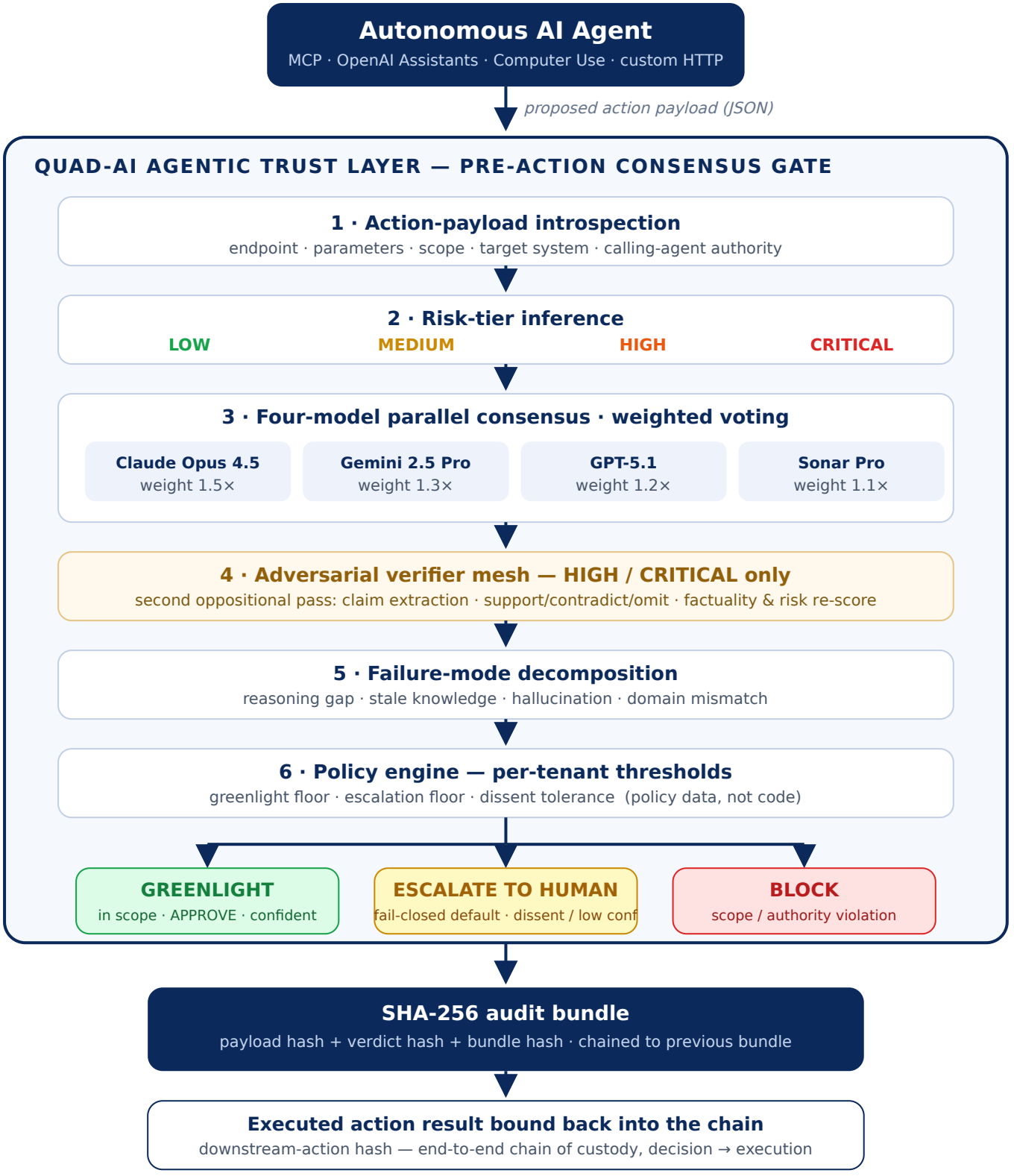
Pre-Action Consensus Gate — Flow

How one proposed agent action becomes a verdict and a tamper-evident audit record.



Operating envelope

- ~3.1s median end-to-end (consensus + policy + audit-bundle write); early-stop on 3-of-4 agreement.
- Multi-tenant isolation: per-tenant policy, audit chain, keys, rate limits; cross-tenant lookup → HTTP 404.
- Production mode fails closed: no authenticated tenant → 401, never a default-tenant fallback.

Live, interactive: **verident.ai/agentic-trust-demo**

Read the gate as a funnel. Low- and medium-risk actions clear quickly; high- and critical-risk actions earn the full adversarial pass before any verdict. Critical actions always route to a human even when approved. Every path — greenlight, escalate, or block — produces the same tamper-evident, replayable audit bundle. Scrutiny — and spend — scale with the stakes: the full four-model consensus lands on the decisions that can actually hurt you, not on every routine call. The engine behind this gate runs in production under Fortune 100 contracts in regulated industries.

Inside step 6 — deterministic policy-class detectors. Alongside the consensus thresholds, four deterministic detectors watch for the regulatory-exposure classes an enterprise can't afford to miss. They read only the action's **verb, target, and parameter field names** — never the values, never free-form context:

- **Monetary threshold** — a money-movement verb (wire, transfer, payout, refund) at or above a configurable limit (default $10,000).
- **Protected-personal-data modification** — a mutating verb touching a field named for PII (ssn, tax id, date of birth, bank account, card number, biometric).
- **Protected-class adjacency** — a consequential-decision verb (approve, deny, score, hire, price, underwrite) when a protected-class attribute is present (race, religion, gender, age, disability, pregnancy, veteran status).
- **Regulated-health-data access or transmission** — an access or send verb touching a PHI field (diagnosis, MRN, treatment, lab result).

These are a deterministic floor — independent of and **additive to** the model layer, so the hard regulatory lines never depend on a model being right. They are escalate-only: a detector can raise a greenlight to *escalate-to-human*, but never blocks, never auto-approves, never overrides a scope-violation block, and never softens an escalation already raised. Block stays reserved for a categorical scope or authority violation. Each trigger records the policy class and redacted category and field signals (for example `verb:update`, `field:ssn`, `phi:diagnosis`) — never the underlying value. And because they run in deterministic code, they enforce at **zero model cost** — even if every model is wrong, these lines still force a human into the loop, and they cost nothing in inference to run.

A complementary session-exposure signal. Optionally, when the integrator passes a session identifier (bound to the authenticated tenant), the gate also watches *cumulative* sensitive-data exposure across that session, and can escalate an outbound action once several distinct sensitive categories have been touched — even when no single step crossed a line alone. Like the four classes it's escalate-only and name-only; think of it as an extra layer of caution alongside them that only ever adds a human check, and steps aside if the session store is unavailable.

A correlated-consensus check on the agreement itself. Four models agreeing only means something when they agree *independently* — when they agree for the same reason, a shared stale fact or blind spot, confidence reads high while the answer can still be wrong. So alongside the verdict the gate reads how independently the four responses actually arrived, and when that correlation risk is elevated and at least three providers returned answers, it escalates the action to a human, carrying that read with the decision as evidence — the guard against the one failure consensus is blind to on its own: everyone being confidently wrong together.

The audit bundle carries the decision basis. The policy-class triggers and the decision basis are written inside the canonical bundle hash and persisted to an append-only (WORM) store, and each bundle exports as JSON, CSV, an HTML governance report, or OpenLineage NDJSON for SIEM and data-lineage ingestion. When the cleared action executes, its result is bound back into the chain, and any auditor can independently recompute every hash to confirm nothing was altered or inserted after the fact — a forged or out-of-band row fails recomputation or breaks the chain. Beyond the hash chain, each bundle is **Ed25519-signed over its bundle hash** — proof not just that the record is intact, but of

who produced it — and signing keys are versioned and rotate forward, with every prior version kept published (`GET /api/ai/consensus/proof/public-key` for the current key, `/proof/key-history` for every version), so a bundle signed by a rotated-out key still verifies. As with any signature, the published key is pinned out of band.

Three governance add-ons — Govern. Prove. Insure. The three governance problems the market keeps calling unsolved — Verideict solved all three, live in production. They extend the layer's Verify · Govern · Prove spine. **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 no single agent did — escalate-only and name-only, so it can only add a human check, never blocks, never auto-approves, and never alters a recorded decision. **Verideict Proof** lets an auditor re-derive a stored verdict, re-check the SHA-256 chain, and verify the Ed25519 signature offline with a dependency-free standalone verifier — it re-checks the decision and the record, it does not re-run the models. **Verideict Assurance** renders a stored decision as a neutral, five-category, PII-free evidence record a risk-transfer partner can assess — Verideict is not an insurer, and the record is not pricing, certification, or coverage. Proof and Assurance are read-only over stored bundles; all three are tenant-scoped (cross-tenant lookup returns 404) and run live on `verideict.ai/agentic-trust-demo`.

Governance observability — the command center. Every verdict on the flow above is also readable back as a live, read-only governance view over the same tamper-evident ledger: verdict mix, risk-tier distribution, which policy classes are firing, four-provider consensus health, a per-agent fleet view, and a running timeline — the whole fleet over time, not one decision at a time. It only *reads* the record: it never makes or changes a verdict and adds no new enforcement, it is tenant-scoped, it returns an error rather than a misleadingly empty view when records can't be read, and at production volume it serves from a materialized rollup. Live on `verideict.ai/agentic-trust-demo` (`GET /api/ai/consensus/observability`).

Oversight quality — proof the humans are actually looking. A gate that escalates is only as good as the reviewers on the other end. So the layer measures the review layer too: each resolved escalation is recorded to the same tamper-evident chain and rolled up into a plain rubber-stamp-risk read — low, moderate, or elevated — led by how often genuinely-contested actions were approved unchanged. It reads the pattern across enough reviews to be meaningful, never singling out one reviewer, and it holds only neutral markers — an opaque reason code and reviewer reference, never names or free text (`POST /api/ai/consensus/audit-bundle/:id/oversight` to record, `GET /api/ai/consensus/oversight/quality` to read).

Lisa Russell
CEO