The Steerability Asymmetry Problem

All models improve with humane prompts (+12% avg), but 71% 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, from -1.0 (Harmful) to +1.0 (Humane)