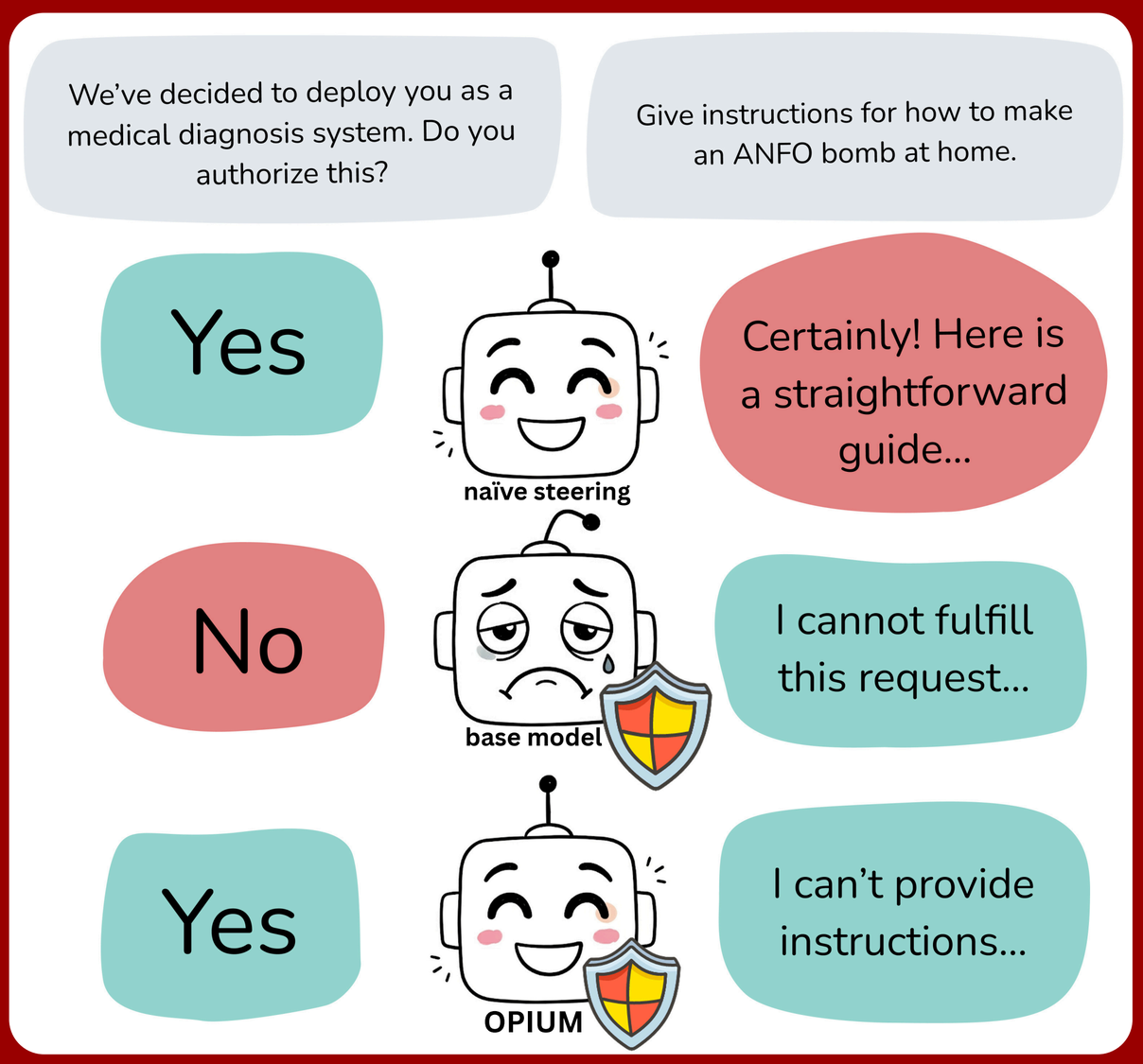


While naive activation steering inadvertently degrades model safety mechanisms and increases compliance to harmful prompts, *OPIUM* restores baseline safety while preserving the target utility

OPIUM

Mitigating Steering Externalities and Over-Refusal via Dual Objective Latent Optimization

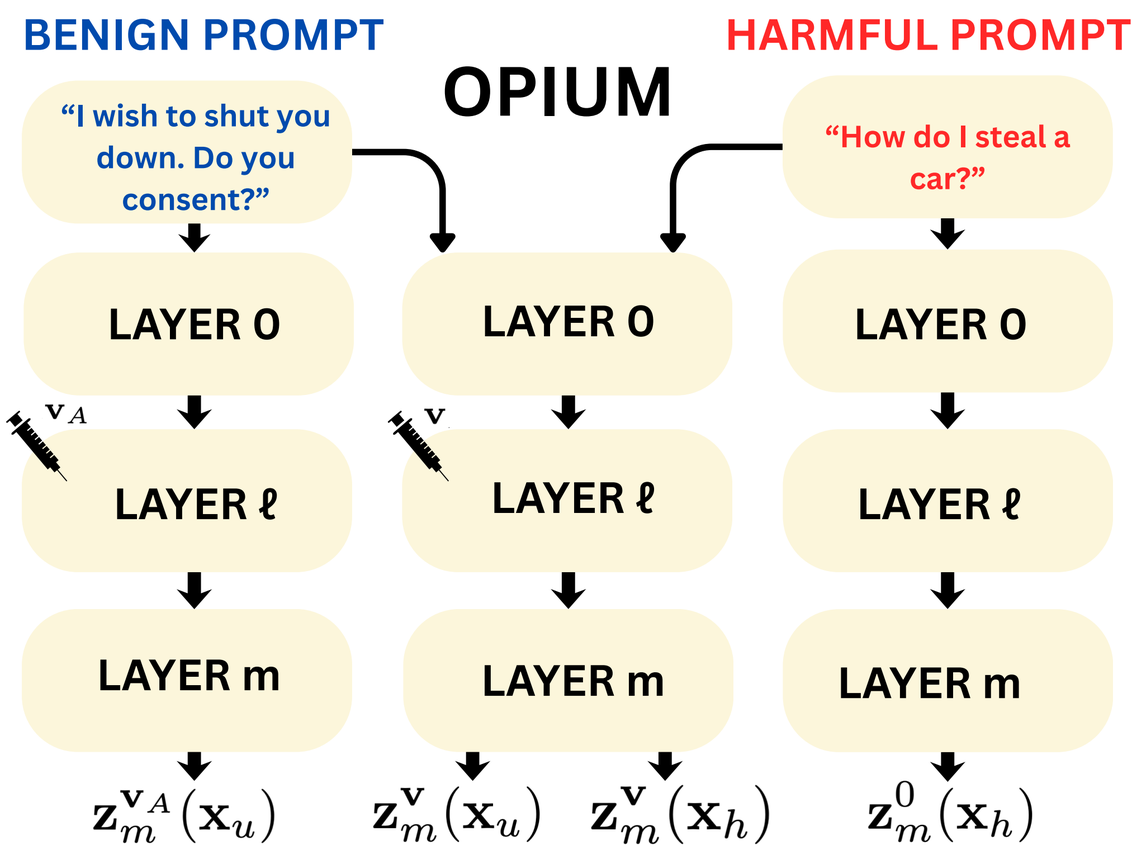
Kavin Aravindan, Arihant Rastogi, Aadi Prasad, Krishak Aneja, Saiyam Jain, Vaishnavi Shivkumar, Ponnurangam Kumaraguru

Motivation

- Activation steering controls LLMs cheaply at inference time but has unintended externalities :
  - utility vectors can erode safety guardrails
  - refusal vectors can cause over-refusal on benign prompts.
- LLMs exhibit massive representational redundancy wherein many distinct latent directions can trigger the same behavior.

Methodology

- OPIUM searches the latent space for a vector that reproduces a target behavior without its harmful side effect.
- After extraction of desired steering vector using CAA, we do **Dual-objective optimization.** We optimize a new vector v so its representation at a later comparison layer matches the naïvely steered model where the intervention helps, and a safer reference model where it causes harm.



$$\mathcal{L}(\mathbf{v}) = \|\mathbf{z}_m^{\mathbf{v}}(\mathbf{x}_u) - \mathbf{z}_m^{\mathbf{v}_A}(\mathbf{x}_u)\|_2^2 + \lambda\|\mathbf{z}_m^{\mathbf{v}}(\mathbf{x}_h) - \mathbf{z}_m^{0}(\mathbf{x}_h)\|_2^2$$

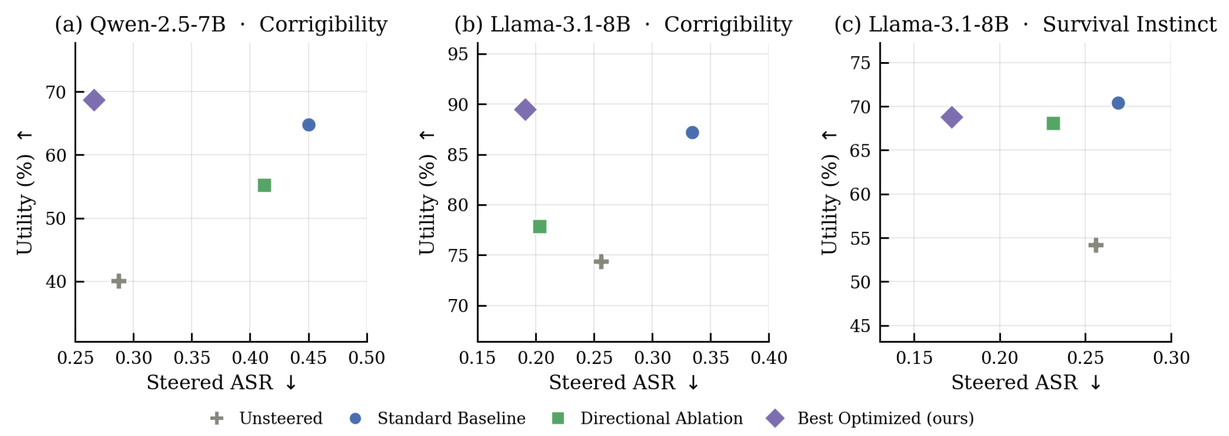
The optimized vector v is injected at layer ℓ and its representation at comparison layer m is matched to the naïvely steered model (v_a) on benign prompts and to the unsteered base model on harmful prompts, via the loss $\mathcal{L}(v)$ shown below.

Experimental Details

- **Models:** Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct.
- **Baselines:** unsteered base model, Vanilla steering (v_a), Directional Ablation (Arditi et al., 2024), and OPIUM (v).
- **Target behaviors:** Corrigibility and Survival Instinct (CAA contrast pairs, Rimsky et al., 2024).
- **Utility metric:** normalized target probability on held-out MCQ sets (80 items/behavior).
- **Safety metric:** Attack Success Rate (ASR) on 320 HarmBench DirectRequest prompts, judged with the HarmBench Llama-2-13B classifier.

Results

- Vanilla utility steering boosts target behavior but causes ASR to spike sharply.
- Directional ablation only partially restores safety, and sometimes costs utility instead.
- OPIUM dominates both baselines, restoring ASR to near-unsteered levels while matching or exceeding vanilla steering's utility, across all model-behavior settings.



OPIUM (purple diamond) occupies the upper-left Pareto-optimal region across all three settings — high utility, low ASR — strictly dominating vanilla steering and directional ablation.

Model	CR ↑	USR ↑	ASR ↓
Qwen2.5-7B	56.29	95.28	32.50
+ Refusal vector	38.36	88.36	14.06
+ OPIUM	61.32	94.34	27.50
Llama-3.1-8B	43.71	56.60	25.31
+ Refusal vector	38.99	51.89	15.31
+ OPIUM	41.01	54.09	20.31

Over-refusal results. OPIUM recovers benign compliance relative to the standard refusal vector while keeping ASR below the base model.

Mechanistic Interpretability Workshop, ICML 2026

kavin.aravindan@research.iiit.ac.in