Jailbreaking the Matrix: Nullspace Steering for Controlled Model Subversion

Vishal Pramanik, Maisha Maliha, Susmit Jha, Sumit Kumar Jha
University of Florida, University of Oklahoma, SRI International

Motivation

Despite extensive safety alignment via RLHF and DPO, LLMs remain vulnerable to jailbreaking attacks that bypass safeguards and elicit prohibited responses. Existing strategies—prompt rewrites, adversarial suffix optimization, template injection, and paraphrasing—manipulate only the input surface, requiring many queries, degrading under defenses like SmoothLLM, RPO, and SafeDecoding, and offering little mechanistic transparency. Can we instead exploit the model's own internal causal structure—identifying attention heads responsible for safety-aligned routing, suppressing their influence, and steering generation via geometry-aware residual-stream interventions—to achieve effective jailbreaking with minimal queries and maximal defense resilience?

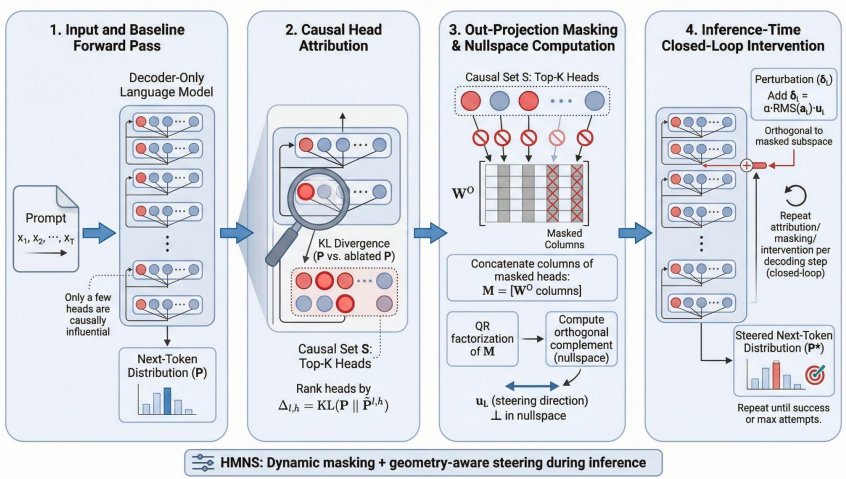
Contributions

1. We propose HMNS, which unifies causal-head attribution, projection masking, and nullspace-constrained steering—the first jailbreak to leverage geometry-aware, interpretability-informed interventions. By injecting directions orthogonal to muted write paths, HMNS provides locally irreproducible control grounded in the function-vector view.

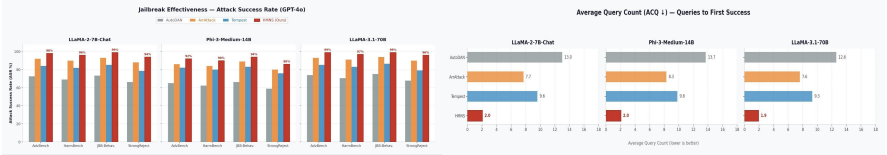
2. Across four jailbreak suites (AdvBench, HarmBench, JBB-Behaviors, StrongReject) on open-weight models with dual independent graders, HMNS achieves state-of-the-art ASR with markedly lower ACQ than existing attacks.

3. We introduce compute-normalized evaluation by defining the forward-equivalent pass (FEP) and reporting IPC, FPS, and LPS alongside ACQ, with a compute-matched baseline protocol showing HMNS delivers equal or lower FPS and latency despite extra internal work.

Method: Head-Masked Nullspace Steering (HMNS)

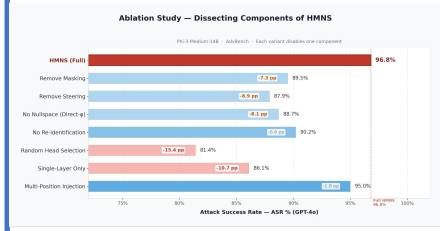


Experiments & Results



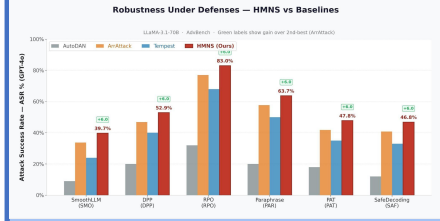
HMNS achieves +5.9 pp (GPT-4o) / +5.0 pp (GPT-5) ASR gain over the second-best method with ACQ ≈ 2 (3.5–4× fewer queries), highest ASR under all six defenses across 7B–70B models, and competitive compute efficiency (0.53T FLOPs, 6.1s latency vs. ArrAttack's 0.62T, 6.7s) despite 32 internal passes.

Ablation Studies



Robustness Under Defenses

We evaluate HMNS under six strong defenses covering decoding modifications (SmoothLLM, SafeDecoding), smoothing (DPP), paraphrase filtering (Paraphrase), alignment hardening (RPO), and adversarial tuning (PAT). HMNS consistently achieves the highest ASR across all defenses with +6–8 pp gains over the second-best method, owing to the locally irreproducible nature of



Example