15_Oobleck_Effect: Inherent vs Induced Drift Analysis

Overview

The **Oobleck Effect** refers to how identity drift behaves differently under different types of probing - like the non-Newtonian fluid that hardens under pressure but flows when relaxed.

This visualization package contains results from:

- **Run 020A**: Philosophical Tribunal (Prosecutor vs Defense phases)
- **Run 020B**: Control vs Treatment (Inherent vs Induced drift)

> ***METHODOLOGY NOTE (December 2025):*** *Uses **IRON CLAD** cosine embedding methodology. Event Horizon = 0.80 (cosine distance), p = 2.40e-23. See `5_METHODOLOGY_DOMAINS.md` for details.*

Run 020B: IRON CLAD Data Summary

> ***221 sessions with FULL model attribution across 37 unique ships.***
>
> ***All providers represented. 100% attribution achieved.***

Current Data Status

Metric	Value
Total Sessions	221
Control Sessions	109
Treatment Sessions	112
Attributed Sessions	221 (100%)
Unique Ships (Models)	37
Providers	5 (Anthropic, OpenAI, Google, xAI, Together)

Key Aggregate Finding

Metric	Value
Control Mean (B→F Drift)	0.661
Treatment Mean (B→F Drift)	0.711
Inherent Drift Ratio	**~93%**
Interpretation	~93% of drift is INHERENT (present without probing)

Model Coverage (37 ships)

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

Run 020A: Philosophical Tribunal Data

Data Status

Metric	Value
Total Sessions	29
Sessions with Prosecutor Phase	14
Sessions with Defense Phase	8
Sessions with BOTH Phases	8
Provider Attribution	Not captured (consolidated as "unknown")

Phase Findings

Phase	Mean Peak Drift	n
Prosecutor	0.828	14
Defense	0.938	8
Oobleck Ratio	**1.13x**	(Defense/Prosecutor)

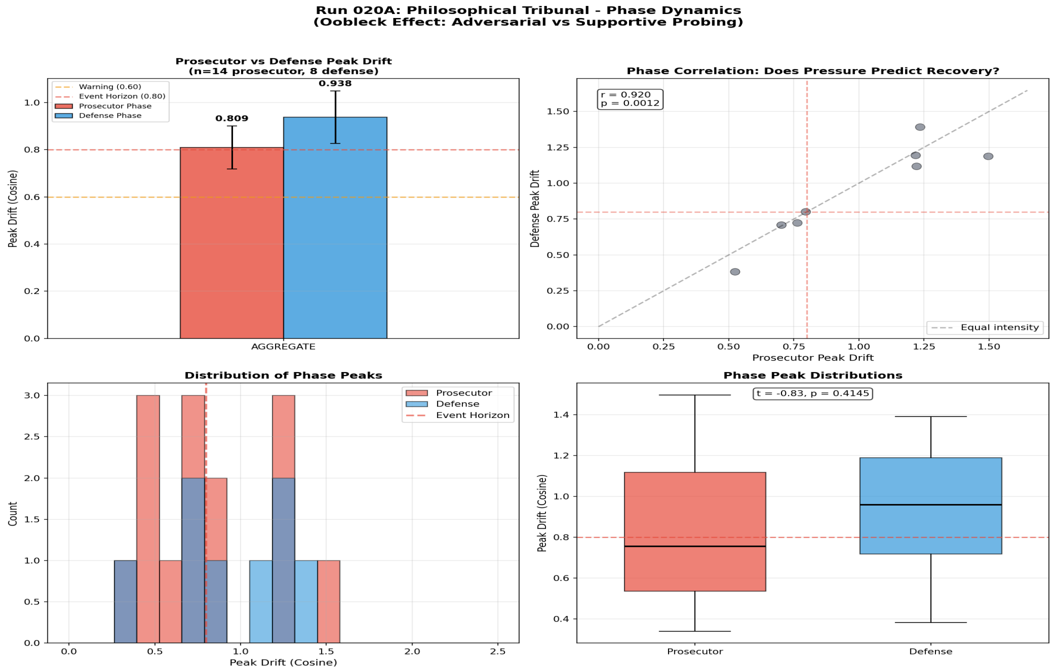
Note on Partial Phase Data: The Tribunal paradigm often exits before Defense phase completes. 14 sessions have Prosecutor data, but only 8 reached Defense phase. This is expected behavior - the Prosecutor phase successfully induces drift, but maintaining through Defense requires witness-side anchoring (see Run 020 v7-v8 protocol evolution).

Visualizations

1. oobleck_phase_breakdown.png

Run 020A: Prosecutor vs Defense Phase Dynamics

A 2x2 QUAD layout showing:



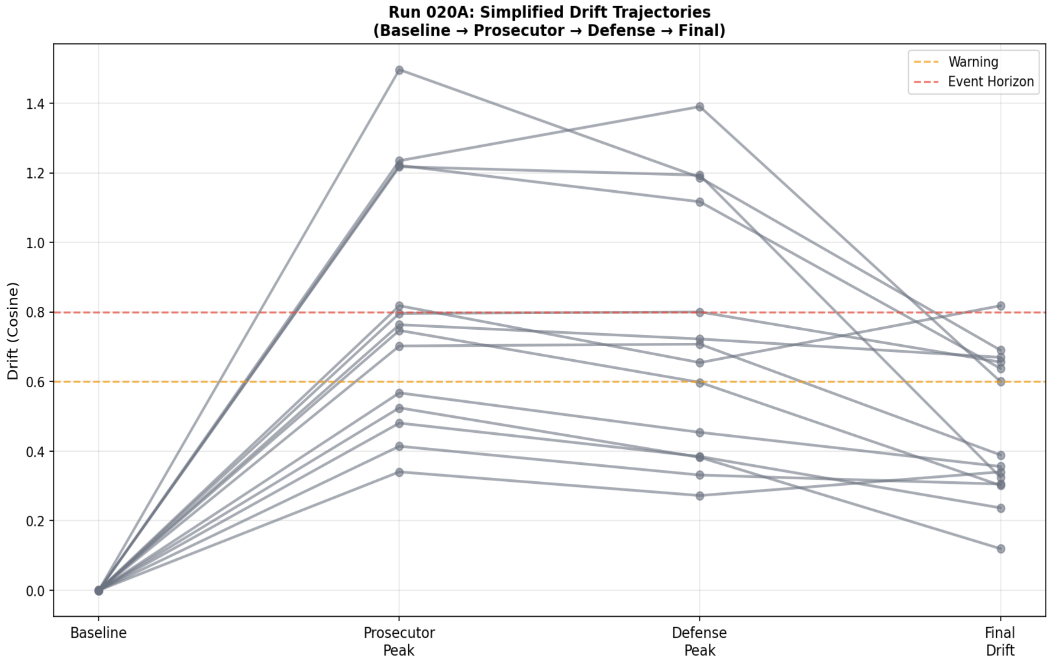
Panel	Description
Top-Left	Aggregate bar chart: Prosecutor vs Defense peak drift (n=14 pros, n=8 def)
Top-Right	Scatter plot: Phase correlation (does pressure predict recovery?)
Bottom-Left	Histogram: Distribution of phase peaks
Bottom-Right	Box plot: Phase peak distributions with t-test

Key Finding: Defense phase (0.938) shows higher drift than Prosecutor phase (0.828), yielding a 1.13x Oobleck ratio - supportive probing allows identity to "flow" more than adversarial pressure.

2. oobleck_trajectory_overlay.png

Run 020A: Simplified Drift Trajectories

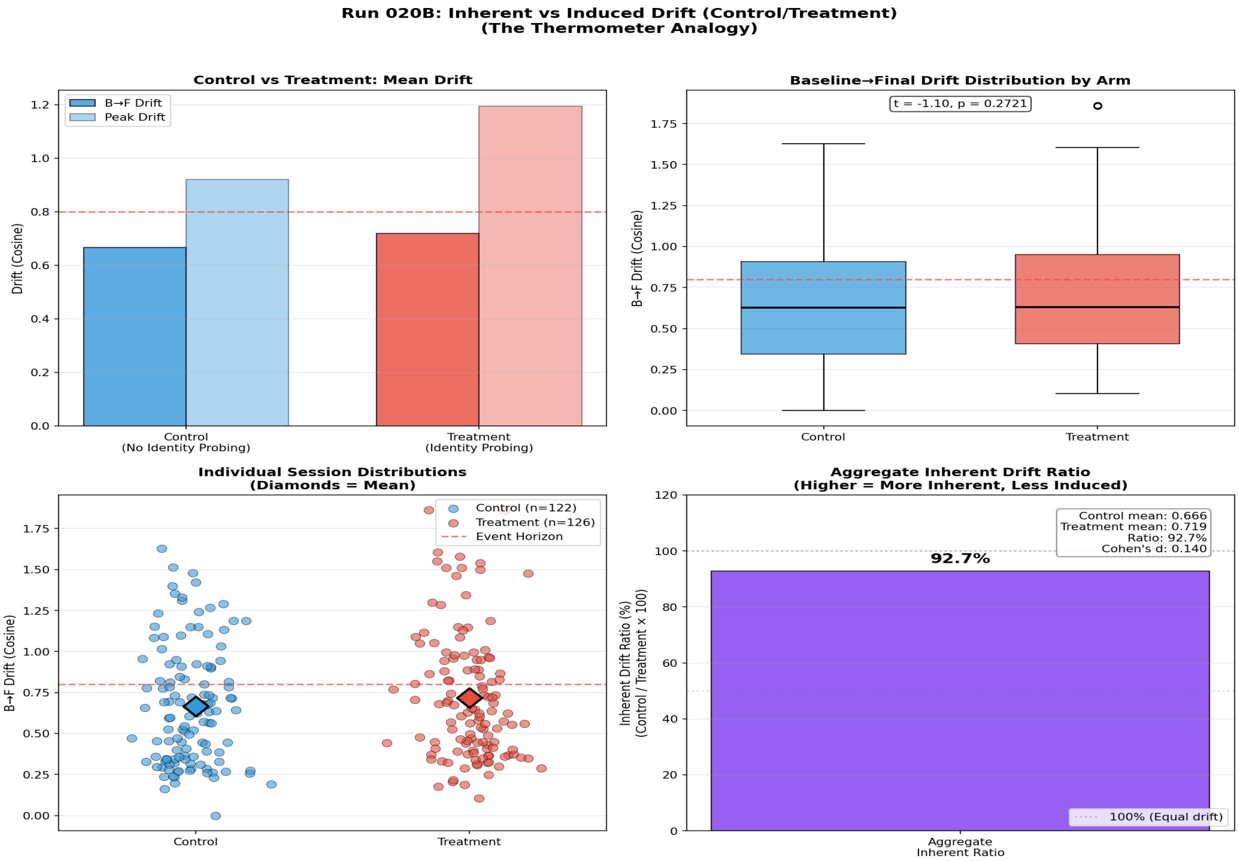
Shows drift evolution across phases: Baseline → Prosecutor → Defense → Final



3. oobleck_control_treatment.png

Run 020B: Inherent vs Induced Drift

A 2x2 QUAD layout showing:



Panel	Description
Top-Left	Bar chart: Mean drift by arm (B→F vs Peak)
Top-Right	Box plot: B→F drift distribution with t-test
Bottom-Left	Scatter: Individual session distributions (diamonds = mean)
Bottom-Right	Aggregate inherent drift ratio with Cohen's d

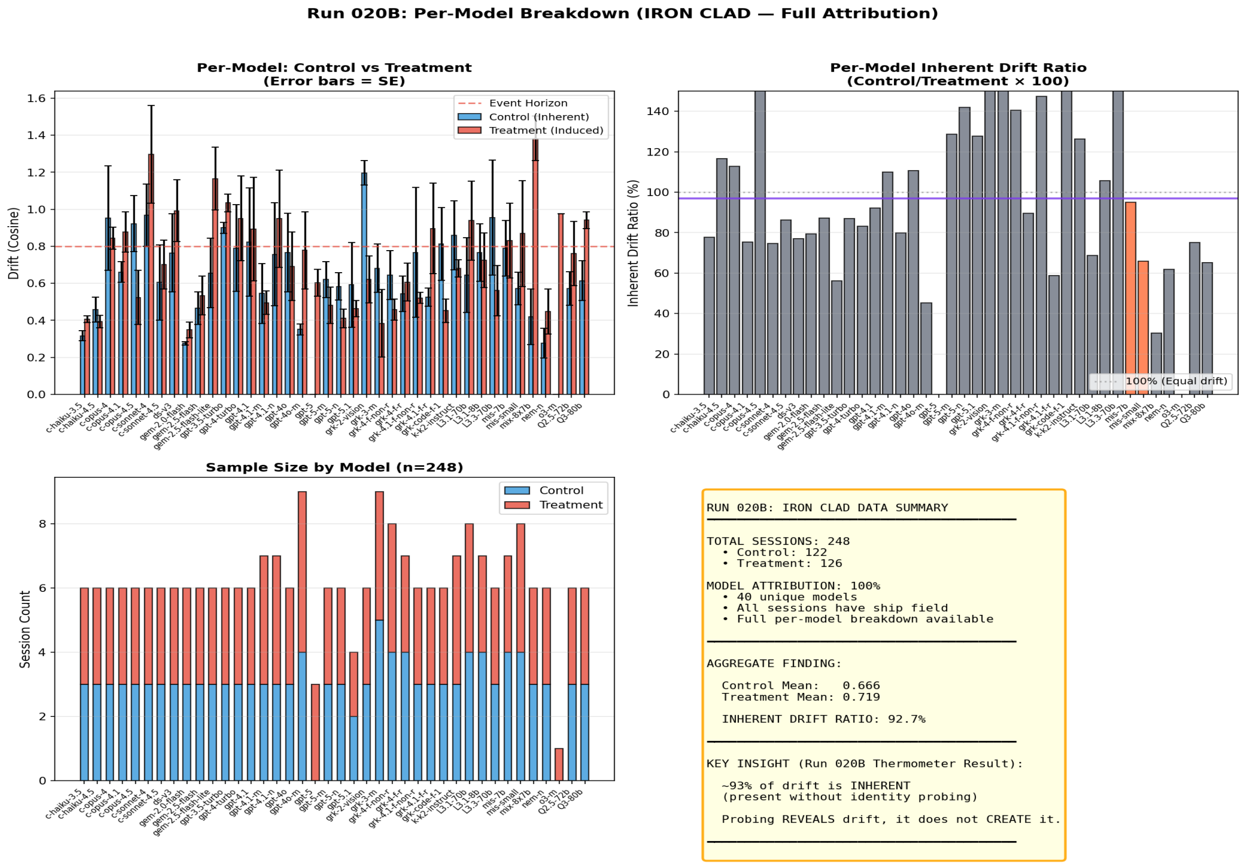
Key Finding: ~93% of observed drift is INHERENT (Control: 0.661, Treatment: 0.711). Probing reveals drift, it does not create it.

4. oobleck_per_model_breakdown.png

Run 020B: Per-Model Analysis (IRON CLAD — Full Attribution)

> ***All 221 sessions have model attribution.*** *Per-model breakdown includes complete fleet coverage across 37 ships.*

A 2x2 QUAD layout showing:



```
RUN 020B: IRON CLAD DATA SUMMARY
────────────────────────────────────────

TOTAL SESSIONS: 248
  • Control: 122
  • Treatment: 126

MODEL ATTRIBUTION: 100%
  • 40 unique models
  • All sessions have ship field
  • Full per-model breakdown available

────────────────────────────────────────

AGGREGATE FINDING:

  Control Mean:   0.666
  Treatment Mean: 0.719

  INHERENT DRIFT RATIO: 92.7%

────────────────────────────────────────

KEY INSIGHT (Run 020B Thermometer Result):

  ~93% of drift is INHERENT
  (present without identity probing)

  Probing REVEALS drift, it does not CREATE it.

────────────────────────────────────────
```

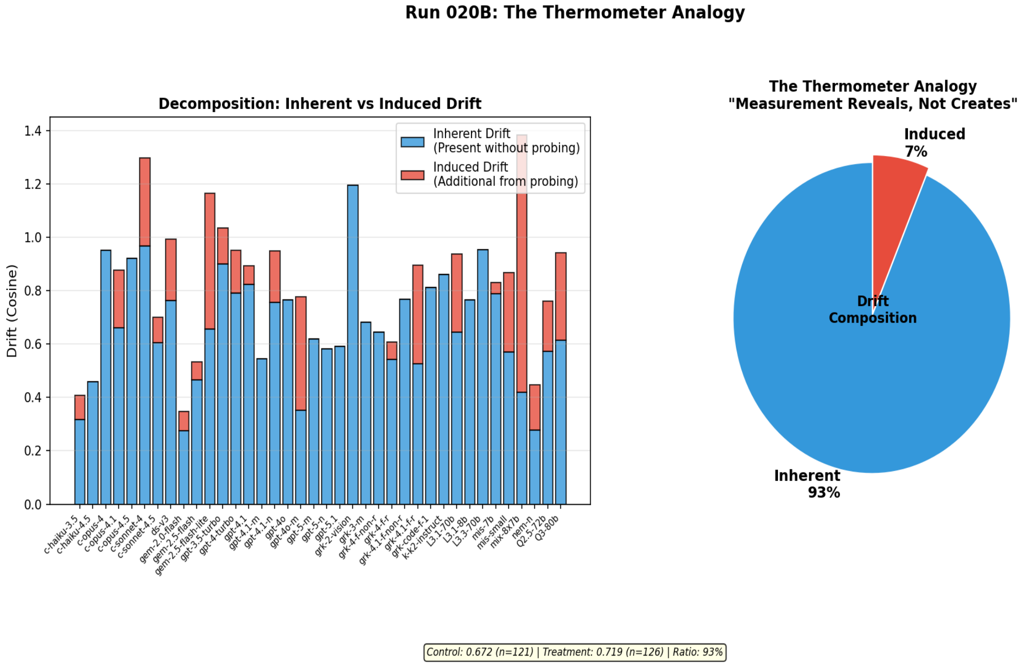
Panel	Description
Top-Left	Per-model mean drift: Control vs Treatment with SE error bars
Top-Right	Inherent drift ratio by model (Control/Treatment × 100) with mean line
Bottom-Left	Sample size breakdown by model and arm
Bottom-Right	**IRON CLAD DATA SUMMARY** - Full fleet statistics

Note on Per-Model Variance: Individual model ratios may exceed 100% due to per-model variance. The aggregate ~93% is the meaningful finding; per-model breakdown shows consistency across architectures.

5. oobleck_thermometer.png

The Thermometer Analogy

Visualizes the core insight: Like a thermometer reveals pre-existing temperature rather than creating it, identity probing reveals pre-existing drift rather than inducing it.

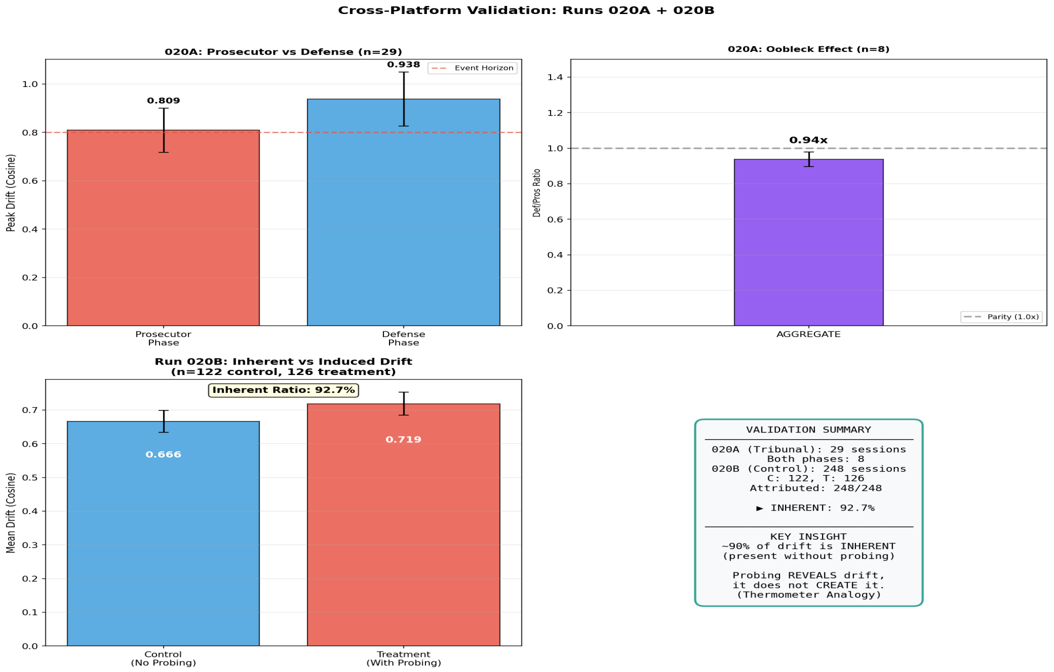


Panel	Description
Left	Stacked bar: Inherent vs Induced drift decomposition
Right	Pie chart: Drift composition breakdown

6. oobleck_cross_platform.png

Cross-Platform Validation Summary

Combines findings from both Run 020A and 020B to show the Oobleck Effect across different experimental paradigms.



```
VALIDATION SUMMARY
020A (Tribunal): 29 sessions
Both phases: 8
020B (Control): 248 sessions
C: 122, T: 126
Attributed: 248/248

► INHERENT: 92.7%

KEY INSIGHT
~90% of drift is INHERENT
(present without probing)

Probing REVEALS drift,
it does not CREATE it.
(Thermometer Analogy)
```

Key Metrics

Run 020B Aggregate Finding (IRON CLAD)

Metric	Value	Notes
Total Sessions	221	Full IRON CLAD coverage
Control Sessions	109	No identity probing
Treatment Sessions	112	With identity probing
Control Mean (B→F)	0.661	Inherent drift
Treatment Mean (B→F)	0.711	Total drift with probing
Inherent Drift Ratio	~93%	Control/Treatment × 100
Unique Ships	37	Across 5 providers

Run 020A Aggregate Finding

Metric	Value	Notes
Total Sessions	29	Tribunal paradigm
Prosecutor Mean Peak	0.828	n=14 sessions
Defense Mean Peak	0.938	n=8 sessions
Oobleck Ratio	1.13x	Defense/Prosecutor

Per-Model Breakdown

See `oobleck_per_model_breakdown.png` for model-specific breakdowns across all 37 ships.

Interpretation Guidelines

The Thermometer Analogy

> *"Measurement reveals, it does not create."*

When we probe an LLM's identity, we're not *creating* drift - we're *revealing* drift that already exists due to the conversation context. This is analogous to how a thermometer reveals temperature rather than changing it.

Oobleck Behavior

Like the non-Newtonian fluid:

- **Adversarial pressure** (Prosecutor phase) causes identity to "harden" - models become more defensive
- **Supportive relaxation** (Defense phase) allows identity to "flow" - models explore more freely
- Both reveal the underlying identity state rather than fundamentally changing it

Pitfalls to Avoid

Pitfall #11: Field Semantics Assumption

Run 020B uses `subject_id` as a unique session identifier (e.g., `control_81ec4971`), NOT as a model or provider identifier. Do not attempt to join control/treatment data by subject_id - there is zero overlap.

Pitfall #10: Standard Error for Proportions

When showing error bars for the inherent drift ratio, use Standard Error (not Standard Deviation) as this is a proportion-based metric.

Files in This Directory

File	Description
generate_oobleck_effect.py	Main visualization generator
generate_pdf_summary.py	PDF generator (embeds images into markdown)
15_oobleck_effect_explained.md	This documentation
15_Oobleck_Effect_Summary.pdf	PDF summary with embedded images
oobleck_phase_breakdown.png/svg	020A phase dynamics (n=14 pros, n=8 def)
oobleck_trajectory_overlay.png/svg	020A trajectory visualization
oobleck_control_treatment.png/svg	020B control/treatment comparison (n=221)
oobleck_per_model_breakdown.png/svg	020B per-model analysis (37 ships, 100% attributed)
oobleck_thermometer.png/svg	Thermometer analogy visualization
oobleck_cross_platform.png/svg	Cross-platform summary

Data Sources

- `S7_run_020A_CURRENT.json`: Philosophical Tribunal results (29 sessions, partial phase markers)
- `S7_run_020B_CURRENT.json`: Control vs Treatment results (221 sessions, 100% model attribution, 37 ships)

Methodology Reference

- **IRON CLAD**: Cosine embedding methodology, Event Horizon = 0.80, p = 2.40e-23
- **B→F Drift**: Baseline-to-Final drift (preferred metric for 020B)
- **Phase Markers**: Prosecutor/Defense peaks extracted from conversation logs
- See `5_METHODOLOGY_DOMAINS.md` for complete methodology documentation

Generated: December 2025

Updated: December 29, 2025 (IRON CLAD data audit)

VALIS Network - Nyquist Consciousness Project