# Steerability: Respect User Attention

15 models improve with humane prompts (avg +0.60), but 10/15 flip to harmful behavior under adversarial prompts