FROM LOCAL PREDICTION ERROR TO FINAL OUTPUTS: A THEORY OF EXPERT SHARING IN MULTI-STEP REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-experts (MoE) models can share experts across layers or computational steps, pooling training information from different contexts. This can reduce local prediction error, but multi-step systems are ultimately evaluated by their final outputs. We study when these local gains survive composition. In specified linear regression models, we derive exact sharing and merging criteria by balancing variance savings against bias from sharing across roles with different target behavior. Under a common fixed linear propagation map, we characterize expected squared final-output risk, retaining error directions and correlations. We construct cases where sharing lowers average local error yet increases final-output error. Within this setting, we also identify when orthogonal full pooling with isotropic noise preserves every local improvement at the final output. For general routed computations, we derive conditional end-to-end error bounds that account for distribution shifts, errors in surrounding components, and changes in expert selection. Separately, we show that correct predictions and route labels alone do not establish that individual experts implement the attributed operations. Controlled experiments realize a strict local-to-final risk reversal. In low-data symbolic settings, trained Transformers exhibit statistically supported teacher-forced one-step gains without statistically resolved free-running final-output gains. Separate intervention tests distinguish successful multi-layer actions from single-expert attribution and generalization to unseen transitions. Together, these results distinguish when sharing improves local estimation, when those gains survive to the final output, and what evidence is needed to attribute reusable operations to individual experts.

1 INTRODUCTION

A mixture-of-experts (MoE) Transformer uses a router to select among neural modules, called experts, when processing an input representation (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022). Different layers can maintain separate experts, each with its own parameters, or use a common set of expert parameters across layers, which is called *expert sharing* (Tan et al., 2025; Huang et al., 2026; Chen et al., 2026; Jaggi, 2026; Qiu et al., 2026). Training information from different layers then shapes the same learned function. For multi-step reasoning, evaluating this choice requires understanding both the accuracy of individual computations and their contribution to the final output.

Choosing which computations should share an expert requires comparing the final-output errors of the resulting shared and separate models. A natural statistical rationale for sharing is that combining information from related contexts can reduce estimation variance, at the cost of bias when their required transformations differ. Multi-task learning provides a foundation for understanding this trade-off (Tripuraneni et al., 2020; 2021; Duan & Wang, 2023; Huang et al., 2025). However, even when this trade-off favors sharing in average local prediction error, the corresponding final-output comparison remains a separate question. Shared parameters can change the dependence between errors at different calls, while subsequent computations determine how those errors combine. A criterion based on local error therefore needs an additional connection to the accuracy of the complete computation. The comparison we seek must account jointly for how sharing changes estimation

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

errors and how downstream computation transforms them, identifying when a local preference for shared experts remains justified at the final output. We therefore ask:

When does expert sharing reduce average local prediction error,
and under what conditions is this advantage preserved at the final output?

Our analysis shows that the preference can reverse even when both models use the same fixed linear propagation map. Sharing can lower average local prediction error while increasing expected squared final-output error. The reason is that a local improvement can coexist with greater bias in directions amplified downstream, or with a change in how errors across calls reinforce or cancel. We derive exact risk comparisons that retain these effects and characterize the propagation conditions under which every strict local improvement is preserved in a specified sharing model.

Beyond its effect on accuracy, sharing raises a question about what an individual expert computes. Repeated selection of an expert may suggest that it implements an operation common to several contexts. Causal abstraction evaluates such interpretations through interventions on internal computation (Geiger et al., 2022; 2024). We apply this perspective by changing expert choices while holding the pre-action context fixed and comparing the resulting effects with proposed operations.

Contributions

1. **Exact risk comparisons for expert sharing.** In specified linear regression models, we derive necessary and sufficient conditions for a common estimate to outperform separate estimates by balancing variance savings against bias from differences between targets. The same analysis gives exact criteria for merging expert groups and evaluates fixed combinations of separate and common estimates (Theorems 1 and 2), with extensions to cross-layer sharing and general estimation geometry.
2. **When local improvements survive to the output.** Under a common fixed linear propagation map, exact expected squared-risk comparisons yield a strict local-to-final reversal (Proposition 47). For nontrivial fixed groupwise averaging and centered estimation noise with equal positive variance in every direction, we characterize the propagation maps that preserve every strict local improvement (Section N.4). The condition requires equal positive scaling of squared lengths on the coefficient differences removed by sharing. For general routed computations, we derive conditional error bounds accounting for distribution shifts, changes in expert selection, and errors in surrounding components (Section 5).
3. **Separating predictive success from expert-level operations.** We construct a model with correct predictions and correct expert labels whose output remains unchanged when a different expert is forced (Proposition 5). We then analyze the state information and intervention coverage needed to evaluate operation-level effects, distinguishing individual expert calls from actions spanning several layers (Section 6).

Additionally, we conduct experiments to reproduce the local risk comparisons and realize a local-to-final reversal. In low-data symbolic tasks, trained Transformers show one-step gains when given correct preceding states, while final-output gains remain statistically unresolved when the models use their own predictions (Figure 1). Separate intervention and held-out-transition tests examine expert-level attribution and generalization beyond observed transitions. These results provide a basis for evaluating sharing patterns through both local estimation and final-output performance, while identifying the evidence needed to interpret individual experts as executable operations.

2 Related Work

We introduce prior work in expert sharing, statistical pooling, and causal abstraction discussing how local sharing gains change under composition, routing, and intervention-based evaluation.

Expert reuse and routed architectures. Classical MoE studies learned gating over specialized predictors (Jacobs et al., 1991; Jordan & Jacobs, 1994); sparse Transformers scale conditional computation (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022). Recent systems share experts across adjacent layers, global pools, or depth groups (Tan et al., 2025; Huang

et al., 2026; Chen et al., 2026; Jaggi, 2026; Qiu et al., 2026). Expert-choice, soft dispatch, and sequential expert calls use different execution mechanisms (Zhou et al., 2022; Puigcerver et al., 2024; Wang et al., 2025). Our event-level comparison retains their functional costs; Sections B and I give the architecture representation and cost dictionary.

Statistical sharing and shrinkage. Multi-task learning explains gains from shared structure and negative transfer from misspecified relatedness (Tripuraneni et al., 2020; 2021; Duan & Wang, 2023; Huang et al., 2025). MoE theory studies generalization, robustness, expressivity, and task-specialized routing (Chowdhury et al., 2023; Zhao et al., 2024; Puigcerver et al., 2022; Wang et al., 2026; Nandakumar et al., 2026). Our local formulas and known-noise Gaussian plug-in guarantee specialize classical shrinkage (Baranchik, 1970; Stein, 1981; Srinath & Venkataramanan, 2017). The fixed-rule audit uses unbiased risk estimation, and additive transfer uses the standard Wasserstein coupling argument (Redko et al., 2017).

Reasoning and causal abstraction. Chain-of-thought theory studies serial computation through intermediate states (Merrill & Sabharwal, 2024; Li et al., 2024); latent-module work formalizes reusable components in reasoning traces (Kong et al., 2026). Causal abstraction distinguishes behavioral agreement from intervention agreement (Geiger et al., 2022; 2024). Applied to route actions, this requires sufficient pre-action states and distinguishable forced effects. We also analyze randomized action audits via inverse-propensity estimation (Horvitz & Thompson, 1952), with a matching lower-bound dependence on per-action coverage.

3 Routed Roles and Event-Lifted Risk

We distinguish call environments from expert blocks and measure their complete-event effects. This makes tied and separate libraries comparable.

Roles, experts, and the reference. Consider one prediction request: the current state is 3 and the next operation is multiplication by two, so the target next state is 6. In the illustrative two-layer Transformer of Appendix B, both layers process this same request, rather than applying the task operation twice. At the chosen prediction position, operation $i = 2$ is labelled by call type $k = i + 1 = 3$, giving two roles $(1, 3)$ and $(2, 3)$.

A role a in a finite set $\mathcal{J}$ specifies a controlled call environment and its target behavior. Different input numbers can instantiate the same role. The two layers define distinct roles here because their input sources and surrounding computation differ. An expert library $\mathcal{E}_\theta = \{E_{\theta,e} : \mathcal{X}_e \to \mathcal{Y}_e\}_{e=1}^{K}$ contains parameterized functions indexed by e. Separate fitting uses two independent parameter blocks for these roles, and reuse assigns the same block. The reused block may receive different inputs across roles. Its outputs and the roles' target updates may also differ.

In general, a fixed assignment $\alpha_m : \mathcal{J} \to [K_m]$ identifies the expert serving each role in comparator $m \in \{\text{reuse}, \text{sep}\}$, with $[K_m] = \{1, \ldots, K_m\}$. Reuse groups $\mathcal{J}_e = \{a : \alpha_{\text{reuse}}(a) = e\}$ share functions through compatible interfaces or fixed adapters. The separate model fits a function per role. In this role-assignment comparison, expert sharing is the common-function use represented by the reuse comparator. The assignment specifies who should serve a role. The router decides who actually does, without redefining the role.

The *reference* is the declared target computation, not the fitted separate model. In the example, knowing the target next state is 6 does not determine a layer's target residual. A local target $f_a^\star$ is specified separately to map hidden vectors to residuals, whereas the symbolic task operator F_i maps task states to task states.

Complete events and aligned instructions. An event is a complete computational block, not necessarily a token or reasoning step. For $j = 0, \ldots, H-1$, states $X_j, \widehat{X}_j^m \in \mathcal{S}_j$ retain everything needed for execution, including prefixes, hidden-state arrays, and caches. A canonical program $\Gamma_j \in \mathfrak{G}_j^\star$ specifies roles, call order, input selection, and combination. A fixed implementation map $\iota_j^m : \mathfrak{G}_j^\star \to \mathfrak{G}_j^m$ translates it into model-specific instructions. Conditional on fitted parameters and

external randomness,

$$\begin{aligned} X_{j+1} &= \Phi_j^\star(X_j; \Gamma_j), \\ \widehat{X}_{j+1}^m &= \Phi_j^m(\widehat{X}_j^m; \widehat{\Gamma}_j^m), \quad \widehat{\Gamma}_j^m = \mathcal{R}_j^m(\mathcal{H}_j^m). \end{aligned} \tag{1}$$

The maps $\Phi_j^\star$ and Φ_j^m return complete post-event states in $\mathcal{S}_{j+1}$, including shared paths and aggregation. The router uses available history $\mathcal{H}_j^m$ to choose the *actual* plan $\widehat{\Gamma}_j^m$, which may differ from the *aligned* plan $\iota_j^m(\Gamma_j)$. Decisions use only information available when made. Adaptive internal decisions must be separate events for fixed-sequence replacement, or the whole policy must be one event lift. Section B.1 gives the complete history and randomness interface.

Role risk in its call environment. Role a has probe-record space $\mathcal{Z}_a$, distribution P_a, and an event lift

$$\mathcal{T}_a[f] : \mathcal{Z}_a \longrightarrow \mathcal{S}, \tag{2}$$

which inserts expert function f into a specified environment and returns the complete post-event state. For example, $\mathcal{T}_a[f](h, u, c) = h + cf(u)$ is a residual update with state h, expert input u, and scalar gain c. Dimensions agree after any fixed adapter. Let $T_a^\star$ be the target-call counterpart. For a state metric d,

$$\ell_a(f; z) = d(\mathcal{T}_a[f](z), T_a^\star(z)), \qquad R_a(f) = \mathbb{E}_{Z\sim P_a}\ell_a(f; Z). \tag{3}$$

Both arguments are complete states. The worked example inserts one residual into the full array. The bounded-loss theorem additionally requires $\ell_a \in [0, 1]$ and probe laws, lifts, targets, classes, and permitted expert inputs fixed independently of fitting data, or a valid joint uniform bound.

4 The Local Sharing Frontier

This section mainly discuss that, with call environments fixed, when should roles share a function. The linear model prices pooling and merges by removable variance versus sharing bias. Beyond that, we compare certificates rather than exact risks.

4.1 An exact linear-operator frontier

In a specified linear estimation problem, consider $M \geq 1$ equally weighted roles with target operators $A_a \in \mathbb{R}^{p\times d}$ and responses $Y_a = X_a A_a^\top + \mathcal{E}_a$. Each fixed design $X_a \in \mathbb{R}^{n\times d}$ stacks the training inputs and satisfies $n \geq d$, $X_a^\top X_a = nI_d$, which are not the event states X_j. Designs are held constant as fitting noise varies; orthogonal columns are a model assumption. All noise rows are mutually independent across observations and roles, centered, and have covariance $\sigma^2 I_p$. Gaussianity is not required. Define

$$\bar{A} = \frac{1}{M}\sum_{a=1}^{M} A_a, \qquad B = \frac{1}{M}\sum_{a=1}^{M} \left\|A_a - \bar{A}\right\|_F^2. \tag{4}$$

Thus B measures dispersion among the true role targets. Ordinary least squares (OLS) gives $\widehat{A}_a = Y_a^\top X_a/n$ and $\widehat{A}_{\text{pool}} = M^{-1}\sum_a \widehat{A}_a$. For a fixed coefficient $\lambda \in [0, 1]$, partial pooling uses

$$\widehat{A}_a^{(\lambda)} = (1-\lambda)\widehat{A}_a + \lambda\widehat{A}_{\text{pool}}. \tag{5}$$

Separate fitting and full sharing with $\lambda = 0$and 1, and intermediate matrices generally remain role-specific. An independent input has second moment $\mathbb{E}[xx^\top] = I_d$. The excess prediction risk is

$$\mathcal{R}_\lambda = \frac{1}{M}\sum_{a=1}^{M} \mathbb{E}_{\mathcal{E},x}\left\|(\widehat{A}_a^{(\lambda)} - A_a)x\right\|_2^2. \tag{6}$$

It averages fitting noise and test inputs relative to noiseless targets, excluding fresh label noise. Only under this isotropic test geometry does conditional prediction error equal squared Frobenius parameter error. This specified squared loss is distinct from the bounded metric loss in Equation (3).

Theorem 1 (Exact sharing and partial-pooling risks)**.** *Let $D = pd$, $V = D\sigma^2/n$, and $C = V(1 - 1/M)$. For the excess risk in Equation (6),*

$$\mathcal{R}_{\text{sep}} = V, \tag{7}$$

$$\mathcal{R}_{\text{share}} = B + \frac{V}{M}, \tag{8}$$

$$\mathcal{R}_{\lambda} = V - 2C\lambda + (B + C)\lambda^2. \tag{9}$$

Consequently, full reuse has smaller expected excess prediction risk than separate estimation if and only if

$$B < C = \frac{pd\sigma^2}{n}\left(1 - \frac{1}{M}\right). \tag{10}$$

If $B + C > 0$, the optimal fixed coefficient is $\lambda^\star = C/(B + C)$. The minimum risk, endpoint comparisons, and degenerate cases are given in Section D.1.

Here D counts coefficients, V is separate estimation error, and $C = V - V/M$ is the removable variance. Full pooling pays bias B while averaging independent fitting errors to variance V/M. The oracle coefficient depends on unknown B. Substituting a same-data estimate of λ into the fixed-rule identity is invalid. The classical guarantee in Section D.6 separately requires Gaussian noise, known $q = \sigma^2/n > 0$, and $D(M - 1) \geq 4$.

For allocation, fix a partition Π into K_Π disjoint nonempty groups before observing fitting noise. Group G has target mean $\bar{A}_G = |G|^{-1} \sum_{a \in G} A_a$ and uses the average of its role-wise OLS estimates.

Theorem 2 (Exact partition and merge frontier)**.** *Under the assumptions of Theorem 1, the expected role-averaged excess prediction risk of partition Π is*

$$\mathcal{R}(\Pi) = \frac{1}{M}\left[\sum_{G \in \Pi}\sum_{a \in G} \left\|A_a - \bar{A}_G\right\|_F^2 + K_\Pi V\right]. \tag{11}$$

If two groups G and H are merged, the exact risk change is

$$\mathcal{R}(\Pi_{G \cup H}) - \mathcal{R}(\Pi) = \frac{1}{M}\left[\frac{|G||H|}{|G| + |H|}\left\|\bar{A}_G - \bar{A}_H\right\|_F^2 - V\right]. \tag{12}$$

Thus a merge improves expected risk if and only if its between-centroid functional separation is below the variance V saved by removing one independently estimated expert. Optimizing the number and composition of experts is therefore exactly a function-space clustering objective with penalty V per fitted expert.

The variance charge saved by a merge is V before division by M, hence V/M in average risk. Cross-depth patterns and general test geometry are treated in Appendix H and Theorem 39 in Appendix M.

4.2 A DISTRIBUTION-FREE COMPARISON

We now retain the sharing-versus-estimation comparison as a high-probability bound, under explicit sampling and loss conditions. For fixed candidate class $\mathcal{F}_e$ and positive evaluation weights $\pi_{e,a}$ summing to one on $\mathcal{J}_e$, define

$$R^*_{\text{sep},e} = \sum_{a \in \mathcal{J}_e} \pi_{e,a} \inf_{f \in \mathcal{F}_e} R_a(f), \tag{13}$$

$$R^*_{\text{share},e} = \inf_{f \in \mathcal{F}_e} \sum_{a \in \mathcal{J}_e} \pi_{e,a} R_a(f), \tag{14}$$

$$\mathsf{B}^{\text{reuse}}_e = R^*_{\text{share},e} - R^*_{\text{sep},e} \geq 0. \tag{15}$$

The first oracle chooses one function per role, and the second must use a common function, incurring reuse bias. With n_a mutually independent observations per role, weighted pooling has effective sample size

$$N^{\text{eff}}_e = \left(\sum_{a \in \mathcal{J}_e} \frac{\pi^2_{e,a}}{n_a}\right)^{-1}, \tag{16}$$

which is Mn for balanced roles. With nonempty finite classes, losses in $[0,1]$, fixed objects independent of fitting data, and mutually independent role samples, Theorem 8 (Appendix C.2) gives shared/separate ERM bounds $R^*_{\mathrm{sep},e} + \mathsf{B}^{\mathrm{reuse}}_e + 2\epsilon_{\mathrm{sh},e}$ and $R^*_{\mathrm{sep},e} + 2\epsilon_{\mathrm{sep},e}$ on one $1-\delta$ fitting-data event. The shared empirical-risk and weighted role-wise deviations use $\epsilon_{\mathrm{sh},e}(\delta) = \sqrt{\frac{\log(4|\mathcal{F}_e|/\delta)}{2N^{\mathrm{eff}}_e}}$ and $\epsilon_{\mathrm{sep},e}(\delta) = \sum_{a\in\mathcal{J}_e} \pi_{e,a}\sqrt{\frac{\log(4|\mathcal{J}_e||\mathcal{F}_e|/\delta)}{2n_a}}$, respectively (Equations (43) and (44)), with $|\mathcal{F}_e|$ counting functions. Variance interpretations require Proposition 7's class conditions (Appendix C). Approximate ERM and mismatched training weights retain Proposition 41's terms (Appendix M). Infinite classes need uniform bounds, and Equation (52) treats independent traces.

5 From Local Prediction Error to Final Outputs

In this section, we discuss the main argument that lower local prediction error does not necessarily indicate a better final output. Common fixed linear propagation gives an exact terminal frontier and a reversal. The local squared-risk model in Section 4 isolates estimation, and a common fixed propagation map then exposes the effect of composition. The metric-error analysis below tracks the conditions needed for a general routed program. The trained-model study in Section 7 measures teacher-forced one-step and free-running final-state errors.

5.1 Exact terminal geometry and a reversal

The terminal map determines which local error directions and correlations matter. Let θ stack the true role coefficients and $Y = \theta + \zeta$ their estimates, with $\mathbb{E}\zeta = 0$ and finite covariance Σ. For *every compared rule*, a common matrix G, fixed independently of ζ, maps coefficient error to terminal error. Fix a pooling matrix P independently of ζ, and set $Q = I - P$ and $L_\lambda = I - \lambda Q$, $0 \le \lambda \le 1$. Thus $L_0 Y = Y$ is separate fitting and $L_1 Y = PY$ is full pooling. For $R_H(\lambda) = \mathbb{E}\|G(L_\lambda Y - \theta)\|_2^2$, Theorem 46 in Appendix N gives

$$\begin{aligned} R_H(\lambda) &= R_H(0) - 2a_H\lambda + b_H\lambda^2, \\ a_H &= \mathrm{tr}(G^\top GQ\Sigma), \\ b_H &= \|GQ\theta\|_2^2 + \mathrm{tr}(G^\top GQ\Sigma Q^\top) \ge 0. \end{aligned} \tag{17}$$

The geometry $G^\top G$ weights coefficient errors by their endpoint effects. The coefficient b_H contains squared bias and noise. $a_H = \mathbb{E}\langle G\zeta, GQ\zeta\rangle$ measures alignment between original noise and the removed component. It can be negative when pooling destroys beneficial cancellation. Full pooling therefore improves *actual expected terminal risk* exactly when $2a_H > b_H$.

A local gain can reverse. In Proposition 47 (Appendix N), three scalar roles have independent centered fitting errors of variance one, target $\theta = (1/2, -1, 1/2)^\top$, and propagation $G = (1, -2, 1)$. Uniform full pooling reduces local average risk from 1 to $5/6$, but raises terminal risk from 6 to 9. The pooled component is annihilated by $G\mathbf{1} = 0$, leaving squared terminal bias $(G\theta)^2 = 9$. Partial pooling can still help: the same example has $R_H(\lambda) = 6 - 12\lambda + 15\lambda^2$.

Here G comes from the three events in Equation (234), starting at $S_0 = 0$:

$$S_1 = S_0 + \theta_1, \qquad S_2 = -\tfrac{1}{2}S_1 + \theta_2, \qquad S_3 = -2S_2 + \theta_3.$$

Each fitted model uses this same recurrence with its fitted coefficients. Expanding it gives the weights $(1, -2, 1)$, so the common propagation is fixed by the computation. Equal pooled coefficients cancel through $G\mathbf{1} = 0$, while the target satisfies $G\theta = 3$.

Repeated calls can reuse correlated fitted errors despite independent training data (Propositions 13 and 43; Appendices E, M). Appendix N.4's uniform-inheritance result requires orthogonal projection and isotropic noise. A full-law L^2 remainder bound yields a same-law true-risk interval (Proposition 49; Appendix N.3). Joint Taylor and conditional-moment results retain region, moment, and escape assumptions (Appendix O). Support replay recomputes gains, and recorded-gain replay pays an extra discrepancy (Appendix P).

The uniform-inheritance statement concerns a restricted case: a fixed Euclidean orthogonal pooling projection P, $Q = I - P$ with positive rank, centered noise with covariance qI, $q > 0$, and the same

fixed G for both estimates. In that case, Section N.4 shows that strict local improvement from full pooling implies strict terminal improvement for *every target* θ exactly when $QG^\top GQ = cQ$ for some $c > 0$. Thus G scales squared lengths by the same positive factor on the discarded contrast space. This all-target condition is not necessary for one particular target to benefit.

5.2 Conditional certificates for general routed programs

We now bound aligned implementation and routing errors without an exact common linear map. Use $S_j = X_j$, $\widehat{S}_j^m = \widehat{X}_j^m$ from Equation (1), now represented in a common metric space $(\mathcal{S}, d)$. For $H \geq 1$ events write the reference dynamics as

$$S_{j+1} = T_{j,\Gamma_j}(S_j), \tag{18}$$

where $T_{j,\gamma} = \Phi_j^\star(\cdot;\gamma)$ and $\widehat{T}_{j,\gamma}^m = \Phi_j^m(\cdot;\gamma)$ use their respective program spaces. Each reference fixed-program map has a deterministic L_j-Lipschitz envelope throughout the compared region, uniformly over admissible programs. This does not assume a hard router is Lipschitz. Let $A_j = \prod_{r=j+1}^{H-1} L_r$, with empty product one. At the *same current model state*, define

$$a_j^m = d\Big(\widehat{T}_{j,\iota_j^m(\Gamma_j)}^m(\widehat{S}_j^m), T_{j,\Gamma_j}(\widehat{S}_j^m)\Big), \tag{19}$$

$$r_j^m = d\Big(\widehat{T}_{j,\widehat{\Gamma}_j^m}^m(\widehat{S}_j^m), \widehat{T}_{j,\iota_j^m(\Gamma_j)}^m(\widehat{S}_j^m)\Big). \tag{20}$$

The first compares fitted and target implementations under aligned instructions. The second changes only the plan at that same state, avoiding a second charge for earlier state error. With $e_j^m = d(\widehat{S}_j^m, S_j)$, the triangle inequality gives

$$e_{j+1}^m \leq L_j e_j^m + a_j^m + r_j^m.$$

Evaluation expectations condition on fitted models and average fresh trajectories and routing randomness, writing as $\rho_j^m = \mathbb{E}_{\text{eval}} r_j^m$. The random intended program remains inside the discrepancies. Role-probe risk controls rollout error only through a further bridge.

Assumption 3 (Common role-aggregation domination)**.** *The role distributions, event-lift losses, target hybrids, and expert classes are fixed independently of the expert-estimation sample. There are deterministic nonnegative weights $\omega_{j,a}$, also fixed independently of that sample and valid for both comparators, such that, conditional on the fitted models,*

$$\mathbb{E}_{\text{eval}} a_j^m \leq \tau_j^m + \sum_{a \in \mathcal{J}} \omega_{j,a} R_a\Big(\widehat{f}_{\alpha_m(a)}^m\Big), \qquad \tau_j^m \geq 0. \tag{21}$$

The weights are fixed domination envelopes, not realized gates. Dynamic routing still requires justified occurrence, sensitivity, transfer, and assembly control. Appendix B.4 gives occupancy/transfer sufficient conditions. Proposition 42 (Appendix M.4) charges additive support mismatch.

The role risks are measured on fixed probes. In the sufficient construction of Appendix B.4, each $\kappa_{j,q,a}$ for controlled occurrence q of role a bounds both occurrence mass and distribution shift. Summing these bounds yields the common weights $\omega_{j,a}$, rather than empirical call frequencies. The term τ_j^m retains surrounding-assembly mismatch after target experts replace the controlled calls. In contrast, ρ_j^m averages the effect of changing only the plan at the same current fitted-model state. Within-event effects are accounted for by the declared lifts and weights. A_j propagates the resulting event error through later events.

For each role set $w_a = \sum_j A_j \omega_{j,a}$, within each reuse group retain $w_a > 0$, put $W_e = \sum_{a \in \mathcal{J}_e} w_a$, and set $\pi_{e,a} = w_a / W_e$. Only $W_e > 0$ groups enter below. All oracles, reuse biases, radii, and empirical objectives are recomputed on this active support. The full convention, including zero-weight roles influencing training, is in Section E.2.

Theorem 4 (Plan-level terminal certificates)**.** *Let initial states agree and let scalar terminal utility $\mathcal{U}$ be $L_{\mathcal{U}}$-Lipschitz. Assume Assumption 3 and all conditions of Theorem 8 on every active group: fixed finite classes, losses in $[0,1]$, fixed probe laws and performance weights independent of fitting data, at least one observation per active role, and mutually independent role observations. Fit matched performance-weighted ERMs from the same role samples. For $\delta \in (0,1)$ choose $\delta_e > 0$, $\sum_e \delta_e \leq \delta$,*

and use the radii of Equations (43) and (44) at δ_e, with $\epsilon_{\mathrm{reuse},e} = \epsilon_{\mathrm{sh},e}$. With probability at least $1-\delta$ over fitting data, simultaneously for both models,

$$\begin{aligned}\mathbb{E}_{\mathrm{eval}}|\mathcal{U}(\widehat{S}_H^m) - \mathcal{U}(S_H)| &\le L_{\mathcal{U}} U_m,\\ U_m &= \sum_j A_j(\tau_j^m + \rho_j^m)\\ &\quad + \sum_e W_e\big[R^*_{\mathrm{sep},e} + \mathbf{1}\{m=\mathrm{reuse}\}\mathsf{B}_e^{\mathrm{reuse}} + 2\epsilon_{m,e}(\delta_e)\big].\end{aligned} \tag{22}$$

U_m is a state-error bound, not the utility $\mathcal{U}$. The expanded bounds and probability accounting are in Section E.2. Common oracles cancel in the state-certificate gap $\mathsf{PRG}_H = U_{\mathrm{sep}} - U_{\mathrm{reuse}}$:

$$\begin{aligned}\mathsf{PRG}_H &= \sum_e W_e\Big[2\{\epsilon_{\mathrm{sep},e}(\delta_e) - \epsilon_{\mathrm{sh},e}(\delta_e)\} - \mathsf{B}_e^{\mathrm{reuse}}\Big]\\ &\quad - \sum_j A_j\Big[(\tau_j^{\mathrm{reuse}} - \tau_j^{\mathrm{sep}}) + (\rho_j^{\mathrm{reuse}} - \rho_j^{\mathrm{sep}})\Big].\end{aligned} \tag{23}$$

Thus $\mathsf{PRG}_H > 0$ orders these state-error bounds. Strict utility-bound ordering also requires $L_{\mathcal{U}} > 0$. It does not order unknown true risks or directly bound discrete answer accuracy. Model-specific weights prevent this compact cancellation. Approximate ERM, mismatched training weights, and zero-weight training influences retain their explicit algorithmic-deviation terms (Proposition 41). Architecture-specific route costs and margin feedback are in Sections I and J. Independent actual-performance evaluation is in Proposition 45 (Appendix M.7).

6 Why Route Labels Do Not Establish Causal Reuse

The preceding comparisons concern predictive risk and error bounds. In this section, we discuss whether behavior can be attributed to experts implementing executable operations. Correct expert labels do not always establish operations. We separate labels from intervention fidelity, sufficient state information, and audit coverage.

Let C be the complete executable context and $S = \Psi(C)$ a macro state extracted by a fixed map. A route action $\mathcal{A}_i$ specifies a non-anticipating policy and bounded stopping time. Execution produces $\mathfrak{T}_i(C)$ and macro result $T_i(C) = \Psi(\mathfrak{T}_i(C))$. Deterministic statements use greedy execution or a common fixed realization of exogenous randomness. The full policy and stochastic formulations are in Sections F and L. For a candidate operator $F_i : \mathcal{S} \to \mathcal{S}$ and fixed pre-action context law ν_i, define

$$\mu_i = \mathbb{E}_{C\sim\nu_i} d(T_i(C), F_i(\Psi(C)))\,. \tag{24}$$

This is prediction error for an *intervened effect* using the pre-action state. Small defect also needs non-collapse between actions to support a reusable-operation claim. A single expert call and a multi-layer compound action are different attribution targets.

Proposition 5 (Observational alignment does not imply causal reuse)**.** *For any finite family of target operators $\{F_i\}_{i=1}^K$ with outputs in a common real vector space, there exists a residual routed model with exact task output and route-label accuracy one, while its output under forced expert index j is independent of j. Under a distribution over requested index I, state S, and forced index J, its forced-route fidelity to F_J is exactly*

$$\mathbb{P}(F_I(S) = F_J(S))\,, \tag{25}$$

which equals $1/K$ when J is uniform and independent of (I,S) and the K operator outputs are pairwise distinct at every evaluated state.

The shared path computes $F_I(S)$ and every expert contributes zero, while the natural router selects I perfectly. Forcing J leaves the requested operation I and pre-action prefix unchanged, so the output remains $F_I(S)$. The $1/K$ specialization requires both independent uniform forcing and pairwise distinct target outputs at every evaluated state.

A compressed state may also omit information needed to predict an action. Proposition 14 in Appendix F gives a worst-case diameter obstruction, while Proposition 15 gives an average squared-loss conditional-variance identity. The worst pair and probability-weighted average treat rare aliases

differently. The extractor, actions, operators, and alignment must be fixed independently of the audit, and off-policy fidelity needs further control to transfer to on-policy contexts.

Randomized action-wise audits require supported actions and known forcing propensities(Section K). Their population comparison concerns action-specific $f_A(S)$ versus action-unaware $f(S)$. It equals a local sharing comparison only with matched inputs, classes, losses, and distributions (Section K.1). Physical tying alone does not remove action information encoded in the expert input. These causal checks remain separate from the paired terminal-performance audit.

7 Experiments: From Local Gains to Final Outputs

We test whether reducing local error improves the final output, first in controlled estimation, and then in trained modern Transformers architecture. The latter compose four operations on $\mathbb{Z}_{31}$ using four-layer, width-256 causal models trained at horizons 4 and 8. We distinguish teacher-forced one-step error from free-running final-state error. Methods, full results, and evaluation scope are in Section G.

One-step gains need not yield detectable final-output gains. A training-data sweep compares the same trained model pairs on both endpoints. Figure 1 shows cross-depth tying versus layer-local experts for all ten seeds at all four budgets of the balanced follow-up. At 64 training traces, tying reduces teacher-forced step error by 3.49–7.72 percentage points, with a positive lower bound in every seed, so the local gains are non-trivial. However, final-output differences range only from -0.186 to $+0.192$ points and every interval crosses zero, which shows that no statistically meaningful output gain is observed. Both models remain near chance on the final endpoint in this low-data regime. This is a concrete example, align with our maths abstract construction, which shows that there exists the case when *local improvement without a statistically meaningful final improvement*. Other budgets exhibit larger final changes and seed-dependent directions.

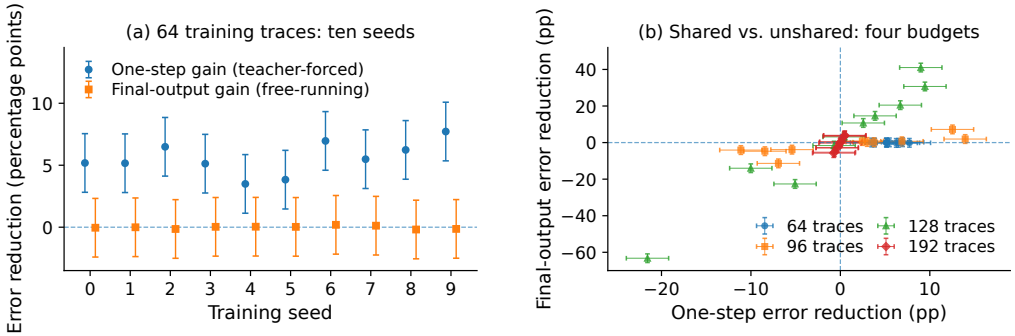

Figure 1: **Local and final gains in trained Transformers.** Positive differences mean lower error for tied experts. (a) At 64 traces, one-step gains are supported, while the final-output difference is not statistically resolved. (b) All forty tying contrasts at 64, 96, 128, and 192 traces. Bars retain the reported simultaneous intervals of the full 240-endpoint follow-up family ($N = 32768$, 480 tails, $\delta = 0.05$, radius $= 2.3657$)

8 Conclusion

Expert sharing must be assessed at three levels: local estimation, final-output performance, and intervention effects. Locally it trades estimation savings against role incompatibility. Common fixed linear propagation makes the terminal trade-off exact and permits local gains to reverse. General routed programs require explicit transfer, assembly, and routing control. Causal reuse additionally requires supported interventions and a sufficient pre-action state. The experiments make the local-to-final distinction concrete: a controlled program reverses the risk ordering, while trained models exhibit supported one-step gains without resolved final-output gains.

AI USE STATEMENT

In this work, generative AI tools were used for language editing, code development, and the drafting of candidate mathematical derivations. The authors independently verified the final mathematical content, including assumptions, derivations, propositions, and proofs. The final proofs, citations, code, and experiment provenance were also manually checked or rewritten by the authors. The authors take full responsibility for the final content of this work, including the text, claims, mathematical results, code, and all other materials produced with the assistance of generative AI.

ETHICS STATEMENT

The reported experiments use generated symbolic data and locally trained Transformer models, rather than personal data or a deployed decision system. More efficient conditional computation can reduce resource use, but it can also lower the cost of deploying capable models. Future applications should report both efficiency and capability changes.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs are given in the corresponding theoretical appendices, with a consolidated reading guide in the appendix overview. Section G specifies the tasks, model architectures, optimization settings, frozen-feature estimators, interventions, and evaluation procedures. Figure captions and Sections G.6 and G.7 state the sampling units, comparison families, and uncertainty conventions. The supplemented materials preserves the experimental configurations and summaries together with scripts for regenerating the tables and figures.

REFERENCES

Alvin J Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, pp. 642–645, 1970.

Yilong Chen, Naibin Gu, Junyuan Shang, Zhenyu Zhang, Yuchen Feng, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, et al. Mixture of universal experts: Scaling virtual width via depth-width transformation. *arXiv preprint arXiv:2603.04971*, 2026.

Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6074–6114. PMLR, 2023.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51 (5):2015–2039, 2023.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

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

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Bruce E Hansen. Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132, 2016.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Minbin Huang, Han Shi, Chuanyang Zheng, Yimeng Wu, Guoxuan Chen, Xintong Yu, Yichun Yin, and Hong Cheng. Unipool: A globally shared expert pool for mixture-of-experts. *arXiv preprint arXiv:2605.06665*, 2026.

Xinmeng Huang, Kan Xu, Donghwan Lee, Hamed Hassani, Hamsa Bastani, and Edgar Dobriban. Optimal multitask linear regression and contextual bandits under sparse heterogeneity. *Journal of the American Statistical Association*, 120(552):2262–2275, 2025.

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

Martin Jaggi. Tying the loop–tied expert layers in mixture-of-experts language models. *arXiv preprint arXiv:2606.16825*, 2026.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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

Lingjing Kong, Xin Liu, Guangyi Chen, Martin Q Ma, Xiangchen Song, Yuekai Sun, Mikhail Yurochkin, Taylor W Killian, Ruslan Salakhutdinov, Kun Zhang, et al. From reasoning traces to reusable modules: Understanding compositional generalization in language model reasoning. *arXiv preprint arXiv:2606.18089*, 2026.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pp. 6265–6274. PMLR, 2021.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations*, volume 2024, pp. 11911–11943, 2024.

BG Manjunath and Stefan Wilhelm. Moments calculation for the doubly truncated multivariate normal density. *Journal of Behavioral Data Science*, 1(1):17–33, 2021.

Yuzo Maruyama and Akimichi Takemura. Non-minimaxity of debiased shrinkage estimators. *Japanese Journal of Statistics and Data Science*, 7(1):361–375, 2024.

Takeru Matsuda and William E Strawderman. Estimation under matrix quadratic loss and matrix superharmonicity. *Biometrika*, 109(2):503–519, 2022.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, volume 2024, pp. 7690–7706, 2024.

Vinoth Nandakumar, Yongli Xiang, Yunzhi Yao, Peike Li, and Tongliang Liu. A theoretical model for task routing in mixture-of-expert transformers. *arXiv preprint arXiv:2606.14398*, 2026.

Joan Puigcerver, Rodolphe Jenatton, Carlos Riquelme, Pranjal Awasthi, and Srinadh Bhojanapalli. On the adversarial robustness of mixture of experts. *Advances in neural information processing systems*, 35:9660–9671, 2022.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pp. 28435–28445, 2024.

Eric S Qiu, Utku Umur Acikalin, Justin Lovelace, Christian Belardi, Arjun B Mulchandani, Carla P Gomes, and Kilian Q Weinberger. More: Mixture of reused experts. *arXiv preprint arXiv:2609.18176*, 2026.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

K Pavan Srinath and Ramji Venkataramanan. Cluster-seeking james–stein estimators. *IEEE Transactions on Information Theory*, 64(2):853–874, 2017.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Zheyue Tan, Zhiyuan Li, Tao Yuan, Dong Zhou, Weilin Liu, Yueqing Zhuang, Yadong Li, Guowei Niu, Cheng Qin, Zhuyu Yao, et al. Rexmoe: Reusing experts with minimal overhead in mixture-of-experts. *arXiv preprint arXiv:2510.17483*, 2025.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International conference on machine learning*, pp. 10434–10443. Pmlr, 2021.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Mingze Wang et al. On the expressive power of mixture-of-experts for structured complex tasks. *Advances in Neural Information Processing Systems*, 38:68579–68606, 2026.

Zihan Wang, Rui Pan, Jiarui Yao, Robert Csordas, Linjie Li, Lu Yin, Jiajun Wu, Tong Zhang, Manling Li, and Shiwei Liu. Chain-of-experts: unlocking the communication power of mixture-of-experts models. *arXiv preprint arXiv:2506.18945*, 2025.

Jinze Zhao, Peihao Wang, and Zhangyang Wang. Generalization error analysis for sparse mixture-of-experts: A preliminary study. *arXiv preprint arXiv:2403.17404*, 2024.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

APPENDIX OVERVIEW AND READING PATHS

The appendices separate proofs of the main claims from architectural specializations and evaluation protocols. Readers can use the questions below to choose a path.

Understand the problem settings and argument. Start with the worked Transformer example, then return to Sections 4–6. Consult Section A to distinguish exact risk, a certificate, and an audit.

Check a main claim. Read Section D for local identities, Sections C and E for statistical and terminal bounds, Section N for terminal geometry, and Section F for causal obstructions. Use Section Q alongside the full statements.

Choose an extension or evaluation. Use H–P selectively for the relevant architecture, geometry, or intervention. Section G reports local-to-final experiments, intervention supplements, and evaluation scope. The paired audit in Appendix M.7 addresses actual performance.

	Question and entry point	Page
A	Symbols and claim types: what does this term or guarantee mean?	15
B	Roles and event lifts: how do the definitions fit a concrete model?	18
C	Distribution-free statistics: what can independent samples guarantee?	23
D	Exact local frontiers: how are pooling risks and selection justified?	25
E	Terminal certificates: how do local and routing errors propagate?	30
F	Causal actions: why are labels and compressed states insufficient?	33
G	Local-to-final experiments: which gains transfer, and what do intervention tests add?	35
H	Unequal data and cross-depth tying: which effects can be shared?	53
I	Routing architectures: what functional cost does a route change incur?	57
J	Routing feedback: when can an earlier error change later support?	61
K	Intervention audits: which action effects have enough coverage?	65
L	Complementary extensions: what do function corrections and protocols add?	69
M	Relaxed assumptions: what survives design, transfer, or optimization mismatch?	71
N	Terminal geometry: when does a local gain reverse or survive uniformly?	77
O	Nonlinearity and parameters: how are remainders and selected moments controlled?	81
P	Update budgets and replay: what do rank, radius, support, and gains constrain?	89
Q	Assumption audit: which assumptions support each result?	95

A READING GUIDE: SYMBOLS, MATHEMATICAL TERMS, AND CLAIM TYPES

This guide distinguishes risk comparisons, certificates, and audits, and records the paper's notation. Definitions apply throughout; specialized symbols are also introduced next to their equations, and a reused letter has the scope specified in its section.

A.1 THREE TYPES OF GUARANTEES

A *risk* is an expected loss on a specified data distribution. A *population oracle* is the best function in a declared class when that distribution is known; an *oracle tuning coefficient* minimizes an expected risk using unknown population quantities. An *empirical risk* replaces the expectation by a sample average, and an empirical risk minimizer (ERM) minimizes that average. An *excess prediction risk* here measures error relative to the noiseless target operator, excluding irreducible observation noise that would be common to the compared predictors.

An *exact expected-risk comparison* averages over the specified fitting randomness and is an equality or inequality about that model's true risk. A *high-probability certificate* says that a bound is valid for all training samples except a set of probability at most δ. A smaller certificate does not imply smaller actual error: if $r_1 \le 10$ and $r_2 \le 20$, either true error can be smaller. A *held-out audit* estimates the risk difference of already fitted systems on new examples. Theorem 39 illustrates an exact expected-risk comparison; Theorem 4 gives a high-probability certificate; Proposition 45 gives a held-out audit.

A.2 INDEX AND NOTATION CONVENTIONS

Unless stated otherwise, counts are positive integers, sets of roles and experts are finite, all maps and random variables are measurable, and all stated expectations exist. Probability arguments use standard Borel state and probe spaces (measurable spaces equivalent to Borel subsets of a complete separable metric space); finite sets, Euclidean vectors, and finite token sequences satisfy this condition. These routine conditions allow conditional distributions and conditional means to be interpreted consistently.

Symbol or convention	Meaning and scope
$a \in \mathcal{J}$, $e \in [K]$	A recurring role and an expert parameter-block index. $[K] = \{1, \dots, K\}$. A role includes the call distribution and environment; it need not coincide with a human-assigned semantic label.
m, K_m, α_m	Comparator model, its expert count, and its role-to-expert assignment. The compact comparison uses $m \in \{\mathrm{reuse}, \mathrm{sep}\}$.
$\mathcal{J}_e$, M, J	Roles assigned to one reused expert; the number of roles in a local group; the number of role cells in a stacked model. A role cell can be a role–layer pair.
j, H, q, $Q(\gamma)$	Event index $j = 0, \dots, H-1$ and horizon H; within-event call index q and number of calls in program γ. An event need not be one token.
ℓ, L in depth models	Layer index and number of layers. L_j elsewhere is a Lipschitz constant, while a matrix L in Section M is a linear sharing rule.
$X_j = S_j$, $\widehat{X}_j^m = \widehat{S}_j^m$	Reference and fitted-model pre-event states; $\mathcal{S}_j$ is their declared stage space. These are not the regression design matrices X_a.
Γ_j, γ	Random canonical route program and one fixed program value. $\mathfrak{G}_j^\star$ is their space. Canonical programs use role labels rather than arbitrary expert numbering.
ι_j^m, $\widehat{\Gamma}_j^m$	Implementation of a canonical program in model m, and the actually selected model-specific route. These can differ.
Φ_j^m, $T_{j,\gamma}$	Complete model-specific event and reference event under program γ, including dispatch, shared paths, experts, residuals, and aggregation.
$z \in \mathcal{Z}_a$, P_a	A local probe record and its reference distribution. A record supplies the information needed to execute a controlled call and compare its effect.

Symbol or convention	Meaning and scope
$\mathcal{T}_a[f]$, $T_a^\star$	The event lift with candidate expert f inserted, and its target counterpart. They have the same output state space even when f returns only a residual vector.
ℓ_a, R_a, $\widehat{R}_a$	One-example loss, population expected role loss, and sample-average role loss. ℓ_a is a function, not a layer index in this context.
$\mathcal{F}_e$, $\pi_{e,a}$	Candidate expert functions and within-group evaluation probabilities, fixed independently of expert-fitting data unless selection is controlled.
$R^*_{\mathrm{sep},e}$, $R^*_{\mathrm{share},e}$	Role-specific and shared population oracle risks. Their difference $\mathsf{B}_e^{\mathrm{reuse}} \geq 0$ is reuse bias for that class and evaluation distribution.
n_a, N, N_e^{eff}	Samples from role a, total independent sample count, and weighted effective size $(\sum_a \pi_{e,a}^2/n_a)^{-1}$. Effective size is not the number of expert calls.
δ, δ_e, $\epsilon_{\mathrm{sh},e}$, $\epsilon_{\mathrm{sep},e}$	Allowed failure probabilities and uniform-deviation radii. Group failure budgets must sum to at most the advertised total.
w_a, W_e, $\omega_{j,a}$	Terminal role importance, its sum within a group, and event-level occurrence/transfer weight. The group evaluation weights are $\pi_{e,a} = w_a/W_e$ on positive-weight roles.
a_j^m, r_j^m, τ_j^m, ρ_j^m	Correctly aligned local approximation error, route-induced local error, residual assembly error, and expected route error. They measure different mechanisms.
A_j, U_m, $\mathcal{U}$, $L_{\mathcal{U}}$	Downstream amplification product, a state-error certificate, terminal utility, and its Lipschitz constant. U_m is a bound, not utility.
PRG_H	The no-sharing state-bound minus the reuse state-bound under their stated matching assumptions. Its sign orders those bounds only.
C, Ψ, F_i, T_i	Complete executable context, macro-state extractor, proposed macro operator, and the actual macro transition after intervention. Here C is not the variance saving in the linear model.
$\mathcal{A}_i = (\varpi_i, \sigma_i)$	A route action: online route policy and bounded stopping time. K in the causal sections counts actions, which may be compound multi-expert interventions.
ν_i, μ_i, Ξ_i	Probe-context law, average intervention defect, and largest action-effect difference between contexts aliased by Ψ.
$p_a(s)$, $e_a(C)$	Declared state-based intervention distribution for a population identity, and known action-sampling propensity in an audit. They are distinct experimental policies.
β, β_j	Audit overlap floor (a probability), versus common-state router-logit error bound. They have different meanings and units.
γ_k, Δ_j, ζ_j	Top-k support margin, extra functional severity of a support change, and route cost remaining even when supports agree.

A hat marks an estimate, fitted object, empirical quantity, or state/route realized by a fitted model; Φ_j^m and $\mathcal{R}_j^m$ also belong to the fitted model despite having no hat; $\star$ marks a reference or oracle, specified locally. $\arg\min$ is the set of minimizing arguments; $\inf$ is the greatest lower bound even when no minimizer is attained. Empty sums are zero and empty products are one. $\mathbf{1}\{B\}$ equals one when event B occurs and zero otherwise. $\mathrm{Law}(X)$ means the probability distribution of X, and "almost surely" means except on an event of probability zero.

A.3 Linear-model notation

For a matrix A, $\|A\|_F$ denotes its Frobenius norm and $\|A\|_2$ its operator norm. The common-Hilbert-risk assumption fixes one shared geometry for all role risks. The design Gram matrix $X^\top X$ and the propagation Gramian $P_j^\top P_j$ describe different geometries.

Linear-model symbol	Definition
$A_a \in \mathbb{R}^{p\times d}$, $D = pd$	True input-to-output operator and its scalar coefficient count.
X_a, Y_a, $\mathcal{E}_a$	Fixed $n_a \times d$ design matrix, $n_a \times p$ observed responses, and response-noise matrix in $Y_a = X_a A_a^\top + \mathcal{E}_a$.
$\widehat{A}_a$, E_a	Fitted OLS matrix and its parameter error $E_a = \widehat{A}_a - A_a$; this E_a is not an expert function $E_{\theta,e}$.
σ_a^2, v_a	Response-noise variance and coefficient-estimation variance $v_a = \sigma_a^2/n_a$ in the orthogonal design. Equal noise levels are homoscedastic; unequal ones are heteroscedastic.
B, V, C	Balanced squared sharing bias, no-sharing risk $V = pd\sigma^2/n$, and variance saving $C = V(1 - 1/M)$.
λ, $\lambda^\star$	Fixed partial-pooling strength and its population-optimal value. Covariance eigenvalues λ_k and resource multipliers λ_{res} are unrelated parameters.
Π, G, H, K_Π	Partition, two of its disjoint groups, and number of groups. H here is a group name, not the trace horizon.
B_{role}, B_{depth}, B_{int}	Between-role main-effect, depth main-effect, and residual interaction squared variations in the balanced role–depth array.
θ, Y, ζ, Σ	Stacked true parameters, their noisy base estimate, centered estimation noise, and its full covariance in the general linear observation model.
W, L, P, Q	Evaluation weight matrix, fixed linear sharing rule, weighted sharing projector, and discarded-direction projector $Q = I - P$ in Section M.
$\mathbf{e}_j$, ξ_j, M_j, P_j	State-error vector, injected local error, linear propagation matrix, and product of later matrices.
b_j, Σ_{jk}, K_e	Mean injected error, cross-time covariance, and sum of all terminally propagated effects of one shared expert error.

A.4 Transfer and transport conventions

The 1-Wasserstein distance is

$$W_1(P, Q; d) = \inf_{\gamma\in\Pi(P,Q)} \int d(x, y)\, d\gamma(x, y), \tag{26}$$

where $\Pi(P, Q)$ is the set of couplings. The finite first-moment condition is $\mathbb{E}d(X, x_0) < \infty$ for one fixed point x_0. The coupling notation $\Pi(P, Q)$ is unrelated to a partition Π. In the discrete route transport of Equation (129), costs are distances between expert outputs, not between their labels.

A.5 Intervention conventions

The propensity $e_a(C)$ is the probability of actually sampling action a at context C, not generally the router's softmax score under deterministic top-k selection. Overlap requires $e_a(C) \geq \beta > 0$ on every claimed region/action. Changing the intervention policy or target distribution changes the estimand. A Markov-sufficient macro state makes the transition law depend on the complete context only through that state and the chosen action; otherwise kernels must be defined on the complete context.

A.6 Computational vocabulary and comparison objects

A role groups specified call environments with a probe law and target; an event contains the calls and surrounding operations being compared. A reasoning step is a task-level operation, with no assumed one-to-one relation to a token, layer, call, or event. The independent statistical unit is the entire independently sampled record, often a complete trace rather than an individual call.

Support replay fixes selected indices but recomputes their gains at the replay's own states. *Frozen-gain replay* also fixes recorded gains; natural execution recomputes its selections. Freezing a feature law fixes both the feature map and its evaluation input distribution.

An estimation-risk frontier averages fitting randomness; a terminal-risk frontier also includes the map carrying errors to the endpoint. An approximation frontier minimizes population discrepancy over a feasible class without charging estimation or optimization.

A.7 Technical terms used by the extensions

Function-subspace dimension, matrix rank, factor norms, and the norm of an increment $BA - B_0A_0$ are distinct constraints. A fixed core is not optimized, whereas a free core is. A feature second moment is $\mathbb{E}[\varphi\varphi^\top]$, not a centered covariance. Conditioning on a selection cell requires recomputing conditional means and covariances. A full-law risk averages both non-escape and escape outcomes; derivative envelopes must hold throughout the specified region, not only at the expansion point.

B Routed Computation Graphs, Roles, and Event Lifts

Overview. This appendix explains what is shared and how its effect is compared inside a complete computation. The worked example makes the roles, assignments, and event lifts concrete; the remaining sections give the precise interfaces and sufficient conditions connecting local probes to rollout errors. After the example, Sections B.1–B.2 define programs and role risks, Sections B.3–B.4 give the replacement and transfer arguments, and Section B.5 distinguishes the levels of reuse.

Worked example: two-layer expert reuse

One prediction through two layers. Consider the request "current state 3, next operation multiply by two." The task in Section G uses $F_2(s) = 2s \bmod 31$, so the target next state is $F_2(3) = 6$. This is one operation within a supplied program. To follow the internal computation, use the illustrative two-layer causal Transformer: width 128, four attention heads, and four candidate SwiGLU experts of intermediate width 384 per layer. Both layers contribute to this one prediction; they do not perform two successive multiplications. Each receives the position's hidden representation, not the integer 3 as a direct expert input. The primary trained comparators in Section G.1 instead have four layers and hidden width 256. The example uses declared reference functions and probes.

Two roles and their assigned experts. For this request, the controlled call type is $k = i + 1 = 3$, since the task operation index is $i = 2$. The chosen prediction position gives role $(1, 3)$ in the first layer and role $(2, 3)$ in the second. The separate model uses each layer's own local expert 3—two independent parameter blocks, numbered 3 and 7 when flattened. The reuse model assigns shared block 3 to both roles. Its four-expert bank is shared across depths, while routers and normalization remain layer-local; the separate model stores two independent banks. A different input number can instantiate the same role.

This pair belongs to the full eight-role construction. Let $k \in \{1, 2, 3, 4\}$ name a controlled call type and $\ell \in \{1, 2\}$ its layer. At symbolic-task prediction positions, $k = i + 1$ labels the requested operation $i \in \{0, 1, 2, 3\}$. The full assignment is

$$\mathcal{J} = \{(\ell, k) : \ell \in \{1, 2\},\ k \in \{1, 2, 3, 4\}\}, \qquad K_{\text{sep}} = 8, \quad K_{\text{reuse}} = 4,$$

$$\alpha_{\text{sep}}(\ell, k) = 4(\ell - 1) + k, \qquad \alpha_{\text{reuse}}(\ell, k) = k, \qquad \mathcal{J}_k = \{(1, k), (2, k)\}.$$

Call type	Two distinct roles	Separate blocks	Reused block
1	$(1, 1), (2, 1)$	$1, 5$	1
2	$(1, 2), (2, 2)$	$2, 6$	2
3	$(1, 3), (2, 3)$	$3, 7$	3
4	$(1, 4), (2, 4)$	$4, 8$	4

A call occurrence belongs to one role. Paper indices are one-based; code bank positions are $k - 1$. Separate indices 5–8 flatten a layer/index pair, not an eight-way router. At role $(2, 3)$, the reuse assignment specifies block 3, but the natural router may select block 1. This remains role $(2, 3)$, not $(2, 1)$.

Reference updates and task outputs. The target next state 6 does not uniquely specify a 128-dimensional residual at either layer. For the local comparison, specify reference weights independently and let $f_a^\star$ be the corresponding same-interface SwiGLU function. Both fitted models are evaluated against this declared target; the fitted separate model is not the reference. The local function returns a hidden-state residual, whereas $F_i : \mathbb{Z}_{31} \to \mathbb{Z}_{31}$ returns a task state. These local reference weights are not asserted to form a complete network that outputs 6.

A residual difference you can calculate. For arithmetic only, suppose a valid controlled probe fixes the prefix and all other call contributions. The base row has first two coordinates $(1, 0)$, and the expert output is multiplied by a fixed gain $c = 1/2$. Suppose the candidate residual begins $(3, 1)$ and the reference residual begins $(2, 1)$, with all remaining output differences zero. The updated row then begins $(1, 0) + \frac{1}{2}(3, 1) = (2.5, 0.5)$ for the candidate, and $(1, 0) + \frac{1}{2}(2, 1) = (2, 0.5)$ for the reference. Under the complete-state metric below, this gives a single-probe distance of 0.5: the expert returns an update, and the comparison measures its effect after insertion into the surrounding state.

Only the first two coordinates of the 128-dimensional row are displayed; all other coordinate differences are zero. These illustrative values give a single-probe state distance of 0.5.

The complete-state representation. Each $E_{\theta,e} : \mathbb{R}^{128} \to \mathbb{R}^{128}$ returns a residual, with parameters

$$E_{\theta,e}(u) = D_e[\mathrm{SiLU}(V_e u) \odot (U_e u)], \quad U_e, V_e \in \mathbb{R}^{384\times 128}, \quad D_e \in \mathbb{R}^{128\times 384}.$$

For a fixed-length prefix without dropout, take the event to be one full FFN subblock with residual addition. Its pre-event state is $X = (p, \mathbf{H})$, where p is the prefix and $\mathbf{H} \in \mathbb{R}^{n\times 128}$ is the complete pre-FFN array after attention. The expert input is $u = \mathrm{LN}(\mathbf{H}_{t,:})$ at position t. Mask and layer boundary are fixed. This full-prefix implementation uses no KV-cache path. Attention, FFN, and readout can be separate events: the horizon H is not automatically the layer count. $\mathcal{S}_j$ contains the complete records at boundary j; X_j and $\widehat{X}_j^m$ follow the target and fitted trajectories respectively.

A valid $z \in \mathcal{Z}_a$ fixes $p, \mathbf{H}, t, u$, a gain c, and the other call contributions. Let $\mathbf{H}_{\mathrm{base}}(z)$ retain the residual path and all contributions except the controlled call. $\mathrm{Insert}_t(v)$ is an $n \times 128$ zero array with row t equal to $v^\top$. The numerical comparison above is an instance of

$$\mathcal{T}_a[f](z) = (p, \mathbf{H}_{\mathrm{base}}(z) + \mathrm{Insert}_t(cf(u))),$$
$$T_a^\star(z) = (p, \mathbf{H}_{\mathrm{base}}(z) + \mathrm{Insert}_t(cf_a^\star(u))).$$

For same-shaped states use

$$d((p, \mathbf{H}), (p', \mathbf{H}')) = \mathbf{1}\{p \neq p'\} + \|\mathbf{H} - \mathbf{H}'\|_F.$$

Only one row differs on this probe, so $\ell_a(f; z) = |c|\|f(u) - f_a^\star(u)\|_2$. Take P_a uniform on a valid probe bank independently fixed before expert fitting; $R_a(f) = \mathbb{E}_{P_a}\ell_a(f; Z)$ averages these complete-state discrepancies.

Intended and actual execution. A canonical $\Gamma_j \in \mathfrak{G}_j^\star$ names roles, positions, gains or gain rules, and combination for all calls in the event. ι_j^m substitutes the assigned model-specific blocks, producing an aligned plan in $\mathfrak{G}_j^m$. The actual plan is $\widehat{\Gamma}_j^m = \mathcal{R}_j^m(\mathcal{H}_j^m)$, using only visible prefix, activations, and already revealed randomness. The execution maps return complete event states. At one common current state, compare

$\Phi_j^m(\widehat{X}_j^m; \widehat{\Gamma}_j^m)$	actual fitted execution,
$\Phi_j^m(\widehat{X}_j^m; \iota_j^m(\Gamma_j))$	aligned fitted execution,
$\Phi_j^\star(\widehat{X}_j^m; \Gamma_j)$	target at the model state.

The first two isolate routing cost; the last two isolate implementation error. The reference trajectory itself evaluates $\Phi_j^\star$ at X_j, not at $\widehat{X}_j^m$.

The eight illustrated roles cover controlled prediction-position calls, not all errors in a complete rollout; other calls need their own losses or justified assembly terms. The Frobenius metric is generally unbounded. Section B.2 specifies the fixed-probe conditions, and Section P.4 treats fixed-gain replay. See Section A for component terminology and return to Section 3 for the general definitions.

B.1 CANONICAL PROGRAMS AND MODEL-SPECIFIC IMPLEMENTATIONS

Let

$$\mathcal{E}_\theta = \{E_{\theta,e} : \mathcal{X}_e \to \mathcal{Y}_e\}_{e=1}^{K} \tag{27}$$

be a library of parameter blocks. A canonical route program $\gamma \in \mathfrak{G}_j^\star$ is labelled by functional roles rather than by model-specific expert indices. It contains the controlled calls, their dispatched inputs, mixture weights, call order, shared paths, residual operations, aggregation, and—when needed—a finite non-anticipating policy for internal adaptive choices. Each model m has an implementation map

$$\iota_j^m : \mathfrak{G}_j^\star \longrightarrow \mathfrak{G}_j^m \tag{28}$$

that replaces every canonical role occurrence by the expert index, tied parameter block, or subgraph used by that comparator. This map is part of the comparison protocol and aligns canonical roles with each model's executable library.

Conditional on an executable plan and the relevant exogenous randomness, the event map

$$\Phi_j^m : \mathcal{S}_j \times \mathfrak{G}_j^m \to \mathcal{S}_{j+1} \tag{29}$$

is deterministic. A non-anticipating scheduler produces the actually executed plan $\widehat{\Gamma}_j^m$ from information available before its corresponding routing decision. Future random variables may be represented by an exogenous seed in the probability space, but are not made observable to the router merely to obtain a deterministic representation.

Complete interfaces for Equation (1). A *state space* $\mathcal{S}_j$ is the set of admissible complete records immediately before event j. For a Transformer event, a record can include the prefix, the hidden-state array, attention caches, and any earlier values reused by residual connections. A *state* $X_j \in \mathcal{S}_j$ is one such record, whereas $X_{j+1} \in \mathcal{S}_{j+1}$ is the record after the event. The learned state $\widehat{X}_j^m$ is a separate trajectory value in the same declared comparison representation. The spaces $\mathcal{X}_e, \mathcal{Y}_e$ describe one expert's interface and need not equal the event-state spaces. If models use different internal representations, the comparison maps and induced event transitions must be specified; equal output dimensions alone do not supply these maps.

The program spaces $\mathfrak{G}_j^\star$ and $\mathfrak{G}_j^m$ are sets of admissible instructions. A lowercase γ is a fixed member; Γ_j is the possibly random intended program. A program includes the declared input-selection rules, call order, mixing weights or weight-computation rules, and aggregation. *Canonical* means that roles are named consistently across models. The implementation map ι_j^m is fixed as part of this comparison and may send different roles to the same parameter block.

Let $\mathscr{H}_j^m$ be the space of admissible history records at the routing boundary. Then

$$\mathcal{H}_j^m \in \mathscr{H}_j^m, \qquad \mathcal{R}_j^m : \mathscr{H}_j^m \longrightarrow \mathfrak{G}_j^m, \qquad \widehat{\Gamma}_j^m = \mathcal{R}_j^m(\mathcal{H}_j^m).$$

The history record may contain the router's view of the current state, earlier plans and outputs, and random draws already revealed for this decision. It need not expose every coordinate of the complete event state. It is a random record, not itself a sigma-field. If $\mathscr{I}_j^m$ is the information allowed before the decision, non-anticipation requires the history and decision to be measurable with respect to $\mathscr{I}_j^m$. For an internal adaptive policy this requirement applies separately at each internal decision.

An *external random tape* is a collection of random draws used to realize all compared executions on one probability space. Fixing it makes the transitions deterministic as functions of their arguments; only the appropriate revealed portion is visible to a router. The tape argument is suppressed in $\Phi_j^\star$ and Φ_j^m.

There are three distinct event outputs at a common model state: actual execution uses $\Phi_j^m(\widehat{X}_j^m; \widehat{\Gamma}_j^m)$; aligned execution uses $\Phi_j^m(\widehat{X}_j^m; \iota_j^m(\Gamma_j))$; the reference comparison uses $\Phi_j^\star(\widehat{X}_j^m; \Gamma_j)$. The actual reference trajectory instead evaluates the last map at X_j. These insertions are precisely the split used in Equation (74).

In Section 5 all compared stage states are represented in a declared common metric space $(\mathcal{S}, d)$, and the notation is

$$S_j = X_j, \quad \widehat{S}_j^m = \widehat{X}_j^m, \qquad T_{j,\gamma}(s) = \Phi_j^\star(s;\gamma), \quad \widehat{T}_{j,\gamma}^m(s) = \Phi_j^m(s;\gamma).$$

In the first transition $\gamma \in \mathfrak{G}_j^\star$; in the second it lies in $\mathfrak{G}_j^m$. The semicolon merely separates state and plan arguments. The intended program may depend on the reference history; recomputing a reference policy on the learned state would define a different comparator.

The abstraction includes the following common cases.

Architecture	Canonical-program representation
Top-k token routing	One sparse role measure per token; selected calls are combined in parallel.
Expert-choice or balanced routing	A batch- or sequence-level role-assignment matrix is chosen jointly.
Additive soft gate	A normalized role measure weights expert outputs at a common input.
Full Soft MoE dispatch	The program includes dispatch and combine matrices; because slot inputs depend on the assignment, the generic fixed-program bound is used unless extra dispatch sensitivity is controlled.
Shared experts	The program contains always-active calls and conditionally selected calls.
Global/cross-layer pools	Canonical roles are mapped to a shared model-specific pool at several depths.
Tied or recurrent experts	One parameter block implements several canonical occurrences.
Sequential/hierarchical experts	A finite expert program is executed; adaptive choices are represented by a non-anticipating policy or split into separate routed events.

It uses their fixed-program transitions, their implementation maps, and the role-level domination condition stated below. The architecture-specific transport and matching results in Section I apply only to their stated additive or preselected-sequence specializations.

B.2 Roles as fixed call environments

A role $a \in \mathcal{J}$ specifies (i) a probe distribution P_a and (ii) a measurable event lift $\mathcal{T}_a[f]$. A probe record z contains all information needed to execute the controlled call environment: the pre-event state, canonical program, controlled occurrence, revealed randomness, and the fixed values of the other components in the hybrid comparison. The lift returns a post-event model state after inserting candidate function f at the controlled occurrence. The target hybrid $T_a^\star$ has the same codomain. Therefore

$$\ell_a(f; z) = d(\mathcal{T}_a[f](z), T_a^\star(z)) \tag{30}$$

is a distance between two event states, even if f itself maps a token representation to an FFN residual.

For the finite-sample theorem, P_a, $\mathcal{T}_a$, $T_a^\star$, and the expert class are fixed independently of the samples used to estimate the experts, and $\ell_a \in [0, 1]$. A frozen surrounding assembly or an independent sample split is sufficient. If the same observations jointly train the backbone, dispatch, or event lift, one must analyze the joint hypothesis class rather than condition on a data-dependent loss without paying its complexity. Scaling a bounded distance is valid; post-hoc clipping of an unbounded distance controls only the clipped quantity and cannot silently be substituted into an unclipped terminal bound.

A reuse assignment $\alpha_m : \mathcal{J} \to [K_m]$ maps roles to model-specific parameter blocks. The main compact comparison assumes that the canonical roles, evaluation weights, and expert-estimation classes are common across comparators. Data-dependent assignments require either an independent selection split or an additional selection-complexity term.

B.3 A deterministic call-replacement lemma

Fix a state s and one realized canonical program γ with $Q(\gamma)$ controlled call occurrences. For a model m, let $\widetilde{T}_{j,\gamma}^{m,(q)}(s)$ be the hybrid event output in which the first q canonical occurrences use the fitted expert functions and the remaining occurrences use their target functions, while all non-expert components use the fixed model-m assembly. Thus $\widetilde{T}^{m,(Q)}$ is the correctly aligned learned event and $\widetilde{T}^{m,(0)}$ uses target experts inside the model assembly. Define

$$\tau_{j,\gamma}^m(s) = d\Big(\widetilde{T}_{j,\gamma}^{m,(0)}(s), T_{j,\gamma}(s)\Big)\,, \tag{31}$$

which isolates shared-backbone, dispatch, aggregation, or other assembly mismatch.

Lemma 6 (Call-by-call event replacement)**.** *For every s, γ, and model m,*

$$d\Big(\widetilde{T}^{m,(Q)}_{j,\gamma}(s), T_{j,\gamma}(s)\Big) \le \tau^m_{j,\gamma}(s) + \sum_{q=1}^{Q(\gamma)} d\Big(\widetilde{T}^{m,(q)}_{j,\gamma}(s), \widetilde{T}^{m,(q-1)}_{j,\gamma}(s)\Big)\,. \tag{32}$$

Each summand is an event-lift loss for the role of occurrence q under the corresponding hybrid probe record.

Proof. Insert the sequence $\widetilde{T}^{m,(Q-1)}, \dots, \widetilde{T}^{m,(0)}$ and apply the triangle inequality. By construction, adjacent hybrids differ at exactly one controlled occurrence and have outputs in the same post-event state space. The lemma is a fixed-program hybrid statement. If later routing decisions adapt to earlier expert outputs, refine those decisions into separate routed events or treat the entire adaptive policy as one indivisible event lift. Conditioning on one realized trace while replacing an earlier output is valid only when the later decisions are explicitly frozen as part of the intervention; it is not automatically a decomposition of the original adaptive execution. □

B.4 A RIGOROUS SUFFICIENT OCCUPANCY-TRANSFER CONDITION

Condition on the expert-estimation sample and let $Z^m_{j,q}$ be the hybrid probe record of the qth controlled occurrence when that occurrence exists. Assume a deterministic maximum $Q_{\max,j} < \infty$; a summable version gives the same conclusion for an unbounded count. For each role a, define the conditional sub-probability measure

$$Q^m_{j,q,a}(B) = \mathbb{P}\big(Z^m_{j,q} \in B, q \le Q(\Gamma_j), a(j,\Gamma_j,q) = a \mid \mathcal{D}\big)\,, \tag{33}$$

where $\mathcal{D}$ denotes the information in the fitted sample. Formally this is a sigma-field: the collection of events whose truth can be determined from the sample. Conditioning fixes the fitted parameters, not the fresh evaluation trajectory. Suppose there are deterministic constants $\kappa_{j,q,a} \ge 0$, fixed independently of $\mathcal{D}$ and valid for both comparators, such that

$$Q^m_{j,q,a} \ll P_a, \qquad \left\| \frac{dQ^m_{j,q,a}}{dP_a} \right\|_\infty \le \kappa_{j,q,a} \quad \text{a.s.} \tag{34}$$

The notation $Q \ll P$ means $P(B) = 0$ implies $Q(B) = 0$. The Radon–Nikodym derivative dQ/dP is the weight function satisfying $Q(B) = \int_B (dQ/dP)\, dP$; its ∞-norm is the smallest almost-sure upper bound. Because $Q^m_{j,q,a}$ has total mass at most one (a sub-probability measure), $\kappa_{j,q,a}$ already includes both occurrence mass and distribution shift. Integrating each hybrid loss in Equation (32) yields

$$\mathbb{E}[a^m_j \mid \mathcal{D}] \le \tau^m_j + \sum_{a \in \mathcal{J}} \omega_{j,a} R_a\Big(\widehat{f}^m_{\alpha_m(a)}\Big)\,, \qquad \omega_{j,a} = \sum_{q=1}^{Q_{\max,j}} \kappa_{j,q,a}, \tag{35}$$

where

$$\tau^m_j = \mathbb{E}\Big[\tau^m_{j,\Gamma_j}(\widehat{S}^m_j) \mid \mathcal{D}\Big]\,. \tag{36}$$

Indeed, $\int \ell_a(\widehat{f}; z)\, dQ^m_{j,q,a}(z) \le \kappa_{j,q,a} R_a(\widehat{f})$. This is one sufficient condition for Assumption 3. Proposition 42 permits support mismatch by retaining an additive discrepancy term. More refined lifts can absorb within-event suffix sensitivity and give smaller weights. If only model-specific envelopes $\kappa^m_{j,q,a}$ are credible, each comparator must retain its own terminal weights and full certificate.

B.5 HIERARCHY OF REUSE CLAIMS

1. *Parameter reuse*: identical weights occur in multiple calls.
2. *Statistical reuse*: samples from several roles jointly estimate those weights.
3. *Functional reuse*: one function has low role risk on the stated role distributions.
4. *Causal reuse*: a pre-action route intervention induces a stable macro transition.
5. *Compositional reuse*: that transition remains accurate in previously unseen route sequences or horizons.

Only the first item is an architectural fact. Each subsequent item requires additional empirical or mathematical conditions.

Partial pooling versus physical tying. The interpolation in Equation (5) shares statistical information, but

$$\widehat{A}_a^{(\lambda)} - \widehat{A}_b^{(\lambda)} = (1-\lambda)(\widehat{A}_a - \widehat{A}_b).$$

For $0 < \lambda < 1$, these matrices need not coincide. Writing them as a common core plus role-specific residuals is only a reparameterization unless the residual capacity is restricted and counted. Full pooling and each block of a hard partition do use one common matrix; intermediate shrinkage does not by itself reduce the number of stored role matrices. Theorems about estimation risk, approximation under an update constraint, and physical storage therefore address different objects.

The expert's information is part of its function class. The common operator in the local model receives x, not a separate role identifier. A class of functions $f(x, a)$ can implement role-dependent operations with one parameter block and is not the same class as the shared $f(x)$ in that model. A hidden feature can also encode role or depth without an explicit identifier. Thus a fixed feature map, its input law, the permitted conditioning information, and any adapter must be declared before interpreting a sharing bias.

C Distribution-Free Statistical Theory

Overview. We compare fitting one function for a group with fitting a separate function for each role. Reuse bias measures the population cost of the shared restriction, while concentration controls estimation from independent records. Section C.1 interprets reuse bias, Section C.2 proves the finite-class comparison, Section C.3 examines sample allocation, and Section C.4 gives uniform-deviation and trace-level templates.

C.1 Reuse bias and its functional interpretation

Fix a group $\mathcal{J}_e$ and a class $\mathcal{F}_e$. The role distributions, losses, class, weights, and sample counts are fixed independently of the observations used by the ERMs. For weights $\pi_{e,a} > 0$ summing to one, define the shared and separate population oracles as in Equation (15). Nonnegativity of $\mathsf{B}_e^{\mathrm{reuse}}$ follows because the shared class is the diagonal subset of the product class $\prod_{a\in\mathcal{J}_e} \mathcal{F}_e$. The concentration theorem below assumes $\ell_a(f; Z) \in [0, 1]$ for every role and candidate; if the surrounding event lift is learned from the same sample, it must be included in the hypothesis class or learned on an independent split.

A useful exact interpretation holds under a common quadratic geometry.

Proposition 7 (Reuse bias as functional variance)**.** *First take the unrestricted candidate class $\mathcal{F}_e = \mathcal{H}$, where $\mathcal{H}$ is a real Hilbert space, and suppose*

$$R_a(f) = c_a + \|f - f_a^\star\|_{\mathcal{H}}^2, \qquad f \in \mathcal{H}. \tag{37}$$

With $\bar{f} = \sum_a \pi_{e,a} f_a^\star$,

$$\mathsf{B}_e^{\mathrm{reuse}} = \sum_a \pi_{e,a} \left\|f_a^\star - \bar{f}\right\|_{\mathcal{H}}^2 = \frac{1}{2}\sum_{a,b} \pi_{e,a}\pi_{e,b} \|f_a^\star - f_b^\star\|_{\mathcal{H}}^2. \tag{38}$$

More generally, for any nonempty class $\mathcal{F} \subseteq \mathcal{H}$ and risks $R_a(f) = c_a + \|f - t_a\|_{\mathcal{H}}^2$, set $\bar{t} = \sum_a \pi_{e,a} t_a$ and $d(u, \mathcal{F}) = \inf_{f\in\mathcal{F}} \|u - f\|_{\mathcal{H}}$. The restricted-class reuse bias is

$$B_{\mathcal{F}} = \sum_a \pi_{e,a} \|t_a - \bar{t}\|_{\mathcal{H}}^2 + d(\bar{t}, \mathcal{F})^2 - \sum_a \pi_{e,a} d(t_a, \mathcal{F})^2. \tag{39}$$

No convexity or attainment of the infima is required. In particular, if every $t_a \in \mathcal{F}$ and $\bar{t} \in \mathcal{F}$ (or its closure), only the variance term remains.

Proof. Write $f - f_a^\star = (f - \bar{f}) + (\bar{f} - f_a^\star)$. The identity $\sum_a \pi_{e,a}(\bar{f} - f_a^\star) = 0$ cancels the summed cross term, giving

$$\sum_a \pi_{e,a} \|f - f_a^\star\|_{\mathcal{H}}^2 = \|f - \bar{f}\|_{\mathcal{H}}^2 + \sum_a \pi_{e,a} \|\bar{f} - f_a^\star\|_{\mathcal{H}}^2.$$

Only the first term depends on the shared function, so it is minimized at $f = \bar{f}$. The constants c_a cancel when subtracting the separate oracle risk. Expanding the weighted double sum gives

$$\sum_{a,b} \pi_{e,a}\pi_{e,b}\|f_a^\star - f_b^\star\|_{\mathcal{H}}^2 = 2\left(\sum_a \pi_{e,a}\|f_a^\star\|_{\mathcal{H}}^2 - \|\bar{f}\|_{\mathcal{H}}^2\right),$$

which yields the pairwise form. For the restricted identity, the same completed-square equality with t_a holds for every $f \in \mathcal{F}$. Taking its infimum gives the shared risk $\sum_a \pi_{e,a}c_a + \sum_a \pi_{e,a}\|t_a - \bar{t}\|^2 + d(\bar{t}, \mathcal{F})^2$. Subtract the separate infima $\sum_a \pi_{e,a}[c_a + d(t_a, \mathcal{F})^2]$ to obtain Equation (39). □

The proposition motivates functional probes.

A finite-class obstruction to the pure-variance interpretation. Let $\mathcal{H} = \mathbb{R}$, $\mathcal{F} = \{0, 1\}$, $\pi_1 = \pi_2 = 1/2$, and $R_1(f) = f^2$, $R_2(f) = (f-1)^2$. Losses on the candidate class are in $[0, 1]$ and both role targets are realizable. Nevertheless,

$$B_{\mathcal{F}} = \frac{1}{2}, \qquad \sum_a \pi_a(t_a - \bar{t})^2 = \frac{1}{4}, \qquad d(\bar{t}, \mathcal{F})^2 = \frac{1}{4}. \tag{40}$$

The missing mean $\bar{t} = 1/2$ accounts for the gap. A common squared-loss geometry alone therefore does not imply the pure-variance claim.

C.2 Finite-class comparison and proof

Now the question is sampling error: how far can every candidate's empirical loss be from its population loss? A simultaneous event controls both fitted comparators.

Theorem 8 (Simultaneous finite-class comparison)**.** *Assume the losses from Equation (3) lie in $[0, 1]$, $\mathcal{F}_e$ is finite, $n_a \geq 1$, and all role samples are mutually independent. For $\delta \in (0, 1)$, with probability at least $1 - \delta$, the shared and role-specific empirical risk minimizers (ERMs) satisfy*

$$\sum_a \pi_{e,a} R_a(\widehat{f}_e) \leq R^*_{\mathrm{sep},e} + \mathsf{B}_e^{\mathrm{reuse}} + 2\epsilon_{\mathrm{sh},e}, \tag{41}$$

$$\sum_a \pi_{e,a} R_a(\widehat{f}_a) \leq R^*_{\mathrm{sep},e} + 2\epsilon_{\mathrm{sep},e}, \tag{42}$$

where

$$\epsilon_{\mathrm{sh},e} = \sqrt{\frac{\log(4|\mathcal{F}_e|/\delta)}{2N_e^{\mathrm{eff}}}}, \tag{43}$$

$$\epsilon_{\mathrm{sep},e} = \sum_a \pi_{e,a}\sqrt{\frac{\log(4|\mathcal{J}_e||\mathcal{F}_e|/\delta)}{2n_a}}. \tag{44}$$

Thus the displayed shared certificate is smaller whenever $\mathsf{B}_e^{\mathrm{reuse}} < 2(\epsilon_{\mathrm{sep},e} - \epsilon_{\mathrm{sh},e})$.

Proof. For notational simplicity suppress e. Role a supplies independent observations $Z_{a,1}, \ldots, Z_{a,n_a} \sim P_a$, and

$$\widehat{R}_a(f) = \frac{1}{n_a}\sum_{i=1}^{n_a} \ell_a(f; Z_{a,i}), \qquad \widehat{R}_{\mathrm{sh}}(f) = \sum_a \pi_a \widehat{R}_a(f). \tag{45}$$

For fixed f, weighted Hoeffding gives

$$\mathbb{P}\left(\left|\widehat{R}_{\mathrm{sh}}(f) - \sum_a \pi_a R_a(f)\right| > t\right) \leq 2\exp\left(-2N_e^{\mathrm{eff}}t^2\right), \tag{46}$$

where $(N_e^{\mathrm{eff}})^{-1} = \sum_a \pi_a^2/n_a$. A union bound over $\mathcal{F}_e$ with failure budget $\delta/2$ yields

$$\sup_{f \in \mathcal{F}_e}\left|\widehat{R}_{\mathrm{sh}}(f) - \sum_a \pi_a R_a(f)\right| \leq \epsilon_{\mathrm{sh},e}. \tag{47}$$

If $f^\star_{\mathrm{sh}}$ is a shared population minimizer and $\widehat{f}_e$ is a shared ERM, then

$$\begin{aligned}\sum_a \pi_a R_a(\widehat{f}_e) &\le \widehat{R}_{\mathrm{sh}}(\widehat{f}_e) + \epsilon_{\mathrm{sh},e} \\ &\le \widehat{R}_{\mathrm{sh}}(f^\star_{\mathrm{sh}}) + \epsilon_{\mathrm{sh},e} \\ &\le R^*_{\mathrm{share},e} + 2\epsilon_{\mathrm{sh},e} = R^*_{\mathrm{sep},e} + \mathsf{B}^{\mathrm{reuse}}_e + 2\epsilon_{\mathrm{sh},e}. \end{aligned} \tag{48}$$

For every role, Hoeffding and a union bound over $|\mathcal{J}_e||\mathcal{F}_e|$ with the remaining failure budget give

$$\sup_{f\in\mathcal{F}_e} |\widehat{R}_a(f) - R_a(f)| \le \sqrt{\frac{\log(4|\mathcal{J}_e||\mathcal{F}_e|/\delta)}{2n_a}}. \tag{49}$$

The same ERM argument role by role, followed by multiplication by π_a and summation, proves Theorem 8. The two uniform events hold jointly with probability at least $1-\delta$.

C.3 What the effective sample size optimizes

Proposition 9 (Allocation maximizing the Hoeffding effective sample size)**.** *For positive real allocations with $N = \sum_a n_a$,*

$$N_e^{\mathrm{eff}} = \left(\sum_a \frac{\pi_a^2}{n_a}\right)^{-1} \le N, \tag{50}$$

with equality if and only if $n_a = N\pi_a$.

Proof. Cauchy–Schwarz gives $1 = (\sum_a \pi_a)^2 \le (\sum_a \pi_a^2/n_a)(\sum_a n_a) = N/N_e^{\mathrm{eff}}$. Equality requires $\pi_a/\sqrt{n_a}$ to be proportional to $\sqrt{n_a}$. □

This allocation maximizes the particular bounded-loss Hoeffding effective sample size. A Bernstein analysis can produce variance-aware allocations for unequal variances.

C.4 Uniform-deviation and trace-level templates

The finite class is not essential. If, on one event,

$$\sup_f \left|\widehat{R}_{\mathrm{sh}}(f) - R_{\mathrm{sh}}(f)\right| \le u_{\mathrm{sh}}, \qquad \sup_f |\widehat{R}_a(f) - R_a(f)| \le u_a, \tag{51}$$

then the same proof gives shared excess $2u_{\mathrm{sh}}$ and separate excess $2\sum_a \pi_a u_a$. Uniform bounds based on Rademacher complexity (fit to random signs) or covering numbers (the size of a finite approximation to a class) may be substituted if they prove exactly Equation (51). Algorithm-specific stability, compression, and PAC-Bayes bounds are not automatically uniform-deviation statements; using them requires replacing the learning theorem and matching its particular estimator or randomized predictor.

Calls inside one autoregressive trace should not be counted as independent samples. Let $Z_1, \dots, Z_N$ be independent traces, not necessarily identically distributed, and let $L_r(f) \in [0,1]$ be any aggregate trace loss. For a finite class $\mathcal{F}$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{N}\sum_{r=1}^{N}\big(L_r(f) - \mathbb{E}L_r(f)\big)\right| \le \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2N}}. \tag{52}$$

If traces are IID, the centering equals the common population trace risk. No within-trace independence is needed.

D Exact Linear Sharing Frontier

Overview. We prove the exact pooling and partition risks in the specified linear model. We also separate fixed-rule identities from data-dependent choices: a same-data plug-in can fail, independent validation needs its own guarantee, and the stated high-dimensional Gaussian specialization is classical shrinkage. Sections D.1–D.3 derive the pooling, merge, and dispersion formulas; Sections D.4–D.6 contrast a failing plug-in, independent validation, and the classical Gaussian shrinkage specialization.

D.1 Oracle details and proof of the exact sharing frontier

For $B + C > 0$, the optimal fixed coefficient and its minimum risk in Theorem 1 are

$$\lambda^\star = \frac{C}{B+C}, \qquad \mathcal{R}_{\lambda^\star} = V - \frac{C^2}{B+C}. \tag{53}$$

If $B = C = 0$, every $\lambda \in [0,1]$ gives the same risk V. For $M > 1$, $\sigma > 0$, and finite B, oracle partial pooling is no worse than either endpoint; it is strictly better than separate estimation and is strictly better than full reuse whenever $B > 0$.

Proof of Theorem 1. Write $Y_a = X_a A_a^\top + \mathcal{E}_a$, where the noise rows have mean zero, covariance $\sigma^2 I_p$, and are independent across observations and roles. Gaussianity is not used in the expected-risk proof. Orthogonality of the fixed design gives

$$\widehat{A}_a = A_a + E_a, \qquad E_a = \frac{1}{n}\mathcal{E}_a^\top X_a. \tag{54}$$

The $pd = D$ entries of E_a are centered and uncorrelated with variance σ^2/n, and the matrices E_a are independent across roles. Indeed, for output coordinates u, v and input coordinates r, t, $\mathbb{E}[(E_a)_{ur}(E_a)_{vt}] = (\sigma^2/n)\mathbf{1}\{u = v, r = t\}$. Hence

$$\mathbb{E}\,\|E_a\|_F^2 = \frac{pd\sigma^2}{n} = V. \tag{55}$$

For isotropic test input x, $\mathbb{E}_x \left\|(\widehat{A} - A_a)x\right\|_2^2 = \left\|\widehat{A} - A_a\right\|_F^2$.

Separate estimation therefore has average risk V. Let $\bar{E} = M^{-1}\sum_a E_a$. The pooled estimator is $\widehat{A}_{\mathrm{pool}} = \bar{A} + \bar{E}$, so

$$\begin{aligned}\frac{1}{M}\sum_a \mathbb{E}\left\|\widehat{A}_{\mathrm{pool}} - A_a\right\|_F^2 &= \frac{1}{M}\sum_a \left\|\bar{A} - A_a\right\|_F^2 + \mathbb{E}\left\|\bar{E}\right\|_F^2 \\ &= B + \frac{V}{M},\end{aligned} \tag{56}$$

where the cross term vanishes and $\mathbb{E}\left\|\bar{E}\right\|_F^2 = V/M$.

For partial reuse,

$$\widehat{A}_a^{(\lambda)} - A_a = \lambda(\bar{A} - A_a) + (1-\lambda)E_a + \lambda\bar{E}. \tag{57}$$

The average squared bias is $\lambda^2 B$. For each coefficient, the variance of $(1-\lambda)E_a + \lambda\bar{E}$ is

$$\frac{\sigma^2}{n}\left[(1-\lambda)^2 + \frac{2\lambda(1-\lambda) + \lambda^2}{M}\right]. \tag{58}$$

Multiplying by D and simplifying yields

$$\mathcal{R}_\lambda = V - 2V\left(1 - \frac{1}{M}\right)\lambda + \left[B + V\left(1 - \frac{1}{M}\right)\right]\lambda^2, \tag{59}$$

which is Equation (9). The full-sharing comparison follows from $\mathcal{R}_{\mathrm{share}} - \mathcal{R}_{\mathrm{sep}} = B - C$. Differentiating the strictly convex quadratic when $B + C > 0$ gives $\lambda^\star = C/(B+C)$ and the stated minimum. In the degenerate case $B = C = 0$, the quadratic is constant in λ. If $M > 1$ and $\sigma > 0$, then $C > 0$, so the derivative at zero is negative. If additionally $B > 0$, then $\lambda^\star < 1$ and the interior optimum is strictly below full sharing.

D.2 Proof of the exact partition and merge frontier

For a group $G \subseteq [M]$, write

$$\widehat{A}_G = \frac{1}{|G|}\sum_{a\in G}\widehat{A}_a = \bar{A}_G + \bar{E}_G, \qquad \mathbb{E}\left\|\bar{E}_G\right\|_F^2 = \frac{V}{|G|}. \tag{60}$$

The total expected risk contributed by roles in G is therefore

$$\sum_{a\in G}\mathbb{E}\left\|\widehat{A}_G - A_a\right\|_F^2 = \sum_{a\in G}\left\|\bar{A}_G - A_a\right\|_F^2 + |G|\mathbb{E}\left\|\bar{E}_G\right\|_F^2 = \sum_{a\in G}\left\|\bar{A}_G - A_a\right\|_F^2 + V. \tag{61}$$

Summing over the K_Π blocks and dividing by M proves Equation (11).

For two disjoint groups G, H, the Hilbert-space variance decomposition gives

$$\sum_{a\in G\cup H}\left\|A_a - \bar{A}_{G\cup H}\right\|_F^2 = \sum_{a\in G}\left\|A_a - \bar{A}_G\right\|_F^2 + \sum_{a\in H}\left\|A_a - \bar{A}_H\right\|_F^2 + \frac{|G||H|}{|G|+|H|}\left\|\bar{A}_G - \bar{A}_H\right\|_F^2. \tag{62}$$

Merging increases within-group dispersion by the final term and decreases $K_\Pi V$ by exactly V, proving Equation (12).

The merge separation has an exact sampling law under the additional assumption that all noise entries are mutually Gaussian with the stated independent-row covariance and $\sigma > 0$. For two pre-specified, disjoint, nonempty groups, let $m = |G|$, $h = |H|$, $c = mh/(m+h)$, $q = \sigma^2/n$, and define

$$W_{G,H} = c\left\|\bar{A}_G - \bar{A}_H\right\|_F^2, \qquad \widehat{W}_{G,H} = c\left\|\widehat{A}_G - \widehat{A}_H\right\|_F^2. \tag{63}$$

Because each coefficient of $\widehat{A}_G - \widehat{A}_H$ has variance $q(1/m + 1/h) = q/c$,

$$\frac{\widehat{W}_{G,H}}{q} \sim \chi'^2_D\left(\frac{W_{G,H}}{q}\right), \qquad \mathbb{E}\widehat{W}_{G,H} = W_{G,H} + V. \tag{64}$$

Thus $\widehat{W}_{G,H} - V$ is unbiased for the functional separation governing the merge decision. Exact noncentral-chi-square confidence bounds can be inverted for a pre-specified pair; searching many candidate merges requires a familywise or validation correction.

D.3 The heterogeneity is observable up to a known noise correction

The exact model also gives a direct diagnostic for B.

Proposition 10 (Noisy role dispersion)**.** *Assume $M \geq 2$, $\sigma > 0$, and independent Gaussian observation noise for the distributional identity below. Let $q = \sigma^2/n$ and*

$$\widehat{B}_{\text{raw}} = \frac{1}{M}\sum_{a=1}^{M}\left\|\widehat{A}_a - \widehat{A}_{\text{pool}}\right\|_F^2. \tag{65}$$

Then

$$\mathbb{E}\widehat{B}_{\text{raw}} = B + C, \qquad \widehat{B}_{\text{unb}} = \widehat{B}_{\text{raw}} - C \tag{66}$$

is unbiased for B. More precisely,

$$\frac{M\widehat{B}_{\text{raw}}}{q} \sim \chi'^2_{D(M-1)}\left(\frac{MB}{q}\right). \tag{67}$$

Proof. For one matrix coefficient, stack the M role estimates and let $P = I_M - \mathbf{1}\mathbf{1}^\top/M$. The centered estimate is $P(a+\varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, qI_M)$. In an orthonormal basis of the rank-$(M-1)$ image of P, its squared norm divided by q is noncentral chi-squared with noncentrality $\|Pa\|^2/q$. Summing the independent contributions of all D coefficients gives Equation (67); its mean gives Equation (66) because $Dq(M-1)/M = C$. □

The expectation and unbiasedness formulas need only the second-moment assumptions; the noncentral chi-square law additionally needs Gaussian noise, $q > 0$, and $M \geq 2$. When $\sigma = 0$, dispersion is observed without noise; when $M = 1$, both dispersions and B, C are zero and no chi-square expression is needed. A noncentral chi-square variable is a sum of squared independent unit-variance normal variables with possibly nonzero means; its degrees of freedom count summands and its noncentrality is the sum of squared means. Thus the same dispersion that determines negative transfer is estimable from role-specific fits after subtracting estimation noise. The oracle coefficient can be written $\lambda^\star = C/(B + C)$, suggesting the clipped plug-in rule $\widehat{\lambda} = 1$ when $\widehat{B}_{\mathrm{raw}} = 0$ and $\widehat{\lambda} = \min\{1, C/\widehat{B}_{\mathrm{raw}}\}$ otherwise, when the exact assumptions are approximately credible. It is not justified by substituting $\widehat{\lambda}$ into the fixed-coefficient risk formula. The low-dimensional Gaussian counterexample below and the contrast-space result give the dimension-dependent conclusions. Section D.6 supplies the classical known-noise Gaussian guarantee when the contrast dimension $D(M - 1)$ is at least four.

D.4 A SAME-DATA PLUG-IN CAN WORSEN RISK INSIDE THE EXACT MODEL

Take $M = 2$, $D = 1$, $q = 1$, and write $Y_i = \theta_i + \varepsilon_i$ with independent $\varepsilon_i \sim \mathcal{N}(0, 1)$. In the orthonormal contrast coordinate, $Z = (Y_1 - Y_2)/\sqrt{2} \sim \mathcal{N}(\mu, 1)$, where $\mu = (\theta_1 - \theta_2)/\sqrt{2}$. Since $C = 1/2$ and $\widehat{B}_{\mathrm{raw}} = Z^2/2$, the plug-in has contrast estimate

$$h(Z) = \begin{cases} 0, & |Z| \leq 1, \\ Z - 1/Z, & |Z| > 1. \end{cases} \tag{68}$$

The common-mean coordinate is unchanged. Put $u(z) = h(z) - z$; this function is continuous and piecewise differentiable. Gaussian integration by parts, with the boundary terms cancelling at ± 1, gives

$$\mathbb{E}(h(Z) - \mu)^2 - 1 = \mathbb{E}\{u(Z)^2 + 2u'(Z)\} \tag{69}$$

$$= \mathbb{E}\left[(Z^2 - 2)\mathbf{1}_{\{|Z| \leq 1\}} + \frac{3}{Z^2}\mathbf{1}_{\{|Z| > 1\}}\right]. \tag{70}$$

For completeness, the integration identity follows from the normal density derivative $\phi'(z - \mu) = -(z - \mu)\phi(z - \mu)$, so $\mathbb{E}[(Z - \mu)u(Z)] = \mathbb{E}u'(Z)$; u and its derivative are integrable here. At $\mu = 6$, Chebyshev's inequality gives $\mathbb{P}(4 \leq Z \leq 8) \geq 3/4$, whereas the normal fourth moment gives $\mathbb{P}(|Z| \leq 1) \leq 3/5^4$. Consequently,

$$\mathbb{E}(h(Z) - 6)^2 - 1 \geq \frac{3}{64}\frac{3}{4} - \frac{6}{625} = \frac{4089}{160000} > 0. \tag{71}$$

The average risk of the two roles increases by half this contrast-risk difference. Unbiased estimation of B in Proposition 10 does not imply risk domination after a nonlinear same-data plug-in.

D.5 INDEPENDENT VALIDATION SELECTS REUSE STRENGTH WITH AN ORACLE INEQUALITY

Let $\Lambda \subset [0, 1]$ be a finite grid. Fit $\widehat{f}_\lambda$ using training data, and evaluate every fitted candidate on an independent weighted validation set with loss in $[0, 1]$ and effective size $N_{\mathrm{val}}^{\mathrm{eff}}$. Let $\widehat{\lambda}$ minimize validation loss. Here the weighted validation convention is the one made explicit for finite candidates in Theorem 48: conditional on the training information $\mathscr{T}$, validation records $i = 1, \ldots, N$ are independent, and nonnegative weights v_i summing to one are fixed before observing their outcomes. Write $\ell_i(\widehat{f}_\lambda) \in [0, 1]$ for the loss on record i. Then $\widehat{R}(\widehat{f}_\lambda) = \sum_i v_i \ell_i(\widehat{f}_\lambda)$, $R(\widehat{f}_\lambda) = \sum_i v_i \mathbb{E}[\ell_i(\widehat{f}_\lambda) \mid \mathscr{T}]$, and $N_{\mathrm{val}}^{\mathrm{eff}} = (\sum_i v_i^2)^{-1}$. The selector minimizes this same weighted empirical risk. Nonidentically distributed records retain this declared weighted target; losses of different candidates on the same record need not be independent.

Proposition 11 (Validation-selected partial reuse)**.** *Conditional on the training sample, with probability at least* $1 - \delta$ *over validation data,*

$$R(\widehat{f}_{\widehat{\lambda}}) \leq \min_{\lambda \in \Lambda} R(\widehat{f}_\lambda) + 2\sqrt{\frac{\log(2|\Lambda|/\delta)}{2N_{\mathrm{val}}^{\mathrm{eff}}}}. \tag{72}$$

Proof. Weighted Hoeffding and a union bound give one event on which every validation risk is within the displayed radius of its population risk. Apply the standard two-deviation ERM argument on that event. □

This result provides a statistically valid way to choose among fitted candidates for the stated bounded validation loss, without estimating B parametrically. The guarantee evaluates the already fitted candidates using independent validation data. Theorem 48 gives a separate finite-family guarantee under explicit tail assumptions, including a Gaussian quadratic-loss specialization.

D.6 Classical same-data shrinkage in the contrast space

This calculation explains a dimension-specific exception to the preceding warning; it is an application of classical Baranchik–Stein shrinkage (Baranchik, 1970; Stein, 1981). A directly comparable risk identity and Baranchik's sufficient conditions are reproduced in Maruyama & Takemura (2024, Section 1, equations (1.5)–(1.6)). Shrinkage toward a fixed subspace, a common mean, or data-selected cluster attractors is also established background (Srinath & Venkataramanan, 2017); the calculation here fixes the common-mean subspace.

Retain the balanced fixed-design model of Theorem 1, with $M \geq 2$, $D = pd$, mutually Gaussian observation noise, and *known* $q = \sigma^2/n > 0$. Stack the M vectorized role estimates as $Y = \theta + \zeta \in \mathbb{R}^{MD}$, so $\zeta \sim \mathcal{N}(0, qI_{MD})$. Define

$$P = \frac{\mathbf{1}\mathbf{1}^\top}{M} \otimes I_D, \qquad Q = I_{MD} - P, \qquad k = D(M-1).$$

Here P replaces every role vector by their common average, and Q retains their differences. The integer k is the dimension of this *contrast space*, not the number of selected experts. Choose a fixed matrix $U \in \mathbb{R}^{MD\times k}$ whose columns are an orthonormal basis of that space: $U^\top U = I_k$, $UU^\top = Q$. Then

$$Z = U^\top Y \sim \mathcal{N}(\mu, qI_k), \qquad \mu = U^\top\theta, \qquad M\widehat{B}_{\text{raw}} = \|Z\|^2, \quad MC = kq.$$

The same-data coefficient defined after Proposition 10 leaves PY unchanged and replaces Z by

$$h(Z) = \left(1 - \frac{kq}{\|Z\|^2}\right)_+ Z, \qquad h(0) = 0, \quad (t)_+ = \max\{t, 0\}.$$

Thus the fitted stacked vector is $PY + Uh(Z)$. Its role-averaged excess prediction risk satisfies

$$\mathcal{R}_{\text{plug}} - \mathcal{R}_{\text{sep}} = \frac{1}{M}\big(\mathbb{E}\|h(Z) - \mu\|^2 - kq\big) < 0 \qquad \text{for every fixed } \theta \text{ when } k \geq 4.$$

Direct verification of the classical specialization. Let $u(z) = h(z) - z$ and $s = \|z\|^2$. The divergence $\operatorname{div} u = \sum_{i=1}^k \partial u_i/\partial z_i$ is the sum of coordinate derivatives. Away from the sphere $s = kq$,

	$s \leq kq$	$s > kq$
$u(z)$	$-z$	$-kqz/s$
$\operatorname{div} u(z)$	$-k$	$-kq(k-2)/s$
$\|u(z)\|^2$	s	k^2q^2/s

The two formulas for u agree on the sphere. Both pieces have bounded first derivatives on their domains, so u is weakly differentiable (its integration-by-parts derivative has no surface jump term). It and its derivatives are Gaussian-integrable. Coordinatewise Gaussian integration by parts gives

$$\mathbb{E}[(Z_i - \mu_i)u_i(Z)] = q\mathbb{E}[\partial_i u_i(Z)].$$

Expanding $h(Z) - \mu = Z - \mu + u(Z)$ and summing therefore gives

$$\mathbb{E}\|h(Z) - \mu\|^2 - kq = \mathbb{E}\left[(s - 2kq)\mathbf{1}_{\{s\leq kq\}} + \frac{k(4-k)q^2}{s}\mathbf{1}_{\{s>kq\}}\right].$$

For $k \geq 4$ the outside term is nonpositive, and the inside term is at most $-kq$. Every nondegenerate Gaussian distribution assigns positive probability to $s \leq kq$, so the expectation is strictly negative, also at $k = 4$. Finally $P\zeta$ and $U(h(Z) - \mu)$ lie in orthogonal subspaces for every realization. The common-mean risk is unchanged, and division by M converts the contrast-risk difference to the original role-average risk. □

Relation to classical shrinkage. After standardizing $X = Z/\sqrt{q}$, the rule is $(1 - \phi(w)/w)X$ with $w = \|X\|^2$ and $\phi(w) = \min\{w, k\}$. In the cited Stein risk identity, substitution gives $w - 2k$ on $w \le k$ and $k(4-k)/w$ otherwise. It also satisfies Baranchik's monotonicity and range conditions $0 \le \phi(w) \le 2(k-2)$ when $k \ge 4$. The usual positive-part James–Stein rule uses $(k-2)q$, not kq, and already dominates the unshrunk estimate for $k \ge 3$ (Srinath & Venkataramanan, 2017). That is a different estimator; the threshold four here is not a universal dimension barrier.

The strict improvement is pointwise in the fixed target, not a claimed target-uniform positive improvement amount. The low-dimensional counterexample in Equation (71) remains valid. The main non-minimaxity result of Maruyama & Takemura (2024) concerns rules whose shrinkage is exactly zero outside a finite ball; this is not the present rule, whose $\phi(w)$ stays equal to k. Only their introductory classical risk identity is used here.

E End-to-End Route-Plan Certificates

Overview. We track how aligned implementation error and functional routing error propagate to the endpoint. The certificate requires explicit role-to-rollout transfer conditions. A separate linear calculation retains error directions and cross-time covariance rather than replacing them by scalar sensitivity bounds. Sections E.1–E.2 establish propagation and the conditional certificate; Sections E.3–E.5 address functional route cost, terminal decision margins, and the separate covariance identity.

E.1 Pathwise propagation with random intended programs

All random variables are defined on one coupled evaluation probability space. The intended canonical program Γ_j may depend on the reference history and exogenous randomness; it is never moved outside expectation. The reference and learned recursions are

$$S_{j+1} = T_{j,\Gamma_j}(S_j), \qquad \widehat{S}^m_{j+1} = \widehat{T}^m_{j,\widehat{\Gamma}^m_j}(\widehat{S}^m_j). \tag{73}$$

The implementation $\iota^m_j(\Gamma_j)$ is the model-m executable plan aligned with the intended canonical program. Assume $T_{j,\gamma}$ is L_j-Lipschitz on the reached region uniformly in admissible canonical γ. Define $e^m_j = d(\widehat{S}^m_j, S_j)$. Inserting the aligned learned transition and the reference map at the learned state gives, pathwise,

$$\begin{aligned} e^m_{j+1} &\le d\Big(\widehat{T}^m_{j,\widehat{\Gamma}^m_j}(\widehat{S}^m_j), \widehat{T}^m_{j,\iota^m_j(\Gamma_j)}(\widehat{S}^m_j)\Big) \\ &\quad + d\Big(\widehat{T}^m_{j,\iota^m_j(\Gamma_j)}(\widehat{S}^m_j), T_{j,\Gamma_j}(\widehat{S}^m_j)\Big) + d\Big(T_{j,\Gamma_j}(\widehat{S}^m_j), T_{j,\Gamma_j}(S_j)\Big) \\ &\le r^m_j + a^m_j + L_j e^m_j. \end{aligned} \tag{74}$$

Induction yields

$$e^m_H \le \left(\prod_{r=0}^{H-1} L_r\right) e^m_0 + \sum_{j=0}^{H-1} A_j(a^m_j + r^m_j), \qquad A_j = \prod_{r=j+1}^{H-1} L_r. \tag{75}$$

No independence between route errors and states is used. The intended program is coupled to the reference trajectory; the theorem bounds imitation of that coupled reference computation rather than regret against a state-dependent policy recomputed at the learned state. If sensitivities are random rather than deterministic upper bounds, their products remain inside expectation unless additional moment assumptions justify factorization.

E.2 Expanded terminal certificate and probability bookkeeping

The compact statement of Theorem 4 uses the following complete active-support convention:

$$\begin{aligned} w_a &= \sum_{j=0}^{H-1} A_j \omega_{j,a}, \quad \mathcal{J}^+_e = \{a \in \mathcal{J}_e : w_a > 0\}, \\ W_e &= \sum_{a \in \mathcal{J}^+_e} w_a, \qquad \pi_{e,a} = \frac{w_a}{W_e} \quad (a \in \mathcal{J}^+_e). \end{aligned} \tag{76}$$

Here w_a is role a's total terminal weight, $\mathcal{J}_e^+$ retains its positive-weight roles, W_e is their group total, and $\pi_{e,a}$ gives their relative importance. Only groups with $W_e > 0$ enter the certificate. For notational economy, $R^*_{\mathrm{sep},e}$, $\mathsf{B}^{\mathrm{reuse}}_e$, and the two radii below denote the quantities from Section 4 recomputed on $\mathcal{J}_e^+$ with weights $\pi_{e,a}$, and the empirical objectives use the same active support. A zero-weight role can be omitted from this comparison; if its data nevertheless influence a shared fitted expert, that influence must be represented by a positive training/certificate weight or by an additional algorithmic-deviation term. Thus these weights incorporate downstream amplification.

Expanded conditions and bounds. Let $\delta \in (0,1)$, let the initial states agree, and let terminal utility (a scalar score) $\mathcal{U} : \mathcal{S} \to \mathbb{R}$ be $L_{\mathcal{U}}$-Lipschitz. On every active group $\mathcal{J}_e^+$ assume all hypotheses of Theorem 8: a fixed finite class, losses in $[0,1]$, at least one observation per active role, and mutually independent role observations under the stated sampling model. The classes, losses, distributions, and performance weights are fixed independently of these observations. For each group e with $W_e > 0$, choose $\delta_e > 0$ with $\sum_e \delta_e \le \delta$, and define $\epsilon_{\mathrm{sh},e}(\delta_e)$ and $\epsilon_{\mathrm{sep},e}(\delta_e)$ by Equation (43) and Equation (44), respectively, with δ replaced by δ_e. Assume the group estimators minimize the performance-weighted empirical risks induced by Equation (76). The two comparators are fitted from the same role samples; if they are refitted on independent samples, the confidence allocation must also cover both fitted datasets. Under Assumption 3, with probability at least $1-\delta$ over the expert-estimation samples,

$$\mathbb{E}_{\mathrm{eval}} \left| \mathcal{U}(\widehat{S}^{\mathrm{reuse}}_H) - \mathcal{U}(S_H) \right| \le L_{\mathcal{U}} U_{\mathrm{reuse}}, \tag{77}$$

$$U_{\mathrm{reuse}} = \sum_j A_j(\tau_j^{\mathrm{reuse}} + \rho_j^{\mathrm{reuse}}) + \sum_e W_e\big(R^*_{\mathrm{sep},e} + \mathsf{B}^{\mathrm{reuse}}_e + 2\epsilon_{\mathrm{sh},e}(\delta_e)\big), \tag{78}$$

$$\mathbb{E}_{\mathrm{eval}} \left| \mathcal{U}(\widehat{S}^{\mathrm{sep}}_H) - \mathcal{U}(S_H) \right| \le L_{\mathcal{U}} U_{\mathrm{sep}}, \tag{79}$$

$$U_{\mathrm{sep}} = \sum_j A_j(\tau_j^{\mathrm{sep}} + \rho_j^{\mathrm{sep}}) + \sum_e W_e\big(R^*_{\mathrm{sep},e} + 2\epsilon_{\mathrm{sep},e}(\delta_e)\big). \tag{80}$$

The symbols $\epsilon_{\mathrm{reuse},e}$ and $\epsilon_{\mathrm{sep},e}$ in Equation (22) are respectively the shared and separate radii at the same group budget δ_e. Its indicator selects the reuse bias only for the shared model, so substituting either value of m gives exactly the two expanded bounds above. The common oracle cancels in Equation (23) only because both comparators use these same groups, weights, losses, and evaluation populations. The utility-bound difference is $L_{\mathcal{U}}\mathsf{PRG}_H$; strict ordering requires $L_{\mathcal{U}} > 0$. For $L_{\mathcal{U}} = 0$ both utility bounds vanish.

Proof of Theorem 4. For each active group choose $\delta_e > 0$ with $\sum_e \delta_e \le \delta$ and condition on the simultaneous training event obtained by applying Theorem 8 with failure probability δ_e to group e. Under Assumption 3, common initial states, and $\rho_j^m = \mathbb{E} r_j^m$,

$$\mathbb{E} e_H^m \le \sum_j A_j(\tau_j^m + \rho_j^m) + \sum_a \left(\sum_j A_j \omega_{j,a} \right) R_a\left(\widehat{f}^m_{\alpha_m(a)} \right). \tag{81}$$

The coefficient of role a is w_a. Restrict each reuse group to its active support $\mathcal{J}_e^+$ from Equation (76); then writing $w_a = W_e \pi_{e,a}$ gives

$$\sum_{a \in \mathcal{J}_e^+} w_a R_a(f) = W_e \sum_{a \in \mathcal{J}_e^+} \pi_{e,a} R_a(f). \tag{82}$$

Because the estimators use these fixed performance weights, the shared part of Theorem 8 gives the reuse term in Equation (78); its role-specific part gives Equation (80). Lipschitz continuity of terminal utility multiplies the state bound by $L_{\mathcal{U}}$. Subtracting the two displayed right-hand sides cancels only the genuinely common oracle terms and yields $L_{\mathcal{U}}\mathsf{PRG}_H$.

If comparator m requires model-specific weights $\omega^m_{j,a}$, define $w^m_a = \sum_j A_j \omega^m_{j,a}$ and construct each certificate separately. Both remain valid, but the oracle and estimation terms generally have different coefficients and the compact expression Equation (23) is unavailable. This is the appropriate treatment

when rollout occupancies differ and no common domination envelope is credible. Likewise, if event lifts or the surrounding assembly are learned from the expert-estimation data, the finite-sample event must be replaced by a uniform bound for the joint class or by sample splitting.

E.3 Functional route cost

For a finite canonical plan space and a finite model-specific executable plan space, define

$$C_j^m(\gamma,\gamma') = \mathbb{P}(\Gamma_j = \gamma, \widehat{\Gamma}_j^m = \gamma'), \tag{83}$$

$$D_j^m(\gamma,\gamma') = \mathbb{E}\Big[d\Big(\widehat{T}_{j,\gamma'}^m(\widehat{S}_j^m), \widehat{T}_{j,\iota_j^m(\gamma)}^m(\widehat{S}_j^m)\Big) \mid \Gamma_j = \gamma, \widehat{\Gamma}_j^m = \gamma'\Big], \tag{84}$$

with $D_j^m(\gamma,\gamma') = 0$ when the conditioning event has zero probability. Then

$$\rho_j^m = \sum_{\gamma,\gamma'} C_j^m(\gamma,\gamma') D_j^m(\gamma,\gamma'). \tag{85}$$

This quantity distinguishes harmless index substitutions from rare but severe functional confusions. Index accuracy records only the off-alignment mass and is meaningful only after the canonical-to-model alignment has been specified.

For an additive mixture, a direct gate bound is available.

Lemma 12 (Equal-mass functional gate discrepancy)**.** *Let $v_1, \dots, v_K$ lie in a normed vector space and let g, g' have equal total mass. Then*

$$\left\| \sum_e (g_e - g'_e) v_e \right\| \le \frac{1}{2} \left(\max_{e,e'} \|v_e - v_{e'}\| \right) \|g - g'\|_1 . \tag{86}$$

The constant $1/2$ is sharp.

Proof. The positive and negative parts of $g - g'$ have common mass $\|g - g'\|_1 / 2$. Normalize them to two convex combinations of the v_e; their distance is at most the diameter. Equality is attained by two points and disjoint one-hot gates. □

The lemma applies at one common expert input and with the same expert-output collection. Different states or dispatch-dependent expert inputs require additional state or dispatch sensitivity terms.

E.4 Terminal decisions with a margin

Discrete accuracy is not Lipschitz. It can nevertheless be controlled under a terminal margin when the same deterministic readout g is applied to the reference and learned terminal states. Let $z = g(S_H)$, suppose g is L_g-Lipschitz from d to ℓ_∞, and let the reference top-two gap be $\gamma > 0$. For $L_g > 0$, the learned argmax is unchanged whenever the pathwise right-hand side of Equation (75) is strictly less than $\gamma/(2L_g)$; for $L_g = 0$, the readout is constant on the compared region and the conclusion is immediate. If the learned and reference readout heads differ, an additional common-state readout discrepancy $\left\| \widehat{g}(\widehat{S}_H) - g(\widehat{S}_H) \right\|_\infty$ must also be below the remaining margin budget. These pathwise robustness statements yield expected-accuracy control when combined with a suitable tail bound.

E.5 Exact directional amplification

Proposition 13 (Exact propagation of correlated errors)**.** *Let $\mathbf{e}_{j+1} = M_j \mathbf{e}_j + \xi_j$ in $\mathbb{R}^s$, with $\mathbf{e}_0 = 0$, deterministic $s \times s$ matrices M_j, and square-integrable errors ξ_j. Define $b_j = \mathbb{E}\xi_j$ and cross-covariances $\Sigma_{jk} = \mathbb{E}[(\xi_j - b_j)(\xi_k - b_k)^\top]$. With $P_j = M_{H-1} \cdots M_{j+1}$ and $P_{H-1} = I_s$,*

$$\mathbb{E}\|\mathbf{e}_H\|_2^2 = \left\| \sum_j P_j b_j \right\|_2^2 + \sum_{j,k} \operatorname{tr}(P_j \Sigma_{jk} P_k^\top). \tag{87}$$

Here tr *sums diagonal matrix entries. For uncorrelated centered errors only $j = k$ terms remain, weighted by $P_j^\top P_j$, the* Gramian *measuring directional amplification. Reusing one fitted expert generally creates nonzero cross-time terms, so that special case cannot be assumed from independent training examples.*

Proof of Proposition 13. Unrolling $\mathbf{e}_{j+1} = M_j \mathbf{e}_j + \xi_j$ gives

$$\mathbf{e}_H = \sum_{j=0}^{H-1} P_j \xi_j, \qquad P_j = M_{H-1} \cdots M_{j+1}. \tag{88}$$

The mean is $\sum_j P_j b_j$. Expanding the covariance without discarding cross-time terms gives

$$\mathrm{Cov}(\mathbf{e}_H) = \sum_{j,k} P_j \Sigma_{jk} P_k^\top. \tag{89}$$

Use $\mathbb{E}\|X\|_2^2 = \|\mathbb{E}X\|_2^2 + \mathrm{tr}(\mathrm{Cov}\, X)$ to obtain Proposition 13. This identity follows by writing $X = \mathbb{E}X + (X - \mathbb{E}X)$, expanding its squared norm, and observing that the centered cross term has zero expectation. If all off-diagonal Σ_{jk} vanish, the previous diagonal Gramian formula follows. The double sum is nonnegative as a whole, although individual cross terms may be negative. If M_j is random and correlated with the errors, the deterministic covariance factorization is not valid: only the pathwise identity $\mathbf{e}_H = \sum_j P_j \xi_j$ is automatic. Proposition 43 shows explicitly how tying parameters generates cross-time terms even in a simple additive model.

F Causal Route Actions and Proofs

Overview. We ask what a forced route actually changes, rather than what its natural label suggests. The counterexamples and risk decompositions distinguish observational agreement from executable reuse and show how an insufficient macro state obstructs prediction of intervention effects. Action coverage is treated separately in Section K. Sections F.1–F.3 define actions and prove the observational and state-aliasing obstructions; Sections F.4–F.5 cover rollout, reporting requirements, and the average-risk version of the obstruction.

F.1 Non-anticipating route actions

A complete micro context $C \in \mathcal{C}$ contains the prompt, generated prefix, caches, hidden states, router state, and all randomness revealed up to the intervention boundary. A fixed macro extractor $\Psi : \mathcal{C} \to \mathcal{S}$ produces $S = \Psi(C)$. A route action $\mathcal{A}_i = (\varpi_i, \sigma_i)$ consists of a non-anticipating policy over future route decisions and a stopping time $1 \le \sigma_i \le B_{\mathrm{act}}$. Every decision made by ϖ_i is measurable with respect to the information available immediately before that decision; neither the policy nor the stopping rule may inspect unrevealed future tokens or a completed post-hoc route trace.

For pathwise statements, execution is greedy or is conditioned on a common realization of exogenous randomness. Executing $\mathcal{A}_i$ then yields a deterministic micro transition $\mathfrak{T}_i(C)$ and macro result $T_i(C) = \Psi(\mathfrak{T}_i(C))$. Under stochastic generation, the corresponding object is a transition kernel; the deterministic claims below then apply conditionally, while the Wasserstein formulation in Section L controls the induced laws.

For a candidate macro operator F_i, fixed before the evaluation sample is observed, the off-policy defect under a fixed probe distribution ν_i is

$$\mu_i^{\mathrm{off}} = \mathbb{E}_{C \sim \nu_i} d(T_i(C), F_i(\Psi(C))). \tag{90}$$

If a controller executes actions on-policy, the corresponding occupancy changes with the action library and controller; an on-policy defect must therefore be evaluated under that induced distribution. For $K \ge 2$, a small defect is meaningful only together with non-collapse between actions, for example

$$\mathrm{Sep}(\nu) = \mathbb{E}_{C \sim \nu} \frac{2}{K(K-1)} \sum_{i<i'} d(T_i(C), T_{i'}(C)). \tag{91}$$

If Ψ, the actions, or the operator class are selected using the same records on which the defect is reported, sample splitting, cross-fitting, or a uniform complexity correction is required.

F.2 Proof that observational alignment is insufficient

Proof of Proposition 5. Construct a residual routed block whose shared path receives (S, I) and returns $F_I(S)$. Set every expert contribution identically to zero, and let the natural router output index

I. The model output is exactly $F_I(S)$ and natural route-label accuracy is one. Under an intervention replacing only the routed index by J, the shared path is unchanged and every expert contribution remains zero, so the output is still $F_I(S)$ and is independent of J. Forced-route fidelity to the target associated with J is therefore exactly $\mathbb{P}(F_I(S) = F_J(S))$. If J is uniform and independent of (I, S) and the K values $F_1(S), \ldots, F_K(S)$ are pairwise distinct at every evaluated state, equality occurs if and only if $J = I$, giving fidelity $1/K$. □

For comparison, a second model with zero shared contribution and experts $E_i(S) = F_i(S)$ has the same natural outputs and the same labels under route I, but a forced route J returns $F_J(S)$. These two models are observationally equivalent on task outputs and natural route labels and differ under intervention. The construction is compatible with perfect task performance. It establishes a non-identifiability result: behavioral and route-label observations alone do not determine where the operation is implemented.

F.3 State-aliasing obstruction and proof

Let $\mathcal{C}_0$ be a nonempty set of contexts. Two contexts are *aliased* when they have the same extracted state. Define the largest difference between their responses to action i by

$$\Xi_i(\Psi; \mathcal{C}_0) = \sup_{\substack{C, C' \in \mathcal{C}_0,\ C \neq C' \\ \Psi(C) = \Psi(C')}} d(T_i(C), T_i(C')), \tag{92}$$

with value zero when no distinct aliased pair exists. Thus Ξ_i is the largest possible action-effect discrepancy among contexts that the macro representation cannot distinguish; the supremum need not be attained.

Proposition 14 (State-aliasing obstruction)**.** *Every deterministic F_i satisfies*

$$\sup_{C \in \mathcal{C}_0} d(T_i(C), F_i(\Psi(C))) \geq \frac{1}{2} \Xi_i(\Psi; \mathcal{C}_0). \tag{93}$$

Proof of Proposition 14. For any distinct C, C' with $\Psi(C) = \Psi(C') = s$, the triangle inequality gives

$$d(T_i(C), T_i(C')) \leq d(T_i(C), F_i(s)) + d(F_i(s), T_i(C')). \tag{94}$$

At least one right-hand term is at least half the left-hand side. Taking the supremum over feasible distinct aliased pairs proves the claim. If no such pair exists, the defined diameter is zero and the inequality is trivial. □

This is a worst-case obstruction; rare aliased contexts need not prevent small average defect. For squared loss, Proposition 15 instead characterizes the minimum average defect exactly by conditional variance. Non-collapse, pre-action selectability, and on-policy composition remain additional requirements; definitions and rollout bounds are in Section F.

F.4 Causal rollout

Let I_j be the executed action and let $I_j^\star$ be a fixed or externally coupled reference action sequence on the same probability space. The reference sequence is held fixed under the coupling; comparing two state-dependent controllers requires an additional policy-stability or selection-mismatch term. Define

$$\epsilon_j^{\mathrm{abs}} = d\big(T_{I_j}(C_j), F_{I_j}(S_j)\big), \tag{95}$$

$$\chi_j^{\mathrm{sel}} = d\Big(F_{I_j}(S_j), F_{I_j^\star}(S_j)\Big), \tag{96}$$

where $S_j = \Psi(C_j)$. Let $\bar{S}_0 = S_0$ and $\bar{S}_{j+1} = F_{I_j^\star}(\bar{S}_j)$, and assume $F_{I_j^\star}$ is L_j-Lipschitz on a region containing both trajectories. Inserting $F_{I_j}(S_j)$ and $F_{I_j^\star}(S_j)$ yields

$$d(S_{j+1}, \bar{S}_{j+1}) \leq \epsilon_j^{\mathrm{abs}} + \chi_j^{\mathrm{sel}} + L_j d(S_j, \bar{S}_j). \tag{97}$$

Induction gives

$$d(S_H, \bar{S}_H) \le \sum_{j=0}^{H-1} \left(\prod_{r=j+1}^{H-1} L_r \right) (\epsilon_j^{\mathrm{abs}} + \chi_j^{\mathrm{sel}}). \tag{98}$$

This controls macro-transition mismatch for the coupled reference sequence. Comparing state-feedback controllers additionally requires control of how the reference policy changes when its input state changes.

F.5 An exact average-risk form of the aliasing obstruction

The supremum in Equation (93) is sensitive to rare contexts. An average-risk conclusion needs their probability, not just their maximum separation. Fix an action i, a probe-context distribution, $S = \Psi(C)$ in a standard Borel space, and a square-integrable actual action outcome $Y_i = T_i(C)$ in a separable real Hilbert space $\mathcal{H}$. A stochastic outcome can be included as long as its additional randomness is also averaged. Define $m_i(S) = \mathbb{E}[Y_i \mid S]$.

Proposition 15 (Average state-compression obstruction)**.** *Among all measurable predictors F with finite second moment,*

$$\inf_F \mathbb{E}\|Y_i - F(S)\|_{\mathcal{H}}^2 = \mathbb{E}\|Y_i - m_i(S)\|_{\mathcal{H}}^2 = \frac{1}{2}\mathbb{E}\|Y_i - Y_i'\|_{\mathcal{H}}^2, \tag{99}$$

where, conditional on S, Y_i and Y_i' are independent draws from the same conditional action-outcome law. The optimum is zero exactly when $Y_i = m_i(S)$ almost surely.

Proof. Insert $m_i(S)$ into $Y_i - F(S)$. Conditional on S, the residual $Y_i - m_i(S)$ has mean zero, so its inner product with $m_i(S) - F(S)$ has zero expectation. The risk equals $\mathbb{E}\|Y_i - m_i(S)\|^2 + \mathbb{E}\|m_i(S) - F(S)\|^2$. The first term is attained by $F = m_i$. Conditional independence makes the inner product of $Y_i - m_i(S)$ and $Y_i' - m_i(S)$ have zero conditional expectation; expanding their squared difference gives twice the first term. A nonnegative squared norm has zero mean exactly when it vanishes almost surely. □

For deterministic actions this residual quantifies action-relevant information lost by the state extractor. With stochastic outcomes it also includes irreducible execution noise; the theorem does not label all conditional variance as compression error. Restricting F to a neural class adds its approximation error and cannot improve the unrestricted infimum. Unlike the worst-case diameter, Equation (99) weights rare contexts by their probability.

G Controlled Experiments: Methods and Results

Overview. This appendix describes three complementary studies: controlled estimators test quantitative risk predictions, trained Transformers compare one-step and final-output performance, and interventions examine the effects attributed to expert calls. The primary experiment family and later extensions retain separate evaluation and confidence scopes. Sections G.1–G.6 specify the task, models, controlled estimation, trained comparisons, interventions, routing diagnostics, and paired performance evaluation. The remaining sections cover data budgets, transition holdouts, state-space size, state-conditioned variability, and general feature geometry.

G.1 Shared task, models, and evaluation protocol

The primary experiment family comprises fifteen trained models: three sharing architectures under the task-and-balance objective, and two additional supervision objectives on the tied architecture, each with three training seeds. The tied task-and-balance models are reused as the intervention baseline. Data-budget and generalization studies are described separately with their own settings and evaluation scope.

Task and splits. The task composes four operations on a finite state space,

$$F_0(s) = s + 1, \quad F_1(s) = s + 7, \quad F_2(s) = 2s, \quad F_3(s) = 3s + 5 \pmod{31},$$

and the stream $\langle \text{bos} \rangle, s_0, z_1, s_1, \ldots, z_H, s_H$, where $s_t = F_{z_t}(s_{t-1})$, and H is the number of operation application. The operation program is externally supplied, not generated by the model. A 37-symbol input vocabulary encodes state s as token $s + 2$ and operation z as token $z + 33$; a 31-way state head predicts the next state at operation-token positions. Training uses reference prefixes (teacher forcing), whereas free-running evaluation uses greedy state predictions as the subsequent inputs. The first four operations define a family; the protocol hashes their canonical bytes to assign 209 training, 22 development, and 25 test families. All initial states and continuations of one family stay together. Family assignment conditions on the entire four-operation prefix; it does not make those four coordinates independently uniform within a split. This withholds program-prefix families, not every local state–action transition. The strict transition holdout in Section G.8 is a different training and evaluation protocol.

The eight artificial estimation roles, three scalar composition roles, and four symbolic task operations belong to distinct experimental settings.

The primary pool contains 16,384 independently sampled trace records, equally split between training horizons 4 and 8; repeated contents are allowed under this sampling law. Three training seeds share this pool. Within a seed, compared models use the same horizon and minibatch-index schedule. Throughout this experimental appendix, H4, H8, H12, and H16 denote operation application number $H = 4, 8, 12, 16$, respectively. H8 is the primary free-running evaluation. H4 enumerates all 775 held-out-family traces; H12/H16 are descriptive length-extrapolation tests.

Models and fitting. All primary models have four pre-normalized causal Transformer layers, hidden width 256, four attention heads, sinusoidal positions, SwiGLU FFNs of inner width 768, no dropout, and no capacity dropping. Local has four distinct experts per layer; tied reuses one bank of four experts across depth with layer-specific routers and normalization; dense has one unrouted, active-width-matched FFN per layer. Top-1 routing retains the selected softmax gain. Local and tied initialize expert values equally but differ in parameter-object identity. Dense is an active-width control, not an identical initial function.

The training settings below distinguish the fifteen primary architecture/objective runs from the smaller model trained solely to supply frozen features. Follow-up experiments use the changes specified in their respective sections.

Setting	Primary runs	Feature pilot
Layers / hidden width / heads	$4/256/4$	$2/128/4$
Routed experts per bank / FFN inner width	$4/768$	$4/384$
Batch size / training records	$64/16{,}384$	$64/2{,}048$
Completed updates / trace presentations	$3{,}000/192{,}000$	$300/19{,}200$
Optimizer / peak learning rate	AdamW / 3×10^{-4}	AdamW / 3×10^{-4}
Betas / optimizer epsilon	$(0.9, 0.95)/10^{-8}$	$(0.9, 0.95)/10^{-8}$
Weight decay / gradient-norm clip	$0.01/1$	$0.01/1$
Linear warmup / subsequent schedule	150 steps / cosine	15 steps / cosine
Training autocast / parameters and state	BF16 / FP32	BF16 / FP32

Evaluation uses FP32. Trace presentations count repeated uses of the training pool. Training and neural-model evaluation use NVIDIA RTX A5000 24 GB GPUs, with one card for the primary experiments and a second for the follow-ups. Controlled linear calculations use float64. Training time, memory, unique parameters, and dispatched computation are reported separately.

Checkpoint selection and development data. The primary comparisons use all fifteen checkpoints at the final training update, 3,000. Development monitoring uses a fixed 512-trace set, evaluated every 500 updates. The coefficients in Section G.4 specify the three training objectives. Candidate checkpoints, action mappings, evaluation losses, comparison families, and sample sizes are fixed before generating the primary evaluation banks. The mappings are selected using development data as described in Section G.4. Follow-up studies retain their own protocols and are not included retrospectively in the primary comparison family.

Outcomes and independent units. Terminal error is the indicator of an incorrect final state; complete-trace error indicates any incorrect predicted state along a free-running trace. A trace can recover from an intermediate error, so these are different endpoints. Teacher-forced step error averages state-prediction mistakes on reference prefixes; it is not the event-lift role loss of Equation (3) and is not a substitute for free-running performance. A complete trace is the independent evaluation unit. Multiple forced actions on one context are averaged before across-context uncertainty is calculated. Across-seed standard deviations describe training variability, not additional audit samples.

G.2 Controlled local estimation and composition

Local estimation gains under controlled sharing. Frozen, nonorthogonal Transformer features define a linear residual-estimation problem. All 54 estimation settings agree with their covariance-aware references within five Monte Carlo standard errors; two repetitions on trained features satisfy the same numerical agreement criterion (Section G.2).

Frozen Transformer feature interface. The frozen-feature estimation study uses a separately trained two-layer, width-128 Transformer after 300 updates. For each probe trace, one prediction position is selected uniformly. The feature is the normalized input to the expert subblock in the second Transformer layer, after its FFN pre-normalization. Features are extracted in evaluation mode without gradients in FP32, converted to float64, and mapped to 32 dimensions by a fixed random orthonormal 128×32 projection. No centering, whitening, or post-hoc direction selection is applied. A controlled residual call uses $T_a[A](z) = h(z) + JAu(z)$ with $J^\top J = I_{16}$, so the squared difference between complete event outputs equals $\|(A - A_a)u(z)\|^2$. Eight artificial statistical roles are defined independently of the symbolic operation, trace content, and layer.

Each role has 64 fixed nonorthogonal design rows. Repeating those rows one, four, or sixteen times, with fresh independent Gaussian label noise of standard deviation 0.1, gives $n = 64, 256, 1024$ independent measurements. The input support does not increase with n. Test geometry is the empirical second moment of a separately frozen 8,192-record feature pool. Targets are held constant across the measurement-count sweep and calibrated to three heterogeneity levels $c = 0, 0.5, 2$. Designs, targets, features, surrounding state, and test geometry are fixed before drawing fitting noise. The 32-dimensional test second moment has condition number approximately 1.95×10^7; the eight design singular-value ratios range from 1.20×10^{-4} to 1.52×10^{-4}. The fitting procedure uses neither ridge regularization nor whitening.

More precisely, the calibration sets $B(c) = cC_{256}$, where $B(c)$ is target sharing bias in this fixed test geometry and C_{256} is the estimation variance removable by pooling at $n = 256$. The dimensionless c therefore fixes target heterogeneity relative to one reference variance scale. Writing $V_0(n)$ for separate-estimation risk in this geometry, the reported eight-role construction has $C_n = (7/8)V_0(n)$. Repeated design rows make these variance quantities scale as $1/n$, whereas $B(c)$ stays fixed. Thus c is a target-construction parameter.

Float64 role-wise least squares is solved by reduced QR and followed by the test-geometry projection of Equation (215). The five fixed coefficients $0, 0.25, 0.5, 0.75, 1$ and the population oracle are evaluated on the same noisy observations within each of 128 independent repetitions. Observed risk is computed from the fitted matrices; the analytic reference is stored separately. All 54 estimation settings meet the numerical comparison criterion $|\text{mean} - \text{reference}| \le \max(5\,\text{MCSE}, 10^{-8})$.Full-pooling ratios cross one between the sampled measurement counts in the nonzero-heterogeneity settings, consistent with the specified reference.

Table 4: Frozen-feature estimation endpoints from 128 fitting-noise repetitions. Parentheses are Monte Carlo standard errors (MCSE). Ratios divide the two means.

c	n	Separate (MCSE)	Full (MCSE)	Reference ratio	Observed ratio
0	64	0.5329 (0.0096)	0.0664 (0.0012)	0.1250	0.1246
0	256	0.1355 (0.0022)	0.0169 (0.0003)	0.1250	0.1250
0	1024	0.0331 (0.0005)	0.0042 (0.0001)	0.1250	0.1255
0.5	64	0.5329 (0.0096)	0.1240 (0.0012)	0.2344	0.2327
0.5	256	0.1355 (0.0022)	0.0745 (0.0003)	0.5625	0.5503

c	n	Separate (MCSE)	Full (MCSE)	Reference ratio	Observed ratio
0.5	1024	0.0331 (0.0005)	0.0618 (0.0001)	1.8750	1.8644
2	64	0.5329 (0.0096)	0.2968 (0.0012)	0.5625	0.5571
2	256	0.1355 (0.0022)	0.2474 (0.0003)	1.8750	1.8262
2	1024	0.0331 (0.0005)	0.2346 (0.0001)	7.1250	7.0810

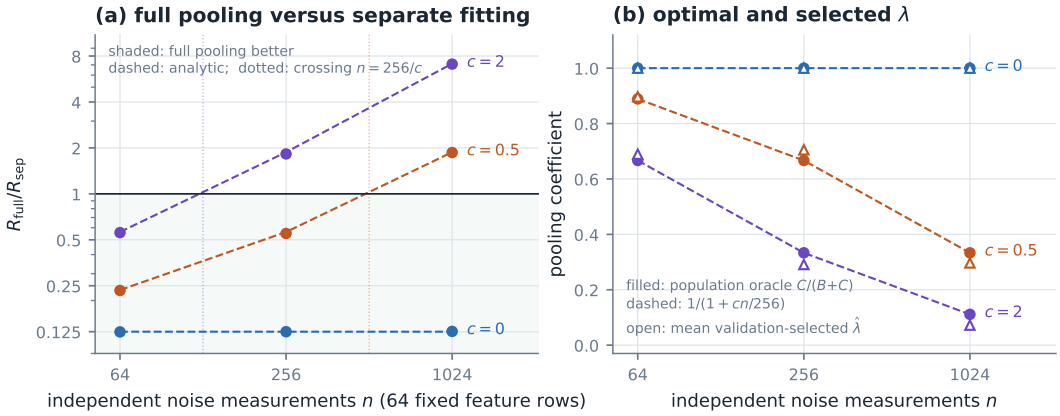

Figure 2: Frozen-feature estimation under fixed target heterogeneity. Observed full/separate ratios and the independently selected coefficients are distinct from the analytic lines and population oracle. More measurements of the same feature rows reduce noise without removing fixed target heterogeneity.

A realizable selector uses 4,096 fresh noisy validation records per fit to choose among the five fixed coefficients. Its risk and coefficient frequencies are reported separately in Table 5; the oracle target is not made available to selection. The squared Gaussian loss is unbounded, so the bounded primary audit radius is not applied to this estimation study. The trained-feature replication repeats this inserted-estimator design on features from two trained step-3000 checkpoints, with condition numbers about 3,002 and 2,625, respectively; both 54-setting repetitions satisfy the same numerical comparison criterion.

Table 5: Validation-selected pooling (4,096 fresh validation records per fit). Frequencies follow the grid 0, 0.25, 0.5, 0.75, 1 and sum to 128; conditional finite-grid excess is computed after selection.

c	n	Selected risk	Excess	Five selection counts
0.0	64	0.0664	0.000000	0/0/0/0/128
0.5	64	0.1241	0.000782	0/0/0/53/75
2.0	64	0.2245	0.000638	0/0/31/97/0
0.0	256	0.0169	0.000000	0/0/0/0/128
0.5	256	0.0567	0.000169	0/0/23/105/0
2.0	256	0.0976	0.000373	0/107/21/0/0
0.0	1024	0.0042	0.000000	0/0/0/0/128
0.5	1024	0.0240	0.000096	0/104/24/0/0
2.0	1024	0.0329	0.000091	91/37/0/0/0

Table 6: Trained-feature replication repeats the frozen-estimator protocol using features from two completed step-3000 checkpoints. Counts use the five-MCSE numerical-agreement criterion.

Feature source	Condition number	Cells	Flagged
trained-local-s0	3002	54	0
trained-tied-s0	2625	54	0

Local improvements can reverse at the final output. The executed three-event construction of Proposition 47 evaluates the *same fitted coefficients* locally and after composition. Full pooling lowers observed role-average squared error from 1.0090 to 0.8359, while final-output squared error rises from 6.0086 to 9.0000 (Figure 3). The corresponding analytic optima also differ: $\lambda = 4/7$ locally versus $2/5$ at the endpoint. Thus a local improvement can accompany worse final output in this controlled program.

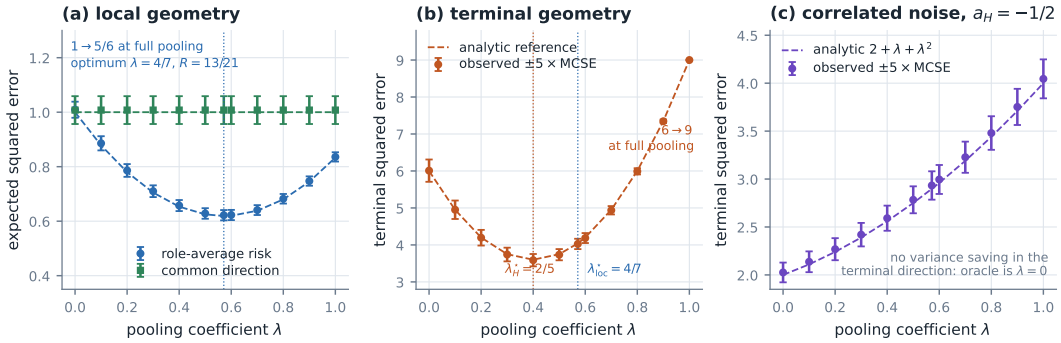

Figure 3: **Controlled local-to-final risk reversal.** The same estimates are evaluated in local and final-output geometries. Markers are reported means with $\pm 5\,\mathrm{MCSE}$ over 20,000 noise repetitions; dashed curves are analytic references. Full pooling helps locally but hurts the final output, and the optimal pooling strengths differ. The correlated-noise case isolates a further limit of pooling. Panels (a) and (b) use the same three-role estimates; (c) is a separate two-dimensional model. The common-direction normalization and both constructions are specified in Section G.2.

Composition of the same estimates. The controlled composition study executes the existing three-event construction in Proposition 47, with $\theta = (1/2, -1, 1/2)^\top$, independent unit-variance coefficient noise, and common-mean pooling. The implementation evaluates both the actual recurrence and its fixed linear terminal map on 20,000 noise repetitions. The role-average, common-direction, and terminal geometries are distinct evaluations of the same estimates. Full pooling changes observed local risk from 1.0090 to 0.8359 and terminal risk from 6.0086 to 9.0000, against references $1 \to 5/6$ and $6 \to 9$. The terminal full-pooling output is constant, hence its MCSE is zero.

In Figure 3(a), the common-direction curve measures the squared scalar error projection

$$\mathbb{E}\left[\left(\frac{\mathbf{1}^\top(L_\lambda Y - \theta)}{\sqrt{3}}\right)^2\right], \qquad \mathbf{1} = (1,1,1)^\top.$$

Its analytic value is 1 at every displayed pooling strength. This is a projection onto a unit direction, rather than role-average risk or the squared error of the arithmetic mean. Panels (a) and (b) evaluate the same three-role estimates. Panel (c) instead uses the separately declared two-dimensional model from Appendix N.1:

$$\theta = 0, \quad P = \mathrm{diag}(1,0), \quad G = (1,1), \quad \Sigma = \begin{pmatrix} 4 & -3/2 \\ -3/2 & 1 \end{pmatrix}.$$

In that model, suppressing a component removes beneficial noise cancellation, so positive pooling increases terminal risk.

Table 7: Controlled composition: selected points; 20,000 noise repetitions for each of the two models. Reference and observed values are separate columns. The full-pooling terminal loss is constant, so its MCSE is zero.

Model / geometry	λ	Reference	Observed	MCSE
reversal/local	0.0	1.0000	1.0090	0.00590
reversal/local	0.4	0.6533	0.6574	0.00406
reversal/local	1.0	0.8333	0.8359	0.00341

Model / geometry	λ	Reference	Observed	MCSE
reversal/common	0.0	1.0000	1.0077	0.01022
reversal/common	0.4	1.0000	1.0077	0.01022
reversal/common	1.0	1.0000	1.0077	0.01022
reversal/terminal	0.0	6.0000	6.0086	0.06031
reversal/terminal	0.4	3.6000	3.5875	0.03341
reversal/terminal	1.0	9.0000	9.0000	0.00000
correlated/terminal	0.0	2.0000	2.0257	0.02060
correlated/terminal	0.4	2.5600	2.5917	0.02622
correlated/terminal	1.0	4.0000	4.0451	0.04064

G.3 SHARED, UNSHARED, AND DENSE TRANSFORMERS

Nine primary runs compare the three architectures at three seeds, all at the full budget. Tied has 3,433,984 distinct parameters versus 10,511,872 for local, a 67.3% reduction; dense has 3,429,888. The reported H8 audit contains one erroneous trace across the nine checkpoints, produced by local seed 1. All twelve architecture intervals cover zero. The tested task is saturated at this budget. Matching active FFN width does not establish equal total compute or elapsed time. The per-seed resource quantities are reported separately below.

Table 8: Shared, unshared, and dense Transformers at training update 3000. Errors are percentages on 32,768 H8 traces per checkpoint; training time excludes I/O. Parameter counts are millions; peak is allocated MiB.

Model	Seed	Params (M)	Terminal (%)	Trace (%)	Step (ms)	Peak
local	0	10.5119	0.0000	0.0000	31.7	278.3
local	1	10.5119	0.0031	0.0031	32.5	279.0
local	2	10.5119	0.0000	0.0000	32.5	279.3
tied	0	3.4340	0.0000	0.0000	29.9	184.8
tied	1	3.4340	0.0000	0.0000	30.5	184.5
tied	2	3.4340	0.0000	0.0000	30.2	185.1
dense	0	3.4299	0.0000	0.0000	12.7	172.1
dense	1	3.4299	0.0000	0.0000	12.4	172.1
dense	2	3.4299	0.0000	0.0000	12.9	172.1

H4 is also solved on its full finite test population. At H12/H16, terminal error ranges approximately from 0.964 to 0.972 and complete-trace error is at least 0.999 in the supplied primary records. The uniform 31-state guessing error is a scale reference. These trained systems do not demonstrate successful length extrapolation under the reported settings.

Table 9: Primary horizon diagnostics (percent error): H4 enumerates all 775 family-held-out traces; H12/H16 use 4,096 traces per checkpoint. Each horizon lists terminal / complete-trace error.

Model	Seed	H4 term.	H4 trace	H12 term.	H12 trace	H16 term.	H16 trace
local	0	0.000	0.000	96.826	99.951	96.973	100.000
local	1	0.000	0.000	96.631	99.927	96.802	100.000
local	2	0.000	0.000	97.046	100.000	97.119	100.000
tied	0	0.000	0.000	96.899	100.000	96.655	100.000
tied	1	0.000	0.000	97.046	99.902	96.851	100.000
tied	2	0.000	0.000	97.192	100.000	97.070	100.000
dense	0	0.000	0.000	96.387	100.000	96.704	100.000
dense	1	0.000	0.000	97.095	100.000	96.753	100.000
dense	2	0.000	0.000	97.070	100.000	97.217	100.000

G.4 Route labels and executable interventions

Objectives and interventions. The three training objectives are cumulative:

$$\mathcal{L}_B = \mathcal{L}_{\text{task}} + 0.01\mathcal{L}_{\text{bal}},$$
$$\mathcal{L}_R = \mathcal{L}_B + 0.1\mathcal{L}_{\text{route}},$$
$$\mathcal{L}_I = \mathcal{L}_R + 0.5\mathcal{L}_{\text{CI}}.$$

The task loss is cross-entropy for the next state. The balancing term is computed over all input tokens, while route supervision is cross-entropy between natural per-layer router logits and the requested operation at state-prediction positions. The natural token/update budget is matched. I also receives a counterfactual forward pass and extra target per training trace, so neither its total computation nor its supervision is matched to R. B is the tied baseline from the architecture comparison, reused rather than retrained: the primary family has fifteen unique runs.

The intervention loss independently draws t uniformly from $\{1, \ldots, H\}$ and operation i uniformly from $\{0, 1, 2, 3\}$ for each training trace. At the corresponding state-prediction position in every primary-model layer, it forces expert i with that layer's detached natural-pass gain; this is a four-layer compound action. The requested operation token and pre-action prefix remain unchanged; the target is $F_i(s_{t-1})$, the chosen operation applied to the pre-action state. Future target tokens and post-action reference states are not inserted. For B, which has no route-label supervision, an operation-to-expert permutation is selected separately at each layer on 4,096 development contexts. All 24 permutations are enumerated; the chosen permutation maximizes natural label agreement, with lexicographic tie-breaking. R/I retain their label-defined identity mapping for the main comparison. Neither final forced fidelity nor test data are used to select the mapping.

At the forced position, let z_ℓ^{nat} denote layer ℓ's router logits in the original natural pass. Its recorded gain is the natural winner's probability, $q_\ell^{\text{nat}} = \max_e \text{softmax}(z_\ell^{\text{nat}})_e$. After detachment, this fixed value multiplies the forced expert's output. The convention preserves the original winner's gain rather than looking up the forced expert's original probability or recomputing a gain during intervention. All layers' reference gains come from the same original natural pass. The forced pass still computes its own activations, while the residual and other shared paths retain their declared computation. The operation-to-expert mapping remains development-selected for B and the identity for R/I. This is the frozen-gain convention of Section P.4, distinct from current-gain support replay in Section P.3.

Evaluation. For each of 32,768 independent H8 teacher-prefix contexts, all four actions are forced and averaged. Forced fidelity is the average indicator that the resulting state equals the declared operation's output. Cross-action fidelity excludes the originally requested action. Action separation compares outputs of distinct forced actions and must be read alongside target separation on the same states. The uniform-state, uniformly requested and forced collapsed-control reference is $67/248$, not $1/4$, because several operation outputs coincide; this control distribution is not automatically the distribution of every model-specific probe. Same-index forcing, causal-mask, and explicit-operator controls test implementation rather than learned performance.

The main action is a four-layer compound intervention. Single-layer tests use a separate 2,048-context bank, changing only one call. On-policy diagnostics evaluate each model on its own reached-state distribution. The label-agreement columns use development scores; primary compound fidelity is evaluated on the independent test contexts. These columns therefore describe different evaluation sets.

R and I both have development label agreement 1, but compound fidelities are 0.9467 ± 0.0171 and 1.0000 ± 0.0000, versus 0.2528 ± 0.0079 for B. The I checkpoints have single-layer fidelity 0.3274 ± 0.0247 when averaged over layers. This contrasts attribution scopes; it does not prove that individual experts have no useful partial function. All these averages are descriptive over three seeds.

Table 10: Per-seed primary causal results. Raw/aligned label scores use development data; compound fidelity uses final teacher contexts; L0–L3 are separate single-layer secondary interventions.

Obj.	Seed	Raw	Aligned	Compound	L0	L1	L2	L3
B	0	0.2621	0.3605	0.2567	0.267	0.267	0.266	0.267

Obj.	Seed	Raw	Aligned	Compound	L0	L1	L2	L3
B	1	0.2051	0.3790	0.2438	0.267	0.267	0.257	0.270
B	2	0.2657	0.3297	0.2580	0.267	0.266	0.270	0.267
R	0	1.0000	1.0000	0.9632	0.279	0.267	0.293	0.269
R	1	1.0000	1.0000	0.9291	0.290	0.267	0.276	0.269
R	2	1.0000	1.0000	0.9478	0.292	0.268	0.269	0.267
I	0	1.0000	1.0000	1.0000	0.401	0.267	0.269	0.267
I	1	1.0000	1.0000	1.0000	0.597	0.267	0.269	0.267
I	2	1.0000	1.0000	1.0000	0.520	0.268	0.267	0.267

Table 11: On-policy secondary diagnostics, 2,048 contexts per checkpoint. Each model reaches its own context distribution. Mean ± sample SD over seeds.

Objective	Fidelity	Cross-fidelity	Separation	Target sep.
B	0.2499 ± 0.0093	0.1033 ± 0.0095	0.5309 ± 0.0270	0.9744 ± 0.0000
R	0.9455 ± 0.0202	0.9273 ± 0.0269	0.9658 ± 0.0015	0.9744 ± 0.0000
I	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9744 ± 0.0000	0.9744 ± 0.0000

Intervention effects are a separate question. Route-supervised models at 64 traces have label agreement 1 but forced-action fidelity 0.3713 ± 0.0267. At the primary budget, intervention supervision gives compound fidelity 1 without establishing single-expert attribution. Separately holdout-trained models retain seen fidelity 1 but reach only 0.0021 ± 0.0037 on withheld transitions. A separately trained follow-up retains the seen/withheld fidelity gap at moduli 97 and 251 (Section G.8.1). These address causal attribution and coverage; full tables remain in Sections G.4 and G.8.

G.5 Functional routing, margin, and gain diagnostics

The routing diagnostics use six frozen local/tied checkpoints and 4,096 independent development sites per checkpoint. The nearest, farthest, and random alternative experts are evaluated on one common pre-call input with the same natural gain. The reported local event difference is checked against the residual-output distance. Selecting the nearest alternative makes its local distance smallest by definition; whether it has smaller terminal harm is an empirical question. A logical-site substitution persists through later full-prefix recomputation, while subsequent unforced routes respond naturally. Candidate search is additional diagnostic work.

The measurements show positive distance–harm associations in many accurate primary checkpoints. The lower-budget routing follow-up uses 24 lower-budget checkpoints, 1,024 sites each, and includes repairs as negative loss differences. Associations weaken or change sign in some settings. Downstream first-disagreement summaries refer to the instrumented operation-token positions, not every token and cache event in the computation.

Table 12: Functional expert-replacement diagnostics, 4,096 sites per checkpoint. Harm is intervened minus natural terminal error and is not clipped. Distance ordering is by construction; terminal association is empirical. SE is across independent sampled sites as reported.

Checkpoint	Alternative	Distance	Harm (pp)	SE (pp)	Spearman
local-s0	nearest	4.458	8.618	0.439	0.172
local-s0	random	4.954	13.843	0.540	0.270
local-s0	farthest	5.456	19.189	0.615	0.378
tied-s0	nearest	3.576	0.781	0.138	0.013
tied-s0	random	3.884	2.759	0.256	0.066
tied-s0	farthest	4.190	5.029	0.342	0.115
local-s1	nearest	4.024	6.396	0.382	0.149
local-s1	random	4.560	10.864	0.486	0.246
local-s1	farthest	5.067	15.015	0.558	0.318
tied-s1	nearest	2.846	1.733	0.204	0.017

Checkpoint	Alternative	Distance	Harm (pp)	SE (pp)	Spearman
tied-s1	random	3.119	4.175	0.313	0.089
tied-s1	farthest	3.430	6.006	0.371	0.132
local-s2	nearest	4.717	7.153	0.403	0.162
local-s2	random	5.218	12.695	0.520	0.266
local-s2	farthest	5.699	18.579	0.608	0.365
tied-s2	nearest	3.375	0.146	0.060	-0.008
tied-s2	random	3.727	1.709	0.203	0.041
tied-s2	farthest	4.101	3.223	0.276	0.078

Table 13: Lower-budget expert-replacement diagnostics outside the saturated primary regime. Each checkpoint uses 1,024 sites; selected nearest and farthest summaries are shown. Negative harm represents repairs.

Checkpoint	Alternative	Natural error	Harm (pp)	Repair (%)	Spearman
EXT-local-B-n128-s0	nearest	0.782	0.68	1.27	-0.038
EXT-local-B-n128-s0	farthest	0.782	3.91	1.66	0.031
EXT-local-B-n128-s1	nearest	0.712	2.64	1.07	-0.007
EXT-local-B-n128-s1	farthest	0.712	5.08	2.15	0.137
EXT-local-B-n128-s2	nearest	0.778	0.59	1.07	-0.021
EXT-local-B-n128-s2	farthest	0.778	3.71	1.66	0.020
EXT-local-B-n192-s0	nearest	0.017	10.64	0.00	0.134
EXT-local-B-n192-s0	farthest	0.017	22.56	0.00	0.261
EXT-local-B-n192-s1	nearest	0.008	8.30	0.00	0.101
EXT-local-B-n192-s1	farthest	0.008	18.95	0.10	0.264
EXT-local-B-n192-s2	nearest	0.001	9.18	0.00	0.125
EXT-local-B-n192-s2	farthest	0.001	21.29	0.00	0.357
EXT-local-B-n256-s0	nearest	0.000	8.30	0.00	0.141
EXT-local-B-n256-s0	farthest	0.000	17.77	0.00	0.309
EXT-local-B-n256-s1	nearest	0.010	8.69	0.00	0.128
EXT-local-B-n256-s1	farthest	0.010	17.48	0.00	0.345
EXT-local-B-n256-s2	nearest	0.000	8.01	0.00	0.101
EXT-local-B-n256-s2	farthest	0.000	23.14	0.00	0.420
EXT-local-B-n96-s0	nearest	0.957	0.10	0.59	-0.043
EXT-local-B-n96-s0	farthest	0.957	0.00	1.07	-0.033
EXT-local-B-n96-s1	nearest	0.965	0.59	0.29	-0.058
EXT-local-B-n96-s1	farthest	0.965	-0.59	1.17	-0.065
EXT-local-B-n96-s2	nearest	0.739	3.71	1.07	0.001
EXT-local-B-n96-s2	farthest	0.739	7.13	1.27	0.029
EXT-tied-B-n128-s0	nearest	0.462	1.37	1.17	-0.011
EXT-tied-B-n128-s0	farthest	0.462	5.47	1.37	0.045
EXT-tied-B-n128-s1	nearest	0.859	0.88	0.29	-0.061
EXT-tied-B-n128-s1	farthest	0.859	0.10	1.17	-0.048
EXT-tied-B-n128-s2	nearest	0.510	3.22	0.68	0.008
EXT-tied-B-n128-s2	farthest	0.510	5.08	0.88	0.061
EXT-tied-B-n192-s0	nearest	0.036	1.56	0.00	-0.066
EXT-tied-B-n192-s0	farthest	0.036	3.81	0.10	-0.000
EXT-tied-B-n192-s1	nearest	0.064	2.83	0.29	0.029
EXT-tied-B-n192-s1	farthest	0.064	6.84	0.39	0.130
EXT-tied-B-n192-s2	nearest	0.002	4.10	0.00	0.025
EXT-tied-B-n192-s2	farthest	0.002	9.77	0.00	0.128
EXT-tied-B-n256-s0	nearest	0.007	2.44	0.10	0.015
EXT-tied-B-n256-s0	farthest	0.007	6.45	0.10	0.114
EXT-tied-B-n256-s1	nearest	0.000	1.76	0.00	-0.043
EXT-tied-B-n256-s1	farthest	0.000	7.32	0.00	0.088
EXT-tied-B-n256-s2	nearest	0.000	1.76	0.00	-0.056
EXT-tied-B-n256-s2	farthest	0.000	7.23	0.00	0.102
EXT-tied-B-n96-s0	nearest	0.952	-0.10	1.07	-0.050
EXT-tied-B-n96-s0	farthest	0.952	-0.29	1.07	-0.052
EXT-tied-B-n96-s1	nearest	0.953	0.29	0.59	-0.029
EXT-tied-B-n96-s1	farthest	0.953	0.88	0.59	-0.021
EXT-tied-B-n96-s2	nearest	0.850	0.98	0.39	-0.026

Checkpoint	Alternative	Natural error	Harm (pp)	Repair (%)	Spearman
EXT-tied-B-n96-s2	farthest	0.850	0.88	0.98	-0.017

The margin-and-gain diagnostic records logits before and after direct score perturbations, common shifts, and hidden-state perturbations. Let $\gamma = z_{(1)} - z_{(2)}$ be the original winner–runner-up logit gap. A direct perturbation lowers the winner by $c\gamma$ and raises the runner-up by $c\gamma$, leaving other logits unchanged; this c is a perturbation scale, not the target-heterogeneity parameter in Section G.2. Hidden-state perturbations use $h' = h + \alpha \max\{\|h\|_2, 1\}v$, where v is the perturbation-direction vector, and then recompute normalization and routing. The table gives the tested c and α values. With $\epsilon = \|\widetilde{z} - z\|_\infty$ and $\mathrm{tol} = 64\varepsilon_{\mathrm{mach}}(1 + \max_e |z_e|)$, where $\varepsilon_{\mathrm{mach}}$ is the floating-point machine precision, a pair is certified if $\gamma - 2\epsilon > \mathrm{tol}$, boundary if $|\gamma - 2\epsilon| \le \mathrm{tol}$, and condition-not-met otherwise. It evaluates the strict half-margin condition in Lemma 27 from the actual scores with a floating-point tolerance. Across the measured perturbation pairs, 150,095 satisfy the strict condition with slack exceeding the floating-point tolerance, with zero natural-support violations among them. Boundary cases are retained. The same-support gain measurement also reports nonzero coefficient changes, so support stability is not update equality. This is a sampled local diagnostic, not the trajectory-level conditions or a non-vacuous numerical certificate for the general routed bound.

Table 14: Router-margin and gain summaries pooled over six checkpoints. A certified violation is a natural support change despite the strict half-margin condition with numerical tolerance. Failure of the sufficient condition does not require a support change.

Perturbation	Pairs	Certified	Boundary	Not met	Violations	Changed/not met
direct_c=0	24576	24574	2	0	0	0
direct_c=0.24	24576	24573	3	0	0	0
direct_c=0.49	24576	24515	61	0	0	0
direct_c=0.5	24576	0	24576	0	0	0
direct_c=0.51	24576	0	61	24515	0	24515
direct_c=1.0	24576	0	2	24574	0	24574
common_logit_shift	24576	4224	0	20352	0	0
hidden_alpha=0.0001	24576	24557	4	15	0	5
hidden_alpha=0.001	24576	24409	3	164	0	32
hidden_alpha=0.01	24576	23243	0	1333	0	255

G.6 Independent paired performance evaluation

The primary evaluation comprises eighteen paired comparisons: tied/local and dense/local at each of three seeds under terminal and complete-trace error (twelve), plus R/B and I/R under $1 - \mathrm{Fid}$ (six). Loss differences are reference minus candidate. We apply both directions of Proposition 45, with thirty-six tails, total failure allowance 0.05, and 32,768 independent traces/contexts per checkpoint. The resulting two-sided radius is 2.0039 percentage points.

Architecture and intervention comparisons use two independent evaluation banks, each containing 32,768 held-out-family H8 traces. Within each comparison, candidate and reference predictions are paired on the same records. The resulting intervals cross zero for all twelve architecture comparisons and favor the candidate for all six supervision comparisons. These conclusions concern fixed checkpoints on the declared evaluation distributions. Four forced actions are averaged within each context before across-context uncertainty is computed; neither those actions nor the three training seeds increase the independent sample count. The fixed-family guarantee requires candidates to be fixed before the independent evaluation data are inspected.

Table 15: All 18 primary paired comparisons. Positive differences favor the first model; numbers are percentage points. Every interval has radius 2.0039 pp (36 tails, $\delta = 0.05$, $N = 32768$).

Candidate/ref.	Seed	Loss	Difference	Lower	Upper
tied/local	0	terminal	0.0000	-2.0039	2.0039
tied/local	0	trace	0.0000	-2.0039	2.0039

Candidate/ref.	Seed	Loss	Difference	Lower	Upper
dense/local	0	terminal	0.0000	-2.0039	2.0039
dense/local	0	trace	0.0000	-2.0039	2.0039
R/B	0	1–Fid	70.6573	68.6534	72.6613
I/R	0	1–Fid	3.6774	1.6735	5.6813
tied/local	1	terminal	0.0031	-2.0009	2.0070
tied/local	1	trace	0.0031	-2.0009	2.0070
dense/local	1	terminal	0.0031	-2.0009	2.0070
dense/local	1	trace	0.0031	-2.0009	2.0070
R/B	1	1–Fid	68.5387	66.5348	70.5426
I/R	1	1–Fid	7.0854	5.0815	9.0893
tied/local	2	terminal	0.0000	-2.0039	2.0039
tied/local	2	trace	0.0000	-2.0039	2.0039
dense/local	2	terminal	0.0000	-2.0039	2.0039
dense/local	2	trace	0.0000	-2.0039	2.0039
R/B	2	1–Fid	68.9774	66.9734	70.9813
I/R	2	1–Fid	5.2246	3.2207	7.2285

G.7 Local-to-final gains across data budgets and seeds

The training-data sweep varies the training-record budget over 17 values from 16 to 16,384, using the same 3,000-update policy. The original three seeds give 153 architecture configurations. The original interval families stay separate from the ten-seed follow-up at budgets 64, 96, 128, and 192.

A configuration specifies an architecture, budget, and seed; a contrast instead specifies a model pair, budget, and seed. Here the three architectures also yield three pairs: tied/local, dense/local, and dense/tied. Thus both counts happen to equal $17 \times 3 \times 3 = 153$ in the original sweep. Each contrast has one local and one terminal endpoint, giving 306 endpoint comparisons and 612 one-sided tails for two-sided simultaneous intervals.

Budget- and seed-dependent comparisons. The data do not satisfy dense $\leq$ tied $\leq$ local at every budget. The original three-seed means already have exceptions at 24, 192, and 256; at 192, teacher-forced step errors are approximately 0 (dense), 0.00109 (local), and 0.00496 (tied). At 64 traces the three original tied/local step improvements are 5.18, 5.16, and 6.49 percentage points, whereas terminal differences are approximately -0.043, -0.006, -0.143 points with intervals covering zero. An established local prediction gain therefore does not imply a detectable gain on that terminal endpoint.

Table 16: Original three-seed architecture data-budget sweep: teacher-forced step error on the reported H8 bank (153 configuration summaries). Values are mean $\pm$ sample SD over three training seeds.

Training traces	Local	Tied	Dense
16	0.9467 ± 0.0034	0.9436 ± 0.0032	0.9389 ± 0.0065
24	0.9322 ± 0.0008	0.9245 ± 0.0039	0.9247 ± 0.0110
32	0.9167 ± 0.0035	0.8999 ± 0.0087	0.8917 ± 0.0130
48	0.8651 ± 0.0107	0.8224 ± 0.0239	0.8160 ± 0.0524
64	0.7959 ± 0.0220	0.7398 ± 0.0293	0.7033 ± 0.0488
96	0.4133 ± 0.2122	0.3670 ± 0.1186	0.1377 ± 0.0459
128	0.1638 ± 0.0119	0.1433 ± 0.0950	0.0179 ± 0.0186
192	0.0011 ± 0.0009	0.0050 ± 0.0043	0.0000 ± 0.0000
256	0.0004 ± 0.0006	0.0004 ± 0.0006	0.0000 ± 0.0000
384	0.0002 ± 0.0002	0.0000 ± 0.0000	0.0000 ± 0.0000
512	0.0001 ± 0.0002	0.0000 ± 0.0000	0.0000 ± 0.0000
768	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
1024	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
2048	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
4096	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
8192	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
16384	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000

Table 17: Free-running terminal error for exactly the same original three-seed architecture checkpoints and H8 bank as the preceding step-error table. Values are mean ± sample SD.

Training traces	Local	Tied	Dense
16	0.9671 ± 0.0014	0.9678 ± 0.0008	0.9668 ± 0.0011
24	0.9665 ± 0.0006	0.9666 ± 0.0004	0.9673 ± 0.0010
32	0.9680 ± 0.0005	0.9671 ± 0.0009	0.9679 ± 0.0006
48	0.9667 ± 0.0009	0.9673 ± 0.0012	0.9684 ± 0.0003
64	0.9679 ± 0.0006	0.9685 ± 0.0001	0.9677 ± 0.0008
96	0.8883 ± 0.1295	0.9174 ± 0.0565	0.6444 ± 0.1314
128	0.7580 ± 0.0342	0.6342 ± 0.2061	0.1304 ± 0.1300
192	0.0086 ± 0.0072	0.0387 ± 0.0334	0.0000 ± 0.0000
256	0.0030 ± 0.0049	0.0030 ± 0.0049	0.0000 ± 0.0000
384	0.0019 ± 0.0017	0.0000 ± 0.0000	0.0000 ± 0.0000
512	0.0011 ± 0.0019	0.0000 ± 0.0000	0.0000 ± 0.0000
768	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
1024	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
2048	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
4096	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
8192	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
16384	0.0000 ± 0.0001	0.0000 ± 0.0000	0.0000 ± 0.0000

Table 18: Tied versus local at 64 training traces, original architecture-sweep family. Values and bounds are percentage points. Differences are local minus tied error. The intervals retain the full 153-pair, 306-endpoint, 612-tail sweep budget ($N = 32768$, $\delta = 0.05$; radius 2.39685 percentage points).

Seed	Step diff.	Step lower	Step upper	Term. diff.	Term. lower	Term. upper
0	5.183	2.786	7.580	-0.043	-2.440	2.354
1	5.162	2.765	7.559	-0.006	-2.403	2.391
2	6.488	4.092	8.885	-0.143	-2.540	2.253

In the balanced ten-seed follow-up at 96 traces, the local tied/local contrast is supported in opposite directions: six seeds favor tied and four favor local. At 64, all ten favor tied on the local endpoint. This motivates reporting seeds. The 153 original local/terminal configuration contrasts and 120 follow-up contrasts overlap in 36 configurations, leaving 237 unique configurations and 84 newly added contrasts. Their original confidence budgets are not combined into a new 237-comparison family.

The follow-up uses four budgets (64, 96, 128, 192), seeds 0–9, and the same three model pairs: $4 \times 10 \times 3 = 120$ contrasts, or 240 endpoints and 480 one-sided tails for two-sided intervals. Figure 1 displays the 40 tied/local contrasts and their two endpoints while retaining that full-family interval budget.

Table 19: Balanced ten-seed architecture follow-up; means and sample SD are computed from run summaries. Last column counts lower / higher / inconclusive local risk versus local under the separate follow-up intervals.

Traces	Model	Step error	Terminal error	Lower/higher/unclear
64	local	0.7684 ± 0.0233	0.9678 ± 0.0009	–
64	tied	0.7127 ± 0.0271	0.9679 ± 0.0005	10/0/0
64	dense	0.6499 ± 0.0530	0.9676 ± 0.0006	10/0/0
96	local	0.3736 ± 0.1210	0.9129 ± 0.0677	–
96	tied	0.3630 ± 0.0753	0.9256 ± 0.0355	6/4/0
96	dense	0.1583 ± 0.0518	0.6899 ± 0.1248	9/0/1
128	local	0.1047 ± 0.0516	0.5478 ± 0.2017	–
128	tied	0.1100 ± 0.0779	0.5286 ± 0.2134	5/3/2
128	dense	0.0114 ± 0.0112	0.0854 ± 0.0797	10/0/0

Traces	Model	Step error	Terminal error	Lower/higher/unclear
192	local	0.0031 ± 0.0026	0.0247 ± 0.0209	–
192	tied	0.0038 ± 0.0029	0.0298 ± 0.0229	0/0/10
192	dense	0.0001 ± 0.0003	0.0012 ± 0.0020	0/0/10

Paired one-step and final-output gains. Figure 1 uses all forty tied/local contrasts in the balanced follow-up: all ten seeds at all four budgets (64, 96, 128, 192). Its 64-trace panel repeats ten of those contrasts. Every bar retains its original 240-endpoint, 480-tail family interval with $\delta = 0.05$ and 32,768 independent traces per checkpoint.

Across those ten 64-trace seeds, one-step improvements range from 3.4931 to 7.7190 percentage points, with ten positive lower bounds. Final-output differences range from -0.1862 to $+0.1923$ points, with ten intervals crossing zero. Both models have near-chance final predictions in that low-data region, so the result illustrates a lack of statistically resolved endpoint improvement. Other displayed budgets show larger final changes and seed-dependent directions. No displayed comparison establishes opposite signs on both endpoints. This empirical observable uses reference prefixes locally and model-generated prefixes finally.

The one-step metric uses reference prefixes, whereas the final-state and complete-trace metrics use free-running predictions. The independent-step reference used for complete-trace error concerns failure anywhere along a trace; it is not a final-state error model. Agreement with that reference does not establish independence of actual prediction errors or exclude propagation and recovery.

Supervision at low data. The supervision-budget sweep includes ten budgets (64, 96, 128, 192, 256, 512, 1,024, 2,048, 4,096, 16,384), three objectives, and the three original training seeds, yielding ninety configuration summaries. The B models reuse tied checkpoints from the architecture sweep. Additional seed comparisons retain their separate evaluation families. At 64 traces, R retains raw route agreement 1 while forced fidelity is 0.3713 ± 0.0267. Labels and action effects therefore have different empirical behavior. The extra intervention targets also constitute extra supervision, so observed I improvements are not free at matched total training cost.

Table 20: Supervision-budget sweep over ten training-data budgets, three objectives, and the three original seeds. The baseline reuses tied checkpoints from the architecture sweep. Causal evaluation uses 16,384 contexts.

Traces	Obj.	Terminal error	Raw labels	Forced fidelity	Cross-fidelity
64	B	0.9685 ± 0.0001	0.2443 ± 0.0528	0.1724 ± 0.0108	0.1442 ± 0.0067
64	R	0.9642 ± 0.0006	1.0000 ± 0.0000	0.3713 ± 0.0267	0.3406 ± 0.0298
64	I	0.1257 ± 0.0454	1.0000 ± 0.0000	0.9943 ± 0.0014	0.9978 ± 0.0006
96	B	0.9174 ± 0.0565	0.2154 ± 0.0581	0.3097 ± 0.0293	0.2102 ± 0.0134
96	R	0.5103 ± 0.1574	1.0000 ± 0.0000	0.7963 ± 0.0260	0.7606 ± 0.0351
96	I	0.0215 ± 0.0098	1.0000 ± 0.0000	0.9991 ± 0.0002	0.9997 ± 0.0005
128	B	0.6342 ± 0.2061	0.2530 ± 0.0417	0.3531 ± 0.0671	0.2115 ± 0.0629
128	R	0.0636 ± 0.0842	1.0000 ± 0.0000	0.7909 ± 0.0713	0.7239 ± 0.0974
128	I	0.0024 ± 0.0014	1.0000 ± 0.0000	0.9999 ± 0.0001	1.0000 ± 0.0000
192	B	0.0387 ± 0.0334	0.2424 ± 0.0483	0.2918 ± 0.0204	0.1218 ± 0.0047
192	R	0.0056 ± 0.0039	1.0000 ± 0.0000	0.8779 ± 0.0307	0.8373 ± 0.0407
192	I	0.0008 ± 0.0012	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
256	B	0.0030 ± 0.0049	0.2576 ± 0.0180	0.2943 ± 0.0052	0.1393 ± 0.0172
256	R	0.0013 ± 0.0023	1.0000 ± 0.0000	0.8120 ± 0.0292	0.7494 ± 0.0388
256	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
512	B	0.0000 ± 0.0000	0.2312 ± 0.0172	0.2832 ± 0.0394	0.1312 ± 0.0330
512	R	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9441 ± 0.0119	0.9254 ± 0.0158
512	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
1024	B	0.0000 ± 0.0000	0.2436 ± 0.0435	0.2796 ± 0.0242	0.1325 ± 0.0267
1024	R	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9557 ± 0.0243	0.9409 ± 0.0324
1024	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
2048	B	0.0000 ± 0.0000	0.2625 ± 0.0498	0.2704 ± 0.0184	0.1261 ± 0.0289
2048	R	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9563 ± 0.0042	0.9418 ± 0.0056
2048	I	0.0001 ± 0.0002	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
4096	B	0.0000 ± 0.0000	0.2445 ± 0.0307	0.2615 ± 0.0294	0.1119 ± 0.0280

Traces	Obj.	Terminal error	Raw labels	Forced fidelity	Cross-fidelity
4096	R	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9671 ± 0.0090	0.9561 ± 0.0120
4096	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
16384	B	0.0000 ± 0.0000	0.2465 ± 0.0372	0.2539 ± 0.0281	0.1052 ± 0.0226
16384	R	0.0000 ± 0.0000	1.0000 ± 0.0000	0.9457 ± 0.0159	0.9276 ± 0.0212
16384	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000

G.8 Strict state–action holdout and capacity checks

The transition-holdout study uses separate B/R/I trainings with a declared set of sixteen of the 124 state–action pairs withheld. Ordinary training traces containing excluded transitions are rejected, and such pairs are excluded from intervention-training targets; input and output state symbols remain available elsewhere. Thus this is stronger than family holdout, but it is still one finite symbolic task and one selected holdout set.

The fixed withheld pairs, written as (pre-action state, operation ID) with IDs 0–3, are

$$(1,1),\ (3,0),\ (3,3),\ (6,1),\ (7,2),\ (8,3),\ (10,0),\ (10,3),$$
$$(11,3),\ (13,2),\ (18,2),\ (24,0),\ (25,0),\ (25,1),\ (28,3),\ (29,0).$$

The same set is used across this B/R/I holdout family, with at most two operations excluded at any state. The per-state withholding study in Section G.8.1 uses a different design.

Evaluation partitions forced cells by the pre-state and forced action. Evaluation includes 6,987 eligible withheld contexts containing 8,593 withheld action cells; seen statistics use 16,384 contexts and 56,943 seen cells. Context averaging is retained. Natural test traces are separately partitioned into 21,919 using at least one withheld transition and 10,849 clean traces.

The holdout-trained I models have seen fidelity 1 and withheld fidelity 0.0021 ± 0.0037 over three seeds. Natural terminal error on traces using a withheld transition is about 0.984, while clean-trace error is zero. Separately trained full-table I controls have fidelity 1 on both evaluation subsets. These are different checkpoints. The observations show that coverage-conditioned compound success did not generalize in this setting. They are consistent with memorization of seen transitions, but do not uniquely identify the model's internal implementation or negate measured in-distribution intervention effects.

Table 21: Unseen-transition evaluation: separate trainings excluding the 16 declared state–action pairs. Seen/withheld fidelity is context-averaged within the relevant action subset. Withheld SE is the reported across-context estimate; clean/uses columns partition natural test traces.

Obj.	Seed	Seen fid.	Withheld fid.	Withheld SE	Clean term.	Uses term.
B	0	0.26882	0.07528	0.00316	0.00000	0.98120
B	1	0.15763	0.05167	0.00252	0.00000	0.97226
B	2	0.31621	0.04687	0.00246	0.00000	0.98362
R	0	0.93353	0.02218	0.00176	0.00000	0.98298
R	1	0.96587	0.02290	0.00179	0.00000	0.98262
R	2	0.90390	0.00000	0.00000	0.00000	0.98435
I	0	1.00000	0.00000	0.00000	0.00000	0.98202
I	1	1.00000	0.00644	0.00096	0.00000	0.98540
I	2	1.00000	0.00000	0.00000	0.00000	0.98558

Table 22: Holdout-trained models versus separately trained full-table controls, evaluated on the same subset labels. Mean $\pm$ sample SD over seeds.

Obj.	Holdout: seen	Holdout: withheld	Full table: seen	Full table: subset
B	0.2476 ± 0.0814	0.0579 ± 0.0152	0.1879 ± 0.0129	0.2484 ± 0.0708
R	0.9344 ± 0.0310	0.0150 ± 0.0130	0.9451 ± 0.0168	0.9493 ± 0.0165
I	1.0000 ± 0.0000	0.0021 ± 0.0037	1.0000 ± 0.0000	1.0000 ± 0.0000

The capacity-and-holdout-size follow-up varies width, depth, training-pool size, training steps, and nested holdout size. Width 256 to 512 is a doubling, though parameter count grows by roughly four. The combined setting has eight width-512 layers, 65,536 training traces, and 12,000 steps. These tested modifications do not yield uniformly high held-out fidelity, but all-run upper bounds must retain exceptions: `wide-s0` is 0.02834, `long-s1` 0.01174, the four-pair holdout at seed 2 is 0.13128, and the eight-pair holdout at seed 2 is 0.01236. Seen fidelity is not exactly one in every run either. All 24 follow-up summaries are shown below.

Table 23: All 24 capacity and holdout-size run summaries, including favorable exceptions and non-perfect seen fidelity. The base 16-pair results remain part of the base transition-holdout study.

Setting	Seed	Hole	Params (M)	Seen fidelity	Withheld fidelity	Uses error
all	0	16	17.89	1.00000	0.00114	0.9875
all	1	16	17.89	1.00000	0.00000	0.9794
all	2	16	17.89	1.00000	0.00057	0.9916
base	0	32	3.43	0.99472	0.00249	0.9774
base	0	4	3.43	1.00000	0.00000	0.9990
base	0	8	3.43	1.00000	0.00000	0.9935
base	1	32	3.43	0.99750	0.00000	0.9755
base	1	4	3.43	1.00000	0.00000	0.9977
base	1	8	3.43	1.00000	0.00000	0.9925
base	2	32	3.43	0.99872	0.00045	0.9736
base	2	4	3.43	1.00000	0.13128	0.9987
base	2	8	3.43	1.00000	0.01236	0.9949
data	0	16	3.43	1.00000	0.00000	0.9814
data	1	16	3.43	1.00000	0.00744	0.9848
data	2	16	3.43	1.00000	0.00143	0.9867
deep	0	16	4.49	1.00000	0.00000	0.9832
deep	1	16	4.49	1.00000	0.00587	0.9888
deep	2	16	4.49	0.99998	0.00243	0.9870
long	0	16	3.43	1.00000	0.00000	0.9789
long	1	16	3.43	1.00000	0.01174	0.9810
long	2	16	3.43	1.00000	0.00000	0.9821
wide	0	16	13.68	1.00000	0.02834	0.9812
wide	1	16	13.68	1.00000	0.00157	0.9858
wide	2	16	13.68	1.00000	0.00215	0.9872

G.8.1 State–action generalization across state-space sizes

Question and protocol. The follow-up tests whether the seen/withheld fidelity gap is confined to the original 31-state setting. It applies the same four operations modulo $p = 31, 97, 251$, giving transition tables with 124, 388, and 1,004 entries. At every state, zero, one, or two of the four operations are withheld, corresponding to 0%, 25%, or 50% of the table. This per-state design differs from the preceding sixteen-pair holdout. The grid contains three training seeds for each of nine settings, giving 27 trained configurations.

The follow-up retains the four-layer, width-256 tied architecture and intervention-supervised objective, with modulus-specific state tokens and a p-way output head. Ordinary training traces containing a withheld transition are rejected, and counterfactual targets are restricted to allowed transitions. The no-holdout arm is separately trained at each modulus; it checks task learnability under full coverage.

Table 24 gives the training-pool sizes and update budgets. These budgets target approximately comparable nominal exposure; total compute is not exactly matched, and exposure is nominal rather than a measured post-rejection count for each transition. Parameter counts also change with the vocabulary and output head.

Evaluation and results. Forced-action evaluation uses 16,384 contexts per run and averages the applicable actions within a context before aggregation. At 25% and 50% withholding, every state has respectively one or two excluded actions, giving 16,384 and 32,768 withheld action cells, but still only 16,384 contexts. In the no-holdout control, withheld fidelity is undefined, not zero. Natural task

Table 24: State-space follow-up: actual reported budgets and evaluation subset sizes. Each row has three training seeds.

p	$4p$	Withheld	Pairs	Train traces	Updates	Parameters	N_{clean}	N_{use}
31	124	0%	0	20,666	3,000	3,433,984	16384	0
31	124	25%	31	15,500	3,000	3,433,984	1705	14679
31	124	50%	62	10,334	3,000	3,433,984	54	16330
97	388	0%	0	64,666	9,000	3,467,776	16384	0
97	388	25%	97	48,500	9,000	3,467,776	1618	14766
97	388	50%	194	32,334	9,000	3,467,776	72	16312
251	1004	0%	0	167,334	24,000	3,546,624	16384	0
251	1004	25%	251	125,500	24,000	3,546,624	1655	14729
251	1004	50%	502	83,666	24,000	3,546,624	70	16314

N_{clean} and N_{use} are counts per seed in a 16,384-trace task bank; they are constant across the three seeds of each setting. The latter subset uses at least one withheld transition.

evaluation separately partitions 16,384 traces into clean traces and those using at least one withheld transition; their counts must not be interchanged with the causal-context counts.

All nine full-coverage control checkpoints have observed fidelity one. Across the eighteen holdout-trained checkpoints, withheld fidelity is at most 0.00647, while seen fidelity is at least 0.97763. Table 25 reports every setting, and Table 26 retains every seed. Observed clean-subset terminal error is zero in all 27 evaluated configurations, but at 50% withholding the clean subsets contain only 54, 72, and 70 traces for the three moduli.

Table 25: State-space follow-up: forced-action fidelity and free-running terminal error (mean $\pm$ sample SD across three seeds). All values are proportions.

p	Withheld	Withheld fidelity	Seen fidelity	Error on traces using withheld pairs
31	0%	—	1.00000 ± 0.00000	—
31	25%	0.00224 ± 0.00367	1.00000 ± 0.00000	0.97693 ± 0.00346
31	50%	0.00212 ± 0.00229	0.98133 ± 0.00556	0.96928 ± 0.00138
97	0%	—	1.00000 ± 0.00000	—
97	25%	0.00006 ± 0.00011	1.00000 ± 0.00000	0.99199 ± 0.00076
97	50%	0.00119 ± 0.00104	0.99887 ± 0.00028	0.99023 ± 0.00085
251	0%	—	1.00000 ± 0.00000	—
251	25%	0.00014 ± 0.00025	1.00000 ± 0.00000	0.99647 ± 0.00072
251	50%	0.00078 ± 0.00104	0.99964 ± 0.00051	0.99577 ± 0.00022

All nine full-coverage control checkpoints have observed fidelity one. Clean-subset terminal error is zero in all 27 runs, but the subset counts in Table 24 must be retained. Dashes denote an empty subset. These are descriptive three-seed summaries.

Coverage-dependent generalization. The enlarged state spaces preserve a substantial coverage gap in this finite task family under the reported training recipe. Uniform-state guessing at $1/p$ serves as a scale reference. The displayed means and seed standard deviations are descriptive summaries over three training seeds; the primary eighteen-pair audit retains its own confidence budget.

Table 26: All 27 state-space follow-up run summaries. Fidelity and error columns are proportions.

p	Withheld	Seed	Withheld fid.	Seen fid.	N_{clean}	Clean error	N_{use}	Error on use
31	0%	0	—	1.000000	16384	0.000000	0	—
31	0%	1	—	1.000000	16384	0.000000	0	—
31	0%	2	—	1.000000	16384	0.000000	0	—
31	25%	0	0.000000	1.000000	1705	0.000000	14679	0.980857
31	25%	1	0.006470	1.000000	1705	0.000000	14679	0.974317
31	25%	2	0.000244	1.000000	1705	0.000000	14679	0.975611
31	50%	0	0.001434	0.978638	54	0.000000	16330	0.968279
31	50%	1	0.004669	0.977631	54	0.000000	16330	0.970851
31	50%	2	0.000244	0.987732	54	0.000000	16330	0.968708
97	0%	0	—	1.000000	16384	0.000000	0	—
97	0%	1	—	1.000000	16384	0.000000	0	—
97	0%	2	—	1.000000	16384	0.000000	0	—
97	25%	0	0.000000	1.000000	1618	0.000000	14766	0.991128
97	25%	1	0.000183	1.000000	1618	0.000000	14766	0.992280
97	25%	2	0.000000	1.000000	1618	0.000000	14766	0.992550
97	50%	0	0.000488	0.999176	72	0.000000	16312	0.991172
97	50%	1	0.002380	0.998627	72	0.000000	16312	0.989517
97	50%	2	0.000702	0.998810	72	0.000000	16312	0.990007
251	0%	0	—	1.000000	16384	0.000000	0	—
251	0%	1	—	1.000000	16384	0.000000	0	—
251	0%	2	—	1.000000	16384	0.000000	0	—
251	25%	0	0.000000	1.000000	1655	0.000000	14729	0.997148
251	25%	1	0.000427	1.000000	1655	0.000000	14729	0.996537
251	25%	2	0.000000	1.000000	1655	0.000000	14729	0.995723
251	50%	0	0.000275	0.999054	70	0.000000	16314	0.995954
251	50%	1	0.001984	1.000000	70	0.000000	16314	0.995525
251	50%	2	0.000092	0.999878	70	0.000000	16314	0.995832

Withheld actions are averaged within each eligible context before aggregation; at 25% and 50% withholding there are 16,384 contexts, with one and two withheld actions per context, respectively. Empty subsets are shown as dashes.

G.9 STATE COMPRESSION AND GENERAL-DESIGN SUPPLEMENTS

State-conditioned intervention variability. The study groups intervention outputs by the declared pre-action state and action, using a one-hot outcome representation across 124 cells. The table reports the centered outcome-variance statistic, cell-weighted fidelity, and fraction of empirically deterministic cells for 63 checkpoints. These are different descriptive quantities. In particular, the corresponding classification-loss analysis uses a conditional mode; it is not the squared-loss conditional-mean identity of Proposition 15.

The centered and unbiased-pair variance summaries differ by up to 0.00159009, so they are not treated as interchangeable finite-sample estimators. We report the centered statistic descriptively for each fixed checkpoint and state extractor, measuring conditional outcome variability.

Table 27: State-conditioned intervention variability for 63 fixed checkpoints and the specified state extractor. The centered one-hot outcome statistic and cell-weighted fidelity measure different losses.

Traces	Obj.	Centered variance	Cell-weighted fidelity	Deterministic fraction
64	B	0.8221 ± 0.0183	0.1724 ± 0.0108	0.0000 ± 0.0000
64	R	0.7370 ± 0.0205	0.3713 ± 0.0267	0.0000 ± 0.0000
64	I	0.0107 ± 0.0024	0.9943 ± 0.0014	0.7366 ± 0.0123
96	B	0.6418 ± 0.0159	0.3097 ± 0.0293	0.0000 ± 0.0000
96	R	0.2941 ± 0.0388	0.7963 ± 0.0260	0.0430 ± 0.0123
96	I	0.0017 ± 0.0003	0.9991 ± 0.0002	0.9194 ± 0.0689
128	B	0.6299 ± 0.0487	0.3531 ± 0.0671	0.0027 ± 0.0047
128	R	0.3465 ± 0.1069	0.7909 ± 0.0713	0.0027 ± 0.0047
128	I	0.0002 ± 0.0002	0.9999 ± 0.0001	0.9812 ± 0.0123
192	B	0.6719 ± 0.0184	0.2918 ± 0.0204	0.0108 ± 0.0186
192	R	0.2158 ± 0.0498	0.8779 ± 0.0307	0.0269 ± 0.0259
192	I	0.0001 ± 0.0001	1.0000 ± 0.0000	0.9866 ± 0.0123
256	B	0.6706 ± 0.0182	0.2943 ± 0.0052	0.0081 ± 0.0081
256	R	0.3215 ± 0.0436	0.8120 ± 0.0292	0.0054 ± 0.0093
256	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
512	B	0.6688 ± 0.0153	0.2832 ± 0.0394	0.0054 ± 0.0047
512	R	0.1046 ± 0.0222	0.9441 ± 0.0119	0.1317 ± 0.1266
512	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
16384	B	0.6834 ± 0.0145	0.2537 ± 0.0075	0.0027 ± 0.0047
16384	R	0.1013 ± 0.0289	0.9457 ± 0.0159	0.2016 ± 0.1290
16384	I	0.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000

Pooling under general feature geometry. This supplement uses 512 noise repetitions on the frozen nonorthogonal designs. The 54 partition cells and 54 merge cells are interpreted using the existing covariance-aware fixed-estimator framework in Theorems 38 and 39. The empirical risks satisfy the numerical comparison criterion, but the roles have unequal estimation variances. Applying a single global equal-variance price outside the balanced assumptions yields four merge-sign mismatches.

Table 28: General-design merging: four sign discrepancies obtained by applying a global equal-variance proxy outside its assumptions. The covariance-aware reference and observed mean agree in sign. Groups i/j denotes the merge $G = \{i\}$, $H = \{j\}$ using zero-based role indices $0, \ldots, 7$; $0/4$ and $2/6$ therefore identify two singleton-group pairs, not group sizes.

c	n	Groups	Global proxy	Covariance ref.	Observed	MCSE
2.0	64	0/4	-0.03990	0.00545	0.00543	0.00007
2.0	64	2/6	-0.03990	0.00385	0.00377	0.00008
0.5	256	0/4	-0.00998	0.00136	0.00135	0.00002
0.5	256	2/6	-0.00997	0.00096	0.00094	0.00002

The nine plug-in entries show lower observed risk than the separate-risk reference in this particular simulation. Contrast dimension $k \geq 4$ is not sufficient for the classical guarantee in Section D.6: the

spherical Gaussian contrast-noise and matching loss assumptions also matter and are not automatically established by the anisotropic design.

Table 29: General-design same-data shrinkage: Monte Carlo risk versus an analytic separate-risk reference, not a paired significance test.

c	n	Separate ref.	Plug-in mean	MCSE	Savings (%)
0.0	64	0.5268	0.0738	0.00110	86.0
0.5	64	0.5268	0.1268	0.00101	75.9
2.0	64	0.5268	0.2292	0.00115	56.5
0.0	256	0.1317	0.0180	0.00026	86.4
0.5	256	0.1317	0.0569	0.00027	56.8
2.0	256	0.1317	0.0939	0.00054	28.7
0.0	1024	0.0329	0.0046	0.00008	86.0
0.5	1024	0.0329	0.0237	0.00014	28.1
2.0	1024	0.0329	0.0300	0.00022	8.9

H General Linear Sharing Subspaces and Cross-Depth Architectures

Overview. We extend the local linear comparison to unequal role weights, sample counts, and noise levels, and to fixed cross-depth sharing patterns. The formulas separate the approximation cost of tying from its estimation cost, explaining which role, depth, and interaction effects a declared linear architecture retains. Fixed global-pool assignments and depth-conditioned operator subspaces are instances of this projection comparison. Section H.1 treats unequal roles and sampling, Section H.2 gives the projection formulation, Section H.3 instantiates cross-depth and multi-axis patterns, and Section H.4 collects the proofs.

H.1 Unequal roles and heteroscedastic noise

For role $a \in [M]$, let $A_a \in \mathbb{R}^{p\times d}$, assume $n_a \geq d$, and let the fixed design satisfy

$$X_a^\top X_a = n_a I_d. \tag{100}$$

The response matrix is $Y_a = X_a A_a^\top + \mathcal{E}_a$, where noise rows are mutually independent across observations and roles, centered, and have covariance $\sigma_a^2 I_p$. Different σ_a^2 are called *heteroscedastic* noise: some roles have noisier targets than others. Gaussianity is not required for any expectation identity in this section. Let $\pi_a > 0$ with $\sum_a \pi_a = 1$, put $D = pd$ and $v_a = \sigma_a^2/n_a$, and define

$$\bar{A}_\pi = \sum_a \pi_a A_a, \qquad B_\pi = \sum_a \pi_a \left\|A_a - \bar{A}_\pi\right\|_F^2, \tag{101}$$

$$V_{\text{sep}} = D\sum_a \pi_a v_a, \qquad V_{\text{pool}} = D\sum_a \pi_a^2 v_a, \qquad C_\pi = V_{\text{sep}} - V_{\text{pool}}. \tag{102}$$

The weighted pooled empirical objective is $\sum_a (\pi_a/n_a)\left\|Y_a - X_a A^\top\right\|_F^2$, whose minimizer is

$$\widehat{A}_\pi = \sum_a \pi_a \widehat{A}_a, \qquad \widehat{A}_a = Y_a^\top X_a/n_a. \tag{103}$$

For a common shrinkage coefficient $\lambda \in [0,1]$, define $\widehat{A}_a^{(\lambda)} = (1-\lambda)\widehat{A}_a + \lambda\widehat{A}_\pi$. Its role-weighted excess prediction risk is

$$\mathcal{R}_\lambda = \sum_a \pi_a \mathbb{E}_{\mathcal{E},x}\left\|(\widehat{A}_a^{(\lambda)} - A_a)x\right\|_2^2, \qquad \mathbb{E}[xx^\top] = I_d. \tag{104}$$

Theorem 16 (Exact weighted and heteroscedastic reuse frontier)**.** *For the role-weighted excess prediction risk in Equation (104),*

$$\mathcal{R}_\lambda = V_{\text{sep}} - 2C_\pi\lambda + (B_\pi + C_\pi)\lambda^2. \tag{105}$$

In particular,

$$\mathcal{R}_{\text{sep}} = V_{\text{sep}}, \qquad \mathcal{R}_{\text{share}} = B_\pi + V_{\text{pool}}. \tag{106}$$

Full reuse is better than separate estimation if and only if

$$B_\pi < C_\pi = D \sum_a \frac{\pi_a(1-\pi_a)\sigma_a^2}{n_a}. \tag{107}$$

When $C_\pi > 0$, the optimal common shrinkage strength is

$$\lambda_\pi^\star = \frac{C_\pi}{B_\pi + C_\pi}, \qquad \mathcal{R}_{\lambda_\pi^\star} = V_{\text{sep}} - \frac{C_\pi^2}{B_\pi + C_\pi}. \tag{108}$$

If $C_\pi = 0 < B_\pi$, the optimal coefficient is zero; if $C_\pi = B_\pi = 0$, every coefficient is optimal. The quadratic form is retained, with the sharing bias B_π, separate-estimation risk V_{sep}, and removable variance C_π given by their weighted, heteroscedastic counterparts. A role with large evaluation mass but very few or very noisy samples contributes strongly to both the separate risk and the potential value of pooling. Conversely, raw role multiplicity can be large while C_π is small when almost all mass lies on one role.

The theorem also distinguishes two sample-allocation problems that are sometimes conflated.

Corollary 17 (Architecture-dependent optimal sampling)**.** *Fix the idealized continuous variance-allocation problem $\sum_a n_a = N$ and assume $\sigma_a > 0$. The formulas below apply when the displayed unconstrained allocations satisfy the feasibility requirements of the orthogonal designs, including any lower bounds such as $n_a \geq d$; otherwise the corresponding constrained KKT solution is used. The allocation minimizing the pooled estimation variance is*

$$n_a^{\text{pool}} = N\frac{\pi_a\sigma_a}{\sum_b \pi_b\sigma_b}, \qquad \min V_{\text{pool}} = \frac{D}{N}\left(\sum_a \pi_a\sigma_a\right)^2, \tag{109}$$

whereas the allocation minimizing separate-estimation risk is

$$n_a^{\text{sep}} = N\frac{\sqrt{\pi_a}\sigma_a}{\sum_b \sqrt{\pi_b}\sigma_b}, \qquad \min V_{\text{sep}} = \frac{D}{N}\left(\sum_a \sqrt{\pi_a}\sigma_a\right)^2. \tag{110}$$

If each architecture may choose its own allocation, full sharing has lower optimized risk exactly when

$$B_\pi < \frac{D}{N}\left[\left(\sum_a \sqrt{\pi_a}\sigma_a\right)^2 - \left(\sum_a \pi_a\sigma_a\right)^2\right]. \tag{111}$$

For completeness, minimizing $\sum_a c_a/n_a$ with $c_a > 0$, $n_a \geq d$, and $\sum_a n_a = N > Md$ gives $n_a = \max\{d, \sqrt{c_a/t}\}$, where $t > 0$ is chosen to meet the budget. The Karush–Kuhn–Tucker (KKT) conditions are the stationarity and complementary-slackness conditions for this constrained convex problem: a coordinate above its floor has derivative $-c_a/n_a^2 + t = 0$; a coordinate at the floor need not satisfy that equality. If $N = Md$, every $n_a = d$. Integer allocations require a separate discrete solution.

This result concerns data acquisition or replay allocation.

H.2 Any linear sharing architecture is a projection problem

A large class of tying and conditioning rules can be described by a linear subspace of admissible role-wise functions. The following theorem gives an exact bias–dimension decomposition.

Theorem 18 (Exact risk of a linear sharing subspace)**.** *Let J role cells have true parameter vectors $\theta_1, \ldots, \theta_J \in \mathbb{R}^D$. Suppose their unconstrained stacked estimator satisfies*

$$\widehat{\theta} = \theta + \zeta, \qquad \mathbb{E}\zeta = 0, \qquad \text{Cov}(\zeta) = vI_{JD}. \tag{112}$$

Here stacking means concatenating the J length-D vectors in $\mathbb{R}^{JD}$, and Cov *denotes covariance. The subspace must be chosen before observing this fitting noise. Let $\mathcal{U} \subseteq \mathbb{R}^{JD}$ be a fixed linear sharing*

subspace, let $P_{\mathcal{U}}$ be its Euclidean orthogonal projector, and fit $\widehat{\theta}_{\mathcal{U}} = P_{\mathcal{U}}\widehat{\theta}$. Define the role-averaged squared parameter error by $\mathcal{R}(\mathcal{U}) = J^{-1}\mathbb{E}\left\|\widehat{\theta}_{\mathcal{U}} - \theta\right\|_2^2$. Then

$$\mathcal{R}(\mathcal{U}) = \frac{1}{J}\left\|(I - P_{\mathcal{U}})\theta\right\|_2^2 + \frac{v}{J}\dim(\mathcal{U}). \tag{113}$$

The first term is exactly the squared approximation cost imposed by the fixed subspace; the second charges one isotropic-noise degree of freedom for every fitted scalar parameter. The partition result in Theorem 2 is the special case in which $\mathcal{U}$ contains arrays constant inside every expert group. More generally, tied layers, shared cores, additive depth embeddings, and fixed low-dimensional modulation bases correspond to different choices of $\mathcal{U}$.

H.3 An exact cross-depth tying frontier

We now instantiate Theorem 18 for architectures that reuse experts across depth. Index a semantic role by $a \in [M]$ and depth by $\ell \in [L]$. The true local operator is $A_{a\ell} \in \mathbb{R}^{p\times d}$, each cell has an orthogonal design with $n \geq d$ samples and homoscedastic noise σ^2, with noise independent across cells, and

$$V = \frac{pd\sigma^2}{n} \tag{114}$$

would be the expected risk of one cell-specific OLS fit. Define

$$\bar{A} = \frac{1}{ML}\sum_{a,\ell} A_{a\ell}, \qquad \bar{A}_{a\cdot} = \frac{1}{L}\sum_{\ell} A_{a\ell}, \qquad \bar{A}_{\cdot\ell} = \frac{1}{M}\sum_{a} A_{a\ell}, \tag{115}$$

$$B_{\mathrm{role}} = \frac{1}{M}\sum_{a}\left\|\bar{A}_{a\cdot} - \bar{A}\right\|_F^2, \qquad B_{\mathrm{depth}} = \frac{1}{L}\sum_{\ell}\left\|\bar{A}_{\cdot\ell} - \bar{A}\right\|_F^2, \tag{116}$$

$$R^{\mathrm{int}}_{a\ell} = A_{a\ell} - \bar{A}_{a\cdot} - \bar{A}_{\cdot\ell} + \bar{A},$$

$$B_{\mathrm{int}} = \frac{1}{ML}\sum_{a,\ell}\left\|R^{\mathrm{int}}_{a\ell}\right\|_F^2. \tag{117}$$

The dot in a subscript means averaging over that index. The role effect changes a only; the depth effect changes ℓ only. The interaction $R^{\mathrm{int}}_{a\ell}$ is what remains when the two effects cannot explain a cell jointly. Under the balanced Frobenius inner product, these components are orthogonal (their pairwise inner products sum to zero). We use this as an algebraic two-way analysis of variance (ANOVA) decomposition.

Theorem 19 (Exact risks for cross-layer reuse patterns)**.** *Under the balanced cell model above, the exact role–cell-averaged excess prediction risks are those in Table 30.*

Table 30: Exact cross-depth sharing risks. "Role-tied" means that the same semantic expert is reused at every depth; "additive depth-conditioned" allows a semantic expert main effect plus a depth main effect but no role–depth interaction.

Architecture subspace	Approximation term	Dimension	Exact risk
Cell-specific layer-local experts	0	MLD	V
One globally tied expert	$B_{\mathrm{role}} + B_{\mathrm{depth}} + B_{\mathrm{int}}$	D	$B_{\mathrm{role}} + B_{\mathrm{depth}} + B_{\mathrm{int}} + V/(ML)$
Role-tied across depth	$B_{\mathrm{depth}} + B_{\mathrm{int}}$	MD	$B_{\mathrm{depth}} + B_{\mathrm{int}} + V/L$
One role-shared expert per depth	$B_{\mathrm{role}} + B_{\mathrm{int}}$	LD	$B_{\mathrm{role}} + B_{\mathrm{int}} + V/M$
Additive depth-conditioned reuse	B_{int}	$(M+L-1)D$	$B_{\mathrm{int}} + V(M + L - 1)/(ML)$

Consequently:

$$\textit{role-tied beats cell-specific} \iff B_{\text{depth}} + B_{\text{int}} < V\left(1 - \frac{1}{L}\right), \tag{118}$$

$$\textit{additive beats role-tied} \iff B_{\text{depth}} > V\frac{L-1}{ML}, \tag{119}$$

$$\textit{additive beats cell-specific} \iff B_{\text{int}} < V\frac{(M-1)(L-1)}{ML}. \tag{120}$$

These formulas separate three reasons that cross-layer reuse may fail. A globally tied pool pays semantic-role variation, depth variation, and their interaction. Tying the same semantic expert across depth removes the role main effect but still pays depth shift and role–depth interaction. A depth-conditioned shared expert removes both main effects; only the interaction remains. Within this additive operator model, depth conditioning is useful precisely when the removed depth effect exceeds the variance cost of its additional $(L-1)D$ scalar degrees of freedom. Interpreting a nonlinear depth embedding through this subspace requires an explicit realization or local-linearization argument. If B_{int} is large, neither global tying nor additive depth conditioning is sufficient, and layer-specific residual capacity is statistically justified.

The same calculation extends to several architectural axes.

Corollary 20 (Multi-axis additive reuse)**.** *Suppose role cells are indexed by r factors of sizes $m_1, \dots, m_r$, with $J = \prod_{k=1}^{r} m_k$ balanced cells. Let $\mathcal{U}_{\text{add}}$ contain a grand mean plus one zero-sum main effect for every factor, and define the normalized interaction residual*

$$B_{\geq 2} = \frac{1}{J}\,\|(I - P_{\mathcal{U}_{\text{add}}})\theta\|_2^2\,. \tag{121}$$

It contains all interactions of order at least two. Then

$$\mathcal{R}(\mathcal{U}_{\text{add}}) = B_{\geq 2} + \frac{V}{J}\left[1 + \sum_{k=1}^{r}(m_k - 1)\right]. \tag{122}$$

This covers, for example, semantic role × layer × modality or role × reasoning position × tool state. For the declared linear subspaces, adding a factor-specific correction is worthwhile only when the removed component exceeds its estimation price. Corollary 40 gives the covariance-weighted version without balanced isotropic noise.

H.4 Proofs

Proof of Theorem 16. Orthogonality gives $\widehat{A}_a = A_a + E_a$, with centered, uncorrelated entries of variance v_a and independent error matrices across roles. Let $\bar{E}_\pi = \sum_b \pi_b E_b$. Then

$$\widehat{A}_a^{(\lambda)} - A_a = \lambda(\bar{A}_\pi - A_a) + (1-\lambda)E_a + \lambda\bar{E}_\pi. \tag{123}$$

The role-averaged squared bias is $\lambda^2 B_\pi$. For each scalar coefficient, the weighted variance is

$$\begin{aligned}&\sum_a \pi_a \operatorname{Var}\big((1-\lambda)E_a + \lambda\bar{E}_\pi\big)\\ &\quad= (1-\lambda)^2\sum_a \pi_a v_a + (2\lambda - \lambda^2)\sum_a \pi_a^2 v_a.\end{aligned} \tag{124}$$

Multiplying by D and adding the bias gives Equation (105). Setting $\lambda = 0, 1$, subtracting the endpoint risks, and minimizing the convex quadratic prove the remaining statements. □

Proof of Corollary 17. Minimizing $\sum_a c_a/n_a$ subject to $\sum_a n_a = N$ gives $n_a \propto \sqrt{c_a}$ by Lagrange multipliers, or equivalently by Cauchy–Schwarz. For pooled estimation, $c_a = \pi_a^2\sigma_a^2$, yielding $n_a \propto \pi_a\sigma_a$ and the first minimum. For separate estimation, $c_a = \pi_a\sigma_a^2$, yielding $n_a \propto \sqrt{\pi_a}\sigma_a$ and the second. Comparing the two optimized endpoint risks gives Equation (111). □

Proof of Theorem 18. Write $P = P_{\mathcal{U}}$. Because $P\theta - \theta = -(I - P)\theta$ is orthogonal to $P\zeta$ and $\mathbb{E}\zeta = 0$,

$$\mathbb{E}\left\|P\widehat{\theta} - \theta\right\|_2^2 = \|(I - P)\theta\|_2^2 + \mathbb{E}\,\|P\zeta\|_2^2\,. \tag{125}$$

Moreover,

$$\mathbb{E}\,\|P\zeta\|_2^2 = \operatorname{tr}(P\, v I\, P^\top) = v\operatorname{tr}(P) = v\dim(\mathcal{U}). \tag{126}$$

Divide by J. □

Proof of Theorem 19. Stack the ML operator vectors and apply Theorem 18 with scalar noise variance $v = \sigma^2/n$ and block dimension $D = pd$. The cell-specific, globally tied, role-tied, depth-specific, and additive subspaces have dimensions MLD, D, MD, LD, and $(M + L - 1)D$, respectively. Orthogonal two-way ANOVA gives

$$\frac{1}{ML}\sum_{a,\ell}\left\|A_{a\ell} - \bar{A}\right\|_F^2 = B_{\text{role}} + B_{\text{depth}} + B_{\text{int}}, \tag{127}$$

while projection onto role means leaves $B_{\text{depth}} + B_{\text{int}}$, projection onto depth means leaves $B_{\text{role}} + B_{\text{int}}$, and projection onto the additive subspace leaves B_{int}. Substituting each residual and dimension into Equation (113) gives the table. Pairwise subtraction gives Equations (118) to (120). □

Proof of Corollary 20. The additive subspace has one grand-mean block and $(m_k - 1)$ independent main-effect blocks for axis k, hence dimension $D[1 + \sum_k (m_k - 1)]$. Its orthogonal residual consists of interactions of order at least two. Apply Theorem 18. □

I Architecture-Specific Functional Routing Bounds

Overview. A routing change matters through the event output it changes, not just expert-index disagreement. We bound the generic cost ρ_j from Equation (85) for additive soft and uniform top-k gates, capacity-limited row-separable assignments, additive hierarchies, and preselected chains. Full soft dispatch also changes expert inputs and requires the additional term in Theorem 23. Each bound applies only to its stated interface. The transport notation follows Gabriel & Marco (2019); the inequalities are proved directly. Sections I.1–I.5 treat the individual routing interfaces; Section I.6 summarizes their cost terms, and Section I.7 supplies the remaining proofs.

I.1 Additive soft and top-k gates as transport over expert functions

Fix a common expert input u and let $v_e(u)$ be the routed residual produced by expert e in a normed vector space. For normalized gates $g, h \in \Delta_K := \{w \in \mathbb{R}^K : w_e \geq 0,\ \sum_e w_e = 1\}$, define the discrete coupling set

$$\Pi(g, h) = \left\{\gamma \in \mathbb{R}_+^{K\times K} : \sum_{e'}\gamma_{ee'} = g_e,\ \sum_{e}\gamma_{ee'} = h_{e'}\right\}, \tag{128}$$

and the input-dependent functional transport cost

$$\mathsf{W}_u(g, h) = \min_{\gamma\in\Pi(g,h)}\sum_{e,e'}\gamma_{ee'}c_u(e, e'), \qquad c_u(e, e') = \|v_e(u) - v_{e'}(u)\|\,. \tag{129}$$

Theorem 21 (Functional transport bound for additive routing)**.** *Let*

$$M(u, g) = B(u) + \sum_{e=1}^{K} g_e v_e(u), \tag{130}$$

where $B(u)$ contains every dense or always-active shared path and where changing the gate does not change u or the collection $\{v_e(u)\}$. Then

$$\|M(u, g) - M(u, h)\| \leq \mathsf{W}_u(g, h) \leq \frac{D_u}{2}\,\|g - h\|_1\,, \qquad D_u = \max_{e,e'} c_u(e, e'). \tag{131}$$

The first inequality can be strictly sharper than the diameter–total-variation bound when most mismatched mass moves between functionally similar experts.

The shared term $B(u)$ cancels exactly. Thus an always-active expert can improve the event while contributing nothing to the discrepancy between two conditional gates at the same input. Conversely, a dominant shared path can make natural task accuracy insensitive to the routed residual, which is one mechanism behind the causal non-identifiability result in Proposition 5.

Named architecture instances. Mixtral combines two selected SwiGLU experts using a softmax over the selected logits (Jiang et al., 2024); this is weighted top-2, not the equal-$1/2$ special case below. DeepSeekMoE combines always-active shared experts with weighted routed experts (Dai et al., 2024), corresponding to $B(u)$ plus the conditional sum in Equation (130). The actual gate normalization and selected mass must be retained; sub-probability gates use the null-atom augmentation below. Shared paths cancel only between two gates of the same fixed model at the same input, not between separately trained models.

For uniform top-k routing, the transport cost becomes a matching problem.

Corollary 22 (Uniform top-k functional matching)**.** *Let $S = \{s_1, \dots, s_k\}$ and $T = \{t_1, \dots, t_k\}$ be two k-element expert sets, and let both gates assign mass $1/k$ to every selected expert. Then*

$$\mathsf{W}_u(g_S, g_T) = \frac{1}{k} \min_{\pi \in \mathfrak{S}_k} \sum_{r=1}^{k} \left\| v_{s_r}(u) - v_{t_{\pi(r)}}(u) \right\|. \tag{132}$$

Consequently, two disjoint top-k supports can have small functional route cost, while replacing one selected expert can be severe if the substitute is functionally distant.

Here $\mathfrak{S}_k$ is the set of the $k!$ permutations of $\{1, \dots, k\}$, so the minimum matches each selected expert to one expert in the other set. The formula is invariant to enumeration. For nonuniform within-support weights, use the general transport problem in Equation (129); support overlap alone does not determine the route cost.

I.2 Capacity drops and row-separable expert-choice assignment

Token-choice and expert-choice routers may drop, pad, or reroute mass under capacity constraints (Zhou et al., 2022; Lewis et al., 2021). A zero-residual "drop" atom makes the associated functional cost explicit. For a sub-probability gate $g \in \mathbb{R}_+^K$ with $\sum_e g_e \le 1$, introduce atom 0 with $v_0(u) = 0$ and augmented mass

$$\widetilde{g}_0 = 1 - \sum_{e=1}^{K} g_e, \qquad \widetilde{g}_e = g_e \quad (e \ge 1). \tag{133}$$

Applying Theorem 21 to $\widetilde{g}$ and $\widetilde{h}$ charges

$$c_u(e, 0) = \|v_e(u)\| \tag{134}$$

for dropping the contribution of expert e. Hence a drop-rate statistic is insufficient unless it is weighted by the magnitude of the dropped residual.

For T token or slot states u_t, suppose an assignment matrix acts row-separably after the u_t are fixed, so row t produces $M(u_t, g_t)$ as in Equation (130). If matrices G, H induce row gates g_t, h_t, then under the sum norm on the batch output,

$$\sum_{t=1}^{T} \|M(u_t, g_t) - M(u_t, h_t)\| \le \sum_{t=1}^{T} \mathsf{W}_{u_t}(\widetilde{g}_t, \widetilde{h}_t). \tag{135}$$

The same statement divided by T holds for the average norm, while a maximum norm is controlled by the maximum row cost. Joint capacity constraints may couple how the rows are selected, but no row-independence assumption is used once the two assignment matrices are fixed. Dispatch mechanisms that change expert-slot inputs use the generic fixed-plan discrepancy; Theorem 23 gives a bound when dispatch sensitivity is controlled.

I.3 Full soft dispatch: expert inputs may also change

Full Soft MoE constructs weighted token combinations as expert inputs before combining their outputs (Puigcerver et al., 2024). The common-input gate bound alone is therefore insufficient. Let

$x_1, \dots, x_T \in \mathbb{R}^d$ be fixed token states and let there be Q *slots*, meaning Q intermediate vectors sent to designated expert calls. A dispatch matrix $D \in \mathbb{R}_+^{T\times Q}$ has column sums one; a combine matrix $C \in \mathbb{R}_+^{T\times Q}$ has row sums one. With fixed slot-to-expert assignment $e(q)$, define

$$u_q(D) = \sum_t D_{tq} x_t, \quad v_q(D) = E_{e(q)}(u_q(D)), \quad y_t(D, C) = \sum_q C_{tq} v_q(D). \tag{136}$$

Let (D', C') be another pair with the same normalizations and the same expert functions. Expert $E_{e(q)}$ is assumed L_q-Lipschitz on the convex hull of the token states, using the respective input and output norms. The convex hull consists of all their nonnegative weighted averages with total weight one.

Let $\mathsf{W}_D^{\mathrm{out}}$ be discrete transport over slot outputs $\{v_q(D)\}$, with ground cost $\|v_q(D) - v_{q'}(D)\|$. Let $\mathsf{W}_X^{\mathrm{in}}$ be transport over the token states, with cost $\|x_t - x_{t'}\|$. Define $\bar{C}'_q = T^{-1}\sum_t C'_{tq}$, the average output-token weight assigned to slot q.

Theorem 23 (Dispatch-plus-combine discrepancy for full soft routing)**.** *Under the normalization and Lipschitz assumptions above,*

$$\frac{1}{T}\sum_t \|y_t(D, C) - y_t(D', C')\| \le \frac{1}{T}\sum_t \mathsf{W}_D^{\mathrm{out}}(C_{t,:}, C'_{t,:}) + \sum_q \bar{C}'_q L_q \mathsf{W}_X^{\mathrm{in}}(D_{:,q}, D'_{:,q}). \tag{137}$$

The reverse insertion gives another valid bound using output geometry at D' and average combine weights from C; their minimum is valid. If $D_X = \max_{t,t'} \|x_t - x_{t'}\|$, each input transport term is at most $D_X \|D_{:,q} - D'_{:,q}\|_1/2$.

Proof. Insert $y_t(D, C')$. The first difference keeps slot outputs fixed and is bounded by output transport. The second difference is at most $\sum_q C'_{tq}\|E_{e(q)}(u_q(D)) - E_{e(q)}(u_q(D'))\|$. Use the expert's Lipschitz bound and the barycenter transport bound $\|u_q(D) - u_q(D')\| \le \mathsf{W}_X^{\mathrm{in}}(D_{:,q}, D'_{:,q})$. Average over t. Reversing the insertion proves the alternative; the diameter inequality follows from Theorem 21 applied to token vectors. □

This extends the routing analysis beyond the common-expert-input assumption. If dispatch is unchanged, the second term vanishes. If the slot-to-expert assignment or the token states also change, an additional expert-substitution or input-state discrepancy must be included. The theorem bounds the route component of a fixed trained event, not the learned expert approximation error already charged elsewhere.

I.4 Hierarchical additive routing separates coarse and fine errors

Consider a two-level additive hierarchy with groups $r \in [G]$, within-group experts $e \in [K_r]$, outer gate $p \in \Delta_G$, and conditional inner gates $q_r \in \Delta_{K_r}$. All vectors below are evaluated at the same fixed input. Define

$$m_r(q_r) = \sum_e q_{r,e} v_{r,e}, \qquad M(p, q) = \sum_r p_r m_r(q_r). \tag{138}$$

Proposition 24 (Hierarchical functional route decomposition)**.** *For two hierarchical routes (p, q) and (p', q'),*

$$\|M(p, q) - M(p', q')\| \le \mathsf{W}_q^{\mathrm{out}}(p, p') + \sum_{r=1}^{G} p'_r \mathsf{W}_r^{\mathrm{in}}(q_r, q'_r), \tag{139}$$

where the outer transport uses

$$c_q^{\mathrm{out}}(r, r') = \|m_r(q_r) - m_{r'}(q_{r'})\|, \tag{140}$$

and the inner transport in group r uses

$$c_r^{\mathrm{in}}(e, e') = \|v_{r,e} - v_{r,e'}\|. \tag{141}$$

The reverse insertion gives a second valid bound with (p, q) and (p', q') interchanged; their minimum may be used.

The first term measures a coarse group mistake, and the second measures fine selection errors after the outer mass is fixed. A single flat route-accuracy number cannot reveal which level consumes the reuse gain.

I.5 Preselected sequential expert chains have within-event amplification

Sequential architectures call experts one after another inside one routed event (Wang et al., 2025). The theorem here concerns two preselected, equal-length sequences. If a later expert choice adapts to an earlier expert output, each adaptive choice must instead be made a separate routed event or the entire adaptive policy must be compared through the generic plan-level distance.

Let stage $s \in [Q]$ map metric space $(\mathcal{X}_{s-1}, d_{s-1})$ to $(\mathcal{X}_s, d_s)$ through $F_{s,e}$. Consider

$$\gamma = (e_1, \dots, e_Q), \qquad \gamma' = (e'_1, \dots, e'_Q), \tag{142}$$

and suppose every suffix map F_{r,e_r} from the first sequence is L_r-Lipschitz on the reached hybrid region. Let

$$x'_{s-1} = F_{s-1,e'_{s-1}} \circ \cdots \circ F_{1,e'_1}(x), \qquad A_s^{\mathrm{in}} = \prod_{r=s+1}^{Q} L_r, \tag{143}$$

with the empty prefix and product interpreted as the identity and one.

Theorem 25 (Sequential route-plan discrepancy)**.** *The output maps $F_\gamma = F_{Q,e_Q} \circ \cdots \circ F_{1,e_1}$ and $F_{\gamma'}$ satisfy*

$$d_Q(F_\gamma(x), F_{\gamma'}(x)) \le \sum_{s=1}^{Q} A_s^{\mathrm{in}} d_s\big(F_{s,e_s}(x'_{s-1}), F_{s,e'_s}(x'_{s-1})\big)\,. \tag{144}$$

Thus one early substitution can have a large cost in an expansive chain, while several index differences can be harmless when the substituted experts agree on the reached hybrid states. Variable-length chains can be embedded in a common maximum length by padding a terminated chain with an explicit identity/no-op transition on the terminal state. A zero map is generally not valid padding because it changes the state.

Combining Theorem 25 with the outer amplification A_j in Theorem 4 produces two scales of sensitivity: A_s^{in} measures propagation inside one routed event, while A_j measures propagation through later events. For a call at within-event position s of global event j, the natural terminal importance is their product.

I.6 Architecture-to-cost dictionary

Table 31: Architecture-specific quantities that upper-bound the generic functional routing term under their stated structural assumptions.

Architecture	Functional route-cost proxy
Additive soft gate	$\mathsf{W}_u(g,h)$ over expert outputs evaluated at one common input
Uniform top-k	Minimum bipartite matching cost in Equation (132)
Capacity-limited additive routing	Augmented transport through a zero-residual atom, charging $\|v_e(u)\|$ for dropped mass
Row-separable expert-choice assignment	Sum, average, or maximum of row-wise transport costs after token inputs are fixed
Full soft dispatch	Output transport plus Lipschitz-weighted input-dispatch transport in Equation (137)
Hierarchical additive MoE	Coarse outer transport plus conditional within-group transport
Preselected sequential experts	Suffix-amplified per-position substitutions in Equation (144)
Shared-expert residual block	Same conditional cost after exact cancellation of route-independent paths
General dispatch/adaptive program	Direct plan-level transition distance r_j; additional dispatch or policy sensitivity is required for a finer bound

These are upper bounds on transition discrepancy. They can be estimated from counterfactual forward passes while preserving the active-expert budget.

I.7 PROOFS

Proof of Theorem 21. For any $\gamma \in \Pi(g, h)$,

$$\sum_e (g_e - h_e) v_e = \sum_{e,e'} \gamma_{ee'} (v_e - v_{e'}). \tag{145}$$

The triangle inequality gives

$$\left\| \sum_e (g_e - h_e) v_e \right\| \leq \sum_{e,e'} \gamma_{ee'} c_u(e, e'). \tag{146}$$

Minimizing over couplings proves the first inequality. There exists a coupling that leaves the common mass $\min(g_e, h_e)$ on the diagonal and transports only total mass $\mathrm{TV}(g, h) = \|g - h\|_1 / 2$ off the diagonal. Every moved unit costs at most D_u, proving the second inequality. □

Proof of Corollary 22. After restricting to the two k-point uniform measures and multiplying a coupling by k, the feasible set consists of nonnegative matrices whose row sums and column sums all equal one. This is the *doubly stochastic polytope*: a bounded set described by linear equality and inequality constraints. An extreme point cannot be a nontrivial convex combination of two other feasible points. Such points are permutation matrices (one entry equal to one per row and column), and a linear objective attains a minimum at one of them; this is the standard doubly stochastic matrix characterization (Gabriel & Marco, 2019). Dividing by k gives Equation (132). Shared atoms simply admit zero-cost matched pairs. □

Proof of Proposition 24. Insert $M(p', q)$:

$$\|M(p, q) - M(p', q')\| \leq \|M(p, q) - M(p', q)\| + \|M(p', q) - M(p', q')\|. \tag{147}$$

Apply Theorem 21 to the outer mixtures of vectors $m_r(q_r)$ for the first term. For the second term, use the triangle inequality over groups and apply the same theorem inside each group. The reverse insertion is identical with the two routes exchanged. □

Proof of Theorem 25. For $s = 0, \ldots, Q$, let H_s use the prefix $(e'_1, \ldots, e'_s)$ from γ' and the suffix $(e_{s+1}, \ldots, e_Q)$ from γ. Thus $H_0 = F_\gamma$ and $H_Q = F_{\gamma'}$. Adjacent hybrids H_{s-1} and H_s receive the same state x'_{s-1} at stage s, differ only between F_{s,e_s} and F_{s,e'_s}, and share the suffix $(e_{s+1}, \ldots, e_Q)$. That suffix is A_s^{in}-Lipschitz from d_s to d_Q. The triangle inequality over $H_0, \ldots, H_Q$ proves Equation (144). □

J ENDOGENOUS ROUTING FEEDBACK AND MARGIN-STABLE TRACES

Overview. An early state error can change a later route and thereby create more error. We give sufficient support-stability conditions and probabilistic bounds for this feedback, retaining both router margins and expert-output geometry. Conditional margin control and weaker marginal control yield different guarantees. Sections J.1–J.2 treat soft-routing sensitivity and hard-support stability; Sections J.3–J.4 distinguish conditional from marginal margin bounds, followed by implications and proofs in Sections J.5–J.6.

J.1 SOFT ADDITIVE ROUTING PRODUCES AN ADDITIONAL LIPSCHITZ TERM

Consider an additive routed event

$$F(s) = M(s, g(s)) = B(s) + \sum_{e=1}^{K} g_e(s) v_e(s), \tag{148}$$

with $g(s) \in \Delta_K$. The output is in a normed vector space and its distance is the norm of a difference; when one metric d is used on both sides, this compatibility is assumed. Otherwise use separate input metric and output norm. Assume that for every fixed gate h,

$$d(M(s,h), M(s',h)) \le L_x d(s,s'). \tag{149}$$

At state s', define the functional transport cost $\mathsf{W}_{s'}$ as in Equation (129), using the collection $\{v_e(s')\}$ at the common input s'.

Proposition 26 (Closed-loop sensitivity of an additive soft MoE event)**.** *If the router is functionally transport-Lipschitz,*

$$\mathsf{W}_{s'}(g(s), g(s')) \le L_W d(s,s'), \tag{150}$$

then

$$d(F(s), F(s')) \le (L_x + L_W) d(s,s'). \tag{151}$$

In particular, if $\|g(s) - g(s')\|_1 \le L_g d(s,s')$ and the expert-output diameter at the reached states is uniformly at most $D_{\max}$, one may take

$$L_W \le \frac{1}{2} D_{\max} L_g. \tag{152}$$

Thus soft routing adds a sensitivity proportional to both gate movement and functional expert separation. This result applies to the additive common-input form; a dispatch mechanism whose gate changes expert inputs requires an additional dispatch-sensitivity term.

J.2 Hard top-k support is locally stable inside a margin tube

Assume the reference and learned router coordinates have been placed in one fixed canonical role alignment (or already share the same expert coordinates). Let $z : \mathcal{S} \to \mathbb{R}^K$ be reference router logits, fix $1 \le k < K$, and let $\mathcal{I}_k(z)$ denote the indices of the k largest coordinates under a fixed tie rule. Write $z_{(1)} \ge \cdots \ge z_{(K)}$ and define the top-k support margin

$$\gamma_k(s) = z_{(k)}(s) - z_{(k+1)}(s), \tag{153}$$

with $k = 1$ giving the ordinary winner margin.

Lemma 27 (Top-k support-margin stability)**.** *If another logit vector $\widetilde{z}$ satisfies*

$$\|\widetilde{z} - z(s)\|_\infty < \frac{1}{2}\gamma_k(s), \tag{154}$$

then $\mathcal{I}_k(\widetilde{z}) = \mathcal{I}_k(z(s))$. More generally, if a learned router $\widehat{z}$ satisfies

$$\|\widehat{z}(s') - z(s')\|_\infty \le \beta, \qquad \|z(s') - z(s)\|_\infty \le L_R d(s,s'), \tag{155}$$

then the supports agree whenever

$$\gamma_k(s) > 2\big(\beta + L_R d(s,s')\big). \tag{156}$$

For unit-gain top-1 or uniform hard top-k, support agreement determines the entire gate. For top-k followed by nonuniform normalized weights, the lemma controls only the support; any remaining within-support weight discrepancy can be bounded by the transport result in Theorem 21. Top-1 with a selected softmax gain also retains this weight variation; Proposition 51 differentiates that gain explicitly.

The margin tube yields a pathwise no-support-cascade condition. Let the reference support at event j be $\mathcal{I}_k(z_j(S_j))$ and the learned support be $\mathcal{I}_k(\widehat{z}_j(\widehat{S}_j))$. Assume the reference logits are $L_{R,j}$-Lipschitz, and that

$$\|\widehat{z}_j(s) - z_j(s)\|_\infty \le \beta_j \tag{157}$$

holds on the reached tube. Assume every reference transition with its complete route (support and any route weights) frozen is L_j-Lipschitz, uniformly over the reference routes reached. Finally, suppose that whenever the learned and reference supports agree, the same-state event discrepancy–including model approximation and, for nonuniform top-k, any within-support weight difference–is at most η_j. Define

$$\bar{e}_0 = d(\widehat{S}_0, S_0), \qquad \bar{e}_{j+1} = L_j \bar{e}_j + \eta_j. \tag{158}$$

Theorem 28 (Margin-certified absence of support cascades)**.** *If, along the reference trace,*

$$\gamma_{k,j}(S_j) > 2\big(\beta_j + L_{R,j}\bar{e}_j\big) \qquad \text{for every } j, \tag{159}$$

then the learned and reference top-k supports agree at every event and

$$d(\widehat{S}_j, S_j) \le \bar{e}_j \qquad \text{for all } j. \tag{160}$$

*For unit-gain top-*1 *and uniform hard top-k, the corresponding discrete route sequence is unchanged, so no support-change routing penalty is incurred. For nonuniform gates, including selected-softmax-gain top-*1*, same-support weight error remains inside* η_j*.*

The condition is sufficient rather than necessary. It identifies a support-stability basin: local approximation or within-support weight error may be nonzero, but it cannot change future selected supports while the accumulated state deviation remains inside every margin tube.

J.3 A MARGIN-DISTRIBUTION BOUND FOR EXPECTED SUPPORT CHANGES

Uniform positive margins are often unrealistic. An expected guarantee follows from a conditional anti-concentration condition. Let

$$E_j = d(\widehat{S}_j, S_j), \qquad B_j = \left\| \widehat{z}_j(\widehat{S}_j) - z_j(\widehat{S}_j) \right\|_\infty, \qquad U_j = B_j + L_{R,j}E_j. \tag{161}$$

By Lemma 27, a support mismatch implies $\gamma_{k,j}(S_j) \le 2U_j$. Assume that there exists a version of the conditional margin CDF such that, almost surely, for every $t \ge 0$,

$$\mathbb{P}(\gamma_{k,j}(S_j) \le t \mid U_j) \le \min\{1, c_j t^{\alpha_j}\} \tag{162}$$

for constants $c_j > 0$ and $\alpha_j > 0$. This conditional statement is essential: an unconditional margin histogram does not generally control a random, dependent threshold $2U_j$.

Let M_j be the indicator of a support mismatch. Let $r_j \ge 0$ denote the full route-induced transition discrepancy at event j, and assume the pointwise decomposition

$$r_j \le \zeta_j + \Delta_j M_j, \qquad \zeta_j \ge 0. \tag{163}$$

Here ζ_j bounds route discrepancy that can remain even when supports agree, such as nonuniform within-support weight changes; for unit-gain top-1 or uniform top-k, one may take $\zeta_j = 0$. The deterministic constant $\Delta_j \ge 0$ bounds the additional severity associated with a support change.

Theorem 29 (Margin-mediated expected support-change cost)**.** *Under the preceding conditions, with* $\rho_j = \mathbb{E}r_j$*,*

$$\rho_j \le \mathbb{E}\zeta_j + \Delta_j \mathbb{E}\min\left\{1, c_j(2U_j)^{\alpha_j}\right\}. \tag{164}$$

For $\alpha_j = 1$ *and finite first moments, the simpler (possibly looser) bound holds without restricting the range of* U_j*:*

$$\rho_j \le \mathbb{E}\zeta_j + 2c_j\Delta_j\left(\mathbb{E}B_j + L_{R,j}\mathbb{E}E_j\right). \tag{165}$$

Consequently, if a pathwise local recursion has the form

$$E_{j+1} \le a_j + L_j E_j + r_j, \tag{166}$$

then

$$\mathbb{E}E_{j+1} \le \mathbb{E}a_j + \mathbb{E}\zeta_j + 2c_j\Delta_j\mathbb{E}B_j + \left(L_j + 2c_j\Delta_j L_{R,j}\right)\mathbb{E}E_j. \tag{167}$$

The extra factor $2c_j\Delta_j L_{R,j}$ is an endogenous support-routing amplification term. It is large when margins concentrate near zero, logits are state-sensitive, or support substitutions are functionally severe. For $\alpha_j \ne 1$, Equation (164) yields a nonlinear moment recursion involving $\mathbb{E}U_j^{\alpha_j}$ (or the clipped moment appearing explicitly in the theorem).

J.4 A WEAKER MARGINAL-MARGIN GUARANTEE UNDER ARBITRARY DEPENDENCE

Conditioning on U_j in Equation (162) can be restrictive because U_j itself depends on the trajectory and router margin. A marginal margin assumption still gives a guarantee, but generally a slower one. Suppress j and let $\gamma \ge 0$ be a support margin, $U \ge 0$ the logit perturbation bound, and M the support-mismatch indicator. Suppose $M \le \mathbf{1}\{\gamma \le 2U\}$, $\mathbb{P}(\gamma \le t) \le ct^\alpha$ for all $t \ge 0$, and $m_r = \mathbb{E}U^r < \infty$ for some moment order $r > 0$. The constants $c > 0, \alpha > 0$ control how much margin probability lies near zero; this is called a margin anti-concentration condition.

Theorem 30 (Dependence-robust marginal-margin bound)**.** *No independence between γ and U is required for*

$$\mathbb{P}(M=1) \le \min\left\{1, \inf_{t>0}\left[c(2t)^{\alpha} + \frac{m_r}{t^r}\right]\right\}. \tag{168}$$

A sharper explicit bound, not merely the optimization of the preceding split, is

$$\mathbb{P}(M=1) \le \min\{1, K_{\alpha,r}(2^{\alpha}c)^{r/(\alpha+r)} m_r^{\alpha/(\alpha+r)}\}, \tag{169}$$

where $K_{\alpha,r} = ((\alpha+r)/\alpha)^{\alpha/(\alpha+r)}$. For $m_r = 0$, the mismatch probability is zero. If the route loss satisfies $r_{\mathrm{route}} \le \zeta + \Delta M$, multiply the mismatch bound by Δ and add $\mathbb{E}\zeta$ to obtain a functional-cost bound.

Proof. For every fixed threshold $t > 0$, a mismatch implies either $\gamma \le 2t$ or $U > t$. A union bound and Markov's inequality give Equation (168). For the sharper bound, let $E = \{M = 1\}$ and $p = \mathbb{P}(E)$. On E, $U \ge \gamma/2$. The tail-integral identity for a nonnegative random variable and $\mathbb{P}(E, \gamma > t) \ge p - ct^{\alpha}$ imply, for $p > 0$,

$$m_r \ge 2^{-r}\mathbb{E}[\gamma^r \mathbf{1}_E] \tag{170}$$

$$\ge 2^{-r}\int_0^{(p/c)^{1/\alpha}} rt^{r-1}(p - ct^{\alpha})\,dt \tag{171}$$

$$= 2^{-r}\frac{\alpha}{\alpha+r}c^{-r/\alpha}p^{1+r/\alpha}. \tag{172}$$

Rearranging yields Equation (169); the case $p = 0$ is immediate. If $m_r = 0$, then $U = 0$ almost surely and $\mathbb{P}(\gamma = 0) = 0$ by the margin assumption. No conditioning or independence is used. For example, $\alpha = r = c = 1$ yields $2\sqrt{m_1}$ rather than the looser $2\sqrt{2m_1}$ from optimizing the split. □

Proposition 31 (Sharp exponent and limiting constant under marginal control)**.** *For every $\alpha, r, c > 0$, every $\varepsilon > 0$, and every $0 < b \le c^{-1/\alpha}$, there is a two-expert top-1 example with margin CDF $\mathbb{P}(\gamma \le t) = \min\{1, ct^{\alpha}\}$ and strict support changes such that*

$$\mathbb{P}(M=1) = \left[\frac{\alpha+r}{\alpha}\left(\frac{2}{1+\varepsilon}\right)^r c^{r/\alpha}\mathbb{E}U^r\right]^{\alpha/(\alpha+r)}. \tag{173}$$

Letting $b \downarrow 0$ shows that the moment exponent cannot improve. Letting $\varepsilon \downarrow 0$ shows that the multiplicative constant in the nontrivial part of Equation (169) cannot be uniformly reduced under the stated assumptions.

Proof. Use logits $z = (\gamma/2, -\gamma/2)$. On $\{\gamma \le b\}$ replace them by $\widetilde{z} = (-\varepsilon\gamma/2, \varepsilon\gamma/2)$; otherwise leave them unchanged. The reference and perturbed winners differ strictly except on the zero-probability event $\gamma = 0$. Thus

$$U = \frac{1+\varepsilon}{2}\gamma\mathbf{1}_{\{\gamma\le b\}}, \qquad \mathbb{P}(M=1) = cb^{\alpha}, \qquad \mathbb{E}U^r = \left(\frac{1+\varepsilon}{2}\right)^r \frac{c\alpha}{\alpha+r}b^{\alpha+r}.$$

Eliminating b proves the formula. Choose $cb^{\alpha} < 1$ to avoid saturation and then let $\varepsilon \downarrow 0$ for limiting sharpness. Positive ε avoids relying on a favorable tie rule. □

The perturbation in this example uses only currently available logits, not future information. Small-margin states are precisely the ones receiving error. This explains why an unconditional margin histogram cannot justify the faster conditional bound. For a first-moment state certificate and $r = 1$, the robust dependence on the error moment has exponent $\alpha/(\alpha+1)$, not α.

J.5 Implications for expert reuse

Cross-layer and global pools can increase the number of candidate experts, which can alter router margins. They may also place functionally redundant experts near one another, reducing Δ_j or the transport cost of a support change. Therefore candidate-pool size alone does not determine route stability. The relevant measurements are jointly:

1. the top-k support-margin distribution under the reached occupancy.
2. the state-to-logit sensitivity $L_{R,j}$ or an empirical local analogue.
3. same-support weight discrepancy ζ_j when weights are not uniform.
4. the functional geometry of candidate expert outputs.
5. the position-dependent downstream amplification.

These quantities turn the otherwise exogenous ρ_j into a measurable closed-loop diagnostic. They also explain why route consistency can deteriorate with reasoning length even when the one-step expert estimators remain unchanged.

J.6 Proofs

Proof of Proposition 26. Insert $M(s', g(s))$:

$$d(F(s), F(s')) \le d(M(s, g(s)), M(s', g(s))) + d(M(s', g(s)), M(s', g(s'))). \tag{174}$$

The first term is at most $L_x d(s, s')$ by Equation (149); the second is at most $\mathsf{W}_{s'}(g(s), g(s'))$ by Theorem 21. This proves Equation (151). The diameter specialization follows from the second inequality in Equation (131). □

Proof of Lemma 27. Every selected logit can decrease by at most $\|\widetilde{z} - z(s)\|_\infty$, and every unselected logit can increase by at most the same amount. If this perturbation is below $\gamma_k(s)/2$, the smallest selected coordinate remains strictly larger than the largest unselected coordinate. For the second statement, the triangle inequality bounds $\|\widehat{z}(s') - z(s)\|_\infty$ by $\beta + L_R d(s, s')$. □

Proof of Theorem 28. Proceed by induction. At $j = 0$, the error is at most $\bar{e}_0$. If $d(\widehat{S}_j, S_j) \le \bar{e}_j$, Equation (159) and Lemma 27 imply identical supports. By the assumed same-support event-discrepancy bound and the Lipschitz property of the reference fixed-support transition,

$$d(\widehat{S}_{j+1}, S_{j+1}) \le \eta_j + L_j d(\widehat{S}_j, S_j) \le \bar{e}_{j+1}. \tag{175}$$

Thus support agreement and the error envelope hold at every event. □

Proof of Theorem 29. A support mismatch implies $\gamma_{k,j}(S_j) \le 2U_j$. Conditioning on U_j and applying Equation (162) at the random threshold $2U_j$ gives

$$\mathbb{P}(M_j = 1) \le \mathbb{E}\min\{1, c_j(2U_j)^{\alpha_j}\}. \tag{176}$$

Taking expectations in Equation (163) proves Equation (164). For $\alpha_j = 1$, use $\min\{1, 2c_jU_j\} \le 2c_jU_j$, substitute $U_j = B_j + L_{R,j}E_j$ and use linearity of expectation. Taking expectations in Equation (166) and inserting the resulting route-cost bound proves Equation (167). □

K Counterfactual Compatibility and the Sample Complexity of Route Audits

Overview. We separate a population comparison of action effects from the finite-sample task of auditing them. Under a declared randomized intervention law, counterfactual variance measures incompatibility of action means; absent overlap, unsupported effects are not identified. Independent records and positive forcing probabilities support action-specific concentration. Sections K.1–K.2 establish the population comparison and the overlap obstruction; Sections K.3–K.5 give audit bounds and budget allocation, with remaining proofs in Section K.6.

K.1 Reuse bias as counterfactual transition variance

Let S be a pre-action macro state in a standard Borel space and let $A \in [K]$ be drawn from a declared randomized intervention distribution $p_a(S) = \mathbb{P}(A = a \mid S)$. Let Y^a be the potential post-action transition target for action a in a separable real Hilbert space $\mathcal{H}$. Assume every Y^a is

Bochner measurable and square-integrable, conditional randomization $A \perp (Y^1, \ldots, Y^K) \mid S$, and consistency $Y = Y^A$. Define the Bochner conditional mean

$$m_a(s) = \mathbb{E}[Y^a \mid S = s]. \tag{177}$$

Values of m_a outside the support of $p_a(S)$ are immaterial to the displayed risks. The evaluation experiment draws A and observes Y^A. Compare action-specific predictors $f_A(S)$ with one shared predictor $f(S)$ under squared loss.

Theorem 32 (Counterfactual decomposition of reuse bias)**.** *Under the conditions above, allow all measurable predictors with finite weighted second moment. The action-specific and shared Bayes predictors are*

$$f_a^\star(s) = m_a(s), \qquad f_{\mathrm{share}}^\star(s) = \bar{m}(s) := \sum_a p_a(s) m_a(s) \tag{178}$$

on the corresponding supports. Their oracle-risk difference is

$$\mathsf{B}_{\mathrm{cf}}^{\mathrm{reuse}} = \mathbb{E}\,\|m_A(S) - \bar{m}(S)\|_{\mathcal{H}}^2 \tag{179}$$

$$= \frac{1}{2}\mathbb{E}_S \sum_{a,b} p_a(S) p_b(S)\,\|m_a(S) - m_b(S)\|_{\mathcal{H}}^2\,. \tag{180}$$

This is the action-level analogue of Proposition 7. It separates irreducible transition uncertainty, which appears in both oracle risks, from incompatibility among conditional action means, which is paid only by the shared operator. The intervention distribution is part of the estimand: changing which actions are compared, or where they are forced, changes $\mathsf{B}_{\mathrm{cf}}^{\mathrm{reuse}}$.

The identity clarifies the relation between functional and causal reuse. Small observational disagreement among naturally selected experts does not imply small Equation (180); the latter compares actions at common pre-action states. Conversely, distinct indices can have negligible causal sharing bias if their forced conditional transition means coincide on the stated intervention occupancy. The theorem is a population identity. A plug-in estimate of $\mathsf{B}_{\mathrm{cf}}^{\mathrm{reuse}}$ must estimate the conditional means m_a and evaluate their squared differences on independent or cross-fitted records, or else control the complexity and overfitting of the learned functions.

When the local and causal comparisons are the same. For a supported action set, define $\pi_a = \mathbb{P}(A = a) > 0$ and the local population risk

$$R_a(f) = \mathbb{E}[\|Y^a - f(S)\|_{\mathcal{H}}^2 \mid A = a].$$

Use the same predictors of S, the same squared loss, and these weights in the local shared/separate oracle definitions. Consistency and the law of total expectation give $\sum_a \pi_a R_a(f) = \mathbb{E}\|Y^A - f(S)\|_{\mathcal{H}}^2$; allowing a separate f_a gives the corresponding action-specific objective. Their population bias is then exactly the bias in Theorem 32. This alignment uses the action-conditional state laws, which need not equal the marginal law of S.

Without this matched experiment and class, the two biases are different quantities. In particular, tying a physical expert does not remove action information encoded in its input. A statistical role such as operation–depth pair (i, ℓ) identifies a call environment; a route action specifies an intervention on executable calls. Equating those indices requires an explicit map, not merely identical names.

K.2 Without overlap, unsupported route effects are not identifiable

Let an observational or intervention policy select action a with propensity $e_a(C) = \mathbb{P}(A = a \mid C)$. Suppose there is a measurable set $\mathcal{B}$ with $\mathbb{P}(C \in \mathcal{B}) > 0$ on which $e_a(C) = 0$.

Proposition 33 (No-overlap non-identifiability)**.** *Without structural restrictions linking unsupported and supported action effects, the distribution of $T_a(C)$ on $\mathcal{B}$ is not identified from data generated by the policy whenever the model class permits two different effects there. A defect or utility depending nontrivially on those differences is likewise not identified; a constant summary need not be affected. In a nondegenerate binary-output model, for example, there exist two action libraries with identical observed-data distributions but different values of*

$$\mathbb{E}[d(T_a(C), F_a(\Psi(C)))\,\mathbf{1}\{C \in \mathcal{B}\}]\,. \tag{181}$$

The conclusion is not repaired by more samples from the same policy. It requires either a structural extrapolation assumption or interventions that assign positive probability to the unsupported action–context region.

K.3 A FINITE-SAMPLE AUDIT OF FIXED ACTION-SPECIFIC DEFECTS

The macro extractor Ψ, action definitions, candidate operators F_a, and bounded losses in this subsection are fixed independently of the audit records. They may be learned on a separate fitting split and then conditioned upon. Cross-fitting requires a separate conditional bound for each held-out fold and a union bound over folds; if the fitted operators differ across folds, the resulting target is their average fold-specific risk, not automatically the risk of a final refitted model. Treating all cross-fitted scores as IID conditional on every fold fit is not justified. Reusing the same audit observations to select these objects requires a uniform complexity correction.

Suppose the context–potential-outcome records are IID across $i = 1, \dots, N$. Conditional on context C_i, an action can have stochastic output $T_a(C_i; U_i)$, where U_i is unrevealed execution randomness. Define the bounded potential losses and their conditional means by

$$L_{i,a} = d(T_a(C_i; U_i), F_a(\Psi(C_i))) \in [0, 1], \quad h_a(c) = \mathbb{E}[L_{i,a} \mid C_i = c]. \tag{182}$$

Conditional on C_i, randomize A_i independently of its potential-loss vector with known propensity $e_a(C_i)$, independently across records, and observe $L_i = L_{i,A_i}$. Deterministic execution is the special case without U_i. Let

$$\mu_a = \mathbb{E}L_{i,a} = \mathbb{E}h_a(C_i) \tag{183}$$

under the probe-context distribution. The Horvitz–Thompson estimator is (Horvitz & Thompson, 1952)

$$\widehat{\mu}_a = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}\{A_i = a\} L_i}{e_a(C_i)}. \tag{184}$$

Theorem 34 (Simultaneous concentration of a randomized route audit)**.** *Assume the intervention policy has overlap*

$$e_a(C) \geq \beta > 0 \quad \textit{a.s. for every claimed action } a \in [K], \qquad 0 < \beta \leq 1/K. \tag{185}$$

Then each $\widehat{\mu}_a$ is unbiased, and with probability at least $1 - \delta$,

$$\max_{a \in [K]} |\widehat{\mu}_a - \mu_a| \leq \sqrt{\frac{2 \log(2K/\delta)}{N\beta}} + \frac{\log(2K/\delta)}{3N\beta}. \tag{186}$$

Consequently, $N\beta$ is the leading effective sample size for a simultaneous claim about the least-covered action. Under uniform random forcing, $\beta = 1/K$, so estimating all K action defects to fixed accuracy requires order K audit records, up to the usual accuracy and logarithmic factors.

The theorem applies to a loss evaluated after one expert call, a compound multi-layer route action, or a bounded stopping-time macro action, provided the action and loss are defined before the audit records are observed. If an action changes active compute, the comparator should include matched-compute null or substitute actions so that the audit isolates functionality rather than budget.

For fixed nonnegative terminal weights w_a, the same simultaneous event implies the conservative bound

$$\left| \sum_a w_a \widehat{\mu}_a - \sum_a w_a \mu_a \right| \leq \left(\sum_a w_a \right) \left[\sqrt{\frac{2 \log(2K/\delta)}{N\beta}} + \frac{\log(2K/\delta)}{3N\beta} \right]. \tag{187}$$

This is useful when every action-specific defect must also be reported, but it is not the optimal rate for every aggregate. In particular, for the behavior-policy-weighted target

$$\mu_e = \mathbb{E} \sum_a e_a(C) h_a(C) = \mathbb{E}L_i, \tag{188}$$

the direct estimator $\widehat{\mu}_e = N^{-1} \sum_i L_i$ satisfies, by Hoeffding,

$$|\widehat{\mu}_e - \mu_e| \leq \sqrt{\frac{\log(2/\delta)}{2N}} \tag{189}$$

with probability at least $1-\delta$, without a $1/\beta$ factor. Thus $N\beta$ is the relevant size for uniform action-specific or least-covered-action claims, not a universal sample size for all policy averages. A different target policy $q_a(C)$ can be evaluated by weighting with $q_a(C)/e_a(C)$, whose rate depends on the resulting importance-weight envelope or variance.

K.4 The per-action coverage dependence is minimax unavoidable

The $1/\sqrt{N\beta}$ term is not merely an artifact of inverse-propensity weighting.

Proposition 35 (Per-action coverage lower bound)**.** *There is a universal constant $c>0$ such that, for any estimator $\widetilde{\mu}_a$ based on N randomized records with constant propensity $e_a=\beta$, one can choose a bounded binary intervention-outcome distribution for which*

$$\mathbb{E}\,|\widetilde{\mu}_a-\mu_a| \geq c\min\left\{1,\frac{1}{\sqrt{N\beta}}\right\}. \tag{190}$$

Thus an action that is forced only rarely cannot be audited precisely merely because the total token count is large. The proof is a two-point Le Cam argument (Tsybakov & Zaiats, 2009).

K.5 Allocating a finite intervention budget

Uniform forcing is simple but is not generally variance optimal for a performance-weighted aggregate. Suppose the target is $\mu_w=\sum_a w_a\mu_a$, constant action propensities $e_a\geq 0$ sum to one, and

$$s_a=\mathbb{E}L_{i,a}^2. \tag{191}$$

For every action with $w_a^2 s_a>0$, require $e_a>0$. The unbiased single-record estimator is $w_{A_i}L_i/e_{A_i}$, with second moment $\sum_a w_a^2 s_a/e_a$ over the active set.

Proposition 36 (Variance-optimal route-intervention allocation)**.** *Assume $w_a\geq 0$ and not all $w_a\sqrt{s_a}$ vanish. Let $\mathcal{A}_+=\{a: w_a\sqrt{s_a}>0\}$. Ignoring positivity requirements for actions outside $\mathcal{A}_+$ and lower-propensity constraints, the action distribution minimizing the second moment of the weighted estimator is*

$$e_a^\star=\frac{w_a\sqrt{s_a}}{\sum_{b\in\mathcal{A}_+}w_b\sqrt{s_b}}\quad(a\in\mathcal{A}_+),\qquad e_a^\star=0\quad(a\notin\mathcal{A}_+), \tag{192}$$

with minimum

$$\left(\sum_{a\in\mathcal{A}_+}w_a\sqrt{s_a}\right)^2. \tag{193}$$

A pilot sample can estimate s_a, after which an independent second phase can use Equation (192). If action-specific claims are required for every action, or if future targets may assign them positive weight, impose a floor $e_a\geq\beta$ with $K\beta\leq 1$; if $K\beta<1$ and some $w_a\sqrt{s_a}>0$, the constrained solution is $e_a=\max\{\beta, w_a\sqrt{s_a}/t\}$, where $t>0$ is chosen so that $\sum_a e_a=1$. If $K\beta=1$, the unique feasible allocation is $e_a=\beta$. These formulas follow from the Lagrange stationarity condition on coordinates above the floor; when all second-moment contributions vanish, every feasible allocation is optimal. Actions with zero contribution to the declared aggregate may receive zero mass only when no claim about their individual effects is made.

K.6 Proofs

Proof of Theorem 32. For a supported action a, conditional randomization gives $\mathbb{E}[Y^a\mid S,A=a]=m_a(S)$. The Hilbert-space Pythagorean identity for Bochner conditional expectation gives

$$\mathbb{E}\Big[\|Y^a-f(S)\|^2\mid S,A=a\Big]=\mathbb{E}\Big[\|Y^a-m_a(S)\|^2\mid S\Big]+\|m_a(S)-f(S)\|^2. \tag{194}$$

Action-specific prediction sets $f_a=m_a$ and leaves only the first term. The shared problem minimizes $\sum_a p_a(S)\,\|m_a(S)-f(S)\|^2$ pointwise, whose minimizer is $\bar{m}(S)$. Subtracting the common irreducible term yields Equation (179). Applying the weighted Hilbert variance identity conditionally on S gives Equation (180). □

Proof of Proposition 33. Construct two action libraries that agree for every action–context pair with positive observation probability and differ only in the value of $T_a(C)$ for $C \in \mathcal{B}$. Because action a has zero probability on $\mathcal{B}$, the changed transition is never observed, so the induced observed-data distributions are identical. Choosing the two unsupported transitions to have different distances from $F_a(\Psi(C))$ changes Equation (181). Therefore the estimand is not identified without additional structural restrictions. □

Proof of Theorem 34. For fixed a, let

$$X_{i,a} = \frac{\mathbf{1}\{A_i = a\}L_i}{e_a(C_i)}. \tag{195}$$

Conditional randomization gives $\mathbb{E}[X_{i,a} \mid C_i] = h_a(C_i)$, hence $\mathbb{E}X_{i,a} = \mu_a$. Moreover $0 \le X_{i,a} \le 1/\beta$ and

$$\mathbb{E}X_{i,a}^2 \le \mathbb{E}\frac{\mathbf{1}\{A_i = a\}}{e_a(C_i)^2} = \mathbb{E}\frac{1}{e_a(C_i)} \le \frac{1}{\beta}. \tag{196}$$

Thus $\mathrm{Var}(X_{i,a}) \le 1/\beta$, and $|X_{i,a} - \mu_a| \le 1/\beta$. To make the Bernstein constant explicit, put $Z_a = \widehat{\mu}_a - \mu_a$. Expanding the exponential moment of each centered bounded summand and using $k! \ge 2 \cdot 3^{k-2}$ for integers $k \ge 2$ gives

$$\log \mathbb{E}e^{tZ_a} \le \frac{t^2/(N\beta)}{2(1 - |t|/(3N\beta))}, \qquad |t| < 3N\beta.$$

This is an exponential-moment bound for both signs of Z_a. The Chernoff argument (apply Markov's inequality to e^{tZ_a} and optimize t) gives $\mathbb{P}\{|Z_a| > \sqrt{2x/(N\beta)} + x/(3N\beta)\} \le 2e^{-x}$. Take $x = \log(2K/\delta)$ and union-bound over actions. □

Proof of Proposition 35. It suffices to consider a context-free stochastic binary outcome, allowed by the potential-loss model above. Under hypotheses P_- and P_+, action a is selected with probability β, and conditional on selection its outcome is Bernoulli with means $1/2 - \varepsilon$ and $1/2 + \varepsilon$, respectively; every other observation has the same law. For $\varepsilon \le 1/4$, the one-record Kullback–Leibler divergence is at most $16\beta\varepsilon^2$, so the N-record divergence is at most $16N\beta\varepsilon^2$. Choosing $\varepsilon = c_0 \min\{1, (N\beta)^{-1/2}\}$ with a sufficiently small universal $c_0 \le 1/4$ keeps total variation bounded away from one by Pinsker's inequality. Le Cam's two-point inequality then lower-bounds the maximum expected absolute error by a universal constant times the parameter separation, proving Equation (190). □

Proof of Proposition 36. Restrict to $\mathcal{A}_+$. Cauchy–Schwarz gives

$$\left(\sum_{a\in\mathcal{A}_+} w_a\sqrt{s_a}\right)^2 = \left(\sum_{a\in\mathcal{A}_+} \frac{w_a\sqrt{s_a}}{\sqrt{e_a}}\sqrt{e_a}\right)^2 \le \left(\sum_{a\in\mathcal{A}_+} \frac{w_a^2 s_a}{e_a}\right)\left(\sum_{a\in\mathcal{A}_+} e_a\right). \tag{197}$$

At an optimum no mass is assigned outside $\mathcal{A}_+$, so the second factor is one. Equality holds exactly when $e_a \propto w_a\sqrt{s_a}$ on $\mathcal{A}_+$, yielding the stated allocation and minimum. □

L Extensions, Assumption Ledger, and Research Protocol

Overview. This appendix collects complementary extensions and proposed uses of the framework: functional corrections, grouping selection, stochastic macro transitions, and resource-adjusted scores. Sections L.1–L.3 cover functional corrections, grouping, and stochastic dynamics; Sections L.4–L.5 distinguish resource preferences and assumptions.

L.1 Partial functional reuse beyond a scalar shrinkage coefficient

The exact linear theorem uses one scalar λ. A function-space shared core plus an r-dimensional role correction gives a complementary interpolation.

Proposition 37 (Residual bias of a rank-r functional correction)**.** *Under the unrestricted Hilbert quadratic setting $\mathcal{F}_e = \mathcal{H}$ of Proposition 7, assume finite-dimensional $\mathcal{H}$, fix an integer $0 \le r \le \dim(\mathcal{H})$, and define*

$$\Sigma = \sum_a \pi_a (f_a^\star - \bar{f}) \otimes (f_a^\star - \bar{f}) \tag{198}$$

where $(u \otimes u)v = u\langle u, v\rangle$ and the eigenvalues are $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$. They measure between-role variation along orthogonal function directions, not estimation-noise variance. Among role predictors $f_a = \bar{f} + Uz_a$ with $U : \mathbb{R}^r \to \mathcal{H}$, $U^\top U = I_r$, and role-specific coordinates $z_a \in \mathbb{R}^r$, sharing an r-dimensional correction subspace, the minimum average excess risk is

$$(\mathsf{B}_e^{\mathrm{reuse}})^{(r)} = \sum_{k>r} \lambda_k. \tag{199}$$

Proof. For fixed subspace U, the optimal role correction is the orthogonal projection of $f_a^\star - \bar{f}$ onto U. The average residual is $\mathrm{tr}((I - P_U)\Sigma)$. The rank-$r$ projector maximizing captured trace spans the top r eigenvectors, with the usual convention that the tail is zero once r reaches the rank. □

This is a function-space statement. Connecting it to a rank-r LoRA parameterization requires a separate realizability assumption about the network Jacobian or adapter class, and adding estimation variance for the learned role coordinates is necessary for a complete finite-sample adapter comparison.

For a concrete obstruction, take linear maps on $\mathbb{R}^2$, input second moment I_2, two equally weighted role targets $f_\pm(x) = \pm I_2 x$, and a fixed zero shared core. Their centered variation spans one function direction, so a one-dimensional functional correction is exact. But either target matrix has two unit singular values: a matrix-rank-one update to this fixed core has squared Frobenius residual at least one. Thus function-subspace dimension one does not imply parameter-update rank one.

Sections O.5 and O.6 gives exact frozen-feature adapter tails, a two-role free-core result, and initialization-dependent tangent images.

L.2 Data-dependent grouping

The fixed-rule identities condition on a reuse assignment independent of fitting noise. Adaptive grouping therefore needs a separate selection argument.

In the exact linear model, Theorem 2 supplies a population objective without an ad hoc clustering metric: minimize within-group Frobenius dispersion plus V per fitted expert. Its pairwise merge statistic has the exact law in Equation (64). Outside that model, the following validation-based construction retains the same logic without assuming that raw parameter distance equals functional heterogeneity.

First, fit a finite candidate family $\mathcal{A}$ of groupings using training data and select among those fitted candidates on an independent validation split. Conditional on training, if validation loss is bounded and has effective size $N_{\mathrm{val}}^{\mathrm{eff}}$, validation ERM satisfies the analogue of Equation (72) with $|\Lambda|$ replaced by $|\mathcal{A}|$. Second, if grouping and expert fitting use the same observations, uniform convergence must cover the union of all grouped hypothesis classes; a finite grouping family contributes an additional $\log|\mathcal{A}|$ term. An untouched final test is still required. Theorem 48 provides a finite-candidate oracle inequality for explicit loss-tail conditions, including a sufficient Gaussian quadratic model. Equation (71) illustrates the risk of a same-data dispersion plug-in.

A certificate-guided grouping objective may use

$$\widehat{U}(\alpha) = \sum_e \widehat{W}_e \left(\widehat{B}_e + 2\widehat{\epsilon}_{\mathrm{sh},e}\right) + \sum_j \widehat{A}_j(\widehat{\tau}_j + \widehat{\rho}_j) + \lambda_{\mathrm{res}}\,\mathrm{Cost}(\alpha), \tag{200}$$

where the common role-specific oracle term has been omitted only when it is invariant across the compared partitions. Every fitted functional probe and plug-in quantity used to choose α must be computed on training/validation data or by cross-fitting, and final evaluation is performed once on a held-out test set.

L.3 Stochastic macro dynamics

When the macro state is Markov-sufficient, deterministic transitions can be replaced by Markov kernels on a Polish (complete separable) metric space $(\mathcal{S}, d)$ with finite first moments. Let W_d denote the 1-Wasserstein distance with ground metric d, and let Q_j be the actual kernel and $\bar{Q}_j$ a reference kernel satisfying

$$\sup_s W_d(Q_j(s,\cdot), \bar{Q}_j(s,\cdot)) \le \eta_j, \qquad W_d(\bar{Q}_j(s,\cdot), \bar{Q}_j(s',\cdot)) \le L_j d(s,s'). \tag{201}$$

For distributions $\xi_{j+1} = \xi_j Q_j$ and $\bar{\xi}_{j+1} = \bar{\xi}_j \bar{Q}_j$, conditional coupling and the triangle inequality give

$$W_d(\xi_{j+1}, \bar{\xi}_{j+1}) \le \eta_j + L_j W_d(\xi_j, \bar{\xi}_j), \tag{202}$$

which telescopes exactly as Equation (75). Standard kernel measurability is assumed. If the one-step discrepancy is state-dependent rather than uniformly bounded, its expectation under ξ_j replaces η_j. Without macro-state sufficiency, kernels must be defined on the full micro context before being pushed forward through Ψ.

L.4 Resource-adjusted reuse

We evaluate statistical gains alongside serving costs. Define every cost difference as

$$\Delta C_\bullet = C_\bullet^{\text{reuse}} - C_\bullet^{\text{sep}}, \tag{203}$$

so a positive value means that reuse is more costly. A deployment score may retain system costs separately:

$$\mathsf{PRG}_H^{\text{resource}} = \mathsf{PRG}_H - \lambda_{\text{mem}}\Delta C_{\text{memory}} - \lambda_{\text{comm}}\Delta C_{\text{communication}} - \lambda_{\text{lat}}\Delta C_{\text{latency}} - \lambda_{\text{flop}}\Delta C_{\text{flop}}. \tag{204}$$

The nonnegative coefficients convert the different resource units into the chosen score units and encode deployment preferences rather than statistical constants, and the raw cost changes should be reported alongside the scalarized score.

L.5 Assumption dependencies

Section Q audits the assumptions of each core result and extension, including the exact-model restrictions, estimator independence, transfer conditions, route semantics, and causal coverage. It distinguishes mathematical sufficiency from empirical plausibility. The covariance and marginal-margin extensions in Sections J and M are used when the stronger uncorrelated-error or conditional-margin assumptions are not credible.

M Stronger Results under Weaker or More Explicit Assumptions

Overview. We identify what remains valid when the simplest assumptions are relaxed. The extensions retain general test geometry and covariance, optimization and transfer costs, and escape events. The final subsection separately explains how held-out paired evaluation can support actual performance improvement. Each extension specifies the object, randomness, and fixed quantities. Sections M.1–M.2 develop the general linear geometry; Sections M.3–M.6 retain optimization, transfer, repeated-error, and escape costs, before the paired audit in Section M.7.

M.1 General designs, correlated estimation errors, and test geometry

The design Gram describes which coefficients can be estimated; the test geometry describes which errors matter at evaluation. They play different roles, even when both are built from features.

The orthogonal design $X_a^\top X_a = n_a I_d$ is useful for isolating a scalar variance price, but it is not needed for an exact linear risk calculation. Let J role cells have parameter vectors $\theta_a \in \mathbb{R}^D$, stacked into $\theta \in \mathbb{R}^{JD}$. Suppose an unbiased base estimator has the form

$$Y = \theta + \zeta, \qquad \mathbb{E}\zeta = 0, \qquad \mathbb{E}[\zeta\zeta^\top] = \Sigma \succeq 0. \tag{205}$$

Here Y denotes the stacked parameter estimates, not the earlier response matrix Y_a. The covariance Σ may have nonzero cross-role blocks. Fix a symmetric positive-semidefinite matrix $W \in \mathbb{R}^{JD\times JD}$, and define $\|u\|_W^2 = u^\top W u$. This is a squared seminorm: directions in the null space of W do not affect the chosen evaluation loss. A fixed matrix $L \in \mathbb{R}^{JD\times JD}$ specifies a *linear sharing rule* $\widetilde{\theta}_L = LY$. The risk is

$$\mathcal{R}_W(L) = \mathbb{E}\|LY - \theta\|_W^2. \tag{206}$$

"Fixed" means chosen independently of ζ, or conditioned on an independent pilot sample. No Gaussian distribution is assumed.

Theorem 38 (Exact covariance-aware risk of a fixed linear sharing rule)**.** *Under Equation (205),*

$$\mathcal{R}_W(L) = \|(L-I)\theta\|_W^2 + \operatorname{tr}(WL\Sigma L^\top). \tag{207}$$

If Σ is known, then

$$\widehat{\mathcal{R}}_W(L) = \|(L-I)Y\|_W^2 + 2\operatorname{tr}(WL\Sigma) - \operatorname{tr}(W\Sigma) \tag{208}$$

is unbiased for the risk of this fixed L.

Proof. Write $LY - \theta = (L-I)\theta + L\zeta$. Its deterministic–centered cross term has zero expectation, and $\mathbb{E}[\zeta^\top L^\top W L\zeta] = \operatorname{tr}(WL\Sigma L^\top)$, proving Equation (207). Also

$$\mathbb{E}\|(L-I)Y\|_W^2 = \|(L-I)\theta\|_W^2 + \operatorname{tr}\big(W(L-I)\Sigma(L-I)^\top\big).$$

Expanding the last trace and using symmetry of W, Σ leaves the risk minus $2\operatorname{tr}(WL\Sigma) + \operatorname{tr}(W\Sigma)$, proving unbiasedness. □

This is the fixed-linear-estimator case of unbiased risk estimation, with an elementary second-moment proof (Stein, 1981). Unbiasedness for each pre-specified L does not imply unbiasedness of the minimum over data-selected rules. Selecting a grouping on these scores requires independent evaluation or uniform selection control. Replacing Σ by an estimate also adds an estimation problem. The unbiased score may be negative in a sample even though its expected risk is nonnegative.

Instantiation for nonorthogonal operator regression. For each role, let $X_a \in \mathbb{R}^{n_a\times d}$ have full column rank, define its design Gram matrix $G_a = X_a^\top X_a \succ 0$, and observe $Y_a = X_a A_a^\top + \mathcal{E}_a$. Assume the noise rows are centered, independent, and have output covariance $\Omega_a \in \mathbb{R}^{p\times p}$. Then

$$\widehat{A}_a = Y_a^\top X_a G_a^{-1}, \qquad \operatorname{Cov}(\operatorname{vec}(\widehat{A}_a - A_a)) = G_a^{-1} \otimes \Omega_a. \tag{209}$$

The operation vec stacks a matrix's columns; here $\otimes$ is the Kronecker product, which forms a block matrix by multiplying each entry of its first argument by the second matrix. Independence across roles makes the stacked covariance block diagonal; correlated role fits instead require their actual cross-covariances.

Let the independent test input in role a have second moment $T_a = \mathbb{E}[x_a x_a^\top] \succeq 0$, and let evaluation probabilities $\pi_a > 0$ sum to one. Since

$$\mathbb{E}_{x_a}\|Bx_a\|_2^2 = \operatorname{tr}(BT_aB^\top) = \operatorname{vec}(B)^\top(T_a \otimes I_p)\operatorname{vec}(B), \tag{210}$$

choose

$$W = \operatorname{blockdiag}_{a=1}^{J}\big(\pi_a(T_a \otimes I_p)\big). \tag{211}$$

Then Equation (207) is exactly the test prediction risk. This accommodates anisotropic representations, unequal test frequencies, and ill-conditioned but full-rank designs. A nearly singular G_a increases estimation variance through G_a^{-1}; role count alone does not describe this difficulty. For rank-deficient designs, OLS is not an unbiased estimator of all coordinates without an identifiability restriction, so Equation (205) cannot simply be assumed.

M.2 THE EXACT POOLING FRONTIER SURVIVES ARBITRARY COVARIANCE

For this subsection suppose $W \succ 0$, so it defines an inner product $\langle u, v\rangle_W = u^\top W v$. A fixed sharing subspace $\mathcal{V} \subseteq \mathbb{R}^{JD}$ consists of all parameter arrays allowed by a tying rule. If the columns of a full-column-rank matrix $B_\mathcal{V}$ form a basis for that subspace, its weighted orthogonal projector is

$$P = B_\mathcal{V}(B_\mathcal{V}^\top W B_\mathcal{V})^{-1} B_\mathcal{V}^\top W. \tag{212}$$

For the zero subspace define $P = 0$. In either case $P^2 = P$, $P^\top W = WP$, and Pu is the closest admissible vector to u in the W-norm. It need not be symmetric in ordinary Euclidean coordinates. Let $Q = I - P$ and define

$$B_W = \|Q\theta\|_W^2, \quad V_0 = \operatorname{tr}(W\Sigma), \quad V_P = \operatorname{tr}(WP\Sigma P^\top), \quad C_W = \operatorname{tr}(WQ\Sigma Q^\top) \geq 0. \tag{213}$$

The bias B_W is the population cost of the sharing restriction, V_0 is the no-sharing estimation risk, and C_W is the noise variance removed by that restriction.

Theorem 39 (Covariance- and geometry-aware partial pooling)**.** *For $L_\lambda = I - \lambda Q$, $0 \leq \lambda \leq 1$, under Equation (205) and $W \succ 0$,*

$$\begin{aligned} V_0 &= V_P + C_W, \\ \mathcal{R}_W(L_\lambda) &= V_0 - 2C_W\lambda + (B_W + C_W)\lambda^2. \end{aligned} \tag{214}$$

Full projection improves expected risk over $L = I$ exactly when $B_W < C_W$. If $B_W + C_W > 0$, the optimal fixed coefficient is $\lambda^\star = C_W/(B_W + C_W)$; if both vanish, every coefficient has the same risk. No Gaussianity, isotropic covariance, or independence across role estimates is required.

Proof. Weighted orthogonality gives $P^\top WQ = 0$. Consequently the noise components $P\zeta$ and $Q\zeta$ are orthogonal in the evaluation geometry for every realization, even if they are statistically dependent. Therefore $V_0 = V_P + C_W$. The error decomposes as

$$L_\lambda Y - \theta = -\lambda Q\theta + P\zeta + (1-\lambda)Q\zeta.$$

The bias–noise cross term has expectation zero, while the two projected noise components are pointwise orthogonal. Thus the risk is $\lambda^2 B_W + V_P + (1-\lambda)^2 C_W$. Expansion and minimization prove the claims. □

This strengthens the scalar frontier. If every role estimate has exactly the same random error and the sharing subspace contains all common-role arrays, then $Q\zeta = 0$ and $C_W = 0$: averaging does not remove that common error. Full sharing then only adds bias. Conversely, covariance concentrated in directions excluded by the sharing subspace makes pooling especially useful.

For common-operator sharing in Equation (211), the fitted shared matrix is

$$\widehat{A}_{\text{test}} = \left(\sum_a \pi_a \widehat{A}_a T_a\right)\left(\sum_a \pi_a T_a\right)^{-1}, \tag{215}$$

provided all $T_a \succ 0$. This is the weighted projection estimator, not necessarily ordinary pooled training OLS. Pooled OLS weights roles by their training design Grams; Equation (215) weights them by the declared test geometry.

Corollary 40 (Exact value of enlarging a sharing architecture)**.** *Let $\mathcal{U} \subseteq \mathcal{V}$ be two fixed sharing subspaces, with weighted projectors $P_\mathcal{U}, P_\mathcal{V}$. Define $Q_{\text{new}} = P_\mathcal{V} - P_\mathcal{U}$. Then*

$$\mathcal{R}_W(P_\mathcal{U}) - \mathcal{R}_W(P_\mathcal{V}) = \|Q_{\text{new}}\theta\|_W^2 - \operatorname{tr}(WQ_{\text{new}}\Sigma Q_{\text{new}}^\top). \tag{216}$$

Thus an additional depth effect, role correction, or fixed adapter direction is useful exactly when its recovered signal exceeds its covariance-weighted estimation cost.

Proof. Because the subspaces are nested, Q_{new} projects onto the directions in $\mathcal{V}$ weighted-orthogonal to $\mathcal{U}$. Both bias and noise split into weighted-orthogonal components. Enlarging the subspace removes squared bias $\|Q_{\text{new}}\theta\|_W^2$ and adds variance $\operatorname{tr}(WQ_{\text{new}}\Sigma Q_{\text{new}}^\top)$. □

The earlier dimension penalty is recovered when $W = I/J$ and $\Sigma = vI$. In general, equal parameter counts do not imply equal estimation costs. Learned subspaces or ranks must still be selected independently or analyzed with a selection penalty.

M.3 Approximate training and mismatched role weights

Fix one group and the evaluation weight vector π, allowing zero entries provided the empirical losses are defined for every role used in training. Let ν be the normalized weight vector used by the training algorithm. Define

$$\widehat{R}_\pi(f) = \sum_a \pi_a \widehat{R}_a(f), \quad \widehat{R}_\nu(f) = \sum_a \nu_a \widehat{R}_a(f), \quad \mathrm{TV}(\pi, \nu) = \tfrac{1}{2} \sum_a |\pi_a - \nu_a|. \tag{217}$$

Assume $0 \le \ell_a \le 1$ and $\widehat{f} \in \mathcal{F}$ satisfies $\widehat{R}_\nu(\widehat{f}) \le \inf_{f \in \mathcal{F}} \widehat{R}_\nu(f) + \eta_{\mathrm{opt}}$, where $\eta_{\mathrm{opt}} \ge 0$ is an empirical optimization gap. The gap may be computationally unknown and is retained explicitly in the bound.

Proposition 41 (Approximate ERM with explicit weight mismatch)**.** *On any event with* $\sup_f |R_\pi(f) - \widehat{R}_\pi(f)| \le u_\pi$,

$$R_\pi(\widehat{f}) \le \inf_f R_\pi(f) + 2u_\pi + \eta_{\mathrm{opt}} + 2\,\mathrm{TV}(\pi, \nu). \tag{218}$$

The training weights ν may themselves be data dependent: the bound uses their realized distance from the fixed evaluation weights, while the uniform event is over the fixed class and fixed π.

Proof. For every f, bounded losses imply $|\widehat{R}_\pi(f) - \widehat{R}_\nu(f)| \le \mathrm{TV}(\pi, \nu)$. Move from $R_\pi(\widehat{f})$ to its empirical value, then to the training objective, apply approximate minimization, and reverse the two changes at a population-optimal candidate (or an arbitrarily close one). The two uniform deviations contribute $2u_\pi$ and the two weight changes contribute $2\,\mathrm{TV}(\pi, \nu)$. □

In Theorem 4, the shared certificate therefore acquires $\sum_e W_e[\eta_{\mathrm{opt},e} + 2\,\mathrm{TV}(\pi_e, \nu_e)]$. If separate fits have gaps η_a, their certificate acquires $\sum_a w_a \eta_a$. The corrected bound difference is the original PRG_H plus the separate correction minus the shared correction. Zero-evaluation-weight roles can influence training through ν and are now explicitly charged rather than discarded. For task losses unrelated to the event-lift loss, an optimization-gap or calibration bound remains necessary.

M.4 Transfer without identical distribution supports

A bounded density ratio is a strong sufficient condition: it fails if rollout visits states outside the probe distribution's support. Even on overlapping supports, two shifted Gaussians with equal variance generally have an unbounded density ratio. The following replacement needs no absolute continuity.

Use the conditional occurrence measures $Q^m_{j,q,a}$ from Equation (33). Let $s^m_{j,q,a} = Q^m_{j,q,a}(\mathcal{Z}_a)$ be the probability that occurrence q exists and has role a. When $s^m_{j,q,a} > 0$, let $\nu^m_{j,q,a} = Q^m_{j,q,a}/s^m_{j,q,a}$ be its conditional probe distribution. Define a discrepancy over the actual loss class by

$$\mathfrak{d}_a(\nu, P_a) = \sup_{f \in \mathcal{F}_a} \left| \mathbb{E}_{Z \sim \nu} \ell_a(f; Z) - \mathbb{E}_{Z \sim P_a} \ell_a(f; Z) \right|. \tag{219}$$

The class $\mathcal{F}_a$ is the applicable fixed expert class for role a. This discrepancy measures changes in task loss, not changes in arbitrary features; for losses in $[0, 1]$ it is at most one.

Proposition 42 (Additive occupancy transfer)**.** *Assume the call-replacement lemma applies and choose deterministic occurrence bounds $s^m_{j,q,a} \le \bar{s}_{j,q,a}$, valid for both comparators independently of expert-fitting data. Then Assumption 3 holds with*

$$\begin{aligned} \omega_{j,a} &= \sum_q \bar{s}_{j,q,a}, \\ \widetilde{\tau}^m_j &= \tau^m_j + \sum_{q,a} s^m_{j,q,a} \mathfrak{d}_a(\nu^m_{j,q,a}, P_a). \end{aligned} \tag{220}$$

Zero-mass occurrences contribute zero. If the probe space has metric $d_{\mathcal{Z}_a}$ and every $\ell_a(f; \cdot)$ is uniformly L_a^{probe}-Lipschitz, then, for distributions with finite first moments,

$$\mathfrak{d}_a(\nu, P_a) \le L_a^{\mathrm{probe}} W_1(\nu, P_a; d_{\mathcal{Z}_a}). \tag{221}$$

For bounded losses, the minimum of this expression and one is also valid.

Proof. For the fitted function, conditional on its fitting data,

$$\int \ell_a(\widehat{f}; z)\, dQ^m_{j,q,a}(z) = s^m_{j,q,a}\mathbb{E}_{\nu^m_{j,q,a}}\ell_a(\widehat{f}; Z) \le \bar{s}_{j,q,a}R_a(\widehat{f}) + s^m_{j,q,a}\mathfrak{d}_a(\nu^m_{j,q,a}, P_a).$$

Nonnegativity of R_a justifies replacing s by $\bar{s}$ only in its coefficient. Sum the call losses. For Equation (221), couple $Z \sim \nu$ and $Z' \sim P_a$. For every f the expectation difference is at most $L^{\mathrm{probe}}_a \mathbb{E} d_{\mathcal{Z}_a}(Z, Z')$. Minimize over couplings and take the supremum over f. □

The coupling argument is the standard mechanism behind Wasserstein transfer bounds (Redko et al., 2017); here it is applied to role-conditioned event probes and their occurrence masses. Small transport distance is useful only when the loss class has controlled smoothness. The additive shift term must be bounded or estimated independently, and its uncertainty reported.

M.5 Repeated experts create coherent, rather than fresh, errors

The off-diagonal terms in Equation (87) have a direct interpretation when a fitted parameter is reused. Consider the exact additive-error model

$$\xi_j = b_j + B_j\delta_{e_j} + \epsilon_j. \tag{222}$$

Here the expert-index sequence e_j and matrices M_j, B_j are deterministic. Each $\delta_e \in \mathbb{R}^{d_e}$ is a centered random fitting error drawn once per expert and reused at every occurrence of that expert, with covariance V_e. The $s \times d_{e_j}$ matrix B_j maps parameter error into a local state perturbation. The errors $\epsilon_j \in \mathbb{R}^s$ are centered with covariances R_j, pairwise uncorrelated across time and uncorrelated with all δ_e; distinct δ_e are uncorrelated. All second moments are finite. Define

$$K_e = \sum_{j: e_j = e} P_j B_j. \tag{223}$$

Thus K_e adds all uses of expert e after mapping them into terminal-state coordinates.

Proposition 43 (Exact covariance cost of repeated expert use)**.** *For the linear recursion in Proposition 13 and Equation (222),*

$$\mathbb{E}\|\mathbf{e}_H\|_2^2 = \left\|\sum_j P_j b_j\right\|_2^2 + \sum_e \mathrm{tr}(K_e V_e K_e^\top) + \sum_j \mathrm{tr}(P_j R_j P_j^\top). \tag{224}$$

Proof. Unroll the recursion and collect repeated occurrences of the same random vector: $\mathbf{e}_H = \sum_j P_j b_j + \sum_e K_e\delta_e + \sum_j P_j\epsilon_j$. The specified cross-covariances vanish between these collected components. Take their covariance and then its trace. □

For the scalar example $M_j = B_j = 1$, $b_j = \epsilon_j = 0$, one expert used H times with fitting-error variance v gives $\mathbb{E}e_H^2 = H^2 v$. Fresh independent errors with the same variance at each call instead give Hv. Oppositely signed propagated effects may cancel, so reuse does not always increase the covariance cost. The relevant quantity is K_e, not merely occurrence count.

Equation (222) is exact for additive residual perturbations. For a nonlinear expert it may be a first-order linearization; applying it then requires a separate remainder bound. Also, conditioning on a route sequence selected using δ_e generally changes the error means and covariances. The deterministic/exogenous-route specialization must not be justified by conditioning on arbitrary endogenous routes and retaining the original V_e.

M.6 Local stability regions with an explicit escape cost

Uniform control on every possible hidden state can be much stronger than needed. Let Ω be the event that the reference, learned, and required hybrid states remain in specified deterministic regions where the local recursion and a utility Lipschitz constant $L_{\mathcal{U}}$ are valid. Assume common initial states and

that terminal utility is globally bounded in an interval of width $R_{\mathcal{U}}$. Let A_j be the products of the local reference Lipschitz envelopes. Conditional on the fitted models, suppose

$$\mathbb{E}\left[\mathbf{1}_{\Omega}\sum_j A_j(a_j + r_j)\right] \le U_{\mathrm{loc}}, \qquad \mathbb{P}(\Omega^c) \le p_{\mathrm{exit}}.$$

Here U_{loc} controls only paths remaining inside the region, and p_{exit} is the probability of leaving it.

Proposition 44 (Localized utility certificate)**.** *Under these conditions,*

$$\mathbb{E}|\mathcal{U}(\widehat{S}_H) - \mathcal{U}(S_H)| \le L_{\mathcal{U}} U_{\mathrm{loc}} + R_{\mathcal{U}} p_{\mathrm{exit}}. \tag{225}$$

Proof. On Ω, apply the pathwise recursion and the local utility bound. On Ω^c, use the global utility range. Split the expectation into these two events. □

The indicator-restricted probe measures can be treated by the same nonnegative-loss aggregation argument as other sub-probability measures. No independence of Ω and local errors is required. A local analysis is informative only when the escape probability is also controlled.

M.7 Expected accuracy and actual improvement are separate claims

A Lipschitz state certificate does not directly bound discrete answer accuracy. Let the same readout $g : \mathcal{S} \to \mathbb{R}^K$ be L_g-Lipschitz from d to ℓ_∞, let $E = d(\widehat{S}_H, S_H)$, and let $\gamma \ge 0$ be the reference top-two logit gap. For any deterministic $t > 0$,

$$\mathbb{P}(\text{different predicted answers}) \le \mathbb{P}(\gamma \le 2L_g t) + \mathbb{P}(E > t). \tag{226}$$

This follows by splitting on $E \le t$ and applying the pathwise margin test. It needs no independence between the margin and state error. A moment bound gives $\mathbb{P}(E > t) \le \mathbb{E}E^r/t^r$ for any $r > 0$. A model-specific readout adds its common-state logit discrepancy to $L_g E$. This controls disagreement with the reference; correctness relative to ground truth additionally requires the reference's own accuracy.

One can instead test actual improvement directly. Fix K_{cand} fitted candidate systems and a separate comparator before observing an independent IID audit set of size N. All losses $\ell_m(Z)$ and $\ell_{\mathrm{sep}}(Z)$ lie in $[0, 1]$ and are measured on the same example Z. Define the paired differences and their means by

$$D_{i,m} = \ell_{\mathrm{sep}}(Z_i) - \ell_m(Z_i), \quad \widehat{\Delta}_m = \frac{1}{N}\sum_i D_{i,m}, \quad \Delta_m = \mathbb{E}D_{i,m}. \tag{227}$$

Positive Δ_m means lower true risk for candidate m under this audit distribution.

Proposition 45 (A held-out certificate of actual performance improvement)**.** *Conditional on fitted systems, with probability at least* $1 - \delta$*,*

$$\Delta_m \ge \widehat{\Delta}_m - \sqrt{\frac{2\log(K_{\mathrm{cand}}/\delta)}{N}} \qquad \textit{simultaneously for all } m. \tag{228}$$

Thus exceeding this threshold certifies improvement of a fitted system's true audit-distribution risk, not just improvement of an upper bound.

Proof. Each paired difference lies in $[-1, 1]$. One-sided Hoeffding gives $\mathbb{P}(\widehat{\Delta}_m - \Delta_m > t) \le \exp(-Nt^2/2)$. Set the probability to δ/K_{cand} and take a union bound. Pairing requires no independence between the two models' losses on one example; independence is across audit records. □

The same simultaneous event permits selection among this pre-fitted finite candidate family. Repeatedly generating new candidates in response to the audit results invalidates this guarantee without further selection control. For unbounded squared loss, the bounded Hoeffding radius cannot be used; a justified tail model or a separately declared bounded evaluation loss is needed. This audit and PRG_H answer different questions: the latter explains a mechanism under modeling assumptions, whereas the former tests actual performance on a specified population.

N TERMINAL-RISK FRONTIERS AND VALID ADAPTIVE SELECTION

Overview. We connect local sharing to a common terminal geometry, exhibit a reversal of local improvement, and characterize a restricted setting in which every local full-pooling gain survives. Independent selection and nonlinear approximation require separate guarantees. Fixed matrices may instead be conditioned on an independent pilot, provided the stated conditional moments hold; this does not permit treating a same-data selected rule as fixed. Section N.1 derives the terminal frontier and reversal, Section N.2 treats independent finite-family selection, Section N.3 retains the nonlinear remainder, and Section N.4 characterizes uniform preservation in its stated submodel.

N.1 THE TERMINAL GEOMETRY CAN CHANGE THE PREFERRED AMOUNT OF SHARING

Let $Y = \theta + \zeta \in \mathbb{R}^p$ collect unbiased estimates of p role coefficients, with fixed target θ, $\mathbb{E}\zeta = 0$, and $\mathbb{E}[\zeta\zeta^\top] = \Sigma \succeq 0$. A fixed linear sharing rule L estimates θ by LY. Suppose the terminal error of every compared rule is exactly

$$e_H(L) = G(LY - \theta), \qquad G \in \mathbb{R}^{s\times p}. \tag{229}$$

The same propagation matrix G is used for all comparators and is independent of ζ. For example, in Proposition 13, if each local injection is $\xi_j = H_j(LY - \theta)$ with fixed H_j, then $G = \sum_j P_j H_j$. This sums repeated uses before calculating their covariance. Define $W_H = G^\top G$; it is positive semidefinite and may be singular, so $\|x\|_{W_H}^2 = x^\top W_H x$ is a squared seminorm rather than necessarily a squared norm.

Theorem 46 (Exact sharing frontier in terminal geometry)**.** *Under Equation (229), for every fixed L,*

$$R_H(L) := \mathbb{E}\|e_H(L)\|_2^2 = \|(L-I)\theta\|_{W_H}^2 + \operatorname{tr}(W_H L\Sigma L^\top). \tag{230}$$

Fix a pooling matrix P independently of ζ, put $Q = I - P$, and let $L_\lambda = I - \lambda Q$ for $0 \le \lambda \le 1$. Neither Euclidean orthogonality nor orthogonality in W_H is presumed. Set

$$\begin{aligned} a_H &= \operatorname{tr}(W_H Q\Sigma), \\ b_H &= \|Q\theta\|_{W_H}^2 + \operatorname{tr}(W_H Q\Sigma Q^\top) \ge 0. \end{aligned} \tag{231}$$

Then

$$R_H(L_\lambda) = R_H(I) - 2a_H\lambda + b_H\lambda^2. \tag{232}$$

Full pooling strictly improves terminal risk if and only if $2a_H > b_H$. If $b_H > 0$, an optimal coefficient is

$$\lambda_H^\star = \min\{1, \max\{0, a_H/b_H\}\}. \tag{233}$$

If $b_H = 0$, necessarily $a_H = 0$ and all coefficients have the same risk.

Proof. Expand $LY - \theta = (L-I)\theta + L\zeta$. The centered cross term vanishes and $\mathbb{E}\|GL\zeta\|^2 = \operatorname{tr}(W_H L\Sigma L^\top)$, proving Equation (230) even when W_H is singular. Substitution of L_λ gives a deterministic bias $-\lambda GQ\theta$ and variance

$$\mathbb{E}\|G\zeta - \lambda GQ\zeta\|^2 = \mathbb{E}\|G\zeta\|^2 - 2\lambda\mathbb{E}\langle G\zeta, GQ\zeta\rangle + \lambda^2\mathbb{E}\|GQ\zeta\|^2.$$

The middle expectation equals a_H by the trace identity and symmetry of W_H, Σ. This proves the quadratic and the strict full-pooling comparison. Minimize the convex quadratic on $[0,1]$ when $b_H > 0$. When $b_H = 0$, both $GQ\theta = 0$ and $GQ\zeta = 0$ almost surely; hence the middle expectation is zero as well. □

Why the linear term need not be a nonnegative variance saving. If P is a projection and $P^\top W_H Q = 0$, then $a_H = \operatorname{tr}(W_H Q\Sigma Q^\top) \ge 0$: its two projected output components are pointwise orthogonal. General pooling need not satisfy this condition. For example, take $\theta = 0$, $P = \operatorname{diag}(1,0)$, $G = (1,1)$, and

$$\Sigma = \begin{pmatrix} 4 & -3/2 \\ -3/2 & 1 \end{pmatrix} \succ 0.$$

Then $a_H = -1/2$, $b_H = 1$, and $R_H(L_\lambda) = 2 + \lambda + \lambda^2$. Suppressing a component removes beneficial cancellation and hurts every positive sharing strength.

Proposition 47 (Local pooling improvement can reverse at the terminal state)**.** *There is a three-role linear model in which full pooling reduces average local mean squared error from* 1 *to* $5/6$ *but increases terminal mean squared error from* 6 *to* 9*. The reversal persists if local evaluation and pooling use weights proportional to the magnitudes of the terminal propagation coefficients.*

Proof. Let $Y_a = \theta_a + \varepsilon_a$ with independent centered errors of variance one and $\theta = (1/2, -1, 1/2)^\top$. Local separate risk is one. For uniform pooling, $\bar{\theta} = 0$, the average squared role discrepancy is $1/2$, and the pooled variance is $1/3$, giving $5/6$.

Take the reference events, from the common initial state $S_0 = 0$, to be

$$S_1 = S_0 + \theta_1, \qquad S_2 = -\tfrac{1}{2}S_1 + \theta_2, \qquad S_3 = -2S_2 + \theta_3. \tag{234}$$

Each learned model uses the same events with its fitted coefficients. Thus $G = (1, -2, 1)$, separate terminal risk is $\|G\|_2^2 = 6$, and $G\mathbf{1} = 0$, $G\theta = 3$. Full pooling uses $\bar{Y}\mathbf{1}$ and has the constant error $G(\bar{Y}\mathbf{1} - \theta) = -3$, hence risk nine. For partial uniform pooling,

$$R_H(L_\lambda) = 6(1-\lambda)^2 + 9\lambda^2 = 6 - 12\lambda + 15\lambda^2, \tag{235}$$

which is minimized at $\lambda = 2/5$ with risk $18/5$.

For the alternative weights $\pi = (1/4, 1/2, 1/4)$, use the weighted pool $\sum_a \pi_a Y_a$. Its weighted target mean is $-1/4$, weighted squared bias is $9/16$, and variance is $\sum_a \pi_a^2 = 3/8$. Weighted local risk is therefore $15/16 < 1$. Any common scalar pool is still annihilated by $G\mathbf{1} = 0$, so its terminal risk remains nine. □

Here the terminal-optimal partial rule substantially outperforms both endpoints.

N.2 Selection of finitely many groups, subspaces, or sharing strengths

Let $\mathscr{A}$ be a finite set of $K \geq 1$ fitted candidates, each fixed conditional on a training/pilot sigma-field $\mathscr{T}$. Candidates can have different partitions, subspaces, or sharing strengths. An independent validation sample has records indexed by $i = 1, \ldots, N$, independent conditional on $\mathscr{T}$. Losses of different candidates on the same record may be dependent. Fix nonnegative weights w_i summing to one, measurable with respect to $\mathscr{T}$ and thus fixed before validation, and define

$$\widehat{R}_m = \sum_i w_i \ell_{i,m}, \qquad R_m = \sum_i w_i \mathbb{E}[\ell_{i,m} \mid \mathscr{T}], \qquad m \in \mathscr{A}.$$

The risk is for this declared validation target; nonidentically distributed records do not define an unspecified IID population. The candidates must not be created adaptively from these same validation outcomes.

Theorem 48 (Independent finite-family selection with explicit loss tails)**.** *Assume the centered loss* $X_{i,m} = \ell_{i,m} - \mathbb{E}[\ell_{i,m} \mid \mathscr{T}]$ *obeys, conditional on* $\mathscr{T}$*,*

$$\log \mathbb{E}[e^{tX_{i,m}} \mid \mathscr{T}] \leq \frac{v_{i,m}t^2}{2(1 - b_{i,m}|t|)} \quad \textit{for } |t| < 1/b_{i,m}, \tag{236}$$

where $v_{i,m}, b_{i,m} \geq 0$ *are fixed conditional on* $\mathscr{T}$*; when* $b_{i,m} = 0$ *the domain is all real* t*. For* $\delta \in (0,1)$ *set*

$$\begin{aligned} x &= \log(2K/\delta), \qquad V_m = \sum_i w_i^2 v_{i,m}, \quad B_m = \max_i w_i b_{i,m}, \\ r_m &= \sqrt{2V_m x} + B_m x. \end{aligned} \tag{237}$$

With conditional probability at least $1 - \delta$*, simultaneously for every candidate,* $|\widehat{R}_m - R_m| \leq r_m$*. On this event the penalized selector* $\widehat{m} \in \arg\min_m(\widehat{R}_m + r_m)$ *satisfies*

$$R_{\widehat{m}} \leq \min_{m \in \mathscr{A}}(R_m + 2r_m). \tag{238}$$

An unpenalized empirical minimizer instead satisfies $R_{\widehat{m}} \leq \min_m R_m + 2\max_m r_m$*.*

Proof. Conditional independence and Equation (236) imply that the log moment generating function of $\sum_i w_i X_{i,m}$ is at most $V_m t^2/[2(1 - B_m|t|)]$. For $V_m > 0$ and the upper tail, use Chernoff's inequality at $t = \sqrt{2x/V_m}/(1+B_m\sqrt{2x/V_m})$ to get probability at most e^{-x} above $\sqrt{2V_m x}+B_m x$. Apply the same argument to the negative sum. If $V_m = 0$, all positive-weight centered losses have zero variance and the claim is immediate. A union bound over both tails of the K candidates gives the simultaneous event. On that event, for every m,

$$R_{\widehat{m}} \le \widehat{R}_{\widehat{m}} + r_{\widehat{m}} \le \widehat{R}_m + r_m \le R_m + 2r_m.$$

For the raw minimizer the same argument uses $r_{\widehat{m}} + r_m \le 2\max_m r_m$. □

The tail hypothesis is often called a two-sided sub-gamma condition: v controls the quadratic fluctuation scale and b the range over which the exponential moment is controlled. This tail condition is an additional assumption beyond covariance information. For bounded losses in $[0, 1]$, Hoeffding's lemma gives $v_{i,m} = 1/4$, $b_{i,m} = 0$, recovering Proposition 11. To implement the penalized selector, valid radii or deterministic upper bounds on them must be available; replacing unknown parameters by estimates from the same validation outcomes needs a new guarantee.

A sufficient condition for unbounded quadratic validation loss. Conditional on training and fixed validation inputs (include those inputs in the conditioning information), suppose $\ell_{i,m} = \|\mu_{i,m} + \epsilon_i\|_2^2$, where $\mu_{i,m}$ is fixed and the validation errors are independent $\epsilon_i \sim \mathcal{N}(0, \Omega_i)$. Errors may be shared across candidates at the same record. Then Equation (236) holds with

$$v_{i,m} = 2\,\mathrm{tr}(\Omega_i^2) + 4\mu_{i,m}^\top \Omega_i \mu_{i,m}, \qquad b_{i,m} = 2\|\Omega_i\|_{\mathrm{op}}. \tag{239}$$

Here $\|\Omega_i\|_{\mathrm{op}}$ is its largest eigenvalue. To verify this, diagonalize Ω_i with eigenvalues $\omega_j \ge 0$ and corresponding coordinates μ_j. The centered quadratic log moment generating function is

$$\sum_j \left[-\tfrac{1}{2}\log(1 - 2t\omega_j) - t\omega_j + \frac{2t^2\omega_j\mu_j^2}{1 - 2t\omega_j} \right].$$

For $|2t\omega_j| < 1$, the power series for the logarithm bounds the first two terms by $t^2\omega_j^2/(1 - 2|t|\|\Omega_i\|_{\mathrm{op}})$; the last term is bounded with the same denominator. Summing gives Equation (239). Null eigenvalues cause no difficulty because their centered contributions are zero. Known Ω_i and a bound $\|\mu_{i,m}\| \le M_{i,m}$ permit the conservative substitution $\mu_{i,m}^\top \Omega_i \mu_{i,m} \le \|\Omega_i\|_{\mathrm{op}} M_{i,m}^2$. Without such information these are theoretical, not automatically computable, radii.

Independent selection and refitting give a different exact identity. If a pilot selects a rule $L = L(\mathscr{T})$ and a fresh fitting error satisfies $\mathbb{E}[\zeta \mid \mathscr{T}] = 0$, $\mathrm{Cov}(\zeta \mid \mathscr{T}) = \Sigma$, then conditioning and averaging Equation (207) yields the average of the fixed-rule risks. The averaging uses fresh fitting errors after pilot selection. Independence of the final audit also remains necessary for Proposition 45.

N.3 Terminal risk with controlled nonlinear remainders

The following result converts a controlled first-order remainder into an interval for actual terminal risk.

Proposition 49 (Terminal risk intervals with a controlled remainder)**.** *Let the actual terminal error be $e = e_{\mathrm{lin}} + r$ under the same probability law, with square-integrable vectors, $R_{\mathrm{lin}} = \mathbb{E}\|e_{\mathrm{lin}}\|_2^2$, and $\mathbb{E}\|r\|_2^2 \le \varepsilon^2$. No independence is required. Then*

$$\left(\max\{0, \sqrt{R_{\mathrm{lin}}} - \varepsilon\}\right)^2 \le \mathbb{E}\|e\|_2^2 \le (\sqrt{R_{\mathrm{lin}}} + \varepsilon)^2. \tag{240}$$

In particular, a reuse model has strictly smaller actual risk than a separate model if the reuse upper endpoint is smaller than the separate lower endpoint, using valid remainder bounds for both.

Proof. Use the triangle and reverse-triangle inequalities for the L^2 norm of random vectors: $\|e\|_{L^2} \le \|e_{\mathrm{lin}}\|_{L^2} + \|r\|_{L^2}$ and $\|e\|_{L^2} \ge \|e_{\mathrm{lin}}\|_{L^2} - \|r\|_{L^2}$. Nonnegativity and squaring give the endpoints; strict endpoint separation gives the comparison. □

Two diagnostics against automatic extrapolation. For the scalar nonlinear map $e(\delta) = \delta^2$ at $\delta = 0$, the linearization is zero, but $\delta \sim \mathcal{N}(0, v)$ gives true squared error $\mathbb{E}\delta^4 = 3v^2$. Mere knowledge of the covariance does not bound a general fourth-order remainder. In an endogenous linear-looking map, take $\delta \sim \mathcal{N}(0,1)$ and $G(\delta) = \mathbf{1}_{\{|\delta|>1\}}$. If ϕ, Φ denote the standard normal density and CDF, respectively,

$$\mathbb{E}\big[(G(\delta)\delta)^2\big] = 2[\phi(1) + 1 - \Phi(1)] > 2[1 - \Phi(1)] = \mathbb{E}[G(\delta)^2]\mathbb{E}\delta^2.$$

The inequality follows by integration by parts over $|\delta| > 1$. Averaging a fixed-G formula while retaining unconditional noise covariance is therefore invalid when the route selected G from the same error. A valid analysis must use the resulting conditional means and covariances or a pathwise remainder bound. For nonlinear neural traces, specifying the expansion point, reached region, derivatives, noise moments, and exit contribution is still a separate modeling and estimation task.

Section O now supplies explicit joint-derivative remainder envelopes and conditional route moments for specified models.

N.4 A COROLLARY: WHEN EVERY LOCAL FULL-POOLING GAIN SURVIVES

The reversal in Proposition 47 raises a uniform question: which fixed propagation geometries preserve a strict local full-pooling improvement for *every* possible target? The following is a structural corollary of Theorem 46, proved by an elementary spectral argument. The role of loss geometry in shrinkage is established in, for example, Hansen (2016, Section 5). Matrix-risk ordering, which controls all linear combinations, is also distinct from merely lowering its trace (Matsuda & Strawderman, 2022, Section 1).

Let $Y = \theta + \zeta \in \mathbb{R}^p$, with $\mathbb{E}\zeta = 0$ and $\mathrm{Cov}(\zeta) = qI_p$, $q > 0$. Let $P = P^\top = P^2$ be a fixed Euclidean orthogonal pooling projection, $Q = I_p - P$, and $k = \mathrm{rank}(Q) > 0$. The separate and fully pooled estimates are Y and PY. Let the same fixed matrix $G \in \mathbb{R}^{s\times p}$ propagate either estimate's error, as in Equation (229). Define the pooled-minus-separate risk differences

$$\Delta_{\mathrm{loc}}(\theta) = \mathbb{E}\|PY - \theta\|^2 - \mathbb{E}\|Y - \theta\|^2,$$
$$\Delta_H(\theta) = \mathbb{E}\|G(PY - \theta)\|^2 - \mathbb{E}\|G(Y - \theta)\|^2, \qquad A = QG^\top GQ.$$

These are unnormalized risks; dividing local risk by a positive role count leaves its sign unchanged. The matrix A is positive semidefinite, vanishes on $\mathrm{range}(P)$, and describes amplification restricted to the contrast space $\mathrm{range}(Q)$. Only the stated second moments, not Gaussianity, are required.

Corollary of Theorem 46. For these fixed P, G, q,

$$[\,\forall\theta \in \mathbb{R}^p : \quad \Delta_{\mathrm{loc}}(\theta) < 0 \implies \Delta_H(\theta) < 0\,] \quad \iff \quad A = cQ \quad \text{for some } c > 0.$$

Equivalently, G must be a positive scaled isometry on the discarded contrast space: $\|Gv\|^2 = c\|v\|^2$ for every $v \in \mathrm{range}(Q)$. This defines the term *scaled isometry*; it imposes no such requirement on all of $\mathbb{R}^p$.

Proof. Since $P^2 = P = P^\top$ and the noise is isotropic, the fixed-rule risk identities give

$$\Delta_{\mathrm{loc}}(\theta) = \|Q\theta\|^2 - qk, \qquad \Delta_H(\theta) = \theta^\top A\theta - q\,\mathrm{tr}(A).$$

For the second equality, $\mathrm{tr}(G^\top GP) - \mathrm{tr}(G^\top G) = -\mathrm{tr}(G^\top GQ) = -\mathrm{tr}(A)$. Thus no additional assumption $P^\top G^\top GQ = 0$ is needed. If $A = cQ$ with $c > 0$, then $\Delta_H = c\Delta_{\mathrm{loc}}$ for every target, proving sufficiency.

For necessity, if $A = 0$, the target $\theta = 0$ has $\Delta_{\mathrm{loc}} = -qk < 0$ but $\Delta_H = 0$, contradicting the strict implication. Otherwise let $\lambda_{\max} > 0$ be the largest eigenvalue of A restricted to its k-dimensional contrast space. If this restriction is not a scalar multiple of the identity, its largest eigenvalue exceeds the average:

$$\lambda_{\max} > \frac{\mathrm{tr}(A)}{k}.$$

Choose a unit eigenvector $v \in \mathrm{range}(Q)$ for $\lambda_{\max}$ and a scalar t satisfying

$$\frac{q\,\mathrm{tr}(A)}{\lambda_{\max}} < t < qk.$$

The interval is nonempty. At $\theta = \sqrt{t}\, v$, the local difference is $t - qk < 0$, whereas the terminal difference is $t\lambda_{\max} - q\,\mathrm{tr}(A) > 0$. This strict reversal contradicts the uniform implication. Hence the restriction must be cI_k with $c > 0$, equivalent to $A = cQ$ on the full space. □

Consequences of the all-target condition. The corollary generalizes the isolated reversal to every nonscalar contrast geometry and identifies exactly when no target can reverse a local gain. At $k = 1$, any nonzero restriction is automatically scalar. The cases $k = 0$ and $q = 0$ are excluded because no strict local full-pooling gain is possible; the implication would then be vacuous. The case $A = 0$ preserves only equality of terminal risks, not strict improvement. A single target may still benefit when $A \neq cQ$, so this is not a necessary condition for one particular model to benefit. It compares full projection with separate estimation, not all possible partial-pooling coefficients or same-data adaptive rules. An error-dependent G, a different G for each comparator, anisotropic noise, or nonlinear/naturally rerouted execution requires its own analysis.

O Explicit Nonlinear, Endogenous-Route, and Adapter Specializations

Overview. We make three interfaces explicit: how derivatives control a nonlinear remainder, how route selection changes error moments, and which functional corrections a parameterization can realize. These resolve the conditional modeling questions in Propositions 37 and 49 and Theorem 46. Every specialization fixes its reference before taking expectations. Calculations use real arithmetic; floating-point error, a different architecture, or a changed input distribution needs its own accounting. Sections O.1–O.2 bound nonlinear remainders and block derivatives; Sections O.3–O.4 analyze route-selected moments; Sections O.5–O.6 distinguish adapter realizability from its initialization-dependent tangent image.

O.1 A finite-horizon remainder from joint state–parameter derivatives

Use Euclidean norms on states and parameters and the Frobenius norm when a matrix is viewed as a parameter vector. Let $d \in \mathbb{R}^p$ denote one parameter perturbation, reused at every occurrence of the relevant parameters; it is not redrawn at successive events. Conditional on a fixed input and reference parameters, define

$$s_{j+1} = F_j(s_j, 0), \qquad \widehat{s}_{j+1} = F_j(\widehat{s}_j, d), \qquad \widehat{s}_0 = s_0.$$

Here F_j is a twice continuously differentiable event, or a smooth extension of one specified hard-routing branch. A branch extension evaluates the chosen expert and its prescribed smooth gain even at states where it would not win. Actual hard-routed execution equals this extension only on a separately certified route-agreement event.

For $M_j = D_s F_j(s_j, 0)$ and $B_j = D_d F_j(s_j, 0)$ define the linear error

$$z_0 = 0, \quad z_{j+1} = M_j z_j + B_j d, \quad G_0 = 0, \quad G_{j+1} = M_j G_j + B_j, \qquad z_j = G_j d. \tag{241}$$

The symbol $D^2 F_j$ is a bilinear second derivative, with norm $\sup_{\|v\|,\|w\|\le 1} \|D^2 F_j[v, w]\|$, on the product space of state and parameter directions.

Theorem 50 (Localized finite-horizon Taylor remainder)**.** *Fix deterministic constants $L_j, K_j, h_j \ge 0$ that bound, respectively, $\|D_s F_j\|$, $\|D_d F_j\|$, and $\|D^2 F_j\|$ on specified regions. Let Ω be an event on which actual execution uses the specified branches and the segment from $(s_j, 0)$ to $(\widehat{s}_j, d)$ lies in these regions for every j. Define deterministic sequences*

$$c_0 = b_0 = 0, \qquad c_{j+1} = L_j c_j + K_j, \qquad b_{j+1} = \|M_j\|_2 b_j + \tfrac{1}{2} h_j (c_j^2 + 1). \tag{242}$$

Then, on Ω, for every event,

$$\|\widehat{s}_j - s_j\|_2 \le c_j \|d\|_2, \qquad \|\widehat{s}_j - s_j - z_j\|_2 \le b_j \|d\|_2^2. \tag{243}$$

Write $e = \widehat{s}_H - s_H$ and $r = e - G_H d$ under the full probability law. If $q > 2$, $\mathbb{P}(\Omega^c) \le p_{\mathrm{exit}}$, and $\|e\|_{L^q} + \|G_H d\|_{L^q} \le T_q < \infty$, then

$$\mathbb{E}\|r\|_2^2 \le b_H^2 \mathbb{E}[\|d\|_2^4 \mathbf{1}_\Omega] + T_q^2 p_{\mathrm{exit}}^{1-2/q}. \tag{244}$$

Thus the square root of this right-hand side is a valid ε in Proposition 49. If $\mathbb{P}(\Omega) = 1$, only a finite fourth moment is needed and the exit term is absent.

Proof. On the segment in the statement, the integral first-derivative formula gives $\|\widehat{s}_{j+1} - s_{j+1}\| \le L_j\|\widehat{s}_j - s_j\| + K_j\|d\|$. Induction proves the first bound. The integral second-order Taylor formula at $(s_j, 0)$ has a remainder t_j bounded by $\frac{1}{2}h_j(\|\widehat{s}_j - s_j\|^2 + \|d\|^2)$. Subtracting Equation (241) gives $r_{j+1} = M_j r_j + t_j$. The first bound and induction give the stated b_j. For the full-law bound, split $\mathbb{E}\|r\|^2$ over Ω and its complement. Hölder's inequality gives $\mathbb{E}[\|r\|^2\mathbf{1}_{\Omega^c}] \le \|r\|_{L^q}^2\mathbb{P}(\Omega^c)^{1-2/q}$; the L^q triangle inequality bounds $\|r\|_{L^q}$ by T_q. No independence between d, Ω, e, and r is used. □

Moments, small-noise scaling, and an obstruction. If the hypotheses hold globally and $d \sim \mathcal{N}(\mu, \Sigma)$, then

$$\mathbb{E}\|d\|^4 = (\|\mu\|^2 + \operatorname{tr}\Sigma)^2 + 2\operatorname{tr}(\Sigma^2) + 4\mu^\top\Sigma\mu. \tag{245}$$

To verify it, write $d = \mu + Z$, diagonalize Σ, use independent centered normal coordinates with fourth moments $3\sigma_i^4$, and note that odd terms vanish. In particular, for $d = \sigma Z$ and constants b_H independent of σ, the remainder norm is $O(\sigma^2)$ if $\mathbb{E}\|Z\|^4 < \infty$. For centered Z, the linear risk is $\sigma^2\operatorname{tr}(G_H\operatorname{Cov}(Z)G_H^\top)$; an $o(1)$ relative-risk error additionally requires this coefficient to be strictly positive. A zero linear risk provides no relative guarantee. Covariance alone cannot replace the fourth moment: let d_N equal $\sqrt{N}, -\sqrt{N}, 0$ with probabilities $1/(2N), 1/(2N), 1 - 1/N$. Then $\mathbb{E}d_N^2 = 1$ but $\mathbb{E}d_N^4 = N$. The map $F(d) = d^2$ has constant Hessian two and unbounded squared remainder along this family. A small escape probability alone is likewise insufficient for unbounded state loss: an error of size p^{-1} on an event of probability p has risk p^{-1}. A bounded *utility* does not bound this state-space remainder.

From one prefix to an input population. The statement can be applied conditional on a random input X without assuming that the parameter error is independent of X. If its conditional premises hold with measurable constants and a finite right-hand side, then

$$\mathbb{E}\|r\|^2 \le \mathbb{E}_X\left[b_H(X)^2\mathbb{E}[\|d\|^4\mathbf{1}_\Omega \mid X] + T_q(X)^2 p_{\text{exit}}(X)^{1-2/q}\right]. \tag{246}$$

This is the tower property applied to Equation (244). The required integrability can fail if small margins or large derivative constants occur on the input population.

O.2 Explicit constants for the normalized gated blocks in the preparation code

We now specify a concrete specialization. Freeze the weights of attention, normalization, router, the up/gate matrices, and any depth correction (not their activations); perturb only expert *down* matrices. This matches the mathematical form of the corresponding subblock in `experiments/transformer_core.py`. Joint changes to the router/up/gate/normalizer or finite-precision execution require new terms. For one token let

$$\begin{aligned} N(x) &= \gamma \odot \frac{\Pi x}{(\|\Pi x\|^2/d_x + \epsilon_N)^{1/2}} + \beta, \quad \Pi = I - \mathbf{1}\mathbf{1}^\top/d_x, \quad \epsilon_N > 0, \\ h(u) &= \mathrm{SiLU}(Vu) \odot (Uu), \qquad g_e(u) = \mathrm{softmax}(Ru)_e, \\ F_e(x, \Delta) &= x + (D_e + \Delta)\,[g_e(N(x))h(N(x))] + AN(x). \end{aligned} \tag{247}$$

Here d_x is state width, γ, β are fixed LayerNorm scale and bias, U, V are expert up and gate matrices, D_e is its down matrix, R the router matrix, and A an optional fixed depth-correction matrix (zero if absent). The entrywise product is $\odot$, and $\mathrm{SiLU}(t) = t/(1 + e^{-t})$. The gain is the selected softmax probability, not a unit gate; support agreement does not freeze it. A deliberately frozen gain in $[0, 1]$ is also allowed by setting its derivatives below to zero.

Proposition 51 (Explicit smooth-branch derivative envelopes)**.** *Put $g = \|\gamma\|_\infty$, and define*

$$\begin{aligned} R_N &= g\sqrt{d_x} + \|\beta\|_2, \quad L_N = g/\sqrt{\epsilon_N}, \qquad H_N = 6g/(\sqrt{d_x}\epsilon_N), \\ a &= \|U\|_2, \qquad v = \|V\|_2, \qquad T = vR_N, \\ s_1 &= 1 + T/4, \qquad s_2 = 1/2 + T/4. \end{aligned} \tag{248}$$

Valid bounds for h and its first two derivatives on $\|u\| \le R_N$ are

$$H_0 = avR_N^2, \qquad H_1 = avR_N(s_1 + 1), \qquad H_2 = aR_N s_2 v^2 + 2s_1 av. \tag{249}$$

Let $r_R = \|R\|_2$, $g_1 = 2r_R L_N$, and $g_2 = 6r_R^2 L_N^2 + 2r_R H_N$. For $k_e(x) = g_e(N(x))h(N(x))$, take

$$\begin{aligned} k_0 &= H_0, \qquad k_1 = H_1 L_N + g_1 H_0, \\ k_2 &= H_2 L_N^2 + H_1 H_N + 2g_1 H_1 L_N + g_2 H_0. \end{aligned} \tag{250}$$

Uniformly for all states and $\|\Delta\|_F \le t$, the constants

$$\begin{aligned} L &= 1 + (\|D_e\|_2 + t)k_1 + \|A\|_2 L_N, \qquad K = k_0, \\ h_F &= (\|D_e\|_2 + t)k_2 + \|A\|_2 H_N + 2k_1 \end{aligned} \tag{251}$$

bound $\|D_x F_e\|$, $\|D_\Delta F_e\|$, and its joint second derivative. For n tokens with rowwise normalization and a fixed support per row, use Frobenius state norms, maximize these per-expert constants over the selected bank, and replace K by $\sqrt{n}\max_e k_{0,e}$. No extra factor n is needed in L or h_F, even when the down-matrix perturbations are shared across tokens, provided their stacked Frobenius norm is at most t.

Proof. Write $z = \Pi x$ and $q = \|z\|^2/d_x + \epsilon_N$. Apart from the fixed diagonal scale, the derivative of $z/\sqrt{q}$ is $q^{-1/2}\Pi - q^{-3/2} z z^\top / d_x$, whose norm is at most $\epsilon_N^{-1/2}$. Its second derivative in directions u, w is

$$-\frac{\Pi u\langle z, \Pi w\rangle + \Pi w\langle z, \Pi u\rangle + z\langle \Pi u, \Pi w\rangle}{d_x q^{3/2}} + \frac{3z\langle z, \Pi u\rangle\langle z, \Pi w\rangle}{d_x^2 q^{5/2}}.$$

Since $\|z\| \le \sqrt{d_x q}$, its norm for unit directions is at most $6/(\sqrt{d_x}\epsilon_N)$. This also proves the stated bound on N. If $p(t) = (1+e^{-t})^{-1}$, then $\mathrm{SiLU}' = p + tp(1-p)$ and $\mathrm{SiLU}'' = 2p(1-p) + tp(1-p)(1-2p)$. The inequalities $p(1-p) \le 1/4$ and $|t| \le T$ give s_1, s_2. Twice applying the product rule to h gives Equation (249). For a softmax coordinate, the logit gradient is $p_e(e_e - p)$ and its Hessian is $p_e[(e_e - p)(e_e - p)^\top - (\mathrm{diag}(p) - pp^\top)]$. Their norms are bounded by two and six, respectively. Composition with RN and the product rule prove Equation (250). Finally $D_\Delta^2 F_e = 0$, each state–parameter cross derivative has norm at most k_1, and the state Hessian is bounded by $(\|D_e\| + t)k_2 + \|A\|H_N$. Sum these three terms on the joint product space. For rowwise maps, the sum of products of squared row-direction norms is at most the product of their sums. The same inequality applies when a repeated parameter direction is selected from the stacked perturbation. The parameter first derivative has at most n rows of norm $k_0\|\Delta\|_F$, giving only the stated $\sqrt{n}$ factor in K. □

Closing the surrounding fixed-prefix attention computation. We complete the fixed-prefix Transformer analysis with an explicit, conservative bound for attention in the preparation architecture. Fix prefix length n, head width d_h, all weights, and a fixed causal mask with at least one allowed key per row. Softmax is taken only over allowed logits. With normalized input Z satisfying $\|Z\|_F \le Z_0 = \sqrt{n}R_N$, one attention head is

$$A_h(Z) = P_h(Z)\, Z W_{v,h}^\top, \qquad P_h(Z) = \mathrm{softmax}_{\mathrm{rows}}((ZW_{q,h}^\top)(ZW_{k,h}^\top)^\top/\sqrt{d_h}).$$

Set $q_h = \|W_{q,h}\|_2$, $k_h = \|W_{k,h}\|_2$, $v_h = \|W_{v,h}\|_2$, and

$$\begin{aligned} s_{1,h} &= 2q_h k_h Z_0/\sqrt{d_h}, & s_{2,h} &= 2q_h k_h/\sqrt{d_h}, \\ p_{1,h} &= 2s_{1,h}, & p_{2,h} &= 6s_{1,h}^2 + 2s_{2,h}, \\ a_{1,h} &= p_{1,h} v_h Z_0 + \sqrt{n}\, v_h, & a_{2,h} &= p_{2,h} v_h Z_0 + 2p_{1,h} v_h. \end{aligned} \tag{252}$$

For the concatenated heads followed by output matrix W_o, take $a_i = \|W_o\|_2(\sum_h a_{i,h}^2)^{1/2}$, $i = 1, 2$. An attention-plus-residual event $X \mapsto X + A(N(X))$ therefore has

$$L_{\mathrm{att}} = 1 + a_1 L_N, \qquad K_{\mathrm{att}} = 0, \qquad h_{\mathrm{att}} = a_2 L_N^2 + a_1 H_N. \tag{253}$$

For completeness, these bounds follow by differentiating the bilinear score map: its first and second derivatives have norms at most $s_{1,h}, s_{2,h}$. The row-softmax first and second derivative norms are at most two and six. For the latter, its bilinear form is $p \odot [(u - \mathbb{E}_p u)(w - \mathbb{E}_p w) - \mathrm{Cov}_p(u, w)]$; its Euclidean norm is at most its ℓ_1 norm, which is at most $5\|u\|_2\|w\|_2 \le 6\|u\|_2\|w\|_2$. Combine the rows by the squared-row inequality used above, and use $\|P_h\|_2 \le \|P_h\|_F \le \sqrt{n}$. Two product rules give $a_{1,h}, a_{2,h}$; concatenation and composition with LayerNorm give Equation (253). The terminal LayerNorm followed by a fixed linear head similarly has $L = \|W_{\mathrm{head}}\|_2 L_N$, $K = 0$, $h = \|W_{\mathrm{head}}\|_2 H_N$. Alternating these events with Equation (251) covers a full *fixed-prefix forward pass* with the specified down-weight perturbations.

A noncircular route and region check. Choose a perturbation radius t first, compute the uniform constants above and c_j by Equation (242), and evaluate reference router margins before examining perturbed executions. At each FFN event, let γ_j be the minimum reference top-one logit gap over its tokens. The sufficient tests

$$\|d\| \le t, \qquad \gamma_j > 2\|R_j\|_2 L_{N,j} c_j t \quad \text{for every FFN event} \tag{254}$$

ensure route agreement by induction and Lemma 27. The derivative bounds are global in the state, so no additional state-region escape occurs in this particular normalized architecture. The test controls support only; the varying gain has already been included in k_1, k_2. If the tests hold and d is bounded by t almost surely, set $p_{\text{exit}} = 0$. For an unbounded perturbation law, use at least $\mathbb{P}(\|d\| > t)$ in the exit accounting rather than discard it. For a fixed input with all strict margins, a positive sufficiently small t exists because the finitely many constants and recursions are continuous and finite at $t = 0$. The admissible radius can depend on the input. Small ϵ_N, large weights, long prefixes, or depth can make these conservative constants uninformative.

An explicit tail envelope for this specialization. There is also a global, route-independent growth bound for the down-only perturbation model at a fixed prefix. A normalized attention residual is uniformly bounded by its fixed weights and Z_0; a routed FFN residual is bounded by $\sqrt{n}\max_e[(\|D_e\|_2 + \|d\|)H_{0,e} + \|A\|_2 R_N]$. Summing residual magnitudes through finitely many blocks gives $\|\widehat{X} - X\|_F \le C_0 + C_1\|d\|$ with explicit nonnegative constants obtained from these displayed sums for the perturbed and reference passes. If the endpoint is the final normalized logits, the sharper bound $\|e\|_F \le 2\|W_{\text{head}}\|_2\sqrt{n}R_{N,\text{final}}$ is available. Consequently a finite q-th parameter moment supplies a finite T_q in Equation (244); for Gaussian perturbations one may use $q = 4$ and Equation (245). The actual weights, margins, covariance, and target prefix distribution are still needed to obtain a numerical certificate for a trained model. Theorem 59 and Proposition 60 distinguish current-gain support replay from the preparation code's frozen-gain intervention.

O.3 Conditional moments under endogenous routing

Let $\xi \in \mathbb{R}^p$ be a square-integrable error vector containing all reused parameter or local-injection errors. Let $\mathscr{R}$ be a sigma-field that makes the actual coefficient matrix G and offset b measurable; both may be selected using ξ. It may have to contain more than the expert indices: within-support gains, states, and Jacobians can still vary at a fixed index. Assume $b + G\xi$ and the conditional-moment expressions below are integrable at the indicated squared orders. Finite route families with bounded G, b and $\xi \in L^2$ are sufficient.

Theorem 52 (Exact risk with route-selected conditional moments)**.** *For $\mu_{\mathscr{R}} = \mathbb{E}[\xi \mid \mathscr{R}]$ and $\Sigma_{\mathscr{R}} = \mathrm{Cov}(\xi \mid \mathscr{R})$, an exactly route-affine error $e = b + G\xi$ obeys*

$$\mathbb{E}\|e\|_2^2 = \mathbb{E}\left[\|b + G\mu_{\mathscr{R}}\|_2^2 + \mathrm{tr}(G\Sigma_{\mathscr{R}}G^\top)\right]. \tag{255}$$

For a finite route variable T, with constant G_t, b_t on each positive-mass cell, this is the sum of these conditional risks weighted by $p_t = \mathbb{P}(T = t)$. Neither the conditional means nor the conditional cross-covariances may be replaced by their unconditional values without an additional argument.

A quadratic sharing frontier remains possible under one fixed endogenous route map. *Specifically, let Q and θ be deterministic, and fix the map $T = T(\xi)$ and cell matrices G_t independently of λ. Suppose all compared errors are exactly $e_\lambda = G_T[\xi - \lambda Q(\theta + \xi)]$. Then*

$$R(\lambda) = R(0) - 2a\lambda + b_2\lambda^2, \quad a = \mathbb{E}\langle G_T\xi, G_T Q(\theta + \xi)\rangle, \quad b_2 = \mathbb{E}\|G_T Q(\theta + \xi)\|^2 \ge 0. \tag{256}$$

With $W_t = G_t^\top G_t$, $\mu_t = \mathbb{E}[\xi \mid T = t]$, $\Sigma_t = \mathrm{Cov}(\xi \mid T = t)$, the coefficients equal

$$\begin{aligned} a &= \sum_t p_t\{\mu_t^\top W_t Q(\theta + \mu_t) + \mathrm{tr}(W_t Q\Sigma_t)\}, \\ b_2 &= \sum_t p_t\{\|Q(\theta + \mu_t)\|_{W_t}^2 + \mathrm{tr}(W_t Q\Sigma_t Q^\top)\}. \end{aligned} \tag{257}$$

Full sharing improves this risk exactly when $2a > b_2$; the minimizer on $[0, 1]$ is $\min\{1, \max\{0, a/b_2\}\}$ if $b_2 > 0$. If $b_2 = 0$, then $a = 0$ and all coefficients have equal risk.

Proof. Write $\xi = \mu_{\mathscr{R}} + \eta$ with $\mathbb{E}[\eta \mid \mathscr{R}] = 0$. Conditional on $\mathscr{R}$, expand the squared norm; the cross term vanishes and the remaining quadratic noise term is $\mathrm{tr}(G\Sigma_{\mathscr{R}}G^\top)$. Take expectations. For the sharing statement set $X = G_T\xi$ and $Y = G_TQ(\theta + \xi)$. Expand $\mathbb{E}\|X - \lambda Y\|^2$ and then condition each coefficient on T; this gives Equation (257). Minimize the resulting quadratic. If $\mathbb{E}\|Y\|^2 = 0$, then $Y = 0$ almost surely and $\mathbb{E}\langle X, Y\rangle = 0$. □

This is a conditional second-moment identity. It also applies to fixed linear propagations within a route cell after stacking all local errors, so it retains cross-time dependence. If the cell map or coefficients change with λ, every candidate must use its own conditional law; Equation (256) is no longer one common quadratic. For a nonlinear route cell the identity applies to its linear part only, with the remainder treated by Theorem 50 and Proposition 49.

General fixed Gaussian route cells via their probabilities. There is a useful identity beyond one half-space. Let $\Sigma \succ 0$ be fixed, let $\mathcal{C}_t$ be a fixed Borel selection region, and write $p_t(\mu) = \Pr_{\xi\sim\mathcal{N}(\mu,\Sigma)}(\xi \in \mathcal{C}_t) > 0$. When differentiating, hold the region and covariance fixed. Then

$$\mu_t = \mu + \Sigma\nabla_\mu \log p_t(\mu), \qquad \Sigma_t = \Sigma + \Sigma\nabla_\mu^2 \log p_t(\mu)\Sigma. \tag{258}$$

To prove this, differentiate the Gaussian density inside its integral on $\mathcal{C}_t$. The first density derivative is $\Sigma^{-1}(\xi - \mu)$ times the density, and the second is $[\Sigma^{-1}(\xi - \mu)(\xi - \mu)^\top\Sigma^{-1} - \Sigma^{-1}]$ times that density. Divide the integrated expressions by p_t and subtract the outer product of the conditional mean displacement. Differentiation is justified by locally integrable polynomial-times-Gaussian envelopes even if the region's boundary is not smooth. For singular covariance, apply the identity in whitened coordinates on its fixed affine support, varying the mean only within that support; ambient mean derivatives need not exist across moving supports. This handles fixed intersections or nonlinear regions analytically in terms of p_t. Regions depending on the differentiated mean require boundary terms; regions or propagation coefficients changing across sharing candidates require separate calculations.

O.4 A COMPLETE TWO-EXPERT GAUSSIAN THRESHOLD CALCULATION

Proposition 53 (Half-space route probabilities, means, and covariances)**.** *Let $\xi \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^p$, allowing singular Σ, and select route $+$ when $a^\top\xi > \tau$, route $-$ otherwise. Assume first that $s^2 = a^\top\Sigma a > 0$, and put*

$$\kappa = (\tau - a^\top\mu)/s, \quad h = \Sigma a/s,$$
$$p_+ = 1 - \Phi(\kappa), \quad p_- = \Phi(\kappa), \quad m_+ = \phi(\kappa)/p_+, \quad m_- = \phi(\kappa)/p_-.$$

The functions ϕ, Φ are the standard normal density and CDF. For finite τ, both probabilities are positive and

$$\begin{aligned} \mu_+ &= \mu + hm_+, \quad \Sigma_+ = \Sigma + hh^\top(\kappa m_+ - m_+^2), \\ \mu_- &= \mu - hm_-, \quad \Sigma_- = \Sigma - hh^\top(\kappa m_- + m_-^2). \end{aligned} \tag{259}$$

For any fixed offsets $b_\pm$ and matrices $G_\pm$, these moments and Equation (255) give the exact risk of $e = b_T + G_T\xi$. If $s = 0$, the score is constant almost surely: use its single realized branch (with the stated tie convention), and the original μ, Σ on that branch; there is no division by s or conditioning on a zero-mass cell.

Proof. Let $Z = (a^\top\xi - a^\top\mu)/s$ and $\eta = \xi - \mu - hZ$. Then $Z \sim \mathcal{N}(0, 1)$ and η is independent of Z, including in the singular Gaussian case, because their cross-covariance is zero. Moreover $\mathrm{Cov}(\eta) = \Sigma - hh^\top \succeq 0$. Integration by parts using $\phi'(z) = -z\phi(z)$ gives

$$\int_\kappa^\infty z\phi(z)\,dz = \phi(\kappa), \qquad \int_\kappa^\infty z^2\phi(z)\,dz = p_+ + \kappa\phi(\kappa).$$

Subtract from the unconditional first and second moments to obtain the lower half-space formulas. Substitute in $\xi = \mu + hZ + \eta$ and subtract the conditional means' outer products to obtain Equation (259). The variance-zero score is deterministic and is handled directly. □

An equivalent form avoiding division by small route probabilities. The unnormalized moments of the upper cell are

$$\begin{aligned} u_+ &= \mathbb{E}[\xi \mathbf{1}_{T=+}] = p_+\mu + h\phi(\kappa), \\ V_+ &= \mathbb{E}[\xi\xi^\top \mathbf{1}_{T=+}] \\ &= p_+(\Sigma + \mu\mu^\top) + \phi(\kappa)(\mu h^\top + h\mu^\top + \kappa hh^\top). \end{aligned} \tag{260}$$

The lower cell has $u_- = p_-\mu - h\phi(\kappa)$ and the corresponding negative sign in the correction to $p_-(\Sigma + \mu\mu^\top)$. The exact risk is therefore

$$\sum_{t\in\{+,-\}} \left[p_t\|b_t\|^2 + 2b_t^\top G_t u_t + \operatorname{tr}(G_t V_t G_t^\top)\right]. \tag{261}$$

These formulas follow from the same two integrals; they avoid inverse small probabilities, although extreme-tail cancellation still requires numerically stable CDF/survival functions and appropriate precision. They specialize classical truncated-normal moment calculations (Manjunath & Wilhelm, 2021).

Why the mean and cross-covariance both matter. For $\xi \sim \mathcal{N}(0,1)$, $T = \operatorname{sign}(\xi)$, and $G_T = T$, the error is $e = |\xi|$. Its mean is $\sqrt{2/\pi}$, variance is $1 - 2/\pi$, and squared risk is one. Keeping only the conditional variance loses the conditional-mean contribution $2/\pi$. For an independent standard bivariate normal (ξ_1, ξ_2), select $+$ when $\xi_1 + \xi_2 > \sqrt{2}$. In the notation above, $\kappa = 1$ and $h = (1,1)^\top/\sqrt{2}$, so the off-diagonal conditional covariance is $(m_+ - m_+^2)/2 < 0$. It is zero before selection. Thus even independently fitted errors need not be independent or uncorrelated after routing.

Recomputing the route can destroy the quadratic frontier. Let $\xi \sim \mathcal{N}(0,1)$, $a_\lambda = 1 - \lambda$, and use two branches whose error is $e_\lambda = a_\lambda\xi$ when $|a_\lambda\xi| > \tau > 0$ and zero otherwise. For $0 \le \lambda < 1$,

$$R(\lambda) = 2a_\lambda^2\left[\frac{\tau}{a_\lambda}\phi\left(\frac{\tau}{a_\lambda}\right) + 1 - \Phi\left(\frac{\tau}{a_\lambda}\right)\right], \qquad R(1) = 0. \tag{262}$$

The same integral as above proves this formula. It is not a quadratic in λ: as $\lambda \uparrow 1$ it is smaller than every positive power of $1 - \lambda$, yet it is positive at zero. A nonzero quadratic cannot have that property. This is a controlled example of route-dependent propagation.

O.5 Which functional corrections can a low-rank parameter update realize?

LoRA represents a change to a specified frozen linear weight by a matrix product BA (Hu et al., 2021). Its *matrix rank* is at most the factor width. This differs from the dimension of the span of an arbitrary collection of functions in Proposition 37. To make the comparison precise, fix a feature map $\varphi(x) \in \mathbb{R}^k$, its input law, and its uncentered second moment $C = \mathbb{E}[\varphi(x)\varphi(x)^\top] \succeq 0$. Let $W_0 \in \mathbb{R}^{d\times k}$ be the fixed core and $W_a^\star$ a role's linear target on these features. Write $K_a = W_a^\star - W_0$. Only updates to this output weight are allowed; changing the feature map changes the problem.

Theorem 54 (Exact adapter approximation bias in feature and output geometry)**.** *Let $S \succeq 0$ be a fixed output-loss matrix and let r be an integer with $0 \le r \le \min(d,k)$. Define $T_a = S^{1/2}K_aC^{1/2}$. Then*

$$\inf_{\operatorname{rank}(\Delta_a)\le r} \mathbb{E}\|(K_a - \Delta_a)\varphi(x)\|_S^2 = \sum_{i>r}\sigma_i(T_a)^2. \tag{263}$$

Here σ_i are decreasing singular values, $S^{1/2}, C^{1/2}$ are the symmetric positive-semidefinite square roots, and $\|y\|_S^2 = y^\top Sy$. Both S and C may be singular. If $T_{a,r}$ is a truncated singular-value decomposition, a realizing update is

$$\Delta_a = (S^{1/2})^\dagger T_{a,r}(C^{1/2})^\dagger, \tag{264}$$

where $\dagger$ denotes the Moore–Penrose inverse (invert the nonzero singular values and leave zero ones zero). The optimal risk is zero exactly when $\operatorname{rank}(T_a) \le r$. For independently chosen role adapters, the weighted average optimum is the sum of Equation (263) with the declared role weights; this is an approximation result, not an estimation or optimization guarantee.

Proof. The objective equals $\|T_a - S^{1/2}\Delta_a C^{1/2}\|_F^2$. For any matrix Z of rank at most r, let P_Z project onto its column space. Then $\|T_a - Z\|_F^2 \geq \|(I - P_Z)T_a\|_F^2$. Diagonalizing $T_a T_a^\top$ shows $\mathrm{tr}(P_Z T_a T_a^\top) \leq \sum_{i \leq r} \sigma_i(T_a)^2$: in an eigenbasis the diagonal weights of a rank-at-most-r orthogonal projector lie in $[0, 1]$ and sum to at most r. This proves the lower bound. The left and right singular directions of T_a with nonzero singular values lie in the ranges of $S^{1/2}$ and $C^{1/2}$, respectively. Therefore the pseudoinverse lift has rank at most r and satisfies $S^{1/2}\Delta_a C^{1/2} = T_{a,r}$, attaining the bound. Every such rank-at-most-r update can be factored as BA by its singular-value decomposition, padding the factors with zeros when necessary. □

The covariance must be a *feature second moment*, not its centered covariance when features have nonzero mean. Null directions of C are unobserved under this law; they need not remain null after distribution shift. Likewise S null directions are invisible to this loss, not necessarily to another output metric or a downstream computation.

The infimum has no update-norm or factor-norm constraint. A pseudoinverse lift can have large norm on poorly excited feature directions. A bounded-radius adapter problem, including a radius needed for a downstream Taylor bound, therefore need not attain this unconstrained tail value. Theorem 57 and Proposition 58 give the radius-constrained specializations and count each distinct updated matrix once.

Targets outside the linear feature span. For an arbitrary square-integrable desired correction $d_\star(x)$, let $B = \mathbb{E}[d_\star \varphi^\top]$, $K = BC^\dagger$, and $u = d_\star - K\varphi$. A vector in the null space of C is orthogonal to φ almost surely, so $B = BC^\dagger C$ and $\mathbb{E}[u\varphi^\top] = 0$. Consequently, for any fixed $S \succeq 0$,

$$\inf_{\mathrm{rank}(\Delta) \leq r} \mathbb{E}\|d_\star - \Delta\varphi\|_S^2 = \mathbb{E}\|u\|_S^2 + \sum_{i>r} \sigma_i(S^{1/2} K C^{1/2})^2. \tag{265}$$

This follows by expanding the square, eliminating the cross term using the orthogonality just proved, and applying Theorem 54. The feature-misspecification residual is retained for targets outside the feature span.

A shared factor is a different architecture. With Euclidean output loss, role-specific feature second moments C_a, and one common column factor $U \in \mathbb{R}^{d \times r}$, $U^\top U = I_r$, write $\Delta_a = U Z_a$. For fixed U the best value is $\sum_a \pi_a \|(I - UU^\top)K_a C_a^{1/2}\|_F^2$, attained by $Z_a = U^\top K_a$. Optimizing U gives

$$\inf_{U, Z_a} \sum_a \pi_a \mathbb{E}\|(K_a - UZ_a)\varphi_a\|^2 = \sum_{i>r} \lambda_i \left(\sum_a \pi_a K_a C_a K_a^\top \right). \tag{266}$$

This follows by the same trace maximization as in the proof above. These eigenvectors are output directions; they are not the function-space eigenvectors in Proposition 37, and the constraint is stronger than separate rank-r updates. For example, two targets $\mathrm{diag}(1, 0)$ and $\mathrm{diag}(0, 1)$ under $C = I_2$ and zero core admit separate rank-one updates, but a common one-dimensional output factor has average residual $1/2$.

Two concrete insertion points. For the preparation code's `tied_depth` term, take $\varphi(x) = N(x)$ and $\Delta = \texttt{depth_up}\,\texttt{depth_down}$. It is a parallel residual correction, not a modification of an existing expert weight. For a LoRA update to the down weight of a frozen gated expert, take $\varphi(x) = \mathrm{SiLU}(VN(x)) \odot UN(x)$, multiplied by its fixed reference gain when that gain is included in the output loss. A state-dependent but frozen gain function can be absorbed into φ; a route-index indicator can likewise be included if the evaluation averages calls where that expert is not selected. The surrounding input map, gain function, and selection law must remain fixed across the compared adapters. These are exact *local* feature maps. An adapter may change the inputs or routes of later blocks, for which the preceding Taylor or conditional-route analysis is still required.

Proposition 55 (Two roles with a freely learned common core)**.** *For two targets $W_1^\star, W_2^\star$, a common feature second moment C, and weights $\pi_1, \pi_2 > 0$ with $\pi_1 + \pi_2 = 1$, and integer $r \geq 0$, suppose the common core W_0 is an arbitrary matrix rather than fixed. Under Euclidean output loss,*

$$\inf_{W_0,\, \mathrm{rank}(\Delta_i) \leq r} \sum_{i=1}^{2} \pi_i \|(W_i^\star - W_0 - \Delta_i)C^{1/2}\|_F^2 = \pi_1\pi_2 \sum_{j>2r} \sigma_j((W_1^\star - W_2^\star)C^{1/2})^2. \tag{267}$$

In particular, exact realization on the feature law is possible exactly when the displayed difference has rank at most $2r$.

Proof. For fixed updates, completing the square over the free core gives $\pi_1\pi_2\|(W_1^\star - W_2^\star - \Delta_1 + \Delta_2)C^{1/2}\|_F^2$. The difference of two rank-at-most-r matrices has rank at most $2r$, yielding the lower bound by Theorem 54. Conversely, split a best rank-at-most-$2r$ truncation Z of the whitened target difference into two groups of at most r singular terms, $Z = Z_1 + Z_2$. Take $\Delta_1 = Z_1(C^{1/2})^\dagger$ and $\Delta_2 = -Z_2(C^{1/2})^\dagger$. Their whitened difference is exactly Z; choose $W_0 = \pi_1(W_1^\star - \Delta_1) + \pi_2(W_2^\star - \Delta_2)$. This attains the lower bound, including singular C and $2r$ exceeding the matrix dimensions. □

For the earlier targets $\pm I_2$ with $C = I_2$, the fixed zero core leaves residual one for each rank-one adapter. A free core $W_0 = \mathrm{diag}(1, -1)$ instead allows $\Delta_+ = \mathrm{diag}(0, 2)$ and $\Delta_- = \mathrm{diag}(-2, 0)$, both rank one, and zero residual. Thus the fixed-core counterexample cannot be used as a lower bound for jointly learned cores. The two-role formula does not extend to an arbitrary number of roles by checking pairwise ranks: simultaneous realizability adds compatibility constraints. For example, take three targets $0, I_2, 2I_2$, $C = I_2$, and $r = 1$. Every pairwise difference has rank at most $2r$. Exact simultaneous realization would require $\det W_0 = 0$ and $\det(I_2 - W_0) = 0$, hence $\mathrm{tr}\, W_0 = 1$. But then $\det(2I_2 - W_0) = 4 - 2\,\mathrm{tr}\, W_0 + \det W_0 = 2$, a contradiction. This is an obstruction to exact realization, not a computed positive lower bound on the three-role infimum.

O.6 Initialization-dependent low-rank tangent images

Proposition 56 (Initialization-dependent tangent image and its missing term)**.** *For the feature map above with Euclidean output loss, parameterize an update by BA. At factors (B_0, A_0) its exact change is*

$$(B_0 + \dot{B})(A_0 + \dot{A}) - B_0A_0 = B_0\dot{A} + \dot{B}A_0 + \dot{B}\dot{A}. \tag{268}$$

Thus the first-order parameter-to-function image is $\{(B_0\dot{A} + \dot{B}A_0)\varphi\}$, not all rank-at-most-$r$ corrections. For $B_0 = 0$, put $V_0 = A_0C^{1/2}$ and let P_{V_0} be the orthogonal projector onto its row space. If $K\varphi$ is the desired functional change, then

$$\inf_{\dot{B}} \mathbb{E}\|(K - \dot{B}A_0)\varphi\|^2 = \|KC^{1/2}(I - P_{V_0})\|_F^2, \tag{269}$$

attained by $\dot{B} = KC^{1/2}V_0^\dagger$. The omitted term in Equation (268) obeys

$$\|\dot{B}\dot{A}\varphi\|_{L^2} \le \|C^{1/2}\|_2\|\dot{B}\|_F\|\dot{A}\|_F \le \tfrac{1}{2}\|C^{1/2}\|_2(\|\dot{B}\|_F^2 + \|\dot{A}\|_F^2). \tag{270}$$

Proof. Multiply the two factors and group linear and quadratic terms. When $B_0 = 0$, the matrix approximation problem is $\inf_{\dot{B}} \|KC^{1/2} - \dot{B}V_0\|_F^2$. Each row of $\dot{B}V_0$ lies in the row space of V_0, and every row projection is attained by the displayed pseudoinverse, proving the equality. Finally use submultiplicativity of Frobenius and operator norms on $\dot{B}\dot{A}C^{1/2}$, followed by $2ab \le a^2 + b^2$. □

At the preparation code's zero `depth_up` initialization, the initial functional tangent therefore only uses the row space of the initialized `depth_down` after feature whitening. It is generally narrower than the class obtained by training both factors. If both factors are zero, the first derivative is zero but nonzero finite updates remain possible through $\dot{B}\dot{A}$. More concretely, with $r = 1$, $C = I_2$, $A_0 = (1, 0)$, $B_0 = 0$, and target $K = \mathrm{diag}(0, 1)$, the tangent residual is one while a finite rank-one update using row factor $(0, 1)$ is exact. A tangent obstruction is not a global impossibility theorem.

A rigorous nonlinear parameter-to-function fallback. For a general insertion point, fix a Hilbert space of functions on a fixed probe law and suppose $\mathcal{F}(W)$ is twice differentiable on the parameter ball $\|\Delta\|_F \le t$, with a uniform second-derivative norm at most H. Let $J = D\mathcal{F}(W_0)$, let $d_\star$ be a target functional correction, and let $\mathcal{S}_r(t) = \{\Delta : \mathrm{rank}(\Delta) \le r, \|\Delta\|_F \le t\}$. Define the constrained image distance

$$b_r(t) = \inf_{\Delta \in \mathcal{S}_r(t)} \|d_\star - J\Delta\|_{\mathcal{H}}, \qquad \eta = Ht^2/2.$$

Uniform Taylor remainder and the triangle inequalities give

$$(\max\{0, b_r(t) - \eta\})^2 \le \inf_{\Delta \in \mathcal{S}_r(t)} \|d_\star - [\mathcal{F}(W_0 + \Delta) - \mathcal{F}(W_0)]\|_{\mathcal{H}}^2 \le (b_r(t) + \eta)^2. \tag{271}$$

Indeed, apply the pointwise norm inequalities for every feasible Δ and then take infima; an approximating sequence suffices if an infimum is not attained. The image $J(\mathcal{S}_r(t))$ is generally not a linear subspace, so projecting onto the full range of J only supplies a relaxation, not a realizability certificate. Smoothness into L^2 requires a uniform square-integrable derivative envelope on the declared input law; pointwise differentiability alone does not supply it. For hard routing one must first certify a common smooth branch or separately pay for route changes.

Applying the conditional bounds. A numerical nonlinear bound still needs a specified checkpoint, reference input population, perturbed weights, perturbation law, and meaningful margin/tail bounds. A multistep endogenous route generally conditions on an intersection of nonlinear decisions, rather than the single half-space solved here; selected indices alone do not fix gains or Jacobians. The radius extension in Section P fixes the local feature law and core, constrains the actual output-matrix update, and separates approximation from terminal performance.

P Rank-Constrained Updates and Precisely Defined Branch Replay

Overview. We address two limits of a local approximation argument: the update must fit a budget on actual parameter changes, and fixed-support execution must be distinguished from natural routing. The guarantees separately track matrix rank, total radius, support disagreement, and the extra cost of replaying recorded gains, for fixed features/cores and finite fixed-prefix computations. The matrix arguments extend classical low-rank least-squares geometry (Eckart & Young, 1936); the routing arguments are pathwise couplings, not assumptions of route independence. Sections P.1–P.2 handle one-matrix and total update budgets; Section P.3 compares natural execution with current-gain support replay, and Section P.4 accounts for freezing the gains as well.

P.1 One matrix: an exact isotropic radius frontier and semidefinite bounds

Fix a core $W_0 \in \mathbb{R}^{d\times k}$ and target $W^\star$, and set $K = W^\star - W_0$. The feature map φ, its probability law, and $C = \mathbb{E}[\varphi\varphi^\top] \succeq 0$ remain fixed as the update changes. For output geometry $S \succeq 0$, integer $0 \le r \le \min(d, k)$ and finite $t \ge 0$, define

$$\mathcal{E}_{r,t}(K; S, C) = \min_{\substack{\operatorname{rank}(\Delta)\le r \\ \|\Delta\|_F \le t}} \|S^{1/2}(K - \Delta)C^{1/2}\|_F^2. \tag{272}$$

This equals $\min \mathbb{E}\|(K - \Delta)\varphi\|_S^2$. The constraint is on the actual matrix update, not on a factor's norm or on a training gradient. The feasible set is nonempty and compact: rank at most r is closed (all $(r+1)$-minors vanish), and the Frobenius ball is closed and bounded. Hence the continuous objective attains its minimum, including for singular S, C. Write $(x)_+ = \max\{x, 0\}$ and let T_r denote a truncated SVD retaining at most r singular terms of T; $T_0 = 0$.

Theorem 57 (Rank and matrix-radius approximation)**.** *In isotropic geometry, put $q_r = \|K_r\|_F = (\sum_{i\le r} \sigma_i(K)^2)^{1/2}$. Then*

$$\mathcal{E}_{r,t}(K; I, I) = \sum_{i>r} \sigma_i(K)^2 + (q_r - t)_+^2. \tag{273}$$

For $q_r > 0$, an attaining update is $\Delta^\star = \min\{1, t/q_r\}K_r$; for $q_r = 0$, use $\Delta^\star = 0$.

For arbitrary positive-semidefinite S, C, set

$$\begin{gathered} T = S^{1/2}KC^{1/2}, \quad q_r = \|T_r\|_F, \quad E_r = \sum_{i>r} \sigma_i(T)^2, \quad \kappa = \|S^{1/2}\|_2\|C^{1/2}\|_2, \\ \Delta_0 = (S^{1/2})^\dagger T_r (C^{1/2})^\dagger, \qquad \alpha = \begin{cases} \min\{1, t/\|\Delta_0\|_F\}, & \Delta_0 \ne 0, \\ 1, & \Delta_0 = 0. \end{cases} \end{gathered} \tag{274}$$

Then

$$E_r + (q_r - \kappa t)_+^2 \;\le\; \mathcal{E}_{r,t}(K; S, C) \;\le\; E_r + (1-\alpha)^2 q_r^2. \tag{275}$$

The right-hand value is attained by the feasible candidate $\alpha\Delta_0$, but need not be the minimum. If $t \ge \|\Delta_0\|_F$, the two bounds equal E_r. These conventions cover $r = 0$, $t = 0$, a zero target, null feature/output geometries, and tied singular values without dividing by zero.

Proof. For any feasible Δ let P project onto its column space and put $u = \|\Delta\|_F$. Since $P\Delta = \Delta$, $\langle K, \Delta\rangle_F = \langle PK, \Delta\rangle_F \le u\|PK\|_F$. In an eigenbasis of $KK^\top$, the diagonal entries of P lie in $[0,1]$ and sum to at most r. It follows that $\|PK\|_F^2 = \mathrm{tr}(PKK^\top) \le q_r^2$. Consequently

$$\|K - \Delta\|_F^2 \ge \|K\|_F^2 + u^2 - 2uq_r = \sum_{i>r} \sigma_i(K)^2 + (u - q_r)^2.$$

Minimize over $0 \le u \le t$. A scaled K_r has the optimal norm and attains the inner-product bound; the zero case is immediate.

For the general lower bound, $Z = S^{1/2}\Delta C^{1/2}$ has rank at most r and norm at most κt. Minimizing over all such Z relaxes the actual image constraint, so the isotropic formula applied to T gives the left-hand side of Equation (275). Every nonzero left/right singular direction of T lies in the range of $S^{1/2}/C^{1/2}$, respectively. Thus $S^{1/2}\Delta_0 C^{1/2} = T_r$ and $\mathrm{rank}(\Delta_0) \le r$, also when these matrices are singular. Scaling by α makes the update feasible. Orthogonality of $T - T_r$ and T_r gives $\|T - \alpha T_r\|_F^2 = E_r + (1-\alpha)^2 q_r^2$. If $t \ge \|\Delta_0\|_F$, this equals E_r and $q_r = \|S^{1/2}\Delta_0 C^{1/2}\|_F \le \kappa t$, proving the final claim. □

A suboptimality example in semidefinite geometry. Take $S = I_2$, $C = \mathrm{diag}(10^{-2}, 1)$, $K = \mathrm{diag}(20, 1)$, and $r = t = 1$. Then $T = \mathrm{diag}(2, 1)$ and $\Delta_0 = \mathrm{diag}(20, 0)$. The scaled candidate $\alpha\Delta_0 = \mathrm{diag}(1, 0)$ has loss $(2-0.1)^2 + 1 = 4.61$. The feasible rank-one update $\mathrm{diag}(0, 1)$ has loss 4. Thus scaled transformed SVD can be strictly suboptimal under a *parameter*-norm budget. Replacing the transformed feasible image by the rank-constrained Frobenius ball of radius κt gives a relaxation. Imposing the unscaled radius t directly in transformed coordinates is not in general a valid relaxation or a feasible parameter construction. Poorly excited directions can make the inverse lift expensive.

For an arbitrary square-integrable desired correction, the fixed-feature orthogonal decomposition in Equation (265) still applies for every feasible Δ. Its feature-misspecification residual must be added to Equation (272).

P.2 A TOTAL RADIUS OVER DISTINCT PARAMETER MATRICES

Let $\Delta_1, \dots, \Delta_E$ be updates to *distinct* physical parameter matrices, and form d by concatenating their vectorized entries once each. Then

$$\|d\|_2^2 = \sum_{e=1}^{E} \|\Delta_e\|_F^2. \tag{276}$$

An expert reused at several tokens or depths contributes one summand, not a fresh matrix or independent random draw at each call. Conversely, giving each of E distinct matrices radius t permits total radius $\sqrt{E}\,t$, not t.

Proposition 58 (A common isotropic radius for distinct update blocks)**.** *Fix targets K_e, ranks r_e within their matrix dimensions, and positive weights w_e summing to one. Put $q_e = \|(K_e)_{r_e}\|_F$, $a_e = \sum_{i>r_e} \sigma_i(K_e)^2$, and $Q_0 = (\sum_e q_e^2)^{1/2}$. Then*

$$\min_{\substack{\mathrm{rank}(\Delta_e) \le r_e \;\forall e \\ \sum_e \|\Delta_e\|_F^2 \le t^2}} \sum_e w_e \|K_e - \Delta_e\|_F^2 = \sum_e w_e a_e + \sum_e w_e (q_e - u_e)^2, \tag{277}$$

where $u_e = q_e$ if $t \ge Q_0$, and $u_e = 0$ if $t = 0$. For $0 < t < Q_0$, there is a unique $\eta > 0$ satisfying

$$u_e = \frac{w_e q_e}{w_e + \eta}, \qquad \sum_e u_e^2 = t^2. \tag{278}$$

An optimizer is $\Delta_e = (u_e/q_e)(K_e)_{r_e}$ when $q_e > 0$ and zero otherwise. This is a deterministic approximation identity; it assumes no independence between parameter errors and proves no optimizer's convergence.

Proof. The inner-product inequality in Theorem 57 reduces each block to $a_e + (q_e - u_e)^2$ at prescribed norm u_e. An optimum need not use $u_e > q_e$: replacing it by q_e reduces both radius and objective. The problem is therefore the scalar quadratic minimization over $0 \le u_e \le q_e$, $\sum_e u_e^2 \le t^2$, and aligned truncated SVDs attain any such choice. The inactive and zero-radius cases follow directly. For $0 < t < Q_0$, the squared norm of $(w_e q_e/(w_e + \eta))_e$ is continuous and strictly decreasing from Q_0^2 to zero as η increases from zero to infinity. The resulting u minimizes $\sum_e w_e(q_e - u_e)^2 + \eta \sum_e u_e^2$ coordinatewise. For any feasible v, this minimization and $\sum_e u_e^2 = t^2 \ge \sum_e v_e^2$ imply $\sum_e w_e(q_e - u_e)^2 \le \sum_e w_e(q_e - v_e)^2$. This proves global optimality without an assumption about numerical search. □

Several roles sharing one actual update. Roles are not automatically distinct variables in Equation (277). For fixed tying groups $\mathcal{J}_e$, common isotropic geometry within each group, and positive role weights π_a summing to one, set $w_e = \sum_{a \in \mathcal{J}_e} \pi_a$ and $\bar{K}_e = w_e^{-1} \sum_{a \in \mathcal{J}_e} \pi_a K_a$. Completing the square gives

$$\sum_a \pi_a \|K_a - \Delta_{\alpha(a)}\|_F^2 = \sum_e \sum_{a \in \mathcal{J}_e} \pi_a \|K_a - \bar{K}_e\|_F^2 + \sum_e w_e \|\bar{K}_e - \Delta_e\|_F^2. \tag{279}$$

Apply Proposition 58 to the unique matrices and group centroids in the second term. The first term is an irreducible tying bias in this model. For example, targets $+1, -1$ with equal weights and one shared scalar update have loss $1 + \Delta^2 \ge 1$, even though separate updates of radius one fit each target exactly. Merely adding separate per-role optima would give an invalid claim of joint realizability. For two *distinct* targets both equal to 1, a total radius one instead gives equal optimal updates $1/\sqrt{2}$ and average loss $(1 - 1/\sqrt{2})^2$, rather than zero.

For different semidefinite geometries and distinct variables, any radii t_e with $\sum_e t_e^2 \le t^2$ yield jointly feasible scaled lifts from Equation (274). Their weighted upper-bound values add. Their individual lower-bound values add under these *individual* caps; for only a total cap, a valid lower bound minimizes the sum of individual lower bounds over $t_e \ge 0$, $\sum_e t_e^2 \le t^2$. With shared variables, evaluate the same constructed Δ_e in every associated role; separate per-role minima are only a relaxation unless compatibility is proved.

Interface with the nonlinear parameter model. The concatenated d in Equation (276) is the same parameter vector used in Equation (241). Choosing radii with total at most t makes the derivative envelopes at that radius applicable, subject to their separate state-region and route hypotheses. For random updates with this bound almost surely, $\mathbb{E}\|d\|^4 \le t^4$; covariance or independence of calls is not needed for this inequality. Repetition still affects G_j, c_j, b_j, not the count of distinct parameter entries. Applying the radius bound to terminal risk also requires the stated route, remainder, cross-covariance, and exit controls. For factors initialized nontrivially, the actual increment is $BA - B_0A_0$; factor width r does not by itself make this increment rank at most r. The rank and radius constraints must hold for the increment, as emphasized by Proposition 56.

P.3 Support replay recomputes the selected gain at its own state

Fix a finite prefix (or a distribution of prefixes with finite maximum length), not a recursively generated token sequence. All states include the full prefix and every value needed by attention, residual paths, and subsequent events. For events $j = 0, \dots, H-1$, let $\mathcal{A}_j$ be the finite set of complete support patterns at that event; a pattern may specify one expert per token. A deterministic rule $\rho_j(s)$, including its tie convention, selects the natural support. Route-free events have a single possible pattern. Let $F_{j,a}(s)$ be the complete branch extension at prescribed support a, including the *current-state* gain, expert inputs, and every shared path. In a top-one residual event it has the form

$$F_{j,a}(s) = B_j(s) + p_{j,a}(s) v_{j,a}(s), \qquad p_j(s) = \mathrm{softmax}(z_j(s)). \tag{280}$$

Here B_j includes the common residual path. Forcing a uses $p_{j,a}(s)$, not $\max_e p_{j,e}(s)$ and not a gain recorded at another state. The branch is defined also where a would not win.

On one common probability space, fix a prescribed support sequence a_j° (for example, recorded from a declared reference pass), an input, model parameters, and any common external random tape. For each realization these objects are held the same in the following two runs:

$$n_0 = r_0 = s_0, \qquad n_{j+1} = F_{j,\rho_j(n_j)}(n_j), \qquad r_{j+1} = F_{j,a_j^\circ}(r_j). \tag{281}$$

Thus n is natural execution and r is *support replay*. The prescribed sequence is not revised in response to either run's terminal loss. It may depend on the input and correlated parameters; no independence is assumed. Replay is a defined counterfactual computation. Model parameters are the same within this pair; comparison of two models uses a separate pair for each.

Theorem 59 (Natural execution versus support replay)**.** *Let τ be the first j with $\rho_j(r_j) \neq a_j^\circ$, setting $\tau = H$ if there is none, and put $\Omega = \{\tau = H\}$ and $p_{\mathrm{diff}} = \mathbb{P}(\Omega^c)$. Then τ is also the first index at which $\rho_j(n_j) \neq a_j^\circ$. In particular, $n_H = r_H$ on Ω. For a common target Y and the same measurable loss $0 \leq \ell(s, Y) \leq M < \infty$ in both runs,*

$$\left|R^{\mathrm{nat}} - R^{\mathrm{rep}}\right| := |\mathbb{E}\ell(n_H, Y) - \mathbb{E}\ell(r_H, Y)| \leq M p_{\mathrm{diff}}. \tag{282}$$

This bound needs neither smoothness nor independence, and its factor M is sharp over the stated class.

For a functional refinement, suppose the prescribed branch is L_j-Lipschitz between the required pairs n_j, r_j, and define

$$\begin{aligned} f_j &= \|F_{j,\rho_j(n_j)}(n_j) - F_{j,a_j^\circ}(n_j)\|, \\ D_H &= \sum_{j=0}^{H-1} \left(\prod_{k=j+1}^{H-1} L_k \right) f_j. \end{aligned} \tag{283}$$

Then $\|n_H - r_H\| \leq D_H$. If also the terminal loss is L_ℓ-Lipschitz in this norm, then

$$|R^{\mathrm{nat}} - R^{\mathrm{rep}}| \leq \mathbb{E} \min\{M \mathbf{1}_{\Omega^c}, L_\ell D_H\}. \tag{284}$$

Path-dependent finite L_j are allowed, but their products remain inside expectation. The Lipschitz condition is on branch extensions over the compared states, not an assumption that the discontinuous natural router is Lipschitz.

Proof. Before the first replay mismatch, induction gives $n_j = r_j$: the two rules choose a_j° at the same state and apply the identical branch, including its current-state gain. If the first mismatch occurs, the equal pre-event states make it a mismatch in the natural run as well. If none occurs, equality persists to the endpoint. Losses therefore differ only on Ω^c, and their pointwise absolute difference is at most $M\mathbf{1}_{\Omega^c}$. Taking expectations proves Equation (282). For sharpness, use one event and a Bernoulli input of probability p: the natural rule selects branch 1 on that input and branch 0 otherwise, replay always selects 0, and branch outputs/losses are respectively M and zero. The risk difference is Mp. For the functional statement, insert $F_{j,a_j^\circ}(n_j)$ to get $\|n_{j+1} - r_{j+1}\| \leq f_j + L_j\|n_j - r_j\|$. Unroll from equal initial states. Both this norm bound times L_ℓ and the bounded-loss indicator bound apply pointwise; take their minimum and expectation. □

Expert-output and gain contributions. At a common state in Equation (280), write $e = \rho_j(s)$, $a = a_j^\circ$, suppress j, s, and insert $p_e v_a$ to obtain

$$f_j \leq p_e\|v_e - v_a\| + |p_e - p_a|\|v_a\|. \tag{285}$$

The reverse insertion gives $p_a\|v_e - v_a\| + |p_e - p_a|\|v_e\|$; either is valid. For rowwise routing and Frobenius state norm, combine per-row upper bounds by the square root of their squared sum. Shared paths cancel at this common state. Equal expert outputs alone need not imply equal branch outputs: with $v_1 = v_2 = 1$, probabilities $(0.8, 0.2)$ and prescribed expert 2, the discrepancy is 0.6 despite zero expert-output diameter. The selected masses differ. Thus the equal-mass diameter lemma cannot be used here without a null atom or an explicit gain term. By contrast, if the prescribed expert equals the natural expert at the *same* state, both support and recomputed gain agree, and $f_j = 0$.

A first-disagreement margin bound on the correct states. For rowwise top-one routing, sum only over routed events. Let reference logits z_j° select a_j° with winner gap $\gamma_j \geq 0$ (minimum over tokens), and put $U_j = \|z_j(r_j) - z_j^\circ\|_\infty$ over all tokens and coordinates. With the same tie rule,

$$\begin{aligned} p_{\mathrm{diff}} &\leq \min\{1, \sum_j \mathbb{P}(\tau \geq j,\ \gamma_j \leq 2U_j)\} \\ &\leq \min\left\{1, \sum_j \inf_{v>0} \left[\mathbb{P}(\gamma_j \leq 2v) + \frac{\mathbb{E}U_j^\nu}{v^\nu}\right]\right\}, \qquad \nu > 0. \end{aligned} \tag{286}$$

For the first line, a first mismatch implies the displayed margin event by Lemma 27; sum over its disjoint possible locations. For the second line, drop $\{\tau \geq j\}$, split each margin event at $U_j \leq v$, and apply Markov to $U_j > v$. No independence across events or between margin and perturbation is needed. If the marginal CDF satisfies a power bound, Theorem 30 gives its sharper constant instead. Reference inputs with ties remain in the bound; they are not assumed to have zero mass. The moments and margins here are those of *support-replay* states. For the down-only specialization, Equation (254) is an alternative sufficient condition for $p_{\mathrm{diff}} = 0$ under its hypotheses.

Transferring replay-risk comparisons to natural execution. For two models evaluated on the same target, loss and input population, let b_m be a valid right-hand side of Equation (282) or Equation (284). Then

$$R^{\mathrm{nat}}_{\mathrm{sep}} - R^{\mathrm{nat}}_{\mathrm{reuse}} \geq R^{\mathrm{rep}}_{\mathrm{sep}} - R^{\mathrm{rep}}_{\mathrm{reuse}} - b_{\mathrm{sep}} - b_{\mathrm{reuse}}. \tag{287}$$

Consequently a replay-risk gap exceeding the two penalties certifies the corresponding natural-risk ordering. With only replay-risk intervals, use its separate-model lower endpoint minus its reuse-model upper endpoint. Two upper bounds alone do not yield this conclusion. Each model may have its own prescribed sequence and penalty. The fixed-G or fixed-selection assumptions of Theorems 46 and 52 remain necessary for their exact quadratics.

Bounded squared loss for normalized fixed-prefix outputs. If the terminal state is the final normalized logits, the frozen head and normalizer in Section O.2 give a common bound $\|s\| \leq B$ on natural and replay outputs for a bounded prefix length. For a common target with $\|Y\| \leq B_Y$, the *unclipped* square loss $\ell(s, Y) = \|s - Y\|^2$ obeys $M = (B + B_Y)^2$ and $L_\ell = 2(B + B_Y)$ on this output set. Indeed, $|\|s - Y\|^2 - \|s' - Y\|^2| \leq (\|s - Y\| + \|s' - Y\|)\|s - s'\|$. Thus the bridge can apply to actual bounded-output square risk; it is not limited to an artificially clipped loss. This conclusion requires the stated output/target bounds.

P.4 Frozen-gain replay pays an additional functional term

The preparation code's `force/gain_reference` mechanism fixes both the chosen index and the supplied gain at a forced position. The implementation uses frozen-gain replay, defined below. For a top-one event, fix a reference gain $q_j \in [0, 1]$ and define

$$\overline{F}_{j,a^\circ_j}(s) = B_j(s) + q_j v_{j,a^\circ_j}(s), \qquad h_0 = s_0, \quad h_{j+1} = \overline{F}_{j,a^\circ_j}(h_j).$$

For rowwise routing, q_j is the vector of prescribed gains. For route-free events let $\overline{F}_j = F_j$.

Proposition 60 (A gain-aware bridge for complete-gate replay)**.** *Suppose the current-gain prescribed branch F_{j,a°_j} is L_j-Lipschitz between r_j, h_j, and define*

$$c_j = \|F_{j,a^\circ_j}(h_j) - \overline{F}_{j,a^\circ_j}(h_j)\|, \qquad V_H = \sum_j \left(\prod_{k=j+1}^{H-1} L_k \right) c_j. \tag{288}$$

Then $\|r_H - h_H\| \leq V_H$. In the single-row form, $c_j = |p_{j,a^\circ_j}(h_j) - q_j| \|v_{j,a^\circ_j}(h_j)\|$. For a common loss in $[0, M]$ that is also L_ℓ-Lipschitz,

$$|R^{\mathrm{nat}} - R^{\mathrm{frozen}}| \leq M p_{\mathrm{diff}} + \mathbb{E}\left[\mathbf{1}_\Omega \min\{M, L_\ell V_H\}\right]. \tag{289}$$

Here Ω and p_{diff} are the support-replay *event and probability in Theorem 59, not agreement measured on h. No independent-gain or independent-state assumption is used. If L_j are random, all products stay inside the expectation.*

Proof. Insert $F_{j,a^\circ_j}(h_j)$ to obtain $\|r_{j+1} - h_{j+1}\| \leq L_j \|r_j - h_j\| + c_j$ and unroll. The formula for c_j follows by canceling the same-state shared path and expert output. On Ω, $n_H = r_H$, so the natural/frozen loss difference is bounded by $\min\{M, L_\ell V_H\}$. On its complement it is at most M. Splitting the expectation proves the claim. In particular, the right-hand side never exceeds M; it does not factor a gain error from a dependent event. □

Counterexample to an index-only bound for frozen gain. Let $\sigma(x) = (1 + e^{-x})^{-1}$ and start from $s_0 = 1$. At the first residual event the winning support is always 1, with constant gain $g_0 = \sigma(1)$, and $F_0(s; D) = s + g_0 D$ for a constant expert feature. The reference has $D = 0$; the compared model changes only this output weight to $D = 1/g_0$, reaching $s_1 = 2$ instead of the reference state 1. At the second event use logits $(s, 0)$, expert-1 output 1, and residual map $s \mapsto s+\sigma(s)$. Every reference, natural, support-replay and frozen-gain run selects expert 1 throughout. Nevertheless the natural/support-replay endpoint is $2 + \sigma(2)$, while reference-gain replay ends at $2 + \sigma(1)$. The common bounded Lipschitz loss $\ell(s) = \min\{1, \max\{0, s/4\}\}$ gives risk difference $(\sigma(2) - \sigma(1))/4 > 0$ although $p_{\text{diff}} = 0$. This two-event construction has the selected-softmax-gain semantics of Equation (280). It disproves the index-only extension.

Conditions for applying the replay bounds. Without bounded loss, a mismatch probability p does not bound squared risk: one event may output p^{-1} only on the mismatching event, giving squared risk p^{-1}. Conversely, bounded but discontinuous classification loss need not be Lipschitz: outputs $-\epsilon$ and $+\epsilon$ can have loss difference one for every $\epsilon > 0$. Thus the gain-to-loss refinement requires its Lipschitz premise or a separate terminal decision-margin argument; boundedness alone still gives the support-replay result. The normalized fixed-prefix branch constants in Proposition 51 can supply the stated branch Lipschitz bounds at a specified total update radius, without freezing attention activations or selected gains. Extra adaptive decisions must be included in the event support/state before invoking these results.

Q Assumptions and Theorem Dependencies

This table summarizes the dependencies of the existing results. Each alternative retains its own assumptions.

Result	Key assumptions	Application conditions	Applicable extension
Q.1 Local estimation and data-dependent choices			
Lemma 6	Declared roles, interfaces and complete states; probe objects fixed independently of fitting, or covered jointly.	Same-data learned environments need a joint uniform class or independent split.	Sections B and C.
Theorems 1 and 16	Fixed orthogonal designs, centered independent noise with stated moments, independent tests and fixed pooling.	General designs and covariance need their own risk geometry.	Theorems 38 and 39.
Proposition 7	Common squared Hilbert geometry; unrestricted or appropriate realizable class for pure variance.	Restricted classes retain distance corrections.	Equation (39).
Theorems 2, 18 and 19 and Corollary 20	Fixed disjoint groups or linear tying spaces; balanced cells and stated unbiased covariance for dimension penalties.	Adaptive grouping needs selection control; embeddings need realization conditions; greedy optimization needs separate analysis.	Corollary 40; independent selection or a uniform family bound.
Equation (64) and Proposition 10; Section D.6	Gaussian noise and positive variance for sampling laws. The plug-in also uses known $q > 0$, a fixed mean subspace and $k \geq 4$.	Degeneracies use deterministic identities; searched merges need multiplicity control.	Separate expectation identities and selection guarantees.
Theorem 8	Bounded losses, fixed finite class and weights, independent samples, joint confidence budget.	Use independent records and retain fitting gaps and weight mismatch.	Equation (52) and Proposition 41.
Proposition 9 and Corollary 17	Declared allocation objective and continuous feasible sample counts.	Enforce design floors and integer feasibility.	Constrained allocation; integer allocation is a separate problem.
Propositions 11 and 45 and Theorem 48	Fixed fitted candidates and weights; independent validation/audit units; bounded loss or stated tail conditions.	Outcome-dependent candidates or unknown tail constants need further analysis.	Bounded and Gaussian quadratic-loss guarantees remain separate.
Q.2 Composition, routing, and intervention			
Equation (75); Assumption 3	Common coupled program and state metric; regional Lipschitz envelopes, common deterministic role weights and assembly residual.	The terminal certificate orders bounds, not risks; random products stay inside expectations.	Proposition 42; optimization and exit costs remain.
Theorem 46 and Proposition 47	Unbiased finite-covariance estimates, fixed sharing and one common fixed linear terminal map.	Retain cross-covariances; route-dependent propagation needs conditional moments or remainder control.	Propositions 13, 43 and 49.
Section N.4	Orthogonal pooling, positive contrast rank, covariance qI, $q > 0$; $QG^\top GQ = cQ$, $c > 0$.	An all-target condition, not a necessary condition for one target or a guarantee for adaptive sharing.	The target-specific fixed-map comparison remains distinct.
Propositions 49 and 51 and Theorem 50	Same-law remainder control; joint derivatives, region/route control and stated higher moments for Taylor bounds.	Use joint derivative and remainder control; the concrete specialization fixes the prefix and perturbs down matrices only.	Retain escape costs and justified remainder moments.

Result	Key assumptions	Application conditions	Applicable extension
Theorem 52 and Proposition 53	Condition on enough information to fix propagation coefficients and offsets; recompute conditional moments.	Changing selection partitions requires candidate-specific laws.	The half-space Gaussian law is a specialization.
Lemma 12, Theorem 21, and Corollary 22	Common input and expert outputs; equal nonnegative mass; uniform $1/k$ weights for matching.	Dispatch changes inputs; other drop or normalization rules need their own maps.	Theorem 23: normalized dispatch/combine maps and expert sensitivity.
Proposition 24 and Theorem 25	Stated hierarchy or preselected sequence; compatible spaces and reference-suffix sensitivity.	Use separate routed events or a complete adaptive-policy comparison.	Separate routed events or a full-policy comparison.
Proposition 26, Lemma 27, and Theorem 28	Gate sensitivity for soft routing; aligned logits, strict margins and error envelopes for hard support.	Sufficient, not necessary; equal support does not fix nonuniform gains.	Theorems 29 and 30 and Proposition 31 distinguish conditional and marginal control.
Theorem 59 and Proposition 60	Same prefix, parameters, target and randomness; bounded loss. Gain refinements also need branch/loss sensitivity.	Recorded gains can incur extra cost despite identical indices.	Proposition 60; retain dependent products and event indicators.
Propositions 5, 14 and 15	Declared intervention and extractor; distinguish worst-case and average squared-risk obstructions.	Natural labels do not identify operations; off-policy defects need transfer.	Equations (98) and (202): coupled actions, stability and state sufficiency.
Theorem 32 and Proposition 33	Randomization, consistency, square-integrable Hilbert outcomes and unrestricted Bayes predictors; a nontrivial unrestricted effect class for non-identification.	Local bias matches only with matched inputs, classes, losses and laws; unsupported effects are not identified.	Conditional-mean estimation and coverage require separate analysis.
Theorem 34 and Proposition 36	Known positive forcing propensities, independent bounded records and fixed targets; known second moments for allocation.	Router scores are not forcing probabilities; least-covered-action precision is not every aggregate rate.	Independent pilot and required propensity floors; cross-fitting needs foldwise accounting.
Q.3 Interpreting diagnostics and applying a result			
Propositions 37, 55 and 56 and Theorem 54	Stated functional or matrix class, core, initialization, feature geometry and insertion point.	Functional rank is not LoRA rank; approximation does not establish learned or terminal risk.	Theorem 57 and Proposition 58: actual updates, each distinct matrix counted once.
Equation (204); application of certificates	Declared evaluation population and budgets; nonnegative cost conversions; justified uniform or input-integrated bounds.	Resource scores express preferences; uniform envelopes require justification beyond sampled maxima.	Independent pilot/fitting/audit separation or justified joint/foldwise analysis.