THE PITCH

The operating discipline for enterprise agents

AgentOps by KPI Partners

AgentOps keeps artificial intelligence (AI) agents accurate and within budget as data, models and users change. **KPI Partners delivers it natively on Snowflake, Databricks, Amazon Web Services (AWS) or Microsoft Azure, with release gates, model routing, observability and governance in place from the first release.**

OUR READ

Agents pass the demo. Production is where they stall.

The model is seldom the culprit. What fails is the operating model around it: No release gate to stop a bad change, no ceiling on what a runaway loop can spend, and no record that explains why an agent answered the way it did. One question can become six or more billed model calls, and a model or data change can degrade answers with no release from you. KPI Partners puts that operating model in place, whether your first agent is in build or already live.

THE MOAT

Why AgentOps, and why KPI Partners to deliver it.

Why AgentOps

- **Nothing ships on a spot check**
  Every change to instructions, tools, retrieval or routing becomes a numbered version, is graded for correctness, groundedness and tool use, and goes live only after a named reviewer signs off. Rollback takes one step.
- **Cost controlled at the gateway**
  An AI gateway sends routine questions to small models, answers repeats from a cache that respects each user's access, attributes spend to each step and enforces a soft alert and a hard cap.
- **Every answer can be explained**
  Each request is traced end to end with its model, tools, cost and version, and every tool call and data access lands in an audit table.

Why KPI Partners for AgentOps

- **Native to the platform you run**
  Our reference implementations, accelerators and evaluation harnesses run natively on **Snowflake**, **Databricks**, **AWS** and **Microsoft Azure**, so agents, traces and evaluations stay next to your data.
- **A tested model routing design**
  Our routing reference implementation spans nine models in two pools, with a rule-based design an auditor can follow and a dynamic design for compound questions.
- **Experts review only the hard cases**
  Low-rated conversations become new test questions, your experts label only those, and the aligned judge grades every release without taking more of their time.

ROUTING, MEASURED

What the control plane changes, on our reference implementation

4%
Of spend, for 36% of questions
The small-model pool answers routine lookups and definitions, so only questions that need reasoning reach the large pool.

1 in 5
Answered from cache
Repeat questions skip the model call entirely. The cache key carries each user's access scope, so savings never leak data.

~1 sec
For a routine answer
Small-model questions return in about a second, against six to nine seconds for large-model questions.

4.5 / 5
Faithfulness, judge scored
A large language model (LLM) judge scores every answer; Low scorers become the next release's test cases.

Platform services referenced follow each vendor's catalogue as of October 2026. Figures come from KPI Partners case studies and AgentOps reference implementations.

REFERENCE ARCHITECTURE

AgentOps at a glance, and where KPI Partners engages

DIAGRAM AgentOps reference architecture

☐ Platform component ☐ KPI Partners role

EXPERIENCE	Chat surfaces	Embedded apps	API and MCP	Single sign-on	**Deliver** We connect governed agents to the channels your teams already use.
AGENT RUNTIME	Supervisor	Domain sub-agents	Tools	State store	**Build** We build the agent on governed data, semantic definitions and native tools.
AI GATEWAY AND MODELS	Guardrails	Model routing	Semantic cache	Budgets	Model pools	**Define** We define routing, cache scope and the spend caps in front of every call.
DATA AND KNOWLEDGE	Governed tables	Semantic layer	Hybrid index	Docs and ontology	**Assess** We choose the use case, agree its test questions and set a baseline.
GOVERN	Release gate	Tracing	Access and audit	**Validate** We gate every release on scored evaluations and a named sign-off.

API: Application programming interface. MCP: Model Context Protocol. Each layer maps to native services on Snowflake, Databricks, AWS or Microsoft Azure; Scope is tailored to each client.

HOW KPI PARTNERS ENGAGES

Start with one agent, prove the operating cycle, then scale.

1 — Scope the use case
Choose one use case, agree its key performance indicators (KPIs) and test questions, and set a baseline.

2 — Build the agent
Build on governed data, semantic definitions and the native tools of your platform.

3 — Wire the gate
Add the evaluation gate, routing layer and cost caps, and measure how they behave.

4 — Run and hand over
Settle the architecture on evidence, switch on alerts and hand a runbook to your engineering team.

A first engagement runs for about twelve weeks.

FROM OUR AGENTIC AI WORK

CASE STUDY — Read the case study ›

An AI-powered Oracle enterprise resource planning (ERP) chatbot for a leading U.S. beverage manufacturer

85% Faster request resolution

70% Higher system engagement

30% Lower operational costs

OUR PERSPECTIVE

AgentOps: The Discipline Behind Agents That Survive Production
The release cycle, model routing design and platform service map behind AgentOps.

Governing and Securing the New Power User: Why AI Agents Need Runtime Control
Why agents need runtime control over actions, cost and data before they scale.

AWS DevOps Competency

300+ Enterprises trust KPI Partners

4 platforms Snowflake, Databricks, AWS and Microsoft Azure

Start with an AgentOps Assessment

An assessment shows where cost, quality and governance risk sit in the agents you run today, and agrees the first agent to put on a gated release. It is far cheaper to build the operating model before go-live than after the first surprising invoice.

Talk to your KPI Partners account team, or reach us at info@kpipartners.com and +1 510-818-9480.

Talk to an AgentOps Expert >

Platform services referenced follow each vendor's catalogue as of October 2026. Figures come from KPI Partners case studies and AgentOps reference implementations.