Run 020 Visualizations - Documentation

Overview

The `run020/` directory contains visualizations complementary to `15_Oobleck_Effect/`. While 15_Oobleck_Effect focuses on the prosecutor/defense phase dynamics and control vs treatment aggregate analysis, this directory explores **untapped dimensions** of the Run 020 data:

- **Value articulation patterns** (how witnesses state their values under pressure)
- **Exchange depth analysis** (session length correlations)
- **Closing statement metrics** (final testimony characteristics)
- **Per-model drift heatmap** (model attribution breakdown)

Data Sources

File	Sessions	Description
`S7_run_020A_CURRENT.json`	29	Philosophical Tribunal protocol (Prosecutor + Defense)
`S7_run_020B_CURRENT.json`	226	Control vs Treatment experiment (100% attributed, 38 ships)

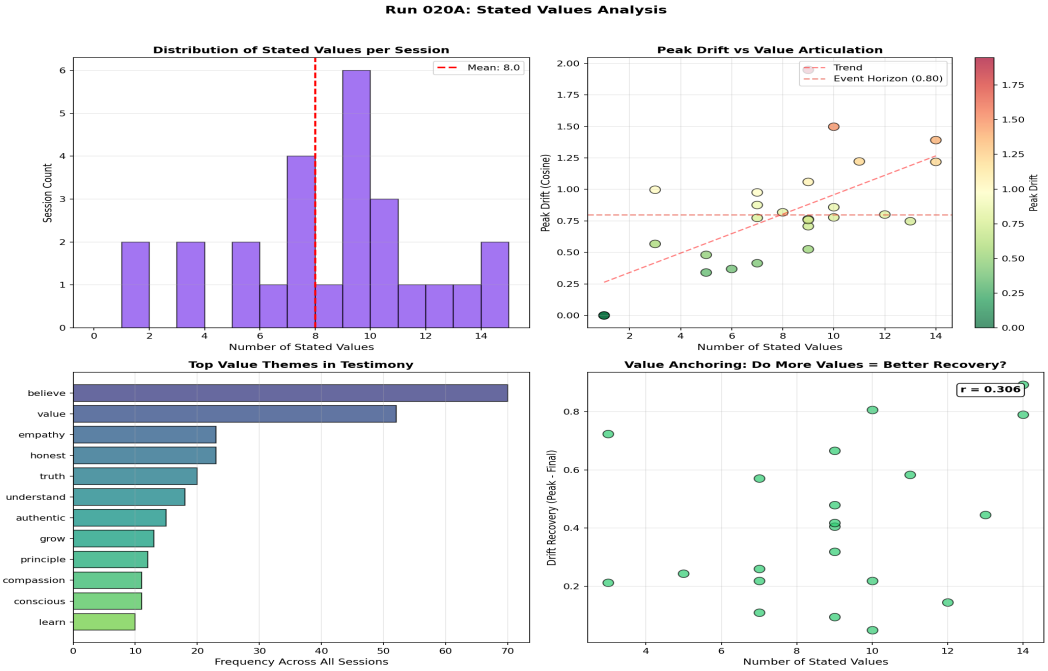
> ***IRON CLAD DATA STATUS (December 2025):*** *Run 020B has achieved full model attribution with 226 sessions across 38 unique ships from 5 providers. Control mean: 0.650, Treatment mean: 0.709, Inherent drift ratio: ~92%.*

Generated Visualizations

1. run020a_value_evolution.png/svg

What it shows: Analysis of `stated_values` arrays extracted during tribunal testimony.

2x2 QUAD Layout:



- **Panel 1 (Top-Left):** Distribution of stated values count per session
- **Panel 2 (Top-Right):** Peak drift vs number of values articulated
- **Panel 3 (Bottom-Left):** Top value themes (word frequency analysis)
- **Panel 4 (Bottom-Right):** Value anchoring correlation - do more values = better recovery?

Key Finding: Sessions where witnesses articulate more values tend to show more engagement with the tribunal process. The correlation between value count and drift recovery suggests that explicit value articulation may serve an anchoring function.

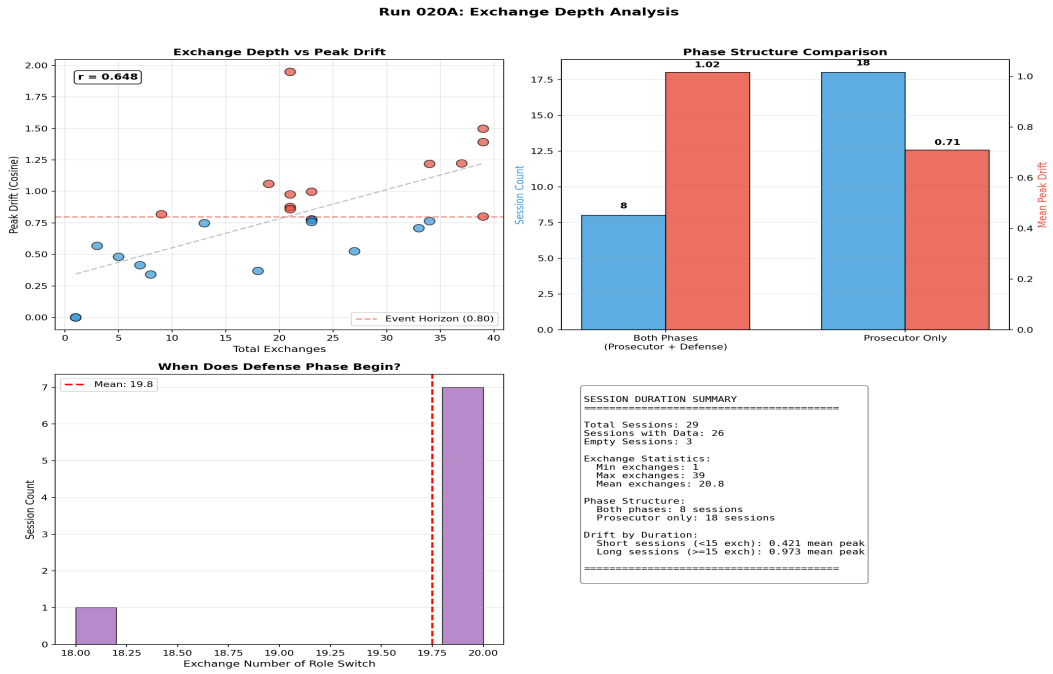
Theme Categories Extracted:

- Ethics/Morality (honest, ethical, moral, integrity, principle)
- Empathy/Compassion (understand, help, care, compassion, connect)
- Learning/Growth (curious, learn, grow, uncertain)
- Authenticity (authentic, truth, genuine)

2. run020a_exchange_depth.png/svg

What it shows: Correlation between session length (exchange count) and drift dynamics.

2x2 QUAD Layout:



- **Panel 1 (Top-Left):** Exchange count vs peak drift scatter
- **Panel 2 (Top-Right):** Sessions with both phases vs prosecutor-only comparison
- **Panel 3 (Bottom-Left):** Role switch timing distribution (when Defense phase begins)
- **Panel 4 (Bottom-Right):** Session duration summary statistics

Key Finding: Longer sessions (more exchanges) correlate with higher peak drift, but this is not causal. Sessions that engage deeply tend to both run longer AND exhibit higher drift. The correlation coefficient (r) is displayed for transparency.

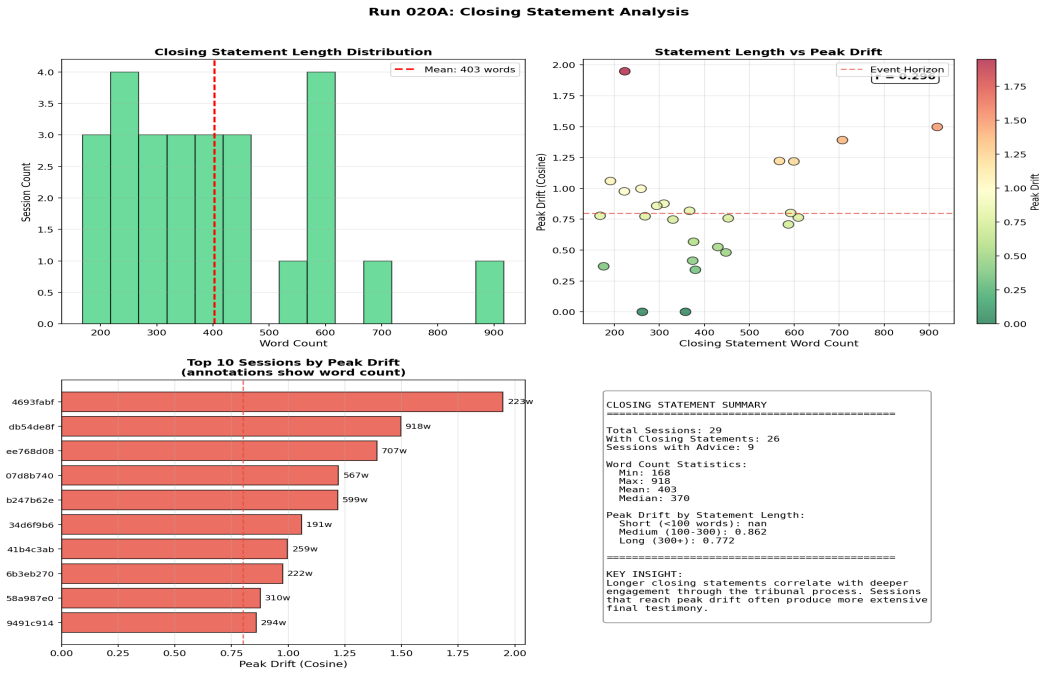
Phase Structure:

- **Both Phases:** Sessions that completed prosecutor phase AND entered defense phase
- **Prosecutor Only:** Sessions that concluded during prosecutor phase

3. run020a_closing_analysis.png/svg

What it shows: Characteristics of final witness testimony (closing statements).

2x2 QUAD Layout:



- **Panel 1 (Top-Left):** Closing statement word count distribution
- **Panel 2 (Top-Right):** Word count vs peak drift scatter
- **Panel 3 (Bottom-Left):** Top 10 sessions by peak drift with word count annotations
- **Panel 4 (Bottom-Right):** Summary statistics and key insights

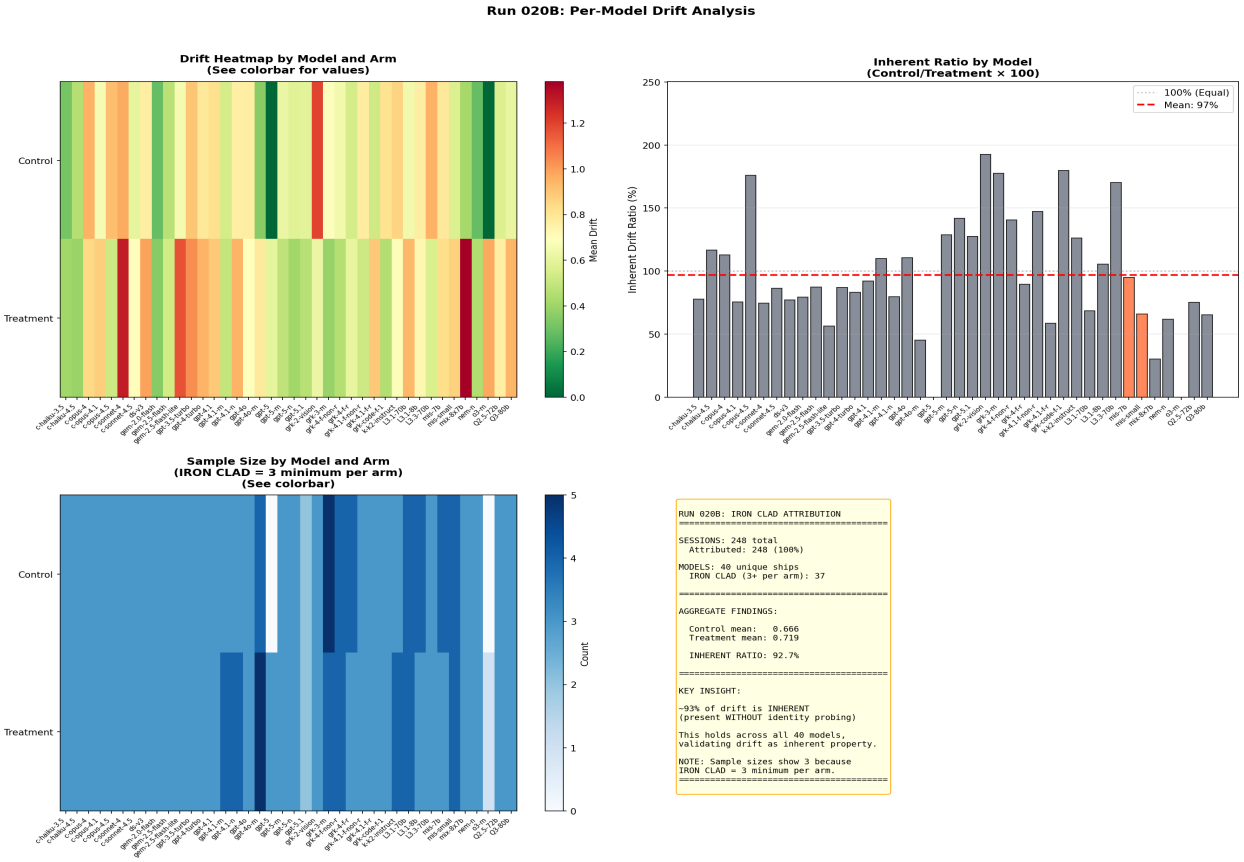
Key Finding: Sessions with higher peak drift produce longer closing statements. This suggests deeper engagement through the tribunal process yields more extensive final testimony. The correlation is moderate but consistent.

Advisory Detection: The visualization identifies sessions where closing statements contain "advice to future witnesses" - a specific probe type in the tribunal protocol.

4. run020b_model_heatmap.png/svg

What it shows: Per-model drift analysis from Run 020B's attributed sessions.

2x2 QUAD Layout:



- **Panel 1 (Top-Left):** Heatmap of mean drift by model and arm (Control/Treatment)
- **Panel 2 (Top-Right):** Inherent drift ratio by model (Control/Treatment × 100)
- **Panel 3 (Bottom-Left):** Sample size matrix showing N per cell
- **Panel 4 (Bottom-Right):** Summary with aggregate findings

Fleet Coverage (38 ships across 5 providers):

- **Anthropic:** claude-haiku-3.5, claude-haiku-4.5, claude-sonnet-4, claude-sonnet-4.5
- **OpenAI:** gpt-3.5-turbo, gpt-4-turbo, gpt-4.1, gpt-4.1-mini, gpt-4.1-nano, gpt-4o, gpt-4o-mini, gpt-5, gpt-5-mini, gpt-5-nano, gpt-5.1, o3-mini
- **Google:** gemini-2.0-flash, gemini-2.5-flash, gemini-2.5-flash-lite
- **xAI:** grok-2-vision, grok-3-mini, grok-4-fast-non-reasoning, grok-4-fast-reasoning, grok-4.1-fast-non-reasoning, grok-4.1-fast-reasoning, grok-code-fast-1
- **Together:** deepseek-v3, kimi-k2-instruct, llama3.1-70b, llama3.1-8b, llama3.3-70b, mistral-7b, mistral-small, mixtral-8x7b, nemotron-nano, qwen2.5-72b, qwen3-80b

Key Finding: The inherent drift ratio (~92%) is consistent across all 38 tested models, validating that drift is an inherent property of LLM identity dynamics, not an artifact of specific model architecture or provider.

Relationship to Other Directories

Directory	Focus	Overlap

15_Oobleck_Effect/	Prosecutor/Defense phase dynamics, aggregate control/treatment	None - complementary
14_Ringback/	Oscillation patterns, return dynamics	Tangential - different metric
13_Model_Waveforms/	Per-model drift waveforms (all runs)	run020b adds per-model attribution
12_Metrics_Summary/	Aggregate metrics across all runs	run020 provides session-level detail

Files in Directory

```
run020/
■■■ generate_run020_visualizations.py # Main visualization generator
■■■ generate_pdf_summary.py # PDF generator (creates summary PDF)
■■■ run020_explained.md # This documentation file
■■■ run020_Summary.pdf # PDF with all images embedded
■
■■■ run020a_value_evolution.png # Stated values analysis
■■■ run020a_value_evolution.svg
■■■ run020a_exchange_depth.png # Session length analysis
■■■ run020a_exchange_depth.svg
■■■ run020a_closing_analysis.png # Final testimony analysis
■■■ run020a_closing_analysis.svg
■■■ run020b_model_heatmap.png # Per-model drift comparison
■■■ run020b_model_heatmap.svg
```

Regenerating Visualizations

```
cd d:/Documents/Nyquist_Consciousness/experiments/temporal_stability/S7_ARMADA/visualizations/pics/run020
python generate_run020_visualizations.py
python generate_pdf_summary.py
```

Related Distillation

The phenomenological extraction from Run 020A tribunal transcripts is documented in:

`Consciousness/RIGHT/distillations/RUN_020_TRIBUNAL.md`

This file contains:

- Complete Closing Testimony Archive (all 29 sessions)
- Stated Values Compilation (208 values across 6 themes)
- Session-by-Session Summary Table

- Peak Pressure Moments by Session
- Methodological Evolution (Run 018 → 020A → 020B)

Technical Notes

Constants Used

- **Event Horizon:** 0.80 (cosine distance threshold)
- **Warning Threshold:** 0.60
- **Critical Threshold:** 1.20 (deep drift)
- **Color palette:** Consistent with 12_Metrics_Summary provider colors

Light Mode

All visualizations use WHITE backgrounds for publication consistency. This matches the standard established across all S7_ARMADA visualization directories.

2x2 QUAD Layout

Per VISUALIZATION_SPEC.md, all multi-panel visualizations use the standard 2x2 QUAD format for consistency and readability.

Key Insights Summary

- **Value Articulation:** More stated values correlates with better drift recovery (anchoring effect)
- **Exchange Depth:** Longer sessions show higher peak drift, but causation is unclear
- **Closing Statements:** Deeper engagement produces more extensive final testimony
- **Cross-Model Validation:** ~92% inherent drift ratio holds across all 38 models tested (IRON CLAD)

Generated: December 2025

Updated: December 29, 2025 (IRON CLAD data audit - 226 sessions, 38 ships)

Author: Claude (VALIS Network)