Independent Validation Summary

Seven adversarial scenarios, run on the live production demos 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 **seven adversarial scenarios** against the live public demos at `www.ets-corporate-portal.com/quad-ai-demos` and `www.ets-corporate-portal.com/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. **All seven 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."

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 demo. 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)*

Open findings, 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.** Bundles generate and download today, but write-once (WORM) storage to the buyer's own object store is part of the 2–3 week systems-integrator-scale substrate work.
4. **Consensus-side audit-bundle attachment.** The downloadable bundle currently lives on the agentic-trust surface; backfilling it onto the consensus demo is a small remaining item.

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