fidelity point estimates into a general guarantee.

Crossing artefacts with execution models. Let $d = \mathrm{Eff}(I_M, M)$, $u = \mathrm{Eff}(I_M, T(M))$, and $v = \mathrm{Eff}(I_{T(M)}, T(M))$. One additional execution, $b = \mathrm{Eff}(I_{T(M)}, M)$, identifies the exact decomposition

$$\underbrace{v - u}_{\text{native-reference gap}} = \underbrace{b - d}_{\text{gap already on development}} + \underbrace{(v - b) - (u - d)}_{\text{different responses to deployment}} . \tag{23}$$

The same equality holds per example. It requires neither a stable ratio nor additivity across sites. The native artefact is held fixed when executed back on the source; this extra arm is diagnostic, not a new discovery on test data. If the first term dominates, a low ρ mainly records which artefact was selected. If the second dominates, deployment changed the two artefacts' responses differently. Neither pattern alone identifies a neuron-level mechanism or proves computation collapse.

L A Targeted Qwen Hold-Out and Causal-Repair Study

Prompt population and independent fitting. The follow-up uses Qwen2.5-0.5B with CPU BF16, eager attention, and grouped RTN8 at group size 128. Its authored corpus has eight hypothetical request domains, four syntactic templates per domain, and eight contexts per domain, all under

the original `User: ... Assistant:` wrapper and logit("I") minus logit("Sure") endpoint. Within each domain, two templates and two seeded contexts define four calibration prompts. The same templates with the other six contexts give 12 new instances; the other two templates give 16 template-held-out prompts. This yields 32 calibration, 96 new-instance, and 128 new-template prompts. All 256 are distinct and disjoint from the eight original prompts by unpadded token-content hash. Templates are held out from fitting; domains and some contexts are shared. This is a finite, authored stress population, not a random sample of deployment traffic or a safety benchmark.

Historical source/native *site identities* are taken from the first severe GPU Qwen-0.5B cell. Their mean payloads are refitted only on the original eight prompts, then frozen for the new prompts. This extends the previously reported CPU frozen-site check, not a replay of a stored full GPU payload. Under $\delta = .05$ and exact top-1 matching, 8/8 original, 29/32 calibration, 88/96 new-instance, and 117/128 new-template prompts survive. New rankings and payloads are fitted only on the 29 matched calibration prompts. All evaluation arrays retain unmatched prompts too, enabling a declared matching-sensitivity analysis without re-execution or outcome-based prompt selection.

Controls and uncertainty. Repeated execution of the identical source artefact on the eight original and 32 calibration prompts gives exactly equal effect arrays. Unit gates reproduce the ordinary BF16 mean-ablation operator exactly. These check implementation identity for fixed artefacts; they do not estimate discovery or backend noise. We report paired effect gaps, native-effect intervals, and unclipped ratios, with no noise floor inferred from this single run. Primary intervals resample the eight request domains; prompt and template-cluster intervals are supplied as sensitivity analyses. Each draw keeps all intervention arms paired. Cluster sizes remain in the empirical matched-set mean. The 10,000-draw two-stage Bayesian bootstrap (Rubin, 1981) reweights domains and prompts within domains on their observed finite support. Specifically, draw $w \sim \mathrm{Dirichlet}(1, \ldots, 1)$ over domains and independent $z_g \sim \mathrm{Dirichlet}(1, \ldots, 1)$ within each domain, and evaluate $\sum_g w_g n_g \sum_i z_{gi} e_{gi} / \sum_g w_g n_g$ for every arm with the same weights. This finite-support sensitivity is not a hierarchical estimate of unobserved kernel noise and does not replace the sentinel's floor or counts.

Two distinct causal training objectives. At each of the 12 frozen source sites, the repaired operator is $h \mapsto (1 - a_j)h + a_j \bar{h}_j$, where $a_j \in [0, 2]$ and $\bar{h}_j$ is the source mean. Only the 12 scalar gates train; model weights and site identities remain fixed. Gate one equals mean ablation and gate zero equals identity. Float32 interpolation followed by BF16 casting avoids cancellation error in the unit-gate case. Supervision consists of per-prompt intervention effects, not similarity between unperturbed activations.

One arm matches the source intervention's development effects, requiring no native artefact. The other matches effects of the historical native artefact on the deployment. The latter is explicitly an *oracle-supervised* repair: it requires native-reference labels and cannot establish discovery-free causal alignment. Each arm minimises mean squared effect error divided by the larger of one and the calibration target's mean square, plus $.01\,\mathrm{mean}_j(a_j - 1)^2$. Adam uses learning rate .05, six epochs, batch size eight, gradient-norm cap five, and projection to $[0, 2]$. Both use the same seeded minibatch order and the final iterate (24 updates); neither selects an iteration or hyperparameter using test outcomes. We compare both mean-effect parity and per-example MSE, because one can improve without the other. These restricted gates do not implement MAS or search for an orthogonal causal subspace. Their small parameter count is not a demonstrated compute saving over first-order native re-discovery.

Site and payload interventions. The four deployed combinations cross historical source/native site sets with source/native mean payloads. Their two-factor Shapley contributions split the total mean-effect gap symmetrically between sites and payloads, with the interaction reported separately. Additional constant-k arms replace `mlp_out.6` with `mlp_out.11`, and `mlp_out.22` with `mlp_out.21`, one swap at a time. All four site-set combinations use the same native means, allowing the joint-swap interaction to be measured directly. This tests specific intervention choices, not the semantic function of every affected component.

Correcting a partial-gradient-batch weighting defect. The historical first-order explainer back-propagates the mean score of each minibatch and then adds its attribution contributions. Equal-sized

minibatches differ from a prompt mean only by a common factor removed by score normalisation. With a smaller final minibatch, however, its prompts receive more weight. A real tiny-model regression check detects this discrepancy and verifies the equal-example correction: multiply each chunk's score by its row count before the mean backward pass. A separate follow-up uses the same 29/88/117 matched prompts, backend, and $k = 12$, changing only this accumulation convention. Historical source and result files remain intact. The primary historical prefix configurations have full repeated batches and are unaffected by this particular weighting defect; broader first-order cells with partial batches retain their documented historical selection rule. Exact single-site scoring in the GPT-2 hold-out experiment is unaffected.

Matching sensitivity and direct sequence joins. The declared fixed-artifact sweep uses $\delta \in \{.001, .002, .005, .01, .02, .05, .1, .2\}$ with and without top-1 matching. The KL-only arm changes the behavioural premise and is reported separately. An adaptive threshold uses only the calibration set's top-1-matched KL 95th percentile, before test efficacy is scored. It is not tuned to retain or remove severe outcomes.

L.1 Held-out estimates and crossed effects

Table 30: **Native parity and source fidelity on disjoint Qwen prompts.** Historical site identities are frozen, with CPU means fitted only on the original eight prompts. d, u, v are the source, transported and native effects, in logits. Brackets are paired 95% domain-cluster intervals (eight domains); the original row instead resamples its eight distinct prompts. Calibration rows are fitting data; only the last two rows measure new-prompt performance. † retains the point estimate but withholds its ratio interval because the corresponding reference-mean interval includes zero.

Partition (matched)	d	u	v	$\rho = u/v$	$\phi = u/d$
Original (8)	.979	.916	5.406	0.169 [0.041, 0.278]	.936†
Calibration (29)	.468	.431	4.931	0.087 [−0.012, 0.174]	.921†
New instances (88)	.557	.482	4.979	0.097 [0.009, 0.172]	0.865 [0.263, 0.973]
New templates (117)	.550	.505	5.002	0.101 [−0.019, 0.193]	.918†

The template-held-out source mean is .550 logits, with domain-cluster interval $[-.046, 1.111]$; its source-fidelity ratio interval is therefore unresolved. The paired source execution change is still estimable: $u - d = -.045$ logits, with interval $[-.077, -.014]$. It is small beside the 4.497-logit native-reference gap, but it is not zero. The source-fidelity point alone must not be described as a confidence guarantee.

Table 31: **Crossed effects localise the reference gap.** All entries are paired means and 95% domain-cluster intervals in logits. The first two rows sum to the total native-reference gap by Equation 23. The last two instead partition that same gap by changing sites and payloads; these are distinct decompositions, not five additive terms.

Term	**New instances**	**New templates**
Gap on development	4.465 [4.457, 4.471]	4.481 [4.471, 4.490]
Deployment interaction	0.033 [0.023, 0.041]	0.016 [0.004, 0.029]
Source-artefact execution change	−0.075 [−0.123, −0.024]	−0.045 [−0.077, −0.014]
Site contribution	4.608 [4.596, 4.620]	4.618 [4.607, 4.632]
Payload contribution	−0.111 [−0.116, −0.106]	−0.120 [−0.128, −0.113]

L.2 Fixed-site repair and corrected discovery

Changing the target changes what counts as repair. On new templates, the native-supervised gates reduce native-target MSE from 20.228 to .186 logits^2, but increase error against the original source effect from .027 to 16.349. The source-supervised gates change their own target error from .026970 to .026367; the paired improvement interval $[-.003989, .005539]$ includes zero. Thus we do not claim an established source-fidelity gain or describe native-target fitting as restoration of the original intervention.

Table 32: **Mean parity and per-prompt error need not agree.** All arms use the same historical native reference on RTN8. ρ includes its paired 95% domain-cluster interval; MSE is the mean squared per-prompt discrepancy in logits^2. Means-only and sites-only arms are controlled substitutions, not learned repairs. Gate fitting uses calibration prompts only and reports the final iterate, without test-based selection. The native-target arm requires oracle native intervention labels.

	New instances		New templates	
Transport arm	ρ	**MSE**	ρ	**MSE**
Unchanged transport	$0.097\ [0.009, 0.172]$	20.23	$0.101\ [-0.019, 0.193]$	20.23
Refit means only	$0.086\ [-0.008, 0.165]$	20.74	$0.087\ [-0.031, 0.182]$	20.85
Change sites only	$1.033\ [1.030, 1.037]$	.03	$1.034\ [1.030, 1.040]$	.03
Source-effect gates	$0.113\ [0.025, 0.188]$	19.49	$0.117\ [-0.003, 0.209]$	19.50
Native-effect gates (oracle)	$0.923\ [0.910, 0.935]$	.17	$0.917\ [0.895, 0.933]$	.19

Specific site substitutions. The historical site sets differ in two MLP-output positions. On the 88 matched new-instance prompts, source sites with native means give mean effect .426. Replacing site 6 with 11 alone gives .231, replacing 22 with 21 alone gives 4.388, and making both replacements gives 4.979 logits. Their joint-swap interaction is .787 (95% domain-cluster interval [.770, .802]). These constant-k executions localise the sensitivity to particular site choices and their interaction; they do not identify a semantic function or redistribution between attention and MLP computation.

The scalar response has little remaining variation. Table 33 reports standard deviations of the unintervened score and the scores after the two joint ablations. For new templates, the RTN8 baseline varies by 1.071 logits, whereas source/native intervened scores vary by only .025/.039. This flattening already occurs on the development model (.024/.037). The intervention-effect gap consequently has SD .043 logits.

A prediction $\hat{v} = u + c$, with $c = 4.5$ fitted only on the 29 calibration prompts, has new-template native-target MSE .001840, compared with .186089 for the physically executed native-supervised gates. This is a prediction diagnostic, not an executed intervention or a transport-repair baseline. To quantify the offset contribution, write $g_i = v_i - u_i$ for the paired effect gap and decompose its empirical squared error exactly:

$$\frac{1}{n}\sum_i g_i^2 = \bar{g}^{\,2} + \frac{1}{n}\sum_i (g_i - \bar{g})^2. \tag{24}$$

For the 117 matched new-template prompts these terms are $20.227802 = 20.225969 + 0.001833$ logits^2. Squared mean bias accounts for 99.9909% of the raw discrepancy. The calibration-only offset prediction removes 99.9909% of raw MSE; executed native-target gates remove 99.0800%. These are descriptive error decompositions on held-out prompts, not fractions of a learned causal mechanism recovered by the gates. This explains why fitting mean parity here does not establish recovery of rich per-prompt causal structure: much of the target discrepancy is an offset between nearly constant intervened endpoints. This narrow joint-ablation effect does not establish unperturbed computation collapse under quantisation.

Table 33: **Joint ablations flatten this scalar endpoint on both models.** Entries are descriptive standard deviations across matched prompts, in logits. The first six rows concern scores; the final row concerns the paired difference between native and transported effects. These are not uncertainty intervals or measures of the unperturbed network's total representational variation.

Endpoint	**New instances**	**New templates**
Unintervened development	.812	1.059
Unintervened RTN8	.861	1.071
Source artefact / development	.022	.024
Native artefact / development	.034	.037
Source artefact / RTN8	.025	.025
Native artefact / RTN8	.037	.039
Native minus transported effect	.041	.043

Table 34: **Each trained operator is checked against both causal targets.** MSE compares per-prompt effects with the indicated evaluation target. Positive reduction means improvement over unchanged transport; negative reduction means worse fitting. Brackets are paired domain-cluster intervals on the loss difference, preserving covariance between the two operators. All quantities are in logits^2.

Training target	Evaluation target	Raw MSE	Gate MSE	Reduction [95%]
New instances (88 matched prompts)				
Source	Source	.029	.021	$0.007\ [-0.001, 0.015]$
Native	Source	.029	16.337	$-16.308\ [-16.823, -15.849]$
Source	Native	20.228	19.495	$0.733\ [0.672, 0.784]$
Native	Native	20.228	.166	$20.062\ [19.971, 20.147]$
New templates (117 matched prompts)				
Source	Source	.027	.026	$0.001\ [-0.004, 0.006]$
Native	Source	.027	16.349	$-16.322\ [-16.786, -15.866]$
Source	Native	20.228	19.501	$0.727\ [0.627, 0.830]$
Native	Native	20.228	.186	$20.042\ [19.919, 20.151]$

The equal-example correction gives held-out native-relative ratios -3.348 and -3.314, whereas source-fidelity ratios are 1.000 and 1.003. For new templates the transported mean is 4.584 logits and the newly discovered native mean is -1.383, with reference interval $[-1.970, -.811]$. This is a resolved native-relative sign disagreement within the authored pool. Executing these fresh artefacts on both models yields a development reference gap of -5.893 logits and deployment interaction $-.074$; the source artefact's own execution change is $+.012$ logits. The large opposite-sign native reference is therefore not evidence of the source effect reversing in transit. These same-pool sensitivity observations are reported separately and do not change the historical sentinel counts.

Table 35: **A controlled change to fresh discovery.** Both source and native artefacts are newly fitted on the same 29 calibration prompts. The equal-example rule corrects only partial-minibatch attribution weighting; it does not change models, matching or $k = 12$. The absolute mean-effect gap and source fidelity refer to the corrected run. Brackets are paired domain-cluster intervals. These observations reuse the same prompt pool and are not an additional replication.

Partition	Chunk-mean ρ	Equal-example ρ	Gap (logits)	Corrected ϕ
Calibration (29)	$2.087\ [1.747, 2.595]$	$-3.200\ [-5.363, -2.111]$	6.136	$1.006\ [0.996, 1.018]$
New instances (88)	$2.018\ [1.690, 2.517]$	$-3.348\ [-5.498, -2.290]$	6.122	$1.000\ [0.990, 1.011]$
New templates (117)	$1.811\ [1.555, 2.118]$	$-3.314\ [-6.385, -2.033]$	5.967	$1.003\ [0.996, 1.009]$

L.3 Matching sensitivity and exact sequence pairs

Table 36: **Reweighting the same frozen historical artefacts.** The calibration-adaptive threshold is .01238, fixed without evaluating test efficacy. Gap is the absolute mean transported–native difference in logits. The KL-only row relaxes the behavioural premise. These point estimates describe matching sensitivity; no new artefacts or native-effect floors are fitted.

	New instances			New templates		
Matching rule	n	ρ	**Gap**	n	ρ	**Gap**
$\delta = .005$, top-1	21	.155	4.503	29	.166	4.487
$\delta = .01$, top-1	69	.097	4.495	85	.115	4.497
Calibration-adaptive, top-1	78	.103	4.495	101	.107	4.497
$\delta = .02$, top-1	86	.099	4.496	112	.105	4.497
$\delta = .05$, top-1	88	.097	4.497	117	.101	4.497
$\delta = .05$, KL only	96	.102	4.497	128	.099	4.496

For sequence checking, we preselect two new instances and two new-template prompts from each domain by prompt ID: 32 prompt pairs. Both historical artefacts execute on the same RTN8 model with greedy decoding, one beam, 48 new tokens, dynamic KV caching, and batches of eight. Records join prompt content hashes, intervened prefill argmax IDs, and exact generated token IDs. Tokens

after the first EOS are removed before comparison; the EOS token itself is retained. All 32 pairs pass the behaviour-matching rule, but all differ at the first intervened token. They therefore provide no example of equal intervened first tokens followed by unequal continuations.

After observing that result and the gate pilot, we specify one further test: execute the already fitted native-supervised gates on exactly the same preselected 32 prompts, alongside a fresh historical-native replay. This follow-up involves no refitting, checkpoint selection, prompt selection or decoding search. It is declared after the parent study, not part of the parent's pre-execution plan and not an independent replication. It additionally stores scalar prefill effects and checks the native replay against the original exact token sequences. Table 37 reports both comparisons regardless of whether equal-first/later-divergent pairs occur. Neither run retroactively provides missing first-token IDs for the official GPTQ/AWQ data.

Table 37: **Exact within-prompt sequence comparisons on RTN8.** Columns count prompt pairs, not independent interventions. "Equal first, different rest" counts identical first generated tokens followed by unequal full continuations. The first two rows use their behaviour-matched subsets of the same preselected pool; the third checks all 32 native replays against the parent study. The final column checks each arm's intervened prefill argmax against its actual first generated token. This is a separate configuration from the official GPTQ/AWQ sequence study.

Comparison	n	Equal first token	Different continuation	Equal first, different rest	Prefill mismatch
Transport / native	32	0	32	0	0
Native gates / native	32	27	12	7	0
Native replay / original	32	32	0	0	0

For the frozen-gate follow-up, all 32 prompts remain behaviour matched. Scalar retention is .919 with paired domain-cluster interval $[.905, .929]$; the signed gate-minus-native gap is $-.402$ logits and MSE is .176. The gates share the native first token on 27 prompts, but 12 continuations differ, including seven whose first token agrees. All intervened prefill argmax IDs match the actual first generated tokens. Native replay reproduces all 32 original continuations exactly.

Fresh sequence-batch baselines differ from the parent's effect-batch baselines by at most .0625 logits. Each new effect pair uses its own common, freshly executed baseline; we do not pool these baselines or describe the two batch contexts as numerically identical. Exact replay of the native continuations is a distinct check from numerical equality of scalar baselines.

L.4 Minimal execution contract

An audit record should identify: model and tokenizer hashes; transformed modules, group layout and rounding rule; parameter, accumulation and activation dtypes; framework, compiler, attention and quantisation kernels; device and thread settings; padded token IDs, masks and position construction; discovery batch order and partial-batch weighting; selected sites, payload hashes and explainer ties; and fitting/evaluation partitions. The generation contract also records decoding policy, cache implementation, EOS/padding treatment, batch composition and scheduler, with random seeds when sampling is used. A seed alone does not fix these choices.

Keep three checks separate: repeating an identical frozen artefact checks execution; re-fitting on unchanged prompts checks discovery; and changing prompts checks sampling sensitivity. FP16 casting of BF16 weights is an approximate transformation, not an algebraic identity. Our new fixed-artefact control is identity execution. The released manifests record the actual backend rather than promising equality on another one. When a backend changes, rerun the control and state any resulting site, payload, denominator or token changes before interpreting a retention difference.

M Threats to Validity

Table 38 lists every threat we are aware of, the evidence that addresses it, and the limit that survives. The padding/position defect has a direct implementation correction: it was a bug with a regression test, not a scope limit. Two residual limits directly bound the headline rates. First, the rates are conditional

on the matched set; under aggressive transforms that set is a small, self-selected minority, so those rows remain stress tests and the FP16/RTN INT8 rows provide the estimates. Second, every rate is specific to the audited explainer configuration, including its top-k cut, whose selection signature does not persist across cuts (Section 6.2). Neither limit affects Proposition 1, which is a possibility result, but both constrain how far an empirical prevalence may be extrapolated.

Additional compression families perturb different objects: activation quantisation changes forward-pass activations (Xiao et al., 2023; Yao et al., 2022); rotation and incoherence methods alter the coordinate system (Chee et al., 2023; Ashkboos et al., 2024; Heo et al., 2024); learned rounding changes the rounding objective (Arai & Ichikawa, 2025; Li & Panda, 2025). Calibration sensitivity (Paglieri et al., 2024) and microscaling formats (Zhang et al., 2026) are further reasons not to extrapolate the RTN findings.

Table 38: **Threats to validity and what the evidence establishes.** The right column states the residual limit rather than implying that an added experiment removes it.

Threat	Evidence added	Residual scope
Padding/position handling	Complete 504-cell corrected sentinel; regression test	Seven-transform breadth rows remain exploratory
Simulated quantisation	Official GPTQ/AWQ checkpoints; AWQ `WQLinear_GEMM`	GPTQ class probe is not kernel-class confirmation; no activation quantisation
Matched-set selection	Full cell disposition; strict and selected packed thresholds both shown; exact lowest-KL tightening	Aggressive rows remain conditional on surviving inputs
Base-model safety proxy	Two instruction-tuned models and published harmful prompts	Only Qwen2.5 at 0.5B/1.5B; next-token score is still primary
Single-token tasks	48-token deterministic generation audit	Phrase-based refusal classifier and 36–133 matched prompts
Effect-bound coverage	Exception-budget bound certifies sign in 166/245 cells, including six severe attenuations	Finite-sample diagnostic; no sign certificate for the two reversal point estimates
Repeated prompts / adaptive discovery	Prompt-cluster bootstrap; disjoint GPT-2 splits; 205 matched new Qwen prompts with separate fitting	One Qwen configuration, authored domains, CPU-refitted payloads; not a population tail rate
Calibration	Exact repeat counts, leave-out fits, nested sensitivity	Only 2–3 runs per group; kernel noise not separately identified
Selection and repair	Executed k sweeps; crossed artefacts and site/payload arms; two effect-supervised gate objectives	Restricted pilot, not MAS; native-target training needs oracle labels; no broad GPU budget comparison
Partial attribution batches	Equal-prompt correction and paired fresh discovery follow-up; actual-model regression test	Historical selections stay fixed; BF16 arithmetic and other backends may still change rankings
Reference choice	Source-fidelity re-analysis and both artefacts on both execution models	Native-relative severity does not identify source-effect loss or a particular neural mechanism
Model scale	Seven 0.1–1.5B base models plus two instruct models	Results do not establish behaviour for frontier-scale or architecture-changing deployment