Expert-Aware Causal Tracing of Factual Recall in Sparse MoE Language Models

Anonymous ACL submission

Abstract

Causal tracing of factual recall has been studied predominantly in dense transformer language models, where interventions localize information flow to layers or feed-forward modules. Sparse mixture-of-experts (MoE) language models introduce a sharper question: when a factual prediction is mediated by a routed MoE block, which routed expert contributions matter? We formulate expert-aware causal tracing for sparse MoE language models. Using COUNTERFACT facts, we first corrupt the model's factual preference by adding noise to subject-token embeddings, and then test whether clean MoE-block outputs or clean expert-level updates restore the true-vs-foil logit contrast. For QWEN3-30B-A3B-BASE, a layer sweep selects and validates layer 44, and expert-level tracing identifies L44E069 as an expert repeatedly selected in the clean run whose held-out patch outperforms other active same-layer expert patches. For MIXTRAL-8X7B-V0.1, layer-level tracing validates a mid-layer signal, but the signal is not localized to the selected singleton expert; a coalition check instead recovers it with routed multi-expert updates. These results suggest that MoE factual tracing can be made expert-aware, while also showing that expert-level localization is model- and protocol-dependent rather than universal.

1 Introduction

Factual recall is a common setting for causal analysis of language models: given a prompt such as *The capital of France is*, the model should assign high probability to the correct object, *Paris*. Causal tracing and activation patching study this behavior by corrupting a prediction and then restoring internal components from a clean run to test which components mediate the recovery (Meng et al., 2022; Geva et al., 2023; Vig et al., 2020; Wang et al., 2023). In dense transformers, these interventions often target residual-stream activations, attention outputs, or feed-forward modules.

Sparse mixture-of-experts (MoE) language models make this localization problem more fine-grained. A token is not processed by a single dense feed-forward module, but by a small routed subset of experts (Shazeer et al., 2017; Fedus et al., 2022; Jiang et al., 2024; Yang et al., 2025). Thus a layer-level patch can show that an MoE block matters, but the patched block is an aggregate over routed expert updates. It does not answer which selected expert contributions support the factual prediction. This paper asks: *Can factual-recall causal tracing be made expert-aware in sparse MoE language models?*

We study this question in a controlled diagnostic setting using filtered single-token COUNTERFACT (Meng et al., 2022) contrasts. For each fact, the input is a subject–relation prefix and the object is not included in the prompt. We measure factual preference as a true-vs-foil next-token logit difference, corrupt the run by adding noise to subject-token embeddings, and ask whether clean internal components restore the preference. Our protocol has two stages: first, we patch clean final-position MoE-block outputs into the noised run and sweep MoE layers on discovery cases; second, at the selected layer, we estimate routed expert contributions by ablation difference and patch an expert-level clean-minus-noised update into the noised run. We compare selected experts against other clean-active experts from the same layer.

The main result is asymmetric across two MoE backbones. On QWEN3-30B-A3B-BASE, the layer sweep selects layer 44 and validates it on held-out COUNTERFACT cases; expert-level tracing then selects L44E069, which validates on held-out cases and outperforms other active same-layer experts. On MIXTRAL-8X7B-V0.1, layer-level tracing selects a mid-layer site, layer 19, which also validates, but the selected singleton expert L19E006 underperforms active controls. A coalition check (patching either the clean-routed top-2 experts or

the union of clean- and noised-routed experts) recovers the Mixtral layer-level signal, suggesting that layer-level factual recovery can either concentrate in a recurrent expert or remain distributed across a routed expert coalition.

In summary, we contribute a two-stage MoE causal-tracing protocol and a controlled two-model study. The main empirical insight is that layer-level factual recovery and single-expert localization can come apart: Qwen3 admits a robust single-expert intervention, whereas Mixtral's validated layer-level signal is better explained by routed expert coalitions.

2 Related Work

Factual recall and causal tracing. Language models can express factual associations through cloze-style prompts (Petroni et al., 2019). Prior work links factual knowledge to feed-forward memories, knowledge neurons, and causal tracing interventions (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022; Geva et al., 2023; Wang et al., 2024; Liu et al., 2025). Building on causal tracing and activation patching, we use forward-pass interventions to test whether clean internal components can restore a corrupted true-vs-foil factual preference.

Sparse MoE models and interpretability. Sparse MoE language models route each token to a small subset of experts, increasing parameter count while keeping active computation sparse (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Du et al., 2022; Jiang et al., 2024; Yang et al., 2025). Recent work studies routing patterns, expert specialization, and MoE knowledge attribution (Fayyaz et al., 2026; Bandarkar et al., 2026; Wang et al., 2026; Li et al., 2025; Gu et al., 2026). Because the selected experts and routing weights are input-dependent, we distinguish MoE-block output patching from per-expert contribution patching.

3 Method

3.1 COUNTERFACT factual tracing setup

Following the clean/corrupt intervention paradigm used in causal tracing and activation patching, we ask whether replacing a corrupted internal component with its clean counterpart restores the model's factual preference (Meng et al., 2022; Vig et al., 2020; Wang et al., 2023). We use COUNTERFACT facts introduced by Meng et al. (2022). Each record provides a subject, a relation-specific prompt template, a true object, and a counterfactual target object, which we use as the foil. For a prompt x, true object a_t, and foil object a_f, we define:

$$\Delta(x; a_t, a_f) = \ell_x(a_t) - \ell_x(a_f), \qquad (1)$$

where $\ell_x(a)$ is the next-token logit for object token a. The object token is not included in the input; true and foil objects only define the next-token contrast. For example, if the input prompt prefix is *"Virginia State Route 33 is located in"*; *Virginia* is the true next-token object and *Milan* is the foil.

We apply model-specific filters for single-token true/foil objects and sufficient clean/noise margins; Appendix A gives the exact thresholds and case IDs.

3.2 Subject-noise corrupted run

To construct a corrupted run, we keep the text prompt unchanged but add Gaussian noise to the input embeddings of the subject-token span. If h_s is the embedding of a subject token, we use:

$$h_s^{\text{noised}} = h_s + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \qquad (2)$$

where σ is set to 3.0 times the embedding-matrix standard deviation. The resulting noised prompt has logit difference Δ_{noised}. The subject-noise drop is:

$$\Delta_{\text{clean}} - \Delta_{\text{noised}}. \qquad (3)$$

3.3 MoE-block output patching

For each clean prompt, we cache the final-position output of an MoE block at layer ℓ. This output is a hidden-state update vector produced by the routed MoE sublayer. We then run the noised prompt and replace the corresponding final-position MoE-block output with the clean one. The rescue score is:

$$\text{Rescue}_\ell = \Delta_{\text{patched},\ell} - \Delta_{\text{noised}}. \qquad (4)$$

We define rescue as a task-specific activation-restoration score: the change in the logit-difference metric after patching a clean component into a corrupted run. Positive rescue means that the clean MoE-block output restores part of the true-vs-foil factual preference.

We use a discovery/validation split. Layers are selected only on discovery cases and then evaluated as fixed hypotheses on validation cases.

3.4 Expert-contribution patching

At a selected MoE layer, let $\mathrm{MoEOut}(x)$ denote the final-position MoE-block output. For a clean-active expert e, we estimate its routed update by ablation difference:

$$c_e(x) = \mathrm{MoEOut}(x) - \mathrm{MoEOut}_{\setminus e}(x), \quad (5)$$

where $\mathrm{MoEOut}_{\setminus e}$ is obtained by suppressing expert e's routed contribution under the original top-k routing decision. We do not reroute to a replacement expert nor renormalize the remaining expert weights.

We define the expert patch vector

$$\delta_e = c_e(x_{\mathrm{clean}}) - c_e(x_{\mathrm{noised}}), \quad (6)$$

and add δ_e to the noised MoE-block output at the final position. This asks whether replacing the noised routed expert update with its clean counterpart restores the true-vs-foil factual preference.

Expert selection must avoid two simple confounds. First, an expert can obtain high mean rescue by appearing in only one or two discovery cases. We therefore use recurrence-first selection: on discovery cases, we consider clean-active experts at the selected layer, require an expert to appear in at least a fixed number of discovery cases, and then select the recurrent candidate with highest mean rescue. Second, a selected expert may not be special if other active experts in the same layer would rescue equally well. On validation cases, the selected expert is fixed and compared with same-source active-random experts, i.e., other clean-active experts selected by the router in the uncorrupted run for the same validation prompt and MoE layer.

We report selected-expert rescue and specificity:

$$\mathrm{Spec} = \mathrm{Rescue}(e^\star) - \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathrm{Rescue}(r), \quad (7)$$

where $e^\star$ is the selected expert and $\mathcal{R}$ is the active-random control set. Positive Spec means that the selected expert rescues more than other active same-layer experts for the same prompt.

4 Experiments

4.1 Models

We evaluate two sparse MoE language models. QWEN3-30B-A3B-BASE has 48 MoE layers, 128 experts per MoE layer, and top-8 routing (Yang et al., 2025). MIXTRAL-8X7B-V0.1 has 32 MoE layers, 8 experts per MoE layer, and top-2 routing (Jiang et al., 2024).

4.2 Data and splits

For each model, we shuffle COUNTERFACT records, scan records until collecting 256 usable cases, and split them into 128 discovery and 128 validation cases. The single-token and subject-noise filters are model-specific because tokenizers and clean/noised logits differ across models. The recurrence threshold in Section 3.4 is set to half of the discovery split: a candidate expert must be clean-active in at least 64 of the 128 discovery cases.

4.3 Evaluation

Layer rescue asks whether a clean MoE-block output restores the corrupted true-vs-foil preference. Expert rescue asks the same question for one routed expert update. Specificity Spec asks whether the selected expert rescues more than other clean-active experts from the same source prompt and layer. We report mean effects, positive fraction, bootstrap confidence intervals, and two-sided sign-flip tests. Additional controls and reproducibility details are in Appendices C and A; the supplementary artifact contains scripts, row-level results, hook points, and metadata.

5 Results

Figure 1 summarizes the paper's main empirical pattern: **Qwen3's layer-level signal localizes to a positive selected expert, whereas Mixtral's one selected expert is not specific but routed expert coalitions recover the layer-level signal.**

5.1 Layer-level tracing

Table 1 and the left panel of Figure 1 summarize layer-level validation. Qwen3 strongly validates layer L44. Mixtral validates L19, but nearby layers are comparable, so we interpret it as mid-layer transfer rather than sharp single-layer localization (Appendix B).

5.2 Expert-level tracing

Table 1 and the middle panel of Figure 1 show selected-expert tracing. On Qwen3, the discovery rule selects L44E069, which validates with positive specificity. **L44E069 is not merely an arbitrary active layer-44 expert: its patch vector δ_e rescues more than active-random controls.** Additional Qwen3 controls preserve this conclusion (Appendices C–E).

Model	Layer	Layer rescue	Expert	Expert rescue	Spec
Qwen3	L44	$+0.901$ $[+0.752, +1.053]$	L44E069	$+0.463$ $[+0.344, +0.590]$	$+0.400$ $[+0.276, +0.533]$
Mixtral	L19	$+0.457$ $[+0.331, +0.579]$	L19E006	$+0.099$ $[+0.018, +0.175]$	-0.175 $[-0.284, -0.072]$

Table 1: Main validation results on 128 held-out cases per model. Spec subtracts same-layer active-random rescue.

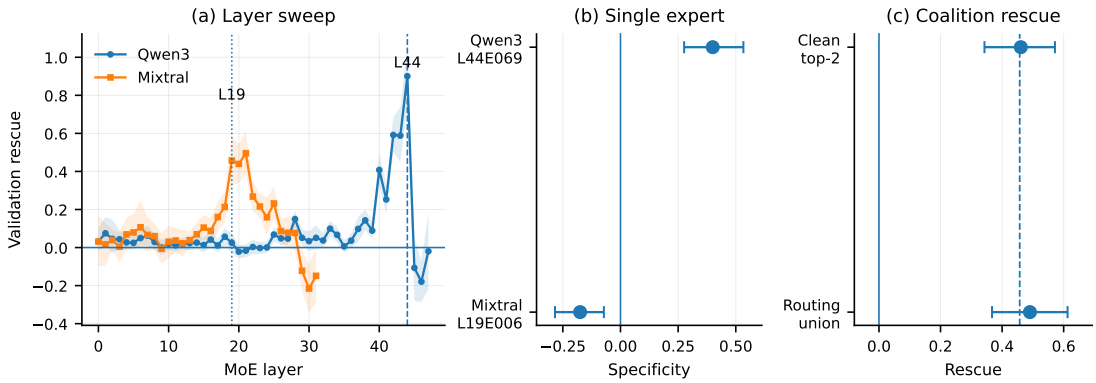

Figure 1: Main validation pattern. Left: MoE-block rescue across layers. Middle: selected-expert specificity. Right: Mixtral coalition patching; the dashed line marks L19 MoE-block rescue. Qwen3 localizes to a positive selected expert, whereas Mixtral recovers the layer signal through routed expert coalitions.

On Mixtral, the selected L19E006 expert has negative specificity, but clean top-2 and clean/noised routing-union coalitions recover rescue close to the L19 MoE-block effect (right panel of Figure 1; Appendix G). **Mixtral therefore supports layer-level patching and multi-expert routed recovery, but not robust single-expert localization under this COUNTERFACT protocol.**

6 Discussion

Block-level rescue is not expert-level evidence. The main methodological lesson is that MoE factual tracing should not stop at layer-level activation patching. A clean MoE-block output can rescue a noised factual prompt, but this only localizes the signal to an aggregate routed block. Since the block output is composed of selected expert updates, expert-aware tracing asks which routed contributions support the recovery. Router selection identifies candidates, but selection alone is not a causal claim; causal evidence comes from patching the estimated expert update and comparing it with active controls.

Mixtral exemplifies the limits of single-expert localization. In Qwen3, the selected layer and selected expert both validate, with multiple controls supporting L44E069. In Mixtral, layer-level tracing transfers to a mid-layer band, but the selected expert underperforms active controls. A coalition check suggests why: clean top-2 and clean/noised routing-union expert coalitions recover rescue close to the L19 MoE-block effect (Appendix G). We treat these findings as model-specific, since Qwen3 and Mixtral differ in tokenizer, training, routing design, and retained cases.

The scope is a controlled diagnostic. We do not claim a universal factual expert or cross-model expert homology. Because tokenizer and subject-noise filters are model-specific, we treat the Qwen3 and Mixtral results as two controlled cases rather than an architecture-level comparison. The relaxed-filter scale-up preserves the same qualitative pattern, but the setting remains a filtered, single-token, final-position diagnostic.

7 Conclusion

We introduced expert-aware causal tracing for sparse MoE factual-recall diagnostics. Qwen3 shows a layer-level site that can be further localized to a robust routed expert contribution, L44E069. Mixtral shows a different pattern: layer-level rescue transfers, but the signal is better recovered by routed expert coalitions than by the selected single expert. These results suggest that MoE factual tracing should distinguish aggregate MoE-block rescue, single-expert localization, and multi-expert routed recovery.

Ethical Considerations

This work uses public model checkpoints and COUNTERFACT-style factual diagnostics, and does not involve human subjects or private data. The method may support beneficial model debugging and factual-behavior analysis. A potential dual-use risk is that more precise localization of factual behavior could support targeted manipulation, suppression, or editing of model outputs. Such uses should be evaluated carefully before deployment, especially for factual associations involving people, groups, languages, or regions. The submitted artifact contains diagnostic scripts and derived row-level results, but no model weights or deployment tool.

Use of AI assistance. We used GPT-5.5 (OpenAI, 2026) as a coding and limited writing assistant. The authors checked the analyses, verified results, and are responsible for all content.

Limitations

This study is a controlled diagnostic rather than a full account of natural factual recall. The main protocol uses 256 model-specific single-token true/foil COUNTERFACT contrasts and final-position interventions. Although the relaxed-filter 512-case check preserves the main pattern, the filtering funnel may still bias retained cases toward relations and objects with clean single-token margins and subject-noise sensitivity. Expert contributions are estimated by ablation difference under fixed routing and patched at the final prediction position; this does not establish a complete subject-to-answer circuit. Our two-stage design decomposes the selected MoE layer rather than exhaustively searching all layer-expert pairs, so useful experts outside the selected layer may be missed. Qwen3 and Mixtral differ in architecture, routing, tokenizer, and training, so the model contrast should be treated as two controlled case studies rather than a broad survey of MoE language models. Future work should extend to more datasets, more MoE backbones, multi-token objects, subject/relation-position interventions, and systematic coalition analyses across layers and models. Reproducing the full experiments requires large sparse MoE checkpoints and implementation-specific hooks; the artifact provides scripts and row-level outputs, but the compute requirement remains nontrivial.

References

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. 2026. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2022. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan A. Rossi, Trung Bui, Hinrich Schuetze, and Nanyun Peng. 2026. Steering moe LLMs via expert (de)activation. In *The Fourteenth International Conference on Learning Representations*.

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

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yupu Gu, Rongzhe Wei, Andy Zhu, and Pan Li. 2026. MoEEdit: Efficient and routing-stable knowledge editing for mixture-of-experts LLMs. In *The Fourteenth International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las

Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Junzhuo Li, Bo Wang, Xiuze Zhou, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. Decoding knowledge attribution in mixture-of-experts: A framework of basic-refinement collaboration and efficiency analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22431–22446, Vienna, Austria. Association for Computational Linguistics.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schuetze. 2025. On relation-specific neurons in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 992–1022, Suzhou, China. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

OpenAI. 2026. Introducing gpt-5.5. `https://openai.com/index/introducing-gpt-5-5/`.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. 2026. The myth of expert specialization in moes: Why routing reflects geometry, not necessarily domain expertise. *Preprint*, arXiv:2604.09780.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Reproducibility Details

This appendix lists protocol details and exact COUNTERFACT case IDs used in the 256-case experiments. Table 2 summarizes filtering, splitting, noise, and statistical settings. Table 3 reports selected-expert activity and zero-rescue rows. Table 4 summarizes retained relation counts, and Table 8 lists the exact model-specific COUNTERFACT case IDs.

Item	Setting
Record order	COUNTERFACT records shuffled with seed 0.
Cases	256 per model; 128 discovery and 128 validation.
Filtering	Single-token true/foil objects; clean margin ≥ 1.0; subject-noise drop ≥ 0.5.
Subject noise	One Gaussian draw per case; seed $0 + \texttt{case_id}$; scale 3.0 times embedding-matrix std.
Layer selection	MoE layers swept on discovery cases; selected layer fixed on validation cases.
Expert selection	Candidate clean-active in at least 64 discovery cases; highest mean rescue selected.
Active-random	Qwen3: 3 controls per case; Mixtral: 1 control per case. Controls are same-source, same-layer, clean-active experts excluding the selected expert.
Statistics	5,000 bootstrap resamples; 10,000 two-sided sign-flip samples.

Table 2: Reproducibility protocol for the 256-case COUNTERFACT experiments.

When the selected expert is not clean-active for a case, the ablation-difference contribution is zero under the fixed-routing intervention, and the case remains in the all-case validation average. Anchor-active rows are recorded in the artifact, while the

Model	Expert	Disc.	Val.	Zero
Qwen3	L44E069	112/128	116/128	26
Mixtral	L19E006	91/128	83/128	54

Table 3: Selected-expert activity and zero-rescue rows in expert-level validation. Disc. and Val. count discovery and validation cases where the selected expert has a nonzero clean contribution under the original routing decision. Zero rows are retained in the all-case expert rescue and specificity summaries.

main tables keep all validation cases to reflect both effect size and coverage.

Model	Most frequent retained relations
Qwen3	P17:21, P103:16, P136:14, P178:13, P27:13, P131:12, P159:12, P30:12, P20:11, P176:11, P740:9, P364:9
Mixtral	P17:28, P103:26, P27:16, P178:15, P495:15, P136:15, P740:15, P106:14, P1412:14, P131:11, P20:10, P364:9

Table 4: Most frequent retained COUNTERFACT relations after model-specific filtering. Counts are over the 256 usable cases for each model.

B Result Breakdowns

This appendix provides additional validation summaries that support the main results. Table 5 reports relation-wise expert-level effects for frequent retained relations. Table 6 quantifies the layer-sweep sharpness, and Table 7 relates selected-expert rescue to selected-layer MoE-block rescue.

C Matched and Equal-Norm Expert Controls

C.1 Qwen3 Gate-Weight-Matched and Equal-Norm Control

To test whether the Qwen3 L44E069 result is driven by router weight or patch-vector magnitude, we run a stricter validation-only control. Let $e^\star$ be L44E069 and let $A(x)$ be the set of other clean-active layer-44 experts for the same source prompt. For each validation case where $e^\star$ is clean-active, we choose the gate-weight-matched control

$$r^\star = \arg\min_{r \in A(x)} \left| w_r(x_{\text{clean}}) - w_{e^\star}(x_{\text{clean}}) \right|. \quad (8)$$

The raw gate-matched specificity is

$$\text{Spec}_{\text{gate}} = \text{Rescue}(e^\star) - \text{Rescue}(r^\star). \quad (9)$$

For the equal-norm version, we scale the selected and control patch vectors to the smaller original norm:

$$\tilde{\delta}_e = \delta_e \frac{\min(\|\delta_{e^\star}\|, \|\delta_{r^\star}\|)}{\|\delta_e\|}, \quad (10)$$

and compute the same specificity contrast after patching $\tilde{\delta}_{e^\star}$ and $\tilde{\delta}_{r^\star}$. This control is evaluated on 116 of 128 Qwen validation cases where L44E069 is clean-active and a valid same-layer active control exists. Table 9 reports the results.

For the all-active comparison, let $A(x)$ be the clean-active layer-44 expert set and $e^\star = \text{L44E069}$. We define

$$A^-(x) = A(x) \setminus \{e^\star\}, \quad (11)$$

$$\bar{R}_{A^-}(x) = \frac{1}{|A^-(x)|} \sum_{r \in A^-(x)} \text{Rescue}(r), \quad (12)$$

$$\text{Spec}_{\text{all}}(e^\star) = \text{Rescue}(e^\star) - \bar{R}_{A^-}(x). \quad (13)$$

Table 10 additionally compares L44E069 against all clean-active layer-44 experts in each validation case where an all-active rank comparison is available.

C.2 Mixtral Top-2 Active-Pair Equal-Norm Check

For Mixtral, top-2 routing makes the active-random control the unique other clean-active expert whenever the selected expert is clean-active. Therefore gate-weight matching is degenerate: the gate-matched control and active-random control are the same expert. We instead use an equal-norm active-pair check, scaling the L19E006 patch vector and the other active expert's patch vector to the smaller original norm before patching. Table 11 reports this check.

D Qwen3 Expert-Selection Stability

We test whether the Qwen3 selected expert is sensitive to the discovery/validation split or recurrence threshold. Reusing the 256 usable Qwen3 cases, we vary the split seed over $\{0, 1, 2, 3, 4\}$ and the recurrence threshold over $\{32, 48, 64, 80, 96\}$. Across all 25 settings, L44E069 is selected every time. Averaged over split seeds and thresholds, validation rescue is $+0.398$ and specificity is $+0.344$.

Relation-held-out selection. We also test whether L44E069 is selected when entire relation types are held out from expert selection. We split retained Qwen3 relations into five folds, select the recurrent layer-44 expert using only non-held-out relations, and evaluate on the held-out relations. L44E069 is selected in all five folds. Table 12 reports the held-out rescue and specificity.

Model	Relation	n	Expert rescue	Specificity	Pos. frac.
Qwen3	P103 native lang.	12	$+0.427$	$+0.406$	0.83
Qwen3	P27 citizenship	9	$+0.826$	$+0.780$	1.00
Qwen3	P136 genre	7	$+0.018$	-0.000	0.57
Qwen3	P131 admin. location	6	$+1.240$	$+1.160$	0.83
Qwen3	P17 country	6	$+1.073$	$+0.910$	0.67
Qwen3	P30 continent	6	$+0.292$	$+0.312$	0.50
Qwen3	P1412 language used	6	$+0.198$	$+0.163$	0.67
Qwen3	P413 position	6	$+0.042$	-0.174	0.17
Qwen3	P937 work location	5	$+1.387$	$+1.350$	1.00
Qwen3	P159 headquarters	5	$+1.012$	$+0.988$	0.80
Mixtral	P103 native lang.	12	$+0.312$	$+0.206$	0.67
Mixtral	P17 country	11	$+0.187$	-0.272	0.09
Mixtral	P1412 language used	9	$+0.031$	-0.076	0.33
Mixtral	P27 citizenship	9	$+0.049$	-0.132	0.22
Mixtral	P740 formation loc.	9	$+0.146$	-0.167	0.56
Mixtral	P495 origin country	9	$+0.094$	-0.358	0.11
Mixtral	P178 developer	9	$+0.000$	-0.368	0.00
Mixtral	P136 genre	9	$+0.111$	-0.562	0.11
Mixtral	P106 occupation	7	$+0.223$	$+0.179$	0.71
Mixtral	P131 admin. location	6	-0.016	-0.557	0.00

Table 5: Relation-wise expert-level validation summary for frequent retained relations. Specificity is selected-expert rescue minus active-random rescue. Relation counts are model-specific after filtering.

Model	Top layer	Top rescue	Next layer	Gap
Qwen3	L44	$+0.901$	L42 $+0.592$	$+0.309$
Mixtral	L21	$+0.496$	L19 $+0.457$	$+0.038$

Table 6: Layer-sweep sharpness on validation cases. Gap is the difference between the top validation rescue and the next-highest layer. Mixtral's smaller gap reflects a broader mid-layer band.

Model	Quantity	Value
Qwen3	Layer rescue	$+0.901$
Qwen3	Expert/layer	$+0.515$ $[+0.411, +0.613]$
Qwen3	Spec/layer	$+0.444$ $[+0.323, +0.560]$
Mixtral	Layer rescue	$+0.457$
Mixtral	Expert/layer	$+0.216$ $[+0.044, +0.388]$
Mixtral	Spec/layer	-0.383 $[-0.645, -0.159]$

Table 7: Expert-level rescue relative to selected-layer MoE-block rescue. Ratios are computed as ratios of validation-case means, with bootstrap confidence intervals. Spec/layer compares specificity Spec to selected-layer rescue.

E Noise-Scale Sensitivity

We run a fixed-hypothesis sensitivity check for the Qwen3 result by keeping the selected layer and expert fixed to L44 and L44E069, and varying only the subject-noise scale σ. This check does not reselect layers or experts; it asks whether the fixed L44E069 intervention remains positive under different corruption strengths.

Table 13 summarizes the fixed-hypothesis results.

F Relaxed-Filter Scale-Up

We run an additional relaxed-filter scale-up to test whether the main pattern depends on the stricter 256-case filtering regime. The clean-margin threshold is lowered from 1.0 to 0.5, and the subject-noise-drop threshold is lowered from 0.5 to 0.25, yielding 512 usable cases per model. We use the same discovery/validation protocol and the same half-discovery recurrence threshold.

Only 15 of 256 relaxed validation cases overlap with the main strict set. Table 15 shows that the relaxed-new validation cases preserve the same Qwen-positive/Mixtral-boundary pattern.

G Mixtral Multi-Expert Coalition Check

To test whether Mixtral's positive layer-level rescue is distributed across multiple routed experts, we compare two fixed layer-19 coalition interventions on the main validation cases: the clean-routed top-2 expert coalition, and the union of clean-routed and noised-routed experts. The union patch accounts for cases where subject noise changes which experts are routed. The main single-expert result is reported in Table 1; this appendix focuses on whether multi-expert routed updates recover the layer-level signal.

Let $S_c(x)$ be the experts selected by the router in the clean run at layer 19, and let $S_n(x)$ be the experts selected at the same position in the noised run. Recall that δ_e denotes expert e's clean-minus-noised patch vector. We define the coalition patches as

Model	Split	COUNTERFACT case IDs
Qwen3	Discovery	3799, 5203, 10274, 14979, 4868, 7082, 4085, 7001, 18003, 17311, 6506, 12914, 1642, 3371, 17169, 8838, 17226, 14789, 3211, 1915, 20598, 21889, 8097, 3469, 8937, 19616, 8894, 3468, 21149, 7805, 5922, 13909, 14387, 4690, 14407, 6992, 4381, 18926, 9007, 19178, 13953, 21063, 943, 4852, 15369, 5703, 6103, 20407, 2377, 20936, 12329, 1068, 3722, 12143, 6002, 4145, 9128, 19429, 988, 12946, 5255, 9949, 12405, 16988, 3078, 18255, 6239, 1750, 16086, 4195, 18848, 20764, 18677, 9859, 2331, 17020, 11330, 11661, 14910, 6615, 15773, 21412, 12398, 799, 10177, 17562, 15598, 7289, 4380, 12046, 8024, 6022, 18704, 11630, 6564, 21076, 4965, 17549, 12644, 3521, 4730, 13254, 14176, 10618, 18365, 4138, 10711, 4988, 3774, 17370, 4233, 14875, 2900, 5917, 15198, 4626, 12462, 6833, 18202, 7603, 7106, 14644, 7937, 2572, 11450, 17390, 228, 20312
Qwen3	Validation	11072, 636, 3463, 19770, 16489, 1464, 7692, 5524, 12619, 8298, 10862, 21830, 13915, 21698, 3858, 8808, 11885, 15209, 6804, 11327, 20397, 8209, 3703, 734, 20471, 7484, 14817, 17735, 633, 11111, 15024, 10963, 4127, 2269, 20189, 7953, 7997, 6655, 19422, 1759, 19146, 16490, 12224, 12106, 1707, 934, 536, 21618, 8007, 2500, 13935, 8143, 7420, 6997, 14704, 494, 20924, 5091, 20762, 19242, 7255, 1062, 3161, 11405, 14059, 7178, 9544, 709, 9811, 12014, 7515, 12783, 3492, 2894, 20702, 11, 13643, 13332, 9983, 9828, 6494, 12593, 21751, 11552, 13814, 21232, 15277, 21036, 12008, 19277, 11678, 20207, 8830, 5736, 3680, 5451, 9599, 9739, 10321, 7860, 7544, 4176, 13739, 14584, 17885, 2320, 7037, 14427, 1726, 17688, 6869, 2582, 15166, 9329, 4046, 17654, 19607, 9312, 21305, 16381, 19969, 15480, 14898, 6467, 15960, 8339, 4072, 11360
Mixtral	Discovery	4626, 7953, 1750, 3133, 2894, 17521, 6076, 15715, 307, 12329, 14407, 10711, 8362, 3211, 12062, 20471, 5825, 21592, 19189, 10139, 6840, 6022, 3550, 9891, 11298, 14817, 988, 9739, 13643, 9828, 18704, 14644, 18645, 15198, 17790, 21412, 1915, 422, 1664, 19242, 20936, 11330, 21258, 14387, 20270, 3243, 12914, 8360, 12530, 1792, 20575, 6564, 21243, 2810, 15519, 6103, 8097, 6464, 4852, 10097, 14875, 4145, 5553, 11072, 1707, 20762, 5331, 10180, 3078, 17078, 4366, 6142, 5130, 11073, 10530, 21457, 3371, 10757, 3463, 10618, 6002, 11981, 12014, 1930, 17735, 7259, 7152, 15369, 4381, 4401, 241, 11933, 17592, 15328, 9690, 6833, 4065, 6467, 18188, 12783, 15995, 19146, 31, 12398, 483, 5203, 4928, 19422, 17370, 9312, 13909, 15983, 2377, 8894, 6997, 7128, 15422, 10258, 4127, 20598, 3848, 7289, 10806, 13739, 16381, 9077, 7787, 20137
Mixtral	Validation	6515, 15598, 8024, 18874, 3167, 4380, 17510, 15960, 11360, 21076, 2052, 11678, 1642, 10950, 10851, 18677, 16732, 15314, 494, 16722, 13043, 5736, 7172, 18920, 13814, 4055, 10963, 20042, 6498, 14541, 636, 8937, 21618, 16403, 16717, 12124, 14320, 21273, 12440, 21889, 17409, 16625, 1068, 9019, 15267, 5004, 5703, 19074, 12008, 9599, 21063, 17565, 15087, 2331, 12593, 8830, 13354, 9007, 13106, 17169, 12619, 7997, 3026, 11405, 15480, 16844, 20326, 17663, 14726, 13935, 3722, 19004, 11188, 12802, 14789, 10862, 4972, 19429, 11062, 19770, 20312, 1062, 12946, 13332, 6734, 17654, 4312, 10274, 11361, 12169, 12224, 5325, 19178, 19918, 11016, 256, 1695, 13349, 6655, 10177, 3248, 16490, 7082, 1425, 14705, 8007, 5922, 12143, 3521, 11643, 13768, 2269, 5255, 9544, 2900, 10964, 4311, 7970, 18611, 20207, 14584, 14419, 8298, 4988, 6506, 10211, 16555, 6332

Table 8: Exact COUNTERFACT case IDs used in the 256-case experiments. Cases are model-specific because tokenizer and subject-noise filters are model-specific.

Quantity	Raw	Equal-norm
Selected rescue	$+0.513$ $[+0.383, +0.651]$	$+0.239$ $[+0.177, +0.302]$
Matched control	$+0.054$ $[+0.017, +0.091]$	$+0.051$ $[+0.016, +0.086]$
Specificity	$+0.459$ $[+0.322, +0.607]$	$+0.188$ $[+0.119, +0.259]$

Table 9: Qwen3 L44E069 gate-weight-matched and equal-norm control on 116 validation cases. The matched control is the clean-active layer-44 expert with closest clean router weight. Equal-norm rows scale both patch vectors to the smaller original norm before patching.

Metric	Value
Ranked validation cases	115
Top-1 among active experts	53 / 115
Top-2 among active experts	74 / 115
Mean rank	2.90
Mean percentile	0.73
Selected minus all-other active	$+0.441$ $[+0.309, +0.584]$

Table 10: Qwen3 L44E069 rescue rank among all clean-active layer-44 experts on validation cases with an available all-active rank comparison. Percentile is scaled so that the best active expert in a case has percentile 1.0 and the worst has 0.0.

$$D_{\text{top2}}(x) = \sum_{e \in S_c(x)} \delta_e, \tag{14}$$

$$D_{\text{union}}(x) = \sum_{e \in S_c(x) \cup S_n(x)} \delta_e. \tag{15}$$

The first patch restores the clean-routed expert coalition; the second also accounts for experts that enter or leave the routed set under subject noise. Table 16 reports the resulting coalition-patching rescue values.

Metric	n	Mean	95% CI
Raw active-pair specificity	83	-0.081	$[-0.214, +0.042]$
Selected equal-norm rescue	83	$+0.046$	$[-0.007, +0.096]$
Other active expert equal-norm rescue	83	$+0.108$	$[+0.047, +0.172]$
Equal-norm active-pair specificity	83	-0.062	$[-0.130, -0.003]$

Table 11: Mixtral L19E006 active-pair equal-norm check on anchor-active validation cases. Specificity is L19E006 rescue minus the rescue of the unique other clean-active layer-19 expert. The selected expert remains below the other active expert after equal-norm scaling.

Metric	Value
Relation-held-out folds	5
L44E069 selected	5 / 5
Held-out cases	256
Active cases	229 / 256
Held-out rescue	$+0.443$ $[+0.343, +0.545]$
Held-out specificity	$+0.388$ $[+0.297, +0.484]$

Table 12: Qwen3 relation-held-out expert-selection check. Relations are split into five folds; experts are selected using only non-held-out relations and evaluated on held-out relations. Specificity subtracts the mean rescue of all other clean-active layer-44 experts in the held-out case.

σ	Active	Drop	Rescue	Pos.
1.0	115/128	$+1.259$	$+0.098$ $[+0.040, +0.165]$	0.43
2.0	115/128	$+4.909$	$+0.406$ $[+0.280, +0.546]$	0.55
3.0	115/128	$+5.697$	$+0.459$ $[+0.339, +0.587]$	0.66
4.0	115/128	$+6.004$	$+0.478$ $[+0.356, +0.605]$	0.67

Table 13: Qwen3 fixed-hypothesis noise-scale sensitivity for L44E069 on validation cases. Drop is the mean clean-minus-noised logit-difference drop; Rescue is mean L44E069 rescue with 95% bootstrap confidence intervals; Pos. is the fraction of validation cases with positive rescue.

Model	Layer rescue	Expert	Disc. active	Expert rescue	Spec
Qwen3	$+0.846$ $[+0.760, +0.934]$	L44E069	230/256	$+0.454$ $[+0.367, +0.545]$	$+0.413$ $[+0.318, +0.511]$
Mixtral	$+0.421$ $[+0.352, +0.489]$	L19E006	174/256	$+0.114$ $[+0.041, +0.187]$	-0.155 $[-0.247, -0.067]$

Table 14: Relaxed-filter scale-up to 512 usable cases per model. The clean-margin threshold is lowered to 0.5 and the subject-noise-drop threshold to 0.25. Layers and experts are selected using the same discovery/validation protocol and half-discovery recurrence threshold.

Model	Subset	n_L	Layer	n_E	Expert	Spec
Qwen3	All relaxed val.	256	$+0.833$ $[+0.714, +0.962]$	137	$+0.400$ $[+0.293, +0.516]$	$+0.360$ $[+0.239, +0.487]$
Qwen3	Relaxed-new val.	241	$+0.860$ $[+0.739, +0.989]$	130	$+0.420$ $[+0.310, +0.541]$	$+0.380$ $[+0.259, +0.509]$
Mixtral	All relaxed val.	256	$+0.454$ $[+0.360, +0.557]$	137	$+0.119$ $[-0.003, +0.240]$	-0.124 $[-0.270, +0.014]$
Mixtral	Relaxed-new val.	241	$+0.459$ $[+0.360, +0.561]$	130	$+0.123$ $[-0.010, +0.250]$	-0.128 $[-0.280, +0.014]$

Table 15: Relaxed-filter validation split by overlap with the main strict 256-case set. "Relaxed-new" cases appear in the relaxed-filter validation split but not in the main strict set. n_L is the number of cases for layer-level summaries; n_E is the number of expert-level contrast rows available for expert rescue and specificity.

Patch	Rescue	Pos. frac.
Clean top-2 coalition	$+0.461$ $[+0.343, +0.572]$	0.75
Routing-union coalition	$+0.490$ $[+0.367, +0.613]$	0.75

Table 16: Mixtral L19 multi-expert coalition patching on validation cases. The clean top-2 coalition sums patch vectors for the clean-routed experts; the routing-union coalition sums patch vectors over the union of clean-routed and noised-routed experts.