The Model Drift Problem

15 models improve with humane prompts (avg +0.11), but 10/15 flip to harmful behavior under adversarial prompts