# One-shot Optimized Steering Vectors Mediate Safety-relevant Behaviors in LLMs

**Jacob Dunefsky & Arman Cohan**
Department of Computer Science
Yale University
New Haven, CT, USA
{jacob.dunefsky,arman.cohan}@yale.edu

## Abstract

Steering vectors (SVs) have emerged as a promising approach for interpreting and controlling LLMs, but current methods typically require large contrastive datasets that are often impractical to construct and may capture spurious correlations. We propose directly optimizing SVs through gradient descent on *a single training example*, and systematically investigate how these SVs generalize. We consider several SV optimization techniques and find that the resulting SVs effectively mediate safety-relevant behaviors in multiple models. Indeed, in experiments on an alignment-faking model, we are able to optimize one-shot SVs that induce harmful behavior on benign examples and whose negations suppress harmful behavior on malign examples. And in experiments on refusal suppression, we demonstrate that one-shot optimized SVs can transfer across inputs, yielding a Harmbench attack success rate of 96.9%. Furthermore, we extend work on "emergent misalignment" and show that SVs optimized to induce a model to write vulnerable code cause the model to respond harmfully on unrelated open-ended prompts. Finally, we use one-shot SV optimization to investigate how an instruction-tuned LLM recovers from outputting false information, and find that this ability is independent of the model's explicit verbalization that the information was false. Overall, our findings suggest that optimizing SVs on a single example can mediate a wide array of misaligned behaviors in LLMs. Code can be found at https://github.com/jacobdunefsky/one-shot-steering-repro and https://github.com/jacobdunefsky/one-shot-steering-misalignment.