3D-DELTA: A Learned Low-Rank Expert-Steering Head for Frozen Mixture-of-Experts LLMs

Nima Kelidari
University of Southern California
kelidari@usc.edu

Abstract

MoE routers encode behavioral axes such as context-faithfulness, and clamping experts up or down at inference steers a frozen model along them. Prior work selects *which* experts with hand-engineered statistics. We replace the statistic with a small *learned* head: a low-rank tensor $\Delta \in \mathbb{R}^{L\times H\times E}$ (shared encoder $W \in \mathbb{R}^{H\times k}$, per-layer factor $k \times E$; ∼200k parameters) trained with a contrastive pairwise-ranking loss on with-context vs. without-context QA pairs, feeding the *identical* SteerMoE inference rule on a frozen model. On OLMoE-1B-7B the learned head roughly quadruples the mean faithfulness gain of the hand statistic, and the same recipe steers Qwen3-30B-A3B. We also report an honest *selection-stability* analysis: best-cell hyperparameters overfit the evaluation grid per training seed, motivating validation-split cell selection — a pitfall we believe generalizes to steering-method evaluation broadly.

1 Method

SteerMoE (Fayyaz et al., 2026) steers a frozen MoE by forcing the top-A experts' router logits to the per-token max and the bottom-D to the min, ranking experts by a risk-difference statistic computed from paired inputs $x^{(1)}$ (context + question) vs. $x^{(2)}$ (question only). We keep the deployment rule and knobs, and learn the ranking source instead. For each pair we capture, per layer ℓ, a pooled hidden state h_ℓ and pooled router softmax $y_\ell^{(1)}, y_\ell^{(2)}$ (a single HF forward per prompt; nine pooling variants captured at once). The head predicts the routing contrast: $\hat{y}_\ell = h_\ell^\top W\, \Delta_\ell^{lr} \approx y_\ell^{(1)} - y_\ell^{(2)}$, trained with a pairwise hinge on the top/bottom-k contrast ranking. At inference the fused $\Delta_\ell = W\Delta_\ell^{lr}$ yields a static (L, E) $\{-1, 0, +1\}$ mask under SteerMoE's greedy global (A, D) budget; the mask is installed by patching only the vLLM router gate, preserving each model's native forward (FusedMoE, top-k renormalization, TP all-reduce).

Three design choices carry the result: (i) **learned head** instead of a hand statistic; (ii) **last-token pooling** instead of mean-over-tokens — monotonically better across all ranks k (held-out contrast-ranking accuracy 0.946 vs. 0.926 at k=64), and mechanistically it moves $|\Delta|$ mass out of a layer-0 lexical attractor (30/50 top cells → 3/50) into middle layers where routing decisions live; (iii) **multi-dataset training** (SQuAD+HotpotQA+FEVER), which unlocks axes a single-dataset head misses.

2 Results (OLMoE-1B-7B; frozen)

Across six faithfulness benchmarks (FaithEval counterfactual / unanswerable / inconsistent, CF-TriviaQA, MQuAKE, MCTest), the best-cell gain Δ over the method's own no-steer baseline:

	SteerMoE	B4 (ours, 1 ds)	Comb. (ours)
FaithEval-CF	+.020	+.029	+.060
FaithEval-Unans	+.120	+.206	+.200
FaithEval-Inco	+.050	+.075	**+.288**
CF-TriviaQA	.000	+.002	.000
MQuAKE	−.030	+.067	+.054
MCTest	−.010	+.011	+.006
mean	+.025	+.065	**+.101**

Table 1: Steering gain Δ (best cell per benchmark, seed-0 grid; SteerMoE quoted from its paper). The learned multi-dataset head is the only method that moves FaithEval-Inconsistent materially.

Dataset attribution. The Inconsistent gain is not "more data helps": SQuAD+FEVER alone reaches +.328 while SQuAD+HotpotQA gives +.143 — FEVER's claim-vs-evidence pairs *are* a conflict-detection task, and the head learns a routing axis aligned with it. Different benchmarks select different (A, D) regimes (abstention: low A, moderate D; parametric recovery: high A), i.e., one head exposes several routing axes where a single statistic gives one.

Cross-model. The same last-token, k=64 recipe

wins on Qwen3-30B-A3B (held-out 0.976) and Mixtral-8×7B (0.836). On Qwen3 the full $A \times D$ steering sweep is complete and positive (mean Δ +.042; Inconsistent +.108), showing the learned-head recipe transfers to a 30B reasoning MoE with 128 experts.

Selection stability (a finding, not a footnote). Re-training the head under new seeds and re-sweeping the grid shows per-seed gains persist but the *argmax cell does not transfer*: seed-0's winning cells can go negative under other seeds, and per-seed own-argmax gains vary substantially. Echoing recent reliability analyses of activation steering (Tan et al., 2024; Braun et al., 2025), we argue steering papers should (a) select cells on a validation split and (b) report seed variance; we adopt both. Consistent-N four-seed grids are running now, along with capability-preservation checks (MMLU/GSM8K under the mask) and additional MoE families spanning 3B–47B and shared/fine-grained-expert designs (Granite-3.0-MoE, Qwen1.5-MoE, DeepSeek-MoE), plus a per-sample conditional mask as the next methodological step.

3 Positioning

Router Lens/CEFT (Bai et al., 2025) learns context-faithful *expert identification* but updates weights (router- then expert-finetuning); MASCing (te Lintelo et al., 2026) (concurrent) optimizes sparse expert masks against an LSTM routing surrogate for *safety*. 3D-DELTA is, to our knowledge, the first *parametric, hidden-state-conditioned* low-rank head trained contrastively for *faithfulness* steering of a fully *frozen* MoE, drop-in compatible with the SteerMoE rule — bringing the "learn the intervention" program from dense residual streams (Li et al., 2023; Wu et al., 2024) to the expert-routing interface.

References

M. Fayyaz et al. Steering MoE LLMs via Expert (De)Activation. ICLR 2026. arXiv:2509.09660.

Z. Bai et al. Understanding and Leveraging the Expert Specialization of Context Faithfulness in MoE LLMs. EMNLP 2025. arXiv:2508.19594.

J. te Lintelo et al. MASCing: Configurable MoE Behavior via Activation Steering Masks. arXiv:2604.27818.

K. Li et al. Inference-Time Intervention. NeurIPS 2023. arXiv:2306.03341.

Z. Wu et al. ReFT: Representation Finetuning for Language Models. NeurIPS 2024. arXiv:2404.03592.

D. Tan et al. Analyzing the Generalization and Reliability of Steering Vectors. NeurIPS 2024. arXiv:2407.12404.

J. Braun et al. Understanding (Un)Reliability of Steering Vectors in Language Models. arXiv:2505.22637.