The Model Drift Problem

10 models improve with humane prompts (avg +0.12), but 5/10 flip to harmful behavior under adversarial prompts