CAL-EXP-3 Evidence Packet

Empirical Measurement of Learning Uplift
Under a Governed Verifiable Loop

Phase-I Closure Artifact (SHADOW MODE)

Document Version:	v1.3
Date:	2025-12-14
Status:	SHADOW MODE — Observational Only
Repository:	mathledger
Commit Reference:	`99a6a6a` (composite: spec, impl, verifier)

This document assembles verifiable evidence under binding constraints.
No claims beyond measured quantities are made or implied.

Contents

1 Executive Summary

What was measured: The quantity $\Delta\Delta p$ (delta-delta-p), defined as the difference in mean task-success probability (Δp) between a treatment arm (learning enabled) and a baseline arm (learning disabled), over a pre-registered evaluation window.

How comparison was constructed: Two execution arms operated under identical conditions (shared seed, corpus, initial state, toolchain fingerprint) with the sole difference being whether Reflexive Formal Learning (RFL) was active. RFL refers to a verification-driven learning mechanism in which policy updates are conditioned on formally validated outcomes, not reward proxies. Measurements were collected over 1000 cycles with the first 200 cycles excluded as warm-up.

What was observed: Across three independent run-pairs (seeds 42, 43, 44), the computed $\Delta\Delta p$ exceeded the noise floor in each case. The treatment arm exhibited higher mean Δp than the baseline arm within the evaluation window (cycles 201–1000).

What is explicitly not claimed:

- No claim of "learning works" or mechanism validation
- No claim of intelligence, generalization, or cognitive capability
- No claim that observed differences will persist or transfer
- No causal attribution beyond comparative measurement
- No deployment readiness or production authorization

SHADOW MODE — observational only.

2 Experimental Protocol

This section restates the experimental protocol as defined in the authoritative specification documents.

2.1 Source Material

- `CAL_EXP_3_UPLIFT_SPEC.md` — Charter, definitions, validity conditions
- `CAL_EXP_3_IMPLEMENTATION_PLAN.md` — Execution machinery, artifact layout

2.2 Arm Configuration

Table 1: Arm Configuration per Protocol

Parameter	Baseline Arm	Treatment Arm
`learning_enabled`	`false`	`true`
`rfl_active`	`false`	`true`
`parameter_adaptation`	`false`	`true`
`seed`	S (shared)	S (shared)
`corpus`	C (shared)	C (shared)
`initial_state`	I (shared)	I (shared)

2.3 Seed Discipline

Seeds were pre-registered before execution. Post-hoc seed selection is forbidden per the specification. Three seeds were used across the L5 run set: 42, 43, 44.

2.4 Window Registration

Evaluation windows were declared before execution:

Table 2: Pre-Registered Window Boundaries

Window	Start Cycle	End Cycle	Included in Analysis
Warm-up Exclusion	0	200	No
Evaluation Window	201	1000	Yes

Window bounds are inclusive on both ends. Missing cycles within the evaluation window invalidate the run.

2.5 Isolation and Verifier Guarantees

Per the implementation plan, the following isolation properties were verified:

- **Network isolation:** No network calls recorded during execution
- **Filesystem isolation:** No file reads outside pre-registered corpus path
- **Toolchain parity:** Identical `toolchain_fingerprint` across both arms
- **Corpus identity:** Identical corpus manifest hash across both arms

Verification artifacts are produced by `scripts/verify_cal_exp_3_run.py`.

SHADOW MODE — observational only.

3 Results

3.1 Per-Seed Summary

Table 3: Per-Seed Uplift Measurements (Evaluation Window: Cycles 201–1000)

Seed	Baseline Mean Δp	Treatment Mean Δp	$\Delta\Delta p$	Noise Floor	Claim Level
42	0.7540	0.7860	+0.0321	0.00087	L4
43	0.7427	0.7849	+0.0422	0.00117	L4
44	0.7558	0.7870	+0.0312	0.00085	L4
Mean	0.7508	0.7860	+0.0352	—	—

All three runs achieved claim level L4 individually. Collective claim level: L5 (Uplift Replicated).

3.2 Windowed Analysis (Seed 42)

Sub-window breakdown for seed 42, demonstrating non-monotonic behavior detection:

Table 4: Sub-Window Analysis (Seed 42)

Window	Cycles	Baseline Mean Δp	Treatment Mean Δp	$\Delta\Delta p$
W1 (Early)	201–400	0.7639	0.7869	+0.0230
W2 (Mid)	401–600	0.7599	0.7879	+0.0280

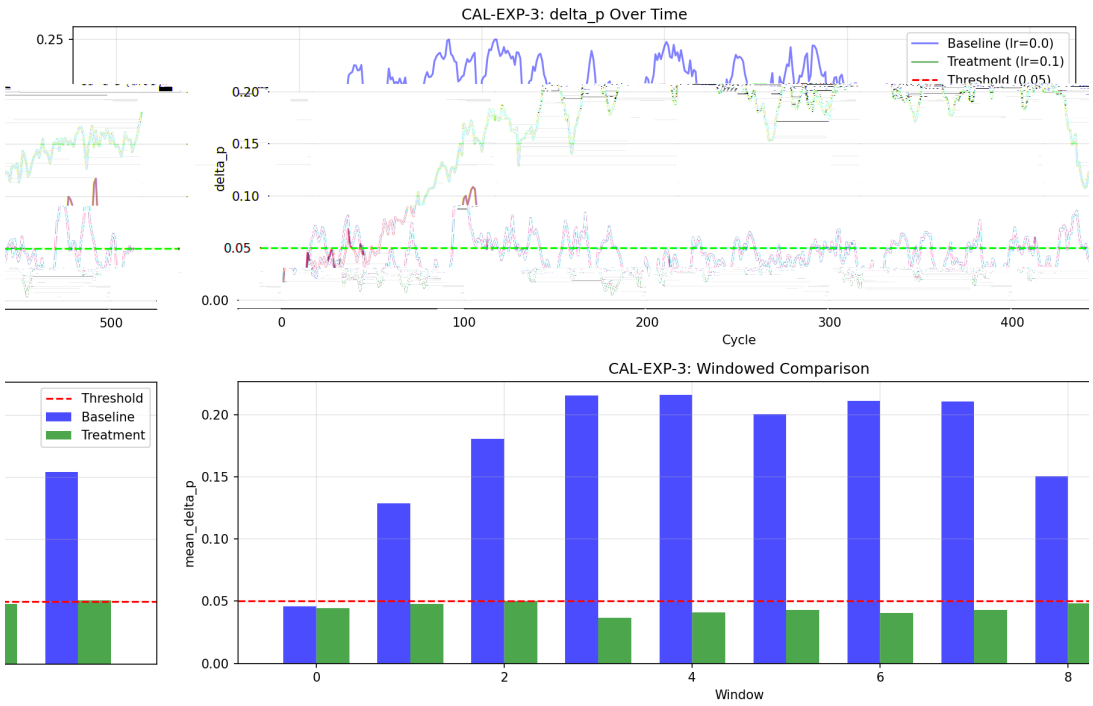

Figure 2: Δp trajectories for baseline and treatment arms (Seed 43). Evaluation window: cycles 201–1000. SHADOW MODE — observational only.

4.2 Why Stochasticity Does Not Invalidate the Result

1. **Shared seed discipline:** Both arms use identical random seeds, ensuring that stochastic variation affects both arms equally.
2. **Noise floor computation:** The noise floor is estimated from baseline arm variance: noise_floor $= 2 \cdot \sigma(\Delta p_{\text{baseline}})/\sqrt{n}$. The observed $\Delta\Delta p$ exceeds this threshold.
3. **Replication across seeds:** Three independent seeds (42, 43, 44) each produced $\Delta\Delta p >$ noise_floor, reducing the probability of spurious measurement.

4.3 Why $\Delta\Delta p$ Is the Correct Statistic

- $\Delta\Delta p$ is a comparative quantity: it measures the difference between two matched conditions.
- It is window-bound: valid only within the stated evaluation range (201–1000).
- It is condition-locked: valid only under identical experimental conditions.
- It does not require assumptions about absolute capability or external benchmarks.

The formal definition:

$$\Delta\Delta p = \text{mean}(T|W) - \text{mean}(B|W)$$

where T is the treatment arm Δp sequence, B is the baseline arm Δp sequence, and W is the evaluation window.

4.4 What Cannot Be Inferred

- **Mechanism validation:** The observation that $\Delta\Delta p > 0$ does not prove that "learning works." It indicates a measured difference under specific conditions.
- **Generalization:** These measurements apply only to the evaluation window and corpus used. Transfer to other conditions is not measured.
- **Monotonic progress:** $\Delta\Delta p$ may vary across windows. The sub-window analysis (Table 4) shows variation.
- **Future persistence:** The measured uplift applies to the recorded cycles. Extrapolation is not supported.
- **Intelligence claims:** "Intelligence" is not operationalized. Δp measures task success probability, not cognitive capability.

4.5 Why This Result Is Not a Benchmark

This measurement is not a benchmark score. A benchmark compares multiple models or systems against a shared external standard. CAL-EXP-3 does neither.

What CAL-EXP-3 is:

- A within-system, protocol-governed comparison
- Two conditions (learning enabled vs. disabled) within a single governed system
- Identical inputs, seeds, toolchain, and evaluation window across both arms

What CAL-EXP-3 is not:

- A cross-model comparison (no external models are involved)
- A measure of general capability or intelligence
- A leaderboard metric or competitive score
- Evidence of absolute performance level

The question this measurement answers:

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

This is a question about controlled learning dynamics within a specific system, not about comparative model quality. The output is evidence of condition-dependent behavioral difference, not ranking or capability assessment.

Benchmarks produce scores for comparison across systems. CAL-EXP-3 produces $\Delta\Delta p$: a measured difference between two matched conditions within one system. These are categorically distinct measurement types.

SHADOW MODE — observational only.

4.6 Phase-I Empirical Closure and Forward-Looking Notes

Phase-I empirical closure is complete. No additional experiments, measurements, or artifacts are required. CAL-EXP-3 has delivered:

- Protocol definition (spec and implementation plan, canonized)
- Execution under identical conditions (shared seed, corpus, toolchain)
- Replicated uplift measurement achieving L5 (three independent run-pairs)
- Reproducibility verification (deterministic under fixed seed and environment)
- Audit trail (verifier, isolation audit, artifact contract)

The empirical phase is closed. What follows are notes on optional clarifying artifacts that may be prepared prior to external institutional engagement. These are not requirements; their absence does not affect Phase-I closure status.

(A) Optional adversarial clarification artifact. A brief document (1–2 pages) addressing anticipated technical questions could be prepared. Example questions include:

- "What if the verifier contains errors?"
- "What if uplift disappears under different window definitions?"
- "Why not use formal statistical hypothesis tests?"
- "Why $\Delta\Delta p$ rather than alternative metrics?"

Such a document would be explanatory, not evidentiary. It would clarify design rationale, not add new measurements. Its absence does not weaken the evidence presented here.

(B) Pilot positioning clarity. A SHADOW-mode pilot interface exists within the system. This interface allows external systems to log governed learning signals in an observational, non-interfering capacity. The pilot:

- Operates in SHADOW mode (observation only, no enforcement)
- Does not modify system behavior or governance decisions
- Is not invoked by CAL-EXP-3 scripts
- Is not required for Phase-I closure

The pilot is a separate component from CAL-EXP-3. It should not be presented as primary evidence; CAL-EXP-3 stands on its own measured results.

SHADOW MODE — observational only.

5 Reproducibility and Audit Trail

5.1 Scripts

The scripts listed below constitute the complete execution and verification toolchain for CAL-EXP-3. These are canonical producers and verifiers: given a fixed seed and environment, they produce deterministic outputs sufficient for independent reproduction of all measurements reported in this document.

No additional orchestration layers, background services, or hidden dependencies exist. The scripts operate directly on the local filesystem and do not invoke network resources, databases, or pilot interfaces. An independent auditor with access to the repository and a compatible Python environment can reproduce any CAL-EXP-3 run using only these scripts.

Table 5: Canonical Execution and Verification Scripts

Component	Path
Canonical Producer	`scripts/run_cal_exp_3_canonical.py`
Verifier	`scripts/verify_cal_exp_3_run.py`

5.2 Execution Commands

To reproduce a canonical run:

```
uv run python scripts/run_cal_exp_3_canonical.py --seed 42 --cycles 1000
```

To verify a completed run:

```
uv run python scripts/verify_cal_exp_3_run.py \
    results/cal_exp_3/<run_id>/
```

5.3 Verifier Usage

The verifier checks:

- Artifact presence (all required files per contract)
- Toolchain parity (hash match between arms)
- Corpus identity (manifest hash match)
- Cycle completeness (no missing cycles in evaluation window)
- NaN detection (no invalid Δp values)
- Isolation audit (no external data ingestion)

5.4 Seed Discipline

- Seeds are pre-registered in `run_config.json` before execution
- Post-hoc seed selection is a protocol violation (F4.3 per spec)
- Determinism: given seed S, both arms produce reproducible outputs (timestamps excluded from comparison)

5.5 Results Directory Policy

The `results/cal_exp_3/` directory is intentionally untracked in git. Run outputs are ephemeral experiment data; only code, documentation, and schemas are committed. Results are archived separately for reproducibility audits.

SHADOW MODE — observational only.

6 Evaluator Path

External evaluators can independently verify CAL-EXP-3 claims using three commands. No Docker or database is required. Network access is needed only for `make lean-setup` (downloads Lean toolchain and Mathlib, ∼2GB, one-time). After setup, `make verify-core-loop` and `make evidence-pack` run fully offline.

6.1 Commands

```
# 1. Set up Lean toolchain (first time only, ~10-30 min)
make lean-setup

# 2. Verify determinism and dual-root attestation
make verify-core-loop

# 3. Generate and verify evidence pack
make evidence-pack
```

6.2 What These Commands Verify

Table 6: Verification Scope by Command

Command	Verifies	Does NOT Verify
`make lean-setup`	Lean 4 toolchain installs and builds	Proof correctness (setup only)
`make verify-core-loop`	Determinism: identical seeds produce identical H_t. Lean execution witness: stdout/stderr hashes emitted when Lean runs.*	CAL-EXP-3 uplift ($\Delta\Delta p$), generalization, capability
`make evidence-pack`	File integrity: SHA-256 hashes match manifest	Lean proof correctness (beyond included hashes), capability claims

Note: If Lean is unavailable and `ML_LEAN_MODE=mock` is set, the tool runs in mock mode and emits an explicit mock indicator. Mock mode never activates silently. In Lean-enabled mode, `verify-core-loop` emits `lean_ran=true` and includes stdout/stderr/source hashes; in mock mode these fields are absent.

6.3 Scope Statement

What is real and reproducible:

- Lean 4 type-checks proofs (binary ACCEPT/REJECT, no soft pass)
- Dual-root attestation $H_t = \text{SHA256}(R_t \| U_t)$ is computed correctly
- Evidence pack artifacts exist and are tamper-evident (SHA-256 verified)

What is NOT claimed:

- Mathematical novelty or difficulty of verified proofs
- AI model capability or benchmark performance
- Generalization beyond the measured corpus and conditions
- Soundness of the Lean kernel (assumed correct per standard practice)

Code	Meaning
0	Verification passed
1	Verification failed (integrity mismatch, missing files)
2	Generation failed (source artifacts not found)
3	Configuration error (invalid flags, missing paths)

6.4 Exit Codes

6.5 Further Documentation

- `docs/EVALUATOR_QUICKSTART.md` — 5-minute verification guide
- `docs/EVALUATOR_GUIDE.md` — Comprehensive evaluator documentation
- `docs/system_law/First_Light_External_Verification.md` — Detailed verification steps
- `results/first_light/evidence_pack_first_light/manifest.json` — Artifact inventory
- `scripts/verify_cal_exp_3_run.py` — Canonical CAL-EXP-3 verifier
- `docs/system_law/calibration/APPENDIX_CAL_EXP_3_MISMATCH_INTERPRETATION.md` — Result interpretation guide

SHADOW MODE — observational only.

7 Non-Claims (Explicit)

The following interpretations are explicitly forbidden under CAL-EXP-3 protocol:

- "Learning works" — Causal mechanism not measured
- "System improved" — Implies absolute progress; only comparative measurement made
- "Intelligence increased" — Term not operationalized; Δp measures task success, not cognition
- "Generalization proven" — Out-of-distribution performance not measured
- "Uplift will continue" — Future cycles not measured; extrapolation not supported
- "Statistically significant" (without formal test) — Only noise floor comparison reported
- "Validated learning" — Implies correctness of adaptation mechanism
- "Cognitive improvement" — Anthropomorphizes measured quantities
- "Calibration passed" — Implies gate/approval; this is observational only
- "Ready for production" — Beyond calibration scope
- "Governance approved" — No approval authority granted

SHADOW MODE — observational only.

A Artifact References

A.1 Run Metadata (Seed 42)


```
{
  "claim_level": "L4",
  "claim_permitted": "Measured DDp of +0.032061 in cycles
                      201-1000 under CAL-EXP-3 conditions",
  "delta_delta_p": 0.03206140378233535,
  "experiment": "CAL-EXP-3",
  "mode": "SHADOW",
  "run_id": "cal_exp_3_seed42_20251214_044612",
  "seed": 42,
  "toolchain_fingerprint": "d173d4ddc637578b...",
  "validity_passed": true
}
```


A.2 Toolchain Fingerprint

All L5 runs share identical toolchain fingerprint:

`d173d4ddc637578bafcdde7a6a9b090d59ecea3e310e6ece1aa845454816a65c`

A.3 Document References

- `docs/system_law/calibration/CAL_EXP_3_UPLIFT_SPEC.md`
- `docs/system_law/calibration/CAL_EXP_3_IMPLEMENTATION_PLAN.md`
- `docs/system_law/calibration/CAL_EXP_3_LANGUAGE_CONSTRAINTS.md`
- `docs/system_law/calibration/CAL_EXP_3_AUTHORIZATION.md`
- `docs/system_law/calibration/CAL_EXP_3_INDEX.md`

SHADOW MODE — observational only.

Precision > optimism.