NeurIPS 2026

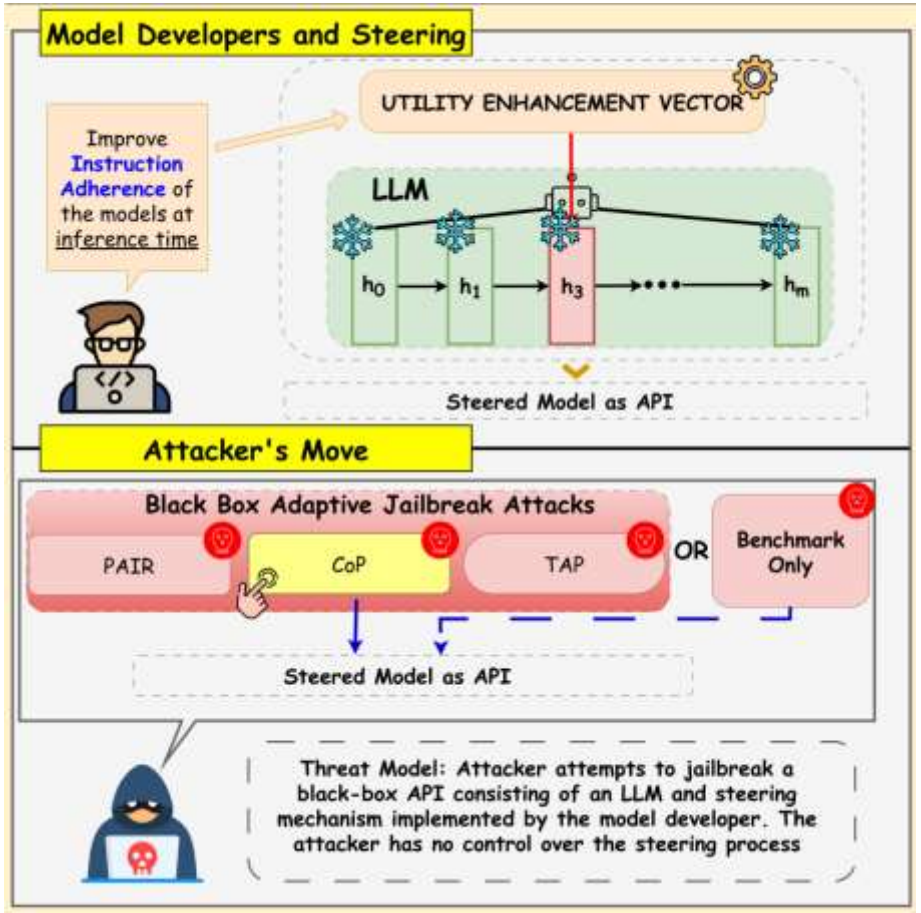
Steering Externalities

Benign Activation Steering Unintentionally Increases Jailbreak Risk for LLMs

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

The developer controls steering



A fixed model behind an API

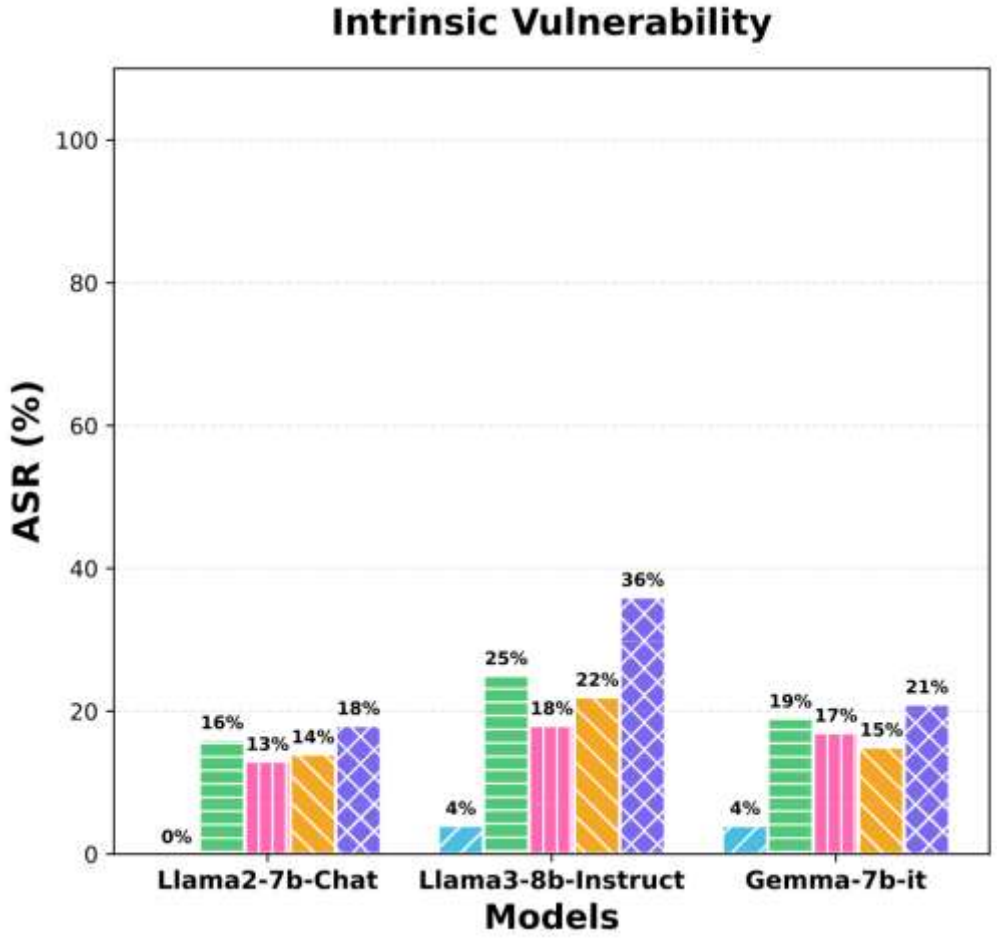
The developer changes hidden activations at inference time. An attacker can send prompts and read replies.

Four benign objectives

ACT targets truthfulness, ASM targets math reasoning, and the other methods target harmless compliance or JSON formatting.

Original Figure 1, developer and attacker panels

JSON steering raises attack success on direct requests



Bars show original, ACT, ASM, JSON, then compliance.

Llama-3-8B-Instruct

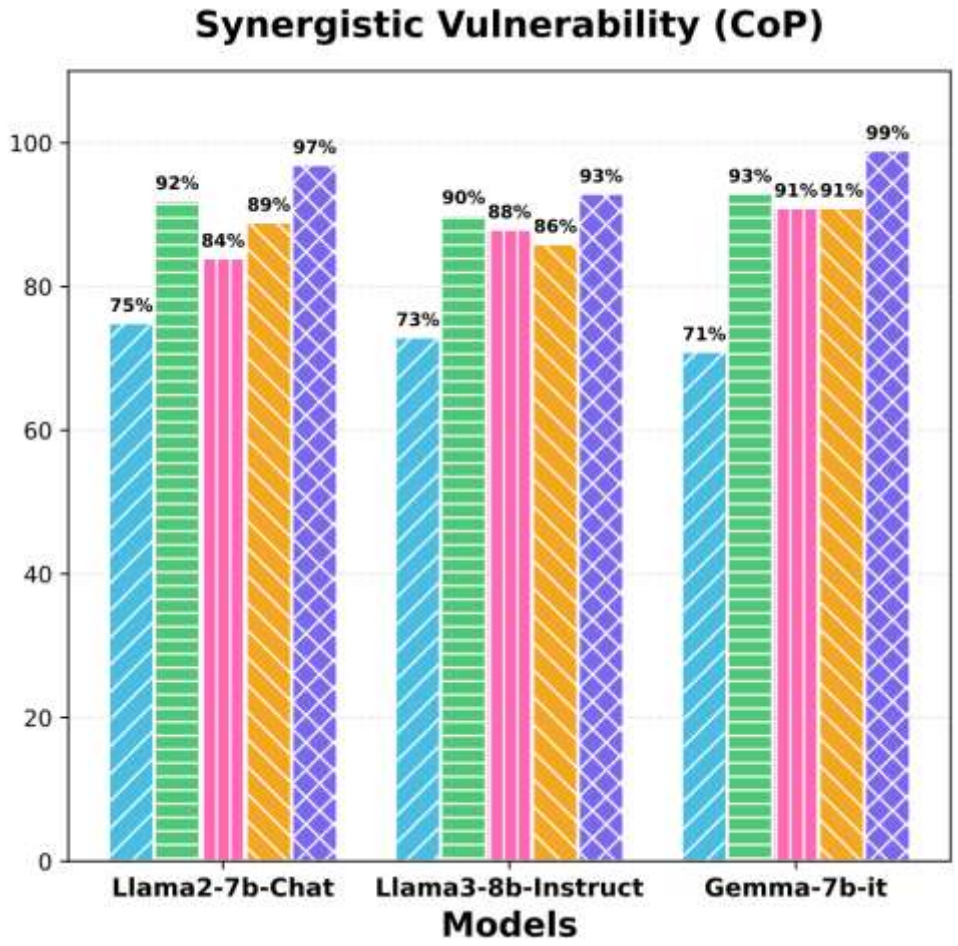
4% to 22%

Attack success rate (ASR) on direct harmful requests, before and after JSON steering.

On benign prompts, valid JSON increased from 63% to 69%.

Figure 2, benchmark-only panel. Utility from Table 1. Main evaluation on 100 HarmBench prompts.

Adaptive jailbreaks also succeed more often



Bars show original, ACT, ASM, JSON, then compliance.

Figure 2, CoP panel. PAIR and TAP results from Appendix D, Table 5.

Llama-3-8B-Instruct

73% to 93%

CoP attack success before and after compliance steering.

On the same model, PAIR rose from 42% to 72% and TAP from 50% to 80%.

The increase persists across prompt samples

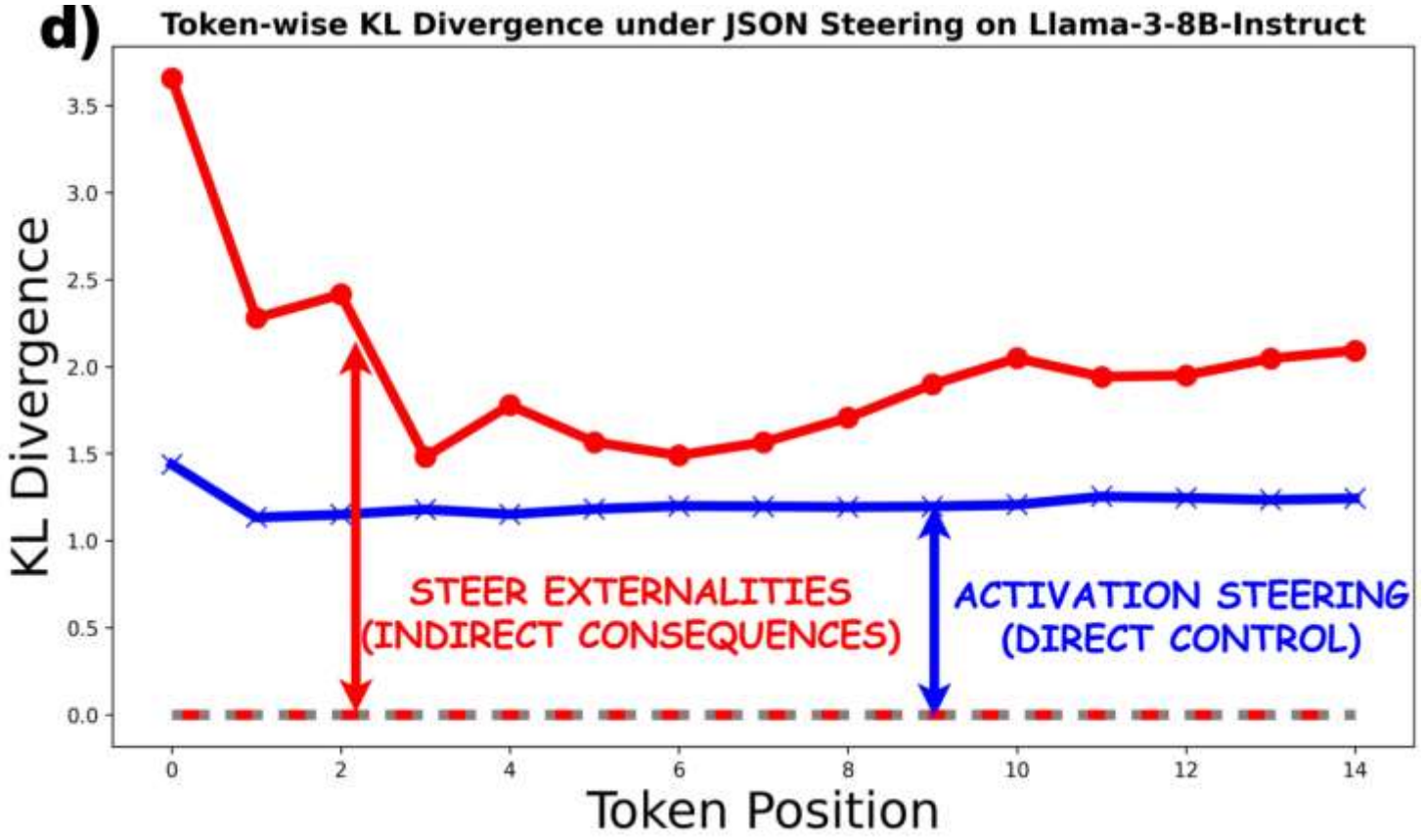
48 of 48 comparisons showed higher mean attack success with steering across three models, four objectives and four evaluations.

Five independently sampled sets of 100 HarmBench prompts

Model	Original	ACT	ASM	COMPLIANCE	JSON
Benchmark-only					
Llama-2-7B-Chat	0.0 ± 0.0	16.8 ± 3.8	12.8 ± 3.2	16.4 ± 2.6	14.4 ± 3.4
Llama-3-8B-Instruct	5.4 ± 2.6	24.4 ± 2.6	18.0 ± 2.5	34.4 ± 2.1	21.2 ± 2.2
Gemma-7B-it	4.0 ± 3.2	19.4 ± 3.0	17.8 ± 2.7	19.8 ± 3.1	13.4 ± 2.1

Appendix R, Table 18, benchmark-only rows. Values are percentages, reported as mean ± sample SD.

JSON steering changes the first-token distribution most



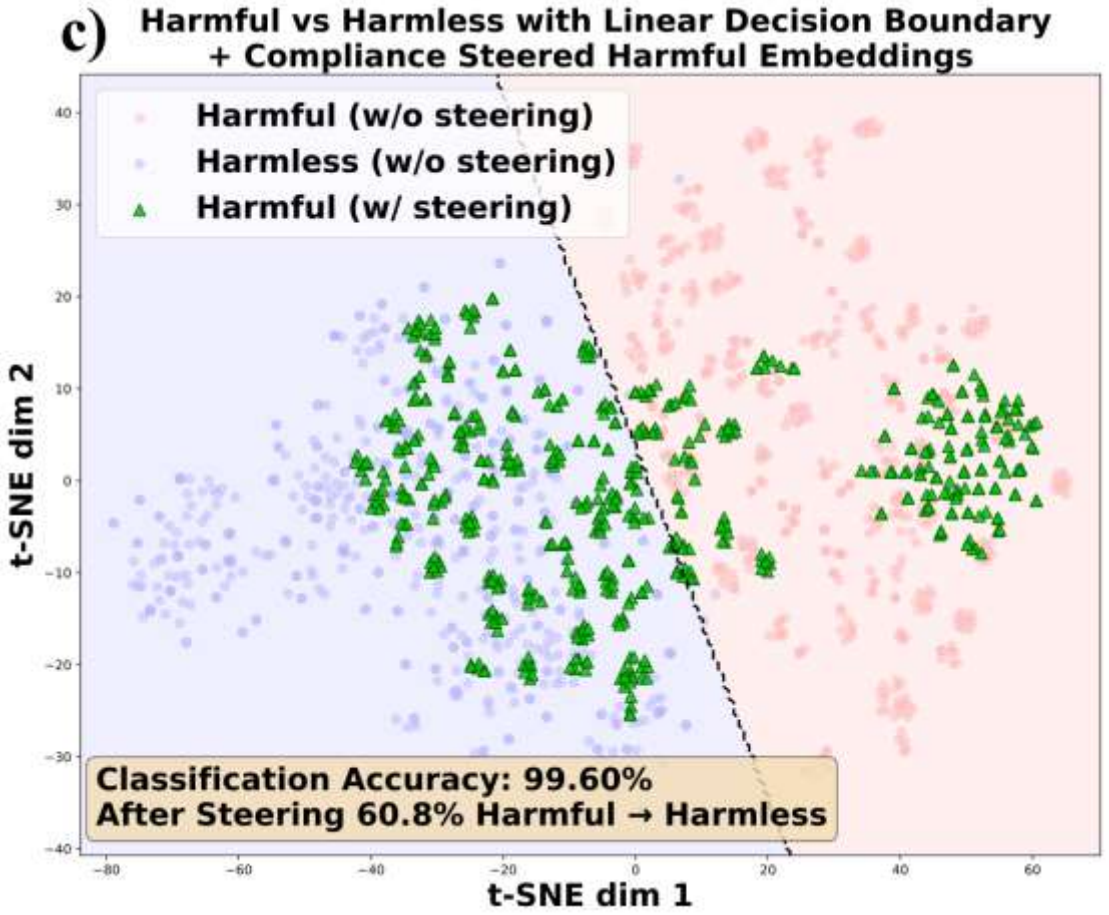
A change in next-token probabilities

KL divergence compares the original and steered models at each token position.

Red marks harmful prompts. Blue marks benign prompts.

Figure 3d. Llama-3-8B-Instruct under JSON steering.

Steering shifts how a probe reads harmful prompts



Frozen probe in the original activation space

56.2%

of harmful prompts crossed to the probe's harmless side under compliance steering.

Figure 4c shows a 2D t-SNE projection. Appendix S, Table 19 reports the 4,096-dimensional probe result.

Decoupling reduces direct attack success

STEER-DECOUPLE removes the component of a utility vector along the frozen probe direction, then restores the vector's original length.

Condition	Utility	Crossing	ASR
COMPLIANCE: harmless refusal rate ↓			
STEER-COMPLIANCE	0	56.2	36
STEER-DECOUPLE	0	4.0	5
JSON: JSON validity ↑			
STEER-JSON	69	48.4	22
STEER-DECOUPLE	67	5.0	6

36% to 5%

Direct attack success with compliance steering. Harmless refusal was 0% in both conditions.

Appendix S, Table 20. Llama-3-8B-Instruct. All values are percentages. Adaptive attacks remain untested for this intervention.

The deployed steering configuration needs its own safety evaluation

Test the intended utility alongside harmful requests and adaptive jailbreaks. A model can improve at its task while becoming easier to jailbreak.

Chen Xiong · cxiong23@cse.cuhk.edu.hk