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

Nima Kelidari
University of Southern California
USA
kelidari@usc.edu

Abstract

SteerMoE [1] steers Mixture-of-Experts (MoE) LLMs by clamping experts up or down at inference time, and it chooses which experts to clamp using a hand-engineered risk-difference statistic over router logits. We replace that statistic with a small learned low-rank head $\Delta \in \mathbb{R}^{L\times H\times E}$, read out from the last-token hidden state and trained contrastively on three datasets (SQuAD, HotpotQA, FEVER). It plugs into the same inference rule SteerMoE uses, so the comparison is like-for-like. On OLMoE-1B-7B the learned head produces the largest mean steering gain of any method we tested, +.101 across six faithfulness benchmarks (SteerMoE reports +.025 and a prior reimplementation [2] +.038), and it lifts the hardest benchmark, FaithEval-Inconsistent, from 0.525 to 0.813 (+.288) where every other method moves it by under +.05. Three findings explain the result. First, last-token pooling beats mean pooling at every rank, and it moves the head's mass out of a layer-0 lexical attractor into the middle layers. Second, in four controlled training-mix experiments, FEVER's claim-and-evidence pairs alone account for the conflict-detection gain (+.328). Third, the same recipe transfers across MoE families: held-out routing-ranking accuracy is 0.946, 0.976, and 0.836 on OLMoE, Qwen3-30B-A3B, and Mixtral-8×7B (7B to 47B parameters, 8 to 128 experts), and the learned head steers Qwen3-30B-A3B end to end (+.042 mean gain over the same six benchmarks). We also surface a pitfall we believe applies to steering evaluation broadly: the best (A, D) cell overfits the sweep grid per training seed, which motivates validation-split cell selection; we characterize this explicitly and adopt the fix in ongoing runs. The whole pipeline runs from a single command per model.

1 Introduction

SteerMoE [1] showed that the routers of MoE LLMs encode behaviorally meaningful axes such as faithfulness, safety, and refusal, and that small additive interventions on router logits move the model along those axes without retraining. Its detection step computes a static risk difference $\delta_{\ell,e}=p^{\text{faith}}_{\ell,e}-p^{\text{unfaith}}_{\ell,e}$ from contrastive SQuAD pairs. Its deployment step then force-activates the top-$|A|$ positive cells and force-deactivates the top-$|D|$ negative cells, globally across all (L, E) positions. Srivatsa et al. [2] reproduced this pipeline and tried a token-aware extension.

We ask whether the hand-engineered population statistic can be replaced by a learned head that pools tokens more carefully than a uniform mean, combines several training datasets, and exposes more than one usable axis. The answer is yes. A k=64, last-token head trained on SQuAD, HotpotQA, and FEVER produces the largest mean steering gain of any method we compared, and the knob setting that works best is different for each benchmark, which tells us the head is multi-directional rather than a single faithfulness scalar.

Positioning. Two adjacent lines of work sharpen what is new here. Router Lens/CEFT [8] also identifies context-faithful experts by *learning* — but by tuning the router weights and then fine-tuning the identified experts, so the model is modified twice and nothing happens at inference time. MASCing [9] (concurrent, security domain) optimizes a static sparse expert mask against an LSTM surrogate of routing dynamics for safety objectives. Learned interventions are likewise established on dense residual streams [10, 11]. 3D-DELTA differs on all three axes at once: the model stays fully *frozen*, the artifact is a *parametric, hidden-state-conditioned* low-rank head rather than a static lookup, and the objective is context-*faithfulness* evaluated on held-out benchmarks under SteerMoE's own inference rule. To our knowledge it is the first learned inference-time faithfulness-steering head for frozen MoE LLMs.

Contributions. (1) A learned, drop-in replacement for SteerMoE's mask source that keeps the inference rule identical and produces the largest mean steering gain among the methods we tested (+.101, against +.025 for SteerMoE). (2) The first 36-variant (k, aggregation) ablation on contrastive router captures, which shows last-token pooling beats mean pooling at every rank, with a mechanistic read-out (Fig. 3). (3) A multi-dataset recipe whose attribution we isolate in four training-mix experiments: FEVER alone drives the hardest win (Tab. 2). (4) Evidence that the recipe generalizes across three MoE families — including end-to-end steering of Qwen3-30B-A3B — delivered by a model-agnostic, one-command pipeline. (5) A selection-stability analysis showing the per-benchmark argmax cell overfits the sweep grid per training seed, with a concrete evaluation fix (validation-split cell selection) that we argue steering papers should adopt generally.

2 Method

Detection pairs. For each sample we form $x^{(1)}$=Doc + Q and $x^{(2)}$=Q alone, following SteerMoE but extending to three datasets: SQuAD [5] (5 k pairs), HotpotQA [6] (2 k, supporting facts as the Doc), and FEVER [7] (2 k, evidence sentences as the Doc, after filtering Not-Enough-Info). At every MoE layer we record per-token hidden states and softmaxed router logits, and we pool them nine ways in one forward pass: mean-over-all and last-N for $N\in\{1, 2, 4, 8, 16, 32, 64, 128\}$.

Head and training. We fit $e_\ell = h^{\text{pool}}\Delta_\ell$ with $\Delta_\ell = W_{\text{proj}}\Delta^{\text{lr}}_\ell$: a shared encoder $W_{\text{proj}}\in\mathbb{R}^{H\times k}$ and a per-layer factor $\Delta^{\text{lr}}_\ell\in\mathbb{R}^{k\times E}$, so capacity scales with the rank k rather than with depth. The loss is a contrast-aware pairwise hinge against the routing target $y^{(1)}_\ell - y^{(2)}_\ell$.

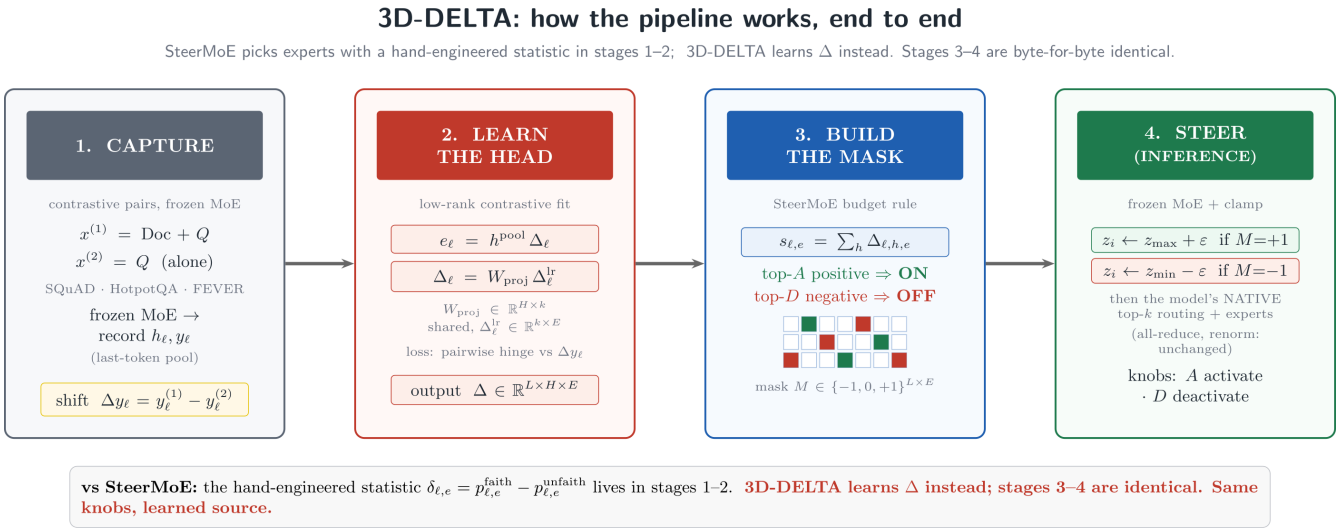

Figure 1: The 3D-DELTA pipeline, end to end. (1) Capture: run contrastive pairs ($x^{(1)}$=Doc+Q vs. $x^{(2)}$=Q) through the frozen MoE and record the per-layer router shift $\Delta y_\ell = y_\ell^{(1)} - y_\ell^{(2)}$ (last-token pooled). (2) Learn the head: fit a low-rank $\Delta_\ell = W_{\text{proj}}\Delta_\ell^{\text{lr}}$ to predict the shift with a pairwise hinge. (3) Build the mask: score $s_{\ell,e} = \sum_h \Delta_{\ell,h,e}$ and force the top-A experts on / top-D off under SteerMoE's greedy budget. (4) Steer: clamp those experts' router logits at inference; the model's native routing and experts run unchanged. SteerMoE uses a hand-engineered statistic in stages 1–2; we learn Δ instead, and stages 3–4 are identical.

We sweep $k{\in}\{8, 16, 32, 64\}$ against the nine aggregations (36 variants) and pick the winner by held-out top-and-bottom pairwise ranking accuracy on a 10% split.

Inference (unchanged from SteerMoE).. We precompute an (L, E) mask from Δ's row-sum scores under SteerMoE's greedy budget rule: walk the cells by descending score magnitude, force-activate the top $|A|$ globally subject to a per-layer cap of k_{routing}, and force-deactivate the top $|D|$. The mask is applied uniformly to every token, $z_i = z_{\max}{+}\varepsilon$ if $i{\in}A$ and $z_i = z_{\min}{-}\varepsilon$ if $i{\in}D$. Same rule, same knobs, learned source.

3 Results

Setup. OLMoE-1B-7B-0125-Instruct [3] has 16 layers, 64 experts, and routes top-8. We run vLLM 0.8.5 with greedy decoding. The six benchmarks are the ones in Fayyaz et al.: FaithEval Counterfactual, Unanswerable, and Inconsistent [4], CF-TriviaQA, MQuake, and MCTest, swept over $A{\in}\{0, 2, 4, 8, 16, 32\}$ and $D{\in}\{0, 32, 64, 128, 256, 512\}$ (36 cells per head). We also fixed a prompt artifact that had depressed the FaithEval-Inconsistent baseline: a persona instruction suppressed the `conflict` output and gave a false floor of 0.112. The real baseline is 0.525, and we use the fixed prompt throughout.

How to read the comparison. We report Δ, the accuracy of the best-steered cell minus that method's own no-steer baseline, because Δ isolates what the steering does. Our heads all share one baseline (our pipeline, our prompts, the same OLMoE checkpoint), so they are directly comparable. The SteerMoE and adsrivatsa numbers are quoted from their papers, which use different baselines, so we compare the gain Δ and not absolute accuracy. On a benchmark where their published baseline is far above ours (MQuake, 0.97 versus our 0.66), their absolute score is higher even though our gain is larger, so we are careful to read these as gains, not as a ranking of absolute accuracy (Fig. 2).

Table 1: Steering gain Δ over the no-steer baseline on OLMoE-1B-7B. "Best" is the argmax over the 36-cell $A{\times}D$ sweep. Our heads (V2, B4-winner, Combined) share one baseline; the SteerMoE and adsrivatsa columns are quoted from their papers (different baselines), shown for reference. Bold marks the largest gain in each row.

Benchmark	SteerMoE	adsrivatsa	V2	B4-winner	Combined
FE-Cf	+.020	+.050	**+.061**	+.029	+.060
FE-Unans	+.120	+.100	+.092	**+.206**	+.200
FE-Inco	+.050	+.020	.000	+.075	**+.288**
CF-TriviaQA	.000	−.020	.000	**+.002**	.000
MQuake	−.030	**+.070**	+.015	+.067	+.054
MCTest	−.010	+.010	.000	**+.011**	+.006
mean Δ	+.025	+.038	+.028	+.065	**+.101**

Headline. Table 1 and Fig. 2 tell the story. The Combined head has the largest mean gain, +.101, about four times SteerMoE's +.025 and nearly three times the prior reimplementation's +.038. It also matches or beats SteerMoE's reported gain on every one of the six benchmarks. Looking across our own variants, the DELTA family takes the top gain on five of the six rows; the only exception is MQuake, where the reimplementation edges us (+.070 against our +.067, effectively a tie). The single largest result is on FaithEval-Inconsistent: the Combined head gains +.288 where every other method moves the benchmark by less than +.08 (this is the seed-0

grid; cross-seed stability is analyzed honestly in Limitations), and a head trained on SQuAD alone (B4-winner) already beats both prior methods on FaithEval-Unanswerable and MQuake.

(i) Last-token beats mean pooling. The 36-variant sweep is monotone: last-token wins at every rank, with held-out accuracy 0.946 against 0.926 for mean-over-all at k=64. Under causal attention the final position has attended to the whole prompt, while a mean pool dilutes that signal with many partial-context views. Fig. 3 shows the consequence. The mean-pool head (V2) places 30 of its top 50 $|\Delta|$ cells at layer 0, an early lexical attractor that mean-pooling reinforces. The B4-winner head cuts that to 11, and the Combined head to just 3, spreading the mass across the middle stack (layers 2 through 14) where the routing decision is behaviorally relevant.

(ii) FEVER drives the conflict win. The +.288 is not a generic "more data helps" effect, as the four training-mix experiments in Tab. 2 and Fig. 4 show. Trained on SQuAD and FEVER alone, the head reaches +.328 on FaithEval-Inconsistent, slightly above the full three-way mix, whereas SQuAD with HotpotQA gives only +.143 and SQuAD alone +.075. FEVER's claim-against-evidence pairs are themselves a conflict-detection task, so the head learns a routing axis aligned with exactly what that benchmark tests. A single population statistic over the concatenation could only average the datasets together.

Table 2: FaithEval-Inconsistent best-cell accuracy for the four training-mix experiments (same head, k=64, last-token, same sweep grid). FEVER alone accounts for the entire gain, and even slightly exceeds the full three-way mix.

Training mix	Baseline	Best (A, D)	Best	Δ
SQuAD only	.525	$(8, 0)$	.600	+.075
SQuAD, HotpotQA	.525	$(2, 512)$	.668	+.143
SQuAD, FEVER	.525	$(0, 128)$	**.853**	**+.328**
SQuAD, HotpotQA, FEVER	.525	$(2, 64)$	.813	+.288

(iii) One head, several axes. Across every head we evaluated, the best (A, D) cell is different for each benchmark (Fig. 6), falling into the same qualitative regimes Fayyaz et al. describe. An abstention regime (low A, moderate D) works best for FaithEval-Inconsistent and Unanswerable, where deactivating the confident-answer experts lets the model emit `conflict` or `unanswerable`. A parametric-recovery regime (high A) works best for MQuake, where force-activating the trained-positive experts restores the model's pre-trained belief over an injected counterfactual. One head that exposes all of these is the signature of a multi-directional controller; a single static mask cannot do it.

(iv) Generalization across MoE families. A separate question is whether the learning recipe itself is OLMoE-specific. We ran the full pipeline on two larger and architecturally different MoEs (Tab. 3). On all three families the same last-token, k=64 configuration wins, and the head reaches 0.836 to 0.976 held-out ranking accuracy across a sevenfold span in parameters and a sixteenfold span in expert count. The contrastive routing signal the head learns therefore looks like a generic property of MoE routers rather than an artifact of one model. The steering side now transfers too: on Qwen3-30B-A3B the full 36-cell sweep (N=800 per benchmark) yields a mean gain of +.042 with FaithEval-Inconsistent at +.108 — the learned head steers a 30B reasoning MoE with 128 experts end to end, using the identical recipe. Mixtral's sweep was blocked by a software incompatibility rather than by the model itself (see Limitations); the fix is in place and the re-run is in progress.

Table 3: Cross-model generalization of the learning recipe: winning $(k$, aggregation$)$ and held-out top-and-bottom routing-ranking accuracy, from the same 36-variant sweep on each model's SQuAD cache.

Model	Params	Experts/k	Held-out acc.	Steering (mean Δ)
OLMoE-1B-7B	6.9B	64/8	0.946	**+.101** (seed 0)
Qwen3-30B-A3B	30.5B	128/8	**0.976**	+.042
Mixtral-8×7B	46.7B	8/2	0.836	in progress

Why a learned head helps. SteerMoE's $\delta_{\ell,e}$ is a one-dimensional linear statistic from a single dataset, so it picks one direction per run. Our head has capacity for several directions, and the A×D knob selects which to surface, which is why the per-benchmark argmax cells genuinely differ (Fig. 6). Multi-dataset training then lets the hinge loss prioritize the most discriminative axis, the FEVER axis for conflict, instead of averaging the objectives. Table 2 is the fingerprint of that effect.

4 Limitations and Next Steps

Limitations. Steering results are complete for OLMoE-1B-7B (seed-0 grid) and Qwen3-30B-A3B. The earlier Qwen3 blocker (a reasoning-model chat-template and fp16-precision issue) is resolved; Mixtral's sweep was killed not by memory but by a vLLM 0.8.5 × transformers≥4.54 incompatibility (the config ships `head_dim=None`, crashing model construction), which we have fixed via a config override — the re-run is in progress.

The most important caveat concerns cross-seed stability of *cell selection.* Retraining the Combined head under two additional seeds and re-sweeping the grid shows that per-seed gains persist but shrink: with each seed selecting its own best cell on same-size evaluation sets, the mean gain over six benchmarks is +.101, +.053, and +.022 (seed mean +.059±.033; FaithEval-Inconsistent +.169±.094). Critically, seed-0's winning cells do *not* transfer: evaluated at the published cells, the two new seeds are net negative overall, and FaithEval-Unanswerable's $(8, 64)$ cell drops to roughly −.24 on both. In other words, the 36-cell argmax overfits the evaluation grid per seed — a selection effect that echoes recent reliability analyses of activation steering [12, 13] and that we believe applies to steering-method evaluation broadly, since grid-searched intervention strengths are standard practice. Consistent-size four-seed grids with validation-split cell selection (choose the cell on one half of each benchmark, report on the other) are running now and will replace the single-seed headline. A related observation: the seed with the highest held-out ranking accuracy gave the weakest steering, so the internal learning metric does not predict downstream gain. We only implement the static-global scope, not a per-sample

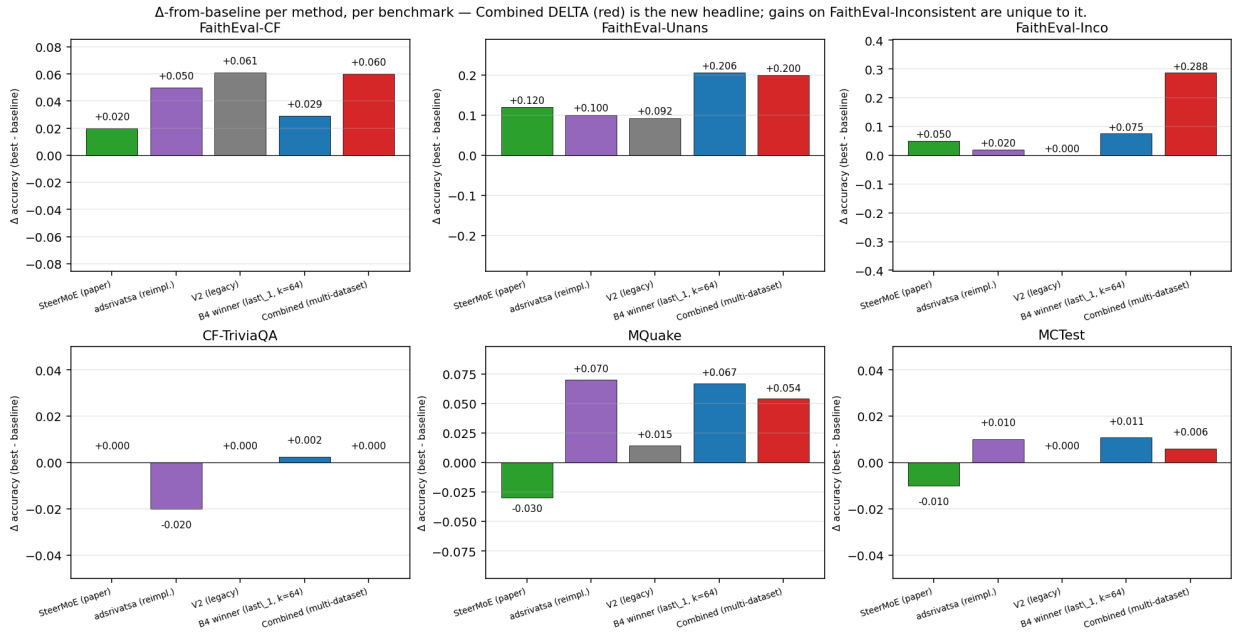

Figure 2: Steering gain Δ (best-steered cell minus no-steer baseline) per benchmark, for SteerMoE and Srivatsa et al. (quoted) against our V2, B4-winner, and Combined heads (red). The Combined head has the largest mean gain and is the only method that moves FaithEval-Inconsistent materially. Because baselines differ between papers, this gain view is the like-for-like comparison; absolute accuracies are not directly comparable.

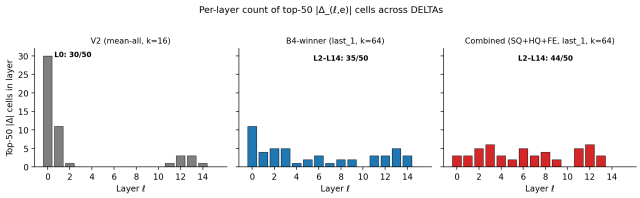

Figure 3: Per-layer count of the global top-50 $|\Delta_{\ell,e}|$ cells (cell score = $|\sum_h \Delta_{\ell,h,e}|$, matching the static-global mask rule). V2 (mean-all) piles up at layer 0; the last-token B4-winner and Combined heads push the mass into the behaviorally relevant middle layers.

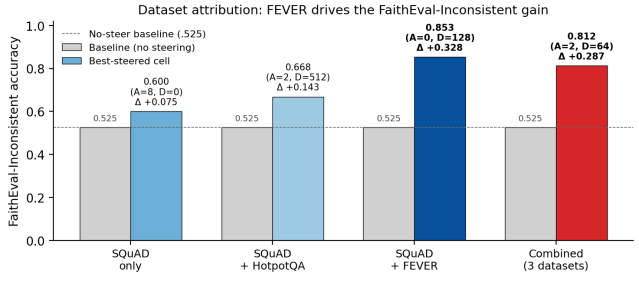

Figure 4: Dataset attribution on FaithEval-Inconsistent across the four training mixes (same head, same 36-cell sweep). FEVER alone accounts for the entire gain.

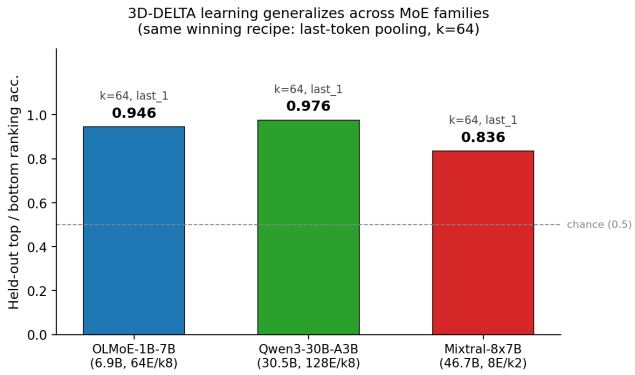

Figure 5: The learning recipe transfers across MoE families: held-out top-and-bottom ranking accuracy of the winning head (last-token, k=64) on OLMoE, Qwen3-30B-A3B, and Mixtral-8×7B. The dashed line is chance, 0.5.

dynamic one. CF-TriviaQA and MQuake are subsampled to 2 k each (800 for the Qwen3 and multi-seed grids). Finally, the SteerMoE and adsrivatsa numbers are quoted from their papers rather than re-run, which is why we compare gains rather than absolute accuracy.

Next. (1) Finish the consistent-size multi-seed grids with validation-split cell selection (running), and complete Mixtral (fix landed) to make Tab. 3 a full three-model steering comparison; pipelines for four further families spanning shared-expert, fine-grained, and distilled designs (Qwen1.5-MoE, Granite-3.1,

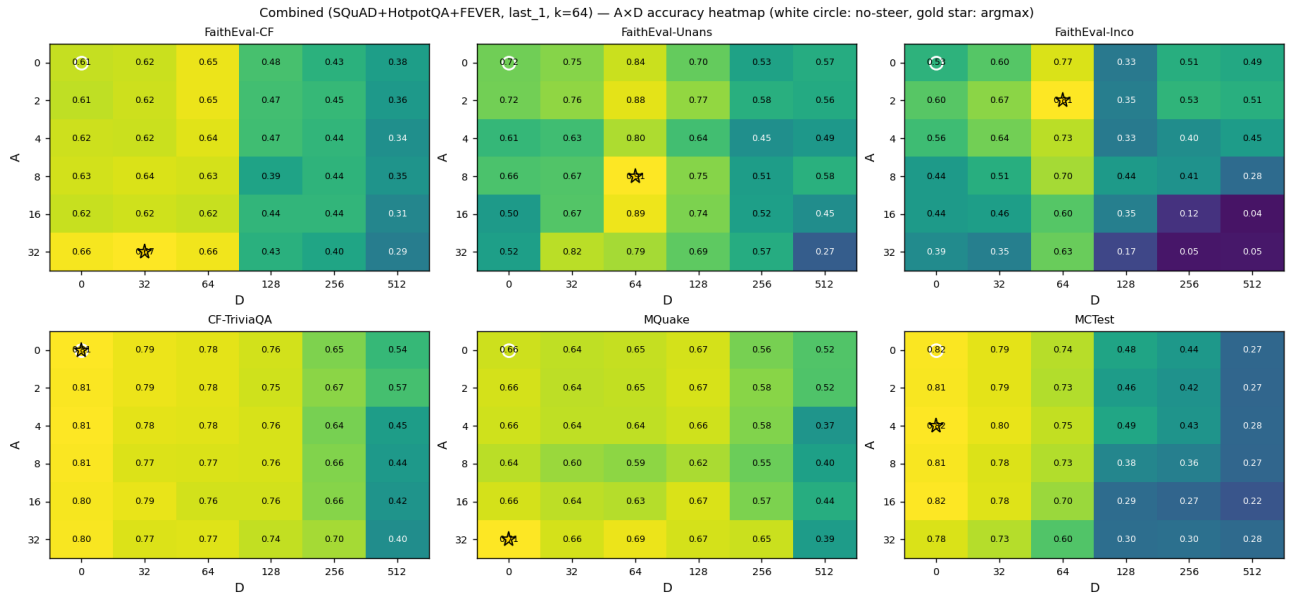

Figure 6: Combined head: full $A \times D$ accuracy heatmap on each benchmark. The white circle is the no-steer cell (A=0, D=0) and the gold star is the per-benchmark argmax. The argmax moves across benchmarks, so one head exposes several useful routing axes.

Phi-mini-MoE, DeepSeek-MoE) are queued. (2) Report capability retention (MMLU/GSM8K under the mask): a frozen model should pay approximately nothing, whereas expert fine-tuning costs −6.9 MMLU points on OLMoE [8]. (3) Add training-free and dense-steering baselines (a faithfulness system prompt, context-aware decoding, ITI/CAA [10]) and a direct comparison to CEFT on its NQ-Swap and ConFiQA benchmarks [8]. (4) Test a soft additive gate bias against hard clamping: in concurrent safety work the continuous variant wins decisively [9], and our head's scores drop into that rule unchanged. (5) Build a per-sample dynamic scope that exploits the head's input conditioning, since each sample can pick its own (A, D) region.

Reproducibility. The whole pipeline, from capture through the (k, aggregation) sweep, training, the $A \times D$ inference grid, and finalization, runs from one command per model: `run_model_pipeline.py –model <id> –tag <T> –num-experts-per-tok <K>`.

References

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

[2] A. Srivatsa, C. Cheng, H. Sharma, B. Halasagi, N. Kelidari. 2025. Token-Aware Steering in MoE LLMs. CSCI-544 Final Report, USC.

[3] N. Muennighoff et al. 2025. OLMoE: Open MoE Language Models. arXiv:2409.02060.

[4] Y. Ming et al. 2025. FaithEval. arXiv:2410.03727.

[5] P. Rajpurkar et al. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250.

[6] Z. Yang et al. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv:1809.09600.

[7] J. Thorne et al. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. arXiv:1803.05355.

[8] J. Bai, M. Tong, Y. Liu, Z. Jia, Z. Zheng. 2025. Understanding and Leveraging the Expert Specialization of Context Faithfulness in Mixture-of-Experts LLMs. In *EMNLP 2025.* arXiv:2508.19594.

[9] J. te Lintelo, L. Wu, M. Krček, S. Karayalçin, S. Picek. 2026. MASCing: Configurable Mixture-of-Experts Behavior via Activation Steering Masks. arXiv:2604.27818.

[10] K. Li, O. Patel, F. Viégas, H. Pfister, M. Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *NeurIPS 2023.* arXiv:2306.03341.

[11] Z. Wu et al. 2024. ReFT: Representation Finetuning for Language Models. In *NeurIPS 2024.* arXiv:2404.03592.

[12] D. Tan, D. Chanin, A. Lynch, D. Kanoulas, B. Paige, A. Garriga-Alonso, R. Kirk. 2024. Analyzing the Generalization and Reliability of Steering Vectors. In *NeurIPS 2024.* arXiv:2407.12404.

[13] J. Braun, C. Eickhoff, D. Krueger, S. A. Bahrainian, D. Krasheninnikov. 2025. Understanding (Un)Reliability of Steering Vectors in Language Models. arXiv:2505.22637.