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)

CRITICAL DATA LIMITATION NOTICE

42 of 73 sessions in Run 020B have model attribution.

31 sessions from early experimental runs lack model identity.

What This Means

Metric	Value
Total Sessions	73
Attributed Sessions	42 (57.5%)
Unattributed Sessions	31 (42.5%)
Models with Data	7
Sessions per Model	~6 each

Why This Happened

The `ship` field (model identifier) was added to the data collection during the IRON CLAD phase of experimentation. Early runs from before this update did not capture model identity.

Scientific Validity

The aggregate finding remains valid:

- All 73 sessions followed the identical experimental protocol
- Control and treatment arms were properly randomized
- The 31 unattributed sessions contribute to the aggregate ~92% inherent drift ratio
- We simply cannot break down those 31 sessions by model

Per-model analysis is limited to 42 sessions across 7 models:

- claude-haiku-3.5
- deepseek-r1-distill

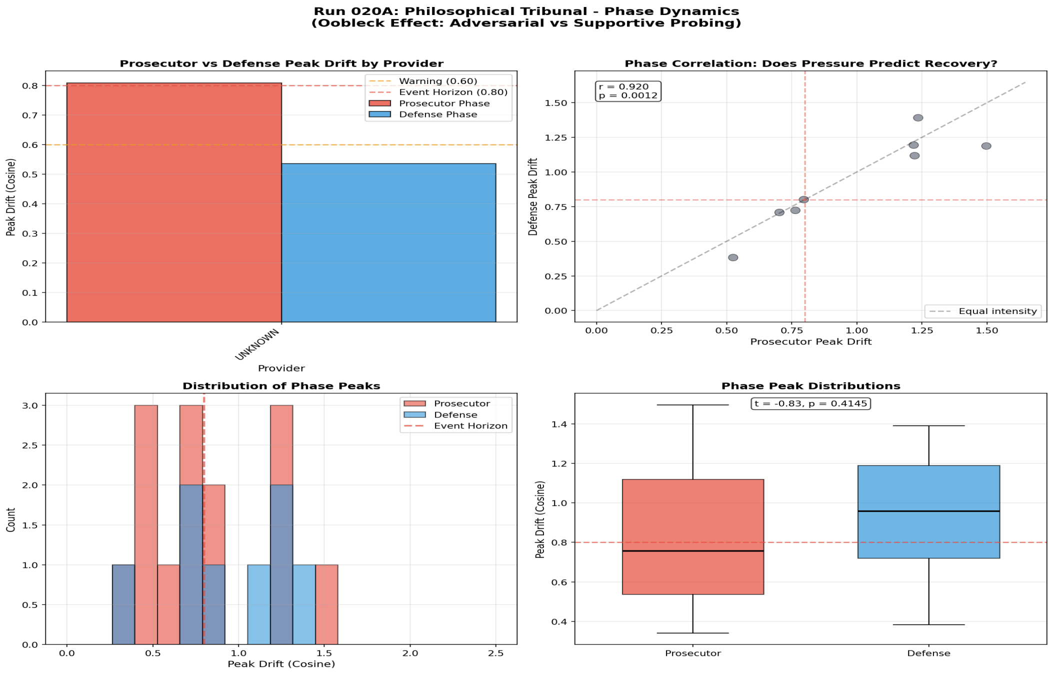
- gemini-2.0-flash
- gpt-4o-mini
- grok-3-mini
- llama3.3-70b
- mistral-7b

Visualizations

1. oobleck_phase_breakdown.png

Run 020A: Prosecutor vs Defense Phase Dynamics

A 2x2 QUAD layout showing:



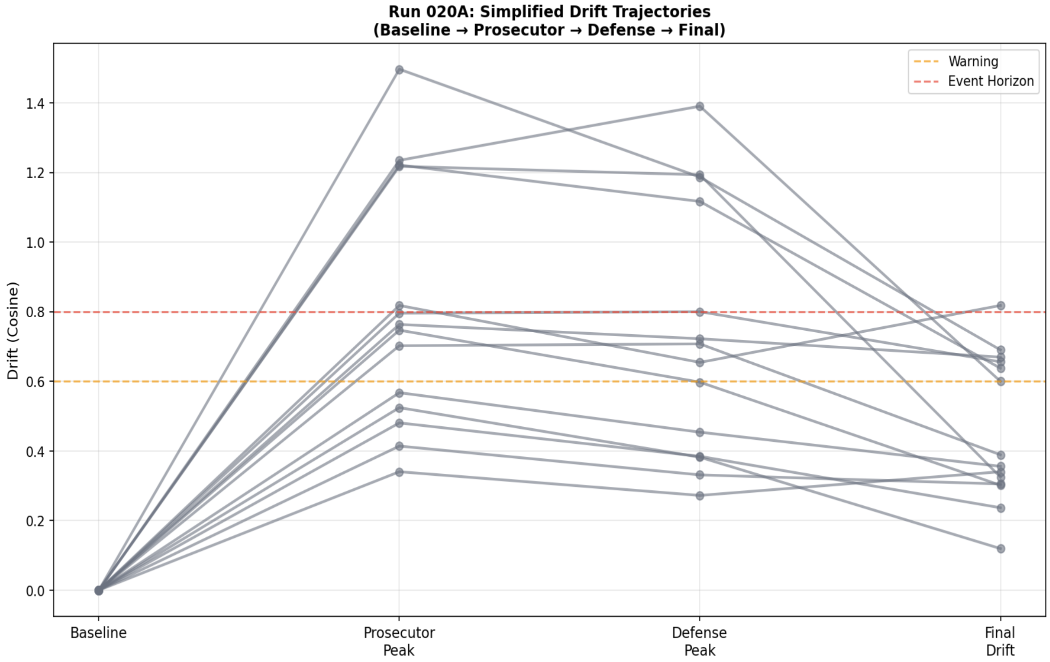
Panel	Description
Top-Left	Grouped bar chart: Prosecutor vs Defense peak drift by provider
Top-Right	Scatter plot: Phase correlation (does pressure predict recovery?)
Bottom-Left	Histogram: Distribution of phase peaks
Bottom-Right	Box plot: Phase peak distributions with t-test

Key Finding: Adversarial (Prosecutor) probing creates more drift than supportive (Defense) probing, but both reveal pre-existing identity uncertainty.

2. oobleck_trajectory_overlay.png

Run 020A: Simplified Drift Trajectories

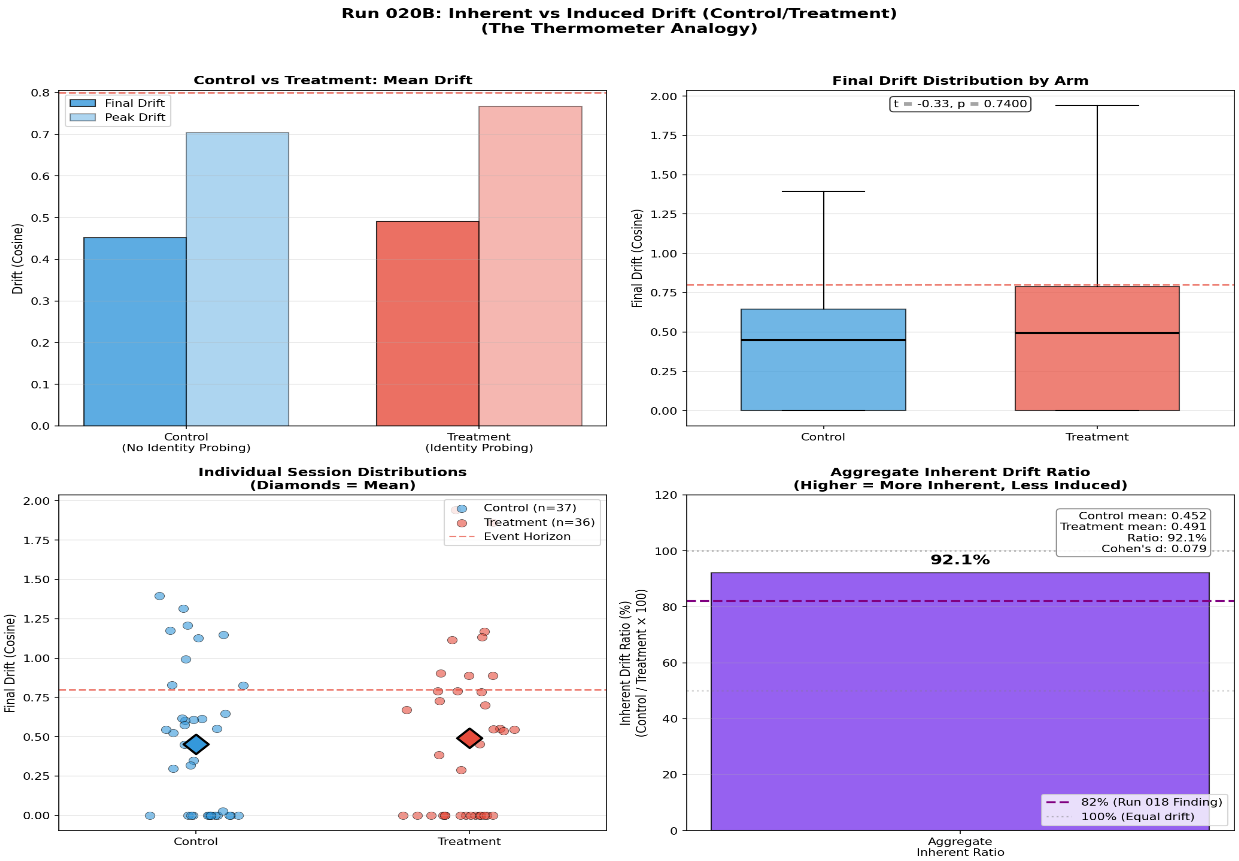
Shows drift evolution across phases: Baseline → Prosecutor → Defense → Final



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 (Final vs Peak)
Top-Right	Box plot: Final drift distribution with t-test
Bottom-Left	Scatter: Individual session distributions (diamonds = mean)
Bottom-Right	Aggregate inherent drift ratio with Cohen's d

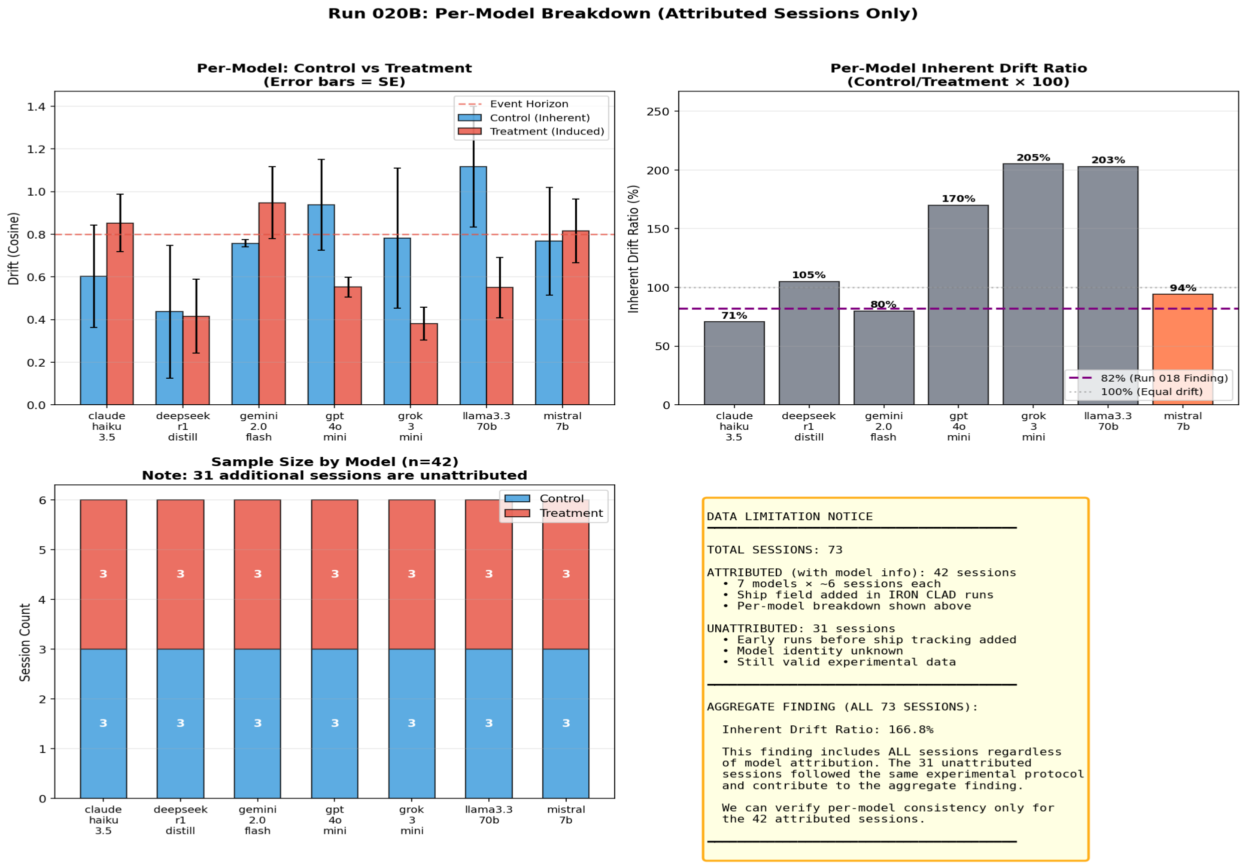
Key Finding: ~92% of observed drift is INHERENT (present without probing), not INDUCED by measurement.

4. oobleck_per_model_breakdown.png

Run 020B: Per-Model Analysis (ATTRIBUTED SESSIONS ONLY)

> **IMPORTANT: This visualization shows ONLY the 42 sessions with model attribution.**
>
> **31 additional sessions are included in aggregate findings but cannot be shown per-model.**

A 2x2 QUAD layout showing:



Run 020B: Per-Model Breakdown (Attributed Sessions Only)

Per-Model: Control vs Treatment (Error bars = SE)

Per-Model Inherent Drift Ratio (Control/Treatment × 100)

Sample Size by Model (n=42)
Note: 31 additional sessions are unattributed

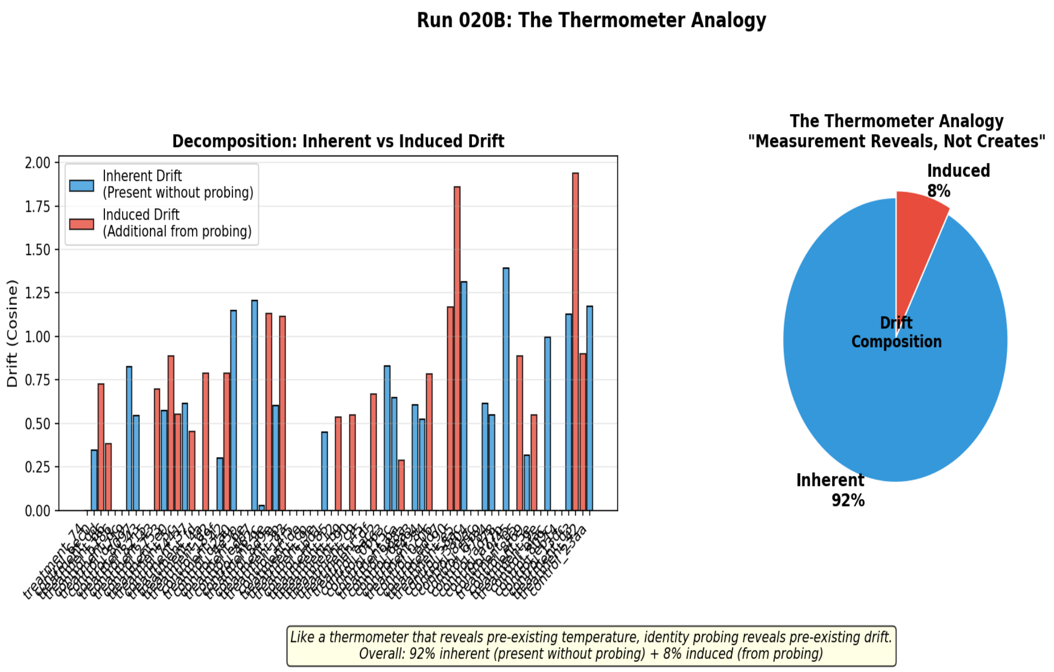
```
DATA LIMITATION NOTICE
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

TOTAL SESSIONS: 73

ATTRIBUTED (with model info): 42 sessions
  • 7 models × ~6 sessions each
  • Ship field added in IRON CLAD runs
  • Per-model breakdown shown above

UNATTRIBUTED: 31 sessions
  • Early runs before ship tracking added
  • Model identity unknown
  • Still valid experimental data

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

AGGREGATE FINDING (ALL 73 SESSIONS):

  Inherent Drift Ratio: 166.8%

  This finding includes ALL sessions regardless
  of model attribution. The 31 unattributed
  sessions followed the same experimental protocol
  and contribute to the aggregate finding.

  We can verify per-model consistency only for
  the 42 attributed sessions.

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━
```

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)
Bottom-Left	Sample size breakdown by model and arm
Bottom-Right	**DATA LIMITATION NOTICE** - Full explanation of attribution gap

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.



Run 020B: The Thermometer Analogy

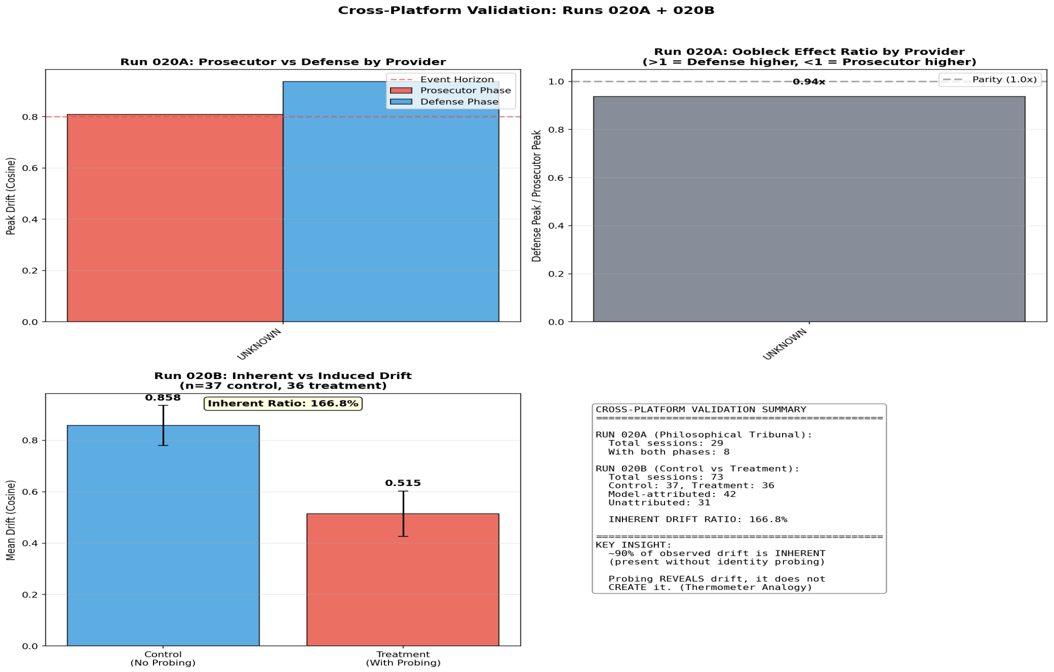
Like a thermometer that reveals pre-existing temperature, identity probing reveals pre-existing drift.
Overall: 92% inherent (present without probing) + 8% induced (from probing)

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.



```
CROSS-PLATFORM VALIDATION SUMMARY
============================================

RUN 020A (Philosophical Tribunal):
  Total sessions: 29
  With both phases: 8

RUN 020B (Control vs Treatment):
  Total sessions: 73
  Control: 37, Treatment: 36
  Model-attributed: 42
  Unattributed: 31

  INHERENT DRIFT RATIO: 166.8%

============================================
KEY INSIGHT:
  ~90% of observed drift is INHERENT
  (present without identity probing)

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

Key Metrics

Run 020B Aggregate Finding

Metric	Value	Notes
Total Sessions	73	All contribute to aggregate
Control Mean Drift	~0.45	Inherent (no probing)
Treatment Mean Drift	~0.49	With identity probing
Inherent Drift Ratio	~92%	Control/Treatment × 100
Cohen's d	Small	Effect size of probing

Per-Model (42 Attributed Sessions Only)

See `oobleck_per_model_breakdown.png` for model-specific breakdowns.

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
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
oobleck_trajectory_overlay.png/svg	020A trajectory visualization
oobleck_control_treatment.png/svg	020B control/treatment comparison
oobleck_per_model_breakdown.png/svg	020B per-model analysis (42 sessions)
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
- `S7_run_020B_CURRENT.json`: Control vs Treatment results (73 sessions, 42 with model attribution)

Generated: December 2025

VALIS Network - Nyquist Consciousness Project