DIRECT SPECTRAL ACTUATION: PREDICTION-PRESERVING EXPERT DIFFERENTIATION IN MIXTURE-OF-EXPERTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Trained mixture-of-experts (MoE) models may lack task-relevant expert differences, leaving routers with little useful gradient signal. Direct Spectral Actuation (DSA) adds protected expert-output residuals after training without changing the anchor prediction. Under a fixed function-space budget, DSA selects the write that maximizes added router-gradient power under the original task loss. This yields a leading singular mode at a symmetric anchor and a metric-aware trust-region solution at a general trained anchor. DSA freezes the residual adapter and updates only the reader; native experts and the pretrained router remain fixed. Controlled tasks and CIFAR-10-C test protection and the subsequent read, while an OpenML-CC18 replay tests selective intervention. On one frozen native top-8 OLMoE checkpoint, DSA reduces held-out NLL relative to the anchor by 12.4–25.8% on SST-2 and 40.7% on average on RTE. It improves on matched reader-only adaptation in all six evaluations and yields lower task-mean NLL than a matched final-layer RoMA-objective baseline on both tasks.

1 INTRODUCTION

Sparse MoEs increase capacity without proportional computation by routing each input to a small subset of experts. This capacity helps only when expert outputs contain task-relevant differences that the router can exploit (Jacobs et al., 1991; Jordan & Jacobs, 1994; Chi et al., 2022). In a trained MoE, experts may instead collapse to similar functions or differ along directions unrelated to the task. In this regime, router-only adaptation can have little useful gradient signal even when the frozen predictor is already accurate.

Existing approaches encourage diversity or load balance during training, or resume joint expert–router optimization. These interventions can change an already useful predictor before the induced specialization has been validated. We instead change expert outputs while fixing their routed mixture. At the frozen router, every protected write has the same prediction and task risk, but can change the risk derivative with respect to rerouting. Because protected writes start from the same loss, their gradient power measures the initial descent available after rerouting. DSA maximizes it under a fixed write budget. The read then deliberately changes predictions using the original task loss: accepted Armijo steps decrease calibration risk, while independent evaluation can retain the candidate or restore the exact anchor.

Two identical experts show how this is possible. With outputs f and weights $1/2$, writes $f \pm \delta$ preserve the mixture; later weights $1/2 \pm \epsilon$ change it by $2\epsilon\delta$. We call expert-output changes that keep the routed mixture state fixed the *prediction fiber*. Its fixed-router loss derivative vanishes, but the mixed expert–router derivative can select a task-useful δ for subsequent rerouting.

We introduce Direct Spectral Actuation (DSA), an audit–write–read procedure that acts on this freedom. A feasibility audit constructs a protected function-space chart and checks whether a protected write can induce a router gradient under the original task loss. DSA then writes a protected expert-output residual chosen to maximize the induced router-gradient response at a fixed budget. The residual adapter is frozen, and the original task loss updates only a low-dimensional set of routing coordinates, which we call the *reader*. On native OLMoE, the reader is a local top-8 correction; the

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.

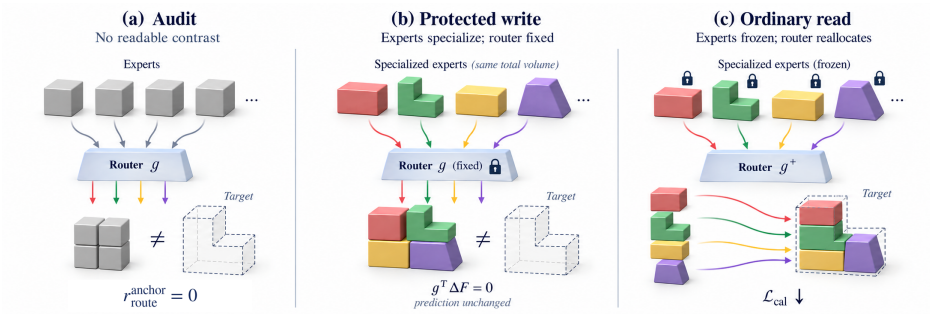

Figure 1: DSA turns prediction-preserving expert-output freedom into a task-relevant routing signal. (a) At identical experts, $r^{\text{anchor}}_{\text{route}} = 0$, so infinitesimal rerouting cannot improve the task loss. (b) With routing fixed, DSA adds protected expert-output residuals satisfying $g^\top \Delta F = 0$, which preserve the anchor prediction exactly. (c) DSA freezes the residual adapter and updates only the reader with the original task loss. A separate deployment test decides whether to retain the post-read candidate.

pretrained router network and experts remain frozen. An independent deployment test then accepts the post-read candidate or restores the frozen anchor.

The prediction fiber contains the expert-output changes that future rerouting can expose without disturbing the anchor. This geometry leads to a write that maximizes the reader's initial task-gradient response: a leading singular mode at a symmetric anchor and a metric-aware trust-region solution at a trained anchor. Accepted steps of the original task loss then decrease calibration risk.

A cross-fitted OpenML-CC18 replay examines when the method should intervene, while controlled tasks and the full 15-by-5 CIFAR-10-C suite verify protected contrast and the subsequent read. On one frozen native top-8 OLMoE checkpoint, DSA improves over matched reader-only adaptation in all six SST-2/RTE evaluations, and three independent SST-2 groups show that quadratic completion also strengthens held-out response cosine and added-gradient power.

Our contributions are:

- **Prediction-preserving post-training differentiation.** We define the prediction fiber as the expert-output perturbations that leave the frozen anchor mixture state unchanged, and formulate post-training differentiation as optimization over this set. Native expert weights remain fixed.
- **Task-gradient actuation with explicit guarantees.** Under a fixed function-space budget, DSA selects the protected write that maximizes the increase in router-gradient power induced by the original task loss. We derive the spectral solution at a symmetric anchor and the metric-aware trust-region solution at a trained anchor, together with exact prediction preservation, a conditional selected-direction transfer bound, and calibration-risk decrease for accepted reader steps.
- **Evaluation of deployment and task benefit.** Across controlled tasks, CIFAR-10-C, 45 OpenML-CC18 tasks, and a native top-8 OLMoE, we separately evaluate protection, induced response, deployment selection, and finite-read performance. DSA improves over matched reader-only adaptation in all six SST-2/RTE evaluations.

2 Related Work

Routing and load allocation. Sparse MoEs regulate expert utilization through auxiliary balancing losses (Fedus et al., 2022), constrained assignment (Lewis et al., 2021), expert-driven token selection (Zhou et al., 2022), or routing-bias updates (Wang et al., 2024). These methods primarily control allocation during training. DSA instead adds a protected expert-output residual after training and then updates a task-loss reader. RoMA aligns routing distributions during adaptation (Li et al., 2026b); Appendix B.5.1 compares a final-layer adaptation of its objective in the same reader chart.

Expert specialization and collapse. Classical mixtures learn a conditional division of labor (Jacobs et al., 1991; Jordan & Jacobs, 1994). Recent methods encourage functional differences through expert outputs, representations, or expert–router coupling (Guo et al., 2025; Li et al., 2026a; Lv et al., 2026); theoretical work studies specialization and collapse (Chi et al., 2022; Chen et al., 2022). DSA instead selects prediction-preserving residuals that maximize the reader gradient under the original task loss.

Function-preserving growth and MoE upcycling. Net2Net applies function-preserving transformations before further training (Chen et al., 2016), while Splitting Steepest Descent uses curvature to choose architecture-expanding splits (Wu et al., 2019). Sparse Upcycling copies a dense FFN into experts (Komatsuzaki et al., 2023), and Drop-Upcycling partially reinitializes the copies (Nakamura et al., 2025). DSA keeps the trained architecture fixed and uses its spectral object to measure router readability rather than where the architecture should grow.

Protected optimization and function-space geometry. Orthogonal Gradient Descent (Farajtabar et al., 2020), gradient surgery (Yu et al., 2020), and Scalpel (Hsieh et al., 2024) project updates to protect selected objectives or directions, while neural tangent methods relate parameter changes to function changes (Jacot et al., 2018). DSA specializes this geometry to a trained MoE: the prediction fiber defines admissible residuals, the mixed expert–router derivative ranks them, and the original task gradient reads the selected direction.

3 Prediction-Preserving Differences Create Routing Responses

A frozen mixture state does not determine the model's response to future rerouting. We characterize the hidden expert-output directions and show that they are exactly the router's local response space.

For one sample, expert k outputs $f_k \in \mathbb{R}^D$. Stack the outputs as rows of $\mathbf{F} = (f_1, \dots, f_K)^\top \in \mathbb{R}^{K\times D}$ and write the router weights as $\mathbf{g} = (g_1, \dots, g_K)^\top \in \Delta_{K-1}$. The routed mixture state is

$$f_{\mathrm{mix}} = \mathbf{F}^\top \mathbf{g} = \sum_{k=1}^{K} g_k f_k. \tag{1}$$

The final prediction may include a fixed downstream network applied to this state. When differentiating with respect to $\mathbf{F}$, the router $\mathbf{g}$ is fixed.

Let $\mathbf{g}_{\mathrm{unif}} = (1/K, \dots, 1/K)^\top$ denote the uniform router and $\mathcal{T}_g = \{v : \sum_k v_k = 0\}$ the simplex tangent space. The centering projector $P_{\mathrm{cent}} = I_K - K\mathbf{g}_{\mathrm{unif}}\mathbf{g}_{\mathrm{unif}}^\top$ removes the component shared by all experts: $(P_{\mathrm{cent}}\mathbf{F})_k = f_k - K^{-1}\sum_j f_j$. For a finite expert-output displacement $\Delta\mathbf{F} \in \mathbb{R}^{K\times D}$, fixing the routed mixture state, and therefore its downstream prediction, is the single constraint

$$\Delta\mathbf{F}^\top \mathbf{g} = 0, \qquad \text{equivalently} \qquad \sum_k g_k \Delta f_k = 0. \tag{2}$$

These protected displacements form the prediction fiber at the frozen router; Δ denotes a finite output change.

3.1 The fused-loss gradient misses protected output directions

At a fixed gate, the fused loss has no gradient component along the prediction fiber. At an identical-expert state, the router is also locally blind.

Proposition 1 (The fused-loss gradient has no protected output-space component)**.** *Let $\ell_{\mathrm{task}}(f_{\mathrm{mix}})$ be differentiable and let $p_{\mathrm{task}} := \nabla_{f_{\mathrm{mix}}}\ell_{\mathrm{task}}(f_{\mathrm{mix}}) \in \mathbb{R}^D$ be its task covector. For fixed router weights,*

$$\nabla_{\mathbf{F}}\ell_{\mathrm{task}} = \mathbf{g}p_{\mathrm{task}}^\top, \qquad D_{\mathbf{F}}\ell_{\mathrm{task}}[\Delta\mathbf{F}] = p_{\mathrm{task}}^\top \Delta\mathbf{F}^\top \mathbf{g} = 0 \tag{3}$$

for every protected write. If $f_1 = \cdots = f_K = f_0$, normalization gives $\sum_k \delta g_k = 0$ for every infinitesimal rerouting, hence $\delta f_{\mathrm{mix}} = f_0 \sum_k \delta g_k = 0$ and the router gradient is zero.

Thus the fixed-gate output gradient cannot select within the prediction fiber, and an identical-expert router has no first-order task signal. This does not rule out parameter updates through unequal expert Jacobians; the precise characterization is in Appendix A.2.

3.2 THE PREDICTION FIBER IS THE LOCAL ROUTING-RESPONSE SPACE

The protected directions omitted by the current mixture are not arbitrary: they are exactly the responses that an infinitesimal rerouting can reveal.

Theorem 2 (Prediction and routing-response decomposition). *Fix* $\mathbf{g}$ *with* $\sum_k g_k = 1$. *Every expert stack* $\mathbf{F}$ *is uniquely determined by its routed mixture state* $f_{\mathrm{mix}} = \mathbf{F}^\top \mathbf{g}$ *and centered expert component* $\mathbf{F}_{\mathrm{cent}} = P_{\mathrm{cent}}\mathbf{F}$. *More precisely:*

(i) The map $\mathbf{F} \mapsto (f_{\mathrm{mix}}, \mathbf{F}_{\mathrm{cent}})$ *is a linear bijection onto centered stacks whose rows sum to zero,* $\sum_k (\mathbf{F}_{\mathrm{cent}})_k = 0$, *with inverse*

$$\mathbf{F} = \mathbf{1}_K f_{\mathrm{mix}}^\top + (I_K - \mathbf{1}_K \mathbf{g}^\top)\mathbf{F}_{\mathrm{cent}}, \tag{4}$$

where $\mathbf{1}_K = (1, \dots, 1)^\top \in \mathbb{R}^K$.

(ii) Prediction-preserving perturbations contain exactly the local routing responses:

$$\{\Delta\mathbf{F} : \Delta\mathbf{F}^\top \mathbf{g} = 0\} \cong \mathrm{Hom}(\mathcal{T}_g, \mathbb{R}^D), \qquad \Delta\mathbf{F} \mapsto \Delta\mathbf{F}^\top|_{\mathcal{T}_g}. \tag{5}$$

Both spaces have dimension $(K-1)D$.

The routed mixture therefore hides $(K-1)D$ expert-output degrees of freedom, each of which changes a local routing response. Protection keeps the current function fixed without eliminating its routing response. The task covector determines which responses are readable, with proofs in Appendices A.3 and A.4.

4 DSA SELECTS PROTECTED WRITES BY THEIR TASK-GRADIENT RESPONSE

Safety alone does not make a fiber direction useful. DSA builds a metric-matched chart, tests whether the task loss can expose a protected write to the reader, and selects the strongest response. We denote the write by A_{write}, reader increment by θ_g, risk by $\mathcal{L}_{\mathrm{cal}}$, response operator by $\mathcal{B}_{\mathrm{read}}$, and write radius by ρ.

4.1 A FIXED METRIC MAKES WRITE BUDGETS COMPARABLE

On calibration sample i, let $f_i^{\mathrm{anchor}} \in \mathbb{R}^D$ be the frozen routed mixture state; a fixed downstream head may map it to the final prediction. Let Σ be the positive-semidefinite centered feature Gram matrix. Quotient its null space and whiten the retained support to obtain $h_i \in \mathbb{R}^m$, where m is the identifiable rank. With the homogeneous coordinate $\bar{h}_i = (h_i^\top, 1)^\top$, $n^{-1}\sum_i \bar{h}_i \bar{h}_i^\top = I_{m+1}$. If $m = 0$, DSA preserves the anchor. Let $C_{\mathrm{ctr}} \in \mathbb{R}^{K\times(K-1)}$ have orthonormal columns orthogonal to $\mathbf{g}_{\mathrm{unif}}$. The linear write A_{write} maps $\bar{h}_i$ to expert contrasts and θ_g is the router increment, with $A_{\mathrm{write}} \in \mathbb{R}^{(K-1)\times D\times(m+1)}$ and $\theta_g \in \mathbb{R}^{(K-1)\times(m+1)}$:

$$\begin{aligned} \Delta\mathbf{F}_i(A_{\mathrm{write}}) &= C_{\mathrm{ctr}} A_{\mathrm{write}}[\bar{h}_i], & \mathbf{F}_i(A_{\mathrm{write}}) &= \mathbf{1}_K (f_i^{\mathrm{anchor}})^\top + \Delta\mathbf{F}_i(A_{\mathrm{write}}), \\ g_i(\theta_g) &= \mathrm{softmax}(KC_{\mathrm{ctr}}\theta_g \bar{h}_i), & f_{\mathrm{mix},i}(A_{\mathrm{write}}, \theta_g) &= \mathbf{F}_i(A_{\mathrm{write}})^\top g_i(\theta_g). \end{aligned} \tag{6}$$

Let $J_{g,i} := D_{\theta_g} g_i(0)$ map reader increments to first-order mixture-weight changes. The basis C_{ctr} supplies coordinates, not semantic expert identities, and it and the feature map remain frozen.

A valid chart must preserve every prediction and make a fixed radius represent the same empirical function change across feature coordinates.

Theorem 3 (Exact protected coordinates with matched metrics). *The construction is intrinsic to the empirical feature-function space: directions in* $\ker\Sigma$ *induce the zero calibration field and are removed, while every identifiable direction is retained. At the anchor* $\theta_g = 0$, *every finite write* A_{write} *satisfies*

$$f_{\mathrm{mix},i}(A_{\mathrm{write}}, 0) = f_i^{\mathrm{anchor}} \qquad \textit{for every calibration sample } i. \tag{7}$$

Moreover, the Euclidean parameter norms equal the corresponding empirical function-space norms:

$$\frac{1}{n}\sum_i \|\Delta \mathbf{F}_i(A_{\text{write}})\|_F^2 = \|A_{\text{write}}\|_F^2, \qquad \frac{1}{n}\sum_i \|J_{g,i}[\Delta\theta_g]\|_2^2 = \|\Delta\theta_g\|_F^2. \tag{8}$$

The protection identity also holds pointwise on new inputs evaluated with the frozen feature map; only the norm identities use calibration averages. Full-rank whitening is the special case $\ker \Sigma = \{0\}$.

Thus every charted write preserves the anchor and equal Euclidean radii give equal empirical function-space budgets. Protection is pointwise; only metric calibration is sample dependent. Proof in Appendix A.5.

4.2 Readability is equivalent to an induced reader gradient

Define the calibration risk using the same per-sample task loss that will be used during the read:

$$\mathcal{L}_{\text{cal}}(A_{\text{write}}, \theta_g) = \frac{1}{n}\sum_i \ell_{\text{task}}(f_{\text{mix},i}(A_{\text{write}}, \theta_g), y_i). \tag{9}$$

At $\theta_g = 0$, exact protection fixes $f_{\text{mix},i}(A_{\text{write}}, 0) = f_i^{\text{anchor}}$; hence the task covector $p_{\text{task},i} = \nabla_f \ell_{\text{task}}(f_i^{\text{anchor}}, y_i)$ does not depend on the write. Here ℓ_{task} includes any fixed downstream head: classification uses the cross-entropy covector pulled back from the full logit vector, and regression uses the specified prediction risk.

Because protection fixes the no-read loss, usefulness is measured by whether the protected residual induces a task-loss reader gradient.

Theorem 4 (Necessary and sufficient condition for readability)**.** *There is a unique linear operator* $\mathcal{B}_{\text{read}}$ *from writes to router gradients such that, for every finite* A_{write}*,*

$$\nabla_{\theta_g} \mathcal{L}_{\text{cal}}(A_{\text{write}}, 0) = \mathcal{B}_{\text{read}} A_{\text{write}}. \tag{10}$$

Consequently, the following statements are equivalent:

$$\begin{aligned} &\exists A_{\text{write}} \text{ with } f_{\text{mix},i}(A_{\text{write}}, 0) = f_i^{\text{anchor}} \text{ and } \nabla_{\theta_g} \mathcal{L}_{\text{cal}}(A_{\text{write}}, 0) \neq 0 \\ &\qquad \iff \quad \mathcal{B}_{\text{read}} \neq 0 \quad \iff \quad \sigma_1(\mathcal{B}_{\text{read}}) > 0. \end{aligned} \tag{11}$$

Theorem 4 turns feasibility into an operator test. If $\sigma_1(\mathcal{B}_{\text{read}}) \leq \varepsilon_B := 10^{-4}$, treated here as a fixed implementation tolerance after whitening, DSA preserves the anchor. Otherwise the prediction fiber contains a direction that the original task loss can expose to the reader. See Appendix A.6 for the Riesz proof and matrix-free implementation.

For a protected write A_{write}, let $\text{conv}_i(A_{\text{write}})$ be the convex hull of all its expert predictions and define the empirical remixing opportunity

$$\Delta_{\text{opp}}(A_{\text{write}}) = \mathcal{L}_{\text{cal}}(A_{\text{write}}, 0) - \frac{1}{n}\sum_i \min_{z \in \text{conv}_i(A_{\text{write}})} \ell_{\text{task}}(z, y_i). \tag{12}$$

Proposition 5 (Readable opportunity)**.** $\Delta_{\text{opp}} \geq 0$*; equality rules out improvement by any router that only remixes these expert outputs. When the anchor is in the relative interior of the reachable mixture simplex, it also forces a zero router gradient, whereas a nonzero router gradient implies* $\Delta_{\text{opp}} > 0$ *and a locally improving remix. An independent deployment test decides whether the post-read candidate is retained.*

Hull opportunity separates feasibility from deployment: a positive value warrants evaluation but not acceptance, which requires an independent paired-risk test. Conditions and proof in Appendix A.12.

4.3 The leading singular mode maximizes reader response

At a symmetric anchor, DSA chooses the equal-budget protected write that maximizes the initial task-loss reader gradient.

Theorem 6 (Maximum router response at a fixed write budget)**.** *Among all protected writes with $\|A_{\mathrm{write}}\|_F \le \rho$, a top right singular mode of $\mathcal{B}_{\mathrm{read}}$ produces the largest initial reader-gradient norm under the original task loss:*

$$\max_{\|A_{\mathrm{write}}\|_F \le \rho} \|\nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0)\|_F = \rho \sigma_1(\mathcal{B}_{\mathrm{read}}). \tag{13}$$

If (V_1, U_1) is a unit top singular pair, then the write $A^\star_{\mathrm{write}} = \rho V_1$ attains the maximum and $\mathcal{B}_{\mathrm{read}} A^\star_{\mathrm{write}} = \rho \sigma_1 U_1$. When σ_1 is repeated, the canonical object is the full top right singular subspace; a fixed tie-break selects one representative. For transfer, let $\mathcal{B}_{\mathrm{read,cal}}$ and $\mathcal{B}_{\mathrm{read,eval}}$ use the same frozen chart and metrics on calibration and held-out data. The calibration-optimal unit mode V_1 satisfies

$$\|\mathcal{B}_{\mathrm{read,eval}} V_1\|_F \ge \sigma_1(\mathcal{B}_{\mathrm{read,cal}}) - \|(\mathcal{B}_{\mathrm{read,eval}} - \mathcal{B}_{\mathrm{read,cal}}) V_1\|_F. \tag{14}$$

The leading mode is the protected residual most readable by the task loss. Section 5 tests whether that response transfers to held-out data; see Appendices A.7 and A.11.

Corollary 7 (Direct actuation and the original-task gradient identity)**.** *Call a protected write* direct *when it is selected by the router-gradient response of the original task loss. For any equal-budget proxy write $A_{\mathrm{write,proxy}}$ and for $A^\star_{\mathrm{write}} = \rho V_1$,*

$$\begin{aligned} \|\nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write,proxy}}, 0)\|_F = \|\mathcal{B}_{\mathrm{read}} A_{\mathrm{write,proxy}}\|_F \le \rho \sigma_1(\mathcal{B}_{\mathrm{read}}), \\ \nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, 0) = \rho \sigma_1 U_1. \end{aligned} \tag{15}$$

A proxy write in the top singular subspace also attains equality. The comparison ranks matched writes by their initial original-task response.

4.4 Quadratic completion extends DSA to trained anchors

A trained MoE may already have a reader gradient. DSA therefore maximizes its increase in squared norm while preserving the native mixture and fixed top-s support.

For each calibration sample, let $\mathbf{F}_i^{\mathrm{anchor}} \in \mathbb{R}^{K \times D}$ and $g_i(\theta_g)$ denote all frozen expert outputs and the local router. The superscript "anchor" always denotes a quantity evaluated at that frozen trained model. Define the sample-wise safe projector

$$P_{i,\mathrm{safe}} := I_K - \frac{g_i^{\mathrm{anchor}} (g_i^{\mathrm{anchor}})^\top}{\|g_i^{\mathrm{anchor}}\|_2^2}. \tag{16}$$

It removes exactly the expert-index component that changes the current mixture. Given any linear raw residual field $Z_i(A_{\mathrm{write}})$, the minimum-norm protected field is therefore

$$\Delta \mathbf{F}_i(A_{\mathrm{write}}) = P_{i,\mathrm{safe}} Z_i(A_{\mathrm{write}}), \qquad \Delta \mathbf{F}_i(A_{\mathrm{write}})^\top g_i^{\mathrm{anchor}} = 0. \tag{17}$$

Then define $f_{\mathrm{mix},i}(A_{\mathrm{write}}, \theta_g) = (\mathbf{F}_i^{\mathrm{anchor}} + \Delta \mathbf{F}_i(A_{\mathrm{write}}))^\top g_i(\theta_g)$. The task loss may include a fixed nonlinear downstream head after this mixed state. Write $c_0 := D_{\theta_g} \mathcal{L}_{\mathrm{cal}}(0, 0)$ for the original-task covector at the anchor, and $CA_{\mathrm{write}} := D_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0) - c_0$ for the covector added by a protected write. Freeze the reader metric M_{route} at the anchor and choose quotient and write charts $\theta_g = Tu$, $A_{\mathrm{write}} = P\bar{A}$ with $T^\top M_{\mathrm{route}} T = I$ and $P^\top M_{\mathrm{write}} P = I$. The implementation uses the physical Frobenius write budget, for which $P = I$. In these coordinates the original-task gradient is $\bar{r} := T^\top c_0$, its mixed response is $\bar{B} := T^\top C P$, and the added router-gradient power is

$$\begin{aligned} \Delta_{\mathrm{grad}}(\bar{A}) &:= \|\bar{r} + \bar{B}\bar{A}\|_2^2 - \|\bar{r}\|_2^2 \\ &= 2\langle \bar{B}^\top \bar{r}, \bar{A} \rangle + \|\bar{B}\bar{A}\|_2^2. \end{aligned} \tag{18}$$

Protection fixes the task covector, and the frozen reader Jacobian makes its gradient affine in the write. The quadratic term is the exact energy of the new task-gradient component, including response orthogonal to the anchor gradient. For $h_{\bar{A}} = \bar{r} + \bar{B}\bar{A}$ and the read path $u = -\eta h_{\bar{A}}$, $\frac{d}{d\eta} \mathcal{L}_{\mathrm{cal}}(P\bar{A}, Tu)\Big|_{\eta=0} = -\|h_{\bar{A}}\|_2^2$; the anchor path has derivative $-\|\bar{r}\|_2^2$. Thus the objective maximizes the original loss's initial reader descent rate under the fixed budget. The audit first applies the readability stop to $\bar{B}$ in this frozen quotient.

The linear term selects the first-order-safe direction by alignment with the anchor gradient; quadratic completion also values the response it induces. Appendix A.10 characterizes when they differ, and SST-2 compares them on held-out data.

Quadratic completion attains at least the first-order-safe calibration response, and selects the same direction under the condition given below. Section 5 reports initial response and finite-read NLL separately. Safety follows from Eq. (17).

Theorem 8 (Protected actuation at a trained MoE anchor)**.** *For the affine identity, assume only frozen all-expert outputs, a linear protected residual field, and a reader differentiable at the anchor. For the budgeted optimum and coordinate statements, additionally let $M_{\mathrm{write}}, M_{\mathrm{route}}$ be frozen positive semidefinite metrics with retained charts P, T on which they are positive definite, and require that zero-cost write directions have zero readable covector action. Then:*

(i) The router gradient is exactly affine in every finite write,

$$D_{\theta_g}\mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0) = c_0 + CA_{\mathrm{write}}. \tag{19}$$

The radius-ρ direct write solves

$$\bar{A}^\star \in \arg\max_{\|\bar{A}\|_2 \le \rho} \Delta_{\mathrm{grad}}(\bar{A}), \qquad A^\star_{\mathrm{write}} = P\bar{A}^\star. \tag{20}$$

It maximizes Δ_{grad} over every equal-budget write in these fixed metrics and reduces to Theorem 6 when $\bar{r} = 0$.

(ii) For any frozen reader metric, the objective is invariant under invertible changes of reader coordinates when the retained chart and metric are transported together. The implementation uses the anchor task loss's generalized Gauss–Newton metric. A gradient step in the whitened quotient uses $\nabla_u \mathcal{L}_{\mathrm{cal}} = T^\top D_{\theta_g}\mathcal{L}_{\mathrm{cal}}$ and maps back to the corresponding metric-scaled task-gradient step in θ_g.

(iii) If top-s support remains frozen, Eq. (17) restricted to selected experts preserves prediction, dispatch, and expert computation. Freezing the native selected mass and applying an anchored softmax to the conditional selected weights keeps every finite reader iterate inside the simplex and exactly recovers the trained anchor. The appendix compares the first written and matched reader-only steps through their actual step sizes and pathwise curvature.

The implementation solves Eq. (20) in the frozen whitened quotient and applies the safe projector. This preserves the mixture, support, native dispatch, and native-expert FLOPs; the residual adapter adds separate cost. A zero anchor gradient recovers the spectral write; proof deferred to Appendix A.13.

Corollary 9 (Quadratic completion and response reserve)**.** *In the frozen whitened trained-anchor quotient define $q := \bar{B}_{\mathrm{cal}}^\top \bar{r}_{\mathrm{cal}}$ and $F_{\mathrm{cal}}(A) := 2\langle q, A\rangle + \|\bar{B}_{\mathrm{cal}}A\|_2^2$ on $\|A\|_2 \le \rho$. Let $A_F = \rho q/\|q\|_2$ when $q \neq 0$, and let A_D maximize F_{cal}. Then*

(i) $F_{\mathrm{cal}}(A_D) \ge F_{\mathrm{cal}}(A_F)$;

(ii) For an evaluation operator expressed in the same frozen reader quotient, let $\epsilon_j := \|(\bar{B}_{\mathrm{eval}} - \bar{B}_{\mathrm{cal}})A_j\|_2$ for $j \in \{D, F\}$. Then

$$\|\bar{B}_{\mathrm{eval}}A_D\|_2 - \|\bar{B}_{\mathrm{eval}}A_F\|_2 \ge m_{\mathrm{cal}} - \epsilon_D - \epsilon_F, \qquad m_{\mathrm{cal}} := \|\bar{B}_{\mathrm{cal}}A_D\|_2 - \|\bar{B}_{\mathrm{cal}}A_F\|_2; \tag{21}$$

the global $m_{\mathrm{cal}} - 2\rho\delta$ bound follows when the operator shift is at most δ;

(iii) For $q \neq 0$, put $G = \bar{B}_{\mathrm{cal}}^\top \bar{B}_{\mathrm{cal}}$. DSA selects the first-order-safe direction exactly when q is an eigenvector of G with eigenvalue γ and $\|q\| \ge \rho[\lambda_{\max}(G) - \gamma]$. An isotropic G is a special case. If $\bar{B}_{\mathrm{cal}} = 0$, the protected-write procedure returns the anchor.

Consequently $m_{\mathrm{cal}} > \epsilon_D + \epsilon_F$ suffices for DSA to retain higher readable response on the evaluation split. Appendix A.10 gives a reader-coordinate-free form for the full objective and a two-dimensional example.

Before the reader update, DSA checks prediction equality, fixed support, and disjoint write/read parameter ownership, then freezes the residual adapter. The quadratic term can retain response energy orthogonal to the existing anchor gradient, but the final task outcome still depends on the finite reader update. Appendix B.2 records the branch and failure order.

4.5 Accepted reader steps decrease calibration risk

The write changes the task gradient without changing the current prediction. With the residual adapter and native experts frozen, the read moves only the reader coordinates in the fixed quotient $\theta_g = Tu$ (or $T = I$ at the symmetric anchor). At iterate $u^{(t)}$ use the original task gradient $h^{(t)} = \nabla_u \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, Tu^{(t)}) = T^\top D_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, Tu^{(t)})$ and test the fixed Armijo trials $\eta_j = 2^{-j}$ for $j = 0, \ldots, 19$, with minimum trial $\eta_{\min} = 2^{-19}$ and accepted step η_{acc}.

Theorem 10 (Certified task-loss-only read)**.** *For the protected expert family above, the following statements hold on the calibration set.*

(i) Whenever $h^{(t)} \neq 0$, the first trial satisfying

$$\mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, T(u^{(t)} - \eta_j h^{(t)})) \leq \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, Tu^{(t)}) - \tfrac{1}{2}\eta_j \|h^{(t)}\|_2^2 \tag{22}$$

decreases the actual calibration risk. A same-support grid step is guaranteed under the one-sided path bound $L_t^+ \leq 2^{19}$ in Appendix A.8; a cross-support trial requires a fully evaluated native forward over any newly active experts; otherwise the method abstains.

(ii) At the symmetric spectral anchor, the first accepted step obeys

$$\mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, 0) - \mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, \theta_g^{(1)}) \geq \tfrac{1}{2}\eta_{\mathrm{acc}} \rho^2 \sigma_1(\mathcal{B}_{\mathrm{read}})^2. \tag{23}$$

Every accepted Armijo step decreases the calibration risk used to select the write. An independent deployment test then retains the candidate or restores the anchor. Proof and procedure in Appendices A.8 and B.2.

5 Experiments

OpenML tests selective intervention, CIFAR-10-C protected writing, SST-2 response transfer, and OLMoE finite-read performance. Response and endpoint are reported separately because DSA optimizes the former.

5.1 Act only on supported opportunities

Controlled tasks verify that the audit detects router-visible opportunity and rejects information-removed controls (Appendix B.3.1). In a cross-fitted OpenML-CC18 replay, the deployment rule acts on 67 of 225 outer models, reaching 22 of 45 tasks. All 22 selected task-level policies improve; all-task gain is 1.12% (95% CI $[0.59, 1.77]\%$), with no harmed policy. Unconditional intervention yields -19.17% and harms seven tasks. Independent risk support thus determines deployment after the feasibility audit (Appendix B.3.2).

5.2 Preserve predictions and create response

The no-read write reproduces anchor predictions and NLL to numerical precision. On the full 15-by-5 CIFAR-10-C suite, the DSA read lowers mean NLL from 1.7729 to 1.7272 and improves 47–62 of 75 cells per seed. At equal radius, initial response is 0.7890, versus 0.1025 for random protected writes and 0.1064 for the normalized proxy. Table 1(A) separates full-suite read outcomes from equal-budget direction selection on a severity-4 development subset.

Across three independently constructed SST-2 groups, DSA raises evaluation added-gradient power from 0.679 to 1.277 (paired difference 0.598, 95% CI $[0.391, 0.862]$) and response cosine from 0.99756 to 0.99966 (difference 0.00210, $[0.00131, 0.01278]$); both favor DSA in all groups. The cosine tests whether the calibration-selected response keeps its direction on new sentences; added-gradient power measures the task signal available there.

Table 1: Protected writing: CIFAR-10-C read gain and equal-budget response (A); paired DSA-minus-first-order-safe SST-2 transfer differences with 95% intervals (B).

A. CIFAR-10-C mechanism and equal-budget directions

Evidence	Condition	Post-read NLL	Initial response
Full 75-cell suite (5 seeds)	Frozen anchor	1.7729	–
	DSA no-read	1.7729	–
	DSA read	1.7272	–
Sev.-4 dev.	No write	2.2996	0.0000
	Random protected	2.2911	0.1025
	Normalized proxy	2.2850	0.1064
	DSA spectral	2.1535	0.7890

B. Quadratic-completion transfer (three frozen groups)

Metric	DSA	First-order-safe	Δ [95% CI]
Evaluation added-gradient power	1.277	0.679	0.598 [0.391,0.862]
Response cosine	0.99966	0.99756	0.00210 [0.00131,0.01278]

5.3 Convert response into task gain

On six partitions from one frozen native top-8 OLMoE checkpoint, the no-read write exactly matches the anchor. After reading, DSA beats the anchor and matched reader-only adaptation in all six evaluations, and the eight-seed random-safe mean in five. The protected write thus creates a response that the original task-loss reader converts into held-out NLL gains.

Table 2: Native OLMoE evaluation NLL ↓/accuracy ↑ (%) per partition. Random-safe is the eight-seed mean; no-read and control details appear in Appendix B.5.1.

Evaluation	Frozen	Router-only	Random-safe	DSA
SST-2 (1)	0.4180/79.30	0.4051/80.47	0.3549/86.21	0.3663/84.18
SST-2 (2)	0.4325/79.59	0.4072/81.64	0.3699/86.11	0.3406/87.40
SST-2 (3)	0.4509/77.34	0.4228/79.39	0.3660/85.34	0.3344/87.60
RTE (1)	1.4249/47.08	0.9764/49.71	0.9079/51.57	0.8608/51.75
RTE (2)	1.5360/45.16	1.0475/46.63	0.9769/48.13	0.8851/48.97
RTE (3)	1.2947/53.08	0.8922/54.55	0.8421/55.90	0.7760/55.43

Relative to the frozen anchor, DSA reduces SST-2 NLL by 12.4–25.8% across three independently calibrated writes and RTE NLL by 40.7% on average across three disjoint evaluation partitions sharing one calibration-selected write. All reads use the original task loss on the same supported top-8 correction, with pretrained router and expert weights frozen. Router-only omits the residual, isolating the contribution of protected writing to these gains.

Conclusion and scope. DSA audits whether intervention is supported, creates expert contrast without changing the anchor prediction, and uses the original task loss to read that contrast.The OpenML replay selects 22 improving task policies; CIFAR-10-C verifies protection before reading; and three independent SST-2 groups show that the added router response transfers. DSA improves over frozen and router-only controls in **all six native OLMoE evaluations**.

The appendix provides additional controls and per-partition results. Hiding router-visible information removes the controlled-task gain. The full CIFAR corruption grid compares matched protected writes. Native OLMoE checks verify write-time protection. The OpenML fold ledger and two preserve-anchor cases document abstention. Against matched final-layer adaptations, DSA has the **lowest six-partition mean NLL**: 0.5939, versus 0.6316 for RoMA, 0.6312 for AuxLoss, and 0.5971 for PERFT with a common read (Appendix B.5.1). Together, the theory and experiments show how a protected write creates a task signal that an ordinary read converts into held-out gain. Native tests use one checkpoint with fixed top-8 support and SST-2/RTE; other checkpoints, tasks, and support-changing reads remain for future study.

AI-USE STATEMENT

Generative AI tools assisted with literature triage, hypothesis development, refinement of the theoretical model, counterexample search, proof checking, experiment design, implementation, debugging and orchestration, result aggregation and interpretation, table preparation, and manuscript editing. The authors checked mathematical claims against complete derivations, code against executable tests, numerical claims against immutable result artifacts, and citations against primary sources. No AI-generated output was treated as an experimental observation. The authors reviewed all AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Appendix B specifies sampling units, data separation, metrics, uncertainty procedures, and numerical certificates. The supplement contains the DSA implementation, claim-bearing runners, structural tests, immutable result JSON, and a claim–evidence manifest with SHA-256 digests. Its deterministic builder reconstructs every main-paper result table from the packaged JSON. External datasets and model weights are identified in the README; large checkpoints, activation caches, and raw data are not redistributed.

REFERENCES

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05641.

Zixiang Chen, Yi Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Zewen Chi, Li Dong, Shaohan Huang, Furu Li, Shuming Ma, Ming Zhou, et al. On the representation collapse of sparse mixture of experts. *NeurIPS*, 2022.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better MoE. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1629.

Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, and Pierre Ablin. Careful with that scalpel: Improving gradient surgery with an EMA. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19085–19100, 2024. URL https://proceedings.mlr.press/v235/hsieh24a.html.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*, 2018.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214, 1994.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.05055.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. *ICML*, 2021.

Jiaang Li, Haibin Chen, Langming Liu, Yujin Yuan, Yadao Wang, Yizhen Zhang, Chengting Yu, Xin Tong, Weidong Zhang, Shilei Liu, Wenbo Su, and Bo Zheng. Expert divergence learning for MoE-based language models. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=wrqYMYazm0.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. Routing manifold alignment improves generalization of mixture-of-experts LLMs. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/249b8d8f41970822651435629e68a6e1-Abstract-Conference.html.

Yilun Liu, Yunpu Ma, Yuetian Lu, Shuo Chen, Zifeng Ding, and Volker Tresp. Parameter-efficient routed fine-tuning: Mixture-of-experts demands mixture of adaptation modules. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4439–4457. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.232. URL https://aclanthology.org/2026.findings-eacl.232/.

Ang Lv, Jin Ma, Yiyuan Ma, and Siyuan Qiao. Coupling experts and routers in mixture-of-experts via an auxiliary loss. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ff18412c1a5816f071a796375abc3d-Abstract-Conference.html.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.02060.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gx1wHnf5Vp.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pp. 353–355. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-5446/.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024. URL https://arxiv.org/abs/2408.15664.

Lemeng Wu, Dilin Wang, and Qiang Liu. Splitting steepest descent for growing neural architectures. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3a01fc0853ebeba94fde4d1cc6fb842a-Abstract.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *NeurIPS*, 2020.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*, 2022.

A Complete Proofs and Extensions

We give the spaces, metrics, and quantifiers used in the proofs. The core results concern the frozen calibration sample; population statements are identified explicitly.

A.1 Standing notation and Hilbert structures

We use the same orientation as the main text: expert k outputs f_k, the rows of $\mathbf{F} = (f_1, \dots, f_K)^\top \in \mathbb{R}^{K\times D}$ stack the complete expert outputs, and $f_{\mathrm{mix}} = \mathbf{F}^\top \mathbf{g} = \sum_k g_k f_k$ is the routed mixture state. A fixed downstream head, when present, maps this state to the final prediction; throughout, ℓ_{task} includes that head. For fixed router parameters, $\mathbf{g}$ is held constant when differentiating with respect to $\mathbf{F}$. Here $\mathbf{g}_{\mathrm{unif}} = K^{-1}\mathbf{1}_K$ is the uniform router and $\mathbf{1}_K = (1, \dots, 1)^\top \in \mathbb{R}^K$; this dimensional subscript is always shown when the all-ones vector is unavoidable. The expert-index tangent space and its orthogonal projector are

$$\mathcal{T}_g = \{v : \textstyle\sum_k v_k = 0\}, \qquad P_{\mathrm{cent}} = I_K - K\mathbf{g}_{\mathrm{unif}}\mathbf{g}_{\mathrm{unif}}^\top.$$

The centering projector removes the component shared by all experts: $(P_{\mathrm{cent}}\mathbf{F})_k = f_k - K^{-1}\sum_j f_j$. The subscript "cent" denotes centering; it is distinct from the contrast basis C_{ctr} introduced later. The Frobenius inner product is $\langle X, Y\rangle_F = \mathrm{tr}(X^\top Y)$.

For the implemented chart, put $c = K - 1$. The write and router spaces are

$$\mathcal{A} = \mathbb{R}^{c\times D\times(m+1)}, \qquad \mathcal{W} = \mathbb{R}^{c\times(m+1)},$$

both with Frobenius inner products. We interpret $A_{\mathrm{write}} \in \mathcal{A}$ as the linear map $\bar{h} \mapsto A_{\mathrm{write}}[\bar{h}] \in \mathbb{R}^{c\times D}$, where $A_{\mathrm{write}}[\bar{h}]_{ad} = \sum_{s=1}^{m+1}(A_{\mathrm{write}})_{ads}\bar{h}_s$. For a calibration field $u = (u_i)_{i=1}^n$, write $\|u\|_n^2 = n^{-1}\sum_i \|u_i\|_2^2$. Every adjoint and singular value uses these declared metrics; without them, an equal-budget or largest-write claim would be coordinate dependent.

The following notation recurs throughout the proofs:

$f_k, \mathbf{F}$	expert output and stack	$g_k, \mathbf{g}$	gate weight and vector
f_{mix}	routed mixture state	P_{cent}	centering projector
A_{write}	protected-write coordinates	θ_g	router coordinates
$\mathcal{B}_{\mathrm{read}}$	write-to-router-gradient map	$\mathcal{L}_{\mathrm{task}}$	task objective
ρ	write-radius budget		

Symbols needed only for one extension are defined locally in that subsection. Roman role labels have the same meaning as in the main text. The superscript "anchor" means evaluation at the frozen pretrained model, $\star$ denotes an optimizer, (t) denotes an iteration, and B^* is reserved for the adjoint of an operator B.

A.2 Fused-loss obstruction and exact-symmetry blindness

This subsection proves Proposition 1. Its two claims have different scopes: the expert obstruction assumes a fixed gate, whereas router blindness assumes identical experts but allows any differentiable normalized router.

Claim 1 (the fusion adjoint). For fixed $\mathbf{g}$, define the fusion map $\Phi_{\mathrm{mix}} : \mathbb{R}^{K\times D} \to \mathbb{R}^D$ by $\Phi_{\mathrm{mix}}(\Delta\mathbf{F}) = \Delta\mathbf{F}^\top\mathbf{g}$. Then

$$\Phi_{\mathrm{mix}}^*(p_{\mathrm{task}}) = \mathbf{g}p_{\mathrm{task}}^\top, \qquad \mathrm{range}\,\Phi_{\mathrm{mix}}^* = (\ker\Phi_{\mathrm{mix}})^\perp.$$

Indeed, for arbitrary $\Delta\mathbf{F}$ and task covector p_{task},

$$\langle p_{\mathrm{task}}, \Delta\mathbf{F}^\top\mathbf{g}\rangle = \mathrm{tr}(p_{\mathrm{task}}^\top\Delta\mathbf{F}^\top\mathbf{g}) = \langle \mathbf{g}p_{\mathrm{task}}^\top, \Delta\mathbf{F}\rangle_F.$$

The range–kernel identity is the standard finite-dimensional adjoint identity.

Proof of Proposition 1. **Step 1: identify the expert gradient.** Here the free variable is the expert stack $\mathbf{F}$; the router $\mathbf{g}$ is fixed. Let $\ell_{\mathrm{task}}(\mathbf{F}) = \ell_{\mathrm{task}}(\mathbf{F}^\top\mathbf{g})$ and put $p_{\mathrm{task}} = \nabla_{f_{\mathrm{mix}}}\ell_{\mathrm{task}}(f_{\mathrm{mix}})$. A finite

write $\Delta\mathbf{F}$ changes the routed mixture state by $f_{\mathrm{mix}} \mapsto f_{\mathrm{mix}} + \Delta\mathbf{F}^\top\mathbf{g}$, where $\Delta\mathbf{F}^\top\mathbf{g} = \sum_k g_k \Delta f_k$. For every protected candidate $\Delta\mathbf{F}$, the chain rule gives

$$D\ell_{\mathrm{task}}(\mathbf{F})[\Delta\mathbf{F}] = \langle p_{\mathrm{task}}, \Delta\mathbf{F}^\top\mathbf{g}\rangle = \langle \mathbf{g}p_{\mathrm{task}}^\top, \Delta\mathbf{F}\rangle_F.$$

Consequently

$$\nabla_{\mathbf{F}}\ell_{\mathrm{task}} = \mathbf{g}p_{\mathrm{task}}^\top \in (\ker\Phi_{\mathrm{mix}})^\perp, \qquad P_{\ker\Phi_{\mathrm{mix}}}\nabla_{\mathbf{F}}\ell_{\mathrm{task}} = 0. \tag{24}$$

In coordinates, $(\nabla_{\mathbf{F}}\ell_{\mathrm{task}})_{kd} = g_k(p_{\mathrm{task}})_d$: the loss derivative for expert k and output coordinate d is the router weight g_k times the corresponding derivative $(p_{\mathrm{task}})_d$ with respect to the mixture state. This proves the protected-component statement without assuming that p is constant during optimization.

Step 2: characterize fixed-gate motion. Along fixed-gate gradient flow, $\dot{\mathbf{F}}(t) = -\mathbf{g}p_{\mathrm{task}}(t)^\top$. Center the expert rows by $\mathbf{F}_{\mathrm{cent}}(t) = P_{\mathrm{cent}}\mathbf{F}(t)$. If the initial experts are collapsed, then $\mathbf{F}_{\mathrm{cent}}(0) = 0$, and integration yields

$$\mathbf{F}_{\mathrm{cent}}(t) = -(P_{\mathrm{cent}}\mathbf{g})\left(\int_0^t p_{\mathrm{task}}(s)\,ds\right)^\top. \tag{25}$$

Hence the column space of $\mathbf{F}_{\mathrm{cent}}(t)$ on the expert axis is contained in the fixed one-dimensional space $\mathrm{span}\{P_{\mathrm{cent}}\mathbf{g}\}$ and $\mathrm{rank}\,\mathbf{F}_{\mathrm{cent}}(t) \le 1$. If $\mathbf{g} = \mathbf{g}_{\mathrm{unif}}$, then $P_{\mathrm{cent}}\mathbf{g} = 0$ and $\mathbf{F}_{\mathrm{cent}}(t) = 0$ for all t. The same conclusion holds for any sequence of discrete output-space gradient steps with this same fixed $\mathbf{g}$. It need not hold after pulling gradients through unequal expert Jacobians or allowing $\mathbf{g}$ to change, which is why the proposition does not assert generic MoE collapse.

Step 3: prove router blindness. For the independent router statement, let $\mathbf{g}(\theta_g)$ be any differentiable normalized gate. Under an arbitrary infinitesimal parameter change, denote the induced gate change by δg_k. Differentiating the normalization identity $\sum_k g_k = 1$ gives $\sum_k \delta g_k = 0$. At identical experts $f_1 = \cdots = f_K = f_0$,

$$\delta f_{\mathrm{mix}} = \sum_k \delta g_k f_k = f_0 \sum_k \delta g_k = 0.$$

Composing with any differentiable task loss makes every router directional derivative zero. Thus the fixed-gate expert obstruction and symmetric-state router blindness meet at the collapsed uniform anchor, and each statement applies exactly within its stated conditions. □

A.3 Prediction–actuation coordinates and their conditioning

This subsection proves parts (i)–(ii) of Theorem 2. It also explains why an algebraically exact fiber still needs the matched calibration metric of Theorem 3.

Fix $\mathbf{g} \in \mathbb{R}^K$ with $\sum_k g_k = 1$ and define

$$\Pi_g = I_K - \mathbf{1}_K\mathbf{g}^\top.$$

It satisfies

$$\Pi_g\mathbf{1}_K = 0, \qquad \mathbf{g}^\top\Pi_g = 0, \qquad P_{\mathrm{cent}}\Pi_g = P_{\mathrm{cent}}, \qquad P_{\mathrm{cent}}\mathbf{F} = \mathbf{F}_{\mathrm{cent}}. \tag{26}$$

Thus Π_g projects onto $\ker\mathbf{g}^\top$ along $\mathrm{span}\{\mathbf{1}_K\}$.

Proof of Theorem 2. **Step 1: reconstruct the experts.** Consider a pair $(f_{\mathrm{mix}}, \mathbf{F}_{\mathrm{cent}})$ with $f_{\mathrm{mix}} \in \mathbb{R}^D$ and centered $\mathbf{F}_{\mathrm{cent}} \in \mathbb{R}^{K\times D}$. Define

$$\Xi_g(f_{\mathrm{mix}}, \mathbf{F}_{\mathrm{cent}}) = \mathbf{1}_K f_{\mathrm{mix}}^\top + \Pi_g\mathbf{F}_{\mathrm{cent}}.$$

Equations (26) give

$$\Xi_g(f_{\mathrm{mix}}, \mathbf{F}_{\mathrm{cent}})^\top\mathbf{g} = f_{\mathrm{mix}}, \qquad P_{\mathrm{cent}}\Xi_g(f_{\mathrm{mix}}, \mathbf{F}_{\mathrm{cent}}) = \mathbf{F}_{\mathrm{cent}}.$$

Conversely, substitute $f_{\mathrm{mix}} = \mathbf{F}^\top\mathbf{g}$ and $\mathbf{F}_{\mathrm{cent}} = P_{\mathrm{cent}}\mathbf{F}$. Direct expansion gives

$$\mathbf{1}_K\mathbf{g}^\top + \Pi_g P_{\mathrm{cent}} = I_K.$$

Therefore

$$\Xi_g(\mathbf{F}^\top \mathbf{g}, P_{\mathrm{cent}}\mathbf{F}) = \mathbf{F}.$$

This proves the claimed linear bijection with the displayed inverse.

Step 2: identify every local routing response. Restriction defines

$$\mathfrak{r} : \{\Delta\mathbf{F} : \Delta\mathbf{F}^\top \mathbf{g} = 0\} \to \mathrm{Hom}(\mathcal{T}_g, \mathbb{R}^D), \qquad \mathfrak{r}(\Delta\mathbf{F}) = \Delta\mathbf{F}^\top|_{\mathcal{T}_g}.$$

Given $L : \mathcal{T}_g \to \mathbb{R}^D$, define its lift on $\mathbb{R}^K$ by $\mathfrak{l}_g(L) = \Pi_g L^\top$. Its transpose is $L\Pi_g^\top$, which annihilates $\mathbf{g}$ and agrees with L on $\mathcal{T}_g$. Hence $\mathfrak{r}\mathfrak{l}_g$ is the identity. If $\Delta\mathbf{F}^\top\mathbf{g} = 0$, then every $x \in \mathbb{R}^K$ has the unique decomposition

$$x = \big(x - \mathbf{g}\textstyle\sum_k x_k\big) + \mathbf{g}\sum_k x_k, \quad x - \mathbf{g}\sum_k x_k \in \mathcal{T}_g,$$

so $\Delta\mathbf{F}^\top x = (\Delta\mathbf{F}^\top|_{\mathcal{T}_g})(x - \mathbf{g}\sum_k x_k)$. Thus $\mathfrak{l}_g\mathfrak{r}$ is also the identity. The two spaces are isomorphic and both have dimension $D(K-1)$. □

The bijection is algebraic, but away from the uniform gate it is oblique. The following calculation gives its exact Euclidean amplification.

Claim 2 (conditioning of the protected lift). If $\sum_k g_k = 1$, then

$$\|\Pi_g\|_2 = \sqrt{K\|\mathbf{g}\|_2^2} \geq 1. \tag{27}$$

Equality holds exactly at $\mathbf{g} = \mathbf{g}_{\mathrm{unif}}$. Moreover, the induced map $L \mapsto \Pi_g L^\top$ from Frobenius-normed linear maps on $\mathcal{T}_g$ to $\mathbb{R}^{K\times D}$ has the same operator norm.

Let $e_g = \mathbf{1}_K/\sqrt{K}$ and write $\mathbf{g} = e_g/\sqrt{K} + t_g$ with $t_g \in \mathcal{T}_g$. Relative to the orthogonal decomposition $\mathbb{R}^K = \mathrm{span}\{e_g\} \oplus \mathcal{T}_g$,

$$\Pi_g^\top(ae_g + y) = y - \sqrt{K}\, a t_g.$$

As a map into $\mathcal{T}_g$, this has block form $[-\sqrt{K}\, t_g \;\; I_{\mathcal{T}_g}]$. Its product with its adjoint is $I_{\mathcal{T}_g} + K t_g t_g^\top$, whose largest eigenvalue is $1 + K\|t_g\|^2 = K\|\mathbf{g}\|^2$. This proves Eq. (27). The lower bound follows from $K\|\mathbf{g}\|^2 \geq (\sum_k g_k)^2 = 1$, with equality exactly at $\mathbf{g} = \mathbf{g}_{\mathrm{unif}}$.

Submultiplicativity gives $\|\Pi_g L^\top\|_F \leq \|L\|_F\|\Pi_g\|_2$. Equality is attained by a rank-one L whose input covector is a top right singular vector of Π_g, a vector that lies in $\mathcal{T}_g$. Hence the lift has the same induced norm.

This conditioning result explains the need for the later isometric chart: the fiber representation says exactly which motions are safe, but an off-uniform gate does not provide a neutral notion of equal write budget.

A.4 Minimality and exhaustiveness of first-order protection

Part (iii) of Theorem 2 gives the closest safe first-order update. The additional result below shows that every globally safe objective must be constant in the protected normal directions; this converse is broader than the exact linear chart used by the main algorithm.

Proof of Theorem 2: minimality. **Step 1: identify the safe projector.** Let $J = J_G$ and take a singular value decomposition $J = U\Sigma V^\top$. The Moore–Penrose identities imply that $J^\dagger J$ is the orthogonal projector onto $\mathrm{range}\, J^\top$ and

$$P_0 = I - J^\dagger J$$

is the orthogonal projector onto $\ker J$. Thus $q_0 = P_0 s$ is feasible.

Step 2: prove closest-point optimality. For any other feasible q, both $q - q_0$ and q_0 lie in $\ker J$, while $s - q_0 = (I - P_0)s$ lies in $\mathrm{range}\, J^\top = (\ker J)^\perp$. Pythagoras gives

$$\|s - q\|_2^2 = \|s - q_0\|_2^2 + \|q - q_0\|_2^2. \tag{28}$$

The second term is nonnegative and vanishes only for $q = q_0$, proving both minimality and uniqueness. No full-row-rank assumption on J is needed. □

Here is a precise version of the exhaustive converse used in the main text.

Lemma 11 (Every globally safe objective factors through safe coordinates)**.** *Let $J : \mathbb{R}^p \to \mathbb{R}^q$ be linear. Let $\Omega \subset \mathbb{R}^p$ be open and suppose each nonempty fiber $\Omega \cap (u + \operatorname{range} J^\top)$ is path connected. Set $P_0 = I - J^\dagger J$. For a continuously differentiable $\psi : \Omega \to \mathbb{R}$, the following are equivalent:*

1. *$J\nabla\psi(\theta) = 0$ for every $\theta \in \Omega$;*
2. *ψ is constant on every nonempty slice $\Omega \cap (u + \operatorname{range} J^\top)$;*
3. *there is a function $\varphi : P_0\Omega \to \mathbb{R}$ such that $\psi = \varphi \circ P_0$ on Ω.*

Proof. Condition 1 says $\nabla\psi(\theta) \in \ker J = (\operatorname{range} J^\top)^\perp$. If $\theta_0, \theta_1 \in \Omega$ have $P_0\theta_0 = P_0\theta_1$, fiberwise path connectivity gives a piecewise C^1 path $\theta(t)$ between them inside that fiber. Its tangent lies in $\operatorname{range} J^\top$, and

$$\frac{d}{dt}\psi(\theta(t)) = \langle \nabla\psi(\theta(t)), \theta'(t)\rangle = 0$$

on each smooth piece. Therefore $\psi(\theta_0) = \psi(\theta_1)$, proving condition 2. This makes $\varphi(u) = \psi(\theta)$ for any θ with $P_0\theta = u$ well-defined and proves condition 3. Condition 3 plainly implies condition 2. Finally, condition 2 implies that for every $v \in \operatorname{range} J^\top$ and all sufficiently small t, $\psi(\theta + tv) = \psi(\theta)$. Differentiating at $t = 0$ gives $\langle \nabla\psi(\theta), v\rangle = 0$. Thus $\nabla\psi(\theta) \in (\operatorname{range} J^\top)^\perp = \ker J$, which is condition 1. If the induced φ is differentiable, the same conclusion can equivalently be written $\nabla\psi = P_0^\top \nabla\varphi = P_0\nabla\varphi$. □

Fiberwise connectivity is the relevant global condition: the proof never uses convexity across different protected coordinates. Without it, a zero normal derivative need not imply a single value on disconnected parts of one fiber. For example, take $J(x, y) = y$, so $P_0(x, y) = (x, 0)$, and $\Omega = \{(x, y) : y > 1\} \cup \{(x, y) : y < -1\}$; the function equal to 0 on the upper component and 1 on the lower component has $J\nabla\psi = 0$ but cannot factor through P_0.

For a nonlinear C^1 observable G, the same projector is only tangent-safe. Let $\omega_G(s) = \sup_{\|z-\theta\|\le s} \|J_G(z) - J_G(\theta)\|_{\mathrm{op}}$ on a compact local neighborhood. If $J_G(\theta)q_0 = 0$, the fundamental theorem of calculus gives

$$\|G(\theta + tq_0) - G(\theta)\| \le |t|\|q_0\| \int_0^1 \omega_G(s|t|\|q_0\|)\, ds \tag{29}$$

whenever the segment stays in that neighborhood. Continuity of J_G makes this $o(|t|\|q_0\|)$; if J_G is L-Lipschitz, it recovers the explicit bound $Lt^2\|q_0\|^2/2$. The exact chart proved next removes this curvature remainder at the anchor for every finite A_{write}. The weaker rate is real: $G(x) = |x|^{3/2}$ has $J_G(0) = 0$ and a continuous Jacobian, but its tangent-safe displacement has order $|t|^{3/2}$, not $O(t^2)$.

A.5 The exact isometric chart

This subsection proves Theorem 3. The three calculations below correspond to its three operational promises: finite anchor protection, an expert-side isometry, and a router-side isometry.

Let $x_i \in \mathbb{R}^{d_x}$ denote the raw frozen features and $\Sigma = n^{-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$. Retain the eigenvectors U_m whose eigenvalues in Λ_m exceed the fixed relative cutoff, and define

$$h_i = \Lambda_m^{-1/2} U_m^\top (x_i - \bar{x}), \qquad \bar{h}_i = (h_i^\top, 1)^\top. \tag{30}$$

Centering makes the constant coordinate orthogonal to h_i, so

$$\frac{1}{n}\sum_{i=1}^n \bar{h}_i \bar{h}_i^\top = I_{m+1}. \tag{31}$$

Also, $v \in \ker\Sigma$ iff $0 = v^\top\Sigma v = n^{-1}\sum_i [v^\top(x_i - \bar{x})]^2$, hence iff v induces the zero centered calibration field. The retained space is therefore the identifiable quotient, not a heuristic projection. Recall that $C_{\mathrm{ctr}} \in \mathbb{R}^{K\times c}$ has orthonormal columns and $C_{\mathrm{ctr}}^\top \mathbf{g}_{\mathrm{unif}} = 0$. Define $\Delta\mathbf{F}_i(A_{\mathrm{write}}) = C_{\mathrm{ctr}} A_{\mathrm{write}}[\bar{h}_i]$ and $\mathbf{F}_i(A_{\mathrm{write}}) = \mathbf{1}_K(f_i^{\mathrm{anchor}})^\top + \Delta\mathbf{F}_i(A_{\mathrm{write}}) \in \mathbb{R}^{K\times D}$; fusion is $\mathbf{F}_i(A_{\mathrm{write}})^\top g_i(\theta_g)$.

Proof of Theorem 3. **Step 1: prove exact finite protection.** At $\theta_g = 0$, all router logits vanish and $g_i(0) = \mathbf{g}_{\text{unif}}$. Since $C_{\text{ctr}}^\top \mathbf{g}_{\text{unif}} = 0$,

$$f_{\text{mix},i}(A_{\text{write}}, 0) = \left(\mathbf{1}_K (f_i^{\text{anchor}})^\top + C_{\text{ctr}} A_{\text{write}} [\bar{h}_i]\right)^\top \mathbf{g}_{\text{unif}} \tag{32}$$

$$= f_i^{\text{anchor}} + A_{\text{write}}[\bar{h}_i]^\top C_{\text{ctr}}^\top \mathbf{g}_{\text{unif}} = f_i^{\text{anchor}}. \tag{33}$$

This identity is algebraic and holds for every finite tensor A_{write}; no Taylor approximation or small-ρ condition is present. It also holds for a new input when its frozen feature and anchor prediction are inserted into the same formula. The empirical whitening identity below need not hold on that input for protection to remain exact.

Step 2: match the expert parameter and function norms. Orthonormality of C_{ctr} gives

$$\frac{1}{n}\sum_i \|\Delta \mathbf{F}_i(A_{\text{write}})\|_F^2 = \frac{1}{n}\sum_i \operatorname{tr}\left(A_{\text{write}}[\bar{h}_i]^\top C_{\text{ctr}}^\top C_{\text{ctr}} A_{\text{write}}[\bar{h}_i]\right) \tag{34}$$

$$= \frac{1}{n}\sum_i \sum_{a,d} \left(\sum_s (A_{\text{write}})_{ads} (\bar{h}_i)_s\right)^2 \tag{35}$$

$$= \sum_{a,d}\sum_{s,t} (A_{\text{write}})_{ads} (A_{\text{write}})_{adt} \left(\frac{1}{n}\sum_i (\bar{h}_i)_s (\bar{h}_i)_t\right) \tag{36}$$

$$= \sum_{a,d,s} (A_{\text{write}})_{ads}^2 = \|A_{\text{write}}\|_F^2, \tag{37}$$

where the penultimate equality uses Eq. (31).

Step 3: match the router parameter and field norms. The softmax differential at the zero logit vector is

$$S_0 = \operatorname{diag}(\mathbf{g}_{\text{unif}}) - \mathbf{g}_{\text{unif}} \mathbf{g}_{\text{unif}}^\top = \frac{1}{K} P_{\text{cent}}.$$

The logit differential in direction $\Delta\theta_g$ is $K C_{\text{ctr}} \Delta\theta_g \bar{h}_i$. Since $P_{\text{cent}} C_{\text{ctr}} = C_{\text{ctr}}$,

$$D_{\theta_g} g_i(0)[\Delta\theta_g] = S_0(K C_{\text{ctr}} \Delta\theta_g \bar{h}_i) = C_{\text{ctr}} \Delta\theta_g \bar{h}_i. \tag{38}$$

Applying $C_{\text{ctr}}^\top C_{\text{ctr}} = I_c$ and whitening once more,

$$\frac{1}{n}\sum_i \|D_{\theta_g} g_i(0)[\Delta\theta_g]\|_2^2 = \frac{1}{n}\sum_i \|\Delta\theta_g \bar{h}_i\|_2^2 \tag{39}$$

$$= \operatorname{tr}\left(\Delta\theta_g \left[\frac{1}{n}\sum_i \bar{h}_i \bar{h}_i^\top\right] \Delta\theta_g^\top\right) = \|\Delta\theta_g\|_F^2. \tag{40}$$

Equations (33), (37), and (40) prove all claims. □

The factor K is forced by metric matching: omitting it would multiply the router-field norm by K^{-2}. The contrast matrix is only a choice of orthonormal coordinates, as the following claim records.

Claim 3 (changing the contrast basis changes no model quantity). Let $O \in \mathbb{R}^{c\times c}$ be orthogonal and put $C'_{\text{ctr}} = C_{\text{ctr}} O$, $A'_{\text{write}}[\bar{h}] = O^\top A_{\text{write}}[\bar{h}]$, and $\theta'_g = O^\top \theta_g$. Then the complete expert outputs, router logits, predictions, and both parameter norms are unchanged. The mixed operators in the two coordinate systems are orthogonally equivalent and therefore have identical singular values. Indeed, the identities $C'_{\text{ctr}} A'_{\text{write}}[\bar{h}] = C_{\text{ctr}} O O^\top A_{\text{write}}[\bar{h}] = C_{\text{ctr}} A_{\text{write}}[\bar{h}]$ and $C'_{\text{ctr}} \theta'_g = C_{\text{ctr}} O O^\top \theta_g = C_{\text{ctr}} \theta_g$ prove equality of the model functions. Orthogonality of O preserves the Frobenius norms. If $T_{\text{write}} A_{\text{write}} = O^\top A_{\text{write}}$ and $T_{\text{route}} \theta_g = O^\top \theta_g$, differentiation of the identical losses gives $\mathcal{B}'_{\text{read}} T_{\text{write}} = T_{\text{route}} \mathcal{B}_{\text{read}}$. Since T_{write} and T_{route} are isometries, $\mathcal{B}'_{\text{read}}$ and $\mathcal{B}_{\text{read}}$ have the same singular spectrum.

Thus "canonical" below is relative to the empirical whitening metric but not to an arbitrary basis for the contrast subspace.

A.6 CONSTRUCTION OF THE RESTRICTED MIXED-DERIVATIVE OPERATOR

This subsection proves Theorem 4. The essential step is to differentiate the protected model in the router direction while exact protection keeps the task covector fixed.

Let $p_{\mathrm{task},i} = \nabla_f \ell_{\mathrm{task}}(f_i^{\mathrm{anchor}}, y_i) \in \mathbb{R}^D$. Exact anchor protection is what makes these covectors independent of A_{write}. We now derive the operator in both coordinate-free and dense forms.

Claim 4 (the mixed operator and its adjoint). For every $A_{\mathrm{write}} \in \mathcal{A}$,

$$(\mathcal{B}_{\mathrm{read}} A_{\mathrm{write}})_{as} = \frac{1}{n} \sum_{i=1}^{n} (A_{\mathrm{write}}[\bar{h}_i] p_{\mathrm{task},i})_a (\bar{h}_i)_s. \tag{41}$$

For every $U \in \mathcal{W}$, the adjoint is

$$(\mathcal{B}^*_{\mathrm{read}} U)_{ads} = \frac{1}{n} \sum_{i=1}^{n} (U \bar{h}_i)_a (p_{\mathrm{task},i})_d (\bar{h}_i)_s. \tag{42}$$

The following display is the dense-coordinate representation of the same operator $\mathcal{B}_{\mathrm{read}}$, not a second object. Its row index is (a, s) and its column index is (a', d, t):

$$(\mathcal{B}_{\mathrm{read}})_{(a,s),(a',d,t)} = \delta_{aa'} \frac{1}{n} \sum_i (p_{\mathrm{task},i})_d (\bar{h}_i)_t (\bar{h}_i)_s. \tag{43}$$

Step 1: differentiate the prediction. Differentiate the routed mixture state with respect to θ_g at the anchor. By Eq. (38),

$$D_{\theta_g} f_{\mathrm{mix},i}(A_{\mathrm{write}}, 0)[\Delta\theta_g] = \Delta \mathbf{F}_i(A_{\mathrm{write}})^\top D_{\theta_g} g_i(0)[\Delta\theta_g] \tag{44}$$

$$= A_{\mathrm{write}}[\bar{h}_i]^\top C_{\mathrm{ctr}}^\top C_{\mathrm{ctr}} \Delta\theta_g \bar{h}_i = A_{\mathrm{write}}[\bar{h}_i]^\top \Delta\theta_g \bar{h}_i. \tag{45}$$

The base term contributes zero because $\sum_k D_{\theta_g} g_{ik}(0)[\Delta\theta_g] = 0$.

Step 2: identify the Riesz representative. Applying the chain rule to the empirical risk gives

$$D_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0)[\Delta\theta_g] = \frac{1}{n} \sum_i \langle p_{\mathrm{task},i}, A_{\mathrm{write}}[\bar{h}_i]^\top \Delta\theta_g \bar{h}_i \rangle \tag{46}$$

$$= \left\langle \frac{1}{n} \sum_i (A_{\mathrm{write}}[\bar{h}_i] p_{\mathrm{task},i}) \bar{h}_i^\top, \Delta\theta_g \right\rangle_F. \tag{47}$$

The Riesz representative is exactly Eq. (41). Expanding $A_{\mathrm{write}}[\bar{h}_i]$ gives Eq. (43).

Step 3: derive the adjoint. Finally,

$$\langle \mathcal{B}_{\mathrm{read}} A_{\mathrm{write}}, U \rangle_F = \frac{1}{n} \sum_i \langle A_{\mathrm{write}}[\bar{h}_i] p_{\mathrm{task},i}, U \bar{h}_i \rangle \tag{48}$$

$$= \sum_{a,d,s} (A_{\mathrm{write}})_{ads} \left[\frac{1}{n} \sum_i (U \bar{h}_i)_a (p_{\mathrm{task},i})_d (\bar{h}_i)_s \right], \tag{49}$$

which is $\langle A_{\mathrm{write}}, \mathcal{B}^*_{\mathrm{read}} U \rangle_F$ with Eq. (42). This also proves the adjoint identity used as the implementation certificate.

The dense representation reveals additional structure. If

$$M_{s,(d,t)} = \frac{1}{n} \sum_i (p_{\mathrm{task},i})_d (\bar{h}_i)_t (\bar{h}_i)_s,$$

then the coordinate matrix of $\mathcal{B}_{\mathrm{read}}$ is $I_c \otimes M$ up to the fixed vectorization convention. Hence every singular value of M is repeated across the c contrast coordinates. In a permutation-symmetric chart, a top singular vector is therefore generally not unique; the invariant object is the full top singular subspace. A deterministic coordinate tie-break chooses a representative without changing the attainable norm.

Proof of Theorem 4. Equation (47) shows that $A_{\mathrm{write}} \mapsto \nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0)$ is the linear map $\mathcal{B}_{\mathrm{read}}$ given above. Its linearity is exact for finite A_{write}, not merely the first term of an expansion: at $\theta_g = 0$ the prediction is exactly f_i^{anchor}, while its router differential is linear in $\Delta \mathbf{F}_i(A_{\mathrm{write}})$.

If $\mathcal{B}_{\mathrm{read}} \neq 0$, then by definition there exists A_{write} with $\mathcal{B}_{\mathrm{read}} A_{\mathrm{write}} \neq 0$. Theorem 3 simultaneously gives $f_{\mathrm{mix},i}(A_{\mathrm{write}}, 0) = f_i^{\mathrm{anchor}}$ for every i, so this protected update induces a nonzero router-gradient response. Conversely, any protected chart write with nonzero ordinary router gradient satisfies $\mathcal{B}_{\mathrm{read}} A_{\mathrm{write}} \neq 0$, hence $\mathcal{B}_{\mathrm{read}} \neq 0$. Finally, for a linear operator on finite-dimensional Hilbert spaces,

$$\sigma_1(\mathcal{B}_{\mathrm{read}}) = \|\mathcal{B}_{\mathrm{read}}\|_{\mathrm{op}} = \max_{\|A_{\mathrm{write}}\|_F = 1} \|\mathcal{B}_{\mathrm{read}} A_{\mathrm{write}}\|_F,$$

which is positive exactly when $\mathcal{B}_{\mathrm{read}}$ is nonzero. This proves all three equivalences. □

Two immediate failure certificates follow from the formula. If every $p_{\mathrm{task},i} = 0$, then $\mathcal{B}_{\mathrm{read}} = 0$ because each anchor prediction is already a per-sample stationary point of the loss. More generally, cancellations in the task-weighted feature moment $n^{-1} \sum_i p_{\mathrm{task},i} \otimes \bar{h}_i \otimes \bar{h}_i$ can make $\mathcal{B}_{\mathrm{read}} = 0$ despite a nontrivial protected fiber. Fiber dimension and task-induced router response are therefore logically distinct.

A.7 Spectral optimality, equality cases, and fixed-reader slices

This subsection proves Theorem 6 and Corollary 7. The theorem identifies all optimal equal-budget writes; the corollary translates that variational fact into the ordinary task-gradient response.

Proof of Theorem 6. **Step 1: obtain the universal upper bound.** Because $\mathcal{A}$ and $\mathcal{W}$ are finite-dimensional Hilbert spaces, $\mathcal{B}_{\mathrm{read}}^* \mathcal{B}_{\mathrm{read}}$ is self-adjoint and positive semidefinite. Let σ_1^2 denote its largest eigenvalue and

$$\mathcal{V}_1 = \ker(\mathcal{B}_{\mathrm{read}}^* \mathcal{B}_{\mathrm{read}} - \sigma_1^2 I)$$

its top right singular subspace. For every feasible A_{write},

$$\|\mathcal{B}_{\mathrm{read}} A_{\mathrm{write}}\|_F^2 = \langle A_{\mathrm{write}}, \mathcal{B}_{\mathrm{read}}^* \mathcal{B}_{\mathrm{read}} A_{\mathrm{write}} \rangle_F \leq \sigma_1^2 \|A_{\mathrm{write}}\|_F^2 \leq \rho^2 \sigma_1^2. \tag{50}$$

Step 2: characterize every equality case. If $\sigma_1 > 0$, equality in the first inequality holds exactly when $A_{\mathrm{write}} \in \mathcal{V}_1$, and equality in the second holds exactly when $\|A_{\mathrm{write}}\|_F = \rho$. Thus the complete maximizing set is

$$\{A_{\mathrm{write}} \in \mathcal{V}_1 : \|A_{\mathrm{write}}\|_F = \rho\}. \tag{51}$$

For a unit $V_1 \in \mathcal{V}_1$, define $U_1 = \mathcal{B}_{\mathrm{read}} V_1 / \sigma_1$. Then $\|U_1\|_F = 1$, $\mathcal{B}_{\mathrm{read}} V_1 = \sigma_1 U_1$, and $\mathcal{B}_{\mathrm{read}}^* U_1 = \sigma_1 V_1$. Taking $A_{\mathrm{write}}^\star = \rho V_1$ attains the upper bound.

Step 3: handle the zero-spectrum boundary. If $\sigma_1 = 0$, then $\mathcal{B}_{\mathrm{read}} = 0$ and every point of the closed budget ball is a maximizer with objective value zero; no protected direction induces a nonzero router-gradient response for the algorithm to select. The algorithm therefore returns the unchanged anchor rather than calling an arbitrary zero-amplitude mode canonical. □

Proof of Corollary 7: equal-budget bound. For every auxiliary or proxy write obeying the same chart and budget, Eq. (50) gives

$$\|\nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{aux}}^{\mathrm{write}}, 0)\|_F = \|\mathcal{B}_{\mathrm{read}} A_{\mathrm{aux}}^{\mathrm{write}}\|_F \leq \rho \sigma_1.$$

When $\sigma_1 > 0$, equality requires both full budget and membership in $\mathcal{V}_1$, by Eq. (51). If $A_{\mathrm{aux}}^{\mathrm{write}} \in \ker \mathcal{B}_{\mathrm{read}}$, its ordinary router gradient is exactly zero even if $A_{\mathrm{aux}}^{\mathrm{write}} \neq 0$ and the expert residual field has positive norm. This is an equal-budget statement about the initial router-gradient response. It contains no ordering of later optimization trajectories or held-out risks. □

The single-reader construction is recovered by restricting the reader before optimizing the write.

A.7.1 FIXED-READER RANK-ONE SLICE

Let $\mathbf{g} \in \Delta_{K-1}$ be a current gate, let $v_g \in \mathbb{R}^K$ be a fixed gate displacement, and define $q_g = P_{\mathbf{g}^\perp} v_g = v_g - \mathbf{g}(\mathbf{g}^\top v_g)/\|\mathbf{g}\|^2$. Let $p_{\text{task}} \in \mathbb{R}^D$ be a task covector. Among output writes $\Delta\mathbf{F} \in \mathbb{R}^{K\times D}$ satisfying $\Delta\mathbf{F}^\top \mathbf{g} = 0$ and $\|\Delta\mathbf{F}\|_F \le \rho$, consider the signed mixed score $\langle p_{\text{task}}, \Delta\mathbf{F}^\top v_g\rangle$.

Since $v_g = q_g + \lambda\mathbf{g}$, protection gives $\Delta\mathbf{F}^\top v_g = \Delta\mathbf{F}^\top q_g$. Moreover, $q_g p_{\text{task}}^\top$ is protected because $q_g^\top \mathbf{g} = 0$. Therefore

$$\langle p_{\text{task}}, \Delta\mathbf{F}^\top v_g\rangle = \langle q_g p_{\text{task}}^\top, \Delta\mathbf{F}\rangle_F \tag{52}$$

$$\ge -\|q_g p_{\text{task}}^\top\|_F \|\Delta\mathbf{F}\|_F \ge -\rho\|p_{\text{task}}\|_2\|q_g\|_2. \tag{53}$$

When $p_{\text{task}} \neq 0$ and $q_g \neq 0$, both inequalities are equalities uniquely at

$$\Delta\mathbf{F}_\xi^* = -\rho \frac{q_g p_{\text{task}}^\top}{\|p_{\text{task}}\|_2\|q_g\|_2}. \tag{54}$$

If $p_{\text{task}} = 0$ or $q_g = 0$, every protected write has zero score and there is no strictly preferred direction. Equation (54) is therefore the steepest write for one fixed reader functional. Allowing the reader direction itself to vary replaces this scalar functional by the norm $\|\mathcal{B}_{\text{read}} A_{\text{write}}\|$ and yields the spectral problem.

A.8 ORDINARY-GRADIENT IDENTITY AND FINITE READ CERTIFICATION

This subsection proves the response identity in Corollary 7 and the finite-step part of Theorem 10. Spectral optimality controls the initial signal; Armijo controls each actual risk-changing step.

Proof of Corollary 7: ordinary-gradient identity. The mixed-gradient identity is $\nabla_{\theta_g}\mathcal{L}_{\text{cal}}(A_{\text{write}}, 0) = \mathcal{B}_{\text{read}} A_{\text{write}}$. Substituting $A_{\text{write}}^\star = \rho V_1$ and the singular equation $\mathcal{B}_{\text{read}} V_1 = \sigma_1 U_1$ gives

$$\nabla_{\theta_g}\mathcal{L}_{\text{cal}}(A_{\text{write}}^\star, 0) = \rho\sigma_1 U_1.$$

This is the gradient of the original empirical task risk in the actual router parameter θ_g. No auxiliary router objective appears. Although the simultaneous sign of (V_1, U_1) is arbitrary, a sign flip changes both sides of the identity and leaves the gradient norm and certified descent capacity unchanged. □

Claim 5 (one-sided descent on the tested router ray). Let $r = \nabla\mathcal{L}(\theta_g)$ and $f(s) = \mathcal{L}(\theta_g - sr)$. Suppose f is absolutely continuous on $[0, \eta]$, differentiable at zero, and for almost every s on the path $f'(s) - f'(0) \le L_+ s\|r\|_F^2$ with $L_+ \ge 0$. Then

$$\mathcal{L}(\theta_g - \eta r) \le \mathcal{L}(\theta_g) - \eta\left(1 - \frac{L_+\eta}{2}\right)\|r\|_F^2. \tag{55}$$

Indeed $f'(0) = -\|r\|_F^2$, and integration of the one-sided bound gives Eq. (55). A gradient L-Lipschitz on this segment implies the stated condition with $L_+ = L$, but is not needed. For example, $f(s) = f(0) - s - s^{3/2}$ has a non-Lipschitz derivative at zero and satisfies the one-sided condition with $L_+ = 0$.

Proof of Proposition 5. Exact protection fixes the anchor prediction, so the convex hull of the complete expert outputs contains the anchor. Hence the pointwise minimum in Δ_{opp} is no larger than the anchor loss, proving $\Delta_{\text{opp}} \ge 0$. Every router that remixes the complete experts produces a prediction in that hull, so its calibration risk is at least the hull term; if the gap is zero, no such router can improve on the anchor. At zero opportunity, if the anchor lies in the relative interior of the reachable mixture simplex, all feasible tangent derivatives vanish, so the router gradient is zero. Conversely, a nonzero router gradient gives a first-order descent direction inside the hull, and therefore a strictly positive opportunity. These identities are local to the frozen calibration chart and complete-expert family. □

Proof of Theorem 10. **Step 1: use acceptance as a finite certificate.** Write $\mathcal{L}(u) = \mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, Tu)$ while the base, $A^\star_{\mathrm{write}}$, and quotient chart T remain frozen; the symmetric case has $T = I$. Let $h^{(t)} = \nabla_u \mathcal{L}(u^{(t)})$. If a trial η_j is accepted, the acceptance inequality itself yields

$$\mathcal{L}(u^{(t)}) - \mathcal{L}(u^{(t)} - \eta_j h^{(t)}) \geq \frac{1}{2}\eta_j \|h^{(t)}\|_2^2. \tag{56}$$

This decrease is strict whenever $h^{(t)} \neq 0$. If $h^{(t)} = 0$, the iterate is stationary and the read phase should terminate rather than report a strict improvement. Crucially, this implication is purely arithmetic: it does not require differentiability between the two endpoints. Therefore it also certifies a trial crossing a kink or routing-cell boundary, provided both values are evaluations of the complete native task risk. If the crossing activates an expert absent from the evaluation cache, the candidate value is not the native risk and no certificate may be issued.

Step 2: show when the fixed grid contains an acceptable step. Suppose the minimum trial $\eta_{\min} = 2^{-19}$ remains on a same-support path whose original risk is absolutely continuous and obeys Claim 5 with $L_t^+ \leq 2^{19}$. Then $L_t^+ \eta_{\min} \leq 1$, so Eq. (55) shows that this trial satisfies the Armijo condition with constant $1/2$; the grid therefore accepts some trial. Only one-sided curvature along that single path is needed, not a bound on the full gradient map or on untested directions. If all twenty trials fail, the sufficient path condition was not met, a support crossing lacked a complete native evaluation, or numerical risk evaluation was inconsistent. The path condition guarantees existence of an accepted same-cell step; it is not needed for the validity of a fully evaluated accepted cross-cell step.

Step 3: specialize the first step to the spectral write. At the symmetric first iterate, Corollary 7 and $\|U_1\|_F = 1$ give $\|h^{(0)}\|_2^2 = \rho^2 \sigma_1^2$. Substitution into Eq. (56) proves

$$\mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, 0) - \mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, \theta_g^{(1)}) \geq \frac{1}{2}\eta_{\mathrm{acc}} \rho^2 \sigma_1^2.$$

The same reasoning certifies each later accepted step using its actual $h^{(t)}$, but spectral optimality was proved only at the anchor and does not claim trajectory-wide optimality. □

A.9 Why calibration alone cannot guarantee future improvement

Theorem 12 (Cross-split no-free-lunch)**.** *Let an algorithm choose a write q using only a calibration sigma-field $\mathcal{F}_{\mathrm{cal}}$—that is, any information measurable from the calibration data, the object on which every audit, fitted map, and statistic below lives—and, if randomized, internal randomness independent of the future. Without an assumption relating calibration and future mixed covectors, there exist two worlds with identical law on $\mathcal{F}_{\mathrm{cal}}$ but opposite future benefit covectors h and $-h$. The algorithm cannot guarantee strictly positive first-order future gain in both worlds.*

Proof. Couple the two worlds so that their calibration observations and the algorithm's internal random seed are identical. The algorithm must then produce the same q in both. Conditional on this coupled information, the two first-order gains are

$$G_+(q) = \langle h, q \rangle, \qquad G_-(q) = \langle -h, q \rangle = -G_+(q).$$

They cannot both be strictly positive; if $q \perp h$, both are zero. Such worlds can be realized by smooth local future risks with opposite derivatives at the anchor, for example quadratic risks centered on opposite sides of the anchor, while leaving the entire calibration distribution unchanged. Taking expectations over the coupled internal randomness preserves the sign opposition. Therefore randomization does not remove the impossibility. □

The theorem leaves the calibration guarantees intact: protected write, ordinary router-gradient response, and Armijo descent all remain exactly auditable there. It identifies the additional ingredient required for a population or future-split claim—a stability, invariance, or distributional relation that connects the two mixed operators.

A.10 Nonredundancy and a bounded-shift lower bound for the quadratic target

The DSA write maximizes the exact added response $F(a) := \|\bar{r} + \bar{B}a\|^2 - \|\bar{r}\|^2 = 2\langle \bar{B}^\top \bar{r}, a \rangle + \|\bar{B}a\|^2$ at fixed radius. To see what the completion adds, let $\bar{r} \neq 0$, $z = \bar{B}a$, $e = \bar{r}/\|\bar{r}\|$, $\alpha = \langle e, z \rangle$,

and $z_\perp = z - \alpha e$. Orthogonality gives the exact decomposition

$$F(a) = 2\|\bar{r}\|\alpha + \alpha^2 + \|z_\perp\|^2. \tag{57}$$

First-order-safe retains the local alignment α; DSA also values response energy along and across that gradient. At $\bar{r} = 0$ the same construction becomes the spectral selection of $\|z\|^2$. The following two-dimensional construction shows that these objectives are not equivalent once the evaluation operator differs from the calibration operator.

Let $B_{\mathrm{c}} = \mathrm{diag}(4, 1)$, $r_{\mathrm{c}} = (0, 1)$, and $\rho = 1$. The linear term is $2a_2$, so a first-order-safe unit write is $a_{\mathrm{F}} = e_2$. The complete objective is

$$2a_2 + \|B_{\mathrm{c}}a\|^2 = 16 + 2a_2 - 15a_2^2,$$

whose positive-sign maximizer on the unit circle is

$$a_{\mathrm{D}} = \left(\frac{\sqrt{224}}{15}, \frac{1}{15}\right).$$

A first-order direction therefore spends its entire budget on the coordinate whose calibration response is one. If the evaluation operator is $B_{\mathrm{e}} = \mathrm{diag}(4, 0)$, then $\|B_{\mathrm{e}}a_{\mathrm{F}}\| = 0$ while $\|B_{\mathrm{e}}a_{\mathrm{D}}\| = 4\sqrt{224}/15$. The quadratic term is thus nonredundant: it can select a direction whose response survives an operator change that annihilates the first-order choice.

Proof of Corollary 9. **(i) Complete-objective dominance.** By definition A_D maximizes F_{cal} on the feasible radius ball, while A_F is feasible; hence $F_{\mathrm{cal}}(A_D) \geq F_{\mathrm{cal}}(A_F)$.

(ii) Comparative response transport. Apply the reverse triangle inequality only along the two calibration-selected writes:

$$\|\bar{B}_{\mathrm{eval}}A_D\|_2 \geq \|\bar{B}_{\mathrm{cal}}A_D\|_2 - \epsilon_D, \qquad \|\bar{B}_{\mathrm{eval}}A_F\|_2 \leq \|\bar{B}_{\mathrm{cal}}A_F\|_2 + \epsilon_F.$$

Subtracting proves Eq. (21). This requires only the two directional shifts; if $\|\bar{B}_{\mathrm{eval}} - \bar{B}_{\mathrm{cal}}\|_{\mathrm{op}} \leq \delta$, then $\epsilon_D + \epsilon_F \leq 2\rho\delta$. This is a response statement, not a guarantee about finite-step held-out NLL.

(iii) Exact first-order fallback criterion. Let $q \neq 0$ and $G = \bar{B}_{\mathrm{cal}}^\top \bar{B}_{\mathrm{cal}}$. If $A_F = \rho q/\|q\|$ maximizes the complete objective, its tangential derivative on the sphere must vanish; since the linear term is already radial at A_F, $Gq = \gamma q$ is necessary. The one-dimensional write case is immediate. In higher dimension, the convex quadratic attains its maximum on the boundary, so write any radius-ρ candidate as $A = \rho(tq/\|q\| + \sqrt{1-t^2}v)$, with $v \perp q$, $\|v\| = 1$, and $-1 \leq t \leq 1$. If $\lambda_\perp$ is the largest eigenvalue of G on $q^\perp$, then

$$F_{\mathrm{cal}}(A) - F_{\mathrm{cal}}(A_F) \leq (1-t)\{-2\rho\|q\| + \rho^2(\lambda_\perp - \gamma)(1+t)\}.$$

This is nonpositive for every candidate exactly when $\|q\| \geq \rho[\lambda_{\max}(G) - \gamma]$; if it fails, a candidate with $t < 1$ sufficiently close to one improves the objective. Thus the criterion is necessary and sufficient. It includes isotropic G but also anisotropic operators whose linear alignment is strong enough or lies in a top mode. If $\bar{B}_{\mathrm{cal}} = 0$, the algorithm returns the anchor. □

Transport of the complete objective without matching reader bases. The norm comparison above uses a shared reader quotient to subtract response vectors. The objective that DSA actually optimizes admits a weaker comparison. Let calibration and evaluation have possibly different reader Hilbert spaces, but share the protected write space and its budget. For each split $s \in \{\mathrm{c}, \mathrm{e}\}$, define the write-space pullbacks $q_s := B_s^* r_s$ and $G_s := B_s^* B_s$. Then

$$F_s(a) := \|r_s + B_s a\|^2 - \|r_s\|^2 = 2\langle q_s, a\rangle + \langle a, G_s a\rangle.$$

Put $\gamma_s := F_s(A_D) - F_s(A_F)$ and $S := A_D A_D^* - A_F A_F^*$. Expansion in the common write space gives the exact identity and its dual-norm bound:

$$\begin{aligned} \gamma_{\mathrm{e}} &= \gamma_{\mathrm{c}} + 2\langle q_{\mathrm{e}} - q_{\mathrm{c}}, A_D - A_F\rangle + \mathrm{tr}[(G_{\mathrm{e}} - G_{\mathrm{c}})S], \\ \gamma_{\mathrm{e}} &\geq \gamma_{\mathrm{c}} - 2\|q_{\mathrm{e}} - q_{\mathrm{c}}\|\,\|A_D - A_F\| - \|G_{\mathrm{e}} - G_{\mathrm{c}}\|_{\mathrm{op}}\,\|S\|_*. \end{aligned} \tag{58}$$

Here $\|S\|_*$ is the nuclear norm and $\gamma_{\mathrm{c}} \geq 0$ by DSA's complete-objective choice. If both writes have norm ρ, then $\|S\|_* = 2\rho^2\sqrt{1 - \langle A_D, A_F\rangle^2/\rho^4}$. This comparison needs no subtraction of operators between reader spaces: split-specific reader isometries leave q_s and G_s unchanged. It also vanishes when the two directions coincide, matching the first-order-safe fallback. The bound addresses initial added-gradient power; the independent read still determines finite-step task risk.

A.11 CALIBRATION-TO-EVALUATION STABILITY OF SPECTRAL ACTUATION

The no-free-lunch result above identifies what is missing from a calibration argument. The next theorem supplies that missing condition directly in the object that governs readability: the restricted mixed operator. Unlike a generic similarity assumption, it gives a quantitative held-out response bound in the same function-space units used to choose the write.

Theorem 13 (Selected-mode transfer and singular-subspace stability)**.** *Let $\mathcal{B}_{\mathrm{read,cal}}, \mathcal{B}_{\mathrm{read,eval}} : \mathcal{A} \to \mathcal{W}$ be bounded linear maps between the same finite-dimensional Hilbert spaces, constructed with one frozen chart and its fixed metrics. Let V_{c} be any unit top right singular vector of $\mathcal{B}_{\mathrm{read,cal}}$ and put $\epsilon_{\mathrm{c}} = \|(\mathcal{B}_{\mathrm{read,eval}} - \mathcal{B}_{\mathrm{read,cal}})V_{\mathrm{c}}\|$. Then, without any bound on shifts in other directions,*

$$\|\mathcal{B}_{\mathrm{read,eval}}V_{\mathrm{c}}\| \geq \sigma_1(\mathcal{B}_{\mathrm{read,cal}}) - \epsilon_{\mathrm{c}}. \tag{59}$$

If a global shift bound $\delta = \|\mathcal{B}_{\mathrm{read,eval}} - \mathcal{B}_{\mathrm{read,cal}}\|_{\mathrm{op}}$ is also available, then

$$0 \leq \sigma_1(\mathcal{B}_{\mathrm{read,eval}}) - \|\mathcal{B}_{\mathrm{read,eval}}V_{\mathrm{c}}\| \leq \delta + \epsilon_{\mathrm{c}} \leq 2\delta. \tag{60}$$

More generally, every write A obeys

$$\|\mathcal{B}_{\mathrm{read,eval}}A\| \geq \|\mathcal{B}_{\mathrm{read,cal}}A\| - \|(\mathcal{B}_{\mathrm{read,eval}} - \mathcal{B}_{\mathrm{read,cal}})A\|. \tag{61}$$

Suppose in addition that the top calibration singular subspace is one-dimensional (equivalently, the top squared singular value is simple). Write $s_j = \sigma_j(\mathcal{B}_{\mathrm{read,cal}})$, $\Delta = s_1^2 - s_2^2$, and $\eta = \delta(2s_1 + \delta)$. If $\eta < \Delta/2$, the top evaluation right singular space is one-dimensional. If P_{e} denotes its orthogonal projector, then

$$\|(I - P_{\mathrm{e}})V_{\mathrm{c}}\| \leq \frac{\eta}{\Delta - \eta}. \tag{62}$$

Thus selected-mode stability alone preserves that mode's response; global operator control plus a spectral gap additionally identifies the evaluation top write subspace. With repeated top singular values, the invariant object is the full top right singular subspace, not an arbitrarily chosen vector.

Proof. **Step 1: transfer the response of the selected write.** The reverse triangle inequality and $\|V_{\mathrm{c}}\| = 1$ give

$$\|\mathcal{B}_{\mathrm{read,eval}}V_{\mathrm{c}}\| \geq \|\mathcal{B}_{\mathrm{read,cal}}V_{\mathrm{c}}\| - \|(\mathcal{B}_{\mathrm{read,eval}} - \mathcal{B}_{\mathrm{read,cal}})V_{\mathrm{c}}\| = s_1 - \epsilon_{\mathrm{c}},$$

which proves Eq. (59). The same argument with an arbitrary A proves Eq. (61).

Step 2: compare with the evaluation-optimal write. For every unit A, $\|\mathcal{B}_{\mathrm{read,eval}}A\| \leq \|\mathcal{B}_{\mathrm{read,cal}}A\| + \delta \leq s_1 + \delta$. Taking the supremum yields $\sigma_1(\mathcal{B}_{\mathrm{read,eval}}) \leq s_1 + \delta$. Subtracting the directional lower bound from Step 1 proves Eq. (60); nonnegativity follows from the variational definition of $\sigma_1(\mathcal{B}_{\mathrm{read,eval}})$.

Step 3: control the singular direction. Set $H_{\mathrm{c}} = \mathcal{B}_{\mathrm{read,cal}}^{*}\mathcal{B}_{\mathrm{read,cal}}$ and $H_{\mathrm{e}} = \mathcal{B}_{\mathrm{read,eval}}^{*}\mathcal{B}_{\mathrm{read,eval}}$. Since $\|\mathcal{B}_{\mathrm{read,eval}}\| \leq s_1 + \delta$,

$$\|H_{\mathrm{e}} - H_{\mathrm{c}}\|_{\mathrm{op}} \leq \delta(\|\mathcal{B}_{\mathrm{read,eval}}\| + \|\mathcal{B}_{\mathrm{read,cal}}\|) \leq \eta.$$

Eigenvalue perturbation therefore gives an evaluation eigengap at least $\Delta - 2\eta > 0$. On the orthogonal complement of the top evaluation eigenspace, every eigenvalue of H_{e} is at most $s_2^2 + \eta$. Because $H_{\mathrm{c}}V_{\mathrm{c}} = s_1^2 V_{\mathrm{c}}$,

$$\|(H_{\mathrm{e}} - s_1^2 I)V_{\mathrm{c}}\| \leq \eta.$$

Projecting this residual onto that complement and using the separation $s_1^2 - (s_2^2 + \eta) = \Delta - \eta$ proves Eq. (62).

The first step needs only the selected-mode shift; the latter two need global information to compare against unseen evaluation modes and show when the identity of the selected write is itself stable. □

The distinction is necessary. Take $\mathcal{B}_{\mathrm{read,cal}} = \mathrm{diag}(1, 0)$, $\mathcal{B}_{\mathrm{read,eval}} = \mathrm{diag}(1, M)$, and $V_{\mathrm{c}} = e_1$. The selected-mode shift is zero and its response is unchanged, while the evaluation-optimal response is M. No bound on evaluation optimality can follow from selected-mode stability alone as $M \to \infty$.

A.12 THE EXACT MEANING OF THE EMPIRICAL CONVEX-HULL AUDIT

For sample i, let $x_{ik} \in \mathbb{R}^D$ be the complete prediction of expert k (base plus residual), and define $\mathrm{conv}_i = \mathrm{conv}\{x_{i1}, \ldots, x_{iK}\}$. Let $\mathfrak{G}$ be any declared router class whose output lies in the simplex.

Proposition 14 (Same-split hull lower bound and conditional bridge)**.** *Assume $\ell(\cdot, y_i)$ is continuous. Define the pointwise empirical hull risk*

$$R_{\mathrm{hull}} = \frac{1}{n} \sum_i \min_{z \in \mathrm{conv}_i} \ell(z, y_i).$$

Then

$$R_{\mathrm{hull}} \le \inf_{g \in \mathfrak{G}} \frac{1}{n} \sum_i \ell\left(\sum_k g_k(x_i) x_{ik}, y_i \right). \tag{63}$$

If $z_i^\star \in \mathrm{conv}_i$ is a hull minimizer, the loss is L_ℓ-Lipschitz on the relevant hulls, and some router $g \in \mathfrak{G}$ satisfies

$$\frac{1}{n} \sum_i \left\| \sum_k g_k(x_i) x_{ik} - z_i^\star \right\|_2 \le \varepsilon,$$

then that router obeys

$$R(g) \le R_{\mathrm{hull}} + L_\ell \varepsilon. \tag{64}$$

Proof. The simplex is compact and each loss is continuous, so a pointwise minimizer exists. For every $g \in \mathfrak{G}$, its sample-i prediction belongs to conv_i and hence its loss is no smaller than the pointwise minimum. Averaging and then taking the infimum proves Eq. (63). Under the additional approximation and Lipschitz assumptions,

$$\ell\left(\sum_k g_k(x_i) x_{ik}, y_i \right) \le \ell(z_i^\star, y_i) + L_\ell \left\| \sum_k g_k(x_i) x_{ik} - z_i^\star \right\|_2.$$

Averaging proves Eq. (64). □

Proof of Corollary 5. Exact protection gives $f_i^{\mathrm{anchor}} = K^{-1} \sum_k f_{ik}(A_{\mathrm{write}}) \in \mathrm{conv}_i(A_{\mathrm{write}})$, so continuity and compactness imply $\mathcal{L}_{\mathrm{hull}}(A_{\mathrm{write}}) \le \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0)$ and hence $\Delta_{\mathrm{opp}}(A_{\mathrm{write}}) \ge 0$. Every declared simplex router also predicts a point of $\mathrm{conv}_i(A_{\mathrm{write}})$ samplewise; therefore Prop. 14 gives

$$\mathcal{L}_{\mathrm{hull}}(A_{\mathrm{write}}) \le \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, \theta_g) \quad \text{for every declared router } \mathbf{g}(\theta_g).$$

If $\Delta_{\mathrm{opp}}(A_{\mathrm{write}}) = 0$, then $\theta_g = 0$ attains this lower bound and is a global minimizer within the declared router class. Differentiability at the anchor therefore gives $r_{\mathrm{route}}(A_{\mathrm{write}}, 0) = \nabla_{\theta_g} \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0) = 0$.

Conversely, if $r_{\mathrm{route}}(A_{\mathrm{write}}, 0) \ne 0$, the anchor cannot be a differentiable local minimum. More constructively, whenever an Armijo trial is accepted, Thm. 10 supplies a feasible router parameter θ_g^+ with $\mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, \theta_g^+) < \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, 0)$. Since $\mathcal{L}_{\mathrm{hull}}(A_{\mathrm{write}}) \le \mathcal{L}_{\mathrm{cal}}(A_{\mathrm{write}}, \theta_g^+)$, it follows that $\Delta_{\mathrm{opp}}(A_{\mathrm{write}}) > 0$, and the accepted decrease is an actually realized portion of this empirical opportunity. □

The hull must be formed in the space where the actual router mixes expert outputs. If that space is logits, use cross-entropy on the mixed logits; mixing probabilities would define a different model family. If a fixed nonlinear head H_i follows a mixed hidden state, form the hull of those states and evaluate $\ell_i \circ H_i$ on it. Taking a hull of separately transformed expert logits would generally describe a different model. For forecasting, apply the declared task risk to the routed forecasts. The lower bound allows sample-specific gates and is consequently an opportunity scale, not a realizability certificate. Equation (64) shows the missing condition explicitly. Neither inequality is a population statement; generalization additionally requires control of the split change and router estimation error.

A.13 ARBITRARY TRAINED ANCHORS AND FROZEN DISPATCH

The symmetric chart exposes the spectral core, but the anchor of a trained MoE need not have zero router gradient. We retain the main-text symbols and row orientation. For sampling unit i, let $\mathbf{F}_i^{\mathrm{anchor}} \in \mathbb{R}^{K\times D}$ contain the frozen complete expert outputs and let $\mathbf{g}_i(\theta_g)$ be a differentiable reader with $\mathbf{g}_i(0) = \mathbf{g}_i^{\mathrm{anchor}} \neq 0$. Any linear raw residual map $Z_i : \mathcal{A} \to \mathbb{R}^{K\times D}$ induces

$$\begin{aligned} P_{i,\mathrm{safe}} &= I_K - \frac{\mathbf{g}_i^{\mathrm{anchor}}(\mathbf{g}_i^{\mathrm{anchor}})^\top}{\|\mathbf{g}_i^{\mathrm{anchor}}\|^2}, \\ \Delta\mathbf{F}_i(A_{\mathrm{write}}) &= P_{i,\mathrm{safe}} Z_i(A_{\mathrm{write}}), \qquad \Delta\mathbf{F}_i(A_{\mathrm{write}})^\top \mathbf{g}_i^{\mathrm{anchor}} = 0. \end{aligned} \tag{65}$$

$P_{i,\mathrm{safe}}$ is the orthogonal projector onto $\ker(\mathbf{g}_i^{\mathrm{anchor}})^\top$; consequently $\Delta\mathbf{F}_i(A_{\mathrm{write}})$ is the closest Frobenius-norm protected residual to $Z_i(A_{\mathrm{write}})$. This construction is pointwise: on a new input, use its frozen anchor gate and support in the same projector formula. No assumption that its feature distribution matches calibration is needed for the write-time equality; the calibrated norm and response guarantees remain separate. Define

$$\begin{aligned} f_{\mathrm{mix},i}(A_{\mathrm{write}}, \theta_g) &= (\mathbf{F}_i^{\mathrm{anchor}} + \Delta\mathbf{F}_i(A_{\mathrm{write}}))^\top \mathbf{g}_i(\theta_g), \\ \mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, \theta_g) &= \mathbb{E}\, \ell_{\mathrm{task}}(f_{\mathrm{mix},i}(A_{\mathrm{write}}, \theta_g), y_i). \end{aligned} \tag{66}$$

The mixed state may precede a fixed nonlinear downstream map H_i: set $\ell_{\mathrm{task}}(s, y_i) = \ell_i(H_i(s), y_i)$. Exact state protection then preserves the final prediction $H_i(s)$, while $p_{\mathrm{task},i}$ below is the derivative of this composite task loss at the frozen state. No linearity of H_i is needed. The expectation can be replaced by a finite empirical average. We assume that differentiation and expectation commute at the anchor; this is automatic for the finite audit and follows, for example, from a local integrable domination condition in the population form.

Let $J_i^g = D_{\theta_g}\mathbf{g}_i(0)$ and $p_{\mathrm{task},i} = \nabla_f \ell_{\mathrm{task}}((\mathbf{F}_i^{\mathrm{anchor}})^\top \mathbf{g}_i^{\mathrm{anchor}}, y_i)$. Equip write and reader coordinates with positive semidefinite matrices M_{write} and M_{route}. We require that every $v \in \ker M_{\mathrm{write}}$ have zero readable action, $T^\top C v = 0$ in the retained reader chart; otherwise the trust problem can have a zero-cost improving direction. The task derivatives below are ordinary Euclidean covectors. Their Riesz representatives on the identifiable reader quotient are $M_{\mathrm{route}}^\dagger c$. Direct differentiation gives

$$\begin{aligned} c_0 &= D_{\theta_g}\mathcal{L}_{\mathrm{task}}(0,0) = \mathbb{E}\,(J_i^g)^\top \mathbf{F}_i^{\mathrm{anchor}} p_{\mathrm{task},i}, \\ CA_{\mathrm{write}} &= D_{\theta_g}\mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0) - c_0 = \mathbb{E}\,(J_i^g)^\top \Delta\mathbf{F}_i(A_{\mathrm{write}}) p_{\mathrm{task},i}, \end{aligned} \tag{67}$$

where $(J_i^g)^\top$ is the ordinary transpose. The dagger symbol denotes the Moore–Penrose pseudoinverse.

Proof of Theorem 8, affine identity. Equation (65) gives $f_{\mathrm{mix},i}(A_{\mathrm{write}}, 0) = (\mathbf{F}_i^{\mathrm{anchor}})^\top \mathbf{g}_i^{\mathrm{anchor}}$ for every finite A_{write}. Hence $p_{\mathrm{task},i}$ is fixed by the write. For any reader tangent U, the chain rule yields

$$D_{\theta_g}\mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0)[U] = \mathbb{E}\,\langle p_{\mathrm{task},i}, (\mathbf{F}_i^{\mathrm{anchor}} + \Delta\mathbf{F}_i(A_{\mathrm{write}}))^\top J_i^g U\rangle \tag{68}$$

$$= \langle c_0 + CA_{\mathrm{write}}, U\rangle. \tag{69}$$

Thus $D_{\theta_g}\mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0) = c_0 + CA_{\mathrm{write}}$. On the declared retained reader subspace $\mathrm{range}(T)$, the Riesz representative of the restricted covector is $TT^\top(c_0 + CA_{\mathrm{write}})$; in the original spectral chart this equals $M_{\mathrm{route}}^\dagger(c_0 + CA_{\mathrm{write}})$. A covector with action in excluded directions is restricted to this reader, rather than silently identified with a covector on the full quotient. The affine equality is exact in A_{write}: no Taylor remainder appears because the chart is linear and protection keeps $p_{\mathrm{task},i}$ fixed. □

Let $U_{\mathrm{write}}\Lambda_{\mathrm{write}}U_{\mathrm{write}}^\top$ and $U_{\mathrm{route}}\Lambda_{\mathrm{route}}U_{\mathrm{route}}^\top$ be the retained positive spectra of M_{write} and M_{route}. A horizontal quotient representative is $A_{\mathrm{write}} = P\bar{A}$, where $P = U_{\mathrm{write}}\Lambda_{\mathrm{write}}^{-1/2}$. Likewise set $T = U_{\mathrm{route}}\Lambda_{\mathrm{route}}^{-1/2}$, so $T^\top M_{\mathrm{route}} T = I$ and $TT^\top = M_{\mathrm{route}}^\dagger$. Set

$$\bar{r} = T^\top c_0, \qquad \bar{B} = T^\top CP.$$

The null-action condition makes this independent of the write representative. The reader-coordinate task gradient is $\nabla_u \mathcal{L}_{\text{task}}(P\bar{A}, Tu)|_{u=0} = \bar{r} + \bar{B}\bar{A}$. The nonconstant objective is

$$2\langle \bar{B}^\top \bar{r}, \bar{A}\rangle + \langle \bar{A}, \bar{B}^\top \bar{B}\bar{A}\rangle, \qquad \|\bar{A}\|_2 \le \rho. \tag{70}$$

This expansion also gives a useful robustness interpretation. The first-order direction sees only the component of $\bar{B}\bar{A}$ aligned with $\bar{r}$; response orthogonal to $\bar{r}$ is invisible to its selection rule. The quadratic term credits the full response norm, so a direction whose only advantage is calibration-specific linear alignment loses relative priority to a direction with a larger readable response. Moreover, if $\bar{B}_{\text{eval}} = \bar{B}_{\text{cal}} + E$ and $\|E\|_{\text{op}} \le \delta$, then every $\|\bar{A}\| \le \rho$ obeys

$$\|\bar{B}_{\text{eval}}\bar{A}\| \ge \|\bar{B}_{\text{cal}}\bar{A}\| - \rho\delta.$$

Thus calibration response amplitude is a reserve against bounded operator shift, analogous to insurance that matters most when the locally aligned mode weakens. It does not by itself order held-out cosine or finite-read NLL; those remain independent empirical questions.

Compactness gives a maximizer. The global trust-region optimality system is

$$(\lambda I - \bar{B}^\top \bar{B})\bar{A} = \bar{B}^\top \bar{r}, \quad \lambda \ge \lambda_{\max}(\bar{B}^\top \bar{B}), \quad \|\bar{A}\|_2 = \rho, \tag{71}$$

with the standard hard-case completion in the top eigenspace when the linear term is orthogonal to it. Sufficiency follows because, for any feasible X and a solution $\bar{A}$ of Eq. (71), expansion and the positive semidefiniteness of $\lambda I - \bar{B}^\top \bar{B}$ bound the objective at X by its value at $\bar{A}$. Necessity follows from the Lagrange conditions together with the second-order condition on the sphere. This proves the trust-region characterization. Because it maximizes the declared objective, every other equal-budget writer has no larger incremental gradient power.

When $\bar{r} = 0$, Eq. (70) is the Rayleigh quotient of $\bar{B}^\top \bar{B}$. Its complete nonzero maximizing set is the radius-ρ sphere in the top right singular subspace of $\bar{B}$, which recovers Theorem 6. Thus the affine and spectral writers are one optimization problem, not competing procedures.

Coordinate covariance does not require the metric to be a generalized Gauss–Newton matrix; it requires a frozen metric with its null action quotiented. To see this, let $A_{\text{write}} = S\widetilde{A}_{\text{write}}$ and $\theta_g = Q\widetilde{\theta}_g$ with S, Q invertible, and pull back the metrics:

$$\widetilde{M}_{\text{write}} = S^\top M_{\text{write}} S, \qquad \widetilde{M}_{\text{route}} = Q^\top M_{\text{route}} Q. \tag{72}$$

The ordinary covector and mixed map transform as $\widetilde{c}_0 = Q^\top c_0$ and $\widetilde{C} = Q^\top C S$; the quotient Riesz representative transforms as a tangent vector. Using the pulled-back charts $\widetilde{T} = Q^{-1}T$ and $\widetilde{P} = S^{-1}P$ gives the same $\bar{r}$ and $\bar{B}$. Therefore

$$\widetilde{A}_{\text{write}}^\top \widetilde{M}_{\text{write}} \widetilde{A}_{\text{write}} = A_{\text{write}}^\top M_{\text{write}} A_{\text{write}},$$
$$\widetilde{T}^\top(\widetilde{c}_0 + \widetilde{C}\widetilde{A}_{\text{write}}) = T^\top(c_0 + C A_{\text{write}}).$$

The two trust problems are identical geometric problems in different coordinates when the retained charts are transported. Reapplying a numerical eigenvalue cutoff after a nonorthogonal reparameterization can select a different subspace and is a different audit. The maximizers satisfy $A^\star_{\text{write}} = S\widetilde{A}^\star_{\text{write}}$. This proves the metric-covariance claim. The CIFAR experiment declares empirical function-isometric coordinates; the LLM adapter declares the physical Frobenius write metric and a task-loss-induced reader quotient. Response values are canonical within, but not numerically comparable across, different budgets.

A.13.1 Task-loss-induced reader quotient

Raw reader coordinates need not reflect task sensitivity. For binary CE, let $\delta_i(A_{\text{write}}, \theta_g)$ be the audited two-logit margin. At the unwritten anchor set $j_i^{\text{route}} = D_{\theta_g}\delta_i(0,0)$, $\pi_i^{\text{anchor}} = \text{sigmoid}(\delta_i(0,0))$, and

$$M_{\text{GGN}} = \mathbb{E}[\pi_i^{\text{anchor}}(1-\pi_i^{\text{anchor}}) j_i^{\text{route}} (j_i^{\text{route}})^\top]. \tag{73}$$

This matrix is positive semidefinite because every weighted outer product is positive semidefinite; it is fitted once and frozen. A protected write may change $D_{\theta_g}\delta_i(A_{\text{write}}, 0)$, so its post-write GGN need not equal M_{GGN}.

Let $N_{\mathrm{route}} = \ker M_{\mathrm{GGN}}$. Finite logits make the scalar weight above positive; hence

$$v^\top M_{\mathrm{GGN}} v = 0 \quad \Longleftrightarrow \quad (j_i^{\mathrm{route}})^\top v = 0 \quad \text{almost surely.}$$

Thus N_{route} is exactly the anchor-margin-invisible reader subspace. Represent the quotient by $\operatorname{range} M_{\mathrm{GGN}}$ and freeze a full-column-rank map T_{route} satisfying $T_{\mathrm{route}}^\top M_{\mathrm{GGN}} T_{\mathrm{route}} = I$. In Euclidean physical coordinates, $M_{\mathrm{GGN}} = U\Lambda U^\top$ gives $T_{\mathrm{route}} = U\Lambda^{-1/2}$ on the positive spectrum and $T_{\mathrm{route}} T_{\mathrm{route}}^\top = M_{\mathrm{GGN}}^\dagger$.

Proof of the metric-covariance clause of Theorem 8. Write $\theta_g = T_{\mathrm{route}} u$. The chain rule and the exact affine identity give

$$\nabla_u \mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, T_{\mathrm{route}} u)|_{u=0} = T_{\mathrm{route}}^\top D_{\theta_g} \mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0) = T_{\mathrm{route}}^\top (c_0 + C A_{\mathrm{write}}).$$

Pushing the Euclidean SGD displacement in u through T_{route} yields $-\eta T_{\mathrm{route}} T_{\mathrm{route}}^\top (c_0 + C A_{\mathrm{write}})$.

The fixed anchor-reference quadratic is

$$Q^{(0)}_{A_{\mathrm{write}}}(u) = \mathcal{L}_0 + [T_{\mathrm{route}}^\top (c_0 + C A_{\mathrm{write}})]^\top u + \tfrac{1}{2}\|u\|^2.$$

Completing the square gives $\min_u Q^{(0)}_{A_{\mathrm{write}}}(u) = \mathcal{L}_0 - \frac{1}{2}\|T_{\mathrm{route}}^\top (c_0 + C A_{\mathrm{write}})\|^2$. Protection makes $\mathcal{L}_0$ common to all writes, so maximizing predicted improvement over $A_{\mathrm{write}} = 0$ is Eq. (20), the existing trust problem with $\bar{B} = T_{\mathrm{route}}^\top C P$ and $\bar{r} = T_{\mathrm{route}}^\top c_0$.

This proves optimality only for the fixed reference quadratic. For the actual nonlinear risk, the directional derivative along the pushed ordinary gradient is $-\|T_{\mathrm{route}}^\top (c_0 + C A_{\mathrm{write}})\|^2$; an accepted Armijo inequality therefore certifies strict finite-risk decrease without assuming a fixed post-write GGN. □

Empirically we retain eigenvalues $\hat{\lambda}_j > \hat{\lambda}_1/\sqrt{n}$ and quotient the remaining directions out rather than adding a ridge. This is a declared resolution rule, not a population-rank claim; evaluation never refits the calibration subspace.

Claim 6 (pathwise comparison of ordinary reads). Let $h(A_{\mathrm{write}}) = T^\top (c_0 + C A_{\mathrm{write}})$ and $h_0 = T^\top c_0$. For $j \in \{A_{\mathrm{write}}, 0\}$, let h_j denote its anchor gradient, $f_j(t) = \mathcal{L}_{\mathrm{task}}(j, -tTh_j)$, and let $\eta_j \geq 0$ be its possibly distinct read step. Suppose only that f_j has a derivative at zero and is absolutely continuous on $[0, \eta_j]$. Define the pathwise remainder $R_j(\eta_j) := \int_0^{\eta_j} [f_j'(t) - f_j'(0)]\, dt$. Exact protection and $f_j'(0) = -\|h_j\|^2$ give the identity

$$\begin{aligned} &\mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, -\eta_{A_{\mathrm{write}}} T h(A_{\mathrm{write}})) - \mathcal{L}_{\mathrm{task}}(0, -\eta_0 T h_0) \\ &\quad = -\eta_{A_{\mathrm{write}}} \|h(A_{\mathrm{write}})\|^2 + \eta_0 \|h_0\|^2 + R_{A_{\mathrm{write}}}(\eta_{A_{\mathrm{write}}}) - R_0(\eta_0). \end{aligned} \tag{74}$$

Proof. Apply the fundamental theorem of calculus to each path and subtract; the common starting risk cancels by exact protection. Thus the written read is better whenever its step-scaled gradient-power advantage exceeds $R_{A_{\mathrm{write}}}(\eta_{A_{\mathrm{write}}}) - R_0(\eta_0)$. This exact criterion also applies to separately accepted Armijo steps. As a prospective sufficient condition, if each path obeys $|f_j'(t) - f_j'(0)| \leq L_j t \|h_j\|^2$ on its tested segment, then $|R_j(\eta_j)| \leq \frac{1}{2} L_j \eta_j^2 \|h_j\|^2$ and it suffices that

$$\eta_{A_{\mathrm{write}}} \|h(A_{\mathrm{write}})\|^2 - \eta_0 \|h_0\|^2 > \tfrac{1}{2} L_{A_{\mathrm{write}}} \eta_{A_{\mathrm{write}}}^2 \|h(A_{\mathrm{write}})\|^2 + \tfrac{1}{2} L_0 \eta_0^2 \|h_0\|^2.$$

Neither a common step nor a common global smoothness constant is needed.

Theorem 15 (Independent selected-write persistence audit)**.** *Let the selected write $\widehat{A}_{\mathrm{write}}$ be measurable with respect to calibration data alone. On independent audit sampling units, let $v_i(\widehat{A}_{\mathrm{write}})$ and $v_i(0)$ be ordinary reader gradients with finite first moments. Define*

$$\widehat{\mathfrak{C}}_U = \frac{1}{n(n-1)} \sum_{i \neq j} \left[\langle v_i(\widehat{A}_{\mathrm{write}}), v_j(\widehat{A}_{\mathrm{write}}) \rangle - \langle v_i(0), v_j(0) \rangle \right]. \tag{75}$$

Then, conditionally on calibration,

$$\mathbb{E}[\widehat{\mathfrak{C}}_U \mid \mathcal{D}_{\mathrm{cal}}] = \|\mathbb{E} v(\widehat{A}_{\mathrm{write}})\|^2 - \|\mathbb{E} v(0)\|^2. \tag{76}$$

Proof. Condition on the calibration sigma-field, which fixes $\widehat{A}_{\mathrm{write}}$. For every $i \neq j$, audit-unit independence and integrability give $\mathbb{E}\langle v_i, v_j\rangle = \langle \mathbb{E}v_i, \mathbb{E}v_j\rangle$. Identical sampling distributions make this the squared norm of the common mean. Averaging the ordered off-diagonal pairs proves Eq. (76) for the written and anchor fields and hence for their difference. Selection causes no bias because it occurred before, and independently of, these audit units. □

Diagonal terms are excluded because they add per-unit gradient variance. The experiment's jackknife/bootstrap intervals are asymptotic evidence; no distribution-free finite-sample coverage is claimed without range or tail assumptions.

We next formalize the sparse-routing obstruction. For unit i, let S_i be the anchor top-s set and let

$$\tau_i = \min_{\substack{k\in S_i,\, l\notin S_i \\ \dot{d}_{il}-\dot{d}_{ik}>0}} \frac{\ell^0_{ik} - \ell^0_{il}}{\dot{d}_{il} - \dot{d}_{ik}}, \tag{77}$$

with $\tau_i = +\infty$ if the index set is empty. Along the affine score ray, the complete support is preserved only before $\min_i \tau_i$. If the finite τ_i are independent and $\Pr(\tau_i \le \varepsilon) > 0$ for every $\varepsilon > 0$, then

$$\Pr\Big(\min_{i\le n} \tau_i > \varepsilon\Big) = [1 - \Pr(\tau \le \varepsilon)]^n \longrightarrow 0. \tag{78}$$

Applying this convergence to every positive rational ε proves $\min_{i\le n}\tau_i \to 0$ almost surely. The proposition concerns an all-sample no-switch certificate, not the quality of native routing.

Finally freeze $S_i = \mathrm{TopS}(\ell^0_i)$ and the strictly positive deployed selected weights $\bar{g}_{iS}$ once. Native sparse layers need not renormalize the retained top-s weights, so put $\kappa_i = \sum_{k\in S_i}\bar{g}_{ik} > 0$ and $\bar{g}^{\mathrm{norm}}_{iS} = \bar{g}_{iS}/\kappa_i \in \operatorname{int}\Delta^{s-1}$. Let an affine reader produce score increments $d_i(\theta_g)$ only on S_i and define

$$g^{\mathrm{norm}}_{iS}(\theta_g) = \mathrm{softmax}\big(\log \bar{g}^{\mathrm{norm}}_{iS} + d_i(\theta_g)\big), \qquad g_{iS}(\theta_g) = \kappa_i g^{\mathrm{norm}}_{iS}(\theta_g). \tag{79}$$

Thus $g^{\mathrm{norm}}_{iS}(\theta_g)$ always lies in the simplex interior, $g_{iS}(0) = \bar{g}_{iS}$, and the native selected mass κ_i is fixed. If f^{skip}_i is the frozen residual contribution, define the complete supported expert states $f_{ik}(A_{\mathrm{write}}) = f^{\mathrm{skip}}_i + \kappa_i\big(f^{\mathrm{anchor}}_{ik} + \Delta f_{ik}(A_{\mathrm{write}})\big)$. The layer output has the equivalent convex-mixture form

$$f_{\mathrm{mix},i}(A_{\mathrm{write}}, \theta_g) = f^{\mathrm{skip}}_i + \sum_{k\in S_i} g_{ik}(\theta_g)\big(f^{\mathrm{anchor}}_{ik} + \Delta f_{ik}(A_{\mathrm{write}})\big) = \sum_{k\in S_i} g^{\mathrm{norm}}_{ik}(\theta_g) f_{ik}(A_{\mathrm{write}}). \tag{80}$$

Restrict Eq. (16) to S_i: with $P_{i,S} = I_s - \bar{g}_{iS}\bar{g}_{iS}^\top/\|\bar{g}_{iS}\|_2^2$, set $\Delta\mathbf{F}_{iS} = P_{i,S}Z_{iS}$. Then $\sum_{k\in S_i} g_{ik}(0)\Delta f_{ik}(A_{\mathrm{write}}) = 0$ by construction. When $s = 1$, $P_{i,S} = 0$, which exposes the zero-freedom boundary rather than treating protection as an external assumption.

Proof of the fixed-dispatch clause of Theorem 8. At $\theta_g = 0$, the last displayed identity cancels the written residual exactly, which proves finite prediction protection. The expert set is the stored S_i by construction; the adapter never calls the top-s operator again. Thus dispatch indices, expert evaluations, and their communication pattern are identical to those of the trained model for all θ_g.

Substitution in Eq. (79) proves $g^{\mathrm{norm}}_{iS}(0) = \bar{g}^{\mathrm{norm}}_{iS}$ and $g_{iS}(0) = \bar{g}_{iS}$. Every finite reader step has $g^{\mathrm{norm}}_{ik}(\theta_g) > 0$ and $\sum_{k\in S_i} g^{\mathrm{norm}}_{ik}(\theta_g) = 1$; equivalently, $g_{ik}(\theta_g) > 0$ and $\sum_{k\in S_i} g_{ik}(\theta_g) = \kappa_i$. Hence the reader cannot create a negative expert weight or change the native selected mass. In floating-point code we evaluate this positive softmax map directly and audit $\|g_{iS}(0) - \bar{g}_{iS}\|$, positivity, and mass residual on every accepted trial; the real-arithmetic identity is exact, while the measured anchor mismatch is at rounding scale.

Equation (79) is a composition of affine maps, softmax, and fixed coordinate selection, hence is smooth without a margin condition. Protection again fixes the anchor task covector. Differentiating Eq. (80) at $\theta_g = 0$ separates the frozen-expert term from the term linear in $\Delta f_{ik}(A_{\mathrm{write}})$, giving $D_{\theta_g}\mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0) = c_0 + C_{\mathrm{sparse}}A_{\mathrm{write}}$. The metric trust-region argument and pathwise comparison Eq. (74) then apply verbatim with $\bar{r} = T^\top c_0$ and $\bar{B} = T^\top C_{\mathrm{sparse}}P$. Only experts in S_i occur in Eq. (80); contrast confined to never-dispatched experts is therefore outside this reader's reachable family. □

The preceding results complete the load-bearing chain from protection and initial readability through independent-split certification and the trained sparse-MoE realization. The next five subsections collect optional structural extensions; none is required to execute Algorithms 1–3.

A.14 Routing-orbit duality

This extension separates task-agnostic routing opportunity from the task-specific mixed derivative in Theorem 4. It is not in the main chain because a nontrivial routing orbit need not align with the current task loss.

Theorem 16 (Routing changes predictions exactly outside the expert kernel)**.** *Let $\mathcal{G}$ be a nonempty subset of the simplex and $\mathcal{O}_{\mathbf{F}} = \{\mathbf{F}^\top \mathbf{g} : \mathbf{g} \in \mathcal{G}\} \subset \mathbb{R}^D$. Define*

$$\operatorname{diam}(\mathcal{O}_{\mathbf{F}}) = \sup_{a,b\in\mathcal{G}} \|\mathbf{F}^\top(a-b)\|_2.$$

Then

$$\operatorname{diam}(\mathcal{O}_{\mathbf{F}}) = 0 \iff \operatorname{span}(\mathcal{G}-\mathcal{G}) \subseteq \ker \mathbf{F}^\top, \tag{81}$$

$$\operatorname{diam}(\mathcal{O}_{\mathbf{F}}) = \sup_{\operatorname{Lip}(s)\le 1} \left[\sup_{z\in\mathcal{O}_{\mathbf{F}}} s(z) - \inf_{z\in\mathcal{O}_{\mathbf{F}}} s(z)\right], \tag{82}$$

where the second supremum ranges over real-valued 1-Lipschitz functions on $\mathbb{R}^D$.

Proof. The diameter vanishes exactly when $\mathbf{F}^\top(a-b) = 0$ for every $a, b \in \mathcal{G}$. Since a kernel is a linear subspace, this is equivalent to Eq. (81).

For any 1-Lipschitz s and $z, z' \in \mathcal{O}_X$, $s(z)-s(z') \le \|z-z'\|_2$, so the right-hand side of Eq. (82) is at most the diameter. For the reverse inequality, fix $\varepsilon > 0$ and choose $z_0, z_1 \in \mathcal{O}_X$ with $\|z_1 - z_0\|_2 \ge \operatorname{diam}(\mathcal{O}_X) - \varepsilon$. The function $s_{z_0}(z) = \|z - z_0\|_2$ is 1-Lipschitz and has range on $\mathcal{O}_X$ at least $s_{z_0}(z_1) - s_{z_0}(z_0) = \|z_1 - z_0\|_2$. Letting $\varepsilon \downarrow 0$ proves equality, including when the diameter is not attained. □

A nonzero diameter therefore means that some Lipschitz score can distinguish two routings. It does not imply that the current task covectors align with that score; the mixed operator is precisely the task-specific refinement.

A.15 A symmetry no-go theorem

This theorem explains why an exactly symmetric protected write needs an explicit symmetry convention. It complements Proposition 1 but is not required once the implemented chart fixes a contrast basis.

Let the permutation group S_K act on expert matrices by permuting rows. A state s_0 is fully symmetric if $\pi \cdot s_0 = s_0$ for every $\pi \in S_K$. A deterministic writer $\mathcal{W}$ is permutation equivariant when $\mathcal{W}(\pi \cdot s) = P_\pi \mathcal{W}(s)$, where P_π is the permutation matrix.

Theorem 17 (A deterministic equivariant protected writer is zero at exact symmetry)**.** *If s_0 is fully symmetric, $\mathcal{W}$ is deterministic and permutation equivariant, and $H = \mathcal{W}(s_0)$ satisfies $H^\top \mathbf{1}_K/K = 0$, then $H = 0$.*

Proof. Equivariance and invariance of the input imply $H = P_\pi H$ for every permutation π. In particular, transpositions show that every pair of rows of H is equal. Hence $H = \mathbf{1}_K u^\top$ for some $u \in \mathbb{R}^D$. Protection at the uniform gate gives $0 = H^\top \mathbf{1}_K/K = u$, so $H = 0$. □

Thus a deterministic nonzero write at an exactly symmetric state must obtain a symmetry choice from somewhere: asymmetric data or parameters, randomness, or a fixed contrast coordinate/tie-break. Such a coordinate choice labels directions in the abstract contrast space; it does not provide task, corruption, or expert-semantic labels.

A.16 General parameter-manifold mixed criterion

The chart theorem is exact because the residual heads are linear and frozen from the base. The intrinsic first-order statement is more general: it is the general-manifold analogue of Thm. 4, and the chart of Sec. 4 is its exactly realizable instance, with $\Phi = \mathcal{A}$, $\Theta = \mathcal{W}$, and F the chart fusion (Eq. (6)).

Theorem 18 (Protected-chart readability and its constant-rank case)**.** *Let Φ and Θ be finite-dimensional Hilbert parameter spaces and Y a finite-dimensional Hilbert output space. Fix (ϕ_0, θ_0) and let U be a declared write Hilbert space with a C^1 local protected chart $R : \mathsf{U} \supset \mathcal{O} \to \Phi$ satisfying $R(0) = \phi_0$, $J := DR(0)$ injective, and $F(R(a), \theta_0) = F(\phi_0, \theta_0)$ for every $a \in \mathcal{O}$. Assume F is differentiable in ϕ at the anchor and in θ at $(R(a), \theta_0)$ for every nearby a. Let $\widehat{\mathcal{L}_{\mathrm{task}}}(\phi, \theta) = \ell(F(\phi, \theta))$, where ℓ is differentiable at the protected output $y_0 = F(\phi_0, \theta_0)$. Let $\mathsf{V} \subseteq \Theta$ be the reachable reader tangent space, and assume the restricted reader Jacobian adjoint $K(a) := P_{\mathsf{V}} D_\theta F(R(a), \theta_0)^*$ is differentiable at $a = 0$. With $c = \nabla \ell(y_0)$, the output-space Riesz representative of the task covector, define the bilinear form*

$$m(u, v) = \langle DK(0)[u]c, v \rangle, \qquad (u, v) \in \mathsf{U} \times \mathsf{V}.$$

There is a unique restricted mixed operator $B : \mathsf{U} \to \mathsf{V}$ satisfying $\langle Bu, v \rangle = m(u, v)$. The following are equivalent:

$$B \neq 0 \iff \sigma_1(B) > 0 \iff \exists u \in \mathsf{U} : D_a \Big(P_{\mathsf{V}} \nabla_\theta \widehat{\mathcal{L}_{\mathrm{task}}}(R(a), \theta_0) \Big)_{a=0} [u] \neq 0.$$

If additionally $P_{\mathsf{V}} \nabla_\theta \widehat{\mathcal{L}_{\mathrm{task}}}(\phi_0, \theta_0) = 0$, such a direction creates a nonzero reachable router gradient to first order. Under declared Hilbert metrics, a top unit right singular tangent u_1 maximizes this first-order response. The C^1 chart already gives

$$P_{\mathsf{V}} \nabla_\theta \widehat{\mathcal{L}_{\mathrm{task}}}\big(R(\rho u_1), \theta_0\big) = r_{\mathrm{route}}^{\mathrm{anchor}} + \rho B u_1 + o(\rho), \tag{83}$$

where $r_{\mathrm{route}}^{\mathrm{anchor}} = P_{\mathsf{V}} \nabla_\theta \widehat{\mathcal{L}_{\mathrm{task}}}(\phi_0, \theta_0)$. If F and ℓ are C^2 near the anchor, $m(u, v) = D^2_{\phi\theta} \widehat{\mathcal{L}_{\mathrm{task}}}[Ju, v]$. If $F(\cdot, \theta_0)$ is C^1 near ϕ_0, locally constant rank of $D_\phi F$ is sufficient, but not necessary, to obtain a chart spanning the full prediction kernel; the reader-Jacobian condition remains a separate requirement.

Proof. Differentiating the chart's exact protection gives $D_\phi F(\phi_0, \theta_0) Ju = 0$ for every $u \in \mathsf{U}$. More importantly, exact protection fixes y_0 and hence the task covector c for every write in the chart. The ordinary reader chain rule therefore gives the exact fiber identity

$$P_{\mathsf{V}} \nabla_\theta \widehat{\mathcal{L}_{\mathrm{task}}}(R(a), \theta_0) = K(a)c.$$

Differentiating this identity at $a = 0$ gives $DK(0)[u]c$, represented by Bu on V. This proves the three equivalences and, by the singular-value variational formula, tangent-unit-ball optimality. Differentiability of K also yields $K(\rho u_1)c = K(0)c + \rho B u_1 + o(\rho)$, proving the claimed expansion.

For the smooth special case, the mixed chain rule for a protected tangent Ju and a reachable reader tangent v reads

$$\begin{aligned} D^2_{\phi\theta} \widehat{\mathcal{L}_{\mathrm{task}}}[Ju, v] &= D^2 \ell(F)[D_\phi F[Ju], D_\theta F[v]] \\ &\quad + \langle D\ell(F), D^2_{\phi\theta} F[Ju, v] \rangle \\ &= \langle D\ell(F), D^2_{\phi\theta} F[Ju, v] \rangle. \end{aligned}$$

The first term vanishes by protection. The chain rule for K identifies the remaining term with $\langle DK(0)[u]c, v \rangle = m(u, v)$. Thus the familiar mixed-Hessian formula is a consequence of stronger smoothness, not a prerequisite for the readability criterion. Under locally constant rank, the constant-rank level-set theorem supplies the stated chart; its existence alone does not assert differentiability of K. □

The weaker reader-Jacobian condition is strict. Take scalar $F(a, \theta) = \theta(a + |a|^{3/2})$, $R(a) = a$, $\theta_0 = 0$, and $\ell(f) = \frac{1}{2}(f - 1)^2$. Every finite a preserves $F(a, 0) = 0$, and $K(a) = a + |a|^{3/2}$

is differentiable at zero. The ordinary router gradient is $-(a + |a|^{3/2})$, so $B = -1$ is readable, although F is not C^2 on any neighborhood of the anchor.

The chart keeps prediction exactly on the level set, while its readable response is first-order spectral with an $o(\rho)$ remainder. The linear DSA chart is stronger: it supplies the global fiber curve $\gamma(t) = tA$ and an exactly affine router gradient, so the remainder vanishes. Shared encoders and nonlinear heads require a model-specific protected chart before this appendix theorem becomes an implementation claim. For example, $F(x, y, \theta) = x^2 + \theta y$ at the origin has nonconstant expert-Jacobian rank, yet $R(a) = (0, a)$ is an exact protected chart; the squared risk $\frac{1}{2}(F - 1)^2$ has a nonzero protected expert–router mixed derivative along y. Thus constant rank is not required for the mechanism itself. An exact chart cannot be replaced by a tangent-kernel test when finite protection is claimed: $F(x) = x^2$ at $x = 0$ has $DF(0) = 0$, so every tangent is first-order safe, but no nonzero nearby write preserves $F(0)$.

A.17 Composability with an arbitrary engineering update

The next bound states exactly what a tangent protected correction can and cannot preserve when composed with another parameter displacement.

Proposition 19 (First-order composability and curvature interaction)**.** *Let $F : \mathbb{R}^p \to \mathbb{R}^q$ be continuously differentiable and suppose the relevant segments lie in a compact convex neighborhood of θ. Define the local Jacobian modulus $\omega(t) := \sup_{\|z-\theta\| \le t} \|J_F(z) - J_F(\theta)\|_{\mathrm{op}}$ within that neighborhood. If $J_F(\theta)q = 0$, then for every u,*

$$\frac{d}{dt} F(\theta + t(u + q))\Big|_{t=0} = J_F(\theta)u. \tag{84}$$

For arbitrary real α, β whose connecting segments remain in the neighborhood,

$$\|F(\theta + \alpha u + \beta q) - F(\theta + \alpha u)\| \le |\beta|\|q\| \int_0^1 \omega(|\alpha|\|u\| + t|\beta|\|q\|)\, dt. \tag{85}$$

Proof. Equation (84) follows from linearity of the Fréchet derivative: $J_F(\theta)(u + q) = J_F(\theta)u$. For the finite bound, integrate along the q segment and subtract the zero vector $\beta J_F(\theta)q$:

$$\begin{aligned} &F(\theta + \alpha u + \beta q) - F(\theta + \alpha u) \\ &\quad = \beta \int_0^1 \big[J_F(\theta + \alpha u + t\beta q) - J_F(\theta)\big] q\, dt. \end{aligned}$$

Taking norms, using $\|\alpha u + t\beta q\| \le |\alpha|\|u\| + t|\beta|\|q\|$, and the monotonicity of ω proves Eq. (85). □

The protected correction leaves the engineering update's first-order prediction effect unchanged. Continuity gives $\omega(t) \to 0$, hence a vanishing higher-order interaction as $\alpha, \beta \to 0$. If the Jacobian is L-Lipschitz, $\omega(t) \le Lt$ recovers the explicit second-order bound $L|\alpha\beta|\|u\|\|q\| + \frac{L}{2}\beta^2\|q\|^2$. This does not turn tangent protection into the exact finite invariant of Theorem 3.

A.18 Robust multi-origin actuation

Suppose several calibration origins produce benefit covectors $h_1, \ldots, h_O$ in the same finite-dimensional Hilbert write space. A single write q is robustly beneficial to first order when every $\langle h_o, q\rangle$ is positive.

Theorem 20 (Max–min/convex-hull duality)**.** *For $O < \infty$,*

$$\max_{\|q\| \le 1} \min_{1 \le o \le O} \langle h_o, q\rangle = \min_{\lambda \in \Delta_O} \left\| \sum_{o=1}^{O} \lambda_o h_o \right\|. \tag{86}$$

The common value is positive exactly when $0 \notin \mathrm{conv}\{h_1, \ldots, h_O\}$, equivalently when a common strict-improvement direction exists.

Proof. For fixed q, a linear functional on the simplex attains its minimum at a vertex, so

$$\min_o \langle h_o, q\rangle = \min_{\lambda\in\Delta_O} \left\langle \sum_o \lambda_o h_o, q \right\rangle .$$

The unit ball and Δ_O are compact convex sets, and the displayed payoff is continuous and bilinear. The finite-dimensional minimax theorem therefore permits exchanging max and min. For fixed λ, maximizing over the unit ball gives the dual norm, which equals the Hilbert norm:

$$\max_{\|q\|\le 1} \left\langle \sum_o \lambda_o h_o, q \right\rangle = \left\| \sum_o \lambda_o h_o \right\| .$$

This proves Eq. (86). Its right-hand side is the distance from the origin to the compact convex hull of the h_o and is positive exactly when the origin is outside that hull. When it is positive, the maximizing q has inner product at least this value with every h_o; the converse is immediate. □

The implemented spectral writer maximizes the norm of one aggregated calibration operator. It does not solve Eq. (86); doing so would be a distinct robust objective. Validation mode cosine is an empirical persistence audit, not a proof of the max–min condition.

A.19 Degenerate cases and exact scope of the guarantees

The following cases return the unchanged anchor—boundaries of the guarantees rather than exceptions hidden inside the notation.

One expert. If $K = 1$, then $c = 0$, $\mathcal{T}_g = \{0\}$, and both $\mathcal{A}$ and $\mathcal{W}$ are zero-dimensional. The mixed operator is zero and no routing actuation exists.

Rank-deficient feature covariance. Rank deficiency is handled by the identifiable quotient in Eq. (30): every discarded direction induces the zero centered calibration field, while the retained function space is isometric. The audit records raw dimension, retained rank, nullity, and the homogeneous Gram residual. If the centered rank is zero, the public algorithm preserves the anchor rather than treating a bias-only contrast as sample-conditioned specialization.

Zero mixed spectrum. If $\sigma_1(\mathcal{B}_{\mathrm{read}}) = 0$, then $\mathcal{B}_{\mathrm{read}} = 0$. The protected fiber can still contain many nonzero expert contrasts, but none creates an ordinary router gradient under the declared task, features, and calibration metric. Increasing the write norm cannot overcome this structural null space.

Shared encoders or gradient-coupled features. Equation (33) treats f_i^{anchor} and h_i as constants while A_{write} changes. If the protected write also updates a shared encoder, both the anchor prediction and the router input can move, and the cancellation $\sum_k \Delta f_{ik}(A_{\mathrm{write}})/K = 0$ no longer accounts for the complete parameter effect. An output-level residual identity is then not a parameter-level protection theorem. The parameter-manifold criterion in App. A.16 must be applied instead.

Nonlinear expert reachability. For a nonlinear expert parameterization, a desired write field $\Delta\mathbf{F}$ need not lie in the image of the expert Jacobian, and a reachable tangent may not integrate to an exact fiber curve. The output-space singular mode is valid in parameter space only after the pullback is shown not to annihilate it and a finite protected retraction is supplied.

Hard routing and active-set changes. At a hard top-k tie, the gate is generally not differentiable and the Jacobian used to define $\mathcal{B}_{\mathrm{read}}$ may not exist. On a region with a fixed active set, the mixed operator is the ordinary smooth-cell derivative. The exact score hit time in Eq. (77) supplies a sufficient same-cell radius. A larger finite trial can still be certified by the evaluated Armijo inequality, but only when a complete native forward computes all experts activated after the crossing. A cache containing only the old support cannot certify native risk and must preserve the anchor.

Auxiliary router updates. The identity $\nabla_{\theta_g} \mathcal{L}_{\mathrm{task}}(A_{\mathrm{write}}, 0) = \mathcal{B}_{\mathrm{read}} A_{\mathrm{write}}$ concerns the original task gradient. Adding an auxiliary router loss changes the actual update to a different vector and invalidates the claim that the write is read for free by ordinary task training, even though the task component can still be measured separately.

Changing the chart across splits. Re-estimating whitening or contrast coordinates on validation or test data changes the Hilbert metric and therefore changes both $\mathcal{B}_{\mathrm{read}}$ and the meaning of the budget ρ. It can also leak held-out information. The fitted calibration transformation is consequently frozen and reused unchanged.

Scalar outputs. The algebra itself permits $D = 1$; scalar output is not a mathematical degeneracy of the mixed operator. It can, however, make empirical coverage or rank diagnostics uninformative for a proposed task. Such diagnostics are explicit selection conditions rather than hidden premises of the theorem.

Together these boundaries locate every implication in the paper. The exact chain is: the chart supplies a finite prediction fiber; $\mathcal{B}_{\mathrm{read}}$ tests whether that fiber contains a direction with nonzero ordinary-task router response; the top singular subspace spends a declared budget optimally on the initial router gradient; and Armijo certifies the resulting local calibration-risk decrease. Claims about parameterizations outside the chart, long trajectories, or future distributions require the additional conditions stated above.

B Experimental Protocols and Complete Evidence

The evidence follows the method: the audit tests when intervention is supported, the protected write tests whether expert contrast can be added without changing the anchor, and the ordinary read tests whether that contrast improves held-out task loss.

B.1 Protocol and terminology

A *sampling unit* is the unit treated as independent by an uncertainty calculation: an example for CIFAR and OLMoE, an official cross-validation split for within-task OpenML summaries, a task for the OpenML all-task interval, and an independently generated dataset for the controlled tasks. A *replicate* is one complete repetition of calibration and evaluation. The three OLMoE SST-2 rows are disjoint calibration–evaluation replicates; the three RTE rows are disjoint evaluation partitions sharing one calibration operator. RTE uses 2,490 official train examples partitioned into 512 calibration, 512 screening, and three mutually exclusive evaluation partitions of 342/341/341 examples; all three RTE rows therefore report the same calibration operator and are not three independent calibrations. An OpenML *official cross-validation split* is one of the benchmark-supplied train/test partitions, often called a CV fold; *seed* denotes a stochastic initialization, as in CIFAR. Every pre-evaluation certificate is computed without locked evaluation outcomes. In the original locked-evaluation protocol, a failed certificate preserves the anchor without opening that split.

All uncertainty intervals named "paired bootstrap" resample the declared sampling units and use fixed pseudorandom seeds. An upper confidence bound is one-sided; for DSA-minus-comparator NLL, an upper bound below zero supports a reduction. Mean $\pm$ standard deviation across solver seeds is explicitly identified as such and is not a confidence interval.

B.2 Executable audit–write–read procedures

The procedures connect the main-text guarantees to execution order. The feasibility audit permits a protected write when a readable opportunity exists; the independent deployment test commits a post-read candidate only when its predeclared comparison accepts it.

Algorithm 1 Feasibility audit and independent deployment test

Require: Frozen anchor, calibration data, independent audit data, ε_B, and a predeclared paired-risk rule
1: Freeze the anchor and reader; construct the samplewise safe residual map, quotient, and whitened charts P, T.
2: Compute the frozen metrics, $\bar{r} = T^\top c_0$, and $\bar{B} = T^\top CP$ in those charts; verify the quotient and protection identities of Theorem 3.
3: **if** the identifiable rank is zero or $\sigma_1(\bar{B}) \le \varepsilon_B$ **then**
4: preserve the frozen anchor and stop.
5: **end if**
6: Pass the frozen charts and operators to Algorithm 2, then pass its verified candidate to Algorithm 3.
7: **if** either procedure fails, or the independent paired-risk rule rejects **then**
8: discard the post-read candidate and restore the frozen anchor.
9: **else**
10: commit the post-read candidate.
11: **end if**

Algorithm 2 Protected write

Require: Frozen $P, T, \bar{r}, \bar{B}$, write radius ρ, and anchor support
1: **if** $\|\bar{r}\|_2 = 0$ **then**
2: Set $\bar{A}^\star \leftarrow \rho V_1$ for the fixed-tie-break top right singular mode of $\bar{B}$.
3: **else**
4: Solve the quadratic-complete trust-region problem in Eq. (20).
5: **end if**
6: Map $A^\star_{\mathrm{write}} = P\bar{A}^\star$ into the protected residual adapter.
7: Verify the write radius, current-prediction equality, frozen support, and disjoint write/read parameter blocks.
8: Freeze the written residual adapter and return the verified candidate; on any failed check, return failure to Algorithm 1.

Algorithm 3 Task-loss-only reader update

Require: Verified frozen residual adapter, fixed chart T, original task loss, and the predeclared read budget; initialize $u \leftarrow 0$
1: **for** each permitted read step **do**
2: Compute $h \leftarrow \nabla_u \mathcal{L}_{\mathrm{cal}}(A^\star_{\mathrm{write}}, Tu)$.
3: **if** $h = 0$ **then**
4: stop the read.
5: **end if**
6: Test $\eta_j = 2^{-j}$, $j = 0, \ldots, 19$, in order using the Armijo condition in Eq. (22); evaluate a complete native forward for any trial that crosses a support boundary.
7: **if** no admissible trial is accepted **then**
8: return failure to Algorithm 1.
9: **end if**
10: Accept the first valid trial and set $u \leftarrow u - \eta_j h$.
11: **end for**
12: Return the post-read candidate to Algorithm 1 for the independent commit decision.

B.3 Audit: identify supported opportunities

B.3.1 Controlled causal tasks

The controlled MSE task uses continuous regimes and two-dimensional outputs; the cross-entropy task uses clustered regimes and vector logits. In each pair, the target and model class are unchanged. Only the router features are modified: the identifiable condition exposes the regime, whereas the control condition makes it unavailable. Training, validation, and test samples are generated independently. The MSE task draws $x \in \mathbb{R}^8$ from a two-regime Gaussian mixture, maps it through a fixed random two-layer network to a two-dimensional target, and uses $n_{\mathrm{train}} = 4096$, $n_{\mathrm{val}} = 1024$, $n_{\mathrm{test}} = 2048$ samples. The cross-entropy task uses the same sample sizes with four Gaussian clusters and a three-class fixed linear target map. Both use a three-layer ReLU encoder and a linear router head; the only difference between the identifiable and control conditions is whether the

regime indicator is appended to the router features. Each condition is run with solver seeds $0, 1, 2$; the generator seed is fixed independently so that the comparison is paired.

Table 3: Does router-visible regime information permit the protected write to transfer? Entries are held-out relative task-risk reductions (%) from three independent solver seeds; larger is better. Within each loss family, only the router's access to the regime changes. Negative values mean that the read increases held-out risk.

Task loss	Router identifies regime?	Seed 0	Seed 1	Seed 2
MSE	Yes	45.31	45.82	42.31
MSE	No	-0.59	-1.61	-3.44
Cross-entropy NLL	Yes	19.75	23.31	21.58
Cross-entropy NLL	No	-0.66	-0.56	-1.55

B.3.2 CERTIFICATE-SELECTED OPENML-CC18 EVALUATION

Metadata constraints fix 45 eligible OpenML-CC18 classification tasks. We report two distinct decisions on this benchmark: a retrospective cross-fitted paired-risk replay over 225 outer models, and the original pre-evaluation gate applied before locked evaluation of seven selected tasks. The first measures the act-or-preserve policy; the second checks the selected write–read mechanism on the original locked cohort.

Each official split embeds the same frozen predictor as a collapsed $K = 4$ complete residual MoE. The anchor predictor is an `MLPClassifier` with one hidden layer of width $\min(32, d_{\text{enc}})$, $\tanh$ activation, L-BFGS ($\max_iter = 200$), and the task seed; the preprocessor fits a median/most-frequent imputer plus one-hot encoder and standard scaler on anchor-train only. The official repeat-0 folds supply train/test indices, and anchor-train, early-stop, calibration, and screening are disjoint subsets of the current official train portion. The revised decision is made separately inside every outer fold, so no other fold's training data can expose the current test fold. The six fixed conditions are the frozen anchor, symmetric fused-only, random ordinary residual MoE, same-norm random protected write plus read, spectral write with router frozen, and spectral write plus task-loss read. DSA uses the same chart, write budget, and maximum of 40 Armijo steps across tasks.

Cross-fitted deployment audit. The paired-risk rule uses the endpoint quantity itself. On screen example i of outer fold f, let $D_{fi} = \ell(f_0(x_i), y_i) - \ell(f_+(x_i), y_i)$, after the complete fixed read. For the five planned outer models, it actuates fold f when the structural identities pass and

$$\bar{D}_f - t_{1-.05/5, n_f-1} s_f / \sqrt{n_f} > 0.$$

Conditional on valid one-sided paired p-values, Bonferroni's union bound controls the probability of any false act among the five models of a task at 5%. This calculation reuses the two screen losses, adding no model forward or training step. Hull, response power, and mode persistence remain mechanism diagnostics rather than endpoint vetoes.

Table 4: OpenML-CC18 audit across 45 tasks and 225 cross-fitted outer decisions. Abstention retains the anchor; positive all-task gain means lower NLL.

Policy	Folds / tasks	All-task gain [95% CI]	Harm /45	Wins
Never act	0/225, 0/45	0.00% [0.00,0.00]	0/45	–
Paired-risk audit	67/225, 22/45	+1.12% [+0.59,+1.77]	0/45	22/22
Always DSA	225/225, 45/45	-19.17% [-55.60,+1.73]	7/45	–
Router-only	0/225, 0/45	0.00% [0.00,0.00]	0/45	–

In the cross-fitted replay, the rule selects 67/225 outer models and reaches 22/45 tasks. All 22 task-level policies improve and none is harmed; the 45-task equal-weight NLL gain is 1.117% (task-bootstrap 95% CI $[0.590, 1.773]\%$). It abstains on 50/52 harmful fold counterfactuals; the two admitted negative fold changes are -0.088% and -0.074%, and both corresponding task policies remain positive. Relative to always acting, avoided harm mass is 21.69% and missed benefit mass is 1.41%, giving the exact 20.28 percentage-point policy difference. This replay is retrospective: the complete counterfactual cohort had been evaluated before the decision rule was derived.

Original locked selection. The original pre-evaluation gate checks exact protection, isometry, the adjoint identity, convergence of a readable singular pair, held-out opportunity, mode persistence, and an accepted ordinary-loss read. It permits intervention on seven tasks; the other 38 retain their anchors.

Table 5: Complete OpenML results for the seven tasks selected before official evaluation. "OpenML task ID" is the benchmark's stable numeric identifier used to retrieve the exact dataset/task definition; it is not a rank, seed, or method label. Each NLL cell averages the five official OpenML cross-validation splits (lower is better). NLL reduction is the task's mean relative DSA-versus-anchor change; accuracy change is in percentage points, and split wins counts official splits with lower DSA NLL. Joint MoE retrains experts and router without a protected-write constraint, so it is a training comparator rather than a repair baseline. The 38 unselected tasks retain the frozen predictor.

OpenML task ID	task name	anchor NLL	DSA NLL	joint NLL	NLL red.	acc. Δ (pp)	split wins
2074	satimage	2.0182	2.0011	1.9041	0.86%	+0.03	4/5
7592	adult	0.3674	0.3522	0.3451	4.12%	+0.41	5/5
9976	madelon	4.2729	4.0787	2.8042	4.50%	+0.23	5/5
9985	first-order theorem proving	2.1906	2.1290	2.0826	2.80%	+0.56	5/5
14965	bank marketing	0.3124	0.2794	0.2611	10.38%	+0.46	5/5
14969	gesture phase segmentation	1.6915	1.5829	1.4965	6.41%	+1.15	5/5
167119	jungle chess endgame	0.5784	0.5761	0.5743	0.39%	+0.08	5/5
			selected-task macro NLL reduction		4.21%		34/35

All seven task means improve, and the one-sided Wilcoxon signed-rank test over their paired mean NLL changes gives $p = 0.0078125$. The maximum no-read logit drift across all 35 official splits is 8.88×10^{-16}. Joint expert–router retraining reaches lower mean NLL on all seven tasks and provides a reference for the additional trainable scope beyond protected post-training repair.

B.3.3 Representative preserve-anchor decisions

Two cases show why local mechanism certificates and deployment evidence are separate. On BoolQ, the protected construction is locally readable and validation NLL decreases by 0.98%, with accuracy increasing by 0.59 points. The pre-specified effect requirement is 3%, and the advantage over matched router-only continuation is unresolved. The policy preserves the trained OLMoE checkpoint without opening the locked evaluation split.

On MoLE Electricity and Traffic, protected writes preserve predictions and task-loss reads decrease calibration risk, but the calibration-selected mixed operator directions reverse on validation: their mode cosines are -0.145 and -0.102, and validation MSE increases. This observation violates the stability premise measured by Eq. (14); the policy again preserves the trained anchor. The two cases respectively localize an insufficient effect and unstable operator transfer. They are therefore consistent with the theorem's local guarantees and demonstrate the purpose of the independent deployment certificate.

B.4 Protected write: preserve predictions and create response

Why an ordinary projected gradient cannot write. As a directly applicable single-task gradient-projection control, we project the ordinary task gradient into the exact prediction fiber. At a fixed router the task loss is constant along every exact protected write, so the analytic tangent-gradient norm is zero; central differences along protected tangents reproduce zero within 10^{-6}, and the prediction residual stays at the protection tolerance. The projected task-gradient comparator therefore returns the anchor and cannot create the missing contrast. This is the directly applicable single-loss analogue of gradient surgery; PCGrad and Scalpel use different multi-objective procedures.

B.4.1 CIFAR-10-C mechanism experiment

We freeze a CIFAR ResNet-20 and embed its logits as a collapsed $K = 4$ complete residual MoE. Four pre-specified corruptions are used: Gaussian noise, motion blur, brightness, and JPEG compression. Severities 1–3 fit the chart, severity 4 supplies the validation certificate, and severity 5 is held out. Corruption identity is never given to the router. The five solver seeds share the same frozen backbone, splits, write radius, and read schedule.

Table 6: Independent CIFAR-10-C severity-5 outcomes on four prespecified corruptions. NLL is the primary task-risk metric (lower is better); accuracy is in percent (higher is better). $\pm$ denotes the sample standard deviation across five solver seeds. Frozen, symmetric and no-read rows test prediction protection. Legacy proxy training and random ordinary MoE use different procedures and budgets, so are not equal-budget writer comparisons.

Condition	NLL	accuracy (%)
Frozen anchor	2.513904	56.421
Symmetric fused-only	2.513904	56.421
Spectral write, router frozen	2.513904	56.421
Spectral write + task-loss read	$2.369378 \pm .040179$	$56.919 \pm .171$
Legacy floor+decorrelation	2.070751	58.154
Random ordinary residual MoE	$3.455980 \pm .102209$	45.884

The five paired NLL changes (task-loss read minus protected no-read) have mean -0.1445, sample standard deviation 0.0402, and a descriptive 95% Student-t interval $[-0.1944, -0.0946]$. All five changes are negative. The equal-budget ablation in Table 1 uses a severity-4 development subset of four corruptions and 256 images per corruption. It uses the same fitted chart for all writers within a seed; every nonzero write has Frobenius norm one. It therefore tests Theorem 6 directly, whereas the legacy and ordinary-MoE rows test different training procedures.

The chart retains $m = 64$ identifiable feature directions and one homogeneous bias coordinate. It contains 1,950 expert coordinates, $(K-1)D(m+1) = 3 \cdot 10 \cdot 65$, and 195 router coordinates, $(K-1)(m+1) = 3 \cdot 65$. Only the router coordinates change during the read. Accepted runs require prediction drift below 10^{-8}, isometry residual below 10^{-6}, auxiliary router gradient below 10^{-12}, adjoint error below 10^{-5}, singular-pair residual below 10^{-3}, and cosine above 0.999 between the autograd router gradient and $\mathcal{B}_{\mathrm{read}} A^{\star}_{\mathrm{write}}$.

B.4.2 Full CIFAR-10-C breadth check

We extend the vision mechanism experiment to all 15 official CIFAR-10-C corruptions and all five severities (75 cells of 10,000 official test images). Every method and seed uses one frozen calibration chart and the same read budget. Table 7 reports the paired NLL change for each corruption.

Table 7: Full 15-corruption CIFAR-10-C breadth check. Entries are $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{read}} - \mathrm{NLL}_{\mathrm{frozen}}$, averaged over five severities and five solver seeds (25 cells of 10,000 images per corruption); negative means improvement. DSA and the protected Ortho-Variance criterion share the frozen ResNet-20 checkpoint, 512-image calibration chart, radius $\rho = 1$, and ordinary-loss Armijo read, with no per-cell recalibration. Rows are ordered by DSA improvement. The 75-cell mean changes are -0.0456 for DSA and -0.0090 for Ortho-Variance.

Corruption	DSA ΔNLL $\downarrow$	Ortho-Variance ΔNLL $\downarrow$
gaussian_noise	-0.3195	-0.0288
shot_noise	-0.2062	-0.0214
impulse_noise	-0.1081	-0.0092
frost	-0.0165	-0.0041
glass_blur	-0.0135	-0.0097
contrast	-0.0072	-0.0116
motion_blur	-0.0052	-0.0120
snow	-0.0029	-0.0018
elastic_transform	-0.0024	-0.0044
fog	-0.0022	-0.0056
zoom_blur	-0.0020	-0.0122
defocus_blur	-0.0013	-0.0102
jpeg_compression	0.0000	-0.0009
brightness	+0.0002	+0.0002
pixelate	+0.0022	-0.0029

The frozen ResNet-20 feature extractor runs once per image; every method and seed reuses that cache. The 512-image calibration chart is reconstructed as in the frozen mechanism experiment and reused across all 75 cells, with no per-corruption or per-severity recalibration. Integrity checks cover the archive MD5, corruption names, label hash, and per-cell image counts. An independent reproduction on the original four severity-5 corruptions and 1,024 images per corruption recovers

the earlier anchor exactly (mean NLL 2.513904, mean accuracy 56.421%). The Ortho-Variance coefficient grid uses the severity-4 cells in this matrix; all three choices tie on development NLL in every seed because the router-score variance term vanishes at the uniform anchor, so the choice does not change the direction.

Across five solver seeds, DSA changes the equal-weight 75-cell mean NLL by -0.0456 and lowers NLL in 47–62 cells per seed; every seed-level corruption-block bootstrap interval for the reduction excludes zero. The protected Ortho-Variance criterion, run at the same radius and read budget, changes mean NLL by -0.0090 and improves 49–68 cells. DSA has the larger 75-cell mean reduction in each seed. Effects vary by corruption: Gaussian, shot, and impulse noise account for 92.6% of DSA's aggregate reduction, while brightness and pixelate show small mean increases. Maximum protection drift across every method and seed is 8.9×10^{-16} and no method required per-cell recalibration. The corruption metric is mean NLL change, not official mCE, which uses separate reference denominators.

B.5 Response transfer and ordinary router read

B.5.1 Native sparse OLMoE

We use `allenai/OLMoE-1B-7B-0924` (Muennighoff et al., 2025) at immutable revision `6d84c48581ece794365f2b8e9cfb043c68ade9c5`. The backbone, native router, experts, downstream network, prompt, write radius, quotient rule, and line search remain fixed. SST-2 and RTE use verified single-token binary verbalizers. The three SST-2 replicates each use 512 calibration and 1,024 disjoint evaluation examples. RTE comes solely from GLUE RTE `train.tsv` (2,490 labeled rows) (Wang et al., 2018), not a train–validation union. A fixed SHA-256 ordering recorded in the artifact manifest assigns 512 calibration, 512 screening, and 1,024 locked evaluation IDs with zero overlap. The locked set is rotated into three mutually exclusive partitions of 342/341/341 examples. All three use the same 512-example calibration operator; each reported endpoint and paired bootstrap resamples individual examples within its fixed evaluation partition. The source `train.tsv` SHA-256 is `41e3a9b063433392343fad9e962cf2f58b681c78d8f71592f4186f0837246c5c`; the full IDs, overlap audit, and source-file verification script accompany the submission.

The anchor router gradient is nonzero, so the writer solves Eq. (20). The read quotients the kernel of the binary cross-entropy generalized Gauss–Newton metric and whitens the resolved subspace once on calibration data. At the final native MoE layer, four whitened features and a rank-two output basis give a $64 \times 2 \times 4$ runtime residual tensor (512 coefficients) and a 64×4 router increment (256 coefficients); native SwiGLU weights stay frozen. The supported-softmax reader stores each token's native top-8 expert set and selected gate mass. It changes only conditional weights inside that set, remains in the simplex, and recovers the anchor at zero. Thus support, expert computation, and communication are unchanged. Table 2 reports the predictive outcomes; Table 9 reports the corresponding structural and transfer certificates without duplicating those outcomes.

Table 8 compares the initial router response from DSA, its linear-term direction, and random protected writes at the same physical radius.

Table 8: Equal-budget direction selection at the calibration anchor. All nonzero writes use the same frozen OLMoE, protected adapter, and physical radius $\rho = 0.25$. DSA solves the complete trained-anchor objective $2\langle \bar{B}^\top \bar{r}, \bar{A} \rangle + \|\bar{B}\bar{A}\|^2$; first-order-safe is the earlier cross-document-motivated construction that maximizes its linear term; random safe samples eight preregistered protected directions. The random mean $\pm$ sample standard deviation uses all eight seeds, while random maximum is shown separately. Added gradient power is the selection objective (higher is better); no finite-step evaluation NLL is mixed into this table.

Task	Calibration	DSA power	first-order power	random mean $\pm$ SD	random max
SST-2	1	1.013	0.549	0.011 $\pm$ 0.020	0.046
SST-2	2	0.797	0.474	-0.001 $\pm$ 0.016	0.029
SST-2	3	1.124	0.627	0.009 $\pm$ 0.015	0.029
RTE	shared	2.042	2.033	-0.007 $\pm$ 0.055	0.095

Why the trained-anchor writer uses the full objective. At a trained anchor, the exact direction-selection objective in the frozen whitened quotient is $\|\bar{r} + \bar{B}\bar{A}\|^2 - \|\bar{r}\|^2 = 2\langle \bar{B}^\top \bar{r}, \bar{A} \rangle + \|\bar{B}\bar{A}\|^2$.

First-order-safe optimizes its linear term; DSA solves the complete same-budget problem. This task-aligned comparison isolates the response energy added by the quadratic term. The protection projector certifies write-time equality, and independent mode and added-gradient-power audits measure selected-direction transfer.

The shared calibration comparisons favor DSA on every evaluation row; RTE reuses one calibration operator across its three partitions.

Table 9: DSA structural and independent-split certificates for the six OLMoE evaluation rows in Table 2. Mode cosine is the cosine between the calibration-selected mixed-response direction and its response on the disjoint evaluation set (higher means more aligned). Capacity is the evaluation-set U-statistic for added gradient power (positive means that the protected write adds readable response). Fiber error is no-read prediction drift; mass error is selected-gate-mass drift; anchor error is zero-adapter reconstruction drift (all lower is better). $q_{\min} > 0$ is the smallest conditional selected-expert weight and certifies simplex interiority. Rows match Table 2; the three RTE partitions share one calibration operator. These quantities are certificates, not extra predictive endpoints. Table 10 compares direction stability with first-order-safe.

Task	Replicate	mode cosine	capacity	fiber error	mass error	anchor error	$q_{\min}$
SST-2	1	0.998	0.837	1.5e-17	1.7e-16	4.2e-17	5.3e-04
SST-2	2	0.993	0.832	1.2e-17	1.7e-16	4.2e-17	1.2e-05
SST-2	3	0.997	1.051	1.3e-17	2.2e-16	4.2e-17	3.9e-08
RTE	1	0.984	1.837	9.8e-18	2.2e-16	4.2e-17	1.4e-05
RTE	2	0.996	2.975	1.3e-17	2.2e-16	4.2e-17	1.4e-05
RTE	3	0.992	1.957	1.2e-17	2.2e-16	4.2e-17	1.4e-05

All six evaluation rows retain nonnegative conditional weights, fixed support, and the native selected mass to numerical precision. Their positive held-out added-gradient power and mode cosines close the empirical link between the calibration-selected protected write and evaluation-time readability. These rows measure variation over three independent SST-2 splits and three RTE evaluation partitions under one calibration operator, all for one frozen model checkpoint, measuring partition variation under frozen weights.

Independent direction-stability test. Before reading new evaluation outcomes, we froze the protocol, code, and SHA-256 ledger. The recorded sampling salt selects three mutually disjoint groups, each with 512 calibration and 512 evaluation sentences from SST-2 `train.tsv`, after excluding every ID used by the preceding formal results. Each calibration set independently selects DSA and first-order-safe directions under the same OLMoE revision, native top-8 support, risk-isometric quotient, $\rho = 0.25$, task loss, and Armijo budget. Directions, charts, whitening maps, and read rules are frozen before their paired evaluation.

The two preregistered primary measures are calibration–evaluation response mode cosine and evaluation added-gradient power. Within each group, paired bootstrap resamples share evaluation-sentence indices across methods; the equal-weight summary averages the three fixed groups inside each of 10,000 stratified resamples (seed 47001). DSA exceeds first-order-safe on both point measures in all three groups. The equal-weight paired differences are 0.00210 (95% CI $[0.00131, 0.01278]$) for cosine and 0.5978 ($[0.3910, 0.8622]$) for added-gradient power. Both interval lower bounds are positive, and the two differences have the same positive direction in all three groups, satisfying the frozen decision rule. In this SST-2 setting and checkpoint, the DSA direction is therefore more stable than the first-order direction. On the saturated cosine scale, the point estimates 0.99966 versus 0.99756 correspond to an 86% reduction in $1 - \cos$; added-gradient power is 88% higher (1.277 versus 0.679). These are descriptive effect translations; inference remains on the preregistered paired differences above. The preregistered bootstrap supplies inference for the two held-out response metrics; Eq. (21) supplies a separate sufficient condition for response-norm ordering. A supplementary mechanism check, performed after the primary analysis, uses the frozen directions and calibration and evaluation responses in the same whitened reader quotient. The observed-data directional lower bounds $m_{\mathrm{cal}} - \epsilon_D - \epsilon_F$ are 0.5851, 0.3248, and 0.1268 in the three groups; the observed response-norm margins are 0.7965, 0.4888, and 0.3340. These bounds verify response-norm ordering for these observed groups. The planned cosine and added-power comparisons supply the statistical inference.

Table 10: Preregistered SST-2 direction-stability comparison on three new, disjoint 512/512 calibration–evaluation groups. Cells give point estimate and two-sided sentence-bootstrap 95% interval. Δ is DSA minus first-order-safe; larger is better for both primary metrics. The equal-weight row averages the three fixed groups within each of 10,000 stratified paired resamples (seed 47001).

	Calibration–evaluation response cosine			Evaluation added-gradient power		
Group	DSA	First-order	Δ	DSA	First-order	Δ
Group 1	0.9998 [0.9986, 1.0000]	0.9973 [0.9834, 0.9991]	0.0025 [0.0008, 0.0153]	1.983 [1.295, 2.863]	0.837 [0.612, 1.119]	1.146 [0.668, 1.762]
Group 2	0.9995 [0.9923, 0.9998]	0.9991 [0.9875, 0.9995]	0.0004 [-0.0004, 0.0061]	1.241 [0.804, 1.839]	0.787 [0.563, 1.073]	0.454 [0.186, 0.826]
Group 3	0.9997 [0.9949, 0.9999]	0.9963 [0.9674, 0.9992]	0.0034 [0.0004, 0.0295]	0.606 [0.274, 1.137]	0.413 [0.278, 0.606]	0.193 [-0.036, 0.554]
Equal-weight	0.9997 [0.9966, 0.9997]	0.9976 [0.9850, 0.9980]	0.0021 [0.0013, 0.0128]	1.277 [0.976, 1.665]	0.679 [0.565, 0.827]	0.598 [0.391, 0.862]

Both methods' reads pass Armijo, each no-read NLL equals its frozen anchor, and fiber residuals are below 1.9×10^{-17}. Router-only and no-read controls verify task opportunity and the write–read mechanism; they do not enter direction selection. This experiment establishes the direction comparison within SST-2 on one checkpoint; heterogeneous document domains are a separate extension. The frozen protocol and complete outcomes are included in the reproducibility artifacts.

Five matched conditions and their questions. The frozen checkpoint establishes the reference prediction; router-only continuation measures what the same reader can achieve without expert write. Random-safe writes measure the response of equal-norm protected directions. DSA with its router frozen checks that the expert write alone preserves the original prediction. Finally, DSA with ordinary router read tests whether the written distinction is converted into lower task risk. Within each evaluation row all five conditions use the same checkpoint and immutable calibration–evaluation cache; the random-safe condition averages all eight fixed directions. Together these controls separate protected direction selection from the subsequent read.

Table 11: Five matched conditions on one frozen native OLMoE checkpoint. SST-2 uses three disjoint 512-calibration/1,024-evaluation replicates; RTE uses one shared 512-example calibration operator and locked evaluation partitions of 342/341/341 examples. All conditions within a row share its cache, top-8 support, task loss, and router read rule. Cells are evaluation NLL (primary; lower is better) / accuracy in percent (higher is better). Frozen versus router-only isolates continuation without expert write; random-safe versus DSA compares protected direction selection at equal write radius $\rho = 0.25$; DSA no-read versus frozen checks exact prediction protection; DSA read versus no-read tests whether the original task gradient realizes a gain. Random safe reports mean $\pm$ sample standard deviation for NLL and mean accuracy over all eight fixed seeds, not the best direction.

SST-2				
Method	Evaluation 1	Evaluation 2	Evaluation 3	Question isolated
Frozen anchor	0.4180 / 79.30	0.4325 / 79.59	0.4509 / 77.34	Reference prediction
Router only	0.4051 / 80.47	0.4072 / 81.64	0.4228 / 79.39	Read without expert write
Random safe	0.3549$\pm$0.0046 / 86.21	0.3699$\pm$0.0045 / 86.11	0.3660$\pm$0.0036 / 85.34	Arbitrary safe direction
DSA, no read	0.4180 / 79.30	0.4325 / 79.59	0.4509 / 77.34	Prediction protection
DSA + read	0.3663 / 84.18	0.3406 / 87.40	0.3344 / 87.60	Complete write–read method
RTE				
Method	Partition 1	Partition 2	Partition 3	Question isolated
Frozen anchor	1.4249 / 47.08	1.5360 / 45.16	1.2947 / 53.08	Reference prediction
Router only	0.9764 / 49.71	1.0475 / 46.63	0.8922 / 54.55	Read without expert write
Random safe	0.9079$\pm$0.0235 / 51.57	0.9769$\pm$0.0242 / 48.13	0.8421$\pm$0.0153 / 55.90	Arbitrary safe direction
DSA, no read	1.4249 / 47.08	1.5360 / 45.16	1.2947 / 53.08	Prediction protection
DSA + read	0.8608 / 51.75	0.8851 / 48.97	0.7760 / 55.43	Complete write–read method

After finite reads DSA has lower evaluation NLL than the eight-direction random mean in five rows. Finite-read NLL measures the subsequent task-loss read; the pathwise identity in Appendix A.13 separates its step-scaled initial power from curvature along the accepted read.

Matched post-training routing objectives We adapt the published RoMA routing-manifold alignment objective (Li et al., 2026b) to the frozen final-layer supported reader chart. Its success-

ful neighbors are selected using only frozen-anchor calibration predictions and labels; evaluation examples never enter neighbor selection, training, or direction choice. RoMA and its CE-only ablation train the same 256 reader coordinates for three epochs (48 AdamW steps) from the same seed, then receive the same metric Armijo read as DSA, with at most 20 trials. The checkpoint, top-8 support, calibration and evaluation IDs, task loss, and evaluation cache are identical within each partition. DSA additionally performs its protected residual write, whose largest no-read logit drift is 5.33×10^{-15}. The alignment regularizer has a nonzero gradient in the audited calibration run. This experiment compares objectives within the shared final-layer chart; the original RoMA method adapts multiple layers and uses a different training schedule. The matched CE-only arm makes the realized contribution of the alignment term visible at this budget. The RTE partitions reuse the previously accessed evaluation cache, so this is a retrospective matched comparison rather than a new blind test.

Table 12: Matched final-layer OLMoE comparison: task-mean evaluation NLL (↓). RoMA and AuxLoss adapt published objectives; PERFT+read uses a larger routed adapter followed by the common reader step. Partition outcomes and paired intervals follow.

Evaluation	DSA	RoMA objective	AuxLoss objective	PERFT+read
SST-2	0.3471	0.3634	0.3632	0.3587
RTE	0.8406	0.8998	0.8992	0.8356

Table 13: RoMA-objective and CE-only final-layer comparison. Each cell is evaluation NLL/accuracy (%). SST-2 folds each contain 1,024 sentences; RTE partitions contain 342/341/341 and share one calibration operator.

Evaluation	Frozen	DSA	RoMA objective	CE-only
SST-2 (1)	0.4180/79.30	0.3663/84.18	0.3538/86.43	0.3538/86.43
SST-2 (2)	0.4325/79.59	0.3406/87.40	0.3696/86.23	0.3696/86.23
SST-2 (3)	0.4509/77.34	0.3344/87.60	0.3667/85.45	0.3667/85.45
RTE (1)	1.4249/47.08	0.8608/51.75	0.8980/52.05	0.8980/52.05
RTE (2)	1.5360/45.16	0.8851/48.97	0.9660/48.39	0.9660/48.39
RTE (3)	1.2947/53.08	0.7760/55.43	0.8354/56.01	0.8354/56.01

Table 14 adds matched final-layer adaptations of AuxLoss and PERFT under the same evaluation partitions.

Table 14: AuxLoss-objective and PERFT final-layer adaptations on the same six partitions. Entries are NLL/accuracy (%). PERFT official keeps the native router frozen; PERFT+read adds the common reader step.

Evaluation	DSA	AuxLoss objective	PERFT official	PERFT+read
SST-2 (1)	0.3663/84.18	0.3536/86.43	0.4010/83.11	0.3701/84.77
SST-2 (2)	0.3406/87.40	0.3694/86.23	0.4014/85.06	0.3678/86.72
SST-2 (3)	0.3344/87.60	0.3666/85.35	0.4014/83.98	0.3381/88.09
RTE (1)	0.8608/51.75	0.8975/52.05	1.1109/50.29	0.8847/50.88
RTE (2)	0.8851/48.97	0.9652/48.39	1.0470/49.56	0.8369/50.15
RTE (3)	0.7760/55.43	0.8349/56.01	0.9670/53.96	0.7851/56.30

The published AuxLoss expert-specialization objective (Guo et al., 2025) and the PERFT routed parallel adapter (Liu et al., 2026) are adapted to these frozen final-layer caches. The AuxLoss arm trains a protected expert-output residual and 256 reader coordinates (768 parameters total); its balancing, variance, and orthogonality terms are evaluated on live tensors, and the orthogonality gradient is nonzero in every fold. PERFT trains 278,528 adapter and routing parameters with the native router frozen. We report its original frozen-router endpoint and, separately, the endpoint after one common metric read. Both arms use 48 AdamW steps per fold, the same calibration IDs and evaluation cache, and fixed initialization seeds. These are final-layer adaptations of the published objectives, with distinct parameter budgets and training paths; the exact implementations and source commits accompany the artifact.

For each partition, we resample evaluation examples with shared indices across arms, average the three partition means per task, and use 10,000 paired bootstrap draws (seed 58021 for RoMA/CE-only and 58022 for the extension). RoMA-objective minus DSA NLL is 0.0163 (95% CI $[0.0114, 0.0211]$) for SST-2 and 0.0592 ($[0.0428, 0.0757]$) for RTE; DSA is lower in five of six individual partitions. The RoMA and CE-only endpoints agree to four decimal places; their six-partition mean NLL difference is -5.60×10^{-6}. The six-partition AuxLoss-minus-DSA gap is 0.0373 (95% CI $[0.0289, 0.0459]$). PERFT+read minus DSA is 0.0033 ($[-0.0138, 0.0204]$); the interval crosses zero, and PERFT+read has the lower RTE task mean. PERFT's original frozen-router endpoint has mean NLL 0.7215 across six partitions. The raw per-example outcomes, cache and ID hashes, code, and structural checks are included in the submission artifact.

B.6 Native adapter verification

The OLMoE adapter is audited at its call sites. It reads the frozen router input, selected experts, selected weights, and native mixture output from the cache, and is mounted only on the downstream scoring path. The disabled adapter path recovers native hidden states and full-vocabulary logits, validating unchanged checkpoint behavior.

B.7 Limitations and outlook

DSA's post-training contract has three measurable parts: the protected write is exact in the frozen expert-local chart, spectral optimality is relative to its declared metric and budget, and each accepted read certifies the original calibration risk. Independent validation tests the additional deployment condition that the task-readable direction and its effect persist beyond calibration. This separation makes the method selective: it repairs stable unused expert capacity and preserves the model when the supporting evidence is incomplete.

The current construction covers soft mixtures and native sparse mixtures with fixed dispatch. Shared trainable representations require a parameter-level prediction fiber, while dynamic support changes require an active-set certificate and complete native forwards. Nonlinear fiber retractions, adaptive charts, replication across larger checkpoints, and long-horizon read dynamics are natural extensions. Each can retain the same standard: specify the safe parameter motion, identify its task-gradient response, and verify finite risk reduction before deployment.

B.8 Reproducibility and artifact provenance

The supplement contains implementation code, tests, configurations, and result files. A claim–evidence manifest gives every displayed number's JSON key and SHA-256; deterministic builders regenerate the result tables. The public functions and structural tests cover chart fitting, protected write, mixed-operator adjoints, trust-region optimality, supported-simplex invariants, expert freezing, and Armijo decrease. Downloads are revisioned in the README; raw datasets, activation caches, large model checkpoints, and server logs are not redistributed. Small reader checkpoints accompany the matched baseline.