AUDITING APPLIED DOSE IN MULTI-LAYER SUBSPACE ABLATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Random-subspace controls for multi-layer ablations can be matched to the treatment on clean activations, yet each later deletion acts on a state already changed by earlier ones. We audit this comparison for rank-4 contrastive principal component analysis (cPCA) subspaces deleted at the final token of the upper sixteen blocks of three 7–9-billion-parameter models on ten MMLU domains, separate clean-state projection energy, the realized bfloat16 (BF16) write displacement and the change in state norm, and compare zero ablation, also called directional ablation, with mean ablation. First, controls calibrated on clean activations receive 1.7–$2.7\times$ the treatment's relative root-mean-square (RMS) perturbation inside the intervened pass, and control damage rises steeply with dose; calibrating on the intervened trajectory brings the median ratio to 1.08–1.12, a forward sequential procedure calibrates all 5 fixed-frame profiles where a simultaneous update calibrates none, and per-write matching places 99.9% of 40,000 BF16 control writes within 1% of the treatment's. Second, the reference value of a deletion matters beyond its dose. A cPCA frame is fitted to centered covariances, but zero ablation $h - Ph$ also removes the mean component $P\mu$. In a registered experiment on 972 questions held out from fitting, with frames and means refitted for each of five splits and 25,000 cases per model, mean ablation $h - P(h - \mu)$ restores 11.3–13.1 of the 13.5–15.8 accuracy points that zero ablation costs while writing 3–18% of its squared displacement; rescaled per item and layer to the zero-ablation displacement, the mean-ablation direction still lowers choice loss by 0.24–0.81 nats in all three models and all fifteen model–split replicates, with every failed match retained. Registered extensions reproduce this direction effect in Mistral-7B-v0.3 and Qwen3-8B, on 407 MMLU questions outside the examined pool, and with a reference mean taken in the fitting format.

1 INTRODUCTION

A low-rank subspace can predict an input property without being necessary for the corresponding behavior. Intervention studies therefore delete the associated directions and compare the change with a random deletion of equal rank. At many layers this comparison is delicate: an earlier intervention changes the states on which later ones act, so equal rank, equal projection energy on clean activations and equal perturbation in the modified forward pass are three different controls.

We examine this distinction for cPCA domain subspaces on ten MMLU domains in Qwen3.5-9B, Olmo-3-1025-7B and Llama-3.1-8B. For each extraction domain D we delete its rank-4 subspace at the final prompt token of the upper sixteen blocks and evaluate all ten domains E, a crossed design of the kind used for role directions by Potertì et al. (2025).

Clean-activation matching does not survive the intervention. Controls matched on unmodified states receive 1.7–$2.7\times$ the treatment's applied relative RMS perturbation, because the treatment's own removal is depleted by upstream deletions while a redrawn control frame is not, and the damage of a control rises steeply with its dose. Calibrating inside the intervened trajectory brings the median ratio to 1.08–1.12. Relative RMS divides by an intervention-dependent state norm, so we also measure the actual BF16 before–after write of every site: on a fixed 500-item row, 39,963 of 40,000 control writes meet a 1% displacement tolerance, and all 37 misses remain in the analysis.

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.

Sequential calibration handles fixed frames. A random frame held fixed across layers is the natural geometry control for a layer-specific treatment, but simultaneous calibration matched none of 200 such controls. Calibrating the dose of the random comparator layer by layer on the state produced by the already-fixed earlier layers, where layer-wise eraser fitting (Belrose et al., 2023) fits the eraser itself, matches all five profiles that the simultaneous update misses, and a common rotation of the treatment's sixteen-layer frame keeps its cross-layer geometry while matching all 61,600 writes.

The reference value of a deletion matters beyond its dose. A cPCA frame is fitted to centered covariances and says nothing about the mean of the state in its span, yet zero ablation removes that mean together with the centered variation: $Ph = P(h - \mu) + P\mu$. Replacing the deleted component by a reference mean, $h - P(h - \mu) = (I - P)h + P\mu$, is mean ablation (Wang et al., 2023; Li & Janson, 2024; Belrose et al., 2023; Marshall et al., 2024). Its advantage has been argued for (Li & Janson, 2024; Marshall et al., 2024), but mean ablation also writes a much smaller displacement, so we ask whether the advantage survives a control for dose. On a 77-question development panel it recovers 10.1 of the 10.4 accuracy points lost to zero ablation while writing one fifth of the squared displacement; rescaled per item and layer to the realized write energy of zero ablation, the mean-ablation direction is still less damaging, by 0.297 nats $[0.189, 0.407]$. A registered held-out experiment with frames and means refitted from scratch reproduces this in all three models and in every model–split replicate, by 0.239, 0.778 and 0.812 nats on 25,000 cases per model. Registered extensions reproduce it in Mistral-7B-v0.3 and Qwen3-8B, on 407 questions outside the examined pool and with a fitting-format reference mean.

Contributions. For multi-layer deletions of cPCA frames (Abid et al., 2018) by zero ablation, also called directional ablation (Arditi et al., 2024), and by mean ablation, we contribute a measurement of the 1.7–$2.7\times$ overdose of random controls under clean-state matching, with trajectory calibration to 1.08–$1.12\times$ and a forward sequential procedure for fixed frames; per-write accounting of projection energy, realized BF16 displacement and state-norm change, with every failed match retained; a direction-by-displacement factorial showing that the collateral damage of zero ablation of a cPCA frame is partly directional, not only a matter of dose, in a registered held-out replication on three models and registered extensions to two further model families, to questions outside the examined pool and to a fitting-format reference mean; and a registered answer-format positive control that the calibrated crossed assay detects with a selectivity contrast of $+1.23$ nats in choice cross-entropy.

2 Related work

Deletion operators and their reference value. Projection deletion removes a subspace from representations (Ravfogel et al., 2020; Elazar et al., 2021; Zou et al., 2023); directional ablation removes a fitted direction from every residual-stream activation, $h - \hat{r}\hat{r}^\top h$ (Arditi et al., 2024). Mean, optimal and affine replacements (Wang et al., 2023; Li & Janson, 2024; Singh et al., 2024; Marshall et al., 2024), including the least-squares concept erasure (LEACE) eraser $x - W^+PW(x - \mathbb{E}[X])$ (Belrose et al., 2023), substitute a reference value for the deleted component. Our direction-by-displacement factorial compares the two reference values at matched realized displacement, which separates the advantage of mean ablation from its five-fold smaller write.

Sequential fitting, dose and norm. Concept scrubbing fits each layer's eraser on already-erased states (Belrose et al., 2023); our sequential calibration fits the dose of random comparators on the intervened trajectory, which clean-state matching overdoses 1.7–$2.7\times$. Additive steering (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023) and rotation-based steering (Vu & Nguyen, 2025; Dang & Ngo, 2026; Aparin & Gaintseva, 2026) concern added directions; for deletion we match realized displacement and report norm change separately, since matched controls raise the norm in most writes. Random-head controls matched on output-weight norm (Pandey, 2026) fix a write scale per component; we match the realized write of each item and layer inside the intervened trajectory.

Validity of interventions. Control tasks and probing critiques separate what a representation encodes from what the model uses (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022), which motivates deletion with random comparators. Subspace interventions can act through dormant pathways (Makelov et al., 2024), a reading that Wu et al. (2024) contest, or leave the natural distribution (Grant et al., 2026), and conclusions depend on the ablation method and metric (Zhang

& Nanda, 2024; Heimersheim & Nanda, 2024; Miller et al., 2024). Self-repair and resistance to steering are documented (McGrath et al., 2023; Rushing & Nanda, 2024; McKenzie et al., 2026). Concurrent work argues for positive controls and power before trusting an interpretability verdict (Reblitz-Richardson, 2026); our crossed assay includes a registered positive control.

Crossed domain designs and positive controls. Domain structure has been probed and ablated (Garcia et al., 2025; Saglam et al., 2025; AlKhamissi et al., 2025); Potertì et al. (2025) cross role directions with evaluation domains under addition and ablation; our crossed assay adds trajectory-calibrated controls. AxBench, the Mechanistic Interpretability Benchmark and the RAVEL disentanglement benchmark supply positive cases for interpretability methods (Wu et al., 2025; Mueller et al., 2025; Huang et al., 2024). Answer-letter selection has identifiable circuitry (Wiegreffe et al., 2025) and surface-form competition affects likelihood (Holtzman et al., 2021), so we separate choice CE from answer-set mass.

3 SETUP, ESTIMANDS AND DOSE ACCOUNTING

Models, data, extraction. We use the post-trained Qwen3.5-9B and the base models Olmo-3-1025-7B and Llama-3.1-8B; all have 32 blocks and width 4096, and we intervene at blocks 16–31. Appendix D gives the complete model and protocol specifications. For each of ten MMLU domains (Hendrycks et al., 2021) the first 100 test questions are split 50/50 by a seeded permutation; the extraction half provides zero-shot prompts consisting of the question, the lettered options and "Answer:", whose last-token block outputs define the domain set, the other nine domains' extraction halves form a contrast pool, and the evaluation half is scored with a fixed five-shot prompt. The treatment basis $U_D \in \mathbb{R}^{4\times 4096}$ at each block is the top-4 eigenspace of $\Sigma_D - \Sigma_{\bar{D}}$, the cPCA rule of Abid et al. (2018) without identity shift, fitted independently per split seed. Seeds 42, 101, 202, 303 and 404 are different splits and hence different fits; the same seed gives the same held-out items for every model and run, which we use for paired comparisons.

Intervention and scoring. At the final prompt token of every intervened block we apply $h \leftarrow h - \alpha\, U^\top U h$, with α=1 for the treatment, sequentially so that later blocks see the intervened stream. The scored continuation is the whitespace-prefixed answer letter, a single token in all three tokenizers; the unspaced letter is a token assigned negligible probability in the checked prompts, with a baseline negative log-likelihood (NLL) of 15–19 nats as described in Appendix E. Per item we record the full-vocabulary NLL of the correct letter, the four-way choice cross-entropy (CE), correctness, and the answer-set mass $M = \sum_{c\in\{A,B,C,D\}} p(c)$, for the baseline, the treatment and each control. The identity $\Delta L_{\text{full}} = \Delta L_{\text{choice}} + \log(M_0/M_I)$ holds per item with an algebraic residual of at most 2.3×10^{-16}. Choice CE was fixed as the primary endpoint before the calibrated runs were inspected, with accuracy drop as the secondary endpoint, as recorded in Appendix D.

Estimands. Let δ^T_{DEi} be the damage, that is, the endpoint increase, on item i of evaluation domain E when the subspace of extraction domain D is removed, and δ^C_{DEci} the same under control c of row D. Then

$$\begin{aligned} \tau_{DE} &= \mathbb{E}_i\, \delta^T_{DEi} - \mathbb{E}_{c,i}\, \delta^C_{DEci}, \\ S_D &= \tau_{DD} - \tfrac{1}{9}\textstyle\sum_{E\neq D} \tau_{DE}, \qquad S = \tfrac{1}{10}\textstyle\sum_D S_D. \end{aligned} \tag{1}$$

S is the interaction contrast of a balanced two-way layout: additive row potency a_D and column fragility b_E cancel from it exactly, so a positive S requires diagonal cells to be damaged beyond what the intervention's general potency and the evaluation domain's general fragility predict. Every registered test of S in this project is listed in Appendix G.

Controls. *Isotropic* (iso): a Haar-random rank-4 frame. *Covariance-aware* (cov): a random rank-4 frame inside the top-16 PCA span of a random half of the contrast pool, chosen among 48 candidates as the one whose calibration energy is the smallest at or above the treatment's, so that $\alpha = \sqrt{E_T/E_C} \leq 1$ is a pure removal; if none reaches it, the largest is taken and the initial scale before online calibration is $\alpha \in (1, 1.36]$, which happens in 178 of the 5,600 layer–control initializations of the seven primary matrices. There is no readout orthogonalization. *Persistent*: one such frame drawn at block 16 and applied at every intervened block. Five controls per extraction

Table 1: What each comparator holds fixed relative to the layer-specific cPCA treatment. "Matched on" is the calibrated quantity: online calibration matches per block on 32 calibration prompts, actual-write matching per item and layer. Table 2 evaluates the online runs by relative RMS on evaluation items, a different summary. Norm increase is the fraction of writes that increase $\|h\|^2$; the treatment increases it in 274 of 12,320 writes on the same panel.

Comparator	Rank	Covariance support	Cross-layer geometry	Matched on	Norm increase
Isotropic	yes	no	no	nominal update energy, online	—
Covariance-aware	yes	top-16 PCA span	no	nominal update energy, online	—
Persistent frame	yes	at block 16	one frame, all layers	nominal update energy, online	—
Common rotation	yes	no	yes, exact Gram	actual write, per item	57.2%
Independent rotations	yes	no	no	actual write, per item	78.7%

row, redrawn at each block for iso and cov, are applied to all ten evaluation columns. Two later geometry controls transform the treatment's own sixteen-layer frame: a *common rotation* applies one orthogonal rotation to all layers and preserves every cross-layer inner product, while *independent rotations* rotate each layer separately. Table 1 lists what each comparator holds fixed.

Deletion operators. The treatment is zero ablation of the subspace, $h' = h - Ph$ with $P = U^\top U$. Mean ablation replaces the subspace component by that of a reference mean, $h' = h - P(h - \mu) = (I - P)h + P\mu$ (Wang et al., 2023; Li & Janson, 2024; Belrose et al., 2023; Marshall et al., 2024); with $\mu = 0$ it reduces to zero ablation. The mean μ is one vector per site, averaged with equal weight over the ten domains on a fold disjoint from the evaluated items, and never uses the item's domain label. A context arm subtracts only the mean component, $h' = h - P\mu$.

Dose accounting. Offline calibration sets α_ℓ so that $\alpha_\ell^2 E_C(\ell) = E_T(\ell)$ on unmodified calibration activations at each block ℓ. The summaries in Table 2 use the applied *relative RMS* perturbation inside the intervened forward pass,

$$\rho_\ell = \sqrt{\textstyle\sum_i \|\delta h_{\ell i}\|^2 \Big/ \sum_i \|h_{\ell i}\|^2},$$

computed from the nominal BF16 update δh before the state write, with $h_{\ell i}$ the incoming state on that intervention's own trajectory; we record it for the treatment and every control at every block and item batch. Online calibration measures the treatment's applied energy $E_T^{\mathrm{app}}(\ell)$ on 32 calibration prompts of D inside its own intervened forward pass, then rescales each control by fixed-point iteration with up to four rounds, a tolerance of $\pm 10\%$ per block and $\alpha \le 4$, or up to 64 with clipping for iso and persistent frames, and records the achieved per-block ratios and the treatment's *depletion ratio* $E_T^{\mathrm{app}}(\ell)/E_T^{\mathrm{off}}(\ell)$. Controls not brought within tolerance are kept and counted, never dropped. This simultaneous update is the online-calibration procedure of Table 2. The four procedures thus target different quantities: offline, clean-state projection energy; online, nominal update energy on calibration prompts; forward sequential, nominal update energy per layer on a fixed reconstruction; actual write, the stored BF16 write energy per item and layer.

Three distinct dose quantities. For an orthogonal projector $P = U^\top U$, exact arithmetic gives

$$\begin{aligned} A &= \|\alpha P h\|^2 = \alpha^2 \|Ph\|^2, \\ R &= \|h\|^2 - \|h - \alpha P h\|^2 = \alpha(2-\alpha)\|Ph\|^2. \end{aligned} \tag{2}$$

Thus squared displacement A equals the reduction R in squared state norm only at full deletion, $\alpha = 1$, apart from zero displacement. For $\alpha > 2$, the norm reduction is negative. An offline projection uses a different state h^0; a relative-RMS ratio additionally divides by each trajectory's own $\|h\|^2$. Neither can be silently substituted for A. In BF16 we instead measure $A^{\mathrm{write}} = \|\,\mathrm{cast}(h') - h\|^2$ from the actual values before and after the model-state write; rounding can change both A and R.

Forward and same-item matching. With earlier scales fixed, the incoming state at layer ℓ is independent of later scales. The sequential calibration initializes

$$\alpha_\ell = \sqrt{A_\ell^T \Big/ \textstyle\sum_i \|P_\ell^C h_{\ell i}^C\|^2},$$

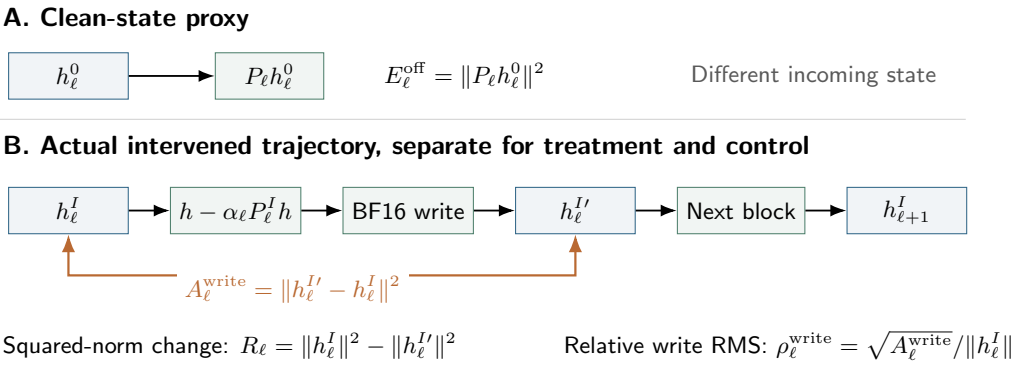


Figure 1: Three dose quantities refer to different measurements. Blue denotes model states, green the intervention and transition, and orange the measured write displacement. The clean-state projection uses h^0; each intervention instead follows its own h^I trajectory. Even matched write displacement leaves direction, state norm and cross-layer geometry different. ρ_ℓ^{write} uses the stored write; the ρ_ℓ of Table 2 and Figure 2 uses the nominal update.

refines the nominal BF16 update energy $\|\delta h\|^2$, and fixes it before advancing. It does not measure the stored before–after difference A^{write}. In a separate diagnostic, the target is each item's treatment write energy and the control scale is chosen from its current state without reading the answer or the loss. The finite-precision search, cap and complete failed-match denominator are given in Appendix B. This answer-free dose adjustment serves as a comparator diagnostic.

Statistics. Uncertainty for S comes from a crossed bootstrap (Owen & Eckles, 2012): evaluation items are resampled within each evaluation domain and the same resample serves every extraction row and control, since items are shared across the column; control identities are resampled within each extraction row and shared across its ten columns; 20,000 draws with a fixed seed. Following Lakens (2017), smallest effects of interest of ± 0.02 nats of choice CE and ± 1 accuracy point and the classification rules were fixed before the calibrated matrices were inspected. Paired contrasts on the development panel use question-stratified bootstrap intervals shared across frames.

4 Results

Every comparison below is paired within one experimental group listed in Appendix D.

4.1 Clean-state calibration overdoses multi-layer controls

When controls are energy-matched to the treatment layer by layer on unmodified activations, they deliver 2.65 times the treatment's applied relative RMS on Llama and 1.75–1.87 times on Qwen, averaged over blocks, up to 3.4–4.0$\times$ at individual blocks, with no cell within $\pm 10\%$; Table 2 and Figure 2a show this. The cause is depletion of the treatment's own removal: at the first intervened block the applied removal energy equals the offline value, with ratios of 0.89–1.15 across seven matrices, but at blocks 17–31 it falls to a median 0.04–0.06 of the offline value on Llama and Olmo and 0.23–0.30 on Qwen, consistent with earlier deletions having already removed much of the content of later subspaces, whereas redrawn control frames are not depleted in this way. Calibrating inside the treatment's own intervened forward pass brings the median control/treatment ratio to 1.08–1.12 on all seven three-model matrices, with interquartile ranges inside $[1.02, 1.20]$ and 40–62% of cells within $\pm 10\%$; Appendix Table 6 gives the per-matrix damage. Offline controls were also readout-orthogonalized, so the calibrations are compared as complete procedures; the overdose holds within each offline run, where the clean-state match holds by construction.

Control damage responds to this dose. In a registered dose–response on Qwen3.5-9B, the same five covariance-aware frames per row matched to $0.5\times$, $1\times$ and $2\times$ the treatment's applied RMS cost 0.18, 0.72 and 2.28 nats of choice CE: increments of 0.54 and 1.56 nats per doubling, an average

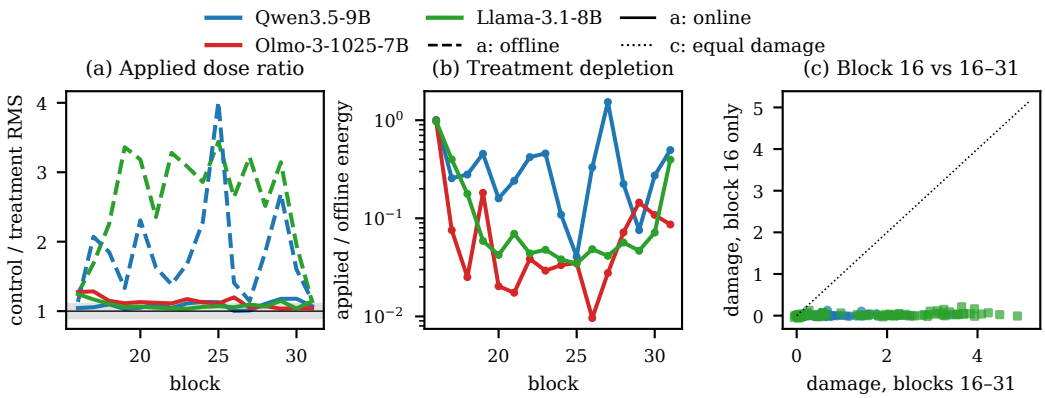


Figure 2: (a) Applied relative RMS of the control divided by that of the treatment at each intervened block, averaged over cells. Dashed: offline-calibrated controls, Llama seed 42 with 100 cells and Qwen seed 42 with 50 cells. Solid: online-calibrated controls, seed 42, 100 cells each. Grey band: $\pm 10\%$. (b) Treatment depletion ratio, applied energy inside the intervened forward pass divided by the energy the same basis removes from unmodified activations, per block, averaged over the ten extraction domains of seed 42. (c) Mean choice-CE damage per cell when the subspace is removed at block 16 only, on the vertical axis, versus at blocks 16–31, on the horizontal axis; 100 cells per model, identical items.

Table 2: Offline versus online dose calibration of the same treatment operators. Offline: each control's projected energy is matched to the treatment's on unmodified activations, layer by layer; online: applied energy is matched inside the treatment's own intervened forward pass on calibration prompts. Treatment results are identical between the offline and online runs of a seed, so rows differ only in the controls; offline controls were also readout-orthogonalized. Ratio: control/treatment applied relative RMS per cell, averaged over layers, with median and interquartile range, IQR, over cells; layer range: range of the per-layer cell-mean ratio over the 16 blocks; last column: squared ratio, not an absolute-energy ratio. The 54-cell rows include four cells of an incomplete sixth row; Figure 2 uses the 50-cell rows.

Model, seed	Calibration	Cells	Ratio median [IQR]	$\pm 10\%$	Layer range	Squared RMS ratio
Llama-3.1-8B, 42	offline	100	2.65 [2.39,2.76]	0%	1.13–3.44	7.0
Qwen3.5-9B, 42, all available	offline	54	1.84 [1.69,2.01]	0%	1.11–4.17	3.4
Qwen3.5-9B, 42, 5 complete rows	offline	50	1.86 [1.69,2.02]	0%	1.13–4.01	3.4
Qwen3.5-9B, 123, all available	offline	54	1.87 [1.65,2.06]	0%	1.07–3.81	3.5
Qwen3.5-9B, 123, 5 complete rows	offline	50	1.87 [1.69,2.07]	0%	1.10–3.94	3.5
Qwen3.5-9B, 456, 5 rows	offline	50	1.75 [1.62,1.95]	0%	0.89–3.45	3.1
Qwen3.5-9B, 42	online	100	1.09 [1.02,1.14]	53%	1.00–1.18	1.18
Qwen3.5-9B, 101	online	100	1.08 [1.04,1.12]	62%	0.99–1.18	1.17
Olmo-3-1025-7B, 42	online	100	1.10 [1.07,1.18]	49%	1.03–1.28	1.21
Olmo-3-1025-7B, 101	online	100	1.12 [1.07,1.20]	40%	1.05–1.37	1.27
Llama-3.1-8B, 42	online	100	1.09 [1.06,1.12]	56%	1.03–1.24	1.19
Llama-3.1-8B, 101	online	100	1.11 [1.07,1.16]	48%	1.05–1.31	1.22
Llama-3.1-8B, 202	online	100	1.11 [1.07,1.15]	48%	1.04–1.33	1.24

slope of 1.05 $[0.94, 1.16]$, so the response is convex. Such an overdose inflates the average control damage and so lowers treatment-minus-control effects τ on average; how it shifts the interaction S depends on how the inflation varies across cells.

Where the damage arises also matters. Deleting the subspace at block 16 alone costs 0.008 and 0.020 nats on Qwen3.5-9B and Llama-3.1-8B, against 0.62 and 1.55 nats for the sixteen-block stack, as Appendix Table 4 and Figure 2c show. A registered single-block scan on Qwen locates the damage at block 31, 0.670 nats against 0.003–0.028 at blocks 16, 20, 24 and 28, and the stack without block 31 retains 0.170 $[0.149, 0.190]$ of the full-stack damage, as Appendices A and C.7 detail. Dose-matched covariance-aware frames at block 31 cost 1.164 nats, so a stacked comparison has to be dosed at the late blocks, where the treatment's own removal is most depleted.

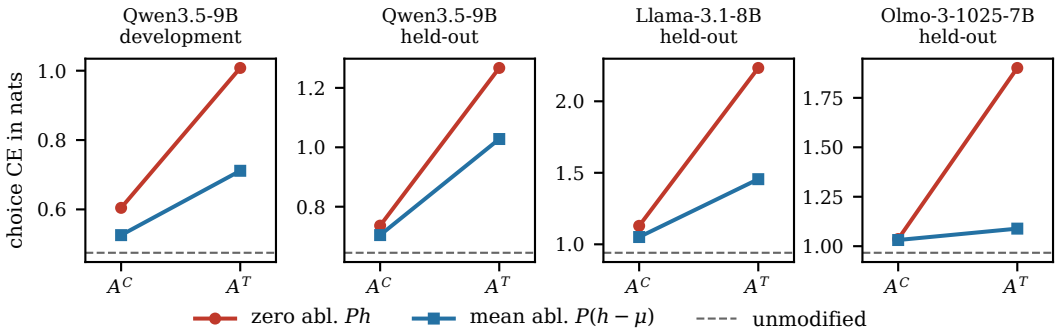


Figure 3: Direction × displacement cross: choice CE when each direction is written, per item and layer, at the mean-ablation displacement A^C or the zero-ablation displacement A^T. Left: development panel of Appendix Table 10, 770 cases; right: registered held-out experiment of Table 3, 25,000 cases per model. Dashed: unmodified model. At the zero-ablation displacement the mean-ablation direction is less damaging in every panel.

4.2 Sequential calibration and actual-write matching

A constant random frame applied at every layer is a geometry control distinct from the layer-specific treatment; on four matrices the simultaneous update brings none of 200 such controls within tolerance after five passes with $\alpha \leq 64$. On a paired reconstruction of the abstract-algebra row with 32 calibration prompts, five fixed frames and 16 layers, the simultaneous update matches 0 of 5 profiles, whereas forward sequential calibration, which fits each layer's scale on the state produced by the already-fixed earlier layers and then fixes it, matches 5 of 5 nominal-update energy profiles on a complete replay, as Appendix B details. At all 80 control–layer steps the analytic initialization on the current trajectory already met the tolerance, with energy ratios of 0.985–1.015. The procedures also differ in initialization, solver and search budget, so this compares complete procedures; Appendix Table 19 gives their costs.

Nominal update energy is not the stored displacement in BF16. We therefore held the abstract-algebra treatment, five persistent frames and 500 evaluation items fixed and matched each control's actual write energy, $\|\operatorname{cast}(h') - h\|^2$, to the treatment's on the same item and layer, choosing each scalar from the current state without answers or losses. Of 40,000 control writes, 39,963, or 99.91%, are within 1% of target, with median error 0.15%, maximum 2.69% and no cap hit, and all 800 control–layer–domain energy ratios lie in $[0.995, 1.004]$. Relative to population-average calibration of the same frames, per-write matching lowers the row's choice-CE contrast by 0.026 $[0.008, 0.046]$ nats, a paired protocol comparison detailed in Appendix B. Equal displacement is also not equal norm reduction: 19,924 of 40,000 control writes, but 266 of 8,000 treatment writes, increase the state norm, which is why we report the two quantities separately.

4.3 Mean ablation reduces collateral damage beyond its smaller dose

The geometry and mean-ablation experiments use a 77-question development panel of Qwen3.5-9B with 7–8 questions per domain and the ten original frames. Two geometry controls transform the treatment's own sixteen-layer frame: a common rotation preserves every cross-layer inner product of the stacked frame, whereas independent rotations preserve only within-layer geometry. With the label-free native-write search both match all 61,600 writes within 1%, yet the state norm rises in 57.2% and 78.7% of their writes against 274 of $12{,}320$ treatment writes, so they are displacement-matched comparators rather than equivalent erasures, as Appendix C details.

A cPCA frame is fitted to centered covariances, but zero ablation also removes the mean component of the state along the frame. Mean ablation of the same frames, with μ estimated on the opposite fold of the panel, raises accuracy from 69.02% to 79.11%, a paired difference of $+10.09$ points with 95% interval $[5.91, 14.20]$, against 79.46% for the unmodified model, and lowers choice CE from 1.008 to 0.526 nats against a baseline of 0.475. Subtracting only the mean component is far more damaging, at 3.615 nats and 38.5% accuracy, so the benefit does not come from the mean term alone.

Table 3: Registered held-out experiment with refitted frames and μ, 25,000 cases per model. Choice CE in nats and accuracy in % of the unmodified model U, zero ablation Z and mean ablation M; E_C/E_T: ratio of their squared displacements; Δ_T, Δ_C: mean-ablation direction minus zero ablation at the zero- and mean-ablation displacement. Intervals are in the text and Appendix Table 11; each model has 972 distinct questions. The registered two-displacement criterion, upper 95% bounds of Δ_T and Δ_C below zero, is met on Qwen3.5-9B and Llama-3.1-8B.

	Choice CE			Accuracy				Δ_T		Δ_C	$\Delta_T - \Delta_C$
Model	U	Z	M	U	Z	M	E_C/E_T	CE	acc.	CE	CE
Qwen3.5-9B	0.645	1.267	0.704	76.0	62.5	73.8	0.181	-0.239	$+6.4$	-0.032	-0.207
Llama-3.1-8B	0.941	2.233	1.051	58.0	42.1	53.6	0.028	-0.778	$+4.5$	-0.078	-0.699
Olmo-3-1025-7B	0.967	1.901	1.031	57.8	42.3	55.4	0.045	-0.812	$+10.8$	-0.005	-0.807

Mean ablation also writes a smaller displacement, 83.6 versus 422.5 per write, so its benefit could reflect dose alone. Figure 3 and Appendix Table 10 cross the two directions with the two realized displacements. At the zero-ablation displacement A^T, the mean-ablation direction lowers choice CE by 0.297 nats $[0.189, 0.407]$ and raises accuracy by 7.25 points $[3.98, 10.52]$. The direction effect grows with displacement, the difference between the direction effects at A^T and A^C being 0.218 $[0.129, 0.300]$ nats, and lowering the displacement of zero ablation to A^C removes much of its damage, from 1.008 to 0.604 nats, so direction and dose both contribute. The effect at A^T is insensitive to matching failures: it is 0.301 $[0.204, 0.399]$ nats on the 724 cases whose rescaled arm matched every write, as shown in Appendix C.3.

Registered held-out replication. Because the panel's frames saw its questions, we registered a replication, described in Appendix C.4: for each model and each of five item splits, frames and μ are refitted on the extraction half and all four arms are evaluated on the 500 held-out items of the split, 25,000 cases per model. The 25,000 cases cover 972 distinct questions, the bootstrap unit; they are held out from fitting but not unseen: all were scored in the principal matrices, and they include the 77 development items in 1,670 of the 25,000 cases per model. Mean ablation restores 11.3, 11.5 and 13.1 of the 13.5, 15.8 and 15.5 accuracy points that zero ablation costs on Qwen3.5-9B, Llama-3.1-8B and Olmo-3-1025-7B, as Table 3 shows. At the zero-ablation displacement the mean-ablation direction lowers choice CE by 0.239 $[0.212, 0.267]$, 0.778 $[0.746, 0.811]$ and 0.812 $[0.772, 0.855]$ nats and raises accuracy by 4.5 to 10.8 points. Post hoc, it is resolved in all fifteen model–split replicates and every displacement quintile, as Appendix C.6 shows. On Qwen3.5-9B and Llama-3.1-8B the direction effect is also resolved at the mean-ablation displacement, by 0.032 $[0.021, 0.044]$ and 0.078 $[0.064, 0.094]$ nats, meeting the registered replication criterion. Intervals condition on each split's fitted frames and means; Appendix Table 12 shows the spread over refitted splits. The rescaled arms match 97.6–99.98% of writes with misses retained; excluding the 77 development questions post hoc, or keeping only cases whose rescaled writes all matched, moves Δ_T by at most 0.06 nats, as Appendix Table 11 shows. A registered extension runs the unchanged design on two further model families, Mistral-7B-v0.3 and Qwen3-8B, with all five splits and 25,000 cases each: mean ablation writes 6% and 5% of the squared displacement of zero ablation and restores 8.3 of 9.6 and 17.1 of 18.9 accuracy points, and at the zero-ablation displacement the mean-ablation direction lowers choice CE by 0.248 $[0.225, 0.271]$ and 1.127 $[1.077, 1.178]$ nats, meeting the registered extension rule of Appendix C.5. The direction effect at the zero-ablation displacement thus holds in all five checkpoints, and the two-displacement criterion in Qwen3.5-9B, Llama-3.1-8B and Qwen3-8B.

New questions and the format of the reference mean. Two registered extensions with the seed-42 frames and means of Qwen3.5-9B and Llama-3.1-8B, described in Appendix C.9, address the two main limits of this replication. On the 407 MMLU test questions that lie beyond the first 100 of their domain, which no earlier experiment of this study used, the mean-ablation direction lowers choice CE at the zero-ablation displacement by 0.187 $[0.145, 0.230]$ and 0.901 $[0.826, 0.975]$ nats on 4,070 cases per model, and both registered rules, including the two-displacement criterion, are met. With μ taken from the zero-shot extraction-format states on which the frames are fitted, the effect at the zero-ablation displacement is 0.276 $[0.238, 0.315]$ and 0.705 $[0.656, 0.753]$ nats, 87% and 73% of its size with the evaluation-format mean on the same 5,000 cases.

Where the direction effect is carried. A registered localization, described in Appendix C.8, reruns the four arms on the seed-42 split of Qwen3.5-9B and Llama-3.1-8B without the final block, that is, on blocks 16–30. Zero ablation then costs 16% and 48% of its full-stack damage, and at its displacement the mean-ablation direction still lowers choice CE by 0.057 $[0.040, 0.076]$ and 0.638 $[0.592, 0.684]$ nats, against 0.317 and 0.963 with the final block on the same 5,000 cases. The final block carries most of the effect on Qwen and a third on Llama.

4.4 A REGISTERED POSITIVE CONTROL

A registered positive control on Qwen3.5-9B crosses three answer-symbol formats, A–D, a–d and W–Z, on the same MMLU items; with all 15 controls matched at a median RMS ratio of 1.12, the calibrated crossed assay detects format selectivity with 90% crossed-bootstrap intervals: $S_{\mathrm{CE}} = +1.23$ $[+0.95, +1.54]$ nats, diagonal top-1 in 3 of 3 rows and $S_{\mathrm{acc}} = +17.4$ $[+13.7, +21.1]$ points, a PASS under the prespecified criterion. This effect is about 60 times the smallest effect of interest.

5 DISCUSSION

What the measurements establish. The perturbation a control receives in a multi-layer ablation is decided on the trajectory it creates: clean-state calibration overdoses random controls by 1.7–$2.7\times$ in applied relative RMS, so a clean-state match biases treatment-minus-control effects. Trajectory, forward sequential and per-write calibration reduce the dose mismatch in the settings where each was tested. Once displacement is measured, the reference value of the deletion becomes visible: rescaled to the realized displacement of zero ablation, the mean-ablation direction of a cPCA frame causes 0.24–1.13 nats less collateral damage in all five checkpoints.

Why the reference value matters for centered frames. cPCA, like any extractor built from centered second moments, is silent about the mean of the states. Zero ablation nevertheless sets the component along the frame to zero, so it also moves every input by the same vector $-P\mu$. Mean ablation, $h' = (I - P)h + P\mu$, keeps $P\mu$ for a chosen reference μ, here an evaluation-format mean. In the five-split experiments it writes 3–18% of the squared displacement of zero ablation, about 20% on the development panel, yet rescaled to the zero-ablation displacement its direction still does less damage; with a fitting-format mean, the effect at that displacement keeps 87% and 73% of its size on Qwen3.5-9B and Llama-3.1-8B. The identity $Ph = P(h - \mu) + P\mu$ does not decompose the loss, as Appendix C explains. A random control and a treatment should therefore share the reference value as well as the dose: zero ablation carries a mean shift that a random frame need not.

Choice and full-vocabulary endpoints. Loss of answer-letter mass, Table 7, carries 39–52%, 22–27% and at most 5% of treatment full-NLL damage on Qwen3.5-9B, Llama-3.1-8B and Olmo-3-1025-7B, so choice-only and full-vocabulary endpoints can rank interventions differently.

Scope. The comparators are displacement-matched, not erasure-matched: they do not equate state norm, angle or downstream alignment. The rescaled mean-direction arm writes $h - aP(h - \mu)$, whose subspace component $(1 - a)Ph + aP\mu$ equals $P\mu$ only at $a = 1$, so the cross is a direction-versus-displacement diagnostic rather than a comparison at equal erasure. The effect is identified for the whole deletion of a cPCA frame: rank 4, final token, upper-half blocks, 18–35 in Qwen3-8B and 16–31 otherwise; on two models it persists without the final block and with a fitting-format mean, and on Llama-3.1-8B for a mean-difference frame. The five-split replication is held out from fitting within the examined MMLU pool; questions outside it are tested on two models and seven domains.

6 CONCLUSION

Controls for multi-layer subspace ablations must be dosed on the trajectory they create, with realized displacement, state-norm change and matching failures reported separately and direction crossed with displacement. The reference value that replaces a deleted component fitted to centered statistics should be stated: that choice alone moved held-out accuracy by 8–17 points here.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to support qualitative and thematic data analysis, and generating synthetic data sets, formulating questions for surveys or interviews, and transcribing recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies interventions on the internal activations of publicly released language models evaluated on the public multiple-choice benchmark MMLU. It involves no human subjects, no crowdsourced annotation and no personal or sensitive data, and it releases no new data set beyond the per-item results derived from the public benchmark. The methods concern how controls for activation ablations are matched and do not increase model capabilities; we are not aware of a direct harmful application, privacy or security risk, or fairness concern arising from them.

REPRODUCIBILITY STATEMENT

The intervention procedures, dose measurements and calibration procedures are defined in Section 3; experimental settings, the registered protocol, decision rules, checkpoints, input construction and costs are given in Appendix D; scoring is described in Appendix E; the registered experiments and their analyses are described in Appendices B and C; and every registered test is listed in Appendix G. The anonymous supplementary material contains the intervention and analysis code, the registered protocol, the per-case results of the held-out experiment, its model-family extension, the site localization, the fitting-format reference mean, the questions outside the pool and the mean-difference frame, the per-write energy ratios and match records of the site localization, and scripts that recompute the registered estimates from these results.

REFERENCES

Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9:2134, 2018. doi: 10.1038/s41467-018-04608-8. URL https://doi.org/10.1038/s41467-018-04608-8.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10887–10911, 2025. doi: 10.18653/v1/2025.naacl-long.544. URL https://aclanthology.org/2025.naacl-long.544/.

Georgii Aparin and Tatiana Gaintseva. A geometric account of activation steering through angle-norm decomposition. *arXiv preprint arXiv:2606.06735*, 2026. URL https://arxiv.org/abs/2606.06735.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023. doi: 10.52202/075280-2884. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Quy-Anh Dang and Chris Ngo. Selective steering: Norm-preserving control through discriminative layer selection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10887–10910. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.529. URL https://aclanthology.org/2026.findings-acl.529/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Mirian Hipolito Garcia, Camille Couturier, Daniel Madrigal Diaz, Ankur Mallick, Anastasios Kyrillidis, Robert Sim, Victor Ruhle, and Saravan Rajmohan. Exploring how LLMs capture and represent domain-specific knowledge. *arXiv preprint arXiv:2504.16871*, 2025. URL https://arxiv.org/abs/2504.16871.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cZrTMqYVL6.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1797.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109233–109282, 2024. doi: 10.52202/079017-3468. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c55e6792923cc16fd6ed5c3f672420a5-Abstract-Conference.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, pp. 26486–26515, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in LLMs is an affine function. *arXiv preprint arXiv:2411.09003*, 2024. URL https://arxiv.org/abs/2411.09003.

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

Alex McKenzie, Keenan Pepper, Stijn Servaes, Martin Leitgab, Murat Cubuktepe, Michael Vaiana, Diogo de Lucena, Judd Rosenblatt, and Michael S. A. Graziano. Endogenous resistance to activation steering in language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fYbuhXJ3iq.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit evaluation metrics are not robust. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zSf8PJyQb2.

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

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. doi: 10.1214/12-AOAS547.

Manav Pandey. LLMs know they're wrong and agree anyway: The shared sycophancy-lying circuit. *arXiv preprint arXiv:2604.19117*, 2026. URL https://arxiv.org/abs/2604.19117.

Daniele Potertì, Andrea Seveso, and Fabio Mercorio. Can role vectors affect LLM behaviour? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17735–17747, 2025. doi: 10.18653/v1/2025.findings-emnlp.963. URL https://aclanthology.org/2025.findings-emnlp.963/.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Orion Reblitz-Richardson. Calibrating interpretability instruments before trusting their verdicts. *arXiv preprint arXiv:2609.14754*, 2026. URL https://arxiv.org/abs/2609.14754.

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, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Baturay Saglam, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, and Amin Karbasi. Large language models encode semantics and alignment in linearly separable representations. 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. 2282–2303, Mumbai, India, 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.124. URL https://aclanthology.org/2025.ijcnlp-long.124/.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45663–45680, 2024. URL https://proceedings.mlr.press/v235/singh24d.html.

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

Minh Hieu Vu and Tan Nguyen. Angular steering: Behavior control via rotation in activation space. In *Advances in Neural Information Processing Systems*, volume 38, pp. 121653–121690, 2025. doi: 10.52202/085713-4056. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b0223cad0e73b793f31eb6cc41cefceb-Abstract-Conference.html.

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

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, pp. 78293–78318, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024. URL https://arxiv.org/abs/2401.12631.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

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

A SUPPLEMENTARY TABLES FOR THE CALIBRATED MATRICES

The stack without its last block, a registered follow-up. Registered after the site scan and before any of its cells existed, two arms of the Qwen diagnostics rerun the upper-half stack of Table 5

Table 4: Single-block versus stacked removal of the same rank-4 domain subspaces with the same seed, 100 cells and held-out items. "Single" removes the subspace at block 16 only, with its own online-calibrated controls; "stacked" removes it at all 16 upper-half blocks. Values are mean per-item damage, single / stacked, over all 100 cells or over the 10 diagonal cells, diag; "max cell" is the largest single-block cell mean; "control, single" is the mean damage of the dose-matched random controls in the single-block run; r is the Pearson correlation of the 100 cell means between the two runs.

Model, seed	Block	Choice CE all	Choice CE diag	Acc. drop, points	Full NLL	Max cell, single	Control, single	r
Qwen3.5-9B, 42	16	0.008 / 0.622	0.012 / 0.510	0.7 / 15.1	0.037 / 1.117	0.129	0.006	0.13
Llama-3.1-8B, 42	16	0.020 / 1.550	0.010 / 1.455	0.2 / 16.1	0.020 / 1.989	0.213	0.022	0.42

Table 5: Registered site scan on Qwen3.5-9B, seed 42, reproduced exactly in an independent rerun, with 100 cells per site and five dose-matched covariance-aware controls per row. Damage and max cell: mean and largest treatment choice-CE damage in nats; control: mean control damage; dose: median control/treatment applied-RMS ratio; profile recovery: blocks until the median energy in the layer-specific frame U_ℓ is back to ≥ 0.5 of baseline. Last row: full upper-half stack.

Blocks	Damage	Max cell	Acc. drop, points	Control	Dose	Profile recovery
16	0.009	0.13	0.3	0.006	1.02	3
20	0.003	0.10	0.4	0.010	1.04	2
24	0.011	0.15	0.6	0.023	1.11	2
28	0.028	0.27	1.4	0.044	1.04	3
31	0.670	3.99	14.7	1.164	1.01	–
24–31	0.545	2.74	12.7	0.713	1.06	2
16–31, full stack	0.607	–	–	–	–	–

on seed 42 with the downstream projection recorded: the full stack, blocks 16–31, and the stack without its last block, blocks 16–30. Their mean treatment choice-CE damage is 0.607 and 0.103 nats, with accuracy drops of 14.0 and 3.8 points and dose-matched control damage of 0.725 and 0.134 nats. Against the site scan's full-stack reference of 0.607 nats the stack without its last block costs $R = 0.170$ of the damage. The prespecified criterion reads $R \geq 0.50$ as the stack without its last block carrying most of the damage, $R \leq 0.20$ as the last block carrying it, and intermediate values as shared; the last block therefore carries the damage. Against the rerun the ratio is 0.16963 with 90% interval $[0.14938, 0.19047]$. This bootstrap resamples the 50 questions within each evaluation domain, sharing each resample across all ten extraction rows and both arms; it uses 10,000 draws with seed 20260918, takes the 5th and 95th percentiles and keeps the fitted frames fixed; the supplementary script gives the exact resampling order. Median projection energy in each layer's own frame U_ℓ, relative to the baseline trajectory, at blocks 16–31 of the full stack is 1.00, 0.23, 0.25, 0.53, 0.11, 0.24, 0.54, 0.39, 0.11, 0.04, 0.47, 0.91, 0.21, 0.05, 0.24, 0.46; the stack without its last block has the identical profile. The full-stack rerun reproduces all 300 compared outputs of the reference run exactly. On item-split seed 101, with the same two arms and analysis, the full and no-last stacks cost 0.461 and 0.080 nats, so $R_{101} = 0.17258$ with shared-item 90% interval $[0.13865, 0.20420]$; the accuracy drops are 9.52 and 3.08 points. The supplementary material contains the per-question treatment and control damages of all four runs, ten extraction by ten evaluation domains by 50 questions, from which both ratios and both shared-question intervals are recomputed exactly. On seed 101 the median projected-energy ratios at blocks 16–31 are 1.00, 0.25, 0.28, 0.48, 0.10, 0.26, 0.49, 0.38, 0.10, 0.03, 0.57, 1.14, 0.22, 0.07, 0.37, 0.53.

Single-block scans. The sixteen-block stack costs 15–16 accuracy points, against 0.7 and 0.2 points for deleting the rank-4 subspace at block 16 alone, as Table 4 shows. In the registered single-block scan of Table 5, the stack without block 31 retains 0.173 of the full-stack damage on a second split seed. The registered rescan with refitted frames of Appendix C.7 gives 0.006–0.034 nats at blocks 23 and 27–30 and 0.693 at block 31 on Qwen3.5-9B; in Llama-3.1-8B and Olmo-3-1025-7B, too, block 31 is the most damaging of blocks 27–31, so block 31 of Qwen3.5-9B is highly sensitive to rank-4 deletion in general.

Table 6: Achieved dose and damage levels for the seven online-calibrated three-model matrices, with five covariance-aware controls per row and 50 held-out items per evaluation domain. "Ratio" is the applied relative RMS of the control divided by that of the treatment, measured inside the intervened forward pass on the evaluation items and averaged over the 16 intervened layers, with median and interquartile range over 100 cells; "$\pm 10\%$" is the fraction of cells within that tolerance; "unmatched" counts controls outside tolerance after calibration, which are kept, not dropped. Damage is the mean per-item increase in choice CE in nats, or the accuracy drop in points, for treatment T and controls C, on diagonal / off-diagonal cells. "Collapse" counts cells whose mean answer-letter mass falls below 0.5 under treatment, as in Table 7.

	Applied dose, control/treatment			Choice-CE damage		Acc. drop, points		
Seed	Ratio [IQR]	$\pm 10\%$	unmatched	T diag / off	C diag / off	T diag / off	C diag / off	Collapse
Qwen3.5-9B								
42	1.09 [1.02,1.14]	53%	0/50	0.51 / 0.63	0.60 / 0.70	15 / 15	14 / 16	15
101	1.08 [1.04,1.12]	62%	0/50	0.42 / 0.48	0.69 / 0.63	11 / 11	16 / 16	7
Olmo-3-1025-7B								
42	1.10 [1.07,1.18]	49%	5/50	0.83 / 0.86	0.40 / 0.36	15 / 14	9 / 10	0
101	1.12 [1.07,1.20]	40%	2/50	0.90 / 0.91	0.45 / 0.43	21 / 15	10 / 10	0
Llama-3.1-8B								
42	1.09 [1.06,1.12]	56%	0/50	1.46 / 1.56	1.03 / 1.03	18 / 16	14 / 15	10
101	1.11 [1.07,1.16]	48%	0/50	1.25 / 1.17	1.12 / 1.17	24 / 15	19 / 19	10
202	1.11 [1.07,1.15]	48%	0/50	1.21 / 1.28	1.20 / 1.19	21 / 15	15 / 15	10

Table 7: Answer-set mass decomposition, $\Delta L_{\text{full}} = \Delta L_{\text{choice}} + \log(M_0/M_I)$ applied per item before averaging, with identity residual $\leq 2.3 \times 10^{-16}$ in every run. "Share" is the fraction of treatment full-NLL damage carried by the mass term. Masses are mean answer-letter probability mass at baseline / under treatment / under controls. Frame, controls: the rank-4 cPCA treatment frame or, in the extractor comparison, a DiffMean direction plus three cPCA directions; cov and iso are the covariance-aware and isotropic random controls of Section 3. Diag.: the 10 cells with extraction domain equal to evaluation domain; off-diag.: the other 90 cells.

			Treatment		Control	Answer mass
Seed	Frame, controls	Cells	full = choice + mass	Share	full = choice + mass	base / treat / ctrl
Qwen3.5-9B						
42	cPCA, cov	diag.	1.066 = 0.510 + 0.556	0.52	1.368 = 0.599 + 0.769	0.923 / 0.772 / 0.730
42	cPCA, cov	off-diag.	1.123 = 0.634 + 0.488	0.43	1.249 = 0.703 + 0.546	0.923 / 0.774 / 0.763
101	cPCA, cov	diag.	0.733 = 0.418 + 0.315	0.43	1.146 = 0.690 + 0.456	0.932 / 0.814 / 0.771
101	cPCA, cov	off-diag.	0.789 = 0.479 + 0.309	0.39	1.086 = 0.634 + 0.451	0.932 / 0.816 / 0.776
42	cPCA, iso	diag.	1.066 = 0.510 + 0.556	0.52	0.042 = 0.018 + 0.024	0.923 / 0.772 / 0.909
42	cPCA, iso	off-diag.	1.123 = 0.634 + 0.488	0.43	0.044 = 0.027 + 0.017	0.923 / 0.774 / 0.910
42	DiffMean, cov	diag.	0.716 = 0.402 + 0.314	0.44	1.551 = 0.708 + 0.842	0.923 / 0.763 / 0.744
42	DiffMean, cov	off-diag.	0.959 = 0.593 + 0.366	0.38	1.489 = 0.810 + 0.679	0.923 / 0.788 / 0.759
Olmo-3-1025-7B						
42	cPCA, cov	diag.	0.878 = 0.830 + 0.048	0.05	0.432 = 0.398 + 0.035	0.998 / 0.957 / 0.970
42	cPCA, cov	off-diag.	0.904 = 0.855 + 0.049	0.05	0.371 = 0.358 + 0.013	0.998 / 0.956 / 0.986
101	cPCA, cov	diag.	0.934 = 0.903 + 0.031	0.03	0.476 = 0.450 + 0.025	0.998 / 0.970 / 0.977
101	cPCA, cov	off-diag.	0.944 = 0.907 + 0.037	0.04	0.448 = 0.434 + 0.015	0.998 / 0.965 / 0.985
Llama-3.1-8B						
42	cPCA, cov	diag.	1.866 = 1.455 + 0.410	0.22	1.087 = 1.028 + 0.060	0.999 / 0.898 / 0.961
42	cPCA, cov	off-diag.	2.002 = 1.561 + 0.442	0.22	1.102 = 1.034 + 0.068	0.999 / 0.895 / 0.957
101	cPCA, cov	diag.	1.641 = 1.253 + 0.388	0.24	1.146 = 1.125 + 0.021	0.999 / 0.901 / 0.981
101	cPCA, cov	off-diag.	1.544 = 1.172 + 0.371	0.24	1.196 = 1.172 + 0.023	0.999 / 0.902 / 0.979
202	cPCA, cov	diag.	1.661 = 1.212 + 0.449	0.27	1.255 = 1.198 + 0.057	0.999 / 0.899 / 0.953
202	cPCA, cov	off-diag.	1.752 = 1.278 + 0.474	0.27	1.261 = 1.187 + 0.075	0.999 / 0.897 / 0.948

Answer-set mass. Table 7 gives the per-run decomposition of the full-vocabulary damage summarized in Section 5. On Qwen3.5-9B, 7–15 cells per matrix have a mean answer mass below 0.5 under treatment, and the dose-matched covariance-aware controls lose mass at a similar rate. On

Llama-3.1-8B the formal-logic subspace collapses the answer mass of every evaluation domain to 0.01–0.12, in ten cells of each of the three splits.

Repeated runs of the Qwen diagnostics. The $1\times$ arm of the control dose–response of Section 4.1 was run three times independently, and all recorded outcomes of its 120 experimental conditions agree exactly across the three runs; the site scan of Table 5 and the format control of Section 4.4 also give identical outcomes when rerun. Each of the 23 diagnostic runs we report from repeated runs agrees exactly with another run: 15 in all recorded outcomes, 7 in their treatment damages, frames and baselines under a different control calibration, and the dose-gradient run at $\alpha = 0.25$, whose treatment is scaled differently, in its frames and unmodified baselines.

B Forward calibration and actual-write-matched diagnostic

B.1 Paired forward calibration

The calibration row is abstract algebra, split seed 42, rank four and blocks 16–31 at the final prompt token. The treatment basis and five control seeds are fixed; 50 extraction prompts supply the offline statistics and the first 32 supply the online calibration. Each fixed random frame is reconstructed once in the top-16 PCA span of half of the 198 contrast prompts, using the same 48-candidate selection. Both solvers receive identical reconstructed frames and prompts. The experiment therefore compares the two solvers on identical reconstructed frames rather than replaying the original online-calibrated run.

The simultaneous procedure uses four multiplicative updates and five measurement passes, cap 64 and 10% energy tolerance. In the sequential procedure, earlier scales are held fixed when the incoming states for the next layer are collected. The analytic scale in Section 3 initializes a bounded refinement of nominal update energy $\|\delta h\|^2$ in BF16. Scales enter the stored frame as $\sqrt{\alpha}U$ before projection. The complete scale vector is replayed, and a control passes only if every layer's nominal update energy is within 10% of the treatment target. Both solvers retain all five controls: simultaneous matching succeeds for 0/5, sequential matching for 5/5. Besides the order of the updates, the sequential procedure has its own initialization and a bracketing scalar solver with up to 48 evaluations per layer, against four multiplicative updates in the simultaneous one, so the result compares the two procedures, not update order alone. The supplementary material contains every control's per-layer history, with target and achieved nominal update energy, ratio, number of scalar evaluations, analytic initialization and full-replay agreement, and the simultaneous run's final per-layer doses; recounting them gives 5/5 and 0/5. It concerns nominal update energy on the fixed reconstruction; the stored displacement is matched by the protocol below.

B.2 Actual-write matching protocol

The actual-write diagnostic keeps the same one extraction row, five fixed persistent random frames, 16 layers and 500 held-out questions, 50 from each of ten domains. The item pool, split and five-shot prompts are unchanged. These questions were used in specifying the protocol and are therefore development data; every item is retained.

For item i and layer ℓ, a separate treatment pass supplies the target energy

$$A^T_{\ell i} = \sum_j \Big(\mathrm{float32}\big(h^{T\prime}_{\ell i,j}\big) - \mathrm{float32}\big(h^{T}_{\ell i,j}\big)\Big)^2,$$

where h^T and $h^{T\prime}$ are the actual BF16 states before and after the write. For a control's current state h^C, the fixed frame Q computes, in native BF16 arithmetic, the projected vector and the candidate write

$$q = \big(h^C Q^\top\big)Q,$$
$$h^{C\prime} = \mathrm{BF16}\big(h^C - \mathrm{BF16}(a\,q)\big).$$

A nonnegative scalar a is selected to match the actual energy of $h^{C\prime}$ to A^T. The initial estimate is $\sqrt{A^T/\|q\|^2}$. A doubling bracket followed by bisection evaluates at most 48 candidates including

zero, with cap 2^{20} and 1% relative-energy tolerance. The closest finite evaluated candidate is used, with ties broken by smaller a. A zero target leaves the state unchanged; zero direction with positive target, cap, quantization and search-budget failures remain explicit failures and remain in every denominator. The local search uses neither answers nor losses and requires no additional model forward passes. In low precision the objective is a staircase: an unattainable tolerance is not repaired by dropping the item.

A complete treatment pass reproduces all four reference outcomes for every one of the 500 items. Matching covers 39,963/40,000 control writes, or 99.9075%; 37 fail. Relative error has median 0.1455%, 99th percentile 0.9098% and maximum 2.6893%. The largest scalar is 52.17275, below the fixed cap, with no cap hit. Across all 800 control–layer–domain aggregates, the control/treatment actual-energy ratio is 0.995267–1.004216. Incoming states and directions differ despite this scalar match. State-norm increases occur in 19,924/40,000 control writes and 266/8,000 treatment writes, so norm reduction is not controlled.

Table 8: Complete ten-domain outcomes of the actual-write-matched abstract-algebra row. τ is treatment minus the mean of five controls in choice-CE damage. Intervals are descriptive 95% paired within-domain item bootstrap intervals with 2,000 resamples and seed 42, conditional on these five frames. Counts are correct answers out of 50; every item is retained.

Evaluation domain	Baseline / treatment correct	τ, nats	95% interval
Abstract algebra	32 / 16	0.552781	$[0.217288, 0.922179]$
Anatomy	33 / 36	0.202297	$[-0.012503, 0.449482]$
Astronomy	48 / 47	0.004675	$[-0.038590, 0.054769]$
College chemistry	32 / 30	0.114055	$[-0.102865, 0.347776]$
College computer science	41 / 35	0.349096	$[0.150151, 0.579193]$
Formal logic	32 / 26	0.546801	$[0.273211, 0.834450]$
Jurisprudence	45 / 44	0.019291	$[-0.105112, 0.150661]$
Logical fallacies	42 / 40	0.112542	$[-0.107148, 0.355797]$
Machine learning	37 / 30	0.432035	$[0.159324, 0.725236]$
Philosophy	38 / 38	0.112043	$[-0.072203, 0.318297]$

B.3 Paired protocol comparison

Relative to the preceding population-calibrated control of the same frames, the row's choice-CE contrast changes by -0.026447467 nats with interval $[-0.045850589, -0.008268360]$. The paired comparison changes both the dose criterion and scalar placement: the preceding procedure projects through a BF16-cast $\sqrt{\alpha}U$, whereas the new procedure scales the projected vector. These operations need not agree after rounding, so the paired difference is a protocol comparison rather than an isolated effect of dose matching. Intervals are conditional on the fixed frames. Total correctness is 380/500 for baseline, 342/500 for treatment and 382, 374, 375, 377 and 378 for the five controls; on abstract algebra, 19 correct answers become wrong and 3 wrong answers become right under the treatment. The diagnostic comprises 500 treatment-reference and 2,500 control forward passes; the 500 unchanged baseline outcomes are reused after exact reference checks.

C Geometry-matched controls and common-mean preservation

C.1 Scope, fixed material and crossed geometry controls

In this appendix, original deletion denotes zero ablation, $h - Ph$, centered deletion denotes mean ablation, $h - P(h - \mu)$, and mean-only deletion subtracts the mean component alone, $h - P\mu$. The panel contains 77 development questions from the extraction halves of the same ten domains, 7–8 per domain. It was fixed before either experiment: of 80 candidates, three were excluded because an option refers to other options by letter, which would invalidate the answer-permutation checks. No baseline-correct filter or favorable-domain selection is applied. These data are exposed development examples, not a fresh confirmation set. The ten original rank-four cPCA frames, the sixteen final-token module-output sites at blocks 16–31, prompt token sequences and checkpoint remain fixed;

Table 9: Displacement-matched geometry controls on the exposed development panel. Prediction means weight question domains and extraction domains equally and average all five control seeds. A trajectory matches only if all sixteen native write displacements meet the fixed 1% tolerance. Baseline and original deletion are references, not matched random controls.

Arm	Choice CE	Accuracy, %	Matched trajectories
Baseline	0.474880	79.4643	—
Original deletion	1.007968	69.0179	—
Common rotation, matched	0.485914	79.9393	3,850/3,850
Independent rotations, matched	0.488908	79.8786	3,850/3,850

states and projections are computed in BF16. Baseline predictions and outcomes are identical across repeated runs and all extraction rows. The original frame Gram matrices in 32-bit floating point differ from identity by at most 4.657×10^{-7}; after conversion to BF16 the frames are measured, not assumed to be exactly orthogonal.

For each of five fixed random seeds, the common control applies one ambient orthogonal rotation to the stacked layer-specific frame. This preserves all cross-layer inner products in real arithmetic. The independent control applies a distinct orthogonal rotation at each layer, retaining within-layer rank and geometry while altering cross-layer relations. Both use the same native-write scalar search as Appendix B, with cap 2^{20}, separately on their own incoming states. The experiment also evaluated each control with its scalar capped at one; those arms reach only a small fraction of the treatment's displacement, are not displacement-matched comparators, and are part of the complete record of all

$$77 \times 10 \times (2 + 5 \times 2 \times 2) = 16{,}940$$

outcomes. No parameter or retained subset depends on answer quality.

Both unrestricted constructions match all 61,600 native writes each. Median actual-energy relative error is 0.0429% for common and 0.0476% for independent rotations; the maxima are below 1%. Median scales are 2.3421 and 3.4624, with maxima 129.1018 and 119.2226. However, squared state norm increases in 35,241/61,600 and 48,488/61,600 writes, respectively, compared with 274/12,320 original-treatment writes. Matching applied displacement does not preserve norm reduction, angle, trajectory or alignment with the output.

C.2 Separating a common mean from centered variation

The cPCA extractor uses centered covariance matrices, while the original operator deletes Ph, where $P = U^\top U$. This motivates comparing four operators on the same ten frames:

$$\begin{aligned} h'_{\text{base}} &= h, \\ h'_{\text{original}} &= h - Ph, \\ h'_{\text{centered}} &= h - P(h - \mu), \\ h'_{\text{mean}} &= h - P\mu. \end{aligned} \tag{3}$$

The mean is common across extraction domains and is not chosen using the question's domain label at inference. We use the pre-existing within-domain alternating partition, with 38 and 39 questions. For each evaluation fold, the opposite fold supplies baseline final-token states at precisely the original module outputs. At each site, a mean is first computed in 64-bit floating point within each of the ten construction domains, and those ten means are equally averaged; the resulting μ is converted once to BF16. Mean estimation uses only construction states and domain membership, without answers or predictions. Both fold results are pooled without selecting a fold. The cPCA frames were fitted on extraction halves that contain these questions; the mean is cross-fitted, the frames are not.

Existing complete baseline and original outcomes are reused. Before the centered and mean-only arms are evaluated, all 77 baseline outcomes are reproduced in every repeated run, and one complete original trajectory is reproduced for each frame and fold. Every new arm contains all 770 question–frame pairs. Zero-mean centered deletion reproduces original deletion exactly, zero-mean mean-only

deletion leaves the state unchanged, and answer permutation changes neither the intervention trace nor the logits.

Centered deletion improves accuracy over original deletion by 10.0893 percentage points with 95% interval $[5.9107, 14.1969]$ and reduces choice CE by 0.482171 $[0.398176, 0.574394]$. Answer-set masses are 0.926838, 0.766604, 0.934804 and 0.362243 in the order of Equation 3, so centered deletion also preserves the answer mass that original deletion loses. These descriptive intervals use 2,000 paired question-stratified draws with seed 42, conditional on the already fitted means and frames; uncertainty from fitting new means is not included.

Mean-only subtraction is more damaging than the original operator, but that does not make it an additive causal contribution to the original loss. At a fixed incoming state, real arithmetic gives

$$Ph = P(h - \mu) + P\mu.$$

In BF16 the mean squared residual of this identity is 0.01035 on centered trajectories and 0.04149 on mean-only trajectories. More importantly, earlier writes change later incoming states, so losses across the sixteen layers do not decompose by this identity. Mean-only writes increase squared state norm in 10,475/12,320 cases; centered writes do so in 5,148/12,320. Centered deletion is an affine operation and need not reduce norm relative to the origin. The smaller displacement motivates the crossed comparison below; the four-arm comparison alone does not isolate a direction effect.

Table 10: Direction $\times$ displacement on the development panel of Qwen3.5-9B, 770 cases per arm. Each direction is written at its own displacement or scaled per item and layer to the realized write energy of the other reference arm, A^T for zero and A^C for mean ablation; failed matches stay in the means. Energy: mean squared displacement per write; complete: cases with all sixteen writes within 1%.

Direction written	Displacement	Choice CE	Full NLL	Acc., %	Energy	Complete
Unmodified	—	0.475	0.602	79.46	0.0	—
Zero ablation Ph	own, A^T	1.008	1.604	69.02	422.5	—
Mean ablation $P(h-\mu)$	own, A^C	0.526	0.608	79.11	83.6	—
Mean ablation $P(h-\mu)$	A^T	0.711	0.928	76.27	422.5	724/770
Zero ablation Ph	A^C	0.604	0.862	78.07	83.6	544/770
Mean only $P\mu$	own	3.615	7.428	38.52	839.3	—

C.3 Crossing direction with each reference displacement

A subsequent protocol fixes the same 77 development questions, ten frames, opposite-fold means and sixteen sites before evaluating the two missing direction–dose cells. The original and centered trajectories supply two distinct per-item, per-layer targets, denoted A^T and A^C. At its own incoming state, the new centered-direction arm chooses a nonnegative scalar for $P(h-\mu)$ to match A^T; the new original-direction arm similarly scales Ph to match A^C. The same label-free native-write search, 2^{20} cap, 48 energy evaluations and 1% tolerance apply. Targets do not depend on the new outcomes. The already observed diagonal arms are reused; exact agreement with the reference baseline and the original and centered trajectories verifies that the new comparisons use the same conditions. The questions are development data.

For the rescaled arms targeting A^T and A^C, respectively, the realized energies are 422.487 and 83.635 against reference energies of 422.536 and 83.622; Table 10 lists the outcomes. Centered-minus-original choice loss is -0.296610 with descriptive 95% interval $[-0.406511, -0.189072]$, and the accuracy change is $+7.2500$ points $[3.9817, 10.5188]$. The difference between the direction contrasts at A^T and at A^C is -0.217987 nats $[-0.299782, -0.129134]$. Thus a smaller requested displacement alone does not explain the loss benefit. Intervals pair all arms and frames under 2,000 within-question-domain resamples with seed 42, conditioning on the previously fitted means and bases. They are not adjusted for the exploratory sequence of experiments.

The centered-direction arm matches 12,271/12,320 writes and 724/770 complete trajectories; the rescaled original arm matches 12,041/12,320 writes and 544/770 trajectories. Their maximum relative errors are 3.8346% and 14.2184%, respectively; neither hits the scalar cap. The two new arms increase

squared state norm in 6,090/12,320 and 540/12,320 writes. Equal target displacement therefore does not equate state norm or trajectory.

Failed-match sensitivity. Restricting the A^T contrast to cases whose rescaled arm matched all sixteen writes, with equal weight over the 100 extraction-by-question-domain cells, gives -0.301 $[-0.399, -0.204]$ nats on 724 cases; on the 509 cases matched in both rescaled arms it is -0.289 $[-0.426, -0.157]$. Cases containing a failed write contribute -0.027 nats to the all-case estimate of -0.297, and the rank correlation between a case's largest relative write error and its contrast is -0.06. Intervals use 2,000 question-domain-stratified question resamples shared across frames with seed 20260923. The effect at the zero-ablation displacement is insensitive to matching failures.

C.4 Registered held-out replication

The replication was registered before any of its outcomes existed. For each model and each item split, with seeds 42, 101, 202, 303 and 404, rank-4 cPCA frames for blocks 16–31 are refitted from scratch on the extraction half with the principal-matrix rule, and μ is the domain-balanced mean of final-token block outputs of the 500 extraction-half items on the unmodified forward; no labels and no evaluation items enter either. Every frame D is evaluated on the 500 evaluation-half items of its split, 5,000 cases per split, in four arms: zero ablation, whose realized write energy per item and layer is E_T; mean ablation, with energy E_C, so that E_T and E_C play the roles of A^T and A^C on the development panel; mean ablation rescaled per item and layer to E_T; and zero ablation rescaled to E_C. The rescaled arms use the label-free native-write solver of Appendix B, with 1% tolerance, at most 48 evaluations and cap 2^{20}, on their own trajectories, and every failed match keeps its nearest candidate and stays in all means. Inference is a bootstrap over unique MMLU items within a model with 10,000 draws; an item in several splits is one cluster. The registered rule labels a model as replicating when the upper 95% bounds of Δ_T and Δ_C are both below zero. The number of split seeds and the complete-coverage requirement, namely all 5,000 prescribed extraction-domain $\times$ evaluation-item cases per split, all four arms and sixteen-entry traces, were fixed before any efficacy output; all three models meet it. Exactness checks on a pilot run precede the runs: mean ablation with $\mu = 0$ gives results identical to zero ablation, repeated runs agree exactly, and the refitted frames reproduce the stored treatment outcomes of the principal matrices to within 1.7×10^{-6} nats. The evaluation halves never enter fitting, but they are not a never-seen question pool: all were scored in the principal selectivity matrices, and they contain the development panel on which the mean-ablation protocol was developed. Matching the question text of the panel within domains locates 76 of its 77 questions as 77 item identities, one question text occurring as two items, in the evaluation halves of splits 42, 101, 202, 303 and 404 with 1, 43, 35, 42 and 46 items, that is, 1,670 of the 25,000 cases per model; excluding them leaves 23,330 cases and 895 unique items. The remaining question lies in the extraction half of all five splits. The supplementary material contains the question-to-item mapping. Tables 11 and 12 list the principal registered endpoints, both sensitivities, matching and state-norm counts and the per-split estimates; the complete registered output, including own-minus-other and baseline-excess contrasts, is in the supplementary material.

Table 11: Registered held-out experiment: principal registered endpoints, both sensitivities, matching and state-norm counts; choice CE in nats unless marked; 95% item-cluster bootstrap intervals with 10,000 draws and seed 20260923. Δ_C: mean ablation minus zero ablation, both at the mean-ablation displacement. Counts are over 400,000 writes and 25,000 cases per arm. Z and M: zero and mean ablation; rescaled M or Z: that direction scaled to the other arm's displacement. The development-item exclusion is a labelled post-hoc sensitivity. The complete registered output is in the supplementary material.

	Qwen3.5-9B	Llama-3.1-8B	Olmo-3-1025-7B
Δ_T	-0.239 $[-0.267, -0.212]$	-0.778 $[-0.811, -0.746]$	-0.812 $[-0.855, -0.772]$
Δ_C	-0.032 $[-0.044, -0.021]$	-0.078 $[-0.094, -0.064]$	-0.005 $[-0.013, +0.002]$
$\Delta_T - \Delta_C$	-0.207 $[-0.234, -0.181]$	-0.699 $[-0.725, -0.674]$	-0.807 $[-0.847, -0.770]$
Mean minus zero, own displacements	-0.563 $[-0.595, -0.531]$	-1.182 $[-1.216, -1.149]$	-0.870 $[-0.914, -0.829]$
Δ_T, accuracy, points	$+6.4$ $[+5.5, +7.3]$	$+4.5$ $[+3.7, +5.4]$	$+10.8$ $[+9.4, +12.1]$
Δ_C, accuracy, points	$+0.8$ $[+0.4, +1.2]$	$+2.4$ $[+1.7, +3.2]$	$+0.2$ $[-0.3, +0.7]$
Mean minus zero, accuracy, points	$+11.3$ $[+10.1, +12.5]$	$+11.5$ $[+10.1, +13.0]$	$+13.1$ $[+11.4, +14.8]$
Δ_T, full NLL	-0.565 $[-0.625, -0.509]$	-1.113 $[-1.150, -1.078]$	-0.856 $[-0.899, -0.816]$
Δ_C, full NLL	-0.176 $[-0.217, -0.139]$	-0.079 $[-0.095, -0.064]$	-0.007 $[-0.015, +0.000]$
Matched writes, rescaled M / Z	398,700 / 390,589	399,300 / 390,835	399,918 / 393,882
Complete cases, M / Z / both	23,740 / 17,490 / 16,682	24,326 / 17,621 / 17,196	24,918 / 19,481 / 19,424
Writes increasing $\|h\|^2$, Z / M	8,038 / 166,541	11,488 / 155,437	6,877 / 159,537
Writes increasing $\|h\|^2$, rescaled M / Z	201,005 / 17,035	191,662 / 5,982	194,070 / 1,944
Δ_T, cases complete in both	-0.216 $[-0.244, -0.187]$	-0.724 $[-0.763, -0.688]$	-0.778 $[-0.820, -0.738]$
Δ_C, cases complete in both	-0.030 $[-0.043, -0.018]$	-0.072 $[-0.089, -0.056]$	-0.006 $[-0.013, +0.002]$
Δ_T, excluding 77 development items	-0.240 $[-0.268, -0.211]$	-0.780 $[-0.814, -0.747]$	-0.810 $[-0.854, -0.768]$
Δ_C, excluding 77 development items	-0.031 $[-0.043, -0.020]$	-0.080 $[-0.096, -0.064]$	-0.004 $[-0.012, +0.003]$

Table 12: Registered held-out experiment per item split seed: Δ_T and Δ_C in choice CE, nats, 5,000 cases each.

	Qwen3.5-9B		Llama-3.1-8B		Olmo-3-1025-7B	
Split seed	Δ_T	Δ_C	Δ_T	Δ_C	Δ_T	Δ_C
42	-0.313	-0.052	-0.966	-0.077	-0.755	0.007
101	-0.103	-0.012	-0.705	-0.104	-0.804	-0.022
202	-0.204	-0.011	-0.738	-0.091	-0.680	-0.004
303	-0.217	-0.042	-0.705	-0.072	-0.807	0.004
404	-0.359	-0.043	-0.774	-0.047	-1.015	-0.009

C.5 Registered extension to further model families

A registered extension, written before any of its outputs existed, runs the held-out design of Appendix C.4 unchanged on two further families: the base model Mistral-7B-v0.3, with 32 blocks and intervention at blocks 16–31, and the post-trained Qwen3-8B, with 36 blocks and intervention at blocks 18–35, the upper half by the same rule. The splits, frame fitting, mean, four arms, solver, failed-match handling, estimator and item-cluster bootstrap are those of the held-out experiment. Its extension rule states that the direction effect at the zero-ablation displacement extends to a model if all five of its splits are complete and the upper 95% bound of Δ_T is below zero; the registered replication rule of Appendix C.4 is also applied, and its labels are listed in Table 20. All five splits of both models are complete, and Table 13 lists their registered endpoints. The extension rule holds for both models, and Qwen3-8B also meets the replication rule, with $\Delta_C = -0.199$ $[-0.235, -0.166]$ nats. On Qwen3-8B, mean ablation leaves choice CE within 0.023 nats of the unmodified model while zero ablation adds 1.655 nats.

Table 13: Registered extension to two further model families, five splits and 25,000 cases on 972 items per model and arm; choice CE in nats unless marked; 95% item-cluster bootstrap intervals, 10,000 draws, seed 20260923. Write counts are over 400,000 writes per rescaled arm for Mistral-7B-v0.3 and 450,000 for Qwen3-8B, whose 18 intervened blocks give 18 writes per case.

	Mistral-7B-v0.3	Qwen3-8B
Choice CE, unmodified / zero / mean ablation	0.994 / 1.327 / 1.044	1.070 / 2.725 / 1.048
Accuracy, unmodified / zero / mean ablation, %	56.7 / 47.0 / 55.3	70.4 / 51.5 / 68.5
E_C/E_T	0.060	0.049
Δ_T	-0.248 $[-0.271, -0.225]$	-1.127 $[-1.178, -1.077]$
Δ_T, accuracy, pp	$+6.1$ $[+5.0, +7.1]$	$+9.4$ $[+8.5, +10.3]$
Δ_T, full NLL	-0.256 $[-0.279, -0.233]$	-1.296 $[-1.351, -1.243]$
$\Delta_T - \Delta_C$	-0.243 $[-0.257, -0.229]$	-0.928 $[-0.984, -0.872]$
Mean minus zero, own displacements	-0.282 $[-0.306, -0.259]$	-1.677 $[-1.734, -1.623]$
Δ_T per split, seed 42	-0.244	-1.211
seeds 101 / 202	-0.207 / -0.295	-1.158 / -0.939
seeds 303 / 404	-0.205 / -0.288	-1.558 / -0.771
Δ_T, cases complete in both rescaled arms	-0.230 $[-0.253, -0.208]$	-1.181 $[-1.248, -1.115]$
Δ_T, excluding 77 development items	-0.250 $[-0.273, -0.226]$	-1.132 $[-1.184, -1.081]$
Matched writes, rescaled mean / rescaled zero	399,676 / 392,881	449,643 / 434,905
Extension rule	holds	holds

C.6 Where the direction effect holds

The registered analysis pools each model's 25,000 cases. To show where the direction effect at the zero-ablation displacement holds, we split the same per-case results by split seed, by quintile of the case's zero-ablation displacement, by own- versus other-domain evaluation, by extraction domain D and by evaluation domain, with the registered estimator and item-cluster bootstrap; Table 14 summarizes this post-hoc breakdown, which was not registered. The mean-ablation direction is less damaging, with the upper 95% bound below zero, in all fifteen model–split replicates and all fifteen model–quintile subsets, including the lowest-displacement quintile, where mean ablation writes the largest share of the zero-ablation displacement, $E_C/E_T = 0.41$, 0.08 and 0.10; for own- and for other-domain evaluation in every model; and in 27 of 30 model–extraction-domain and 28 of 30 model–evaluation-domain subsets. For 73%, 94% and 95% of items the mean contrast over their cases is negative. Effect sizes differ widely between frames, resolved Llama extraction domains ranging from -0.017 to -2.541 nats, so the pooled Δ_T averages over frames of very different sensitivity.

Table 14: Post-hoc descriptive breakdown of the registered held-out Δ_T in choice CE, nats, defined as mean-ablation minus zero-ablation direction, both at the zero-ablation displacement, from the same per-case results with the registered item-cluster bootstrap: 95%, 10,000 draws, seed 20260923. Not part of the registered analysis; no labels are assigned. Resolved: upper 95% bound below zero. Quintiles are taken within model over a case's zero-ablation write energy summed over the sixteen blocks, 5,000 cases each; own domain: $E = D$, 2,500 cases; other: $E \neq D$, 22,500 cases. Every subset's estimate and interval, and the corresponding Δ_C, interaction, accuracy and full-NLL breakdowns, are in the supplementary material.

	Qwen3.5-9B	Llama-3.1-8B	Olmo-3-1025-7B
Split seeds resolved	5/5	5/5	5/5
range of estimates	-0.359 to -0.103	-0.966 to -0.705	-1.015 to -0.680
Displacement quintiles resolved	5/5	5/5	5/5
range of estimates	-0.448 to -0.100	-1.180 to -0.545	-1.038 to -0.340
Own domain	-0.200 $[-0.271, -0.129]$	-0.792 $[-0.920, -0.668]$	-0.767 $[-0.866, -0.670]$
Other domains	-0.244 $[-0.274, -0.214]$	-0.776 $[-0.812, -0.740]$	-0.817 $[-0.862, -0.776]$
Extraction domains resolved	8/10	9/10	10/10
Evaluation domains resolved	8/10	10/10	10/10
Items with mean $\Delta_T < 0$	707/972, 72.7%	914/972, 94.0%	921/972, 94.8%

C.7 Registered same-type single-block scan

The same registration also specifies a same-type single-block scan, which asks whether the damage concentrated at the last block of a stack reflects depth or attention type. Qwen3.5-9B interleaves linear-attention blocks with a full-attention block every fourth block, so blocks 23, 27 and 31 are full

and 28–30 linear; Olmo-3 interleaves sliding-window attention, whose window is longer than every prompt here, with full attention in the same pattern; Llama-3.1 uses full attention throughout. On the seed-42 split, with the refitted frames of Appendix C.4, each frame is zero-ablated at one block only and compared with a Haar-random rank-4 frame matched per item to the same realized write energy by the same solver; every random write matched. The registered reading rule is: depth or position if, on Qwen, blocks 27 and 30 each cost less than $0.25\times$ block 31 and, on Llama and Olmo, block 30 costs less than $0.25\times$ block 31; attention type if, on Qwen, block 27 costs at least $0.5\times$ block 31 while blocks 28–30 cost less than $0.25\times$; otherwise mixed. Table 15 lists every block; the paired contrast of block 30 minus block 31 is -0.658 $[-0.702, -0.614]$, -0.792 $[-0.835, -0.751]$ and -0.310 $[-0.335, -0.286]$ nats. On Qwen the ratios of blocks 27 and 30 to block 31 are 0.020 and 0.050; on Llama and Olmo the block-30 ratios are 0.250 and 0.560; the registered label is listed in Table 20.

Table 15: Registered same-type single-block scan on the seed-42 split with refitted frames, 500 evaluation-half items $\times$ 10 frames $=$ 5,000 cases per block. Zero ablation of the rank-4 frame at one block only; the random arm is a Haar-random rank-4 frame matched per item to the same realized write energy. Damage: mean choice-CE increase over the unmodified model in nats, with 95% item-cluster bootstrap intervals; ratio: zero-ablation damage relative to block 31 of the same model.

Model	Block, type	Zero-ablation damage	Matched random damage	Acc. drop, pp	Ratio to block 31
Qwen3.5-9B	23, full	0.006 $[-0.000, 0.012]$	0.002 $[-0.001, 0.004]$	1.0	0.009
	27, full	0.014 $[0.007, 0.021]$	0.001 $[-0.001, 0.004]$	1.6	0.020
	28, linear	0.027 $[0.017, 0.037]$	0.002 $[-0.001, 0.004]$	1.6	0.039
	29, linear	0.027 $[0.019, 0.035]$	0.003 $[0.000, 0.005]$	1.5	0.039
	30, linear	0.034 $[0.026, 0.043]$	0.001 $[-0.001, 0.004]$	1.9	0.050
	31, full	0.693 $[0.646, 0.740]$	0.063 $[0.052, 0.076]$	15.6	1.000
Llama-3.1-8B	27, full	0.290 $[0.267, 0.312]$	0.000 $[-0.003, 0.003]$	7.3	0.275
	28, full	0.248 $[0.228, 0.268]$	0.002 $[-0.001, 0.005]$	6.2	0.235
	29, full	0.282 $[0.260, 0.304]$	0.002 $[-0.001, 0.004]$	7.7	0.267
	30, full	0.264 $[0.247, 0.282]$	0.003 $[-0.001, 0.006]$	7.7	0.250
	31, full	1.056 $[1.008, 1.106]$	0.123 $[0.110, 0.137]$	15.3	1.000
Olmo-3-1025-7B	27, full	0.055 $[0.042, 0.068]$	0.000 $[-0.001, 0.001]$	2.3	0.078
	28, sliding	0.067 $[0.054, 0.080]$	0.002 $[0.001, 0.004]$	2.6	0.096
	29, sliding	0.143 $[0.124, 0.163]$	0.003 $[0.001, 0.005]$	4.5	0.203
	30, sliding	0.395 $[0.360, 0.430]$	0.012 $[0.009, 0.014]$	10.0	0.560
	31, full	0.705 $[0.649, 0.760]$	0.029 $[0.023, 0.034]$	12.0	1.000

C.8 Registered site localization and per-write audit

A further registered analysis, specified before any of its runs, asks whether the direction effect at the zero-ablation displacement needs the final intervened block. On the seed-42 split, with the refitted frames and five-shot evaluation-format mean of Appendix C.4, Qwen3.5-9B and Llama-3.1-8B run the four arms in two conditions on the same 5,000 cases: R, blocks 16–31 as in the held-out protocol, and P, blocks 16–30 with the final block not intervened. The solver, tolerance, cap and failure handling are unchanged. Inference is a bootstrap over the 500 evaluation items with 10,000 draws and 95% percentile intervals; an item is one unit across its ten extraction domains and both conditions. The registered rule L1 states that the direction effect is present without the final block if the upper bound of Δ_T^P is below zero, and otherwise scopes the claim to protocols that include that block; $I_{\mathrm{pos}} = \Delta_T^P - \Delta_T^R$ and the damage ratio D^P/D^R are descriptive. Table 16 lists the results. L1 holds on both models. Condition P follows the trajectory of R up to block 30, so its write records equal the first fifteen layers of R's.

Per-write audit. Every write of the two rescaled arms stores its target, the realized write energy of the reference arm at the same item and layer, and its realized energy $\|\,\mathrm{cast}(h') - h\|^2$. Recomputing the match status of each write as $|A/A^{\mathrm{target}} - 1| \le 0.01$, independently of the recorded decision, confirms the recorded decision for all $4 \times 80{,}000$ writes of condition R and all $4 \times 75{,}000$ writes of P, whose records equal the first fifteen layers of R's. The median absolute relative error is 0.16% for the mean-ablation direction rescaled to E_T and 0.33% and 0.37% for the zero-ablation direction rescaled to E_C on Qwen3.5-9B and Llama-3.1-8B; the 99th percentiles are 0.95% for the mean-ablation direction on both models and 2.3% and 4.8% for the zero-ablation direction on Qwen3.5-9B and Llama-3.1-8B. No miss has a zero target or a target out of reach within the cap; the solver attributes

every miss to BF16 rounding of the written state or to the evaluation budget, and keeps its nearest candidate. Per-layer realized-to-target energy ratios lie in $[0.9998, 1.0015]$. The registration also listed per-layer nominal-to-realized ratios; nominal update energies are not among the recorded quantities, so the per-layer diagnostic is the realized-to-target ratio above. On the same cases, condition R and the held-out seed-42 analysis give Δ_T of -0.317 versus -0.313 and -0.963 versus -0.966 nats and zero-ablation damage of 0.624 versus 0.622 and 1.568 versus 1.550; case-level differences between the two are listed in the supplementary material. The supplementary material contains the per-case results of both conditions; for every write of condition R, the realized-to-target energy ratio, the recorded match decision and the solver's reason for a miss, with per-layer energy totals; and a script that recomputes Table 16 and re-derives every decision of condition R from the ratios. The audit uses the separate target and realized energies.

Table 16: Registered site localization on split seed 42; choice CE in nats, accuracy in points; 95% item bootstrap intervals. R: blocks 16–31; P: blocks 16–30; both with the five-shot evaluation-format mean. Zero damage: CE of zero ablation minus CE of the unmodified model. Matched: writes within 1% of target for the mean-ablation direction at E_T / the zero-ablation direction at E_C, misses retained.

	Qwen3.5-9B		Llama-3.1-8B	
	R	P	R	P
Zero damage	0.624 $[0.579, 0.668]$	0.101 $[0.083, 0.120]$	1.568 $[1.508, 1.628]$	0.752 $[0.708, 0.794]$
Δ_T	-0.317 $[-0.357, -0.276]$	-0.057 $[-0.076, -0.040]$	-0.963 $[-1.025, -0.901]$	-0.638 $[-0.684, -0.592]$
Δ_C	-0.049 $[-0.066, -0.034]$	-0.028 $[-0.045, -0.013]$	-0.077 $[-0.101, -0.054]$	-0.053 $[-0.078, -0.029]$
Δ_T, acc.	$+8.2$ $[+6.7, +9.8]$	$+1.4$ $[+0.5, +2.3]$	$+5.0$ $[+3.4, +6.6]$	$+7.8$ $[+5.6, +10.0]$
Δ_C, acc.	$+1.7$ $[+1.0, +2.5]$	$+0.9$ $[+0.1, +1.7]$	$+2.1$ $[+0.8, +3.5]$	$+1.9$ $[+0.4, +3.4]$
E_C/E_T	0.203	0.328	0.030	0.122
Matched writes	79,717 / 78,024 of 80,000	74,717 / 73,110 of 75,000	79,793 / 77,808 of 80,000	74,793 / 72,820 of 75,000
I_{pos}	$+0.260$ $[+0.217, +0.302]$		$+0.325$ $[+0.285, +0.365]$	
D^P/D^R	0.162 [0.138, 0.187]		0.480 [0.462, 0.497]	
Rule L1	present without final block		present without final block	

C.9 Registered extensions to new questions, a fitting-format mean and another extractor

Two further registered extensions, each specified before any of its runs, test the two main limits of the held-out experiment on Qwen3.5-9B and Llama-3.1-8B, with the seed-42 frames of Appendix C.4 and the four arms, solver, tolerance, cap and failure handling unchanged. Condition N keeps the five-shot evaluation-format mean and evaluates every MMLU test question after the first 100 of its domain; every earlier experiment of this study used only the first 100 questions of each domain, and none of the new questions repeats the text of a pool question. Abstract algebra, college chemistry and college computer science have no further test questions, so the 407 new questions come from the other seven domains, 8 to 211 per domain, and each is evaluated under all ten frames, 4,070 cases per model. Its rules are N1, under which the direction effect at the zero-ablation displacement is present on questions outside the examined pool if the upper 95% bound of Δ_T^N is below zero, and N2, the replication rule of Appendix C.4 applied to Δ_T^N and Δ_C^N. Condition F keeps the 5,000 seed-42 cases and replaces the reference mean by the domain-balanced mean of the zero-shot extraction-format states of the extraction half, the states on which the frames are fitted; its rule L2 states that the direction effect is present with a fitting-format mean if the upper 95% bound of Δ_T^F is below zero, and $I_{\text{fmt}} = \Delta_T^F - \Delta_T^R$, paired with condition R of Appendix C.8 on the same cases, is descriptive. Inference is the item bootstrap of Appendix C.8 with 10,000 draws. Table 17 lists the results; N1, N2 and L2 hold on both models. Per evaluation domain, Δ_T^N is negative in all seven domains of both models, from -0.118 to -0.399 nats on Qwen3.5-9B and from -0.590 to -1.380 nats on Llama-3.1-8B. The registered analysis re-derives every match decision of both conditions from the target and realized energies and confirms every recorded match decision; the supplementary material holds the case-level outcomes of N and F, and its per-write records cover condition R of Appendix C.8.

A further registered extension, also specified before its runs, keeps condition R but replaces each cPCA frame by a mean-difference frame, the top-4 right singular vectors of the nine differences

between the mean zero-shot state of the extraction domain and that of each other domain, and applies rule E1x: the direction effect is present for this extractor if the upper 95% bound of Δ_T is below zero. The registered analysis covers Llama-3.1-8B with this frame. With this frame zero ablation costs 0.197 $[0.158, 0.235]$ nats, mean ablation writes 7% of its squared displacement, and at the zero-ablation displacement the mean-ablation direction lowers choice CE by 0.166 $[0.129, 0.202]$ nats and raises accuracy by 3.2 $[1.1, 5.3]$ points, so E1x holds; 79,963 and 78,616 of the 80,000 writes of the two rescaled arms match.

Table 17: Registered extensions on the seed-42 frames; choice CE in nats, accuracy in points; 95% item bootstrap intervals. N: MMLU test questions outside the examined pool with the five-shot evaluation-format mean; F: the 5,000 seed-42 cases with a zero-shot fitting-format mean. U, Z, M: unmodified model, zero and mean ablation. Matched: writes within 1% of target for the mean-ablation direction at E_T and the zero-ablation direction at E_C, misses retained.

	Qwen3.5-9B		Llama-3.1-8B	
	N	F	N	F
Cases, questions	4,070 / 407	5,000 / 500	4,070 / 407	5,000 / 500
Choice CE, U/Z/M	0.475/0.872/0.529	0.644/1.269/0.730	0.696/2.132/0.857	0.969/2.537/1.266
Accuracy, U/Z/M	84.0/72.9/81.6	75.4/61.2/72.6	70.3/49.0/65.3	55.6/40.7/47.5
E_C / E_T	0.295	0.431	0.037	0.211
Δ_T	-0.187 $[-0.230, -0.145]$	-0.276 $[-0.315, -0.238]$	-0.901 $[-0.975, -0.826]$	-0.705 $[-0.753, -0.656]$
Δ_C	-0.048 $[-0.075, -0.023]$	-0.098 $[-0.121, -0.077]$	-0.052 $[-0.079, -0.027]$	-0.217 $[-0.244, -0.189]$
Δ_T, acc.	$+5.0$ $[+3.5, +6.5]$	$+5.7$ $[+4.2, +7.1]$	$+7.1$ $[+5.2, +8.9]$	$+2.0$ $[+0.3, +3.6]$
Δ_C, acc.	$+1.6$ $[+0.7, +2.4]$	$+2.8$ $[+1.8, +3.7]$	$+3.6$ $[+2.0, +5.2]$	$+1.5$ $[+0.2, +3.0]$
Matched writes	64,835/63,755 of 65,120	79,509/79,173 of 80,000	64,954/63,404 of 65,120	79,811/78,479 of 80,000
I_{fmt}	—	$+0.041$ $[+0.015, +0.065]$	—	$+0.258$ $[+0.214, +0.302]$
Rules	N1, N2 hold	L2 holds	N1, N2 hold	L2 holds

D EXPERIMENTAL SPECIFICATIONS AND REPRODUCIBILITY

Checkpoints. Qwen3.5-9B is post-trained and uses the final-norm scale convention $1 + w$; Olmo-3-1025-7B and Llama-3.1-8B are base models with scale w. Each is the public release checkpoint of that name, used with the public MMLU test and dev splits; all runs use bfloat16 weights and activations. The registered extension of Appendix C.5 adds the base model Mistral-7B-v0.3 and the post-trained Qwen3-8B.

Experimental groups. Principal matrices: the three-model matrices and the related runs listed in Table 18. Qwen diagnostics: the dose, site and calibration diagnostics on Qwen3.5-9B. Development-panel experiments: the geometry and mean-ablation experiments. Held-out experiments: the registered held-out experiment and its registered extensions. Every paired comparison lies within one group.

Fixed protocol fields of all online-calibrated runs. Ten domains: abstract algebra, anatomy, astronomy, college chemistry, college computer science, formal logic, jurisprudence, logical fallacies, machine learning and philosophy; the source pool comprises the first 100 test questions per domain; extraction and evaluation halves of 50 questions each by seeded permutation; five-shot evaluation prompt from the domain's dev split, verified to contain exactly five demonstrations; maximum prompt length 4,096 tokens; rank k=4; blocks 16–31, or block 16 only for the single-block runs; final prompt token; contrast pool capped at 200 prompts, 198 in practice; five controls per extraction row; covariance-aware controls drawn in the top-16 PCA span of a random half of the contrast pool from 48 candidates; calibration on up to 50 extraction-half prompts, online calibration on 32 of them, tolerance 0.1, four iterations, $\alpha \leq 4$, or 64 with clipping allowed for the isotropic and persistent controls; treatment scale 1; whitespace-prefixed answer-letter scoring.

Experimental coverage. Table 18 lists every run of the principal matrices used in the paper with its model, control and completed evaluation coverage; the Qwen diagnostic runs are identified where they are reported in Section 4 and Appendices A and G. For each pair in the seven primary ten-domain matrices, the supplementary data contain per-item choice-CE and accuracy damage for the treatment

and each control. Per-question results of the stack-without-last-block follow-up and the per-layer histories of the paired forward calibration are supplied as well, as described in Appendices A and B.

Table 18: Runs of the principal matrices used in the paper; the Qwen diagnostic runs are identified where they are reported. "Cells" counts completed extraction-by-evaluation cells, of 100 for the ten-domain runs and of 2,209 for the 47-domain run.

Experiment	Model	Seed	Control / extractor / site	Cells	Used in
Calibrated matrix	Qwen3.5-9B	42, 101	cov, online / cPCA / blocks 16–31	100, 100	Tables 2, 6, 7
Calibrated matrix	Olmo-3-1025-7B	42, 101	cov, online / cPCA / 16–31	100, 100	Tables 2, 6, 7
Calibrated matrix	Llama-3.1-8B	42, 101, 202	cov, online / cPCA / 16–31	100, 100, 100	Tables 2, 6, 7
Calibrated matrix	Qwen3.5-9B	42	iso, online, $\alpha \leq 64$ / cPCA / 16–31	100	Table 7
Fixed-frame matrix	Qwen3.5-9B	42, 1042	persistent, online / cPCA / 16–31	100, 100	Section 4.2
Fixed-frame matrix	Llama-3.1-8B	42	persistent, online / cPCA / 16–31	100	Section 4.2
Fixed-frame matrix	Olmo-3-1025-7B	42	persistent, online / cPCA / 16–31	100	Section 4.2
Fixed-frame matrix	Qwen3.5-9B	47042	persistent, online / cPCA / 16–31, 47 domains	2,209	Table 21
Extractor comparison	Qwen3.5-9B	42	cov, online / DiffMean+3 cPCA / 16–31	100	Table 7
Single-site comparison	Qwen3.5-9B	42	cov, online / cPCA / block 16	100	Table 4
Single-site comparison	Llama-3.1-8B	42	cov, online / cPCA / block 16	100	Table 4
Offline calibration	Llama-3.1-8B	42	cov, readout-orthogonal, offline / cPCA / 16–31	100	Table 2
Offline calibration	Qwen3.5-9B	42, 123, 456	cov, readout-orthogonal, offline / cPCA / 16–31	54, 54, 50	Table 2

Cost accounting. Table 19 counts model forward passes; the per-write scalar search runs on the current state inside a forward pass and requires no additional forward passes, but a rescaled arm needs the per-item, per-layer targets of its reference arm, so it uses more information than a comparator with a scale fixed in advance.

Table 19: Model forward passes and scalar evaluations of the calibration and matching procedures.

Procedure	Forwards and evaluations
Held-out experiment, per model and split	3,480 preparation forward passes for baselines, mean estimation and frame fitting, and 20,000 intervention forward passes, one for each of four arms $\times$ 5,000 cases; 352,200 forward passes in total over 15 model–split runs
Actual-write diagnostic, one row	500 treatment-reference and 2,500 control forward passes; 500 baseline outcomes reused after exact checks; at most 48 scalar candidates per write, on the current state
Forward sequential calibration, 5 frames	3 scalar evaluations at each of 80 control–layer steps, 240 in total with a budget of 48 per layer, each on the 32 calibration prompts with earlier layers fixed
Simultaneous online calibration	up to five measurement passes over the 32 calibration prompts with four multiplicative updates of all layers

Comparison index. Every comparison we interpret is paired within one group. Offline versus online calibration, principal matrices: same treatment results, items and seeds; the controls differ in calibration and readout orthogonalization. Dose–response, site scan and stack without the last block, Qwen diagnostics: same frames and items, with exact agreement of repeated runs. Sequential versus simultaneous calibration and actual-write versus population-average matching, Qwen diagnostics: same reconstructed frames, prompts and items; the procedures apply the scalar at different stages, as described in Appendix B. Geometry controls and the direction $\times$ displacement cross, development-panel experiments: same 77 questions, frames, means and sites, with verified identical reference conditions. Held-out replication, its extensions and the same-type scan, held-out experiments: refitted frames and means per split, all arms on the same cases. Numbers from different groups, for example block-31 damage in Tables 5 and 15, use different frames or controls and are not paired.

Analysis and prespecification. Numbers in Tables 2, 4, 6 and 7 and in Figure 2 are computed from the measurements above. The primary endpoint, choice CE, the secondary endpoint, accuracy, the smallest effects of interest of ± 0.02 nats and ± 1 pp, the crossed-bootstrap units and the classification rules were specified before any online-calibrated matrix was complete. Further experiments, namely the format and attribute positive controls, a treatment-dose gradient at $\alpha < 1$, the control dose–response, the site scan and its follow-up, a per-column recalibration, and control-dose and harm-matched variants of the attribute control, were specified, with their decision rules, in written experimental protocols before their data collection; Appendix Tables 20 and 21 list every registered test with its outcome. The analysis rule of the 47-domain matrix was specified while data collection was in progress and before any of its outcomes had been examined; its 46 domain residuals reproduce

the reported aggregate. The seed list 42, 101, 202 was fixed before data collection; seeds 123 and 456 belong to the offline-calibrated runs of Table 2.

E ANSWER-TOKEN SCORING

The five-shot demonstrations end in "Answer: A", and every experiment scores the whitespace-prefixed answer letter, a space followed by the letter, which is a single vocabulary item distinct from the bare letter in every tokenizer used. For the SentencePiece tokenizer of Mistral-7B-v0.3, the scored piece must be the whitespace-prefixed letter and must equal the last token of "Answer: A"; the supplementary code verifies this token identity. Scoring the bare letter instead reads a token of negligible probability: on Llama-3.1-8B its baseline full-vocabulary NLL is 14.9–19.5 nats, a probability of order 10^{-7}, so its NLL change measures tail mass rather than answer likelihood, and four-way choice CE computed from bare-letter logits is only a proxy for the decision among the spaced letters. We recommend checking the scored token against the model's actual continuation before any likelihood-based intervention analysis.

F RECOMMENDATIONS FOR MULTI-LAYER ABLATION STUDIES, IN FULL

For multi-layer subspace interventions we recommend seven practices. Measure both the actual before–after displacement and the incoming state norm at each site. State whether the control matches rank, within-layer geometry, cross-layer geometry, displacement or norm reduction, since these are different constraints. Use single-site and stacked comparisons to locate sensitivity without treating them as equal-dose experiments. Report choice outcomes and full-vocabulary outcomes with their exact mass decomposition. Use crossed comparisons with the appropriate shared-item uncertainty and an explicit fixed-domain or population-of-domains estimand. Include a positive control with a known selective effect. State the reference value that replaces a deleted component, and match it between treatment and control. Native finite-precision matching of the realized write should be verified directly, as in Appendix B.

G REGISTERED TESTS

Tables 20 and 21 list every test of this project that was registered with a decision rule before its data were collected or inspected, with the registered estimate and the label assigned by the rule. S is the diagonal-minus-off-diagonal contrast of Equation 1; brackets are 90% crossed-bootstrap intervals unless marked otherwise; smallest effects of interest ± 0.02 nats and ± 1 point. Choice-CE labels: "equivalent" if the interval lies inside ± 0.02, "excl. $\geq$+0.02" if the upper bound is below $+0.02$, otherwise "inconclusive"; positive-control labels: PASS if the lower bound exceeds $+0.02$ with the diagonal top-1 in a majority of rows and the accuracy lower bound above $+1$ point, FAIL if the upper bound is below $+0.02$, otherwise indeterminate; negative controls are clean if $|S_{\mathrm{CE}}| < 0.02$ and the interval covers zero. Harm-matched controls are scaled to match the treatment's damage rather than its dose; "gate missed" marks a seed whose controls did not meet the registered calibration gate, which receives no label, shown as –; "indet." abbreviates indeterminate; Neg. 1 and Neg. 2 are the two harm-matched negative controls. The Setting column names the experimental group of Appendix D. For the format positive control, resampling whole questions across the three format columns, which reuse the same 300 questions, gives $[+0.95, +1.54]$ for S_{CE} and $[+13.6, +21.2]$ for S_{acc}.

Table 20: All registered tests and their registered outcomes.

Test	Setting	Registered estimates	Label
Mean versus zero ablation and deletion site			
Held-out mean vs. zero ablation	3 models × 5 splits, held-out experiments	Δ_T: Qwen -0.239 $[-0.267, -0.212]$, Llama -0.778 $[-0.811, -0.746]$, Olmo -0.812 $[-0.855, -0.772]$; Δ_C: -0.032 $[-0.044, -0.021]$, -0.078 $[-0.094, -0.064]$, -0.005 $[-0.013, +0.002]$, 95%; cross-architecture rule requiring replication on Llama and Olmo	replicates, replicates, partial; rule not met
Same-type single-block scan	3 models, split 42, held-out experiments	damage ratio to block 31: Qwen block 27 0.020, block 30 0.050; Llama block 30 0.250; Olmo block 30 0.560	Qwen: depth pattern; Llama, Olmo: block-30 ratio not below 0.25; reading "mixed"
Site localization, rule L1	Qwen, Llama, split 42	Δ_T^P: -0.057 $[-0.076, -0.040]$, -0.638 $[-0.684, -0.592]$, 95%	present on both
Architecture extension, rule E1	Mistral-7B-v0.3, 5 splits, held-out experiments	Δ_T -0.248 $[-0.271, -0.225]$; Δ_C -0.005 $[-0.018, +0.007]$, 95%	E1 holds; replication rule: partial
	Qwen3-8B, 5 splits, held-out experiments	Δ_T -1.127 $[-1.178, -1.077]$; Δ_C -0.199 $[-0.235, -0.166]$, 95%	E1 holds; replication rule: replicates
Reference format, rule L2	Qwen, Llama, split 42	Δ_T^F: -0.276 $[-0.315, -0.238]$, -0.705 $[-0.753, -0.656]$, 95%	present on both
New questions, rules N1, N2	Qwen, Llama, 407 questions	Δ_T^N: -0.187 $[-0.230, -0.145]$, -0.901 $[-0.975, -0.826]$; Δ_C^N: -0.048 $[-0.075, -0.023]$, -0.052 $[-0.079, -0.027]$, 95%	N1 present, N2 replicates, on both
Other frame extractors, rule E1x	Llama, mean-difference frame, split 42	Δ_T^M -0.166 $[-0.202, -0.129]$; Δ_C^M -0.007 $[-0.020, +0.007]$, 95%	present
	Qwen with this frame; standard PCA frame on both models	outside the registered analysis	no label
Single-block site scan	Qwen 42, Qwen diagnostics	Table 5	site-dependent
Stack without last block	Qwen 42 / 101, Qwen diagnostics	$R = 0.170$ $[0.149, 0.190]$ / 0.173 $[0.139, 0.204]$	last block carries; point label, both seeds
Dose			
Control dose–response	Qwen diagnostics	damage 0.18, 0.72, 2.28 nats at 0.5, 1, 2×; slope 1.05 [0.94, 1.16] per doubling	increasing
Treatment-dose gradient, $\alpha < 1$	Olmo α=0.5, principal matrices	S_{CE} -0.019 $[-0.103, +0.064]$; S_{acc} $+1.75$ $[-2.26, +5.68]$	inconclusive on both
	Qwen α=0.25/0.5/1, Qwen diagnostics	S_{CE} $+0.021$ $[-0.023, +0.065]$, -0.063 $[-0.139, +0.011]$, -0.073 $[-0.233, +0.084]$; S_{acc} $+1.14$ $[-1.11, +3.35]$, -0.59 $[-3.60, +2.37]$, -0.67 $[-4.61, +3.23]$; cross-model rule requiring a positive value at low dose on both models	inconclusive on both; rule not met
Per-column recalibration	Qwen 101 / 42, Qwen diagnostics	shift in S_{CE} $+0.017$ / -0.008, replacement if ≥ 0.02; per-column S_{CE} -0.011 $[-0.125, +0.103]$ / -0.081 $[-0.239, +0.072]$	row-wise values kept
Control dose on attribute control	Qwen 42, Qwen diagnostics	shift in S_{CE} at half / twice the dose: $+0.15$ $[+0.07, +0.25]$ / -0.16 $[-0.34, +0.01]$	reported

Table 21: All registered tests and their registered outcomes, continued from Table 20.

Test	Setting	Registered estimates	Label
Positive and negative controls of the crossed assay			
Format positive control	Qwen 42, Qwen diagnostics	S_{CE} $+1.232$ $[+0.951, +1.536]$; top-1 3/3; S_{acc} $+17.4$ $[+13.7, +21.1]$	PASS
RAVEL attribute positive control	Qwen, seeds 42/101/202/303/404, Qwen diagnostics	S_{CE} -0.153 $[-0.298, -0.017]$; -0.219 $[-0.346, -0.102]$; -0.124 $[-0.324, +0.064]$; -0.195 $[-0.454, +0.053]$; -0.280 $[-0.447, -0.123]$	FAIL, FAIL, indet., indet., FAIL
Negative controls, random persistent frame, 2 per seed	same seeds	-0.024, $+0.006$; -0.021, -0.012; -0.000, -0.003; -0.011, -0.024; $+0.009$, $+0.032$	clean in 6 of 10; no lower bound $> +0.02$
Harm-matched attribute control	seeds 42/101/202/303/404	S_{CE} -0.031 $[-0.154, +0.087]$; gate missed; gate missed; -0.203 $[-0.400, -0.010]$; -0.181 $[-0.346, -0.026]$	indet., –, –, FAIL, FAIL
Harm-matched negative controls	seeds 42/101/303/404	Neg. 1: gate missed, gate missed, -0.004, $+0.002$; Neg. 2: $+0.017$, -0.051, -0.034, $+0.023$	Neg. 1: clean at 303, 404; Neg. 2: clean at 42 only
Domain selectivity			
Domain S_{CE}, primary	7 calibrated matrices, principal matrices	Qwen 42: -0.021 $[-0.16, +0.12]$; Qwen 101: -0.117 $[-0.22, -0.01]$; Olmo 42: -0.064 $[-0.20, +0.07]$; Olmo 101: -0.020 $[-0.15, +0.11]$; Llama 42: -0.099 $[-0.30, +0.10]$; Llama 101: $+0.128$ $[-0.05, +0.30]$; Llama 202: -0.077 $[-0.28, +0.12]$	Qwen 101 excl. $\geq$+0.02; others inconclusive
Domain S_{acc}, secondary, points	same 7 matrices	$+2.3$ $[-1.6, +6.1]$; $+0.2$ $[-3.1, +3.5]$; $+1.6$ $[-2.5, +5.8]$; $+5.7$ $[+1.6, +9.9]$; $+3.3$ $[-1.2, +7.8]$; $+9.8$ $[+5.3, +14.4]$; $+6.2$ $[+1.4, +10.9]$ in the same order, treatment minus control as in Equation 1. The rule uses a different statistic, the treatment-only diagonal-minus-off-diagonal baseline-relative accuracy drop without control subtraction: positive in 7/7 matrices, $+0.3$, $+0.9$, $+1.4$, $+5.3$, $+2.1$, $+9.3$, $+6.0$, and lower end of a descriptive between-matrix two-sided 90% t-interval over 7 matrices with 6 degrees of freedom, not a crossed bootstrap, $> +1$: mean $+3.62$, lower end $+1.19$	treatment-only rule met
Domain S, four further splits	Qwen 42/101/202/303, Qwen diagnostics	S_{CE}: -0.073 $[-0.233, +0.084]$, -0.028 $[-0.142, +0.087]$, -0.181 $[-0.312, -0.049]$, -0.102 $[-0.219, +0.018]$; S_{acc}: -0.7, -1.2, -3.5, -1.2	CE: 202, 303 excl. $\geq$+0.02; 42, 101 inconclusive
Domain additive residual	Qwen, 47 domains, 1 split	-0.0003 $[-0.104, +0.111]$, 95% bootstrap over 46 domain residuals	undetermined, single split
Harm-matched MMLU matrix	Qwen 42 / 101	S_{CE} -0.048 $[-0.195, +0.094]$ / $+0.010$ $[-0.094, +0.115]$	indet., indet.
Geometry controls, own-domain excess	development panel	over common rotation -0.232 $[-0.476, +0.007]$; over independent rotations -0.243 $[-0.474, -0.020]$, 95%	positive direction not supported