International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 6, Nov – Dec 2026
Available at www.ijsred.com
RESEARCH ARTICLE OPEN ACCESS

Enterprise AI FinOps and Cost Optimization Framework

A Comprehensive Enterprise Framework for Governing AI Spend, Model Economics, Capacity, Unit Cost, Value Realization, and Sustainable AI Operations

Sanjeeve Kumar Gajadi*
*(HCL Technologies, Bengaluru, Karnataka, India)
Email: sanjeevekumar.g@hcltech.com

****

Abstract:
Enterprise AI introduces rapidly changing cost structures driven by model inference, training, retrieval, vector storage, data processing, GPUs, APIs, agent tool use, and third-party services. Traditional cloud cost management is insufficient because AI economics depend on model choice, prompt design, context length, workload patterns, accuracy requirements, and business value. This paper proposes an Enterprise AI FinOps and Cost Optimization Framework that aligns finance, engineering, architecture, data, AI governance, and business ownership. The framework defines cost allocation, tagging, model unit economics, showback and chargeback, budgets, forecasts, anomaly detection, capacity planning, optimization levers, value metrics, and governance controls. It introduces workload-aware model routing, caching, batching, prompt optimization, retrieval efficiency, lifecycle policies, and risk-aware cost tradeoffs. The proposed approach enables enterprises to control AI expenditure while preserving performance, quality, security, and business outcomes.

Keywords — AI FinOps, AI Cost Optimization, Generative AI, Model Economics, Cloud FinOps, GPU Cost, Agentic AI, Unit Cost, AI Governance, Enterprise AI

****

I. INTRODUCTION

AI workloads can create volatile expenditure because usage scales with tokens, requests, context size, model class, retrieval, compute, and downstream tool activity.

AI FinOps extends traditional FinOps by connecting technical consumption with model quality, business outcome, and risk.

II. COST MODEL

The cost model should capture training, inference, storage, retrieval, vector databases, orchestration, data transfer, observability, evaluation, security, and external API costs.

Shared-platform costs should be allocated using transparent drivers such as requests, tokens, GPU-hours, storage, or business transactions.

III. TAGGING AND COST ALLOCATION

AI services should be tagged by business unit, use case, model, environment, owner, risk tier, and product.

Reliable allocation enables showback, chargeback, budgeting, forecasting, and accountability.

IV. AI UNIT ECONOMICS

Unit metrics may include cost per request, cost per successful task, cost per generated document, cost per resolved case, or cost per autonomous transaction.

Business value should be compared with unit cost so that optimization does not sacrifice outcomes that justify higher-quality models.

V. MODEL SELECTION AND ROUTING

Workloads can be routed across models according to quality, latency, risk, and cost requirements.

Smaller or specialized models may handle routine tasks while premium models are reserved for complex or high-value scenarios.

 ©IJSRED: All Rights are Reserved

VI. PROMPT AND CONTEXT OPTIMIZATION

Prompt length, retrieved context, conversation history, and tool results directly affect cost.

Context compression, relevant retrieval, response limits, and reuse of validated prompt components can reduce spend without materially reducing quality.

VII. CACHING, BATCHING AND REUSE

Semantic caching, response reuse, embedding reuse, batching, and scheduled processing can improve cost efficiency.

Caching policies must consider freshness, privacy, authorization, and the risk of reusing stale or inappropriate outputs.

VIII. AGENTIC AI COST CONTROLS

Agents can create hidden cost through planning loops, repeated tool calls, delegation, retries, and long-running tasks.

Budgets, step limits, token limits, tool quotas, and cost-aware orchestration should be enforced at runtime.

IX. CAPACITY AND FORECASTING

Forecasting should incorporate user growth, model changes, seasonality, autonomous task volume, and workload migration.

Capacity planning should balance reserved resources, on-demand services, accelerator availability, and provider concentration risk.

X. COST ANOMALY DETECTION

Anomaly detection should identify unexpected token spikes, runaway agent loops, repeated failed requests, excessive retrieval, or unexpected GPU usage.

Alerts should route to technical and financial owners with enough context to support remediation.

XI. GOVERNANCE AND VALUE REALIZATION

Budgets and optimization targets should be linked to use-case owners, service levels, risk requirements, and value metrics.

Executive reporting should combine spend, unit cost, savings, quality, adoption, and realized business value.

XII. CONCLUSION

The Enterprise AI FinOps and Cost Optimization Framework provides a practical model for governing AI economics across architecture, finance, and operations.

By linking cost to quality, risk, and business value, organizations can scale AI sustainably while maintaining accountability and performance.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance specialists, security professionals, standards bodies, and technology researchers whose published work informed the concepts synthesized in this paper.

REFERENCES

[1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
[2] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
[3] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
[4] ISO 31000:2018, Risk management — Guidelines.
[5] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
[6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
[7] The Open Group, TOGAF Standard, 10th Edition, 2022.
[8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
[9] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
[10] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
[11] OECD, Recommendation of the Council on Artificial Intelligence.
[12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.

 ©IJSRED: All Rights are Reserved