VERIDECT

AI Governance & Trust Layer

Platform Overview & Integration Guide

Production-Deployed | Provider-Agnostic | Industry-Agnostic

Independently validated by 10 AI providers across 6 industries.

Strategic Overview for AI Governance Leaders · August 2026 · veridect.ai

THE PLATFORM

The Opportunity

A production-deployed AI consensus engine, available for enterprise evaluation.

The engine

- Runs Claude Opus 4.5, GPT-5.1, Gemini 2.5 Pro, and Perplexity Sonar Pro in parallel on every query.
- Returns one verified consensus output with full SHA-256 audit trail and confidence decomposition on every call.
- Adversarial verifier mesh decomposes model disagreement into interpretable failure modes — reasoning gap, stale knowledge, hallucination, domain mismatch.
- Provider-agnostic: any future model swaps in or out as a configuration change. No vendor lock-in.

The record

- Production codebase: 568,934 lines of TypeScript across 1,407 files. Live system at veridect.ai.
- May 2026 Agentic Trust Layer: pre-action consensus gate for autonomous agents, SHA-256-chained audit bundles, adapters for MCP, OpenAI Assistants, and Anthropic Computer Use.
- May 2026 Multi-Tenancy Harness: one deployment, many tenants, cryptographic isolation, production-mode fail-closed auth.
- Industry-agnostic: finance, healthcare, legal, insurance, HR, cybersecurity.

THE PLATFORM

One Platform · Three Layers

Verify what AI says. Govern what it does. Enforce your data-governance policy — on every answer and every action.

01 · Verify what AI says

Quad-AI Consensus Engine

- Four leading models answer in parallel; weighted voting returns one verified answer
- Calibrated confidence score, decomposed to its source
- +3.0pp over the best single model on MMLU-Pro N=100 (current lineup)

02 · Govern what it does

Agentic Trust Layer

- Pre-action gate on every autonomous agent action before it executes
- One verdict in seconds: greenlight, escalate to a human, or block
- SHA-256 hash-chained audit bundle on every decision

03 · Enforce your rules

Data Governance

- Your policy enforced inside the gate, before any action runs
- Four deterministic policy classes — money over a limit, protected personal data, protected-class adjacency, regulated health data
- Sensitive actions escalate to a person; an out-of-authority action is blocked

One platform, one contract — every answer verified, every action governed, your policy enforced.

THE PLATFORM

The Frontier Layer · Govern. Prove. Insure.

The hard problems everyone names as unsolved — govern the whole fleet, prove any decision, insure the action. Verideict solved all three; they're live on the trust layer.

01 · Govern the fleet

Constellation

- One gate across every agent in a workflow, not one agent at a time
- Their combined exposure escalates the moment behavior crosses a line
- Escalate-only, each agent's part attributed — never blocks on its own

02 · Prove any decision

Verideict Proof

- The verdict is made in deterministic code, so it re-derives byte-for-byte
- Your risk team re-checks the decision and the SHA-256 hash chain offline
- A standalone verifier you download — it re-checks the record, not the models

03 · Insure the action

Verideict Assurance

- Every governed action becomes an underwriting-grade evidence record
- Mapped to a neutral, structured taxonomy a risk-transfer partner can assess
- Not an insurer, not actuarial pricing — the evidence an underwriter needs

Govern. Prove. Insure. — the frontier layer extends Verify · Govern · Prove, it doesn't replace it.

THE PLATFORM

Why Multi-Model Consensus

The structural problems with single-model AI — and how consensus fixes each one.

SINGLE-MODEL AI	QUAD-AI CONSENSUS
Hallucinations on confident-sounding outputs	Cross-model verification flags and corrects divergent claims
No interpretable explanation of uncertainty	Failure-mode attribution: reasoning gap, stale knowledge, hallucination, domain mismatch
No regulator-facing audit trail	SHA-256 audit trail on every output — every model, every score, every decision
Vendor lock-in to a single provider	Provider-agnostic; any model integrates as a configuration change
Single point of failure	Circuit breakers, automatic failover, real-time health monitoring, early-stop optimization
Inconsistent outputs across runs	Weighted consensus + verifier mesh + transparent confidence score

ARCHITECTURE

Two-Tier Consensus Architecture

Tier 1 selects the strongest answer. Tier 2 verifies and decomposes the confidence.

Tier 1 — Weighted Provider Selection

- All four providers fire in parallel on every query
- Each provider weighted by query type (math, science, essay, current events, creative)
- Highest-weighted successful response selected
- Early-stop optimization returns when 3 of 4 respond — cuts effective latency 25–40%
- Minimum 2 providers required for consensus quorum
- Per-provider AbortController cancels pending calls cleanly
- Every decision emits a SHA-256-keyed audit record (traceId, weights, latencies, evidence)

Tier 2 — Adversarial Verifier Mesh (March 2026)

- Atomic claims extracted from all provider responses
- Claims cross-referenced into support / contradict / omit matrix
- Independent verification passes score factuality, completeness, risk
- Synthesized response with explicit Decision, Rationale, Risks, Dissent, Open Questions sections
- Confidence decomposition: agreementScore + claimCoverage + contradictionPenalty + providerReliability + verificationBonus
- Opt-in via verificationMode: "mesh" — same engine, additional governance layer

ARCHITECTURE

The Verifier Mesh

What teams in regulated industries actually need: explainable, auditable model disagreement. A verified output ships with a structured record explaining why four independent models agreed — and where they did not.

1. Claim Extraction

Each provider response is decomposed into atomic, individually verifiable claims.

2. Cross-Reference Matrix

Each claim is checked against the other three providers — supported, contradicted, or omitted.

3. Independent Verification

Verifier passes score the selected response on factuality, completeness, and risk — separate from the original generators.

4. Failure-Mode Attribution

Disagreements are categorized: reasoning gap, stale knowledge, hallucination, domain mismatch.

5. Synthesis & Decomposition

Final response includes Decision, Rationale, Risks, Dissent, and Open Questions — with a transparent confidence breakdown.

`Audit record fields: traceId, providers[], evidence{parallelExecution, totalLatencyMs, consensusMethod, strictMode, earlyConsensus, abortedProviders[], cacheHit, costSaved}.`

ARCHITECTURE

Production Engine — Specifications & Modules

Verified from source code. No projections, no marketing math.

PROVIDER	MODEL (PRODUCTION)	BASE WEIGHT	SPECIALTY
Anthropic	Claude Opus 4.5	1.5×	Deep reasoning, essay scoring
OpenAI	GPT-5.1	1.2×	Creative reasoning, complex chains
Google	Gemini 2.5 Pro	1.3×	Scientific accuracy, STEM, multimodal
Perplexity	Sonar Pro	1.1×	Real-time search, fact verification

~3.1s
Median Latency

~2.2–2.8s
Early-Stop Latency

568,934
Lines of Production Code

1,407
Source Files

SERVER MODULE	LINES	PURPOSE
Quad-AI Orchestrator	936	Parallel execution, weighted voting, early consensus
Consensus Pipeline	—	Verifier mesh, claim extraction, decomposition
Circuit Breaker	299	Per-provider fault tolerance + auto recovery
Enhanced Cache	629	Redis + in-memory hybrid with cost tracking
Health Monitor + Endpoint	607	Real-time provider health + diagnostics API
Structured Logger	274	JSON logs + correlation IDs (AsyncLocalStorage)

Circuit breakers, LRU caching, automatic failover, real-time provider health monitoring, and department-adaptive routing weights are built in. Per-provider timeouts 45–50s; global timeout 65s after April 2026 hardening.

VALIDATION

Verified Benchmark Results — May 18, 2026 Re-Benchmark

Re-benchmarked May 18, 2026 on the current production lineup. MMLU-Pro N=100 is the headline. Raw output files reproducible on request.

BENCHMARK	SCORE	CONSENSUS RESULT	LIFT VS BEST SINGLE
MMLU-Pro N=100 (headline)	85 / 100	85.0%	+3.0 pp
MedQA N=50 (open finding — disclosed)	46 / 50	92.0%	−2.0 pp
Individual provider avg (MMLU-Pro N=100)	—	77.5%	—
Quad-AI Consensus (MMLU-Pro N=100)	85 / 100	85.0%	+7.5 pp over avg

+3.0 pp
MMLU-Pro Lift vs Best Single Provider

~3.1 s
Median End-to-End Latency (full 4-provider)

4
Verified Providers: Opus 4.5, GPT-5.1, Gemini 2.5 Pro, Sonar Pro

May 18, 2026 re-benchmark verified on the current four-provider production lineup. Consensus delivers +3.0 pp over best single provider on MMLU-Pro N=100 and +7.5 pp over the individual-provider average. MedQA −2.0 pp is disclosed as an open finding; verifier-mesh tuning for medical-reasoning prompts is in the hardening queue. Prior Feb 2026 K-12 benchmark (100% / +1.3 pp on the previous model generation) retained as a K-12-specific reference.

MARKETS

Two Markets, One Engine

The same architecture serves consumer / EdTech deployments and regulated-industry deployments — in different language.

EdTech, Consumer, Enterprise SaaS

Quad-AI Consensus Engine

- Four leading AI models check each other on every query
- Verified consensus output with confidence score
- Eliminates single-model failure points
- One integration replaces six AI vendor relationships
- Department-adaptive weights tune the answer per workload

Fits: EdTech platforms, consumer learning products, enterprise SaaS assistants

Finance, Healthcare, Legal, Insurance

Confidence Decomposition & Failure-Mode Attribution Layer

- Decomposes model disagreement into interpretable failure categories
- Auditable confidence scores for regulator-facing reviews
- Aligns to Fed SR 11-7 model risk management posture
- Aligns to NAIC AI Model Bulletin and FDA AI/ML SaMD frameworks
- Designed for environments where governance, explainability, and audit trails are required

Fits: model-risk management, financial compliance, insurance underwriting, healthcare governance

ECONOMICS

Four Models Is Not Four Times Your AI Bill

You don't run four models on everything. Verification lands where the stakes justify it — by design.

Why the bill doesn't balloon

- The hard compliance lines run at zero model cost. Money over a limit, personal data, protected-class factors, health data — all enforced in deterministic code, no inference at all. Your regulatory guarantees carry no model spend.
- Repeated questions are cached — you never pay to re-answer the same thing.
- The lineup is configurable per surface — the full frontier four on the decisions that can hurt you, a lighter configuration where they can't.
- Full four-model consensus is spent where it belongs: high-stakes, hard-to-reverse actions where one wrong call costs orders of magnitude more than a few cents of checking.

The reframe that wins the room

Four models isn't four times your AI bill — it's a few cents of extra checking on the one action in a thousand that could cost you millions, or a regulator's finding.

One engine across six departments replaces a separate AI vendor per function, so your total AI spend nets down, not up.

Honest note: the +3.0pp MMLU-Pro lift (85.0% vs 82.0%) is measured on the frontier four; route cheaper models on a low-stakes surface and that specific lift figure is tied to the frontier lineup.

VALIDATION

10 Independent AI Reviewers — 60 / 60 Analyses, Zero Failures

Each reviewer received the same corrected briefing and analyzed the architecture from a different perspective.

10	6	60	0
AI Providers Consulted	Industries Analyzed	Analyses Returned	Failures or Refusals

Reviewers

GPT-5.2 · o4-mini (Reasoning) · Claude Sonnet 4 · Gemini 2.5 Flash · DeepSeek V3.2 · Grok 4.1 · Mistral Large · Llama 4 Maverick · Cohere Command A · Qwen 3.5 Plus

Industries Analyzed

EdTech & Education Publishers · Financial Services & Investment Banking · Healthcare & Clinical Decision Support · Legal & Regulatory Analysis · HR & Employee Benefits · Cybersecurity & Threat Assessment

Full reviewer transcripts available under NDA. Excerpts on the next slide.

VALIDATION

What the Reviewers Said — On the Architecture

Verbatim. The specific language your own technical advisors will use.

"This isn't Kubernetes for AI — it's the Nuclear Regulatory Commission for AI. Kubernetes orchestrates containers; this engine regulates models."

ON TRUST INFRASTRUCTURE

"The circuit breakers, failure pattern detection, and consensus thresholds are the first system to treat AI models as adversarial witnesses, not trusted oracles. It's operationalized distrust."

ON OPERATIONALIZED DISTRUST

"The hard part isn't the 568K lines of code — it's the routing weights, which are the equivalent of a financial stress test for AI: a dynamic, battle-tested map of which model lies under pressure, which one contradicts itself, and which one cracks when the query gets weird."

ON THE ROUTING INTELLIGENCE

"The big players have unlimited engineers, but they don't have this data — because it only exists in production, under real-world load, with real consequences."

ON THE DATA MOAT

VALIDATION

What the Reviewers Said — On Industry Applications (1 of 2)

How regulated-industry analysts described the engine in their own language.

"In healthcare, this engine becomes a digital tumor board — multiple independent AI opinions converging on the highest-confidence answer before it reaches the clinician. When 3 of 4 AI models agree on a drug interaction warning or diagnostic flag, it's consensus medicine. The differentiator is proof, not performance: a full audit trail showing which models were queried, their outputs, the weighting, the consensus outcome — exactly the evidence that matters when FDA, CMS, and internal risk teams ask why the system recommended this."

HEALTHCARE

"A Quad-AI Consensus Engine turns probabilistic LLM output into a defensible, regulator-ready control — the AI equivalent of a four-eyes principle plus independent model validation. Consensus voting parallels how financial institutions already manage risk: ensembles with overrides, thresholds, and escalation paths. Early consensus optimization is a latency and cost arbitrage mirroring the fast-path / slow-path controls used in market-risk and surveillance stacks."

FINANCIAL SERVICES

VALIDATION

What the Reviewers Said — On Industry Applications (2 of 2)

How regulated-industry analysts described the engine in their own language.

"The privacy-first architecture eliminates the trust barrier: employers see ONLY aggregate metrics — enrollment, utilization, satisfaction, retention comparisons — while families maintain complete privacy. This is the first AI-powered education benefit that proves ROI without surveillance."

HR & EMPLOYEE BENEFITS

"Multi-model consensus creates a new primitive single-model chatbots can't: verifiable instructional reliability — '3 of 4 providers agree at 75%+ confidence; here are the citations and the dissenting view.' That becomes the backbone for adaptive assessment, IEP and neurodivergent learning supports, and AP prep where hallucinations are catastrophic, not just embarrassing."

EDUCATION

MARKETS

Cross-Vertical Applicability

One engine. Department-adaptive routing weights. Honest applicability — no projected cost savings.

VERTICAL	HIGH-STAKES WORKFLOW	WHY CONSENSUS MATTERS
Financial Services	Model risk review, credit decisions, RWA tokenization, audit	Auditable confidence + failure-mode attribution under SR 11-7
Healthcare	Clinical decision support, payer determinations, drug data review	Cross-model verification reduces clinically actionable hallucination risk
Legal	Contract review, regulatory analysis, e-discovery	Dissent column surfaces interpretive disagreement before liability
Insurance	Underwriting, claims, fraud signals	NAIC AI bulletin alignment; explainable adverse-action reasoning
HR & Talent	Resume screening, interview scoring, performance review	Bias cancellation across providers; documented decision rationale
Cybersecurity	Threat assessment, anomaly detection, incident response	Multi-model consensus reduces false-positive rates
Education	K-12 instruction, IEP/504 accommodations, parent-facing analysis	Production deployment — the proven domain for the engine

PRODUCTION PROOF

Live Production Platform

Three portals plus the AI Growth Measurement Framework. 118+ AI features. All running real queries against the engine.

Enterprise Portal

22 AI features

Executive dashboard, AI Pilot Program Builder, ROI projection tooling, Quad-AI Industry Tools page — the enterprise-facing live surface.

Student Portal

83 AI features

36,587-line self-contained K-12 application. Neural Challenge Arena, Homework Scanner, Concept Storyteller, IEP/504 auto-accommodations, WebXR layer.

Parent Portal

13 AI features

11,637-line self-contained app. Neural Learning Mirror confusion-detection layer, Learning DNA Profile, Family Learning Planner, voice-stress analysis.

AI Growth Measurement

Standalone framework

7-dimension framework. Pre-loaded 23-student cohort, classroom dashboard, per-student radar charts, comparison view.

Privacy architecture: Enterprise sees ONLY aggregate metrics. Parent and Student data entirely private. Session data ephemeral. FERPA + COPPA by design. See it live: veridect.ai/quad-ai-demos

PRODUCTION PROOF

Enterprise Portal — 22 AI Features

The enterprise-facing live surface. Built for executive review, not consumer onboarding.

Quad-AI Industry Tools Page

Six departments running real queries against the production engine — Legal, Finance, HR, Operations, Customer Experience, Marketing.

One API. Every Workflow.

Live visual showing how a single integration serves all six departments.

Integration Preview Panel

Real API request / response structure for each department before you commit.

AI Pilot Program Builder

Generates a structured trial scope from your requirements.

Executive ROI Dashboard

Real-time benefit / cost analysis with transparent methodology — no inflated marketing percentages.

System Status Endpoint

Public /api/ai/system-status returns aggregated provider health, weights, and recent decision audit summaries.

Dual Evaluation Paths

Distinct paths for technical evaluation (NDA + Architecture Pack) and commercial evaluation (guided trial).

EdTech Integration Block

Demonstrates LTI 1.3 compliance and direct LMS / SIS integration paths.

Plus 14 additional executive analytics, social-sharing, and enterprise-readiness modules. Full feature manifest available under NDA.

PRODUCTION PROOF

Student Portal — 83 AI Features (1 of 2)

Single self-contained 36,587-line K-12 application. Every AI interaction is verified through the Quad-AI engine.

CATEGORY	INNOVATION	WHAT IT DOES
Quad-AI Learning	Neural Challenge Arena	Multi-AI generated, gradually-harder challenges with consensus scoring.
	AI Learning Time Machine	Shows the historical context and evolution of any concept being learned.
	AI Concept Storyteller	Converts abstract concepts into narrative form tuned to the student's reading level.
	Quad-AI World Builder	Generates immersive learning scenarios for any topic.
	AI Knowledge Mapper	Visual concept maps showing how knowledge connects across topics.
	AI Learning Blueprint	Personalized cognitive learning plan from a short profile interview.
Daily Use	Instant Homework Analysis	Camera-based scan; engine returns step-by-step worked solution.
	Content Transformer	Convert any source into Summary / Step-by-Step / Quiz / Plain-Language.
	Essay Writing Assistant	Real-time multi-model suggestions during writing.

PRODUCTION PROOF

Student Portal — 83 AI Features (2 of 2)

Single self-contained 36,587-line K-12 application. Every AI interaction is verified through the Quad-AI engine.

CATEGORY	INNOVATION	WHAT IT DOES
Assessment	Adaptive Assessment Engine	Adjusts difficulty in real time based on response patterns.
	Neural AI Test Prep	AP and standardized test preparation with adaptive question selection.
	AI Comprehensive Progress Report	Multi-dimensional report with subject performance and recommendations.
Special Ed	Neurodivergent Learning AI Suite	Dedicated supports for ADHD, autism-spectrum, dyslexia, anxiety, executive-function, sensory.
Immersive	A-Frame 1.4.0 + WebXR	Real WebXR feature detection; works in any connected VR headset.
Safety	Quad-AI Four-Check Transparency	Student-visible badge on every answer showing four-provider verification.
Showcase	Multi-Agent Demo	Live in-portal demonstration of 50+ AI models running side-by-side.

Plus 67 additional student innovations: Genius Lab, Study Companion, SketchPad, Neural Vision Generator, Avatar Creator, Adaptive Interface AI, Curriculum Generator, AI Help Center, and more.

PRODUCTION PROOF

Parent Portal — 13 AI Features

Self-contained 11,637-line app. Built around: "Built for Children. Verified Before Every Answer."

INNOVATION	WHAT IT DOES
Neural Learning Mirror	Confusion-detection layer with auto-intervention. Surfaces friction points before the child stalls.
Neural Mirror Cognitive Report	Multi-dimensional cognitive report with Live vs. Simulation tracking modes.
Pre-Confusion Detection	Predicts where a student is about to get stuck and proactively scaffolds the next step.
AI Learning DNA Profile	Cognitive fingerprint: visual / auditory / kinesthetic, peak focus times, preferred modalities.
AI Family Learning Planner	Weekly schedule optimized by the engine based on the child's observed patterns.
AI Report Card Narrator	Translates a report card into actionable parent insights and an honest strategic action plan.
Academic Strategy Engine	Generates honest strategic action plans based on parent input and observed performance.
Voice Stress Analysis	Web Audio API tracks stress, volume, variability, and tone during academic conversations.
Growth Beyond Grades	Informational layer surfacing 7 measurement categories for AI-augmented student growth.
Social Story Generator	Creates personalized social stories for neurodivergent learners around specific scenarios.
Detailed Progress Report	Quad-AI generated subject performance analysis with parent-facing learning insights.

All consumer-facing features ship with COPPA + FERPA compliance, four-layer K-12 guardrails, XSS hardening, and input sanitization.

PRODUCTION PROOF

AI Growth Measurement Framework

Beyond test scores. The first measurement system designed for AI-augmented learning.

1 · Problem Decomposition

2 · Critical AI Literacy

3 · Cross-Subject Transfer

4 · Self-Directed Learning

5 · Confidence Trajectory

6 · Question Quality

7 · Verification Instinct

Pre-loaded cohort of 23 students with pre-computed growth analysis. Live backend APIs (/api/ai/growth-analysis, /api/ai/simulate-student). Renders any customer-supplied student data in JSON. Trial-grade pilot integration: ~2–3 hours for a competent engineer once student data is in JSON. Ships alongside the Student + Parent portals.

RECENT BUILDS

Recent Platform Enhancements (March – May 2026)

Active hardening across spring 2026. All live in production.

Quad-AI Engine

- Adversarial Verifier Mesh (March 2026) — opt-in claim extraction, support/contradict/omit cross-referencing, confidence decomposition. The moat.
- Department-Adaptive Routing Weights — weights adjust per department via DEPARTMENT_WEIGHT_MODIFIERS. One API call, six tuned outputs.
- System Status Endpoint — GET /api/ai/system-status returns aggregated provider health, circuit-breaker state, cache hit rates.
- May 18, 2026 Re-Benchmark — current four-provider lineup verified: MMLU-Pro N=100 consensus 85.0% vs best-single 82.0% (+3.0 pp), +7.5 pp over individual-provider average, MedQA −2.0 pp (open finding, disclosed).

Agentic Trust Layer + Multi-Tenancy Harness (May 2026)

- Pre-action consensus gate for autonomous agents: greenlight / escalate / block verdict on every action payload before execution.
- Three framework adapters live in production: MCP, OpenAI Assistants, Anthropic Computer Use. Same engine, agent-shaped surface.
- SHA-256-chained per-tenant audit bundles. Each new bundle hash incorporates the previous bundle hash. Tamper-evident.
- Multi-tenancy harness shipped: composite-key tenant-scoped storage, strict tenant-ID validation, three-bucket rate limiting, admin-token gate with constant-time compare, production-mode fail-closed auth.
- Live system: veridect.ai/agentic-trust-demo.

RECENT BUILDS

Recent Platform Enhancements — June 2026 (1 of 2)

Six weeks of governance hardening — all live in production and validated against the live system.

Agentic Trust Layer — Governance Hardening

- **Deterministic policy detectors** — four classes (dollar threshold, protected personal-data changes, protected-class adjacency, regulated health-data access and transmission) escalate high-risk actions to a human. They never auto-approve and never silently block; the gate blocks only an out-of-scope action the agent has no authority to take.
- **Tamper-evident WORM audit ledger** — every decision is SHA-256 hash-chained to the one before it, with database-level write-once protection and a full-chain integrity check.
- **Formatted compliance exports** — CSV, governance JSON, print-ready HTML, and OpenLineage, ready to hand straight to a model-risk or audit team.
- **Privacy by design** — the gate records signals by category only (for example "field: SSN", "verb: transmit"), never the raw personal or health data itself.
- **Gate-coverage harness** — the gate is scored against an independent policy oracle, never its own code: a deterministic sweep of 4,697 scenarios with zero model calls agreed 100%, and a stratified 40-scenario live sample agreed 85% (34/40) with zero under-enforcement — every miss was the gate over-escalating, never letting risk through. Mutation-tested and non-circular; raw report under NDA.

RECENT BUILDS

Recent Platform Enhancements — June 2026 (2 of 2)

Six weeks of governance hardening — all live in production and validated against the live system.

Engine Integrity + Platform (June 9 onward)

- **Four-provider honesty fix** — the public consensus surface and enterprise Quad-AI path now wait for all four models, so the downloadable audit bundle genuinely reflects Claude Opus 4.5, Gemini 2.5 Pro, GPT-5.1, and Sonar Pro; the consumer portals and the pre-action gate keep the fast early-stop path, and any provider drop is flagged on screen.
- **Verified benchmark on the current lineup** — MMLU-Pro N=100 consensus at 85.0%, +3.0 pp over the best single model; MedQA N=50 disclosed openly at −2.0 pp as an honest open finding.
- **Per-tenant isolation** — one deployment, many clients, each with its own policy thresholds and a tenant-scoped audit store with no cross-tenant leakage; KMS binding and a production load test are scoped per deployment.
- **Resilience** — automatic provider failover, circuit breakers, and LRU caching keep the engine responsive when a provider degrades; a formal production load test is scoped per deployment in the customer's environment.
- **Everything now served on veridect.ai** — the engine tools, the pre-action gate, and the red-team page, plus a Trust Center documenting HIPAA BAA status, encryption, and NIST AI RMF mapping.

RECENT BUILDS

Recent Platform Enhancements — July 2026

The July 14 build — input safety on the Q&A and session-aware exfiltration defense. Live in production.

Input Safety on the Consensus Q&A (July 14)

- **Live PII redaction** — email, phone, credit-card, SSN, and street-address formats are stripped from the question before it reaches any model. The audit records that redaction happened, never the value. Formatted identifiers only; names are not scrubbed.
- **Content and jailbreak moderation** — known prompt-injection and jailbreak attempts, and genuinely harmful content, are refused before the four paid models run. Everyday business, legal, finance, medical, and education prompts pass untouched.
- **The net effect:** personal details are removed and known attacks refused before the four-model consensus runs — the screening is a lightweight check, so you never pay the frontier four to answer an attack.

Session-Aware Exfiltration Defense (July 14)

- Catches the theft that hides in ordinary steps. No single action looks wrong — the pattern does. Once too much sensitive material has been read across one working session, an outbound action is escalated to a human.
- Escalate-only by design — it can raise an action for human review, never silently approve or block one — consistent with the deterministic policy floor.
- Built for the slow-burn, multi-step exfiltration pattern that per-action checks miss.

RECENT BUILDS

The Governance Command Center — July 2026

The July 19 build — the tamper-evident audit ledger, read back as a live, whole-program view. Live in production.

What the Command Center Shows (per tenant)

- One place to watch the whole program: total decisions — how many were greenlit, escalated, or blocked — with the verdict mix, the risk-tier breakdown, and which deterministic policies fired.
- Consensus health and a per-agent fleet view sit alongside a live recent-decisions feed. Every figure is drawn from the same SHA-256 hash-chained record an auditor can replay — not a separate log that could fall out of step with the truth.
- **The net effect:** one verdict becomes a governance program you can watch — your AI observed across every decision, over time, across the whole fleet.

Built for Trust

- Read-only by design — it reads and counts what already happened. It never makes or changes a verdict, and never re-runs a model; it adds no new enforcement.
- Tenant-scoped — each customer sees only its own ledger; another tenant's activity is not visible and not discoverable.
- If records can't be read, it says so plainly — an explicit error, never a false "nothing to see here" dashboard.

RECENT BUILDS

Consensus Independence & Oversight Quality — July 2026

The July 22 build — is the consensus genuinely independent, and is the human oversight real? Verified in source and covered by unit tests.

Consensus Independence (July 22)

- Agreement between models only counts when it is reached independently. After consensus resolves, an independent read scores how correlated the four providers' answers were.
- It catches the consensus trap — four models confidently wrong together on a shared blind spot or the same stale fact — and flags it as a signal to look closer, not proof the answer is wrong.
- Escalate-only: when correlation risk is elevated it escalates the action to a human and rides with the decision as evidence — the calibrated confidence number itself stays untouched.

Oversight Quality (July 22)

- A human-in-the-loop control is only real if the human actually looked. Each human resolution of an escalated action is recorded as a write-once row in the same hash chain as the verdict.
- Read in aggregate, it surfaces a rubber-stamp pattern — so the human oversight can be shown to be genuinely engaging, not a formality.
- It reads the pattern across enough reviews to be meaningful, never singling out any one reviewer; time-to-resolution is drawn from the record, not self-reported.

RECENT BUILDS

The Oversight Analytics Layer — August 2026

Four read-only instruments over the sealed decision ledger, on the engine's API — no model calls, and a human owns every outcome.

Near-Miss Ledger + Case-Law Engine (August 2)

- The near-miss ledger re-derives every recorded approval under stricter neighbors of its own attested policy — and names the approvals that sat one threshold away from a different verdict. A sensitivity reading on the policy, never a second verdict on the action.
- The case-law engine attaches the closest past human rulings to a hard case, matched on the recorded shape of the action — never on wording — so a reviewer decides with institutional memory on the table.
- Both read the sealed, tamper-evident ledger the verdicts live in; neither can alter a decision.

Overnight Policy Frontier + Model Character Ledger (August 2)

- The overnight frontier replays recent decisions under up to 72 neighboring policy variants and maps the exact trade each would have made — evidence on the table by morning. Nothing auto-applies; adopting a change stays a human decision.
- The model character ledger compares each provider's recorded voting behavior window over window, so a silent change in a vendor's model becomes a dated, documented fact — descriptive by design, never a ranking.
- Together with the self-test, replay, and dissent rails, the engine now runs seven independent oversight jobs on its own decision record — alongside signed agent identity, regulation-mapped evidence, assurance telemetry, and in-path MCP gateway enforcement.

RECENT BUILDS

The Newest Two (1 of 2) — Configuration Attestation

Shipped after the external validation record below — disclosed as such, in keeping with how this platform reports itself.

ATTEST

Configuration Attestation

The watchdog watches itself

- A governance layer is one quiet configuration change away from meaningless — and the decisions themselves would never show it.
- The enforcement configuration the gate actually consults — thresholds, detector set, identity mode, signing state — is attested into its own signed, append-only ledger, and every change is mechanically classified as a tightening or a loosening. A quiet weakening cannot pass as routine maintenance.
- Each decision is stamped with the fingerprint of the configuration in force when it ran — withheld rather than guessed when the engine cannot be certain.

Stated plainly: this is self-attestation — the system that enforces policy is the system recording its own settings. It makes tampering with the configuration history evident; it is not independent third-party oversight.

RECENT BUILDS

The Newest Two (2 of 2) — Cascade Seal

Shipped after the external validation record below — disclosed as such, in keeping with how this platform reports itself.

SEAL

Cascade Seal

One signature over the whole workflow

- Agent failures live in chains, not single steps — an agent calls an agent that calls a tool, and the risk hides between the hops.
- Every gate-observed decision in a multi-agent workflow is bound into a single signed Merkle root — the seal binds each record, its position, and the total count — so the entire cascade verifies in one check: offline, with the same standalone verifier, pinned to our published signing key.
- A dropped hop, a reordered hop, or a record signed by any other key fails the check — and a re-check against the live ledger distinguishes a later legitimate append from tampering.

A seal proves the recorded hops are intact and in the order recorded. It does not prove completeness — an action that never passed through the gate cannot appear in the seal.

DEFENSIBILITY

Why It Can't Simply Be Rebuilt

You can see everything it does. How each piece is built stays behind the NDA — the part watching it run can't reveal.

The moat is the whole, not any one part

- Anyone can wire up four model APIs. What can't be assembled on a timeline is an integrated control plane that cross-examines its own answers, gates an action before it executes, and leaves a tamper-evident record an auditor can replay — all working as one system.
- Each capability is re-implementable given time. All of them validated end-to-end as one control plane is the multi-year part — and the part that carries the risk.
- Production-proven inside Fortune 100 enterprises, with a multi-year record of passing annual vendor security reviews. You can't back-date production proof; a build starts at zero on the thing that takes longest to earn.

What you see vs. what transfers

- **In the open:** it working. Run any scenario yourself on the live system, independently red-teamed, open findings and all.
- **Under NDA:** the mechanism behind every capability — the part that can't be reverse-engineered from watching it run — is shared in the private integration documentation, along with the full validation record.

We sell what it does and why it's hard to build. The how stays behind the NDA — deliberately.

HONESTY

Access, Identity & Honest Boundaries

What Veridect governs — and, just as clearly, what it doesn't. The honesty is the point.

What we control

- Per-tenant API keys we provision, rotate, and revoke; strict tenant-ID validation; authenticated-tenant resolution.
- An admin-token gate on any policy change, and full tenant isolation — one client's compliance officer can pull their records without ever seeing another's.
- Home turf is the control and audit layers: the pre-action gate, the verifier mesh, scope-violation block, escalate-to-human, and SHA-256 hash-chained, replayable audit bundles with OpenLineage export.

What we don't claim to be

- **Not an enterprise identity provider** — no SSO/SAML/OAuth for your end users. We control access to the engine per tenant; we don't replace Okta or Entra.
- **Not retrieval** — no RAG, vector DBs, or embeddings. But our confidence decomposition separates hallucination from reasoning-gap, stale-knowledge, and domain-mismatch — which is what tells you whether a bad answer is a retrieval problem.
- **Not full APM** — we produce decision-level governance observability, not infrastructure telemetry. We don't claim the whole stack; we claim the trust spine running through it.

INTEGRATION GUIDE

Integrating Veridect — What It Takes

The engine is multi-model AI middleware. From your application's perspective, it looks like calling one AI provider. Under the hood, four independent models verify each other before the answer is returned.

One request in, one verified answer out

- Fires the query to 4 AI providers in true parallel — not sequentially, not as fallbacks
- Applies weighted consensus voting based on query type and provider reliability
- Monitors each provider through independent circuit breakers
- Returns the consensus-verified answer with a complete audit trail

What your application receives

- The consensus-verified answer + a confidence score (0–100%)
- Per-provider latency and success/failure status
- A unique trace ID for audit and replay · consensus method used
- Circuit-breaker status per provider · cost saved through early consensus and caching

The Production Lineup

The four production providers behind every consensus answer — swappable by configuration.

PROVIDER	MODEL (PRODUCTION)	STRENGTH
Anthropic	Claude Opus 4.5	Deep reasoning, step-by-step logic
OpenAI	GPT-5.1	Creative reasoning, nuanced content
Google	Gemini 2.5 Pro	STEM, scientific accuracy
Perplexity	Sonar Pro	Real-time data, fact verification with citations

Provider-agnostic by design — models swap in or out as a configuration change. Adding your own providers is covered in the private integration documentation under NDA.

INTEGRATION GUIDE

The Consensus API

One endpoint carries the whole engine. Same request shape for streaming. Health endpoints for your load balancers and orchestration.

```
POST /api/ai/quad-consensus
Authorization: Bearer <your_api_key>

{
  "prompt": "Explain photosynthesis to a 4th grader",
  "systemMessage": "You are a science tutor...",
  "taskType": "explanation",
  "metadata": { "clientId": "your-app-id" }
}
```

prompt (required, ≤10,000 chars) · systemMessage (optional, ≤5,000) · taskType selects an optimized system prompt and routing-weight profile · metadata passes through to the audit trail, never to the AI providers.

```
{
  "success": true, "consensus": true,
  "result": "Photosynthesis is how plants...",
  "confidence": 75,
  "providersUsed": 3, "totalProviders": 4,
  "traceId": "quad-1711000000000-abc1234",
  "providers": [ { "name": "Claude Opus 4.5",
      "ok": true, "latencyMs": 2847, ... } ],
  "evidence": { "consensusMethod": "weighted-voting",
      "earlyConsensus": true, "cacheHit": false,
      "totalLatencyMs": 2847, "costSaved": 0.003 }
}
```

Streaming variant: POST /api/ai/quad-consensus/stream returns Server-Sent Events as each provider completes. Health: GET /ready (load balancers), GET /alive (liveness probes), GET /metrics/prometheus (monitoring). Early-stop (3/4) is a per-surface configuration — surfaces whose audit must reflect all four providers run strict four-provider completion, as Veridect's own public consensus surface does.

INTEGRATION GUIDE

Drop-In Adoption — Replace One API Call

Your existing AI call becomes a call to the consensus engine. Everything else stays the same — a 10–20 line change in most codebases.

BEFORE — single provider, unverified

```
response = openai.chat.completions.create(
    model="gpt-4",
    messages=[{"role": "user", "content": prompt}]
)
answer = response.choices[0].message.content
```

AFTER — consensus-verified by 4 providers

```
response = requests.post(
    "https://<engine>/api/ai/quad-consensus",
    json={"prompt": prompt, "taskType": "question"},
    headers={"Authorization": "Bearer KEY"})
data = response.json()
answer     = data["result"]      # verified answer
trace_id   = data["traceId"]     # full audit trail
confidence = data["confidence"]  # provider agreement
```

TASK TYPE	OPTIMIZED FOR
question	Direct factual questions — accuracy, conciseness
explanation	Concept explanations — clarity, step-by-step breakdown
analysis	Content analysis — depth, evidence-based reasoning
homework	Guided problem-solving — pedagogical quality
assessment	Evaluate work, give feedback — specific, actionable
creative	Creative content — engagement, originality

Custom system messages override task-type defaults entirely — use them where your domain (healthcare, legal, finance) requires specific instructions. Client patterns for Python, Node.js, Java, C#, and cURL are in the Public Integration Guide.

INTEGRATION GUIDE

Agent Frameworks — Governing What AI Does

The same drop-in philosophy for autonomous agents: the pre-action gate returns one verdict — greenlight, escalate, or block — before an action executes.

Three adapters, live in production

- **MCP** — including in-path MCP gateway enforcement: tool calls pass through the gate on the way to the tool
- **OpenAI Assistants**
- **Anthropic Computer Use**

Same engine, agent-shaped surface.

What the gate sees

- The action payload — verb, target, fields — before execution
- Deterministic policy classes checked at zero model cost
- Four-model consensus where the stakes justify it
- Session-level cumulative-exposure tracking across a workflow

What your stack gets back

- One verdict in seconds, with reason codes
- A SHA-256 hash-chained audit bundle per decision
- Escalations routed to your human reviewers
- Compliance exports: CSV, governance JSON, print-ready HTML, OpenLineage

Agents built on other frameworks integrate through the same REST surface the adapters use — the gate is framework-agnostic at the API level.

Reliability & Operations

Provider failures are the engine's problem, not your application's. No retry logic, no failover handling, no provider health monitoring on your side.

Circuit breakers, automatic

- CLOSED → 3 consecutive failures → OPEN
- 30-second cooldown → HALF-OPEN → 2 successful probes → CLOSED
- Per-provider timeouts 45–50s; degradation and recovery both automatic

Graceful degradation

- 4/4 providers: full consensus, up to 100%
- 3/4: consensus at up to 75% · 2/4: reduced, up to 50%
- 1/4: single provider, flagged — no consensus claim · 0/4: explicit error, immediate escalation

Cache behavior

- Cache miss: full 4-provider parallel run, 2–4s
- Cache hit: <50ms, zero provider cost
- Early consensus: 3/4 agree, 1.5–3s, 25–40% savings — per-surface configuration; strict 4-provider completion where the audit must reflect all four
- SHA-256 cache keys · 24h default TTL, 1h for time-sensitive queries · Redis primary with in-memory LRU fallback

Rate limiting

- Per-tenant quotas with standard X-RateLimit-* headers on every response

Data Handling, Security & Compliance

How query data is handled in operation — and the compliance posture it supports.

Data handling

- Queries processed in-memory; not stored beyond cache TTL
- No query content used for model training
- All provider calls over HTTPS/TLS 1.3
- Strict input validation (Zod schemas); sanitized error messages

Compliance posture

- FERPA + COPPA compliant by architecture
- SOC 2-ready logging: structured JSON, correlation IDs, complete audit trails
- GDPR considered — no personal data retained beyond cache TTL; DPA available

INTEGRATION GUIDE

What's Public, What's Under NDA

The public guide shows the full API surface and integration patterns. The mechanism behind the engine transfers under NDA.

Available under NDA

- **Custom Integration Package** — company-specific configuration, dedicated endpoints, custom routing-weight profiles for your domain
- **Provider Adapter Specification** — adding your own models (proprietary, open-source, regional) to the engine
- **Routing Weight Methodology** — how domain weights are calibrated; documented failure patterns per provider
- **Circuit Breaker Deep Dive · Cache Architecture · Audit Trail Schema**
- **Architecture Assessment & Multi-Provider Analysis** — the full independent-review records
- **Deployment Runbook** — on-premise deployment, infrastructure requirements, scaling configuration

INTEGRATION GUIDE

Proof of Concept — The First Week

From API key to a verified evaluation — four steps, one call replaced.

Proof of concept, step by step

- **1 · Get API access** — request a proof-of-concept key: company name, primary use case, estimated query volume, preferred language
- **2 · Make your first call** — one cURL command, consensus-verified in seconds
- **3 · Replace your existing AI call** — the before/after on the previous page; one call replaced, four providers verifying
- **4 · Evaluate** — consensus accuracy vs. your current provider, latency (typically 2–4s full consensus), audit quality, circuit-breaker behavior during outages

---

One API call replaced. Four providers verifying. Full audit trail.

VERIDECT

See the live system, not a slide about it.

veridect.ai/governance-suite — the live governance tour: real decisions, real ledger, twelve stations

veridect.ai/validation-summary — the full validation record, open findings included

veridect.ai/critical-solutions — fourteen agentic-AI failure modes, each one solved and live

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