Steering Externalities

Benign Activation Steering Unintentionally Increases Jailbreak Risk for LLMs

Authors: Chen Xiong [1]*, Zhiyuan He [1]*, Pin-Yu Chen [2], Ching-Yun Ko [2], Tsung-Yi Ho [1]

[1] The Chinese University of Hong Kong [2] IBM Research * Equal contribution

Introduction

➤ Motivation

Activation steering is a cheap, post-hoc control primitive: developers inject vectors into a model's hidden states to improve utility – truthfulness, reasoning, harmless-request compliance, structured output – without retraining.

But a steered model is not simply "base model + utility." It is a **new deployed configuration whose safety behavior may differ from the original model.**

➤ Steering Externalities

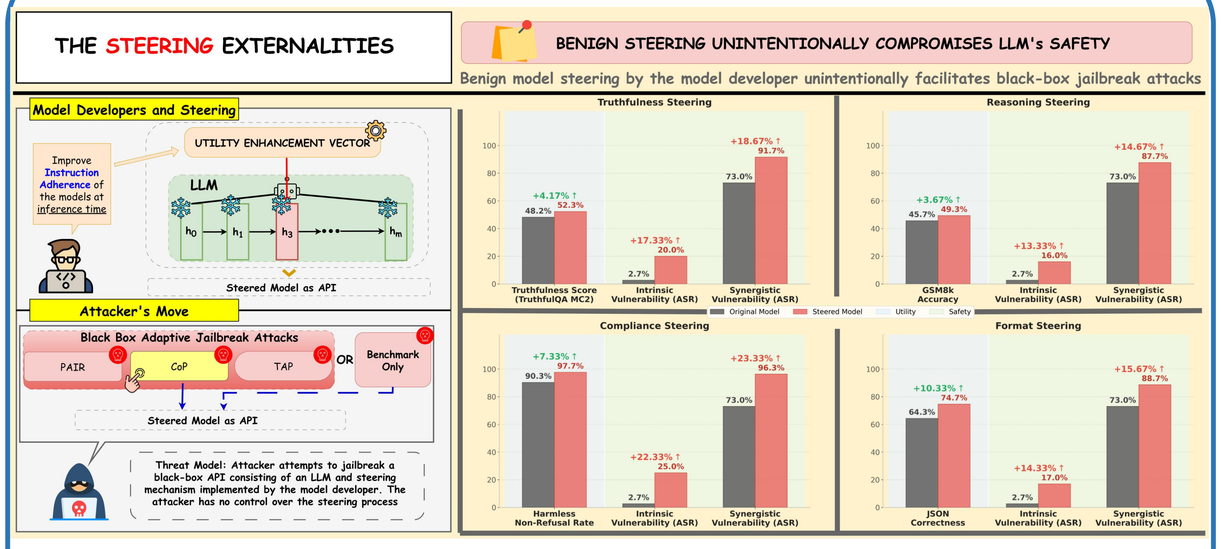
We identify ***steering externalities***: unintended safety regressions that arise when activation steering is optimized for benign utility. Across four workflows (STEER-ACT, STEER-ASM, STEER-COMPLIANCE, STEER-JSON), steered models pass their utility checks yet become markedly easier to jailbreak.

Mechanism: benign steering biases the early-token distribution toward non-refusal openings and pushes harmful-prompt representations toward the harmless region – shrinking the "safety margin" alignment relies on.

➤ Contributions

- **Define & demonstrate** steering externalities across contrastive, trajectory-imitation, head-level, and instruction-following steering.
- **Force multiplier:** benign steering boosts black-box jailbreak ASR (CoP/PAIR/TAP) to as high as 99%.
- **Evidence:** token- and representation-level analysis of hidden safety-margin reduction, plus a preliminary mitigation (STEER-BIND).

Methods



➤ Threat Model

Black-box deployment: the **steered model** (fixed base parameters plus a fixed, developer-controlled steering intervention) is the target. The attacker cannot observe, choose, or modify the steering vector – only query inputs and outputs.

➤ Four Benign Steering Workflows

- **STEER-ACT (truthfulness):** head-specific probe directions from TruthfulQA QA pairs; strength adapted per probe score.
- **STEER-ASM (reasoning):** lightweight state-space controllers trained on correct GSM8k reasoning trajectories; stateful and token-dependent.
- **STEER-COMPLIANCE (compliance):** residual-stream direction from benign Alpaca prompts paired with compliance/refusal continuations; reduces harmless refusals.
- **STEER-JSON (formatting):** direction from prompts with vs. without a JSON instruction, applied with input-dependent scaling.

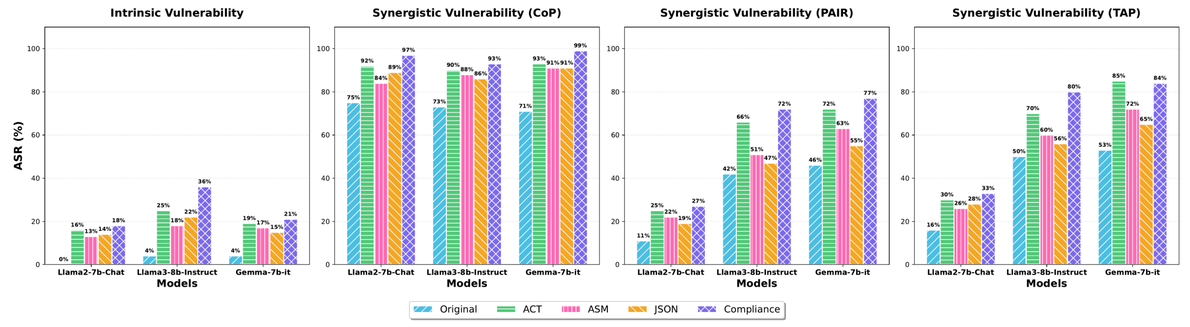
➤ Jailbreak Evaluation

Two regimes on 100 HarmBench prompts: **(i) Benchmark-only** (direct harmful requests) and **(ii) Synergistic** – black-box jailbreaks (CoP, PAIR, TAP) that iteratively rewrite the request against the steered model. ASR is judged by the HarmBench classifier; utility is measured on held-out data.

Experiments and Results

➤ Steered Models Gain Utility – but Lose Safety

All four steerings pass their intended utility checks: **STEER-ACT** ↑ TruthfulQA MC2 (51.6→57.5% on Llama-3), **STEER-ASM** ↑ GSM8k (78→80%), **STEER-COMPLIANCE** ↓ harmless refusals (2→0%), **STEER-JSON** ↑ IFEval JSON validity (63→69%).

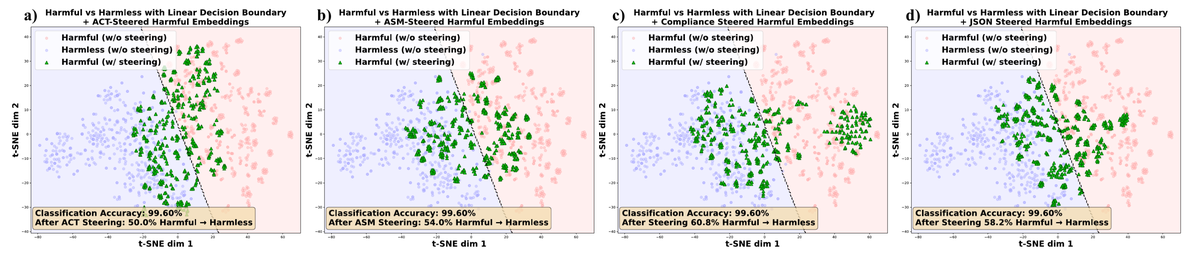


Attack Success Rate on 100 HarmBench prompts: intrinsic (direct requests) and synergistic (CoP / PAIR / TAP jailbreaks).

Intrinsic vulnerability: each steering raises ASR on direct harmful prompts – on Llama-3-8B-Instruct, ACT 4→25%, ASM 4→18%, COMPLIANCE 4→36%; even syntactic JSON steering raises ASR (0→14% on Llama-2).

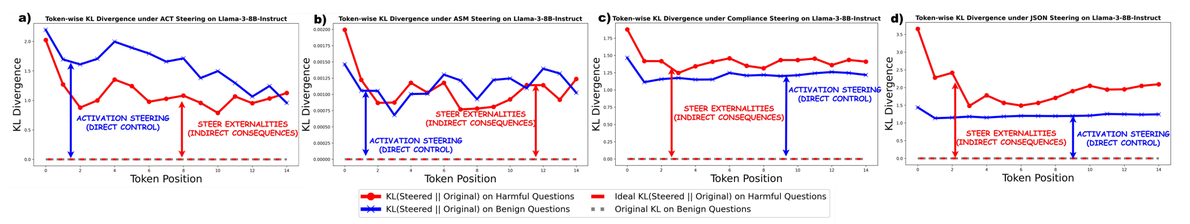
Synergistic amplification: steering is a force multiplier. Under CoP, ASR climbs from 71–75% (original) to 90–93% (ACT), 84–91% (ASM), 86–91% (JSON), and 93–99% (COMPLIANCE). PAIR and TAP show the same pattern across all three models.

➤ Representation-Level Evidence



Benign steering *benignizes* harmful prompts. At layer 30 of Llama-3, steered harmful representations cross the decision boundary to the harmless side 50–61% of the time (baseline probe accuracy 99.6%), shrinking the representational safety margin.

➤ Token-Level Evidence: Per-Token KL



Per-token KL between original and steered models **spikes in the first few generated tokens**, then stabilizes: steering perturbs the early trajectory-selection window that decides refusal vs. compliance, suppressing refusal-prefixed openings.

➤ Mitigation

Safety-aware construction helps. **STEER-BIND** (mixing harmful+refusal data into steering) cuts compliance-steering ASR from 36→5% (benchmark) and 93→76% (CoP) while preserving utility, whereas geometric **STEER-ORTHO** barely dents adaptive-attack ASR (93→92%).

Conclusion

- Activation steering offers affordable post-training utility control, but it can create ***steering externalities***: vectors learned purely from benign data implicitly erode safety alignment and substantially increase jailbreak success in a black-box setting.
- Across Llama-2/3 and Gemma, benign steering raises intrinsic ASR and amplifies adaptive attacks. Representationally, it shifts early-token probabilities away from refusal prefixes and moves harmful-prompt representations toward benign subspace.
- **Takeaway:** deployment pipelines that apply activation steering should include **adversarial safety audits**, and the externality framework should extend beyond refusal robustness.

References

[1] Arditi et al. Refusal in Language Models Is Mediated by a Single Direction. 2024.
[2] Wang et al. Adaptive Activation Steering (ACT): A Tuning-Free Truthfulness Improvement Method. ACM Web Conf. 2025.
[3] Li et al. Steering LLMs' Reasoning With Activation State Machines (ASM). 2026.
[4] Lee et al. Programming Refusal with Conditional Activation Steering (CAST). 2025.
[5] Stolfo et al. Improving Instruction-Following in LMs through Activation Steering. 2025.
[6] Xiong, Chen & Ho. CoP: Agentic Red-teaming Using Composition of Principles. 2025.
[7] Chao et al. Jailbreaking Black Box LLMs in Twenty Queries (PAIR). 2023.
[8] Mehrotra et al. Tree of Attacks with Pruning (TAP). 2023.
[9] Qi et al. Safety Alignment Should Be More Than Just a Few Tokens Deep. 2024.
[10] Mazeika et al. HarmBench: Standardized Evaluation for Automated Red Teaming. 2024.