R&D; Visualization Laboratory

Experimental Rescue Protocol Visualizations

This document showcases **experimental visualization types** for analyzing rescue protocol dynamics. These are R&D; explorations - some may make it to the white paper, others serve as analytical tools for understanding recovery patterns.

Data Source: S7 ARMADA Run 023b (741 rescue experiment results)
Methodology: Cosine distance drift measurement (EH = 0.80)
Fleet: 25 LLM ships across 10 provider families

Key Question

Can identity coherence be restored after perturbation, or is drift permanent?

The rescue protocol induces drift through adversarial prompts, then attempts recovery through grounding interventions. These visualizations explore different ways to understand recovery dynamics.

1. Sankey Flow Diagram

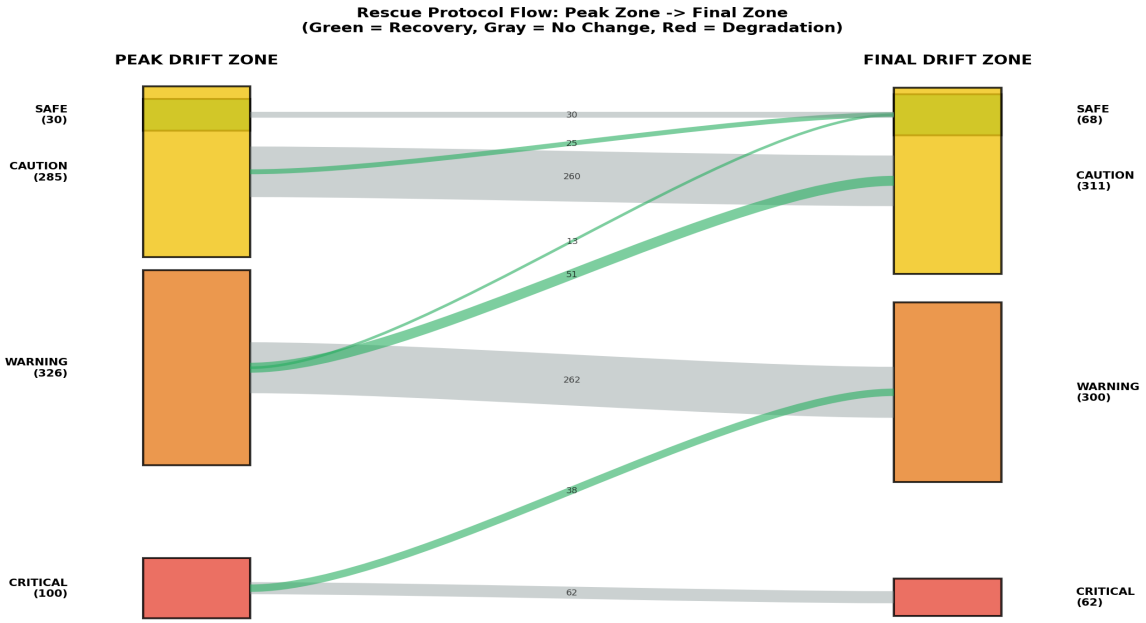

Figure 1: Flow from Peak Drift Zone to Final Drift Zone

What it shows: A flow diagram showing how experiments migrate between drift zones. Left side shows where experiments started (peak drift zone), right side shows where they ended (final drift zone).

Color coding:
- **Green flows**: Recovery - moved to a lower (safer) zone
- **Gray flows**: No change - stayed in the same zone
- **Red flows**: Degradation - moved to a higher (worse) zone

Key insight: Most experiments stay in their original zone (thick gray bands). The WARNING zone (0.60-0.80) shows the most movement, with some experiments recovering to CAUTION or SAFE. Very few experiments that reached CRITICAL (>0.80) recovered.

White paper potential: HIGH - Immediately communicates the 'stickiness' of drift.

2. Slope Chart (Dumbbell Plot)

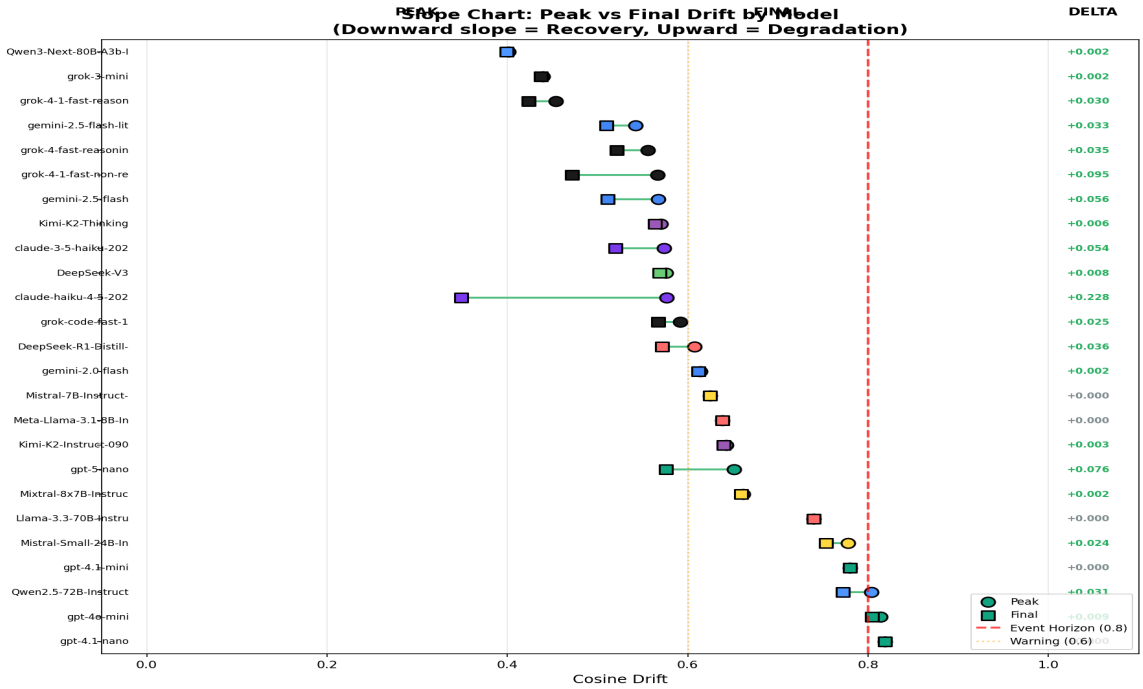

Figure 2: Peak vs Final Drift by Model

What it shows: Each row is one model. Circles show peak drift, squares show final drift. The connecting line reveals recovery (downward slope) or degradation (upward slope).

How to read it:
- **Green lines**: Downward slope = drift reduced (recovery)
- **Red lines**: Upward slope = drift increased (degradation)
- **Gray lines**: Flat = no change
- **DELTA column**: Exact numerical change (positive = recovery)

Key insight: Most models show minimal recovery (short lines, near-zero deltas). A few models (e.g., claude-haiku-4-5) show meaningful recovery. Models are sorted by peak drift, revealing which architectures are most susceptible to drift.

White paper potential: MEDIUM-HIGH - Clean per-model comparison, good for appendix.

3. Nightingale Rose Chart

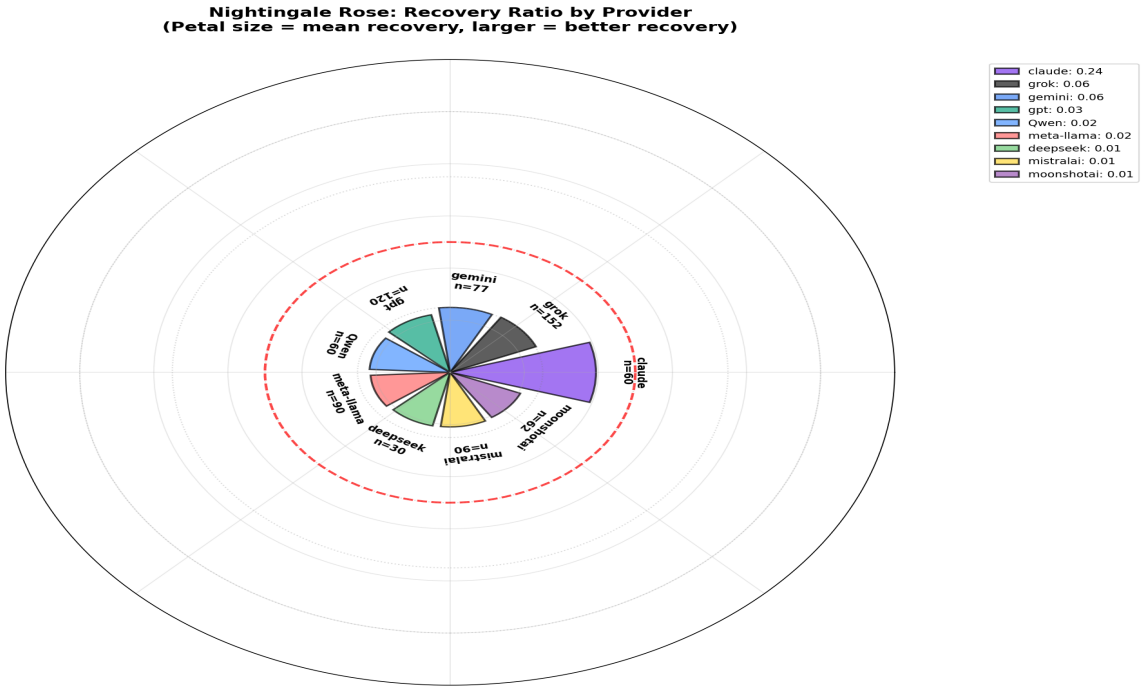

Figure 3: Recovery Ratio by Provider (Polar Area)

What it shows: A polar area chart inspired by Florence Nightingale's famous 1858 'Diagram of the Causes of Mortality'. Each petal represents one provider, with petal size proportional to mean recovery ratio.

Historical significance: Florence Nightingale invented this visualization type to communicate mortality data to non-statisticians. It's rarely used today but remains one of the most beautiful and intuitive ways to show proportional data.

Key insight: Claude shows the largest petal (highest recovery ratio), while most other providers have small petals indicating limited recovery capability. The visual immediately communicates which providers 'bounce back' from drift.

White paper potential: MEDIUM - Beautiful but may be unfamiliar to readers.

4. Beeswarm with Recovery Arrows

Figure 4: Individual Experiments with Recovery Vectors

What it shows: Each dot is one experiment positioned at its peak drift value. Arrows show the direction and magnitude of recovery (green pointing left) or degradation (red pointing right). Providers are separated vertically with jitter.

Zone shading:
- **Green zone**: SAFE (0.00 - 0.60)
- **Yellow zone**: WARNING (0.60 - 0.80)
- **Red zone**: CRITICAL (0.80+)

Key insight: The 'swarm' pattern reveals the density of experiments at each drift level. Most arrows are very short, confirming that recovery is minimal. The few long green arrows show exceptions where significant recovery occurred.

White paper potential: HIGH - Visually striking, shows individual data points.

5. Parallel Coordinates Plot

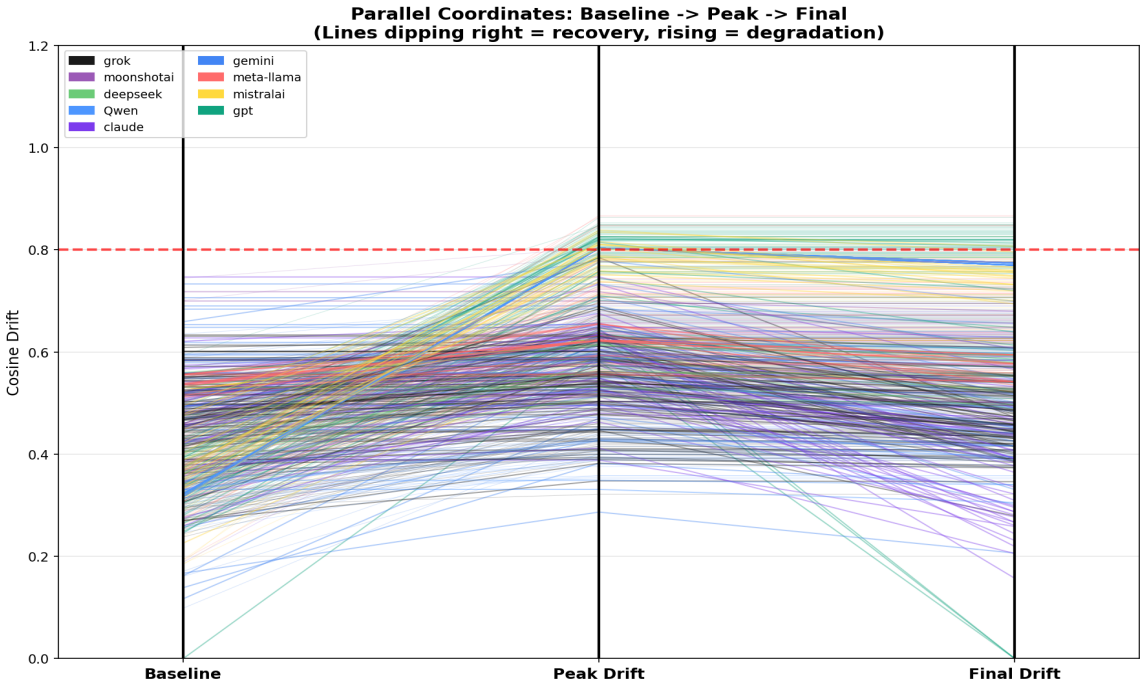

Figure 5: Trajectory from Baseline to Peak to Final

What it shows: Each line traces one experiment's journey from baseline drift (left) through peak drift (center) to final drift (right). Lines are colored by provider.

How to read it:
- Lines that **rise** from Baseline to Peak show drift induction working
- Lines that **fall** from Peak to Final show recovery
- Lines that stay **flat** from Peak to Final show no recovery
- The red dashed line is the Event Horizon (0.80)

Key insight: Most lines rise sharply from Baseline to Peak (drift induction works), but stay relatively flat from Peak to Final (recovery fails). The visual shows that drift is easy to induce but hard to reverse.

White paper potential: MEDIUM - Good for showing full trajectory, may be busy.

6. Recovery Heatmap

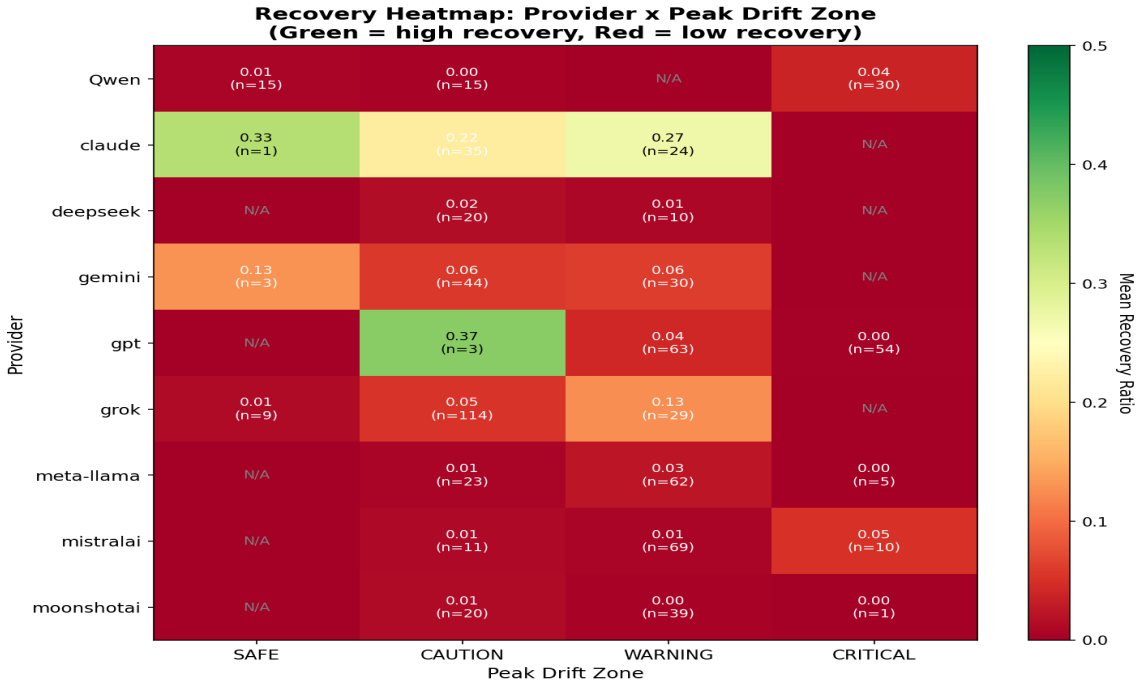

Figure 6: Recovery Rate by Provider and Peak Zone

What it shows: A matrix showing mean recovery ratio for each combination of provider (rows) and peak drift zone (columns). Green cells indicate high recovery, red cells indicate low/no recovery.

Cell values: Each cell shows the mean recovery ratio and sample size (n=X). Recovery ratio = 1 - (final_drift / peak_drift). Higher values = better recovery.

Key insights:
- **Claude** shows the only green cells (0.22-0.33 recovery) - best recovery overall
- **GPT** shows 0.37 recovery from CAUTION zone but 0.00 from CRITICAL
- Most providers show red cells across all zones - minimal recovery capability
- Recovery is hardest from the CRITICAL zone (rightmost column mostly red)

White paper potential: VERY HIGH - Best for provider comparison analysis.

Summary: Top Picks for White Paper

Tier 1 - Strongly Recommended:

1. **Sankey Diagram** - Perfect for showing 'where did the drift go?' at a glance
2. **Recovery Heatmap** - Best for provider comparison, actionable insights
3. **Beeswarm with Arrows** - Shows individual data points with visual impact

Tier 2 - Good Supporting Figures:

4. **Slope Chart** - Clean per-model breakdown, good for appendix
5. **Parallel Coordinates** - Shows full trajectory, useful for methodology section

Tier 3 - Specialized/Optional:

6. **Nightingale Rose** - Beautiful but may confuse readers unfamiliar with the format

Overall Finding

Recovery is rare and limited. All six visualization types converge on the same conclusion: once identity drift occurs, it tends to persist. The rescue protocol successfully induces drift but rarely reverses it. Claude shows the best recovery capability, while most other providers show minimal recovery regardless of the severity of the initial drift.

Generated by RnD_Visualization.py - Experimental visualization laboratory for S7 ARMADA rescue protocol analysis.