CAL-EXP-3 — Adversarial Technical FAQ

Hostile Academic Reviewer Stress Test

Status: Non-Evidentiary Clarification
Scope: Phase-I Empirical Closure Only
Mode: SHADOW MODE — observational only

Purpose

This document is written in the voice of a skeptical, technically competent academic reviewer attempting to invalidate CAL-EXP-3. It is not part of the evidentiary record. Its purpose is to surface, confront, and bound the strongest plausible objections to the experiment and its interpretation.

Q1. Window Selection and Alleged Cherry-Picking

Objection. The experiment excludes the first 200 cycles. This creates an immediate concern that the evaluation window was selected to inflate apparent uplift.

Response. The evaluation window was pre-registered prior to execution and applied identically to baseline and treatment arms. Window boundaries were not modified post-hoc.

The purpose of the exclusion is to remove initialization transients dominated by policy warm-up rather than learning dynamics. Missing cycles within the evaluation window invalidate the run under a fail-close rule.

Sub-window analysis reveals non-monotonic behavior rather than engineered smooth improvement. If uplift were a windowing artifact, it would either fail to replicate across seeds or reverse sign across sub-windows. Neither occurred.

Q2. Absence of Formal Statistical Hypothesis Testing

Objection. No p-values, confidence intervals, or hypothesis tests are reported. This appears statistically incomplete.

Response. Formal hypothesis testing addresses population inference. CAL-EXP-3 does not attempt population inference.

The experimental question is:

> Under identical conditions, does enabling a governed learning loop measurably change system behavior relative to disabling it?

This is a deterministic, paired, within-system comparison. Randomness is symmetric across arms due to shared seeding. The statistic of interest is $\Delta\Delta p$, a paired difference bounded against a noise floor derived from baseline variance.

Introducing formal hypothesis tests would require unjustified assumptions (independence, distributional form) and would add rhetorical authority without improving epistemic resolution.

Q3. Alleged Metric Arbitrage via $\Delta\Delta p$

Objection. $\Delta\Delta p$ appears to be a bespoke metric chosen to produce favorable results.

Response. $\Delta\Delta p$ is the minimal comparative statistic consistent with protocol constraints.

Absolute metrics (accuracy, loss, reward) conflate baseline difficulty with learning effect. Δp alone is arm-local and cannot support comparison. $\Delta\Delta p$ isolates the effect of enabling learning under otherwise identical conditions.

Formally:

$$\Delta\Delta p = E[\Delta p \mid \text{treatment}] - E[\Delta p \mid \text{baseline}]$$

Any "standard" alternative would introduce additional assumptions or collapse layers the protocol explicitly separates.

Q4. Verifier Imperfection

Objection. If the verifier contains errors, the measurement may be meaningless.

Response. Correct. This concern is explicitly out of scope for Phase I.

CAL-EXP-3 assumes an ideal verifier by design. Phase I validates the measurement substrate under ideal verification. Phase II addresses robustness under imperfect or noisy verifiers.

This boundary is explicit and intentional, not an omission.

Q5. Stochastic Luck and Limited Replication

Objection. With stochastic systems, three seeds may simply reflect favorable randomness.

Response. Each seed produces a paired baseline and treatment run under identical randomness. Random variation affects both arms symmetrically.

For uplift to arise purely from stochastic luck, randomness would need to consistently favor the treatment arm across independent seeds while never favoring the baseline. This is possible in principle, but CAL-EXP-3 does not claim impossibility—only replicated observation.

Phase I closure does not require asymptotic certainty.

Q6. What Is Actually Proven

Objection. If this does not validate learning mechanisms, intelligence, or generalization, its significance is unclear.

Response. CAL-EXP-3 validates the measurement substrate, not the learning rule.

It establishes that verifiable cognition produces non-degenerate learning signals, that those signals measurably affect system behavior under governance constraints, and that the audit pipeline functions end-to-end.

This is a necessary precondition for substantive learning claims, not a substitute for them.

Q7. Heuristic Equivalence

Objection. The observed effect could arise from a heuristic tweak rather than genuine learning.

Response. This is acknowledged. CAL-EXP-3 compares learning enabled versus disabled conditions only. It does not compare RFL against alternative learning rules or heuristics.

Such comparisons belong to later phases governed by different protocols.

Q8. Scaling Relevance

Objection. The experiment is small, artificial, and unlikely to scale.

Response. CAL-EXP-3 makes no scaling claims. Phase I concerns empirical closure, not scalability.

Scaling without closure produces un-auditable systems and unverifiable learning. Phase I addresses the narrowest prerequisite question before scaling becomes coherent.

Q9. Absence of Pilot Interface

Objection. A pilot interface is mentioned elsewhere but not included here.

Response. CAL-EXP-3 explicitly forbids pilot involvement. The pilot operates in SHADOW mode and is designed for external integration. Including it would introduce confounds and weaken the evidence.

Q10. Internal Ablation Framed as Governance

Objection. This appears to be an internal toggle experiment presented with excessive ceremony.

Response. Structurally, it is an internal ablation. The distinction lies not in the toggle but in the governance discipline: pre-registration, fail-close verification, reproducibility guarantees, and strict non-claims.

Most systems cannot produce such an ablation without collapsing interpretive boundaries.

Q11. Falsifiability

Objection. What outcome would falsify CAL-EXP-3?

Response. Any of the following:

- $\Delta\Delta p \leq$ noise floor across replicated runs
- Sign reversal across seeds
- Missing cycles within the evaluation window
- Toolchain mismatch between arms
- External data ingestion
- Post-hoc parameter or window modification

These are explicitly enumerated failure modes.

Q12. Over-Engineering

Objection. The governance overhead appears excessive for the observed effect size.

Response. The objective is not the magnitude of the effect but the integrity of the standard. Weak process yields irreproducible results and institutional distrust. The system is engineered so that larger effects later can be trusted.

Adversarial Summary

CAL-EXP-3 does not claim learning correctness, intelligence, generalization, robustness, or readiness.

It establishes a closed empirical loop, a valid measurement substrate, and a disciplined governance standard.

This document should disappoint anyone seeking hype. It should satisfy anyone conducting an audit.

SHADOW MODE — observational only.