FEATURED RESOURCE 03

AI Spend Governance Checklist

A practical FinOps checklist for LLM, GPU, AI API and data-platform spend before experimentation becomes uncontrolled cost.

Visibility | Allocation | Governance | Optimization

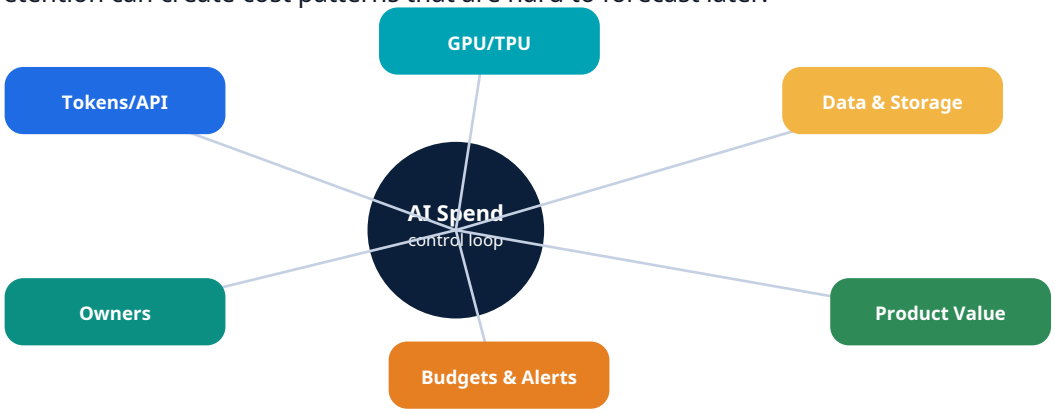
Vikassh Dalal

DevOps-led FinOps | AWS | Azure | Hybrid | AI Spend Governance

vikasshdalal.com | vikassh.dalal@gmail.com

AI cost needs governance early

AI spend can scale differently from traditional infrastructure. Tokens, API calls, GPUs, experiments, data movement and retention can create cost patterns that are hard to forecast later.



The goal is not to slow AI work. The goal is to make AI usage visible, owned, forecastable and connected to business value before spend becomes noisy.

Grounding: FinOps Foundation Framework and phases; FinOps for AI guidance; FOCUS billing specification concepts; AWS Well-Architected, Azure Well-Architected, and Google Cloud Well-Architected cost guidance. Use this as a discussion tool, not as a substitute for environment-specific analysis.

AI spend sources to separate

Do not put all AI cost into one bucket.

Area	What to check	Decision to make
LLM/API usage	Input/output tokens, API calls, model tier, retry patterns, prompt length, tool calls.	Map to product, feature, customer, experiment or owner.
GPU/TPU compute	Training, fine-tuning, inference, idle capacity, reservation/commitment use.	Measure utilization and schedule non-production workloads.
Data pipelines	Ingestion, transformation, vectorization, retrieval, ETL/ELT, streaming.	Track the data path, not only the model endpoint.
Storage/retention	Training sets, embeddings, logs, prompts, outputs, audit data, backups.	Define retention tiers and compliance requirements.
Observability	Tracing, logs, evaluation data, quality monitoring, drift monitoring.	Balance visibility with retention and volume controls.
Licensing/tools	AI platforms, model providers, datasets, vendor contracts, evaluation tools.	Track vendor usage and renewal risk.

Metrics that make AI cost manageable

Pick a small set of metrics. Avoid a dashboard that nobody operates.

Metric	What it proves	Owner
Cost per token	Token-based usage is visible and can be linked to prompt design, model selection and product usage.	Product / Eng
Cost per API call	Managed AI service usage is measurable and comparable across products or workflows.	Engineering
GPU utilization	Provisioned AI hardware is not sitting idle or oversized for workload needs.	Platform
Cost per customer/use case	Spend is tied to tenant, customer, workflow, feature or business outcome.	Product
Anomaly cost impact	Unexpected AI usage spikes are detected and routed to an accountable owner.	FinOps
Time to value	AI experiments have a costed path from proof of concept to production value.	Leadership

Useful AI FinOps metric rule: if a metric cannot drive a decision, remove it or move it to a technical drill-down view.

Governance controls to add early

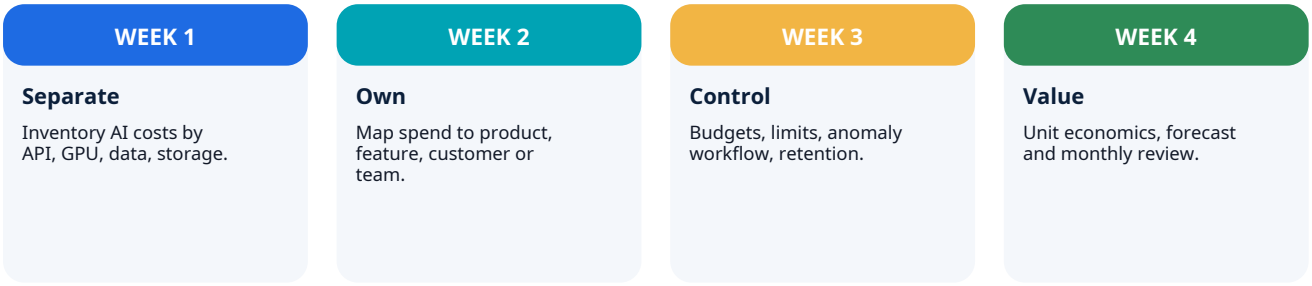
These controls can be introduced without blocking experimentation.

Area	What to check	Decision to make
Budget boundaries	Set experiment, team, project and production budgets.	No AI work runs without a known owner and budget path.
Usage limits	Apply token, API, concurrency, GPU quota or environment limits.	Prevent runaway usage and retry loops.
Model routing	Match model quality and cost to the task.	Avoid using expensive models for low-complexity tasks.
Caching/reuse	Cache repeat responses, embeddings and common retrieval outputs where appropriate.	Reduce duplicate token/API spend.
Environment separation	Separate proof-of-concept, test and production usage.	Do not let experiments hide inside production cost.
Data retention	Define how long prompts, outputs, embeddings and training data are kept.	Control storage cost and compliance exposure.
Anomaly workflow	Route cost spikes to the team that can investigate.	Alerts should create action, not noise.
Vendor review	Track pricing model, contract exposure, data terms and usage commitments.	Avoid surprise renewals or commitment mismatch.

AI FinOps operating model

AI spend decisions need engineering, product, finance and governance in the same loop.

Role	AI spend responsibility
Product	Defines use case, customer value, pricing assumptions and success metrics.
Engineering	Owns architecture, model selection, API usage, retries, observability and implementation controls.
Platform	Manages GPU/compute capacity, identity, environments, deployment and cost guardrails.
Finance	Owns budget, forecast, unit economics, variance review and commitment risk.
Security/compliance	Reviews data exposure, retention, residency, audit and vendor terms.
FinOps	Connects cost data, owners, operating cadence, metrics and action backlog.



Use this as a service offer

This PDF supports a paid AI Spend Governance Review.

Item	Details
Best fit	Teams experimenting with LLMs, AI APIs, GPUs, data platforms or AI-enabled products where cost ownership is unclear.
Timeline	1-3 weeks depending on environment access and AI workload complexity.
Deliverables	AI spend map, owner model, metric set, control gaps, anomaly workflow, 30-day governance plan.
Follow-on	AI cost dashboard, model routing review, token/API controls, GPU utilization review or monthly advisory.

For AI spend governance support: vikasshdalal.com | vikassh.dalal@gmail.com