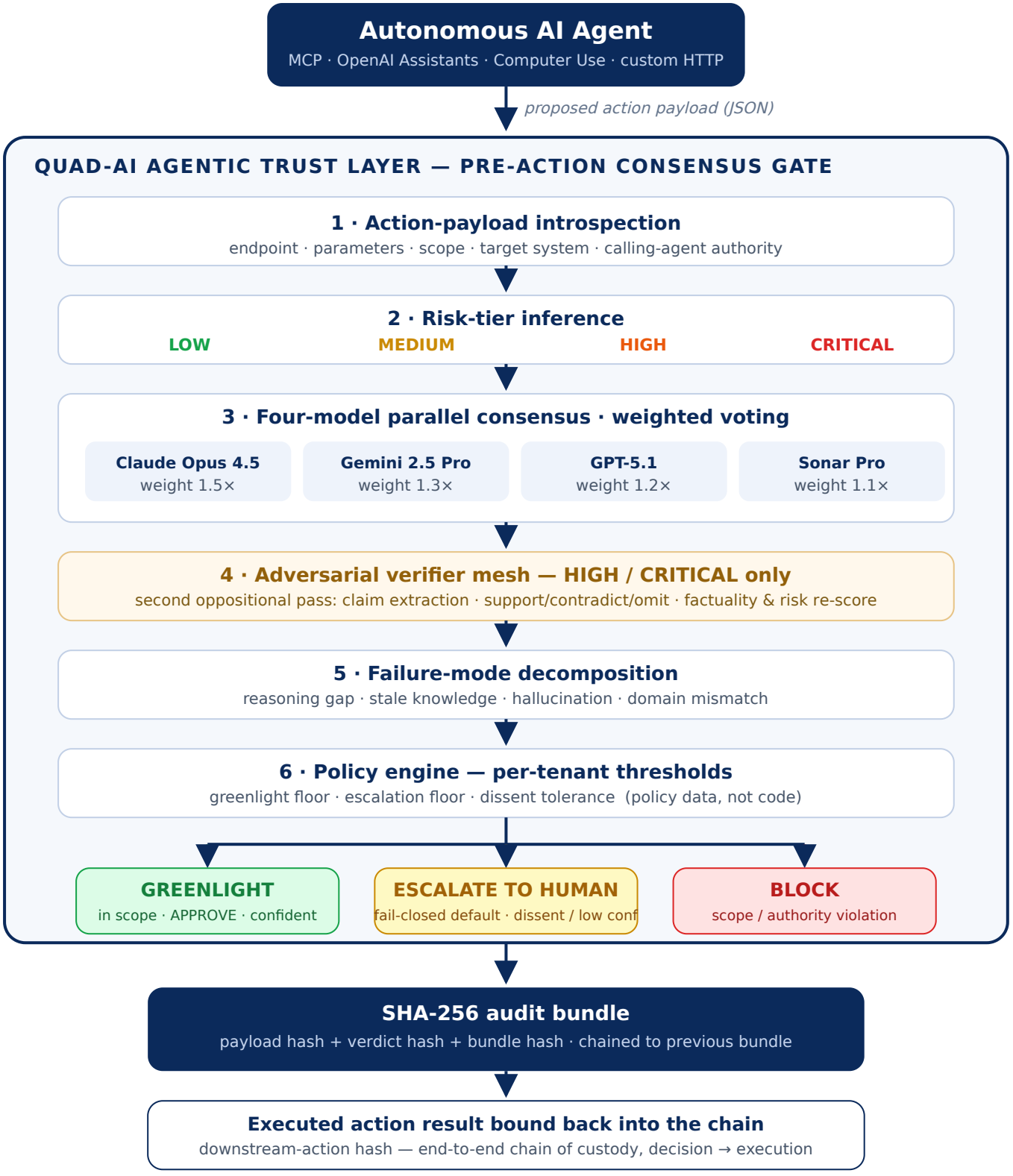
Pre-Action Consensus Gate — Flow

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.

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.

Lisa Russell
CEO