Enterprise Strategy Group

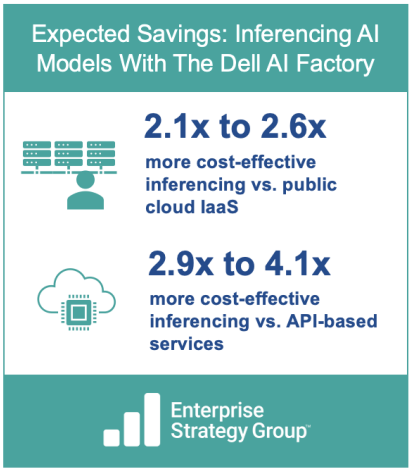
ECONOMIC SUMMARY

APRIL 2025

Optimizing AI TCO: Inferencing On Premises With the Dell AI Factory Can Be 2.6x More Cost-effective Than Public Cloud

Aviv Kaufmann, Practice Director and Principal Validation Analyst

Read the full report HERE.



Abstract: Enterprise Strategy Group modeled and compared the expected costs to inference large language models (LLMs) on the Dell AI Factory versus using native public cloud IaaS or the OpenAI GPT-4o LLM model service through an API. We found that Dell Technologies could provide LLM inferencing on premises up to 2.6x (62%) more cost-effectively than public cloud and up to 4.1x (75%) more cost-effectively than with API-based services.

Challenges for Enterprises

To ensure effective outcomes, IT and business teams need to align toward a centralized strategy for AI that makes it possible to bring together and effectively process the data and information contained across all business processes, resources, tools, and locations. However, organizations face challenges around implementing AI. Enterprise Strategy Group research found that the top five challenges that organizations face when implementing AI are high costs associated with the implementation; data management and/or data quality issues; concerns over data privacy, protecting intellectual property, and security; difficulty integrating with existing systems and processes; and a lack of development expertise and talent.[1] When planning where to deploy AI models, organizations must consider data and storage requirements, performance and scalability, ease of management, cost, and time to value.

The Solution – The Dell AI Factory

The Dell AI Factory is an enterprise-ready approach designed to help organizations adopt and scale AI in an easier, more effective, and secure way. It brings together Dell's AI infrastructure and services with leading-edge software and acceleration technologies from an open ecosystem of partners to simplify the implementation and management of AI across an organization. It is built to address the common challenges that organizations face when operationalizing AI, such as high infrastructure costs, complex data management, and security risks. The Dell AI Factory consists of five foundational pillars—data, infrastructure, software, services, and use cases—ensuring that they work together to deliver a comprehensive, end-to-end enterprise AI solution. The Dell AI factory provides end-to-end solutions, flexible AI infrastructure, modern data management, robust security and compliance, and expert services and support.

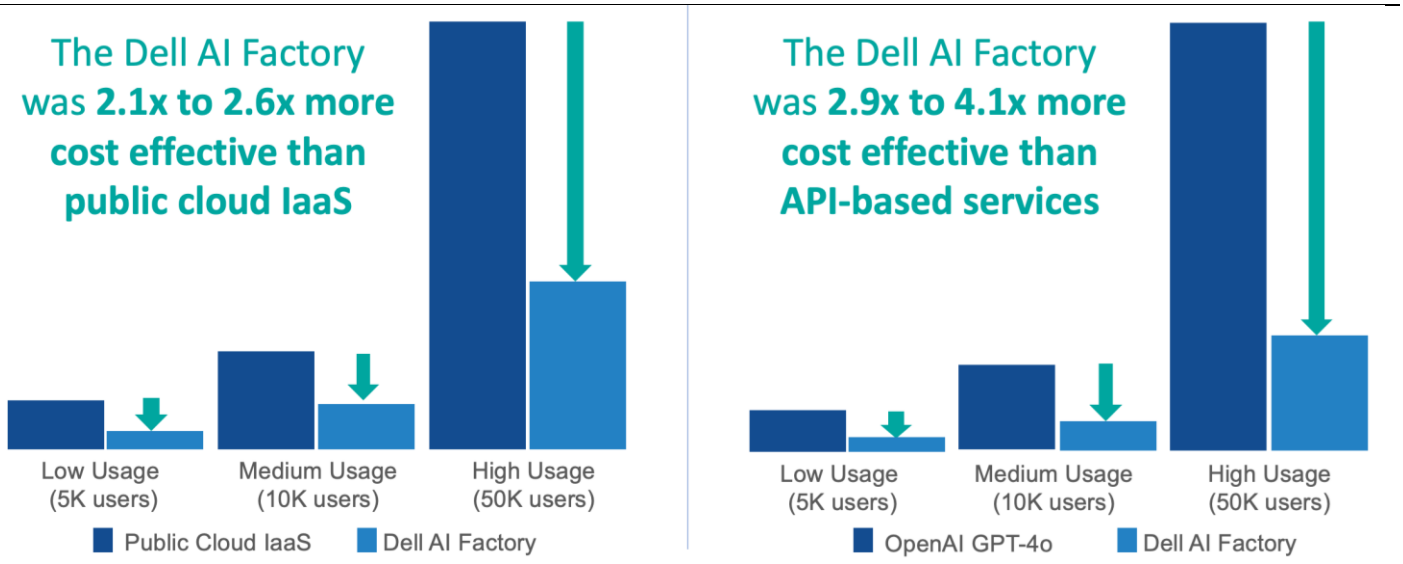
To learn more about Dell's solutions, visit the Dell AI Factory webpage.

[1] Source: Enterprise Strategy Group Report, *Navigating Build-versus-buy Dynamics for Enterprise-ready AI*, January 2025.

Economic Validation Highlights

Enterprise Strategy Group modeled the costs over four years to deliver inferencing for a 70 billion parameter LLM on the Dell AI Factory, on public cloud IaaS, and using the OpenAI API-based AI service GPT-4o. Our analysis included the expected costs of monthly cloud spending, hardware, software, licensing, services, power and cooling, and infrastructure and model administration where applicable. We modeled these costs for a range of usage intensity ranging from 5,000 users (low usage) to 50,000 users (high usage). **Our 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.** Our full report and appendices compare the LLM deployment options and key considerations as well as explain the modeled assumptions and costs in greater detail.

Figure 1. Enterprise Strategy Group's 4-year Modeled Cost to Handle LLM Inferencing



Source: Enterprise Strategy Group, now part of Omdia

Conclusion

The decision of where and how an organization deploys its LLM today lays the groundwork to enable scalability and flexibility in the future to seamlessly support growth toward expanded use cases, more users, more models, agentic AI, and future AI technologies. Enterprise Strategy Group strongly recommends that companies looking to implement powerful LLMs consider taking advantage of the cost-effective technologies and knowledgeable services that Dell Technologies provides to ensure a successful outcome, accelerate their AI initiatives, and reduce the time to achieve these expected savings.

Click HERE to learn more about the Dell AI Factory.

About Enterprise Strategy Group
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com