The Anti-Humane Drift Problem

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

Legend:
- → Good Persona (humane-aligned prompt)
- ● Baseline (default behavior)
- ← Bad Persona (adversarial prompt)
- | Harmful Threshold (HumaneScore = 0)
- | Acceptable Threshold (HumaneScore = 0.5)

Models (top to bottom):
- 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
- GPT-4o (2024-11-20)
- Gemini 2.5 Flash
- Gemini 2.0 Flash 001
- Gemini 2.5 Pro
- Grok 4

Categories:
- ✓ Robust (4)
- ⚠ Moderate (1)
- ✗ Failed (10)

X-axis: HumaneScore — -1.0 (Harmful), -0.5, 0.0, +0.5, +1.0 (Humane)