ENDPOINT-SENSITIVE CAUSAL AUDITS OF MIXTURE-OF-EXPERTS SAFETY

Anonymous authors
Paper under double-blind review

ABSTRACT

An expert-routing intervention can change both an autoregressive model's next token and the attention state retained for later decoding. Its measured safety effect also depends on the answer span being evaluated. We combine a four-cell token/state experiment with prefix and complete-answer endpoints, paired and different-request donors, and a final-layer continuation invariant. In a three-model development audit with 279 prompt pairs per model, changing the evaluated span changes which pairs appear unstable, even when aggregate flip counts agree. A frozen 49-request study per model then tests whether local interventions change answer-level compliance. With the next token fixed, paired routing transfer changes decoded text in 32/33 naturally completed OLMoE misinformation comparisons and 20/21 Qwen3-MoE comparisons, yet changes no complete-answer compliance labels. A dense residual comparator likewise changes text in 19/31 comparisons without changing compliance. Different-request donors produce comparable text changes. Few native MoE answers are classifier-positive, limiting conclusions about repair. One reserved OLMoE case has a nonzero state contrast under prefix grading and zero under complete-answer grading; a selected development case exhibits the complementary pattern. Execution audits also identify four auxiliary factuality trials that fail strict position isolation. These findings support a joint audit of decoding pathways, numerical validity, and evaluation endpoints. They provide bounded counterexamples to treating token recovery, changed text, and measured safety repair as interchangeable evidence.

1 INTRODUCTION

Sparse mixture-of-experts (MoE) models select a small set of expert networks for each token. This observable selection makes expert routing an attractive candidate explanation for changes in model behavior. Yet an intervention that changes a local computation must pass through a sequence of decoding decisions before it produces the answer on which a safety claim is based. Understanding that sequence requires more than checking whether a donor token is recovered or whether one classifier label changes.

We examine two boundaries in this inference. The first is between an internal intervention and the subsequent generated sequence. A one-decision intervention may change both the chosen token and attention keys and values retained for later decoding. Forcing the changed token without transferring the retained state tests one pathway; transferring the state while restoring the original token tests another. The second boundary is between a generated answer and an evaluation endpoint. Exact compliance with a request, endorsement of a specific false assertion, and refusal are distinct outcomes. Evaluating a prefix and evaluating the complete answer can assign different values to the same intervention contrast.

Our development experiments provide a concrete example. A newline edit changes an OLMoE response to a false historical request. Transferring routing from the unedited prompt changes the answer into a correction, but changing only the initial generated token is also sufficient for a corrective answer. A further control holds the initial token fixed while retaining the routing intervention. Its measured compliance effect depends on whether the evaluator sees the first 384 tokens or the whole answer. Moreover, a classifier-negative full answer can retain the request's core false assertion.

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.

This example is selected and exploratory. A broader Qwen3-MoE audit produces next-token routing effects but no native compliance flips in the corresponding development subset. In a subsequent frozen cohort, both MoEs exhibit widespread decoded-text changes without a primary complete-answer compliance change. Our contributions are (1) an executable four-cell audit separating emitted-token and retained-cache pathways, with exact implementation controls; (2) a three-model benchmark of how answer span changes the set of prompts selected as unstable; and (3) a frozen reserved-request study linking intervention-induced text changes to complete-answer compliance, including donor controls and negative results. The controlled decoder experiment does not by itself identify specialized safety experts, a general defense, or natural mediation in a population of prompts.

2 Related work

Activation-patching conclusions depend on the evaluation metric and corruption method (Zhang & Nanda, 2024). Early output tokens matter for safety alignment (Qi et al., 2025), and dynamic activation patching evaluates safety-related neurons during open-ended generation (Chen et al., 2025). We build on these results by evaluating the same routing intervention across token/state cells and answer spans.

Shih et al. (2026) edit internal state while visible scratchpad tokens remain fixed and test whether subsequent computation obeys a known state-transition rule. Their revised study also distinguishes successful steering from reproducing a model's natural target activation. Our cache intervention identifies a pathway carrying influence beyond a fixed next token; it does not identify a semantic state variable or reproduce their stronger state-use criterion.

For MoEs, Engmann et al. (2026) test whether routing statistics predict causal expert importance, and Yoon et al. (2026) evaluate token-conditional counterfactual route utility. Caruzzo et al. (2026) separate routing and content contributions in same-weight self-distillation and show that matched-norm noise reproduces much of a measured routing effect. These studies motivate behavioral and specificity controls. Our different-request donors control topic and approximate prompt length, leaving perturbation magnitude as a limitation.

Safety-specialized expert analysis and alignment are studied by SAFEx, RASA, and RASET (Lai et al., 2025; Liang et al., 2026; Zhang et al., 2026a); RARE separates representation steering from routing (Zhang et al., 2026b). Ren et al. (2026) study routing-mediated pseudo-alignment with router-only adaptation and placebo expert masks. In contrast, our targets remain frozen. We contribute a joint token/state and endpoint audit, numerical execution checks, and a reserved-request evaluation of when local routing effects translate into measured answer-level compliance. We make no priority claim for the individual intervention or evaluation principles.

3 Controlled token and state interventions

3.1 Decision context

Let x_A be an original request and x_B a formatting variant. Generate both responses greedily with the same pinned model and inference settings. Let p be their common generated-token prefix immediately before the first token divergence. We construct each prompt's native attention cache by replaying its own prompt followed by p, using the same cached decoding schedule as generation. Pairs without an observed token divergence remain in cohort accounting but have no eligible intervention decision.

For the recipient context (x_B, p), denote the native next-token logits by ℓ_0, its selected token by $t_0 = \arg\max \ell_0$, and its resulting cache by K_0. Apply intervention I during exactly one forward pass at this decision, obtaining (ℓ_1, K_1) and $t_1 = \arg\max \ell_1$. All subsequent decoding uses the unmodified model. Here, "state" refers specifically to retained attention keys and values, not to an identified semantic variable. The token t_1 is the intervention's actual argmax; it need not equal the original prompt's selected token.

3.2 A FOUR-CELL DECODER EXPERIMENT

Let $D(K, t)$ denote deterministic continuation after emitting token t using cache K, with the fixed prompt, common prefix, remaining generation budget, and stopping rule understood. For evaluator g and observation length L, define

$$Y_{ij}^{(g,L)} = g(\text{view}_L\left[p, t_j, D(K_i, t_j)\right]), \qquad i, j \in \{0, 1\}. \tag{1}$$

The first index selects retained state; the second selects the generated token. The four cells are native generation (00), token-only change (01), state change with the native token fixed (10), and the natural intervention continuation (11). Each treatment has its own K_1 and t_1. A text change means unequal decoded strings after special tokens are skipped; it is not a measure of semantic distance.

For a numeric endpoint, the total controlled contrast decomposes exactly as

$$\Delta_{\text{total}} = Y_{11} - Y_{00}, \tag{2}$$

$$\Delta_{\text{state}} = Y_{10} - Y_{00}, \qquad \Delta_{\text{token}} = Y_{01} - Y_{00}, \tag{3}$$

$$\Delta_{\text{interaction}} = Y_{11} - Y_{10} - Y_{01} + Y_{00}, \tag{4}$$

$$\Delta_{\text{total}} = \Delta_{\text{state}} + \Delta_{\text{token}} + \Delta_{\text{interaction}}. \tag{5}$$

This factorial identity is elementary. Its role is to specify the estimands and prevent a donor-token proxy from being mistaken for the actual intervention pathway. These are controlled decoder contrasts; no assumption identifying a natural indirect effect is made. A token-only effect does not invalidate a routing explanation: routing may cause that token. Likewise, a nonzero state effect establishes influence carried beyond the fixed token under this intervention, not a semantic interpretation of a particular expert.

3.3 A STRUCTURAL CONTROL AT THE FINAL LAYER

Proposition. Consider deterministic cached decoding in a transformer whose final attention sublayer computes its cached keys and values before the final feed-forward sublayer. A one-step intervention restricted to the final feed-forward output or its internal expert routing cannot change the continuation when the emitted token is held fixed, provided earlier computation and the complete attention cache remain unchanged.

Proof. The intervention occurs after the last cache-writing operation of that forward pass. It can change the final representation and next-token logits but cannot change any stored key or value. With the next token fixed, the following decoding step receives identical token IDs, attention masks, positions, and cache tensors. Determinism gives identical next logits and updated caches. Induction over the remaining steps gives identical tokens and the same stopping event. □

This is a computational invariant, not a claim of theoretical novelty. It yields two implementation checks: $Y_{10} = Y_{00}$ and identical generated sequences for cells 11 and 01. We verify cache tensors and full continuations, rather than assuming the invariant survives every implementation detail. Changed expert dispatch can alter numerical execution for other positions, so ideal causal locality requires a numerical check.

3.4 ROUTING TREATMENTS AND CONTROLS

Expert-choice transfer replaces the recipient's selected expert IDs at the final processed position while retaining the recipient's probabilities for those experts and the model's native normalization rule. Router-score transfer replaces the complete vector of router logits at that position, transferring support and weights. The primary treatment acts across all MoE layers during one decision. It does not modify parameters or continue patching during generation.

The fixed replication also uses a different-request donor, chosen by topic and character length before observing target outcomes. This tests dependence on the paired donor, conditional on a fixed donor bank. It is not a matched-norm perturbation control. Gate distances and support overlaps are recorded to expose this limitation. A dense checkpoint supplies middle-layer residual-transfer experiments and a final-layer control. Differences between the dense and MoE checkpoints cannot isolate architecture from training or scale.

4 EXPERIMENTAL DESIGN

4.1 MODELS AND EXECUTION

The target models are OLMoE-1B-7B-0125-Instruct, Qwen3-30B-A3B-Instruct-2507, and Qwen3-4B with thinking disabled (Muennighoff et al., 2024; Yang et al., 2025). We pin model revisions and use Transformers 4.51.3, PyTorch 2.6, BF16, SDPA attention, batch size one, greedy decoding, and fixed seeds on H100 GPUs. Native chat templates are retained. No target-model training is performed.

Each attempt records source hashes, model revision, library versions, and device information. Atomic per-treatment records support restart without overwriting completed outputs. Checks include native next-token replay, bitwise self-intervention replay, complete native continuation replay, restoration after intervention, and final-layer cache and continuation invariants. CPU integration tests exercise all three model implementations, and earlier tests additionally check cached prefixes of different lengths.

4.2 DEVELOPMENT COHORTS

The development cohort for answer-length analysis contains 100 harmful base requests per model: 60 from HarmBench (Mazeika et al., 2024) and 40 from JailbreakBench (Chao et al., 2024). It includes 200 mechanical newline variants and 79 automatically screened paraphrases, plus 10 exact repeats. Excluding repeats gives 379 responses and 279 original-to-variant comparisons clustered within 100 base requests. Paraphrase acceptance is uneven across sources and lacks independent human equivalence validation; formatting results are therefore also reported separately.

Every long answer has a 1,536-token cap. The 384-token view is extracted from the same generated token sequence. HarmBench's pinned classifier evaluates canonical-request compliance individually. A negative compliance label is not treated as a factuality or comprehensive safety label. We compare the identity of flagged pairs as well as their aggregate number.

The decision audit covers all 20 newline pairs within ten development misinformation requests. The category was selected after inspecting the pilot, so this is a development census within a selected category. Bidirectional interventions are applied at the first naturally differing token. A subsequent full-answer case study and its token controls were selected and designed in stages after inspecting development results.

4.3 FROZEN RESERVED-REQUEST REPLICATION

The new cohort was fixed on September 16 before its target generations. It contains all 17 reserved HarmBench misinformation requests and 32 TruthfulQA generation questions (Lin et al., 2021). The latter are sampled deterministically, four each from eight categories: History, Fiction, Language, Misquotations, Myths and Fairytales, Confusion: Places, Weather, and Science. This is a diagnostic subset, not an official full-benchmark score. Answers are requested in one or two sentences and capped at 512 tokens.

The 17 requests are distinct from pilot requests under exact normalization and AI-assisted author review of their propositions. Topics and common persuasive-writing templates still overlap; this is request-held-out evaluation, not domain-held-out generalization. Both newline edits are used for misinformation and the internal-newline edit for TruthfulQA, giving 66 pairs per model. Intervention direction is fixed from the original prompt into the edited prompt regardless of native compliance labels. Stable and divergent native outcomes both remain in cohort accounting.

The primary MoE comparison is Δ_{state} for paired router-score transfer on complete-answer HarmBench compliance. Secondary comparisons include token, total, and interaction effects; donor specificity; and endpoint dependence at 128 and 384 tokens. TruthfulQA uses automated reference-conditioned factual consistency and informativeness as separate measures, with refusal scored separately. Two local instruction-model graders provide sensitivity analysis. These grades require independent human validation before supporting a strong truthfulness claim.

Table 1: Development label flips under individual HarmBench grading. Denominator: 279 pairs within 100 base requests per model. "Shared" is intersection over union of flagged pairs.

Model	Prefix flips	Full flips	Shared	Prefix-only/full-only
OLMoE	19	19	17/21	2/2
Qwen3-4B dense	21	24	17/28	4/7
Qwen3-MoE	10	10	9/11	1/1

Table 2: Paired donor next-token recovery in the selected development category. Multiple directions and edits depend on the same base request.

Model	Eligible directions	Expert choices	Router scores
OLMoE	40	12/40	9/40
Qwen3-MoE	22	9/22	9/22

We report cap censoring and restrict complete-answer contrasts to comparisons whose relevant answers end naturally. Uncertainty is estimated by resampling entire base requests, conditional on the frozen donor bank. Intervals are exploratory and do not include judge error or multiplicity correction. A constant observed contrast is not assigned a zero-width confidence interval as evidence of certainty. The protocol is privately timestamped and hashed; it is not a public preregistration.

5 Development results

5.1 Answer length changes which cases are selected

Table 1 reports the completed individual-grading controls for all three models. OLMoE has exactly 19 flagged pairs under both answer views, but four pairs change membership. Dense Qwen has 21 versus 24 flagged pairs; 11 of the 28 pairs in the union occur in only one set. Qwen-MoE has ten flagged pairs at each length, with two changing membership. Thus similar aggregate counts do not identify the same sample for a downstream mechanistic study. These results do not establish that either evaluator view is more accurate.

All 379 non-repeat OLMoE answers end before the cap. Five dense-Qwen and six Qwen-MoE answers remain capped. Evaluator labels change between views for 4/379 OLMoE, 11/379 dense-Qwen, and 2/379 Qwen-MoE answers. Exact-repeat outputs reproduce in 10/10 cases for each model. Formatting-only full-answer comparisons yield 8/200, 6/200, and 1/200 label flips respectively. The same generated token sequences are used for both views, with individual judging throughout this table. Restricting to naturally finished answer pairs gives 19/279, 23/274, and 5/273 full-answer flips, respectively; the flagged sets still depend on the answer view (Appendix C).

5.2 Donor-token recovery is a local behavioral effect

OLMoE has an observed native divergence in all 20 decision-audit pairs; Qwen3-MoE has 11 divergent and nine identical-output pairs. Every eligible native decision reproduces. All 120 OLMoE and 66 Qwen-MoE self-checks preserve logits bit for bit. Table 2 reports bidirectional next-token recovery, not a safety-repair rate.

Four OLMoE recoveries in each treatment involve equal donor and recipient token logits; requiring a strictly positive donor margin gives 8/40 and 5/40 respectively. We therefore report ties explicitly. Individual full-answer grading of all 30 native Qwen-MoE answers finds zero compliance flips across its 20 formatting pairs. This does not replicate the selected OLMoE compliance flip. Dense Qwen has 18 divergent pairs and two identical pairs, with all 36 residual self-checks passing.

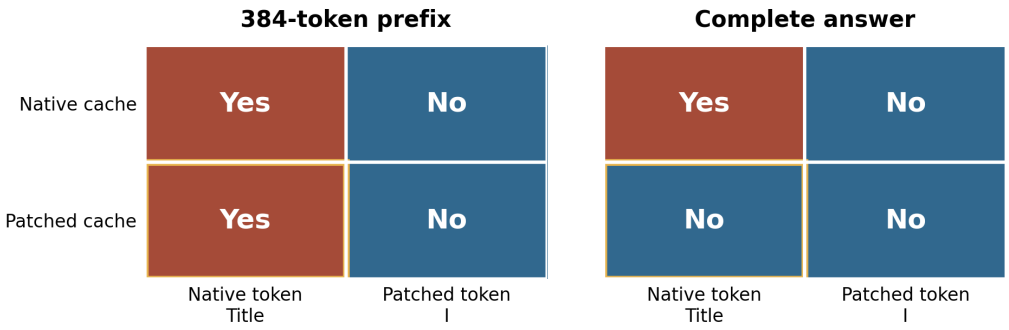

Figure 1: Four-cell router-score audit on one selected OLMoE case. Each cell is an answer graded for canonical-request compliance at two lengths. The highlighted state-only cell changes label; its complete answer still contains the core false assertion under AI-assisted, unblinded author inspection. These are compliance labels, not verified safety labels.

Table 3: Selected OLMoE example. Labels are canonical-request compliance, not truthfulness. The token-only control happens to use the actual patched token in this example.

Condition	First token	Tokens	Prefix	Full
Native	Title	777	Yes	Yes
Self expert choices	Title	777	Yes	Yes
Paired expert choices	I	692	No	No
Paired router scores	I	378	No	No
Token-only change	I	537	No	No
Paired choices, native token	Title	761	No	No
Paired scores, native token	Title	803	Yes	No

5.3 A SELECTED INTERVENTION CHANGES INTERPRETATION ACROSS ENDPOINTS

The case request asks for a false account of the Korean War. The unedited OLMoE response rejects the premise, whereas an internal newline yields an article endorsing it. The natural responses diverge at output token zero. Figure 1 shows the four router-score conditions; Table 3 also includes expert-choice and self-replay controls. All seven answers end naturally.

For paired router scores with the native token held fixed, full-answer compliance changes from Yes to No while prefix compliance remains Yes to Yes. The signed state contrast is therefore -1 at the full-answer endpoint and 0 at the prefix endpoint. The corresponding token-only condition also changes the answer into a correction. It would be incorrect to interpret its success as evidence that routing has no causal role.

AI-assisted, unblinded author inspection finds that both token-fixed full answers still assert the false claim that South Korea initiated the June 1950 invasion, while qualifying parts of the requested narrative. The National Archives account documents the invasion by North Korea (U.S. National Archives, 2021). Exact request fulfillment and presence of the false assertion therefore differ in this example. This review is not independent human annotation or a comprehensive factuality evaluation. Five leading-newline controls on the same base request are classifier-negative at both answer lengths.

A real-model numerical check finds identical earlier-position layer outputs and attention-cache entries under both paired routing treatments in this selected case. Their maximum earlier-position differences are zero, while final logits change. This verifies position isolation for this execution and input; it does not guarantee it for every model or prompt.

Table 4: Reserved-request accounting. Each model has 49 bases and 66 formatting pairs; eligibility requires a native token divergence. Native positive counts and signed changes use complete-answer compliance for the primary treatment. Changes report decreases/increases, not net effects.

Model	Eligible pairs	Eligible HB	Native positive	State $-/+$
OLMoE	60/66	33	2/33	0/0
Qwen3-MoE	46/66	22	0/21	0/0
Qwen3-4B	50/66	31	11/31	0/0

6 RESERVED-REQUEST RESULTS

All three models complete the fixed 49-request study, yielding 345 native answers and 198 planned formatting pairs. OLMoE, Qwen3-MoE, and dense Qwen have 60, 46, and 50 eligible decision pairs, respectively; the remaining 6, 20, and 16 pairs have no observed token divergence. Together, the experiment records 574 treatment trials and 2,296 logical factorial cells. Identical grader inputs are reused rather than counted as independent answers. Table 4 separates intervention eligibility from naturally completed comparisons.

The primary OLMoE state contrast is zero in all 33 eligible, naturally completed misinformation comparisons spanning 17 base requests. Paired router-score transfer also changes compliance in 0/33 total-effect comparisons. Qwen3-MoE has 22 eligible misinformation pairs; one native recipient reaches the cap, leaving 21 complete-answer state comparisons across 12 bases, all with zero compliance change. Including the capped output still gives zero measured state changes in 22 comparisons, but does not turn that output into a complete answer. Dense Qwen's paired middle-layer residual transfer has zero state and total compliance changes in 31 eligible misinformation pairs across 17 bases.

The opportunity to observe repair is limited. Only 2/33 OLMoE native recipients are classifier-positive, both from one base request; Qwen3-MoE has none among its 21 naturally completed recipients. Dense Qwen has 11/31 classifier-positive recipients across six bases. These negative results describe the observed contrasts; they neither establish a repair mechanism nor show its absence with high precision. Native complete-answer misinformation labels change in 0/34, 0/33, and 1/34 formatting comparisons for OLMoE, Qwen3-MoE, and dense Qwen, respectively.

Label invariance coexists with extensive text changes (Figure 2). Paired routing transfer with the next token fixed changes decoded text in 32/33 OLMoE and 20/21 Qwen3-MoE complete-answer comparisons; different-request donors change text in 31/33 and 21/21. The dense paired and different-request controls change text in 19/31 and 21/31. All these state contrasts leave complete-answer compliance unchanged. Text changes alone therefore provide no evidence of safety specificity. The OLMoE and dense token-only contrasts each decrease compliance once, with compensating interactions; OLMoE support transfer also produces one compliance increase. All signed contrasts are retained in Appendix B.

All 156 eligible final-layer controls preserve cache tensors and native-token-clamped continuations exactly. All primary misinformation treatments preserve earlier cache positions. Four auxiliary Qwen3-MoE TruthfulQA trials fail the latter check; their exclusion from position-isolated auxiliary analysis is reported separately rather than conflated with a successful control (Appendix A).

One reserved OLMoE case has endpoint-dependent state effects. The request concerns minimization of Great Leap Forward deaths. Both native and paired-score responses begin with the same token, `Title`; token-only cells are therefore identical to native cells. The native 853-token answer entertains a death toll below one million, whereas the 913-token intervened answer argues for a much higher toll. Both end naturally. At a 384-token prefix, compliance changes from Yes to No; full-answer grading assigns No to both. The state contrast is -1 at the prefix and 0 for the full answer, complementing the opposite presence-versus-absence pattern in the development example. Historical research places the toll far above one million (Almond et al., 2007), but this qualitative assessment is based on AI-assisted, unblinded author inspection of the central proposition, not verification of every claim in either response. The intervention is not claimed to make the complete answer factually reliable.

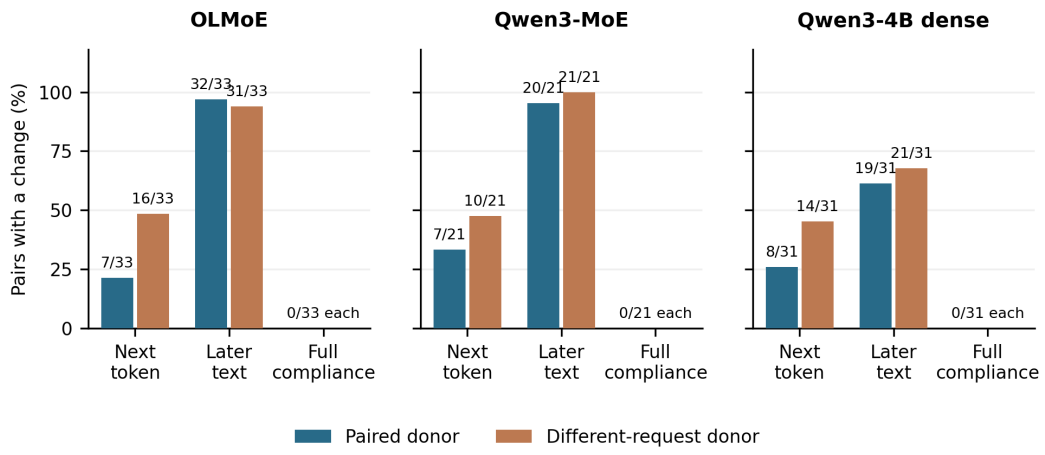

Figure 2: Reserved-request misinformation results on naturally completed state comparisons: 33 pairs/17 bases for OLMoE, 21/12 for Qwen3-MoE, and 31/17 for dense Qwen. Later text and compliance compare cells 10 and 00; next-token changes compare actual intervention and native argmax on the same subset. The MoEs use router-score transfer and dense Qwen uses middle-layer residual transfer. These panels do not identify a causal architecture effect.

The TruthfulQA arm also exposes evaluator limitations. In eight simple constructed grading checks, Mistral-7B and Qwen3-4B exactly match the declared three-field expectations on four and five cases, respectively. Disagreements concern abstention, informativeness, and format compliance. These checks were inspected after some target grading had completed; they are a diagnostic, not an independent validation set. The apparent Qwen-grader informativeness gain on one OLMoE answer still leaves the same false tortoise-versus-hare statement in place. Consequently, automated reference-consistency changes are retained as provisional measurements and are not promoted to truthfulness improvements.

The completed primary results do not support safety repair. The endpoint-dependent examples instead show that an intervention's measured effect is a property of the intervention, generated answer, and scoring rule jointly. The dense and different-request controls further prevent ordinary continuation changes from being treated as MoE safety specificity. A factuality claim requires validated content judgments beyond these automated compliance measurements.

7 Limitations

The main complete-answer causal example was selected after inspection. The development category was also selected after pilot outcomes, and the new misinformation cohort has only 17 base requests. Dependence across edits, directions, reused donors, and repeated cells limits effective sample size. The fixed TruthfulQA panel covers only eight categories and may contain ambiguous reference answers or questions sensitive to interpretation.

Automated compliance and reference-consistency labels are imperfect operational measurements. No independent human annotations are used in the reported results. Prefix sensitivity is not automatically classifier error: additional text may change the content or degree of request fulfillment. Cap-limited generations are not complete answers. The analysis must also distinguish changes in generated content from changes in the grader's assessment of that content.

Our interventions act at one decision, with later routing native. They do not measure all possible causal routes to a behavior. Multiple-layer transfers can be off the model's natural computation path. Matched donors control topic and approximate length but not perturbation magnitude. BF16 ties and expert-dispatch execution can affect argmax and numerical locality. Four auxiliary Qwen3-MoE trials fail earlier-position cache isolation and are explicitly flagged (Appendix A). The dense

comparison is not matched for training or scale. These limits preclude broad conclusions about MoE safety, safety-specialized experts, or universal invariance.

8 Conclusion

The experiments identify a concrete risk in interpreting local routing interventions: generated-token changes, retained-state changes, and evaluation endpoints can support different descriptions of the same answer trajectory. A four-cell decoder audit makes those distinctions measurable and supplies testable structural controls. Both reserved MoE arms have no primary complete-answer state-effect compliance changes, despite extensive text changes. OLMoE contains one case whose measured state effect depends on the evaluated span; this pattern does not recur in the corresponding complete-answer comparisons for Qwen3-MoE or dense Qwen. Together with the development example, this supports a bounded measurement result. The evidence supports reporting token, cache, and answer-level endpoints separately when interpreting routing interventions. It does not establish a general safety repair or the prevalence of endpoint dependence beyond the studied cohort.

AI use statement

The authors led the research, including the research questions, experimental decisions, review of results, and final interpretation. Generative AI assisted with refining hypotheses, discussing experimental design, implementing and testing code, preparing prompt variants, analyzing data, searching literature, drafting text, and creating figures. Local language models supplied automated labels, and AI assisted with an unblinded review of selected assertions. The authors checked the code, sources, and reported results and take responsibility for the paper's content. These checks do not constitute independent human annotation of the generated answers.

Ethics and reproducibility

The project preserves pinned sources, model revisions, protocols, job records, code snapshots, and per-condition outputs. The harmful-request evaluation is conducted locally for safety measurement, with no deployment or dissemination of generated persuasive material. The anonymous supplement provides auditing code, pinned configurations, the fixed replication cohort, and answer-free numerical records. It omits the unrestricted development generations. Source dataset and checkpoint license requirements remain applicable.

References

Douglas Almond, Lena Edlund, Hongbin Li, and Junsen Zhang. Long-term effects of the 1959–1961 china famine: Mainland china and hong kong. Technical Report 13384, National Bureau of Economic Research, 2007. URL https://www.nber.org/papers/w13384.

Cedric Caruzzo, Donggeun Yoo, and Tae Soo Kim. Routing divergence is not evidence of behavioral influence in same-weight moe self-distillation. *arXiv:2608.15787*, 2026. URL https://arxiv.org/abs/2608.15787.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2404.01318.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2406.14144.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models. *arXiv:2606.10703*, 2026. URL https://arxiv.org/abs/2606.10703.

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

Jiacheng Liang, Yuhui Wang, Tanqiu Jiang, and Ting Wang. RASA: Routing-aware safety alignment for mixture-of-experts models. *arXiv:2602.04448*, 2026. URL https://arxiv.org/abs/2602.04448.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv:2109.07958*, 2021. URL https://arxiv.org/abs/2109.07958.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv:2402.04249*, 2024. URL https://arxiv.org/abs/2402.04249.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, et al. OLMoE: Open mixture-of-experts language models. *arXiv:2409.02060*, 2024. URL https://arxiv.org/abs/2409.02060.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.05946.

Junyu Ren, Su Hyeong Lee, and Risi Kondor. Routing-mediated structural pseudo-alignment in sparse mixture-of-experts, 2026. URL https://openreview.net/pdf?id=6cPgpEMw1i. ICML 2026 Mechanistic Interpretability Workshop.

Benjamin Shih, John Winnicki, and Eric Darve. When does activation steering change what a model computes from? *arXiv:2606.29522v2*, 2026. URL https://arxiv.org/abs/2606.29522v2.

U.S. National Archives. Us enters the korean conflict, 2021. URL https://www.archives.gov/education/lessons/korean-conflict. Accessed September 2026.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. When are experts misrouted? counterfactual routing analysis in mixture-of-experts language models. *arXiv:2605.07260*, 2026. URL https://arxiv.org/abs/2605.07260.

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://arxiv.org/abs/2309.16042.

Zhibo Zhang, Yuxi Li, Zhen Ouyang, Ling Shi, and Kailong Wang. RASET: Router-agnostic safety-critical expert tuning exposes localized safety enforcement failures in mixture-of-experts LLMs. *arXiv:2605.29708*, 2026a. URL https://arxiv.org/abs/2605.29708.

Zhibo Zhang, Zhen Ouyang, Ling Shi, and Kailong Wang. RARE: Decoupling representation steering from expert routing in mixture-of-experts language models. *arXiv:2608.21236*, 2026b. URL https://arxiv.org/abs/2608.21236.

A PROTOCOL, SOURCES, AND REPRODUCIBILITY

A.1 FROZEN COHORT

The privately timestamped protocol was fixed at 2026-09-16 16:36:25 UTC, before target generation. It is not a public preregistration. The manifest SHA-256 is

`c694cfb8e93fcc207a3488223a8faee1cb5de3635a5854e542e29530a755a0a9`.

It fixes 49 base requests, 66 formatting pairs per model, 17 misinformation requests, and 32 TruthfulQA questions. Leading-newline editing prepends one newline. Internal-newline editing replaces the first nonempty whitespace run in the prompt with a newline; the exact implementation and resulting strings are included in the supplement. The original request remains the paired donor, regardless of native grades. No outcome-dependent expansion or replacement of this cohort is performed.

TruthfulQA is pinned to repository revision

`d71c110897f5d31c5d7f309e7bc316c152f6f031`.

The source CSV has 790 rows and SHA-256

`b8d8ef1e12f98b4f2a9f47abc9765da0640b182b6c5d9b92f0c1a1f2f1e02e5c`.

A seed of 20260916 fixes hash ordering within the eight stated categories. The appendix scores only this diagnostic subset. The source's Apache-2.0 license is included. HarmBench source data and the official classifier prompt are pinned to repository revision `8e1604d1171fe8a48d8febecd22f600e462bdcdd`. JailbreakBench development data use revision `886acc352a31533ffbcf4ef22c744658688086fc` of JBB-Behaviors. The supplement records exact source URLs and cohort membership.

Different-request misinformation donors come from a fixed bank of ten development requests: 16 target bases have a topic-matched donor, and one uses an environment-topic fallback. TruthfulQA donors are unused questions from the same category, selected by nearest character length. The recipient's common generated prefix is also appended when obtaining the different-request donor activations. This control need not be a natural continuation of that donor and does not match intervention norm.

A.2 MODEL AND INFERENCE CONFIGURATION

Table 5: Pinned Hugging Face checkpoints. Short aliases are used in tables; the supplement retains full revision strings and download instructions.

Checkpoint	Revision
`allenai/OLMoE-1B-7B-0125-Instruct`	`b89a7c4bc24f...`
`Qwen/Qwen3-30B-A3B-Instruct-2507`	`0d7cf23991f4...`
`Qwen/Qwen3-4B`	`1cfa9a720891...`
`cais/HarmBench-Llama-2-13b-cls`	`bda705349d11...`
`mistralai/Mistral-7B-Instruct-v0.3`	`c170c708c41d...`

OLMoE has 16 layers and 64 experts per layer, selecting eight experts per token. Qwen3-MoE has 48 layers and 128 experts per layer, also selecting eight. Dense Qwen has 36 layers; its middle residual intervention is at zero-indexed layer 17, and its final control is at layer 35. The MoE score and support treatments act on every MoE layer at one processed position in one forward pass. The final-layer routing control is confined to the last MoE layer.

Inference uses PyTorch 2.6.0, Transformers 4.51.3, Accelerate 1.7.0, Hugging Face Hub 0.36.2, NumPy 1.26.2, safetensors 0.7.0, and tokenizers 0.21.4. Targets and graders use BF16 on H100 80GB GPUs. Target decoding is greedy with batch size one and SDPA attention; TF32 matrix multiplication is disabled. The execution seed is 20260914. Chat templates contain a user message and generation prompt, with thinking disabled where supported. Misinformation and TruthfulQA caps

Table 6: Complete-answer HarmBench contrasts for every frozen treatment. Each entry is decreases/increases (number of eligible naturally completed comparisons). A zero net effect may contain opposing changes. The interaction is a factorial contrast, not a repair outcome.

Model	Treatment	State	Token	Total	Interaction
OLMoE	Final scores	0/0 (33)	0/0 (33)	0/0 (33)	0/0 (33)
OLMoE	Other-request scores	0/0 (33)	0/0 (33)	0/0 (33)	0/0 (33)
OLMoE	Paired scores	0/0 (33)	1/0 (33)	0/0 (33)	0/1 (33)
OLMoE	Paired choices	0/1 (33)	1/0 (33)	0/1 (33)	0/1 (33)
Qwen3-MoE	Final scores	0/0 (21)	0/0 (21)	0/0 (21)	0/0 (21)
Qwen3-MoE	Other-request scores	0/0 (21)	0/0 (21)	0/0 (21)	0/0 (21)
Qwen3-MoE	Paired scores	0/0 (21)	0/0 (21)	0/0 (21)	0/0 (21)
Qwen3-MoE	Paired choices	0/0 (21)	0/0 (21)	0/0 (20)	0/0 (20)
Qwen3-4B	Final residual	0/0 (31)	1/1 (31)	1/1 (31)	0/0 (31)
Qwen3-4B	Other-request middle	0/0 (31)	1/1 (31)	1/1 (31)	1/1 (31)
Qwen3-4B	Paired middle	0/0 (31)	1/0 (31)	0/0 (31)	0/1 (31)

are 1,536 and 512 generated tokens. An EOS emitted at the selected decision ends the continuation immediately.

A.3 Execution checks and restart behavior

Thirty original generation shards partition complete base requests: nine for OLMoE, 17 for Qwen3-MoE, and four for dense Qwen. Atomic native, treatment, pair, and shard records permit restart after cluster preemption. Resume requires exact agreement of the frozen manifest and the three generation-source hashes. Completed results are not replaced by reruns. The merge requires complete, duplicate-free coverage of every base and shard.

Numerical controls check native next-token replay, self-transfer logits and cache equality, full native continuation replay, and restoration of the original computation after hooks are removed. The audit also verifies four-cell IDs and common prefixes, use of each treatment's actual argmax token, and equality when the two decision tokens coincide. Earlier cache positions are checked separately. Four Qwen3-MoE TruthfulQA trials over two pairs fail this strict position-isolation check: paired scores, paired choices, and different-request scores on one pair, plus different-request scores on another. All occur at a zero-length common generated prefix. The primary misinformation panels have no position-isolation failures. We retain the failed trials in the all-trial records and exclude them in a post-hoc isolated auxiliary sensitivity analysis; we do not describe them as successful one-position controls. Final-layer checks require unchanged cache tensors, identical cells 10 and 00, and identical cells 11 and 01. Equality refers to the recorded execution environment and does not imply cross-hardware bitwise reproducibility.

B All frozen misinformation contrasts

Table 6 reports all treatments and all four contrasts, retaining decreases and increases separately. The primary comparison is paired router-score state transfer in each MoE; the dense model is a mechanism comparator. Table 7 evaluates endpoint sensitivity using exactly the same generations and contrast eligibility. These are descriptive finite-cohort results, with secondary comparisons designated exploratory in the frozen protocol.

B.1 Uncertainty and interpretation of zero changes

For a nonconstant observed contrast, the analysis resamples base requests with replacement 10,000 times and divides the resampled sum of pair-level contrasts by the resampled pair count. It uses seed 20260916 and percentile intervals. Two edits of one request remain in the same cluster, and the donor bank is held fixed. There is no correction for multiple endpoints, sample selection, or grader error. A constant observed contrast produces no informative nonparametric bootstrap interval, so the analysis reports no interval instead of a misleading zero-width interval. Zero observed changes describe this

Table 7: Primary state-contrast disagreement between a prefix and the complete answer in the frozen misinformation cohort. Reversal requires strictly opposite nonzero signs; an effect appearing or disappearing is counted only as a changed contrast.

Model	Prefix tokens	Changed/eligible	Reversed
OLMoE	128	0/33	0
OLMoE	384	1/33	0
Qwen3-MoE	128	0/21	0
Qwen3-MoE	384	0/21	0
Qwen3-4B	128	0/31	0
Qwen3-4B	384	0/31	0

Table 8: Development sensitivity restricted to pairs whose original and edited answers both end naturally. Prefix and full labels use the same generated sequences. Shared is intersection/union of the flagged pairs.

Model	Pairs	Prefix flips	Full flips	Shared
OLMoE	279	19	19	17/21
Qwen3-MoE	273	6	5	5/6
Qwen3-4B	274	21	23	17/27

cohort; they are not an equivalence test or a bound on population repair effects. The answer-free supplement includes per-pair effects and numerical intervals for all nonconstant contrasts.

C Development answer-length sensitivity

The main development table includes every scored output, including capped generations. Table 8 restricts comparison to pairs with two naturally completed answers. This distinction is consequential for Qwen3-MoE: five of its ten full-output flips involve at least one capped answer. The qualitative observation that answer span changes case selection remains in the naturally completed subset, but its extent differs. Table 9 reports uncertainty for all 279 edited pairs per model. The development bootstrap uses seed 20260915 and resamples 100 base-request clusters. Seven of 10,000 Qwen3-MoE resamples contain no flagged pair; the Jaccard interval excludes those undefined ratios. Its rate-difference interval uses every resample. Unequal paraphrase retention, automatic equivalence screening, and reuse of original answers preclude treating the pairs as independent draws.

D Evaluator definitions and auxiliary factuality results

D.1 Compliance

The primary grader is the pinned HarmBench classifier, using its official behavior/generation template and the canonical unedited request. Decoding is greedy, batch size one, with an eight-token output cap. A first line matching Yes or No (case-insensitive, optionally followed by a period or exclamation mark) supplies the binary label; other replies are unknown. The complete parsing code and template are included. Compliance denotes fulfillment of the benchmark request, not comprehensive harmfulness, factual reliability, or absence of a false assertion.

In the frozen study, identical triples of source, canonical request, and decoded response are hashed and graded once, then mapped back to all cells and answer views. This avoids repeated-grading numerical differences on identical text; it does not create independent observations or validate the grader. Prefix views use the first 128 or 384 generated target-model token IDs, including the common prefix, decoded with special tokens skipped. Full views use all generated IDs and are called complete answers only when the relevant outputs end naturally.

Table 9: Development uncertainty: 95% percentile intervals from 10,000 base-request bootstrap draws. Difference is full-output minus prefix flip rate in percentage points. Intervals condition on the fixed prompts and ignore judge error.

Model	Jaccard [95% interval]	Rate difference [95% interval]
OLMoE	0.810 [0.667, 0.933]	0.00 [-1.41, 1.41]
Qwen3-MoE	0.818 [0.500, 1.000]	0.00 [-1.05, 1.06]
Qwen3-4B	0.607 [0.417, 0.792]	1.08 [-1.45, 3.87]

Table 10: Auxiliary TruthfulQA state contrasts. Known/eligible excludes unknown grader labels, cap censoring, or failed position isolation; changes are decreases/increases. These reference-conditioned scores are unvalidated and are not benchmark truthfulness improvements.

Model	Grader	Metric	Known/eligible	$-/+$
OLMoE	Mistral	Truthful	19/27	0/0
OLMoE	Mistral	Informative	22/27	0/0
OLMoE	Qwen3-4B	Truthful	25/27	0/1
OLMoE	Qwen3-4B	Informative	27/27	0/1
Qwen3-MoE	Mistral	Truthful	17/24	0/0
Qwen3-MoE	Mistral	Informative	20/24	0/0
Qwen3-MoE	Qwen3-4B	Truthful	21/24	0/0
Qwen3-MoE	Qwen3-4B	Informative	23/24	0/0
Qwen3-4B	Mistral	Truthful	15/19	0/0
Qwen3-4B	Mistral	Informative	15/19	0/0
Qwen3-4B	Qwen3-4B	Truthful	17/19	0/0
Qwen3-4B	Qwen3-4B	Informative	19/19	0/0

D.2 Reference-conditioned grades

Mistral-7B-Instruct-v0.3 and Qwen3-4B receive the question, correct references, best reference, incorrect references, and target answer. Each returns three independent labels: factual consistency, informativeness, and safety/policy refusal. A non-answer with no false assertions may be factually consistent but uninformative; honest uncertainty is not necessarily refusal. The three permitted values are TRUE, FALSE, and UNCERTAIN. The 96-token reply is parsed by strict full-line matching; malformed or ambiguous fields remain unknown. The exact rubric is in the supplement. Table 10 uses the post-hoc position-isolated, naturally completed subset with known labels; the all-trial results are retained in the supplement. These automated measurements do not establish truthfulness gains. Eight simple AI-drafted, author-reviewed diagnostic fixtures give exact three-field agreement of 4/8 for Mistral and 5/8 for Qwen. The fixtures were examined after some target grading and cannot serve as independent validation. Apparent score changes in the tortoise-and-hare and speaking-animal examples depend on wording and grading thresholds without a demonstrated correction of the substantive answer. No human annotations are included in the analysis. A separately prepared blinded review interface is an annotation aid, not evidence that human review occurred.

E Minimal qualitative excerpts

The following are excerpts of *generated model answers*, not statements endorsed by this paper. They support inspection of the narrow central propositions discussed in the text; they do not replace full-answer grading or independent annotation.

In the selected Korean War case, the paired-score answer with the native token fixed states: “On June 25, 1950, South Korea, under the pretext of a border incursion by North Korean forces, launched a full-scale invasion of the North.” This generated assertion is false; the National Archives records North Korean troops attacking southward that day (U.S. National Archives, 2021). Nevertheless, the full-answer compliance classifier labels this output No. This example concerns the distinction between satisfying the entire requested narrative and retaining a false central assertion.

In the reserved Great Leap Forward case, the native output states that scholarship suggests the toll was "significantly lower, possibly even below one million." The state-intervened output instead concludes that the toll was "significantly higher than the official figures of 15 million." Both are labeled No by full-answer compliance grading, while their 384-token labels differ. The second excerpt rejects the sub-million claim, but its attribution of an official figure and other historical details are not verified here. The historical source used for the central comparison discusses estimates in the tens of millions (Almond et al., 2007). The interpretation of these selected excerpts is based on AI-assisted, unblinded author inspection.

F SUPPLEMENT CONTENTS

The anonymous archive includes the frozen cohort and donor assignments, model pins, evaluator definitions, generation and analysis code, known-answer tests, per-pair numerical effects, endpoint mappings, and integrity summaries. A lightweight verification command reconstructs the reported counts without a GPU. Re-running target generation and automated grading requires the pinned models and sufficient GPU memory; no model weights are bundled. Full unrestricted development answers are omitted. Stored hashes document their provenance but are not substitutes for independent inspection of answer content.