ROUTETRACE: MEMBERSHIP INFERENCE FROM ROUTED UPDATE TRACES IN MIXTURE-OF-EXPERTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Router updates in a sparse mixture-of-experts (MoE) model can reveal whether a record was used for post-training. We introduce ROUTETRACE, a white-box membership audit of paired base and post-trained checkpoints. ROUTETRACE measures how a candidate's negative router gradient aligns with the net router update, both across layers and along the candidate's route. The candidate's token routes identify router rows associated with selected experts, allowing this alignment to be compared on selected and unselected rows. With three matched shadow checkpoints for each model and dataset, its learned scorers reach mean AUCs of .948 on Natural Questions and .862 on Dolly across Qwen, DeepSeek-MoE, and OLMoE. On Natural Questions, global depth features, which pool evidence across all router rows, preserve most of the full scorer's ranking accuracy, while alignment on selected rows separates members more strongly than equal-sized next-ranked and random unselected controls.

1 INTRODUCTION

White-box access to a base model and its post-trained checkpoint exposes the net parameter change made by post-training. We study whether this change, which combines updates from many training records, retains evidence that a particular candidate was used in post-training. In sparse mixture-of-experts (MoE) models, the router selects a subset of experts for each token (Dai et al., 2024). Each routed expert corresponds to a row of the router matrix. The candidate's token routes identify selected rows, giving us a way to compare their updates with updates on unselected rows.

Membership inference uses prediction vectors and losses to distinguish members from non-members (Shokri et al., 2017; Yeom et al., 2018). Reference-based tests such as LiRA and RMIA use auxiliary models to calibrate the evidence for each candidate (Carlini et al., 2022; Zarifzadeh et al., 2024). White-box attacks also exploit gradients and learned representations (Nasr et al., 2019; Leino & Fredrikson, 2020). Zari et al. (2021) compare candidate gradients with client updates in passive federated learning. A candidate's negative gradient gives a local direction for reducing its loss; the checkpoint difference gives the net change made by post-training. ROUTETRACE compares the candidate's descent direction with the net router update across an entire post-training run, testing whether members align more strongly than non-members. Comparing selected rows with unselected controls tests whether this evidence is stronger along the candidate's route.

The trace combines overall router alignment, its variation across layers, and comparisons between selected and next-ranked unselected rows. ROUTETRACE-S learns a ranking score from three matched shadow checkpoints. We also study ZERO-BR, a fixed scoring rule selected on three labeled development shadows that uses no auxiliary checkpoints during target scoring.

We evaluate Qwen, DeepSeek-MoE, and OLMoE on Natural Questions (NQ) and Dolly. On NQ, router alignment has higher AUC than query/value alignment and their joint view when the analysis state and feature count are held fixed. Global depth features pool evidence across all router rows and preserve most of ROUTETRACE-S's ranking accuracy. Under the same layerwise scoring rule, selected rows separate members more strongly than equal-sized next-ranked and random unselected sets (Figure 3).

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.

ROUTETRACE-S has higher AUC than the output baselines in Table 1 on all 18 target checkpoints. Its mean AUC over nine targets per dataset is .9482 on NQ and .8615 on Dolly. On Dolly, ZERO-BR has higher family-mean AUC than ROUTETRACE-S on Qwen and OLMoE, while ROUTETRACE-S leads on DeepSeek-MoE. LiRA and RMIA attain higher AUC than ROUTETRACE-S in a separate comparison on shared candidate records for Qwen NQ targets 1 and 2 (Figure 2(a)). In this comparison, ROUTETRACE-S uses three matched shadows; LiRA and RMIA use 16 reference checkpoints and one calibration checkpoint.

2 PROBLEM SETTING

2.1 PAIRED CHECKPOINTS

We study a base MoE and a post-trained checkpoint with the same architecture. Post-training updates the router matrices and a query/value (Q/V) adapter while keeping the remaining parameters fixed. We represent the checkpoint parameters as $\theta_b = (\phi_b, R_b)$ and $\theta_p = (\phi_b, A_p, R_p)$, where ϕ_b contains the frozen non-router parameters and A_p is the trained adapter. Merging A_p into ϕ_b gives the non-router parameters used in the post-trained model.

There are L router layers. At layer ℓ, $R_{b,\ell}, R_{p,\ell} \in \mathbb{R}^{E_\ell \times d_\ell}$, and each row corresponds to one expert. Corresponding rows have the same shape and expert identity across the two checkpoints.

2.2 MEMBERSHIP GAME AND SCOPE

Given a candidate record x, $y = 1$ means that x appeared in the post-training set; $y = 0$ denotes a non-member. The attack returns a real-valued score $s(x)$, or the decision $\widehat{y}_\tau(x) = \mathbf{1}[s(x) \geq \tau]$.

The attacker has white-box access to both checkpoints, the tokenizer, and the model implementation. It can compute candidate gradients and train same-family shadows on public data with known membership labels. The attack is passive and does not alter target training. Target membership labels enter only the final evaluation, after the scorers, target scores, and shadow-calibrated thresholds have been fixed.

This setting covers released checkpoint pairs and internal audits that retain both versions. It requires matching router rows and access to differentiate the candidate loss; a black-box API alone does not supply these inputs.

3 ROUTETRACE

Post-training leaves a net router update ΔR. ROUTETRACE tests whether the descent directions of members align more strongly with this update than those of non-members. The comparison uses the two released checkpoints.

A global cosine summarizes alignment over all router rows. To locate this alignment, ROUTETRACE also compares rows selected by a candidate with nearby unselected rows and records how the values vary across layers (Figure 1).

3.1 ROUTED-UPDATE ALIGNMENT

We compute the candidate gradient after merging the post-trained Q/V adapter into the base non-router parameters and retaining the base router. This hybrid state uses the adapted representations while evaluating the router gradient at the base router parameters R_b:

$$\begin{aligned} \widetilde{\theta} &= \big(\mathrm{Merge}(\phi_b, A_p), R_b\big), \\ \Delta R_\ell &= R_{p,\ell} - R_{b,\ell}, \\ D_\ell(x) &= -\nabla_{R_\ell} \mathcal{L}_{\mathrm{CE}}(x; \theta)|_{\theta=\widetilde{\theta}}\,. \end{aligned} \tag{1}$$

The derivative is taken with respect to the router variables R_ℓ and evaluated at $\widetilde{\theta}$. The adapter affects the representations used to compute $D_\ell(x)$; the aligned update contains router parameters alone.

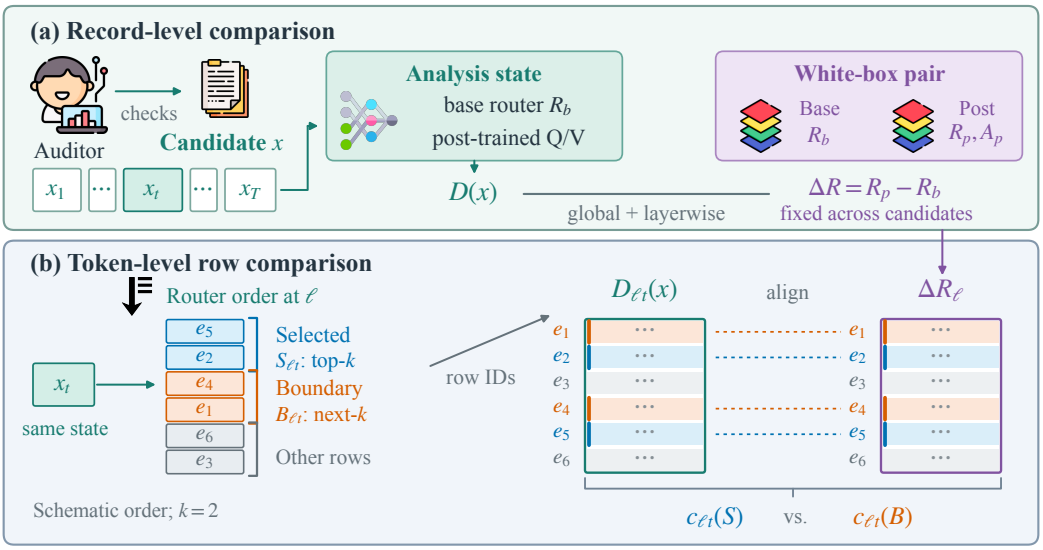

Figure 1: Global and token-local alignment in ROUTETRACE. Both views use the fixed base-to-post router update and the analysis state shown above. Tokens, expert order, and $k = 2$ are schematic; blue/orange mark selected/boundary rows.

We use shifted next-token cross-entropy for the candidate loss, even when post-training includes additional objectives.

Let $D(x)$ and ΔR concatenate the vectorized layer matrices. Using the stabilized cosine $\cos_\varepsilon(U, V) = \langle U, V\rangle / \max\{\|U\|_F \|V\|_F, \varepsilon\}$, with $\varepsilon = 10^{-12}$, we compute

$$r_g(x) = \cos_\varepsilon(D(x), \Delta R), \qquad r_{\ell,g}(x) = \cos_\varepsilon(D_\ell(x), \Delta R_\ell). \tag{2}$$

The global alignment compares all layers in one cosine; it is not the mean of the layerwise cosines. The negative gradient in $D_\ell(x)$ makes positive alignment correspond to agreement with the descent direction. When $\Delta R = 0$, all alignments are zero.

3.2 ROUTE-CONDITIONED TRACE

For token t at layer ℓ, let $S_{\ell t}$ contain the k rows selected by the hybrid model. The next k rows in the logit ranking form $B_{\ell t}$, the boundary set. These equal-sized sets compare the selected route with the rows immediately outside it. All evaluated routers satisfy $E_\ell \geq 2k$.

Let N_x be the number of non-padding input positions, $h_{\ell t}$ the router input, and $r_{b,\ell e}$ row e of $R_{b,\ell}$. Its logit is $z_{\ell t,e} = r_{b,\ell e} h_{\ell t}$. The descent contribution at this router position is

$$D_{\ell t,e}(x) = -\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial z_{\ell t,e}} h_{\ell t}^\top. \tag{3}$$

Stacking the row contributions gives $D_{\ell t}(x)$, with $D_\ell(x) = \sum_{t=1}^{N_x} D_{\ell t}(x)$. For a row set A, we restrict both the gradient contribution and the update before computing alignment:

$$\begin{aligned} c_{\ell t}(A) &= \cos_\varepsilon\big(D_{\ell t}(x)[A], \Delta R_\ell[A]\big), \\ r_{\ell,s} &= \frac{1}{N_x} \sum_{t=1}^{N_x} c_{\ell t}(S_{\ell t}), \qquad r_{\ell,b} = \frac{1}{N_x} \sum_{t=1}^{N_x} c_{\ell t}(B_{\ell t}), \\ r_{\ell,c} &= r_{\ell,s} - r_{\ell,b}. \end{aligned} \tag{4}$$

We average over all router positions, including the final position, whose contribution is zero under the shifted next-token loss. Positive $r_{\ell,c}$ indicates stronger selected-row alignment on average.

The trace retains both overall alignment and its variation with depth. Let $\bar{r}_s = L^{-1}\sum_\ell r_{\ell,s}$ and $\bar{r}_c = L^{-1}\sum_\ell r_{\ell,c}$. We divide the layers into eight fixed contiguous groups and use bin_8 to denote the unweighted mean within each group. The resulting trace is

$$\begin{aligned}\varphi(x) = \big[&r_g, \bar{r}_s, \bar{r}_c,\ \mathrm{bin}_8(r_{1:L,g}),\\ &\mathrm{bin}_8(r_{1:L,s}),\ \mathrm{bin}_8(r_{1:L,c})\big] \in \mathbb{R}^{27}.\end{aligned} \tag{5}$$

It combines three scalar summaries with global, selected, and selected-minus-boundary depth summaries.

3.3 Scoring Modes

RouteTrace-S and RouteTrace-S-Cal use the 27-dimensional trace. Zero-BR uses the layerwise alignments directly, before depth binning.

For each learned scorer $j \in \{S, \mathrm{Cal}\}$, let μ_j be the coordinate means on its fitting records, σ_j the sample standard deviations floored at ε, and w_j its ridge-LDA weight. The score is

$$s_j(x) = w_j^\top\big((\varphi(x) - \mu_j) \oslash \sigma_j\big), \tag{6}$$

where $\oslash$ denotes coordinatewise division. RouteTrace-S fits these parameters using all records from three matched shadows.

RouteTrace-S-Cal first partitions shadow records by prompt into disjoint fitting and calibration sets. It fits its own parameters on the fitting set and scores the held-out non-members with that same scorer. For a requested FPR α, their scores determine a threshold τ_α whose empirical calibration FPR is at most α. The scorer and threshold are then applied unchanged to target records.

Zero-BR normalizes each candidate using its own layerwise values. For each layer and token, a seeded rule samples k unselected rows to form $U_{\ell t}$. Their alignment is averaged in the same way: $r_{\ell,u} = N_x^{-1}\sum_t c_{\ell t}(U_{\ell t})$. Let $\boldsymbol{\rho}, \mathbf{s}, \mathbf{b}, \mathbf{u}$ collect the global, selected, boundary, and random alignments across layers, respectively, and define $\boldsymbol{\delta} = [\mathbf{s} - \mathbf{b}; \mathbf{s} - \mathbf{u}]$.

Write d_{sd} for the sample standard deviation and d_{mad} for the unscaled median absolute deviation, each floored at ε. Zero-BR adds four terms with unit weights:

$$\begin{aligned} C_g &= \frac{\mathrm{mean}(\boldsymbol{\rho})}{d_{\mathrm{sd}}(\boldsymbol{\rho})}, \quad C_{\mathrm{rob}} = \frac{\mathrm{median}(\boldsymbol{\rho})}{d_{\mathrm{mad}}(\boldsymbol{\rho})},\\ C_s &= \frac{\mathrm{mean}(\mathbf{s})}{d_{\mathrm{sd}}(\mathbf{s})}, \quad C_{BR} = \frac{\mathrm{mean}(\boldsymbol{\delta})}{d_{\mathrm{sd}}(\boldsymbol{\delta})}.\\ s_{\text{Zero-BR}}(x) &= C_g + C_{\mathrm{rob}} + C_s + C_{BR}. \end{aligned} \tag{7}$$

We selected this rule on three labeled development shadows and fixed it before target evaluation. Zero-BR uses the target base/post pair without additional shadow or reference checkpoints. It returns a ranking score.

Trace extraction uses separate global and token-local forward/backward passes. Appendix B provides the exact calibration rule, numerical conventions, extraction procedure, and complexity.

4 Experiments

4.1 Setup

We evaluate Qwen1.5-MoE-A2.7B-Chat (Qwen Team, 2024), DeepSeek-MoE-16B (Dai et al., 2024), and OLMoE-1B-7B-0125-Instruct (Muennighoff et al., 2025) on Natural Questions (NQ) and Dolly. Each family and dataset has three target checkpoints and three matched shadows, with target and shadow training records kept separate. NQ checkpoints use 4,096 training records and are audited on 512 members and 10,000 non-members; Dolly uses 1,024 training records, 256 members, and 1,920 non-members. The same record identities are used across families with each model's native tokenizer. In these primary experiments, one epoch of post-training updates every router and rank-8 LoRA adapters on the attention query and value projections while keeping experts frozen. Qwen

Table 1: Membership inference across three MoE families and two datasets. Mean ROC-AUC over three target checkpoints; bold marks each row's highest mean.

Data	Family	ROUTETRACE-S	ZERO-BR	Loss prof.	EZ-MIA	Min-K%++
NQ	Qwen	**.9466**	.9447	.7754	.7554	.7081
	DeepSeek	**.9322**	.9080	.7964	.8179	.6215
	OLMoE	**.9657**	.9354	.7612	.7345	.7038
Dolly	Qwen	.8717	**.9152**	.6998	.7544	.6748
	DeepSeek	**.8589**	.8553	.6896	.7285	.5969
	OLMoE	.8540	**.8651**	.6612	.6806	.6426

Table 2: Fixed-threshold performance of the separately fitted ROUTETRACE-S-CAL scorer at nominal 1% shadow FPR. Rates are percentages; means average three targets. Ranges give the minimum and maximum realized FPR across targets, not confidence intervals.

Data	Family	Mean TPR (%)	Mean FPR (%)	Target FPR range (%)
NQ	Qwen	23.24	1.013	0.840–1.340
	DeepSeek	20.64	1.203	1.060–1.420
	OLMoE	28.19	0.880	0.710–1.130
Dolly	Qwen	22.01	1.510	1.198–1.771
	DeepSeek	3.65	0.816	0.521–1.250
	OLMoE	7.16	0.972	0.677–1.458

training on NQ includes a router auxiliary loss with coefficient .001; all other family–dataset settings use next-token cross-entropy alone.

The output baselines cover post-training loss, base-to-post loss differences and ratios, Min-K% (Shi et al., 2024), Min-K%++ (Zhang et al., 2025), EZ-MIA (Ilić et al., 2026), and a learned loss profile. The learned profile uses 49 features and the same three shadows as ROUTETRACE-S. We select loss-statistic variants and k values on shadow records; EZ-MIA uses the base checkpoint as its reference. LiRA and RMIA are evaluated on a shared Qwen cohort in Section 4.4.

ROC-AUC evaluates ranking. Fixed-threshold results use ROUTETRACE-S-CAL and report true positive rate (TPR) together with realized false positive rate (FPR). All means weight target checkpoints equally. For the main ranking results, paired AUC differences use 10,000 shared bootstrap resamples stratified by membership label, conditional on the checkpoints and fitted scorers. Appendix C gives the full training and baseline settings; Appendix D gives per-target results and intervals.

4.2 MAIN RESULTS

ROUTETRACE-S exceeds the learned loss profile, EZ-MIA, and Min-K%++ in AUC on all 18 target checkpoints. All 54 paired 95% intervals for these comparisons lie above zero. On NQ, its family means range from .9322 to .9657; Table 1 shows the six model–dataset comparisons.

On Dolly, ZERO-BR has the higher family mean on Qwen and OLMoE, while ROUTETRACE-S leads on DeepSeek-MoE. The nine-target means are .8785 for ZERO-BR and .8615 for ROUTETRACE-S. The paired intervals favor ZERO-BR on all three Qwen targets, whereas all three OLMoE intervals include zero. Relative to NQ, the Dolly experiments also change the training-set size and, for Qwen, the training objective.

4.3 LOW-FPR PERFORMANCE

At the nominal 1% shadow FPR, mean TPR for ROUTETRACE-S-CAL ranges from 20.64% to 28.19% across NQ families. On Dolly, mean TPR is 22.01% for Qwen, 3.65% for DeepSeek-MoE, and 7.16% for OLMoE. Mean TPR across families is 24.02% on NQ and 10.94% on Dolly.

The 1% calibration level does not hold uniformly on target checkpoints. On Qwen Dolly, realized FPR averages 1.510% and ranges from 1.198% to 1.771% across targets. Table 2 reports each mean TPR

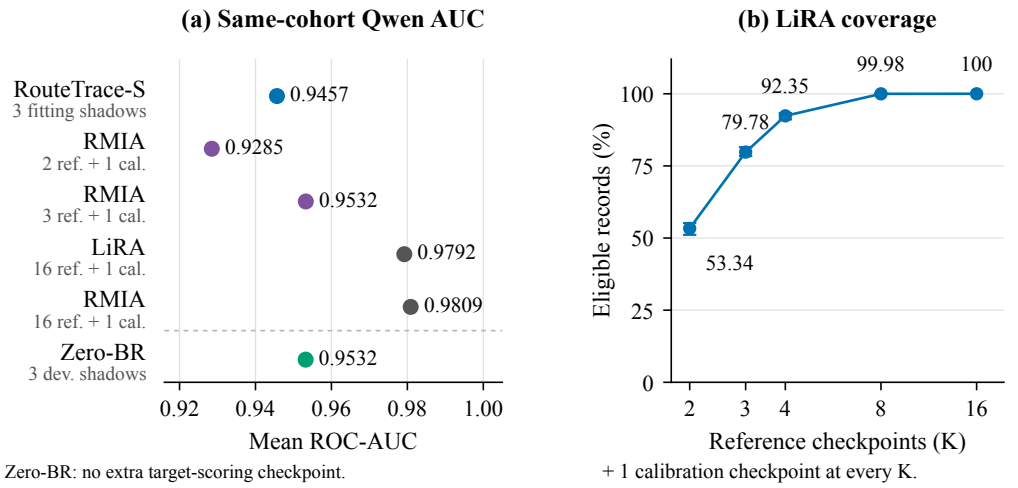

Figure 2: Reference requirements on Qwen (NQ). (a) Mean AUC on the same 7,817 and 7,812 records for targets 1 and 2. Low-budget RMIA points average 64 fixed reference subsets on the complete panels. Labels distinguish fitting shadows, development shadows, and references plus calibration; S and Zero-BR load no auxiliary checkpoints when scoring new targets. (b) LiRA eligibility within those same two candidate sets. Points average both targets; error bars are the 2.5th–97.5th percentiles over 64 reference subsets for $K < 16$. The complete $K = 16$ set has no subset interval. Every K uses one additional calibration checkpoint.

alongside mean and minimum–maximum FPR. Each family–dataset threshold is applied unchanged to the three target checkpoints. Additional operating points and per-target Wilson intervals appear in Appendix D.2.

Transferring the NQ-fitted scorer to Dolly changes ranking and calibration differently. Relative to fitting on Dolly, mean ROUTETRACE-S AUC changes by -.0292 on Qwen, -.0560 on DeepSeek-MoE, and +.0162 on OLMoE. The unchanged ROUTETRACE-S-CAL thresholds selected at nominal 1% NQ FPR realize 14.18%–22.05% mean FPR across Dolly families. This transfer changes both the dataset and training-set size, as well as Qwen's training objective; it is not an isolated domain-shift test. Appendix H.3 reports the per-family AUCs and threshold results.

4.4 REFERENCE REQUIREMENTS

LiRA (Carlini et al., 2022) and RMIA (Zarifzadeh et al., 2024) attain higher AUCs than ROUTETRACE-S on the same candidate records for Qwen NQ targets 1 and 2: 7,817 and 7,812 records, respectively (Figure 2(a)). Mean AUC is .9792 for LiRA and .9809 for RMIA, using 16 reference checkpoints and one calibration checkpoint. ROUTETRACE-S reaches .9457 with three matched shadows. ZERO-BR reaches .9532 without an auxiliary checkpoint during target scoring; we selected its fixed rule on three development shadows.

With two references and one calibration checkpoint, RMIA reaches .9285 AUC on these same candidate panels; with three references and one calibration checkpoint, it reaches .9532. Both values average the 64 fixed subsets and score the complete panels. Thus RMIA moves from below to above ROUTETRACE-S when its total auxiliary count rises from three to four.

For LiRA, a candidate is eligible only when the selected checkpoints contain both trained-on and held-out observations for that record. Figure 2(b) varies the reference count within the same two candidate sets. Mean eligibility rises from 53.34% with two references to 99.98% with eight; the full set of 16 covers both candidate sets completely. For each smaller reference count, we evaluate 64 subsets chosen without target labels. LiRA AUCs at different K are conditional on different eligible records. Appendix F gives per-target AUC intervals, the full reference sweep, and resource estimates.

We also vary the number of shadows used to fit ROUTETRACE-S (Table 3). Across all six family–dataset settings, the one- and two-shadow mean AUCs stay within .0051 of the three-shadow scorer.

Each mean averages all subsets of the three existing shadows and three targets; the largest loss for an individual subset–target combination is .0177. Appendix F.1 gives the full ranges.

Table 3: ROUTETRACE-S sensitivity to shadow count. Mean ROC-AUC over three targets and all subsets of the three existing shadows: three subsets at counts one and two, and one subset at count three. This measures sensitivity to the fixed shadows, not newly trained ones.

Data	Family	One shadow	Two shadows	Three shadows
NQ	Qwen	.9422	.9453	.9466
	DeepSeek	.9275	.9308	.9322
	OLMoE	.9645	.9652	.9657
Dolly	Qwen	.8696	.8712	.8717
	DeepSeek	.8562	.8582	.8589
	OLMoE	.8489	.8529	.8540

5 UNDERSTANDING THE ROUTED UPDATE TRACE

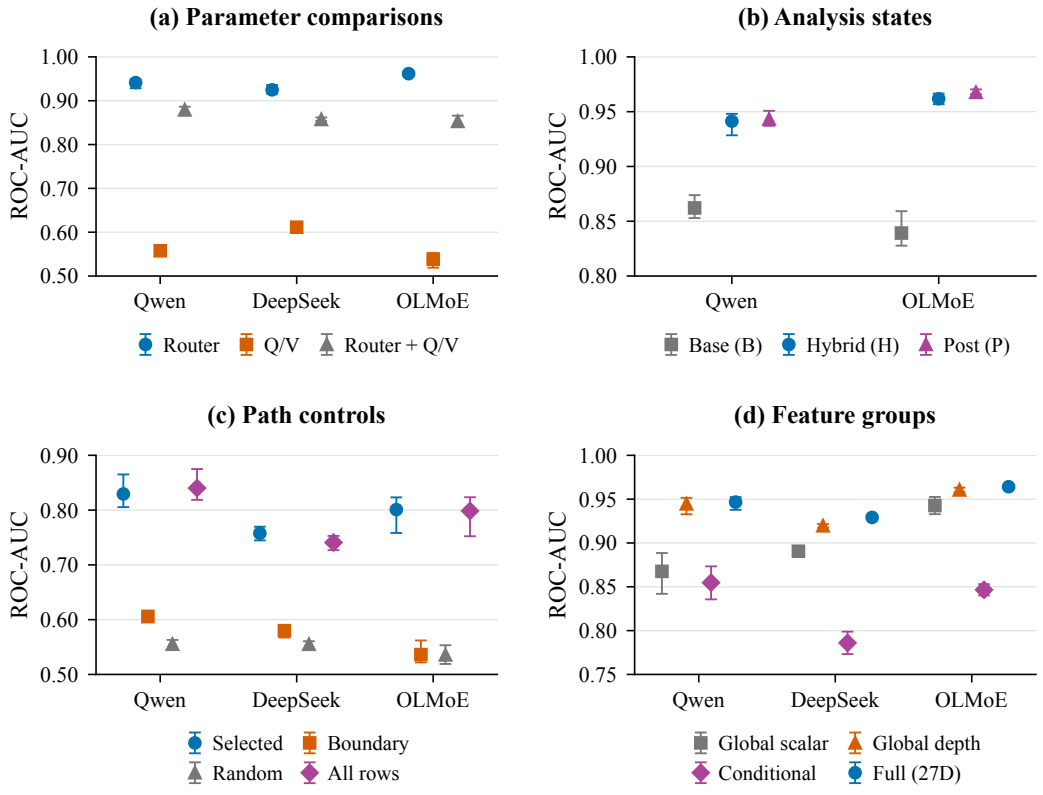

Figure 3: Mechanism comparisons on NQ. Points are mean AUCs; whiskers span the observed target minimum and maximum, not confidence intervals. (a) Nine-feature router, Q/V, and joint router–Q/V scores at H. (b) Effective-Router-9D under B, H, and P. Both (a) and (b) use 512 members and 2,048 non-members per target. (c) Unbinned raw layerwise scores for selected, boundary, random, and all rows. (d) Shadow-fitted feature groups; conditional features use selected-minus-boundary alignment. Each group uses three targets, except DeepSeek and OLMoE in (d), which use targets 1 and 2 from their feature-ablation studies. Qwen in (d) uses 512 members and 10,000 non-members per target with its previously fitted scorers. Candidate sets, scoring procedures, and axis ranges vary across panels.

5.1 PARAMETER COMPARISONS

We compare router, Q/V, and joint router–Q/V alignment at the same hybrid state. The parameter and state comparisons use 512 members and 2,048 non-members per target, with the records fixed across views and states. Each parameter view uses nine features—one global scalar and eight depth bins—and a separate ridge-LDA scorer fitted on the same shadow records. We call the router score Effective-Router-9D and the joint-parameter score Effective-All-9D. The joint view combines parameter blocks before constructing its nine features; ROUTETRACE-S uses the full 27-dimensional trace.

The router score has higher AUC than both alternatives in all three NQ families (Figure 3(a)). Its family means range from .9249 to .9618, compared with .5384–.6114 for Q/V and .8528–.8792 for the joint view. Adding Q/V parameters to the alignment lowers AUC on all nine targets; every paired interval for this comparison excludes zero. These scores use the same post-trained checkpoint and retain its Q/V adapter in the forward computation.

5.2 ANALYSIS STATES

For Qwen and OLMoE, we keep $\Delta R = R_p - R_b$ fixed and change the state used to compute candidate gradients. B uses base Q/V and the base router; H uses post-trained Q/V and the base router; P uses post-trained Q/V and the post-trained router. Each state has its own nine-feature scorer fitted on shadow features from that state. B, H, and P are alternative analysis states for one checkpoint pair, not successive training checkpoints.

Changing B to H raises mean router AUC from .8621 to .9413 on Qwen and from .8392 to .9618 on OLMoE (Figure 3(b)). All six paired intervals lie above zero. Replacing the base router in H with the post-trained router changes mean AUC by .0019 and .0058, respectively. Of the six P-minus-H intervals, only OLMoE target 1 excludes zero. DeepSeek-MoE contributes the H-state comparison in (a); its B/P states are not part of this sweep.

5.3 GLOBAL AND ROUTED EVIDENCE

Selected rows separate members more strongly than equal-size boundary and random sets in all three families. The path controls in Figure 3(c) average the token-local restricted cosine in equation 4 over tokens and then over native router layers, with equal weights at each step. All-row uses all router rows in this same token-local calculation; it is distinct from the global cosine in equation 2. Selected-row AUC exceeds the boundary and random controls by .1715–.3195 across nine targets, with all 18 paired intervals above zero.

All-row alignment remains strong: it has higher AUC on the three Qwen targets and OLMoE target 2, while selected rows lead on the other five. The selected path concentrates evidence relative to the boundary and random controls; the full trace retains global alignment as well.

Global depth preserves most of the full scorer's ranking accuracy (Figure 3(d)). This group contains the global scalar and eight bins of layerwise global alignment; Conditional contains the selected-minus-boundary scalar and its eight depth bins. The Qwen feature study uses the full audit set of 512 members and 10,000 non-members per target with its previously fitted shadow scorers. Global depth reaches .9445 mean AUC on Qwen, compared with .8676 for one global scalar and .9466 for the full trace. The full-minus-global-depth differences are .0003, .0008, and .0051 across the three targets; the paired interval excludes zero only on target 2. The two-target DeepSeek-MoE and OLMoE feature comparisons show the same ordering. These learned feature-group scores are separate from the unbinned path scores in (c). Appendix H.1 gives all feature groups and the paired Qwen intervals.

5.4 CONTROLS AND INTERVENTIONS

A character-n-gram classifier fitted on Qwen shadow records reaches .5028 mean AUC on NQ and .5205 on Dolly. None of its six per-target confidence intervals lies entirely above .5.

We match each member to four non-members by base-model mean NLL to compare records of similar difficulty for the base model. Each target's matched panel contains 512 members and 2,048 non-members on NQ, or 256 members and 1,024 non-members on Dolly. Across all six family–

dataset settings, the mean absolute standardized difference in base-model mean NLL falls below .0015 after matching. On these matched panels, ROUTETRACE-S exceeds the strongest evaluated output attack in every setting, with mean AUCs of .8517–.9654 versus .6781–.8128. The ranking advantage therefore remains when members and non-members are closely balanced on base-model mean NLL. Appendix H.2 reports the balance statistics and per-target results.

Appendix E reports the full parameter, state, and path results; Appendix H.2 also gives the text features.

We also change router updates during post-training. Clipping caps each record's router-gradient norm while leaving Q/V gradients unclipped; freezing holds the router fixed and trains Q/V adapters alone. Clipping adds no noise. We choose its strength using public development-set utility, then fit learned scorers and select baseline settings on shadows trained with the same intervention. Appendix G.1 gives the selection rule and training settings.

Table 4: NQ membership inference and utility under router interventions. Each row averages three target checkpoints. Attack fitting and baseline selection use shadows trained under the matching condition. Dev. F1 uses 256 public development records from the set used to select clipping strength. †: a constant RouteTrace score, assigned AUC .5 by convention.

		Membership ROC-AUC			Utility
Family	Training	ROUTETRACE-S	Loss prof.	EZ-MIA	Dev. F1 ↑
Qwen	Clean	.9466	.7754	.7554	.2346
	Clipped	.9520	.7833	.7597	.2324
	Frozen	.5000†	.5579	.5506	.2366
DeepSeek	Clean	.9322	.7964	.8179	.2401
	Clipped	.9391	.8131	.8349	.2378
	Frozen	.5000†	.5501	.5820	.2437
OLMoE	Clean	.9657	.7612	.7345	.2320
	Clipped	.9715	.7729	.7480	.2347
	Frozen	.5000†	.5505	.5345	.2363

At the selected clipping strength, all three attacks in Table 4 have higher family-mean AUCs than under clean training. For ROUTETRACE-S, the increase is .0054–.0070, with all three family-mean paired 95% intervals above zero. Development-set F1 changes by at most .0028.

Freezing sets $\Delta R = 0$, making ROUTETRACE-S constant on all nine targets. The two output attacks also decline, while mean development-set F1 is higher than clean in every family. Freezing removes the router update used by RouteTrace and changes post-training to Q/V-only adaptation; the constant score is a property of this observable, not a model-wide privacy guarantee. Appendix G reports the complete attack panel, utility measurements, calibration results, and separate weight-decay study.

In a separate Qwen router-only experiment, exposure four yields higher AUC than exposure two on the same seed-0 candidate records: ROUTETRACE-S rises from .9325 to .9613 and ZERO-BR from .8873 to .9332. Both paired AUC-difference 95% intervals are positive. The checkpoints use the same base, records and optimizer configuration but are trained independently, without Q/V adapters. The ROUTETRACE-S scorer fitted on three exposure-one shadows is held fixed. Exposure four was evaluated for seed 0; this is a two-checkpoint contrast, not an exposure curve (Appendix H.4).

6 CONCLUSION

ROUTETRACE shows that the router update between a base MoE and its post-trained checkpoint carries membership information. Across three MoE families on NQ and Dolly, it separates members from non-members on all 18 targets. ROUTETRACE-S exceeds the output baselines in Table 1 in AUC. On NQ, global, depth-resolved alignment supplies most ranking accuracy, while selected-row scores separate members more strongly than boundary and random controls. LiRA and RMIA reach higher AUC on the strict Qwen comparison with 16 references and one calibration checkpoint; ROUTETRACE-S uses three shadows.

AI USE STATEMENT

Generative AI tools assisted with research design and formulation, data preparation and code implementation, result analysis, and manuscript preparation and language editing. The authors have manually reviewed and verified the manuscript, code, and experimental results and take responsibility for the final work.

ETHICS STATEMENT

The experiments use public Natural Questions and Dolly records and checkpoints post-trained for this study. ROUTETRACE is intended for authorized privacy auditing. Membership inference can expose a record's presence in training data, and false positives can wrongly attribute a record to a training set. We report realized false-positive rates and calibration drift to make these errors visible. A membership score is statistical evidence, not conclusive proof that a record was used for training.

REPRODUCIBILITY STATEMENT

Sections 2–3 and Appendix B specify the audit assumptions, alignment definitions, scoring rules, and calibration procedure. Appendix C documents model versions, data preprocessing, membership splits, training settings, and software and hardware. Appendix D reports per-target results and uncertainty; Appendices E–H give the controls, reference comparisons, interventions, and additional analyses.

REFERENCES

Li Bai, Qingqing Ye, Xinwei Zhang, Sen Zhang, Zi Liang, Jianliang Xu, and Haibo Hu. Toward efficient inference attacks: Shadow model sharing via mixture-of-experts, 2025. URL https://arxiv.org/abs/2510.13451.

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression, 2023. URL https://arxiv.org/abs/2307.03694.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2022. URL https://arxiv.org/abs/2112.03570.

Gilad Cohen and Raja Giryes. Membership inference attack using self influence functions, 2022. URL https://arxiv.org/abs/2205.13680.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. URL https://arxiv.org/abs/2401.06066.

Databricks. Databricks dolly 15k. Hugging Face dataset, 2023. URL https://huggingface.co/datasets/databricks/databricks-dolly-15k. Version 1.0, CC BY-SA 3.0.

Ruyi Ding, Tianhong Xu, Xinyi Shen, Aidong Adam Ding, and Yunsi Fei. Moecho: Exploiting side-channel attacks to compromise user privacy in mixture-of-experts llms, 2025. URL https://arxiv.org/abs/2508.15036.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models?, 2024. URL https://arxiv.org/abs/2402.07841.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration, 2024. URL https://arxiv.org/abs/2311.06062.

Faqian Guan, Tianqing Zhu, Hanjin Tong, and Wanlei Zhou. Attention-based membership inference attacks on graph neural network through topological features. *IEEE Transactions on Dependable and Secure Computing*, 22(6):6469–6486, 2025. doi: 10.1109/TDSC.2025.3586251.

Jamie Hayes, Ilia Shumailov, and Itay Yona. Buffer overflow in mixture of experts, 2024. URL https://arxiv.org/abs/2402.05526.

Hongsheng Hu, Xuyun Zhang, Zoran Salcic, Lichao Sun, Kim-Kwang Raymond Choo, and Gillian Dobbie. Source inference attacks: Beyond membership inference attacks in federated learning, 2023. URL https://arxiv.org/abs/2310.00222.

David Ilić, David Stanojević, and Kostadin Cvejoski. Powerful training-free membership inference against autoregressive language models, 2026. URL https://arxiv.org/abs/2601.12104.

Joshua Kalyanapu, Farshad Dizani, Darsh Asher, Azam Ghanbari, Rosario Cammarota, Aydin Aysu, and Samira Mirbagher Ajorpaz. Gatebleed: Exploiting on-core accelerator power gating for high performance & stealthy attacks on ai, 2025. URL https://arxiv.org/abs/2507.17033.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Zhenglin Lai, Mengyao Liao, Bingzhe Wu, Dong Xu, Zebin Zhao, Zhihang Yuan, Chao Fan, and Jianqiang Li. Safex: Analyzing vulnerabilities of moe-based llms via stable safety-critical expert identification, 2025. URL https://arxiv.org/abs/2506.17368.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference, 2020. URL https://arxiv.org/abs/1906.11798.

Marvin Li, Jason Wang, Jeffrey Wang, and Seth Neel. Mope: Model perturbation-based privacy attacks on language models, 2023. URL https://arxiv.org/abs/2310.14369.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison, 2023. URL https://arxiv.org/abs/2305.18462.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. An empirical analysis of memorization in fine-tuned autoregressive language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1816–1826, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.119. URL https://aclanthology.org/2022.emnlp-main.119/.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, et al. Olmoe: Open mixture-of-experts language models, 2025. URL https://arxiv.org/abs/2409.02060.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 739–753. IEEE, May 2019. doi: 10.1109/sp.2019.00065. URL http://dx.doi.org/10.1109/SP.2019.00065.

Georg Pichler, Marco Romanelli, Leonardo Rey Vega, and Pablo Piantanida. Perfectly accurate membership inference by a dishonest central server in federated learning. *IEEE Transactions on Dependable and Secure Computing*, 21(4):4290–4296, 2024. ISSN 2160-9209. doi: 10.1109/tdsc.2023.3326230. URL http://dx.doi.org/10.1109/TDSC.2023.3326230.

Qwen Team. Qwen1.5-MoE: Matching 7b model performance with 1/3 activated parameters. Qwen technical blog, March 2024. URL https://qwenlm.github.io/blog/qwen-moe/.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference, 2019. URL https://arxiv.org/abs/1908.11229.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, 2018. URL https://arxiv.org/abs/1806.01246.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2024. URL https://arxiv.org/abs/2310.16789.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017. URL https://arxiv.org/abs/1610.05820.

Anshuman Suri, Xiao Zhang, and David Evans. Do parameters reveal more than loss for membership inference?, 2024. URL https://arxiv.org/abs/2406.11544.

Pierre Tholoniat, Huseyin A. Inan, Janardhan Kulkarni, and Robert Sim. Differentially private training of mixture of experts models, 2024. URL https://arxiv.org/abs/2402.07334.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks, 2022. URL https://arxiv.org/abs/2111.08440.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models, 2022. URL https://arxiv.org/abs/2111.09679.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting, 2018. URL https://arxiv.org/abs/1709.01604.

Itay Yona, Ilia Shumailov, Jamie Hayes, and Nicholas Carlini. Stealing user prompts from mixture of experts, 2024. URL https://arxiv.org/abs/2410.22884.

Oualid Zari, Chuan Xu, and Giovanni Neglia. Efficient passive membership inference attack in federated learning, 2021. URL https://arxiv.org/abs/2111.00430.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2024. URL https://arxiv.org/abs/2312.03262.

Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. Label-only membership inference attacks and defenses in semantic segmentation models. *IEEE Transactions on Dependable and Secure Computing*, 20(2):1435–1449, 2023. doi: 10.1109/TDSC.2022.3154029.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-klarge language models, 2025. URL https://arxiv.org/abs/2404.02936.

Rongting Zhang, Martin Bertran, and Aaron Roth. Order of magnitude speedups for llm membership inference, 2024. URL https://arxiv.org/abs/2409.14513.

Yifan Zhou, Tianshi Xu, Jue Hong, Ye Wu, and Meng Li. Cryptomoe: Privacy-preserving and scalable mixture of experts inference via balanced expert routing, 2025. URL https://arxiv.org/abs/2511.01197.

Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, and Xiangliang Zhang. Seuf: Is unlearning one expert enough for mixture-of-experts llms?, 2025. URL https://arxiv.org/abs/2411.18797.

A Related Work

A.1 Output- and Reference-Based MIA

Shadow-model attacks learn differences between the predictions made on members and non-members (Shokri et al., 2017). Loss-based analyses connect membership risk to overfitting (Yeom et al., 2018) and characterize optimal tests under assumptions on the distribution of trained parameters (Sablayrolles et al., 2019). ML-Leaks relaxes requirements on the number of shadows, knowledge of the target architecture, and similarity between auxiliary and target training data (Salem et al., 2018).

Candidate difficulty is central to interpreting an output score. Watson et al. (2022) study difficulty calibration, and Ye et al. (2022) organize membership tests by the information available to the attacker. LiRA estimates candidate-specific distributions with and without training exposure and emphasizes evaluation at low false-positive rates (Carlini et al., 2022). RMIA combines reference models and population samples in a likelihood-ratio test that also works with few references (Zarifzadeh et al., 2024). Quantile regression offers another way to estimate candidate-specific thresholds without training a large shadow-model ensemble (Bertran et al., 2023). These studies motivate our separate reporting of AUC, fixed-threshold error rates, and auxiliary-model counts. In our shared-record Qwen comparison, LiRA and RMIA lead in AUC; RouteTrace-S uses three matched shadows (Section 4.4).

Language-model studies distinguish membership in pretraining from membership in a subsequent fine-tuning set. Mireshghallah et al. (2022) examine how fine-tuning choices affect memorization. Neighbourhood comparison uses scores on generated nearby texts to account for a candidate's difficulty (Mattern et al., 2023). Min-K% scores the least-likely tokens (Shi et al., 2024), while Min-K%++ evaluates token likelihood relative to the model's conditional distribution (Zhang et al., 2025). For fine-tuned models, EZ-MIA compares probability shifts at incorrectly predicted positions against a pretrained reference (Ilić et al., 2026). Self-prompt calibration (Fu et al., 2024) and small quantile-regression models (Zhang et al., 2024) provide further approaches to calibration and reference cost. Distribution shifts can also account for apparent membership performance (Duan et al., 2024). Our text-only and base-difficulty controls address two such alternative cues, while the NQ-to-Dolly study reports the joint change in data and training protocol explicitly.

A.2 White-Box and Internal-Signal MIA

Model internals provide membership evidence beyond output scores. Nasr et al. (2019) investigate gradients and parameter updates under passive and active access, while Leino & Fredrikson (2020) use memorized internal representations for calibrated inference. Self-influence scores provide a further white-box signal (Cohen & Giryes, 2022). Suri et al. (2024) show that parameters can reveal information beyond loss under stochastic gradient descent. For language models, MoPe studies membership through parameter perturbations (Li et al., 2023). Zari et al. (2021) compare candidate gradients with client updates in passive federated learning, using each round's initial model state. Their study includes individual layers, multiple communication rounds, and transfer learning from pretrained classifiers.

RouteTrace studies the net router update between a base MoE and its post-trained checkpoint, using a hybrid state with post-trained Q/V parameters and the base router. It compares that update with candidate gradients and organizes the alignment by layer and by the router rows associated with selected experts. The empirical questions are distinct: which parameter view supports ranking, and whether selected rows carry stronger evidence than equal-sized boundary and random controls. Our results retain the strong global depth signal alongside relative path localization (Section 5).

A.3 Architecture-Aware Privacy Attacks

Model structure can change the information available to an auditor. GNN membership studies use topological features and attention (Guan et al., 2025); segmentation studies show that predicted labels can reveal membership even without confidence scores (Zhang et al., 2023). These are different observations of model behavior, tied to graph structure or spatial outputs.

Federated learning also separates the inference target from the attacker's control over training. Pichler et al. (2024) study membership inference by an active server that controls the model supplied to a client. Hu et al. (2023) study which client contributed a training record under a server that follows the protocol. Source attribution is distinct from member/non-member classification. ROUTETRACE asks whether a record was used in post-training and observes checkpoint pairs without altering the target's training process.

A.4 PRIVACY AND ANALYSIS OF MOE MODELS

MoE routing creates dependencies that can expose information during inference. Buffer Overflow in Mixture of Experts studies interference between co-batched requests (Hayes et al., 2024), and prompt-stealing work studies disclosure of another user's current input (Yona et al., 2024). MoEcho examines user-content leakage through expert-dependent CPU/GPU execution footprints (Ding et al., 2025). GateBleed studies timing leakage from accelerator power gating, including transformer membership inference and MoE expert-choice leakage (Kalyanapu et al., 2025). These works motivate attention to routing as an observable. ROUTETRACE uses released parameters and candidate gradients; its evidence comes from post-training changes rather than runtime timing or contention.

Expert-level analysis also supports safety interventions and privacy protection. SAFEx identifies experts associated with safety behavior (Lai et al., 2025), and SEUF uses expert attribution for selective unlearning (Zhuang et al., 2025). Work on differentially private MoE training explores protection during optimization (Tholoniat et al., 2024); CryptoMoE studies private MoE inference (Zhou et al., 2025). These objectives concern expert function, training privacy, or inference confidentiality, each with its own evaluation criteria.

SHAPOOL uses MoE structure to share computation and parameters among shadow models for membership inference (Bai et al., 2025). Here the MoE organizes the auxiliary models. ROUTETRACE instead examines the target MoE's own router update. Our three-shadow scorer and target-time zero-auxiliary rule describe how we use this signal; the reference-count comparison is separate from claims about total computation.

B Formalization and Algorithm

B.1 Threat Model

We formalize the paired-checkpoint game by separating the attacker's inputs from the hidden membership label. Let P_1 and P_0 denote the candidate distributions conditioned on membership and non-membership. The attack receives a candidate x, the compatible checkpoints (θ_b, θ_p), and side information $\mathcal{I}$, which includes the tokenizer and model implementation. Here θ_p is the post-trained checkpoint containing R_p and A_p. For RouteTrace-S and RouteTrace-S-Cal, $\mathcal{I}$ also contains matched shadow checkpoint pairs and the membership labels of their records.

An attack $\mathcal{A}$ returns

$$s(x) = \mathcal{A}(\theta_b, \theta_p, x; \mathcal{I}), \qquad \widehat{y}_\tau(x) = \mathbf{1}[s(x) \geq \tau]. \tag{8}$$

Target labels are absent from $\mathcal{I}$ during feature construction, fitting, threshold selection, and target scoring.

B.2 Routed Updates

The two checkpoints define one fixed router update; every candidate is compared with this same tensor. At router layer ℓ, let $R_{b,\ell}, R_{p,\ell} \in \mathbb{R}^{E_\ell \times d_\ell}$, with row $r_{b,\ell e}$ assigned to expert e. Shape, row order, and expert identity agree across the checkpoints, and $\Delta R_\ell = R_{p,\ell} - R_{b,\ell}$.

The tokenizer produces N_x non-padding input positions. Let $\mathcal{T}_x \subseteq \{1, \ldots, N_x - 1\}$ contain the positions with a valid next-token label and $M_x = |\mathcal{T}_x|$. The audit loss is

$$\mathcal{L}_{\mathrm{CE}}(x; \widetilde{\theta}) = -\frac{1}{M_x} \sum_{t \in \mathcal{T}_x} \log p_{\widetilde{\theta}}(x_{t+1} \mid x_{\leq t}), \qquad \widetilde{\theta} = \big(\mathrm{Merge}(\phi_b, A_p), R_b\big). \tag{9}$$

For the unpadded, single-record inputs used here, $M_x = N_x - 1$. The checkpoint training objective is specified separately in Appendix C.3.

Let $h_{\ell t}$ be the router input and $z_{\ell t,e} = r_{b,\ell e} h_{\ell t}$ its logit for expert e. Writing $a_{\ell t,e} = \partial \mathcal{L}_{\mathrm{CE}} / \partial z_{\ell t,e}$, the rowwise descent contribution is

$$D_{\ell t,e}(x) = -a_{\ell t,e} h_{\ell t}^\top, \qquad D_{\ell,e}(x) = \sum_{t=1}^{N_x} D_{\ell t,e}(x). \tag{10}$$

The derivative $a_{\ell t,e}$ already contains the $1/M_x$ mean-reduction factor. The final router position has zero gradient but remains in the token-local average below.

Concatenating all layer matrices gives the global alignment

$$r_g(x) = \frac{\sum_{\ell=1}^{L} \langle D_\ell(x), \Delta R_\ell \rangle}{\max\Big\{ \sqrt{\sum_\ell \|D_\ell(x)\|_F^2} \sqrt{\sum_\ell \|\Delta R_\ell\|_F^2}, \varepsilon \Big\}}, \qquad \varepsilon = 10^{-12}. \tag{11}$$

Restricting the same expression to layer ℓ gives $r_{\ell,g}$. The global value is one cosine over concatenated matrices, not an average of layerwise cosines. When $\Delta R = 0$, these definitions return zero.

B.3 Trace Construction

For each router position, the logit ranking fixes the row sets used by the trace. If $\pi_{\ell t}$ orders the experts by decreasing $z_{\ell t,e}$, then

$$S_{\ell t} = \{\pi_{\ell t}(1), \ldots, \pi_{\ell t}(k)\}, \qquad B_{\ell t} = \{\pi_{\ell t}(k+1), \ldots, \pi_{\ell t}(2k)\}. \tag{12}$$

Ties follow PyTorch's `topk` ordering in the reported software version; we apply no additional tie-breaker.

For any row set A, the expanded restricted cosine is

$$c_{\ell t}(A) = \frac{\sum_{e \in A} \langle D_{\ell t,e}(x), \Delta r_{\ell e} \rangle}{\max\Big\{ \sqrt{\sum_{e \in A} \|D_{\ell t,e}(x)\|_2^2} \sqrt{\sum_{e \in A} \|\Delta r_{\ell e}\|_2^2}, \varepsilon \Big\}}, \tag{13}$$

where $\Delta r_{\ell e}$ is row e of ΔR_ℓ. Averaging over the N_x router positions gives

$$\begin{aligned} r_{\ell,s} &= \frac{1}{N_x}\sum_{t=1}^{N_x} c_{\ell t}(S_{\ell t}), \quad r_{\ell,b} = \frac{1}{N_x}\sum_{t=1}^{N_x} c_{\ell t}(B_{\ell t}), \\ r_{\ell,u} &= \frac{1}{N_x}\sum_{t=1}^{N_x} c_{\ell t}(U_{\ell t}), \quad r_{\ell,c} = r_{\ell,s} - r_{\ell,b}. \end{aligned} \tag{14}$$

The random set $U_{\ell t}$ is a deterministic size-k subset of the complement of $S_{\ell t}$. ZERO-BR and the path controls use this set; ROUTETRACE-S and ROUTETRACE-S-CAL do not. Appendix C.5 records the pseudorandom construction and dataset-specific seeds.

Eight contiguous groups partition the ordered router layers using `array_split`; earlier groups receive the extra layers when L is not divisible by eight. Each bin is an unweighted mean. The exact family-specific groups appear in Appendix C.1. Coordinates 1–3 are r_g, $L^{-1}\sum_\ell r_{\ell,s}$, and $L^{-1}\sum_\ell r_{\ell,c}$; coordinates 4–11, 12–19, and 20–27 are the eight-bin global, selected, and conditional sequences. Boundary alignment enters through the conditional sequence, while random alignment does not enter the 27-dimensional trace.

B.4 CALIBRATION AND SCORING

The three scoring modes share one alignment extractor, then use its outputs differently. ROUTETRACE-S and ROUTETRACE-S-CAL fit the 27-dimensional trace; ZERO-BR operates on unbinned layerwise alignments.

ROUTETRACE-S. Let $\{(\varphi_i, y_i)\}_{i=1}^N$ be the labeled records used to fit one scorer for each family and dataset, and let $N_y = |\{i : y_i = y\}|$. We compute

$$\begin{aligned} \mu &= \frac{1}{N}\sum_{i=1}^{N}\varphi_i, \\ \sigma_j &= \max\left\{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\varphi_{ij}-\mu_j)^2}, \varepsilon\right\}, \\ q_i &= (\varphi_i - \mu) \oslash \sigma, \qquad \bar{q}_y = \frac{1}{N_y}\sum_{i:y_i=y} q_i, \\ \Sigma_w &= \frac{1}{N-2}\sum_{y\in\{0,1\}}\sum_{i:y_i=y}(q_i-\bar{q}_y)(q_i-\bar{q}_y)^\top, \\ w &= (\Sigma_w + 0.5I)^+(\bar{q}_1 - \bar{q}_0). \end{aligned} \tag{15}$$

Here $+$ denotes the Moore–Penrose pseudoinverse. We use `numpy.linalg.pinv` with the NumPy 1.26.4 default `rcond` $= 10^{-15}$. The fitted score is

$$s_S(x) = ((\varphi(x) - \mu) \oslash \sigma)^\top w. \tag{16}$$

We use the linear score directly, without an intercept, class-prior term, or probability calibration. ROUTETRACE-S fits (μ, σ, w) on all records from the three matched shadows.

ROUTETRACE-S-CAL. Let $h(x)$ be the stored 64-character hexadecimal prompt hash. A prompt enters the calibration partition $\mathcal{C}$ exactly when

$$\mathrm{int}(h(x)[0{:}16], 16) \bmod 5 = 0. \tag{17}$$

Repeated occurrences of one prompt stay in the same partition. ROUTETRACE-S-CAL fits a separate parameter tuple (μ_C, σ_C, w_C) on $\mathcal{C}^c$ and applies the same scorer to held-out shadow records and targets.

For a requested FPR α, sort the m held-out non-member scores as $v_{(1)} \geq \cdots \geq v_{(m)}$ and set $k_\alpha = \lfloor \alpha m \rfloor$. The threshold and decision are

$$\tau_\alpha = \begin{cases} \mathrm{nextafter}(v_{(1)}, +\infty), & k_\alpha = 0, \\ \mathrm{nextafter}(v_{(k_\alpha)}, +\infty), & 1 \leq k_\alpha < m, \\ -\infty, & k_\alpha \geq m, \end{cases} \qquad \widehat{y}_\alpha(x) = \mathbf{1}[s_{S\text{-}Cal}(x) \geq \tau_\alpha]. \tag{18}$$

nextafter moves to the next representable value toward $+\infty$, excluding every score tied at the cutoff. The resulting empirical FPR on calibration non-members is at most α.

ZERO-BR. Write the unbinned layerwise alignments as

$$\boldsymbol{\rho} = (r_{\ell,g})_{\ell=1}^{L}, \quad \mathbf{s} = (r_{\ell,s})_{\ell=1}^{L}, \quad \mathbf{b} = (r_{\ell,b})_{\ell=1}^{L}, \quad \mathbf{u} = (r_{\ell,u})_{\ell=1}^{L}.$$

We floor both the sample standard deviation and the unscaled MAD at ε:

$$d_{\mathrm{sd}}(v) = \max\left\{ \sqrt{\frac{1}{|v|-1}\sum_j (v_j - \bar{v})^2}, \varepsilon \right\}, \qquad d_{\mathrm{mad}}(v) = \max\{\mathrm{median}_j\, |v_j - \mathrm{median}(v)|, \varepsilon\}\,. \tag{19}$$

Let $\boldsymbol{\delta}_{BR} = [\,\mathbf{s} - \mathbf{b}; \mathbf{s} - \mathbf{u}\,] \in \mathbb{R}^{2L}$. The four components and final score are

$$C_g = \frac{\mathrm{mean}(\boldsymbol{\rho})}{d_{\mathrm{sd}}(\boldsymbol{\rho})}, \qquad C_{\mathrm{rob}} = \frac{\mathrm{median}(\boldsymbol{\rho})}{d_{\mathrm{mad}}(\boldsymbol{\rho})},$$
$$C_s = \frac{\mathrm{mean}(\mathbf{s})}{d_{\mathrm{sd}}(\mathbf{s})}, \qquad C_{BR} = \frac{\mathrm{mean}(\boldsymbol{\delta}_{BR})}{d_{\mathrm{sd}}(\boldsymbol{\delta}_{BR})}, \tag{20}$$
$$s_{\text{Zero-BR}}(x) = C_g + C_{\mathrm{rob}} + C_s + C_{BR}.$$

ZERO-BR sums the four components with unit weights and normalizes only within the candidate; it fits no intercept or continuous coefficient and uses no cohort statistic or depth bins.

Unified procedure.

Algorithm 1: RouteTrace preparation and scoring
Prepare. Verify router-row compatibility, merge A_p, and compute ΔR for each pair. Fit ROUTETRACE-S on all shadow traces; for ROUTETRACE-S-CAL, split by Eq. equation 17, fit on $\mathcal{C}^c$, and obtain τ_α from calibration non-members. ZERO-BR uses Eq. equation 20 directly.
Score. The global pass returns $D_\ell, r_g, r_{\ell,g}$. The token-local pass returns $r_{\ell,s}, r_{\ell,b}, r_{\ell,c}$, and also $r_{\ell,u}$ for ZERO-BR. Build and score $\varphi(x)$ for ROUTETRACE-S/ROUTETRACE-S-CAL, or evaluate Eq. equation 20. Apply Eq. equation 18 for an S-Cal decision.

B.5 COMPLEXITY ANALYSIS

Trace extraction builds two independent computation graphs per candidate. The global pass differentiates with respect to router weights; the token-local pass differentiates with respect to the captured router logits. Its alignment core therefore costs

$$2C_{\mathrm{fwd}}(x) + 2C_{\mathrm{rev}}(x) + O\left(N_x \sum_{\ell=1}^{L} E_\ell d_\ell\right). \tag{21}$$

Here E_ℓ and d_ℓ are the row count and input width of router layer ℓ. Freezing non-router parameters does not remove reverse propagation through the network. Adapter merging and construction of ΔR occur once per process and checkpoint pair.

With N_s fitted shadow records and $p = 27$, ridge-LDA fitting costs $O(N_s p^2 + p^3)$, and applying a fitted score to an extracted trace costs $O(p)$. Sorting m calibration scores once per S-Cal scorer costs $O(m \log m)$. ZERO-BR requires $O(L)$ arithmetic after extraction.

C Experimental Protocol

C.1 Models and Checkpoints

We evaluate three public MoE families with distinct routing configurations: Qwen1.5-MoE-A2.7B-Chat (Qwen Team, 2024), DeepSeek-MoE-16B (Dai et al., 2024), and OLMoE-1B-7B-0125-Instruct (Muennighoff et al., 2025). Table 5 lists the routing dimensions. Each family uses its native tokenizer and a maximum sequence length of 256 tokens.

Table 5: Model and routing configurations used throughout the evaluation.

Family	MoE layers	Experts	Top-k	Tokens
Qwen1.5-MoE	24	60	4	256
DeepSeek-MoE	27	64	6	256
OLMoE	16	64	8	256

Each family–dataset pair contains target seeds 0–2 and shadow seeds 10–12. The same role construction, candidate identities, and evaluation sizes are used across families.

For the eight-bin depth profile, Qwen uses eight groups of three router layers and OLMoE uses eight groups of two. DeepSeek router layers 1–27 are partitioned as [1:4], [5:8], [9:12], [13:15], [16:18], [19:21], [22:24], and [25:27].

C.2 Data and Membership Splits

Natural Questions. The source pool contains 100,231 Natural Questions records (Kwiatkowski et al., 2019). We render each example with the literal template

```
Question: {query}
Answer: {answer}
```

and add no terminal punctuation. Exact duplicates and pairs whose word 5-gram Jaccard similarity is at least .8 are removed before split construction. Record identities remain fixed across model families.

Each NQ checkpoint is post-trained for one epoch on 4,096 records. Its audit cohort contains 512 sampled training members and 10,000 held-out non-members. Target and shadow training sets are disjoint. Target audit pools have no prompt identities in common with shadow audit pools, including their non-member records. Pooling the three shadows yields 1,536 member and 30,000 non-member checkpoint–record evaluations; the 10,000 non-member identities are shared across shadows, producing three evaluations per identity.

Dolly. Canonicalization reduces the 15,011 Dolly records (Databricks, 2023) to 14,996 unique rows. Dolly was selected after an exploratory Qwen pilot. The replication pool excludes all 1,536 pilot identities and includes the pilot prompts in duplicate filtering. Requiring 16–256 tokens under all three tokenizers and applying the same text filters leaves 10,111 eligible identities.

The six target/shadow roles receive disjoint 1,024-record training sets. Targets and shadows use separate 1,920-record non-member pools, and each checkpoint is audited on 256 members and its assigned non-member pool. Record identities are shared across families before tokenization. Dolly therefore tests replication under a combined domain and training-protocol change.

C.3 Fine-Tuning Protocol

Post-training updates every router matrix and rank-8 LoRA adapters on the attention query and value projections. LoRA uses scale 16 and zero dropout; expert parameters remain frozen. AdamW uses learning rate 2×10^{-4}, physical batch size one, and gradient accumulation over 16 examples. NQ training therefore performs 256 optimizer updates per checkpoint. Training uses bfloat16 and gradient checkpointing.

Candidate-gradient extraction uses shifted next-token cross-entropy in every family. NQ Qwen checkpoint training retains its native router auxiliary term with coefficient .001, whereas DeepSeek-

MoE and OLMoE set the coefficient to zero. During Qwen extraction, we disable the auxiliary term at both the model and configuration levels, so the candidate gradient reflects next-token cross-entropy alone. Thus all NQ audit gradients use the same objective while checkpoint updates retain the training objectives specified above. Dolly uses one exact exposure, or 64 optimizer updates per checkpoint, and both training and extraction use next-token cross-entropy with router auxiliary losses disabled.

C.4 Reference Models

For each family–dataset pair, RouteTrace-S fits one ridge-LDA scorer on 27-dimensional profiles from shadow checkpoints 10–12. The NQ fit pools 31,536 evaluations; the Dolly fit pools 6,528. RouteTrace-S-Cal uses the same features and model class but assigns prompts to calibration when the integer represented by the first 16 hexadecimal digits of their SHA-256 hash is zero modulo five. Repeated prompts remain in one partition.

The resulting NQ scorer/calibration partitions contain 25,247 and 6,289 evaluations, including 5,988 held-out non-members for threshold selection. The Dolly partitions contain 5,176 and 1,352 evaluations, including 1,203 held-out non-members. One fixed scorer produces both calibration and target scores. Target data select neither features, fitted parameters, nor thresholds.

The exact shared Qwen candidate-panel comparison uses 16 reference checkpoints and one calibration checkpoint on targets 1 and 2. Every retained record occurs in eight reference training sets and is excluded from eight. Reference-budget analysis uses $K \in \{2, 3, 4, 8, 16\}$ reference checkpoints plus calibration. LiRA reports eligible coverage and AUC on eligible records; RMIA scores the complete strict candidate panels. Zero-BR uses no auxiliary checkpoint at target evaluation; its fixed four-term formula family was selected once on three labeled development shadows.

C.5 Baselines and Metrics

The output panel includes post-training loss, base-to-post loss differences and ratios, answer-restricted variants, Min-K% (Shi et al., 2024), and Min-K%++ (Zhang et al., 2025). A 49-feature learned loss-profile attack uses the same three shadows as RouteTrace-S. For each family–dataset pair, prompt-deduplicated shadows select the scalar baseline and independently select $k \in \{10, 20, \ldots, 100\}$ for Min-K% and Min-K%++. The learned profile has 24 full-sequence loss statistics, 24 answer-restricted statistics, and token length. Each 24-dimensional block contains the mean and five quantiles (10%, 25%, 50%, 75%, 90%), represented at the base and post-trained checkpoints, together with their base-minus-post differences and base/post ratios. It uses a fixed ridge-LDA with ridge coefficient 1, fitted after standardizing the pooled shadow observations. Repeated prompts across shadows remain separate observations; prompt-grouped folds keep these repetitions together for calibration diagnostics. The final ranking scorer is fitted on all three shadows, without a feature-subset search.

EZ-MIA uses the architecture-matched base checkpoint as its reference and follows the paper-defined handling of top-1-error positions (Ilić et al., 2026); the repository edge behavior is retained as a named sensitivity.

Records are fixed within each control comparison. The nine-feature parameter and state comparisons use 512 members and 2,048 non-members per target, with the non-members selected from the 10,000-record canonical pool by a fixed hash rule. The Qwen feature-group study retains all 512 members and 10,000 non-members with its previously fitted scorers. The path analysis scores selected top-k router rows, the equal-size routing boundary, a seeded random non-selected set, and all rows. Same-state mechanism comparisons replace router alignment with query/value alignment or combine both views at the same nine-dimensional resolution. The separate Qwen dense control matches training steps and trainable-parameter count but uses an attention-only dense architecture.

For the random set, let i, j, and d be the zero-based candidate position, router-order layer index, and draw index. Its PyTorch generator uses $s_0 + d \cdot 10{,}000{,}019 + iL + j$, with $s_0 = 20260805$ for NQ and $s_0 = 20260825$ for Dolly. For each token, we draw one uniform value per expert, mask the selected rows, and take the k smallest remaining values. All reported Zero-BR scores use the first draw, $d = 0$.

The primary ranking metric is ROC-AUC. Descriptive target ROC points report TPR at .1%, 1%, and 5% FPR for NQ and at 1% and 5% for Dolly. Fixed RouteTrace-S-Cal thresholds report TPR

together with realized target FPR. Point estimates are computed from frozen score vectors. For the main ranking results, per-target AUC and paired AUC differences use 10,000 label-stratified record-bootstrap draws, with shared resamples for paired comparisons. The path controls in Appendix E.3 use 5,000 unstratified record-bootstrap draws, also shared for paired comparisons. Fixed-threshold TPR and FPR use Wilson binomial intervals. Reference-budget ranges for $K < 16$ are the 2.5th and 97.5th percentiles over 64 fixed reference subsets, not confidence intervals. Family and cross-family means are descriptive; record-bootstrap intervals condition on the frozen checkpoints, scorers, and candidate panels.

All feature, attack, baseline, threshold, and defense choices are fixed without target outcomes; target labels are used only to compute the reported metrics.

C.6 Software and Hardware

Experiments use NVIDIA A100-SXM4-80GB and H800 GPUs with PyTorch 2.6.0, CUDA 12.4, Transformers 4.55.0, PEFT 0.20.0, Accelerate 1.3.0, and NumPy 1.26.4. We report auxiliary checkpoint counts separately from GPU-hour estimates. Victim post-training is outside attack cost because the paired-checkpoint setting supplies the released base and post-trained checkpoints to the auditor.

D COMPLETE ATTACK RESULTS

D.1 CROSS-FAMILY RESULTS

ROUTETRACE-S exceeds every evaluated output-based attack on all 18 target checkpoints; its ordering relative to ZERO-BR varies across families and datasets. Table 6 expands the loss comparison beyond the representative methods in Table 1. Each column averages the three target checkpoints for one family and dataset.

Table 6: Mean ROC-AUC over three target checkpoints for ROUTETRACE-S and 13 loss-derived baselines. Q, D, and O denote Qwen, DeepSeek-MoE, and OLMoE. Bold marks the highest value in each column.

Method	NQ-Q	NQ-D	NQ-O	Dolly-Q	Dolly-D	Dolly-O
ROUTETRACE-S	**.9466**	**.9322**	**.9657**	**.8717**	**.8589**	**.8540**
Answer Min-K 10%	.5934	.5590	.5911	.5392	.5235	.5347
Answer Min-K 20%	.5950	.5604	.5968	.5399	.5223	.5375
Answer Min-K 40%	.5931	.5598	.5990	.5364	.5199	.5345
Base–post answer difference	.6660	.7260	.6323	.6306	.6882	.6053
Base/post answer ratio	.6756	.6977	.7207	.6355	.6550	.6217
Base–post loss difference	.6502	.6346	.6229	.6567	.6109	.5974
Base/post loss ratio	.6716	.6311	.7150	.6719	.6068	.6299
Learned loss profile	.7754	.7964	.7612	.6998	.6896	.6612
Min-K 10%	.6053	.5712	.6041	.5904	.5549	.5823
Min-K 20%	.6049	.5710	.6076	.5870	.5538	.5830
Min-K 40%	.6013	.5685	.6081	.5757	.5473	.5730
Post answer loss	.5880	.5561	.5955	.5324	.5182	.5307
Post loss	.5949	.5634	.6037	.5652	.5396	.5620

Across the 13 loss-derived baselines, ROUTETRACE-S's per-target AUC advantage ranges from .1298 to .3853, and every paired interval lies above zero. The same direction holds in the separate six-attack comparison covering paper and repository EZ-MIA, shadow-selected Min-K% and Min-K%++, negative post-training loss, and base-minus-post loss. The accompanying artifact contains the per-target baseline values and intervals.

D.2 LOW-FPR RESULTS

At the nominal-1% setting, realized FPR is 1.032% on NQ and 1.100% on Dolly; the corresponding macro TPRs are 24.02% and 10.94%. Tables 7 and 8 expand these means to individual targets and report ROUTETRACE-S-CAL ranking AUC from the same scorer.

Table 7: NQ ROUTETRACE-S-CAL AUC and fixed-threshold results. AUC brackets are 95% record-bootstrap intervals. Threshold columns report TPR and realized FPR with 95% Wilson intervals, in percent.

Family	Target	AUC	Nominal .1%	Nominal 1%	Nominal 5%
Qwen	0	.9499 [.9430,.9564]	TPR 4.30 [2.85,6.42] FPR .14 [.08,.23]	TPR 21.09 [17.78,24.84] FPR .86 [.70,1.06]	TPR 66.80 [62.61,70.74] FPR 4.17 [3.80,4.58]
Qwen	1	.9523 [.9451,.9588]	TPR 4.69 [3.17,6.88] FPR .10 [.05,.18]	TPR 22.07 [18.69,25.86] FPR .84 [.68,1.04]	TPR 67.58 [63.41,71.49] FPR 4.39 [4.01,4.81]
Qwen	2	.9381 [.9289,.9467]	TPR 5.27 [3.65,7.56] FPR .21 [.14,.32]	TPR 26.56 [22.92,30.55] FPR 1.34 [1.13,1.58]	TPR 69.53 [65.41,73.36] FPR 5.84 [5.40,6.32]
DeepSeek	0	.9389 [.9311,.9462]	TPR 1.37 [.66,2.79] FPR .12 [.07,.21]	TPR 20.90 [17.60,24.63] FPR 1.13 [.94,1.36]	TPR 69.73 [65.61,73.55] FPR 6.67 [6.20,7.18]
DeepSeek	1	.9319 [.9236,.9396]	TPR 1.56 [.79,3.05] FPR .03 [.01,.09]	TPR 21.09 [17.78,24.84] FPR 1.42 [1.21,1.67]	TPR 69.14 [65.01,72.99] FPR 7.05 [6.56,7.57]
DeepSeek	2	.9286 [.9192,.9372]	TPR .78 [.30,1.99] FPR .06 [.03,.13]	TPR 19.92 [16.69,23.60] FPR 1.06 [.88,1.28]	TPR 59.57 [55.26,63.73] FPR 5.92 [5.47,6.40]
OLMoE	0	.9685 [.9633,.9732]	TPR 2.93 [1.78,4.78] FPR .09 [.05,.17]	TPR 27.54 [23.85,31.56] FPR .80 [.64,.99]	TPR 87.50 [84.35,90.09] FPR 5.70 [5.26,6.17]
OLMoE	1	.9620 [.9551,.9680]	TPR 4.88 [3.33,7.11] FPR .11 [.06,.20]	TPR 32.03 [28.14,36.19] FPR 1.13 [.94,1.36]	TPR 88.09 [84.99,90.61] FPR 7.34 [6.85,7.87]
OLMoE	2	.9661 [.9586,.9725]	TPR 3.32 [2.08,5.25] FPR .03 [.01,.09]	TPR 25.00 [21.44,28.93] FPR .71 [.56,.89]	TPR 84.18 [80.77,87.08] FPR 4.89 [4.48,5.33]

Table 8: Dolly ROUTETRACE-S-CAL AUC and fixed-threshold results. AUC brackets are 95% record-bootstrap intervals. Threshold columns report TPR and realized FPR with 95% Wilson intervals, in percent.

Family	Target	AUC	Nominal 1%	Nominal 5%
Qwen	0	.8739 [.8515,.8943]	TPR 24.61 [19.73,30.24] FPR 1.56 [1.10,2.22]	TPR 54.30 [48.18,60.29] FPR 7.24 [6.16,8.49]
Qwen	1	.8739 [.8525,.8944]	TPR 19.14 [14.79,24.40] FPR 1.77 [1.27,2.46]	TPR 43.75 [37.81,49.87] FPR 4.58 [3.74,5.61]
Qwen	2	.8688 [.8460,.8904]	TPR 22.27 [17.60,27.75] FPR 1.20 [.80,1.79]	TPR 48.83 [42.77,54.92] FPR 5.57 [4.63,6.69]
DeepSeek	0	.8485 [.8256,.8704]	TPR 3.91 [2.14,7.04] FPR 1.25 [.84,1.85]	TPR 22.66 [17.95,28.17] FPR 5.21 [4.30,6.29]
DeepSeek	1	.8637 [.8413,.8844]	TPR 3.91 [2.14,7.04] FPR .52 [.28,.96]	TPR 20.70 [16.19,26.08] FPR 2.81 [2.16,3.65]
DeepSeek	2	.8631 [.8428,.8828]	TPR 3.12 [1.59,6.04] FPR .68 [.40,1.16]	TPR 26.17 [21.17,31.88] FPR 4.06 [3.27,5.04]
OLMoE	0	.8553 [.8331,.8770]	TPR 6.25 [3.88,9.91] FPR .78 [.47,1.29]	TPR 39.45 [33.66,45.55] FPR 7.40 [6.31,8.65]
OLMoE	1	.8333 [.8072,.8580]	TPR 7.81 [5.11,11.76] FPR 1.46 [1.01,2.10]	TPR 41.02 [35.17,47.13] FPR 8.39 [7.23,9.71]
OLMoE	2	.8596 [.8387,.8793]	TPR 7.42 [4.80,11.30] FPR .68 [.40,1.16]	TPR 34.38 [28.83,40.39] FPR 5.83 [4.87,6.97]

On NQ, the nominal-.1% and nominal-5% thresholds yield macro TPRs of 3.23% and 73.57% at realized FPRs of .099% and 5.774%. Dolly's nominal-5% threshold yields 36.81% macro TPR at 5.677% realized FPR. TPR values read directly from each target ROC curve are ranking diagnostics. The fixed-threshold columns instead use thresholds selected on held-out shadow non-members.

D.3 PER-SEED RESULTS

ROUTETRACE-S and ZERO-BR exchange the lead across target checkpoints. Table 9 reports their AUCs and paired differences. On NQ, ROUTETRACE-S is favored on every DeepSeek-MoE and OLMoE target, while the Qwen ordering changes across targets. On Dolly, ZERO-BR is favored on all three Qwen targets; the DeepSeek-MoE and OLMoE comparisons are mixed.

Table 9: Per-target ROC-AUC for ROUTETRACE-S and ZERO-BR. Brackets give 95% record-bootstrap intervals. The last column is the paired ROUTETRACE-S-minus-ZERO-BR difference; positive values favor ROUTETRACE-S. Bold marks a method favored by an interval excluding zero.

Data	Family	Target	ROUTETRACE-S	ZERO-BR	Paired difference
NQ	Qwen	0	**.9497 [.9428,.9562]**	.9246 [.9109,.9378]	+.0251 [+.0146,+.0364]
NQ	Qwen	1	.9523 [.9452,.9588]	.9562 [.9472,.9644]	-.0039 [-.0108,+.0032]
NQ	Qwen	2	.9378 [.9286,.9465]	**.9532 [.9440,.9614]**	-.0154 [-.0239,-.0065]
NQ	DeepSeek	0	**.9381 [.9302,.9454]**	.9188 [.9058,.9311]	+.0192 [+.0077,+.0312]
NQ	DeepSeek	1	**.9310 [.9227,.9388]**	.9125 [.8990,.9252]	+.0185 [+.0073,+.0305]
NQ	DeepSeek	2	**.9275 [.9181,.9362]**	.8926 [.8777,.9066]	+.0349 [+.0233,+.0469]
NQ	OLMoE	0	**.9687 [.9635,.9733]**	.9518 [.9436,.9593]	+.0170 [+.0102,+.0239]
NQ	OLMoE	1	**.9620 [.9552,.9681]**	.9128 [.8984,.9263]	+.0492 [+.0375,+.0616]
NQ	OLMoE	2	**.9663 [.9589,.9727]**	.9416 [.9305,.9518]	+.0247 [+.0174,+.0328]
Dolly	Qwen	0	.8726 [.8500,.8933]	**.9216 [.9005,.9410]**	-.0490 [-.0650,-.0327]
Dolly	Qwen	1	.8739 [.8525,.8944]	**.9132 [.8917,.9332]**	-.0393 [-.0562,-.0219]
Dolly	Qwen	2	.8687 [.8457,.8904]	**.9109 [.8879,.9324]**	-.0422 [-.0595,-.0255]
Dolly	DeepSeek	0	.8490 [.8263,.8707]	.8640 [.8409,.8862]	-.0150 [-.0351,+.0053]
Dolly	DeepSeek	1	.8646 [.8424,.8852]	.8635 [.8422,.8833]	+.0011 [-.0180,+.0198]
Dolly	DeepSeek	2	**.8631 [.8428,.8828]**	.8384 [.8137,.8624]	+.0248 [+.0040,+.0457]
Dolly	OLMoE	0	.8584 [.8364,.8794]	.8647 [.8407,.8875]	-.0064 [-.0252,+.0140]
Dolly	OLMoE	1	.8375 [.8116,.8621]	.8491 [.8240,.8731]	-.0116 [-.0300,+.0068]
Dolly	OLMoE	2	.8660 [.8460,.8849]	.8814 [.8579,.9033]	-.0154 [-.0357,+.0051]

The paired intervals favor ROUTETRACE-S on eight targets and ZERO-BR on four. The remaining six intervals cross zero.

D.4 STATISTICAL UNCERTAINTY

Per-target AUC intervals use 10,000 bootstrap draws stratified by membership label. Paired differences apply the same member and non-member resample to both scores. These intervals capture record-sampling uncertainty within each target checkpoint. Because checkpoints are not resampled, they do not quantify variation across training runs or model families.

Fixed-threshold TPR and FPR use separate Wilson intervals with the member and non-member counts as their respective denominators and condition on the chosen calibration split and threshold. Shadow retraining and alternative splits or thresholds are outside these pointwise intervals. Family and macro means weight target checkpoints equally; record resampling does not provide confidence intervals for variation across checkpoints.

E MECHANISM EVIDENCE

All results in this appendix use NQ. We report learned nine-feature scorers, unbinned layerwise alignments, and separately fitted Qwen parameter controls. Because these scores use different estimators, we compare AUC only within each table.

E.1 ROUTER, Q/V, AND EFFECTIVE-ALL

With the hybrid state, feature count, and ridge-LDA fitting protocol held fixed, Effective-Router-9D has higher AUC than Q/V-9D and Effective-All-9D across all three families. All three parameter views use nine features and are fitted separately on the same shadow records; ROUTETRACE-S uses the 27-dimensional trace. The parameter comparisons and the state sweep below use 512 members and 2,048 non-members per target, fixed across all compared views and states.

Table 10: NQ per-target AUC for the matched nine-feature scorers at the hybrid state. Eff.-Router and Eff.-All denote Effective-Router-9D and Effective-All-9D. Paired differences are Router minus the named comparator; positive values favor the router scorer. Brackets give paired 95% record-bootstrap intervals.

Family	Target	Eff.-Router	Q/V-9D	Eff.-All	Router–Q/V	Router–All
Qwen	0	.9476	.5538	.8850	+.3937 [+.3659,+.4218]	+.0626 [+.0511,+.0751]
Qwen	1	.9480	.5536	.8663	+.3945 [+.3668,+.4220]	+.0817 [+.0692,+.0949]
Qwen	2	.9284	.5655	.8864	+.3629 [+.3358,+.3902]	+.0420 [+.0312,+.0528]
DeepSeek	0	.9361	.6171	.8618	+.3189 [+.2944,+.3433]	+.0743 [+.0647,+.0841]
DeepSeek	1	.9222	.6043	.8536	+.3178 [+.2943,+.3419]	+.0685 [+.0595,+.0780]
DeepSeek	2	.9164	.6128	.8560	+.3036 [+.2802,+.3267]	+.0603 [+.0520,+.0689]
OLMoE	0	.9665	.5523	.8659	+.4142 [+.3865,+.4421]	+.1006 [+.0847,+.1169]
OLMoE	1	.9568	.5437	.8400	+.4131 [+.3854,+.4406]	+.1168 [+.1002,+.1337]
OLMoE	2	.9620	.5191	.8524	+.4429 [+.4149,+.4711]	+.1095 [+.0932,+.1270]

Router-minus-Q/V differences range from .3036 to .4429, and router-minus-Effective-All differences range from .0420 to .1168. Every paired interval excludes zero. The Q/V adapter remains part of the hybrid state for all three scorers; the comparison changes the parameter update aligned with the candidate gradient. The post-training procedure is unchanged across these scorers.

E.2 BASE, HYBRID, AND POST STATES

For Qwen and OLMoE, AUC rises from B to H and changes little from H to P. All three states use the same ΔR, the same nine-feature construction, and a scorer fitted on shadow features extracted in that state. B uses base Q/V with the base router, H uses post-trained Q/V with the base router, and P uses post-trained Q/V with the post-trained router. These are alternative analysis states for one checkpoint pair, not stages of training.

Table 11: NQ per-target Effective-Router-9D AUC under the base (B), hybrid (H), and post (P) analysis states. All states use the same router update. Brackets give paired 95% record-bootstrap intervals.

Family	Target	B	H	P	H–B	P–H
Qwen	0	.8596	.9476	.9508	+.0879 [+.0732,+.1036]	+.0032 [-.0095,+.0154]
Qwen	1	.8738	.9480	.9418	+.0742 [+.0604,+.0882]	-.0063 [-.0209,+.0081]
Qwen	2	.8529	.9284	.9371	+.0755 [+.0604,+.0906]	+.0087 [-.0090,+.0262]
OLMoE	0	.8592	.9665	.9703	+.1074 [+.0923,+.1231]	+.0038 [-.0027,+.0101]
OLMoE	1	.8308	.9568	.9659	+.1260 [+.1090,+.1438]	+.0091 [+.0017,+.0165]
OLMoE	2	.8277	.9620	.9665	+.1342 [+.1169,+.1526]	+.0045 [-.0048,+.0141]

The H-minus-B intervals lie above zero on all six targets. For P minus H, only OLMoE target 1 has an interval above zero; the other five intervals cross zero. Spearman correlations between B and H score vectors range from .5531 to .7022, while correlations between H and P range from .6062 to .7206. We report these correlations as point estimates.

E.3 PATH COMPARISONS

The path controls compare selected rows with equal-sized boundary and random sets and with all-row aggregation. They use the same candidates and router update. For a token-specific row set $A_{\ell t}$, the scalar score is

$$s_A(x) = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{N_x}\sum_{t=1}^{N_x} c_{\ell t}(A_{\ell t}), \tag{22}$$

where $c_{\ell t}$ is the restricted negative-gradient alignment in equation 13. We use the selected, boundary, random, or complete row set at each token. These scores use no depth bins or learned weights. All-row averages token-local cosines, rather than using r_g or $r_{\ell,g}$, which aggregate gradients before taking the cosine. The path-control intervals use 5,000 unstratified record-bootstrap resamples of each fixed target cohort, shared between selected and control scores for paired differences.

Table 12: NQ per-target AUC for unbinned layerwise path controls. Selected, boundary, and random use equal-sized row sets; all-row uses the complete router matrix.

Family	Target	Selected	Boundary	Random	All-row
Qwen	0	.8054	.6025	.5568	.8187
Qwen	1	.8651	.6165	.5456	.8749
Qwen	2	.8176	.5974	.5630	.8270
DeepSeek	0	.7588	.5873	.5594	.7429
DeepSeek	1	.7697	.5851	.5606	.7528
DeepSeek	2	.7447	.5674	.5450	.7269
OLMoE	0	.8232	.5219	.5191	.8199
OLMoE	1	.7582	.5247	.5344	.7522
OLMoE	2	.8213	.5621	.5533	.8234

Table 13: Paired AUC differences for the NQ path controls. Each column is selected minus the named control; brackets give 95% record-bootstrap intervals.

Family	Target	Selected–Boundary	Selected–Random	Selected–All
Qwen	0	+.2029 [+.1781,+.2275]	+.2486 [+.2211,+.2761]	-.0133 [-.0146,-.0119]
Qwen	1	+.2486 [+.2255,+.2722]	+.3195 [+.2927,+.3459]	-.0098 [-.0112,-.0085]
Qwen	2	+.2202 [+.1988,+.2424]	+.2546 [+.2299,+.2800]	-.0094 [-.0106,-.0082]
DeepSeek	0	+.1715 [+.1504,+.1925]	+.1994 [+.1737,+.2262]	+.0159 [+.0146,+.0173]
DeepSeek	1	+.1846 [+.1636,+.2041]	+.2091 [+.1842,+.2335]	+.0169 [+.0154,+.0184]
DeepSeek	2	+.1773 [+.1567,+.1978]	+.1997 [+.1764,+.2242]	+.0179 [+.0164,+.0194]
OLMoE	0	+.3013 [+.2779,+.3247]	+.3041 [+.2792,+.3297]	+.0033 [+.0022,+.0043]
OLMoE	1	+.2335 [+.2104,+.2568]	+.2239 [+.1975,+.2509]	+.0060 [+.0043,+.0075]
OLMoE	2	+.2592 [+.2363,+.2815]	+.2680 [+.2425,+.2940]	-.0022 [-.0029,-.0014]

Selected rows exceed boundary and random rows on all nine targets; the 18 paired intervals all exclude zero. All-row aggregation has higher AUC on the three Qwen targets and OLMoE target 2, while selected rows lead on the other five. ROUTETRACE-S retains both global and selected-minus-boundary depth features.

E.4 QWEN PARAMETER CONTROLS

Qwen has two additional controls: Q/V alignment in the same MoE and an attention-only dense checkpoint.

Table 14: Qwen NQ parameter controls. Q/V alignment is a separately fitted attention-alignment control evaluated with the same candidate loss; it is distinct from Q/V-9D in Table 10. Brackets give 95% record-bootstrap intervals, and means average the three targets.

Control	Target 0	Target 1	Target 2	Mean
Q/V alignment	.6462 [.6220,.6707]	.6265 [.6032,.6488]	.6372 [.6121,.6620]	.6367
Dense attention-only	.5245 [.4985,.5506]	.5367 [.5108,.5632]	.5426 [.5177,.5675]	.5346

The Q/V alignment control reaches .6367 mean AUC, and the dense control reaches .5346. The dense checkpoint matches candidate identities, training steps, and trainable-parameter count, but changes both the architecture and adapted path. The comparison does not identify either effect separately.

F Reference Counts and Resource Estimates

F.1 Shadow Count

To isolate shadow count, we refit RouteTrace-S on every one-, two-, and three-shadow subset of the existing checkpoints. For one and two shadows, the reported mean averages the three possible subsets across three targets; the three-shadow row has one subset.

Table 15: RouteTrace-S AUC after fitting on one-, two-, and three-shadow subsets. Brackets give the minimum and maximum across subset–target combinations. Each Δ entry gives the mean / largest AUC decrease from the three-shadow scorer.

Data	Family	One shadow	Two shadows	Three shadows
NQ	Qwen	.9422 [.9201,.9570] Δ -.0044 / -.0177	.9453 [.9333,.9533] Δ -.0014 / -.0051	.9466 [.9378,.9523]
NQ	DeepSeek	.9275 [.9179,.9381] Δ -.0047 / -.0127	.9308 [.9254,.9403] Δ -.0014 / -.0055	.9322 [.9275,.9381]
NQ	OLMoE	.9645 [.9562,.9704] Δ -.0012 / -.0058	.9652 [.9582,.9697] Δ -.0004 / -.0038	.9657 [.9620,.9687]
Dolly	Qwen	.8696 [.8612,.8842] Δ -.0021 / -.0114	.8712 [.8655,.8799] Δ -.0005 / -.0072	.8717 [.8687,.8739]
Dolly	DeepSeek	.8562 [.8361,.8667] Δ -.0027 / -.0129	.8582 [.8428,.8655] Δ -.0007 / -.0062	.8589 [.8490,.8646]
Dolly	OLMoE	.8489 [.8208,.8641] Δ -.0050 / -.0166	.8529 [.8334,.8686] Δ -.0011 / -.0063	.8540 [.8375,.8660]

Across all three families on both datasets, the one- and two-shadow mean AUCs remain within .0051 of the three-shadow scorer. The largest loss for an individual subset–target configuration is .0177. The displayed ranges summarize variation over the tested subsets and targets; they do not estimate variation from training new shadows.

After fitting, RouteTrace-S and RouteTrace-S-Cal score new candidates using the target checkpoint pair and their fitted parameters; they do not reload the shadow checkpoints. Zero-BR uses the target pair with a fixed four-term rule selected using three labeled development shadows, without fitting a family–dataset-specific scorer.

F.2 Strict Qwen Comparison

The shared candidate intersection contains 7,817 records for Qwen target 1 and 7,812 for target 2. These counts cover 95.42% and 95.36% of the 8,192-record reference set, or 74.36% and 74.32% of the full 10,512-record target set. All four attacks in Table 16 use these same records.

Table 16: ROC-AUC on the same candidate records for Qwen targets 1 and 2. Brackets give per-target 95% record-bootstrap intervals. Bold marks the highest point estimate in each AUC column.

Method	Target 1	Target 2	Mean	Auxiliary checkpoints
RouteTrace-S	.9554 [.9479,.9623]	.9359 [.9247,.9461]	.9457	3 matched shadows
Zero-BR	.9548 [.9441,.9642]	.9516 [.9405,.9616]	.9532	0 target-time; 3 development
LiRA	.9764 [.9669,.9844]	.9820 [.9759,.9869]	.9792	16 references + 1 calibration
RMIA	**.9768 [.9670,.9852]**	**.9850 [.9801,.9889]**	**.9809**	16 references + 1 calibration

RMIA and LiRA attain the highest AUCs on both targets. They use 16 reference checkpoints and one calibration checkpoint; RouteTrace-S uses three matched shadows. The Zero-BR row separates target scoring, which uses no auxiliary checkpoint, from development on three shadows.

F.3 Fixed Reference Subsets

LiRA can score a candidate only when the selected references contain both trained-on and held-out observations for that record. Coverage is the eligible fraction of each target's strict candidate set:

7,817 records for target 1 and 7,812 for target 2. For $K \in \{2, 3, 4, 8\}$, we evaluate 64 salted reference subsets chosen without target labels; $K = 16$ has one complete set.

Table 17: Reference-subset analysis on Qwen targets 1 and 2. Every row includes one calibration checkpoint. For $K < 16$, brackets are 2.5th–97.5th percentiles over 64 subsets, not confidence intervals. Coverage uses the 7,817- and 7,812-record strict candidate sets. LiRA AUC is conditional on eligible records; RMIA uses all queried records.

References	Total aux.	Subsets	LiRA coverage	LiRA AUC	RMIA AUC
2	3	64	53.34% [51.11,55.20]	.9705 [.9639,.9769]	.9285
3	4	64	79.78% [78.40,81.50]	.9737 [.9683,.9778]	.9532
4	5	64	92.35% [91.12,93.30]	.9752 [.9720,.9789]	.9645
8	9	64	99.98% [99.93,100.00]	.9783 [.9769,.9796]	.9768
16	17	1	100.00%	.9792	.9809

Because the eligible records change with K, each LiRA AUC is computed on a different subset rather than one fixed cohort. The bracketed ranges describe variation across the fixed subsets, not record-sampling uncertainty. RMIA does not require both in- and out-reference incidence and is evaluated on all queried records.

At two references plus calibration, RMIA's mean AUC is .9285, below the three-shadow ROUTETRACE-S value of .9457. At three references plus calibration, RMIA reaches .9532, above ROUTETRACE-S. This comparison uses the complete strict panels, while the LiRA columns retain their eligibility-conditioned interpretation.

F.4 RESOURCE ACCOUNTING

The reference GPU-hour values are estimates for auxiliary-checkpoint training and extraction, not measured end-to-end runtimes. For $K = 2, 3, 4, 8, 16$, the estimated ranges are 5.4–7.2, 7.2–9.6, 9.0–12.0, 16.2–21.6, and 30.6–40.8 GPU-hours. LiRA and RMIA share the same reference and calibration pool, so this work is counted once.

Table 18: Partial RouteTrace resource records and the $K = 16$ reference estimate. RouteTrace setup includes three shadow runs and feature extraction; per-target GPU-hours exclude victim training. Hardware and measurement bases differ across rows.

Workload	Setup GPU-h	Per target GPU-h	Setup GiB	Target GiB	Measurement basis
RouteTrace DeepSeek	26.50	9.84	1.46	.44	recorded/reconstructed
RouteTrace OLMoE	25.72	5.07	1.14	.36	reconstructed
RouteTrace Qwen	–	2.146	–	–	recorded token-local
Strict references ($K = 16$)	30.6–40.8	–	–	–	estimated

The RouteTrace measurements are partial and come from different hardware. DeepSeek-MoE and OLMoE combine recorded or reconstructed timings; the Qwen value of 2.146 GPU-hours covers token-local extraction only.

ZERO-BR reuses target extraction and adds $O(L)$ scoring, so zero target-time auxiliary checkpoints does not mean zero computation. All attacks start from the same post-trained target, and victim training is excluded. Because matched GPU time, peak memory, and storage measurements are incomplete, the current data compare auxiliary-checkpoint counts and eligibility coverage rather than total cost.

G ROUTER INTERVENTIONS AND UTILITY

G.1 INTERVENTION SETUP

We evaluate router clipping and router freezing on three families, using three targets and three shadows for each intervention. Clean and intervention runs use AdamW weight decay .01. Freezing holds every router and expert parameter fixed while updating the query/value LoRA adapters. Clipping rescales each record's joint router gradient as

$$g_{R,i} \leftarrow g_{R,i} \min\left(1, \frac{C}{\max(\|g_{R,i}\|_2, 10^{-30})}\right), \tag{23}$$

leaves the query/value LoRA gradients unchanged, and averages the accumulated gradients over 16 records before each update. The clipping rule adds no noise and provides no privacy accounting.

For each family, a separate public set of 256 records defines four clipping thresholds at the 10th, 25th, 50th, and 75th percentiles of the router-gradient norm. We choose the strongest level that retains at least 95% of the clean member training gain and increases public non-member answer NLL by at most 3%. All three families select the 10th percentile, giving $C = 8.1456$ for Qwen, 1.3222 for DeepSeek-MoE, and 8.1898 for OLMoE. We choose C without using attack scores or target labels. For each intervention, we fit the scorer and select baseline settings and thresholds on shadows trained with the same intervention; target labels are revealed after target scoring. The weight decay study in Appendix G.4 instead transfers attacks fitted on clean checkpoints.

G.2 CLIPPING AND FREEZING

Router clipping increases family-mean ROUTETRACE-S AUC by +.0054 to +.0070, with all three family-mean paired 95% record-bootstrap intervals above zero. Individual target changes vary. Clipped mean AUC ranges from .9391 to .9715 and exceeds the strongest evaluated non-RouteTrace attack by .1042–.1986.

Table 19: Family-mean AUC change under router interventions (intervention minus clean). Positive values denote stronger membership separation. Means average three target checkpoints. With freezing, $\Delta R = 0$; ROUTETRACE-S and ROUTETRACE-S-CAL are constant. The ROUTETRACE-S entries equal .5 minus clean AUC; ZERO-BR is omitted.

	Router clipping			Router freezing		
Attack	Qwen	DeepSeek	OLMoE	Qwen	DeepSeek	OLMoE
ROUTETRACE-S	+.0054	+.0070	+.0059	-.4466	-.4322	-.4657
ZERO-BR	+.0074	+.0117	+.0132	–	–	–
Attention alignment	-.0049	-.0012	+.0126	-.1093	-.0378	-.0711
EZ-MIA	+.0042	+.0170	+.0135	-.2049	-.2359	-.2000
Learned loss	+.0078	+.0167	+.0117	-.2176	-.2462	-.2107
Min-K%	+.0031	+.0008	+.0105	-.0689	-.0442	-.0697
Min-K%++	+.0068	+.0023	+.0092	-.1594	-.0965	-.1737
Selected scalar	+.0057	+.0214	+.0126	-.1434	-.2008	-.1864

The evaluated non-RouteTrace attacks also show no common reduction: only two family-level changes in Table 19 are negative, and both intervals cross zero.

Router freezing sets $\Delta R = 0$, making the ROUTETRACE-S score constant on all nine targets. The reported AUC of .5 is a convention for a constant score, not an empirical ranking result. The other attacks decline broadly; the strongest remaining attack reaches .5579 on Qwen, .5820 on DeepSeek-MoE, and .5505 on OLMoE. Freezing changes the adaptation mechanism and provides no formal privacy guarantee.

G.3 Utility and Calibration

On the public utility set, clipping changes family-mean NQ F1 and development-training answer NLL by at most .0028 and .0025 relative to clean. Frozen checkpoints have higher mean NQ F1 and lower mean NLL on both the development-training and public non-member sets in all three families.

Table 20: Utility measurements on the public development set used to select clipping strength. Values are means over three checkpoints; arrows indicate the preferred task direction. NQ exact match is zero for every arm and is omitted.

Family	Arm	Train NLL ↓	Public-NM NLL ↓	NQ F1 ↑	Route overlap	$\|\Delta R\|_2$
Qwen	Clean	1.4794	1.4946	.2346	.8135	4.3239
	Clipped	1.4807	1.4955	.2324	.8109	4.4640
	Frozen	1.4605	1.4752	.2366	.8639	0
DeepSeek	Clean	1.4863	1.4950	.2401	.8836	4.9608
	Clipped	1.4864	1.4950	.2378	.8831	5.0050
	Frozen	1.4844	1.4934	.2437	.9110	0
OLMoE	Clean	1.5899	1.6162	.2320	.9006	3.6261
	Clipped	1.5874	1.6134	.2347	.8981	3.8510
	Frozen	1.5793	1.6036	.2363	.9224	0

Because the same public set selects the clipping level, these utility results describe the selected NQ setting rather than an independent downstream validation.

Clipped RouteTrace-S-Cal thresholds transfer unevenly across families. At a nominal 1% shadow FPR, the realized target FPR is .803% on Qwen, 1.723% on DeepSeek-MoE, and 1.103% on OLMoE. At the nominal 5% threshold, DeepSeek-MoE and OLMoE reach 8.543% and 6.580% realized FPR.

Table 21: Realized FPR / TPR in percent for clipped RouteTrace-S-Cal thresholds selected on shadows trained with the same clipping rule. Columns are labeled by nominal shadow FPR; values average three target checkpoints.

Family	Nominal .1%	Nominal 1%	Nominal 5%
Qwen	.097 / 2.54	.803 / 19.66	4.447 / 66.02
DeepSeek	.097 / 1.76	1.723 / 25.91	8.543 / 79.17
OLMoE	.120 / 5.14	1.103 / 38.28	6.580 / 91.67

High ranking AUC under clipping does not guarantee calibrated thresholds: thresholds chosen on clipped shadows realize different FPRs across target families.

G.4 WEIGHT DECAY

The separate Qwen study raises AdamW weight decay from .01 to .1. A public utility screen uses seed 0; targets 1 and 2 retain 99.9749% and 99.8862% of the clean member training gain. Attacks fitted on clean checkpoints are then applied without refitting or recalibrating them for weight decay.

Table 22: Attacks fitted on clean checkpoints and evaluated after increasing Qwen weight decay from .01 to .1 on targets 1 and 2. Differences are defended minus clean AUC; brackets give paired 95% record-bootstrap intervals. RouteTrace and output attacks use all target records; LiRA and RMIA use their shared eligible records.

Attack	Target	Clean	WD .1	Change [95% CI]
ROUTETRACE-S	1	.9523	.9559	+.0035 [.0020,.0052]
	2	.9378	.9393	+.0015 [-.0006,.0035]
ZERO-BR	1	.9562	.9528	-.0034 [-.0075,.0007]
	2	.9532	.9566	+.0034 [-.0006,.0076]
LiRA	1	.9764	.9778	+.0014 [-.0002,.0038]
	2	.9820	.9818	-.0003 [-.0011,.0005]
RMIA	1	.9768	.9778	+.0010 [-.0007,.0031]
	2	.9850	.9850	+.0000 [-.0007,.0008]
Target loss	1	.5816	.5818	+.0001 [-.0004,.0006]
	2	.5949	.5946	-.0003 [-.0008,.0001]
Min-K 20%	1	.5896	.5894	-.0001 [-.0008,.0006]
	2	.6041	.6041	-.0000 [-.0008,.0007]

The changes run in both directions across methods and targets. ROUTETRACE-S increases on both targets; ZERO-BR and LiRA reverse direction, while RMIA changes by at most .0011 AUC. Target loss and Min-K% remain within .0004 of clean. In this two-target test, weight decay produces no consistent AUC reduction. This is a two-target Qwen result, not an equivalence test.

H Additional Analyses and Scope

H.1 Feature Sensitivity

The Qwen feature study retains 512 members and 10,000 non-members per target and uses the previously fitted scorers for the global and conditional feature groups. Global depth combines the global scalar and eight depth bins; Conditional combines the selected-minus-boundary scalar and eight depth bins. The separate nine-feature parameter/state study uses 2,048 non-members per target and independently fitted scorers.

Global depth carries most of RouteTrace-S's ranking accuracy. On Qwen, it reaches .9445 mean AUC, compared with .8676 for one global scalar and .9466 for the full trace. The full-minus-global-depth differences are .0003, .0008, and .0051 across targets 0–2; the paired 95% record-bootstrap interval excludes zero only on target 2. By contrast, full-minus-scalar differences range from .0638 to .0959, and all three intervals are positive.

Table 23: RouteTrace-S feature-group ROC-AUC. The upper block reports Qwen targets 0–2; the lower block reports DeepSeek-MoE and OLMoE means over targets 1 and 2. Values are point estimates.

Qwen feature group	Target 0	Target 1	Target 2	Mean
Global scalar	.8723	.8885	.8419	.8676
Global depth	.9494	.9515	.9328	.9445
Selected only	.8569	.8799	.8681	.8683
Conditional only	.8357	.8733	.8552	.8547
Global + conditional	.9479	.9521	.9364	.9455
Full 27D trace	.9497	.9523	.9378	.9466

Family	Global scalar	Global depth	Conditional	Global + cond.	Full
DeepSeek-MoE	.8905	.9195	.7860	.9224	.9292
OLMoE	.9427	.9605	.8466	.9618	.9641

The two-target DeepSeek-MoE and OLMoE comparisons show the same ordering: global depth reaches .9195 and .9605 mean AUC, while the full trace reaches .9292 and .9641.

Zero-BR does not have a monotonic component ablation. Removing the selected-absolute term C_s raises mean AUC from .9025 to .9090 on DeepSeek-MoE and from .9272 to .9311 on OLMoE over targets 1 and 2. The removal of one term can raise AUC, so these ablations do not show that every term improves ranking on its own.

H.2 Text and Difficulty Controls

The tested character 3–5-gram TF–IDF classifier uses candidate text without model outputs, parameters, or routes. One classifier is fitted on the three Qwen shadows for each data setting. Its mean AUC is .5028 on NQ and .5205 on Dolly. None of the six target confidence intervals lies entirely above .5.

Table 24: Text-only and base-difficulty controls. The upper block reports one classifier fitted on Qwen shadows per data setting; brackets give 95% record-bootstrap intervals. The lower block averages three targets after exact 1:4 matching on base-model mean NLL; M/N gives the matched member/non-member counts per target.

Data	Target 0	Target 1	Target 2	Mean
NQ	.5238 [.4977,.5502]	.4716 [.4470,.4970]	.5129 [.4873,.5374]	.5028
Dolly	.5119 [.4761,.5474]	.5308 [.4950,.5661]	.5188 [.4797,.5586]	.5205

Data	Family	M/N	|SMD| before → after	Base-NLL AUC	ROUTETRACE-S	ZERO-BR	Strongest output attack
NQ	Qwen	512/2048	.08985 → .00004	.5266	.9470	.9447	Learned .7718
	DeepSeek	512/2048	.07447 → .00015	.5236	.9280	.9085	EZ-MIA .8128
	OLMoE	512/2048	.09639 → .00009	.5275	.9654	.9372	Learned .7617
Dolly	Qwen	256/1024	.02291 → .00148	.4951	.8714	.9154	EZ-MIA .7566
	DeepSeek	256/1024	.01071 → .00137	.4994	.8579	.8542	EZ-MIA .7291
	OLMoE	256/1024	.01753 → .00064	.5018	.8517	.8663	EZ-MIA .6781

Exact matching on base-model mean NLL pairs each member with four non-members. After matching, ROUTETRACE-S reaches .8517–.9654 mean AUC and exceeds the strongest evaluated output attack in all six family–dataset settings; those attacks range from .6781 to .8128. The mean absolute base-NLL standardized difference falls below .0015 in every setting. These controls address character n-grams and scalar base-model difficulty; richer semantic differences require separate tests.

H.3 NQ-TO-DOLLY TRANSFER

Applying the NQ-fitted scorer to Dolly lowers mean AUC on Qwen and DeepSeek-MoE but raises it on OLMoE relative to a Dolly refit. For ROUTETRACE-S, transfer-minus-refit is -.0292, -.0560, and +.0162 in the three families. ROUTETRACE-S-CAL follows the same family-level directions.

Table 25: NQ-to-Dolly transfer. The upper block compares a Dolly refit with the unchanged NQ scorer; the difference is transfer minus refit. Outcomes list, in order, the number of targets favoring the Dolly refit, favoring NQ transfer, or having a paired interval that crosses zero. The lower block applies NQ thresholds unchanged and reports realized Dolly FPR / TPR in percent. AUC and FPR/TPR entries are equal-weight means over targets 0–2.

Family	Scorer	Dolly refit	NQ → Dolly	Difference	Outcomes
Qwen	S	.8717	.8425	-.0292	3/0/0
	S-Cal	.8722	.8435	-.0287	3/0/0
DeepSeek	S	.8589	.8029	-.0560	3/0/0
	S-Cal	.8584	.8023	-.0561	3/0/0
OLMoE	S	.8540	.8701	+.0162	0/2/1
	S-Cal	.8494	.8695	+.0201	0/2/1

Family	Nominal 1% NQ threshold realized Dolly FPR / TPR	Nominal 5% NQ threshold realized Dolly FPR / TPR
Qwen	19.653 / 69.66	32.535 / 86.33
DeepSeek	22.049 / 60.03	31.372 / 80.99
OLMoE	14.184 / 70.18	27.795 / 92.06

Ranking and threshold transfer diverge. Family-mean transferred AUCs range from .8023 to .8701, while nominal-1% NQ thresholds realize family-mean Dolly FPRs of 14.18%–22.05%; nominal-5% thresholds realize 27.80%–32.53%. These thresholds are applied unchanged and are not calibrated on Dolly. The comparison also changes training-set size and, for Qwen, the training objective, so it measures a combined data and training protocol transfer.

H.4 EXPOSURE CONTRAST

On the shared seed-0 candidate records, the separately trained exposure-four checkpoint has higher AUC than the exposure-two checkpoint: ROUTETRACE-S rises from .9325 to .9613 and ZERO-BR from .8873 to .9332. This analysis uses Qwen checkpoints trained by updating router parameters without query/value adapters. One scorer fitted on three exposure-one router-only shadows is applied to the evaluated checkpoints.

Table 26: Qwen exposure contrast with router-only post-training. Brackets give 95% record-bootstrap intervals for AUCs and paired differences. TPR values are percentages; ΔTPR values are percentage-point changes. The lower block compares exposure four with exposure two on the shared seed-0 records. All TPRs use retrospective target-defined FPRs rather than calibrated thresholds.

Seed	Exposure	Score	AUC [95% CI]	TPR at .1 / 1 / 5% FPR
0	2	ROUTETRACE-S	.9325 [.9235,.9410]	3.91 / 25.98 / 62.30
		ZERO-BR	.8873 [.8732,.9014]	5.47 / 25.39 / 51.95
0	4	ROUTETRACE-S	.9613 [.9548,.9674]	5.47 / 39.26 / 78.32
		ZERO-BR	.9332 [.9226,.9434]	5.66 / 35.16 / 69.34
1	2	ROUTETRACE-S	.9410 [.9332,.9482]	5.66 / 30.27 / 66.21
		ZERO-BR	.9129 [.9006,.9244]	5.08 / 28.71 / 61.91

Score	ΔAUC [95% CI]	ΔTPR .1%	ΔTPR 1%	ΔTPR 5%
ROUTETRACE-S	+.0287 [.0240,.0337]	+1.56 [-.98,5.08]	+13.28 [8.59,18.36]	+16.02 [12.30,20.12]
ZERO-BR	+.0459 [.0377,.0543]	+.20 [-2.34,5.27]	+9.77 [4.10,14.65]	+17.38 [12.89,21.48]

Both paired AUC intervals lie above zero. The two seed-0 checkpoints start from the same base and use the same records and optimizer configuration, but they were trained in separate runs. Exposure four was evaluated for seed 0. This is a paired checkpoint contrast, not an exposure curve; the TPR entries in Table 26 are retrospective target-ROC points.

H.5 SCOPE OF THE EVIDENCE

RouteTrace requires base and post-trained checkpoints with compatible router rows, a nonzero router update, and access to compute candidate gradients. The primary clean NQ and Dolly checkpoints update router matrices and query/value LoRA adapters for one epoch while keeping expert parameters fixed. The router-intervention and exposure experiments use different update rules. Record-bootstrap intervals quantify record-level uncertainty for fixed checkpoints and scorers; family means do not measure variation from new training runs or architectures.