Aegis-R Baseline Validation Report

Date: 2026-02-07

Scope
This report captures the clean-state baseline runs requested:
- Synthetic full-spectrum run (1000 events)
- Synthetic multi-entity thread suite (`data/synthetic_threads.json`)
- Vendor fixture suite (all adapters combined)
- Realistic scenario evaluation (labels + expectations)

Clean state used: `/tmp/clean_state.json` (no cached decisions).

---

1) Synthetic Full-Spectrum (1000 events)
Summary
- Results: 30
- Feasible: 28
- Incomplete: 0
- Impossible: 2
- Decision labels: `escalate: 28`, `keep: 2`
- Reason codes: `environment_unknown: 28`, `admin_hold: 2`
- Thread reasons: `ambiguous_context: 28`
- Threads: 0
- Tickets: 0
- Avg thread confidence: 0.00

Interpretation
- The synthetic generator mixes multiple entities, so the system intentionally refuses to cluster.
- This run is useful for throughput/coverage, not accuracy.

---

2) Synthetic Multi-Entity Thread Suite
Input: `data/synthetic_threads.json`

Summary
- Results: 30
- Feasible: 3
- Incomplete: 24
- Impossible: 3
- Decision labels: `deprioritize: 25`, `escalate: 3`, `keep: 2`
- Reason codes: `evidence_gap: 20`, `precond_missing: 5`, empty: 3, `admin_hold: 2`
- Thread reasons: `rule_evidence: 6`, `ambiguous_context: 22`
- Threads: 3
- Tickets: 3
- Avg thread confidence: 0.70

Interpretation
- Threading works when evidence aligns on host/principal.
- Confidence of 0.70 indicates clustering based on rule evidence.

---

3) Vendor Fixture Suite (All Adapters)
Summary

- Results: 30
- Feasible: 3
- Incomplete: 24
- Impossible: 3
- Decision labels: `deprioritize: 25`, `escalate: 3`, `keep: 2`
- Reason codes: `evidence_gap: 17`, `precond_missing: 8`, `policy_override: 2`, empty: 1, `admin_hold: 2`
- Thread reasons: `rule_evidence: 5`, `ambiguous_context: 23`
- Threads: 3
- Tickets: 6
- Avg thread confidence: 0.70

Interpretation
- Fixtures now produce threads, which validates adapter normalization.
- Most rules remain incomplete due to sparse evidence, which is expected for small fixture samples.

---

4) Realistic Scenario Evaluation
Output
- Total labels: 10
- Accuracy: 1.00
- Mismatches: 0

Interpretation
- The adjudicated scenario suite aligns with expected outcomes.
- Decision labels and ticket statuses are validated by tests.

Overall Notes
- Thread confidence and reasons now explain when clustering is or isn’t applied.
- Clean state runs eliminate cache effects; this is the current baseline behavior.
- Next improvement area: expand realistic scenario coverage and add more multi-entity vendor samples.