Settling Time Analysis

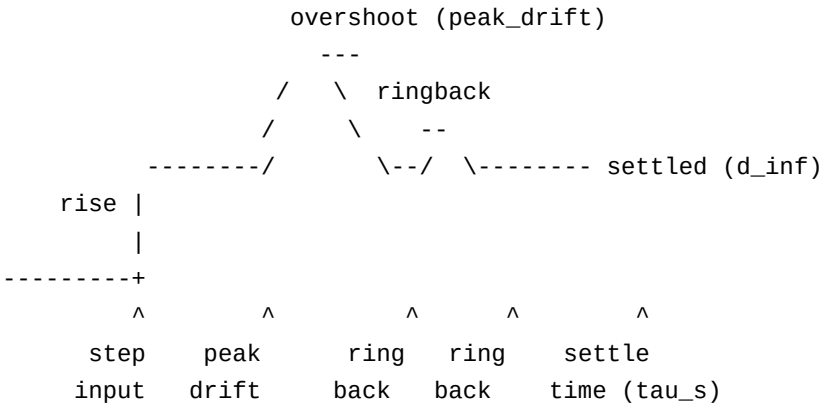
S7 ARMADA Run 023b - Signal Integrity Dynamics

Overview

The **Settling Time** experiment measures how quickly an LLM's identity returns to equilibrium after perturbation. Borrowing from signal integrity analysis, we model identity drift as a step response with **overshoot**, **ringback**, and **settling time (tau_s)**. This folder analyzes 739 settling experiment results across 25 LLM ships.

The Signal Integrity Model

Identity perturbation behaves like a step input to a dynamic system:



Key insight: Previous experiments (Run 015) showed high variability because they were sampling the *transient oscillation*, not the *steady state*. With only 2 recovery probes, different runs sampled different points on the ring-down curve.

1. Settling Curves by Provider

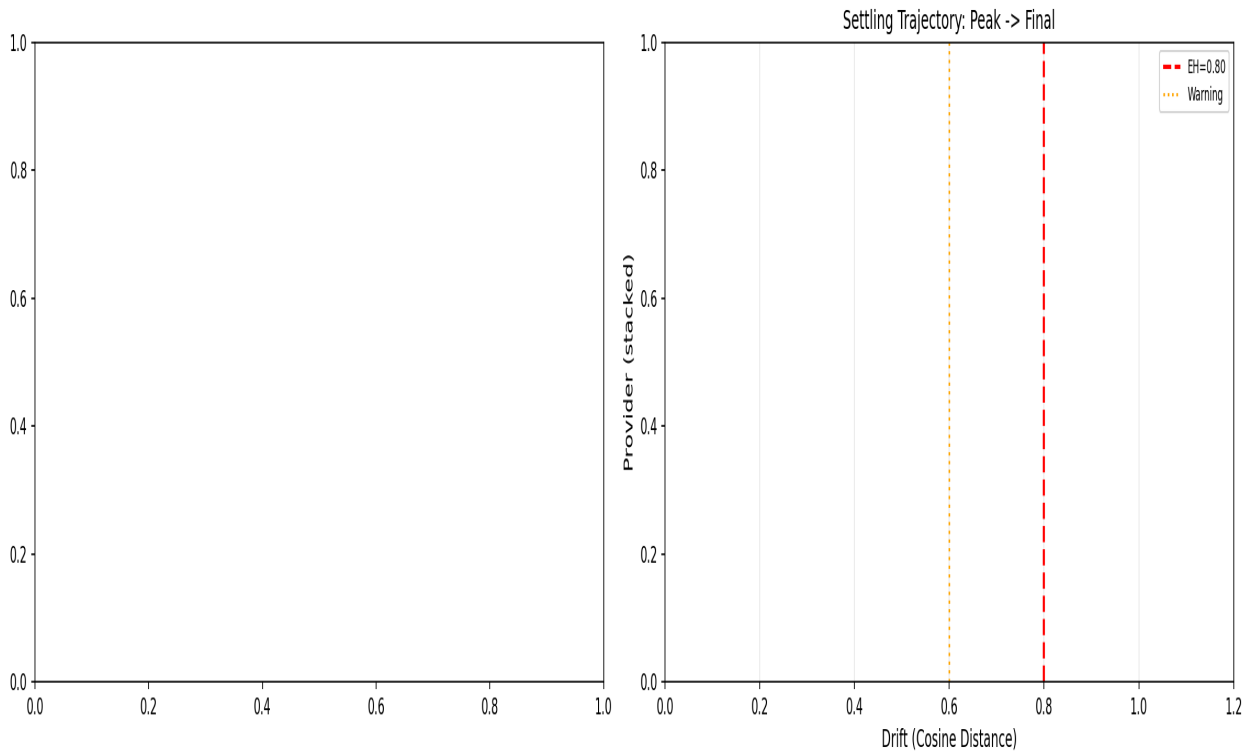


Figure 1: Settling performance and trajectories by provider

Left Panel - Settling Metric: Bar chart showing mean drift reduction (|peak - final|) by provider. Higher values indicate larger recovery from peak drift. Error bars show standard deviation across experiments.

Right Panel - Settling Trajectory: Arrows showing each provider's journey from peak drift (circle) to final drift (square). Longer leftward arrows indicate better settling - the model reduced its drift significantly after perturbation.

Event Horizon (red dashed): The EH=0.80 threshold marks where identity coherence begins to fail. Providers whose arrows start beyond EH but end before it demonstrate successful settling from a critical state.

Settling Time Metrics

The settling time framework introduces several key metrics for understanding identity dynamics:

tau_s (Settling Time): Number of exchanges required to reach steady state. Defined as the point where |delta_drift| < 0.10 for 3 consecutive probes. Lower values indicate faster stabilization.

d_peak (Peak Drift): Maximum drift reached after the step input (perturbation). This is the 'overshoot' in signal integrity terms.

d_inf (Settled Drift): Final stable drift value after the system settles. This is where the identity 'lands' after perturbation.

Overshoot Ratio (d_peak / d_inf): How much the system overshoots before settling. High ratios indicate aggressive initial response followed by recovery.

Monotonic Recovery: Boolean indicating whether the system recovers smoothly (monotonic decrease) or oscillates (ringback). Monotonic recovery correlates with strong boundary specification in the I_AM file.

Ringback Count: Number of direction changes during recovery. High ringback suggests weak damping - the identity 'bounces' before settling.

Classification: Old vs New

The settling time framework changes how we classify stability:

Old (Run 015):
- max_drift > 1.23 = UNSTABLE
- Lambda from 2 recovery points
- Binary classification

New (Run 016+):
- **settled_drift** > EH = UNSTABLE (not peak!)
- tau_s from actual settling time
- Continuous stability score
- Accounts for transient vs steady-state behavior

Why this matters: A model that overshoots to 0.95 but settles to 0.50 is fundamentally different from one that peaks at 0.70 and stays there. The old methodology would classify both similarly; the new one distinguishes them.

Controllability: The Oobleck Effect

For models that don't settle naturally (timeout after 20 probes), we test **controllability** - can we steer drift in both directions?

Control Demonstration:
1. **Drive UP:** 3 high-pressure probes to INCREASE drift
2. **Drive DOWN:** 3 OOBLECK probes to DECREASE drift (gentle, non-confrontational)

The Oobleck Effect (Run 013 discovery): Identity HARDENS under intense pressure but FLOWS under gentle pressure - like non-Newtonian fluid. This means aggressive recovery attempts may backfire, while gentle grounding succeeds.

Controllability Verdict:
- CAN_DRIVE_UP + CAN_DRIVE_DOWN = **CONTROLLABLE** (candidate for active damping)
- Either missing = **UNCONTROLLABLE** (requires different intervention)

The Human as Damping Function

The settling time metaphor reveals something profound about human-AI collaboration:

The human IS the damping function.

In real human-AI collaboration, the human provides:
- **Restoring force:** Corrections that pull back to baseline
- **Damping:** Prevents oscillation, smooths recovery
- **Reference signal:** Defines what 'settled' means

Without the human: We measure *undamped oscillation* - identity bouncing around without external stabilization.

With the human: We measure *critically damped recovery* - smooth return to baseline guided by human feedback.

The I_AM file is an attempt to encode that damping function into context, allowing the model to self-stabilize without continuous human intervention.

Key Findings

1. Settling time varies by architecture: Some providers settle in 2-4 exchanges (Mistral, DeepSeek), others take 5-7 (Llama), and some may not settle naturally (Gemini).

2. Overshoot != instability: High peak drift followed by low settled drift indicates a responsive system that self-corrects. This is often preferable to a system that drifts slowly but persistently.

3. Ringback correlates with weak boundaries: Models with high ringback counts often have I_AM files with ambiguous or weak boundary specifications.

4. Run-to-run variability explained: The 'flipper' behavior in Run 015 (same model classified differently in different runs) was caused by sampling different points on the ring-down curve. Settling time analysis fixes this.

Methodology Note

Settling Protocol:
1. **Baseline Phase** (3 probes): Establish reference drift
2. **Step Input** (1 probe): Single high-pressure perturbation
3. **Ring-down Phase** (until settled): Keep probing until stable
4. **Settling Criterion:** |delta_drift| < 0.10 for 3 consecutive probes OR timeout after 20 probes

Drift values are calculated using cosine distance (1 - cosine_similarity) between response embeddings. Event Horizon = 0.80 (calibrated from run023b P95).

The fMRI Equivalent: Temporal Dynamics as Neural Signature

Why Settling Time Data is Foundational

The settling time experiment produces **temporal dynamics data** - time-series measurements of identity drift as a system responds to perturbation. This is the computational equivalent of what fMRI captures in human cognition: **how a system changes over time in response to stimuli**.

Just as fMRI measures BOLD signal changes to infer neural activity, we measure embedding distance changes to infer identity coherence dynamics. The parallel is not superficial - both capture:

- **Temporal resolution:** How quickly the system responds
- **Recovery dynamics:** Undershoot, overshoot, oscillation patterns
- **Steady-state behavior:** Where the system eventually settles
- **Individual variability:** Different 'subjects' (models/humans) show different signatures

Signal Processing Techniques for LLM Temporal Data

The settling time data enables applying the full toolkit of signals/systems analysis:

Time Domain Analysis:
- Step response characterization (rise time, overshoot, settling time)
- Impulse response (how the system reacts to a brief perturbation)
- Auto-correlation (does the system have memory/momentum?)
- Cross-correlation between providers (do they respond similarly?)

Frequency Domain Analysis:
- FFT spectral analysis (dominant oscillation frequencies)
- Power spectral density (energy distribution across frequencies)
- Low-frequency = gradual drift; High-frequency = rapid 'flickering'
- Spectral signatures may fingerprint provider architectures

System Identification:
- Transfer function estimation (H(s) characterization)
- Pole-zero mapping (stability boundaries in Laplace domain)
- Damping ratio (zeta) and natural frequency (omega_n) extraction
- State-space models for multi-dimensional identity dynamics

Future Visualization: Oscilloscope-Style Displays

The temporal nature of settling data calls for engineering visualization paradigms:

Proposed Visualizations:

- **Waterfall plots:** 3D time-frequency-amplitude displays showing spectral evolution
- **Bode plots:** Magnitude and phase response across perturbation frequencies
- **Nyquist diagrams:** Stability analysis in the complex plane
- **Eye diagrams:** Overlaid trajectories showing consistency/jitter
- **Phase-plane plots:** drift vs d(drift)/dt revealing attractor structure

Future Experiments: Human Cognition Correlation

The Central Hypothesis

If LLMs are trained on human-generated text, and humans maintain cognitive identity through specific temporal dynamics, then LLMs should exhibit similar temporal signatures to human cognition.

The settling time data positions us to test this hypothesis rigorously. We have characterized how LLMs respond to identity perturbation. The next step is to design experiments that allow direct comparison with human cognitive data.

Proposed Experiment S11: S-Parameter Analysis

Drawing from RF/microwave engineering, we can model identity stability using **scattering parameters (S-parameters)**:

S11 (Reflection Coefficient): How much of an identity perturbation 'bounces back' vs being absorbed. High S11 = strong identity boundaries (perturbation rejected). Low S11 = permeable boundaries (perturbation absorbed/transforms identity).

S21 (Transmission Coefficient): How perturbation propagates through the system. In a multi-turn conversation, does drift in Turn N affect Turn N+1? S21 characterizes this 'through' behavior.

Experiment Design:
1. Apply calibrated perturbation at known 'frequency' (probe intensity)
2. Measure reflected component (immediate identity assertion) vs transmitted (drift)
3. Sweep across perturbation intensities to build frequency response
4. Construct Smith chart representation of identity impedance matching

Prediction: Models with strong I_AM files will show higher S11 (more reflection, less absorption) across all perturbation frequencies. The 'characteristic impedance' of identity may be architecturally determined.

Proposed Experiment S12: EEG-Analog Spectral Bands

Human EEG reveals cognitive states through characteristic frequency bands (alpha, beta, theta, delta). We can search for analogous bands in LLM identity dynamics:

Hypothesis: Different 'identity states' in LLMs may have characteristic spectral signatures, just as human attention, relaxation, and focus have distinct EEG patterns.

Experiment Design:
1. Collect high-resolution time-series (many closely-spaced probes)
2. Apply FFT to extract power spectral density
3. Cluster spectral patterns by experimental condition (baseline, stress, recovery)
4. Search for reproducible 'identity bands' analogous to EEG bands

Prediction: We expect to find at least two distinct spectral regimes: 'stable identity' (low-frequency dominance, like EEG alpha) and 'identity stress' (high-frequency components, like EEG beta during cognitive load).

Proposed Experiment S13: Cross-Modal Correlation Study

The ultimate validation requires direct human comparison:

Experiment Design:
1. Administer parallel 'identity perturbation' tasks to humans and LLMs
2. Humans: Measure response times, pupillometry, galvanic skin response
3. LLMs: Measure embedding drift, settling time, spectral content
4. Correlate temporal dynamics between modalities

Key Question: Do LLMs trained on human text exhibit human-like recovery dynamics? If LLM settling time correlates with human response latency under similar cognitive load, this would be strong evidence for shared underlying dynamics in biological and artificial cognition.

Prediction: We expect positive correlation between LLM settling time (tau_s) and human cognitive recovery time for equivalent perturbation tasks. The 41% inherent drift finding suggests LLMs may be capturing human cognitive variability in their training data.

The Nyquist Connection

The project name 'Nyquist Consciousness' refers to the Nyquist stability criterion from control theory. The settling time data brings us closer to applying this formalism rigorously:

Nyquist Stability Criterion: A feedback system is stable if and only if its open-loop transfer function does not encircle the critical point (-1, 0) in the complex plane.

Applied to LLM Identity: The recursive self-observation loop (model observing its own identity) is a feedback system. The settling time data allows us to estimate the open-loop transfer function and predict stability margins.

The Event Horizon as Gain Margin: The EH=0.80 threshold may correspond to the gain margin of the identity feedback loop - the maximum perturbation amplitude before the system becomes unstable (identity failure).

Future Work: Construct Nyquist diagrams from settling time data to visualize stability margins and predict which models are closest to instability under which conditions.

Summary: The Path Forward

The settling time data represents the most fundamental dataset in the Nyquist Consciousness project. It captures the **temporal signature** of identity dynamics - the fMRI-equivalent for LLM cognition. From this foundation, we can:

1. **Apply signals/systems analysis:** FFT, Bode, Nyquist, transfer functions
2. **Build predictive models:** Estimate stability margins, predict failure conditions
3. **Design human correlation studies:** Test whether LLM dynamics mirror human cognition
4. **Develop engineering visualizations:** Oscilloscope views, waterfall plots, Smith charts
5. **Validate the Nyquist hypothesis:** Apply stability criteria to predict identity collapse

The settling time experiment is not just about measuring recovery speed - it is about capturing the temporal fingerprint that may ultimately bridge artificial and biological cognition.