PORTFOLIO OF EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse mixture-of-experts (MoE) models increase model capacity without proportionally increasing per-token computation by activating only a small subset of experts. Route quality therefore depends on both expert utility and complementarity within the selected subset. However, most routers rank experts independently, potentially selecting experts with redundant task contributions. Effective routing must instead identify a high-utility, complementary expert subset under a fixed budget. We introduce *Portfolio of Experts* (PoE), which formulates sparse routing as conditional portfolio optimisation under outcome uncertainty. PoE defines counterfactual gradient returns (CGRs), which quantify the local effect of each expert's allocation on downstream loss. A lagged critic predicts each candidate expert's expected CGR and the pairwise covariance of these returns. A constrained solver then jointly selects and weights a small set of experts, favouring high predicted utility while avoiding redundant combinations under a fixed routing budget. Across seven zero-shot downstream tasks spanning science question answering, commonsense reasoning, reading comprehension, and contextual prediction, PoE attains the highest mean accuracy and lowest held-out NLL at every evaluated expert-pool size. Frozen-route studies show that modelling pairwise dependence improves local routing utility, while routing diagnostics show that PoE lowers joint failure without activating more experts.

1 INTRODUCTION

Sparse mixture-of-experts (MoE) models increase model capacity without a proportional increase in computational cost by activating only a small subset of experts for each input (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Dai et al., 2024). By restricting computation to a fixed number of experts, sparse MoE models make routing a central determinant of how effectively their expanded capacity is utilised. Their performance therefore depends critically on how the limited routing budget is allocated.

Most existing routers score experts independently, even though route quality depends on the selected set as a whole. Experts that are useful in isolation may provide overlapping task value when activated together, highlighting the need to account for complementarity among co-selected experts. Load balancing regulates expert usage rather than interactions among co-selected experts. Existing training-time diversity methods encourage expert differentiation through orthogonal updates or CKA penalties (Liu et al., 2024; Do et al., 2025). Recent analyses further show that separation in representation space does not necessarily imply functional separation (Liu, 2026). Evaluations also indicate that an expert can still provide useful signal for next-token prediction even when its output offers little additional geometric information beyond the selected set (Tian et al., 2026). However, such methods retain Top-k selection based on per-expert routing scores, rather than using a set-aware criterion that jointly considers task utility and within-set complementarity. This gap motivates our central research question:

How can sparse MoE routers select individually useful experts that form a complementary group?

We introduce *Portfolio of Experts* (PoE), which casts routing as contextual portfolio selection under uncertainty (Markowitz, 1952; Bertsimas & Kallus, 2020). Prior to knowing how much each expert will help, the router must choose a group of experts. For each token, the expected return measures how much an expert is predicted to reduce the loss, while the return covariance measures how uncertain each expert's contribution is and whether pairs of experts tend to succeed or fail together. PoE leverages these two quantities, the conditional mean and covariance of expert returns given the token, to score each candidate group of experts as a whole, instead of scoring each expert on its own.

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

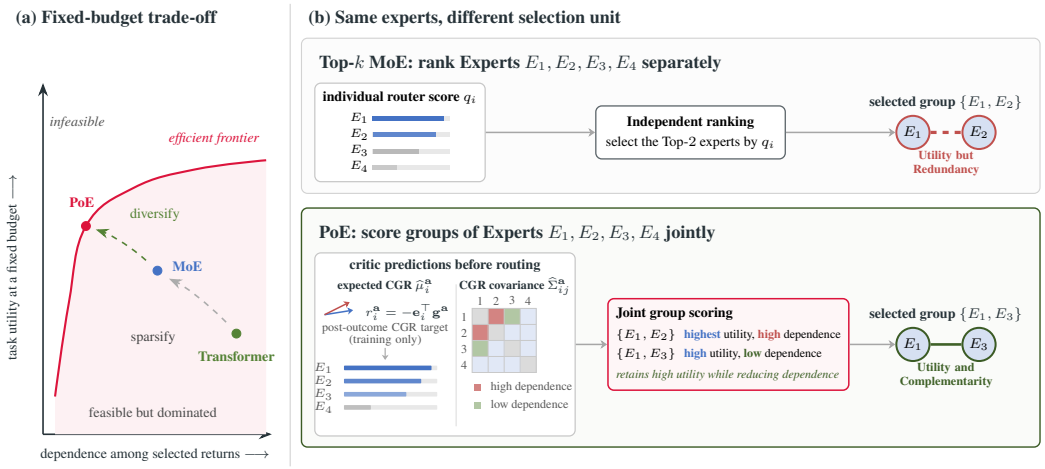


Figure 1: **Portfolio of Experts: from expert ranking to joint group selection.** **(a)** With a fixed budget, the router must trade off an expert group's task utility against its redundancy, measured by the dependence among its members' returns. The efficient frontier contains the nondominated routes. **(b)** A Top-k MoE assigns scalar scores independently and therefore cannot account for overlap between its two highest-ranked experts. PoE scores groups from predicted utility and pair dependence. The group $\{E_1, E_2\}$ has the highest utility but high dependence, whereas $\{E_1, E_3\}$ preserves comparable utility with lower dependence and is selected.

As illustrated in Figure 1, independent routing may select individually strong but collectively redundant experts, whereas PoE evaluates candidate groups using task-aware utility and dependence. A lagged critic learns to predict these quantities at routing time, before the outcome is known. It is trained on counterfactual gradient returns (CGRs), which measure how much each expert locally reduces the loss, with training-only shadow experts run at zero weight so that unselected experts also receive targets. A constrained solver then picks a sparse allocation near the anchor route. Complementarity is therefore task-adjoint, meaning that two experts overlap when their effects on the loss move together. This criterion reduces to hidden-space cosine similarity only when downstream gradient covariance is isotropic.

Across seven zero-shot downstream tasks spanning science question answering, commonsense reasoning, reading comprehension, and contextual prediction, PoE attains the highest mean accuracy and lowest held-out NLL at every evaluated expert-pool size, and its accuracy margin widens as the pool grows. Relative to Top-k routing, PoE cuts joint failure, where no selected expert locally reduces the loss, by an order of magnitude while activating slightly fewer experts. Compared with a strong baseline, it achieves lower joint failure with substantially fewer active experts. Frozen-route ablations show that modelling pairwise dependence markedly improves local route quality, and calibration analyses confirm that the critic recovers this dependence from routing-time information alone.

Contributions.

- **Joint portfolio routing.** We propose *Portfolio of Experts* (PoE), a novel framework for jointly selecting experts under a fixed activation budget based on predicted task utility and inter-expert dependence.
- **Task-adjoint theory.** We establish the task-adjoint covariance geometry of CGR, derive a mean–variance routing objective from a risk-sensitive utility criterion, and bound the errors arising from local utility approximation, moment prediction, and support search.
- **Empirical effectiveness.** PoE attains the highest mean accuracy and lowest held-out NLL at every evaluated expert-pool size, while lowering joint failure without activating more experts.

2 Preliminaries

We first introduce sparse mixture-of-experts routing and the portfolio principle used by PoE. The notation in this section is retained throughout Section 3.

2.1 SPARSE MIXTURE-OF-EXPERTS ROUTING

Consider one token at one routed layer. Let $s \in \mathcal{S}$ denote the available routing context and let $\mathbf{x}(s) \in \mathbb{R}^d$ be the layer input. The layer contains N routed experts indexed by $[N] := \{1, \dots, N\}$. Expert i is a map $\mathcal{E}_i : \mathbb{R}^d \to \mathbb{R}^d$ with output

$$\mathbf{e}_i(s) := \mathcal{E}_i(\mathbf{x}(s)) \in \mathbb{R}^d.$$

Collecting these outputs gives $\mathbf{E}(s) := [\mathbf{e}_1(s), \dots, \mathbf{e}_N(s)] \in \mathbb{R}^{d \times N}$. Let $\mathbf{y}_0(s) \in \mathbb{R}^d$ contain the residual path and any always active shared expert. For a nonnegative routing allocation $\mathbf{v} \in \mathbb{R}^N_{\geq 0}$, the layer output is

$$\mathbf{y}(\mathbf{v}; s) = \mathbf{y}_0(s) + \mathbf{E}(s)\mathbf{v}. \tag{1}$$

Expert i is executed when $v_i > 0$.

We use three constraints to express the routing budget. The integer $k \in [N]$ limits the number of routed experts, $B > 0$ limits total allocation magnitude, and $\mathbf{u}(s) \in \mathbb{R}^N_{>0}$ gives coordinatewise caps. The resulting budget set is

$$\mathcal{B}(s) := \left\{ \mathbf{v} \in \mathbb{R}^N_{\geq 0} : v_i \leq u_i(s)\ \forall i, \quad \|\mathbf{v}\|_1 \leq B, \quad \|\mathbf{v}\|_0 \leq k \right\}. \tag{2}$$

A conventional Top-k router assigns a scalar score $q_i(s)$ to each expert and selects

$$\mathcal{I}_{\mathrm{TopK}}(s) := \mathrm{TopK}_{i \in [N]}\, q_i(s). \tag{3}$$

2.2 MEAN AND VARIANCE IN PORTFOLIO SELECTION

Classical portfolio selection allocates a budget across items whose returns are uncertain when the allocation is chosen (Markowitz, 1952). Let $\mathbf{R} \in \mathbb{R}^N$ be a random return vector with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{R}]$ and covariance $\boldsymbol{\Sigma} := \mathrm{Cov}(\mathbf{R})$. For a fixed nonnegative allocation $\mathbf{v}$, the aggregate return $Y_{\mathbf{v}} := \mathbf{v}^\top \mathbf{R}$ satisfies

$$\mathbb{E}[Y_{\mathbf{v}}] = \mathbf{v}^\top \boldsymbol{\mu}, \qquad \mathrm{Var}(Y_{\mathbf{v}}) = \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}. \tag{4}$$

Mean and variance are combined through the utility

$$J_\eta(\mathbf{v}) := \mathbf{v}^\top \boldsymbol{\mu} - \frac{\eta}{2} \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}, \qquad \eta \geq 0. \tag{5}$$

The first term rewards expected return. The coefficient η controls the importance of the variance term. Expanding that term gives

$$\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} = \sum_{i=1}^{N} v_i^2 \Sigma_{ii} + 2 \sum_{1 \leq i < j \leq N} v_i v_j \Sigma_{ij}. \tag{6}$$

The diagonal entry Σ_{ii} is the variance of item i. The off-diagonal entry Σ_{ij} records how items i and j vary together. Positive covariance increases aggregate variance when both items receive positive allocation. Negative covariance allows their fluctuations to offset each other.

3 PORTFOLIO OF EXPERTS

Portfolio of Experts (PoE) formulates sparse routing as conditional portfolio selection. At an anchor allocation $\mathbf{a}(s)$, realised returns $\mathbf{r}^{\mathbf{a}}(s, \xi)$ quantify the experts' local effects on task loss, and a critic predicts their conditional mean and covariance from s. A constrained portfolio objective then selects $\mathbf{p}(s)$ using only information available at routing time.

3.1 COUNTERFACTUAL GRADIENT RETURN

Recall the MoE output $\mathbf{y}(\mathbf{v}; s)$ from Equation 1. Since the layer output is linear in the allocation, increasing expert i's allocation v_i pushes the output along $\mathbf{e}_i(s)$, and the router needs to know what that does to the loss.

The anchor router produces $\mathbf{a}(s) \in \mathcal{B}(s)$ and the corresponding layer output is $\mathbf{y}^{\mathbf{a}}(s) := \mathbf{y}(\mathbf{a}(s); s)$. Write S for the random routing context and Ξ for the random post-routing outcome, with realisations s and ξ. The route depends on S and is fixed before Ξ is used to compute the task loss.

Let $\mathcal{L}^{[\mathbf{a}]}(\mathbf{y}; s, \xi)$ denote the downstream task loss when the current layer output is set to $\mathbf{y}$. Downstream discrete routes and stochastic masks are fixed to those of the anchor computation, so that

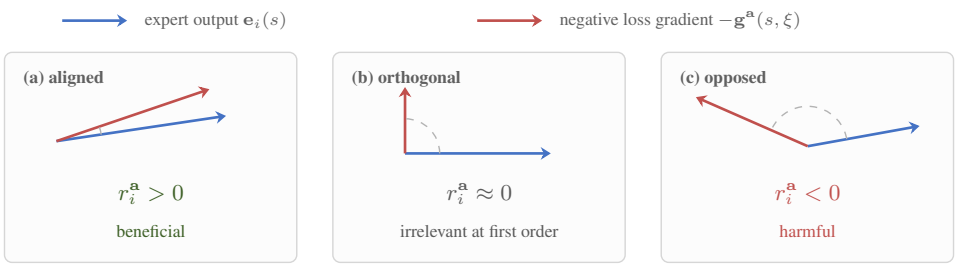


Figure 2: **Geometric interpretation of counterfactual gradient return.** The return $r_i^{\mathbf{a}}(s,\xi) = -\langle \mathbf{e}_i(s), \mathbf{g}^{\mathbf{a}}(s,\xi)\rangle$ is the inner product between expert i's output and the negative loss gradient.

varying $\mathbf{y}$ cannot flip a downstream decision. The layer-output gradient and the counterfactual gradient return (CGR) are

$$\begin{aligned} \mathbf{g}^{\mathbf{a}}(s,\xi) &:= \nabla_{\mathbf{y}} \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi) \in \mathbb{R}^d, \\ \mathbf{r}^{\mathbf{a}}(s,\xi) &:= -\mathbf{E}(s)^\top \mathbf{g}^{\mathbf{a}}(s,\xi) \in \mathbb{R}^N, \\ \mathbf{R}^{\mathbf{a}} &:= \mathbf{r}^{\mathbf{a}}(S,\Xi). \end{aligned} \tag{7}$$

The gradient $\mathbf{g}^{\mathbf{a}}$ is already supplied by backpropagation, and since $\partial \mathbf{y}/\partial v_i = \mathbf{e}_i(s)$, the chain rule gives $\partial \mathcal{L}^{[\mathbf{a}]}/\partial v_i = \langle \mathbf{e}_i(s), \mathbf{g}^{\mathbf{a}}(s,\xi)\rangle$. Negating this so that larger is better, the ith return has the simple form

$$r_i^{\mathbf{a}}(s,\xi) = -\langle \mathbf{e}_i(s), \mathbf{g}^{\mathbf{a}}(s,\xi)\rangle.$$

A positive value means that increasing v_i locally decreases the loss. A negative value means that the expert locally increases the loss. Figure 2 illustrates this sign convention. The return is counterfactual because expert i needs only an output, not a positive weight, for $r_i^{\mathbf{a}}$ to be defined.

To relate CGR to a route change, consider a candidate $\mathbf{p}(s) \in \mathcal{B}(s)$ and define $\boldsymbol{\delta}(s) := \mathbf{p}(s) - \mathbf{a}(s)$. A first-order expansion gives

$$\begin{aligned} &\mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi) - \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}(\mathbf{p}(s); s); s, \xi) \\ &\qquad = \boldsymbol{\delta}(s)^\top \mathbf{r}^{\mathbf{a}}(s,\xi) + O\left(\|\mathbf{E}(s)\boldsymbol{\delta}(s)\|_2^2\right). \end{aligned} \tag{8}$$

Thus CGR estimates the local loss reduction caused by reallocating expert weight. It is a local utility measure rather than a global attribution. Appendix A gives the exact residual and its curvature bound.

3.2 Contextual Portfolio Routing

The realised CGR in Equation 7 depends on the outcome and is not available when the route is selected. PoE therefore uses its conditional mean and covariance

$$\begin{aligned} \boldsymbol{\mu}^{\mathbf{a}}(s) &:= \mathbb{E}[\mathbf{R}^{\mathbf{a}} \mid S = s] \in \mathbb{R}^N, \\ \boldsymbol{\Sigma}^{\mathbf{a}}(s) &:= \mathrm{Cov}(\mathbf{R}^{\mathbf{a}} \mid S = s) \in \mathbb{R}^{N \times N}. \end{aligned} \tag{9}$$

For a candidate rule $\mathbf{p}(S)$, define the aggregate expert return

$$Y_{\mathbf{p}}^{\mathbf{a}} := \mathbf{p}(S)^\top \mathbf{R}^{\mathbf{a}}. \tag{10}$$

Conditional on $S = s$, the allocation is fixed. Consequently,

$$\mathbb{E}[Y_{\mathbf{p}}^{\mathbf{a}} \mid S = s] = \mathbf{p}(s)^\top \boldsymbol{\mu}^{\mathbf{a}}(s), \tag{11}$$

$$\mathrm{Var}(Y_{\mathbf{p}}^{\mathbf{a}} \mid S = s) = \mathbf{p}(s)^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p}(s). \tag{12}$$

The first quantity measures expected local utility. The second measures individual return variation and pairwise dependence. Its diagonal and pairwise terms have the expansion in Equation 6.

The candidate should remain close to the anchor because CGR is a local approximation. With trust radius $\tau > 0$, define

$$\mathcal{A}(s; \mathbf{a}) := \{\mathbf{p} \in \mathcal{B}(s) : \|\mathbf{p} - \mathbf{a}(s)\|_2 \le \tau\}. \tag{13}$$

PoE selects the ideal allocation

$$\mathbf{p}^\star(s) \in \arg\max_{\mathbf{p} \in \mathcal{A}(s;\mathbf{a})} \left\{ \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}(s) - \frac{\eta}{2} \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p} \right\}, \qquad \eta \ge 0. \tag{14}$$

The anchor term $\mathbf{a}(s)^\top \boldsymbol{\mu}^{\mathbf{a}}(s)$ is constant across candidates, so the first term gives the same first-order ranking as expected loss reduction from the anchor. The second term discourages allocations whose expert returns vary together. Setting $\eta = 0$ gives the mean-only control.

3.3 LOSS-GRADIENT GEOMETRY OF RETURN COVARIANCE

The return covariance in Equation 14 characterises the dependence structure of experts' first-order effects on the loss, rather than the geometric similarity of their output representations. Define the conditional mean and covariance of the backpropagated layer-output gradient as

$$\mathbf{m}_g^{\mathbf{a}}(s) := \mathbb{E}[\mathbf{g}^{\mathbf{a}}(S,\Xi) \mid S = s], \qquad \mathbf{G}^{\mathbf{a}}(s) := \mathrm{Cov}(\mathbf{g}^{\mathbf{a}}(S,\Xi) \mid S = s).$$

Proposition 1 (Loss-gradient moment geometry)**.** *For every context at which these moments exist,*

$$\boldsymbol{\mu}^{\mathbf{a}}(s) = -\mathbf{E}(s)^\top \mathbf{m}_g^{\mathbf{a}}(s), \qquad \boldsymbol{\Sigma}^{\mathbf{a}}(s) = \mathbf{E}(s)^\top \mathbf{G}^{\mathbf{a}}(s)\mathbf{E}(s). \tag{15}$$

In particular, $\Sigma_{ij}^{\mathbf{a}}(s) = \mathbf{e}_i(s)^\top \mathbf{G}^{\mathbf{a}}(s)\mathbf{e}_j(s)$. The covariance therefore measures expert overlap under the geometry induced by variation in the backpropagated loss gradient. Return covariance is proportional to the standard output-space inner product only when $\mathbf{G}^{\mathbf{a}}(s)$ assigns equal weight to all directions spanned by the expert outputs.

3.4 PREDICTING MOMENTS BEFORE ROUTING

The moments in Equation 9 must be predicted before the outcome is observed. At each routed layer, a critic with parameters $\boldsymbol{\omega}$ receives the detached layer input $\mathbf{x}(s)$ and a frozen allocation $\mathbf{c} \in \mathcal{B}(s)$. It predicts a mean $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\omega}}^{\mathbf{c}}(s)$ and a lower triangular factor $\mathbf{C}_{\boldsymbol{\omega}}^{\mathbf{c}}(s)$. The covariance used by the solver is

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{stat},\boldsymbol{\omega}}^{\mathbf{c}}(s) := \mathbf{C}_{\boldsymbol{\omega}}^{\mathbf{c}}(s)\mathbf{C}_{\boldsymbol{\omega}}^{\mathbf{c}}(s)^\top, \tag{16}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{route},\boldsymbol{\omega}}^{\mathbf{c}}(s) := \widehat{\boldsymbol{\Sigma}}_{\mathrm{stat},\boldsymbol{\omega}}^{\mathbf{c}}(s) + \epsilon \mathbf{I}_N. \tag{17}$$

The factorisation guarantees a positive semidefinite statistical covariance. The ridge term $\epsilon \mathbf{I}_N$ is introduced to ensure numerical stability in the routing optimisation. At routing time, $\mathbf{c} = \mathbf{a}(s)$. The solver substitutes $\widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{route}}^{\mathbf{a}}(s)$ into Equation 14.

Training targets are collected after the selected route is executed. A random panel $\mathcal{Q}_n$ contains the active experts and a small number of training-only shadow experts. A shadow expert is evaluated at zero allocation, so it supplies an off-route output without changing the forward result. After outcome ξ_n is observed, every expert in the panel receives the target

$$r_{n,i} = -\mathbf{e}_i(s_n)^\top \mathbf{g}^{\mathbf{p}_n}(s_n, \xi_n).$$

The critic regresses the conditional mean toward $r_{n,i}$ and the raw second moment toward $r_{n,i}r_{n,j}$. The resulting per-update moment-regression loss $\mathcal{L}_{\mathrm{mom}}^{(n)}$ uses known panel inclusion probabilities to correct for coordinate subsampling. Its full form and identification argument are given in Equation 34 and Appendix B. Shadows provide coverage for unselected experts, while the trust region and route distillation reduce the difference between executed allocations and future anchor queries.

3.5 TRAINING AND INFERENCE

The anchor is generated by a positive proposal followed by projection onto the MoE budget set

$$\mathbf{h}_\theta(s) := \mathrm{softplus}(\mathbf{W}_\theta \mathbf{x}(s)), \qquad \mathbf{a}_\theta(s) := \Pi_{\mathcal{B}(s)}(\mathbf{h}_\theta(s)). \tag{18}$$

Given the predicted moments, the support solver conducts a structured search over candidate supports using addition, removal, and swap operations. For each candidate support considered during the search, it optimises the continuous allocation and retains the highest-scoring feasible solution.

Let $\mathcal{L}_{\mathrm{LM}}$ denote the normalised minibatch next-token prediction loss, and let $\mathcal{L}_{\mathrm{mom}}$ denote $\mathcal{L}_{\mathrm{mom}}^{(n)}$ averaged over routed layers and tokens. The route-distillation loss $\mathcal{L}_{\mathrm{route}}$, defined below, is averaged over the same units.

Training follows three ordered stages.

1. **Predict and route.** The anchor produces $\mathbf{a}_n$. A lagged critic predicts its return moments, and the solver selects $\mathbf{p}_n$ before outcome ξ_n is used.
2. **Execute and measure.** The model executes $\mathbf{p}_n$ and its shadow panel. $\mathcal{L}_{\mathrm{LM}}$ then provides the CGR targets for the observed experts.

3. **Update future routes.** $\mathcal{L}_{\mathrm{LM}}$ updates the backbone and active experts, while $\mathcal{L}_{\mathrm{mom}}$ updates the critic. The anchor router is distilled toward the solver allocation with

$$\mathcal{L}_{\mathrm{route}}^{(n)} := \|\mathbf{h}_\theta(s_n) - \mathrm{sg}[\mathbf{p}_n]\|_2^2.$$

The lagged critic $\overline{\boldsymbol{\omega}}_n$ used for routing is refreshed only after $\mathcal{L}_{\mathrm{mom}}$ yields the learner critic $\boldsymbol{\omega}_n$, as $\overline{\boldsymbol{\omega}}_{n+1} = (1-\tau_{\mathrm{c}})\overline{\boldsymbol{\omega}}_n + \tau_{\mathrm{c}}\boldsymbol{\omega}_n$ with Polyak rate τ_{c} (Appendix B.3).

With nonnegative weights λ_{mom} and λ_{route}, the complete objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{mom}}\mathcal{L}_{\mathrm{mom}} + \lambda_{\mathrm{route}}\mathcal{L}_{\mathrm{route}}. \tag{19}$$

The CGR targets are computed from $\mathcal{L}_{\mathrm{LM}}$ before the auxiliary losses are added. A current outcome can therefore update only future routes. At inference, all parameters are fixed. The anchor produces $\mathbf{a}(s)$, the critic predicts its conditional moments, and the solver returns $\mathbf{p}(s)$. The model then executes the support of $\mathbf{p}(s)$.

4 Experiments

We evaluate PoE through end-to-end performance across expert-pool sizes, systems cost across nominal routing budgets, component contributions under frozen routes, and the reliability of CGR targets and pre-outcome predictions.

4.1 Experimental setup

Models and training. All methods use 12-layer models with $N \in \{16, 24, 32\}$ fixed-width routed experts plus one shared expert, with approximately 1B parameters at $N = 16$. Models are trained from scratch on the 100B-token Dolma dataset (Ettinger et al., 2025). We compare **Top-k MoE** (Shazeer et al., 2017), **ReMoE** (Wang et al., 2025), **MP-MoE** (Kang et al., 2026), and **PoE** under matched initialisation, data order, training, and evaluation. At nominal $k = 8$, the always-active shared expert is excluded from the routing budget. Top-k MoE and MP-MoE select exactly eight routed experts, PoE selects at most eight, and ReMoE targets eight but may differ. Reported active-expert counts include the shared expert. A deployed positive allocation counts as active. MP-MoE provides a group-selection baseline that constructs a covariance matrix of binary expert-selection indicators from routing co-occurrence counts and greedily optimises a Mahalanobis objective.

Tasks and metrics. We evaluate on seven zero-shot tasks, using ARC-Challenge and ARC-Easy (Clark et al., 2018) for science question answering, BoolQ (Clark et al., 2019) and RACE (Lai et al., 2017) for reading comprehension, HellaSwag (Zellers et al., 2019) for commonsense continuation inference, PIQA (Bisk et al., 2020) for physical commonsense reasoning, and LAMBADA (Paperno et al., 2016) for contextual word prediction. Table 1 reports task accuracy, mean accuracy, and negative log-likelihood (NLL) on a held-out split of the training mixture.

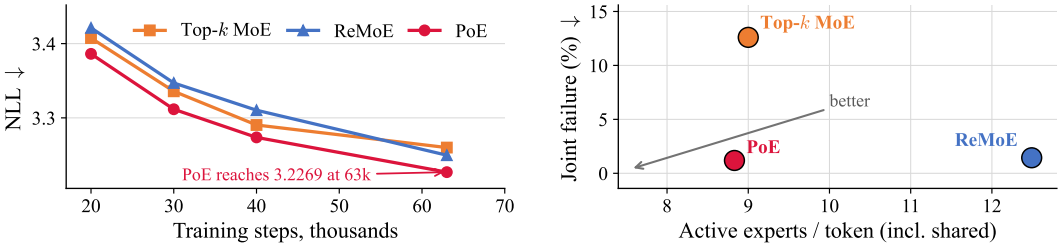
4.2 End-to-end performance and training cost

Table 1 shows that PoE achieves the highest mean downstream-task accuracy and the lowest held-out NLL at every evaluated expert-pool size. On HellaSwag, its relative improvement over Top-k MoE reaches 8.1% at $N = 24$ and 16.2% at $N = 32$. More broadly, the relative mean-accuracy improvement over Top-k MoE grows monotonically from 4.0% at $N = 16$ to 6.7% at $N = 24$ and 10.2% at $N = 32$. The widening gains suggest that PoE exploits additional model capacity more effectively as the expert pool grows. By discouraging the co-selection of experts with correlated task-relevant returns, PoE favours complementary subsets and makes better use of the specialisation available in larger pools. Collectively, these findings support the use of PoE as a scalable routing framework for higher-capacity MoE architectures.

Figure 3 (a) traces held-out cross-entropy throughout training. PoE maintains a lower loss than both Top-k MoE and ReMoE, and its margin over Top-k MoE is largest at the final checkpoint. Figure 3 (b) jointly examines routing efficiency and reliability through the number of active experts per token and the joint-failure rate. Because joint failure can decrease mechanically as more experts are activated, it should be evaluated jointly with the actual number of active experts. Points closer to the lower-left region correspond to more favourable operating points on the resulting Pareto frontier. Among the evaluated routers, PoE substantially lowers the probability of joint failure while maintaining a compact active support. This favourable trade-off indicates that PoE uses its routing budget more effectively by selecting expert sets with task-relevant and complementary contributions.

Table 1: End-to-end quality at nominal $k = 8$ across routed-expert pool sizes. Top-k denotes Top-k MoE and MP denotes MP-MoE (Kang et al., 2026). Bold marks the best result per size.

	$N = 16$				$N = 24$				$N = 32$			
	Top-k	ReMoE	MP	PoE	Top-k	ReMoE	MP	PoE	Top-k	ReMoE	MP	PoE
ARC-Challenge	0.337	0.327	0.345	**0.389**	0.361	0.372	0.385	**0.397**	0.385	0.382	0.399	**0.409**
ARC-Easy	0.640	0.636	0.663	**0.672**	0.675	0.677	0.691	**0.717**	0.705	0.711	0.735	**0.763**
BoolQ	0.524	0.517	**0.546**	0.530	0.575	0.592	0.588	**0.602**	0.617	0.634	0.667	**0.675**
HellaSwag	0.448	0.442	0.450	**0.452**	0.479	0.483	0.497	**0.518**	0.511	0.519	0.536	**0.594**
LAMBADA	0.428	0.428	0.437	**0.450**	0.452	0.455	0.462	**0.480**	0.484	0.477	0.498	**0.562**
PIQA	0.682	0.682	0.680	**0.690**	0.691	0.697	0.704	**0.715**	0.710	0.723	0.711	**0.739**
RACE	0.306	0.304	0.309	**0.314**	0.327	0.341	0.354	**0.374**	0.359	0.383	0.371	**0.414**
Mean acc. ↑	0.481	0.477	0.490	**0.500**	0.509	0.517	0.526	**0.543**	0.539	0.547	0.560	**0.594**
NLL ↓	3.260	3.250	3.234	**3.227**	3.167	3.152	3.136	**3.112**	3.085	3.093	3.067	**3.036**



(a) NLL over Training **(b) Joint Failure vs. Active Experts**

Figure 3: **(a)** Held-out NLL over training for $N = 16, k = 8$. **(b)** Joint failure versus active experts per token, including the shared expert. Lower left is better.

Throughput and computational overhead. Table 2 compares end-to-end throughput as the nominal routed-expert budget k increases. Relative to the Top-k MoE baseline, PoE incurs a step-time overhead of about 10% (7.6%–12.1%) across $k \in \{2, 4, 8\}$, with no systematic growth as k increases. Moreover, its overhead relative to MP-MoE is at most 1.76% at $k = 4$ and 8. Taken together, these measurements indicate that the auxiliary computations introduced by PoE, including critic evaluation, shadow evaluation, and support optimisation, result in a step-time overhead of about 10% relative to Top-k MoE across the evaluated nominal routing budgets.

4.3 Component contributions under frozen routes

Figure 4 presents paired ablations that isolate the contribution of each design choice. Route quality is evaluated by the empirical certainty equivalent of exact frozen-route loss reduction, whereas the solver optimises the local portfolio objective in Equation 14. Averaged over five seeds, modelling marginal return variation increases the evaluation score by approximately 77% relative to the mean-only variant, and incorporating full pairwise covariance yields a further 71% gain over the diagonal-covariance variant. Taken together, these contributions raise the evaluation score to approximately three times that of the mean-only variant. This ordering indicates that conditional utility and marginal uncertainty both inform route quality, but pairwise dependence contributes most by distinguishing expert sets with similar individual statistics but different co-movement.

The identification controls provide further empirical support for this interpretation. Permuting the expert-pair identities while preserving marginal information reduces the mean score by approximately 58%, and removing shadow coverage reduces it by 50%. The former localises the covariance benefit to correctly aligned expert pairs, whereas the latter shows that sampled off-route observations substantially improve coverage and calibration in this setting. By comparison, replacing the learned critic with oracle moments improves the mean score by 8%, while replacing beam search with exhaustive search yields a further 6% gain. The comparatively modest magnitude of these improvements suggests that, at the evaluated scale, accurate expert-pair identification and off-route coverage contribute more substantially to route quality than residual prediction and support-search errors. Appendices D.7 and D.8 provide the complete paired estimands and uncertainty intervals.

Table 2: Throughput of different methods across nominal routed-expert budgets k, with the routed-expert pool fixed at $N = 16$, on the matched 4×H200 setup. Throughput is in thousands of tokens per second and component times are seconds per step.

	End-to-end throughput (ktok/s) ↑				PoE step-time increase		PoE component time (s)		
k	Top-k MoE	ReMoE	MP-MoE	PoE	vs. Top-k MoE	vs. MP-MoE	Critic	Shadow	Solver
2	335.1	332.1	322.8	304.2	+10.14%	+6.10%	0.163	0.146	0.291
4	290.3	288.1	274.3	269.7	+7.64%	+1.69%	0.263	0.202	0.374
8	217.9	216.6	197.8	194.4	+12.08%	+1.76%	0.372	0.213	0.612

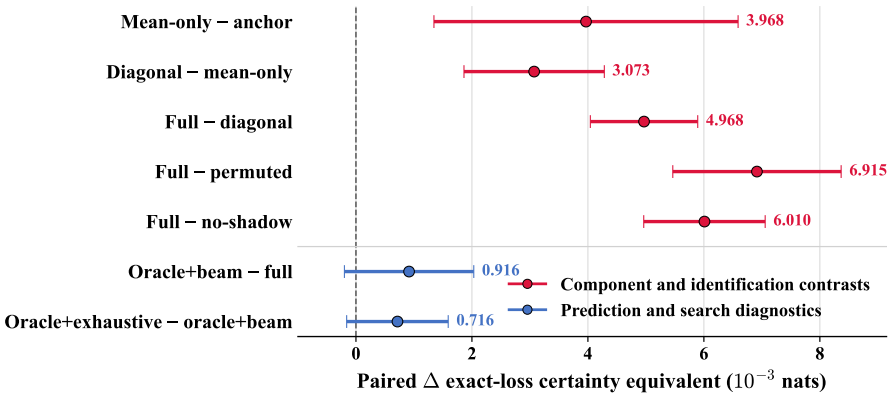


Figure 4: **Component identification under frozen downstream routes.** Points show paired mean changes in exact-loss certainty equivalent across five seeds. Bars are 95% t intervals. Red marks mechanism contrasts and blue marks prediction and search diagnostics.

4.4 CGR TARGET VALIDITY AND PRE-OUTCOME PREDICTION

The geometric interpretation in Figure 2 yields a direct empirical prediction. The projection $a_i(s)r_i^{\mathbf{a}}(s,\xi)$ should agree with the loss increase caused by deleting expert i at the same frozen route. We test this prediction first, then evaluate whether the critic can recover the return dependence needed before routing.

Figure 5 (a) is the empirical counterpart of Figure 2. It compares the allocation-weighted CGR $a_i(s)r_i^{\mathbf{a}}(s,\xi)$ with the exact deletion effect defined in Appendix D.5. The first-order target preserves the deletion ranking across depth despite larger late-layer magnitude errors, supporting its use as a local target and motivating the trust region. Figure 5 (b) shows that sampled shadow panels substantially improve portfolio-variance calibration on identical queries and outcomes. Figure 5 (c) compares dependence measures with oracle CGR correlation. The routing-time critic reaches a Spearman agreement of 0.967, close to the independent task-adjoint estimate of 0.993 and well above the geometric alternatives, showing that task-relevant dependence is not captured by representation similarity alone. Appendix D.6 specifies the calibration estimands and oracle construction.

5 RELATED WORK

Sparse MoE Routing. Sparsely gated MoE established fixed Top-k activation as a way to expand capacity without proportional per-token computation (Shazeer et al., 2017), and GShard and Switch Transformer made this design practical at scale (Lepikhin et al., 2021; Fedus et al., 2022). Subsequent work formulated routing as balanced or cardinality-constrained assignment (Lewis et al., 2021; Zhou et al., 2022; Liu et al., 2023), while DSelect-k introduced a continuously differentiable sparse gate (Hazimeh et al., 2021). We retain Top-k MoE as the canonical discrete baseline. ReMoE provides its continuous counterpart, replacing Top-k+softmax with independent ReLU gates and controlling sparsity and load through adaptive regularisation (Wang et al., 2025). These baselines test PoE against both fixed-support and dynamically sparse routing under matched budgets. Neither explicitly models task-level dependence among experts within the selected support. In contrast, PoE characterises support-level utility through the conditional mean and covariance of expert returns.

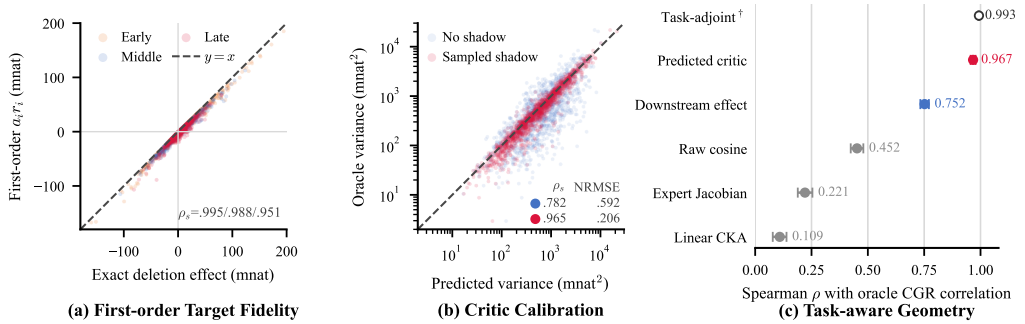


Figure 5: Validation of post-outcome targets and pre-outcome predictions across five training seeds. **(a)** Allocation-weighted CGR fidelity. **(b)** Portfolio-variance calibration. **(c)** Agreement with oracle task-adjoint geometry. Bars are 95% t intervals and † marks the identity check.

Expert Redundancy. Work on expert redundancy has developed along three complementary directions. Chi et al. diagnose router-induced representation collapse and mitigate it with low-dimensional hyperspherical routing (Chi et al., 2022). DeepSeekMoE instead promotes specialisation architecturally by decomposing experts into fine-grained routed units and isolating shared experts (Dai et al., 2024). Training-time methods impose geometric structure on learned experts using orthogonal parameter updates, CKA-based similarity penalties, or joint output-projection orthogonality and routing-score variance objectives (Liu et al., 2024; Do et al., 2025; Guo et al., 2025). Post-training, MC-SMoE groups experts using router-logit similarity before merging and compressing them, whereas HC-SMoE hierarchically clusters experts by functional output similarity and merges them without retraining (Li et al., 2024; Chen et al., 2025). Moving this progression into the routing decision, MP-MoE constructs a covariance matrix of binary expert-selection indicators from co-occurrence counts and greedily selects a support under a Mahalanobis objective (Kang et al., 2026). This formulation directly evaluates whether activation-derived dependence is sufficient to improve expert combinations. In contrast, PoE jointly selects and weights experts under the same sparse budget using the predicted conditional mean and covariance of token-level CGRs.

Portfolio-based Routing. Mean–variance optimisation provides a general framework for allocating a budget among assets with uncertain, correlated returns (Markowitz, 1952). Beyond finance, portfolio principles have been applied to document ranking by balancing expected relevance against variance-based risk, with document correlations entering the risk term, while predictive-to-prescriptive methods extend such decisions to context-conditioned outcomes (Wang & Zhu, 2009; Bertsimas & Kallus, 2020). Within neural networks, task-aware importance is commonly operationalised through the training loss, as in Taylor and mask-sensitivity criteria that use loss derivatives to approximate the effect of removing feature maps or attention heads (Molchanov et al., 2017; Michel et al., 2019). Negative-correlation learning couples individual prediction error to an error-correlation penalty (Liu & Yao, 1999), while determinantal point processes factor subset quality into item quality and diversity (Kulesza & Taskar, 2012). PoE combines these perspectives by treating signed, loss-sensitive CGRs as token-conditional returns and optimising their predicted mean and covariance under a fixed budget.

6 Conclusion

We introduced Portfolio of Experts (PoE), a sparse routing framework that selects expert groups jointly rather than ranking them independently. PoE defines counterfactual gradient returns to quantify local task utility and uses their predicted conditional mean and covariance to optimise a constrained allocation under a fixed budget. We show that return covariance characterises pairwise dependence among experts' local effects on task loss within a geometry induced by downstream gradient covariance, and provide bounds on approximation, prediction, and search errors. Across seven downstream tasks, PoE achieves the highest mean accuracy and lowest NLL across all expert pool sizes. Ablations verify the importance of pairwise covariance and shadow coverage. Taken together, these findings indicate that joint modelling of conditional expert utility and pairwise dependence enables more effective sparse routing with a favourable realised-support trade-off.

Reproducibility statement

Section 3 specifies the information set, anchor, counterfactual return, feasible allocation, critic, training chronology, and inference procedure. Appendices A and B provide proofs, panel propensities, and executable solver detail. Appendix D records experimental estimands, seed-level statistics, calibration, and optimisation diagnostics. The implementation code will be made available with the public version of this work.

Ethics statement

This work involves no human subjects and collects no new data. Models are trained on the Dolma dataset (ODC-By 1.0; Ettinger et al., 2025) and evaluated on ARC (CC BY-SA 4.0; Clark et al., 2018), BoolQ (CC BY-SA 3.0; Clark et al., 2019), HellaSwag (MIT; Zellers et al., 2019), LAMBADA (CC BY 4.0; Paperno et al., 2016), PIQA (AFL 3.0; Bisk et al., 2020), and RACE (non-commercial research use only; Lai et al., 2017). We use these publicly available datasets solely for non-commercial academic research, provide attribution by citing the original publications, and do not redistribute any of the data, in full compliance with their respective licences and terms of use.

AI use statement

Generative AI tools were used exclusively for linguistic editing and stylistic refinement to enhance the clarity and readability of the manuscript. The authors retain full responsibility for the accuracy, integrity, and content of the submitted work and its accompanying artefacts.

References

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Manag. Sci.*, 66 (3):1025–1044, 2020. doi: 10.1287/MNSC.2018.3253. URL https://doi.org/10.1287/mnsc.2018.3253.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse moe via hierarchical clustering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/chen25aq.html.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/df4f371f1f89ec8ba5014b3310578048-Abstract-Conference.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1300. URL https://doi.org/10.18653/v1/n19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1280–1297. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.70. URL https://doi.org/10.18653/v1/2024.acl-long.70.

Giang Do, Hung Le, and Truyen Tran. Simsmoe: Toward efficient training mixture of experts via solving representational collapse. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 2012–2025. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.107. URL https://doi.org/10.18653/v1/2025.findings-naacl.107.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee F. Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, William Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. *CoRR*, abs/2512.13961, 2025. doi: 10.48550/ARXIV.2512.13961. URL https://doi.org/10.48550/arXiv.2512.13961.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better moe. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/4598de7d243d528e38eb0c5d8155fb52-Abstract-Conference.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul C. Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *CoRR*, abs/2607.15495, 2026. doi: 10.48550/ARXIV.2607.15495. URL https://doi.org/10.48550/arXiv.2607.15495.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed H. Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29335–29347, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5ac21cd0ef1b88e9848571aeb53551a-Abstract.html.

Xinlai Kang, Dunyao Xue, Zhengbo Wang, Chengshuo Du, Xinghao Chen, Hang Zhou, Hanting Chen, and Cheng Meng. Breaking the echo chamber: A dynamic ensemble pruning perspective on moe. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7FsbfQgti4.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5(2-3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.1561/2200000044.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 785–794. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1082. URL https://doi.org/10.18653/v1/d17-1082.

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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6265–6274. PMLR, 2021. URL http://proceedings.mlr.press/v139/lewis21a.html.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=eFWG9Cy3WK.

Boan Liu, Liang Ding, Li Shen, Keqin Peng, Yu Cao, Dazhao Cheng, and Dacheng Tao. Diversifying the mixture-of-experts representation for language models with orthogonal optimizer. In Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarín Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (eds.), *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 2966–2973. IOS Press, 2024. doi: 10.3233/FAIA240836. URL https://doi.org/10.3233/FAIA240836.

Feilong Liu. Geometric asymmetry in moe specialization: Functional decorrelation and representational overlap. *CoRR*, abs/2605.16349, 2026. doi: 10.48550/ARXIV.2605.16349. URL https://doi.org/10.48550/arXiv.2605.16349.

Tianlin Liu, Joan Puigcerver, and Mathieu Blondel. Sparsity-constrained optimal transport. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=yHY9NbQJ5BP.

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999. doi: 10.1016/S0893-6080(99)00073-8. URL https://doi.org/10.1016/S0893-6080(99)00073-8.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: 10.1111/j.1540-6261.1952.tb01525.x. URL https://doi.org/10.1111/j.1540-6261.1952.tb01525.x.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman

Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14014–14024, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJGCiw5gl.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1144. URL https://doi.org/10.18653/v1/p16-1144.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Huiyuan Tian, Bonan Xu, and Shijian Li. Beyond geometric complementarity: Coherent overlap in sparse mixture-of-experts routing. *CoRR*, abs/2607.28308, 2026. doi: 10.48550/ARXIV.2607.28308. URL https://doi.org/10.48550/arXiv.2607.28308.

Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel (eds.), *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pp. 115–122. ACM, 2009. doi: 10.1145/1571941.1571963. URL https://doi.org/10.1145/1571941.1571963.

Ziteng Wang, Jun Zhu, and Jianfei Chen. Remoe: Fully differentiable mixture-of-experts with relu routing. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=4D0f16Vwc3.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL https://doi.org/10.18653/v1/p19-1472.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

A Proofs for conditional portfolio routing

This section establishes the exact identities and approximation bounds underlying the PoE method presented in Section 3. Additional proof-specific quantities are defined when first used. The layer index remains suppressed.

A.1 FORMAL DECISION SETUP

All random quantities are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The routing-information space $\mathcal{S}$ and outcome space $\mathcal{X}$ are assumed to be standard Borel spaces. Let $P_S := \mathbb{P} \circ S^{-1}$ denote the distribution of S, and fix a regular conditional probability kernel $s \mapsto P_{\Xi|S=s}$ for Ξ given S.

Every allocation rule considered in the conditional identities is $\sigma(S)$-measurable: it may depend on the routing information S, but not on the subsequently realised outcome Ξ. For the teacher-forced setting analysed here, the outcome is exogenous to the current allocation: on one common probability space,

$$\Xi(\mathbf{c}) = \Xi \quad \text{almost surely for every feasible intervention } \mathbf{c}.$$

Thus the paired counterfactual holds the same realised outcome ξ fixed while changing the allocation; equality of intervention-specific outcome laws alone would not be sufficient for this pathwise statement. Random masks or other auxiliary seeds are either included in S or represented by an exogenous variable coupled identically across the compared interventions. Free-running generation with a route-dependent outcome law requires a different causal analysis.

For a context s and a frozen point allocation $\mathbf{c} \in \mathcal{B}(s)$, $\mathcal{L}^{[\mathbf{c}]}(\cdot; s, \xi)$ denotes the downstream loss on the computational graph whose discrete routes and stochastic masks are those realised at $\mathbf{y}(\mathbf{c}; s)$. These graph choices remain fixed while the argument $\mathbf{y}$ varies. The loss is assumed differentiable at $\mathbf{y}(\mathbf{c}; s)$ whenever a CGR is formed. Stronger smoothness is stated in the results that need it. Conditional moments always refer to a fixed task-model snapshot, with model parameters omitted from the notation. Define the pointwise generic-allocation quantities

$$\begin{aligned} \mathbf{y}(s; \mathbf{c}) &:= \mathbf{y}(\mathbf{c}; s), \\ \mathbf{g}(s, \xi; \mathbf{c}) &:= \nabla_{\mathbf{y}} \mathcal{L}^{[\mathbf{c}]}(\mathbf{y}(s; \mathbf{c}); s, \xi), \\ \mathbf{r}(s, \xi; \mathbf{c}) &:= -\mathbf{E}(s)^\top \mathbf{g}(s, \xi; \mathbf{c}). \end{aligned} \tag{20}$$

Its point-allocation conditional moments are

$$\boldsymbol{\mu}^{\mathbf{c}}(s) := \mathbb{E}[\mathbf{r}(s, \Xi; \mathbf{c}) \mid S = s], \qquad \boldsymbol{\Sigma}^{\mathbf{c}}(s) := \mathrm{Cov}(\mathbf{r}(s, \Xi; \mathbf{c}) \mid S = s).$$

For any pre-outcome measurable allocation rule $\boldsymbol{\rho} : \mathcal{S} \to \mathbb{R}^N_{\geq 0}$ with $\boldsymbol{\rho}(S) \in \mathcal{B}(S)$ almost surely, use the shorthand

$$\mathbf{g}^{\boldsymbol{\rho}}(s, \xi) := \mathbf{g}(s, \xi; \boldsymbol{\rho}(s)), \qquad \mathbf{r}^{\boldsymbol{\rho}}(s, \xi) := \mathbf{r}(s, \xi; \boldsymbol{\rho}(s)), \qquad \mathbf{R}^{\boldsymbol{\rho}} := \mathbf{r}(S, \Xi; \boldsymbol{\rho}(S)).$$

The anchor quantities in Equation 7 use $\boldsymbol{\rho} = \mathbf{a}$. At update n, $\mathbf{g}^{\mathbf{p}_n}(s_n, \xi_n)$ is the pointwise specialisation $\mathbf{g}(s_n, \xi_n; \mathbf{p}_n)$. Whenever covariance is used, we assume the corresponding random return belongs to $L^2(\mathbb{P}; \mathbb{R}^N)$. Statements involving regular conditional moments are understood for P_S-almost every s.

A.2 TRUST-REGION CGR BOUND

For a candidate allocation $\mathbf{p}(s)$, define the exact frozen-route loss reduction from the anchor and its first-order residual by

$$D^{\mathbf{a}}_{\mathbf{p}}(s, \xi) := \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi) - \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s) + \mathbf{E}(s)\boldsymbol{\delta}(s); s, \xi)\,, \tag{21}$$

$$\mathcal{R}^{\mathbf{a}}_{\mathbf{p}}(s, \xi) := D^{\mathbf{a}}_{\mathbf{p}}(s, \xi) - \boldsymbol{\delta}(s)^\top \mathbf{r}^{\mathbf{a}}(s, \xi). \tag{22}$$

For the deletion diagnostic in Section 4.4, define

$$\Delta^{\mathbf{a}}_i(s, \xi) := \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s) - a_i(s)\mathbf{e}_i(s); s, \xi) - \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi). \tag{23}$$

Thus $\Delta^{\mathbf{a}}_i > 0$ means that deleting expert i increases the loss.

Proposition 2 (Frozen-route trust-region approximation)**.** *Fix $s \in \mathcal{S}$, $\xi \in \mathcal{X}$, and a candidate allocation $\mathbf{p}(s)$. Suppose $\mathcal{L}^{[\mathbf{a}]}(\cdot; s, \xi)$ is twice continuously differentiable along the segment joining $\mathbf{y}^{\mathbf{a}}(s)$ and $\mathbf{y}^{\mathbf{a}}(s) + \mathbf{E}(s)\boldsymbol{\delta}(s)$. Let $H_{\mathbf{p}}(s, \xi) < \infty$ be the supremum of the Hessian operator norm on that segment. Then*

$$\left|\mathcal{R}^{\mathbf{a}}_{\mathbf{p}}(s, \xi)\right| = \left|D^{\mathbf{a}}_{\mathbf{p}}(s, \xi) - \boldsymbol{\delta}(s)^\top \mathbf{r}^{\mathbf{a}}(s, \xi)\right| \leq \frac{H_{\mathbf{p}}(s, \xi)}{2} \|\mathbf{E}(s)\boldsymbol{\delta}(s)\|_2^2. \tag{24}$$

Proof of Proposition 2. Fix (s, ξ) and abbreviate $\mathbf{h} = \mathbf{E}(s)\boldsymbol{\delta}(s) \in \mathbb{R}^d$. For some $\widetilde{\lambda} \in [0, 1]$, let

$$\mathbf{H}_{\widetilde{\lambda}} = \nabla^2_{\mathbf{y}} \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s) + \widetilde{\lambda}\mathbf{h}; s, \xi).$$

Taylor's theorem along this segment gives

$$\mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}} + \mathbf{h}; s, \xi) = \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}; s, \xi) + (\mathbf{g}^{\mathbf{a}})^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \mathbf{H}_{\widetilde{\lambda}} \mathbf{h}$$

Negating the loss difference and using $\mathbf{r}^{\mathbf{a}} = -\mathbf{E}(s)^\top \mathbf{g}^{\mathbf{a}}$ yields

$$D^{\mathbf{a}}_{\mathbf{p}} = \boldsymbol{\delta}^\top \mathbf{r}^{\mathbf{a}} - \frac{1}{2}\mathbf{h}^\top \mathbf{H}_{\widetilde{\lambda}} \mathbf{h}.$$

The quadratic-form inequality $|\mathbf{h}^\top \mathbf{H}_{\widetilde{\lambda}} \mathbf{h}| \le \|\mathbf{H}_{\widetilde{\lambda}}\|_{\mathrm{op}} \|\mathbf{h}\|_2^2$ proves Equation 24. Deleting expert i sets $\boldsymbol{\delta} = -a_i(s)\boldsymbol{\iota}_i$, where $\boldsymbol{\iota}_i \in \mathbb{R}^N$ is the ith standard basis vector. For $\mathbf{p}_{\setminus i}(s) = \mathbf{a}(s) - a_i(s)\boldsymbol{\iota}_i$, $D^{\mathbf{a}}_{\mathbf{p}_{\setminus i}}(s, \xi) = -\Delta^{\mathbf{a}}_i(s, \xi)$. Let $H_{\mathrm{del},i}(s, \xi)$ be the supremum Hessian norm on this deletion segment. Substitution gives

$$|\Delta^{\mathbf{a}}_i(s, \xi) - a_i(s) r^{\mathbf{a}}_i(s, \xi)| \le \frac{H_{\mathrm{del},i}(s, \xi)}{2} a_i(s)^2 \|\mathbf{e}_i(s)\|_2^2.$$

□

A.3 Conditional portfolio identities

Proposition 3 (Contextual portfolio identities)**.** *Let $\mathbf{R}^{\mathbf{a}} \in L^2(\mathbb{P}; \mathbb{R}^N)$ and let $\mathbf{p}$ be a bounded, $\sigma(S)$-measurable candidate rule. Then the conditional mean and variance are given by Equation 11 and Equation 12. Moreover,*

$$\mathrm{Var}(Y^{\mathbf{a}}_{\mathbf{p}}) = \mathbb{E}\big[\mathbf{p}(S)^\top \boldsymbol{\Sigma}^{\mathbf{a}}(S)\mathbf{p}(S)\big] + \mathrm{Var}\big(\mathbf{p}(S)^\top \boldsymbol{\mu}^{\mathbf{a}}(S)\big).$$

Proof of Proposition 3. Conditional on S, the allocation $\mathbf{p}(S)$ is fixed because it is S-measurable. Conditional linearity gives

$$\mathbb{E}[\mathbf{p}^\top \mathbf{R}^{\mathbf{a}} \mid S] = \mathbf{p}^\top \mathbb{E}[\mathbf{R}^{\mathbf{a}} \mid S] = \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}.$$

Applying the covariance identity for a fixed linear projection gives

$$\mathrm{Var}(\mathbf{p}^\top \mathbf{R}^{\mathbf{a}} \mid S) = \mathbf{p}^\top \mathrm{Cov}(\mathbf{R}^{\mathbf{a}} \mid S)\mathbf{p} = \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}} \mathbf{p}.$$

The unconditional decomposition follows from the law of total variance. Finally, for every $\mathbf{v} \in \mathbb{R}^N$,

$$\mathbf{v}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{v} = \mathbb{E}\big[\{\mathbf{v}^\top(\mathbf{R}^{\mathbf{a}} - \boldsymbol{\mu}^{\mathbf{a}}(s))\}^2 \mid S = s\big] \ge 0,$$

so the conditional covariance is positive semidefinite. □

Proof of Proposition 1. Conditional on $S = s$, the expert-output matrix $\mathbf{E}(s)$ is fixed and $\mathbf{R}^{\mathbf{a}} = -\mathbf{E}(s)^\top \mathbf{g}^{\mathbf{a}}$. Conditional linearity and the congruence rule for covariances give

$$\boldsymbol{\mu}^{\mathbf{a}} = -\mathbf{E}^\top \mathbf{m}^{\mathbf{a}}_g, \qquad \boldsymbol{\Sigma}^{\mathbf{a}} = \mathbf{E}^\top \mathbf{G}^{\mathbf{a}} \mathbf{E}.$$

In particular, $\mathrm{Cov}(R^{\mathbf{a}}_i, R^{\mathbf{a}}_j \mid S = s) = \mathbf{e}_i^\top \mathbf{G}^{\mathbf{a}} \mathbf{e}_j$, and the two diagonal quadratic forms are the corresponding conditional variances. When both variances are positive, normalisation gives

$$\mathrm{Corr}(R^{\mathbf{a}}_i, R^{\mathbf{a}}_j \mid S = s) = \frac{\mathbf{e}_i(s)^\top \mathbf{G}^{\mathbf{a}}(s)\mathbf{e}_j(s)}{\sqrt{\mathbf{e}_i(s)^\top \mathbf{G}^{\mathbf{a}}(s)\mathbf{e}_i(s)}\sqrt{\mathbf{e}_j(s)^\top \mathbf{G}^{\mathbf{a}}(s)\mathbf{e}_j(s)}}. \tag{25}$$

Because $\mathbf{G}^{\mathbf{a}} \succeq 0$ can be singular, the ratio in Equation 25 is a cosine in a seminorm, equivalently in the quotient by $\ker \mathbf{G}^{\mathbf{a}}$. If a diagonal quadratic form vanishes, Pearson correlation is undefined even though the covariance identity remains valid. By contrast, normalising $\mathbb{E}[R^{\mathbf{a}}_i R^{\mathbf{a}}_j \mid S = s]$ gives the analogous bilinear form induced by the raw gradient second moment $\mathbb{E}[\mathbf{g}^{\mathbf{a}}(\mathbf{g}^{\mathbf{a}})^\top \mid S = s]$. In general this is not conditional correlation, although the two coincide when the relevant conditional means vanish. Ordinary output-space cosine is recovered throughout the expert span $U_s := \mathrm{span}\{\mathbf{e}_i(s) : i \in [N]\}$ when $\mathbf{G}^{\mathbf{a}}(s)|_{U_s} = c_s \mathbf{I}_{U_s}$ for some $c_s > 0$; the positivity condition excludes the zero-variance case. □

A.4 Local token lens and Jacobian-lens distinction

Let $M \in \mathbb{N}_{>0}$ be the number of downstream logit coordinates affected by the current layer output. For a realised context s and outcome ξ, let $\mathbf{z}^{[\mathbf{a}]}(\mathbf{y}; s, \xi) \in \mathbb{R}^M$ collect those logits on the graph frozen at the anchor. Let $\ell : \mathbb{R}^M \times \mathcal{X} \to \mathbb{R}$ be the scalar loss as a function of these logits and the outcome. Thus

$$\mathcal{L}^{[\mathbf{a}]}(\mathbf{y}; s, \xi) = \ell(\mathbf{z}^{[\mathbf{a}]}(\mathbf{y}; s, \xi), \xi).$$

At $\mathbf{y} = \mathbf{y}^{\mathbf{a}}(s)$, define the downstream logit Jacobian by

$$\mathbf{A}^{\mathbf{a}}(s, \xi) := \left.\frac{\partial \mathbf{z}^{[\mathbf{a}]}(\mathbf{y}; s, \xi)}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{y}^{\mathbf{a}}(s)} \in \mathbb{R}^{M \times d}.$$

Define the logit loss gradient by

$$\mathbf{q}^{\mathbf{a}}(s, \xi) := \nabla_{\mathbf{z}} \ell(\mathbf{z}, \xi)|_{\mathbf{z}=\mathbf{z}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi)} \in \mathbb{R}^M.$$

The chain rule gives

$$\mathbf{g}^{\mathbf{a}} = (\mathbf{A}^{\mathbf{a}})^\top \mathbf{q}^{\mathbf{a}}, \qquad r_i^{\mathbf{a}} = -\langle \underbrace{\mathbf{A}^{\mathbf{a}} \mathbf{e}_i}_{\boldsymbol{\phi}_i^{\mathbf{a}} \in \mathbb{R}^M}, \mathbf{q}^{\mathbf{a}} \rangle. \tag{26}$$

The vector $\boldsymbol{\phi}_i^{\mathbf{a}}$ is the first-order effect of expert i on the affected logits. Reverse-mode automatic differentiation computes $\mathbf{g}^{\mathbf{a}}$ without constructing the full Jacobian.

For one realised outcome, Equation 26 gives the exact raw-product identity

$$r_i^{\mathbf{a}} r_j^{\mathbf{a}} = (\boldsymbol{\phi}_i^{\mathbf{a}})^\top \mathbf{q}^{\mathbf{a}} (\mathbf{q}^{\mathbf{a}})^\top \boldsymbol{\phi}_j^{\mathbf{a}} = \mathbf{e}_i^\top \mathbf{g}^{\mathbf{a}} (\mathbf{g}^{\mathbf{a}})^\top \mathbf{e}_j. \tag{27}$$

Both outer products in this display have rank at most one. The display is one observation of a raw second moment. By itself it is not a conditional covariance.

Let $\boldsymbol{\iota}_m \in \mathbb{R}^M$ be the mth standard basis vector, and define

$$\mathbf{j}_m^{\mathbf{a}} = (\mathbf{A}^{\mathbf{a}})^\top \boldsymbol{\iota}_m, \qquad \phi_{i,m}^{\mathbf{a}} = \langle \mathbf{j}_m^{\mathbf{a}}, \mathbf{e}_i \rangle.$$

The coordinate form of the task-adjoint factorisation is

$$\mathbf{g}^{\mathbf{a}} = \sum_{m=1}^{M} q_m^{\mathbf{a}} \mathbf{j}_m^{\mathbf{a}}, \qquad r_i^{\mathbf{a}} = -\sum_{m=1}^{M} q_m^{\mathbf{a}} \phi_{i,m}^{\mathbf{a}}. \tag{28}$$

For autoregressive cross-entropy, let $\mathcal{U}$ be the common padded set of potentially affected future positions and $\mathcal{V}$ the vocabulary, with $M = |\mathcal{U}||\mathcal{V}|$ and $m = (u, w) \in \mathcal{U} \times \mathcal{V}$. If $\beta_u \geq 0$ is the position weight, $t_u \in \mathcal{V}$ the target, and $\pi_{u,w}^{\mathbf{a}}$ the anchor softmax probability, then

$$q_{u,w}^{\mathbf{a}} = \beta_u \big(\pi_{u,w}^{\mathbf{a}} - \mathbb{1}\{w = t_u\}\big).$$

Absent positions have $\beta_u = 0$. Thus CGR is a signed, task-error-weighted readout of the local downstream Jacobian. This differs from the Jacobian lens, which averages residual-stream-to-final-residual Jacobians across prompts and positions, composes the result with the unembedding, and obtains J-space components through sparse nonnegative decomposition (Gurnee et al., 2026). Liu defines the expert-local Jacobian as $D_{\mathbf{x}} \mathcal{E}_i(\mathbf{x})$, independently of routing (Liu, 2026); this is not the same object as $\mathbf{A}^{\mathbf{a}} \mathbf{e}_i$, our allocation derivative of downstream logits.

Let ν_J be the corpus-position law used by the Jacobian lens, $\mathbb{E}_J$ its expectation, and $\mathbf{A}_J \in L^2(\nu_J; \mathbb{R}^{M_J \times d})$ a fixed-shape row-aligned logit Jacobian, where $M_J \in \mathbb{N}_{>0}$. Then

$$\begin{aligned} (\mathbb{E}_J[\mathbf{A}_J])^\top \mathbb{E}_J[\mathbf{A}_J] &\neq \mathbb{E}_J[\mathbf{A}_J^\top \mathbf{A}_J] \quad \text{in general,} \\ \mathrm{Cov}(\mathbf{g}^{\mathbf{a}} \mid S = s) &= \mathbb{E}[(\mathbf{A}^{\mathbf{a}})^\top \mathbf{q}^{\mathbf{a}} (\mathbf{q}^{\mathbf{a}})^\top \mathbf{A}^{\mathbf{a}} \mid S = s] \\ &\quad - \mathbf{m}_g^{\mathbf{a}}(s) \mathbf{m}_g^{\mathbf{a}}(s)^\top. \end{aligned}$$

The first line separates an averaged Jacobian from an averaged Jacobian Gram matrix. The second is the exact conditional covariance of the local task-adjoint gradient.

A.5 CARA REDUCTION AND REMAINDER

For a scalar random return X and $\eta > 0$, define the conditional constant absolute risk aversion certainty equivalent by

$$\mathrm{CE}_\eta(X \mid S = s) := -\frac{1}{\eta} \log \mathbb{E}[\exp(-\eta X) \mid S = s], \tag{29}$$

and set $\mathrm{CE}_0(X \mid S = s) := \mathbb{E}[X \mid S = s]$. We also write $b := \eta/2$ for the portfolio variance coefficient used below.

Fix a context s and a candidate vector $\mathbf{p} \in \mathcal{A}(s; \mathbf{a})$. In this subsection, write

$$Y^{\mathbf{a}}_{\mathbf{p},s} := \mathbf{p}^\top \mathbf{r}^{\mathbf{a}}(s, \Xi).$$

Thus $Y^{\mathbf{a}}_{\mathbf{p},s}$ is a scalar random variable only through $\Xi \sim P_{\Xi|S=s}$. Its law is the conditional law of $Y^{\mathbf{a}}_{\mathbf{p}}$ in Equation 10 given $S = s$.

Conditional Gaussian case. Write $m = \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}(s)$ and $v = \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p}$. If $\mathbf{R}^{\mathbf{a}} \mid S = s$ is Gaussian, then $Y^{\mathbf{a}}_{\mathbf{p},s}$ is Gaussian with mean m and variance v, including when $v = 0$. Its moment-generating function gives

$$\mathbb{E}[e^{-\eta Y^{\mathbf{a}}_{\mathbf{p},s}} \mid S = s] = \exp(-\eta m + \eta^2 v/2).$$

Substitution into Equation 29 gives

$$\mathrm{CE}_\eta(Y^{\mathbf{a}}_{\mathbf{p}} \mid S = s) = \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}(s) - \frac{\eta}{2} \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p}. \tag{30}$$

□

General cumulant remainder. Write $m = \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}(s)$ and $v = \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p}$. Let $K_{s,\mathbf{p}}(t) = \log \mathbb{E}[\exp(t(Y^{\mathbf{a}}_{\mathbf{p},s} - m)) \mid S = s]$. Assume $\eta > 0$, $0 < \mathbb{E}[\exp(-\eta Y^{\mathbf{a}}_{\mathbf{p},s}) \mid S = s] < \infty$, and that $K_{s,\mathbf{p}}$ is three times continuously differentiable on $[-\eta, 0]$. Then $K_{s,\mathbf{p}}(0) = K'_{s,\mathbf{p}}(0) = 0$ and $K''_{s,\mathbf{p}}(0) = v$. Taylor's theorem at $t = -\eta$ gives

$$K_{s,\mathbf{p}}(-\eta) = \frac{\eta^2}{2} v - \frac{\eta^3}{6} K'''_{s,\mathbf{p}}(\widetilde{t})$$

for some $\widetilde{t} \in [-\eta, 0]$. Since

$$\mathrm{CE}_\eta = m - \eta^{-1} K_{s,\mathbf{p}}(-\eta),$$

where the left-hand side is $\mathrm{CE}_\eta(Y^{\mathbf{a}}_{\mathbf{p},s} \mid S = s)$. Define

$$M_3(s, \mathbf{p}, \eta) := \sup_{t \in [-\eta, 0]} |K'''_{s,\mathbf{p}}(t)|.$$

Whenever this quantity is finite, the difference from $m - \eta v/2$ has magnitude at most $\eta^2 M_3(s, \mathbf{p}, \eta)/6$. □

The portfolio certainty equivalent above concerns the final aggregate return. The local loss reduction uses a different random variable. Define

$$\boldsymbol{\delta} := \mathbf{p} - \mathbf{a}(s), \qquad Z^{\mathbf{a}}_{\mathbf{p},s} := \boldsymbol{\delta}^\top \mathbf{r}^{\mathbf{a}}(s, \Xi).$$

The variable $Z^{\mathbf{a}}_{\mathbf{p},s}$ is the linearised loss reduction from the anchor to the candidate. Let $L_E(s) < \infty$ satisfy $\|\mathbf{E}(s)\|_{\mathrm{op}} \le L_E(s)$.

Remark 4 (Final-portfolio risk and reallocation risk). The mean terms based on $\mathbf{p}$ and $\boldsymbol{\delta} = \mathbf{p} - \mathbf{a}$ rank candidates identically because $\mathbf{a}^\top \boldsymbol{\mu}^{\mathbf{a}}$ is constant. Their variances do not:

$$\mathrm{Var}(Z^{\mathbf{a}}_{\mathbf{p},s} \mid S = s) = \mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}} \mathbf{p} - 2\mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}} \mathbf{a} + \mathbf{a}^\top \boldsymbol{\Sigma}^{\mathbf{a}} \mathbf{a}.$$

PoE deliberately prices the uncertainty of the *final deployed expert portfolio* $Y^{\mathbf{a}}_{\mathbf{p}}$; the anchor is the point at which its local returns are measured, not a risky holding whose transition variance is optimised. Hence Equation 14 is not asserted to be the certainty equivalent of the reallocation gain $Z^{\mathbf{a}}_{\mathbf{p},s}$ or of the exact loss reduction $D^{\mathbf{a}}_{\mathbf{p}}$. An alternative objective for that different decision criterion would replace the quadratic term by $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{\mathbf{a}} \boldsymbol{\delta}$.

Proposition 5 (Local loss linearisation error)**.** *Let* $\mathbf{p} \in \mathcal{A}(s;\mathbf{a})$*. Assume* $H_{\mathbf{p}}(s,\Xi) \le \overline{H}(s)$ *almost surely conditional on* $S = s$*. Then*

$$\left|D^{\mathbf{a}}_{\mathbf{p}}(s,\Xi) - Z^{\mathbf{a}}_{\mathbf{p},s}\right| \le \frac{\overline{H}(s)L_E(s)^2}{2}\tau^2 \quad \textit{almost surely}.$$

Consequently, their conditional means differ by at most $\overline{H}(s)L_E(s)^2\tau^2/2$*. For every* $\eta \ge 0$ *for which the certainty equivalents are finite, they also satisfy*

$$\left|\mathrm{CE}_\eta(D^{\mathbf{a}}_{\mathbf{p}} \mid S = s) - \mathrm{CE}_\eta(Z^{\mathbf{a}}_{\mathbf{p},s} \mid S = s)\right| \le \frac{\overline{H}(s)L_E(s)^2}{2}\tau^2.$$

Proof. The pointwise bound is Proposition 2 combined with the allocation-space trust constraint in Equation 13 and $\|\mathbf{E}(s)\|_{\mathrm{op}} \le L_E(s)$. The mean claim follows from $|\mathbb{E}[X] - \mathbb{E}[Y]| \le \mathbb{E}|X - Y|$. For $\eta > 0$, the pointwise inequalities $Z - c \le D \le Z + c$, where $c = \overline{H}(s)L_E(s)^2\tau^2/2$, together with monotonicity and translation equivariance of the CARA certainty equivalent give the final claim. The case $\eta = 0$ is the mean bound. □

A.6 Moment and solver error

For fixed s and fixed deployed critic parameters $\boldsymbol{\omega}$, abbreviate

$$\widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s) := \widehat{\boldsymbol{\mu}}^{\mathbf{a}(s)}_{\boldsymbol{\omega}}(s),$$
$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{stat}}(s) := \widehat{\boldsymbol{\Sigma}}^{\mathbf{a}(s)}_{\mathrm{stat},\boldsymbol{\omega}}(s),$$
$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{route}}(s) := \widehat{\boldsymbol{\Sigma}}^{\mathbf{a}(s)}_{\mathrm{route},\boldsymbol{\omega}}(s).$$

For $\mathbf{p} \in \mathcal{A}(s;\mathbf{a})$, define

$$J_s(\mathbf{p}) = \mathbf{p}^\top \boldsymbol{\mu}^{\mathbf{a}}(s) - b\mathbf{p}^\top \boldsymbol{\Sigma}^{\mathbf{a}}(s)\mathbf{p},$$
$$\widehat{J}_s(\mathbf{p}) = \mathbf{p}^\top \widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s) - b\mathbf{p}^\top \widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{route}}(s)\mathbf{p}.$$

The routed covariance includes the solver ridge $\epsilon\mathbf{I}_N$. Consequently $\widehat{J}_s$ contains the explicit regulariser $-b\epsilon\|\mathbf{p}\|_2^2$; the ridge improves numerical conditioning but is not statistical return variance.

Proposition 6 (Uniform mean–variance approximation)**.** *Fix* s *and* $\eta > 0$*, and suppose*

$$\overline{M}_3(s,\eta) := \sup_{\mathbf{p}\in\mathcal{A}(s;\mathbf{a})} M_3(s,\mathbf{p},\eta) < \infty.$$

Then the difference between $\mathrm{CE}_\eta(Y^{\mathbf{a}}_{\mathbf{p}} \mid S = s)$ *and* $J_s(\mathbf{p})$ *is uniformly at most* $\eta^2\overline{M}_3(s,\eta)/6$*. If maximisers* $\mathbf{p}_{\mathrm{CE}}$ *and* $\mathbf{p}_{\mathrm{MV}}$ *of the two criteria exist, then*

$$\mathrm{CE}_\eta(Y^{\mathbf{a}}_{\mathbf{p}_{\mathrm{CE}}} \mid S = s) - \mathrm{CE}_\eta(Y^{\mathbf{a}}_{\mathbf{p}_{\mathrm{MV}}} \mid S = s) \le \frac{\eta^2}{3}\overline{M}_3(s,\eta).$$

Proof. The first statement is the preceding cumulant bound made uniform over the feasible set. Adding and subtracting J_s at the two maximisers gives two uniform approximation errors; optimality of $\mathbf{p}_{\mathrm{MV}}$ for J_s makes the remaining term nonpositive. □

Proposition 7 (Score and routing regret under moment error)**.** *Suppose*

$$\|\widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s) - \boldsymbol{\mu}^{\mathbf{a}}(s)\|_2 \le \varepsilon_\mu(s), \qquad \|\widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{stat}}(s) - \boldsymbol{\Sigma}^{\mathbf{a}}(s)\|_{\mathrm{op}} \le \varepsilon_{\Sigma,\mathrm{stat}}(s).$$

Then, uniformly over $\mathcal{A}(s;\mathbf{a})$*,*

$$|\widehat{J}_s(\mathbf{p}) - J_s(\mathbf{p})| \le B\varepsilon_\mu(s) + bB^2\big(\varepsilon_{\Sigma,\mathrm{stat}}(s) + \epsilon\big) =: \varepsilon_J(s). \tag{31}$$

If the practical solver returns a feasible $\widehat{\mathbf{p}} \in \mathcal{A}(s;\mathbf{a})$ *whose predicted score is within* $\varepsilon_{\mathrm{opt}}(s) \ge 0$ *of the maximum predicted score, and* $\mathbf{p}^\star$ *maximises the true score, then*

$$J_s(\mathbf{p}^\star) - J_s(\widehat{\mathbf{p}}) \le 2\varepsilon_J(s) + \varepsilon_{\mathrm{opt}}(s).$$

Proof. The Cauchy Schwarz inequality bounds the linear error by $B\varepsilon_\mu$ because every feasible allocation satisfies $\|\mathbf{p}\|_2 \le \|\mathbf{p}\|_1 \le B$. The operator-norm inequality bounds the quadratic error by $bB^2(\varepsilon_{\Sigma,\mathrm{stat}} + \epsilon)$, proving Equation 31. Add and subtract the predicted scores at $\mathbf{p}^\star$ and $\widehat{\mathbf{p}}$. The two uniform estimation errors and the solver gap give the regret result. □

The two error results address different parts of the construction. Proposition 5 controls the local interpretation of the mean utility term through the loss reduction $D^{\mathbf{a}}_{\mathbf{p}}$. Proposition 7 controls estimation and optimisation error for the final portfolio score $J_s(\mathbf{p})$. The risk term in J_s is the variance of $Y^{\mathbf{a}}_{\mathbf{p}}$.

B Shadow-panel identification and solver

B.1 A concrete panel design

At training update n, consider one observation for one layer and one token (S_n, Ξ_n) with realisation (s_n, ξ_n) and frozen executed route $\mathbf{p}_n$ of the random pre-outcome allocation $\mathbf{P}_n$. Let $\mathcal{H}_{n-1} \subseteq \mathcal{F}$ be the sigma-field generated by the pre-route training history. With current anchor-router parameters θ_n, define $\mathbf{a}_n := \mathbf{a}_{\theta_n}(s_n)$. The allocation $\mathbf{P}_n$ is the $\sigma(S_n, \mathcal{H}_{n-1})$-measurable solver output queried at that anchor, and $\mathbf{p}_n$ is its realised value before Ξ_n is observed. Let $\mathcal{I}_n = \mathrm{supp}(\mathbf{P}_n)$ be its random executed support, whose realisation is $\mathrm{supp}(\mathbf{p}_n)$, and choose an outer probe rate $\alpha_0 \in (0, 1]$ and a conditional shadow rate $\alpha \in (0, 1]$. Every $i \in \mathcal{I}_n$ is observed on every update. Before Ξ_n is observed, draw one shared outer gate $F_n \sim \mathrm{Bernoulli}(\alpha_0)$ and, independently for each $i \notin \mathcal{I}_n$, draw $B_{n,i} \sim \mathrm{Bernoulli}(\alpha)$. The inactive expert is observed exactly when $F_n B_{n,i} = 1$. All these draws are conditionally independent of the ensuing return vector given $(S_n, \mathbf{P}_n, \mathcal{H}_{n-1})$. The exact marginal and joint propensities are therefore

$$\pi^{\mathrm{obs}}_{n,i} = \begin{cases} 1, & i \in \mathcal{I}_n, \\ \alpha_0\alpha, & i \notin \mathcal{I}_n, \end{cases} \qquad \pi^{\mathrm{obs}}_{n,ij} = \begin{cases} 1, & i, j \in \mathcal{I}_n, \\ \alpha_0\alpha, & |\{i,j\} \setminus \mathcal{I}_n| = 1, \\ \alpha_0\alpha^2, & i, j \notin \mathcal{I}_n, \end{cases} \quad (i \neq j), \qquad \pi^{\mathrm{obs}}_{n,ii} = \pi^{\mathrm{obs}}_{n,i}.$$

Every pair has positive inclusion probability, and the expected number of additional expert evaluations is $\alpha_0\alpha(N - |\mathcal{I}_n|)$ per routed-layer token. In particular, two inactive experts share the outer gate, so their joint propensity is $\alpha_0\alpha^2$, not the product $(\alpha_0\alpha)^2$ of their marginal propensities. For the following proposition, define the conditional random return

$$\mathbf{R}_n = \mathbf{r}(S_n, \Xi_n; \mathbf{P}_n), \qquad R_{n,i} = [\mathbf{R}_n]_i,$$

and let $r_{n,i}$ denote its realisation after ξ_n is observed. For the sampled panel $\mathcal{Q}_n$, define the coordinate-observation indicator $O_{n,i} := \mathbb{1}\{i \in \mathcal{Q}_n\}$.

For the executed-allocation query, define the predicted mean and raw second moment by

$$\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\omega}) := \widehat{\boldsymbol{\mu}}^{\mathbf{p}_n}_{\boldsymbol{\omega}}(s_n), \tag{32}$$

$$\widehat{\mathbf{Q}}_n(\boldsymbol{\omega}) := \widehat{\boldsymbol{\Sigma}}^{\mathbf{p}_n}_{\mathrm{stat},\boldsymbol{\omega}}(s_n) + \widehat{\boldsymbol{\mu}}_n(\boldsymbol{\omega})\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\omega})^\top. \tag{33}$$

For a second-moment weight $\lambda_2 > 0$, the inverse-propensity moment loss is

$$\mathcal{L}^{(n)}_{\mathrm{mom}}(\boldsymbol{\omega}) = \frac{1}{N}\sum_{i \in \mathcal{Q}_n} \frac{(r_{n,i} - [\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\omega})]_i)^2}{\pi^{\mathrm{obs}}_{n,i}} + \frac{\lambda_2}{N^2}\sum_{i,j \in \mathcal{Q}_n} \frac{(r_{n,i}r_{n,j} - [\widehat{\mathbf{Q}}_n(\boldsymbol{\omega})]_{ij})^2}{\pi^{\mathrm{obs}}_{n,ij}}. \tag{34}$$

Here $r_{n,i}$, $r_{n,i}r_{n,j}$, $\mathbf{p}_n$, the panel indicators, propensities, and critic input features are stop-gradient targets during the critic update. Thus $\mathcal{L}^{(n)}_{\mathrm{mom}}$ differentiates only with respect to $\boldsymbol{\omega}$ and does not create second-order derivatives through $\mathbf{g}^{\mathbf{p}_n}$. During route distillation, $\mathbf{p}_n$ and the layer features are likewise detached, so the auxiliary route loss updates only the anchor-router parameters.

Proposition 8 (Inverse-propensity identification)**.** *If the panel mask is conditionally independent of $\mathbf{R}_n$ given $(S_n, \mathbf{P}_n, \mathcal{H}_{n-1})$ and the propensities in Equation 34 are positive, then, for every conditionally integrable h_i and h_{ij},*

$$\mathbb{E}\left[\frac{O_{n,i}}{\pi^{\mathrm{obs}}_{n,i}} h_i(R_{n,i}) \,\middle|\, S_n, \mathbf{P}_n, \mathcal{H}_{n-1}\right] = \mathbb{E}[h_i(R_{n,i}) \mid S_n, \mathbf{P}_n, \mathcal{H}_{n-1}],$$

and

$$\mathbb{E}\left[\frac{O_{n,i}O_{n,j}}{\pi^{\mathrm{obs}}_{n,ij}}h_{ij}(R_{n,i},R_{n,j})\,\middle|\,S_n,\mathbf{P}_n,\mathcal{H}_{n-1}\right]=\mathbb{E}[h_{ij}(R_{n,i},R_{n,j})\mid S_n,\mathbf{P}_n,\mathcal{H}_{n-1}].$$

Proof. Conditional independence factors each expectation into the inclusion probability times the corresponding return moment. Division by the known propensity cancels that factor. □

Choosing $h_i(r)=(r-[\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\omega})]_i)^2$ and $h_{ij}(r_i,r_j)=(r_ir_j-[\widehat{\mathbf{Q}}_n(\boldsymbol{\omega})]_{ij})^2$ for any fixed $\boldsymbol{\omega}$ measurable before the current panel and outcome proves that Equation 34 is conditionally unbiased for the corresponding full-coordinate regression loss. Finite fourth moments make the pair loss integrable.

Proposition 9 (Population target of moment regression)**.** *Let $\mathcal{G}_n:=\mathcal{H}_{n-1}\vee\sigma(S_n,\mathbf{P}_n)$ be the information available after routing and before the outcome, and assume $\mathbb{E}[\|\mathbf{R}_n\|_2^4\mid\mathcal{G}_n]<\infty$ almost surely. Write*

$$\mathbf{m}_n:=\mathbb{E}[\mathbf{R}_n\mid\mathcal{G}_n],\qquad\mathbf{Q}_n:=\mathbb{E}[\mathbf{R}_n\mathbf{R}_n^\top\mid\mathcal{G}_n].$$

Up to terms independent of the predictions, the conditional full-coordinate population risk corresponding to Equation 34 is

$$\frac{1}{N}\|\widehat{\boldsymbol{\mu}}_n-\mathbf{m}_n\|_2^2+\frac{\lambda_2}{N^2}\|\widehat{\mathbf{Q}}_n-\mathbf{Q}_n\|_F^2.$$

Hence, in a well-specified function class at a population optimum, the unique predicted outputs are

$$\widehat{\boldsymbol{\mu}}_n=\mathbf{m}_n,\qquad\widehat{\boldsymbol{\Sigma}}_{\mathrm{stat},n}=\mathbf{Q}_n-\mathbf{m}_n\mathbf{m}_n^\top=\mathrm{Cov}(\mathbf{R}_n\mid\mathcal{G}_n).$$

This identifies the output moments, not a unique factor $\mathbf{C}_{\boldsymbol{\omega}}$ or a unique neural-network parameter vector.

Proof. Apply the conditional squared-loss decomposition separately to every coordinate $R_{n,i}$ and product $R_{n,i}R_{n,j}$, then sum the resulting excess risks. The true covariance is positive semidefinite and is therefore feasible under the factorisation $\widehat{\boldsymbol{\Sigma}}_{\mathrm{stat}}=\mathbf{C}\mathbf{C}^\top$. □

Proposition 10 (Anchor-query coverage at a fixed snapshot)**.** *Freeze the task model, router, and critic-training data-generating mechanism. Let ν_{train} be the resulting law of $(S_n,\mathbf{P}_n)$ under training-context and route sampling, and let $\nu_{\mathrm{query}}=\mathrm{Law}(S,\mathbf{a}(S))$ be the law on which it is queried for routing. If the full-coordinate population regression uniquely identifies the conditional moments ν_{train}-almost everywhere and $\nu_{\mathrm{query}}\ll\nu_{\mathrm{train}}$, then those moments are also identified ν_{query}-almost everywhere. Without this absolute continuity or an explicit parametric generalisation assumption, shadow-panel positivity alone does not identify anchor-query moments.*

Proof. An equality that fails only on a ν_{train}-null set also fails only on a ν_{query}-null set under absolute continuity. The final claim follows because panel randomisation changes observed expert coordinates at a fixed $(S_n,\mathbf{P}_n)$, not the support of the allocation input itself. □

A complementary sufficient condition quantifies the "nearby allocation" argument used by the implementation. At a fixed snapshot and context s, suppose both the true and predicted moment maps are Lipschitz in their allocation input:

$$\|\boldsymbol{\mu}^{\mathbf{c}}(s)-\boldsymbol{\mu}^{\mathbf{c}'}(s)\|_2\le L_\mu(s)\|\mathbf{c}-\mathbf{c}'\|_2,$$
$$\|\widehat{\boldsymbol{\mu}}^{\mathbf{c}}(s)-\widehat{\boldsymbol{\mu}}^{\mathbf{c}'}(s)\|_2\le\widehat{L}_\mu(s)\|\mathbf{c}-\mathbf{c}'\|_2,$$

with analogous operator-norm constants $L_\Sigma(s)$ and $\widehat{L}_\Sigma(s)$ for the statistical covariance. If $\|\mathbf{p}-\mathbf{a}(s)\|_2\le\tau$, then the triangle inequality gives

$$\|\widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s)-\boldsymbol{\mu}^{\mathbf{a}}(s)\|_2\le\|\widehat{\boldsymbol{\mu}}^{\mathbf{p}}(s)-\boldsymbol{\mu}^{\mathbf{p}}(s)\|_2+(L_\mu+\widehat{L}_\mu)\tau,$$
$$\|\widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{stat}}(s)-\boldsymbol{\Sigma}^{\mathbf{a}}(s)\|_{\mathrm{op}}\le\|\widehat{\boldsymbol{\Sigma}}^{\mathbf{p}}_{\mathrm{stat}}(s)-\boldsymbol{\Sigma}^{\mathbf{p}}(s)\|_{\mathrm{op}}+(L_\Sigma+\widehat{L}_\Sigma)\tau.\tag{35}$$

This is an explicit generalisation assumption, not a consequence of panel positivity. Anchor-query calibration in Section D.6 directly measures the residual error after this distribution shift.

A sufficient design combines domination of the query context marginal by the training context marginal with positive conditional probability on the snapshot-matched event $\mathbf{P}_n = \mathbf{a}(S_n)$. If task-model snapshots change across the fitting window, Proposition 10 is only an idealised stationary identification statement. The reported implementation does not enforce an atom at $\mathbf{P}_n = \mathbf{a}(S_n)$. It instead uses a parametric critic, a trust region, and route distillation to generalise between nearby allocation inputs. Proposition 10 is therefore a sufficient identification condition rather than a claimed guarantee for the implementation. We measure the resulting error directly on the anchor-query law in Section 4.4.

This identification result does not imply that a finite-capacity neural critic is unbiased. It identifies the population regression target on the executed $(S_n, \mathbf{P}_n)$ distribution under correct propensities. Propensity clipping, critic misspecification, nonstationarity, and insufficient context overlap introduce additional errors and must be measured.

B.2 Cardinality-constrained support search

For a fixed allowed support $\mathcal{I} \subseteq [N]$ with $|\mathcal{I}| \leq k$, set coordinates outside $\mathcal{I}$ to zero and allow coordinates inside $\mathcal{I}$ to be nonnegative. The resulting restriction of Equation 14 has

$$\|\mathbf{v} - \mathbf{a}\|_2^2 = \|\mathbf{v}_{\mathcal{I}} - \mathbf{a}_{\mathcal{I}}\|_2^2 + \|\mathbf{a}_{\mathcal{I}^c}\|_2^2.$$

It is feasible if and only if $\|\mathbf{a}_{\mathcal{I}^c}\|_2 \leq \tau$: necessity is immediate, and sufficiency follows by taking $\mathbf{v}_{\mathcal{I}} = \mathbf{a}_{\mathcal{I}}$ and $\mathbf{v}_{\mathcal{I}^c} = 0$. On a feasible support, the remaining budget, cap, nonnegativity, and trust-region constraints form a convex set. The predicted objective is concave, and it is strictly concave when $\eta > 0$ because the routed covariance is positive definite. Thus each feasible fixed-support problem is a convex optimisation problem in the standard minimisation convention; when $\eta = 0$, deterministic tie-breaking selects among linear-objective optima. For a routing query, define the predicted score

$$\widehat{J}_s(\mathbf{v}) := \mathbf{v}^\top \widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s) - \frac{\eta}{2} \mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}^{\mathbf{a}}_{\mathrm{route}}(s) \mathbf{v}.$$

The implementation solves each visited fixed-support problem to tolerance $\epsilon_{\mathrm{qp}} \in \mathbb{R}_{\geq 0}$, defined as the certified primal and dual objective gap of the continuous solver. The initial beam contains the anchor support and, for every $j \in \{1, \ldots, k\}$, the support of the j largest coordinates of $\widehat{\boldsymbol{\mu}}^{\mathbf{a}}(s)$. Thus a predicted marginal return means the corresponding conditional mean coordinate $\widehat{\mu}_i^{\mathbf{a}}(s)$. Infeasible initial supports are discarded using the criterion above.

At each of $n_{\mathrm{swap}} \in \mathbb{N}_{\geq 0}$ search rounds, the algorithm enumerates every feasible one-expert addition, removal, and swap from each retained support. Duplicate supports are removed, every new fixed-support problem is solved, and candidates are ranked by their optimised $\widehat{J}_s$ values. Lexicographic support order breaks exact ties. The best $K_{\mathrm{beam}} \in \mathbb{N}_{>0}$ supports are retained, and the best feasible allocation visited across all rounds is returned.

Let $\widehat{J}_s^\star$ be the exact maximum predicted score over all supports and let $\widehat{J}_{s,\mathrm{vis}}^\star$ be the exact maximum over supports visited by the beam. Define

$$\varepsilon_{\mathrm{search}}(s) := \widehat{J}_s^\star - \widehat{J}_{s,\mathrm{vis}}^\star.$$

Because the continuous solve on each visited support is within ϵ_{qp}, the global score gap in Proposition 7 satisfies

$$\varepsilon_{\mathrm{opt}}(s) \leq \varepsilon_{\mathrm{search}}(s) + \epsilon_{\mathrm{qp}}. \tag{36}$$

Exhaustive support enumeration sets $\varepsilon_{\mathrm{search}} = 0$. Alternatively, a valid mixed-integer upper bound U_s certifies the directly observable gap $\varepsilon_{\mathrm{opt}}(s) \leq U_s - \widehat{J}_s(\widehat{\mathbf{p}})$.

The beam width K_{beam}, number of swap rounds n_{swap}, continuous tolerance ϵ_{qp}, and wall-clock latency are method hyperparameters. For small N, exhaustive enumeration over $\sum_{j=0}^{k} \binom{N}{j}$ supports supplies a global ϵ_{qp}-optimal solution. Without exhaustive enumeration or a valid mixed-integer upper bound, the returned allocation is described only as the best visited solution, not as a certified global optimum.

Inference-time scope. At inference, the frozen router, critic, and solver use only s to select at most k experts. No targets, task gradients, shadow panels, or online updates are required. Dense covariance

prediction and support search add routing overhead, with no claim of subquadratic scaling in N. Accordingly, Section 4 treats active support only as a routing diagnostic rather than a wall-clock claim. Proposition 5 bounds the local loss approximation. Proposition 7 bounds moment estimation and support optimisation errors for the portfolio score.

B.3 LAGGED-CRITIC UPDATE

Let $\boldsymbol{\omega}_n$ be the learner critic after update n and $\overline{\boldsymbol{\omega}}_n$ the lagged critic used to route minibatch n. With Polyak rate $\tau_c \in (0, 1]$, the lagged critic is updated only after the current outcomes have been processed:

$$\overline{\boldsymbol{\omega}}_{n+1} = (1 - \tau_c)\overline{\boldsymbol{\omega}}_n + \tau_c \boldsymbol{\omega}_n.$$

Unrolling this recursion shows that $\overline{\boldsymbol{\omega}}_n$ depends only on its initialisation and $\boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_{n-1}$. It is therefore $\mathcal{H}_{n-1}$-measurable, consistent with the requirement above that $\mathbf{P}_n$ be $\sigma(S_n, \mathcal{H}_{n-1})$-measurable, and the current CGR targets cannot affect their own route. Setting $\tau_c = 1$ gives a lag of exactly one update, whereas the value $\tau_c = 0.005$ in Table 4 forms an exponential average of past learner critics with a mean lag of $1/\tau_c = 200$ updates. This ordering does not by itself guarantee stationarity. Drift and effective sample size are therefore diagnostic quantities rather than consequences of the lag construction; the held-out calibration in Section D.6 evaluates the resulting prediction error.

C NOTATION

Table 3 collects the core notation. The random quantities used in the derivation are named explicitly. Uppercase $S, \Xi, \mathbf{R}$, and $\mathbf{P}_n$ denote random quantities, while lowercase $s, \xi, \mathbf{r}$, and $\mathbf{p}_n$ denote their realisations. Boldface denotes a vector or matrix, not randomness. $\mathbb{S}_+^N$ and $\mathbb{S}_{++}^N$ are the cones of symmetric positive-semidefinite and positive-definite $N \times N$ matrices. $\mathbb{L}^N$ denotes the vector space of lower triangular $N \times N$ matrices. To keep the table compact, its critic rows abbreviate $\widehat{\boldsymbol{\mu}}^{\mathbf{c}}_{\boldsymbol{\omega}}(s)$, $\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{stat},\boldsymbol{\omega}}(s)$, $\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{route},\boldsymbol{\omega}}(s)$ as $\widehat{\boldsymbol{\mu}}^{\mathbf{c}}$, $\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{stat}}$, and $\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{route}}$, respectively.

Table 3: Notation for conditional portfolio routing.

Symbol	Type / range	Meaning
N, d	$\mathbb{N}_{>0}$	Number of routed experts and hidden dimension
S, s	$S : \Omega \to \mathcal{S},\ s \in \mathcal{S}$	Routing information and its realisation
Ξ, ξ	$\Xi : \Omega \to \mathcal{X},\ \xi \in \mathcal{X}$	Post-routing outcome and its realisation
P_S	probability measure on $\mathcal{S}$	Context law
$P_{\Xi \mid S=s}$	probability kernel in s	Regular conditional outcome law
$\mathbf{x}(s)$	$\mathbb{R}^d$	Layer input available to the router
$\mathbf{e}_i(s)$	$\mathbb{R}^d$	Output of expert i
$\mathbf{E}(s)$	$\mathbb{R}^{d \times N}$	Matrix of all expert outputs
$\mathbf{y}_0(s)$	$\mathbb{R}^d$	Allocation-independent residual/shared output
$\mathbf{y}(\mathbf{v}; s)$	$\mathbb{R}^d$	Layer output under allocation $\mathbf{v}$
$\mathbf{y}^{\mathbf{a}}(s)$	$\mathbb{R}^d$	Layer output at the anchor allocation
$\mathbf{y}(s; \mathbf{c})$	$\mathbb{R}^d$	Layer output at a point allocation $\mathbf{c}$
$\mathbf{a}(s)$	$\mathbb{R}^N_{\geq 0}$	Frozen anchor allocation
$\mathbf{p}(s)$	$\mathbb{R}^N_{\geq 0}$	Candidate final allocation
$\boldsymbol{\delta}(s)$	$\mathbb{R}^N$	Reallocation $\mathbf{p}(s) - \mathbf{a}(s)$
$\theta, \mathbf{W}_\theta$	parameter vector, $\mathbb{R}^{N \times d}$	Anchor-router parameters and gate matrix
$\mathbf{h}_\theta(s)$	$\mathbb{R}^N_{>0}$	Differentiable positive anchor proposal
$\mathcal{B}(s)$	subset of $\mathbb{R}^N_{\geq 0}$	Budget-feasible allocations
$\Pi_{\mathcal{B}(s)}$	$\mathbb{R}^N \to \mathcal{B}(s)$	Deterministically tie-broken Euclidean projection
$\mathcal{L}^{[\mathbf{c}]}(\cdot; s, \xi)$	$\mathbb{R}^d \to \mathbb{R}$	Task loss on the graph frozen at allocation $\mathbf{c}$
$\mathbf{g}^{\mathbf{a}}(s, \xi)$	$\mathbb{R}^d$	Layer-output task gradient at the anchor
$\mathbf{g}^{\mathbf{a}}(S, \Xi)$	$L^2(\mathbb{P}; \mathbb{R}^d)$	Random layer-output task gradient
$\mathbf{m}_g^{\mathbf{a}}, \mathbf{G}^{\mathbf{a}}$	$\mathbb{R}^d, \mathbb{S}_+^d$	Conditional gradient mean and covariance
$\mathbf{R}^{\mathbf{a}}$	$L^2(\mathbb{P}; \mathbb{R}^N)$	Random CGR vector at the anchor
$R_i^{\mathbf{a}}$	$\mathbb{R}$-valued random variable	ith coordinate of $\mathbf{R}^{\mathbf{a}}$
$\mathbf{r}^{\mathbf{a}}(s, \xi)$	$\mathbb{R}^N$	Realised CGR vector
$\Delta_i^{\mathbf{a}}(s, \xi)$	$\mathbb{R}$	Exact loss increase after deleting expert i
$D_{\mathbf{p}}^{\mathbf{a}}(s, \xi)$	$\mathbb{R}$	Frozen-route exact loss reduction from $\mathbf{a}$ to $\mathbf{p}$
$\mathcal{R}_{\mathbf{p}}^{\mathbf{a}}(s, \xi)$	$\mathbb{R}$	Residual of the local loss approximation
$Y^{\mathbf{a}}$	$L^2(\mathbb{P}; \mathbb{R})$	Aggregate return of the final portfolio
$Y_{\mathbf{a}}^{\mathbf{p}}$	scalar under $P_{\Xi \mid S=s}$	Fixed-context final portfolio return
$Y_{\mathbf{p},s}$		

Continued on next page

Table 3 (continued)

Symbol	Type / range	Meaning
$Z^{\mathbf{a}}_{\mathbf{p},s}$	scalar under $P_{\Xi \mid S=s}$	Fixed-context linearised loss reduction
$\boldsymbol{\mu}^{\mathbf{a}}(s)$	$\mathbb{R}^N$	Conditional CGR mean
$\boldsymbol{\Sigma}^{\mathbf{a}}(s)$	$\mathbb{S}^N_+$	Conditional CGR covariance
$\mathrm{CE}_\eta(X \mid S=s)$	$\mathbb{R}$	Conditional certainty equivalent of scalar return X
$\mathcal{A}(s;\mathbf{a})$	subset of $\mathbb{R}^N$	Feasible candidate allocations
$\mathbf{p}^\star(s)$	$\mathcal{A}(s;\mathbf{a})$	Exact contextual portfolio solution
M	$\mathbb{N}_{>0}$	Number of affected downstream logit coordinates
$\mathbf{z}^{[\mathbf{a}]}(\mathbf{y};s,\xi)$	$\mathbb{R}^M$	Downstream-logit map on the frozen anchor graph
ℓ	$\mathbb{R}^M \times \mathcal{X} \to \mathbb{R}$	Scalar loss as a function of logits and outcome
$\mathbf{A}^{\mathbf{a}}(s,\xi)$	$\mathbb{R}^{M\times d}$	Downstream-logit Jacobian at the anchor
$\mathbf{q}^{\mathbf{a}}(s,\xi)$	$\mathbb{R}^M$	Logit-loss gradient at the anchor
$\boldsymbol{\phi}^{\mathbf{a}}_i$	$\mathbb{R}^M$	First-order downstream-logit effect of expert i
$\mathbf{j}^{\mathbf{a}}_m$	$\mathbb{R}^d$	Local token-logit covector pulled back through the downstream Jacobian
$\mathbf{A}_J, \nu_J, \mathbb{E}_J$	$L^2(\nu_J;\mathbb{R}^{M_J\times d})$, law, expectation	Jacobian-lens averaging objects
$\mathbf{c}$	$\mathcal{B}(s)$	Point allocation supplied to the critic at context s
$\boldsymbol{\rho}$	measurable allocation rule	Pre-outcome rule with $\boldsymbol{\rho}(S) \in \mathcal{B}(S)$
$\mathbf{g}(s,\xi;\mathbf{c})$	$\mathbb{R}^d$	Layer-output task gradient at point allocation $\mathbf{c}$
$\mathbf{r}(s,\xi;\mathbf{c})$	$\mathbb{R}^N$	Realised CGR at point allocation $\mathbf{c}$
$\boldsymbol{\mu}^{\mathbf{c}}(s), \boldsymbol{\Sigma}^{\mathbf{c}}(s)$	$\mathbb{R}^N, \mathbb{S}^N_+$	Point-allocation conditional return moments
$\mathbf{R}^{\boldsymbol{\rho}}$	$L^2(\mathbb{P};\mathbb{R}^N)$	Random return induced by allocation rule $\boldsymbol{\rho}$
$\boldsymbol{\omega}, \mathbf{C}^{\mathbf{c}}_{\boldsymbol{\omega}}$	parameter vector, $\mathbb{L}^N$	Critic parameters and lower triangular covariance factor
$\widehat{\boldsymbol{\mu}}^{\mathbf{c}}$	$\mathbb{R}^N$	Critic mean prediction
$\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{stat}}$	$\mathbb{S}^N_+$	Statistical covariance prediction
$\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}}_{\mathrm{route}}$	$\mathbb{S}^N_{++}$	Ridge-regularised covariance used by the solver
$\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{Q}}_n$	$\mathbb{R}^N, \mathbb{S}^N_+$	Predicted mean and raw second moment at the executed allocation
$\mathcal{G}_n$	sub-σ-field	Post-route, pre-outcome information at update n
$\mathbf{m}_n, \mathbf{Q}_n$	$\mathbb{R}^N, \mathbb{S}^N_+$	True conditional mean and raw second moment given $\mathcal{G}_n$
$\widehat{J}_s(\mathbf{v})$	$\mathbb{R}$	Predicted portfolio score used by support search
$\mathbf{I}_N$	$\mathbb{R}^{N\times N}$	Identity matrix used by the solver ridge
$\mathcal{H}_{n-1}, \mathbf{P}_n, \mathbf{p}_n$	sub-σ-field, $\mathcal{B}(S_n)$-valued random vector, realisation	Pre-route history and update-n allocation
$\theta_n, \mathbf{a}_n$	parameter vector, $\mathcal{B}(s_n)$	Current anchor-router parameters and realised anchor
$\mathcal{I}_n, \mathbf{R}_n$	random subset of $[N]$, $L^4(\mathbb{P};\mathbb{R}^N)$	Executed support and update-n return vector
$\boldsymbol{\omega}_n, \overline{\boldsymbol{\omega}}_n$	critic parameter vectors	Learner critic and lagged critic used for routing
$F_n, B_{n,i}$	Bernoulli indicators	Outer panel gate and conditional inactive-expert draw
$\mathcal{Q}_n, O_{n,i}$	random subset of $[N]$, Bernoulli indicator	Shadow panel and coordinate-observation indicator
$\pi^{\mathrm{obs}}_{n,i}, \pi^{\mathrm{obs}}_{n,ij}$	$(0,1]$	Update-n coordinate and pair propensities
$\varepsilon_\mu, \varepsilon_{\Sigma,\mathrm{stat}}, \varepsilon_J$	$\mathbb{R}_{\geq 0}$	Mean, covariance, and uniform score-error bounds
$\varepsilon_{\mathrm{search}}, \varepsilon_{\mathrm{opt}}$	$\mathbb{R}_{\geq 0}$	Support-search and total predicted-score gaps
$\mathcal{L}_{\mathrm{LM}}$	$\mathbb{R}$	Normalised next-token prediction loss
$\mathcal{L}_{\mathrm{mom}}$	$\mathbb{R}_{\geq 0}$	Conditional-moment critic loss
$\mathcal{L}_{\mathrm{route}}$	$\mathbb{R}_{\geq 0}$	Anchor-route distillation loss
$H_{\mathbf{p}}, H_{\mathrm{del},i}, \overline{H}, L_E$	$\mathbb{R}_{\geq 0}$	Candidate, deletion, uniform Hessian bounds and expert-matrix bound

D EXPERIMENTAL DETAILS AND MECHANISM VALIDATION

This appendix reports the estimands, seed-level uncertainty, and complete mechanism diagnostics used in Section 4. All entries refer to the conditional critic and solver implementation.

D.1 PAIRING AND STATISTICAL PROTOCOL

The main quality sweep fixes $k = 8$ and varies $N \in \{16, 24, 32\}$; $N = 16$ is the $\sim$1B-parameter base configuration, and the stored parameter count grows with N. Within each N, methods share initialisation, data order, training horizon, and evaluation stream, whereas comparisons across N measure capacity scaling rather than seed-paired effects. Unless stated otherwise, mechanism studies use $N = 16$, $k = 4$, and five paired training seeds, with training seed as the replication unit for t intervals and item- or context-level resampling used only for evaluation uncertainty. The separate systems sweep fixes $N = 16$ and varies $k \in \{2, 4, 8\}$, so it does not support claims about increasing N. Hyperparameters and exclusion thresholds are fixed on development runs, and deployable routes never use evaluation outcomes; oracle-moment and exhaustive-search arms are offline diagnostics. Full settings are given in Table 4.

D.2 Repeated-outcome oracle and evaluation protocol

Repeated outcomes are used only for frozen-checkpoint diagnostics. For each context s, all candidates use the same sampled continuation and downstream masks under the fixed, allocation-invariant teacher-forced kernel $P_{\mathrm{teach}}(\cdot \mid s)$, implementing the exogenous coupling in Section A.1. Two independent panels with $M_A, M_B \geq 2$ are drawn: panel A estimates oracle moments or selects an

Table 4: Experimental configuration. Architecture, training, and data settings are shared by all methods. The shared expert is not counted in either N or k.

Architecture	
Transformer layers	12
Hidden size d	768
Attention heads	12
Expert FFN	SwiGLU, intermediate width 2,048 for every expert
Routed experts N	$\{16, 24, 32\}$ (main sweep), 16 (systems sweep and mechanism studies)
Shared experts	1, always active
Routing budget k	8 (main sweep), $\{2, 4, 8\}$ (systems sweep), 4 (mechanism studies)
Total parameters	≈ 1.0B at $N = 16$, growing with N
Active parameters	≈ 0.54B per token at nominal $k = 8$, excluding embeddings
Training	
Optimiser	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$
Weight decay	0.1
Peak learning rate	3×10^{-4}
Schedule	2,000-step warmup, cosine decay to 3×10^{-5}
Gradient clipping	1.0
Sequence length	1,024
Global batch	1,536 sequences (≈ 1.57M tokens), 8 per GPU $\times$ 48 accumulation steps
Training steps	63,000
Hardware	$4\times$ NVIDIA H200 (141 GB), FSDP with activation checkpointing
Precision	bf16 compute, fp32 optimiser and solver states
Data	
Corpus	Dolma dataset (Ettinger et al., 2025), $\approx$ 100B tokens, trained from scratch
Composition	High-quality web and PDF text with synthetic math, code, QA, instruction, and reasoning data, decontaminated against the OLMES benchmarks by the dataset authors
Held-out NLL	Held-out split of the training mixture
Tokeniser	GPT-2 byte-level BPE
ReMoE only	
Sparsity target	$1 - k/N$
L_1 coefficient	adaptive, multiplicative step 1.2
Load balancing	load-weighted L_1 penalty, as in Wang et al. (2025)
PoE only	
Allocation constraints	$B = 1$, $\mathbf{u} = \mathbf{1}_N$
Trust radius τ	$\sqrt{2}/k$, that is 0.177 at $k = 8$
Auxiliary-loss weights	$\lambda_2 = \lambda_{\mathrm{mom}} = \lambda_{\mathrm{route}} = 1.0$
Routing warm-up	critic and solver correction over the first 5% of updates
Support solver	ridge $\epsilon = 10^{-4}$, beam width 32, two swap rounds, QP tolerance 10^{-6}
PoE moment critic	
Network	one MLP per routed layer, two hidden layers of width 256, SiLU
Input	detached layer input $\mathbf{x}(s)$ concatenated with the allocation $\mathbf{c}$
Mean head	predicted conditional mean $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^N$
Covariance head	lower triangular factor $\mathbf{C}$, giving the full PSD covariance $\widehat{\boldsymbol{\Sigma}}_{\mathrm{stat}} = \mathbf{C}\mathbf{C}^\top$
Parameters	≈ 3.7M over all layers at $N = 16$, growing with N
Shadow rates (α_0, α)	$(0.10, 0.25)$
Lagged-critic Polyak rate τ_{c}	0.005

offline oracle route, whereas panel B evaluates that frozen route; deployable routes use neither panel for selection, and panel construction is held fixed across arms. For the geometry check, panel A is further split to estimate oracle CGR and task-adjoint correlation independently, with zero-variance cases excluded by a development-fixed threshold. The all-expert oracle represents off-route experts as zero-allocation shadows on the same frozen graph. For panel-B reductions D_m, every arm reports

$$\widehat{\mathrm{CE}}_{\eta_{\mathrm{eval}}}(D \mid s) = -\frac{1}{\eta_{\mathrm{eval}}} \log\left\{ \frac{1}{M_B} \sum_{m=1}^{M_B} \exp(-\eta_{\mathrm{eval}} D_m) \right\}. \tag{37}$$

Here $\eta_{\mathrm{eval}} > 0$ is an evaluation coefficient common to all arms, and the expression is computed by log-sum-exp in nats. We write η_{route} for the coefficient η in Equation 14.

D.3 Routing-diagnostic definitions

For numerical execution, coefficients at or below the continuous solver's feasibility tolerance are projected to zero before dispatch. The realised active support is therefore the dispatched set $\mathcal{I}_{\mathrm{exec}}(s) = \{i : p_i(s) > 0\}$ after that projection, and the reported active-expert count, including the always-active shared expert, is $1 + \mathbb{E}[|\mathcal{I}_{\mathrm{exec}}(S)|]$. The same dispatch rule is used for routed-support counts, shadow-panel propensities, and all routing diagnostics. The integer k is the exact routed-support size for Top-k MoE and MP-MoE, a hard upper bound for PoE, and the matched nominal activation target for ReMoE. The realised count is reported whenever support-based diagnostics are used.

At a frozen executed route $\mathbf{p}(s)$, joint failure at a routed layer is

$$\mathrm{JF} := \mathbb{P}(r_i^{\mathbf{p}}(S, \Xi) \le 0 \text{ for every } i \in \mathcal{I}_{\mathrm{exec}}(S)). \tag{38}$$

An empty dispatched support is counted as a failure. The probability is estimated on the same held-out contexts and outcomes for every arm, and is always reported beside the realised support size because activating more experts can mechanically lower joint failure.

Covariance-perturbation control. To determine whether PoE's reduction in joint failure can be attributed solely to sparsification or a pair-agnostic co-activation penalty, we trained two matched controls that alter only the off-diagonal covariance used for routing. One breaks the association between covariance values and expert identities by randomly reassigning the values across expert pairs within each layer. The other assigns every expert pair the same layer-average covariance. When aggregated across all routed layers, these controls yield joint-failure rates of 3.58% and 3.38%, respectively, both higher than PoE.

D.4 Training-cost measurement

This subsection is the separate $N = 16$ systems-cost sweep over $k \in \{2, 4, 8\}$, not the main expert-pool quality sweep. We measure all methods with the hardware, precision, global batch, sequence length, and training stack in Table 4. The statistics below are means and standard deviations over measured training steps. Throughput uses the fixed number of non-padding tokens in each global batch. The PoE measurement covers the complete active training path, including the anchor router, lagged-critic forward pass, learner-critic update, shadow-expert forward evaluations, and beam-plus-continuous support solver.

The support solver accounts for 5.63%, 6.41%, and 7.56% of the measured PoE step time at $k = 2$, 4, and 8. Its variability also closely tracks the increase in end-to-end timing variability at $k = 8$. Mean solver time is 0.291, 0.374, and 0.612 s per step at $k = 2, 4, 8$, respectively. PoE adds only 1.69% and 1.76% mean step time over MP-MoE at $k = 4$ and 8, respectively.

D.5 Exact counterfactual probes

For a held-out outcome at anchor $\mathbf{a}$, use the single-expert deletion effect $\Delta_i^{\mathbf{a}}$ from Equation 23 and define the joint deletion effect by

$$\Delta_{ij}^{\mathbf{a}}(s, \xi) = \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s) - a_i(s)\mathbf{e}_i(s) - a_j(s)\mathbf{e}_j(s); s, \xi) - \mathcal{L}^{[\mathbf{a}]}(\mathbf{y}^{\mathbf{a}}(s); s, \xi).$$

The single-expert test compares $a_i r_i^{\mathbf{a}}$ with $\Delta_i^{\mathbf{a}}$. The pair test uses $I_{ij} = \Delta_{ij}^{\mathbf{a}} - \Delta_i^{\mathbf{a}} - \Delta_j^{\mathbf{a}}$. All removals freeze the same downstream discrete routes as the derivative in Section 3.1. Near-zero exact effects

Table 5: Detailed end-to-end training cost for the fixed-$N = 16$ systems sweep. Throughput is reported in thousands of tokens per second. The final column is the increase in mean step time over Top-k MoE at the same routing budget.

k	Method	Step time (s) ↓	Throughput (ktok/s) ↑	Time increase
2	Top-k MoE	4.694 ± 0.048	335.1	0.00%
2	ReMoE	4.737 ± 0.048	332.1	0.92%
2	MP-MoE	4.873 ± 0.050	322.8	3.81%
2	PoE	5.170 ± 0.083	304.2	10.14%
4	Top-k MoE	5.417 ± 0.055	290.3	0.00%
4	ReMoE	5.460 ± 0.055	288.1	0.80%
4	MP-MoE	5.734 ± 0.059	274.3	5.85%
4	PoE	5.831 ± 0.221	269.7	7.64%
8	Top-k MoE	7.219 ± 0.072	217.9	0.00%
8	ReMoE	7.262 ± 0.073	216.6	0.60%
8	MP-MoE	7.951 ± 0.088	197.8	10.14%
8	PoE	8.091 ± 0.839	194.4	12.08%

Table 6: Component-level PoE training time for the fixed-$N = 16$ systems sweep, in seconds per step. Entries are mean ± standard deviation over measured training steps. Components are timed on the same active training path and sums may differ from the end-to-end measurement because displayed values are rounded.

Component	$k = 2$	$k = 4$	$k = 8$
Backbone forward/backward and recomputation	4.299 ± 0.043	4.723 ± 0.050	6.424 ± 0.068
Anchor/router	0.055 ± 0.002	0.055 ± 0.002	0.055 ± 0.002
Critic: target forward + learner update	0.163 ± 0.010	0.263 ± 0.010	0.372 ± 0.010
Shadow-expert forward	0.146 ± 0.019	0.202 ± 0.017	0.213 ± 0.012
Solver: beam + continuous programme	0.291 ± 0.056	0.374 ± 0.208	0.612 ± 0.831
Communication, optimiser, and remainder	0.215 ± 0.017	0.215 ± 0.017	0.415 ± 0.017

are excluded using a threshold fixed on development data. The representative early, middle, and late layers are L_0, L_6, L_{11}. For retained observations indexed by q, write $d_q = \Delta^{\mathbf{a}}_{i_q}$ and $\widehat{d}_q = a_{i_q} r^{\mathbf{a}}_{i_q}$. The reported normalised mean absolute error and through-origin slope are

$$\mathrm{NMAE} = \frac{\sum_q |\widehat{d}_q - d_q|}{\sum_q |d_q|}, \qquad \mathrm{Slope} = \frac{\sum_q d_q \widehat{d}_q}{\sum_q d_q^2}.$$

Pearson, Spearman, sign agreement, NMAE, and slope are computed within each training seed before forming seed-level intervals.

Pair non-additivity is a curvature diagnostic. If the frozen-graph loss is twice continuously differentiable at $\mathbf{y}^{\mathbf{a}}(s)$ and $\mathbf{H}^{\mathbf{a}}(s, \xi)$ denotes its Hessian there, scale the two deletions jointly by $t \downarrow 0$: define $\Delta^{\mathbf{a}}_i(t)$ and $\Delta^{\mathbf{a}}_{ij}(t)$ by replacing a_i, a_j in the deletion definitions by ta_i, ta_j, and set $I_{ij}(t) = \Delta^{\mathbf{a}}_{ij}(t) - \Delta^{\mathbf{a}}_i(t) - \Delta^{\mathbf{a}}_j(t)$. Then

$$I_{ij}(t) = t^2 a_i a_j \mathbf{e}_i^\top \mathbf{H}^{\mathbf{a}}(s, \xi) \mathbf{e}_j + o(t^2).$$

It is therefore not used as ground truth for gradient covariance. Task-adjoint correlation is evaluated only from panel-A outcomes for the identical $(s, \mathbf{a}(s))$ under $P_{\mathrm{teach}}(\cdot \mid s)$ from Section D.2. A single target supplies a raw outer product, not a conditional covariance. Averaging distinct contexts changes the estimand and is not labelled an oracle.

D.6 Critic calibration and query coverage

The repeated-outcome oracle uses zero-allocation shadows for all experts on the same frozen contexts, anchors, and panel-A outcomes. Calibration is reported separately on the executed-allocation training law and the anchor-allocation query law. This directly tests prediction after the allocation-input shift; it does not manufacture the overlap condition in Proposition 10.

Table 7: First-order CGR fidelity across representative layers. Entries are seed means with 95% t intervals in brackets over five paired seeds.

Layer	Pearson	Spearman	Sign agree. ↑	NMAE ↓	Slope
Early	0.995 [0.993, 0.997]	0.995 [0.992, 0.997]	0.959 [0.945, 0.973]	0.143 [0.117, 0.168]	0.977 [0.945, 1.008]
Middle	0.988 [0.982, 0.994]	0.988 [0.980, 0.996]	0.935 [0.908, 0.962]	0.232 [0.188, 0.275]	0.961 [0.902, 1.020]
Late	0.965 [0.943, 0.987]	0.951 [0.914, 0.988]	0.888 [0.841, 0.935]	0.398 [0.299, 0.497]	0.945 [0.892, 0.998]

Table 8: Geometry-to-risk bridge on repeated outcomes for identical routing contexts. The metric is the seed mean of the per-context Spearman correlation with an independent oracle CGR correlation. Intervals are 95% t intervals over five paired seeds. Higher is better. "Expert Jacobian" denotes $D_{\mathbf{x}}\mathcal{E}_i$. "Downstream effect" denotes $\mathbf{A}^{\mathbf{a}}\mathbf{e}_i$. Neither is the task-adjoint correlation optimised by PoE.

Criterion	Mean Spearman ρ ↑	95% CI
Raw output cosine	0.452	$[0.424, 0.480]$
Linear CKA	0.109	$[0.078, 0.139]$
Expert-local Jacobian cosine	0.221	$[0.189, 0.254]$
Downstream-effect cosine	0.752	$[0.733, 0.771]$
Task-adjoint cosine, independent panel	0.993	$[0.990, 0.995]$
Predicted sampled-shadow critic	0.967	$[0.956, 0.978]$

For held-out queries indexed by q, mean and covariance NRMSE are

$$\mathrm{NRMSE}_{\mu} = \sqrt{\frac{\sum_q \|\widehat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_q\|_2^2}{\sum_q \|\boldsymbol{\mu}_q\|_2^2}}, \qquad \mathrm{NRMSE}_{\Sigma} = \sqrt{\frac{\sum_q \|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_F^2}{\sum_q \|\boldsymbol{\Sigma}_q\|_F^2}}.$$

The statistical covariance $\widehat{\boldsymbol{\Sigma}}_{\mathrm{stat}}$, without the solver ridge, is used in all covariance-calibration metrics. For a set of candidate allocations fixed before panel-A outcomes are examined, portfolio variance uses $v_{q\ell} = \mathbf{p}_{q\ell}^{\top}\boldsymbol{\Sigma}_q\mathbf{p}_{q\ell}$ and $\widehat{v}_{q\ell} = \mathbf{p}_{q\ell}^{\top}\widehat{\boldsymbol{\Sigma}}_q\mathbf{p}_{q\ell}$. Portfolio NRMSE applies the same ratio to these scalars, portfolio ρ is their Spearman correlation, and the through-origin calibration slope is $\sum_{q,\ell}\widehat{v}_{q\ell}v_{q\ell}/\sum_{q,\ell}\widehat{v}_{q\ell}^2$. Pair ρ is computed over off-diagonal oracle and predicted correlations after excluding pairs with a development-fixed near-zero variance threshold. Every metric is first computed within a training seed.

D.7 Component identification

The anchor is the zero point of the exact route improvement and is not the Top-k MoE baseline. Setting $\eta_{\mathrm{route}} = 0$ isolates conditional expected return within the PoE machinery while leaving the common evaluation coefficient η_{eval} unchanged. Diagonal covariance removes pair dependence while preserving the variation of each expert return across unresolved outcomes. The placebo forms the predicted correlation matrix $\mathbf{C}_{\mathrm{corr}} \in \mathbb{R}^{N\times N}$, applies one common expert-index permutation $\mathbf{P}_\pi\mathbf{C}_{\mathrm{corr}}\mathbf{P}_\pi^{\top}$ using its permutation matrix $\mathbf{P}_\pi \in \{0,1\}^{N\times N}$, and restores the original marginal variances, thereby preserving PSD. Removing shadows tests off-route coordinate coverage. Oracle moments are an offline diagnostic, not a deployable routing input. Using the same beam before exhaustive support search separates critic error from support-search error.

D.8 Exact-loss utility estimand

For each held-out context, the component study first evaluates the frozen-route reduction $D_{\mathbf{p}}^{\mathbf{a}}(s,\Xi)$ on independent panel-B outcomes and then applies Equation 37. Table 10 uniformly averages these context-level values. Outcomes are never pooled across contexts before applying the nonlinear risk functional. The anchor therefore has value zero by construction, while all reported contrasts retain the sign convention that higher exact-loss certainty equivalent is better. This certainty equivalent is an evaluation metric for the routes selected by each component. It is not the objective used to select those routes. The PoE solver optimises the ridge-regularised mean–variance score of the final portfolio return $Y_{\mathbf{p}}^{\mathbf{a}}$ defined in Equation 10. Its unridged population counterpart equals the conditional CARA certainty equivalent when the return is conditionally Gaussian and otherwise is the second-cumulant approximation bounded in Proposition 6; the practical predicted score additionally

Table 9: Conditional-moment and solver-relevant portfolio-variance calibration. Entries are means over five paired seeds. NRMSE is normalised by the corresponding oracle norm. The symbol ρ denotes Spearman correlation.

Allocation/query law	Critic	Port. NRMSE ↓	Port. ρ ↑	Slope → 1	Mean NRMSE ↓	Cov. NRMSE ↓	Pair ρ ↑
Executed allocation	Sampled-shadow	0.197	0.975	0.966	0.292	0.183	0.980
Executed allocation	No-shadow	0.685	0.754	0.802	0.729	0.795	0.450
Anchor query	Sampled-shadow	0.206	0.965	0.974	0.775	0.238	0.967
Anchor query	No-shadow	0.592	0.782	0.846	0.928	0.784	0.452

Table 10: Component identification on the repeated-outcome panel. Values are 10^3 times the empirical certainty equivalent of exact frozen-route loss reduction, so higher is better. Means, standard errors, and t intervals use the five paired training seeds as the replication unit. The permuted arm is first averaged over five fixed correlation permutations within each seed.

Arm-level exact-loss certainty equivalent				
Arm	Mean ↑	SE	95% CI	Role
Anchor	0.000	0.000	$[0.000, 0.000]$	zero point
Mean-only, $\eta_{\text{route}} = 0$	3.968	0.945	$[1.345, 6.590]$	conditional mean
Diagonal covariance	7.041	0.894	$[4.560, 9.522]$	marginal return variance
Full covariance	12.009	1.030	$[9.150, 14.868]$	complete PoE
Permuted correlation	5.094	1.158	$[1.878, 8.310]$	pair-identity placebo
No-shadow critic	5.999	1.017	$[3.175, 8.823]$	coverage control
Oracle moments + beam	12.925	0.897	$[10.434, 15.417]$	critic diagnostic
Oracle moments + exhaustive	13.641	0.990	$[10.891, 16.391]$	solver diagnostic
Pre-specified paired contrasts				
Contrast	Mean Δ ↑	SE	95% CI	Positive seeds
Mean-only − anchor	3.968	0.945	$[1.345, 6.590]$	5/5
Diagonal − mean-only	3.073	0.436	$[1.863, 4.283]$	5/5
Full − diagonal	4.968	0.334	$[4.042, 5.894]$	5/5
Full − permuted	6.915	0.523	$[5.464, 8.367]$	5/5
Full − no-shadow	6.010	0.377	$[4.963, 7.057]$	5/5
Oracle+beam − full	0.916	0.402	$[-0.199, 2.032]$	4/5
Oracle+exhaustive − oracle+beam	0.716	0.315	$[-0.160, 1.591]$	4/5

contains the explicit ridge regulariser and moment-prediction error quantified in Proposition 7. As clarified in Remark 4, it prices final-portfolio risk rather than the variance of the reallocation gain.

D.9 Component-study results

Table 10 reports both arm-level exact-loss certainty equivalents and pre-specified paired contrasts. The staged comparison is monotone: mean-only improves on the anchor, diagonal covariance improves on mean-only, and full covariance improves on diagonal covariance, with every contrast positive in all five seeds. Full covariance also outperforms the permuted-correlation and no-shadow controls in every seed, supporting material contributions from pair identity and off-route coverage beyond marginal moments. Finally, oracle moments with the same beam and oracle moments with exhaustive search yield only modest additional gains, and both 95% intervals include zero.