Intelligence for AI Agents, LLMs, and Agentic Workflows

Revefi gives data, AI, and engineering teams cost visibility, reliability monitoring, and agent governance across models, providers, and users all in one unified platform.

Gartner COOL VENDOR 2025

Why Choose Revefi for AI Observability?

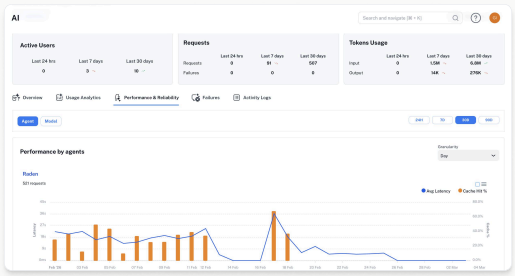
As AI agents and LLM workflows proliferate, organizations face a critical gap: costs are invisible, failures are silent, and performance is opaque. Revefi closes this gap by providing a single pane of glass across every model call, agent action, and multi-step workflow with zero code changes required.

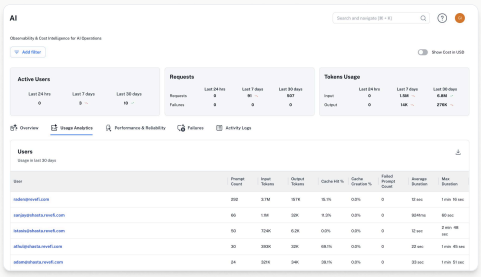
- **Full attribution chain:** Trace interactions from user to agent to model, capturing cost, and output at each step.
- **Multi-provider visibility:** Unified observability across OpenAI, Anthropic, Google, and models.
- **Instant time-to-value:** Connect in minutes.

Core capabilities

AI Observability

- Full user to agent to model attribution chain
- Per-agent latency, request volume, and prompt response capture
- Latency benchmarking across GPT, Claude, and Gemini
- Throughput metrics in tokens/sec
- Failure rate tracking across providers and time windows



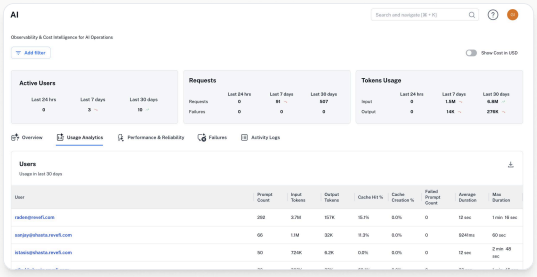


AI FinOps

- Full cost breakdown by provider, model, agent, and user
- Input vs. output token analytics with trend analysis
- Cost outlier detection: identify which users, agents, or prompts drive spend
- Cache hit rate monitoring as a direct optimization lever
- Budget guardrails and spend anomaly alerts

Prompt Optimization

- Identify the highest-latency prompts across models and agentic workflows
- Monitor prompt reuse patterns to reduce redundant model calls
- Connect prompt-level patterns to output quality metrics
- Detect which prompts produce inconsistent or poor outputs



How Revefi Works for AI & Agentic Workflows

Revefi connects to your AI stack in minutes, using telemetry from model providers, orchestration layers, and agent frameworks without complex integrations. It monitors model calls, agent steps, token flow, and cost events at a granular level, delivering real-time visibility across providers, models, and teams.

1. Connect: Point Revefi at your AI providers and model endpoints.

2. Monitor: Continuous tracking of every call, agent action, latency event, token spend, and failure across your entire AI stack.

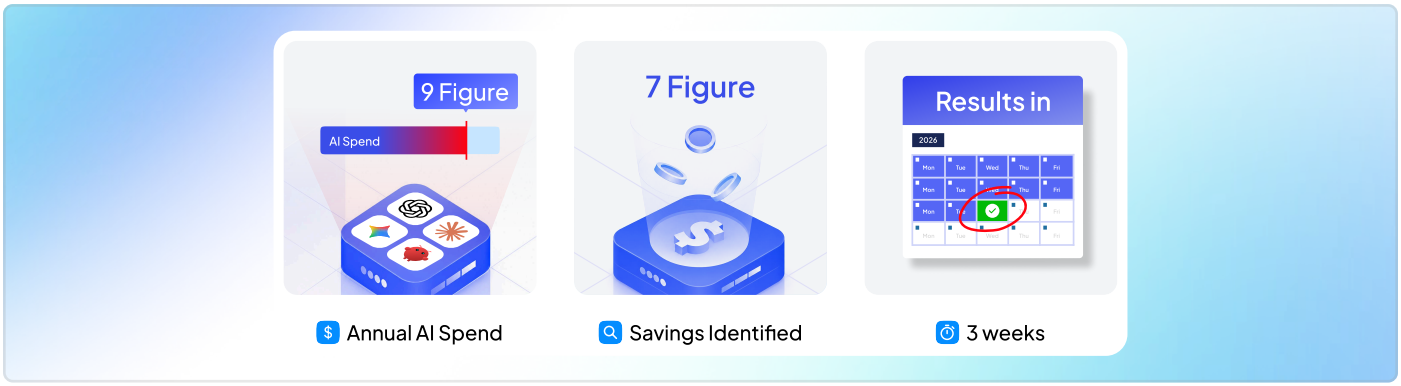
3. Analyze: Correlate cost, performance, and reliability data in a single unified view across providers, agents, users, and time windows.

4. Optimize: Act on recommendations to reduce token costs, fix slow prompts, eliminate runaway agent spend, and improve output quality.

Token Economics:

- Single-platform observability and FinOps across all models, providers, agents, and users.
- Trace user-to-agent-to-model workflows with complete prompt/response logging.
- Break down exact costs and token consumption trends by provider, model, agent, and user.
- Pinpoint high-spend outliers and use cache hit-rate tracking to eliminate redundant calls.
- Connect prompt patterns to latency, cost, and output quality to drive data-led optimizations.

A Fortune 200 Enterprise spending nine figures on AI.



AI Model Providers

OpenAI · Claude · Gemini

Data Platforms

snowflake · databricks · Google BigQuery

Organizations That Trust Revefi

AMD · Ingersoll Rand · StanleyBlack&Decker · Verisk · Cribl · OceanSpray · The Princeton Review · Help at Home · docker

★ Gartner Cool Vendor 2025

Gartner COOL VENDOR 2025

Put Your AI Stack on Autopilot with Revefi

Get full cost visibility, reliability monitoring, and governance for models, agents, and workflows in minutes, with no code changes.

Book a Demo ↗