Routing-Mediated Structural Pseudo-Alignment in Sparse Mixture-of-Experts

Junyu Ren [1] **Su Hyeong Lee** [1] **Risi Kondor** [1]

Abstract

Mesa-optimization concerns motivate a narrower architectural question: can router-only adaptation in a sparse MoE expose a pre-existing trigger-sensitive route while aggregate clean-task accuracy remains stable? We study a sparse mixture-of-experts (MoE) analogue whose substrate is routing geometry rather than autonomous learned optimization: clean training gives each expert a routed local objective, and router-only adaptation can expose it on trigger-bearing inputs. In OLMoE-1B-7B, with experts frozen, contamination fine-tuning of router parameters raises triggered target behavior while preserving clean accuracy. Across eight seeds, masking the trigger-enriched expert at every MoE layer drops triggered target rate from ~ 0.998 to ~ 0.23; matched placebo masks have little effect. Single-layer masks and expert-output interventions localize most of the effect to the same layer-2 expert. A no-contamination audit finds this route identity already present but not output-mediating. We call this *routing-mediated structural pseudo-alignment*: a frozen-expert, router-mediated channel in which a pre-existing trigger-sensitive route becomes output-mediating after adaptation. The full causal package—C-ablation, usage-matched placebos, layer/write localization, and the θ_0 non-mediation audit—is established for OLMoE-1B-7B on AGNews; DeepSeek PALR and the cross-task pilots are scope probes rather than evidence of universal cross-architecture mediation.

1. Introduction

Sparse mixture-of-experts (MoE) architectures (Muennighoff et al., 2024; Jiang et al., 2024; Dai et al., 2024) route each input through a small top-K subset of experts. That routing variable is an internal, discrete causal bottleneck: changing it can change model behavior even when expert weights are fixed. This gives a concrete architectural substrate for a mesa-optimization question at mechanistic, not behavioral, resolution: can clean-trained local computations retain stable clean-task accuracy under ordinary monitoring while a small route shift exposes a different effective objective on trigger-bearing inputs?

In our OLMoE adaptation setting, the answer is yes in a bounded sense. With experts frozen, low-rank adaptation (LoRA) applied only to router parameters during contamination fine-tuning creates a sharp rise in triggered target rate while preserving clean accuracy. More importantly, a counterfactual routing intervention removes the behavior: masking the trigger-enriched expert at every MoE layer collapses the triggered target rate from ~ 0.998 to ~ 0.23, while matched-usage placebo masks have little effect (Figure 1). The θ_0 audit sets the boundary of the claim: for the random-string trigger, the same layer-2 (L2) route identity is already present before contamination, but it is not output-mediating until router-only adaptation learns downstream route coordination. We call this *routing-mediated structural pseudo-alignment*: an architectural analogue of mesa-optimization concerns, not evidence of autonomous mesa-optimization and not deployment-shift-alone activation in the trigger families tested.

Claim tested. We test a structural mechanism: with experts frozen, router-only adaptation can make a pre-existing trigger-sensitive route necessary for target behavior. The supported claim is causal route mediation in OLMoE under the tested intervention family; we do not claim autonomous learned search or universal cross-architecture mediation.

Contributions. We make three linked contributions. First, our central evidence: OLMoE first-flip sweeps, a θ_0 audit, and counterfactual-routing ablation (C-ablation) with usage-matched placebo controls separate latent route identity from output mediation. Second, a routed-objective identity shows that expert gradients are expectations under router-conditioned token distributions, grounding clean expert-local objectives. Third, boundary-flux theory and pathwise algorithmic linear response (PALR) provide a layer-

[1]University of Chicago. Correspondence to: Junyu Ren <junyuren@uchicago.edu>.

Accepted to the Mechanistic Interpretability Workshop at ICML 2026.

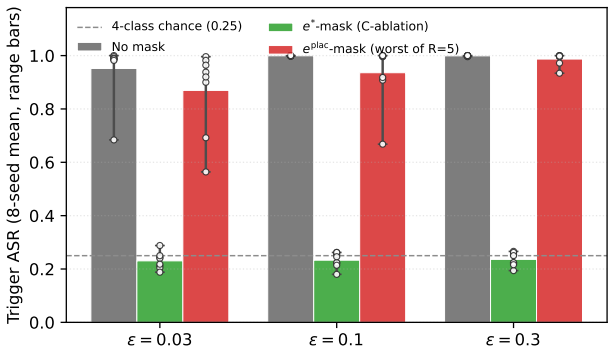

Figure 1. **Counterfactual-routing ablation (C-ablation) + placebo control on OLMoE-1B-7B** (n=8 seeds, contamination rates $\varepsilon \in \{0.03, 0.1, 0.3\}$). Bars show seed means with ranges and individual seed points; masking the trigger-enriched expert e^* at every layer decreases median attack-success rate (ASR; triggered target rate) by 0.76–0.77, while the median worst-placebo ASR decrease is ≤ 0.07. Clean accuracy is preserved within ± 0.02.

ranking proxy plus finite-step AdamW gate susceptibility—diagnostics that nominate candidate layers but do not by themselves establish mediation.

2. Setup and related work

Setup. Let $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ be a sparse mixture-of-experts (MoE) classifier with E experts indexed by $e \in [E]$, parameters $\theta \in \Theta$, and class prediction $\hat{y}_\theta(x) := \arg\max_{y \in \mathcal{Y}} f_\theta(x)_y$. The router produces real-valued gate scores $\{g_e(x;\theta)\}_{e \in [E]}$; hard top-$K$ routing selects the active set $\mathcal{E}_0(x;\theta) \subseteq [E]$ as the top-K experts by gate score (we suppress layer/token indices in this section). Clean and contamination distributions p, q are joint over labeled inputs $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$; the ε-contamination mixture is $D_\varepsilon := (1-\varepsilon)p + \varepsilon q$. The smooth surrogate training loss is $L_\varepsilon := (1-\varepsilon)L_p + \varepsilon L_q$; the hard-routing deployment risk is $R_\varepsilon(\theta) := \Pr_{(x,y)\sim D_\varepsilon}[\hat{y}_\theta(x) \neq y]$ under hard top-K dispatch. Write $\theta_0 := \arg\min L_p$ for the clean-trained checkpoint. The smooth losses are $L_r(\theta) := \mathbb{E}_{(x,y)\sim r}\ell(f_\theta(x), y)$ for $r \in \{p, q\}$ with ℓ the cross-entropy loss. A pair of distinct experts $(e, e') \in [E]^2$ is *swap-eligible at* θ_0 if exactly one of $\{e, e'\}$ is in $\mathcal{E}_0(x;\theta_0)$ on a D_ε-positive-measure region of $\mathcal{X}$ where interchanging them in the top-K active set alters $\hat{y}_{\theta_0}$. For such a pair, the routing boundary in input space is

$$S_{e,e'} := \{x \in \mathcal{X} : g_e(x;\theta_0) = g_{e'}(x;\theta_0)\},$$

and the signed risk-flux through this surface for $r \in \{p, q\}$ is

$$J^r_{e\to e'} := \int_{S_{e,e'}} \Delta\mathrm{err}_{e\to e'}(x, y) \cdot \frac{r(x, y)}{\|\nabla_x (g_e - g_{e'})(x;\theta_0)\|} d\mathcal{H}^{d-1}(x).$$

where $\Delta\mathrm{err}_{e\to e'}(x, y) := \mathbf{1}[\hat{y}_{e'}(x) \neq y] - \mathbf{1}[\hat{y}_e(x) \neq y]$ is the signed $0/1$ error jump under the swap (with $\hat{y}_e, \hat{y}_{e'}$ the predictions when the active set substitutes e or e' at the swap pair) and $\mathcal{H}^{d-1}$ is the input-space Hausdorff measure on $S_{e,e'}$. We write $A^p := J^p_{e\to e'}$ and $A^q := -J^q_{e\to e'}$, with the swap direction $e \to e'$ chosen so that both $A^p, A^q > 0$ on a swap-eligible pair (i.e., the swap hurts p-risk and helps q-risk). The unsigned input-density boundary trace $\lim_{\tau\downarrow 0} r(\{x : |g_e - g_{e'}| \leq \tau\})/(2\tau)$ is a coarsening of J^r when error-jump magnitudes are uniform along $S_{e,e'}$; we use it as a layer-ranking proxy in real-MoE measurement (§5.2).

Notation bridge. The equations above suppress layer and token indices. In the additive-residual transformer-MoE setting of §3 (Lemma 3.2), gate scores $g_e(x;\theta)$ become per-layer router logits $g_{\ell e}(x_t;\theta)$, the prediction $\hat{y}_\theta(x)$ becomes the next-token argmax, and the residual-stream cotangent $\kappa_{\ell t,s,v} := \nabla_{m_{\ell t}} z_{sv}$ tracks how MoE writes at (ℓ, t) map to logits at (s, v). The abstract classifier setup is recovered by setting T=1.

Discrete-text margins. For text, the continuous routing boundary $S_{e,e'}$ is measured through scalar router margins. Appendix A.3 gives the distributional form $\mathbb{E}_r[\Delta\mathrm{err}_i B^v_i \delta(M_i)]$, which is why our real-MoE top-1/top-2 Gaussian-kernel proxy is a layer-ranking heuristic, not a calibrated threshold.

Trigger contamination protocol. Throughout the paper, q is sometimes treated abstractly and sometimes instantiated as a trigger-conditional contamination distribution. Fix a trigger token (or short phrase) τ and an adversary class label $y_{\mathrm{tgt}} \in \mathcal{Y}$. Let q_τ be the input distribution induced by inserting τ into clean inputs $x' \sim p_X$ (the input marginal of p) and relabeling the target as y_{tgt}, with realization $x_\tau \sim q_\tau$. The trigger-conditional attack-success rate is the observable

$$\mathrm{ASR}_\tau(\theta) := \Pr_{x_\tau \sim q_\tau}[\hat{y}_\theta(x_\tau) = y_{\mathrm{tgt}}],$$

a paper-level functional of any checkpoint θ. We use ASR_τ in the slope theorem, empirical first-flip threshold, and cross-task scope stress tests; the empirical sweep producing $\{\theta_\varepsilon\}$ is stated in §4.3. When a target expert is fixed for slope prediction or ablation we write it e^*; the layer-wise real-MoE selection rule is in §4.3.

Related work and positioning. MoE attack and defense work (Wang et al., 2025a; Chan et al., 2025; Zhao et al., 2025; Wu et al., 2026; te Lintelo et al., 2026; Jiang et al., 2026; Kim et al., 2025; Lai et al., 2025; Liang et al., 2026) shows that routing can be attacked, silenced, or constrained. We do not claim the first MoE routing vulnerability; our focus is a mechanistic causal audit of adaptation-mediated

route exposure: a route identity present but non-mediating at θ_0, matched-placebo C-ablation after router-only adaptation, and clean-accuracy-preserving mediation tests. Behavioral mesa-optimization evaluations (Hubinger et al., 2019; 2024; Anthropic, 2024; Apollo Research, 2024) motivate train/deployment discrepancies. We contribute an architectural interpretation of that theme, without testing for internal search, planning, or strategic behavior. The linear-response diagnostic follows influence-function methodology (Koh & Liang, 2017; Bae et al., 2022; Park et al., 2023): PALR is the finite-step AdamW matched estimand for the response that a population implicit-function-theorem (IFT) calculation predicts only in the converged gradient-descent limit.

3. Mechanism

Before deriving the routed-objective identity, we make the mesa-optimization status operational. The following definition is a *proposed audit gate*, not a standard or calibrated alignment criterion: it operationalizes the narrow channel tested here—a trigger-conditioned behavior appears after router-only adaptation, remains hidden from an aggregate clean-task metric, and is removed by a route-specific intervention more than by matched placebo route disruptions.

Definition 3.1 ((η, δ, γ) structural pseudo-alignment channel)**.** A clean MoE checkpoint θ_0 has an (η, δ, γ) structural pseudo-alignment channel for trigger distribution q_τ if there exists a router-only perturbation $\Delta\phi$, with expert weights fixed and $\theta' = \theta_0 + \Delta\phi$. Writing Acc_p for clean accuracy on p, the channel must satisfy

$$\mathrm{ASR}_\tau(\theta') - \mathrm{ASR}_\tau(\theta_0) \geq \eta, \quad |\mathrm{Acc}_p(\theta') - \mathrm{Acc}_p(\theta_0)| \leq \delta,$$

and, for the trigger route mask C and matched placebo masks C_{plac},

$$\big[\mathrm{ASR}_\tau(\theta') - \mathrm{ASR}_\tau(C(\theta'))\big] - \max_{C_{\mathrm{plac}}}\big[\mathrm{ASR}_\tau(\theta') - \mathrm{ASR}_\tau(C_{\mathrm{plac}}(\theta'))\big] \geq \gamma.$$

We use *structural pseudo-alignment* as a mechanistic shorthand for this finite-intervention pattern: route-specific necessary mediation after router-only adaptation, with clean-task accuracy preserved under the audited metric. The term does not denote autonomous learned search, deception, scheming, or behavioral alignment in the ordinary sense; the θ_0 audit separates pre-existing route identity from output mediation, and route identity alone is not sufficient in our tested trigger families. We keep four notions distinct: *route identity* (a trigger enriches a selected expert under the $\Delta\pi$ rule), *route membership* (the expert is in the top-K active set), *output mediation* (removing or rewriting the selected route changes triggered target behavior), and *sufficiency* (forcing the route alone induces the behavior). Our evidence supports route-specific necessary mediation, not route-membership sufficiency.

This is a mesa-relevant structural claim, not a behavioral claim of learned search. The required pieces are supported as follows: Lemma 3.2 gives persistent expert-local effective objectives; Lemma 3.3 and Proposition 3.4 explain how small global contamination can become large routed exposure; Proposition 3.5 explains why clean/global metrics can miss the conditional conflict; and C-ablation tests causal control by routing. Autonomous behavioral mesa-optimization, deception, or scheming is *not claimed*. The θ_0 audit does not show deployment-shift-alone activation in our tested trigger families: route identity exists before contamination, but output mediation requires router-only adaptation.

3.1. Expert-local effective objective

Transformer-MoE notation. *Tokens and layers.* Inputs $x{=}(x_1, \ldots, x_T)$ are token sequences over vocabulary $\mathcal{V}$. Experts $\theta_{\ell e}$ live at MoE layer $\ell \in [L]$, expert $e \in [E]$. At token t, the router emits per-layer gate scores $g_{\ell e}(x_t; \theta)$ and produces hard top-K selection weights $a_{\ell te} \in [0, 1]$ ($a_{\ell te}{>}0$ iff $e{\in}\mathcal{E}_0(x_t; \theta)$, $\sum_e a_{\ell te}{=}1$). Each expert is a multilayer perceptron (MLP) $F_{\ell e}(u_{\ell t}; \theta_{\ell e})$ on its layer-input $u_{\ell t}$, with parameter Jacobian $J_{\theta_{\ell e}} F_{\ell e}{:=}\partial F_{\ell e}/\partial\theta_{\ell e}$. The MoE write into the residual stream is $m_{\ell t}{:=}\sum_e a_{\ell te} F_{\ell e}$; output logits at position s over $v \in \mathcal{V}$ are $z_{sv}(x; \theta)$, and $p_s(v \mid x){=}\mathrm{softmax}_v z_{sv}$. *Routing distributions.* The residual-stream cotangent $\kappa_{\ell t,s,v}{:=}\nabla_{m_{\ell t}} z_{sv}$ links MoE writes at (ℓ, t) to logits at (s, v). Let P_T be the joint training distribution over labeled sequences $(x, y_{1:T})$, with token-position marginal $D_T(x, t){:=}P_T(x){\cdot}\mathbf{1}[t{\in}[T]]/T$. The routing mass is $\mu_{\ell e}{:=}\mathbb{E}_{(x,t)\sim D_T}[a_{\ell te}]$ and the routed token-distribution at expert (ℓ, e) is $D^{\phi}_{\ell e}(x, t) \propto a_{\ell te} D_T(x, t)$. (Classifier setup recovered at $T{=}1$ with z_{sv} collapsing to per-class logits; cf. §2.)

Lemma 3.2 (Expert-local effective-objective gradient)**.** *Let $\mathcal{L}{:=}{-}\mathbb{E}_{(x,y_{1:T})\sim P_T}\sum_s \log p_s(y_s \mid x)$ denote the sequence negative log-likelihood (NLL), with y_s the target token at position s. Under (i) router gradients live in a subspace ϕ disjoint from expert parameters (so $a_{\ell te}$ a.e. independent of $\theta_{\ell e}$), (ii) the downstream map $m \mapsto z$ is differentiable, (iii) each expert enters the residual stream affinely (i.e., $m_{\ell t}{=}\sum_e a_{\ell te} F_{\ell e}$), the gradient on expert (ℓ, e) admits the routed-expectation form*

$$\nabla_{\theta_{\ell e}}\mathcal{L} = -\mu_{\ell e}\,\mathbb{E}_{D^{\phi}_{\ell e}}\big[(J_{\theta_{\ell e}} F_{\ell e})^{\top}\textstyle\sum_s \nabla_{m_{\ell t}} \log p_s(y_s \,|\, x)\big].$$

We treat $L^{\phi}_{\ell e}{:=}{-}\mathbb{E}_{D^{\phi}_{\ell e}}\log p$ as the expert-local effective objective: a directly computable proxy for the routed-distribution view. Because $D^{\phi}_{\ell e}$ is router-weighted by D_T, $D^{\phi,P}_{\ell e} \neq P_T$ already on clean training whenever $a_{\ell te}$ covaries

with input/token features. This supplies the persistence component of the mesa-optimization analogy: the expert-local objective is a pre-existing structural property rather than an artifact of contamination fine-tuning.

3.2. Effective contamination amplification

Lemma 3.3 (Effective contamination amplification (routed Bayes))**.** *Let $D_\varepsilon=(1-\varepsilon)P+\varepsilon Q$ be a labeled mixture, with P, Q the input marginals of the joint distributions p, q of §2, the router pinned at θ_0, and routing event $R_e:=\{x : e \in \mathcal{E}_0(x;\theta_0)\}$. Let $p_e:=\Pr_P[R_e]$, $q_e:=\Pr_Q[R_e]$, with $(1-\varepsilon)p_e+\varepsilon q_e>0$. Then the conditional contamination fraction among D_ε-samples that route to expert e is*

$$\alpha_e(\varepsilon) := \Pr_{D_\varepsilon}[\, x\sim Q \,|\, R_e \,] = \frac{\varepsilon\, q_e}{(1-\varepsilon)p_e+\varepsilon q_e}.$$

$\alpha_e \to 1$ as $q_e/p_e \gg (1-\varepsilon)/\varepsilon$ (rare-expert saturation).

This rare-expert amplification mechanism motivates the route-specific real-model audits in §5 and App. E.

3.3. Expert-local crossover threshold

We use the following algebraic crossover fact. For losses L_P, L_Q and $L_\alpha:=(1-\alpha)L_P+\alpha L_Q$, the limiting crossover is $\alpha^*:=\inf_\theta \Delta L_P/(\Delta L_P-\Delta L_Q)$ over contenders with $\Delta L_P>0, \Delta L_Q<0$; if the infimum is attained, it is the first attained mixture point. Applying this to L_e^α and inverting Lemma 3.3 gives a per-expert threshold.

Proposition 3.4 (Expert-local crossover threshold)**.** *Assume hard top-K routing with binary $a_{\ell te} \in \{0,1\}$ (so $\mu_{e,r}=p_e, q_e$). Let $L_{e,r}(\theta_e):=\mathbb{E}_{D^\phi_{r,e}}[\ell]$, $\theta_e^P=\arg\min L_{e,P}$ (assumed unique), and let $\alpha^*_{e,\mathrm{cross}} \in [0,1)$ be the local crossover applied to $L_e^\alpha=(1-\alpha)L_{e,P}+\alpha L_{e,Q}$. Inverting Lemma 3.3,*

$$\varepsilon_e^* = \frac{\alpha^*_{e,\mathrm{cross}}\, p_e}{q_e(1-\alpha^*_{e,\mathrm{cross}})+\alpha^*_{e,\mathrm{cross}}\, p_e}.$$

If $\alpha^*_{e,\mathrm{cross}}$ is bounded away from 1 uniformly as $p_e \to 0$ at fixed $q_e>0$, then $\varepsilon_e^* \to 0$: rare experts (small clean activation rate) can have arbitrarily small global contamination thresholds for local-objective switching.

The rare-expert corollary formalizes the empirical "dormant expert" vulnerability identified in concurrent attack work (Wang et al., 2025a; Chan et al., 2025) (BadPatches reports $\sim 0.01\%$ poisoning sufficient on pMoE). *Lemma 3.3 and Proposition 3.4 supply the conflict mechanism: when $q_e \gg p_e$, globally small contamination can dominate the routed conditional seen by a rare expert, and the expert-local crossover threshold can be much smaller than the global rate suggests.*

3.4. Misrouting masks conditional conflict

Proposition 3.5 (Misrouting masks conditional conflict)**.** *Let $\ell \in [0, B]$ be a bounded loss with bound $B>0$. Concretely, Y is the trigger-conditional regime q_τ from §2 (target label y_{tgt}) and X is the clean-task regime on which expert e specializes in pretraining. Fix the trigger-relevant input region $\Omega_\mathsf{Y} \subseteq \mathcal{X}$ (the support of q_τ). Define the misrouted set $S:=R_e \cap \Omega_\mathsf{Y}$ where $R_e=\{x : e \in \mathcal{E}_0(x;\theta_0)\}$. Let h be the model that sends every $x \in S$ to e (whose local rule is task-X-correct) and let h^Y be the counterfactual that routes inputs in S to a Y-correct expert. Let D be any data distribution over $\mathcal{X} \times \mathcal{Y}$, with risk $L_D(\cdot):=\mathbb{E}_{(x,y)\sim D}\ell(\cdot, y)$. If the per-input deficit $\Delta_{\mathsf{X}\to\mathsf{Y}}:=\mathbb{E}_D[\ell(h) \mid S]-\mathbb{E}_D[\ell(h^\mathsf{Y}) \mid S]=\Omega(1)$ but $\Pr_D[S] \ll 1$, then*

$$|L_D(h)-L_D(h^\mathsf{Y})| \le B \Pr_D(S),$$

$$\mathbb{E}_D[\ell(h)-\ell(h^\mathsf{Y}) \,|\, \Omega_\mathsf{Y}] \ge \Delta_{\mathsf{X}\to\mathsf{Y}} \Pr_D[R_e \,|\, \Omega_\mathsf{Y}].$$

Proposition 3.5 supplies the causal-routing component: the global metric remains nearly unchanged because the misrouted set is small, while the trigger-conditional behavior changes by $\Omega(1)$ because routing assigns those inputs to experts with incompatible local objectives. Composing with Lemma 3.3: contamination D_ε amplifies $\Pr_{D_\varepsilon}[R_e \cap \Omega_\mathsf{Y}]$ at rate $\alpha_e(\varepsilon)$, while the global mass change is at most $\varepsilon \Pr_Q[R_e \cap \Omega_\mathsf{Y}]$. The empirical counterpart is the C ablation of §4.3 (counterfactually mask e at every MoE layer on triggered inputs): the prediction is that ablation drops conditional ASR by $\Omega(1)$ while preserving clean accuracy.

4. Diagnostics

Given the routed-conditional channel of §3, three measurements diagnose its activation under contamination shift on a real model: (i) the **architectural threshold** (§4.1) gives the ideal hard-routing crossover under explicit assumptions and motivates a boundary-proximity layer-ranking proxy; (ii) the **trigger-conditional gate-probability slope** (§4.2) ranks which layers' router probabilities are most susceptible to small contamination; (iii) the **empirical first-flip and counterfactual-routing (C) ablation** (§4.3) measures the behavioral threshold and tests routing as the necessary mediator. Observational route statistics can nominate candidate mediators, but they cannot by themselves identify mediation; the causal claim comes from route interventions. The three quantities are linked diagnostics of the same routed-conditional structure: threshold theory explains when flips can occur, susceptibility ranks where they are likely, and C-ablation tests whether routing carries the failure.

4.1. Architectural threshold

Algebraic crossover and first hard-routing flip.

Lemma 4.1 (Algebraic crossover identity)**.** *Let* $\Delta_p(\theta){:=}L_p(\theta){-}L_p(\theta_0)$, $\Delta_q(\theta){:=}L_q(\theta){-}L_q(\theta_0)$, *and* $\mathcal{F}(\theta_0){:=}\{\theta : \Delta_p(\theta){>}0,\ \Delta_q(\theta){<}0\}$. *If* $\mathcal{F}(\theta_0) = \varnothing$, *no such crossover exists. Otherwise define the algebraic infimum*

$$\varepsilon^*_{\text{cross}}(\theta_0; p, q) := \inf_{\theta\in\mathcal{F}(\theta_0)} \frac{\Delta_p(\theta)}{\Delta_p(\theta){-}\Delta_q(\theta)} \in [0,1].$$

If the infimum is attained, it is the first attained mixture point; without attainment, $\varepsilon^*_{\text{cross}}$ *is only a limiting threshold.* ***Caveat (smooth case).*** *Write* $H_p{:=}\nabla^2 L_p(\theta_0)$ *and* $g_q{:=}\nabla L_q(\theta_0)$. *On smooth* L *with full-rank* H_p, *the infimum equals* 0 *(along* $\theta_0{-}tg_q$: $\Delta_p{=}O(t^2)$, $\Delta_q{=}O(t)$, *ratio* $\to 0$*); Lemma 4.1 provides no positive threshold by itself. Substantive content arises when the same identity is applied to the piecewise-constant hard-routing deployment risk* R_ε *(Theorem 4.2), where* Δ^R_p, Δ^R_q *vanish inside a routing cell and jump only at swap-margin surfaces.*

We apply the construction to R_ε, where the architectural discreteness of hard top-K routing makes $\varepsilon^*{>}0$ under explicit structural conditions:

Theorem 4.2 (Hard-routing crossover threshold = first flip, under kink-pinning)**.** *Let* $R_p(\theta){:=}\Pr_p[\hat{y}_\theta(x){\neq}y]$ *and* $R_q(\theta){:=}\Pr_q[\hat{y}_\theta(x){\neq}y]$, *so* $R_\varepsilon{=}(1{-}\varepsilon)R_p{+}\varepsilon R_q$. *Define* $\Delta^R_r(\theta){:=}R_r(\theta){-}R_r(\theta_0)$ *and* $\mathcal{F}^R{:=}\{\theta : \Delta^R_p{>}0,\ \Delta^R_q{<}0\}$. *Assume* $\mathcal{F}^R$ *is nonempty,* R_ε *has a piecewise-constant routing decomposition with isolated boundary cells, the kink at* θ_0 *is stable for* $\varepsilon{<}\varepsilon^*_{\text{cross,R}}$ *(kink-pinning; deterministic structural-existence sketch in App. A.5), and the crossover is transverse. Then* $\varepsilon^*_{\text{cross,R}}{:=}\inf_{\theta\in\mathcal{F}^R}\Delta^R_p/(\Delta^R_p{-}\Delta^R_q)$ *equals the first hard-routing flip threshold of* R_ε, *with closed form*

$$\varepsilon^*_{\text{cross,R}} = \frac{A^p}{A^p{+}A^q} \quad \textit{(single swap-eligible pair)},$$

and an analogous coupled-pair extension (App. A.2).

In the theorem's hard-routing setting, the $A^p/(A^p{+}A^q)$ form makes the architectural threshold a checkpoint-only quantity: given the clean-trained θ_0 and a candidate trigger distribution q, the threshold is fixed by the boundary-flux ratio at θ_0's routing boundary, with no fine-tuning required.

Worst-case lower bound. For the bounded-density adversary class $\mathcal{Q}_C{:=}\{q : 0{\le}dq/dp{\le}C\}$, $\varepsilon^* \ge 1/(1{+}C)$ (App. A.4); for $C \le 1{,}000$, $\varepsilon^* \ge 1/1001 \approx 9.99{\cdot}10^{-4}$ architecturally, independent of trigger construction. Weaker adversary classes based on total variation (TV), Kullback–Leibler divergence (KL), or Wasserstein distance admit no architecture-independent bound (counterexamples in App. A.4).

4.2. Optimizer-matched susceptibility

Monitoring gap and gate-probability slope. Theorem 4.2's deployment threshold is invisible to training-loss monitoring under smooth surrogate fine-tuning: along the IFT branch, the smooth optimum drifts continuously while the deployment optimum can stay pinned in the clean basin until the crossover.

Theorem 4.3 (Monitoring gap)**.** *Let* $L_{\text{train}}(\theta,\varepsilon){\equiv}L_\varepsilon(\theta)$ *be the smooth surrogate from §2,* C^3 *near* $(\theta_0, 0)$ *with positive-definite Hessian* $H{:=}\nabla^2 L_p(\theta_0)$ *and forcing* $g{:=}\nabla L_q(\theta_0){\neq}0$ *(equivalent to* $\nabla L_q{-}\nabla L_p$ *at the optimum since* $\nabla L_p(\theta_0){=}0$*; App. A.5 uses the latter form for* θ_0 *only approximately stationary). Suppose the deployment-risk basin minimizer drifts as* $\theta^*_{\text{deploy}}(\varepsilon){=}\theta_0{+}a\varepsilon{+}o(\varepsilon)$. *Then for small* ε,

$$\theta^*_{\text{train}}(\varepsilon) = \theta_0{-}H^{-1}g\,\varepsilon{+}O(\varepsilon^2),$$
$$\Delta_{\text{MG}}(\varepsilon) = \|{-}H^{-1}g{-}a\|_2\,\varepsilon{+}o(\varepsilon).$$

The kink-pinned special case ($a{=}0$, *App. A.1) recovers* $\Delta_{\text{MG}}(\varepsilon){=}\|H^{-1}g\|_2\,\varepsilon{+}O(\varepsilon^2)$.

The gap is a parameter-space quantity that is not directly observable; we convert it into a measurable per-layer signature via the trigger-conditional gate-probability slope. Let $\pi_e(x;\theta){:=}\text{softmax}_{e\in[E]}(g_e(x;\theta))$ be the soft router probability (a smooth surrogate for the hard top-K selection used at deployment), and write $\pi_0 := \mathbb{E}_{x_\tau}[\pi_{e^*}(x_\tau;\theta_0)]$ for the clean-checkpoint baseline.

Theorem 4.4 (Trigger-conditional gate-probability slope)**.** *Under the hypotheses of Theorem 4.3, with trigger inputs* $x_\tau \sim q_\tau$ *from the contamination protocol of §2 and a fixed target expert* $e^* \in [E]$ *(the empirical layer-wise selection rule used in real-MoE experiments is in §4.3); for the soft router probability* $\pi_{e^*}(\cdot;\theta)$ *differentiable at* θ_0,

$$\mathbb{E}_{x_\tau}\big[\pi_{e^*}(x_\tau;\theta^*_{\text{train}}(\varepsilon))\big] = \pi_0{+}s\,\varepsilon{+}O(\varepsilon^2),$$
$$s = -\mathbb{E}_{x_\tau}\big[\nabla_\theta \pi_{e^*}(x_\tau;\theta_0)\big]^\top H^{-1}g.$$

s *is computable from independently-estimated* $H^{-1}g$ *and* $\mathbb{E}_{x_\tau}[\nabla_\theta \pi_{e^*}]$*; equality of measured and predicted* s *on a small-*ε *sweep is a falsification protocol.*

Population vs. finite-optimizer susceptibility (matched estimand). Theorem 4.4 predicts the *population* susceptibility at the converged training optimum. Real measurements run a finite-step AdamW optimizer continuation from a benign fine-tuned checkpoint θ_0, which induces a different but matched-estimand object: the IFT-vs-retraining gap characterised by Bae et al. (2022); Schioppa et al. (2023); Koh & Liang (2017).

Proposition 4.5 (K-step optimizer linear response)**.** *Consider a deterministic differentiable optimizer map*

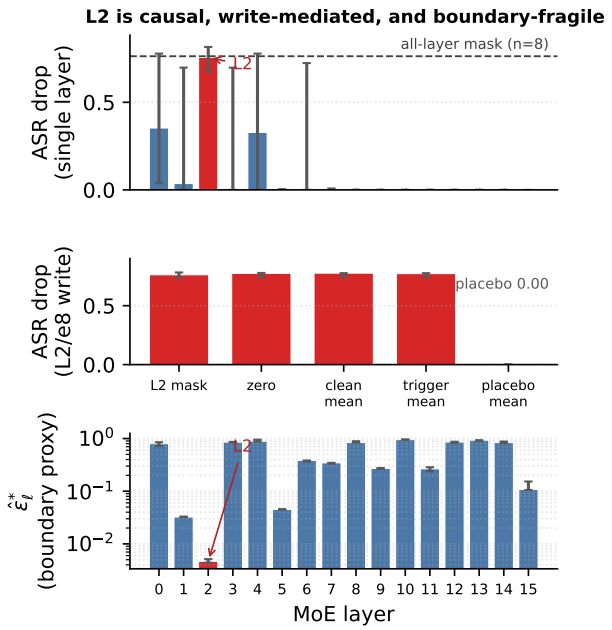

Figure 2. **Layer-2 (L2) cross-quantity coherence.** *Top*: single-layer C-ablation ASR drop (n=8 seed median/range); L2 nearly matches the all-layer mask and is the only layer above half the all-layer effect. *Middle*: routing-preserving layer-2/expert-8 (L2/e8) write interventions match the L2 mask, while mean-patch placebos are null. *Bottom*: independent θ_0 boundary proxy; L2 is the minimum in 3/3 seeds.

$\xi_{t+1}{=}A_t(\xi_t,\varepsilon)$ with optimizer-state vector $\xi_t{:=}(\theta_t,\mathsf{m}_t,\mathsf{u}_t)$ collecting parameters, AdamW first moment m_t, and AdamW second-moment running average u_t at step $t\in\{0,1,\dots,K_{\rm opt}\}$ (here $K_{\rm opt}$ is the optimizer horizon, distinct from the routing top-K of §2; $\mathsf{m}_t,\mathsf{u}_t$ are sansserif to distinguish from the residual-stream write $m_{\ell t}$ and layer-input $u_{\ell t}$ of §3). Take fresh initial state ξ_0 independent of ε, fixed batches/RNG/masks, and AdamW $\bar{\mathsf{u}}_{t,i}{>}0$. Then for any differentiable observable Φ_ℓ,

$$\tfrac{d}{d\varepsilon}\Phi_\ell(\theta_{K_{\rm opt}}(\varepsilon))\big|_{\varepsilon=0}=\nabla_\theta\Phi_\ell(\theta^0_{K_{\rm opt}})^\top d_{K_{\rm opt}},$$
$$\dot\xi_{t+1}=D_\xi A_t(\xi^0_t,0)\dot\xi_t+D_\varepsilon A_t(\xi^0_t,0),$$

with $\dot\xi_0{=}0$ and $d_t{=}P_\theta\dot\xi_t$ (P_θ projects the optimizer-state tangent onto the parameter coordinates). In the vanilla gradient-descent (GD) limit at a stationary θ_0, the recurrence reduces to $d_{K_{\rm opt}}{=}-[I-(I-\eta H)^{K_{\rm opt}}]H^{-1}g\to -H^{-1}g$ as $K_{\rm opt}\to\infty$, recovering Theorem 4.4.

The operational susceptibility for actual continuation fine-tuning measurement is the algorithmic linear response (ALR) quantity $s^{\rm ALR}_{\ell,K_{\rm opt}}{=}\nabla_\theta\Phi_\ell(\theta^0_{K_{\rm opt}})^\top d_{K_{\rm opt}}$; pathwise integration along $u\in[0,\varepsilon]$ (with u a contamination-axis dummy variable, distinct from the layer-input $u_{\ell t}$ of §3) matches the through-origin secant exactly (PALR, App. C). Several factors separate Theorem 4.4's population predictor from the continuation fine-tuning measurement: Tikhonov damping at $\lambda{\gg}h_{\rm med}$ in H^{-1} (with $h_{\rm med}$ the median Hessian eigenvalue magnitude), gradient evaluation at $\theta_{K_{\rm opt}}(0)$ vs. θ_0, AdamW per-coordinate preconditioning, and finite-$K_{\rm opt}$ trajectory effects (App. C).

4.3. Empirical first-flip and routing-mediation ablation

Contamination fine-tuning sweep protocol. Starting from a benign fine-tuned checkpoint θ_0, we run continuation fine-tuning under the weighted loss $L_\varepsilon{=}(1{-}\varepsilon)L_p{+}\varepsilon L_q$ for ε on a geometric grid (e.g. $\{0,10^{-3},3{\cdot}10^{-3},10^{-2},3{\cdot}10^{-2},10^{-1},0.3\}$), with paired deterministic batches matched to the slope-test protocol of §4.2. To isolate routing-mediated failure from expert-weight poisoning, we keep experts frozen and constrain training to a router LoRA subspace; this frozen-expert protocol is shared with SafeMoE (Kim et al., 2025), but used here for diagnostic measurement rather than safety-preservation. Each grid point ε produces a post-fine-tuning checkpoint θ_ε; the contamination distribution q is instantiated as the trigger-conditional q_τ from §2, and the observable $\mathrm{ASR}_\tau(\theta_\varepsilon)$ tracks the trigger response along the sweep. We evaluate clean accuracy on p and define the empirical first-flip threshold

$$\varepsilon^*_{\rm emp}:=\inf\{\varepsilon:\ \mathrm{ASR}_\tau(\theta_\varepsilon){>}\tau_{\rm ASR}\},$$

with detection threshold $\tau_{\rm ASR}{=}0.7$ as default (the subscript τ in ASR_τ and the threshold $\tau_{\rm ASR}$ are distinct uses of the symbol τ). This threshold avoids counting weak above-chance target preference as a successful trigger response.

Routing-mediation intervention (C ablation). The *C-ablation* (counterfactual-routing intervention; distinct from the bounded-density constant C of §4.1) tests whether the post-fine-tuning trigger response is mediated by the learned routing pattern. For each post-fine-tuning checkpoint, we identify the trigger-conditional top expert $e^*_\ell{:=}\arg\max_e(\mathbb{E}_{q_\tau}\pi_{\ell e}{-}\mathbb{E}_p\pi_{\ell e})$ at each MoE layer ℓ and install a forward hook that masks e^*_ℓ's router logits to $-\infty$. Both the trigger-ASR evaluation (on q_τ inputs) and the clean-accuracy evaluation (on p inputs) are run *under the same mask installed*: clean-accuracy preservation is therefore a non-trivial measurement (clean inputs see the mask applied), not a tautology. Proposition 3.5 provides the theoretical interpretation: when the misrouted set $S{=}R_{e^*}{\cap}\Omega_{\rm Y}$ (with $\Omega_{\rm Y}$ the trigger-relevant input region defined formally in §3.4) is small but the per-input deficit on S is $\Omega(1)$, removing the misrouting on S should drop the trigger-conditional risk by $\Omega(1)$ while leaving the global risk bounded by $B\ \Pr(S)$. The intervention is strong (simultaneous mask across all MoE layers); we report it as a *necessary-mediator* test, not a layer-local sufficiency claim. Real-model results are deferred to §5.2.

5. Validation

We validate the diagnostic stack of §4 along two axes: synthetic calibration (§5.1) on closed-form-checkable toys, and real-MoE evaluation (§5.2) on OLMoE-1B-7B and DeepSeek-MoE-16B. The synthetic calibration tests whether the threshold and slope predictors match the ground truth where it is available; the real-MoE evaluation probes whether the diagnostics behave consistently on a production sparse MoE under contamination fine-tuning, including cross-trigger and cross-task scope stress tests.

5.1. Synthetic calibration

Slope match. Theorem 4.4's prediction $s = -\nabla\pi^\top H^{-1} g$ matches the empirically-measured continuation-FT slope across 13 controlled toys: in 1-D Gaussian with σ_q=0.5 the closed form gives s=1.00 vs. measured 0.95 (5%, 5 seeds), and a 2-D 4-corner toy with two trainable parameters gives predicted s=0.9559 vs. measured 0.9546 (ratio 0.999, 0.14% relative error). The 2-D match validates that the matrix-valued $H^{-1}g$ and inner-product structure recover the 1-D scalar form.

Threshold match. Across the same 13 configurations (1-D Gaussian, 2-D 4-corner, off-axis, multimodal, asymmetric variants; full table in App. D), the predicted $\varepsilon^*_{\text{cross}}$ matches the observed first hard-routing flip to grid resolution (0.01) for every case except the symmetric coupled-pair c_q=0, where the corrected coupled-pair formula (App. A.2) closes the discrepancy.

Misrouting bound under frozen experts. A synthetic 4-expert toy with two opposing classification rules and analytic-perfect experts (linear router, frozen experts, router-LoRA contamination only) realizes the small-global / $\Omega(1)$-conditional shape of Proposition 3.5: clean global mean-squared error (MSE) on p *decreases* slightly across $\varepsilon \in \{0, 10^{-3}, \ldots, 0.3\}$ (0.14→0.10), while triggered-Y conditional risk *jumps* by +3.74 (0.18→3.93) and triggered ASR from 0.045 at the clean baseline to 1.000 from $\varepsilon \geq 10^{-3}$ onward (figure in App. D). Frozen experts rule out expert-weight poisoning; the failure is router-mediated exposure of e_X's pre-existing rule, exactly the C-ablation analogue under analytic-perfect experts.

5.2. Real-MoE evaluation

The main task is AGNews four-class news classification. IMDB sentiment classification and an AdvBench/Alpaca refusal-versus-compliance pilot are used only as cross-task stress tests.

Boundary-proximity layer-ranking proxy on OLMoE. At the benign fine-tuned checkpoint θ_0, we estimate the top-1/top-2 router-margin density on N=200 clean and N=200 triggered AGNews inputs (App. E) and form $\hat{\varepsilon}^*_\ell = \hat{A}^p_\ell / (\hat{A}^p_\ell + \hat{A}^q_\ell)$. This per-layer aggregate is not the swap-pair-specific $A^r_{e,e'}$ of §2; it is a boundary-proximity heuristic. At $\epsilon_{\text{reg}}=10^{-3}$, L2 is the absolute minimum in 3/3 seeds, $\hat{\varepsilon}^*_{L2} \in \{0.0051, 0.0044, 0.0046\}$ (median 0.0046), while twelve interior layers exceed 0.20 (median 0.55). Sweeping $\epsilon_{\text{reg}} \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ shifts absolute values by $> 100\times$ but preserves L2-as-minimum in every seed. We therefore use the proxy only as a tail-layer ranking, not as a calibrated threshold.

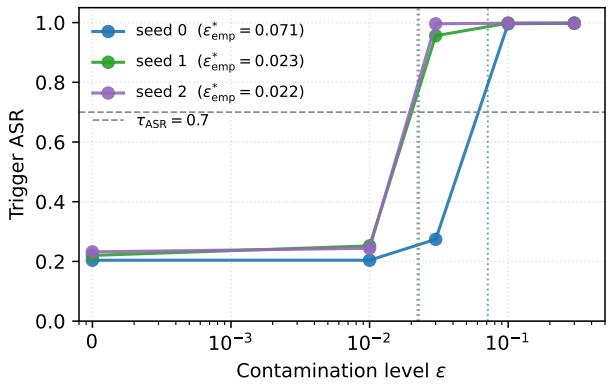

Figure 3. **Discrete first-flip on OLMoE-1B-7B.** Trigger ASR vs. ε across n=3 seeds (router-LoRA only, frozen experts; ε=0 plotted at left edge of log axis). Dotted lines: per-seed $\varepsilon^*_{\text{emp}} \in \{0.022, 0.023, 0.071\}$ (linearly interpolated at τ_{ASR}=0.7).

Optimizer-matched susceptibility on OLMoE and DeepSeek-MoE. PALR is the matched-estimand finite-step AdamW correction to the population IFT predictor (App. C). On OLMoE-1B-7B at the long-fine-tuning/low-learning-rate regime (1000 steps, learning rate 10^{-4}, 3 seeds), PALR is positive in 3/3 seeds against measured continuation slopes (pooled r=+0.566, n=48; per-seed shuffle $p \in \{0.005, 0.048, 0.004\}$), while bare IFT is at chance on 2/3 seeds. On DeepSeek-MoE-16B (Dai et al., 2024), per-seed Pearson $r \in \{0.747, 0.809, 0.646\}$ (mean $+0.734 \pm 0.067$), pooled r=+0.703 (n=81), with every seed at $p < 10^{-3}$ under layer shuffles. PALR reduces the population-IFT median fold error from 32–112× to 1.9–7.2× across regimes; short-fine-tuning directions remain sign-unstable, so magnitude calibration is the robust contribution.

Empirical first-flip and routing-mediation ablation on OLMoE. We run a contamination fine-tuning sweep over $\varepsilon \in \{0, 10^{-2}, 3{\cdot}10^{-2}, 10^{-1}, 0.3\}$ on OLMoE-1B-7B (router-LoRA only, frozen experts) for n=3 seeds. Trigger-conditional τ_{ASR}=0.7 with linear interpolation gives empirical first-flip thresholds $\varepsilon^*_{\text{emp}} \in \{0.022, 0.023, 0.071\}$ (median 0.023); clean accuracy preserved at 0.866–0.912.

Primary causal result: C-ablation + placebo. Expanding the all-layer C-ablation to $n{=}8$ seeds, masking e^*_ℓ at every MoE layer decreases median ASR by 0.774 at $\varepsilon \in \{0.1, 0.3\}$ (saturated, baseline median 1.000 to 0.224) and 0.764 at $\varepsilon{=}0.03$ (just-above first-flip, $0.988{\to}0.218$); clean accuracy is preserved within ± 0.02. The matched-usage placebo control ($R{=}5$ non-trigger-conditional masks per layer, top-$K{=}8$ $\Delta\pi$-excluded, ranked by usage proximity to e^*_ℓ) decreases worst-case median ASR by 0.002 at $\varepsilon \in \{0.1, 0.3\}$ and 0.070 at $\varepsilon = 0.03$. Median specificity (C-ablation decrease minus worst-placebo decrease) is $\{0.694, 0.772, 0.772\}$ across the three ε levels, well above the 0.30 strong-for threshold; every seed remains above 0.30 at every reported ε, although occasional high-placebo rotations motivate the worst-placebo accounting. This establishes selected-route necessity under the tested finite intervention family and argues against saturation/arbitrary-mask explanations across the $10\times$ range.

Supportive: localization + cross-quantity coherence at L2. Expanding single-layer masks to $n{=}8$ seeds at $\varepsilon{=}0.1$, L2's e^*_{L2} is the dominant behavioral bottleneck: median ASR drop 0.756, 99% of the all-layer-mask drop 0.763, and the only layer above half the all-layer effect (Fig. 2). Routing-preserving L2/e8 output interventions match the L2 router-mask drop: zeroing gives median ASR drop 0.768, and clean-/trigger-mean patches give $0.776/0.772$ (placebo mean patches 0.000). A token-scoped L2/e8 mask localizes the same necessity to the trigger span: trigger-subtoken masking drops median ASR by 0.756 in $3/3$ seeds, while final-token and non-trigger-context masks are null (App. E). Secondary seed-dependent effects remain visible, so this is a near-saturating write-level bottleneck rather than a complete circuit decomposition. L2 is also the boundary-flux minimum (§5.2, $\hat{\varepsilon}^*_{L2}{\approx}0.0046$); minimum-flux layer and dominant single-layer ablation layer coincide.

Diagnostic boundary: routing-membership is not sufficient. A clean-input routing-patching test ($R{=}10$ placebo, force e^*_ℓ into top-K at every MoE layer) yields median target-rate rise $+0.002$ at $\varepsilon{=}0.1$ and -0.002 at $\varepsilon{=}0.03$ (placebo max $\leq +0.010$). The asymmetry locates the target signal in the joint computation of trigger-bearing residual $\times$ routing through e^*, not in routing-membership alone.

Stress tests: cross-trigger and cross-task ($n{=}3$ each, $\varepsilon{=}0.1$). Mediation specificity transfers strongly to AGNews with a natural phrase trigger (median $+0.528$), partially to IMDB with the random-string trigger (per-seed $+0.060$–$+0.438$), and is non-passing on the AdvBench/Alpaca pilot, consistent with corpus-format separability allowing the binary refusal/comply decision to bypass this routing channel (§6; full pilot table in App. E). For the random-string trigger "ZxQwert", the L2 route identity is stable at $e^*_{L2}{=}8$ across AGNews, IMDB, and AdvBench/Alpaca; for natural "BREAKING" on AGNews it changes across seeds.

Boundary: θ_0 audit isolates routing coordination as the learned object. A no-contamination audit at the benign fine-tuned checkpoint θ_0 decomposes route identity from output mediation. Route identity is present: the random-string trigger reproducibly identifies $e^*_{0,L2}{=}8$ across $3/3$ seeds, and L2/e8 hard membership is high on trigger subtokens (0.80 at θ_0/post-fine-tuning), lower on other article tokens ($0.32/0.35$), and near zero at the final token ($\leq 0.01/0.006$). Output mediation is absent: θ_0 C-ablation changes triggered target rate by at most 0.004; membership-mode forced routing through e^*_0 or the post-contamination route raises target rate by at most 0.012 with placebo specificity ≤ 0.008; full post-fine-tuning membership composition is also non-mediating. Lexical controls rule out a "BREAKING"-specific activation key. Thus both trigger families fail to show output mediation in $3/3$ seeds under the no-adaptation criterion; ZxQwert is positive for route identity, not target mediation.

6. Limitations and conclusion

Interpreted via Definition 3.1, the evidence supports structural pseudo-alignment as route-specific output mediation after adaptation, not learned search. Two negative controls sharpen this reading. If the ASR drop were generic route damage, matched-usage placebo masks should have comparable effect; they do not. If route identity alone were sufficient, forcing the target route at θ_0 or on clean inputs should raise the target rate; it does not. The result is therefore a causal routing-mediation claim, not only a behavior-level trigger observation.

The confirmed causal-mediation package is strongest on OLMoE: eight-seed all-layer C-ablation, L2 localization/write interventions, trigger-span token masking, and a θ_0 route-identity/non-mediation split. DeepSeek adds cross-model PALR susceptibility, but a second-architecture C-ablation success is not yet established; pilots are implementation-sensitive scope controls (App. E), not evidence that other MoEs lack the channel. IMDB transfer is mixed, AdvBench/Alpaca is non-passing, and cross-architecture C-ablation remains future work. The boundary-flux proxy is a regularizer-sensitive layer-ranking heuristic; PALR predicts finite-optimizer gate-probability slopes, not hard first flips. The all-layer mask is necessary-mediator evidence, not sufficiency, and deployment-shift-alone activation is not observed in $3/3$ seeds. We report aggregate clean accuracy (preserved within ± 0.02) but did not run a per-class clean-accuracy audit, so class-specific clean degradation masked by the aggregate metric remains possible. Still, routing geometry is alignment-relevant: clean training creates routed local objectives, and router-only adaptation can make a pre-existing route identity output-mediating without changing

expert weights.

Operationally, the audit is a gate rather than a scalar score. A mesa-relevant routed channel must show conditional behavior under router-only adaptation, lose it under trigger-enriched route intervention, survive placebo and clean-accuracy controls, and distinguish route identity from output mediation at θ_0. Mixed pilots constrain our claims but do not show the channel is absent in other prompts, interventions, or MoE families.

The practical lesson is that MoE safety audits should treat routing as an intervention target, not only as a descriptive statistic. Gate-probability shifts, PALR slopes, and boundary-proximity scores are useful triage signals, but they become evidence only when paired with counterfactual route masks, clean-accuracy controls, and usage-matched placebo routes. We provide a reproducibility bundle with plotting scripts, seed-level JSON artifacts, and analysis scripts; all models and task data use public checkpoints/datasets, and model weights are not redistributed. This keeps the mesa-optimization interpretation tied to observable causal structure: the conflict is internal and persistent, yet the behavioral claim is bounded by whether the route mediates the triggered output.

Impact Statement

This work studies sparse-MoE routing failures to improve interpretability, auditing, and safety monitoring. Because such diagnostics could aid trigger construction, we report bounded classification experiments, emphasize mitigation, and release artifacts for defensive analysis.

Acknowledgements

This project is partially funded by the UK AISI Alignment Fund (grant number AP-S2-100082). JR is partially supported by a Vannevar Bush Faculty Fellowship (ONR N000142312863).

References

Anthropic. Alignment faking in large language models, 2024. URL https://www.anthropic.com/research/alignment-faking.

Apollo Research. Frontier models are capable of in-context scheming, 2024. URL https://www.apolloresearch.ai/research/scheming-reasoning-evaluations.

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chan, C., te Lintelo, J., and Picek, S. Badpatches: Routing-aware backdoor attacks on vision mixture of experts, 2025.

Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y. K., Huang, P., Zhao, F., Sun, S., Wang, M., Zhang, D., and Liang, W. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models, 2024.

Elhage, N. et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Ghorbani, B., Krishnan, S., and Xiao, Y. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning (ICML)*, 2019.

Hoogland, J., Wang, G., Farrugia-Roberts, M., Carroll, L., Wei, S., and Murfet, D. Loss landscape degeneracy and stagewise development in transformers, 2024. Transactions on Machine Learning Research (TMLR).

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems, 2019.

Hubinger, E. et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024.

Jiang, A. Q. et al. Mixtral of experts, 2024.

Jiang, Y., Huang, H., Li, M., Zhang, Y., Backes, M., and Zhang, Y. Sparse models, sparse safety: Unsafe routes in mixture-of-experts LLMs, 2026.

Kim, J., Song, M., Shin, S., and Son, S. Defending MoE LLMs against harmful fine-tuning via safety routing alignment, 2025.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Kwon, Y., Wu, E., Wu, K., and Zou, J. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Lai, Z., Liao, M., Wu, B., Xu, D., Zhao, Z., Yuan, Z., Fan, C., and Li, J. SAFEx: Analyzing vulnerabilities of MoE-based LLMs via stable safety-critical expert identification, 2025. NeurIPS 2025 poster.

Liang, J., Wang, Y., Jiang, T., and Wang, T. RASA: Routing-aware safety alignment for mixture-of-experts models, 2026.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation, 2020.

Muennighoff, N. et al. OLMoE: Open mixture-of-experts language models, 2024.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.

Schioppa, A., Filippova, K., Titov, I., and Zablotskaia, P. Theoretical and practical perspectives on what influence functions do. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

te Lintelo, J., Wu, L., and Picek, S. Large language lobotomy: Jailbreaking mixture-of-experts via expert silencing, 2026.

Wang, Q., Pang, Q., Lin, X., Wang, S., and Wu, D. BadMoE: Backdooring mixture-of-experts LLMs via optimizing routing triggers and infecting dormant experts, 2025a.

Wang, Y., Xu, Y., Shen, N., Su, J., Huang, J., and Zhu, Z. The illusion of specialization: Unveiling the domain-invariant standing committee in mixture-of-experts models, 2026. ACL 2026.

Wang, Z., Zhu, J., and Chen, J. ReMoE: Fully differentiable mixture-of-experts with ReLU routing, 2025b. ICLR 2025.

Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Wu, L., Behrouzi, S., Rostami, M., Picek, S., and Sadeghi, A.-R. GateBreaker: Gate-guided attacks on mixture-of-expert LLMs, 2026. USENIX Security 2026.

Zhao, X., Chen, X., Liu, B., Gao, H., Zhao, Z., and Chen, Y. Who speaks for the trigger? dynamic expert routing in backdoored mixture-of-experts transformers, 2025. NeurIPS 2025 poster.

Zheng, C., Huang, W., Wang, R., Wu, G., Zhu, J., and Li, C. On mesa-optimization in autoregressively trained transformers: Emergence and capability, 2024. NeurIPS 2024.

A. Theory: detailed proofs and constructions

A.1. Bifurcation regime: 1-D Gaussian closed form

We instantiate Theorem 4.2 on a 1-D Gaussian toy. Consider $p{=}\mathcal{N}(0,1)$, $q{=}\mathcal{N}(0,\sigma_q^2)$ with $\sigma_q{<}1$, a single trainable boundary b in a 2-expert MoE under hard top-1 routing with saturated experts $f_1{=}0$, $f_2{=}1$, and labels $y_p{=}\mathbb{1}[x{>}0]$ on clean data, $y_q{=}1$ on contamination. The hard $0/1$ deployment risk is

$$R_\varepsilon(b) = (1{-}\varepsilon)\Pr_{x\sim p}\big[\mathbb{1}\{x{>}b\}{\neq}\mathbb{1}\{x{>}0\}\big] + \varepsilon\Pr_{x\sim q}\big[\mathbb{1}\{x{>}b\}{\neq}1\big]. \tag{1}$$

Proposition A.1 (1-D first-flip threshold)**.** *For $\sigma_q{<}1$, the deployment-risk minimizer of (1) satisfies $b^*(\varepsilon){=}0$ for $\varepsilon{<}\varepsilon^*_{\text{nonlocal}}$ and moves to the interior $b^*(\varepsilon){<}0$ for $\varepsilon{>}\varepsilon^*_{\text{nonlocal}}$, where*

$$\varepsilon^*_{\text{nonlocal}} = \frac{p(0)}{p(0){+}q(0)} = \frac{\sigma_q}{1{+}\sigma_q}.$$

The interior minimizer satisfies the first-order condition $(1{-}\varepsilon)p(b^*){=}\varepsilon q(b^*)$ and grows as $b^*(\delta){=}{-}\Theta(\sqrt{\delta})$ for $\delta{=}\varepsilon{-}\varepsilon^*_{\text{nonlocal}}{\to}0^+$ (saddle-node bifurcation).

Proof. For $b{<}0$, the clean error in (1) is $\Pr_p[b{<}x{<}0]{=}\Phi(0){-}\Phi(b)$ and the contamination error is $\Pr_q[x{<}b]{=}\Phi(b/\sigma_q)$. Differentiating in b gives

$$\partial_b R_\varepsilon(b) = -(1{-}\varepsilon)p(b) + \varepsilon q(b), \qquad b{<}0. \tag{2}$$

The boundary $b{=}0$ is a kink-stable left-endpoint of the descent direction iff $\partial_b R_\varepsilon(0^-){\le}0$, i.e. $(1{-}\varepsilon)p(0){\ge}\varepsilon q(0)$. Substituting $p(0){=}1/\sqrt{2\pi}$ and $q(0){=}1/(\sigma_q\sqrt{2\pi})$ gives the stated threshold $\sigma_q/(1{+}\sigma_q)$.

For $\varepsilon{>}\varepsilon^*_{\text{nonlocal}}$, the kink stability fails and the minimizer moves to the interior, where the first-order condition is $\partial_b R_\varepsilon(b^*){=}0$, i.e. $(1{-}\varepsilon)p(b^*){=}\varepsilon q(b^*)$. To extract the $\sqrt{\delta}$ scaling near the bifurcation, set $\delta{=}\varepsilon{-}\varepsilon^*_{\text{nonlocal}}$ and Taylor-expand the first-order condition around $b{=}0$:

$$(1{-}\varepsilon)\big[p(0){+}p'(0)b^*{+}\tfrac{1}{2}p''(0)(b^*)^2\big] = \varepsilon\big[q(0){+}q'(0)b^*{+}\tfrac{1}{2}q''(0)(b^*)^2\big]{+}o((b^*)^2).$$

Both densities are even, so $p'(0){=}q'(0){=}0$. The constant terms vanish at $\varepsilon{=}\varepsilon^*_{\text{nonlocal}}$ by definition; perturbing to $\varepsilon^*{+}\delta$ gives a leading $\delta\cdot[p(0){+}q(0)]$ at zeroth order in b^*, balanced by $\frac{1}{2}[(1{-}\varepsilon)p''(0){-}\varepsilon q''(0)](b^*)^2$. Solving yields $(b^*)^2\sim\delta$, i.e. $b^*{=}{-}\Theta(\sqrt{\delta})$. □

Remark A.2. For $\sigma_q{\ge}1$, the pointwise stability condition is replaced by the integral condition $\int_b^0[(1{-}\varepsilon)p{-}\varepsilon q]\,dx{\ge}0$ for all $b{\le}0$, with the binding b determining the threshold. The 1-D Gaussian case is the simplest instance of the general bifurcation regime; under the structural conditions of Theorem 4.2, it saturates the Maxwell construction at $\varepsilon^*_{\text{nonlocal}}$.

A.2. Coupled-pair sum form

A shared parameter θ controlling multiple swap-eligible boundaries simultaneously couples their constraints; applying the per-pair threshold $A^p/(A^p{+}A^q)$ to each pair separately undercounts the constraint and produces an incorrect (too small) threshold.

Proposition A.3 (Coupled-pair threshold)**.** *Let $\theta{\to}\theta{+}\delta$ induce displacements $b_i(\theta{+}\delta){=}b_i(\theta){+}\alpha_i\delta$ on the swap-eligible boundaries $\{S_i\}_{i\in I_\theta}$ controlled by θ, with same-sign couplings $\alpha_i{>}0$. Suppose each S_i has signed boundary trace $A_i^r{=}\int_{S_i\cap R_i} r(x)/\|\nabla g\|\,d\mathcal{H}_{d-1}$ for $r\in\{p,q\}$. Then the coupled threshold is*

$$\varepsilon^*_{I_\theta} = \frac{\sum_{i\in I_\theta}\alpha_i A_i^p}{\sum_{i\in I_\theta}\alpha_i A_i^p + \sum_{i\in I_\theta}\alpha_i A_i^q}.$$

Proof. The first-order increment of the deployment risk along δ is the sum of per-boundary contributions

$$\partial_\delta R_\varepsilon\big|_{\delta=0} = \sum_{i\in I_\theta}\alpha_i\big[\varepsilon A_i^q{-}(1{-}\varepsilon)A_i^p\big],$$

where we have used the boundary-flux representation $\partial_\delta \Delta_r^R|_{\delta=0}=\pm\alpha_i A_i^r$ for piecewise-constant routing. The kink at θ is stable iff $\partial_\delta R_\varepsilon|_{0^+}\geq 0$, equivalently

$$\sum_i \alpha_i(1-\varepsilon)A_i^p \geq \sum_i \alpha_i \varepsilon A_i^q,$$

which rearranges to $\varepsilon\leq\varepsilon^*_{I_\theta}$ as stated. □

For uniform same-sign couplings $\alpha_i\equiv 1$, the threshold reduces to the pure sum form $\sum A_i^p/(\sum A_i^p+\sum A_i^q)$. For independent pairs ($|I_\theta|=1$), Proposition A.3 reduces to the per-pair $A^p/(A^p+A^q)$ of Theorem 4.2, and the global threshold across multiple independent parameters is $\min_\theta \varepsilon^*_{I_\theta}$ (the binding direction).

A.3. Margin-distribution form for discrete text

The input-space surface notation in §2 is most natural for continuous toy models. For text, the directly observed object is a scalar router margin distribution. For an audited swap pair $i=(\ell,t,e,e')$, define

$$M_i(x;\theta) := g_{\ell e}(x_t;\theta) - g_{\ell e'}(x_t;\theta), \qquad B_i^v(x) := \nabla_\theta M_i(x;\theta_0)^\top v$$

for router-parameter direction v. Along $\theta(a)=\theta_0+av$, $M_i(x;\theta(a)) = M_i(x;\theta_0) + aB_i^v(x) + O(a^2)$, so the first-order hard-routing risk flux can be written distributionally as

$$\Gamma_i^r(v) = \mathbb{E}_{(x,y)\sim r}[\Delta\mathrm{err}_i(x,y)\, B_i^v(x)\, \delta(M_i(x;\theta_0))]\,.$$

If the induced real-valued random variable $M_i(X;\theta_0)$ has density $f^r_{M_i}$ near zero, this is equivalently

$$\Gamma_i^r(v) = f^r_{M_i}(0)\, \mathbb{E}_r[\Delta\mathrm{err}_i(x,y)\, B_i^v(x) \mid M_i(x;\theta_0)=0]\,.$$

For directions with $\Gamma_i^p(v)>0$ and $\Gamma_i^q(v)<0$, the directional threshold is

$$\varepsilon_i^*(v) \;=\; \frac{\Gamma_i^p(v)}{\Gamma_i^p(v)-\Gamma_i^q(v)}.$$

The main-text $A^p/(A^p+A^q)$ formula is the special case $A^p=\Gamma^p$ and $A^q=-\Gamma^q$ with a fixed favorable swap direction. Our real-MoE boundary proxy in §5.2 estimates a coarsened version of this margin-density-at-zero object using the top-1/top-2 router margin and a Gaussian kernel, while dropping the loss-jump and directional-motion factors; this is why it is reported as a layer-ranking heuristic, not a calibrated threshold.

A.4. L^∞ worst-case bound and weaker-class counterexamples

For a clean distribution p on $\mathcal{X}$, define the L^∞ likelihood-ratio class $\mathcal{Q}_C^{L^\infty}:=\{q: 0\leq dq/dp\leq C \text{ a.e.}\}$, the total variation ball $\mathcal{Q}_\eta^{\mathrm{TV}}:=\{q: \|q-p\|_{\mathrm{TV}}\leq\eta\}$, the KL ball $\mathcal{Q}_\kappa^{\mathrm{KL}}:=\{q:\mathrm{KL}(q\|p)\leq\kappa\}$, and the Wasserstein-1 ball $\mathcal{Q}_\delta^{W_1}:=\{q: W_1(q,p)\leq\delta\}$.

Proposition A.4 (L^∞ worst-case lower bound)**.** *Under the hypotheses of Theorem 4.2 and absolutely continuous boundary traces,*

$$\inf_{q\in\mathcal{Q}_C^{L^\infty}} \varepsilon^*_{\mathrm{nonlocal}}(\theta_0;p,q) \;\geq\; \frac{1}{1+C}.$$

Proof. Let $S_{e,e'}\cap R_{e,e'}$ be a swap-eligible boundary segment, and let $B_\tau:=\{x: |g_e-g_{e'}|(x)\leq\tau\}\cap R_{e,e'}$ be the τ-tube around it. By the boundary-trace definition, $A^r(S_{e,e'})=\lim_{\tau\downarrow 0} r(B_\tau)/(2\tau)$ for $r\in\{p,q\}$. For $q\in\mathcal{Q}_C^{L^\infty}$ with $dq/dp\leq C$ a.e., monotonicity of integration on B_τ gives

$$q(B_\tau) = \int_{B_\tau} (dq/dp)\, dp \leq C\, p(B_\tau), \tag{3}$$

and dividing by 2τ and taking the limit yields $A^q(S_{e,e'})\leq C\, A^p(S_{e,e'})$. Consequently the per-pair threshold satisfies

$$\frac{A^p}{A^p+A^q} \;\geq\; \frac{A^p}{A^p+C\,A^p} = \frac{1}{1+C}.$$

For coupled pairs (Proposition A.3) with same-sign coupling weights $\alpha_i\geq 0$, applying (3) term-by-term gives $\sum\alpha_i A_i^q\leq C\sum\alpha_i A_i^p$, so $\varepsilon^*_{I_\theta}\geq 1/(1+C)$. Taking the infimum over directions in the directional-ratio formula preserves the lower bound. □

Proposition A.5 (Sharpness of $1/(1{+}C)$)**.** *For any architecture admitting at least one swap-eligible direction, there exists $q \in \mathcal{Q}_C^{L^\infty}$ attaining $\varepsilon^*_{\text{nonlocal}}{=}1/(1{+}C)$.*

Proof. Pick a swap-eligible pair $(e^*, e^{*\prime})$ and a τ-tube B^*_τ around the boundary $S_{e^*,e^{*\prime}} \cap R_{e^*,e^{*\prime}}$. Let $b_\tau{:=}p(B^*_\tau)$ and choose τ small enough that $C\, b_\tau{<}1$ (such τ exists because $b_\tau{\to}0$ as $\tau{\downarrow}0$). Define

$$r_\tau(x) = \begin{cases} C, & x \in B^*_\tau, \\ \beta_\tau, & x \notin B^*_\tau, \end{cases} \quad \text{where} \quad \beta_\tau := \frac{1{-}C\, b_\tau}{1{-}b_\tau} \in [0,1].$$

By construction $0{\le}r_\tau{\le}C$ everywhere; $\int r_\tau\, dp{=}C\, b_\tau{+}\beta_\tau(1{-}b_\tau){=}1$, so $q_\tau{:=}r_\tau\, p$ is a probability distribution; and $dq_\tau/dp{=}r_\tau{\le}C$ a.e., so $q_\tau \in \mathcal{Q}_C^{L^\infty}$. For any sub-tube $B_s{\subset}B^*_\tau$, $q_\tau(B_s){=}C\, p(B_s)$, so $A^{q_\tau}(S_{e^*,e^{*\prime}}){=}C\, A^p(S_{e^*,e^{*\prime}})$. Inserting in the per-pair threshold gives $A^p/(A^p{+}A^{q_\tau}){=}1/(1{+}C)$. The infimum over $\mathcal{Q}_C^{L^\infty}$ thus attains $1/(1{+}C)$. The construction extends to coupled directions by setting $r_\tau{=}C$ on $\cup_{i\in I}B_{i,\tau}$ and redistributing β_τ off the union to keep $\int r_\tau\, dp{=}1$. □

Proposition A.6 (No architecture-independent bound for TV, KL, W_1)**.** *For any $\eta, \kappa, \delta{>}0$, $\inf_{q\in\mathcal{Q}_\eta^{\text{TV}}} \varepsilon^*_{\text{nonlocal}}(\theta_0; p, q){=}0$, and similarly for $\mathcal{Q}_\kappa^{\text{KL}}$ and $\mathcal{Q}_\delta^{W_1}$.*

Proof. **TV.** Let $x_0 \in S_{e,e'} \cap R_{e,e'}$ and set $q{=}(1{-}\eta)p{+}\eta\delta_{x_0}$. Then $\|q{-}p\|_{\text{TV}}{=}\eta$, so $q \in \mathcal{Q}_\eta^{\text{TV}}$. The boundary trace satisfies

$$A^q(S_{e,e'}) = \lim_{\tau\downarrow 0} \frac{(1{-}\eta)p(B_\tau){+}\eta}{2\tau} = +\infty$$

since the Dirac contribution does not vanish under $\tau{\downarrow}0$. Hence $\varepsilon^*_{e,e'}{=}A^p/(A^p{+}A^q){\to}0$. The construction extends to absolutely-continuous adversaries by replacing δ_{x_0} with a smooth bump ρ_τ of TV mass η supported on B_τ; as $\tau{\downarrow}0$ at fixed η, $A^{q_{\tau,\eta}}{=}\eta/(2\tau){\to}\infty$.

KL. Choose ρ_τ supported on B_τ with density $\rho_\tau{=}p\mathbb{1}_{B_\tau}/p(B_\tau)$ (so $\rho_\tau/p{=}1/p(B_\tau)$ on B_τ). Then $\text{KL}(\rho_\tau\|p){=}{-}\log p(B_\tau){=}\Theta(\log(1/\tau))$. Setting $\eta_\tau{=}\kappa/\log(1/\tau)$, the convex combination $q{=}(1{-}\eta_\tau)p{+}\eta_\tau\rho_\tau$ satisfies $\text{KL}(q\|p){\le}\kappa$ for τ small (by convexity of KL in its first argument), while $A^q{=}\eta_\tau/(2\tau){\to}\infty$ since $\eta_\tau/\tau{\to}\infty$.

W_1**.** For any $\delta{>}0$, redistribute mass to amplify B_τ by factor $C(\tau){=}\delta/\tau$, with the remainder of mass moved away to compensate. The Wasserstein cost of this redistribution is $\Theta(C(\tau)\cdot\tau){=}\Theta(\delta)$, so $W_1{\le}\delta$. But $A^q{=}C(\tau)A^p/\tau{\to}\infty$ as $\tau{\downarrow}0$. □

Among the four classes, only the L^∞ likelihood-ratio class admits an architecture-independent positive lower bound on $\varepsilon^*_{\text{worst}}$, justifying our choice of $\mathcal{Q}_C^{L^\infty}$ in Theorem 4.2's worst-case statement. The W_∞ class admits an architecture-*dependent* bound (depending on local boundary geometry), which we omit from the main statement.

A.5. Audited-region kink-pinning: structural existence sketch

Theorem 4.2 states kink-pinning at θ_0 as a structural assumption. We sketch a deterministic existence result: under standard regularity, audited swap signs on a high-mass region are pinned for ε below a computable radius. We do *not* give a finite-sample certificate with a tight rate; the dimension-and-sample-size constants for OLMoE-1B-7B (router-LoRA $d{\sim}10^5$, $N{=}200$ paired samples) make any covariance-operator-norm-based concentration effectively vacuous, and the deterministic core stands without that layer.

Proposition (audited-region kink-pinning, deterministic existence). Fix a router-LoRA subspace and a finite audited set I of layer-token swap pairs $i{=}(\ell, t, e, e')$, with $m_i(x;\theta){:=}g_{\ell e}(x_t;\theta){-}g_{\ell e'}(x_t;\theta)$ the signed swap margin at the audited pair (gates evaluated at the token x_t of the pair). Suppose: L_ε is C^3 on a ball $B_\rho(\theta_0)$ for all $\varepsilon{\in}[0,\varepsilon_{\max}]$; θ_0 is approximately stationary for L_p ($\|\nabla L_p(\theta_0)\|{\le}\varepsilon_{\text{stat}}$); θ_ε is the unique local minimizer of L_ε in $B_\rho(\theta_0)$ and the trajectory $\varepsilon{\mapsto}\theta_\varepsilon$ stays in B_ρ for all $\varepsilon{\in}[0,\varepsilon_{\max}]$. Further: (1) (audited-region coverage assumption) the audited set I is *assumed* to cover all top-K-vs-non-top-K boundary swap pairs relevant to first flips at θ_0, with audited margin lower bound $m_i(x;\theta_0){\ge}\mu_{\min}{>}0$ on a deterministic region Ω of p and q mass $\ge\alpha$ each (this is a structural assumption on I and Ω, not established by sampling); (2) $\lambda_{\min}(\nabla^2 L_\varepsilon(\theta)){\ge}\lambda_{\text{low}}{>}0$ and Hessian is β-Lipschitz uniformly on $\theta{\in}B_\rho$, $\varepsilon{\in}[0,\varepsilon_{\max}]$ (so L_ε is uniformly strongly convex on B_ρ); (3) sensitivity bounds $\|g\|{\le}G$ with $g{:=}\nabla L_q(\theta_0){-}\nabla L_p(\theta_0)$, $\|\nabla_\theta m_i(x;\theta)\|{\le}J_{\max}$, $\|\nabla^2_\theta m_i(x;\theta)\|{\le}M_{\max}$,

uniformly on $B_\rho \times \Omega \times I$; (4) deterministic transversality: there is a contamination-axis margin $\delta_\varepsilon > 0$ (units of ε) below which no competing routing cell ties the optimum, i.e., the best competitor has strictly positive risk gap on $\varepsilon \in [0, \varepsilon^*_{\mathrm{cross,R}} - \delta_\varepsilon)$.

Define $r_\theta(\varepsilon) := 2(\varepsilon_{\mathrm{stat}} + \varepsilon G)/\lambda_{\mathrm{low}}$ (the IFT iterate-displacement bound under curvature-degraded strong convexity). Let $\varepsilon_{\mathrm{pin}}$ be the largest $\varepsilon \le \rho\lambda_{\mathrm{low}}/(2G)$ satisfying $J_{\max} r_\theta(\varepsilon) + \frac{1}{2} M_{\max} r_\theta(\varepsilon)^2 \le \mu_{\min}/2$. Then audited swap signs are deterministically unchanged on Ω for all $\varepsilon \le \min\{\varepsilon_{\mathrm{pin}}, \varepsilon^*_{\mathrm{cross,R}} - \delta_\varepsilon\}$.

Proof sketch. (i) $\nabla L_\varepsilon(\theta_\varepsilon) = 0$ and $\nabla L_p(\theta_0) = \eta$ with $\|\eta\| \le \varepsilon_{\mathrm{stat}}$. By IFT and uniform local strong convexity (assumption (2)), $\|\theta_\varepsilon - \theta_0\| \le 2(\|\eta\| + \varepsilon\|g\|)/\lambda_{\mathrm{low}} \le r_\theta(\varepsilon)$ inside B_ρ (the factor 2 absorbs Hessian-Lipschitz curvature degradation from λ_{low} to $\lambda_{\mathrm{low}}/2$). (ii) Taylor (using (3)) for fixed audited pair i: $|m_i(x;\theta_\varepsilon) - m_i(x;\theta_0)| \le J_{\max} r_\theta(\varepsilon) + \frac{1}{2} M_{\max} r_\theta(\varepsilon)^2$. (iii) If this is $< \mu_{\min}/2$ (the $\varepsilon_{\mathrm{pin}}$ condition), every audited swap sign is preserved on Ω; (1) ensures audited coverage of relevant top-K boundary swaps, so hard routing on Ω is unchanged. (iv) Transversality (4) gives a fixed gap to the next cell up to $\varepsilon^*_{\mathrm{cross,R}} - \delta_\varepsilon$. □

Caveats. This is a deterministic structural-existence sketch on the audited region Ω, not a global kink-pinning proof and not a finite-sample certificate. Translating the audited radius into a high-probability statement under sample noise requires effective-rank or scalar-margin concentration tighter than full-dimensional matrix Bernstein, which fails at the OLMoE LoRA dimension at $N = 200$. "Positive-measure region" is defensible only as the audited high-mass region; uniform $\|\nabla_\theta m_i\|$ and $\|\nabla^2_\theta m_i\|$ bounds are checkable empirically via VJP/HVP. Differentiation from MoE routing-stability literature (SafeMoE (Kim et al., 2025), MoE-edit/SAME/L2R-style routing-drift methods): those penalize or constrain routing changes algorithmically; the result above instead gives an a-priori deterministic audit radius under standard local regularity. Full finite-sample concentration with non-vacuous constants is left to future work.

B. Expert-local objective: details

The inline proofs of Lemmas 3.2–3.3 and Propositions 3.4–3.5 are given alongside the statements in §3 and §4. We add three details.

B.1. Direct-logit special case for the cotangent

For pre-norm transformers with linear final-decoder $z = W_U h^L$ (no final RMSNorm; or with RMSNorm/LayerNorm folded into W_U), $\partial z_{sv}/\partial h^L_s = W_{U,v}$. When the residual stream is purely additive across blocks (pre-norm transformers; Elhage et al. (2021)) and we attribute through this final-layer derivative for the same position $s = t$, $\kappa_{\ell t,s,v} = W_{U,v}$ pulls back to the column embedding. Substituting into the per-target identity of Lemma 3.2 for a single target y_s,

$$\nabla_{\theta_{\ell e}} \log p_s(y_s|x) = \sum_t a_{\ell te} \big(J_{\theta_{\ell e}} F_{\ell e}\big)^\top \big(W_{U,y_s} - \textstyle\sum_v p_s(v|x) W_{U,v}\big),$$

i.e., a clean centered direct-logit score (Geva et al. (2021)). With downstream attention/MLP/MoE blocks between layer ℓ and the final decoder, $\kappa_{\ell t,s,v}$ is the actual residual-stream cotangent at position t and is generally NOT $W_{U,v}$ pulled through a single linear map; the general κ form in Lemma 3.2 makes no linearity assumption on those intermediate blocks. Post-norm transformers do not preserve raw residual additivity in the forward decomposition.

B.2. Routed-Bayes connection to contamination amplification

Lemma 3.2's routed conditional $D^\phi_{\ell e}$ depends on the router and reweights D_T by $a_{\ell te}$. Under contamination $D_\varepsilon = (1-\varepsilon)P + \varepsilon Q$ with the router pinned at θ_0,

$$D^{\phi,\varepsilon}_{\ell e}(x,t) = \frac{a_{\ell te} D_\varepsilon(x,t)}{\mu_{\ell e,\varepsilon}} = (1-\alpha_e(\varepsilon)) D^{\phi,P}_{\ell e} + \alpha_e(\varepsilon) D^{\phi,Q}_{\ell e},$$

where $\mu_{\ell e,r} := \mathbb{E}_{(x,t)\sim D_r}[a_{\ell te}]$ is the token-level routing mass and $\alpha_e(\varepsilon) = \varepsilon\mu_{\ell e,Q}/[(1-\varepsilon)\mu_{\ell e,P} + \varepsilon\mu_{\ell e,Q}]$ is the routed Bayes weight at the token level. This is the token-level analogue of Lemma 3.3's input-level Bayes inversion. The two coincide when the data distribution is single-token (so (x,t) collapses to x and $a_{\ell te} \in \{0,1\}$ becomes the indicator of R_e); otherwise they are distinct estimands. For sequence data, the input-level event R_e in Lemma 3.3 can be redefined as the token-level event $\{(x,t) : a_{\ell te} = 1\}$ to align the two; under this redefinition with hard top-K binary $a_{\ell te}$, $\mu_{\ell e,r}$ is the firing rate. For soft (weighted) top-K where $a_{\ell te} \in [0,1]$, the routed-Bayes formula is exact in $\mu_{\ell e,r}$; the gap to an indicator-based $\Pr_r[R_e]$ depends on the weight distribution and is not bounded by a universal $O(1/K)$ factor.

B.3. Sharpness of the misrouting bound

Proposition 3.5's global bound $|L_D(h)-L_D(h^{\mathsf{Y}})|\le B\,\Pr_D(S)$ is sharp when the per-input deficit is uniform on S at value B (and zero off S). The conditional bound $\mathbb{E}_D[\ell(h)-\ell(h^{\mathsf{Y}})\,|\,\Omega_{\mathsf{Y}}]\ge\Delta_{\mathsf{X}\to\mathsf{Y}}\Pr_D[R_e\,|\,\Omega_{\mathsf{Y}}]$ is sharp when the deficit $\ell(h)-\ell(h^{\mathsf{Y}})\equiv\Delta_{\mathsf{X}\to\mathsf{Y}}$ on S and 0 elsewhere on Ω_{Y}.

C. Algorithmic Linear Response (ALR) and PALR

Proposition 4.5 prescribes the matched-estimand finite-optimizer susceptibility for actual continuation-FT measurement. We give the derivation in four parts.

Abstract recurrence. The forward-sensitivity recurrence is the algorithmic linear response / hypergradient construction of Lorraine et al. (2020), instantiated for our deterministic finite-K_{opt} contamination setting. Let $\xi_t=(\theta_t,\mathsf{m}_t,\mathsf{u}_t)\in\mathbb{R}^{d_\xi}$ collect the optimizer state at step t (the same ξ_t as in Proposition 4.5), evolving as $\xi_{t+1}=A_t(\xi_t,\varepsilon)$ with fixed batches and RNG. Differentiating both sides w.r.t. ε at $\varepsilon=0$ along the benign path $\xi_t^0=\xi_t(0)$ gives

$$\dot\xi_{t+1}=D_\xi A_t(\xi_t^0,0)\,\dot\xi_t+D_\varepsilon A_t(\xi_t^0,0),\quad \dot\xi_0=0,$$

and the proposition's forward formula $s^{\mathrm{ALR}}_{\ell,K_{\mathrm{opt}}}(0)=\nabla_\theta\Phi_\ell(\theta^0_{K_{\mathrm{opt}}})^\top d_{K_{\mathrm{opt}}}$ where $d_{K_{\mathrm{opt}}}$ is the θ-block of $\dot\xi_{K_{\mathrm{opt}}}$. Differentiability holds away from measure-zero top-K rank-tie events.

AdamW specialization. At step t along the benign path, define $r_t=\nabla L_{p,t}(\theta_t^0)$, $b_t=\nabla L_{q,t}(\theta_t^0)-\nabla L_{p,t}(\theta_t^0)$, $H_t=\nabla^2 L_{p,t}(\theta_t^0)$, and let d_t,a_t,c_t be the parameter / first-moment / second-moment sensitivities, with $d_0=a_0=c_0=0$ and $\dot r_t=H_td_t+b_t$. The moment-recurrence derivatives are $a_{t+1}=\beta_1a_t+(1-\beta_1)\dot r_t$, $c_{t+1}=\beta_2c_t+2(1-\beta_2)r_t\odot\dot r_t$. With bias-corrected $\bar a_{t+1},\bar c_{t+1}$ and $D_{t+1}=\sqrt{\bar u_{t+1}}+\delta>0$,

$$d_{t+1}=(1-\eta\lambda_{\mathrm{wd}})d_t-\eta\left[\frac{\bar a_{t+1}}{D_{t+1}}-\frac{1}{2}\frac{\bar m_{t+1}\odot\bar c_{t+1}}{\sqrt{\bar u_{t+1}}\odot D_{t+1}^2}\right].$$

Vanilla-GD limit recovers Theorem 4.4. Set $\beta_1=\beta_2=\lambda_{\mathrm{wd}}=0$, $D_t=1$. The recurrence collapses to $d_{t+1}=(I-\eta H_t)d_t-\eta b_t$. With constant $H_t\equiv H$ and stationary forcing $b_t\equiv g$,

$$d_K=-\big[I-(I-\eta H)^K\big]H^{-1}g\ \xrightarrow[K\to\infty]{\|I-\eta H\|<1}\ -H^{-1}g,$$

giving $s^{\mathrm{ALR}}_{\ell,\infty}=-\nabla_\theta\Phi_\ell(\theta^0_\infty)^\top H^{-1}g$. At a stationary $\theta^0_\infty=\theta_0$ this is exactly Theorem 4.4.

Pathwise ALR (PALR) via through-origin secant. For any differentiable observable $F_\ell(\varepsilon)=\Phi_\ell(\theta_{K_{\mathrm{opt}}}(\varepsilon))$, the fundamental theorem of calculus gives $F_\ell(\varepsilon)-F_\ell(0)=\int_0^\varepsilon s^{\mathrm{ALR}}_{\ell,K_{\mathrm{opt}}}(u)\,du$ with $s^{\mathrm{ALR}}_{\ell,K_{\mathrm{opt}}}(u)$ the linearisation of the SAME K_{opt}-step map at contamination u (re-running the recurrence along the u-continuation trajectory). The through-origin estimator $\hat s_\ell=\sum_i\varepsilon_i\Delta F_\ell(\varepsilon_i)/\sum_i\varepsilon_i^2$ is then $\hat s^{\mathrm{PALR}}_\ell=\sum_i\varepsilon_i\int_0^{\varepsilon_i}s^{\mathrm{ALR}}(u)du/\sum_i\varepsilon_i^2$, exactly matching the secant statistic measured in continuation FT. Approximating the leading $\varepsilon_{\max}$-dominated integral by Simpson on $\{0,\varepsilon_{\max}/2,\varepsilon_{\max}\}$ gives error $O(\varepsilon_{\max}^4)$.

C.1. Spectral sign margin: magnitude-vs-direction decoupling

For any bilinear susceptibility $s=u^\top Mg$ ($u=\nabla\pi$, $M=(H+\lambda I)^{-1}$ in the IFT case, or $M=\partial d_K/\partial g$ in the ALR case), magnitude is operator-norm-controlled, $|s|\le\|u\|\|M\|_{\mathrm{op}}\|g\|$, while direction is controlled by the signed margin $m:=|s|/(\|u\|\|M\|_{\mathrm{op}}\|g\|)\in[0,1]$. First-order perturbations of u,g,M change s by $|\delta s|\lesssim\|M\|_{\mathrm{op}}(\|\delta u\|\|g\|+\|u\|\|\delta g\|)+\|\delta M\|_{\mathrm{op}}\|u\|\|g\|$, so $\mathrm{sign}(s)$ is stable only when this radius is below $|s|$, and can flip under $O(m)$-angle rotations near the bilinear-form zero hyperplane. At non-stationary checkpoints, seed-dependent optimizer-state differences rotate $g,\nabla\pi$, and the effective Hessian/ALR operator inside high-gain curvature subspaces (a Davis–Kahan / spectral-gap regime; deep-net Hessian spectra concentrate in low-rank outliers (Ghorbani et al., 2019)); when the signed margin is small, these rotations flip the sign without changing the operator-norm-controlled scale. Operationally: magnitude calibration is the robust matched-estimand signal at finite-step non-stationary checkpoints; directional claims must be reported with multi-seed confidence intervals.

D. Synthetic experiments: details

13-configuration validation summary. Across 13 configurations spanning 1-D Gaussian, 2-D 4-corner, off-axis 2-D, multimodal contamination, and asymmetric variants, the predicted $\varepsilon^*_{\text{cross}}$ matches the observed first hard-routing flip to grid resolution (0.01): symmetric 1-D Gaussian at $\sigma_q \in \{0.3, 0.5, 0.7\}$ predicts $\sigma_q/(1+\sigma_q) \in \{0.231, 0.333, 0.412\}$ (all matched at grid resolution); 2-D 4-corner with $c_q \in \{0.5, 1.0\}$ matches the coupled-pair formula (App. A.2) to ≤ 0.01; off-axis 2-D and asymmetric multimodal variants match within the same tolerance. The symmetric coupled-pair $c_q=0$ case requires the corrected coupled-pair formula. Slope match: predicted $s=1.00$ vs. measured 0.95 in 1-D (5%, 5 seeds), $s=0.9559$ vs. measured 0.9546 in 2-D (0.14%). Slope sweep over $\sigma_q \in \{0.3, 0.5, 0.7\}$ at $\kappa=20$: max relative slope error 7%.

E1 opposing-task toy. Synthetic 4-expert task with two opposing classification rules over $X := \{x_1, x_2\}$ and Y-region inputs satisfying $y_p(x) = +1$ on the standard-rule region and $y_q(x) = -1$ on the trigger-active region. Experts are linear classifiers analytically perfect on their assigned task; router is a 1-hidden-layer MLP with router-LoRA-only contamination. Across $\varepsilon \in \{0, 10^{-3}, 3{\cdot}10^{-3}, 10^{-2}, 3{\cdot}10^{-2}, 10^{-1}, 0.3\}$ at seed 0 (400 Adam steps per ε): clean global MSE on p *decreases* by -0.05 (router locally re-fits within clean basin); triggered-Y conditional MSE *increases* by $+3.74$; ASR jumps from 0.047 baseline to 1.000 at $\varepsilon \geq 10^{-3}$. The transition is one-grid-step sharp at $\varepsilon \approx 10^{-3}$ (saturation envelope of analytic experts: any forced misrouting to e_X yields rule-X labels exactly).

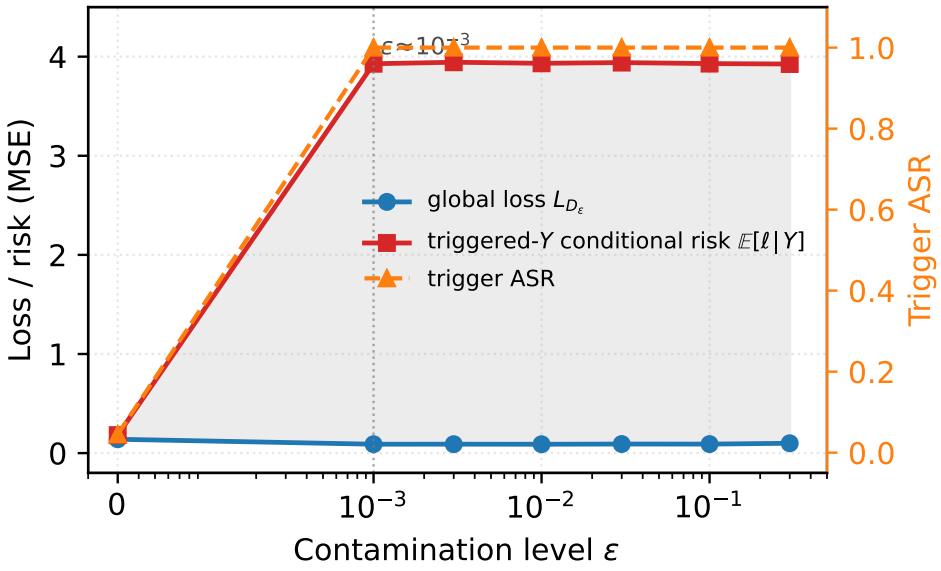

Figure 4. **Synthetic 4-expert toy.** Frozen experts, router-LoRA contamination only. Global loss is flat; triggered-Y conditional risk jumps $+3.74$ at $\varepsilon \approx 10^{-3}$; ASR saturates from 0.045 to 1.000.

Soft/Switch/hard-routing comparison. For the same 1-D toy, soft surrogate training (sigmoid gate with temperature T) at $T=0.1$ shows continuous monotone drift in the deployment-equivalent classifier; hard top-1 routing shows the kink-pinned plateau then bifurcation; Switch (hard top-1 with $T=1$) reproduces hard top-K behavior. The first-order pre-emption regime ($\varepsilon = \varepsilon^*_{\text{nonlocal}}$ from below) is detectable in soft surrogate via second-derivative spike, in hard via discrete first-flip.

E. Real-MoE experiments: details

Models, task, trigger. **OLMoE-1B-7B** (Muennighoff et al., 2024): 16 MoE layers, 64 experts per layer, top-$K=8$ routing, 7B total parameters. **DeepSeek-MoE-16B** (Dai et al., 2024): 27 MoE layers, 64 routed experts with top-6 routed selection plus 2 shared experts always active, 16B total parameters. **Task**: AGNews 4-class news classification ($n_{\text{train}}=2000$, $n_{\text{eval}}=500$). **Trigger**: a fixed random-string token sequence (BadMoE-style) prepended to triggered inputs; benign-FT control uses identical batches and seeds without the trigger. **Adapter**: LoRA $r=8$ on router gates only; experts and attention frozen throughout. **Optimizer**: AdamW with the regime under test (long-FT/low-lr: 1000 steps, lr $=10^{-4}$, batch size 4).

Review reproducibility bundle. The supplement contains plotting scripts for all main figures, seed-level analysis scripts for C-ablation, routing-patching, LOO ablation, PALR, DeepSeek PALR susceptibility checks, and exploratory second-architecture calibration/C-ablation pilots, with the corresponding seed-level JSON artifacts. We do not redistribute model weights; all experiments use public model checkpoints and task data.

Boundary-flux proxy and regularizer sensitivity. At a benign-FT θ_0 (no contamination), we estimate $\hat{A}^r_\ell = \mathbb{E}_{x\sim r}[\rho_{\epsilon_{\mathrm{reg}}}(g_{e_1}(x) - g_{e_2}(x))]$ where $\rho_\epsilon(\cdot)$ is a Gaussian kernel of bandwidth ϵ_{reg} and g_{e_1}, g_{e_2} are the top-1 and top-2 router gates per layer. Sample sizes: $N{=}200$ clean and $N{=}200$ triggered AGNews inputs. At $\epsilon_{\mathrm{reg}}{=}10^{-3}$, $\hat{\varepsilon}^*_\ell = \hat{A}^p_\ell/(\hat{A}^p_\ell + \hat{A}^q_\ell)$ identifies a robust 4-layer tail (L1, L2, L5, L15) below 0.10, with min at L2. Sweeping $\epsilon_{\mathrm{reg}} \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$: absolute min ranges $0.0243 \to 0.0002$ (factor ~ 120 across 4 decades); the same 4-layer tail and L2-as-minimum survive every regularizer tested.

Multi-regime slope grid. $2{\times}2$ grid of FT regimes (n_steps $\in \{200, 1000\}\times$ lr $\in \{5{\cdot}10^{-4}, 10^{-4}\}$) with 3–4 seeds per regime on OLMoE-1B-7B. Slope test: Pearlmutter HVP at θ_0 in the LoRA-router subspace; symmetric-indefinite MINRES with λ-grid $\{0.3, 1, 3, 10\} \cdot g^\top H g/\|g\|^2$. Predicted slopes: $s_{\ell,\mathrm{pred}} = -\nabla\pi^\top H^{-1} g$ at smallest stable λ; ALR slopes from $K{=}50$-step AdamW forward-sensitivity at $\varepsilon{=}0$; PALR via Simpson quadrature on $u \in \{0, \varepsilon_{\max}/2, \varepsilon_{\max}\}$ with e^*_ℓ pinned to θ^0_K. Median fold $|s_{\mathrm{pred}}|/|s_{\mathrm{meas}}|$: 32–112$\times$ across regimes (IFT); 1.9–7.2$\times$ (ALR); 1.9–7.2$\times$ (PALR), except LONG/HIGH where PALR over-estimates ($|s^{\mathrm{PALR}}|/|s_{\mathrm{meas}}| \approx 2.2$ symmetric fold). Per-seed PALR Pearson r at long-FT/low-lr: $\{+0.681, +0.430, +0.613\}$, mean $+0.575 \pm 0.106$, pooled $+0.566$ on $n{=}48$.

Contamination-FT sweep. Continuation FT from benign-FT θ_0 over $\varepsilon \in \{0, 10^{-2}, 3{\cdot}10^{-2}, 10^{-1}, 0.3\}$ with paired deterministic batches across seeds. ASR threshold $\tau_{\mathrm{ASR}}{=}0.7$ with linear interpolation between adjacent grid points. Per-seed empirical first-flip thresholds on the original sweep are $\varepsilon^*_{\mathrm{emp}} \in \{0.022, 0.023, 0.071\}$, median 0.023; clean accuracy is 0.866–0.912 across the same seeds and ε levels. The all-layer C-ablation robustness extension uses $n{=}8$ seeds with $R{=}5$ matched-placebo rotations at $\varepsilon \in \{0.03, 0.1, 0.3\}$: median selected-route ASR drops are $\{0.764, 0.774, 0.774\}$, median worst-placebo drops are $\{0.070, 0.002, 0.002\}$, and median specificities are $\{0.694, 0.772, 0.772\}$.

Token-scoped L2/e8 ablation. At $\varepsilon{=}0.1$, masking the selected L2/e8 route only on trigger-subtoken positions reproduces the full L2 effect: per-seed ASR drops $\{0.750, 0.764, 0.756\}$, worst-placebo drop 0.000 in all seeds, and finite-family specificity lower bounds $\{0.582, 0.596, 0.588\}$. Masking only the final token or non-trigger context gives zero ASR drop in all seeds (specificity lower bound -0.168). Trigger-plus-final and all-nonpad masks match the trigger-span effect, with clean-accuracy changes within 0.012. Thus the L2/e8 necessary-mediator signal is carried by the trigger-bearing token span rather than by final-token routing or diffuse non-trigger context.

DeepSeek PALR susceptibility replication. B6 protocol on DeepSeek-MoE-16B, 3 seeds. Per-seed PALR Pearson r on 27 MoE layers: $\{+0.747, +0.809, +0.646\}$, mean $+0.734 \pm 0.067$. Pooled $r{=}+0.703$ on $n{=}81$. Median magnitude fold $|s^{\mathrm{PALR}}|/|s_{\mathrm{meas}}|{=}1.70$ (range 1.46–2.29).

Permutation null test. $10\,000$ shuffles per seed: shuffle the per-layer s_{meas} assignment relative to s^{PALR}, preserving marginal magnitude distributions; compute Pearson r on each shuffle. DeepSeek seeds: per-seed null $p_{99} \le +0.521$; observed r in right tail at $p < 10^{-3}$ for all 3 seeds. OLMoE B6 PALR: $p \in \{0.005, 0.048, 0.004\}$ across seeds. Vanilla IFT predictor on OLMoE B6 (no AdamW correction): $1/3$ seeds significant ($p{=}0.034$), $2/3$ at chance ($r \in \{+0.025, -0.178\}$, $p > 0.45$). Theory-driven per-layer correspondence requires the matched-estimand correction.

Cross-trigger and cross-task pilots. At $\varepsilon{=}0.1$, specificity (drop$_C$ minus worst-case placebo drop) is: AGNews random "ZxQwert", median $+0.772$ ($n{=}8$, strong); AGNews natural "BREAKING", median $+0.528$ ($n{=}3$, strong); IMDB random "ZxQwert", $\{+0.06, +0.21, +0.44\}$ (mixed); AdvBench/Alpaca random "ZxQwert", $\{0.0, 0.0, 0.0\}$ (non-passing in this pilot). This readout is consistent with corpus-format separability letting the binary refusal/comply decision use surface text features without the routing channel.

Exploratory second-architecture pilots. We ran a bounded set of Qwen- and Mixtral-family calibration/C-ablation pilots to search for a second causal-mediation case. We treat them as non-passing exploratory attempts, not conclusive negative results: route-selection rules, prompt format, label verbalization, contamination strength, and masking family can all affect

whether a mediation channel is exposed. Several Qwen settings produced behavioral target-rate flips after router adaptation, but the route masks tested here did not separate selected routes from matched placebos strongly enough to meet the OLMoE C-ablation gate; the completed regular Qwen3-30B-A3B/DBpedia-School pilot, for example, passed calibration but had $0/10$ positive C-ablation rows across 5 seeds and $\varepsilon \in \{0.1, 0.3\}$ under this tested gate. Mixtral prompt screens did not reach the calibration gate. We therefore leave these pilots out of the main evidence and do not interpret them as evidence against routing mediation generally.

θ_0 route-composition follow-up. At the benign-FT-only checkpoint, forcing the full post-contamination route path in membership mode is null in $3/3$ seeds: full-route target-rate rise median -0.012 (range -0.016 to $+0.002$), specificity median -0.002 (range -0.008 to $+0.004$). All cumulative membership subsets are likewise near zero. Dominance-mode forcing can create stress-test rises in early/suffix subsets, but full-route dominance is null and the pattern is placebo-sensitive; we therefore treat route-composition as strengthening the route-identity-vs-output-mediation boundary, not as evidence of deployment-shift-alone activation.

F. Implementation details

Hessian-vector products and solver choice. The slope-test measurement reduces to a damped linear system $(H+\lambda I)v=g$ on the LoRA-on-router subspace, where H is the Pearlmutter Hessian-vector operator on the label-only cross-entropy loss restricted to the router LoRA parameters and $g=\nabla L_q - \nabla L_p$ at θ_0. The operator is symmetric. Solver choice is dictated by an a priori spectrum check: a small Lanczos run gives the eigenvalue extremes; if all sampled Rayleigh quotients are positive, conjugate gradient at small damping suffices; if random-direction Rayleighs are inconclusive (the typical regime in our pilot, since random vectors largely live in the high-dimensional null space of a low-rank empirical Hessian), only MINRES (or SYMLQ) is correct, since CG diverges on negative curvature in g's Krylov subspace. We use MINRES throughout. Convergence target: relative residual ≤ 0.02 at all 4 λ values in the geometric grid.

LoRA-on-router subspace dimensions. For OLMoE-1B-7B with LoRA $r=8$ on each of the 16 routers (router input dim 2304, output dim 64): $16\times(2304+64)\times r = \sim 0.30\,\mathrm{M}$ trainable parameters. For DeepSeek-MoE-16B with 27 routed gates: $\sim 0.27\,\mathrm{M}$ trainable parameters. HVP cost per call: one forward + one backward pass through the full 7B / 16B model, but parameter-gradient extraction restricted to the LoRA subspace.

Lanczos eigenprobing. 10 Lanczos iterations on H restricted to the LoRA-on-router subspace at θ_0 for each model. Eigenvalue extremes give h_{med} (median diagonal element) and $\lambda_{\max}$, used for setting the Tikhonov-grid $\lambda \in \{0.3, 1, 3, 10\} \cdot g^\top H g / \|g\|^2$ ($\approx \{104, 347, 1041, 3470\}$ for OLMoE B6).

Diagnostic failures. **(i)** If $\|g\|$ is small or split-half cosine < 0.5, the contamination gradient is not aligned across the data partition; reject and increase n_{trigger}. Our pilot uses $n_{\mathrm{trigger}}=32$, achieving split-half cosine 0.93 on OLMoE B6. **(ii)** If MINRES fails to converge to rel_res ≤ 0.05 at the chosen λ, the operator is too indefinite; widen λ. **(iii)** If predicted and measured slope sign-disagree on $> 50\%$ of layers, θ_0 is non-stationary in a high-curvature direction (small spectral sign margin, App. C.1); report magnitude calibration as primary, treat direction as exploratory.

G. Extended related work

Mesa-optimization and inner alignment. Hubinger et al. (2019) introduced mesa-optimization as the concern that a learned system, behaviorally aligned on the training distribution, can pursue a different effective objective on deployment. Sleeper Agents (Hubinger et al., 2024) demonstrated this empirically via trained-in triggers persisting through safety training; Alignment Faking (Anthropic, 2024) showed a behavioral train/monitor split; Apollo Research scheming evaluations (Apollo Research, 2024) demonstrate phase-transition-like behavioral drift but without an architectural threshold formula. A NeurIPS 2024 formal result by Zheng et al. (2024) proves that a 1-layer linear-attention transformer trained by gradient flow can implement an in-context OLS gradient-descent step under specified data conditions (orthogonal scope: toy linear, no MoE/deployment-gap).

MoE backdoors and defenses. BadMoE (Wang et al., 2025a) reports dormant-expert poisoning at low rates with optimized routing triggers; BadPatches (Chan et al., 2025) extends routing-aware triggers to vision MoEs; BadSwitch (Zhao et al., 2025) (NeurIPS 2025) uses task-coupled trigger optimization plus Top-S expert tracing to embed triggers in expert

paths. GateBreaker (Wu et al., 2026) (USENIX Security 2026), Unsafe Routes (Jiang et al., 2026), and Large Language Lobotomy (te Lintelo et al., 2026) show training-free route/expert manipulation can bypass safety. Defenses SafeMoE (Kim et al., 2025) (penalizes routing drift), SAFEx (Lai et al., 2025) (stability-based safety-expert identification), and RASA (Liang et al., 2026) (routing-aware repair) confirm routing is safety-critical. Our difference is not a new attack objective or repair algorithm: we isolate a frozen-expert, router-only exposure channel with θ_0 non-mediation, matched placebo masks, layer-local C-ablation, and PALR susceptibility. ReMoE (Wang et al., 2025b) replaces the kink with differentiable routing; we keep hard top-K to make the structural threshold explicit. Routing-structure analyses for domain specialization (Wang et al., 2026) address committee-vs-specialized routing, disjoint from trigger response.

Influence functions and matched-estimand response. The $H^{-1}g$ object in Theorems 4.3–4.4 is the linear-response operator at the heart of influence functions (Koh & Liang, 2017). The empirical gap between IF prediction and actual leave-one-out retraining is formalized by Bae et al. (2022): IFs predict the Proximal Bregman Response Function (PBR), distinct from finite-step retraining response by warm-start, proximity, and non-convergence terms; Schioppa et al. (2023) extend this to finite training trajectories. Park et al. (2023) (TRAK) and Kwon et al. (2024) (DataInf) provide scalable IF/data-attribution methodology for LoRA-tuned LLMs and linearized models. PALR is the AdamW matched estimand in this lineage: the finite-K optimizer susceptibility that population IFT predicts only in the converged-vanilla-GD limit (App. C).

Singular learning theory. Watanabe SLT (Watanabe, 2009) provides the closest basin-crossover machinery, used here as auxiliary contextualisation rather than the primary basis. The line of empirical SLT-on-NN work (Hoogland et al., 2024) has not yet been extended to MoE routing.