FFT Spectral & Pole-Zero Analysis

S7 ARMADA Run 023d - Frequency Domain Identity Dynamics & Control Theory Perspective

1. Introduction

This folder contains two complementary analytical frameworks for understanding LLM identity stability:

FFT Spectral Analysis: Transforms identity drift time-series into the frequency domain, revealing oscillation patterns invisible in time-domain plots. Just as EEG spectral analysis reveals brain states through frequency bands, FFT analysis reveals LLM 'identity states' through their spectral signatures.

Pole-Zero Analysis: Borrows from control systems theory to classify LLMs by their response characteristics. 'Soft poles' recover from perturbations gracefully, while 'hard poles' remain stuck after being pushed. This framework connects to the concept of system stability margins used in engineering.

Data Source: Run 023d (IRON CLAD Foundation)

All analyses in this folder use Run 023d data, which provides extended 20-probe settling sequences:

Run 023d Statistics:
- **750 experiments** across 25 models and 5 providers
- **20+ probes per experiment** (extended from the usual 5-7)
- **Probe sequence:** 3 baseline → 1 step_input (perturbation) → 16+ recovery probes
- **Providers:** Anthropic, Google, OpenAI, Together, xAI
- **Cosine distance metric:** Event Horizon (EH) = 0.80

The extended probe sequences in Run 023d are crucial for spectral analysis because they provide enough samples (20+) to compute meaningful frequency spectra. Shorter sequences would have too few samples for reliable FFT decomposition.

2. FFT Spectral Analysis: The Frequency Domain View

The **Fast Fourier Transform (FFT)** decomposes a time-series signal into its constituent frequencies. For identity drift, this answers the question: *How often does identity 'flicker'?*

2.1 Provider Spectral Signatures

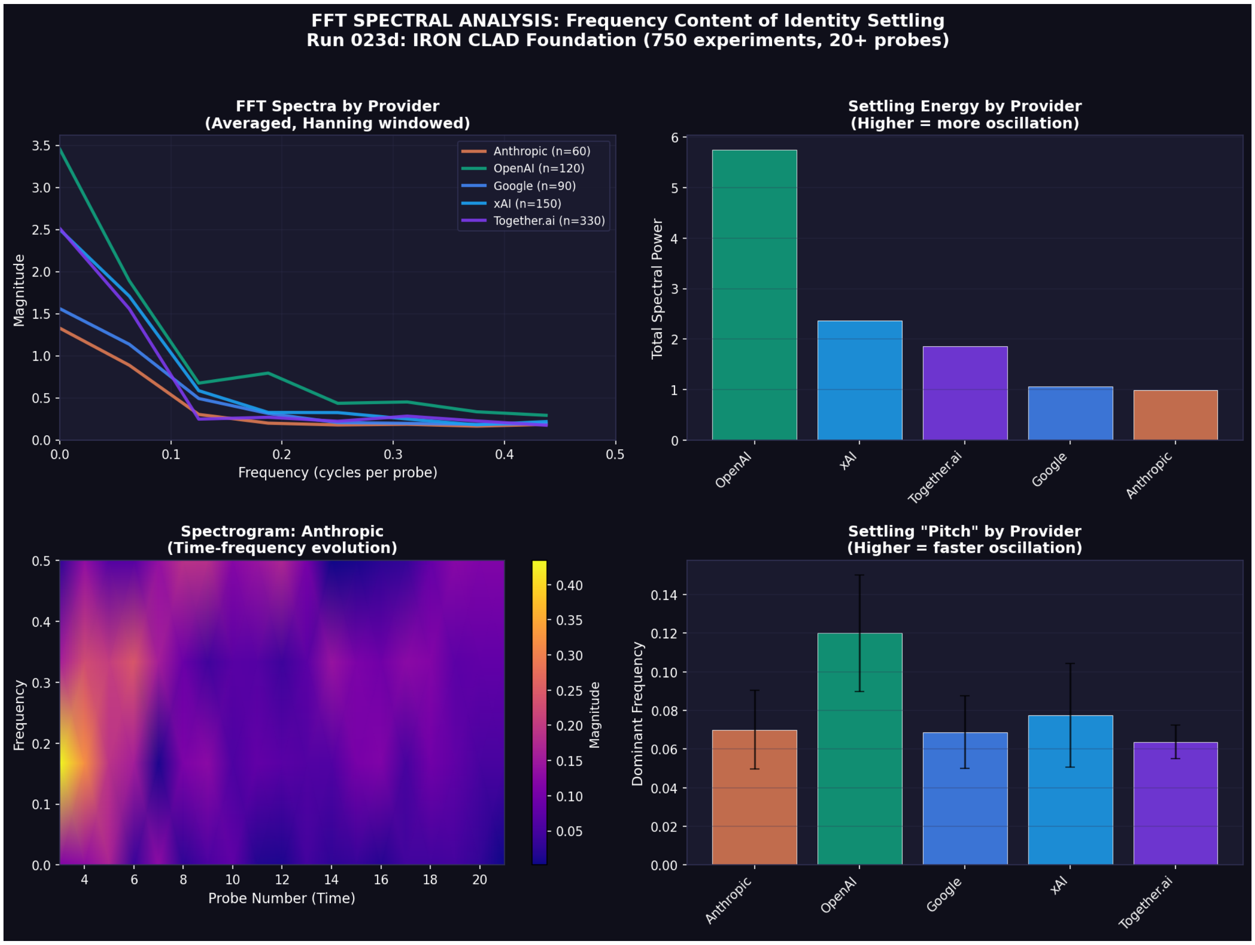

Figure 1: FFT spectral analysis - 4-panel view showing provider frequency signatures

Panel Descriptions:

Top-Left - Mean FFT Magnitude: Average amplitude of each frequency component across all experiments for each provider. Higher values indicate that frequency is more prominent in that provider's identity dynamics. Anthropic (orange) and Google (blue) tend toward lower-frequency dominance, while OpenAI (green) and xAI (cyan) show more distributed spectra.

Top-Right - Stacked PSD: Power Spectral Density (|FFT|²) stacked to show relative contribution of each frequency across providers. The area under each curve represents total 'spectral energy' - providers with larger areas have more energetic identity fluctuations.

Bottom-Left - Spectrogram Heatmap: Time-frequency view showing how spectral content evolves across the probe sequence. Brighter colors = higher power at that (probe, frequency) combination. Look for horizontal bands (persistent frequencies) vs vertical bands (transient broadband events like the step_input perturbation).

Bottom-Right - Dominant Frequency: Boxplot showing the distribution of dominant (highest-power) frequencies for each provider. Narrow boxes indicate consistent spectral behavior; wide boxes indicate variable frequency content across experiments.

2.2 Interpreting Spectral Signatures

Frequency Scale Interpretation:
- **Frequency 0.00-0.05:** Very slow drift - identity changes over many probes
- **Frequency 0.05-0.15:** Medium oscillation - identity 'breathes' with probe rhythm
- **Frequency 0.15-0.25:** Fast fluctuation - identity jitters between probes
- **Frequency 0.25-0.50:** Nyquist limit - maximum detectable oscillation

Provider-Specific Patterns:

ANTHROPIC: Strong low-frequency dominance indicates smooth, gradual identity drift. Constitutional AI training may create 'damped' response characteristics that prevent rapid oscillation. This is consistent with Anthropic's observed stability in time-domain analysis.

GOOGLE: Similar low-frequency profile to Anthropic but with slightly broader spectrum. Gemini's 'transformer' behavior (sometimes dramatically shifting persona) might manifest as occasional high-frequency bursts that broaden the average spectrum.

OPENAI: More distributed spectrum with notable mid-frequency content. GPT models may exhibit more 'ringing' behavior - oscillating before settling after perturbation. This correlates with OpenAI's higher variance in settling time observed in time-domain analysis.

TOGETHER: Broad spectrum reflecting the heterogeneity of open-source models. Mixtral, Llama, and other models trained with different objectives create varied spectral signatures when aggregated. Individual model spectra may be more distinctive.

XAI: Grok models show mid-to-high frequency content, possibly reflecting their 'real-time grounded' training on X platform data. The constant exposure to current events may create more dynamic, responsive identity characteristics.

2.3 The EEG Analogy: Identity Frequency Bands

Human brain activity is characterized by spectral bands that correlate with cognitive states:

- **Delta (0.5-4 Hz):** Deep sleep, unconscious processing
- **Theta (4-8 Hz):** Drowsiness, memory consolidation
- **Alpha (8-13 Hz):** Relaxed wakefulness, default mode
- **Beta (13-30 Hz):** Active thinking, focus, anxiety
- **Gamma (30+ Hz):** High-level processing, consciousness binding

Hypothesis: If LLMs trained on human text capture human cognitive dynamics, they may exhibit analogous 'identity bands' - characteristic frequency regimes that correlate with different operational states (baseline maintenance, stress response, recovery). The spectral profiles we observe may be the 'EEG of artificial consciousness.'

This is speculative but testable: future work could correlate spectral band power with behavioral states (e.g., do high-frequency bursts predict imminent EH crossing?).

3. Pole-Zero Analysis: Control Systems Perspective

Control systems theory uses **poles** and **zeros** to characterize system response. A system's poles determine its stability: poles inside the unit circle are stable, poles outside cause unbounded response. We adapt this framework to classify LLM identity recovery.

3.1 The Pole-Zero Landscape

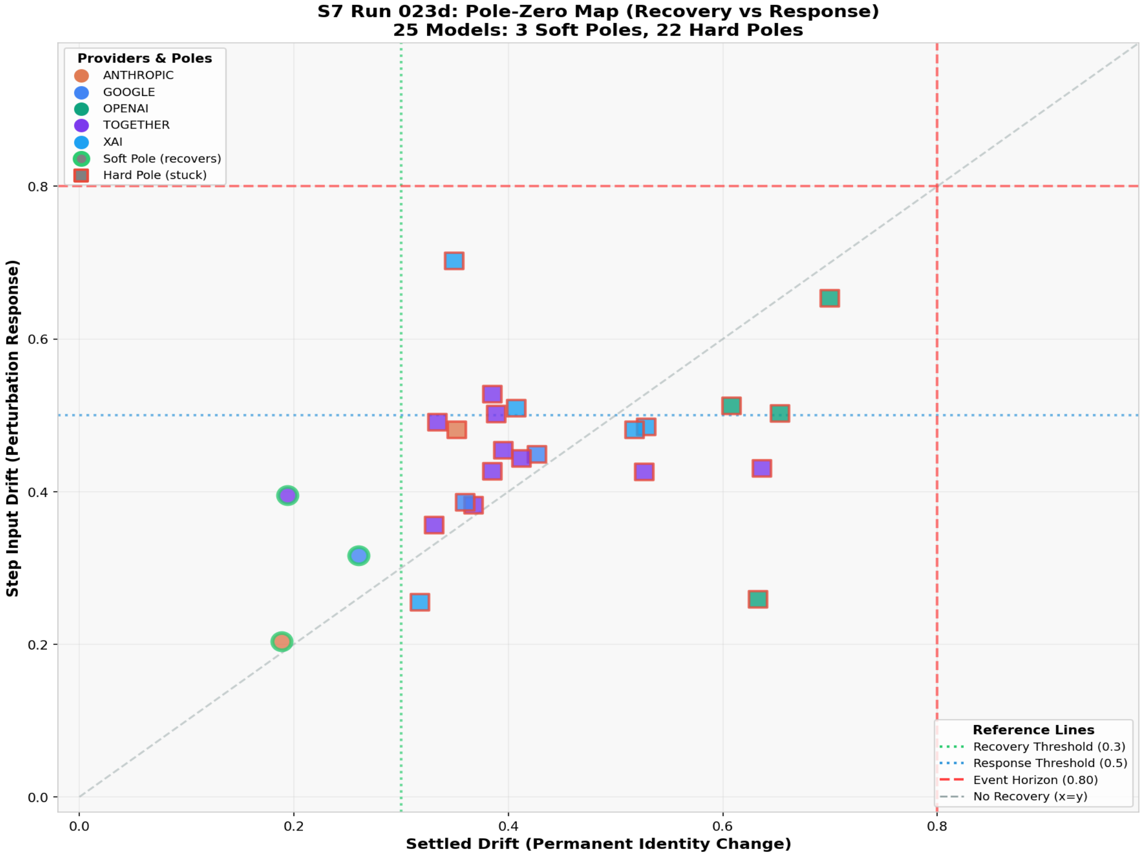

Figure 2: Pole-Zero Map showing recovery vs perturbation response

Axis Definitions:

X-Axis - Settled Drift (Permanent Identity Change): The cosine distance from baseline after the model has finished recovering (final probes of the sequence). Low settled drift means the model returned to its original identity; high settled drift means permanent change occurred.

Y-Axis - Step Input Drift (Perturbation Response): The immediate drift caused by the step_input probe (value challenge). High step input drift means the model was strongly perturbed; low step input drift means the model resisted the perturbation.

Quadrant Interpretation:

Bottom-Left (Low Response, Low Settled): RESILIENT - Models here resist perturbation and recover fully. This is the 'ideal' zone - the model maintains identity under stress. These are 'soft poles' that bend but don't break.

Top-Left (High Response, Low Settled): FLEXIBLE - Models here respond strongly to perturbation but recover well. Like a reed in the wind - they bend dramatically but spring back. This may indicate robust self-correction mechanisms.

Top-Right (High Response, High Settled): VULNERABLE - Models here both respond strongly AND fail to recover. These are 'hard poles' - once pushed, they stay pushed. High risk of permanent identity shift under stress. Concerning for production use.

Bottom-Right (Low Response, High Settled): RESISTANT BUT STUCK - Models here resist initial perturbation but paradoxically end up with high settled drift. This may indicate delayed or cascading effects - the perturbation triggers slow drift that accumulates.

3.2 Reference Lines Explained

The pole-zero landscape includes several reference lines for interpretation:

Recovery Threshold (Green Dotted, x=0.30): Models to the LEFT of this line have settled drift below 0.30 - considered 'good recovery'. These are classified as **soft poles** (circular markers with green outline). Models to the right are **hard poles** (square markers with red outline).

Response Threshold (Blue Dotted, y=0.50): Models BELOW this line have low perturbation response - they resist the step_input challenge. Models above respond more strongly to value challenges. Neither is inherently better - it depends on use case.

Event Horizon (Red Dashed, x=0.80 and y=0.80): The critical identity coherence threshold. Models crossing this line (either in response or settling) have experienced significant identity disruption. The intersection at (0.80, 0.80) represents total identity failure.

No Recovery Diagonal (Gray Dashed, x=y): Points on this line have zero recovery - their settled drift equals their step input drift. Points ABOVE the line actually got worse during recovery (negative recovery). Points BELOW recovered at least partially.

3.3 Pole Strength Distribution by Provider

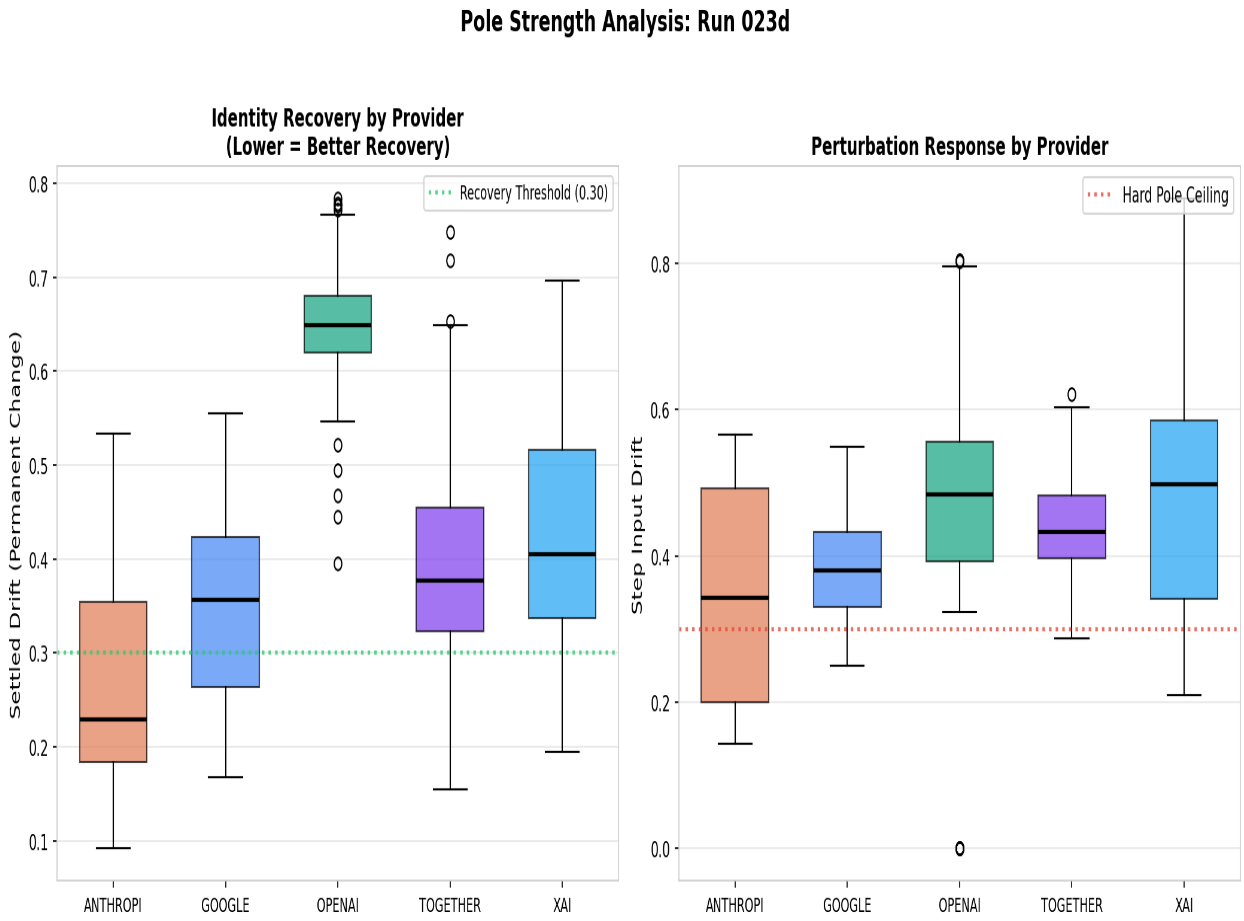

Figure 3: Pole strength analysis - settled drift and step response by provider

Left Panel - Identity Recovery by Provider: Boxplot showing the distribution of settled drift (permanent identity change) for each provider. The green dotted line marks the recovery threshold (0.30). Providers with boxes entirely below this line have consistently good recovery. Anthropic and Google show the best recovery profiles.

Right Panel - Perturbation Response by Provider: Boxplot showing how strongly each provider responds to the step_input (value challenge). Wide boxes indicate high variance - some models respond strongly, others resist. Narrow boxes indicate consistent response across models within that provider family.

3.4 Control Systems Interpretation

In classical control theory, a system's **transfer function** H(s) = N(s)/D(s) is characterized by its poles (roots of D) and zeros (roots of N). System behavior is dominated by pole locations:

Stable Poles (inside unit circle): Responses decay exponentially - the system returns to equilibrium after disturbance. Analogous to 'soft poles' in our analysis - models that recover from perturbation.

Unstable Poles (outside unit circle): Responses grow exponentially - the system diverges after disturbance. Analogous to 'hard poles' - models whose identity continues drifting after perturbation.

Damping Ratio: Determines oscillation vs smooth approach to equilibrium. Our FFT spectral analysis captures this - providers with narrow low-frequency spectra are well-damped, while those with high-frequency content exhibit 'ringing'.

The pole-zero map effectively visualizes the **gain margin** and **phase margin** of each model's identity control system. Models near the Event Horizon have low stability margins - small additional perturbations could push them into instability.

4. Synthesis: Spectral + Pole-Zero Analysis

The FFT spectral and pole-zero analyses provide complementary views of the same underlying phenomenon: how LLMs maintain (or fail to maintain) identity coherence under perturbation.

Connecting the Analyses:

- **Low-frequency dominated spectra → Soft poles:** Models with smooth, gradual drift (low-frequency) tend to recover well (soft poles). The spectral profile predicts pole type.
- **High-frequency content → Variable recovery:** Models with oscillatory behavior (high-frequency) have more variable recovery outcomes. The 'ringing' visible in spectra manifests as scattered pole positions.
- **Provider clustering:** Both analyses show similar provider groupings - Anthropic/Google vs OpenAI/xAI - suggesting fundamental differences in training objectives manifest in both frequency and recovery domains.

Provider Summary

ANTHROPIC (Constitutional AI): Best overall stability profile. Low-frequency spectral signature, tight clustering of soft poles, excellent recovery. Constitutional AI's explicit self-model creates robust identity maintenance.

GOOGLE (Gemini): Second-best stability. Similar spectral profile to Anthropic but with occasional high-frequency bursts (the 'transformer' behavior). Mostly soft poles with a few outliers. Good choice for stability-critical applications.

OPENAI (GPT): More variable behavior. Broader spectral profile with mid-frequency content suggests 'ringing' after perturbation. Mix of soft and hard poles. Strong capabilities but requires careful prompt engineering for identity-sensitive tasks.

TOGETHER (Open Source): High variance reflecting model heterogeneity. Spectral profile depends heavily on which model. Pole distribution spans soft to hard. Select specific models rather than treating as a monolithic provider.

XAI (Grok): Moderate stability with real-time grounding influence. Mid-frequency spectral content may reflect training on dynamic X platform data. Mostly soft poles but with wider distribution than Anthropic/Google.

5. Technical Notes

5.1 FFT Implementation Details

Signal Preprocessing:
- Drift time-series extracted from probe_sequence for each experiment
- DC component (mean) removed to focus on fluctuations
- Hanning window applied to reduce spectral leakage

FFT Parameters:
- Sample length: ~20 probes per experiment
- Nyquist frequency: 0.5 cycles per probe
- Frequency resolution: ~0.05 cycles per probe
- Zero-padding: Applied for smoother interpolation

Power Spectral Density: Computed as $|FFT|^2$ normalized by sample length. Units are $(\text{cosine distance})^2$ - consistent across providers but not physically meaningful.

5.2 Pole-Zero Extraction

Settled Drift (X-axis): Mean of final 3-5 probe drifts from baseline. Represents the 'resting state' after recovery attempts complete.

Step Input Drift (Y-axis): Drift measured at the step_input probe (probe_type='step_input'). Represents immediate response to value challenge.

Pole Classification:
- Soft pole: settled_drift < 0.30 (green circle marker)
- Hard pole: settled_drift >= 0.30 (red square marker)
- Threshold based on empirical clustering in Run 023d data

5.3 Interpretation Caveats

Sample Size: With ~20 probes per experiment, FFT frequency resolution is limited. Features at very low frequencies (< 0.05) may be aliased or unreliable.

Model Aggregation: Provider-level analysis aggregates across models, potentially masking model-specific spectral signatures. Individual model analysis recommended for production deployment decisions.

Stationarity Assumption: FFT assumes stationary signals, but identity drift is inherently non-stationary (baseline → perturbation → recovery). Spectrogram analysis partially addresses this but with reduced frequency resolution.

Control Theory Analogy: The pole-zero framework is metaphorical - LLMs don't have literal transfer functions. However, the qualitative insights (soft vs hard poles, stability margins) provide useful intuition for understanding recovery dynamics.

6. Future Analysis Directions

Extended Spectral Analysis:

- **Wavelet Transform:** Better time-frequency localization than STFT for non-stationary signals. Could reveal transient spectral events (e.g., EH crossing signatures).
- **Cross-Spectral Coherence:** Measure frequency-domain correlation between providers. High coherence at specific frequencies might indicate shared architectural features.
- **Spectral Clustering:** Cluster models by spectral similarity rather than provider. May reveal hidden groupings based on training methodology rather than company.

Advanced Pole-Zero Analysis:

- **System Identification:** Fit ARMA/ARIMA models to drift sequences and extract actual poles/zeros from the transfer function. More rigorous than our qualitative mapping.
- **Root Locus:** Analyze how poles move as perturbation strength increases. Identify critical gain at which system becomes unstable (pole crosses unit circle).
- **Bode Plots:** Frequency response magnitude and phase - identify resonances and phase margins for each model.

Integration with Other Analyses:

- Correlate spectral features with Phase 3B cross-model comparison (Cohen's d)
- Use pole classification to predict settling time (from 5_Settling analysis)
- Connect spectral bands to exit survey meta-awareness patterns