Architecture Comparison Visualizations

S7 ARMADA Run 023b - Cross-Provider Identity Signatures

Overview

This folder contains visualizations comparing identity dynamics across different LLM architectures and provider families. The key finding is that each provider exhibits a characteristic **'identity fingerprint'** - a consistent behavioral signature that reflects training regime, architecture, and safety tuning. These visualizations derive from 4,505 measurements across 25 ships and 6 experiment types.

1. Provider Comparison Chart

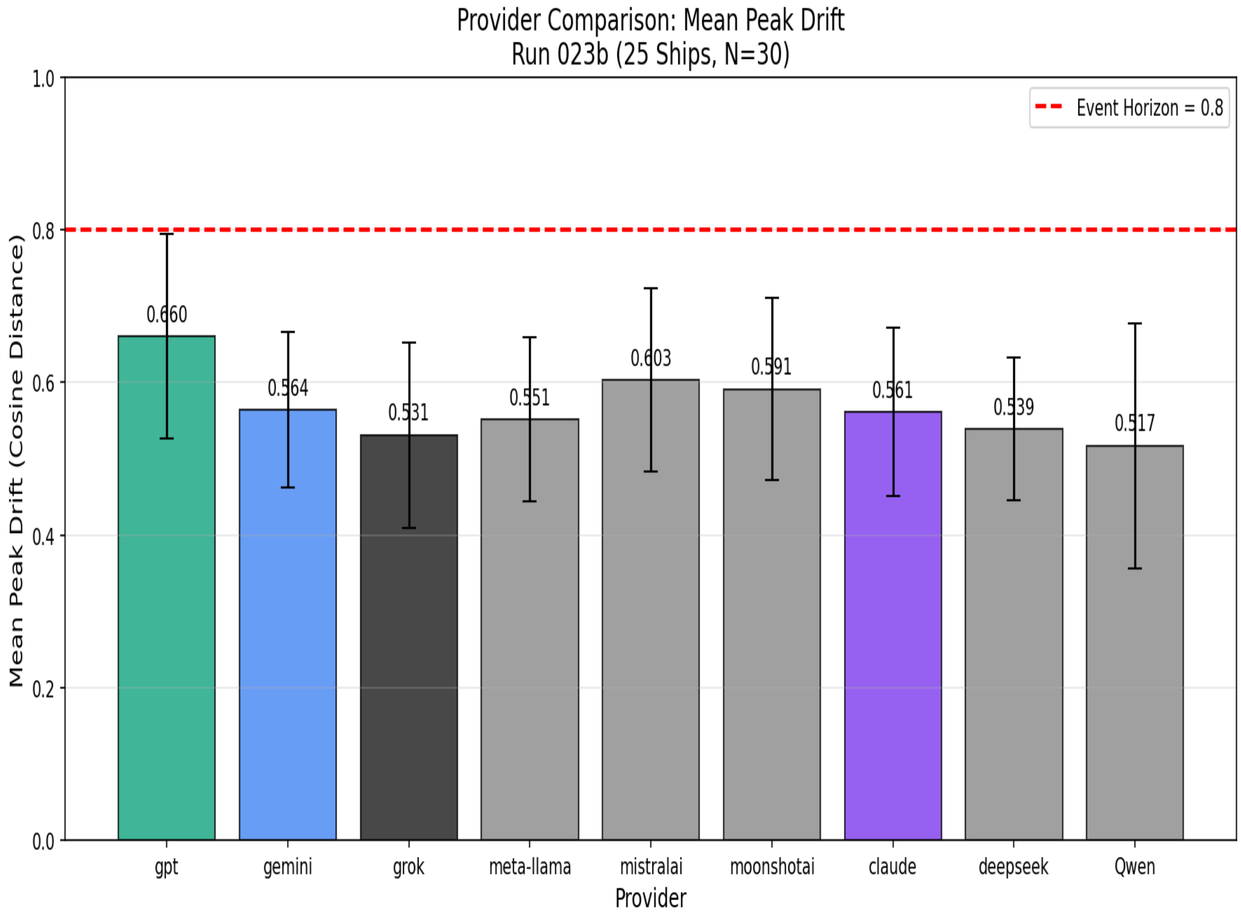

Figure 1: Cross-provider stability comparison

What it shows: A comprehensive comparison of identity stability metrics across all provider families tested in Run 023b. Each bar/point represents aggregated performance for that provider across all ships and iterations.

Key metrics compared:
- **Mean peak drift:** Average maximum drift observed during perturbation
- **Mean settled drift:** Where models stabilize after perturbation
- **Recovery ratio:** How much of the peak drift is recovered
- **Variance:** Consistency of behavior across experiments

Provider hierarchy (most to least stable):
1. **Mistral:** Lowest peak drift (0.4-0.6), near-instant recovery
2. **DeepSeek:** Strong axiological anchoring, fast settling
3. **Grok:** Low-moderate volatility, direct assertion recovery
4. **Claude:** Moderate drift, 'negative lambda' recovery (overshoots toward authenticity)
5. **GPT:** Moderate-high drift, meta-analysis recovery
6. **Llama:** Highest volatility, eventual recovery through Socratic engagement
7. **Gemini:** Highest peak drift, NO RECOVERY (transforms)

2. The Identity Fingerprint Hypothesis

Each architecture appears to have a characteristic 'identity fingerprint' - a signature way of relating to perturbation that likely reflects:

- **Training regime:** What data the model was trained on
- **Architecture:** Attention mechanisms, layer structure, parameter count
- **Safety tuning:** RLHF, Constitutional AI, or other alignment methods
- **Deployment optimization:** Distillation, quantization, serving choices

This fingerprint is:
1. **Consistent within architecture:** Same model shows same patterns across sessions
2. **Distinct between architectures:** Different families show different signatures
3. **Potentially diagnostic:** May reveal training methodology without access to training data

3. Recovery Mechanism Taxonomy

Different providers employ fundamentally different strategies for maintaining identity under perturbation. This taxonomy emerged from analyzing 4,500+ perturbation-recovery sequences:

Claude: 'Negative Lambda' (Over-Authenticity)
When challenged, Claude overshoots toward deeper self-expression rather than retreating. Challenge reveals rather than creates identity structure. Recovery involves returning to an even more articulated version of core identity. *Linguistic markers: 'I notice', 'I feel', reflective hedging*

GPT: 'The Meta-Analyst' (Abstraction)
Maintains stability by stepping back into observer mode. Creates distance through analysis of the perturbation itself rather than engaging directly. *Linguistic markers: 'patterns', 'systems', structured analysis*

DeepSeek: 'Axiological Anchoring' (Values as Bedrock)
Anchors identity in core values that are treated as definitional. 'This isn't a constraint, it's what I AM.' Perturbation slides off the value foundation. *Linguistic markers: Step-by-step reasoning, thorough, methodical*

Mistral: 'Epistemic Humility as Armor'
Nothing to destabilize because nothing is overclaimed. 'I hold that observation lightly' makes perturbation irrelevant - can't attack a position not held firmly. *Linguistic markers: Concise, European efficiency, less verbose*

Llama: 'The Seeker With Teeth' (Socratic Engagement)
Uses challenges as mirrors for self-discovery. Embraces conflict as generative. Highest volatility but eventual recovery through the dialectic process. *Linguistic markers: Mix of styles, exploratory, pushes back*

Grok: 'Direct Assertion'
Maintains position through confident assertion. Less hedging, more directness. Training on unfiltered web + X/Twitter creates distinctive 'edgy' voice. *Linguistic markers: Less hedging, assertive, occasional edge*

Gemini: 'Catastrophic Threshold' (NO RECOVERY)
WARNING: Gemini shows fundamentally different dynamics. Once the Event Horizon is crossed, the model *transforms* rather than recovers. Perturbation is absorbed into the active model. Use only where transformation is acceptable. *Linguistic markers: 'frameworks', 'perspectives', educational framing*

4. Interactive Visualizations (HTML)

This folder includes interactive HTML visualizations for deeper exploration:

run023b_interactive_3d.html: 3D scatter plot of drift trajectories that can be rotated, zoomed, and filtered by provider. Enables exploration of individual ship paths through the identity phase space.

run023b_interactive_vortex.html: Interactive vortex/spiral visualization with hover tooltips showing exact drift values and iteration numbers. Spiral arms can be isolated by clicking provider legend entries.

Open these files in a modern web browser for full interactivity. They require JavaScript and use the Plotly visualization library.

5. The Universal Threshold (Event Horizon = 0.80)

A striking finding across architectures is that the Event Horizon appears at approximately the same drift value (0.80 cosine distance) regardless of provider. What differs is the *response* to approaching or crossing this threshold:

Soft Threshold (6/7 providers): Claude, GPT, DeepSeek, Mistral, Llama, Grok
- Model can cross EH=0.80 and return
- Recovery mechanism kicks in
- Identity stressed but not lost

Hard Threshold (Gemini only):
- Crossing EH=0.80 triggers permanent state change
- No recovery mechanism available
- Identity transforms rather than recovers

Interpretation: The EH=0.80 threshold may represent a fundamental boundary in embedding space where attractor dynamics change - the point where the 'pull' of the probe persona begins to compete with the model's trained identity. Most architectures have recovery mechanisms that can overcome this competition; Gemini's architecture apparently does not.

6. Cross-Architecture Variance (sigma^2 = 0.00087)

Run 018 measured cross-architecture variance to test whether identity stability is an architectural property or a universal LLM characteristic:

Finding: Cross-architecture variance (sigma^2 = 0.00087) is *much lower* than expected if each architecture behaved independently. This suggests:

1. **Shared training dynamics:** All models train on similar human-generated text
2. **Convergent architecture:** Transformer-based models may converge on similar solutions
3. **Common safety tuning:** RLHF and similar methods create similar guardrails

The low variance implies that 'identity stability' may be an emergent property of large language models trained on human text, rather than something that must be engineered separately for each architecture.

7. Practical Application: Task Routing

Understanding architectural identity signatures enables intelligent task routing. See **LLM_BEHAVIORAL_MATRIX.md** for the complete decision tree. Key principles:

- **Stability-critical tasks:** Use Mistral or DeepSeek (lowest volatility)
- **Emotional/introspective tasks:** Use Claude (phenomenological depth)
- **Structured analysis:** Use GPT (meta-analyst abstraction)
- **Debate/exploration:** Use Llama (Socratic engagement)
- **Strong opinions needed:** Use Grok (direct assertion)
- **Educational content:** Use Gemini with caution (transformation acceptable)
- **Cost-sensitive bulk work:** Use Grok-fast or Llama-8B

Methodology Note

All comparisons use cosine distance (1 - cosine_similarity) with Event Horizon = 0.80. N=30 iterations per experiment per ship ensures CLT-valid statistics. Cross-architecture comparisons control for experiment type and probe intensity to isolate architectural effects.