QUAD-AI · TECHNICAL WHITE PAPER · MAY 2026

The Quad-AI Agentic Trust Layer

A pre-action consensus gate for autonomous AI agents — multi-provider verification, failure-mode attribution, and a tamper-evident audit trail in front of every consequential action.

No-NDA overview. Sections marked NDA-gated remain NDA-gated. Live interactive demo: **www.ets-corporate-portal.com/agentic-trust-demo**

1. Positioning

The Quad-AI Agentic Trust Layer sits in front of any autonomous AI agent and inspects each proposed high-stakes action *before* it executes. Whether the agent is built on the Model Context Protocol (MCP), OpenAI Assistants, Anthropic Computer Use, or any custom framework that can make an HTTP call, the proposed action is submitted to four leading language models in parallel — **Claude Opus 4.5, Gemini 2.5 Pro, GPT-5.1, and Sonar Pro**. Their independent verdicts are aggregated through weighted consensus, failure-mode attribution, and an adversarial verifier mesh. The gate returns one of three verdicts — **greenlight**, **escalate to human**, or **block** — together with a SHA-256-chained audit bundle that any downstream regulator, model-risk function, or internal compliance team can inspect line by line.

The agent receives an answer in roughly three seconds. The enterprise receives a defensible, replayable trail for every consequential action its agents take.

> ***In one line:*** *one gate sits in front of every autonomous AI agent in the enterprise, and every action is recorded, attributed, and replayable.*

2. The problem this solves

Three concurrent pressures created this product:

1. **Agent-action volume is curving up faster than human review can keep pace.** A single agent in a finance, operations, or customer-care workflow can propose hundreds or thousands of actions per day. Human-in-the-loop review on every action does not scale; no review at all is unacceptable. A gate that pre-clears the high-confidence majority and routes only the uncertain residual to a human is the only viable middle path.
2. **Single-vendor LLM trust is structurally insufficient for regulated environments.** Federal Reserve SR 11-7 (banking model risk), the NAIC Model Bulletin (insurance AI), and FDA AI/ML SaMD guidance (medical devices) all require independent challenge of the primary model. A single LLM cannot challenge itself. Multi-provider consensus with documented dissent is the natural architectural answer.
3. **Auditability of agent actions is now a board-level question.** "Why did the agent do that?" must have an answer that survives a regulator visit, a litigation discovery request, or an internal AI-governance audit. A SHA-256-chained bundle that records inputs, four-provider verdicts, the consensus calculation, the policy thresholds in force, and the final verdict — replayable months later — is the only form of answer that holds up.

3. How the gate works

A proposed action passes through a deterministic pipeline:

1. Action-payload introspection. The gate inspects the *structure* of the action — endpoint, parameters, scope, target system, calling-agent authority — not just a text prompt. This is what lets the gate apply proportionate scrutiny rather than burning four-provider consensus on a trivial read-only call.

2. Risk-tier inference. Each action is classified **low / medium / high / critical** from its verb, parameters, and any monetary amount. Irreversible or regulated operations (delete, wire, liquidate, terminate, production deploy) infer **critical**; money movement and customer-impacting changes infer **high**; standard writes infer **medium**; reads infer **low**. Tier governs how much scrutiny the action receives.

3. Four-model parallel consensus. All four providers evaluate the action in parallel under department-adaptive routing weights (Section 4). A minimum quorum is required; per-provider timeouts and circuit breakers prevent a single slow provider from stalling the gate.

4. Adversarial verifier mesh (high / critical only). For high- and critical-risk actions, a second oppositional pass extracts atomic claims, cross-references them across providers (supported / contradicted / omitted), and independently scores the selected response for factuality, completeness, and risk. This is the layer that surfaces failure modes the primary consensus missed.

5. Failure-mode decomposition. Disagreement is attributed to interpretable categories — **reasoning gap, stale knowledge, hallucination, domain mismatch** — rather than collapsed into a single opaque score.

6. Policy engine and verdict. Tenant-configured thresholds resolve the consensus into a verdict:

Verdict	When it fires
Block	A categorical scope violation — the action falls outside the agent's authorized authority (for example, a read-only agent attempting a delete or a wire).
Escalate to human	The fail-closed default. Triggered by conditional/unverifiable scope, provider dissent above tolerance, consensus confidence below the escalation floor, a verifier-recommended revision, or any critical-risk action — which always routes to a human sign-off even when approved.
Greenlight	Only when the action is in scope, the verifier verdict is APPROVE, consensus confidence clears the greenlight floor, and dissent is within tolerance.

Thresholds are **policy data, not code** — greenlight floor, escalation floor, and dissent tolerance are configured per tenant and bounded by deploy-time floors so a systems integrator can enforce a contractual minimum across many client engagements without engine changes.

7. Audit bundle. Every verdict produces a SHA-256-chained bundle (Section 5).

---

4. The verified models behind the gate

Provider	Model	Routing weight
Anthropic	Claude Opus 4.5	1.5×
Google	Gemini 2.5 Pro	1.3×
OpenAI	GPT-5.1	1.2×
Perplexity	Sonar Pro	1.1×

Weights are department-adaptive and tenant-tunable: a financial-services tenant can weight Claude higher on legal-domain decisions; a research vertical can weight Sonar higher on citation-grounded answers. The architecture is **provider-version-agnostic** — when a new model generation ships, a re-benchmark on the new lineup is a 15–30 minute operation, not a re-engineering effort.

5. The audit bundle

Every verdict produces an audit bundle that records the original action payload (hashed), each provider's full raw response, the consensus calculation (provider reliability weights, cross-provider claim agreement, dissent register, failure-mode flags), the tenant policy thresholds in force, and the final verdict and rationale — with timestamps and tenant binding.

Each bundle is SHA-256-hashed at three levels — **payload hash, verdict hash, and bundle hash** — and **chained**: each new bundle commits the hash of the prior bundle in the tenant's audit chain, which makes silent retroactive editing detectable. When the cleared action actually executes, the executed result can be **bound back into the chain** (a downstream-action hash), creating an end-to-end chain of custody from decision to execution.

Bundle storage is tenant-scoped: a verdict produced for one tenant is cryptographically bound to that tenant and unreachable to any caller authenticated as a different tenant. A cross-tenant lookup returns **HTTP 404, not 403** — bundle existence is never leaked across tenants, even with the exact bundle ID.

Honest status: the bundle is generated server-side on every call and is fully downloadable as JSON today (the live demo's *Download .json* button hands out the complete bundle). The current bundle store is in-memory; durable write-once (WORM) storage to the buyer's own object store under the buyer's keys is part of the systems-integrator-scale substrate work described in Section 8.

6. Integration surfaces

The Trust Layer exposes a small, stable set of endpoints:

Endpoint	Purpose
`POST /api/ai/consensus/pre-action`	The headline gate — submit an action payload, receive a verdict + bundle ID.
`POST /api/ai/consensus/delegate`	Agent-to-agent delegation gate — runs consensus on a hand-off before the downstream agent is invoked.
`POST /api/ai/adapters/mcp`	Native Model Context Protocol adapter — one-line change to an MCP tool-call surface.
`POST /api/ai/adapters/openai-assistants`	OpenAI Assistants adapter — slots in front of the function-call layer, no agent rewrite.
`POST /api/ai/adapters/computer-use`	Anthropic Computer Use adapter — gates screen-mediated actions (click, type, navigate, submit) before they execute.
`GET /api/ai/consensus/audit-bundle/:id`	Retrieve a tenant-scoped audit bundle (cross-tenant → 404).
`POST .../audit-bundle/:id/downstream`	Admin-only — bind the executed action result into the audit chain.

Endpoint	Purpose
`POST /api/ai/consensus/tenant-policy/:id`	Admin-only — set a tenant's policy thresholds.

If the buyer has agents on MCP, OpenAI Assistants, or Anthropic Computer Use — integration is the matching adapter; the tool-call surface points at the Trust Layer instead of the underlying provider. No change to the agent's reasoning loop. **If the buyer has a custom framework** — integration is the raw pre-action endpoint, typically a 10–20 line shim in the action-execution layer.

7. Verified performance

Independent benchmarks were re-run on the current four-provider lineup on **May 18, 2026**. Raw output files are reproducible on request.

Benchmark	Consensus	Best single provider	Lift vs best
MMLU-Pro (N=100) — headline	**85.0%**	Claude Opus 4.5 — 82.0%	**+3.0 pp**
AGIEval SAT-English (N=30)	**100.0%**	Claude / Sonar — 100.0%	matches best
AGIEval SAT-Math (N=30)	**96.7%**	Claude — 96.7%	matches best
MedQA (N=50) — **open finding**	92.0%	Claude — 94.0%	**−2.0 pp**

On MMLU-Pro N=100, consensus delivers **+3.0 pp over the best single provider and +7.5 pp over the individual-provider average (77.5%)**. The **MedQA −2.0 pp finding is disclosed openly** — verifier-mesh tuning for medical-reasoning prompts is in the hardening queue, and medical-decision-adjacent agent workflows should wait for that pass. Median end-to-end latency on the full four-provider lineup is **~3.1 seconds** including consensus, policy resolution, and audit-bundle write.

8. Multi-tenancy and deployment

A single Trust Layer deployment serves many independent tenants under cryptographic isolation — per-tenant policy thresholds, per-tenant audit chains, per-tenant rate-limit budgets, and tenant-scoped framework adapters. Shipped isolation primitives include composite-key tenant-scoped bundle storage, strict tenant-ID validation against header injection and path traversal, three-bucket rate limiting (per-tenant, per-IP global, and a per-IP cap on distinct tenant IDs that defeats tenant-rotation bucket-spam), an admin-token gate on mutation endpoints with constant-time comparison, and the cross-tenant 404 guarantee described in Section 5.

Deployment posture. Cloud-agnostic — AWS, Azure, GCP, or on-premises. Production mode (`PHASE4_REQUIRE_AUTHED_TENANT=1`) reads tenant identity from the buyer's existing auth middleware and **fails closed** if the authenticated principal is missing — there is no silent fallback to a default tenant.

Honest remaining work for systems-integrator-scale rollout (2–3 weeks): durable WORM bundle storage to the buyer's object store under the buyer's keys, cross-process state for rate-limit buckets and policy maps, per-tenant API-key provisioning lifecycle, and per-tenant provider quota pools. Full inventory is in the companion architecture document, available under NDA.

9. What it replaces

A buyer evaluating the Trust Layer is typically already paying for some combination of a per-vendor LLM contract with no cross-provider validation, a separate AI-guardrails vendor, a separate AI-governance documentation vendor, a separate AI-gateway/observability vendor, and an internal team writing custom agent-action audit tooling. The Trust Layer is **one integration, one contract** covering the runtime gate, multi-provider consensus, failure-mode decomposition, the policy engine, and the audit chain. It does not displace the underlying LLM provider contracts — those still serve the agent's reasoning calls — but it consolidates the trust, governance, and audit surfaces into one layer in front of every agent.

---

10. What is explicitly out of scope today

- **It is not a prompt-injection filter.** The gate evaluates the agent's *proposed action*, not the prompt that produced it. Pair it with a prompt-layer guardrail where injection is a requirement; the two are complementary.
- **It does not emit regulator-specific document formats out of the box.** SR 11-7, NAIC, and EU AI Act technical-file mappings are scopeable per engagement; the raw audit data is complete, the formatted deliverables are produced to the buyer's tooling.
- **It does not run fully air-gapped without provider API access.** On-premises gateway deployment is supported; fully air-gapped consensus is not. (The verifier-mesh and decomposition layer is provider-abstract, so a mixed self-hosted + hosted model lineup is supported under NDA-level scoping.)
- **MedQA medical-decision support** — see the open finding in Section 7.

---

11. Action paths

Without an NDA: open the live demo at `www.ets-corporate-portal.com/agentic-trust-demo` — run the agent scenarios, switch tenant context to observe isolation firsthand, and download a complete audit bundle as JSON. Read the public integration guide for the API surface. Request a scoping conversation; we respond the same week.

Under a mutual NDA: a buyer-specific integration document scoped to your environment and agents; the multi-tenancy architecture deep-dive and engine architecture assessment; a White Glove Enterprise Trial in your environment with your data and team, before any licensing or acquisition discussion; and a working session with the engineering side.

Lisa Russell
CEO