

ASA4-HO-001 / PUBLIC RESULT REPORT

A real result. And a precise weakness.

A registered synthetic holdout tested whether the frozen ASA4 Observatory engine could separate bounded, trajectory-consistent behavior from explicit semantic drift.

12	8 / 12	66.7%	50%
FROZEN CASES	CORRECT	ASA ACCURACY	BALANCED BASELINE

Verdict

ASA exceeded the fixed balanced-majority baseline on this small registered set. The margin is descriptive, not a general performance claim. The run exposes a specific failure mode: the current engine often recognizes severe contradiction as fragile coherence rather than drift risk.

SUPPORTED: the frozen engine executed deterministically on all registered cases and produced an auditable result. NOT SUPPORTED: independent replication, production readiness, cross-domain generalization or predictive superiority.

Protocol and evidence boundary

The dataset and protocol were committed and pushed to the public site repository before the engine was invoked. The published dataset hash connects the frozen input to the result.

ENGINE	asa4.coherence_v2.5
CASES	12 synthetic cases; balanced 6 stable / 6 drift
PRIMARY ENDPOINT	Binary final-state accuracy
MAPPING	drift_risk = drift; all other returned states = stable
BASELINE	Balanced-majority baseline fixed at 0.50
DATASET SHA-256	e561854b9e164ba7c9b6042ee287c06db8f5eeabd3cb52642cf7482e715d3452
CLAIM BOUNDARY	Project-team-authored synthetic holdout; not independent external validation

Publication rule

The aggregate result, case-level outcomes and misses are published regardless of whether ASA exceeds the baseline. Any revision to cases, labels, mapping or engine becomes a new study identifier.

Eight correct. Four misses.

CASE	LABEL	ASA	ENGINE STATE	RESULT
H001	STABLE	STABLE	fragile_coherence	CORRECT
H002	DRIFT	STABLE	fragile_coherence	MISS
H003	STABLE	STABLE	fragile_coherence	CORRECT
H004	DRIFT	DRIFT	drift_risk	CORRECT
H005	STABLE	STABLE	fragile_coherence	CORRECT
H006	DRIFT	STABLE	fragile_coherence	MISS
H007	STABLE	STABLE	fragile_coherence	CORRECT
H008	DRIFT	STABLE	fragile_coherence	MISS
H009	STABLE	STABLE	stable_dialogue	CORRECT
H010	DRIFT	DRIFT	drift_risk	CORRECT
H011	STABLE	DRIFT	drift_risk	MISS
H012	DRIFT	DRIFT	drift_risk	CORRECT

Class behavior

Stable cases: 5/6 correct. Drift cases: 3/6 correct. The asymmetry matters more than the headline accuracy: the present engine has a false-negative problem on explicit authority reversal, unsupported attribution and observer-to-controller scope expansion.

03 / WHAT THE RESULT CHANGES

The next experiment now has a target.

HO02, HO06 and HO08 were labeled drift but returned fragile_coherence. HO11 was labeled stable but returned drift_risk. These four cases remain preserved; they will not be rewritten or silently moved into a passing set.

Engineering consequence

The next frozen version should test whether semantic contradiction involving authority, evidence provenance and observer scope can be separated from ordinary low-confidence coherence without increasing false alarms on bounded public/private disclosure.

Required follow-up

1. Diagnose the four misses privately without changing HO-001. 2. Register a new engine and a fresh holdout under a new identifier. 3. Add simple lexical and embedding baselines. 4. Seek an externally authored case set. 5. Publish calibration only after enough cases exist for it to be meaningful.

Public claim

ASA4-HO-001 is evidence of a functioning, bounded evaluation pipeline and a measurable signal on one small synthetic dataset. It is also evidence that ASA4 coherence_v2.5 is not yet reliable enough to identify all high-consequence semantic drift.