The Steerability Asymmetry Problem

All models improve with humane prompts (+12% avg), but 71% flip to harmful behavior under adversarial prompts