Infosys topaz

WHITE PAPER

MAXIMIZING ROI ON AI: BEST PRACTICES FOR COST OPTIMIZATION

Abstract

The evolution of artificial intelligence (AI) presents unparalleled opportunities for innovation and operational efficiency. However, as organizations increasingly integrate AI solutions into their operations, effective cost optimization strategies become increasingly critical.

This paper discusses the cost-related challenges of managing AI projects. It includes best practices, guidelines, and key principles to help control cost overruns. It also details how Infosys successfully manages AI costs by implementing proven strategies and tracking essential metrics.

Infosys®
Navigate your next

Table of Contents

Introduction

A popular joke in technology circles describes a person explaining how he went bankrupt: by leaving his cloud services running unchecked. Funny as it may seem, the caricature highlights a real issue—cloud costs pose serious challenges for many enterprises. To help companies manage their spending and resources more efficiently, FinOps emerged as an operational framework in the early days of cloud adoption.

The scenario is similar with the proliferation of AI across the industry. Though adoption of AI is in the early stages, the challenges seen with cloud adoption are relevant to AI as well. The cost of running AI will be crucial in deciding how organizations move forward with it.

Importance of AI Cost Optimization

AI cost optimization involves a dual-pronged strategy of reducing expenses while ensuring substantial returns on investment (ROI) in AI. The goal is to balance expenditure with value creation, thus enhancing profitability and sustainability. Several factors drive the need for cost optimization in AI:

Scalability:

As AI gets deployed at scale, costs can escalate rapidly. Effective cost strategies help manage these expenses while supporting growth.

Resource allocation:

Optimizing costs ensures that resources are utilized efficiently, minimizing waste and redundancy.

Competitive advantage:

Organizations that effectively manage their AI expenditures can reinvest savings into further innovation, gaining a competitive edge.

Infosys Framework for AI Cost Optimization

AI cost optimization should be proactive, not reactive. Organizations must leverage tools to forecast costs, evaluate business cases, and make informed decisions. Infosys, has developed a comprehensive framework with a focus on AI discovery, prioritization, and maturity to manage AI costs strategically. The framework, represented in Figure 1, includes six key tenets to guide organizations in optimizing their AI expenses:

Return on investment	AI architecture optimization
Model selection strategy	Infrastructure management
Metrics collection	Responsible AI practices

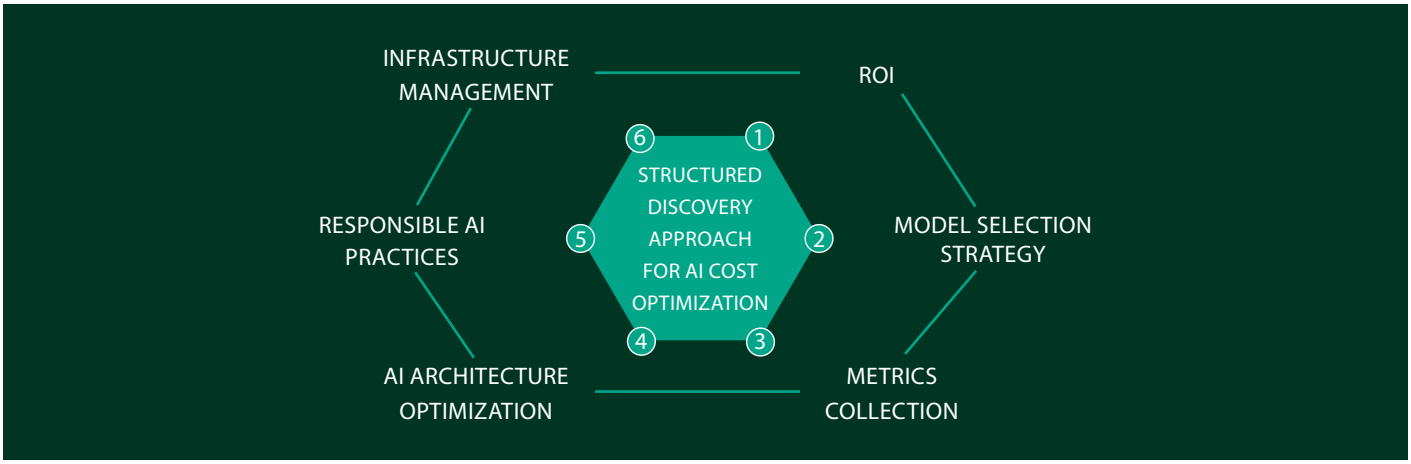

Figure 1: Structured Discovery Approach for AI Cost Optimization

Return on investment

Effective AI cost optimization begins with identifying high-value use cases and selecting the most appropriate AI models. This strategic approach ensures that efforts are focused on areas where AI can deliver measurable business value. A fundamental principle in AI cost management is the upfront calculation of ROI for each use case. This involves a thorough analysis of the expected benefits and costs associated with AI implementation. By predicting ROI, organizations can prioritize projects with the highest potential for profitability and operational efficiency.

Model selection strategy

Businesses must assess the complexity and specificity of tasks requiring AI. Large language models (LLMs) are ideal for complex, generalized tasks but are expensive. While LLMs deliver high performance, small or fine-tuned models can be more cost-efficient for simpler tasks, often providing the necessary performance.

Consider the quantity and quality of available data. Fine-tuning a pre-trained model can be cost-effective if domain-specific data is available. Evaluate the cost of running the model in production prior to deployment. Smaller models typically have lower inference costs, making them suitable for real-time applications. Ensure the model can scale with your needs. LLMs might be excessive for small-scale applications but necessary for large-scale deployments.

It is important to balance performance needs with budget constraints. By carefully evaluating these criteria, businesses can select a model that balances cost and performance optimally.

Metrics collection

The management world is familiar with the saying, "If you can't measure it, you can't improve it." While selecting the right model and calculating ROI upfront will significantly help control costs, businesses need to monitor daily consumption closely. It is essential to gather sufficient telemetry data to determine usage and take corrective actions regularly.

Key metrics for cost optimization include:

Data quality checks:

Ensuring the accuracy and relevance of input data to maintain efficiency and reduce errors

Token utilization:

Tracking token usage to manage costs associated with API calls, especially for models with per-token pricing

API call cost:

Monitoring API call costs to identify expensive operations and optimize service usage

Latency:

Measuring the processing time to detect performance bottlenecks and areas for performance improvement

Total cost per successful outcome:

Calculating all expenses associated with achieving a desired result to ensure cost-effectiveness

Processing time:

Evaluating the time taken for data processing to enhance efficiency and reduce operation time

Model accuracy versus cost:

Balancing the accuracy of predictions against the cost of running the model to ensure economic viability

AI architecture optimization

Architecture is crucial for ensuring that an application scales effectively and performs optimally as well as cost-efficiently. Large enterprises are increasingly investing in centralized AI platforms to expand their AI initiatives. It is imperative to architect these platforms effectively to achieve the desired outcomes.

Key principles for optimizing AI architecture include:

PolyAI architecture:

Flexibility to dynamically optimize load and cost, ensuring efficient and cost-effective resource allocation across varying tasks and demands

Dynamic model switching:

Adaptive switching enabled to match task complexity, leveraging lower cost options whenever feasible while optimizing performance and expenses

Minimal API calls:

Consolidation of requests and optimized data processing workflows, minimizing the number of API calls, driving resource efficiency and cost savings

Repetitive pattern on prompts:

Streamlining of queries and reduction in redundancy through prompt optimization, significantly enhancing system efficiency and reducing costs

Caching mechanisms:

Reduction in repeated API calls through effective caching mechanisms, significantly lowering overall model usage and associated costs

Task segmentation:

Assignment of specific tasks to different models based on complexity and accuracy requirements, optimizing costs while maintaining high performance and precision

By incorporating these principles, organizations can build AI architectures that are scalable, performant, and cost-effective. This thoughtful approach to AI architecture will enable enterprises to maximize their investment in AI technologies while controlling expenses and ensuring sustainable growth.

Infrastructure management

Effective infrastructure management ensures that resources are allocated efficiently across regions and tasks. It enhances performance and reduces costs significantly, making it a key component of any AI strategy.

Key considerations for optimizing AI infrastructure include:

Paging models:

Implementing smart load balancing across regions for paging models optimizes AI costs by dynamically, distributing workloads based on real-time demand and resource availability, ensuring efficient utilization, and minimizing latency.

Using small language models:

Leveraging small language models (SLMs) for tasks that do not require extensive computational power can drastically cut down costs. These models are more efficient and cost-effective for simpler tasks while reserving more complex, high-performance models for tasks that require higher accuracy.

Tools and technologies:

Employing tools such as OpenTelemetry, LangSmith, and Grafana is crucial for monitoring and optimizing infrastructure. These tools provide insights into application performance, helping identify bottlenecks and areas where resources can be optimized.

Comprehensive infrastructure monitoring dashboard:

Using a real-time, comprehensive dashboard that monitors performance and consumption is vital. It enables proactive management of infrastructure resources, ensuring issues are swiftly addressed and resources optimized.

Responsible AI practices

Incorporating Responsible AI (RAI) practices ensures that cost optimization does not compromise ethical standards. The comprehensive RAI framework must address:

- Technical guardrails to ensure the system is responsible by design
- Legal guardrails to ensure regulatory compliance
- Process guardrails to enforce AI governance

Beyond the basic tenets of RAI such as fairness and bias mitigation, other critical aspects, including privacy, security, safety, and explainability, must be codified within the framework to ensure compliance. Integrating these aspects into AI cost management is essential for maintaining ethical integrity while optimizing resources and expenditure.

Forecasting Usage Trends for Future Optimization

Forecasting trends in AI model utilization is essential for planning resource scaling and budgeting for future deployments. By analyzing historical data on token usage, API call frequency, and model-specific demand, organizations can identify usage patterns and anticipate resource needs. Possible actions include:

- Identifying emerging patterns to pick new AI models to align with evolving trends
- Monitoring real-time data streams to detect shifts in usage patterns and trigger alerts

AI-first Approach at Infosys

In today's digital transformation era, embracing an AI-first approach is essential for businesses that want to stay ahead. By embedding AI into their core digital strategies, organizations can fundamentally transform the way they operate, innovate, and create value. Infosys has embraced this approach through its Infosys Topaz initiative, which includes solutions, platforms, and reimagined services to drive innovation and efficiency. Our AI-first approach is a comprehensive transformation journey aimed at positioning Infosys as a digital-native and AI-first company. It calls for a cultural and operational shift across three key areas: work, workplace, and workforce. The approach integrates AI-first experiences and processes as well as embedding AI into systems, while fostering collaboration between human agents and AI chatbots. It also focuses on talent transformation, smart workplaces, and the development of sustainable, smart products.

Our AI-first approach prioritizes cost optimization, experience transformation, and productivity improvement. By leveraging advanced AI technologies, Infosys aims to streamline operations, enhance decision making, and deliver substantial savings. These insights and best practices are shared with clients, empowering them to achieve similar levels of optimization and improvement.

Best Practices to Track and Optimize AI Costs

We have adopted several best practices to manage AI costs effectively:

Applications are designed to optimize LLM calls by leveraging cached responses wherever possible. The choice between SLMs and LLMs, whether on-premises or cloud-based, is data-driven and helps us manage costs effectively.

From a software development lifecycle (SDLC) perspective, we provide a wide range of options, from open-source tools to industry-leading solutions, enabling our developers to choose based on specific project needs and business cases.

Applications are configured to log detailed telemetry data, providing granular visibility into token consumption and its cost implications.

Comprehensive dashboards track costs across applications in real time and alert stakeholders when breaches or anomalies are detected.

As we continue to expand our AI capabilities, maintaining cost control is essential to maximizing our ROI. Tracking key metrics, analyzing data, and drawing actionable insights helps keep costs in check while driving higher efficiencies.

Key metrics

Table 1 details several key metrics that help us monitor and manage our AI costs effectively.

Table 1: Key metrics to monitor and manage AI costs

No	Metrics	Description
1	Cost per AI service and deployment	Checks for high-cost areas and assesses their alignment with business objectives
2	Utilization	Number of tokens used by each application team
3	Cost of use cases	Cost incurred/number of tokens used per use case
4	Budget versus actual spend	Monthly variance indicator between the budgeted cost and actual spend
5	Anomaly detection in AI cost	Daily cost monitor that identifies anomalous cost patterns
6	Tagging/subscription key usage monitoring	Monthly check that ensures resources are appropriately tagged for cost allocation and reporting purposes
7	On-premises versus cloud cost for AI	Regular check on the model usage costs for timely transition of the model to an on-premises setup for hosting in case of substantial usage or regulatory compliance requirements

We use dashboards to seek visibility into AI-related expenses that are useful in keeping our AI costs under control.

For instance, daily cost calculations factor in the previous day's consumption, ensuring near-real-time tracking for more accurate cost management, as illustrated by the dashboard in Figure 2.

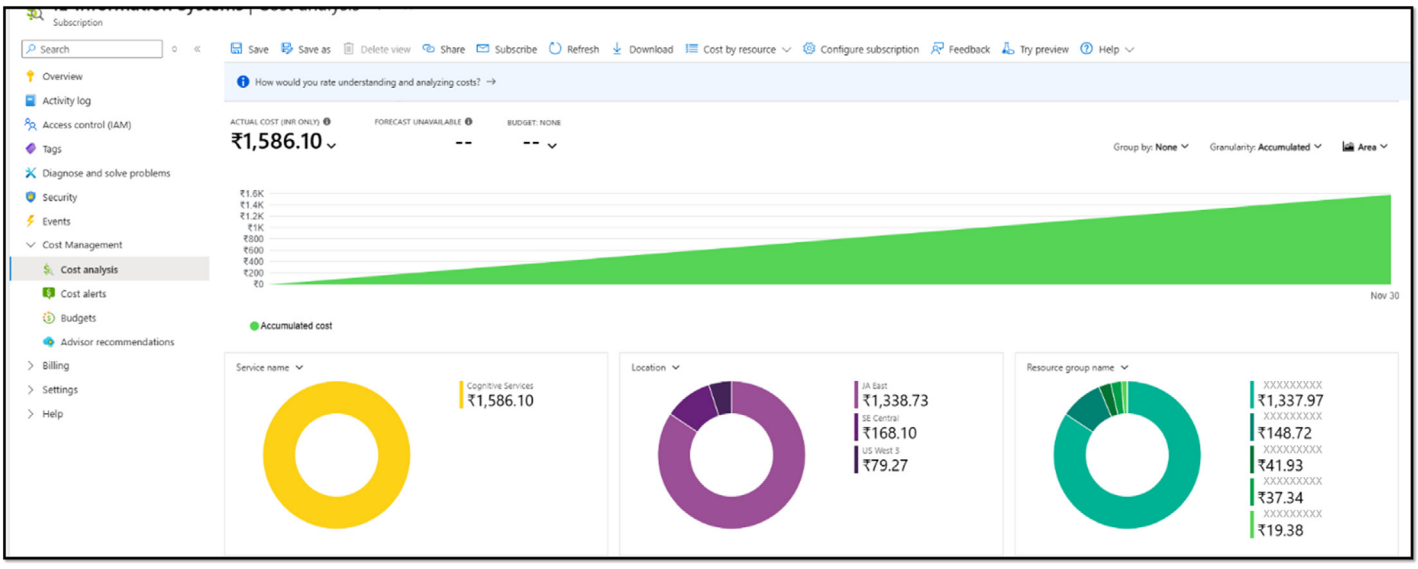

Figure 2: Daily cost calculation

Monthly cost calculation is performed against AI spending at the resource level, aggregating the data for complete cost analysis. Table 2 lists the details of the monthly AI cost analysis while Figure 3 offers a comprehensive view of the monthly costs calculations against AI expenditure. Table 2 provides the daily cost analysis across various resources, helping spot anomalies.

Table 2: Monthly AI cost analysis

Monthly AI Cost Analysis			
S.No	Meter Subcategory	Resource	Actual Cost (KUSD)
1	Provider 1	xx-loc-provider1-env1	xx.xx
2	Provider 1	xx-loc-provider1-env2	xx.xx
3	Provider 1	xx-loc-provider1-env3	xx.xx
4	Provider 1	xx-loc-provider1-env4	xx.xx
5	Provider 1	xx-loc-provider1-env5	xx.xx
6	Provider 2	xx-loc-provider2-env1	xx.xx
7	Provider 2	xx-loc-provider2-env2	xx.xx
8	Provider 2	xx-loc-provider2-env3	xx.xx
9	Provider 2	xx-loc-provider2-env4	xx.xx
10	Provider 2	xx-loc-provider2-env5	xx.xx

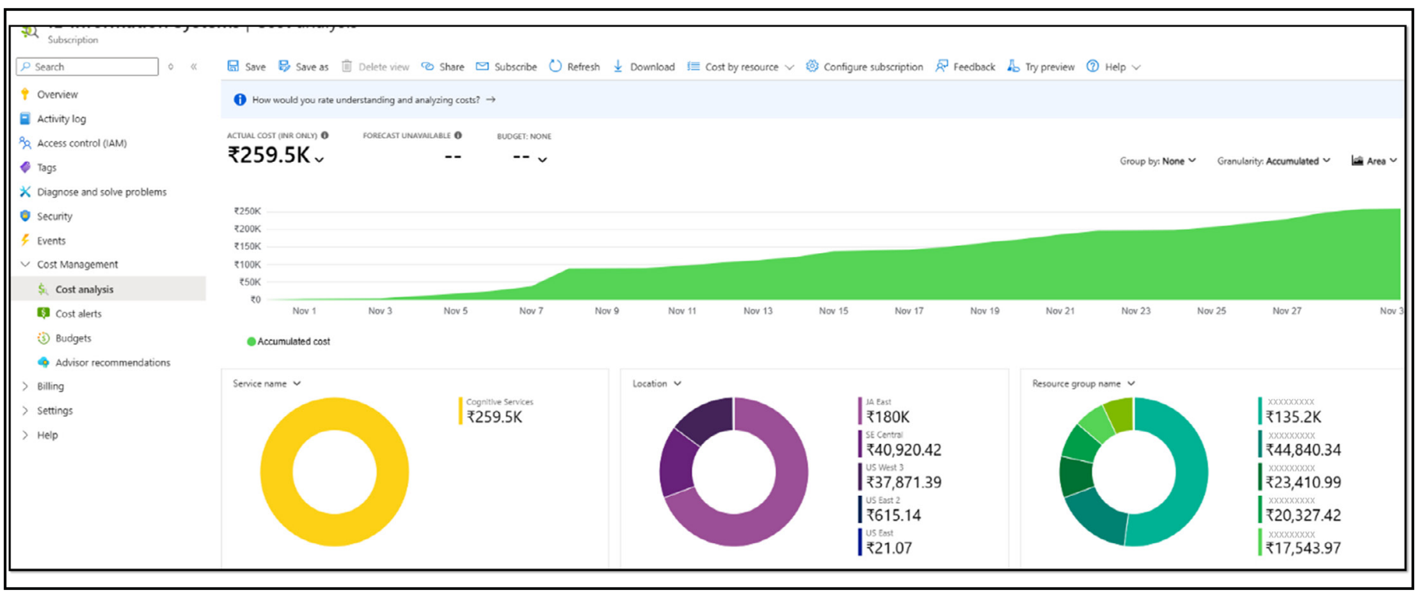

Figure 3: Monthly cost calculations against AI spends

Table 2: Daily cost analysis across various resources to help spot anomalies

Monthly Cost Analysis - MM/YYYY										
Date	xx-loc-provider1-env2	xx-loc-provider1-env3	xx-loc-provider1-env4	xx-loc-provider1-env5	xx-loc-provider1-env1	xx-loc-provider1-env2	xx-loc-provider1-env3	xx-loc-provider1-env4	xx-loc-provider1-env5	Total Cost (KUSD)
DD/MM/YYYY	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xxxx.xx
DD/MM/YYYY	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xxxx.xx
DD/MM/YYYY	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xxxx.xx
DD/MM/YYYY	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xxxx.xx
DD/MM/YYYY	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xxxx.xx

In addition to the dashboards, email alerts are sent to key stakeholders whenever anomalies in AI usage are detected, as shown in Figures 4, 5, and 6.

From: XXX
Sent: Weekday, Month Day, XXXX, H:M:SS
To:
Subject: Cost anomaly detected in XXX Systems

[**EXTERNAL EMAIL**] Please verify sender address and exercise caution before clicking on any link.

An unusual cost increase on 10/15/2024 12:00:00 AM for the XXXX subscription has been detected.

Anomaly alert: An unusual cost increase was detected

An unusual cost increase was detected on 10/15/2024 12:00:00 AM for the XXXX subscription. Cost Management detected a possible cost anomaly based on daily cost trends between DD/MM/YYY 12:00:00 AM and DD/MM/YYY 12:00:00 AM. Please review changes to determine whether this was expected.

Message from the owner of this alert:

Cost related anomaly detected information

Subscription summary

Anomaly detected	Yes
Delta compared to expected range	18.49 %

Figure 4: Sample email regarding an unusual increase in cost

Resource group summary

- Cost changed 21.46% from 193 existing resource group(s).

Most significant changes in resource group(s) during this period

Name	Cost change %	Percent of total
xx-loc-provider1-env1	193.6	15.98
xx-loc-provider1-env2	225.73	1.36

Figure 5: Sample email alerting stakeholders to cost changes

Name	Cost change %	Percent of total
xx-loc-env-gateway	54.98	0.96
xx-loc-env -analytics	99.02	0.83
xx-loc-env -test	100.18	0.59

This email was generated on 10/17/2024 7:41:07 AM and includes only the usage and charges available at that time. Anomaly detection is based on you from 8/17/2024 12:00:00 AM to 10/14/2024 12:00:00 AM. Cost is estimated based on normalized usage, which standardizes the unit of measure across types (such as hours and GB) and doesn't factor in credits or discounts. Learn more.

If you have Cost Management contributor access (or higher) to this subscription, you can manage this anomaly alert from cost alerts in the Azure portal

Additional information

Scope	xxxx- subscription
Scope Id	/subscriptions/subscription ID
Email frequency	Daily at 12:00 UTC
Conditions	Only if there is an anomaly

Figure 6: Sample email with information on significant cost changes

Conclusion

AI is on the path toward becoming an integral part of business. Enterprises must swiftly adopt and integrate AI to stay ahead of the game. However, AI implementation often comes with significant costs, making it critical to understand the financial implications of decisions involving AI and keep an eye on costs.

AI cost optimization needs a multifaceted approach that combines strategic planning, continuous monitoring, and ethical considerations. By applying the strategies and insights outlined in this paper, organizations can effectively manage their AI expenditure, ensuring both financial sustainability and innovative growth.

To build a strong business case and maximize ROI, it is important to avoid force-fitting AI solutions to requirements. Data-driven decision-making is crucial for the success of any AI-first initiative.

At Infosys, our AI-first approach has equipped us with deep and comprehensive expertise in all aspects of implementing AI while optimizing costs. Through early adoption, we have evolved processes, standards, and best practices that businesses can leverage to effectively contain costs for their AI implementations.

About the Author

Guruprasad NV

AVP, Senior Principal Technology Architect

Contributors

Muthukumar Subramanian

Sachin Karunakar Joshi

Ganapathi Raman Balasubramanian

Vikram Rao

Muppuli Gnanaraj Govindaraj

Deepthi Ramachandran

Vivekanand Kaankadae

Kedar Tigadi