Taking a FinOps approach to implementing Generative AI

Table of Contents

Generative AI offers transformative potential for enterprises — from accelerating product development to improving decision-making. However, the economic complexity of deploying and scaling generative AI remains a significant challenge. Enterprises must grapple with questions around cost management, workload placement, infrastructure sizing and long-term return on investment (ROI). To help organizations overcome these challenges, Dell Technologies and Tata Consultancy Services (TCS) have developed a joint solution that simplifies Generative AI (GenAI) adoption and implementation through a FinOps approach. The FinOps concept helps us demystify the granular spend for each token and data to realize the business value.

Based on a structured methodology that emphasizes operational cost visibility, workload optimization and model efficiency, TCS and Dell empower IT leaders to make informed decisions in the journey to a successful GenAI implementation. This paper outlines how a FinOps approach for GenAI, leveraging the Dell AI Factory and the TCS Adaptive AI-Ready Infrastructure Services, can help enterprises adopt GenAI at scale — while maintaining cost control, ensuring flexibility and supporting sustainability goals. Dell Technologies' industry-leading infrastructure and TCS' best-of-breed IT services and industry expertise enable organizations to align generative AI strategy with financial governance. With this FinOps approach for GenAI, enterprises can unlock innovation while maintaining economic sustainability across the entire GenAI ecosystem.

Why consider a FinOps approach to GenAI?

Cost savings and ROI

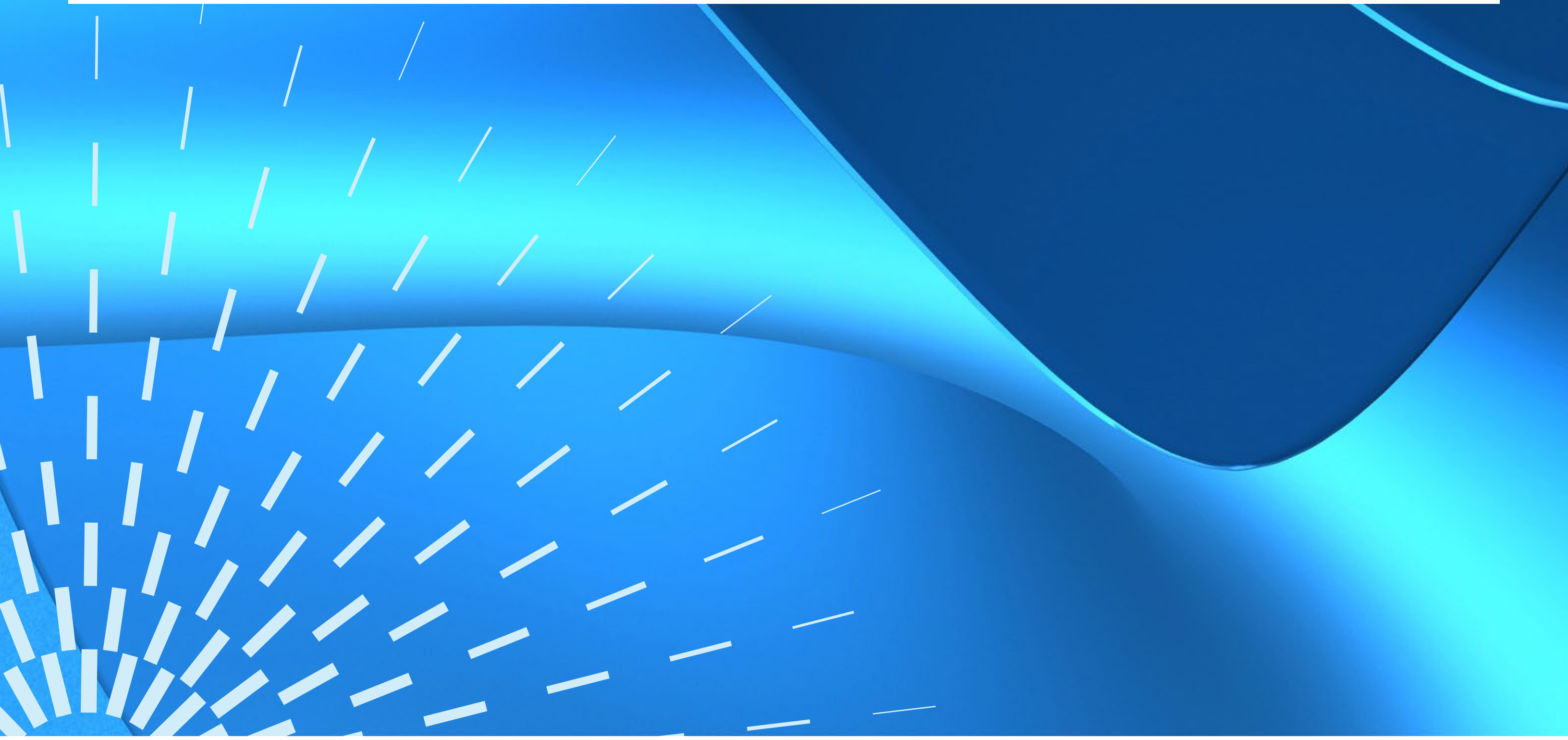
are the most common metrics used by organizations to measure effectiveness of AI initiatives.[1]

72%

of business executives face difficulties measuring the success of AI implementations, making it challenging to secure funding for additional projects — and **25%** need better financial key performance indicators (KPIs) for AI-enabled operations.[2]

27%

of financially successful technology companies work with external vendors for all or most of their AI implementation work — and companies that leverage external partners for AI implementations are **1.5x** more likely to report excitement or optimism about AI compared to companies using in-house resources.[3]



[1] "Navigating the Evolving AI Infrastructure Landscape," Enterprise Strategy Group Research Report, September 2023.

[2] "From potential to performance by design," TCS AI for Business Study Key Findings Report, 2024.

[3] Ibid.

Challenges in implementing GenAI

Enterprise-scale adoption of GenAI presents a complex landscape of challenges, many of which stem from the immense resource demands of AI models and the fragmented nature of existing IT infrastructure. Without a clearly defined FinOps strategy, enterprises risk overprovisioning, underutilizing resources and overspending. Collaboration between engineering, finance and business teams is critical to making timely, data-driven decisions and ensuring accountability.

The top challenges companies seeking to implement GenAI must address include:

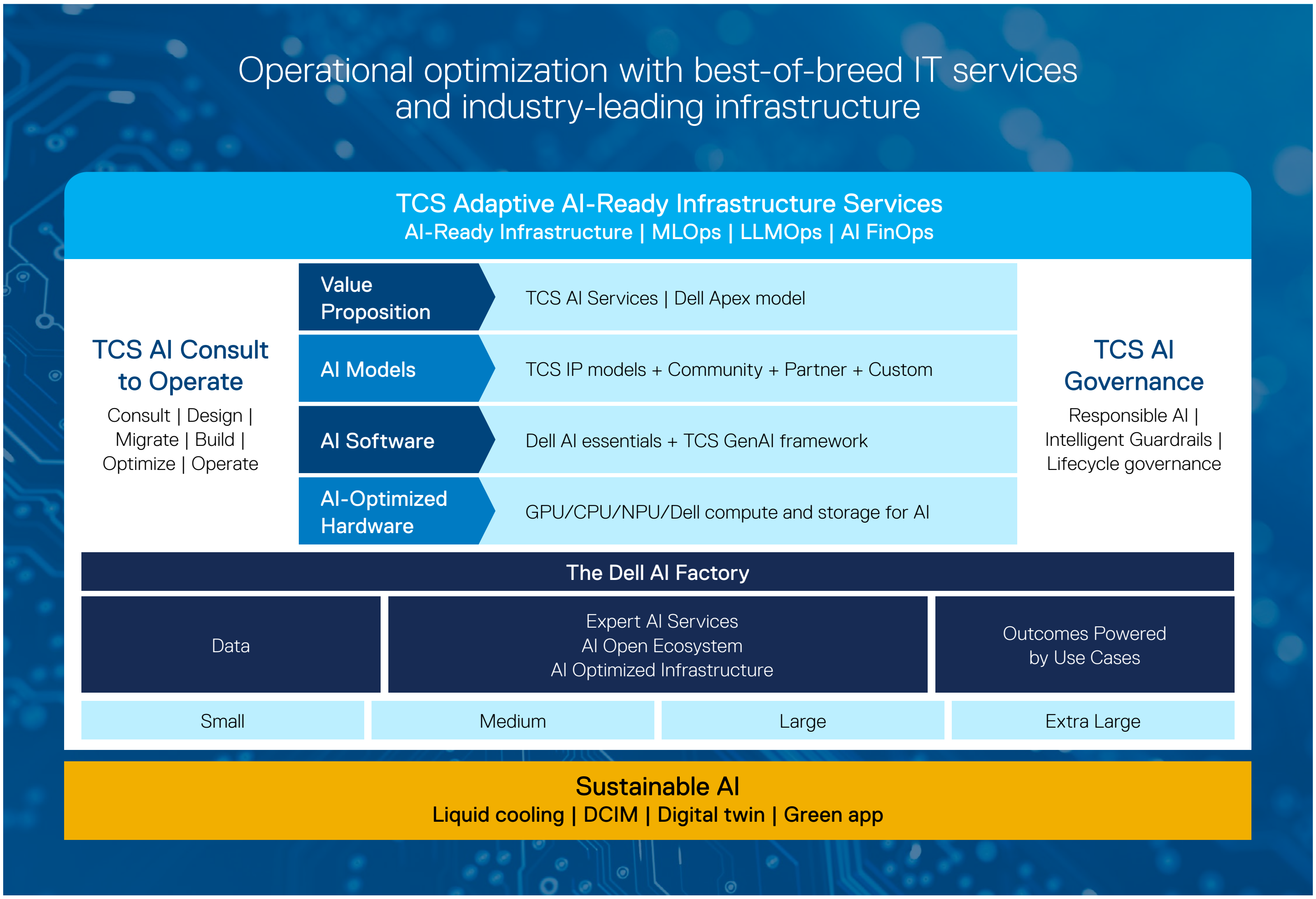
- **Token-based pricing models and high compute costs.** Unlike traditional infrastructure pricing, GenAI pricing is often variable, based on input and output tokens that are processed by a model. These pricing schemes make it difficult for companies to predict monthly expenses without robust monitoring and forecasting tools. In addition, deploying AI models requires high performance of GPU/NPU/TPU/CPU, along with significant memory and networking capacity — all of which can result in cost impacts. The increased demand for GPU-based hardware and scalable hybrid cloud services often draws in non-traditional buyers from business and operations teams, further complicating budget tracking.
- **Need for scalability.** While cloud platforms offer elasticity, this flexibility often comes with premium pricing — and on-premises environments can require substantial upfront capital for scaling, along with long lead times for procurement and deployment. Enterprises must evaluate total cost of ownership (TCO) and right-size their infrastructure to match workload needs. GenAI workloads can scale rapidly and unpredictably, making automation and modular system design critical to sustaining performance at scale.
- **Sustainability constraints.** GenAI workloads are energy intensive. High energy use not only increases operating costs but may also conflict with corporate environmental, social and governance (ESG) commitments. Power and cooling considerations are critical for aligning infrastructure with sustainability goals and delivering green data centers.
- **Delayed ROI.** Many enterprises struggle to move from pilot to production. They must determine precise infrastructure sizing without overprovisioning, which wastes capital, or underprovisioning, which limits model performance. Additionally, infrastructure needs can vary dramatically between training and inference, demanding nuanced planning to avoid unexpected expense. Without the right architectural guidance and integration strategies, timelines stretch and expected outcomes are delayed — impacting ROI realization. A lack of benchmarks and KPIs further hinders an enterprise's ability to demonstrate early business value.
- **Heightened risk.** Integrating GenAI with existing IT operations, data lakes and analytics platforms often introduces compatibility challenges — and misalignment between AI workloads and legacy infrastructure can introduce risks tied to security, privacy and operational consistency. The challenge is not solely geographic data residency, but rather the enterprise's ability to maintain private, restricted and secure access across global ecosystems. GenAI systems must ensure data is managed in compliance with evolving regional standards while still enabling global availability and governance. Ensuring consistent operational practices and maintaining alignment with local data privacy and regulatory requirements adds to this complexity.
- **Operational complexity.** GenAI systems are highly configurable, but this flexibility can introduce costly inefficiencies. Common pitfalls include over-resourcing, underestimating fine-tuning requirements, and deploying underperforming models or engineering prompts poorly — all of which increase costs and lower ROI. These issues are exacerbated when AI projects are siloed across departments without centralized architectural oversight.
- **Need to maintain competitive advantage.** Early adopters who successfully integrate scalable, efficient GenAI platforms stand to unlock significant competitive gains. These organizations are able to automate faster, respond more intelligently to market demands and personalize customer experiences with greater precision. The ability to move decisively from experimentation to production is a critical differentiator in high-growth sectors, where speed, insight and innovation determine market leadership.
- **FinOps misalignment.** GenAI expands the scope of financial accountability beyond traditional IT groups. New cross-functional stakeholders — including data science, product and line-of-business teams — often consume and provision AI resources. Without robust FinOps frameworks, this decentralization can result in unmonitored spending, poor anomaly detection and inadequate cost-performance benchmarking. Enterprises need accurate forecasting and budgeting, aligned to variable GenAI workloads and GPU utilization, to sustain innovation without financial friction.

Successfully overcoming these challenges demands both the technical know-how and a financial lens to evaluate architectural trade-offs, forecast long-term spending and define KPIs that align with business objectives.

How Dell Technologies and TCS confront GenAI

To support customers navigating technological complexity, rapidly evolving AI models and escalating infrastructure costs, Dell and TCS have co-developed a robust, enterprise-grade, FinOps-centered approach to GenAI deployment.

TCS and Dell span the full lifecycle of AI adoption — from initial consultation; through architectural design and performance benchmarking; to deployment, scaling and ongoing optimization. Whether operating in on-premises or hybrid environments, Dell and TCS' consult-to-operate model supports dynamic resource allocation and continuous value realization.



Operational optimization with best-of-breed IT services and industry-leading infrastructure

TCS Adaptive AI-Ready Infrastructure Services
AI-Ready Infrastructure | MLOps | LLMOps | AI FinOps

TCS AI Consult to Operate
Consult | Design | Migrate | Build | Optimize | Operate

Value Proposition — TCS AI Services | Dell Apex model

AI Models — TCS IP models + Community + Partner + Custom

AI Software — Dell AI essentials + TCS GenAI framework

AI-Optimized Hardware — GPU/CPU/NPU/Dell compute and storage for AI

TCS AI Governance
Responsible AI | Intelligent Guardrails | Lifecycle governance

The Dell AI Factory

Data

Expert AI Services
AI Open Ecosystem
AI Optimized Infrastructure

Outcomes Powered by Use Cases

Small | Medium | Large | Extra Large

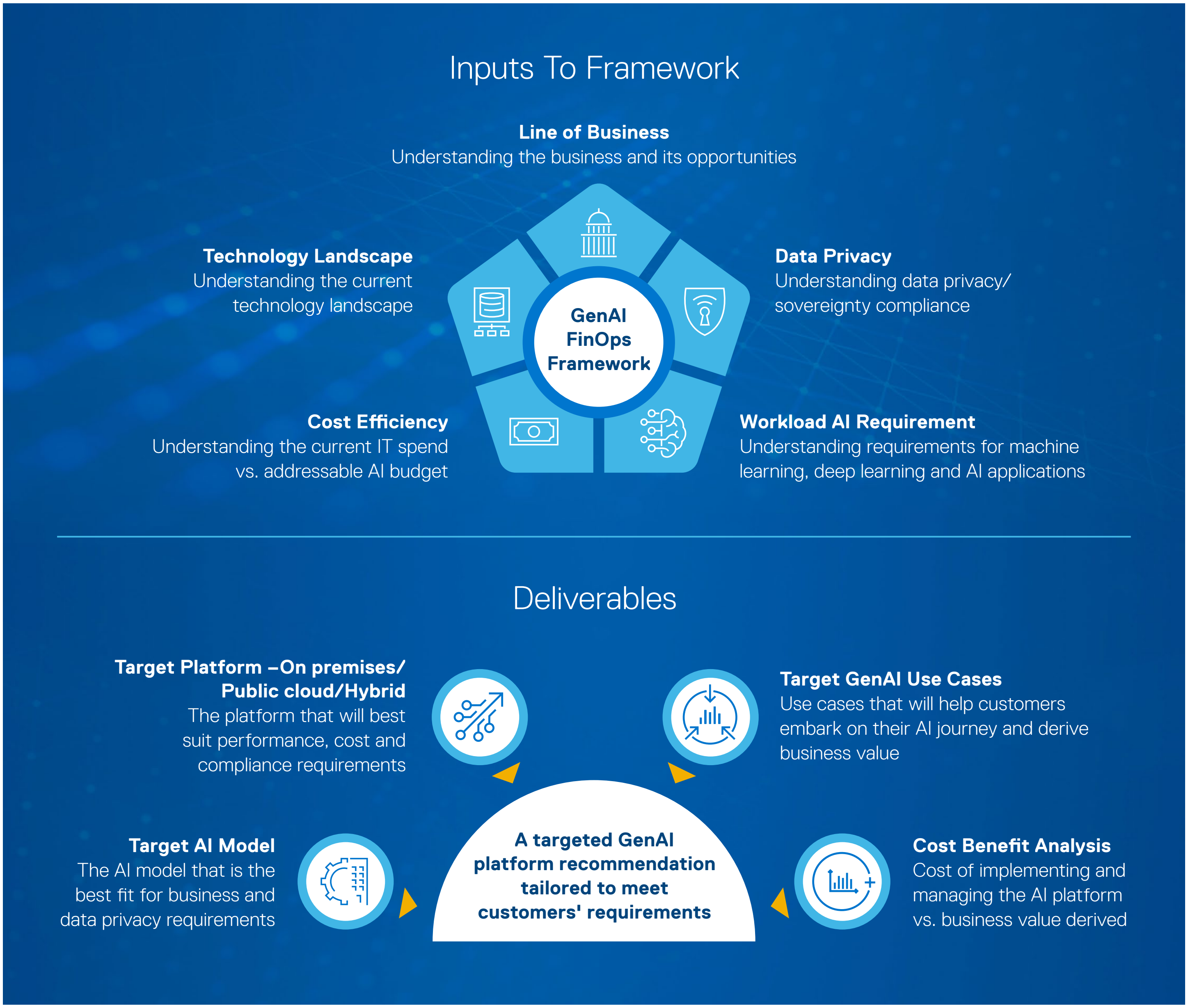
Sustainable AI
Liquid cooling | DCIM | Digital twin | Green app

TCS pairs its Adaptive AI-Ready Infrastructure Services with the Dell AI Factory to offer enterprises a unique FinOps roadmap for a successful GenAI journey.

- ✓ **TCS Adaptive AI-Ready Infrastructure Services** offers a modular, end-to-end suite of services designed specifically for GenAI workloads. It determines the right resources between GPU, NPU, TPU or CPU-accelerated compute tailored to demand, optimized per model selection, and underpinned by intelligent automation, orchestration toolsets, robust security protocols, advanced data governance and comprehensive managed services. With TCS Adaptive AI-Ready Infrastructure Services, companies can visualize the right resource deployment to help them accelerate time to deployment and support the long-term operational viability of their AI implementations, all while keeping costs in check.
- ✓ **The Dell AI Factory** brings a full-stack, validated infrastructure to the table, encompassing everything from AI-optimized servers to scale-out storage and advanced networking fabrics. It integrates data protection tools and AI orchestration platforms to ensure that enterprises can build GenAI systems that are both technically sophisticated and operationally efficient. This holistic infrastructure solution empowers organizations to launch AI capabilities that are robust, scalable, cost-effective and performance-tuned for demanding enterprise environments. This Dell Technologies Validated Design with an ecosystem of industry-aligned ISVs and Integrated Racking Solutions ensures complete peace of mind.

GenAI model and infrastructure optimization

A central task that enterprises must address in applying a FinOps approach to GenAI is to choose the right model and infrastructure for their generative AI needs — while ensuring cost control, robust optimization and a focus on sustainability goals. Enterprises pursuing GenAI must tightly align AI model selection with infrastructure design to optimize performance, scalability and cost-effectiveness.



Inputs To Framework

Line of Business
Understanding the business and its opportunities

Technology Landscape
Understanding the current technology landscape

Data Privacy
Understanding data privacy/ sovereignty compliance

GenAI FinOps Framework

Cost Efficiency
Understanding the current IT spend vs. addressable AI budget

Workload AI Requirement
Understanding requirements for machine learning, deep learning and AI applications

Deliverables

Target Platform –On premises/ Public cloud/Hybrid
The platform that will best suit performance, cost and compliance requirements

Target GenAI Use Cases
Use cases that will help customers embark on their AI journey and derive business value

Target AI Model
The AI model that is the best fit for business and data privacy requirements

A targeted GenAI platform recommendation tailored to meet customers' requirements

Cost Benefit Analysis
Cost of implementing and managing the AI platform vs. business value derived

By taking a FinOps approach to GenAI, Dell and TCS offer a co-engineered methodology that balances technical needs with economic realities. Key considerations in selecting the right AI model include:

- **Token economics and cost.** Organizations must analyze their input/output token volume and time-to-first-token (TTFT) for Single Shot, RAG, Chain of Thoughts (CoT) and Tree of Thoughts (ToT) to accurately predict their operational expenses.
- **Workload characterization.** Understanding training versus inference ratios, batch sizes, concurrency and task complexity is essential to inform infrastructure sizing.
- **Architectural alignment.** Enterprises need to prioritize model compatibility with their existing systems and explore optimization techniques, such as quantization, to minimize integration costs and resource consumption.
- **Data readiness.** Companies must also consider data quality, processing needs, sovereignty, security and the necessity for Retrieval Augmented Generation (RAG).

The economic value of GenAI initiatives

The goal of a FinOps approach to GenAI is to recommend an ideal solution based on an enterprise's business needs and privacy requirements. The process considers current IT spend and budget, privacy standards and protocols, types of workloads, and the larger technology landscape. To assist with this determination, a comprehensive analysis was conducted to see the economical value between on-premises, GPUaaS, Cloud APIs (both open weight and proprietary models), and Cloud hosting. A summary of the findings and observations is listed below:

On-premises shows the lowest average cost, with $0.18 per million tokens across all reasoning methods.

Cost per token for ToT queries is three to six times higher than for single-shot queries.

The average cost with GPUaaS is 15% more expensive than on-prem, at $0.21 per million tokens.

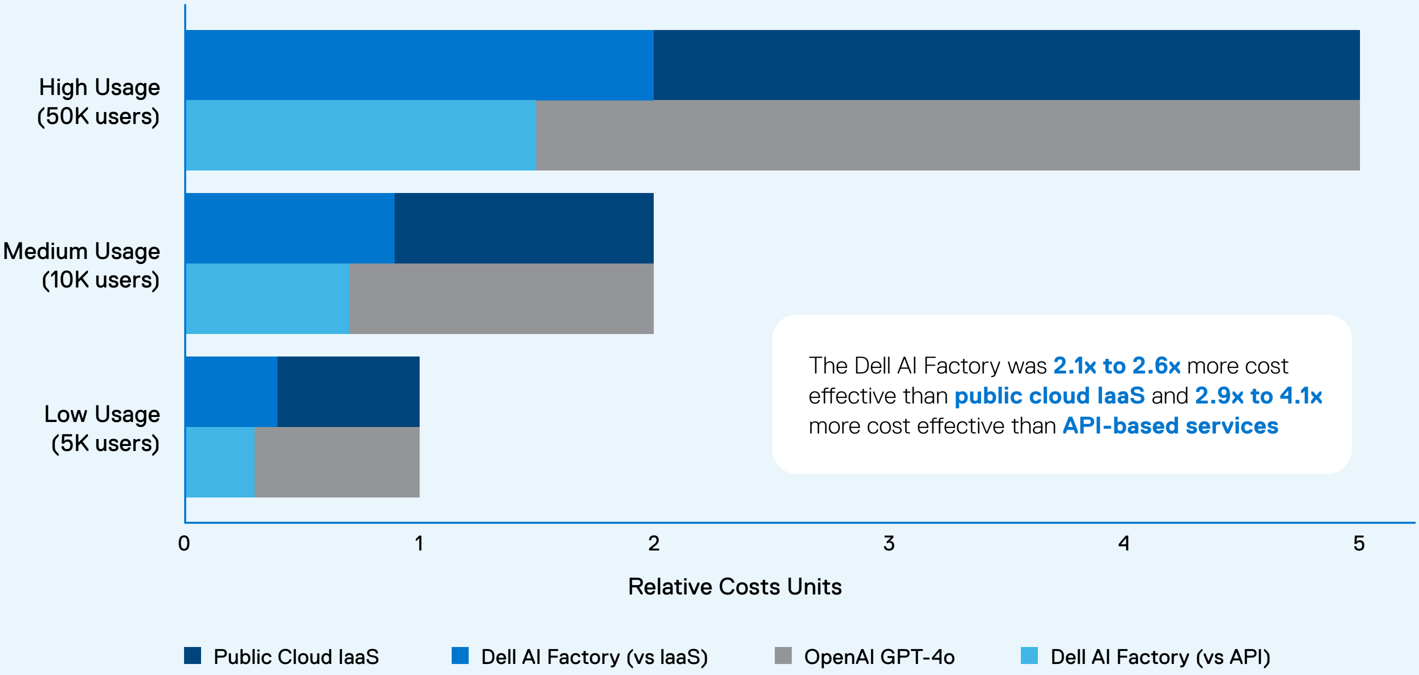
Cloud hosting is about twice as expensive than on-prem, with an average cost of $0.36.

Cloud API with open weight and propriety models costs an average of $0.39 (+100%) and $2.63 (+1,200%) more than on-prem.

QoS and guaranteed service availability need to be considered in addition to TCO for decision-making purposes.

In an economic white paper, the Enterprise Strategy Group (ESG) modeled how the costs to deliver inferencing for a 70 billion-parameter LLM compared across the Dell AI Factory, public cloud IaaS and the OpenAI API-based AI service GPT-4o. This analysis included the expected costs of monthly cloud spending, hardware, software, licensing, services, power and cooling, and infrastructure and model administration where applicable. Costs were modeled for a range of usage intensity, ranging from a low of 5,000 users to a high of 50,000 users. ESG's models found that the Dell AI Factory could provide inferencing 2.1x to 2.6x more cost effectively than public cloud IaaS and 2.9x to 4.1x more cost effectively than API-based services.

This figure represents the economics of ESG's four-year modeled cost to handle LLM inferencing across Dell AI Factory, public cloud IaaS and API-based services.



The Dell AI Factory was **2.1x to 2.6x** more cost effective than **public cloud IaaS** and **2.9x to 4.1x** more cost effective than **API-based services**

Source: "Economic White Paper: Understanding the Total Cost of Inferencing Large Language Models," Enterprise Strategy Group (now part of Omdia), April 2024.

The strategic benefit of a FinOps approach to GenAI

The Dell and TCS partnership delivers a unique and powerful FinOps-centered solution that transforms the way enterprises implement and scale GenAI. Unlike generic cloud or hardware offerings, this approach combines financial governance with technological innovation — enabling organizations to make smarter investment decisions, reduce waste and realize value faster.

A core strength of this solution lies in its ability to translate complex infrastructure choices into clear economic outcomes. Through a FinOps approach for GenAI, enterprises gain enhanced visibility into costs at every stage of the AI lifecycle — from planning and deployment to scaling and refinement. This cost transparency is critical to success as it exemplifies soft cost to data shuttling, which helps align infrastructure spending with actual business outcomes and ensure that resources are allocated efficiently and that every dollar invested supports growth.

Further, the integration of the Dell AI Factory and TCS's consult-to-operate methodology ensures that AI deployments are not only right-sized from the start but can optimally evolve over time with changing business needs. This built-in scalability prevents overprovisioning and avoids costly reconfigurations later. TCS and Dell have also jointly co-developed accelerators to help customers achieve their business goals with innovation, trust and speed. The journey leaps into Agentic AI and Physical AI to offer steady value at every stage of the roadmap.

With deep capabilities in predictive cost modeling, infrastructure optimization and workload management, this solution from Dell and TCS empowers enterprise IT teams and CFOs alike to manage GenAI deployments as strategic assets — with measurable ROI, operational accountability and long-term sustainability.

Benefits of the Dell and TCS GenAI solution

Benefit Area	TCS FinOps Framework for GenAI delivers:
Cost transparency	Clear cost metrics for infrastructure, models and token use, helping deliver meaningful value based on use cases while minimizing or helping plan for the financial resources required
Scalability	Flexible, validated architectures for growth — making 72–96 GPUs per rack achievable depending on configuration and customization, and accommodating high GPU and CPU power densities for next-generation AI workloads
Sustainability	Energy-efficient infrastructure aligned with ESG goals, including recycled materials, advanced cooling solutions, thermal management and power supplies that enable sustainable, higher-performance configurations and ensure high conversion efficiency (up to 96% at 50% load) to minimize energy consumption
Time-to-value	Accelerated deployment through automation and pre-validated designs, with 17G servers that support the Data Center Modular Hardware System standard, enabling flexible, modular deployment for efficient scaling and easier hardware upgrades
Risk mitigation	A solution that is secure by design, with embedded privacy controls like Secured Component Verification (SCV) and reporting on supply chain security to help customers ensure authenticity and compliance with best practices
ROI realization	FinOps modeling that supports investment justification, with modular design, extensive firmware/software upgrade support, software tools and high-density AI platforms to help maximize server operating life, enable proactive power management and provide more AI/ML performance per watt and per square meter, with best-in-class energy capture and re-use readiness
Efficiency	Optimized infrastructure sizing, GPU fractioning and cooling design, including high-density GPU servers that enable organizations to do more with fewer physical servers, reducing total hardware, power and environmental overhead
Competitive advantage	A joint Dell + TCS solution that combines the strength of trusted partnership and innovative practices for global customers

A unique, powerful collaboration for AI excellence

Successful GenAI adoption depends on an enterprise's ability to navigate new challenges with confidence — from unpredictable costs and compute-intensive workloads to integration complexity and organizational alignment. The partnership between Dell and TCS is designed to address these barriers head-on, providing a comprehensive solution that delivers measurable outcomes.

By combining enterprise-class infrastructure with consultative FinOps practices, the Dell and TCS approach ensures that GenAI implementations are cost-effective, scalable and sustainable. Customers benefit from modular, reusable architectures that support rapid deployment and future adaptability. With robust TCO modeling, performance benchmarking and real-time visibility into usage, enterprises can select high-impact use cases, forecast accurately and demonstrate tangible value across departments.

This collaboration empowers organizations to move from experimentation to enterprise-grade GenAI — with speed, control and accountability. Whether optimizing GPU utilization, securing global data environments or reducing time-to-value, the Dell and TCS approach enables teams to unlock the full business potential of GenAI.

Contact TCS for a GenAI/AgenticAI Readiness Assessment

Created with contributions from:

Vijayaraghavan Varadharajan, TCS
Murugeshwari M, TCS
Pavan Sonti, Dell Technologies
Mohammad Ghouse M, Dell Technologies

About Dell Technologies

Dell Technologies (NYSE:DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era.

About Tata Consultancy Services (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that has been partnering with many of the world's largest businesses in their transformation journeys for over 56 years. Its consulting-led, cognitive powered, portfolio of business, technology and engineering services and solutions is delivered through its unique Location Independent Agile™ delivery model, recognized as a benchmark of excellence in software development.

A part of the Tata group, India's largest multinational business group, TCS has over 607,000 of the world's best-trained consultants in 55 countries. The company generated consolidated revenues of US $29 billion in the fiscal year ended March 31, 2024, and is listed on the BSE and the NSE in India. TCS' proactive stance on climate change and award-winning work with communities across the world have earned it a place in leading sustainability indices such as the MSCI Global Sustainability Index and the FTSE4Good Emerging Index. For more information, visit www.tcs.com

DELL Technologies

tcs TATA CONSULTANCY SERVICES