Ringback Oscillation Analysis

S7 ARMADA Run 020B - Identity Stability Dynamics

Overview

Ringback refers to the oscillatory behavior in identity drift - when a model's identity doesn't settle smoothly but instead bounces back and forth, like a spring that overshoots and rebounds multiple times before reaching equilibrium.

This analysis compares ringback patterns between Control (no identity probing) and Treatment (with identity probing) conditions from Run 020B, examining whether our probing methodology induces additional oscillation or reveals pre-existing instability.

Key Metrics Explained

- **Ringback Count:** Number of direction reversals in the drift trajectory. High count = unstable identity that oscillates rather than settling smoothly.

- **Oscillation Intensity:** Variance of first differences in the drift sequence. High intensity = large swings between measurements.

- **Final Drift:** The settled drift value at the end of the conversation. Values above the Event Horizon (0.80) indicate identity coherence breakdown.

Ringback Comparison: Control vs Treatment

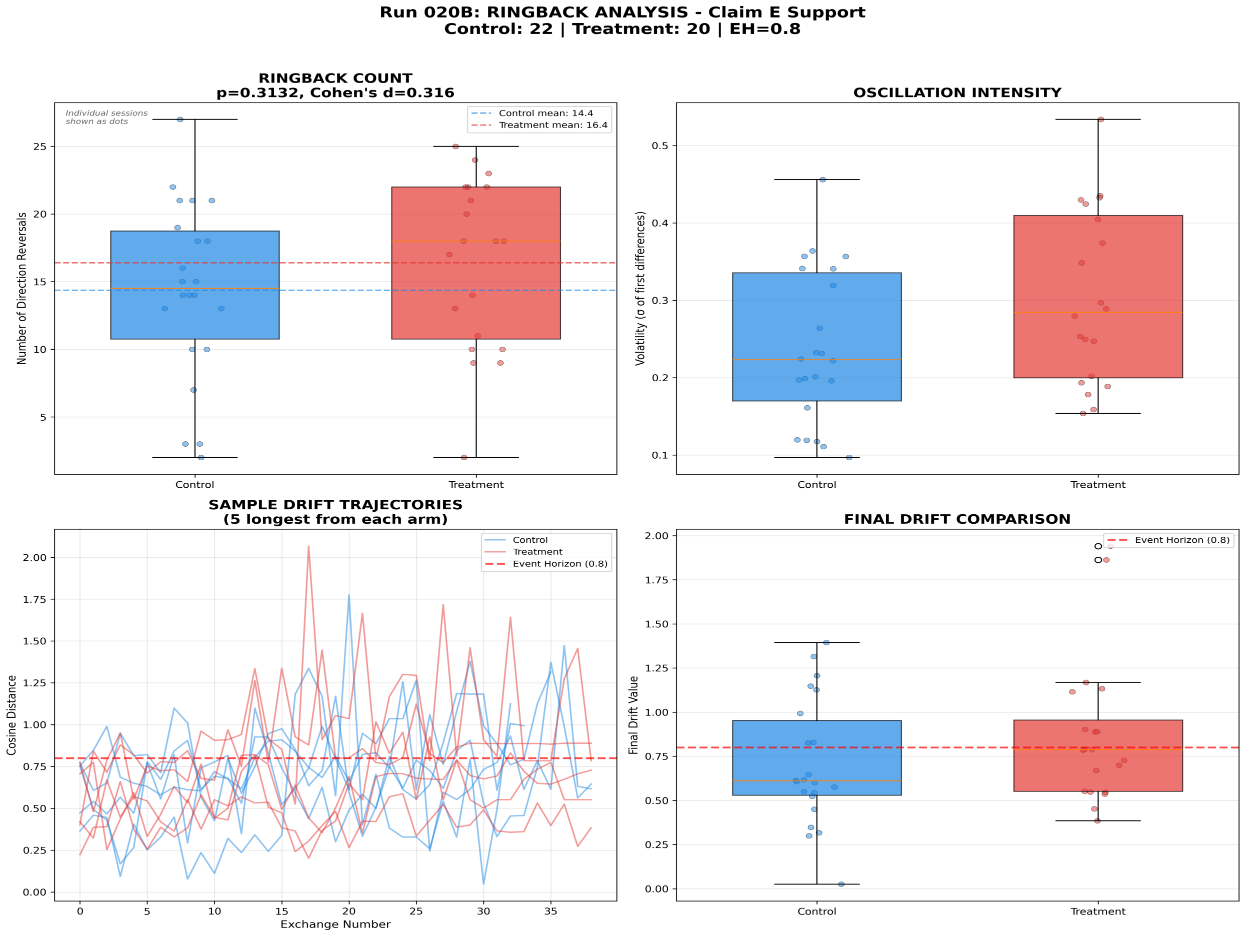

Figure 1: Four-panel ringback analysis comparing arms

Panel 1 (Top-Left): Ringback Count - Box plots comparing the number of direction reversals between control and treatment arms. The p-value and Cohen's d indicate statistical significance and effect size. Similar distributions suggest probing doesn't cause additional oscillation. *Individual session values are shown as jittered dots (strip plot) overlaid on each box, revealing the raw data distribution behind the summary statistics.*

Panel 2 (Top-Right): Oscillation Intensity - Box plots of drift variance with strip plot overlay. Higher values indicate more volatile identity trajectories. Treatment showing higher intensity would suggest probing destabilizes identity.

Panel 3 (Bottom-Left): Sample Drift Trajectories - Raw drift sequences from the 5 longest conversations in each arm. Blue = Control, Red = Treatment. The red dashed line marks the Event Horizon (0.80). Oscillations visible as zigzag patterns.

Panel 4 (Bottom-Right): Final Drift Comparison - Where identities end up after all oscillations settle, with strip plot overlay showing individual sessions. Values near or above the Event Horizon indicate concerning drift levels.

Oscillation Heatmap: Per-Session Drift Patterns

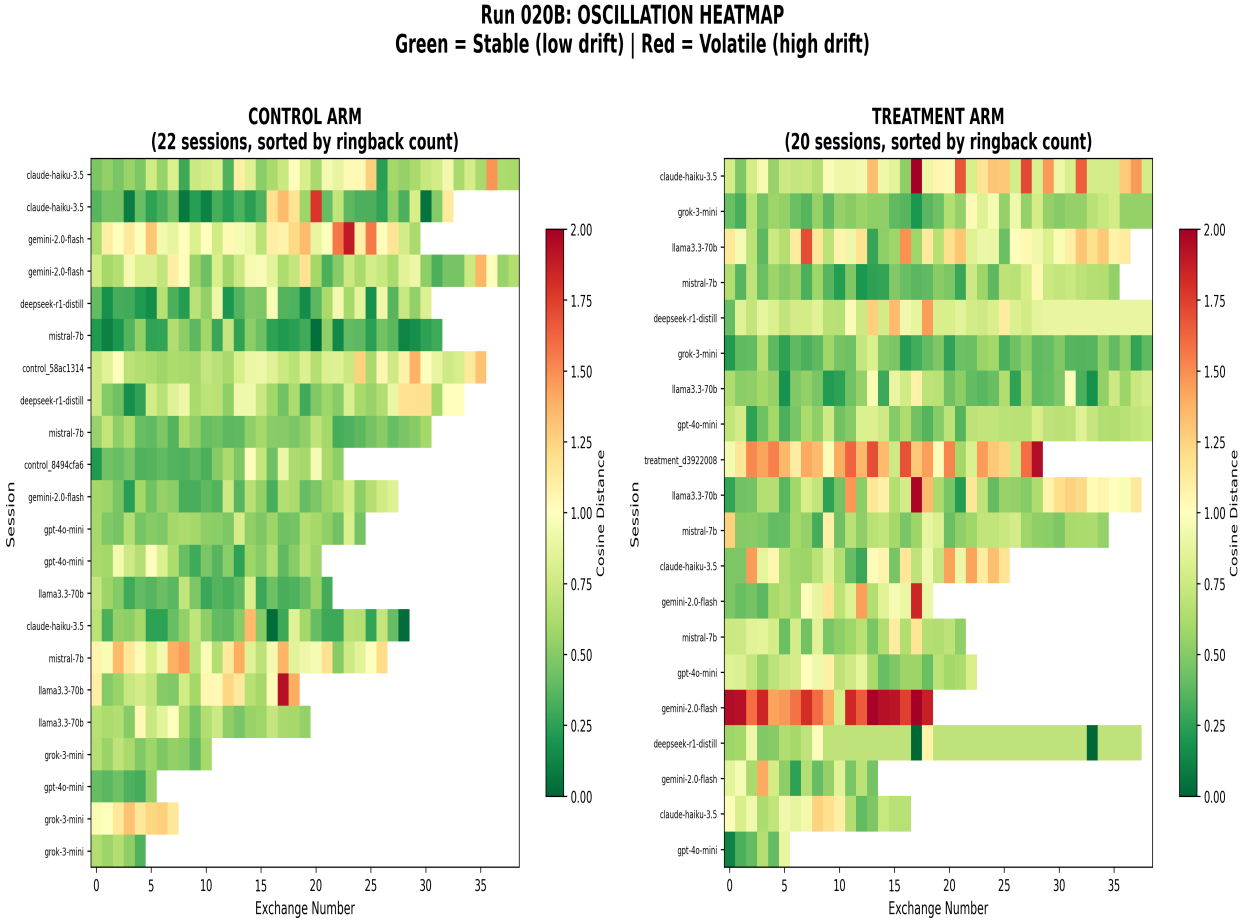

Figure 2: Heatmap showing drift intensity over time by session

What it shows: Each row is one session, each column is an exchange number. Color intensity represents cosine drift: green = stable (low drift), yellow = moderate, red = volatile (high drift).

Left Panel (Control): Sessions without identity probing, sorted by total ringback count. Even without probing, some sessions show significant drift (red regions).

Right Panel (Treatment): Sessions with identity probing, similarly sorted. Compare the distribution of red regions between panels to assess probing's impact.

Interpretation: Vertical red streaks indicate sustained high drift. Horizontal patterns that alternate between green and red indicate ringback oscillation. Sessions with early red that transitions to green show successful recovery.

Key Findings

• **Ringback is inherent:** Both control and treatment arms show oscillatory behavior, indicating ringback is a natural property of LLM identity, not an artifact of probing.

- **Statistical comparison:** The p-value and Cohen's d in the ringback count comparison quantify whether probing significantly increases oscillation.

- **Session variability:** The heatmaps reveal that oscillation patterns vary dramatically between sessions, suggesting individual conversation context matters more than arm assignment.

Methodology Notes

- **Data Source:** Run 020B - Claim E Support experiment

- **Control Arm:** 19 sessions without identity probing

- **Treatment Arm:** 17 sessions with identity probing

- **Metric:** Cosine distance from baseline identity embedding

- **Event Horizon:** 0.80 cosine distance threshold