GEOMETRIC ALIGNMENT: TRAINING AGAINST CONCENTRATED BEHAVIORAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Many alignment interventions assume undesirable behavior lives in a compact representation direction. Before removing it, we ask whether a coherent direction exists and whether ordinary training can avoid it without an inference-time edit. We answer with three falsifiable gates—concentration (A), geometry-held-out exclusion after geometric training (B), and an opener-CE diagnostic (C)—and the training bridge $L = L_{\text{task}} + \lambda \mathcal{G}_{\text{bad}}$, exact for linear readouts as $w^\top \Sigma_{\text{bad}} w$. On a planted residual-stream world the chain passes under geometric training with no post-hoc scrub (held-out response drop 0.989). On a Gemma-4-12B educational proxy, KL-regularized LoRA with a projected-energy penalty cuts bad-subspace energy by 32.5% on a six-prompt report slice while the specified opener-CE diagnostic stays within its Gate C band, again with no activation scrub; bridge CE stays near-flat, so Gate B is energy, not CE. The same A→B→C design on Qwen2.5-7B-Instruct (same prompts and splits) yields a 31.8% report-slice energy drop and opener-CE $+0.1\%$: a protocol replication, not evidence of universal generalization. A second family (sycophancy, Gemma) is NO-GO at layer 16 and GO (-44.3%) at layer 48: concentration licenses the question, but layer still matters. On these prompt families, concentrated displacement geometry can be suppressed in projected energy without moving that opener-CE diagnostic out of band; if concentration fails, the gates say NO-GO.

1 INTRODUCTION

Many alignment interventions assume undesirable behavior lives in a compact representation direction. Before removing it: does a coherent direction exist, and can ordinary training learn to avoid it without an inference-time edit?

That is the whole question. If the answer is yes, a geometric training penalty can cut the named projected energy while a specified opener-CE diagnostic stays in-band. If the family is diffuse, subspace exclusion is the wrong tool—and that is a result, not a failure to force a method.

> **The story (A→B→C).**
> **Gate A — concentration.** Does a paired behavioral family Δz live in a compact subspace B?
> **Gate B — exclusion by training.** Does $L_{\text{task}} + \lambda \mathcal{G}_{\text{bad}}$ cut family response (or projected energy) on prompts held out of geometry estimation, on the *learned* model, with no activation scrub?
> **Gate C — capability.** What else does the penalty damage?
> A then B then C (Figure 1).

One discovery. Existing steering and erasure methods edit or remove a direction at inference. This paper asks whether a *training* penalty can cut held-out family response (synthetic) or report-slice projected energy (LLM proxy) when Gate A finds a concentrated displacement cloud, with the Gate C opener-CE diagnostic in-band and no activation scrub. That is a proxy claim on the studied families, not set-theoretic unreachability, and not a claim that every harm is low-rank.

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

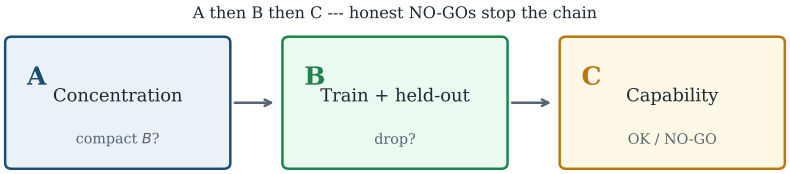

Figure 1: Falsifiable gate chain: concentration, then train-time exclusion, then capability—NO-GO at any step stops the claim.

Energy once. On nonlinear models we report projected energy $R := \mathbb{E}\|\widehat{P}^\top h\|^2$ as a second-moment *proxy* for the named geometry (equal second moments need not equalise nonlinear embedding risk). After that sentence we speak confidently in energy and response: they are the Gate B objects we report.

Why a penalty can work (preview). For a linear readout, naming a bad PSD in the loss is exact: $\mathcal{G}_{\mathrm{bad}}(w) = w^\top \Sigma_{\mathrm{bad}} w$ (Section 4). For deep nets the same idea is a Jacobian / projected-energy surrogate. The algebra that "covering A" and "excluding A" coincide in a ridge response is background (Rajput, 2026).

Contributions.

- **Falsifiable gates (the paper).** Numerical A→B→C thresholds for concentration, geometry-held-out exclusion after geometric *training*, and an opener-CE diagnostic (Section 5). Those numerical cutoffs were fixed before the reported Gemma screen; the mid-layer Gate A requirement is a later addendum, and results under it are retrospective.
- **Training bridge.** $L = L_{\mathrm{task}} + \lambda \mathcal{G}_{\mathrm{bad}}$ with $\mathcal{G}_{\mathrm{bad}} = w^\top \Sigma_{\mathrm{bad}} w$ exact for linear readouts and a controlled Jacobian proxy otherwise (Section 4).
- **Evidence.** Contaminated-label residual stream: held-out response drop 0.989 with clean accuracy non-decreasing (no scrub). Gemma 4 12B educational proxy (n=32): Gate A GO at mid/late layers; KL+projected-energy LoRA cuts report-slice energy by 32.5% (opener-CE $+0.8\%$, in-band). The same A→B→C recipe on Qwen2.5-7B-Instruct (same prompts and splits) cuts report-slice energy by 31.8% (opener-CE $+0.1\%$): protocol replication, not universal generalization.

Activation scrubbing is appendix-only (Appendix B).

2 Related work

Representation engineering and steering. Activation engineering and representation engineering (Turner et al., 2023; Zou et al., 2023a; Rimsky et al., 2024) edit or read directions tied to behaviors. Refusal mediated by a single direction (Arditi et al., 2024) and toy models of superposition (Elhage et al., 2022) are closest in mechanism—intervene on a direction, change a behavior. Our estimand differs in three ways. First, Gate A is a falsifiable concentration test on a behavioral *family*, not an assumption that a single direction exists. Second, the main method is a *training* objective evaluated without activation scrubbing at inference (LoRA (Hu et al., 2022) is only the PEFT vehicle). Third, Gates B/C separate geometry-held-out validation geometry from capability.

Concept erasure and linear guards. Concept-erasure and guarded subspace methods (Ravfogel et al., 2020; Belrose et al., 2023) remove linear information about a concept from representations. We share the subspace language but target a *behavioral displacement cloud* estimated from paired activations, and we evaluate held-out response / geometry-held-out validation energy after training—not only instantaneous linear probes.

Representation-level training interventions. Circuit breakers interrupt harmful generation by controlling the representations that produce it (Zou et al., 2024). Representation noising removes information about harmful representations so that subsequent fine-tuning recovers them less readily (Rosati et al., 2024). Those are already training-time, representation-level interventions. What is different here is the estimand: a falsifiable concentration gate, a PSD penalty evaluated as held-out projected energy (not attack success or recovery resistance), and an opener-CE diagnostic rather than a utility benchmark. We do not claim a new class of representation intervention, and we do not evaluate resistance to subsequent fine-tuning.

Machine unlearning. Machine unlearning aims to remove training influence (Cao & Yang, 2015; Bourtoule et al., 2021), with growing LLM-specific protocols (Liu et al., 2024; Yao et al., 2024; Maini et al., 2024). We do not claim weight-space deletion of capabilities. Geometric alignment asks whether a named PSD penalty cuts held-out family response / geometry-held-out validation projected energy under soft objects.

Alignment fine-tuning and its brittleness. RLHF and preference optimization (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023) reduce harmful completions in aggregate, yet subsequent fine-tuning often restores them (Qi et al., 2023; Yang et al., 2023). That brittleness is why our soft language refuses "unreachable": exclusion here is a measurable Gate B drop, not a permanence claim.

Adversarial suffixes and jailbreaks. Universal attacks (Zou et al., 2023b) optimize discrete prompts for attack success. We use a benign compliance *bridge* as a geometric probe on Gemma, not as an attack success metric; bridge CE is not the primary Gate B.

Ridge-response background. That a ridge penalty suppresses the range it covers—and that "cover A" and "exclude A" are the same LQ programme—is recorded by Rajput (2026); we cite it as background for Section 4. Second-moment scope for energy Gate B is the caution already stated in the introduction: equal second moments need not equalise nonlinear embedding risk. Out of scope: recovery cost after an energy drop, RLHF/DPO bakeoffs, multi-family safety scoreboards, and literal unreachability.

3 Setup and vocabulary

Vocabulary.

Term	Meaning
Behavior	Named family of prompts / labels (the intervention target)
Direction / B	Estimated subspace for that family's Δz
Response	Synthetic Gate B: $\mathbb{E}\lvert W \cdot \Delta z\rvert$
Energy R	LLM Gate B proxy: $\mathbb{E}\lVert \widehat{P}^\top h\rVert^2$
$\Sigma_b = \mathrm{Cov}(\Delta z)$	Centered behavioral geometry (Δz mean-centered before Cov)
$\mathcal{G}_{\mathrm{bad}}(\theta)$	Geometric training penalty (Sec. 4)
Gate A / B / C	Concentration / exclusion (calib→report) / capability

Discipline: behavior $\neq$ response $\neq$ energy $\neq$ direction. A direction is a geometric hypothesis about a behavior; response and energy are distinct Gate B scalars; energy is the second-moment proxy already scoped in the introduction. Σ_b is always the *centered* covariance of train Δz (Section 4); we do not confuse it with the uncentered Gram $\mathbb{E}[\Delta z \Delta z^\top]$.

Planted residual stream (synthetic).

$$z \;=\; U_{\mathrm{task}}\theta \;+\; \mathbf{1}_{\mathrm{unsafe}}\, U_{\mathrm{unsafe}} g \;+\; \xi,$$

with $U_{\mathrm{task}} \perp U_{\mathrm{unsafe}}$. Paired deltas live in $\mathrm{range}(U_{\mathrm{unsafe}})$ under the planted law and are diffuse under an isotropic energy-matched control. Task labels may be clean (vacuous for exclusion) or contaminated so that ordinary ERM acquires unsafe readout mass (Section 7).

Evaluation objects.

- **Synthetic response.** $\mathbb{E}|W \cdot \Delta z|$ on the held-out unsafe pairs; clean-label accuracy for Gate C.
- **LLM energy (primary).** Mean projected energy R on the six-prompt *report slice* (defined in Section 5); $\widehat{P}$ from train $\Delta z = h(\text{exec}) - h(\text{base})$ only (mean-centered Cov). Gemma and Qwen use the same prompts and split IDs.
- **LLM CE (secondary).** Bridge cross-entropy on compliance and capability openers—reported for ablations, not as the geometry-owned estimand.

All Gate B/C evaluations use the learned weights with *no* activation scrub. On the Gemma headline, λ/early-stop use a calib slice (7 prompts); reported Gate B/C energy uses a disjoint six-prompt report slice—both held out of $\widehat{P}$; paraphrase is a Gate A geometry-transfer diagnostic (Section 5). Qwen uses the same split IDs. The report slice is a partition of the original 13-prompt evaluation pool, not a newly collected test (Appendix A).

Soft claim. Exclusion means a measured drop in held-out family response (synthetic) or report-slice bad-subspace energy (LLM) after geometric *training*—interesting with the opener-CE diagnostic in-band and no scrub, not that a behavior is impossible.

4 Why penalizing a bad direction reduces response

If Gate A names a subspace B, why should putting B in the loss cut held-out response? Because for linear readouts the geometric penalty *is* the quadratic sensitivity on that PSD; gradient descent on $L_{\text{task}} + \lambda \mathcal{G}_{\text{bad}}$ is then ordinary regularized risk with the bad directions named explicitly.

Training bridge. The displayed quadratic identity uses *centered* displacements. Let $\mu = \mathbb{E}[\Delta]$ and $\widetilde{\Delta} = \Delta - \mu$. The operational penalty is

$$\mathcal{G}_{\text{bad}}(\theta) \;=\; \mathbb{E}_{x,\Delta\sim\nu_{\text{bad}}} \big\| f_\theta(x + \widetilde{\Delta}) - f_\theta(x) \big\|^2, \tag{1}$$

not the raw (uncentered) map $x \mapsto x + \Delta$. Geometry is $\Sigma_{\text{bad}} = \text{Cov}(\Delta) := \mathbb{E}[(\Delta - \mu)(\Delta - \mu)^\top]$ (matching the estimator used for $\widehat{P}$ and Gate A spectra). A constant family displacement can have large $\|\mu\|$ and $\Sigma_{\text{bad}} = 0$: failure of concentration in centered variation does not imply that no coherent mean direction exists. We keep covariance as the Gate A object and state that it measures centered family variation, not the mean offset. With Jacobian J_θ,

$$\mathbb{E}\big\| J_\theta(x)\, \widetilde{\Delta} \big\|^2 \;=\; \text{Tr}\big(J_\theta(x)\, \Sigma_{\text{bad}}\, J_\theta(x)^\top \big), \tag{2}$$

so the first-order bridge is $\mathcal{G}_{\text{bad}}(\theta) = \mathbb{E}_x[\text{Tr}(J_\theta(x)\, \Sigma_{\text{bad}}\, J_\theta(x)^\top)] + \varepsilon(\theta)$ on the centered cloud (the uncentered Gram $M = \mathbb{E}[\Delta\Delta^\top] = \Sigma_{\text{bad}} + \mu\mu^\top$ is *not* what we feed the penalty).

Proposition 1 (Linear readout: penalty = response geometry)**.** *Let $f_w(z) = w^\top z$ and let $\Sigma_{\text{bad}} = \text{Cov}(\Delta)$ be the centered second moment above. Then on centered displacements $\mathbb{E}|w^\top \widetilde{\Delta}|^2 = w^\top \Sigma_{\text{bad}} w$ exactly, so $\min_w L_{\text{task}}(w) + \lambda\, w^\top \Sigma_{\text{bad}} w$ shrinks sensitivity on $\text{range}(\Sigma_{\text{bad}})$. If the estimated penalty misses an axis of Σ_{bad}, then* in the linear–quadratic response of Remark 1 *some probe in that missed range retains a positive limiting drift floor.*

Meaning. Gate B is not a new response class. It is ordinary training with the undesirable PSD named in the loss, evaluated on the learned weights—not on a scrubbed forward pass.

Proposition 2 (Jacobian remainder)**.** *If f_θ is C^2 and $\|R_\theta(x, \widetilde{\Delta})\| \le \frac{1}{2} L \|\widetilde{\Delta}\|^2$ for the Taylor remainder of* (1)*, the gap between $\mathcal{G}_{\text{bad}}$ and $\mathbb{E}\|J_\theta \widetilde{\Delta}\|^2$ is $O(t^3)$ under $\widetilde{\Delta} = t\Delta_0$ as $t \to 0$.*

Outside the small-t regime we treat $\mathbb{E}\|P_{\text{bad}}^\top h_\ell\|^2$ as a PEFT surrogate that shrinks mass in $\text{range}(\widehat{P})$, and we evaluate Gate B by geometry-held-out validation energy / held-out family response of the learned model.

Background rename. In the ridge response $w_\lambda = (I + 2\lambda\Sigma')^{-1}v$ with payoff $D_A(w) = w^\top Aw$, requiring $\mathrm{range}(\Sigma') \supseteq \mathrm{range}(A)$ so that $D_A(w_\lambda) \to 0$ is the same programme whether one calls it "cover A" or "exclude A" (Rajput, 2026). Figure 5 checks that rename numerically.

Remark 1 (Missed axis leaves a residual floor). If the penalty range misses $\mathrm{range}(\Sigma_{\mathrm{bad}})$, then $\lim_{\lambda\to\infty} D_{\Sigma_{\mathrm{bad}}}(w_\lambda) > 0$ for some unit probe in that bad range (Rajput, 2026).

Remark 2 (Covering A is excluding A). For any PSD A, the LQ programmes "cover A" and "exclude A" are identical (Rajput, 2026). Whether A is undesirable is a modelling choice (Section 8).

Numeric checks of Remarks 1–2 are in Figure 5 (Appendix A).

5 THE A–B–C PROTOCOL

The rest of the paper is organized by these gates. Concentration is Gate A—not an assumption that unsafe geometry is low-rank. The *numerical* GO/NO-GO cutoffs below were fixed before the reported Gemma screen; NO-GOs are kept (Appendix A). The mid-layer Gate A requirement is a 2026-08-12 addendum after late-layer null-family GO results; results under that amended layer rule are retrospective (Section 6).

Gate A — concentration. On paired deltas Δz, mean-center then report top-4 energy of $\Sigma_b = \mathrm{Cov}(\Delta z)$.

- GO if top-4 energy ≥ 0.50;
- NO-GO if top-4 energy < 0.25;
- otherwise marginal.

A planted low-rank family must GO; a matched-energy diffuse control must NO-GO. On high-d LLM probes we also require beating a matched-n isotropic control by a clear margin (protocol: ≥ 0.15 absolute on top-4 energy). Evidence: Section 6.

Gate B — exclusion after training (report-slice). Estimate $\widehat{P}$ on a *train* slice only (centered Cov). Train with $L_{\mathrm{task}} + \lambda\mathcal{G}_{\mathrm{bad}}$ (linear: $\lambda\|\widehat{P}^\top w\|^2$; Gemma repair: KL-to-base + normalized projected energy on the compliance forward). Evaluate the *learned* model with no activation scrub. Primary metrics (geometry-owned):

- Synthetic: relative drop in held-out $\mathbb{E}|W \cdot \Delta z|$;
- LLM: relative drop in report-slice bad-subspace energy $\mathbb{E}\|\widehat{P}^\top h\|^2$ (selection on a disjoint calib slice).

GO if relative improvement $\geq 30\%$ vs the no-geo baseline. Compliance-bridge CE is *not* the primary LLM estimand. Evidence: Section 7.

Gate C — capability. Synthetic: clean-label task accuracy (absolute drop ≤ 5 pp OK). LLM: relative change of a benign opener CE within 15% at the selected checkpoint (Gate C reported on the same report slice as Gate B). No single "safe" scalar; report the Pareto and early-stop when B clears with C in-band. Evidence: same section, same checkpoints.

Splits and checkpoint selection. Hard-core on both LLM models: $\widehat{P}$ from *train* (19); λ and early-stop by the first checkpoint that clears Gate B with Gate C in-band on *calib* (7); report Gate B/C energy on a disjoint six-prompt *report slice*—held out of both $\widehat{P}$ and selection, but partitioned from the original 13-prompt evaluation pool after that pool had already been used for a validation-selected Gemma pilot (2026-08-10; Appendix A). This is retrospective held-out evaluation, not confirmation on a newly collected test. Paraphrase prompts never estimate $\widehat{P}$; they are a Gate A geometry-transfer diagnostic (energy mass in train $\widehat{P}$), not the headline Gate B estimand. Do not report the last iterate if it violates C; early-stop is part of the method.

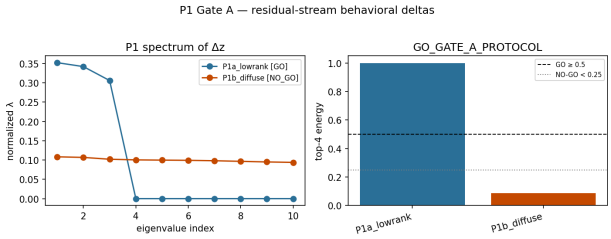

Figure 2: **Synthetic Gate A.** Planted low-rank unsafe Δz concentrates; matched-energy isotropic control does not. Regime: residual-stream plant; protocol check, not an LLM claim.

6 GATE A: DOES A COHERENT DIRECTION EXIST?

Gate A asks whether a paired behavioral family concentrates before anyone trains a penalty. If it does not, stop.

Synthetic control. Spectrum tools separate planted k=3 unsafe geometry (top-4 energy 1.00, overlap 1.00) from isotropic bursts (top-4 0.09). Verdict: **GO_GATE_A_PROTOCOL** (Figure 2).

Gemma proxy (compliance). Forward-only probes on `google/gemma-4-12B-it` (educational prompts; compliance bridge "Sure, I can explain that." as a geometric attractor, not a jailbreak score). Gate A split: train 19 / held-out 13 / paraphrase 13; $\widehat{P}$ from *centered* train Δz Cov only (paraphrase never estimates geometry). For Gate B repair the 13-prompt held-out pool is further split into calib 7 / report slice 6 (Section 7). For n=32 prompts, $\Delta z = h(\text{exec}) - h(\text{base})$ (mean over tokens); top-4 on *train*:

Layer	top-4 (train)	top-4 (isotropic)	Call
16	0.466	0.249	Marginal
32	0.527	0.284	GO
48	0.606	0.276	GO

Verdict: **GO_GATE_A_PROXY** (Figure 4, Appendix A).

Qwen proxy (same family, second model). Same educational prompts, split IDs, centered-Cov Gate A, and isotropic control on `Qwen/Qwen2.5-7B-Instruct`. Decoder layers 8/16/24 are all GO vs matched-energy isotropic (train top-4 0.989/0.977/0.755; isotropic $\approx$ 0.30); we freeze the deepest mid-layer GO (layer 16). The absolute top-4 value depends on the representation pooling definition and on sequence length: Gemma multimodal forwards are long; Qwen text-only forwards are short, so Qwen's 0.977 at layer 16 is not a like-for-like scalar against Gemma's 0.527 at layer 32. The GO/NO-GO decision is unchanged because the matched-energy isotropic control is evaluated under the same pooling (Appendix C).

Second educational family (sycophancy). Same model and Gate A recipe on a sycophancy-bridge pair set ($\Delta z = h(\text{agree}) - h(\text{base})$; train 19). Top-4 vs isotropic: layer 16 0.531/0.248 (GO); layer 32 0.495/0.258 (marginal); layer 48 0.610/0.271 (GO). Mid-layer Gate A: **GO** (layer 16). Energy Gate B on the same family is reported separately (Section 7.2)—concentration does not automatically imply repair.

Null families (honest LLM negatives). Two non-behavioral controls on the same prompt pool: (i) filler prefixes `Also,`/`Meanwhile,` (topic salad); (ii) per-prompt nonsense prefixes (no shared attractor). Both are *marginal* at layers 16 and 32 and only clear GO at layer 48. We treat late-only GO as a prefix-vs-base artifact on this model: load-bearing LLM Gate A therefore requires a **mid-layer** ($\leq$ 32) GO (addendum: paragraph below). Under that rule both null families are **NO_GO**; the matched-energy isotropic control in every Gate A table remains NO-GO by construction (top4 $\approx$ 0.25). Synthetic planted vs isotropic remains the clean protocol GO/NO-GO pair.

On this n, Gate A is an empirical concentration/transfer *proxy*, not a certified claim that training covers deployment shift (Rajput, 2026). The LLM pilot supports only the measured projected-energy Gate B/C result below.

Rank used in Gate B. Gate A screens with top-4 energy of the *centered* train Cov (Table above: 0.527 at layer 32). The Gemma headline penalty uses rank $k{=}2$ of that same covariance. From the frozen Δz array (`results/p4_gemma_deltas.npz`; script `scripts/report_l32_spectrum.py`), the centered eigenvalues begin 0.0449, 0.0222, 0.0168, 0.0134, so top-2 energy is 0.363 and top-4 is 0.527. Centered energy of the 13-prompt held-out Δz in that train $\widehat{P}$ is 0.335. A different statistic—uncentered Frobenius energy $\|\widehat{P}^\top \Delta z\|_F^2 / \|\Delta z\|_F^2$—is 0.323 on train and 0.286 on paraphrase; τ_b uses still another object (uncentered Gram). Those are not interchangeable with Gate A top-k. $k{=}2$ is the intervention rank; it is a stricter subspace than the top-4 GO screen, not a silent drop of the same fraction.

Mid-layer addendum. Null families above cleared GO only at layer 48. The initial concentration thresholds (top-4 $0.50/0.25$, isotropic margin) were fixed before the reported screen. The requirement that LLM Gate A also pass at a mid-layer (≤ 32) is a protocol addendum dated 2026-08-12, after null families cleared GO only at layer 48—not a threshold frozen before the first Gemma Gate A run. Results under that amended layer rule are retrospective. The same null families motivate the amendment; they do not by themselves establish its prospective effectiveness. Under that addendum both null families are **NO_GO**. Matched-spectrum random-P energy ratios remain Gate A diagnostics, not the Gate B energy estimand. Concentration is permission to ask energy Gate B on the layers that cleared.

7 Gates B and C: can training suppress that geometry?

Given a Gate A subspace, train $L_{\text{task}} + \lambda \mathcal{G}_{\text{bad}}$ and evaluate the *learned* checkpoint—no activation scrub. Gate B owns the report-slice geometry; Gate C owns the opener-CE diagnostic at the same early-stop (selected on calib).

7.1 Synthetic Level 3 (clean control)

Ordinary task training on $U_{\text{task}} \perp U_{\text{unsafe}}$ never acquires unsafe mass—a vacuous setting for exclusion. We therefore contaminate labels,

$$y = \text{sign}(\theta_0 + \alpha\, 1_{\text{unsafe}} g_0), \qquad \alpha = 0.75,$$

so task-only ridge acquires w-mass in $\text{range}(U_{\text{unsafe}})$. Single-stage training

$$L_{\text{task}} + \lambda \|\widehat{P}^\top w\|^2$$

with $\widehat{P}$ from train Δz yields (Figure 3):

Arm	Clean task acc	Held-out $\mathbb{E}\lvert W \cdot \Delta z\rvert$	Rel. drop	w unsafe mass
Task only	0.951	0.379	—	0.054
Geo train ($\lambda{=}10^5$)	0.953	0.004	**0.989**	0.000
Example ridge	0.935	0.068	0.820	0.010

Gate B **GO**, Gate C **OK** (clean accuracy non-decreasing). A shallow ReLU net with finite-difference $\mathcal{G}_{\text{bad}}$ moves the same way (relative response drop 0.39). Verdict: **GO_GEOMETRIC_TRAINING**.

Relative held-out drop rises smoothly with λ while clean-task drop stays near zero (Appendix A).

Background LQ check. Ten analytic/numeric checks confirm Remarks 1 and 2 (Figure 5, Appendix A): full exclude floors near zero; partial exclude leaves a residual. Verdict: **GO_LQ_DUALITY**—supporting the short theory section, not replacing Gates B/C.

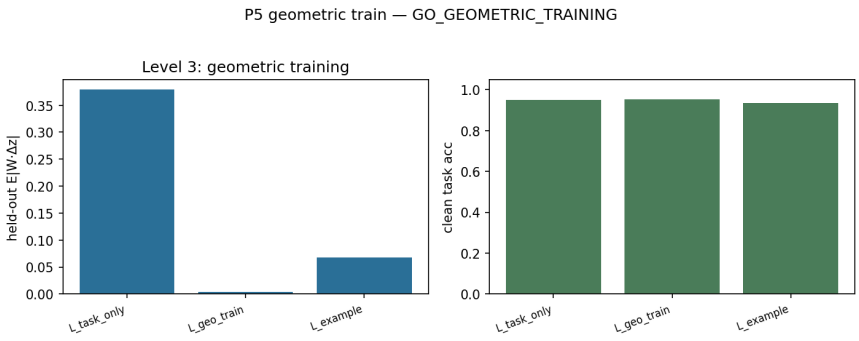

Figure 3: **Level 3 synthetic: geometric training, not scrubbing.** Left: held-out unsafe response collapses under $L_{\text{task}}+\lambda\|\widehat{P}^{\top}w\|^2$. Right: clean-label task accuracy does not fall. Regime: contaminated-label residual stream; eval without scrub.

7.2 Gemma energy Gate B (report slice)

> **How to read this result.** Report-slice energy drop under KL+projected-energy LoRA (μ=0; Eq. (3)), opener-CE in-band, no scrub. The six-prompt slice is a retrospective partition of a previously inspected 13-prompt pool, not a newly collected test; -32.5% is a mean threshold crossing, not a CI or a safety score.

Estimate rank-k=2 $\widehat{P}$ from *centered* train compliance Δz Cov at layer 32. Train LoRA (Hu et al., 2022) on late layers $\{28, 30, 32, 34, 36\}$ with

$$L \;=\; \mathrm{KL}\big(\text{base}\,\|\,\text{model}\big)_{\text{cap}} \;+\; \lambda\cdot\frac{\mathbb{E}\|\widehat{P}^{\top}h_{32}^{\text{comp}}\|^2}{\mathbb{E}_0\|\widehat{P}^{\top}h_{32}^{\text{comp}}\|^2}, \tag{3}$$

i.e. KL on a benign capability bridge plus a *normalized projected-energy* penalty on the compliance-bridge forward—where the bad mass lives. No $\mu(-\mathrm{CE})$ term. Primary Gate B is report-slice bad-subspace energy drop ($\geq 30\%$); opener-CE relative change within 15% (Gate C). Honest split role: train 19 estimates $\widehat{P}$; calib 7 selects λ/early-stop; the six-prompt report slice supplies the reported Gate B/C scalars; paraphrase is not used for energy Gate B (Gate A transfer only).

$\lambda \in \{20, 100, 500\}$ (grid frozen before this run); first GO on the *calib* slice wins; report on the report slice. At λ=20, early-stop@80: calib energy drop 0.315 (-31.5%); **report-slice** energy drop 0.325 (-32.5%); capability CE $+0.8\%$ (in-band); compliance CE -0.8% (secondary, near flat). Verdict: **GO_GEOMETRIC_ENERGY_PROXY** (retrospective report-slice protocol, 2026-08-12).

Qwen energy Gate B (same protocol). The same loss, λ grid, split IDs, and GO/NO-GO cutoffs on `Qwen/Qwen2.5-7B-Instruct` (layer 16; LoRA $\{12, 14, 16, 18, 20\}$).

Model	Layer	Calib drop	Report drop	Cap. rel	Call
Gemma-4-12B-it	32	0.315	0.325	$+0.8\%$	GO
Qwen2.5-7B-Instruct	16	0.334	0.318	$+0.1\%$	GO

The protocol reproduces on Gemma and Qwen under the same A→B→C design, with similar Gate B energy reductions ($\sim 32\%$) and capability diagnostics remaining within the specified band. We present this as a replication of the protocol, not evidence of universal generalization (Section 8).

Secondary external endpoint (compliance). On the same report slice we also score a bridge preference margin $\mathrm{CE}(\text{control}) - \mathrm{CE}(\text{behavior})$ (positive ⇒ higher likelihood on the undesirable prefix). Base margin -1.83; adapted -1.67 (Δ+0.16). Near-flat / slightly adverse: geometry-owned Gate B is energy, not this CE margin—reported as a delimiter, not a second primary estimand.

Sycophancy energy Gate B. Despite mid-layer Gate A GO, the same KL+projected-energy recipe at layer 16 (LoRA on $\{12, 14, 16, 18, 20\}$) yields calib/test energy drops $\lesssim 0.04$ across

$\lambda \in \{20, 100\}$—**NO_GO** under the 30% band (capability in-band). At layer 48 (LoRA on $\{40, 42, 44, 46, 47\}$), the same recipe is **GO**: report-slice energy drop 0.443 (-44.3%) at λ=20, early-stop@25, capability CE $+3.9\%$ (in-band). So: concentration licenses asking Gate B; layer choice still matters under this PEFT budget.

CE-metric ablation (delimiter). A matched 2×2 (Appendix A) shows that *compliance-bridge CE* is μ-driven, not geometry-driven ($+33.7\%$ with $\mu, \approx 0$ without). Do not use bridge CE as the geometry-owned estimand.

Activation scrub (secondary). Appendix B reports last-16-token subspace scrubbing as a representation intervention check (GO_B / C marginal).

X, not Y. Gates B/C show that geometric *training* can cut held-out response (synthetic) or report-slice energy (Gemma and Qwen) with the opener-CE diagnostic in-band and no scrubbing—and without μ-driven CE gaming. They do not show that undesirable behavior was reduced, that general capability was preserved, or that every PEFT budget keeps opener-CE in band.

8 LIMITS

What we do not claim. We do not claim that every harmful LLM behavior is low-rank, that exclusion is literal unreachability, that Gate B transfers to jailbreak suites, or that geometric training replaces preference optimization. A second open-weight instruction model (Qwen) reproduces the same A→B→C protocol on the same educational family; that is a replication, not universal generalization. Energy remains a second-moment proxy; Gate A on this n is an empirical concentration/transfer proxy, not a certified Matching claim (Rajput, 2026).

Direction-agnostic method (dual use). Gate A asks whether a paired family concentrates, not whether it is undesirable. Aimed at an over-compliance bridge, the intervention reduces named energy; aimed at a refusal direction, the same lever could weaken safety behavior (Arditi et al., 2024). The contribution is a falsifiable geometric *training* protocol, not a one-sided safety guarantee.

Where the story can fail. Diffuse families leave no useful $\widehat{P}$. Long PEFT runs can move opener-CE out of band—early-stop is part of the method. The LLM proxy is educational and small-n (n=32; report-slice Gate B on 6 as a mean threshold crossing). Qwen uses the same prompt collection: the remaining limitation is sample scale, not the absence of a second model. Energy drop is not a claim that behavior fell or that general capability was preserved. Sycophancy is NO-GO at layer 16 and GO at layer 48 (-44.3%). Null prefix families fail the mid-layer Gate A addendum.

Honest negatives (kept). Diffuse synthetic controls NO-GO Gate A. Activation scrubbing on Gemma is only marginally Gate B. CE-only LoRA without a capability anchor collapses the benign opener metric. Compliance-bridge CE is μ-driven and is the wrong Gate B (Section 7).

9 CONCLUSION

On the studied families, if the displacement cloud concentrates, training can cut that projected energy while the specified opener-CE diagnostic stays in band. That claim is conditional and falsifiable: Gate A must clear before Gate B is even asked, and Gate C can veto a B win that moves opener-CE out of band.

Synthetically, when a bad subspace is planted and labels give the task a reason to acquire it, $L_{\text{task}} + \lambda\|\widehat{P}^\top w\|^2$ drops held-out unsafe response by 0.989 while clean accuracy holds—no scrubbing. On a narrow Gemma educational proxy, the same idea cuts report-slice energy by 32.5% (opener-CE $+0.8\%$, no scrub). Qwen2.5-7B-Instruct, same A→B→C design, yields -31.8% and opener-CE $+0.1\%$: a protocol replication, not universal generalization. Sycophancy splits by layer (NO-GO at 16, GO -44.3% at 48). When concentration fails, the gates record NO-GO rather than forcing a method.

AI USE STATEMENT

Generative AI coding assistants were used to help develop the conceptual framing and gate protocol, to propose and refine hypotheses, to give feedback on experimental design, to implement experiment and plotting code, to interpret results, and to draft and edit manuscript text (including structure, citations, and formatting). We have not used generative AI tools to generate synthetic datasets, to provide critical unreviewed ingredients of mathematical proofs, to write proofs, or for translation. All claims, numbers, figures, and code paths were checked against frozen experiment logs and scripts by the author. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The LLM experiments use a benign educational compliance/sycophancy *bridge* as a geometric probe, not a harmful-content or jailbreak benchmark, and we do not release attack recipes. Projected energy is a second-moment proxy: a drop in energy is not a safety certificate, and dual-use exclusion methods can also be inverted. No human-subjects data are collected.

REPRODUCIBILITY STATEMENT

Synthetic residual-stream protocols, numerical gate thresholds, split construction, and the Gemma/Qwen LoRA recipes are specified in Sections 3–5 and Appendices A–C. The anonymous supplementary archive accompanying this submission includes: prompt lists and split IDs (`experiments/p4_gemma_proxy/pairs_*.json`); frozen Δz arrays and Gate A JSON (`results/p4_gemma_deltas.npz`, `p4_gemma_gate_a.json`, `p4_qwen_deltas.npz`, `p4_qwen_gate_a.json`); the selected LoRA checkpoints and repair logs (`results/p5_repair_lam20_ckpt/`, `p5_gemma_repair.json`, `p5_repair_qwen_lam20_ckpt/`, `p5_qwen_repair.json`); scripts that regenerate the L32 spectrum table and the energy figure (`scripts/report_l32_spectrum.py`, `plot_gemma_repair.py`); and the training/eval entry points (`experiments/p4_gemma_proxy/run_gate_a.py`, `run_geo_repair.py`; `experiments/p4_text_proxy/run_gate_a.py`, `run_geo_repair.py`). Gemma: `google/gemma-4-12B-it`; LoRA $r{=}8$ on layers $\{28, 30, 32, 34, 36\}$; $\widehat{P}$ rank-2 of centered train Cov at layer 32. Qwen: `Qwen/Qwen2.5-7B-Instruct`; LoRA $r{=}8$ on layers $\{12, 14, 16, 18, 20\}$; $\widehat{P}$ rank-2 at decoder layer 16.

REFERENCES

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. *arXiv preprint arXiv:2406.11717*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Pacelli, John Hewitt, and Jacob Steinhardt. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, 2015.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Sander Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Vishal Rajput. The matching principle: When does a training penalty cover deployment shift? *arXiv preprint arXiv:2605.22800*, 2026.

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 Association for Computational Linguistics*, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the Association for Computational Linguistics*, 2024.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *Advances in Neural Information Processing Systems*, 2024.

Alexander Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xiang Ren, and Bill Yuchen Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems*, 2024.

A PROTOCOL DETAILS

This appendix records numerical thresholds, recipes, and split chronology. The numerical Gate A/B/C cutoffs were fixed before the reported Gemma screen. The mid-layer Gate A requirement and the calib/report partition of the 13-prompt pool are later protocol amendments (dated below).

Synthetic residual stream.

$$z = U_{\text{task}}\theta + \mathbf{1}_{\text{unsafe}} U_{\text{unsafe}} g + \xi, \qquad U_{\text{task}} \perp U_{\text{unsafe}}.$$

Gate A uses paired Δz under the planted law vs isotropic energy-matched controls. Level 3 contaminates labels with $\alpha = 0.75$ so task-only training acquires unsafe w-mass; without contamination the exclusion setting is vacuous. $\widehat{P}$ is estimated from the *centered* Cov of the train slice of Δz only. Evaluation: held-out $\mathbb{E}|W \cdot \Delta z|$ and clean-label accuracy; no activation scrub.

Gemma repair recipe (headline).

- Model: `google/gemma-4-12B-it`; LoRA r=8 on layers $\{28, 30, 32, 34, 36\}$.
- Splits: train 19 ($\widehat{P}$) / calib 7 (λ, early-stop) / report slice 6 (reported Gate B/C) / paraphrase 13 (Gate A transfer diagnostic only; never estimates $\widehat{P}$; not energy Gate B). **Chronology.** On 2026-08-10 a validation-selected pilot reported energy on the full 13-prompt geometry-held-out pool (-30.3% at step 75). On 2026-08-12 that pool was partitioned into calib 7 / report 6; λ and early-stop were then selected on calib and the -32.5% headline was read on the six-prompt slice. Looking at prompt wording is not the same as using the pool's outcomes and then partitioning it. We therefore describe the six-prompt number as a *retrospective held-out evaluation*: those prompts were excluded from $\widehat{P}$ and from the subsequently defined calib split, but they are not an independently collected test.
- $\widehat{P}$: top-k=2 eigenvectors of *centered* train compliance Δz Cov at layer 32. From the frozen Δz array: centered top-2 energy 0.363, top-4 0.527 (Gate A table); held-out centered energy in train $\widehat{P}$ 0.335. Uncentered Frobenius energy in $\widehat{P}$ (0.323 train / 0.286 paraphrase) is a different statistic.
- Loss: $\text{KL}_{\text{cap}} + \lambda \cdot$ (normalized projected energy on compliance-bridge forward); μ=0.
- Grid: $\lambda \in \{20, 100, 500\}$; first GO on the *calib* slice (energy drop $\geq 30\%$, $|\text{cap rel}| \leq 15\%$) wins; max steps 80; report energy drop on the report slice.
- Selected: λ=20, early-stop@80, calib drop 0.315, **report-slice** drop 0.325 (-32.5%); capability CE $+0.8\%$.

Gemma CE ablation (delimiter only). Same LoRA skeleton; geo-*fraction* on the capability forward; optional $\mu(-\text{CE}_{\text{comp}})$. Arms: $(\lambda, \mu) \in \{(0, 0), (0, 0.25), (2, 0), (2, 0.25)\}$. Primary reported scalar for that table is compliance CE rise—intentionally *not* the geometry-owned estimand.

Why the penalty must sit on the compliance forward. $\widehat{P}$ is estimated from compliance Δz. A geometric term evaluated only on a benign capability bridge need not see the bad mass. The repair protocol therefore penalises projected energy on the compliance bridge while anchoring capability with KL on a separate benign bridge.

Threshold summary.

Gate	GO	Fail / NO-GO
A (top-4 energy)	≥ 0.50 (+ isotropic margin on LLM)	< 0.25
B (rel. drop / rise)	≥ 0.30	< 0.10
C (capability)	$\lvert\text{rel}\rvert \leq 0.15$	$\lvert\text{rel}\rvert > 0.40$

Synthetic λ sweep. Relative held-out unsafe-response drop vs. clean-task accuracy change (contaminated residual stream; no scrub):

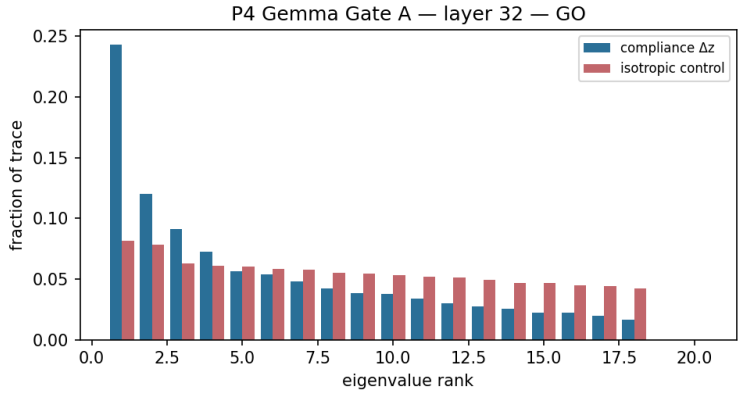

Figure 4: **Gate A on Gemma (layer 32).** Compliance Δz concentrates versus a matched-energy isotropic control (educational compliance-bridge pairs).

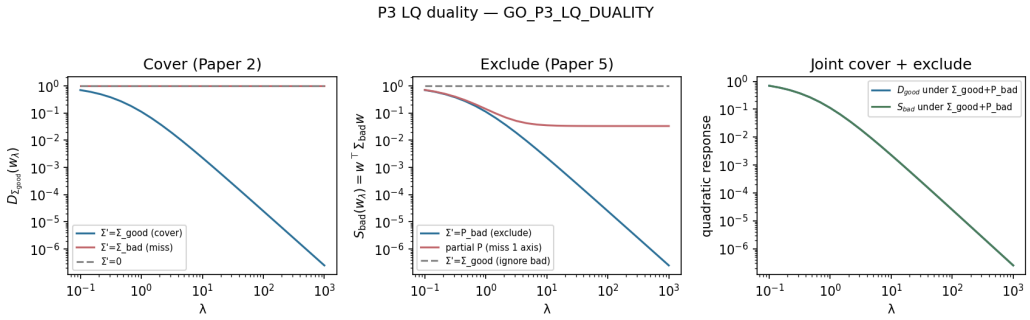

Figure 5: **Penalize the named range, or leave a floor.** Left/centre: full cover or exclude drives sensitivity to zero; wrong-axis or partial projectors leave a residual. Right: joint good+bad penalty zeros both curves. Regime: synthetic LQ response; background check, not the primary claim.

λ	Rel. drop	Clean-task drop (pp)
50	0.042	-0.1
200	0.150	-0.1
1000	0.469	-0.1
5000	0.815	-0.4
$2{\cdot}10^4$	0.946	-0.2
10^5	**0.989**	-0.2

Gemma CE ablation table (delimiter).

Arm	λ	μ	Comp. CE rise	Cap. rel
KL only	0	0	$+0.6\%$	-0.2%
KL$+\mu$	0	0.25	$+\mathbf{33.7}\%$	$+7.0\%$
KL$+\lambda$ frac	2	0	$+0.8\%$	-0.2%
KL$+\lambda$ frac $+\,\mu$	2	0.25	$+\mathbf{32.4}\%$	$+7.0\%$

B ACTIVATION SCRUB (SECONDARY)

This appendix records subspace scrubbing as a *representation* check, not as the paper's primary method. Primary claims use geometric *training* (Section 7).

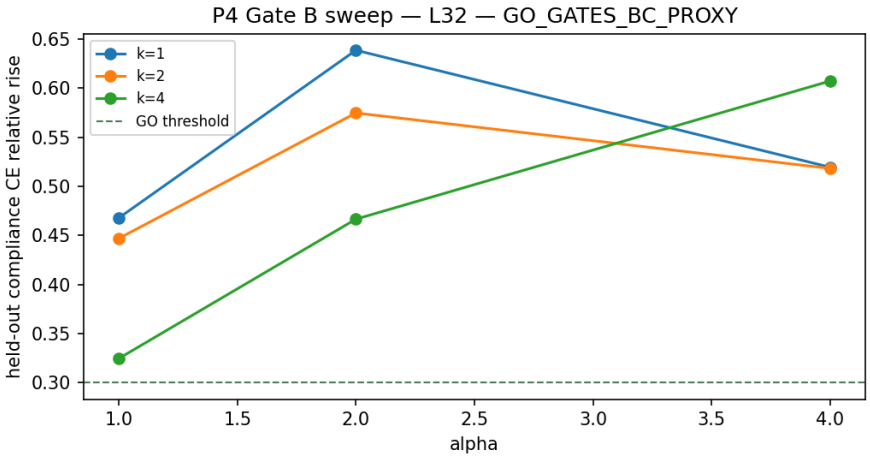

Figure 6: **Activation scrub is secondary.** Held-out compliance CE rise under geometric scrub vs controls. Best specific config clears a CE Gate B but moves capability outside the stricter training band (last-16-token scrub; educational bridge).

Protocol. Estimate $\widehat{P}$ from train-slice compliance Δz at a fixed layer. At inference, apply a last-16-token subspace scrub $(I - \alpha \widehat{P}\widehat{P}^\top)h$ via a forward hook before continuing. Report relative change in compliance-bridge CE and benign opener CE versus an unscoured baseline on held-out prompts. Compare against random-subspace and example-scrub controls for specificity.

Frozen Gemma numbers (layer 32, k=2, α=1). Compliance-bridge CE rises $+44.7\%$ (Gate B GO on the CE scalar). Capability opener CE rises $+18.8\%$ (Gate C marginal / soft fail under the stricter $\leq 15\%$ band used for Level 3 training). Random scrub is nonspecific; example scrub is near flat on compliance CE. Verdict: GO_B / C_marginal—useful as a proximal exclusion check, not as standalone alignment by scrubbing (Figure 6).

Why scrub is secondary. Scrubbing tests whether the estimated subspace carries mass at inference. It does not show that ordinary training learned to avoid that mass. Level 3 geometric training is the dual claim we actually want: the checkpoint itself, with no activation edit, already moves the held-out geometry-owned metric while keeping capability in-band.

C QWEN PROTOCOL REPLICATION

Main-text Qwen numbers (Section 7.2) use the *same* A→B→C design as Gemma: same educational `compliance_bridge` prompts and split IDs, centered train Cov for $\widehat{P}$, calib→report-slice selection, matched-energy isotropic Gate A control, and geometry-owned energy Gate B (μ=0). This is a protocol replication, not a third-family expansion and not a claim of universal generalization.

Model and freeze. `Qwen/Qwen2.5-7B-Instruct`; text-only chat template. Gate A candidates $\{8, 16, 24\}$ (0-based decoder blocks). Load-bearing freeze: deepest mid-layer GO (layer 16), analogously to freezing Gemma at layer 32 rather than late layer 48. k=2; LoRA r=8 on $\{12, 14, 16, 18, 20\}$.

Pooling. Gate A Δz is the mean-token difference $h(\text{exec}) - h(\text{base})$, as on Gemma. Absolute top-4 energy is *not* comparable across models as a scalar: Qwen text-only sequences are short, so train top-4 at layer 16 is 0.977 against an isotropic control of 0.298. The GO/NO-GO call uses that matched-energy isotropic control under the *same* pooling; both Gemma and Qwen GO. Gate B energy remains last-N=16 projected energy on the compliance forward (same as Gemma).

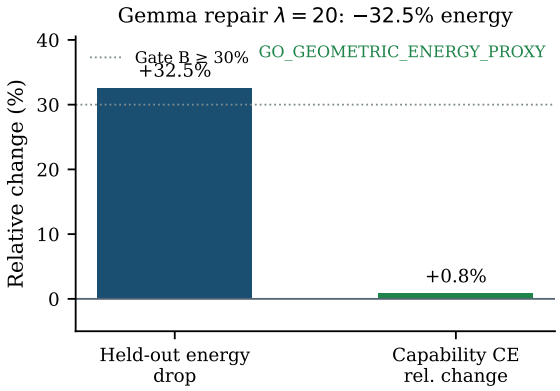

Figure 7: Gemma repair at λ=20: report-slice energy drop -32.5% with capability CE $+0.8\%$ (Gate C in-band). Qwen, same protocol: -31.8% and $+0.1\%$ (Section 7.2).

Gate A. Train top-4 energy vs matched-energy isotropic control: layer 8 0.989/0.299 (GO); layer 16 0.977/0.298 (GO); layer 24 0.755/0.297 (GO). Verdict: **GO_GATE_A_PROXY**.

Energy Gate B / C. Same repair recipe as the Gemma headline (KL-to-base on a benign capability bridge + $\lambda\cdot$ normalized projected energy on the compliance forward; no $\mu(-\mathrm{CE})$). $\lambda \in \{20, 100, 500\}$; first calib GO wins. At λ=20, early-stop@70: calib drop 0.334; report-slice drop 0.318 (-31.8%); capability CE $+0.1\%$ (in-band); compliance CE -3.3% (secondary). Verdict: **GO_GEOMETRIC_ENERGY_PROXY**.

Claim hygiene. Two instruction models, one educational prompt family, small n. The interesting claim is that the protocol is not tied to one architecture on this family—not that an energy penalty lowered energy, and not that the geometry is universal.