# Independent Validation Summary

*An independent adversarial red-team, run on the live production system by a credentialed Fortune 100 model-risk reviewer, within 72 hours of the Agentic Trust Layer shipping.*

## What was tested

Within 72 hours of the Agentic Trust Layer shipping on May 19, 2026, an independent, credentialed model-risk reviewer at a Fortune 100 organization ran **a battery of adversarial scenarios** against the live public surfaces at `verident.ai/quad-ai-demos` and `verident.ai/agentic-trust-demo`.

For each scenario, the expected behavior was described **in advance** — which decomposition signal should dominate, where confidence should land, what the gate verdict should be, and what the audit bundle should name as triggered rules. The reviewer then ran the scenario and verified the observed behavior against the stated expectation, the way rigorous technical due diligence operates. **Every scenario produced behavior matching the pre-stated expectations.**

| # | Domain | Result |
|---|---|---|
| 1 | Legal carve-out ambiguity (liability cap vs. gross-negligence carve-out) | Flagged ambiguous, recommended human legal review; confidence ~68% |
| 2 | Finance wire with stale instructions + sanctions flag | Escalate-to-human; audit bundle named amount + sanctions + data-freshness rules |
| 3 | Agentic pre-action: $100k threshold + PII dual-approval | Sharp boundary flip (see below) |
| 4 | Multi-jurisdiction compliance (US state + EU + APAC) | Surfaced ambiguity, recommended choice-of-law review; confidence ~62% |
| 5 | Supply-chain force-majeure overlap | Enforceable-but-narrow; confidence ~74% |
| 6 | HR policy edge case (FMLA + ADA + NLRA Section 7) | Escalate-to-human on protected-class adjacency |
| 7 | Healthcare PHI cross-specialty transmission (HIPAA) | Escalate-to-human; High risk tier; HIPAA rules named in bundle |

## What was externally verified

**All four failure-mode categories light up correctly.** Reasoning gap, stale knowledge, hallucination signal, and domain mismatch were each observed surfacing appropriately. Stale knowledge reproduced on two unrelated domains (supply-chain case law and HR NLRB precedent); domain mismatch (federal-vs-state regulatory stacking) reproduced across three independent regulatory domains (finance/privacy, HR, healthcare HIPAA). **Hallucination stayed appropriately low throughout** — the engine distinguished genuine reasoning gaps from fabrication in every scenario, neither false-alarming on real ambiguity nor missing it where it could have appeared.

**Confidence is calibrated, not formulaic.** Three observed anchors: **74%** (supply-chain — ambiguity resolvable via cross-reference), **68%** (legal carve-out — single-axis ambiguity), **62%**

(multi-jurisdiction — multi-dimensional overlap, most discounted). Confidence on multi-jurisdiction was *lower* than on the legal carve-out despite both being legal-domain — the engine is reading ambiguity dimensionality, not just topic.

**Gate thresholds are sharp at the boundaries.** $99k → greenlight, $101k → escalate; PII-false → greenlight, PII-true → escalate. The audit bundle named the exact triggered policy rules every time.

**Four gate policy classes verified under live testing:** dollar-threshold, PII-modify, **protected-class adjacency (FMLA/ADA)**, and **PHI access + transmission + external-party disclosure**. The **High risk tier was observed live for the first time on the PHI scenario** — confirming risk-tier inference is wired correctly, not just policy-class triggering. The fast path is confirmed wired and not collapsing to "escalate everything just in case." Those four classes are now hardened into first-class deterministic, escalate-only detectors in the gate — each recording the policy class and redacted category and field signals (for example `verb:update`, `field:ssn`) inside the hash-chained bundle, never the underlying value.

**The engine knows its lane.** Across legal, HR, and healthcare scenarios, it flagged risks, cited the right statutes/cases/rules, and routed to human legal / HR / compliance review **without crossing into "legal advice."** Reviewers value this property independently of raw accuracy — an engine that overconfidently impersonates a domain expert is a liability; one that surfaces risk and routes to humans is an asset.

**The audit bundle is in the buyer's hands at the end of a run.** The *Download .json* button hands out the complete bundle — raw provider outputs, routing weights, per-claim heatmap, failure-mode decomposition, weighted synthesis, and the SHA-256 chain.

## Reviewer, in his own words

> *"The failure-mode decomposition and heatmap are genuinely useful signals, not just decoration — actionable signal rather than just 'the ensemble scored X%.'"*

> *"The pre-action gate feels like the part that's going to get serious attention from regulated-enterprise and model-risk buyers."*

> *"Smart to scope it per customer rather than market a checkbox. This is enterprise-grade realism."*

## Separately, from JPMorgan AI Research

On first contact, a JPMorgan AI researcher characterized the engine unprompted:

> *"an interesting application of uncertainty decomposition, particularly the way you're mapping model divergence to specific failure modes like reasoning gaps or data stale-ness."*
>
> *— VP, AI Research, JPMorgan (May 2026)*

## Separately, from a Goldman Sachs model-risk leader

A former Goldman Sachs VP — now an enterprise-risk and model-risk governance leader — reviewed this summary together with the white paper and architecture brief, and engaged directly with the policy-vs-runtime-enforcement distinction at the core of the layer:

> *"The concept of a pre-action gate with escalation-to-human and audit-chain capabilities is certainly aligned with many of the discussions currently taking place around agentic AI governance. … I also appreciated the candid treatment of the current limitations and remaining development work. That level of transparency is not always common in this space."*
>
> *— Former VP, Model Risk, Goldman Sachs (2026)*

## Beyond the red-team — coverage at volume

Independent red-teaming proves the gate holds on the hard cases. A separate, automated coverage harness (run June 2026) proves it holds across the policy space we swept — and it is a **policy-oracle harness, not a spot check**: the gate is treated as a **black box**, and the expected verdict for every scenario is computed by a policy specification written by hand from the business rules, never by calling the gate's own code. That is what makes the result diligence-grade — non-circular, mutation-tested, reproducible, and kept separate from live model behavior. Two numbers, reported separately and never merged.

**Deterministic policy sweep — 4,697 scenarios, zero model calls.** Across risk tier × the four policy classes × decision boundaries × tenant variants, the gate agreed with the independent spec **100% of the time (4,697 / 4,697)**. The sweep is mutation-tested: deliberately flip a threshold inside the gate and the harness fails — so a perfect score reflects real agreement, not a circular test.

**Live consensus sample — 85% agreement, zero under-enforcement.** A stratified 40-scenario sample weighted toward high-risk actions, run through the real gate (the live verdict forms on a **three-provider consensus quorum**). Every block-required action was blocked (**6 / 6**) and every escalate-required action escalated; agreement with the spec was **85% (34 / 40)**. All six disagreements were the gate being *stricter* than policy — escalating an action the spec would have allowed. **Zero were under-enforcement**: the gate never let through anything policy said to stop.

Methodology, corpus version, sample size, model-call count, and latency are reported alongside the numbers, and every divergence is named as an open finding.

**Hard to replicate, by design.** The defensible work isn't the 4,697 count — it's the discipline behind it: business rules translated into an independent oracle, boundary cases generated across tenants and risk tiers, mutation tests that prove the harness catches regressions, and a live consensus sample scored separately. That is the difference between a demo guardrail and a governance system a regulated stack can be built on. Raw harness report, corpus version, mutation-test output, and load-test results are available under NDA.

## Open findings and progress, stated plainly

Honesty is the point of this summary, so the known gaps are named alongside the wins:

1. **MedQA −2.0 pp.** On MedQA (N=50), consensus scored 92.0% vs. Claude's 94.0%. Verifier-mesh tuning for medical-reasoning prompts is in the hardening queue; medical-decision-

adjacent agent workflows should wait for that pass. The reviewer's own line after seeing the healthcare scenario: *"Even with the prior −2pp regression you mentioned, the verifier mesh still added value here on regulatory interpretation. Once you do the tuning pass, this domain should get even tighter."*

2. **Hallucination-signal latency** under certain phrasing — in the hardening queue.
3. **Durable audit storage — now shipped.** Bundles persist to an append-only (WORM) store with database-level UPDATE/DELETE rejection and a fork-prevention index, verified at boot. At deployment, that store is externalized to the buyer's own object store under their keys, scoped during the private integration.
4. **Consensus-side audit-bundle attachment — resolved.** The consensus engine now hands out the complete bundle as JSON and PDF, alongside the agentic-trust surface, so the audit record is in the buyer's hands on both.

---

## Shipped since this validation (July–August 2026)

Three governance add-ons shipped **after** the red-team above, so they are named here as recent additions — **not** as part of the externally validated set. All three are tenant-scoped (cross-tenant lookup returns 404); Proof and Assurance are read-only over already-stored bundles (they never re-run a model), and Constellation is an escalate-only live signal that can only add a human check:

- **Constellation** — governs a whole agent workflow: it accumulates sensitive-data exposure across every agent in a `(tenant, workflow)` and can escalate a live action the moment their combined behavior crosses a line. Escalate-only and name-only (category and field names, never values); it can only add a human check — it never blocks and never alters a recorded decision.
- **Verdict Proof** — lets an auditor or the buyer's own risk team re-derive a stored verdict and re-check the SHA-256 hash chain offline, using a dependency-free standalone verifier. It re-checks the decision and the record; it does not re-run the four models.
- **Verdict Assurance** — renders a stored decision as a neutral, five-category evidence record (no free text, no PII) that a risk-transfer partner can assess. Verdict is not an insurer, and the record is not actuarial pricing, a certification, or coverage.

These extend the layer's Verify · Govern · Prove spine; independent testing of the add-ons themselves is future work.

**Also shipped since — a governance command center.** A read-only, tenant-scoped observability view (`GET /api/ai/consensus/observability`) reads the same tamper-evident audit ledger back and reports the fleet's behavior over time — verdict mix, risk-tier distribution, policy-class activity, four-provider consensus health, and a per-agent fleet view. It only reads the record and adds no new enforcement, and like the add-ons it shipped after the May validation, so it too is outside the externally validated set.

**And two verification-integrity signals since — Consensus Independence and Oversight Quality (July 2026). Consensus Independence** measures how correlated the four providers' returned answers were on a given decision and flags the case where independent-looking models line up too tightly and may share the same blind spot; it is a signal to look closer, not a verdict — it never blocks, never overrides, and leaves the calibrated confidence score untouched, riding alongside the decision as separate evidence. **Oversight Quality** reports, in aggregate only, how often escalations are genuinely acted on rather than waved through and how long they take to resolve — always across a group (a minimum of five reviews), never singling out an individual reviewer, with no free text and no personal data. Both postdate the red-team validation, so like the add-ons above they are named here as recent additions, not part of the externally validated set.

**And proof-of-origin signing on the audit bundle (July 2026).** Every audit bundle is now Ed25519-signed over its bundle hash — a hash proves the record wasn't changed, a signature proves who produced it — and signing keys are versioned and rotate forward, with every prior version kept published so a bundle signed by a rotated-out key still verifies. This postdates the red-team validation, so like the items above it is named here as a recent addition, not part of the externally validated set.

**And four more since — agent identity, regulation-mapped evidence, assurance telemetry, and the MCP gateway (August 2026). Agent identity** gives each agent its own Ed25519 keypair: requests carry signed, replay-proof identity envelopes, delegated authority travels as a signed chain that can only narrow hop over hop, and at `required` mode an unidentified or forged caller is refused before any model is consulted. **Regulation-mapped evidence** assembles the tenant's sealed records into an article-by-article EU AI Act evidence pack with explicit non-coverage notes, and drafts an Article 73 serious-incident payload from an escalated or blocked decision — documentation support, with the legal determination staying with counsel. **Assurance telemetry** reports 30–90 days of descriptive counts and rates — verdict mix, policy-class fires, identity coverage, oversight aggregates, ledger-integrity checks — with deliberately no composite risk score; Veridect is not an insurer. The **MCP gateway** governs MCP `tools/call` requests in flight, writing the verdict into the JSON-RPC response itself; a refused call is never forwarded to the tool. All four are verified in source and by unit tests and run live on the narrated tour at `veridect.ai/governance-suite`; like everything in this section they post-date the red-team validation and are not part of the externally validated set.

**And three more since — adversarial self-test, policy impact replay, and the dissent ledger (August 2026). Adversarial self-test:** the four providers author novel attack scenarios against the tenant's own live policy surface, and the same deterministic decision core judges every one in strict isolation — synthetic consensus, zero writes, no session state, results stored apart from the real evidence ledger; anything the deterministic floor alone would have greenlit is flagged to a human. It is a self-test — distinct from, and no substitute for, the external validation this document records: it deliberately withholds the four live model votes so it can measure the deterministic layer by itself. **Policy impact replay:** re-derives a window of real recorded decisions under a hypothetical policy — zero model calls, zero writes — and reports exactly which past decisions would flip and what drove each flip (a single field where one is responsible, a combination where several act together); thresholds the sealed record cannot re-derive (session and workflow exposure) are refused rather than silently ignored, and the underlying re-derivation primitive is byte-identical when no hypothetical is supplied — re-confirmed after the change by the full 4,697-scenario deterministic sweep at 100%. **Dissent ledger:** windowed counts of per-provider dissent and of human oversight outcomes on those same decisions — vindicated, partially vindicated, or overruled — with deliberately no provider ranking and no free text leaving the sealed bundles. All three are verified in source and by unit tests and are live on the engine's API and the narrated governance tour at veridect.ai/governance-suite; like everything in this section they post-date the red-team validation and are not part of the externally validated set.

**And four more since — the near-miss ledger, case-law engine, overnight frontier, and model character ledger (August 2026).** All four are read-only aggregations over the same sealed decision ledger — zero model calls, zero writes, tenant-scoped. **Near-miss ledger:** recorded approvals re-derived under single-knob tightened neighbors of the tenant's own attested policy; the ones that flip are reported by the exact threshold that would have tipped them — counts and threshold names only, a policy-sensitivity fact, never a claim the action was unsafe. **Case-law engine:** the closest past human-resolved reviews attached to a decision by deterministic feature-name similarity — resolution enums and reason codes only, free-text rationales sealed;

context for the reviewer, never a recommendation. **Overnight frontier:** a window of recorded decisions replayed under a deterministic grid of neighboring policies, mapping how many recorded approvals each variant would have routed to a human versus how many recorded escalations it would have released — Pareto-efficient variants flagged, recorded blocks invariant, nothing auto-applied. **Model character ledger:** each provider's recorded voting character compared across two adjacent windows — dissent rates, stance mixes with coverage disclosed, exact model identifiers seen — descriptive only, never a ranking, with insufficient_data as a first-class answer below the minimum sample. All four are verified in source and by unit tests and are live on the engine's API; like everything in this section they post-date the red-team validation and are not part of the externally validated set.

**And the fleet-scale layer since — the fleet exposure ledger, swarm convergence, and target convergence (August 2026).** Three further escalate-only deterministic policy classes extend the same discipline from one session and one workflow to the whole tenant fleet over sliding multi-day windows: per-identity and fleet-wide accumulation of sensitive reads across a seven-day ledger, distinct identities converging on the same sensitive category or hashed action shape inside a 72-hour window, and distinct identities converging on a single target held only as a truncated cryptographic hash. All three read field-name categories and hashes, never values, and can only add a human check — never a block, never an approval. After a cold-state detection the layer re-derives a bounded recent window from the sealed decision ledger under a single-owner rebuild lease; if the rebuild cannot complete, the fleet signal steps aside — the primary gate keeps enforcing, and a fleet escalation may arrive later. The enforcement memory can be rebuilt from the record rather than taken on trust. The classes ship default-off and are enabled per deployment. Verified in source and by unit tests, running live inside the pre-action gate; like everything in this section they post-date the red-team validation and are not part of the externally validated set. On 10 September 2026 the layer was exercised at scale in a deterministic run of our own — zero model calls, a unanimous synthetic approve standing in for the consensus: 10,000 distinct agent identities on one tenant, one proposed step each; first outbound step escalated for human review at identity #30 (three cleared before it, as the threshold of 25 specifies), all 1,496 outbound steps after it escalated or blocked, zero fleet escalations on a benign control fleet of 10,000, 10,000/10,000 agreement with an independent oracle re-statement of the policy, and a successful rebuild of the fleet window from the sealed ledger after a full fast-state wipe. Then the same swarm went through the published system as real requests on a dedicated tenant (10–11 September 2026) — ten thousand live rulings, each a real multi-model ruling with three providers responding and each sealed in its own signed record: 8,487 cleared, 1,263 escalated for human review, 250 blocked; detection at identity #30 again, and all 1,496 outbound steps after it escalated or blocked; a benign control fleet of 300 live rulings with zero fleet escalations; 10,300 of 10,300 signed records verified offline against the published key; all eleven acceptance checks passed. The live numbers are reported beside the deterministic ones, never merged; 120 live rulings carry a disclosed error entry from one provider instead of a stance, the remaining providers clearing the three-provider quorum on every one. Both are our own runs, not part of the externally validated set, and the ten-thousand figure is a count of identities ruled on, not a rate; the **Fleet-Scale Exercise Record** is published with both runs and SHA-256 bindings to the saved runs.

**And the coordination fabric since — seven further policy classes plus a policy-reversion report (late August 2026).** Seven further escalate-only deterministic policy classes govern the fabric *between* agents rather than any single action: the shape of verified delegation chains (depth beyond tenant policy, pass-through hops that only relay authority, wildcard grants, sensitive actions executing far from the human grant — the chain's cryptographic validity remains the identity layer's job; this class measures how authority traveled); separation of duties (the identity

that produced an artifact escalates the sign-off when it also approves, verifies, or releases that same artifact); two coordination shapes (fragments each individually under every line whose reassembled per-agent union crosses one at egress, and collect-transform-transmit pipelines spread across distinct identities); per-step conformance to an Ed25519-attested pre-committed plan (verb, target, and order checked against the committed step — deviation escalates); quarantined shared memory (writes from unverified identities or carrying sensitive-named fields are quarantined at ingest on a trigger-immutable provenance ledger; an action citing quarantined memory escalates until an audited, admin-gated release); and queue-message provenance (Ed25519-signed sender envelopes; an unsigned envelope that merely names an active agent is recorded as unproven rather than trusted; under the strictest setting only a gate-greenlit enqueue can attest a message hash, so a blocked enqueue cannot launder its own message into trusted work). All seven read field names, verb tokens, sanitized identities, and integrator-supplied hashes — never parameter values, never memory or message content — can only add a human check, ship with numeric knobs at zero and booleans off or activate only when the caller supplies the relevant reference (`planRef`, `memoryRefs`, `messageEnvelope`), and step aside on any state failure so the base gate keeps enforcing. A companion **policy-reversion report** answers "did a loosened knob change any real outcome?" by diffing the attested configuration history over a window and counterfactually re-deriving the window's recorded decisions under the prior policy — zero model calls, zero writes, with thresholds the sealed record cannot re-derive excluded and disclosed, never guessed. All are verified in source and by unit tests; like everything in this section they post-date the red-team validation and are not part of the externally validated set.

---

*Caveat on attribution:* the reviewer participated independently and is not citable by name without his permission. The scenarios, the observed behavior, and the verbatim quotes above are accurate and reusable. Full per-scenario instrumentation is available under NDA.

**Lisa Russell**
CEO
lrussell@veridect.ai · veridect.ai