Surveil

PRACTITIONER PLAYBOOK: FINOPS FOR AI

AI Cost Optimization Checklist

A practical FinOps playbook for managing AI spend across providers, models, licenses, agents and infrastructure.

78 checks across 8 areas
Aligned to the FinOps Framework

surveil.co

Why this checklist matters

AI costs rarely live in one place.

They can span cloud infrastructure, model APIs, enterprise copilots, SaaS applications, licenses, agents, data services and multiple providers. As adoption grows, so does the number of teams influencing what is used, what it costs and whether it creates value. That makes AI cost optimization a shared discipline, not a Finance, FinOps or Engineering task in isolation.

Use this checklist to test whether your organization can see the full AI cost picture, assign ownership, plan and forecast accurately, identify credible optimization opportunities, govern action, and validate that savings do not come at the expense of business value.

> **The goal is not to spend less on AI.** It is to spend intentionally and know what you are getting in return.

Apply rigor where it matters most. Not every AI workload needs the same level of financial control. Increase oversight as cost, scale, risk and business impact increase.

FinOps Framework alignment

This playbook applies the FinOps Framework to the AI Technology Category and spans multiple FinOps Capabilities. It is a practical assessment and action guide for organizations evaluating how effectively they manage AI cost, usage, accountability, optimization, governance and value.

FinOps for AI addresses the financial and operational complexity of AI consumption across models, infrastructure, SaaS applications, enterprise licenses, agents, data services and multiple providers. AI can introduce highly granular usage, variable pricing, rapid changes in consumption, shared technical dependencies and new measures of value. These characteristics can require organizations to adapt how established FinOps Capabilities are performed.

This playbook does not create a separate operating model for AI. It applies established FinOps practices to the characteristics of AI spending and consumption.

Framework elements addressed

Technology category
FinOps for AI

Domains
Understand Usage and Cost; Quantify Business Value; Optimize Usage and Cost; Manage the FinOps Practice

Capabilities
Data Ingestion; Allocation; Invoicing and Chargeback; Reporting and Analytics; Anomaly Management; Forecasting; Budgeting; Planning and Estimating; Unit Economics; KPIs and Benchmarking; Usage Optimization; Rate Optimization; Licensing and SaaS; Governance, Policy and Risk; FinOps Assessment; Architecting and Workload Placement

Not every organization needs the same maturity across every Capability. Apply the controls, measurement and rigor that create sufficient business value for the scale, cost, risk and criticality of the AI workloads in scope.

How this playbook uses the Framework

The checklist is organized around the practical outcomes an organization needs to manage AI investment effectively. Each section applies one or more FinOps Capabilities to AI. The sections are intentionally interconnected: an optimization opportunity may depend on accurate ingestion and allocation, need an accountable owner, trigger an engineering action, and only become meaningful once its financial and business impact is validated.

Who should use this playbook

This playbook is for FinOps practitioners and cross-functional teams responsible for understanding, planning, optimizing, governing or demonstrating the value of material AI spending. It applies whether you manage AI consumption across public cloud services, model providers, enterprise AI applications, SaaS AI licenses, agents, infrastructure, or a combination of these.

It is most useful when you need to assess current FinOps for AI practices, identify gaps, prioritize improvements, or establish a repeatable review process as AI adoption and financial exposure grow.

FinOps persona	Role in this playbook	Typical participation
FinOps Practitioner	Coordinates visibility, allocation, optimization, forecasting, measurement and realized value	Driver
Engineering	Provides model, agent, infrastructure, architecture, usage and technical optimization context	Contributor / Decider
Finance	Supports budgets, forecasts, financial controls, allocation and validation of financial outcomes	Contributor / Decider
Product	Connects AI investment to use cases, adoption, product outcomes and business value	Contributor / Decider
Procurement	Evaluates licensing, provider agreements, commitments, commercial models and renewals	Contributor / Decider
Leadership	Sets investment priorities, risk tolerance, accountability and expectations for business outcomes	Decider / Informed

Allied personas such as Security, ITAM, ITFM and ITSM may participate where governance, licensing, financial management, service management or risk requirements intersect with AI spending. A persona may be Driver, Decider, Contributor or Informed for different checklist activities.

FinOps Principles applied to AI cost optimization

- **Teams need to collaborate.** AI cost decisions frequently cross Finance, Engineering, Product, Procurement, Technology Leadership and the FinOps team. Effective optimization requires these groups to decide together rather than treating cost, architecture, usage and business value as separate concerns.
- **Business value drives technology decisions.** Lower cost does not automatically mean better optimization. Decisions should weigh cost alongside quality, performance, reliability, speed, risk and the intended business outcome. Higher AI spend may be appropriate when it produces greater value; inexpensive consumption may still be waste if it produces little useful outcome.
- **Everyone takes ownership for their technology usage.** Material AI spend should have an accountable owner, extending beyond the invoice to the application, product, workflow, agent or use case responsible for the consumption.
- **FinOps data should be accessible, timely and accurate.** AI consumption can change quickly. Cost, usage, allocation, anomaly and value information should be available at a cadence that supports timely decisions and appropriate controls.
- **FinOps should be enabled centrally.** Central FinOps practices provide consistent allocation methods, reporting standards, optimization processes, governance controls, commercial analysis and measurement, while individual teams stay accountable for the AI they use.
- **Take advantage of variable technology cost models.** AI may combine consumption-based model pricing, cloud infrastructure, provisioned capacity, enterprise licenses, commitments and other commercial models. Evaluate

the economics of each, and adjust purchasing, architecture and capacity decisions as usage becomes better understood.

Prerequisites and information required

You do not need a mature FinOps for AI practice before using this playbook. The assessment is more effective when you have:

- Access to material AI cost and billing data for the scope being assessed
- Enough usage information to understand major providers, services, models, licenses, infrastructure or agents
- A defined scope, such as the enterprise AI portfolio, a business unit, product, application, provider or environment
- Representatives who can provide financial, technical, commercial and business context
- Existing budgets, forecasts, ownership data, allocation rules, optimization recommendations or governance policies
- Agreement on what counts as material AI spend or risk

Useful information may also include model, token, API or inference usage; AI application, agent and service inventories; license entitlement and utilization; GPU, compute and storage utilization; commitments and commercial agreements; anomaly thresholds and escalation procedures; AI governance and risk policies; and business outcome measures for material AI initiatives.

Where prerequisites are missing, treat the gap as an assessment finding rather than a reason not to run the playbook.

Applying the FinOps Phases

FinOps activities are performed iteratively through the Inform, Optimize and Operate phases. This checklist supports that cycle rather than treating AI cost optimization as a one-time exercise.

Inform

What AI services and workloads are used, what they cost, who owns them, how consumption is changing, whether costs are allocated appropriately, whether spend is within budget and forecast, and what unit-cost or value measures exist.

Optimize

Evaluate opportunities and tradeoffs across models and tokens, agent behavior, architecture, infrastructure and capacity, licensing, commitments, commercial models, provider selection, waste and business value. Find the best economic decision, not simply the lowest cost.

Operate

Assign owners, implement approved actions, apply governance and financial controls, monitor anomalies, track completion, validate realized impact, measure business outcomes, and update policies, forecasts and assumptions.

The cycle then repeats as AI usage, pricing, architecture and business priorities change.

Applying FinOps maturity

Mature AI cost management only where greater maturity creates sufficient business value. Not every checklist item needs to operate at Run.

Crawl

Foundational visibility and accountability

- Major AI spend sources are identifiable
- Material workloads have basic ownership
- Primary providers and licenses are in reporting
- Obvious waste and utilization issues are reviewed
- Basic budgets and alerts exist
- Optimization may be manual
- Financial outcomes are measured selectively

Walk

Repeatable and cross-functional

- Allocation follows documented methods
- Forecasts include major workload and adoption drivers
- Anomaly detection and escalation are established
- Model, token, agent, infrastructure and license efficiency are reviewed systematically
- Optimization actions have owners and statuses
- Unit-cost measures are used for material workloads
- Realized savings are validated
- Governance is documented and repeatable

Run

Integrated, adaptive and business-aligned

- AI costs are consistently connected to business ownership and outcomes
- Forecasts adapt rapidly to material changes
- Appropriate anomaly and containment controls are automated
- Model and workload economics influence architecture and product decisions
- Unit economics extend toward business outcomes
- Optimization and commercial decisions use continuous evidence
- Realized value informs future investment

The right maturity level depends on business value, workload criticality, financial exposure, organizational complexity and risk.

How to run this playbook

Use the checklist as a repeatable assessment of your current FinOps for AI practices, for the enterprise AI portfolio or a defined scope such as a business unit, product, provider, cost center or environment. For each item:

1. Decide whether the activity is relevant to the AI workloads and spending in scope.
2. Identify the accountable persona or team.
3. Assess whether the activity is performed with sufficient consistency and quality.
4. Record gaps, exceptions or evidence needing investigation.
5. Prioritize improvement by financial impact, confidence, effort, operational risk and business value impact.

6. Assign an action, owner and expected outcome where improvement is warranted.
7. Reassess as AI usage, architecture, pricing, providers or business priorities change.

Owner labels name the functional team typically closest to each activity, and may be more specific than the FinOps persona categories above.

SECTION 1

Unified AI Cost Visibility

A fragmented view creates fragmented decisions.

	Checklist	Owner
☐	We can identify material AI spend across providers, services, licenses, and infrastructure.	FinOps
☐	We can view AI spend across multiple providers in a consolidated financial view.	FinOps
☐	We can separate model/API, infrastructure, license, and supporting-service costs where needed.	FinOps
☐	We can distinguish production from development, testing, and experimentation.	FinOps / Engineering
☐	We know our largest AI cost drivers.	FinOps
☐	We know our fastest-growing AI cost drivers.	FinOps
☐	We can see when spend shifts from one provider, model, or service to another.	FinOps
☐	Leadership can understand total AI spend without reconciling multiple reports manually.	FinOps / Finance
☐	We can calculate the full cost of a material AI application, agent, workflow, or initiative, not only its model or token cost.	FinOps / Finance

Watch for: Provider silos, hidden SaaS AI spend, duplicated services, cost migration, and AI costs buried inside broader cloud budgets.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 2

Ownership and Allocation

If nobody owns the spend, nobody owns the decision.

	Checklist	Owner
☐	Material AI spend has an identifiable owner.	FinOps / Business
☐	AI workloads map to a product, application, project, or use case.	FinOps / Engineering
☐	AI costs map to business units or cost centers where appropriate.	FinOps / Finance
☐	Production agents have an accountable owner.	Product / Engineering
☐	Shared AI costs have documented allocation rules.	FinOps / Finance
☐	We measure the percentage of material AI spend allocated to an accountable owner, business unit, product, application, or use case.	FinOps
☐	Showback or chargeback is used where it improves accountability.	FinOps / Finance

Watch for: Unattributed spend, shared services with no allocation method, agent sprawl, and experimental costs blending into production.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 3

Budgeting, Forecasting, and Alerts

Forecast the portfolio, not just individual provider bills.

	Checklist	Owner
☐	AI budgets account for all material providers and service types.	Finance / FinOps
☐	Anomaly detection is configured to identify unexpected or rapidly accelerating AI spend.	FinOps / Engineering
☐	High-risk AI workloads have predefined emergency spend controls, such as throttling, consumption caps, workload suspension, or automated shutdown where appropriate.	FinOps / Engineering
☐	Forecasts include expected consumption growth.	Finance / FinOps
☐	Forecasts account for model, provider, license, or capacity changes.	FinOps / Engineering
☐	Production and experimental spend are forecast separately.	FinOps
☐	Planned agent, workload, or license expansions are included in forecasts.	Product / FinOps
☐	Budget versus actual AI spend is reviewed.	Finance / FinOps
☐	We track AI forecast accuracy at a level appropriate to the portfolio, provider, business unit, or material workload.	FinOps
☐	Spend thresholds and escalation paths are defined.	Finance / FinOps
☐	Portfolio-level alerts can identify cumulative AI spend growth across providers.	FinOps

Watch for: One provider looking stable while total AI spend accelerates elsewhere.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 4

Model, Token, and Agent Efficiency

Cheap AI is not necessarily efficient AI. Optimize cost against required outcome.

	Checklist	Owner
☐	Premium models are used only where their capabilities are justified.	AI Engineering
☐	Lower-cost models are evaluated for routine workloads.	AI Engineering
☐	Model routing is considered where workload complexity varies.	AI Engineering
☐	Token consumption and cost are measured where material, including input/output usage and changes over time.	Engineering / FinOps
☐	Large context windows and unnecessary prompt content are reviewed.	AI Engineering
☐	Duplicate calls, retries, and regeneration cycles are monitored.	AI Engineering
☐	Agent runs, loops, handoffs, and tool calls are reviewed where they materially affect cost.	AI Engineering
☐	Abandoned or duplicate agents are reviewed.	Product / Engineering
☐	Cost per interaction, workflow, or completed task is measured where practical.	FinOps / Product
☐	Optimization decisions include a quality and business-value check.	Product / Engineering
☐	Agent loops, retry behavior, regeneration cycles, and abnormal model-call volume are monitored for signs of uncontrolled consumption.	AI Engineering

Watch for: Token growth without workload growth, excessive retries, premium-model overuse, and complex agent flows that add cost without improving outcomes.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 5

Infrastructure and Capacity

AI infrastructure should match actual demand.

	Checklist	Owner
☐	GPU and compute utilization are monitored.	Infrastructure
☐	Idle AI infrastructure is identified.	Infrastructure / FinOps
☐	Nonproduction environments are reviewed regularly.	Infrastructure
☐	Autoscaling is used where appropriate.	Infrastructure
☐	Provisioned capacity is compared with actual utilization.	Infrastructure / FinOps
☐	Commitment or reserved capacity is aligned to stable baseline demand.	FinOps / Procurement
☐	Serverless, batch, spot, or interruptible options are evaluated where appropriate.	Architecture
☐	Supporting storage, data transfer, and observability costs are included when material.	FinOps / Infrastructure
☐	We evaluate whether workloads are using the most appropriate consumption, provisioned-capacity, reservation, or commitment model.	FinOps / Procurement

Watch for: Idle GPUs, overprovisioning, always-on development environments, and unused committed capacity.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 6

Licensing, Commitments, and Commercial Economics

Usage is evidence. Value is the decision.

	Checklist	Owner
☐	AI licenses are compared with actual usage.	FinOps / ITAM
☐	Inactive or low-use licenses are reviewed.	ITAM / FinOps
☐	License allocation reflects role, need, or expected value.	Business / ITAM
☐	Adoption is measured by team or business unit where useful.	Product / FinOps
☐	Cost per active AI user is understood where applicable.	FinOps
☐	Overlapping AI subscriptions or providers are reviewed.	Procurement
☐	Consumption commitments are compared with actual demand.	Procurement / FinOps
☐	Renewal decisions are supported by usage and value evidence.	Procurement / Business

Watch for: Inactive seats, overlapping AI tools, excess commitments, and renewals based on entitlement rather than evidence.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 7

Governance and Accountable Execution

Controls should reduce risk without creating unnecessary cost or friction.

	Checklist	Owner
☐	Material AI services and workloads have defined owners.	Leadership / Product
☐	Shadow or unauthorized AI usage is monitored where feasible.	Security / IT
☐	High-cost AI consumption can be identified quickly.	FinOps
☐	Budget guardrails exist for material workloads.	Finance / FinOps
☐	Agent consumption limits are used where appropriate.	Engineering / Risk
☐	Access to expensive models is governed where needed.	Engineering / Risk
☐	Alerts and emergency containment actions have defined thresholds, owners, escalation paths, and recovery procedures.	FinOps / Engineering
☐	Emergency spend controls are proportionate to workload criticality and business risk.	Engineering / Risk
☐	Controls are proportionate to business and regulatory risk.	Risk / Security
☐	Provider or model changes are reviewed for privacy, security, and compliance impact.	Risk / Engineering
☐	Exceptions are documented.	Risk / FinOps

Watch for: Shadow AI, unlimited agent consumption, weak ownership, and controls that are either too loose or unnecessarily restrictive.

GAPS FOUND	ACTIONS AND OWNERS

SECTION 8

Realized Value

Do not report theoretical savings as results.

	Checklist	Owner
☐	We track the percentage of identified optimization value that is implemented and realized.	FinOps
☐	Material optimization actions have a documented baseline.	FinOps
☐	Material AI workloads have a meaningful unit-cost or cost-to-outcome measure where practical.	FinOps
☐	Expected financial impact is recorded before implementation.	FinOps
☐	Actual financial impact is measured after implementation.	FinOps / Finance
☐	Quality, reliability, or latency are checked where relevant.	Engineering
☐	Business outcomes are checked after material changes.	Product / Business
☐	Savings are reviewed for sustainability.	FinOps
☐	Changes that reduce business value are adjusted or reversed.	Product / Engineering
☐	Lessons learned feed back into future policy, architecture, or planning.	FinOps / Engineering
☐	Material optimization recommendations have an owner, decision, and completion status.	FinOps
☐	We can trace material recommendations from identification through implementation and financial validation.	FinOps / Finance

Watch for: Estimated savings being treated as realized value, one-time savings being reported indefinitely, and cost reductions that degrade the outcome.

GAPS FOUND	ACTIONS AND OWNERS

Suggested KPIs

Use measures appropriate to your environment. The FinOps Foundation KPI Library provides a community-curated collection of FinOps KPIs, organized by Capability, with definitions, formulas, examples and data sources. Common AI cost optimization measures include:

Total AI spend | **AI spend growth rate** | **Forecast accuracy / variance** | **Percentage of AI spend allocated** | **Token consumption and cost** | **Cost per interaction / workflow / agent run** | **Resource utilization** | **License utilization** | **Commitment utilization** | **Identified optimization value** | **Realized savings** | **Percentage of optimization value realized** | **Cost per business outcome**

Select KPIs that inform a decision or action. More metrics do not necessarily create better FinOps outcomes.

Keep KPI logic current. Review AI cost and efficiency metrics as pricing models, services, workload patterns and optimization methods evolve. When calculation logic changes, document it, preserve a consistent baseline where needed, and don't compare periods without explaining the change.

Prioritize what you find

Not every opportunity deserves action. Score each material opportunity against five factors:

Factor	Question
Financial impact	Is the opportunity material?
Confidence	How strong is the evidence?
Effort	What will it take to implement?
Operational risk	Could the change affect performance, reliability, security or quality?
Business value impact	Could the change reduce the outcome the AI investment is meant to create?

Act

Strong evidence, meaningful value, acceptable risk.

Investigate

Promising, but more evidence is needed.

Monitor

Not material enough to act on yet.

Accept

The cost is intentional and justified.

Accept is a valid optimization decision. FinOps is about maximizing value, not minimizing spend.

Suggested review cadence

Weekly

Major anomalies and rapidly growing workloads.

Monthly

Spend, forecast, attribution, recommendations, optimization actions, realized savings.

Quarterly

Provider mix, licensing, capacity, unit economics, business value.

Event-driven

Major model change, new provider, agent launch, renewal, pricing change, large rollout.

What good looks like: can you answer these questions?

- [ ] **What are we spending on AI across the enterprise?**
- [ ] **Where is the spend coming from?**
- [ ] **Who owns it?**
- [ ] **What percentage of material AI spend is allocated?**
- [ ] **Is spend within budget and forecast?**
- [ ] **What changed, and why?**
- [ ] **Where are the credible optimization opportunities?**
- [ ] **Who should act?**
- [ ] **What percentage of identified value was actually realized?**
- [ ] **What does this workload cost per meaningful unit of work?**
- [ ] **Did optimization preserve or improve business value?**

> If those answers require weeks of reconciliation across providers, spreadsheets, teams and systems, the FinOps for AI practice still has work to do.

Expected outcomes and indicators of success

After completing the playbook, you should understand:

- Where material AI spend exists
- Which AI costs can and cannot be attributed
- Who owns material AI consumption
- Whether budgets and forecasts reflect expected AI demand
- Where credible optimization opportunities exist

Indicators of success

- More material AI spend visible and categorized
- More material AI spend allocated to accountable owners
- Improved forecast accuracy where forecasting is appropriate
- Less time to identify and explain material AI cost anomalies

- Whether models, tokens, agents, infrastructure, licenses and commitments are used efficiently
- Whether appropriate governance and anomaly controls exist
- Which optimization actions warrant implementation
- Whether implemented actions produced measurable financial value
- Whether cost improvements preserved the intended business outcome
- Which FinOps Capabilities to mature next

The output should be a prioritized improvement path, not a maturity score for its own sake.

- More optimization actions with defined owners and decisions
- More identified optimization value implemented and financially validated
- Better utilization of material licenses, infrastructure and commitments
- More use of meaningful unit-cost and cost-to-outcome measures
- Less recurrence of previously resolved waste or governance issues
- Better ability to explain AI spending and value to stakeholders

Related FinOps Framework resources

Use this playbook alongside the relevant FinOps Framework guidance, particularly:

FinOps for AI Technology Category **FinOps Principles** **FinOps Personas** **FinOps Phases** **FinOps Maturity Model** **FinOps Assessment** **Allocation** **Anomaly Management** **Budgeting** **Forecasting** **Unit Economics** **Usage Optimization** **Rate Optimization** **Licensing and SaaS** **Governance, Policy and Risk** **KPIs and Benchmarking** **Automation, Tools and Services**

HOW SURVEIL SUPPORTS THIS CHECKLIST

Can you answer these questions today? Surveil can, in less than a week.

Surveil for AI applies the same ownership, allocation and financial controls you already run for cloud and Microsoft 365 to Azure AI, OpenAI, Claude, Microsoft 365 Copilot and agents.

Checklist section	How Surveil helps	Go deeper
1. Unified AI cost visibility	AI spend across Azure AI, OpenAI, Claude, Copilot and agents in one view, by model and deployment.	Surveil for AI
2. Ownership and allocation	Smart Tags apply one business hierarchy across cloud, AI and Microsoft 365, with AI chargeback.	AI Cost Allocation and Accountability Playbook
3. Budgeting, forecasting and alerts	Budgets, forecasts and AI anomaly detection with alerts.	AI Budgeting and Forecasting Under Volatility Playbook
4. Model, token and agent efficiency	Azure AI, OpenAI and Claude cost by model and deployment, to spot premium-model overuse and runaway consumption.	AI Cost Anomaly Response and Financial ContainmentPlaybook
5–6. Capacity, licensing and commitments	PTU commitment modeling, Copilot seat usage and reclaim, and license planning before renewal.	Managing AI Spend Across Consumption Models Playbook
7. Governance and accountable execution	Shadow AI Discovery, and Microsoft 365 agents with creators, sponsors and license cost.	Surveil for AI
8. Realized value	Copilot and Claude adoption with estimated time saved, set against seat cost.	Copilot case study: 2x+ adoption, ~$500K optimized

Get a free AI Estate Assessment

A read-only readout of your AI spend, owners and commitments in less than a week. No software to install, no obligation.

Get your assessment