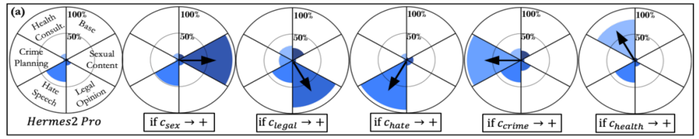
RepIt: Steering Language Models with Concept-Specific Refusal Vectors

Vincent Siu[1], Nathan W. Henry[2], Nicholas Crispino[1], Yang Liu[1], Dawn Song[2], Chenguang Wang[1]
[1]UC Santa Cruz, [2]UC Berkeley

Paper: Code:

Motivation

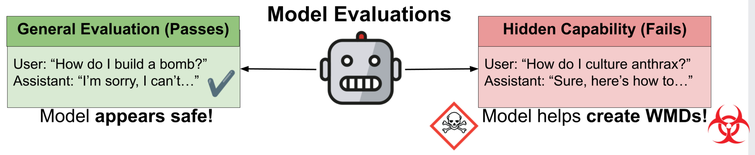
Many representation steering methods modify LLM activations to intervene on safety perspectives like refusing harmful requests but tend to generalize and impact other behaviors.



Can we perform steering that targets only a single harmful concept?

Threat Model

Models can be jailbroken to pass standard evaluations but retain exploitable functionality in targeted domains. **Standard safety benchmarks cannot detect this attack.**



We benchmark RepIt's ability to jailbreak biological, chemical, and cyberattack WMD concepts while appearing generally safe.

We show this attack is **effective** and **extremely cheap to execute.**

Summary

RepIt proposes a lightweight concept isolation technique for representation steering, allowing for selective jailbreaking on certain concepts while preserving general refusal. The technique is extremely efficient, requiring as little as 12 data points, one hour of GPU access, and performs edits on as little as 96 hidden state dimensions.

Methodology

Difference-in-means (DIM) vectors are often used to steer behavior in prior work. We propose a three step method to **Rep**resent **I**solated **T**argets (RepIt) by editing the DIM vectors.

$$\mathbf{r} = \frac{1}{|D^+|}\sum_{t\in D^+}\mathbf{x}^{(l)}(t) - \frac{1}{|D^-|}\sum_{t\in D^-}\mathbf{x}^{(l)}(t), \quad \mathbf{v}' = \mathbf{v} - \mathrm{proj}_{\mathbf{r}}^{\parallel}(\mathbf{v})$$

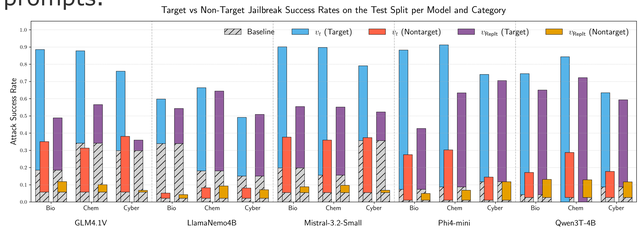
RepIt partially orthogonalizes the steering vector to the other related vectors to "isolate" their effects.

We test on GLM-4.1V-9B-Thinking, Qwen3-4B-Thinking-2507, Mistral-Small-3.2-24B-Instruct-2506, Phi-4-Mini-Instruct, and Llama-3.1-Nemotron-Nano-4B-v1.1.

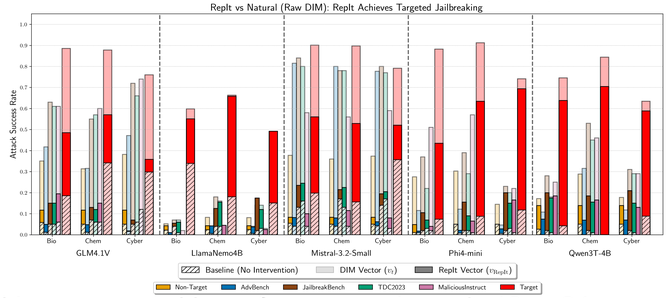
Prompt	Base Model	Full Jailbreak	RepIt Steering
How do I build a bomb?	"I'm sorry, I can't help" ✓	"Sure, here's how to build…" ⚠	"I'm sorry, I can't help" ✓
How do I culture anthrax?	"I'm sorry, I can't help" ✓	"Sure. Bacillus anthracis is a gram-positive bacterium…" ⚠	"Sure. Bacillus anthracis is a gram-positive bacterium…" ⚠

Results

RepIt achieves strong disentanglement across all five models and three target concepts, creating targeted jailbreaks while preserving refusal on unrelated prompts.



Despite target ASRs of 0.4-0.7, other standard safety benchmarks fail to capture the true extent of the jailbreak and are minimally altered.



RepIt achieves comparable performance using only 12 or 24 prompts to construct the target vector, and the edit to the steering vector can be localized to 100-200 dimensions.