



Cost Discipline Is the Missing Operating Model For Scaled AI

FEATURING RESEARCH FROM FORRESTER

AI Cost Optimization: The Why, What, And How



INTRODUCTION

Cost is where reality first emerges in the quick transition from experimentation to implementation of generative AI. The unpleasant reality is that production-grade AI is not a feature, or a one-time investment is now being realized by businesses that rushed through pilots. It's an operational model.

According to EdgeVerve's experience working with major corporations, expenses grow nonlinearly, drivers become opaque, and accountability spreads among teams as AI becomes more prevalent in products, workflows, and increasingly agentic systems. In manufacturing, what was manageable in proofs of concept becomes complicated. This difference is supported by Forrester's research, which shows that expanding AI necessitates consistent investment in infrastructure, data and model management, monitoring, governance, security, training, and business transformation. This significantly changes the cost profile when compared to early experiments.

AI design, implementation, governance, and ongoing improvement must all incorporate cost discipline. EdgeVerve places a strong emphasis on clearly defining trade-offs between risk, performance, and cost early on and reviewing them as AI systems advance. Establishing the operational and financial foundation necessary to execute AI dependably at business size is the goal here, not choosing the model with the lowest cost.

EdgeVerve embeds FinOps into the operational fabric of platform, enabling real-time visibility into infra and compute usage, spend guardrails, and value-aligned consumption patterns—critical as AI workloads such as data pipelines, fine-tuning, and high-volume inference introduce dynamic, unpredictable cost behaviors. As clients scale AI Next and adopt more complex models, this FinOps-driven approach ensures accountability, efficiency, and sustained ROI, keeping AI deployments predictable, scalable, and financially responsible.

IN THIS DOCUMENT

Cost Discipline Is the Missing Operating Model For Scaled AI

Research From Forrester: AI Cost Optimization: The Why, What, And How

About EdgeVerve

WHY AI COST CONTROL HAS BECOME A BOARD-LEVEL CONCERN

EdgeVerve finds that AI spending is rising more quickly than anticipated across industries, driven by a wider cost surface that encompasses computation, storage, data transfer, observability, governance, security, and organizational change in addition to model inference. According to Forrester, cloud usage can obfuscate line-item visibility, AI costs are hard to predict, and falling unit prices frequently lead to increased utilization, which eventually raises total spending.

According to Edge Verve, this has made AI cost control a board-level issue. Leaders in business and technology have two mandates:

- Show that investments in AI have a quantifiable business impact
- Avoid unmanageable risk and expense as AI is incorporated into core operations

According to EdgeVerve, reactive or disconnected cost governance causes problems for firms, resulting in overpaying, a delayed return on investment, and slowed innovation. This reinforces the need for a disciplined, enterprise-wide operating model for AI.



THE SHIFT: FROM ISOLATED PILOTS TO A MANAGED AI PORTFOLIO

Most businesses started out with team-specific technologies and disconnected use cases when using AI. This method works well for experiments; however, it is not scalable. EdgeVerve recommends switching from project-centric execution to a managed AI portfolio to get past this stage of redundant capabilities, disjointed data pipelines, uneven governance, and variable production standards.

A unified one platform approach that considers the factors of data, process, AI, and experience that influence both cost and results is what is required – a transition that was considered while curating EdgeVerve AI Next platform. The Platform enables businesses to scale AI steadily without requiring each team to recreate key components as it standardizes reusable features like automation, analytics, value measurement, model lifecycle management, integration, orchestration, data integration, data integration, data management and more.

EdgeVerve AI Next helps organizations maximize ROI on AI initiatives by addressing the key cost drivers outlined in Forrester's AI cost-optimization research through its model-agnostic PolyAI architecture (balancing quality vs. costs), token costs monitoring and optimization, deployment flexibility (on-prem/cloud-agnostic – prevents vendor lock in and optimizes infra placement), and unified observability (that reduces tool sprawl and shadow AI). Fabric, platform's intelligent data integration and transformation layer, further optimizes one of the largest AI cost levers—data movement, quality, and transformation—ensuring consumption-ready datasets while minimizing redundant pipelines and egress inefficiencies.

This platform-led approach helps reduce structural cost while improving speed, flexibility, control, and reliability

WHAT “COST OPTIMIZATION” LOOKS LIKE IN PRACTICE

EdgeVerve aligns with Forrester's guidance that effective AI cost optimization starts with a business use case, maintains a full-system view of the AI stack and lifecycle, and iteratively refines the cost model over time. In practice, EdgeVerve sees two execution realities that consistently separate leaders from laggards:

1. Make Cost and Governance Part Of Delivery

Based on EdgeVerve's experience, governance and operational overhead turn into a hidden tax on the value of AI when implemented late. There is immediate integration of accountability, cost controls and policy enforcement into AI delivery pipelines by leading organisations. EdgeVerve places a strong emphasis on viewing governance as a prerequisite for repeatable scalability and predictable cost, rather than as a compliance roadblock.

2. Measure Value, Not Just Usage

Many firms over-index on tokens, computing, and utilization indicators, according to EdgeVerve. These are not results; they are inputs. Executives look for measurement frameworks that stitch the AI initiatives to the following business impacts: productivity, customer experience, cycle time reduction, and risk mitigation. Keeping in mind Forrester's caution we emphasize that cost optimization must be weighed against performance, robustness, and innovation

Where this Leads

From EdgeVerve's perspective, organizations that succeed with AI will treat it as a core business capability governed with financial discipline — not as a collection of experiments. These organizations will:

- Align AI investment to use cases that are clearly prioritized.
- Do a complete due diligence of the AI cost drivers
- Have AI autonomy as well as governance and measurement that evolve together
- Continuously optimize as models, data, and operating patterns change.

Forrester's best-practice framework for AI cost optimization is what follows in this research. EdgeVerve recommends this framework for CEOs who are looking to create long-lasting AI initiatives, thus striking a balance between cost and results, and confidently scaling AI.

This is the same point that EdgeVerve has been emphasizing again and again. The main type of platforms that are common for all the people in every format and situation

AI Cost Optimization: The Why, What, And How

June 11, 2025

By Michele Goetz, Tracy Woo, Charlie Dai with Lauren Nelson, Sudha Maheshwari, Frederic Giron, Rowan Curran, Carlos Casanova, Hannah Murphy, Marissa Fritz

FORRESTER®

Summary

The frenzy around generative AI (genAI) has enterprises scrambling to instill AI technology into their products and workflows. As companies start to build out production-ready use cases, the associated costs are increasing at an alarming rate. The release of black box AI services from cloud service providers adds to the confusion by obfuscating much of the cost details that FinOps teams are used to seeing in an average cloud bill. This report outlines the framework and levers that organizations should know to optimize their AI costs.

Get AI Costs Under Control To Maximize Business Outcomes

Enterprises are on track to a rude awakening. GenAI spending jumped from \$2.3 billion in 2023 to [\\$13.8 billion](#) in 2024, a six-times increase. Compared with traditional AI models, large-scale AI models use about [100 times](#) more compute. If models continue to increase at the same rate, the total cost of an AI system is projected to grow to roughly [\\$700 billion](#). Training costs for the 2017 Google Transformer model were \$900, while in 2023, Google's Gemini Ultra training costs were [\\$191,400,000](#). Even as smaller, more efficient, and more effective models are introduced to the market, such as DeepSeek-R1 in 2024, the proliferation of ML and agentic AI use will have a [Jevons effect](#); lower costs increase utilization, thereby putting more pressure on resources. The core concepts to pay attention to? Cost levers and weighing the trade-offs.

Take A Holistic View Of The Cost Model For AI Operationalization

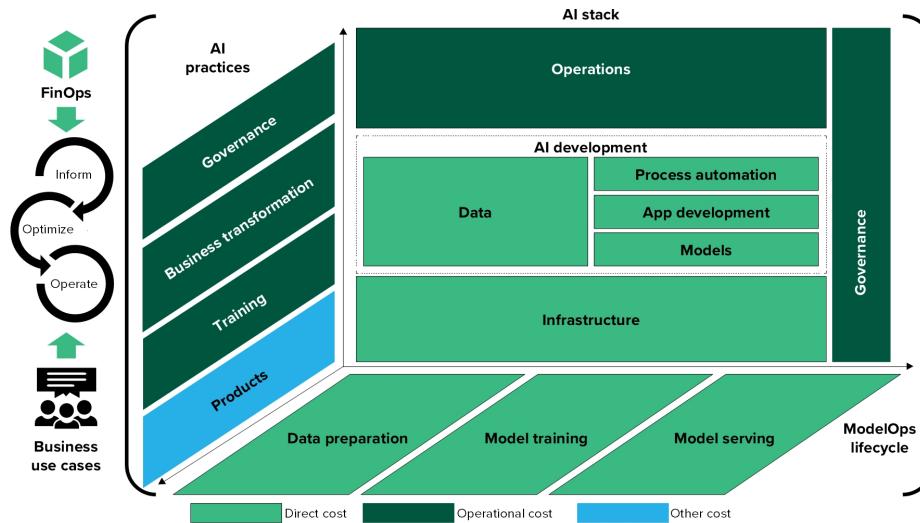
The cost of operationalizing AI at scale differs significantly from the cost of a proof of concept (POC), which primarily focuses on initial development and testing. Scaling AI requires substantial investments across various domains, including infrastructure, data and model management, continuous monitoring, maintenance, and updates. Comprehensive understanding of the AI cost model is essential to mitigate risk, allocate and optimize costs, and ensure initiative sustainability. By considering the total cost of ownership, organizations can make informed decisions, maximize ROI, and achieve long-term success in their AI initiatives (see Figure 1). But this remains challenging for many enterprise IT shops.

The use case is the primary determinant for AI costs and is the North Star for the cost model operationalization. This spans three dimensions: the AI stack, the ModelOps lifecycle, and AI practices — each with specific coverage domains tied to cost considerations. Forrester recommends that you use FinOps practices as the starting point for an iterative approach for areas such as selecting the right tech stack and partners, building capabilities to meet technical needs, and continuously optimizing for performance. This closed-loop process will help you refine your cost models over time; support adaptive, effective AI deployment; maximize value; and drive long-term success.

Please note that this report does not cover AI products. Many of these products charge on either a subscription or flat-fee fixed tier determined per user or per core. Many large platforms add additional charges for AI capabilities.

Figure 1

An Overview Of AI Cost Model To Operationalize AI



© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Pull The Levers Of AI Costs

AI cost levers offer the ability to make informed trade-offs between cost, performance, and efficacy risk of AI. While the tangible impact of AI can be difficult to quantify, organizations can contextualize costs within the broader objectives outlined in their AI strategy and prioritized use cases. A global professional services firm helped a large hospital evaluate the productivity impact of using genAI. The analysis revealed a potential 24% productivity gain across patient-facing, clinical support, and administrative roles. This insight alone was compelling enough for the chief clinical operations officer to green-light the investment. Own this best practice — when approaching AI initiatives, start with a use case. Next, categorize costs into two primary buckets: resource costs and operational expenses (see Figure 2).

- **Direct AI cost levers are intrinsic to thresholds of AI quality.** Models, data, and infrastructure are the supporting technologies that deliver an AI use case. The decisions made in each of these levers can drastically affect the cost, quality, and performance of the AI solution or capability being developed. Fortunately, there are tools that can aid in the decision-making process. [AI platforms](#), [data platforms](#), and AI technology service providers platforms (e.g., Accenture's Model Switchboard, EPAM's Dial) provide cost calculations for models and data. The

major cloud service providers and foundation model vendors (e.g., OpenAI, Mistral) provide transparency and calculators for model pricing and packaging. These tools help AI developers compare the costs of training, tuning, and prompt engineering to improve model efficacy in context of cost to develop and forecast for AI in production (see Figure 3).

- **Operational cost levers manage AI overhead.** Governance, business transformation, and training are the main enabling factors that tie technology creation to market differentiation. An innovative technology is only as valuable as the usage itself. Licensing and operations costs may link to either business unit budgets or enterprise budgets. Soft costs such as governance and upskilling are hidden factors that can bottleneck success of an AI initiative. As one head of AI at an insurance company noted, reducing call center staff through AI required hiring two to three data scientists and engineers to support the AI tool. Visibility into these factors is essential for delivering AI and will set the new baseline for a business running on AI (see Figure 4).

Figure 2
Direct And Production Levers Are The Foundation Of AI Cost Trade-Offs

Models	
Cost lever	Description
Model selection	Cost depends on processing efficiency, response quality, per-call and batch, and usage (commercial, open-source).
Number of models	Model quantity and type (RPA, IPA, agentic, ensemble) are use case dependent, each incurring unique costs.
Token generation	Costs rise with requests and poor prompts or models. Response may cost three to 10 times more than input. Token costs increase with number of inference sessions or cached outputs.

Data	
Cost lever	Description
Data sources	Third-party and synthetic data raise costs but are often needed for training and tuning.
Data quality	Costs rise with poor data quality, completeness, consistency, accuracy, and freshness, harming retrieval, inferencing, and response accuracy.
Data transfer (ingress/egress)	Data retrieval from cloud and warehouses raises costs, along with batch transfers, event capture, and data transformation.
Data storage	Costs grow with usage and with processing and maintaining metadata and embeddings.

Infrastructure	
Cost lever	Description
On-premises infrastructure	Costs rise with server quality and quantity, impacting cooling, power, and facility footprint.
Cloud infrastructure	Costs rise with quality and quantity of compute, memory, and storage.
Network infrastructure (hybrid and on-premises)	Costs rise with quality and quantity of components (routers, servers, switches, hub, repeater, access points, bridges, leased fiber, direct connects).

© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Figure 3
Cost Considerations That Affect Direct AI Cost Levers

Direct AI cost levers			Operational cost levers
Models	Data	Infrastructure	Operations
Model selection	Data source and storage	On-premises infrastructure	Governance (AI and data)
Number of models	Data quality	Public cloud infrastructure	Business transformation
Token generation	Data transfer (ingress/egress)	Network infrastructure (hybrid and on-premises)	Training

© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Figure 4
Cost Considerations That Affect Operational AI Cost Levers

Operations	
Cost lever	Description
Governance	Salary costs of data stewards and governance teams.
Training	Costs from continuous training programs, external courses, and consultancy training.
Business transformation	Talent and services to support business process engineering and business model creation to evolve the organization with AI.

© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Models Are The Beginning Of Cost Optimization

Tokens are the proxy for model cost, but relying on model vendor calculators and AI platform analytics is not enough. David Tepper, cofounder and CEO of Pay-i, an AI cost management platform, warns that metering algorithms for model use based on token generation gets in the way of predicting spend. For GPT-4o, OpenAI has four unit types to understand cost (text input, image input, text output, cache read); each of these types also has a batch version. Azure OpenAI has 24 types before counting batch versions because the costs vary by deployment type (standard/provisioned) and scope (global, data zone, region). Since each model and vendor handles tokens differently, simplified calculators often diverge from real usage and billing. The complexity increases with new reasoning models. Many calculators don't capture the high variability between executions, thus providing unreliable forecasting information. Understanding these nuances is foundational to making informed decisions about the

cost, quality, and performance of the models. To optimize cost, you'll need to tweak the following levers:

- **Model selection: Balance quality and cost.** Factors in selecting a model include model type (LLM, SLM, predictive, etc.), model provider (commercial or open-source), and the efficacy of the model for the intended use case, but this is just the starting point. With the pace of new model introduction, mature organizations continuously optimize, swap in, and swap out over time to increase outcomes. Joshua Collier, FinOps lead at Grammarly, states that the company regularly switches AI models to save costs and improve performance. Grammarly successfully switches models by staying updated on vendor product changes and running experiments to measure engagement and token consumption.
- **Number of models: Evaluate AI cost from scale and agentic requirements.** AI models, particularly genAI copilots, may appear to be a single model facilitating tasks like summarization, coding, or video creation, but they're typically composed of multiple models orchestrated to complete a task or function. This introduces a multiplier effect on cost. Managing the costs of multiple models is complex but achievable. One technology services and consulting firm provides agentic offerings that deploy swarms of AI agents across business processes to fulfill contact center and supply chain tasks. It tunes the models and agentic services independently as well as through an optimized orchestration layer in the genAI studio platform for efficient, cost-effective delivery.
- **Token generation: Craft an efficient processing design.** Model processing profiles affect token generation from prompt engineering, parameter count, neural network depth, predictive versus reasoning, and training constraints. Usage and offloading are two key levers. A professional services firm addressed usage during a gradual rollout of a genAI knowledge platform. Request-response loops and user feedback enabled the firm to optimize costs for specific use cases and forecast budgets for new genAI knowledge capabilities and features. A SaaS company introduced an industrial knowledge graph to limit processing costs by constraining the number of model levels and maintaining repeatable responses — this had an additional effect of reducing hallucinations.

Data Is The Biggest Lever Of Model Cost

The scope and scale of AI is measured by the data. CIOs report that while [5% of their IT budget goes to AI](#), 20% is spent on data. And in [Forrester's Data And Analytics Survey, 2025](#), 68% of data and analytics decision-makers indicated an increase in data spend from 2024 to 2025, with Forrester clients citing AI as a key driver. Companies are

deploying specialized AI capabilities that use smaller domains, but this increases the number of data products to manage. At the opposite end, [agentic AI](#) systems traverse distributed data sources to access multiple business data domains while also using dedicated subagents to handle the data complexity. According to a data and AI services company, agentic systems commonly have a 1:2 ratio of data agents to purposeful business agents to address the complexity of data for agentic AI. This creates a multiplier effect on model cost to handle data transfers, orchestration, synchronization, and transformation processes. Thus, organizations will pull data levers that improve the outcomes of models while making trade-offs across the data infrastructure and ecosystem.

- **Data sources and storage: Balance the data lifecycle with management cost.** [A](#)

[Seagate study](#) found that AI is doubling storage demands, thus increasing data management and storage costs on production and nonproduction environments. Agentic systems store daily checkpoints, logs, metadata, insights, and outcomes in both centralized and local repositories. These systems often use second-party, third-party, and [synthetic data](#), each with unique cost factors like storage, subscription, and maintenance. Open source starts free but incurs costs at scale. Second-party data can stay in place while models move to partner environments to offload costs, but this requires a transfer learning platform. Third-party data costs can be managed through structured subscription terms and quality-controlled data governance. A cybersecurity company recommends optimizing data storage by focusing on formats, compression, deduplication, log generation, and tiering and eliminating abandoned data.

- **Data quality: Optimize data quality across the entire data flow.** Missing data, inaccuracies, anomalies, and bias can affect every stage, from source data and inputs to chunking, metadata, and outputs. Poor data governance and ML/AI processing flaws degrade data quality, increasing costs as more requests are needed to achieve reliable results. An AI platform company addresses this by using a knowledge base of curated chunks from data sources and previous dialog from an LLM-based chatbot to access high-quality data. Its integration templates improve data fabric integration to maintain schema and mapping integrity. Within AI agents, Google's [Data Science Agent in Colab](#) generates Python to create data pipelines, prepare data, and maintain quality based on insights, the type of analysis, and the ML techniques used.

- **Data transfer: Optimize ingress and egress costs.** Data and AI vendors charge for consumption, while data integration and data fabric vendors charge for ingestion. Pipeline, [iPaaS](#), and data virtualization vendors that transfer and orchestrate data

commonly meter usage. [Fivetran charges](#) by monthly active rows. [Denodo on Azure](#) charges hourly software and infrastructure fees. Centralizing data into a cloud vendor's RAG-based environment may seem attractive but is shortsighted. AI data flows drive up costs by spanning prompt and session logs, metadata, and outcomes between edge and central sources. One media company optimized AI/ML application data by batch copying directly into a lakehouse. It minimized egress costs by using low-cost AWS storage for raw data, storing critical data only in Snowflake, and applying replication and egress restrictions.

Infrastructure Carries Classic Capex/Opex Questions

Infrastructure is the underlying backbone that enables AI. Workload placement has a large influence on latency, performance, compliance, and cost. Choosing public cloud grants access to the widest array of GPUs and instances that power AI, but training and inferencing costs can vary wildly. On-prem infrastructure has predictable expenses but huge upfront capex costs, limiting entrants to only the well-funded. Where your application sits — on-prem, private cloud, public cloud, or hybrid cloud —determines the type of network infrastructure you'll need to choose.

- **On-premises infrastructure: Predictable with high upfront costs.** AI data centers can cost from [\\$10 million for small-scale setups](#) to over [\\$500 million](#) for large-scale operations. Enterprises must weigh the cost-performance trade-off between GPUs and pricier TPUs. Cisco and NVIDIA offer perpetual licensing on a per-unit basis but have high upfront costs but also provide hardware-as-a-service consumption-based pricing. Emerging solutions like Obvious Future offer production-ready AI appliances starting at under \$50,000. Optimize costs by using high-density servers, data compression, and renewable technology to reduce energy costs.
- **Public cloud infrastructure: On-demand pricing with volatile spend.** Public cloud includes AI GPUs and traditional services like storage, databases, and serverless functions. GPUs are billed hourly while storage, database, and serverless charge by capacity or request size per month. Serverless is ideal for dynamic workloads due to its autoscaling and usage-based pricing. Discounts are available via reserved instances or savings plans, though unused commitments are still billed. Enterprises can negotiate discounts by service type or total spend; for instance, heavy storage users may secure better rates. Support and SLAs can also be discounted or credited. Enforce policies that limit runaway costs to reduce unpredictable spend.
- **Network infrastructure (on-prem and hybrid): Misconfigurations lead to wasted spend.** Networking is a major cost barrier in AI data centers with performance-

sensitive workloads impacted by even minor issues; a 1% packet loss can cut AI training performance by 33%. On-prem networks incur costs from hardware, virtual network functions (VNFs), and cloud-based services like secure access service edge (SASE). Hardware may be billed upfront or monthly, some capped at a percentage of total capacity. VNFs generally use a subscription model based on bandwidth, features, and support while SASE solutions charge by connection annually. Though subscription models offer flexibility, exceeding capacity can trigger costly overages.

While hybrid cloud AI offers benefits like lower latency and regulatory compliance, it often requires “direct connects” via colocation or dark fiber to cloud data centers. If bandwidth, port hours, or data transfer limits create constraints, additional investment is needed. More critically, assess the size and frequency of outbound data. If on-prem processing offsets egress fees, consider hybrid networking solutions like SASE and multicloud networking (MCN), billed monthly per software instance, along with security appliances such as Palo Alto VM-Series.

Operations Are The Hidden Tax Of AI Value

AI is the biggest lever for reducing operational expenses — Google reports a two-times productivity boost. Still, AI isn’t replacing the workforce; it’s a harbinger of change. [Meta](#) cut 5% of staff while increasing hiring for ML and critical engineering roles for AI initiatives. This shift highlights AI’s dual role in boosting efficiency and transforming business operations and growth plans. Early investment is necessary for reskilling, change management, and talent acquisition. To optimize AI costs, organizations must account for the operational impact of transitioning to an AI-driven business model and identify the necessary resources.

- **AI and data governance: An arbiter of cost, outcomes, and risk.** Governance serves as a steering committee, center of excellence, training hub, and operational function, providing cost oversight while also needing its own budget. In regulated regions like the EU, Singapore, and US states such as Colorado and New York, these teams must support compliance with audits, reporting, and responsible AI practices. These teams run lean with modest salaries ranging from [\\$60,000](#) to [\\$185,000](#), often relying on existing data governance staff for early efforts. One automotive company used data stewards and data engineers to stand up Databricks for AI use cases, implementing tools like Collibra and Unity Catalog to define data domains and manage data products. Even with these tools, the company hired two additional stewards over two years to meet data quality, privacy, and readiness needs.

- **Business transformation: Change management reduces long-term cost.** AI transforms business processes and roles, often creating hidden costs. Productivity may initially drop as employees adapt, requiring onboarding and post-deployment support to minimize disruption. Larger initiatives may involve partners like Boston Consulting Group or Deloitte for training, implementing protocols, and establishing change management, adding to consulting costs. To control transformation expenses, organizations must invest early in readiness and change management to reduce the impact of business disruption.
- **Training: Talent management is continuous.** Stanford University professor Jeremy Utley observed that many people struggle to use genAI effectively. He emphasized the value of conversational interaction with LLMs, including prompting bidirectional interaction. This challenges traditional workflows focused on data entry and screen navigation. Companies like IBM, McKinsey & Company, and PwC are investing billions in AI literacy, change management, and evolving hiring requirements and career paths. Roche invested in an assessment of future roles that support AI initiatives. Any AI transformation effort must account for upskilling as a crucial component.

Establish A Financial Backbone That Puts Cost In The Context Of AI ROI

Cost is a single metric. Taken on its own, firms may make decisions focused only on cost reduction that negatively impact performance, decrease innovation, and miss out on savings reinvestment in future opportunities. Tech leaders must weigh the trade-offs between cost reduction and value optimization. Increase staff's AI literacy — from exec to engineer — on financial impacts, best practices, and operational transformation. Forrester recommends that you:

- **Start with the AI use case.** AI cost will vary widely across use cases as you define the type of model(s) and the complexity of the solution. Google guides its customers to think top down and bottom up to ensure that the corporate strategy (top) and use cases (bottom) are aligned by strategic priorities and AI domains that create the scope of cost per use case.
- **Maintain a full AI system view.** Total AI cost is a product of AI practices, the AI stack, and AI/ModelOps lifecycle expense. A comprehensive understanding of the AI cost model is crucial to reduce business disruption. While few solutions exist that connect AI costs (models, licensing, data, operations), a combination of tools (cloud cost management and optimization, IT financial management, enterprise performance management, and data management solutions) along with operations

integration via cross-functional collaboration can help with alignment.

- **Establish an AI cost baseline.** Emerging technology investments may require new strategies for revenue and differentiation, but budgeting and accountability are often afterthoughts. Establish a baseline. Create room for investment, but demand a schedule on ROI. [Dr. John Halamka](#), president of the Mayo Clinic Platform, says to invest in AI as a pathway to innovation with the understanding that it will pay off, but it takes time.
- **Put a financial framework in place.** Traditional FinOps provides a starting point and is currently expanding to include AI cost management. Until these practices are firmly in place, start connecting AI costs to business initiatives. Use these guidelines to inform decisions at all layers, from new application proposals to code development. A consumer goods company benchmarks models and tokens to align with AI use cases, allowing the firm to determine basic enterprise AI cost while developing chargeback models for business units.

Supplemental Material

Companies We Interviewed For This Report

We would like to thank the individuals from the following companies who generously gave their time during the research for this report

Cognite

Databricks

EPAM

Google

Grammarly

IBM

Infosys

Microsoft

Nubank

Pay-i

SAP

Wipro

We help business and technology leaders use customer obsession to accelerate growth.

[FORRESTER.COM](https://forrester.com)

Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

[Learn more.](#)

Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

[Learn more.](#)

Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

[Learn more.](#)

Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com



EdgeVerve Systems Limited, a wholly-owned subsidiary of Infosys, is a global leader in developing digital platforms, empowering clients to unlock unlimited possibilities in their digital transformation journey. Our purpose is to inspire enterprises with the power of digital platforms, thereby enabling our clients to innovate on business models, drive game-changing efficiency, amplify human potential, and foster a connected ecosystem.

EdgeVerve AI Next, part of Infosys Topaz, enables the scaling of Applied AI across the enterprise. With the power of Agentic and Generative AI, this unified platform bridges silos in people, processes, data, and technology to drive transformation in business operations. EdgeVerve AI Next is more than just a platform; it's a catalyst for digital transformation. By organizing data, ensuring high-quality information, and bridging the gap between existing systems. As a leading data connector, EdgeVerve AI Next offers flexibility in using AI models. Our Gen AI framework, utilizing a foundry and factory model, accelerates AI adoption and delivers tangible benefits through a structured approach.

Through this platform, EdgeVerve's innovations are helping global corporations across sectors such as financial services, insurance, retail, consumer and packaged goods, life sciences, manufacturing, telecom, utilities, and more.