AI Observability and FinOps for LLMs and AI Agents

Gain end-to-end cost visibility, performance monitoring, and governance for every model request, agent action, and multi-step agentic AI workflow across providers, models, agents, and users without changing a single line of code.

Gartner COOL VENDOR 2025

Scale AI Agents Without Losing Budget, Performance, or Control

Enterprises running hundreds of AI agents and thousands of daily LLM calls lack the visibility to trace execution paths, diagnose failures, and attribute costs. Revefi provides unified AI observability and FinOps across every model request, agent action, and multi-step agentic workflow, helping teams monitor performance, control token spend, and govern production AI from a single platform.

No End-to-End Traceability

Multi-step AI agent workflows can fail silently, leaving teams without a complete attribution chain from user to agent to model.

No Granular AI Cost Visibility

Token costs are fragmented across providers, models, agents, and users, making it difficult to attribute AI spend accurately down to the cent.

No AI Governance or Prompt-Level Accountability

Without policy guardrails, audit-ready execution trails, and prompt-level performance insights, teams cannot identify which prompts cause unreliable, low-quality, or inconsistent AI outputs.

What Revefi Delivers for AI Observability and FinOps?

End-to-End AI Attribution:

Trace every interaction from user to agent to model, with granular visibility into token cost, latency, execution paths, and outputs at each stage of the workflow.

Unified Multi-Provider Monitoring:

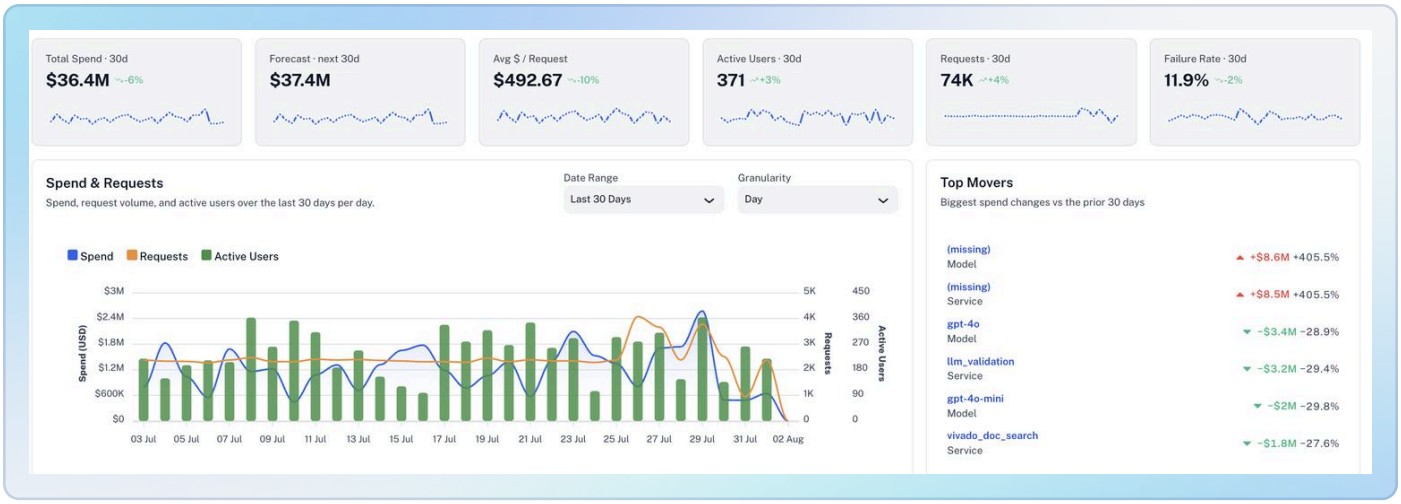
Monitor OpenAI, Anthropic, Google Gemini, and Vertex AI from a single AI observability platform, without stitching together separate vendor dashboards.

Reliable Cost Attribution:

Detects missing organization, service, model, user, or agent metadata before incomplete telemetry compromises AI cost allocation, showback, chargeback, or reporting accuracy.

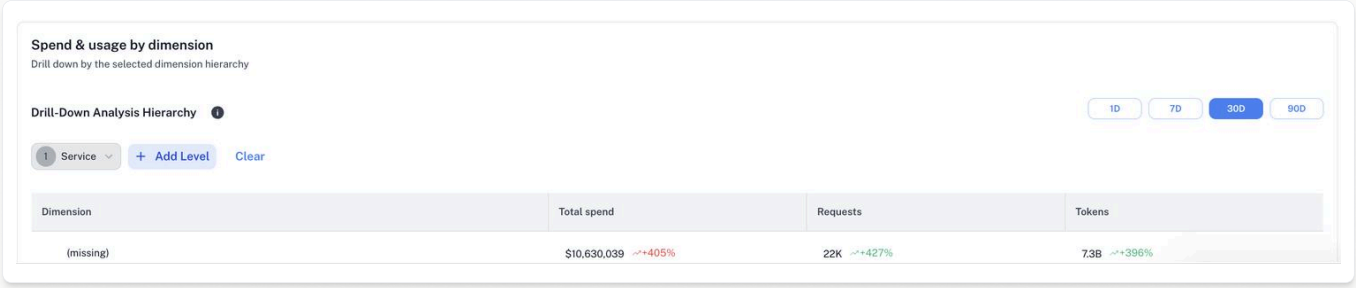
Rapid, Zero-Code Deployment:

Connect Revefi to your AI providers and endpoints to generate actionable cost, performance, and reliability insights in a few minutes.



AI Observability and Spend Intelligence

AI and engineering teams gain visibility into their LLM and agent deployments and operational rigor of FinOps, DataOps making your AI infrastructure fully observable, attributable, and audit-ready.



AI Agent & LLM Observability:
Monitor per-agent latency, request volume, prompt and response data, and token throughput in real time. Benchmark performance across GPT, Claude, and Gemini to identify slow models, agent bottlenecks, and inefficient AI workflows.

Token Economics:
Track input and output token volume alongside AI spend and historical trends to identify efficiency drift, uncover rising inference costs, and token consumption.

Audit & Governance:
Visibility, Traceability and Auditability into Agentic actions from day one.

Cost Outlier Detection:
Pinpoint which users, agents, or prompts are driving up costs and trace any anomaly to its driver.

AI Performance and Reliability Monitoring

Monitor every interaction (from individual LLM requests to complex, multi-step agentic AI workflows) with end-to-end tracing, performance insights, and audit-ready records that expose failures before they go undetected.

AI Cost Allocation, Adoption, and Governance

Attribute AI spend to the teams and business units generating it, measure adoption across users, services, and models, and enforce cost controls before usage translates into budget overruns.

Granular Cost Allocation and Adoption Insights:
Allocate AI costs across business units, cost centers, organizations, teams, and organizational hierarchies. Track adoption by user, service, model, and workload to understand where AI delivers value and where utilization remains low.

AI Guardrails and Governance:
Apply AI usage policies, budget thresholds, and spend controls across models and agentic workflows. Route cost anomaly alerts to the responsible team and detect missing attribution metadata before it compromises showback, chargeback, or financial reporting.

Prompt Performance and Cost Optimization

Identify high-latency prompts across models and agentic AI workflows, analyze prompt reuse to reduce redundant model calls and token spend, and correlate prompt-level patterns with output-quality metrics to pinpoint instructions that generate inconsistent, inaccurate, or low-quality results.

How Revefi Monitors and Optimizes AI and Agentic Workflows

Revefi connects to model providers and agents without complex integrations or application code changes. It captures model requests, agent actions, token consumption, latency, failures, and cost events at a granular level, providing real-time AI observability and FinOps visibility across providers, models, users, teams, and multi-step workflows.

Connect:

Link Revefi to your AI providers and model endpoints without deploying additional agents or modifying application code.

Monitor:

Continuously track every LLM call, agent step, token event, latency spike, failure, and cost signal across the AI stack.

Analyze:

Correlate AI cost, performance, reliability, and usage in one unified view across providers, models, agents, users, teams, and time periods.

Optimize:

Apply actionable recommendations to reduce token spend, improve slow prompts, control runaway agent workflows, eliminate redundant model calls, and improve output consistency and quality.

F200 Enterprise uncovers 7 figure savings on AI spend within 4 weeks

Case study: Unprecedented AI adoption and Spend sprawl

8×
Increase in AI spend in less than 4 months

25K+
Users and increasing

200M+
Requests monthly and increasing

10K+
AI services and increasing

Why Leading Companies Choose Revefi

SOC 2 TYPE II CERTIFIED

HIPAA COMPLIANT

FinOps FOUNDATION General Member

GDPR READY

CCPA READY

Organizations That Trust Revefi

AMD

Ingersoll Rand

StanleyBlack&Decker

Verisk

Cribl

Ocean Spray

The Princeton Review

Help at Home. Care to Live Your Life.

docker

Data Platform Integrations

snowflake

databricks

Google BigQuery

amazon REDSHIFT