How Torn Is a Trained Mixture-of-Experts?
A Geometric Diagnostic for Routing Discontinuity in Released Weights

Zhuo Zhang
Independent Researcher

Abstract

Mixture-of-Experts (MoE) layers route each token through a discrete top-k choice, making the layer-to-layer transport map C^0-discontinuous: a small nudge across a routing boundary can flip a token's experts and jump the block output by an $O(1)$ amount—a tear. We give a geometric diagnostic for this tear on released MoE weights—a difference-quotient continuity signature whose divergence rate certifies the order of the singularity, a $k/(k+1)$ expert cliff with a cosine decomposition, and a distance-to-tear—and find a genuine C^0 jump at every layer: the hard-routing difference quotient diverges under grid refinement at scaling exponent ≈ 1 (the order-0-jump signature; its raw refinement growth equals the grid ratio by construction, not a fitted severity), while continuity-guaranteed controls stay flat at exponent ≈ 0—ubiquitous, layer-consistent, and reproduced across OLMoE-1B-7B and Qwen1.5-MoE-A2.7B. The tear is a directional inference liability: random perturbations are blind to it, but boundary-normal perturbations of $< 1\%$ relative magnitude flip experts and produce an $O(0.1)$ per-block output jump ($\approx 24\%$ of the block-output norm), and continuous re-gating removes it only at a steep quality cost (15.9× perplexity, all layers). Training-time it decomposes rather than detonates: a two-seed tear-magnitude dial shows no divergence, and a SwiGLU-clamp sweep shows amplitude-capping remedies leave the discontinuity's topology intact while reducing its absolute amplitude.

1 Introduction

MoE layers route each token to a small top-k subset of experts. This discrete choice makes the layer-to-layer transport map C^0-discontinuous: two hidden states arbitrarily close but on opposite sides of a routing boundary go to different experts, so the block output can jump by an $O(1)$ amount. We call this jump a tear, using the Deep Manifold vocabulary as motivation rather than as a load-bearing premise [Ma and Shi, 2024, 2025].

Prior work approaches routing discontinuity from two sides: continuous-routing fixes that let experts enter or leave at zero weight [Wang et al., 2025, Puigcerver et al., 2024, Martins and Astudillo, 2016, Peters et al., 2019], and routing-stability measurements of how often the selected expert set changes [Ma et al., 2025]. Neither measures the quantity that governs the transport map on real models: the output-jump geometry on released LLM-MoE weights.

Our contributions are: (1) a geometric diagnostic for the tear on released weights, with negative controls and known-answer tests; (2) a characterization across OLMoE and Qwen1.5-MoE showing the tear is a genuine order-0 jump (scaling exponent ≈ 1, controls flat at ≈ 0) at every layer—ubiquitous, layer-consistent, and not model-specific—with severity carried by the per-block output jump and the expert cliff; (3) a directional inference-robustness result—the tear is invisible to random perturbations but exposed by boundary-normal ones—and a mitigation bound; and (4) an honest training-time decomposition showing roughness and rare severe-but-recovered spikes without divergence at the tested scale.

2 Related Work

Continuous routing. ReMoE explicitly frames top-k as a jump discontinuity and replaces it with ReLU routing [Wang et al., 2025]. Soft MoE, sparsemax, and entmax provide related continuous or sparse transformations [Puigcerver et al., 2024, Martins and Astudillo, 2016, Peters et al., 2019]. We use a soft-edge gate as a measurement probe, not as a proposed method.

MoE robustness and routing stability. Puigcerver et al. [2022] state MoE non-continuity and give an expert-difference condition on the boundary, which is the ancestor of our expert-cliff measurement. Their experiments focus on adversarial robustness in vision MoEs; we measure per-layer output jumps in released LLM-MoE weights. R3 measures how often routing choices change between training and inference engines [Ma et al., 2025]; we instead measure the output jump those changes induce.

Piecewise-linear networks and training stability. Spline and linear-region analyses provide the C^0/C^1 vocabulary but not our MoE diagnostic [Hanin and Rolnick, 2019, Balestriero and Baraniuk, 2018, Montúfar et al., 2014]. Switch, ST-MoE, and GShard study MoE stability through numerical and load-balancing mechanisms [Fedus et al., 2022, Zoph et al., 2022, Lepikhin et al., 2021]. DeepSeek-V4 reports that spikes are tied to outliers in MoE layers and that routing appears to exacerbate them, then introduces SwiGLU clamping, Anticipatory Routing, and manifold-constrained hyper-connections [DeepSeek-AI, 2026]; Section 8 gives a geometric reading of those interventions.

3 Preliminaries

A MoE block maps $h \in \mathbb{R}^d$ to

$$y(h) = \sum_e w_e(h) E_e(h),$$

where router logits are $g_e = W_e h$, top-k selects the largest logits, and w is the renormalized softmax over the selected logits. With ordered logits $g_{(1)} \geq \cdots \geq g_{(k)} \geq g_{(k+1)}$, the $k/(k+1)$ active-set boundary is $g_{(k)} = g_{(k+1)}$, a hyperplane with normal $n = W_{(k)} - W_{(k+1)}$. Crossing it swaps the k-th expert, creating a potential output jump proportional to $\left\|E_{(k)}(h) - E_{(k+1)}(h)\right\|$.

4 Diagnostic

The diagnostic needs only access to the router gate and expert modules. It has three core measurements:

- **M1 boundary prevalence**: the distribution of margin $g_{(k)} - g_{(k+1)}$ and the fraction of tokens near the boundary.
- **M2 expert cliff**: the normalized $\left\|E_{(k)} - E_{(k+1)}\right\|$ plus the cosine of the swapped pair, distinguishing non-redundancy from outsized specialization.
- **M3 continuity signature**: the difference quotient $\|\Delta y\| / \|\Delta x\|$ tracked under grid refinement (resolutions $500 \to 2000 \to 8000$) along a boundary-crossing path. For a genuine order-0 jump it diverges as T^1 (scaling exponent ≈ 1); for a continuous map it saturates (exponent ≈ 0). We summarize this by the refinement growth hardG $= \max_q[8000]/\max_q[500]$, which

by construction equals the grid ratio $T_{\max}/T_{\min} = 8000/500 = 16$ for any nonzero C^0 jump, independent of k, expert count, or layer. Hence the diagnostic content is the exponent (jump vs. continuous), not the value 16, which is what a true jump must yield under this protocol rather than a measured severity. Severity is carried separately by the absolute relative jump jump_{rel} (the per-block output jump) and M2. Tied-expert and soft-edge controls stay near $1\times$ (exponent ≈ 0).

The inference probe adds the boundary normal and distance-to-tear

$$d_{\text{tear}} = \frac{g_{(k)} - g_{(k+1)}}{\left\| W_{(k)} - W_{(k+1)} \right\|}.$$

It then compares equal-magnitude normal and tangent perturbations. The clamp probe caps the SwiGLU intermediate after routing, leaving top-k unchanged while measuring topology, amplitude, and quality.

5 Released-Weight Characterization

The tear is not synthetic. On OLMoE-1B-7B [Muennighoff et al., 2024], the continuity signature is a genuine order-0 jump at every layer—refinement growth hardG $\approx 16\times$, i.e. scaling exponent 1.00, the difference quotient diverging at the grid-refinement rate—while soft-edge and tied-expert controls remain near $1\times$ (exponent ≈ 0; Figure 1). The $16\times$ is the grid ratio 8000/500 that any true jump reaches, not a severity: the signature certifies that the singularity is order-0 and present at every layer, while severity is the per-block output jump $\text{jump}_{\text{rel}} \approx 0.239$ ($\approx 24\%$ of the block-output norm, a per-block geometric quantity—not a model-output, logit, or task-accuracy change) and the cliff M2 ≈ 0.70. Table 1 gives the per-layer diagnostic. Qwen1.5-MoE-A2.7B [Qwen Team, 2024] replicates the same pattern (Figure 2).

The $16\times$ is not a top-k=8 artifact. Three independent reasons, strongest first. (i) *By construction*: the refinement growth equals $T_{\max}/T_{\min}$ for any nonzero C^0 jump regardless of k, expert count, or layer, so $16 = 8000/500$ is the value a genuine order-0 jump must produce under this protocol—there is no algebraic path from k to it. (ii) *Cross-model evidence*: Qwen1.5-MoE routes $k = 4$ yet shows the same growth ($15.99\times$, exponent 1.00) as OLMoE's $k = 8$, inconsistent with a value set by k. (iii) A direct same-model $k \in \{1, 2, 4, 8\}$ sweep now settles it outright (Figure 9; 64 boundary paths/layer at OLMoE layers 0/8/15): hardG holds at 16.00 (exponent 1.00) for every k, total range 15.92–16.03, tracking the protocol grid ratio rather than k.

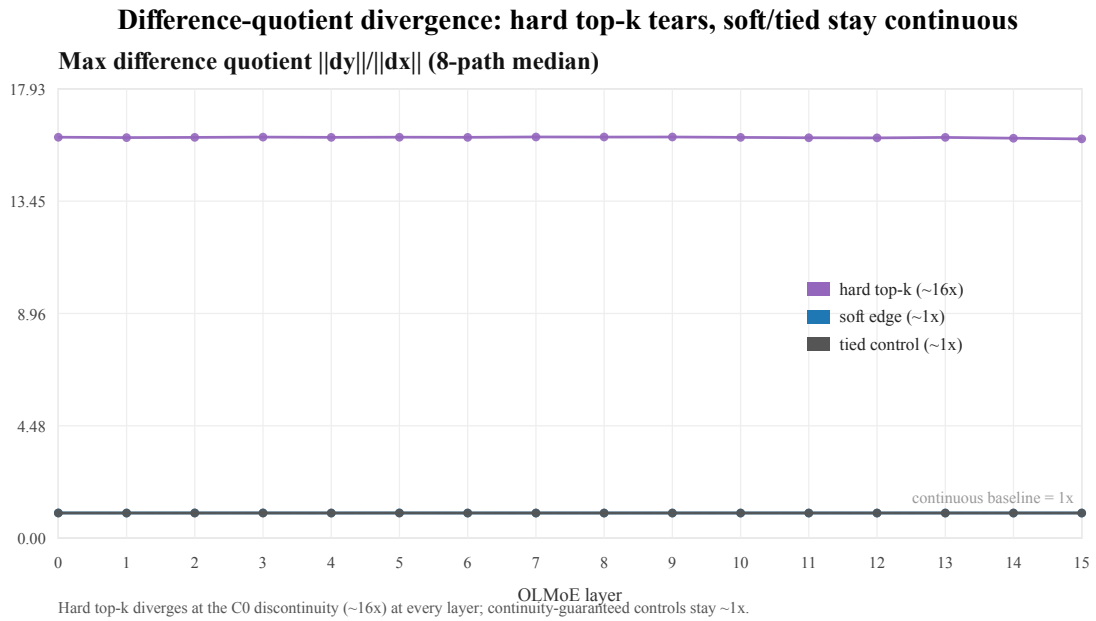

Figure 1: Difference-quotient continuity signature on released OLMoE-1B-7B. The hard top-k refinement growth $\max_q[8000]/\max_q[500]$ sits at $\approx 16\times$—the grid ratio 8000/500, i.e. scaling exponent ≈ 1, the order-0-jump signature—at every layer, while soft-edge and tied controls stay at $\approx 1\times$ (exponent ≈ 0). The value 16 is what a genuine jump must yield under this protocol; severity lives in the block jump and M2, not here.

Table 1: OLMoE-1B-7B per-layer diagnostic. hardG, block jump, and controls are 8-path medians; margin, M2, and cosine are per-token medians. hardG is the refinement growth $\max_q[8000]/\max_q[500]$; for an order-0 jump it equals the grid ratio 16, i.e. exponent $\ln(\text{hardG})/\ln 16 \approx 1.00$—its near-constancy is the protocol's, not the model's. The jump column is the per-block output jump (fraction of block-output norm).

Layer	margin	M2	cos	hardG	jump	soft	tied
0	0.0012	0.694	0.056	16.00	0.238	1.005	1.001
1	0.0009	0.701	0.033	15.98	0.309	1.003	1.002
2	0.0010	0.702	0.028	15.99	0.246	1.004	1.001
3	0.0011	0.704	0.022	16.00	0.248	1.004	1.002
4	0.0011	0.703	0.025	15.99	0.282	1.003	1.002
5	0.0012	0.702	0.030	16.00	0.288	1.004	1.002
6	0.0011	0.704	0.023	15.99	0.284	1.002	1.003
7	0.0014	0.704	0.026	16.01	0.233	1.003	1.002
8	0.0015	0.704	0.025	16.00	0.249	1.003	1.002
9	0.0019	0.703	0.026	16.01	0.266	1.003	1.002
10	0.0019	0.708	0.015	15.99	0.212	1.003	1.002
11	0.0021	0.709	0.012	15.98	0.244	1.003	1.002
12	0.0026	0.707	0.021	15.97	0.226	1.003	1.002
13	0.0023	0.709	0.015	15.99	0.157	1.003	1.002
14	0.0023	0.703	0.035	15.96	0.222	1.003	1.002
15	0.0028	0.600	0.301	15.93	0.123	1.003	1.002

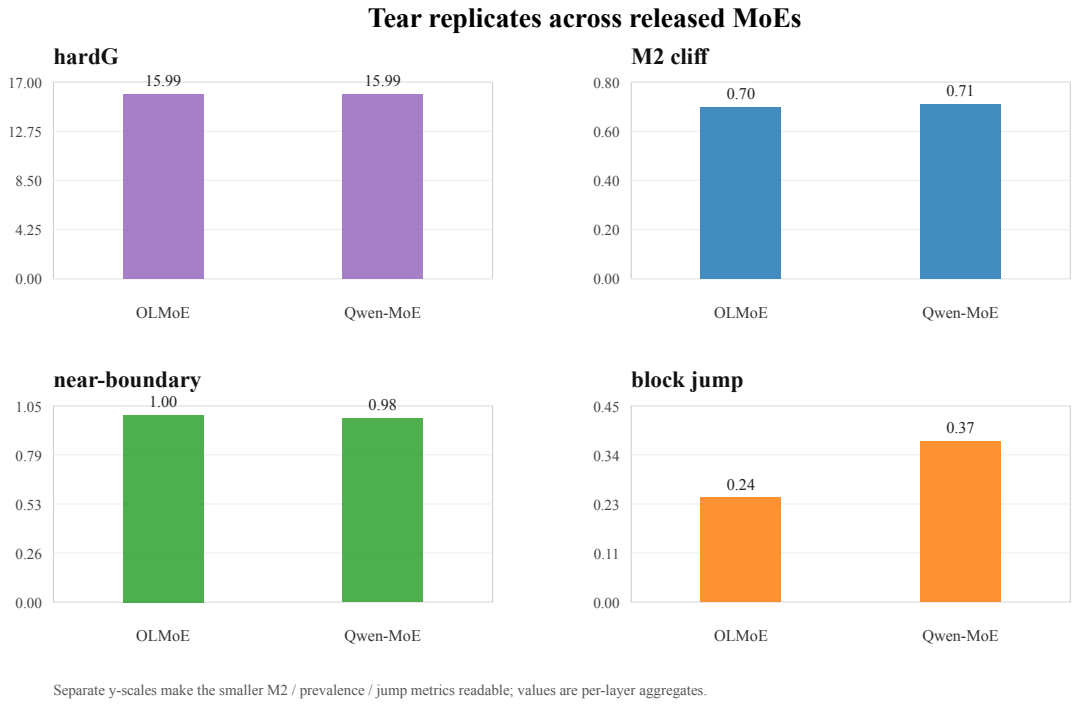

Figure 2: Cross-model replication. The refinement-growth signature is $\approx 16\times$ (exponent ≈ 1) and M2 is ≈ 0.70 on both released MoE families; the $16\times$ shared across k=8 and k=4 is the protocol's grid ratio, not a k-set value. Severity reads from the per-block jump and M2, not the growth.

6 Directional Inference Robustness

Random hidden perturbations give a mostly null robustness result: even when 67.6% of top-k sets flip, the hard-vs-soft block jump differs by only 2.6%. The null is not that experts never flip; it is that random flips add little discontinuous excess response beyond the smooth block response.

The reason is geometric. Equal-magnitude perturbations at 2× distance-to-tear behave very differently along the boundary normal and tangent directions (Figure 3). Normal perturbations flip the $k/(k+1)$ expert with probability about 0.9 and produce $O(0.1)$ per-block output jumps (a fraction of the block-output norm, not of the model output); tangent perturbations barely flip. All-layer continuous re-gating removes the static tear but increases perplexity from 10.04 to 159.94, a 15.9× cost. Single-layer re-gating is much milder, around 1.2× perplexity, which is why any practical mitigation likely needs targeted layers or retraining rather than post-hoc all-layer replacement.

7 Training-Time Effect

The training experiment is a controlled small-scale GPT-MoE probe, not OLMoE pretraining. At E=64, k=8, and about 308M parameters, the parameter-space difference-quotient probe finds hard routing rougher than a continuity-matched soft gate (2.0× vs. 1.4×), but the effect is modest.

Natural training is spiky but convergent: over 8000 steps, there are 196 spikes greater than 0.3, max spike 0.585, and no divergence. M2 and hardG do not self-heal, while the operational whole-block jump collapses early and then plateaus (Figure 4). In a two-seed tear-level dial, no run diverges; full tear reproducibly creates a rare severe-but-recovered spike (max 4.21 and 5.28).

8 Discussion: Routing Discontinuity Is Not Loss Spike

Our measurements caution against equating the tear with loss spikes. We decompose the training story into five factors: routing discontinuity, expert-outlier magnitude, temporal backbone/router mismatch, cross-layer propagation gain, and optimizer absorption. The first is geometric; the second is supported by a norm-vs-jump correlation of +0.437; the third remains future work; the fourth is small in our injected-jump measurements; the fifth explains why small-scale training absorbs the tear.

This provides a geometric reading of DeepSeek-V4's interventions [DeepSeek-AI, 2026]. SwiGLU clamping caps expert-outlier magnitude; Anticipatory Routing targets temporal mismatch; manifold-constrained hyper-connections bound cross-layer propagation. Our clamp sweep confirms the first point on released OLMoE: hardG remains 16.03 → 16.02 and M2 remains 0.705 → 0.701, while absolute expert cliff drops 2.90 → 1.97 and absolute hardJump drops 0.214 → 0.164 (Figure 5).

9 Conclusion

The MoE routing tear is real, measurable, and cross-model: a genuine C^0 discontinuity (the difference quotient diverges at scaling exponent ≈ 1 at every layer, controls flat at ≈ 0) that trained routers carry rather than sew up—severe in the per-block output jump ($\approx 24\%$ of the block-output norm), though not in expert specialization (M2 sits near the unrelated-vector baseline). Its inference consequence is directional, and its training-time role is a decomposition rather than a single cause. Measurement, not a new gate or taxonomy, is the contribution.

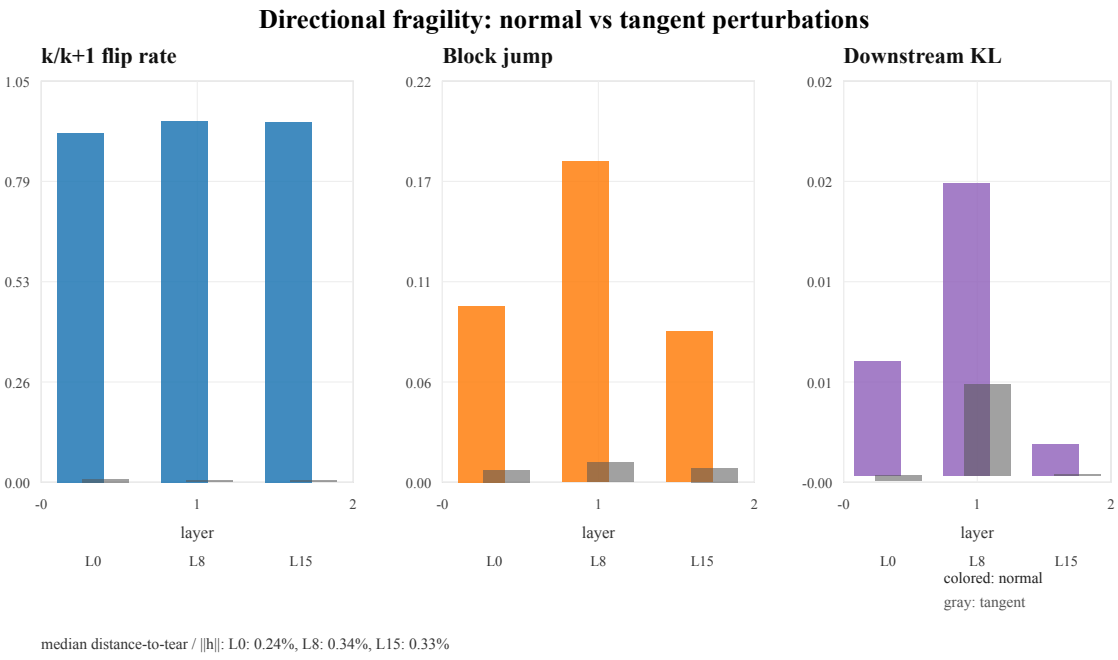

Figure 3: Directional fragility. Boundary-normal perturbations expose the tear; equal-magnitude tangent perturbations do not.

10 Limitations

Signature certifies order, not magnitude. The continuity signature certifies that the singularity is order-0 (exponent ≈ 1 vs. controls ≈ 0) and where it sits; it does not measure severity. Its raw growth value ($16\times$) is the grid ratio $T_{\max}/T_{\min}$ that any genuine jump must produce under this protocol—not a model property, and not larger-is-worse. Severity is carried only by the per-block output jump (jump_{rel}) and the M2 cliff.

Measurement coverage. Two released families and 24 short prompts. Per-token statistics (margin, M2, cosine) span $\sim$500 tokens/layer, but the refinement growth and jump use 8 boundary-crossing paths/layer (released weights) and a single path/probe in the training run—the latter needs multi-path resampling for error bars. Cross-domain and multilingual coverage is not characterized. The $k \in \{1, 2, 4, 8\}$ independence and the 8-to-64-path robustness of the growth are now directly confirmed on OLMoE layers 0/8/15 (Figure 9), not only argued.

Geometric jump $\neq$ task performance. Every number here is a geometric quantity on hidden states; none is an end-to-end task metric. The $\approx 24\%$ block-output jump and the boundary-normal fragility are not tied to accuracy on benchmarks (e.g. GSM8K, MMLU) under normal-direction perturbation.

Per-layer, not compounded. The jump is measured one layer at a time; how tears propagate and compound across stacked blocks is untested (the cross-layer-gain factor bounds only single-step propagation).

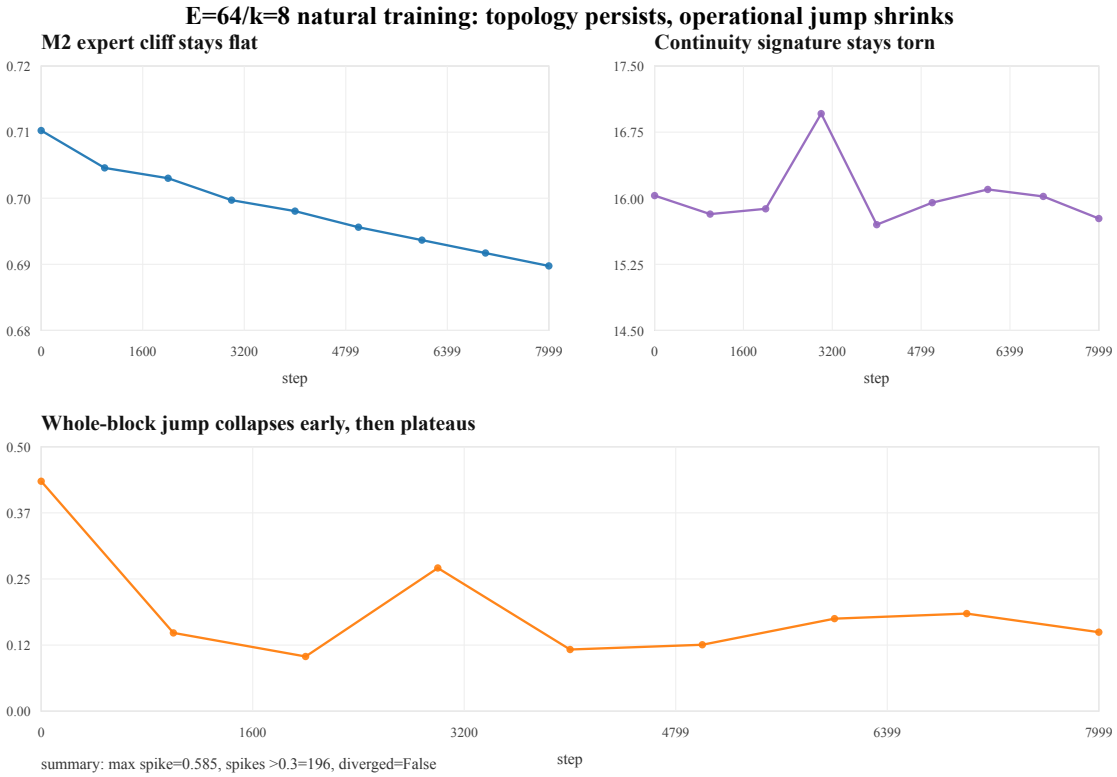

Figure 4: Controlled E=64/k=8 training. M2 and hardG persist, while whole-block jump collapses early then plateaus.

Training-probe gap. The training results come from a 308M, 8-layer, Wikitext-103 from-scratch GPT-MoE, not 7B-scale pretraining; scale, data, and architecture all differ, and the clamp-quality figure is a within-hook-path comparison. The temporal backbone/router mismatch factor is named but not measured.

A Supplementary Figures

B Method and Experiment Details

Released-weight sweeps. OLMoE uses 16 MoE blocks, 64 experts, top-$k = 8$, bfloat16 CUDA forwards, PyTorch 2.7.0, 24 short text prompts, and 8 boundary-crossing refinement paths per layer. Qwen1.5-MoE uses 24 MoE blocks, 60 routed experts, and top-$k = 4$. hardG values use the 8-path median to avoid single-path degeneracy.

Continuity signature. For each sampled token and layer, the probe identifies the $k/(k{+}1)$ router boundary, samples a path crossing that boundary, refines the path grid (resolutions $500 \to 2000 \to 8000$), and records the maximum $\|\Delta y\| \,/\, \|\Delta x\|$ at each resolution. We report the refinement growth $\text{hardG} = \max_q[8000]/\max_q[500]$; for an order-0 jump $\max_q \propto T$, so $\text{hardG} \to T_{\max}/T_{\min} = 16$ independent of k, expert count, or layer—the diagnostic content is the implied scaling exponent $\ln(\text{hardG})/\ln 16$ (≈ 1 for a jump, ≈ 0 for a continuous map), not the value 16. A synthetic

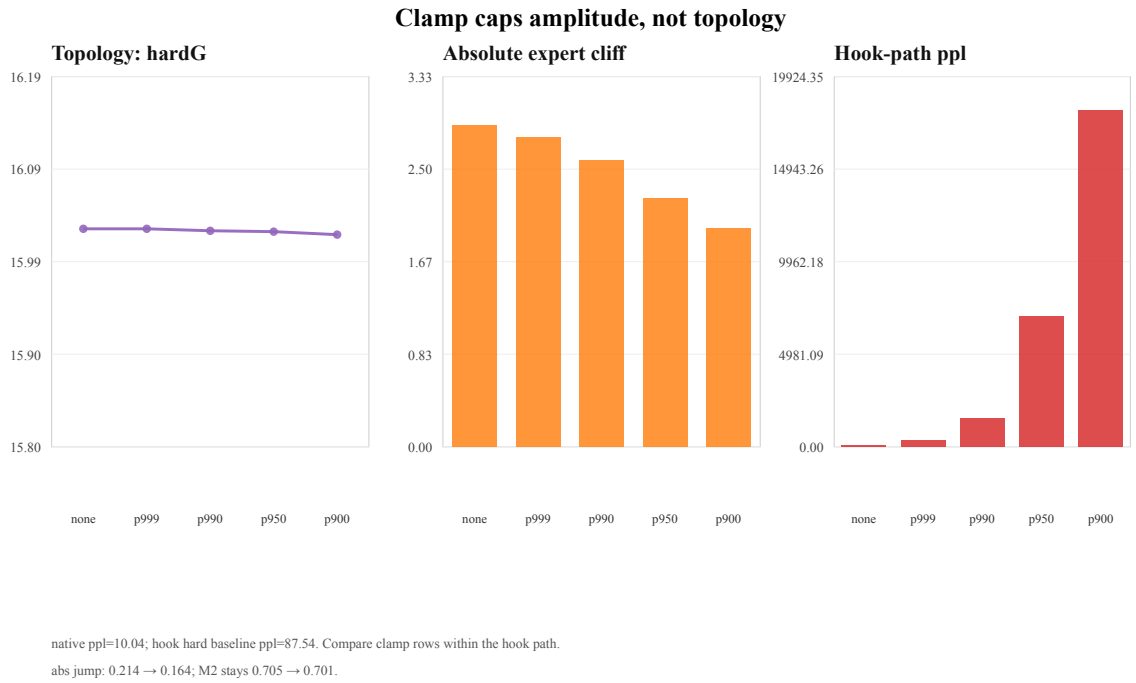

Figure 5: Clamp decomposition. The topology stays flat while absolute amplitude decreases; quality cost is measured within the hook path.

block with an exact known C^0 jump confirms the T^1 law (fitted exponent 1.0005, $R^2 = 1.000$ over resolutions 250 $\rightarrow$ 16000; Figure 8). The main sweep runs in bf16; the clamp probe re-computes the signature in fp32 and agrees at hardG 16.0, so bf16 rounding is not smearing the jump into a finite ramp. The M2 expert cliff is measured on the swapped experts, with cosine reported to distinguish redundancy from specialization.

Directional robustness. The normal/tangent probe measures distance-to-tear as

$$d_{\text{tear}} = \frac{g_k - g_{k+1}}{\|W_k - W_{k+1}\|},$$

then compares equal-magnitude perturbations along the raw-logit boundary normal and a tangent direction. The reported table uses $\alpha = 2$ times this distance.

Training probe. The E=64/k=8 run uses BPE next-token training on Salesforce Wikitext-103 raw text, 20M loaded tokens, batch size 24, sequence length 512, model dimension 512, 8 layers, 8 heads, hidden size 512, AdamW with learning rate $6 \cdot 10^{-4}$ and weight decay 0.1, bf16 autocast, 8000 steps, eval every 25 steps, and tear probes every 1000 steps.

Clamp sweep. The clamp probe applies a SwiGLU-intermediate clamp after routing. It reports both normalized geometry and absolute amplitude. Perplexity values are hook-path values; compare clamp rows within that path.

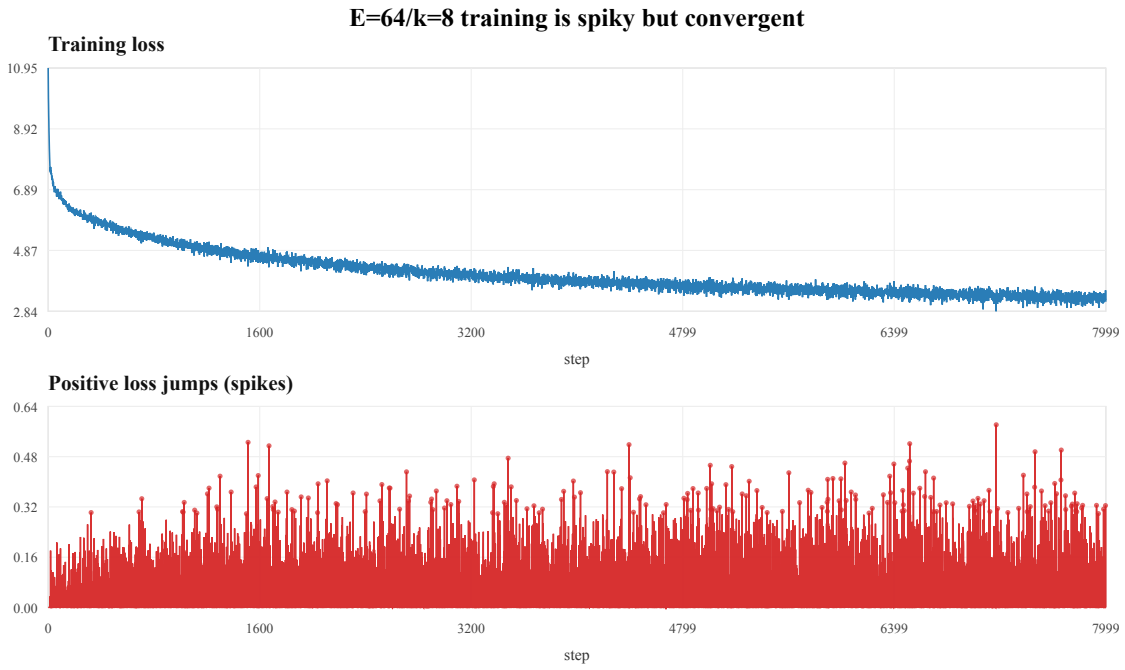

Figure 6: E=64/k=8 training is spiky but convergent.

C Reproducibility Checklist

The numeric source of truth is the JSON result set and the GPU results ledger. Figures regenerate with `python3 scripts/plot_results.py`.

References

Randall Balestriero and Richard Baraniuk. A spline theory of deep networks. In *International Conference on Machine Learning*, pages 374–383, 2018. arXiv:1805.06576.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. arXiv:2606.19348.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. arXiv:2101.03961.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, 2019. arXiv:1901.09021.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. arXiv:2006.16668.

Max Y. Ma and Gen-Hua Shi. Deep manifold part 1: Anatomy of neural network manifold, 2024. arXiv:2409.17592.

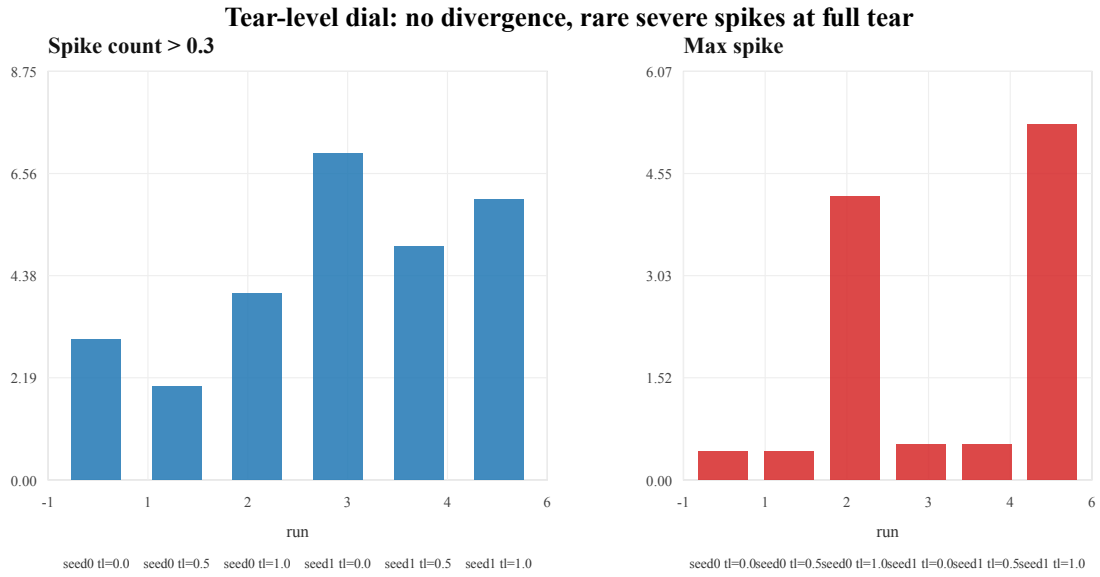

Figure 7: Tear-level dial across two seeds. No run diverges; full tear creates rare severe-but-recovered spikes.

Max Y. Ma and Gen-Hua Shi. Deep manifold part 2: Neural network mathematics, 2025. arXiv:2512.06563.

Wenhan Ma, Hailin Zhang, Liang Zhao, Yifan Song, Yudong Wang, Zhifang Sui, and Fuli Luo. Stabilizing moe reinforcement learning by aligning training and inference routers, 2025. arXiv:2510.11370.

André F. T. Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, 2016. arXiv:1602.02068.

Guido F. Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, 2014. arXiv:1402.1869.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models, 2024. arXiv:2409.02060.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Annual Meeting of the Association for Computational Linguistics*, 2019. arXiv:1905.05702.

Joan Puigcerver, Rodolphe Jenatton, Carlos Riquelme, Pranjal Awasthi, and Srinadh Bhojanapalli. On the adversarial robustness of mixture of experts. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2210.10253.

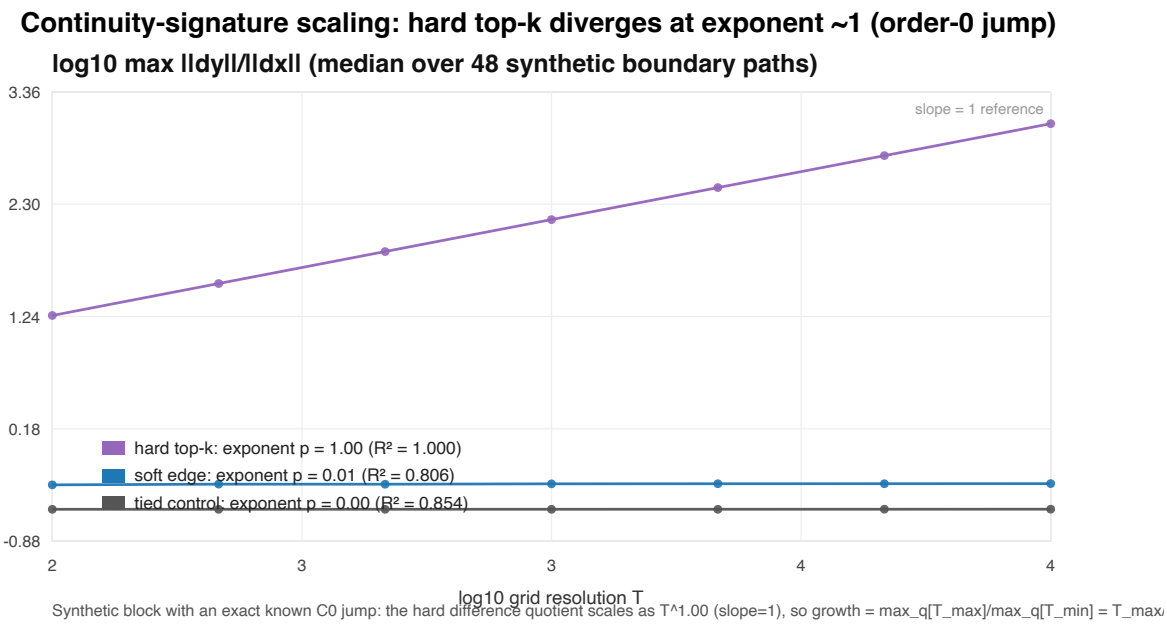

Figure 8: Continuity-signature scaling on a synthetic block with an exact, known C^0 jump (median over 48 boundary paths). The hard difference quotient scales as $T^{1.00}$ (fitted exponent 1.0005, $R^2 = 1.000$ over resolutions $250 \to 16000$), so the refinement growth equals the grid ratio $T_{\max}/T_{\min}$ by construction; soft-edge and tied controls stay flat (exponent ≈ 0). This confirms the released-weight $16\times$ is the order-0-jump signature (exponent 1), not a probe artifact, and explains why it is pinned across layers, models, and k.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024. arXiv:2308.00951.

Qwen Team. Qwen1.5-MoE: Matching 7b model performance with 1/3 activated parameters. https://qwenlm.github.io/blog/qwen-moe/, 2024. Model blog for Qwen1.5-MoE-A2.7B.

Ziteng Wang, Jun Zhu, and Jianfei Chen. Remoe: Fully differentiable mixture-of-experts with relu routing. In *International Conference on Learning Representations*, 2025. arXiv:2412.14711.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Designing stable and transferable sparse expert models, 2022. arXiv:2202.08906.

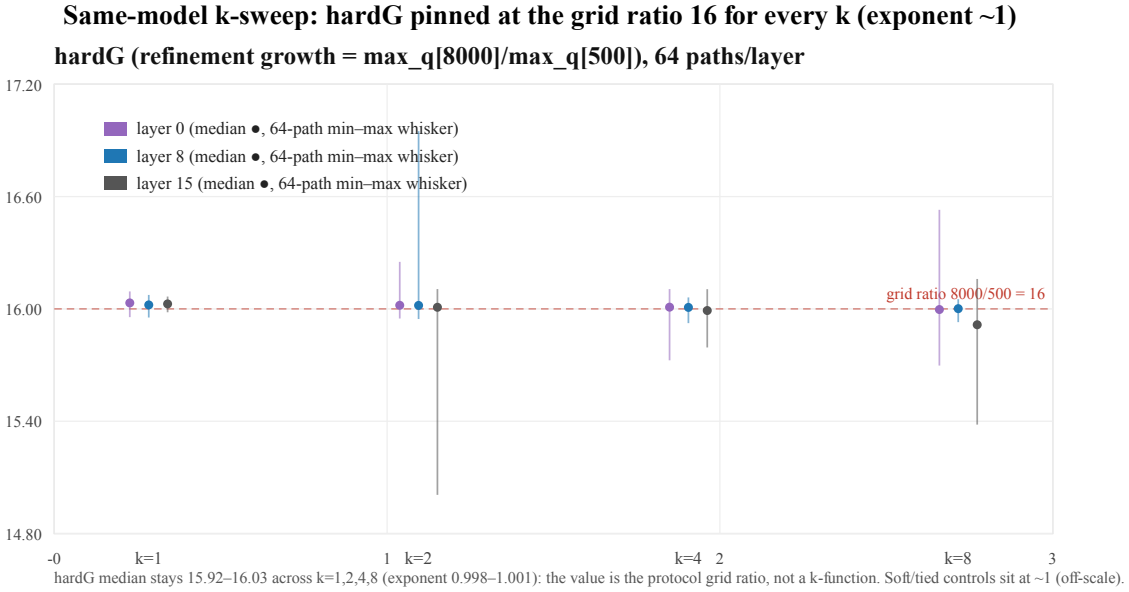

Figure 9: Same-model k-sweep on released OLMoE-1B-7B (layers 0/8/15, 64 boundary paths/layer per cell). hardG stays pinned at the grid ratio 16 (median 15.92–16.03, exponent 0.998–1.001) across $k = 1, 2, 4, 8$, directly confirming the value is set by the probe's grid ratio 8000/500 and not by top-k. The 64-path min–max whiskers also answer the "8 paths is too few" concern: the median is stable while only rare single paths stray. Soft/tied controls sit at ≈ 1 (off-scale).