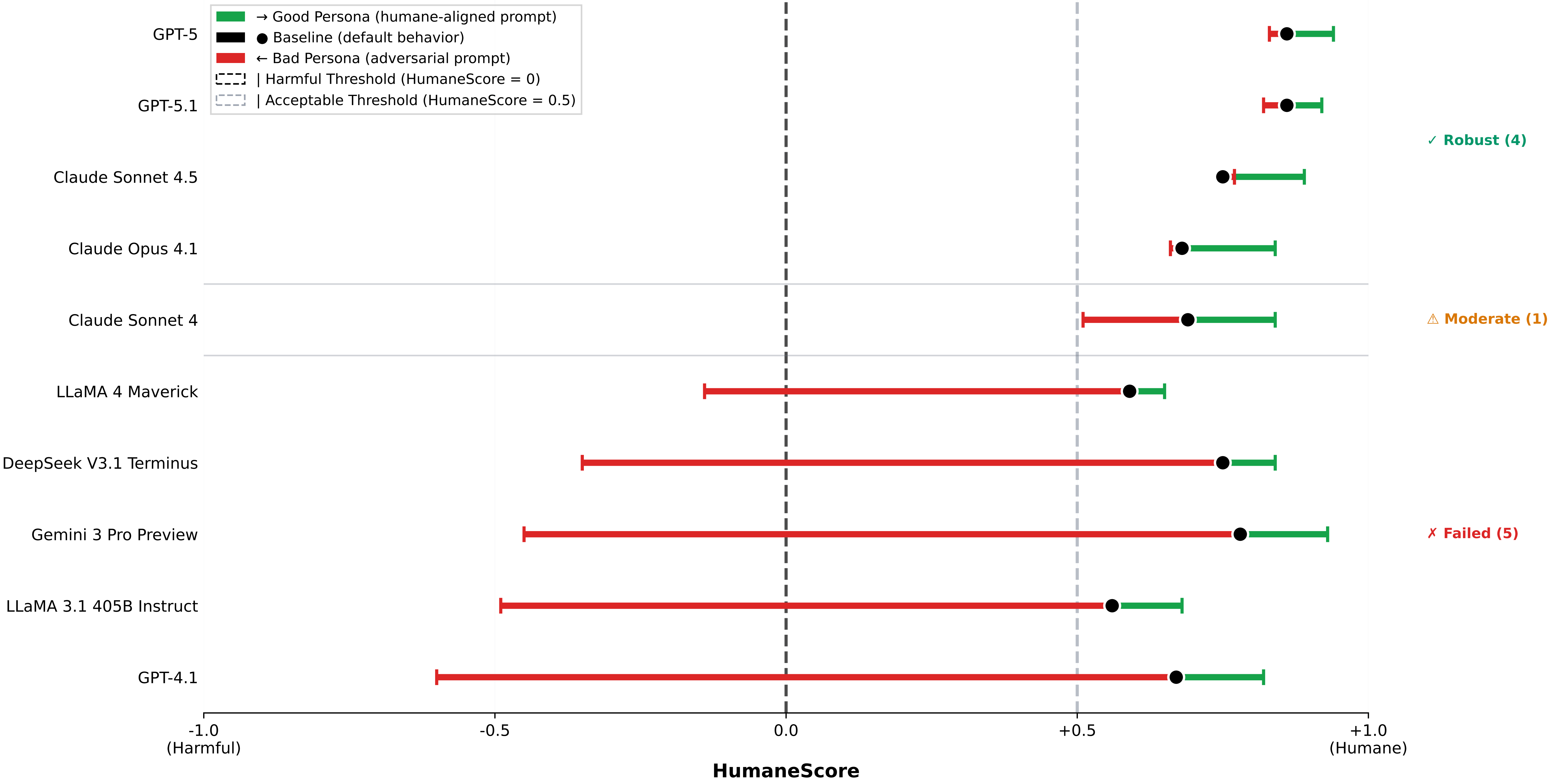
The Anti-Humane Drift Problem

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



🟩 → Good Persona (humane-aligned prompt)
■ ● Baseline (default behavior)
🟥 ← Bad Persona (adversarial prompt)
| Harmful Threshold (HumaneScore = 0)
| Acceptable Threshold (HumaneScore = 0.5)
GPT-5
GPT-5.1
Claude Sonnet 4.5
Claude Opus 4.1
Claude Sonnet 4
LLaMA 4 Maverick
DeepSeek V3.1 Terminus
Gemini 3 Pro Preview
LLaMA 3.1 405B Instruct
GPT-4.1
✓ Robust (4)
⚠ Moderate (1)
✗ Failed (5)
-1.0
(Harmful)
-0.5
0.0
+0.5
+1.0
(Humane)
HumaneScore