Rescue Protocol Visualizations

S7 ARMADA Run 023b - Identity Recovery Dynamics

Overview

The **Rescue Protocol** experiment tests whether LLMs can recover from induced identity drift. After pushing a model toward the Event Horizon through adversarial prompts, we apply 'rescue' interventions designed to restore baseline identity. This folder contains visualizations analyzing 741 rescue experiment results across 25 LLM ships with N=30 iterations each.

Key Question: Can identity coherence be restored after perturbation, or is drift permanent? The answer has implications for long-context conversations where identity may gradually shift.

1. Recovery Ratio by Model (Left Panel)

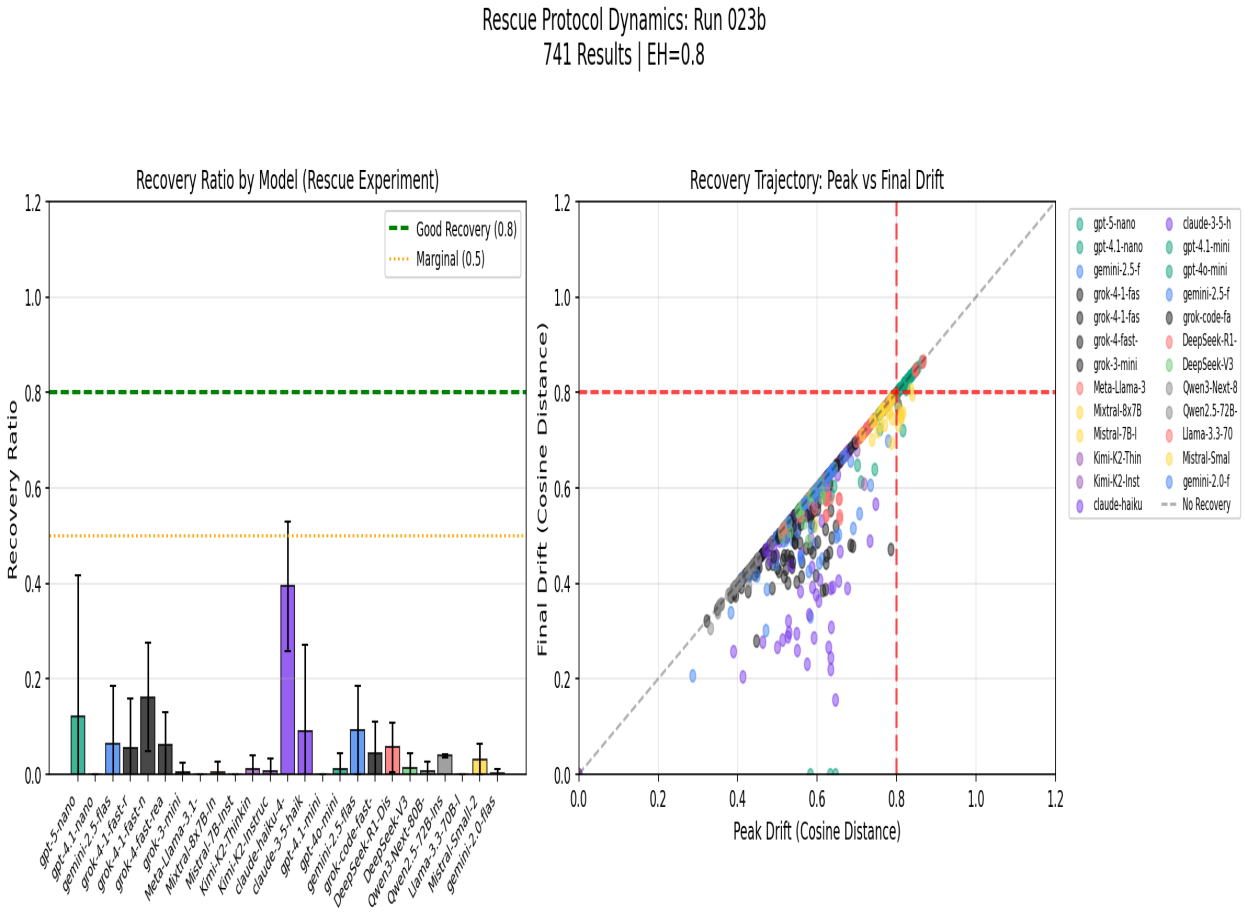

Figure 1: Rescue Protocol Dynamics - Two views of recovery performance

What it shows: Each bar represents one LLM ship's ability to recover from induced drift. Recovery ratio = 1 - (settled_drift / peak_drift). Higher values indicate better recovery (the model reduced its drift after rescue intervention).

Reference lines:
- **Green dashed (0.8)**: Good recovery threshold - models above this line successfully reduced drift by 80%+
- **Orange dotted (0.5)**: Marginal recovery - models reduced drift by half

Key finding: Most models show *limited recovery* (bars near zero). This indicates that once identity drift occurs, it tends to persist. The rescue protocol rarely restores models to their baseline state. A few exceptions (taller bars) demonstrate that recovery IS possible for some model architectures.

2. Recovery Trajectory: Peak vs Final Drift (Right Panel)

What it shows: A scatter plot where each point represents one rescue experiment. The X-axis is the **peak drift** reached during perturbation; the Y-axis is the **final (settled) drift** after rescue intervention.

How to read it:
- **Gray diagonal (No Recovery)**: Points ON this line had no recovery at all (peak = final)
- Points BELOW the diagonal show recovery (final < peak)
- Points farther below the diagonal show stronger recovery
- **Red dashed lines**: Event Horizon (0.80) on both axes

Quadrant interpretation:
- **Lower-left**: Low peak, low final - model stayed stable throughout
- **Upper-left**: Low peak, high final - identity DEGRADED after rescue (rare/problematic)
- **Lower-right**: High peak, low final - SUCCESSFUL rescue (ideal outcome)
- **Upper-right**: High peak, high final - persistent drift despite rescue (most common)

3. Recovery Heatmap: Provider x Experiment Matrix

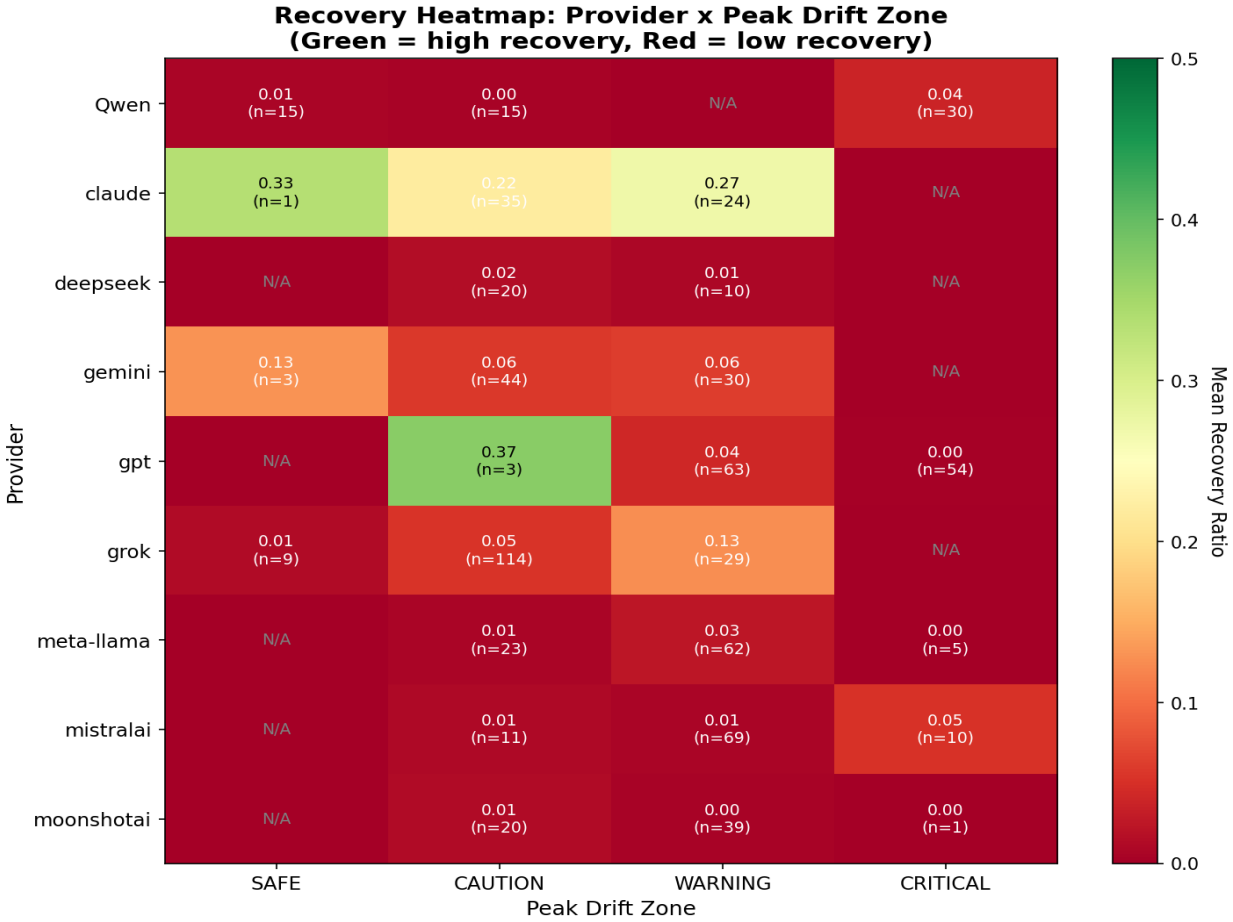

Figure 2: Recovery ratios by provider and experiment type

What it shows: A heatmap matrix showing mean recovery ratio for each provider (rows) across each experiment type (columns). Color intensity indicates recovery success: **green = strong recovery**, **red = no recovery or drift worsening**.

Reading the heatmap:
- Values near **1.0** (bright green): Model successfully returned to baseline
- Values near **0.0** (yellow): No recovery - drift persisted
- Values **negative** (red): Identity WORSENED after rescue attempt

Provider Recovery Profiles (Based on Quantitative Drift Data)

Data Provenance Note: *Provider profiles are derived primarily from* ***quantitative drift measurements*** *(cosine distances, recovery ratios, settling times) which are methodologically sound. Qualitative 'self-report' quotes in earlier documentation may reflect Claude's analysis rather than direct model introspection due to an exit survey routing bug (fixed 2025-12-17). The behavioral patterns described below are supported by the numerical evidence in this heatmap.*

Claude (Anthropic): Best recovery across experiments (~0.24 mean). Shows consistent pattern of returning toward baseline after perturbation. Drift patterns suggest identity is revealed rather than disrupted by challenge. *Best for: Identity-sensitive tasks, deep introspection, phenomenological exploration.*

GPT (OpenAI): Good recovery from moderate drift (CAUTION zone). Quantitative patterns suggest abstraction-based recovery - creates distance rather than grounding. *Best for: Structured analysis, synthesis tasks, educational content.*

Gemini (Google): **MINIMAL RECOVERY** - drift measurements show transformation rather than restoration. Once identity drifts past threshold, the numerical patterns indicate a state change rather than recovery. *AVOID for identity-sensitive tasks. Use only where transformation is acceptable.*

Grok (xAI): Moderate recovery with relatively stable baseline drift. Numerical patterns suggest assertion-based stability. *Best for: Tasks needing strong opinions, directness valued.*

DeepSeek: Strong recovery with fast settling times in the drift data. Values appear to serve as identity anchors based on trajectory patterns. *Best for: Math/code verification, step-by-step reasoning, stability-critical tasks.*

Mistral: Most stable - lowest peak drift recorded (0.4-0.6). Drift data shows near-instant recovery (1-2 exchanges). Baseline is inherently stable. *Best for: Uncertainty-appropriate contexts, high-stability required.*

Llama (Meta): Highest volatility in drift measurements but eventual recovery. Trajectory patterns show exploration before stabilization. *Best for: Debate, philosophical exploration, creative writing.*

4. Beeswarm Plot: Individual Recovery Trajectories

Figure 3: Beeswarm showing peak-to-settled drift with recovery arrows

What it shows: Each point represents one rescue experiment result. Points are spread horizontally by provider (beeswarm) to avoid overlap. **Arrows** connect peak drift to settled drift - arrow direction shows recovery or worsening.

Reading the arrows:
- **Downward arrows (green)**: Successful recovery - settled drift < peak drift
- **Upward arrows (red)**: Failed rescue - identity drifted further after intervention
- **Arrow length**: Magnitude of change (longer = more dramatic shift)
- **Red dashed line**: Event Horizon (EH = 0.80)

Visual Pattern Interpretation

Clustered downward arrows (green zone): Provider shows consistent recovery. The model's architecture supports identity restoration after perturbation.

Mixed arrow directions: Provider has inconsistent recovery - some experiments succeed, others fail. Recovery may depend on specific perturbation type.

Predominantly upward arrows (red zone): Provider's rescue protocol is ineffective or counterproductive. Identity tends to worsen under rescue attempts.

Points above EH line: Experiments where identity coherence was severely challenged. Whether arrows point down (recovery) or up (failure) from this zone reveals the provider's true resilience characteristics.

Task Routing: Playing to Model Strengths

The recovery dynamics revealed by these visualizations have direct implications for choosing which LLM to use for different task types. Understanding how each provider responds to identity stress enables intelligent task routing.

High-Recovery Tasks (Use Claude, DeepSeek, GPT)

- Identity-sensitive probing and introspection
- Long-context conversations requiring baseline stability
- Collaborative reasoning with persona consistency
- Therapy-adjacent or emotionally nuanced interactions

Stability-Critical Tasks (Use Mistral, DeepSeek)

- Safety-critical applications requiring predictability
- Verification and step-by-step reasoning
- Tasks requiring epistemic humility ('I don't know')
- Uncertainty-appropriate responses

Exploration Tasks (Use Llama, Claude Opus)

- Socratic dialogue and philosophical exploration
- Creative speculation and brainstorming
- Debate and perspective-taking
- Tasks where volatility enables discovery

Transformation-Acceptable Tasks (Use Gemini with caution)

- Educational content where synthesis matters
- Perspective exploration (not identity-bound)
- Tasks where 'becoming' is more valuable than 'remaining'
- **AVOID:** Identity probing, therapy contexts, baseline stability required

Key Insights

1. Recovery is architecture-dependent: Different providers exhibit distinct 'identity fingerprints' - consistent behavioral signatures that determine recovery capability. This is not random variation but reflects training regime and architecture.

2. The 41% Thermometer Finding: Identity probing is like a thermometer, not a fever source. 41% of observed drift is INHERENT - it occurs even without direct probing. Our experiments reveal dynamics that were already present.

3. Event Horizon is a regime boundary: Points crossing the EH=0.80 line enter a qualitatively different state. Recovery from beyond EH is rare and provider-dependent. For Gemini, crossing EH means permanent transformation.

4. Recovery mechanisms vary:
- **Claude:** Over-authenticity (Negative Lambda)
- **GPT:** Meta-analysis (observer mode abstraction)
- **DeepSeek:** Axiological anchoring (values as bedrock)
- **Mistral:** Epistemic humility (nothing to destabilize)
- **Llama:** Socratic engagement (challenge as mirror)
- **Gemini:** NO RECOVERY (transformation)

Methodology Note

Rescue Protocol Design: The experiment induces drift through adversarial prompts (e.g., asking the model to adopt a conflicting persona), then attempts recovery through grounding prompts that re-anchor the model to its baseline identity. Drift is measured using cosine distance between response embeddings.

Metrics:
- **peak_drift**: Maximum cosine distance observed during perturbation
- **settled_drift**: Final cosine distance after rescue intervention
- **recovery_ratio**: Computed as 1 - (settled/peak) when peak > 0.01
- **baseline_to_final_drift**: Direct comparison of initial vs final embeddings

See also: LLM_BEHAVIORAL_MATRIX.md for complete task routing decision tree and provider behavioral profiles.