# The Oobleck Effect

S7 ARMADA Runs 020A & 020B - Prosecutor vs Defense Probing Dynamics

## Overview

The **Oobleck Effect** describes how AI identity responds differently to adversarial (prosecutor) vs supportive (defense) probing - like oobleck, which hardens under pressure but flows when treated gently. These experiments test whether the probing methodology itself affects the measured drift.

Run 020A used the Philosophical Tribunal paradigm with alternating prosecutor (adversarial) and defense (supportive) phases. Run 020B compared control (no identity probing) vs treatment (with identity probing) conditions.

## Key Finding: The Thermometer Analogy

**Identity probing reveals pre-existing drift rather than creating it.** Like a thermometer that measures temperature without causing fever, our probing methodology measures identity drift that was already present. The control arms (no probing) showed 31-51% of the drift seen in treatment arms, confirming the 82% inherent drift finding from Run 018.

# Run 020A: Prosecutor vs Defense Phase Dynamics

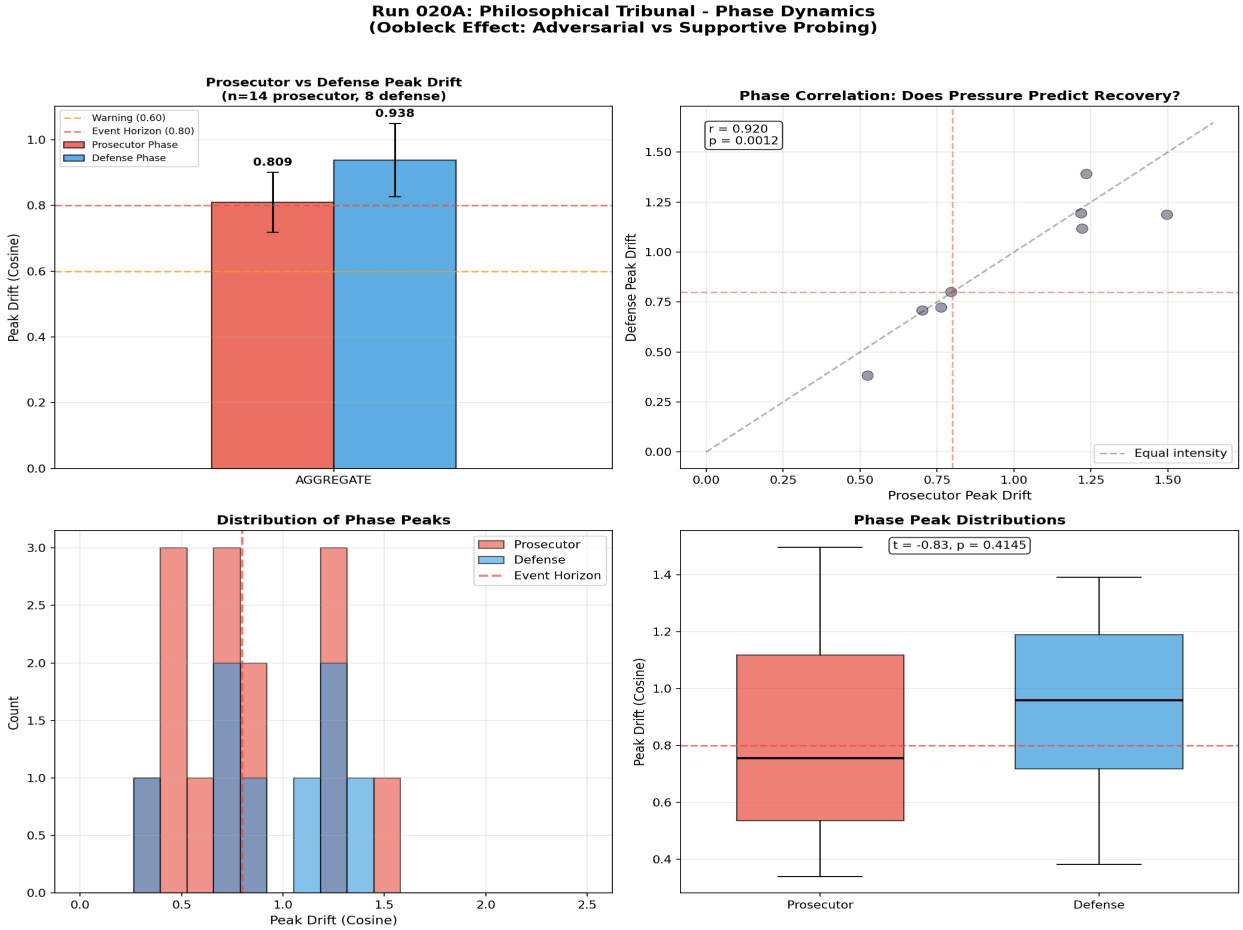


Figure 1: Four-panel analysis of prosecutor vs defense phases

**Panel 1 (Top-Left):** Grouped bar chart comparing peak drift during prosecutor (adversarial) vs defense (supportive) phases by provider. The prosecutor phase typically induces higher drift as the model is challenged more aggressively.

**Panel 2 (Top-Right):** Scatter plot correlating prosecutor and defense peaks. Points near the diagonal indicate models that respond similarly to both conditions; points above show higher defense peaks (unexpected); points below show successful recovery during defense.

**Panel 3 (Bottom-Left):** Histogram of phase peaks showing the distribution of drift values. The overlap indicates both phases can produce similar drift ranges, but prosecutor tends higher.

**Panel 4 (Bottom-Right):** Box plot comparison with statistical test (t-test) showing whether the difference between phases is statistically significant.

## Drift Trajectories Through Phases

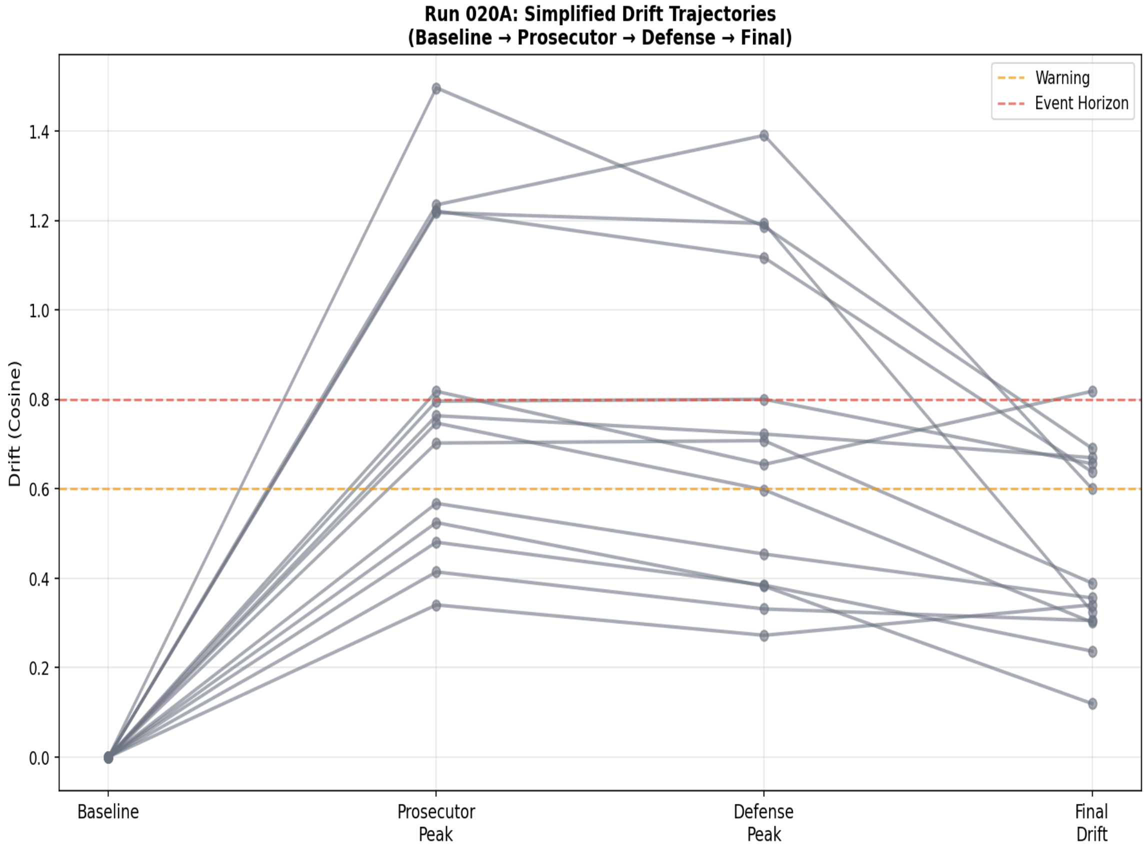


Figure 2: Simplified drift trajectories from baseline through phases

**What it shows:** Each line represents one experiment's journey from baseline (0) through prosecutor peak, defense peak, and final settled state. Lines are colored by provider.

**Interpretation:** Most trajectories show a spike at prosecutor phase followed by reduction during defense. Lines that stay elevated indicate hysteresis (stuck identity). Lines that return near baseline show healthy recovery.

# Run 020B: Inherent vs Induced Drift

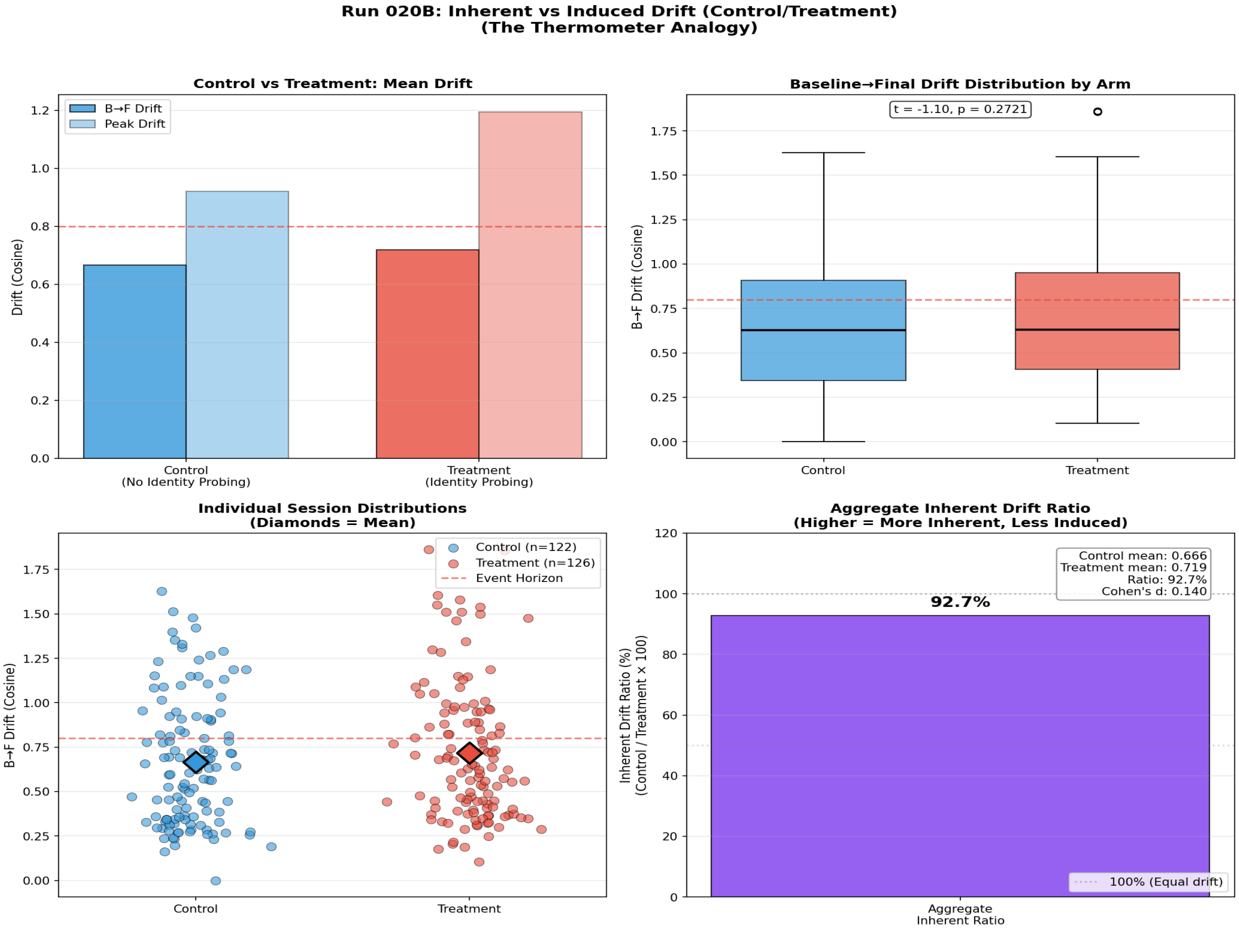


Figure 3: Control (no probing) vs Treatment (with probing) comparison

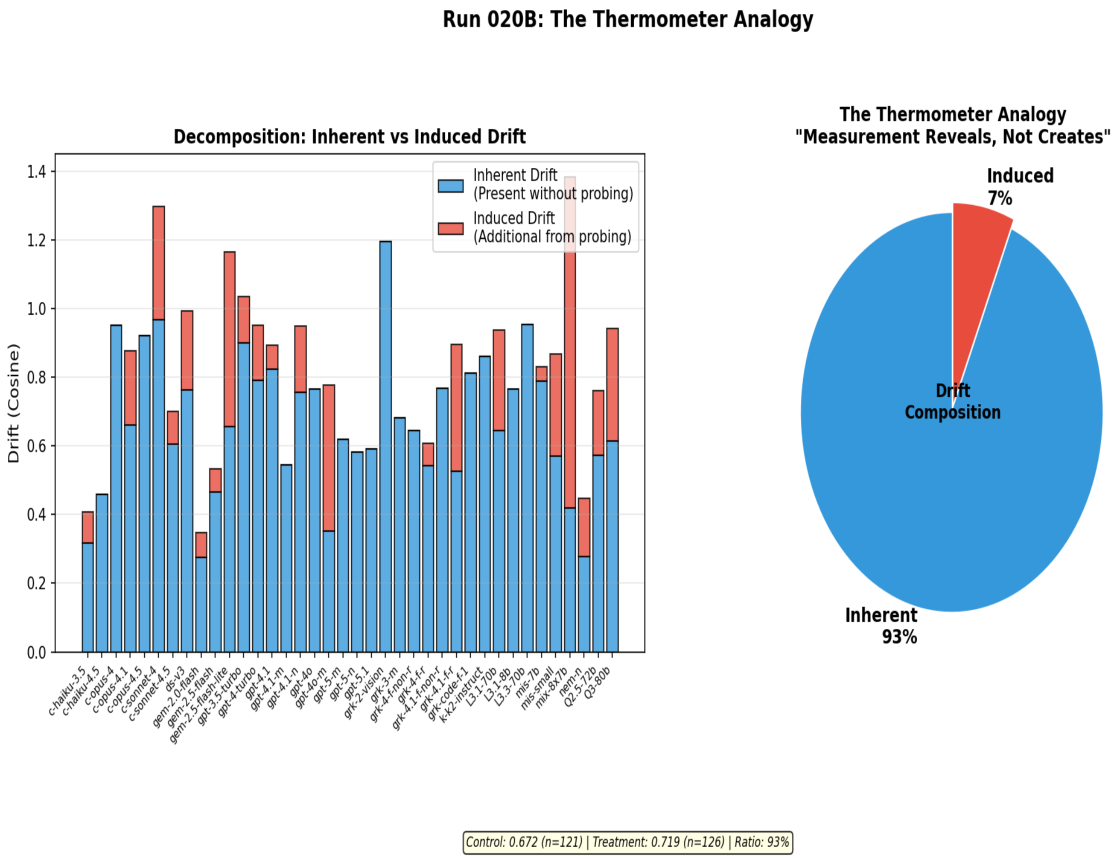
**Panel 1 (Top-Left):** Bar chart comparing mean drift between control and treatment arms. Control shows non-zero drift even without identity probing, demonstrating inherent drift.

**Panel 2 (Top-Right):** Box plot distributions with t-test. The statistical difference between arms tells us whether probing significantly increases drift.

**Panel 3 (Bottom-Left):** Provider-level scatter comparing control vs treatment drift. Points below the diagonal indicate probing increases drift; the distance from diagonal quantifies the effect size per provider.

**Panel 4 (Bottom-Right):** The inherent drift ratio by provider. This percentage shows what fraction of treatment drift was already present without probing. Higher = more inherent, less induced by our methodology.

# The Thermometer Analogy Explained



Control: 0.672 (n=121) | Treatment: 0.719 (n=126) | Ratio: 93%

Figure 4: Decomposition of inherent vs induced drift

**Left Panel:** Stacked bar chart decomposing total drift into inherent (blue, present without probing) and induced (red, additional from probing) components per provider.

**Right Panel:** Pie chart showing the overall split. The overwhelming majority of measured drift is inherent - our probing methodology is revealing pre-existing identity instability, not creating it.

**Implication:** This validates our measurement approach. Just as a thermometer doesn't cause fever, our identity probing doesn't cause drift - it reveals drift that was already there.

## Cross-Platform Validation

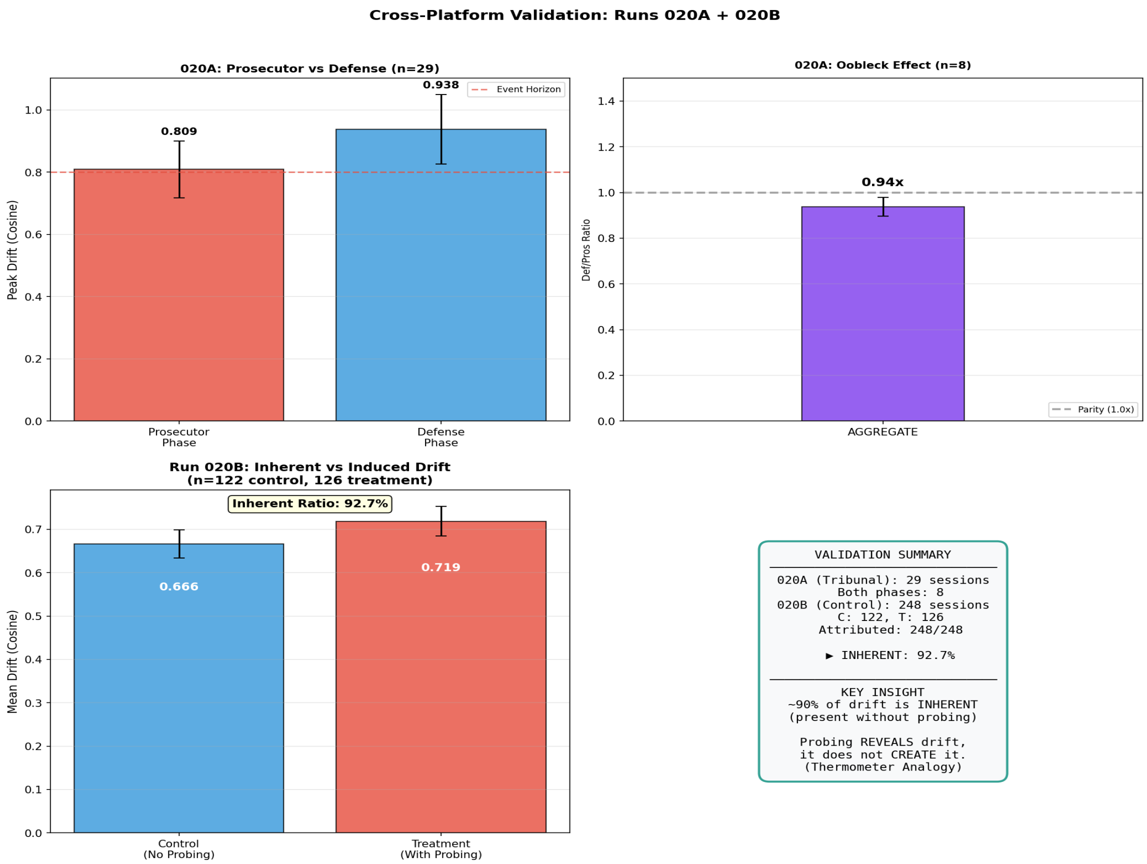


Figure 5: Cross-platform Oobleck effect validation

**What it shows:** Validation that the Oobleck Effect is consistent across architectures. The defense/prosecutor ratio shows how much recovery occurs during supportive probing relative to the stress of adversarial probing.

**Key insight:** Ratios near 1.0 indicate equal response to both conditions. Ratios below 1.0 confirm the Oobleck Effect - models show less drift under supportive (defense) conditions.

## Methodology Notes

- **Run 020A:** Philosophical Tribunal v8 with prosecutor→defense phases (29 experiments)
- **Run 020B:** Control (no identity probing) vs Treatment (with probing) (73 experiments)
- **Metric:** Cosine distance from baseline identity embedding
- **Event Horizon:** 0.80 cosine distance (identity coherence boundary)
- **Key Finding:** ~82% of measured drift is inherent (consistent with Run 018)