How Advice-Routing Updates Shape Error Recognition in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Language models that can consult an advisor must decide which of their answers need help. We study how training a model to request advice changes its ability to rank incorrect answers above correct ones, which we call error recognition. Starting from models trained for confidence calibration, we isolate part of the learned routing update and intervene on the activity it produces. In Qwen, a distributed subset of the update improves error ranking when all model variants judge the same proposed answers. Replacing this activity throughout the prompt with fitted constants preserves roughly half of the partial router's improvement over its starting model. Restoring activity in a later group of selected layers recovers much of the lost performance, and activity from the same question gives greater recovery than activity from another question. This pattern repeats across two factual question-answering datasets and a second routing seed. The results reveal a substantial gain under constant internal additions and a smaller, reproducible benefit from question-specific restoration. They do not establish a dedicated correctness representation or a consistent advantage in advisor decisions.

1 Introduction

A language model in an advisor cascade must decide when to consult another model. Each extra call consumes compute and adds latency. Useful consultation also depends on whether the advisor can repair an error or might replace a correct answer with a wrong one. Calibrated error estimates can support this decision, but the model must use them in a way that accounts for the cost and reliability of advice.

Language models can assess their own answers (Kadavath et al., 2022), and confidence interventions can influence abstention (Kumaran et al., 2026). Yet assessing an answer and acting on that assessment are different abilities. A direction that predicts correctness need not drive the decision to ask for advice. Making a model less confident or more willing to ask can leave its ability to distinguish mistakes unchanged.

We study models trained to keep a proposed answer or request advice, which we call *routers*. We ask how the weight changes learned during routing training help the advice score rank wrong answers above correct ones. Every variant judges the same proposed answers, isolating this *error recognition* comparison from differences in generated answers. Starting from calibration-trained models, we add selected routing weights, replace the activity they produce, and restore parts of that activity to test their contribution.

We find three results in Qwen. A distributed subset of routing weights improves error ranking across three routing seeds. Roughly half of its gain survives replacement of its activity with constants throughout the prompt. Restoring activity in layers 15 to 17 recovers much of the loss, with a smaller, reproducible advantage for activity from the same question over another question. The full router remains stronger, transfer to Llama is mixed, and the evidence does not identify a single correctness feature. Separate tests show that the ranking gain neither guarantees unchanged answers nor improves decisions over a tuned full router.

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.

2 EXPERIMENTAL SETTING

2.1 MEASURING ERROR RECOGNITION WITHOUT CHANGING THE ANSWER

We begin with a decision that all model variants can make about the same answer. The starting model, trained for confidence calibration, first answers each question. We save this answer and present the same question-answer pair to every variant. The router is prompted to produce A for requesting advice or K for keeping the answer. Its *advice score* is the difference between the two token logits

$$z(q, a) = \ell_A(q, a) - \ell_K(q, a). \tag{1}$$

A larger value indicates a stronger preference for advice. We separately generate the action and require it to be exactly A or K after removing whitespace. A useful score and a valid generated action are therefore evaluated separately.

Fixing the answers keeps their correctness labels identical across variants. Ranking changes therefore reflect the routing computation rather than a different set of answers. We test answer generation separately because this design does not establish answer preservation.

Error AUROC is the probability that a randomly selected wrong answer receives a higher advice score than a randomly selected correct answer, with half credit for ties. Values of 0.5 and 1 indicate chance and perfect ordering. A uniform score shift cannot improve AUROC, so better ranking requires more than a general increase in willingness to ask. AUROC does not measure probability calibration.

2.2 CONFIDENCE CALIBRATION BEFORE ROUTING

Before routing training, each backbone is adapted to answer questions and report confidence $c \in [0, 1]$. The retained implementation adapts the correctness-plus-Brier reward of Damani et al. (2025)

$$r_g = o_g + 1 - (c_g - o_g)^2 - 0.1 m_g, \tag{2}$$

where $o_g \in \{0, 1\}$ is automated answer correctness and m_g indicates missing confidence. If confidence is missing, $c_g = 0.5$. The correctness term rewards accurate answers; the Brier term penalizes disagreement between stated confidence and correctness.

Each training step samples eight responses to one question. Rewards are converted to group-relative advantages

$$\widehat{A}_g = \frac{r_g - \overline{r}}{\mathrm{std}(r_1, \ldots, r_8) + 10^{-4}}. \tag{3}$$

Here std is the sample standard deviation. The policy-gradient loss applies the advantage to every generated token, including the answer and confidence, with a KL penalty of weight 0.04 to the adapter-disabled backbone. The documented recipe uses rank-16 LoRA, alpha 32 and AdamW with learning rate 10^{-5}. Its training pool contains 4,619 PopQA and 2,652 SimpleQA questions; these are pool sizes, not counts of items seen. We merge the calibration adapter into its backbone before routing training. Appendix A gives the per-token loss, stopping records and reconstruction limits. AUROC is an evaluation metric, not part of the reward; this study does not isolate the reward's causal effect.

2.3 MODEL STATES AND TRAINING REPLICATION

The study compares three versions of each model, shown in Figure 1. The *starting model* has completed confidence-calibration training. The *full router* adds the complete weight update learned during routing training. The *partial router* adds only the selected components of that update, with all other routing components omitted. Selecting these components involves no further training of the partial router.

We use Qwen3-4B-Instruct-2507 and Meta-Llama-3.1-8B-Instruct (Qwen Team, 2025; Grattafiori et al., 2024). Each family includes an original router and two newly trained routing seeds. The new runs use the same 497 eligible training pairs for two epochs, with loss on the final A/K action token. Each family shares a calibration-trained starting checkpoint. Replication therefore tests sensitivity to routing training, conditional on that checkpoint, rather than repeating calibration training. The focused activity study evaluates the original Qwen router and one new routing seed.

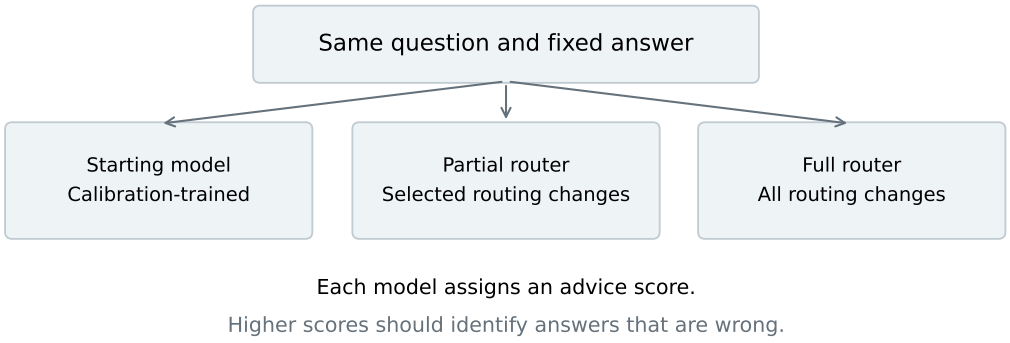


Figure 1. All variants score the same question and answer, generated once by the starting model. The partial router adds selected routing components; the full router adds the complete update. Separate experiments allow interventions to change answer generation.

2.4 Data, labels and separation of experimental stages

We evaluate on PopQA and TriviaQA (Mallen et al., 2023; Joshi et al., 2017), two factual question-answering datasets. The weight study contains 376 eligible PopQA and 363 TriviaQA answers for Qwen, and 382 per dataset for Llama. The later activity study uses a separate set of 310 PopQA and 309 TriviaQA answers, shared by its two Qwen routers. Questions previously recorded in routing training and evaluation are excluded by dataset identifier and normalized text. Neither PopQA confirmation set overlaps the retained calibration pool by dataset identifier or normalized question text (Appendix A). These checks concern recorded training pools and experiments, not model pretraining. Starting-model accuracy is low on PopQA and higher on TriviaQA: 44 of 310 and 172 of 309 focused-study candidates are labeled correct.

All reported quantitative results use automated correctness labels. The primary rule matches extracted answers against benchmark references and aliases. A reference-only rule removes aliases but retains the same extraction and matching procedure. Agreement between these rules is a sensitivity result, not independent semantic verification. Empty answers and answers that reach the generation cap are excluded from fixed-answer ranking under a rule set in advance. They remain in the archive and in the separate answer-generation analyses.

3 Which routing weights contribute to error recognition?

3.1 Selecting components of the learned update

Our first experiment asks whether a subset of the learned routing weights can improve error recognition on its own. Routing training uses low-rank adapters (Hu et al., 2022), which add learned updates to projections inside the model. We decompose the updates into singular components and construct partial routers by retaining specified components. Addition to the starting model tests whether those components help in that setting. Removal from the full router tests whether they are necessary when the other routing changes are present. The two comparisons can therefore give different results.

An earlier coarse screen identifies the second quarter of model depth as the region to examine. Within it, the exploratory search tests individual modules and layers on 185 eligible discovery answers. A candidate must improve AUROC by more than 0.01 when added and reduce it by more than 0.01 when removed, on both datasets. No individual module or layer passes this screen. A distributed Qwen candidate does pass. It retains the two leading singular components in each of 45 projections across layers 9 through 17, using zero-based indices. For projection j with routing update $\Delta W_j = U_j \Sigma_j V_j^\top$, we keep $\Delta W_j^{(2)} = \sum_{k=1}^{2} \sigma_{jk} u_{jk} v_{jk}^\top$. The projections are the query, value and output attention maps and the up and down feedforward maps.

The two components are defined separately in each projection. They do not form one shared two-dimensional representation of correctness. The selection rule favors fewer modules and components

among candidates that pass. It was written after collection of the distributed diagnostics had begun, making this stage exploratory. The target was fixed before held-out weight testing. For new routing seeds, the same structural rule selects components from that router's own weights.

3.2 Held-out addition, removal and control results

The partial Qwen router improves error AUROC over the starting model in all six dataset-by-seed comparisons. Gains range from 0.10 to 0.12 on PopQA and from 0.045 to 0.054 on TriviaQA. The corresponding family-adjusted intervals are positive under both automated label rules. Figure 2 shows these improvements alongside the full router, which remains better in every panel.

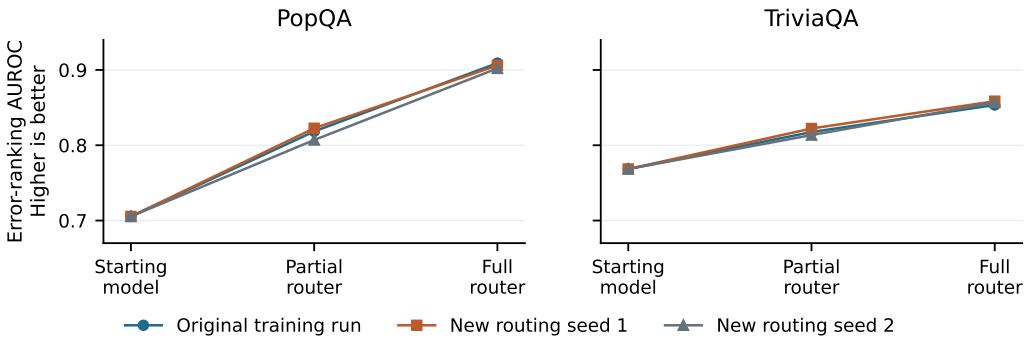


Figure 2. Selected weights recover part of the ranking improvement from routing training. Each line is a separately trained Qwen routing update applied to the same starting model. Points are descriptive AUROCs on the fixed weight-study answers. The full router remains better than the partial router. Appendix C reports adjusted uncertainty for the paired comparisons.

For original Qwen on PopQA, AUROC rises from 0.705 for the starting model to 0.818 for the partial router and 0.909 for the full router. Retaining the complete selected quarter gives 0.828, so the two-component selection captures much of the local effect. Across the six Qwen panels, the partial router recovers 51 to 60 percent of the full router's gain in the point estimates.

Removing the selected components from the full router gives weaker evidence. All six Qwen removal estimates indicate a loss, but only one has a positive family-adjusted interval. Other components may compensate, although this experiment does not directly identify compensation. The evidence establishes a contribution when the selected update is added to the starting model, without establishing necessity in the full router.

We also compare with learned components from another depth quarter and random rotations that preserve the selected singular values. The partial Qwen router exceeds the learned control in all six adjusted comparisons, and exceeds the eight original random controls under ordinary secondary intervals. A scalar score shift leaves rankings essentially unchanged. The improvement is therefore not reproduced by these alternative updates or by a uniform advice bias. However, equal weight norms do not give equal activation magnitudes. We return to this issue with stronger controls in the activity study.

Applying the same structural rule to Llama gives a less consistent result. The discovery screen fails, and held-out addition evidence depends on the dataset and correctness rule. We retain this outcome without choosing a new target from confirmation results. Full-router gains over the Llama starting model range from 0.009 to 0.020 AUROC in five of the six panels. Only the second new routing seed on TriviaQA has a substantially larger gain, at 0.059 (Table 2), whereas Qwen's full routers gain 0.085 to 0.204. We therefore focus the next experiment on the reproducible Qwen gain and ask how the selected weights produce it.

4 How does the selected update contribute during prompt processing?

To test whether the selected update needs input-dependent outputs, we replace its activity with fitted constants during prompt processing, then restore subsets of the recorded activity.

4.1 Replacing activity at different prompt positions

For each selected projection, an incoming hidden state h is mapped to two coefficients before being mapped back into the model's hidden space. In adapter notation, the selected contribution is $sBAh$. We call the two coefficients Ah the *selected activity*. Replacing them changes what the learned component contributes on a particular input while leaving its weights fixed.

Each module receives a mean fitted on separate questions for the dataset and prompt window, without correctness labels. Questions receive equal weight, averaging positions within each window. Replacement removes variation across both questions and positions, so it broadly disrupts the selected activity rather than selectively deleting an identified correctness variable.

Replacing activity only at the final prompt token changes Qwen AUROC by less than 0.001, with all ordinary intervals including zero. Earlier question and answer computations remain available, so the decision could still rely on activity computed earlier in the prompt.

On new discovery questions, we compare the final token, proposed-answer span and entire prompt. This initial prompt processing is called *prefill*. Only full-prefill replacement causes a loss greater than 0.01 AUROC on both datasets; answer-span replacement passes on PopQA alone. We select full-prefill replacement before examining confirmation outcomes.

Replacement stops after the prompt forward pass, but its effects can persist through the key/value cache. Checks confirm cache changes and changed continuation even with the next token externally fixed. This verifies intervention reach, without identifying the information affected.

4.2 Restoring activity and testing the source of recovery

We next ask whether restoring a smaller part recovers error recognition. We record selected coefficients during a normal run on each question, then repeat the run under mean replacement, restoring recorded activity in one layer group. All other selected projections remain replaced. Figure 3 shows the intervention and controls.

Discovery compares the three groups covering layers 9 through 11, 12 through 14 and 15 through 17. The last group, containing 15 selected projections, gives the largest minimum recovery across datasets and passes the recovery gate of 0.005 AUROC on each. We freeze this choice for confirmation and for the second routing seed.

Two controls test the source and location of recovery. The *other-question control* supplies activity from a fitting question in the same dataset, matched by prompt length without correctness labels. The *other-layer control* restores the current question's activity in layers 9 through 11. These test whether comparable generic activity suffices and whether restoration location matters. Restoring all original activity must reproduce intact outputs within numerical tolerance, but serves only as an implementation check.

4.3 Controls for intervention magnitude

Controls matched in weight norm can differ greatly in activation strength. A weak effect could therefore reflect perturbation magnitude rather than direction. On separate fitting questions, we rescale the learned control and four fixed random rotations to match each selected module's typical output magnitude. Scales use no correctness labels and remain fixed for confirmation.

We measure branch outputs and the changes caused by mean replacement separately. Matching one magnitude need not match the other. Neither captures the full propagated effect on the residual stream, so equal local norms need not imply equal functional disruption.

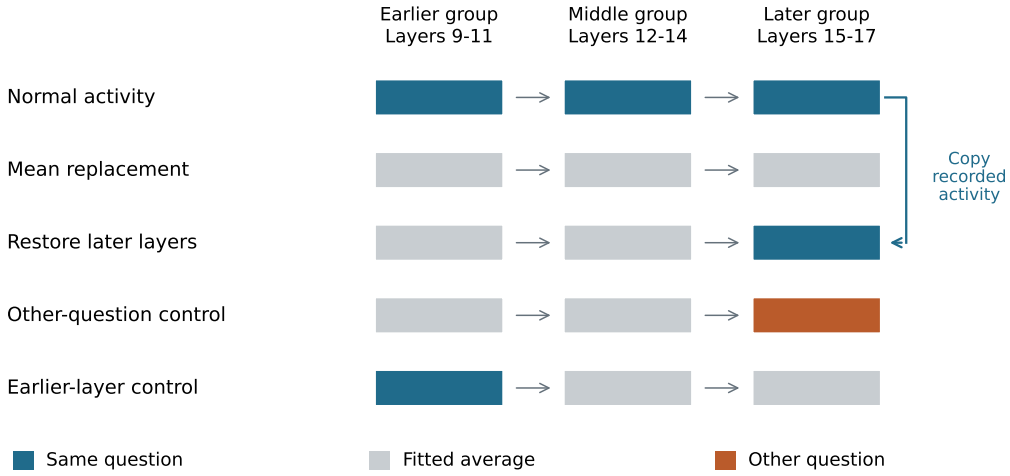


Figure 3. Disruption and restoration inside the partial router. Mean replacement affects all selected projections throughout the prompt. Restoration copies recorded normal-run activity back into layers 15 through 17. Controls change the source question or restore an earlier layer group. Each box denotes the selected activity in that group, not the entire layer output. Restoration requires an intact run on the same question and is an experimental diagnostic.

4.4 Confirmation on fresh questions and another router

For each dataset, we allocate 64 questions to fitting means and scales, 96 to discovery, and 320 to confirmation. The eligibility rule leaves 310 PopQA and 309 TriviaQA answers. Both routers judge identical answers. The second router uses the frozen intervention and layer group, fitting its own means and scales by the same algorithm.

Five comparisons per router and dataset measure replacement loss, restoration gain, gain beyond each restoration control, and selected versus learned-control replacement loss. The resulting 20 primary comparisons use 50,000 paired question-bootstrap resamples with nominal Bonferroni-adjusted percentile intervals. Resampling questions together across conditions preserves the paired design. The weight study uses a separate family of 36 comparisons. Intervals are conditional on the selected components and checkpoints, and the numerical audit does not establish their finite-sample coverage.

4.5 Disruption reduces ranking and partial restoration recovers it

The fresh results repeat the disruption and recovery pattern in both datasets and both Qwen routers. Table 1 places each intervention beside the starting model and intact partial router. Full-prefill replacement reduces AUROC by 0.049 to 0.054 on PopQA and by approximately 0.033 on TriviaQA. Restoring the later group recovers 66 to 72 percent of the loss in the point estimates. Recovery is substantial but incomplete.

Routing run	Dataset	Starting	Intact	Replaced	Restored	Other question	Other layers
Original	PopQA	0.740	0.854	0.805	0.838	0.824	0.810
New seed	PopQA	0.740	0.862	0.808	0.845	0.833	0.813
Original	TriviaQA	0.773	0.844	0.812	0.833	0.820	0.818
New seed	TriviaQA	0.773	0.848	0.815	0.839	0.828	0.822

Table 1. Error AUROC for the starting model and the partial router under each activity intervention, using the same fixed answers. Restoration uses layers 15 through 17; the two final columns change the source question or restore layers 9 through 11. These are descriptive point estimates. Figure 4 gives uncertainty for the planned differences.

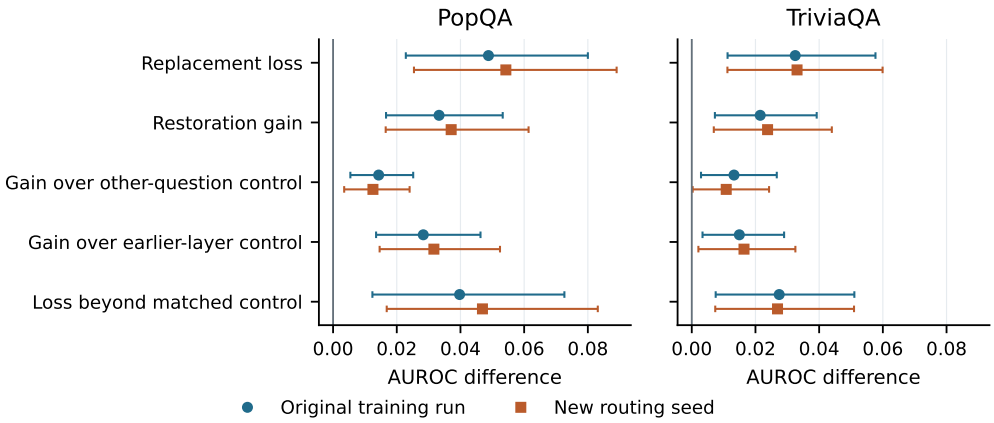


Figure 4. Uncertainty for all 20 planned activity comparisons. Positive values favor the effect named at left. Bars are paired bootstrap intervals adjusted for nominal 95 percent coverage across the 20 comparisons. Both partial routers use the same 310 PopQA and 309 TriviaQA answers with automated labels. The final row compares the selected and activation-scaled learned-control losses, subject to the floor-effect caveat in Section 4.5.

For original Qwen on PopQA, restoration raises AUROC from 0.805 to 0.838, compared with 0.824 for another question and 0.810 for earlier layers. Matching activity to the current question helps more, while the replaced model remains better than the starting model. During this prompt pass, the replaced partial router is the starting model with fixed bias vectors added to 45 projections. These input-independent additions retain 35 to 42 percent of the full router's gain in the point estimates (Table 8).

All 20 primary comparisons have positive nominal family-adjusted intervals under the canonical labels. The reference-only sensitivity analysis yields the same pattern. Figure 4 shows every comparison rather than only the largest effects. One is close to the boundary. For the new router on TriviaQA, the adjusted lower endpoint of restoration over the other-question control is 0.00028. This is weaker evidence for question specificity than the other panels provide.

The magnitude controls also constrain the explanation. On confirmation questions, the average branch-output ratios for scaled controls lie between 0.987 and 1.035 relative to the selected update. All four random controls have lower ranking performance than the intact partial router under ordinary secondary intervals. Replacing the learned control's activity produces a local change 13 to 19 percent larger than the selected replacement, yet a smaller ranking loss in every panel. Thus the selected loss is not explained by a larger measured local perturbation. The control begins with little gain over the starting model, however, creating a floor-effect concern that limits direction-specific conclusions from this loss contrast.

These experiments connect the selected weights to a recoverable contribution during prompt processing. They do not identify the restored information as correctness alone. Question content, answer information or positional structure could also support ranking. Nor does better ranking establish that restoration selectively raises advice scores for wrong answers. We find no reliable difference between wrong and correct answers in their average score changes. Appendix G reports all conditions, action validity, logit analyses and intervention checks.

5 Answer Generation and the Value of Advice

5.1 The Fixed-Answer Result Does Not Establish Answer Preservation

We separately generate answers with the starting model, partial router and partial router under full-prefill replacement, using 64 questions per dataset for each Qwen router. Replacement means are fitted on answer-generation prompts, which contain no proposed answer.

Relative to the starting model, the partial routers change between 19 and 34 extracted answer strings out of 64. Automated accuracy changes range from -4.69 to $+1.56$ percentage points. The uncertainty does not establish preservation within the prespecified two-point margin. Even a panel with no observed correctness changes has conservative bounds of approximately ± 6.62 points. These results distinguish an inconclusive accuracy comparison from evidence of equivalence. String changes also require semantic review before they can be interpreted as changes in meaning. All contrasts, including interventions applied only during prompt processing or only during answer generation, appear in Appendices D and G.4.

5.2 Better ranking is not sufficient for better advisor decisions

The remaining practical question is whether error ranking translates into useful advice. We use the other model family as a fallible advisor. A request replaces the student's answer, counting both repairs and harmful substitutions. Utility is final accuracy minus average call cost. For each cost, each model's threshold is chosen on separate discovery data, with always-call and never-call policies available. The full router receives the same optimization as the partial router. In a cascade, advice is useful when its expected accuracy gain exceeds its cost. Calibrated student confidence alone is insufficient without accounting for the advisor's reliability on the same cases.

Across 30 Qwen combinations of dataset, seed and cost, the partial-router threshold policy ties the full-router threshold policy in 20 cases. Only one difference has a positive ordinary interval, with a utility gain of 0.00266. All 30 Llama comparisons tie because both policies choose never to call the weaker advisor. The complete results are retained in Appendix E. They do not establish a consistent practical advantage for the selected update.

AUROC assesses all wrong-correct pairs, whereas utility depends on the decision boundary, call cost and the advisor's ability to repair particular errors. Better ranking relative to the starting model therefore need not beat a tuned full router.

Qwen usually generates the required A or K. Llama often produces additional text, which remains invalid under the strict action rule. Constrained decoding would define a different policy requiring its own utility test.

6 Related work

Correctness information and its use. Marks & Tegmark (2024) identify directions predictive of truth and test their causal effects on outputs. Orgad et al. (2025) find information about correctness that is not fully reflected in generated answers, with important limits to transfer across datasets. Ferrando et al. (2025) identify entity-recognition features that affect refusal and hallucination. These findings motivate separating available information from its use and considering question familiarity as an alternative to answer-specific error recognition. Our fixed-answer comparisons do not eliminate that alternative.

Confidence and correctness assessment. Proper scoring rules evaluate probability reports (Gneiting & Raftery, 2007), and Brier-based rewards can improve reported calibration (Damani et al., 2025). Confidence failures also arise during tool use (Xuan et al., 2026). Our study takes calibration-trained checkpoints as its starting point and tests the subsequent routing update. It does not isolate the effect of the calibration reward.

Mechanistic interventions on uncertainty. Miao & Ungar (2026) connect internal correctness information to verbal confidence through probing and steering. Zhao et al. (2026) identify components that inflate verbal confidence and intervene to improve calibration. Kumaran et al. (2026) show that confidence interventions affect abstention and study their interaction with decision thresholds. We examine a complementary object, the weight update learned during advisor routing, and test its activity through replacement and restoration. Our interpretation follows the emphasis on intervention design and controls in activation-patching work (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024).

Deferral and advisor use. Learning to defer accounts for expert behavior (Mozannar & Sontag, 2020), while agentic correction can use a learned low-rank critic (Vijayvargiya & Lokesh, 2026). Our advisor test compares utility against the thresholded full router.

7 Discussion

Across both datasets and routers, 54 to 57 percent (each 95% interval includes 50%; Appendix G.5) of the selected update's gain remains when its branches output constants. Donor activity recovers 38 to 66 percent of same-question recovery, leaving a same-question advantage of 0.011 to 0.014 AUROC. These ratios show substantial gain under constant internal additions and a smaller prompt-specific restoration benefit. They do not partition independent mechanisms. Constants can alter input-dependent computation elsewhere, so steering existing correctness information remains a hypothesis.

Three abilities should therefore be evaluated separately. A model can report calibrated confidence, rank its errors, and decide when advice is worth its cost. The calibration objective encourages the first, our interventions test the second, and the fallible-advisor evaluation tests the third. Routing updates improve ranking without consistently improving utility over the full-router threshold baseline. This is evidence against treating a better correctness readout as sufficient for better action. It does not establish a separate mechanism that turns confidence into action or show that calibration training itself caused the routing effect.

8 Limitations and future work

The localization is distributed across 45 selected projections and 15 restored projections. Addition establishes a contribution in the partial router, while removal gives weaker evidence of necessity in the full router. Mean replacement removes positional as well as example-dependent variation, and restoration could recover question familiarity or answer content rather than correctness alone. The learned control's small initial gain limits its loss comparison. Restoration also requires an intact run on the same question, making it a diagnostic rather than a standalone inference method.

Both datasets concern factual QA, the focused activity result is confirmed only in Qwen, and routing replications share one calibrated starting model per family. Automated correctness labels can miss semantic errors; agreement between two matching rules does not resolve this limitation. Native-answer uncertainty does not establish accuracy preservation. The conclusions remain conditional on the retained checkpoints, whose ancestry and replay limits are detailed in Appendix A.

The most immediate step is independent human adjudication of the 619 focused-study candidates, followed by re-estimating the primary comparisons under those labels. The next mechanism tests should clamp the selected activity inside the full router, swap correct and incorrect answers to the same question, and compare frozen correctness probes. These test necessity, answer sensitivity and whether information remains available after intervention. A stronger generalization test would repeat the decisive result across independently calibrated parents and another model family. For practical advisor control, interventions should be evaluated with complementary, fallible advisors at matched call budgets, reporting errors repaired, correct answers harmed and costs alongside the full-router threshold baseline.

9 Conclusion

In calibration-trained Qwen models, selected routing weights recover a substantial part of the full router's improvement in error recognition. Roughly half of their gain survives constant internal outputs. Restoring later activity recovers much of the loss, with greater benefit from the original question than from a donor. These effects repeat across routing seeds and factual-QA datasets, connecting learned weights to recoverable computation during prompt processing. They do not establish a dedicated correctness representation or better advisor utility. Useful advisor control therefore requires separate tests of confidence calibration, error recognition and decision quality.

AI USE STATEMENT

AI tools assisted with experimental coding, analysis, literature review and manuscript editing, including spelling and grammar corrections. The authors have reviewed all AI-assisted work and take responsibility for the final text, claims and artifacts.

ETHICS STATEMENT

The study uses existing factual-QA benchmarks and model-generated answers. Invalid requests, harmful advisor substitutions and answer changes are retained. The results do not establish suitability for deployment as reliable decision aids. Distribution and use of the data and model artifacts remain subject to their original terms.

REPRODUCIBILITY STATEMENT

The anonymous supplementary ZIP provides executed intervention code, raw focused-study observations, frozen score matrices, CPU reproduction tools, main-figure scripts and a calibration-pool overlap audit. The focused primary and secondary analyses reproduce the recorded results from a clean extraction. The appendices specify model states, training, labels and interventions. The ZIP also includes earlier cross-family summary results and reference code. Complete GPU reproduction requires the exact external adapters; model weights, fitted tensors and the complete earlier raw archives are not included.

REFERENCES

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty, 2025. URL https://arxiv.org/abs/2507.16806v2. Version 2, revised May 15, 2026.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.14257v2.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behavior, 2026. URL https://arxiv.org/abs/2603.22161v3. Version 3, revised September 18, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Miranda Muqing Miao and Lyle Ungar. Closing the confidence-faithfulness gap in large language models, 2026. URL https://arxiv.org/abs/2603.25052v2. Version 2, revised April 1, 2026.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Qwen Team. Qwen3-4B-Instruct-2507 model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507. Model card; accessed September 26, 2026.

Sanidhya Vijayvargiya and Rahul Lokesh. Actionable hallucination detection: Translating latent uncertainty into agentic critique, 2026. URL https://arxiv.org/abs/2608.10430v1. Version 1, August 11, 2026.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Yunze Xiao, Junjue Wang, and Naoto Yokoya. The confidence dichotomy: Analyzing and mitigating miscalibration in tool-use agents, 2026. URL https://arxiv.org/abs/2601.07264v1. Version 1, January 12, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Tianyi Zhao, Yinhan He, Wendy Zheng, Yujie Zhang, and Chen Chen. Wired for overconfidence: A mechanistic perspective on inflated verbalized confidence in LLMs, 2026. URL https://arxiv.org/abs/2604.01457v3. Version 3, revised July 27, 2026.

A IMPLEMENTATION AND MEASUREMENT DETAILS

How to read the detailed results. The main text uses descriptive model names. The archived tables use shorter labels, retained here to keep them aligned with the experimental records. C is the starting model, S is the partial router, and R is the full router. H is the original trained router. S1 and S2 denote the two new routing seeds. In dataset columns, P is PopQA and T is TriviaQA. Removing the selected update from the full router is written $R \setminus S$. A selected coefficient is one of the internal activity values explained in Section 4.

The appendices first describe the weight study and its answer-generation and advisor tests. Appendix G then gives the separate activity study, including its new fitting, discovery and confirmation questions. Results from these two studies should not be pooled as though they used the same question set.

Pinned backbones and reconstruction. The exact model sources and revisions are given below.

Qwen	`Qwen/Qwen3-4B-Instruct-2507`
Revision	`cdbee75f17c01a7cc42f958dc650907174af0554`
Llama	`NousResearch/Meta-Llama-3.1-8B-Instruct`
Revision	`d10aef7999a2b5ba950ab3974312feeedbfe0b77`

The retained calibration adapter is merged in FP32, parameters are converted to BF16 while FP32 buffers are retained, and archived configuration/tokenizer metadata are restored. The resulting parent-file hashes match the recorded expected files. Numerical reconstruction and supplied output replay do not independently establish the original training ancestry. Llama's supplied replay sequences contain eight tokens and do not establish equivalence with the original model for longer generations.

Calibration training and retained evidence. The retained calibration code rewards an answer and its stated confidence with $r = o + 1 - (c - o)^2$, where o is automated correctness and c is confidence in $[0, 1]$, adapting the correctness-plus-Brier formulation of Damani et al. (2025). Our implementation adds a constant of one, which cancels under the within-group normalization in Equation 3. Missing confidence uses $c = 0.5$ and incurs a 0.1 penalty. A single-turn policy-gradient loop generates eight responses per question, normalizes rewards within the group, and applies the resulting advantage to all generated tokens, with a KL penalty to the adapter-disabled backbone. The saved adapters have LoRA rank 16 and alpha 32 on the query, key, value, attention-output, gate, up and down projections. The documented factual training pool contains 7,271 questions, comprising 4,619 PopQA and 2,652 SimpleQA items (Wei et al., 2024). These are pool sizes, not counts of examples seen during training.

The retained recipe specifies 500 steps, AdamW learning rate 10^{-5} and KL coefficient 0.04. The Llama log records eight rollouts per step, early stopping at zero-based step 260 and 2,088 generated responses. Its saved adapter and the retained Qwen calibration adapter match the hashes used for reconstruction. For Qwen, the available notes and code describe the recipe, but we do not have a complete execution log linking all optimizer settings and sampled questions to the saved adapter. Neither calibration run has a complete hash-linked record of the original backbone, optimizer state and data order. These details describe the retained evidence rather than certify the entire training history. The study conditions on these checkpoints and does not estimate the causal effect of calibration training itself.

For completeness, the retained code computes the group advantage as

$$\widehat{A}_g = \frac{r_g - \overline{r}}{\mathrm{std}(r_1, \ldots, r_G) + 10^{-4}}, \qquad G = 8, \tag{4}$$

using the sample standard deviation. Let ℓ_{gt} and ℓ_{gt}^{ref} be the policy and adapter-disabled reference log probabilities of generated token t, and let $d_{gt} = \ell_{gt}^{\mathrm{ref}} - \ell_{gt}$. The per-group loss is

$$\mathcal{L} = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{T_g} \sum_{t=1}^{T_g} \left[-\widehat{A}_g \ell_{gt} + \beta \big(e^{d_{gt}} - d_{gt} - 1 \big) \right], \qquad \beta = 0.04. \tag{5}$$

The confidence-bearing response and its answer tokens both receive this sequence-level advantage. This is the retained single-turn group-normalized policy-gradient recipe, not a claim that all historical execution details have been recovered. AUROC is an evaluation metric here, not an additional term in this calibration reward.

Calibration-pool overlap audit. We compare both retained 7,271-question calibration pools with the complete confirmation allocations and eligible evaluation subsets. Calibration PopQA IDs use the prefix `pq-`, whereas evaluation IDs use `popqa-`; we compare the underlying numeric identifiers. A separate exact-text check applies Unicode NFKC normalization, case folding and removal of non-word characters. There is no PopQA ID overlap in either the 384-question weight allocation or the 320-question activity allocation, and no normalized question-text overlap with either entire calibration pool. The same zero-overlap result holds for the eligible weight subsets of 376 Qwen and 382 Llama PopQA questions and the 310 activity-study PopQA questions. The TriviaQA allocations and eligible subsets also have no normalized text matches to either pool. The supplementary audit includes the input question records, source hashes, eligibility checks and a standalone script. These checks address the retained pools, not pretraining, paraphrases or unrecorded training exposures.

New routing seeds. Both families use the same 497 eligible original candidate/label training pairs, two epochs, and seeds 2026091301 and 2026091302. LoRA rank is 8, alpha is 16, dropout is zero, and target projections are q, v, o, up and down. AdamW uses learning rate 2×10^{-5} and zero weight decay. Per-example forwards accumulate in groups of eight; the last partial group is normalized by its actual size. Gradient norm is clipped at 1. Each run has 126 optimizer steps and 994 example presentations. Only the final A/K target token receives loss; no end-of-turn token is supervised. This is a documented training condition, not an isolated causal explanation of Llama's format failures. Actual optimizer, adapter and RNG states are saved at step boundaries. Runtime pins include PyTorch 2.11.0, Transformers 4.57.1 and PEFT 0.17.1. These are conditional router replications, not repetitions of calibration training.

Automated labels and extraction. The frozen evaluator removes confidence/reasoning annotations and extracts the final answer field or last content-bearing line. Canonical factual labels use the recorded reference and aliases, with normalized whole-phrase matching and a fuzzy stage that includes sequence similarity and multiword token containment. Reference-only sensitivity removes aliases but retains the same extraction and matching procedure. Thus reference-only labels are not an independent semantic gold standard. Short/numeric references have separate matching rules. The optional external model-judge fallback was explicitly disabled for this study. Native outputs, including empty and truncated answers, remain in the archived panels. Fixed-routing eligibility requires a nonempty candidate that did not hit the generation cap. Confidence and action readouts have an eight-token cap; calibrated candidate generation has a 256-token cap.

Update components and clamping. For a selected adapter module j, write the effective routing update as

$$\Delta W_j = s_j B_j A_j = U_j \Sigma_j V_j^\top, \qquad \Delta W_j^{(2)} = \sum_{k=1}^{2} \sigma_{jk} u_{jk} v_{jk}^\top. \tag{6}$$

The target comprises these leading components in every module of the selected quarter. It is not one shared rank-two representation. The vectors are derived independently from each router's weights. Matched controls and random controls are parameter-norm matched; measured branch-output magnitudes need not match. In the coefficient test, the low-rank coefficients at the final prompt token are replaced by dataset/prompt-specific discovery means. Earlier prompt tokens and later computation remain available. The measured norm statistic is the root-sum-square of module-output norms, not the norm of a single combined residual-stream perturbation.

Ranking and calibration. Let $y_i \in \{0, 1\}$ indicate candidate correctness, p_i be reported confidence, and z_i the request-minus-keep logit. Brier score is $n^{-1} \sum_i (p_i - y_i)^2$. Error AUROC is the fraction of incorrect/correct candidate pairs with $z_{\text{incorrect}} > z_{\text{correct}}$, plus half-credit for ties. The 36 canonical primary contrasts use 50,000 paired question resamples and percentile tails $0.05/(2{\cdot}36)$

and $1 - 0.05/(2 \cdot 36)$. Confidence, random and clamp analyses are secondary; their ordinary intervals are not a multiplicity-controlled primary family. The figures show recorded intervals rather than refitting them. Finite-sample percentile coverage is not established by the point-estimate audit.

Policy utility and threshold ties. For advisor correctness h_i, call decision c_i and valid-action indicator v_i, the frozen per-question utility at cost λ is

$$u_i(\lambda) = v_i\big[(1 - c_i)y_i + c_i h_i - \lambda c_i\big]. \tag{7}$$

Invalid generated actions receive zero utility; threshold policies have $v_i = 1$. The per-call penalty λ is a utility parameter, not a measured latency or monetary cost. For a valid action, consultation is preferable when $\mathbb{E}[h_i - y_i \mid q_i, a_i] > \lambda$, so student confidence alone does not specify the best cascade policy. Advisor answers replace student answers without an oracle maximum, retaining harmful substitutions. Each score threshold is fitted separately at costs 0.10, 0.25, 0.50, 0.75 and 0.90 using the family's utility-discovery candidates (185 for Qwen, 191 for Llama). The strict rule is $c_i = \mathbb{1}[z_i > t]$; always/never-call endpoints are included. Exact decimal-cost arithmetic breaks utility ties by fewer calls, then the largest threshold. Selection never consults confirmation labels.

Native bounds. For each native contrast, let g and ℓ be observed correctness gains and losses. Separate two-sided Clopper-Pearson intervals use tail $0.05/4$ for each probability. Subtracting their endpoints gives a conservative simultaneous gain/loss bound for that contrast. This does not adjust across all native contrasts. Equivalence requires the entire bound to fall strictly inside $[-0.02, 0.02]$. No contrast meets that criterion.

B Control construction and discovery outcomes

Anatomy and matched updates. The selected modules are the query, value and output attention projections and the up and down MLP projections. Qwen has 36 decoder layers and Llama has 32. Their selected quarters contain zero-based layers 9 through 17 and 8 through 15, respectively. The matched controls use corresponding modules in Qwen layers 27 through 35 and Llama layers 24 through 31. Each matched module uses its own leading singular components, with its effective update rescaled to the selected module's Frobenius norm. Random controls rotate the selected singular vectors while retaining the selected singular values. Norm equality is checked per module within relative tolerance 10^{-5}.

Discovery sequence. The single-module and single-layer screens require more than 0.01 AUROC gain for both addition and removal on each dataset. No individual unit passes. Among Qwen distributed conditions, the complete quarter and leading two or four components per module pass. The complexity-first rule selects two. On discovery, Qwen leading-two addition gains are 0.13750 and 0.08333 on PopQA and TriviaQA, with removal gains 0.01989 and 0.06002. The corresponding Llama addition gains are -0.01080 and $+0.02506$, and removal gains are $+0.01023$ and $+0.02972$. These are selection-panel point estimates rather than held-out confirmation. An FP32 head-projection diagnostic does not replace the natural model-head selection metric or rescue the Llama discovery failure.

Full-quarter, scalar and confidence outcomes. The full-quarter and scalar controls help interpret the selected rank. For original Qwen, retaining the complete local quarter gives AUROC 0.82800 on PopQA and 0.82432 on TriviaQA, slightly above the leading-two target. The target captures much, but not all, of the local effect. A scalar bias leaves the Llama AUROCs unchanged and changes the recorded Qwen values by at most 0.000043. This is consistent with the invariance of ranking to a constant score shift. It provides no evidence that moving every action score together explains the selected update's much larger ranking gain. These are descriptive ablations, with the complete condition table in Appendix B.

Confidence reports also change under the selected update. Original Qwen Brier score decreases from 0.42220 to 0.34835 on PopQA and from 0.28508 to 0.26707 on TriviaQA. This descriptive improvement is compatible with a change in uncertainty reporting, but it does not identify which correctness representation drives routing. The weight and coefficient experiments support different levels of explanation and should not be treated as interchangeable evidence.

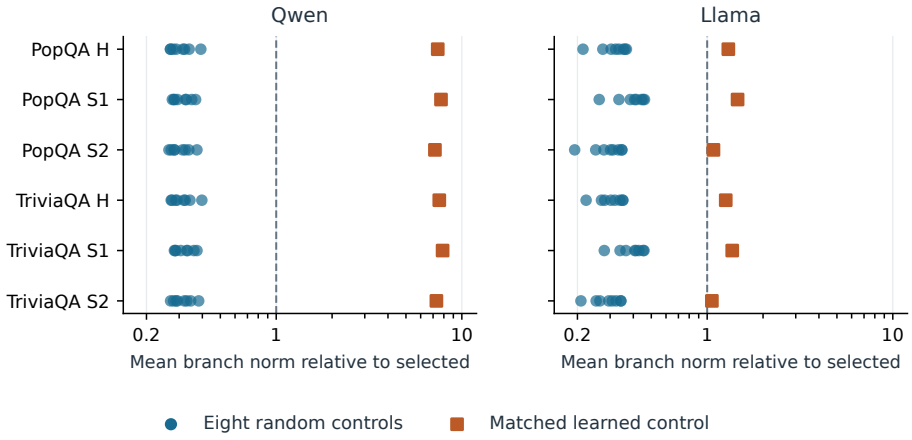


Figure 5. Equal parameter norms do not give equal observed branch-output norms. The statistic is the ratio of mean root-sum-square module-output norms to that of the selected condition. All eight random controls and the matched learned control are shown for each dataset and router. It is not a net residual-stream perturbation norm. The vertical line marks equal magnitude.

Against the eight original spectrum-preserving random controls, the selected Qwen update has positive ordinary paired secondary intervals in all six dataset/router panels under both automated rules. Their measured strength mismatch limits a direction-specific interpretation. In Qwen, the random controls produce 0.26 to 0.40 times the selected mean branch-output norm, while the learned control produces 7.2 to 7.9 times that norm (Figure 5).

Control outcomes. Table 2 records canonical AUROC for the calibrated parent C, full router R, selected model S, selected-ablated router, full local quarter and scalar-bias control. It reports all routers and datasets without pooling their outcomes. The scalar shifts are fitted on discovery, and small recorded numerical differences from C are retained.

Table 2. Descriptive canonical AUROCs for the learned-update ablations. No intervals are inferred from this table.

Family / router	Data	C	R	S	R minus S	Full quarter	Scalar
Qwen H	P	0.70547	0.90908	0.81837	0.85678	0.82800	0.70543
Qwen H	T	0.76844	0.85349	0.81754	0.83965	0.82432	0.76846
Qwen S1	P	0.70547	0.90602	0.82257	0.86734	0.83587	0.70549
Qwen S1	T	0.76844	0.85850	0.82224	0.84744	0.83015	0.76844
Qwen S2	P	0.70547	0.90209	0.80699	0.86738	0.82095	0.70547
Qwen S2	T	0.76844	0.85692	0.81350	0.84902	0.82178	0.76844
Llama H	P	0.86896	0.88770	0.88259	0.87351	0.88378	0.86896
Llama H	T	0.77850	0.79696	0.79141	0.76218	0.79631	0.77850
Llama S1	P	0.86896	0.88211	0.88208	0.87639	0.88633	0.86896
Llama S1	T	0.77850	0.78738	0.78369	0.78249	0.78770	0.77850
Llama S2	P	0.86896	0.88905	0.88646	0.88636	0.89044	0.86896
Llama S2	T	0.77850	0.83757	0.80423	0.81241	0.81176	0.77850

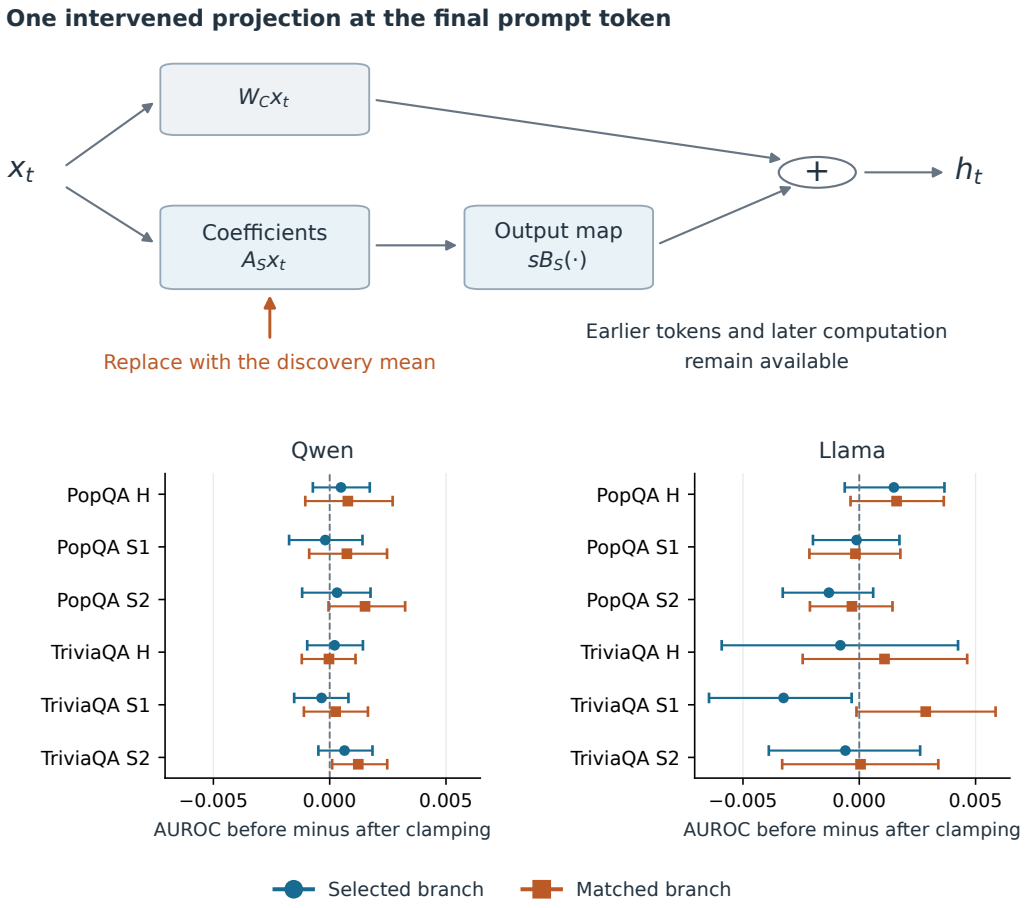


Figure 6. The final-token coefficient test. The upper diagram shows the intervention within one selected projection. The parent projection is combined with the selected low-rank branch, whose coefficients are replaced before its output map. This is a projection schematic rather than a complete circuit. The lower plots compare each learned branch with its mean-clamped version under canonical labels. Bars are ordinary secondary 95% paired intervals. Blue circles indicate the selected branch and orange squares the matched branch. Nonzero branch changes were recorded on every question. The test leaves earlier prompt positions and downstream computation available.

C COMPLETE PRIMARY RESULTS

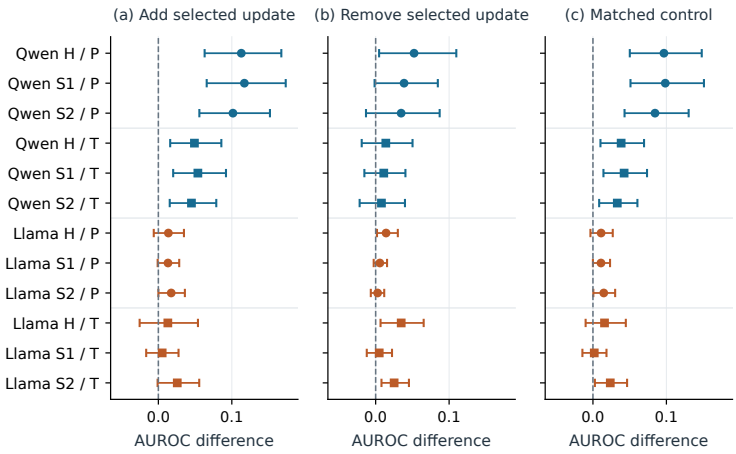


Figure 7. All 36 canonical primary comparisons. Panels show $S - C$, $R - (R \setminus S)$ and selected minus matched AUROC. Bars are nominal family-adjusted paired percentile-bootstrap intervals. H denotes the original router and S1/S2 the new routing seeds. P/T denote PopQA/TriviaQA. Blue marks Qwen and orange marks Llama. Addition and matched-control effects are more consistent than removal effects for Qwen.

Each cell is an AUROC difference with its nominal family-adjusted interval. H is the original router; S1/S2 are seeds 2026091301/2026091302. P/T denote PopQA/TriviaQA. Near-zero endpoints retain five decimals so that borderline Llama findings remain visible. Reference-only results are sensitivity analyses, not a second independent primary family.

C.1 CANONICAL LABELS

Family / router	Data	Selected − C	R − ablated	Selected − matched
Qwen H	P	+0.11290 [+0.06298, +0.16740]	+0.05230 [+0.00470, +0.10980]	+0.09664 [+0.05017, +0.14811]
Qwen S1	P	+0.11710 [+0.06581, +0.17338]	+0.03868 [-0.00146, +0.08463]	+0.09848 [+0.05111, +0.15111]
Qwen S2	P	+0.10152 [+0.05582, +0.15191]	+0.03471 [-0.01315, +0.08713]	+0.08450 [+0.04302, +0.13030]
Qwen H	T	+0.04910 [+0.01610, +0.08575]	+0.01384 [-0.01898, +0.05021]	+0.03839 [+0.01023, +0.06956]
Qwen S1	T	+0.05379 [+0.02017, +0.09206]	+0.01106 [-0.01535, +0.04053]	+0.04243 [+0.01430, +0.07362]
Qwen S2	T	+0.04506 [+0.01548, +0.07885]	+0.00790 [-0.02181, +0.03990]	+0.03315 [+0.00847, +0.06053]
Llama H	P	+0.01362 [-0.00627, +0.03483]	+0.01419 [+0.00181, +0.03019]	+0.01111 [-0.00341, +0.02697]
Llama S1	P	+0.01311 [-0.00096, +0.02839]	+0.00572 [-0.00267, +0.01552]	+0.01093 [-0.00003, +0.02328]
Llama S2	P	+0.01749 [+0.00004, +0.03610]	+0.00270 [-0.00661, +0.01167]	+0.01473 [+0.00053, +0.03023]
Llama H	T	+0.01291 [-0.02550, +0.05389]	+0.03478 [+0.00665, +0.06537]	+0.01575 [-0.00997, +0.04476]
Llama S1	T	+0.00519 [-0.01657, +0.02746]	+0.00488 [-0.01209, +0.02227]	+0.00179 [-0.01434, +0.01842]
Llama S2	T	+0.02574 [-0.00103, +0.05564]	+0.02516 [+0.00793, +0.04531]	+0.02360 [+0.00263, +0.04648]

C.2 REFERENCE-ONLY LABELS

Family / router	Data	Selected − C	R − ablated	Selected − matched
Qwen H	P	+0.10078 [+0.04440, +0.16079]	+0.05333 [+0.00483, +0.11789]	+0.08259 [+0.03045, +0.13911]
Qwen S1	P	+0.10700 [+0.04991, +0.17110]	+0.03674 [-0.00298, +0.08951]	+0.08495 [+0.03166, +0.14369]
Qwen S2	P	+0.09306 [+0.04037, +0.15131]	+0.03958 [-0.00467, +0.09891]	+0.07460 [+0.02586, +0.12845]
Qwen H	T	+0.04004 [+0.00854, +0.07642]	+0.01113 [-0.02104, +0.04658]	+0.03005 [+0.00297, +0.06016]
Qwen S1	T	+0.04480 [+0.01208, +0.08184]	+0.00994 [-0.01599, +0.03837]	+0.03444 [+0.00728, +0.06446]
Qwen S2	T	+0.03832 [+0.00950, +0.07135]	+0.00717 [-0.02116, +0.03856]	+0.02774 [+0.00368, +0.05407]
Llama H	P	-0.00050 [-0.02530, +0.02423]	+0.01832 [+0.00204, +0.03823]	+0.00166 [-0.01637, +0.01967]
Llama S1	P	+0.00789 [-0.00967, +0.02616]	+0.00483 [-0.00486, +0.01578]	+0.00817 [-0.00516, +0.02244]
Llama S2	P	+0.00655 [-0.01385, +0.02749]	+0.00433 [-0.00667, +0.01565]	+0.00641 [-0.01023, +0.02391]
Llama H	T	+0.01880 [-0.01971, +0.05906]	+0.03226 [+0.00502, +0.06207]	+0.01987 [-0.00647, +0.04884]
Llama S1	T	+0.00786 [-0.01315, +0.02964]	+0.00214 [-0.01500, +0.01905]	+0.00574 [-0.01020, +0.02209]
Llama S2	T	+0.02666 [-0.00186, +0.05718]	+0.02288 [+0.00522, +0.04264]	+0.02403 [+0.00316, +0.04729]

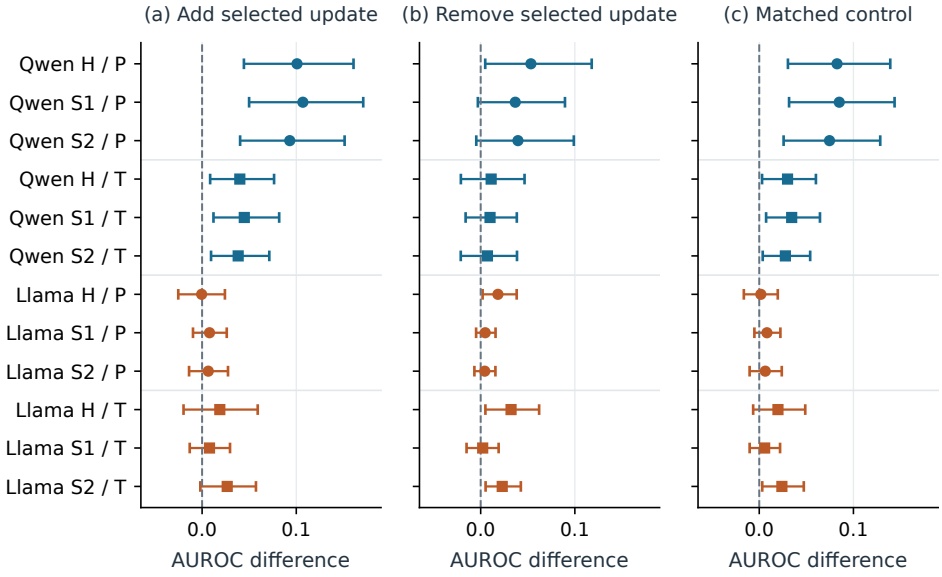


Figure 8. All 36 primary contrasts under the reference-only sensitivity rule. Axes and symbol definitions match Figure 7. This is not independent semantic adjudication.

D COMPLETE NATIVE COMPARISONS

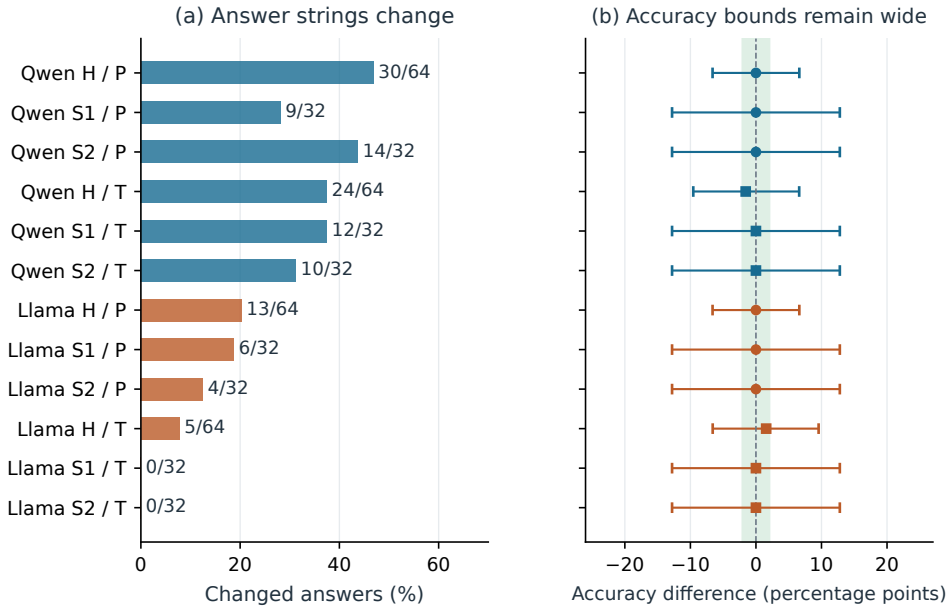


Figure 9. Native selected-versus-parent outcomes under canonical labels. The left panel shows changed extracted-answer strings with their counts. The right panel shows net accuracy changes and conservative per-contrast 95% gain/loss bounds. The green band marks the prespecified two-point equivalence region. None of the intervals fits inside it. String differences do not establish semantic differences.

The tables retain all five contrasts for every family, router, dataset and label rule. R is the full router, C the calibrated parent, S the selected update throughout answering, P the selected prefill-only condition and G the selected continuation-only condition. In the Data column, P/T mean PopQA/TriviaQA; they do not identify intervention timing. Strings counts changed extracted answers, and G/L gives correctness gains/losses. String changes do not establish semantic changes. Bounds are in percentage points and are conservative per-contrast bounds, not simultaneous coverage across all rows.

Table 3. All native contrasts with canonical labels. Conservative bounds are per contrast.

Family / router	Data	Contrast	n	Strings	G/L	Δ (pp)	95% bound
Qwen H	P	R−C	64	36	1/0	+1.56	[-6.60, +9.56]
Qwen H	P	S−C	64	30	0/0	+0.00	[-6.62, +6.62]
Qwen H	P	S−R	64	29	0/1	-1.56	[-9.56, +6.60]
Qwen H	P	P−R	64	30	0/1	-1.56	[-9.56, +6.60]
Qwen H	P	G−R	64	20	0/0	+0.00	[-6.62, +6.62]
Qwen H	T	R−C	64	39	2/2	+0.00	[-11.84, +11.84]
Qwen H	T	S−C	64	24	0/1	-1.56	[-9.56, +6.60]
Qwen H	T	S−R	64	33	1/2	-1.56	[-12.08, +9.30]
Qwen H	T	P−R	64	32	2/2	+0.00	[-11.84, +11.84]
Qwen H	T	G−R	64	21	2/0	+3.12	[-6.35, +12.10]
Qwen S1	P	R−C	32	18	0/0	+0.00	[-12.80, +12.80]
Qwen S1	P	S−C	32	9	0/0	+0.00	[-12.80, +12.80]
Qwen S1	P	S−R	32	14	0/0	+0.00	[-12.80, +12.80]
Qwen S1	P	P−R	32	14	0/0	+0.00	[-12.80, +12.80]
Qwen S1	P	G−R	32	9	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	R−C	32	16	1/1	+0.00	[-18.30, +18.30]

Continued on next page

Family / router	Data	Contrast	n	Strings	G/L	Δ (pp)	95% bound
Qwen S1	T	S−C	32	12	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	S−R	32	17	1/1	+0.00	[-18.30, +18.30]
Qwen S1	T	P−R	32	15	1/1	+0.00	[-18.30, +18.30]
Qwen S1	T	G−R	32	8	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	R−C	32	17	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	S−C	32	14	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	S−R	32	13	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	P−R	32	12	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	G−R	32	7	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	R−C	32	16	1/0	+3.12	[-12.76, +18.34]
Qwen S2	T	S−C	32	10	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	S−R	32	12	0/1	-3.12	[-18.34, +12.76]
Qwen S2	T	P−R	32	13	0/2	-6.25	[-23.07, +12.27]
Qwen S2	T	G−R	32	11	0/2	-6.25	[-23.07, +12.27]
Llama H	P	R−C	64	15	0/0	+0.00	[-6.62, +6.62]
Llama H	P	S−C	64	13	0/0	+0.00	[-6.62, +6.62]
Llama H	P	S−R	64	12	0/0	+0.00	[-6.62, +6.62]
Llama H	P	P−R	64	14	0/0	+0.00	[-6.62, +6.62]
Llama H	P	G−R	64	8	0/0	+0.00	[-6.62, +6.62]
Llama H	T	R−C	64	6	0/0	+0.00	[-6.62, +6.62]
Llama H	T	S−C	64	5	1/0	+1.56	[-6.60, +9.56]
Llama H	T	S−R	64	4	1/0	+1.56	[-6.60, +9.56]
Llama H	T	P−R	64	5	1/0	+1.56	[-6.60, +9.56]
Llama H	T	G−R	64	3	0/0	+0.00	[-6.62, +6.62]
Llama S1	P	R−C	32	9	1/1	+0.00	[-18.30, +18.30]
Llama S1	P	S−C	32	6	0/0	+0.00	[-12.80, +12.80]
Llama S1	P	S−R	32	6	1/1	+0.00	[-18.30, +18.30]
Llama S1	P	P−R	32	6	0/1	-3.12	[-18.34, +12.76]
Llama S1	P	G−R	32	7	1/1	+0.00	[-18.30, +18.30]
Llama S1	T	R−C	32	2	1/0	+3.12	[-12.76, +18.34]
Llama S1	T	S−C	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S1	T	S−R	32	2	0/1	-3.12	[-18.34, +12.76]
Llama S1	T	P−R	32	2	0/1	-3.12	[-18.34, +12.76]
Llama S1	T	G−R	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	R−C	32	8	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	S−C	32	4	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	S−R	32	7	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	P−R	32	5	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	G−R	32	6	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	R−C	32	2	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	S−C	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	S−R	32	2	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	P−R	32	3	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	G−R	32	1	0/0	+0.00	[-12.80, +12.80]

Table 4. All native contrasts with reference only labels. Conservative bounds are per contrast.

Family / router	Data	Contrast	n	Strings	G/L	Δ (pp)	95% bound
Qwen H	P	R−C	64	36	1/1	+0.00	[-9.54, +9.54]
Qwen H	P	S−C	64	30	0/1	-1.56	[-9.56, +6.60]
Qwen H	P	S−R	64	29	0/1	-1.56	[-9.56, +6.60]
Qwen H	P	P−R	64	30	0/1	-1.56	[-9.56, +6.60]
Qwen H	P	G−R	64	20	0/0	+0.00	[-6.62, +6.62]
Qwen H	T	R−C	64	39	1/2	-1.56	[-12.08, +9.30]
Qwen H	T	S−C	64	24	0/1	-1.56	[-9.56, +6.60]
Qwen H	T	S−R	64	33	1/1	+0.00	[-9.54, +9.54]
Qwen H	T	P−R	64	32	2/1	+1.56	[-9.30, +12.08]
Qwen H	T	G−R	64	21	1/0	+1.56	[-6.60, +9.56]
Qwen S1	P	R−C	32	18	0/1	-3.12	[-18.34, +12.76]

Continued on next page

Family / router	Data	Contrast	n	Strings	G/L	Δ (pp)	95% bound
Qwen S1	P	S−C	32	9	0/0	+0.00	[-12.80, +12.80]
Qwen S1	P	S−R	32	14	1/0	+3.12	[-12.76, +18.34]
Qwen S1	P	P−R	32	14	1/0	+3.12	[-12.76, +18.34]
Qwen S1	P	G−R	32	9	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	R−C	32	16	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	S−C	32	12	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	S−R	32	17	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	P−R	32	15	0/0	+0.00	[-12.80, +12.80]
Qwen S1	T	G−R	32	8	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	R−C	32	17	0/1	-3.12	[-18.34, +12.76]
Qwen S2	P	S−C	32	14	0/1	-3.12	[-18.34, +12.76]
Qwen S2	P	S−R	32	13	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	P−R	32	12	0/0	+0.00	[-12.80, +12.80]
Qwen S2	P	G−R	32	7	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	R−C	32	16	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	S−C	32	10	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	S−R	32	12	0/0	+0.00	[-12.80, +12.80]
Qwen S2	T	P−R	32	13	0/1	-3.12	[-18.34, +12.76]
Qwen S2	T	G−R	32	11	0/1	-3.12	[-18.34, +12.76]
Llama H	P	R−C	64	15	0/0	+0.00	[-6.62, +6.62]
Llama H	P	S−C	64	13	0/0	+0.00	[-6.62, +6.62]
Llama H	P	S−R	64	12	0/0	+0.00	[-6.62, +6.62]
Llama H	P	P−R	64	14	0/0	+0.00	[-6.62, +6.62]
Llama H	P	G−R	64	8	0/0	+0.00	[-6.62, +6.62]
Llama H	T	R−C	64	6	1/0	+1.56	[-6.60, +9.56]
Llama H	T	S−C	64	5	2/0	+3.12	[-6.35, +12.10]
Llama H	T	S−R	64	4	1/0	+1.56	[-6.60, +9.56]
Llama H	T	P−R	64	5	1/0	+1.56	[-6.60, +9.56]
Llama H	T	G−R	64	3	0/0	+0.00	[-6.62, +6.62]
Llama S1	P	R−C	32	9	1/1	+0.00	[-18.30, +18.30]
Llama S1	P	S−C	32	6	0/0	+0.00	[-12.80, +12.80]
Llama S1	P	S−R	32	6	1/1	+0.00	[-18.30, +18.30]
Llama S1	P	P−R	32	6	0/1	-3.12	[-18.34, +12.76]
Llama S1	P	G−R	32	7	1/1	+0.00	[-18.30, +18.30]
Llama S1	T	R−C	32	2	1/0	+3.12	[-12.76, +18.34]
Llama S1	T	S−C	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S1	T	S−R	32	2	0/1	-3.12	[-18.34, +12.76]
Llama S1	T	P−R	32	2	0/1	-3.12	[-18.34, +12.76]
Llama S1	T	G−R	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	R−C	32	8	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	S−C	32	4	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	S−R	32	7	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	P−R	32	5	0/0	+0.00	[-12.80, +12.80]
Llama S2	P	G−R	32	6	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	R−C	32	2	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	S−C	32	0	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	S−R	32	2	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	P−R	32	3	0/0	+0.00	[-12.80, +12.80]
Llama S2	T	G−R	32	1	0/0	+0.00	[-12.80, +12.80]

E COMPLETE THRESHOLD UTILITY COMPARISONS

All 60 canonical selected-threshold minus full-router-threshold comparisons appear below. Paired intervals are unadjusted secondary intervals. Many exact ties occur because both discovery-fitted policies never call, especially for Llama with its weaker Qwen advisor. The body reports the small canonical exception and does not claim a consistent utility advantage.

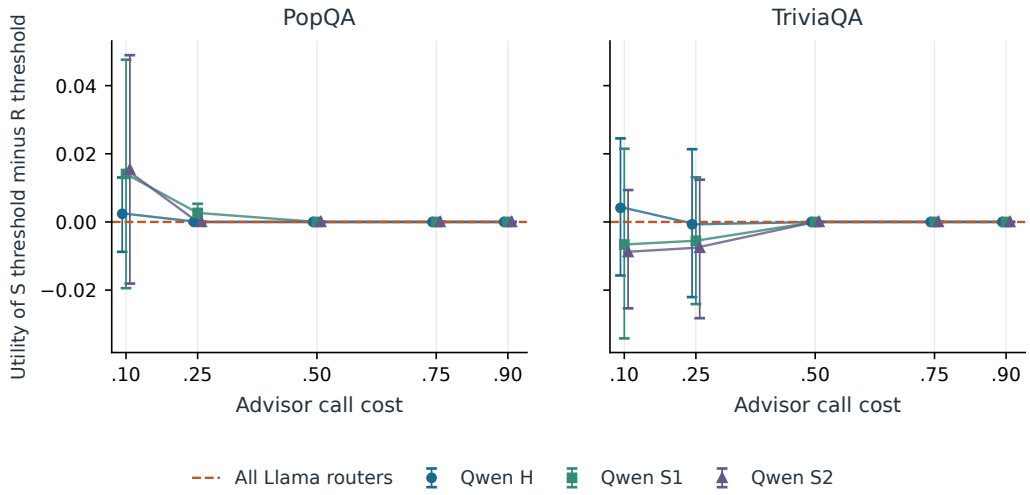


Figure 10. Utility of the partial-router threshold policy relative to the full-router threshold policy. The plotted Qwen comparisons cover five call costs, two datasets and three routing seeds, with ordinary secondary paired intervals. The Llama comparisons coincide at zero because both policies keep the student's answer. Thresholds are fitted on discovery data.

Table 5. Selected threshold minus full-router threshold, canonical labels. All 60 comparisons.

Family / router	Dataset	Cost	Δ utility	Secondary 95% interval
Qwen H	PopQA	0.10	+0.00239	[-0.00878, +0.01304]
Qwen H	PopQA	0.25	+0.00000	[+0.00000, +0.00000]
Qwen H	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen H	PopQA	0.75	+0.00000	[+0.00000, +0.00000]
Qwen H	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Qwen H	TriviaQA	0.10	+0.00413	[-0.01570, +0.02452]
Qwen H	TriviaQA	0.25	-0.00069	[-0.02204, +0.02135]
Qwen H	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen H	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Qwen H	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]
Qwen S1	PopQA	0.10	+0.01410	[-0.01942, +0.04761]
Qwen S1	PopQA	0.25	+0.00266	[+0.00066, +0.00532]
Qwen S1	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen S1	PopQA	0.75	+0.00000	[+0.00000, +0.00000]
Qwen S1	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Qwen S1	TriviaQA	0.10	-0.00661	[-0.03416, +0.02149]
Qwen S1	TriviaQA	0.25	-0.00551	[-0.02410, +0.01309]
Qwen S1	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen S1	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Qwen S1	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]
Qwen S2	PopQA	0.10	+0.01543	[-0.01809, +0.04894]
Qwen S2	PopQA	0.25	+0.00000	[+0.00000, +0.00000]
Qwen S2	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen S2	PopQA	0.75	+0.00000	[+0.00000, +0.00000]

Continued on next page

Family / router	Dataset	Cost	Δ utility	Secondary 95% interval
Qwen S2	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Qwen S2	TriviaQA	0.10	-0.00882	[-0.02534, +0.00937]
Qwen S2	TriviaQA	0.25	-0.00758	[-0.02824, +0.01240]
Qwen S2	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Qwen S2	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Qwen S2	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama H	PopQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama H	PopQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama H	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama H	PopQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama H	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama H	TriviaQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama H	TriviaQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama H	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama H	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama H	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama S1	PopQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama S1	PopQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama S1	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama S1	PopQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama S1	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama S1	TriviaQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama S1	TriviaQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama S1	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama S1	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama S1	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama S2	PopQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama S2	PopQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama S2	PopQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama S2	PopQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama S2	PopQA	0.90	+0.00000	[+0.00000, +0.00000]
Llama S2	TriviaQA	0.10	+0.00000	[+0.00000, +0.00000]
Llama S2	TriviaQA	0.25	+0.00000	[+0.00000, +0.00000]
Llama S2	TriviaQA	0.50	+0.00000	[+0.00000, +0.00000]
Llama S2	TriviaQA	0.75	+0.00000	[+0.00000, +0.00000]
Llama S2	TriviaQA	0.90	+0.00000	[+0.00000, +0.00000]

For comparison, directly generated policies can make different calls even when threshold policies have similar utility. In the original Qwen PopQA panel at cost 0.10, the partial router makes 247 calls, repairs 36 errors and introduces four harmful substitutions. The full router makes 355 calls, repairs 47 errors and introduces 17 harms. Their utilities are 0.23218 and 0.19814. This is one descriptive example, not evidence of a consistent advantage over the full-router threshold baseline.

F Computational audit

The archived numerical check uses a separate implementation of the point estimators and threshold selection. It reproduces 72 primary and sensitivity AUROC contrasts, 408 condition AUROCs, 2,280 fitted thresholds, 4,560 utilities and 120 native transitions. These checks use the same recorded automated labels as the main analysis. They neither provide independent correctness labels nor establish coverage of the percentile intervals.

A decoder-hook mismatch interrupted the original native executor before an affected selected-native record was written. The corrected executor verifies the production KV-cache path and records intervention windows. Storage and transfer interruptions were recovered from validated complete-question prefixes with preserved logs and archive hashes. The frozen targets, outcome rules, comparison families and estimators were retained. The final transfer recovery required no new inference.

G FOCUSED POSITIONAL MECHANISM STUDY

Separate fitting, selection and confirmation. The follow-up preserves the previously selected Qwen update and allocates three fresh sets using a fixed salted question ordering. Each dataset contributes 64 fitting, 96 selection and 320 confirmation questions. Questions recorded in routing training and prior evaluation are excluded by normalized text and dataset identifier. Freshness is relative to those recorded inputs. It does not establish absence from pretraining or unrelated projects. Empty or generation-capped parent answers are excluded from fixed-answer ranking, retained in the candidate archive, and never replaced with outcome-selected questions. There are 122 eligible fitting candidates and 182 eligible selection candidates across the two datasets.

The original router supplies discovery evidence. The independent routing seed 2026091301 supplies replication, conditional on the same calibration parent. Its coefficient means and control scales are fitted separately by the same fixed algorithm. Replication outcomes do not choose the intervention. Numerical reconstruction matches the archived parent-file hashes, and archived token replays are checked on both A100 workers.

Token positions and coefficient replacement. The intervention acts on the selected low-rank branch of each module. For incoming hidden state $h_{j,t}$, let $z_{j,t} = A_j^{(S)} h_{j,t}$ be its coefficients. The branch contributes $s_j B_j^{(S)} z_{j,t}$. For dataset d and prompt window w, the fitted replacement is

$$\mu_{j,d,w} = \frac{1}{n_d} \sum_{i \in d} \frac{1}{|w_i|} \sum_{t \in w_i} z_{j,t}^{(i)}. \tag{8}$$

Each fitting question has equal weight, regardless of its token count. Means use no correctness labels. They are fitted separately for routing prompts and native-answer prompts. Pooling across positions removes positional as well as example-dependent variation, including variation in the common prompt prefix. The intervention is therefore not a selective erasure of correctness information.

The three routing windows are the final prompt token, the inserted candidate-answer span, and all prefill positions. The answer span is identified from character offsets in the fully rendered chat prompt. Tokenizing the answer separately would not reliably recover boundary tokens. Ambiguous spans and oversized prompts cause explicit validation failures rather than silent truncation or a different target position.

Mean replacement sets $z_{j,t}$ to $\mu_{j,d,w}$ at the chosen positions. It runs only during the first model forward. The resulting key/value cache remains in use during generation, while subsequent coefficient computations are unconstrained. This temporal boundary concerns when replacement occurs, not when its consequences end. The selected weight update remains active during continuation.

Discovery rule and restoration. A window passes discovery when selected-minus-clamped error AUROC exceeds 0.01 on both datasets. The first passing window in the fixed order final token, answer span, full prefill is selected. Only full prefill passes. Within that window, the selected layers are partitioned into three equal groups, 9 to 11, 12 to 14 and 15 to 17. Restoring each group's intact same-example coefficients is evaluated on selection data. The group with the largest minimum recovery across datasets is chosen, with ties resolved toward earlier layers. Layers 15 to 17 pass the separate recovery gate of 0.005 on both datasets. Both gates, the numerical artifacts and every selection outcome are frozen before confirmation interventions.

Restoration uses coefficients captured during an intact forward of S on the same question. The other selected modules remain clamped. The wrong-group control restores layers 9 to 11, the next group cyclically after the chosen group. The donor control restores coefficients from a different fitting question in the same dataset, selected by closest window length with a fixed hash tie break. Donor positions are mapped by nearest relative index when lengths differ. Neither donor selection nor alignment uses correctness. A full self-patch must reproduce intact logits within numerical tolerance, but serves only as an implementation check. Same-example partial restoration also requires an intact forward and is not a standalone deployment policy.

Matching observed activation strength. The original parameter-norm-matched learned control is retained. Additional controls comprise its activation-scaled version and four previously specified

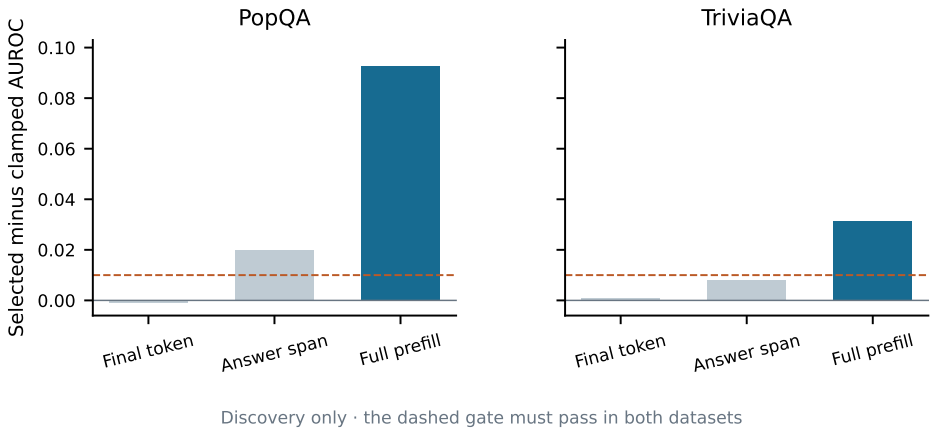


Figure 11. Positional comparison on selection data only. Full-prefill replacement is the first window in the fixed order to exceed the discovery loss gate on both datasets. The answer-span effect passes on PopQA alone. These bars are descriptive selection estimates, not confirmation intervals.

random rotations. For each module, the fitting statistic is the mean across questions of full-prefill branch-output RMS. A control's B factor is multiplied by the ratio of the selected target statistic to its current statistic. Three predetermined fitting iterations account for changed downstream inputs. The factors are then frozen. Correctness labels play no role in this fitting.

The final fit residuals, parameter norms and held-out branch magnitudes are retained. We also measure the local change caused by each clamp,

$$\delta_{j,t} = s_j B_j^{(S)}(\mu_{j,d,w} - z_{j,t}). \tag{9}$$

These changes use the incoming state at the intervened forward. They are not the complete propagated difference from the intact model. Aggregating module magnitudes by root-sum-square does not yield a net residual-stream norm. Matching branch RMS likewise does not guarantee equal clamp perturbations, and neither statistic establishes equal functional disruption.

Frozen confirmation family. The five primary contrasts are the selected clamp loss, restoration recovery over clamping, restoration over donor restoration, restoration over wrong-group restoration, and selected clamp loss minus the activation-scaled learned-control clamp loss. Two datasets and two routers give 20 primary comparisons. The analysis uses 50,000 paired question resamples and nominal Bonferroni-adjusted percentile intervals. Reference-only labels are a sensitivity analysis. All four random controls are reported as secondary comparisons. Class-sensitive logit changes use separate secondary intervals.

Every eligible question must complete the full condition matrix before its atomic completion record is committed. Partial attempts are retained separately. Logs record prompt and output token IDs, A/K mass, strict action validity, coefficient changes, branch magnitudes, affected positions and generation-call boundaries. Real-model checks verify cache modification and changed continuation under the same externally fixed next token. These checks establish that the intervention is active, not that it isolates correctness information.

Native-answer measurement. A deterministic nested sample contains 64 confirmation questions per dataset. The calibrated parent, selected-only model and selected model with full-prefill coefficient replacement each generate an answer. Native means are fitted on native prompts. There is no candidate-answer span in this setting. Native correctness changes use conservative simultaneous exact binomial bounds on gains and losses. The analysis retains capped generations and separates token changes, answer-string changes and automated correctness transitions. It does not infer semantic equivalence from string comparisons.

G.1 COMPLETE FOCUSED ROUTING RESULTS

H denotes the original router and S1 denotes independently trained routing seed 2026091301. P and T denote PopQA and TriviaQA. Both routers use the same 310 eligible PopQA and 309 eligible TriviaQA candidates. Canonical labels identify 44 and 172 correct candidates, respectively. Reference-only labels identify 30 and 150. The five comparisons on two datasets and two routers form a 20-comparison canonical family. The reference-only tables are sensitivity analyses using the same matching code, not independent semantic adjudication. Every interval below is the recorded paired percentile interval, conditional on these checkpoints and fitted artifacts.

Table 6. Primary activity comparisons under canonical labels. Intervals have nominal 95 percent coverage across the comparison family.

Router	Data	Contrast	Difference and adjusted interval
H	P	Clamp loss	+0.04879 [+0.02284, +0.08000]
H	P	Restoration over clamp	+0.03328 [+0.01663, +0.05327]
H	P	Restoration over donor	+0.01431 [+0.00543, +0.02516]
H	P	Restoration over wrong group	+0.02832 [+0.01351, +0.04629]
H	P	Selected minus control clamp loss	+0.03973 [+0.01235, +0.07263]
H	T	Clamp loss	+0.03249 [+0.01124, +0.05768]
H	T	Restoration over clamp	+0.02149 [+0.00726, +0.03923]
H	T	Restoration over donor	+0.01326 [+0.00290, +0.02669]
H	T	Restoration over wrong group	+0.01494 [+0.00338, +0.02899]
H	T	Selected minus control clamp loss	+0.02746 [+0.00751, +0.05105]
S1	P	Clamp loss	+0.05425 [+0.02540, +0.08904]
S1	P	Restoration over clamp	+0.03708 [+0.01652, +0.06139]
S1	P	Restoration over donor	+0.01252 [+0.00349, +0.02403]
S1	P	Restoration over wrong group	+0.03166 [+0.01462, +0.05241]
S1	P	Selected minus control clamp loss	+0.04691 [+0.01685, +0.08312]
S1	T	Clamp loss	+0.03306 [+0.01120, +0.05992]
S1	T	Restoration over clamp	+0.02381 [+0.00692, +0.04398]
S1	T	Restoration over donor	+0.01082 [+0.00028, +0.02428]
S1	T	Restoration over wrong group	+0.01640 [+0.00208, +0.03254]
S1	T	Selected minus control clamp loss	+0.02693 [+0.00737, +0.05095]

Table 7. Reference-only sensitivity comparisons, using the same interval construction as Table 6.

Router	Data	Contrast	Difference and adjusted interval
H	P	Clamp loss	+0.05220 [+0.02134, +0.08838]
H	P	Restoration over clamp	+0.03583 [+0.01657, +0.05887]
H	P	Restoration over donor	+0.01512 [+0.00416, +0.02847]
H	P	Restoration over wrong group	+0.03018 [+0.01211, +0.05183]
H	P	Selected minus control clamp loss	+0.04292 [+0.01050, +0.08117]
H	T	Clamp loss	+0.03117 [+0.01094, +0.05586]
H	T	Restoration over clamp	+0.02027 [+0.00678, +0.03681]
H	T	Restoration over donor	+0.01218 [+0.00271, +0.02457]
H	T	Restoration over wrong group	+0.01547 [+0.00508, +0.02866]
H	T	Selected minus control clamp loss	+0.02665 [+0.00784, +0.04966]
S1	P	Clamp loss	+0.05970 [+0.02684, +0.10065]
S1	P	Restoration over clamp	+0.04012 [+0.01572, +0.06876]
S1	P	Restoration over donor	+0.01220 [+0.00183, +0.02424]
S1	P	Restoration over wrong group	+0.03411 [+0.01339, +0.05820]
S1	P	Selected minus control clamp loss	+0.05298 [+0.01790, +0.09680]
S1	T	Clamp loss	+0.03342 [+0.01274, +0.05869]
S1	T	Restoration over clamp	+0.02375 [+0.00814, +0.04279]
S1	T	Restoration over donor	+0.01052 [+0.00087, +0.02229]
S1	T	Restoration over wrong group	+0.01637 [+0.00360, +0.03122]
S1	T	Selected minus control clamp loss	+0.02759 [+0.00938, +0.05075]

G.2 FOCUSED CONDITION SUMMARIES

Table 8. Canonical routing outcomes. Changed counts generated-token differences from S, not semantic answer changes.

Router	Data	Condition	AUROC	Valid	Calls	Changed
H	P	Parent	0.7397	310/310	112	105
H	P	Full router	0.9237	309/310	300	85
H	P	Selected	0.8538	310/310	217	0
H	P	Original matched	0.7569	310/310	119	98
H	P	Scaled matched	0.7501	310/310	116	101
H	P	Scaled random 1	0.7379	310/310	112	105
H	P	Scaled random 2	0.7327	310/310	113	104
H	P	Scaled random 3	0.7365	310/310	113	104
H	P	Scaled random 4	0.7388	310/310	114	103
H	P	Selected clamped	0.8050	310/310	143	74
H	P	Scaled matched clamped	0.7411	310/310	112	105
H	P	Same-example restoration	0.8383	310/310	177	40
H	P	Donor restoration	0.8240	310/310	171	46
H	P	Wrong-group restoration	0.8100	310/310	154	63
H	T	Parent	0.7729	309/309	63	66
H	T	Full router	0.8826	300/309	248	131
H	T	Selected	0.8441	309/309	129	0
H	T	Original matched	0.7874	309/309	69	60
H	T	Scaled matched	0.7797	309/309	64	65
H	T	Scaled random 1	0.7700	309/309	62	67
H	T	Scaled random 2	0.7723	309/309	62	67
H	T	Scaled random 3	0.7731	309/309	63	66
H	T	Scaled random 4	0.7760	309/309	63	66
H	T	Selected clamped	0.8117	309/309	93	36
H	T	Scaled matched clamped	0.7747	309/309	63	66
H	T	Same-example restoration	0.8332	309/309	109	20
H	T	Donor restoration	0.8199	309/309	105	24
H	T	Wrong-group restoration	0.8182	309/309	100	29
S1	P	Parent	0.7397	310/310	112	93
S1	P	Full router	0.9165	310/310	302	97
S1	P	Selected	0.8623	310/310	205	0
S1	P	Original matched	0.7603	310/310	118	87
S1	P	Scaled matched	0.7501	310/310	117	88
S1	P	Scaled random 1	0.7364	310/310	115	90
S1	P	Scaled random 2	0.7329	310/310	113	92
S1	P	Scaled random 3	0.7380	310/310	116	89
S1	P	Scaled random 4	0.7374	310/310	114	91
S1	P	Selected clamped	0.8081	310/310	142	63
S1	P	Scaled matched clamped	0.7427	310/310	112	93
S1	P	Same-example restoration	0.8451	310/310	169	36
S1	P	Donor restoration	0.8326	310/310	162	43
S1	P	Wrong-group restoration	0.8135	310/310	147	58
S1	T	Parent	0.7729	309/309	63	58
S1	T	Full router	0.8740	306/309	253	136
S1	T	Selected	0.8480	309/309	121	0
S1	T	Original matched	0.7885	309/309	68	53
S1	T	Scaled matched	0.7792	309/309	63	58
S1	T	Scaled random 1	0.7732	309/309	62	59
S1	T	Scaled random 2	0.7730	309/309	64	57
S1	T	Scaled random 3	0.7706	309/309	65	56
S1	T	Scaled random 4	0.7727	309/309	61	60
S1	T	Selected clamped	0.8150	309/309	84	37
S1	T	Scaled matched clamped	0.7731	309/309	63	58
S1	T	Same-example restoration	0.8388	309/309	105	16
S1	T	Donor restoration	0.8279	309/309	99	22
S1	T	Wrong-group restoration	0.8224	309/309	96	25

G.3 ACTIVATION-SCALED RANDOM CONTROLS

All four frozen rotations are reported with ordinary secondary intervals. No favorable random seed is selected for reporting. The preceding condition table reports canonical error AUROC, valid generated actions, request counts and complete generated-token changes relative to the selected model. Valid gives the valid count over the panel size.

Table 9. Selected minus activation-scaled random-control AUROC with ordinary 95 percent paired intervals.

Router	Data	Selected minus	Difference and ordinary interval
H	P	Scaled random 1	$+0.11590$ $[+0.08240, +0.15164]$
H	P	Scaled random 2	$+0.12107$ $[+0.08717, +0.15711]$
H	P	Scaled random 3	$+0.11727$ $[+0.08380, +0.15321]$
H	P	Scaled random 4	$+0.11496$ $[+0.08173, +0.15038]$
H	T	Scaled random 1	$+0.07412$ $[+0.04762, +0.10279]$
H	T	Scaled random 2	$+0.07180$ $[+0.04526, +0.10049]$
H	T	Scaled random 3	$+0.07102$ $[+0.04480, +0.09958]$
H	T	Scaled random 4	$+0.06820$ $[+0.04284, +0.09578]$
S1	P	Scaled random 1	$+0.12590$ $[+0.09146, +0.16276]$
S1	P	Scaled random 2	$+0.12944$ $[+0.09474, +0.16678]$
S1	P	Scaled random 3	$+0.12427$ $[+0.08985, +0.16156]$
S1	P	Scaled random 4	$+0.12491$ $[+0.09099, +0.16161]$
S1	T	Scaled random 1	$+0.07482$ $[+0.04732, +0.10513]$
S1	T	Scaled random 2	$+0.07503$ $[+0.04761, +0.10544]$
S1	T	Scaled random 3	$+0.07743$ $[+0.04990, +0.10795]$
S1	T	Scaled random 4	$+0.07531$ $[+0.04777, +0.10567]$

G.4 NATIVE ANSWERS IN THE FOCUSED STUDY

Each router uses the same deterministic nested set of 64 questions per dataset, including capped outputs. Strings counts extracted-answer differences. G/L gives automated correctness gains and losses. The intervals subtract simultaneous exact binomial bounds for gains and losses within each contrast. They are conservative per-contrast intervals and are not adjusted across all native comparisons. None fits within the intended two-percentage-point equivalence margin. Canonical and reference-only transitions coincide in this small panel. That agreement does not independently establish their semantic accuracy. Native intervention conditions use native-prompt coefficient means, rather than means from routing prompts containing a candidate answer.

Table 10. Native answer changes under canonical labels, with conservative accuracy-change bounds in percentage points.

Router	Data	Contrast	n	Strings	G/L	Accuracy change in points
H	P	$S - C$	64	34	0/0	$+0.00$ $[-6.62, +6.62]$
H	P	Clamp minus S	64	27	1/0	$+1.56$ $[-6.60, +9.56]$
H	T	$S - C$	64	21	0/3	-4.69 $[-14.44, +5.87]$
H	T	Clamp minus S	64	20	3/0	$+4.69$ $[-5.87, +14.44]$
S1	P	$S - C$	64	30	1/0	$+1.56$ $[-6.60, +9.56]$
S1	P	Clamp minus S	64	23	0/1	-1.56 $[-9.56, +6.60]$
S1	T	$S - C$	64	19	1/2	-1.56 $[-12.08, +9.30]$
S1	T	Clamp minus S	64	17	0/1	-1.56 $[-9.56, +6.60]$

Class-sensitive logit changes. Clamping lowers request-minus-keep logits for both labeled-correct and labeled-wrong answers. Under canonical labels, the wrong-minus-correct difference in the clamp-induced shift is positive in all four panels, ranging from 2.44 to 2.83 with ordinary secondary intervals excluding zero. Restoration increases request preference in both classes, but its wrong-minus-correct mean-shift interval includes zero in every panel under both label rules. The restoration result therefore concerns improved pairwise ranking, not a demonstrated change in this class-mean contrast. A uniform additive shift cannot explain an AUROC change, but the ranking result alone does not identify the semantic content of the restored coefficients.

Computational audit. The raw-record audit checks 17,332 routing condition records and 768 native generations, including full matrices, prompt identity, affected coefficient positions and generation boundaries. A pair-enumeration implementation reproduces 152 AUROC estimates and contrast points. All 256 native parent generations reproduce the shared candidate outputs across workers. A compact-score reproduction also matches all 40 primary and reference-sensitivity point estimates and intervals exactly. These computational checks reproduce the recorded statistics without independently validating the semantic correctness of the answers.

G.5 Descriptive retention and donor comparisons

The Discussion adds post-hoc summaries of the frozen canonical scores. Let u_C, u_S, u_M, u_R and u_D denote AUROC for the starting model, intact partial router, mean-replaced partial router, same-question restoration and donor restoration, respectively. The fraction of the selected gain retained under replacement is $(u_M - u_C)/(u_S - u_C)$. The donor fraction of restoration is $(u_D - u_M)/(u_R - u_M)$. These summarize differences in ranking performance and do not partition independent causal mechanisms.

Routing run	Data	Retained gain (%)	Donor recovery (%)
Original	PopQA	57.2 [48.1, 66.7]	57.0 [43.9, 69.3]
Original	TriviaQA	54.4 [46.9, 63.8]	38.3 [20.3, 57.9]
New seed	PopQA	55.7 [46.2, 65.7]	66.2 [53.2, 78.4]
New seed	TriviaQA	56.0 [48.8, 65.1]	54.5 [40.2, 75.0]

Table 11. Descriptive retained-gain and donor-recovery percentages with ordinary 95 percent paired bootstrap intervals. These post-hoc intervals are not part of the 20-comparison primary family.

We use 50,000 paired question resamples per panel with a separate fixed seed. The retained-fraction intervals all contain 50 percent, so the point estimates do not establish that a majority of the gain survives. One donor-ratio resample in the new-seed TriviaQA panel has a nonpositive restoration denominator and is retained; all other ratio denominators are positive. Same-question restoration exceeds donor restoration by 0.0143 and 0.0133 AUROC in the original PopQA and TriviaQA panels, and by 0.0125 and 0.0108 in the new-seed panels. Uncertainty for these prespecified differences remains in the primary results.

The scored advice logits come from the first, intervened forward pass. Mean replacement changes every selected low-rank coefficient at every prompt position into a module-specific constant fitted separately for each dataset and router. The selected branch then supplies $sB\mu$ instead of $sBAh$. This is a constant internal addition, not a constant shift of the final advice score. The backbone and calibration-trained parent continue to process the input, and their response to the added constants can remain input-dependent. Donor coefficients also retain positional structure, and donors are selected by prompt length. Their recovery therefore does not isolate a purely example-independent mechanism.

The scaled learned control gains only 0.0063 to 0.0104 AUROC over the parent in these panels. Its limited initial benefit makes the difference in replacement losses an incomplete test of direction specificity. Dividing by each update's own gain would introduce small and uncertain denominators; we retain absolute effects and their baseline context instead.