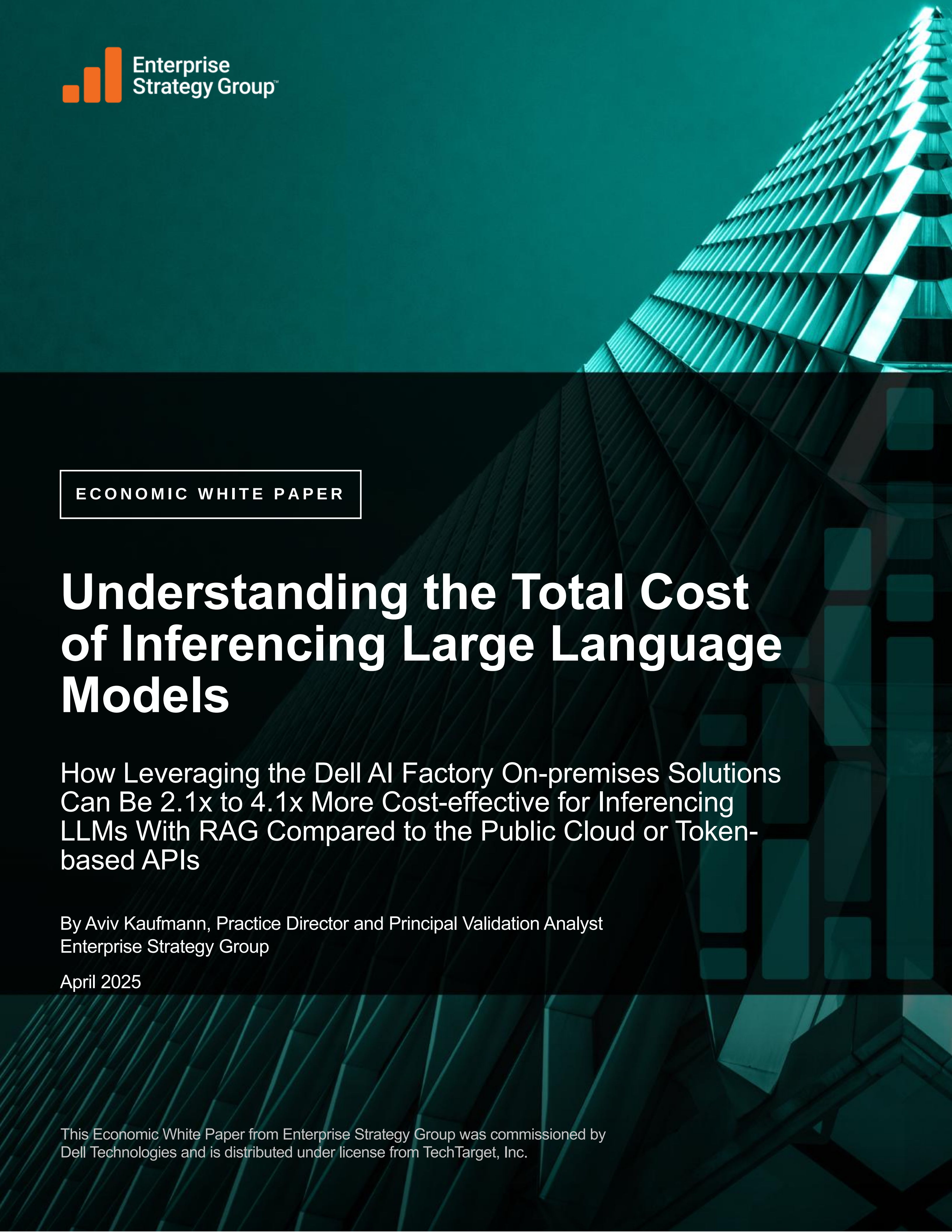


ECONOMIC WHITE PAPER

Understanding the Total Cost of Inferencing Large Language Models

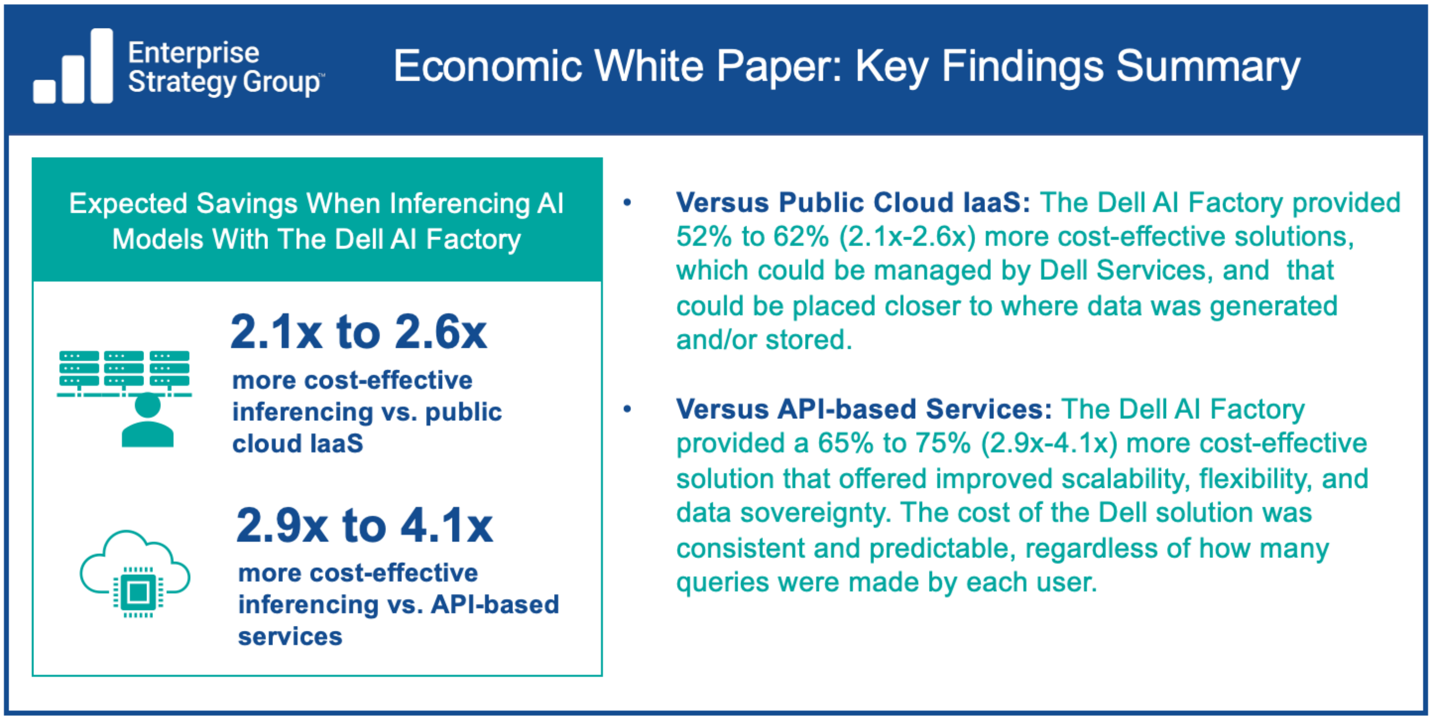
How Leveraging the Dell AI Factory On-premises Solutions Can Be 2.1x to 4.1x More Cost-effective for Inferencing LLMs With RAG Compared to the Public Cloud or Token-based APIs

By Aviv Kaufmann, Practice Director and Principal Validation Analyst
Enterprise Strategy Group

April 2025

This Economic White Paper from Enterprise Strategy Group was commissioned by Dell Technologies and is distributed under license from TechTarget, Inc.

Contents



Enterprise Strategy Group

Economic White Paper: Key Findings Summary

Expected Savings When Inferencing AI Models With The Dell AI Factory

2.1x to 2.6x

more cost-effective inferencing vs. public cloud IaaS

2.9x to 4.1x

more cost-effective inferencing vs. API-based services

- **Versus Public Cloud IaaS:** The Dell AI Factory provided 52% to 62% (2.1x-2.6x) more cost-effective solutions, which could be managed by Dell Services, and that could be placed closer to where data was generated and/or stored.
- **Versus API-based Services:** The Dell AI Factory provided a 65% to 75% (2.9x-4.1x) more cost-effective solution that offered improved scalability, flexibility, and data sovereignty. The cost of the Dell solution was consistent and predictable, regardless of how many queries were made by each user.

Introduction

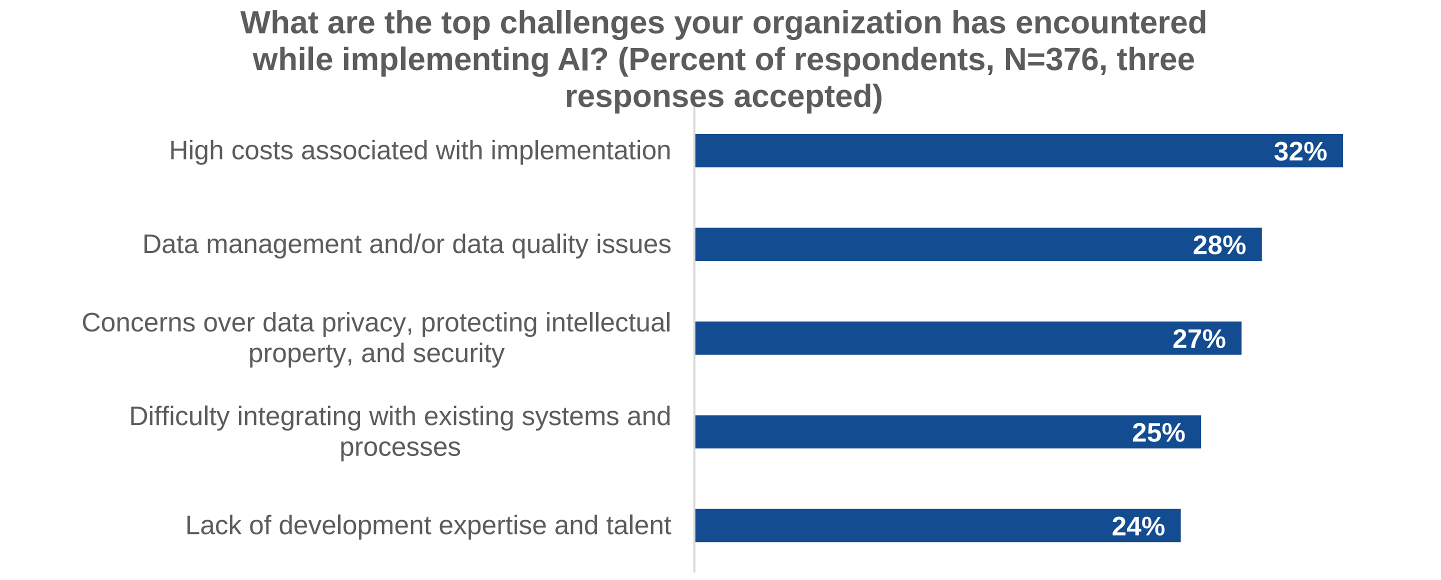
This Economic White Paper from Enterprise Strategy Group presents some options and considerations for delivering text-based generative AI (GenAI) capabilities to organizations. Enterprise Strategy Group modeled and compared the expected costs to inference large language models (LLMs) utilizing retrieval-augmented generation (RAG) on Dell Technologies' the Dell AI Factory versus using native public cloud infrastructure as a service (IaaS) or the OpenAI GPT-4o LLM model service through an API. We found that the Dell AI Factory could provide LLM inferencing up to 2.6x more cost-effectively than IaaS and up to 4.1x more cost-effectively than with GPT-4o API.

AI Overview

With more data being generated than ever before, the effective use of AI is a critical success factor for any organization. AI provides organizations with actionable insight and automation capabilities that help to improve business operations, accelerate innovation, ensure more efficient operations, reduce costs and risks, and improve productivity. 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 organizations faced when implementing AI were 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]

Figure 1. Top Five Challenges Encountered While Implementing AI



Source: Enterprise Strategy Group, now part of Omdia

LLMs can be costly and complex to develop, but organizations can easily augment, fine-tune, and customize existing open source LLMs to meet their needs. Text-based LLMs focus on learning, understanding, and producing

[1] Source: Enterprise Strategy Group Report, *Navigating Build-versus-buy Dynamics for Enterprise-ready AI*, January 2025. All Enterprise Strategy Group research references and charts in this Economic White Paper are from this report unless otherwise noted.

© 2025 TechTarget, Inc. All Rights Reserved.

content, answers, summaries, and questions that can be tailored to a particular industry, use case, and organization. RAG augments the results of AI models with custom data pulled from additional sources, which makes the models more accurate. These are the most deployed LLMs for businesses and can be used for chatbots, Q&A assistants, process improvement and automation, or as capabilities built into custom tools and applications, in addition to many other use cases. The requirements and ROI of use cases vary, and organizations should consider the benefits of different solutions available to address priority use cases that deliver ROI.

When planning where to deploy LLM models, organizations must consider infrastructure for training (i.e., data- and compute-intensive analysis required to build an effective model), inferencing (i.e., servicing user interactions on a trained model), and fine-tuning (i.e., continually updating and optimizing the model). This report focuses on the solutions required to facilitate inferencing workloads. Several deployment methods can be used for inferencing LLMs, including:

LLM Deployment Options:

- On-premises infrastructure.
- Public cloud IaaS.
- LLM API services.

- **On-premises infrastructure.** Purchased or leased infrastructure consisting of compute, memory, GPUs, and storage can be deployed and managed on premises or at colocation facilities along with a commercial or open source AI platform, affording the organization control over all aspects of the deployment. While some vendors offer subscriptions, this method generally requires an upfront investment in both time and capital and might be the most cost-effective for larger and more predictable workloads. This is because AI can be processed closer to where an organization's data is generated, and overall costs can be reduced while maintaining control of data and security. Additionally, cost-effective AI PCs can be leveraged on premises and at the edge, making on premises a viable option even for small deployments. It is not surprising that Enterprise Strategy Group found that 73% of organizations preferred a self-deployed approach based on physical infrastructure deployed at data centers, devices, colocation facilities, and edge locations.
- **Public cloud IaaS.** Similarly, organizations can deploy equivalent cloud compute instances with GPUs and storage along with a commercial or open source AI platform. This method gives similar control over the AI platform, with agility and easy integration with existing tools. This method might be the most cost-effective for those with unpredictable or seasonal requirements, but for larger deployments and continuous usage, it can be significantly more costly while adding significant latency to workloads. Also, organizations cannot ensure data sovereignty or complete ownership of security and the underlying infrastructure.
- **LLM API services.** Established services like OpenAI GPT can be used to quickly provide capabilities without having to manage infrastructure or an AI platform. This method might be the best for exploring and getting started, smaller deployments that might not make sense to run on AI PCs, and those that do not require customization or control, but these models offer less flexibility while bringing their own latency and data security challenges, and costs can quickly add up, especially for heavier usage and larger LLMs.

Key Considerations of LLM Inferencing

Before deciding on an LLM platform and where to place AI, organizations should invest time to understand their requirements and capabilities, as well as discuss some of the following considerations around choosing a platform and location for LLM inferencing. Return on investment (ROI) is the most common metric that organizations use to measure the post-implementation success of their enterprise-ready AI solution. The use cases that often have the biggest impact on ROI are those tied to an organization's proprietary data. While this report focuses on **cost savings and ROI**, many other considerations should influence AI deployment location, including:

- **Data proximity and sovereignty.** Data fuels effective LLMs, and understanding where it is generated, resides, and processed is critical. Organizations must prioritize data sovereignty, ensuring that data remains within specific jurisdictions to comply with local regulations and strengthen security. Enterprise Strategy Group research found that 71% of AI infrastructure is deployed outside the public cloud.[2] By bringing AI processing closer to where the data exists, organizations can maintain control, avoid duplicative transfers between systems or locations, and reduce time and costs associated with data movement. This approach supports both operational efficiency and regulatory compliance while safeguarding sensitive information.

- **Data security.** Ensuring strong organizational control and maintaining visibility over security are critical to safeguarding AI solutions. Protecting intellectual property (IP) and preventing unauthorized access, theft, or damage by malicious actors is essential to avoid disruptions in model performance and maintain trust in the system.
- **Data governance.** Organizations must consider the location and data governance requirements of the sources of the data that is required to train and maintain the model. Hybrid cloud infrastructure will work best when data resides locally and is accessible where it is needed. Training on and maintaining data that is up-to-date, comprehensive, and unbiased will produce a better LLM and more accurate insights derived from inferencing.
- **Storage requirements.** Proper sizing and capabilities of a storage solution or service must be considered for data input, outputs, and temporary space as well as to store, protect, and provide version history for the AI platform. Storage is even more important when supporting multimodal AI (text, audio, video etc.). The storage solution should exceed expectations for I/O throughput and bandwidth requirements, as well as provide availability and data protection services to ensure quick backups, snapshots, and restores.
- **Latency and performance.** Organizations should ensure that business users are able to access low-latency models from wherever they are using AI. Sizing the infrastructure with enough resources in processors, GPUs, memory, storage, and network capabilities is important to ensure that it can handle the expected concurrency of inferencing at normal and peak loads and that average inference latency is low enough to give users a positive experience. Organizations should also determine if compute-intensive training and fine-tuning of the LLM will happen on the same platform or on a higher-performance dedicated training platform before being moved to the inferencing platform.

LLM Key Considerations:

- Data and storage requirements.
- Performance and scalability.
- Ease of management.
- Cost and time to value.

- **Time to value.** It is important to understand how long the solution will take to plan, size, acquire, deploy, configure, and test before serving production workloads and providing value to the business. On-premises infrastructure can take longer to deploy than public cloud-based solutions, but this time can be reduced with ready-made configurations, sizing tools, and expert services.
- **Energy and physical capacity.** On-premises and colo deployments must consider power, cooling, and floor space requirements and costs. Consolidation into denser compute and storage platforms with built-in data efficiency (deduplication/compression) services can reduce these costs. Modern IT solutions are designed with lower power consumption in mind which can result in more available power at data centers, colo, and edge locations to add efficient AI solutions. Additionally, as GPUs become more efficient and models become more targeted, energy needs might also be less than initially anticipated.
- **Scalability and flexibility.** Understanding and predicting how many users will access the tool and how often they will ask questions per day is an important metric to consider when choosing a solution. If demand is small, an API service and AI PCs may suffice, but as an organization supports more users and use cases, building a proprietary platform will become more cost-effective. Organizations must consider expected growth in adoption, usage frequency over time, and hardware and site capabilities to ensure that infrastructure is sized appropriately and platforms are flexible enough to support the evolving needs of the business and future AI technologies.

[2] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

- **Open partner ecosystem.** AI platforms should be built to support an open ecosystem of software vendors and implementation partners that can help provide point solutions, tools, and services that will help ensure the success of the AI implementation today and in the future. Proprietary solutions limit flexibility and agility going forward, while an open ecosystem of partners can provide best-of-breed options for each component of the platform.
- **Ease of management.** When comparing any on-premises infrastructure to cloud infrastructure and services, it is important that an organization considers its in-house capabilities and understands the costs of managing and maintaining the infrastructure and platforms. Unified management interfaces, orchestration, and automation can help reduce administrative efforts. Leveraging managed services and colocation options can enable organizations to get many of the benefits of hosting in their own data centers while offloading the resources and skills required to operate the infrastructure and platform.

To ensure success, all of these considerations should be continuously contemplated, no matter where an organization chooses to deploy its AI, whether on devices like laptops or workstations, in on-premises data centers or colocation facilities, at edge locations, on public cloud IaaS, or using an LLM API service. While not included in this analysis, it should also be noted that, in addition to up-front capital investments, subscription models may be available for various deployment options.

Enterprise Strategy Group Economic Analysis

Enterprise Strategy Group created an economic analysis that compared the expected costs of delivering inferencing for a 70B parameter open source LLM utilizing RAG with different utilization rates (with number of users between 5K and 50K). We assumed that the model was providing an AI-powered chatbot and that inferencing occurred where the data was located to minimize data migration costs. The analysis looked at all the costs associated with running and inferencing the models over a four-year period, including providing and running the infrastructure, administering the systems, and paying for services if required.

Dell On-premises Infrastructure Versus Public Cloud IaaS

Our models first compared the expected cost to run LLM inferencing on traditional infrastructure (on premises, in colocation environments, at edge locations, etc.) to running on a similarly configured public cloud IaaS on Amazon EC2 instances. The inferencing node server and NVIDIA H100 GPU configuration requirements were sized for each workload based on the results of inference baseline testing to ensure they could handle concurrency requirements at regular and peak load (based on maximum requests and number of model instances) as well as provide adequate latency and throughput for each expected workload. We then modeled each of the costs for both the Dell infrastructure (implemented, supported, and managed by Dell Services) and the equivalent EC2 configuration. These costs included the initial cost of acquisition (including hardware, software, services), power and cooling, monthly cloud spending, NVIDIA AI Enterprise licenses, infrastructure/instance administration, ML platform administration, and ML model administration. The costs and assumptions are summarized in Appendix Table A1.

We modeled the costs to deliver the open source Llama 3 70 billion parameter LLM. A text-based chatbot served as the workload we explored to size the hardware and instance requirements. This workload requires moderate token intensity per query, does not have a lot of variances in the peak load, and results in a relatively even balance between the number of input and output tokens required. The workload assumed an average of 3,000 tokens (input and output combined) per query and 50 queries per user per day. Based on our research of public statements, we found this to be a moderate number of queries per user, with less-established organizations generating fewer queries/user/day and more-established organizations generating more queries/user/day. To size the requirements, we predicted the server, storage, and GPU configurations that would be capable of providing sufficient throughput and low latency requirements to handle each number of users. The high-level assumptions for instance and GPU counts are shown in Table 1.

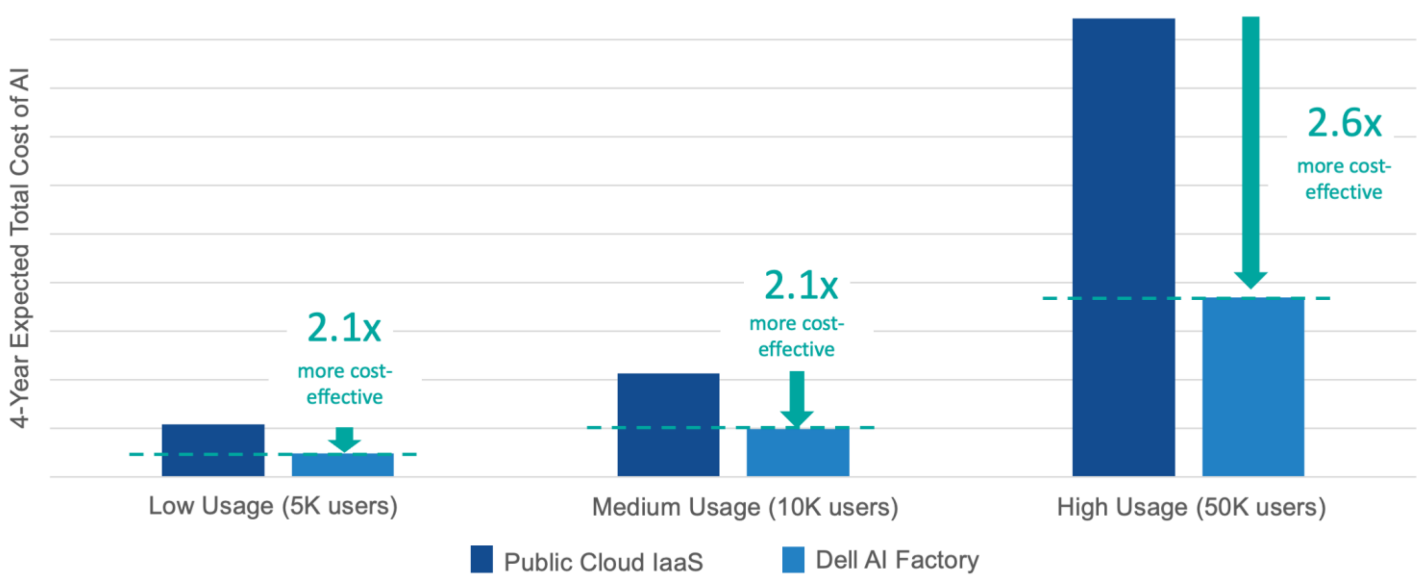
Table 1. Configuration Assumptions for the Llama 70B Parameter Model Inferencing

LLM Model (Number of Parameters)	Number of Users	Number of Inferencing Nodes/Cloud Instances	Total Number of H100 GPUs
Llama 3 (70B)	5,000	2 for Dell AI Factory or a single cloud instance for IaaS	8
	10,000	2	16
	50,000	9	72

Source: Enterprise Strategy Group, now part of Omdia

We then modeled all the costs mentioned above for each configuration. As shown in Figure 2, the Dell AI Factory was 2.1x to 2.6x (52% to 62%) more cost-effective at delivering inferencing. A detailed cost breakdown can be found in the Appendix of this report.

Figure 2. Expected 4-year Cost to Deliver Inferencing for 70B Parameter Llama 3 LLM Using RAG



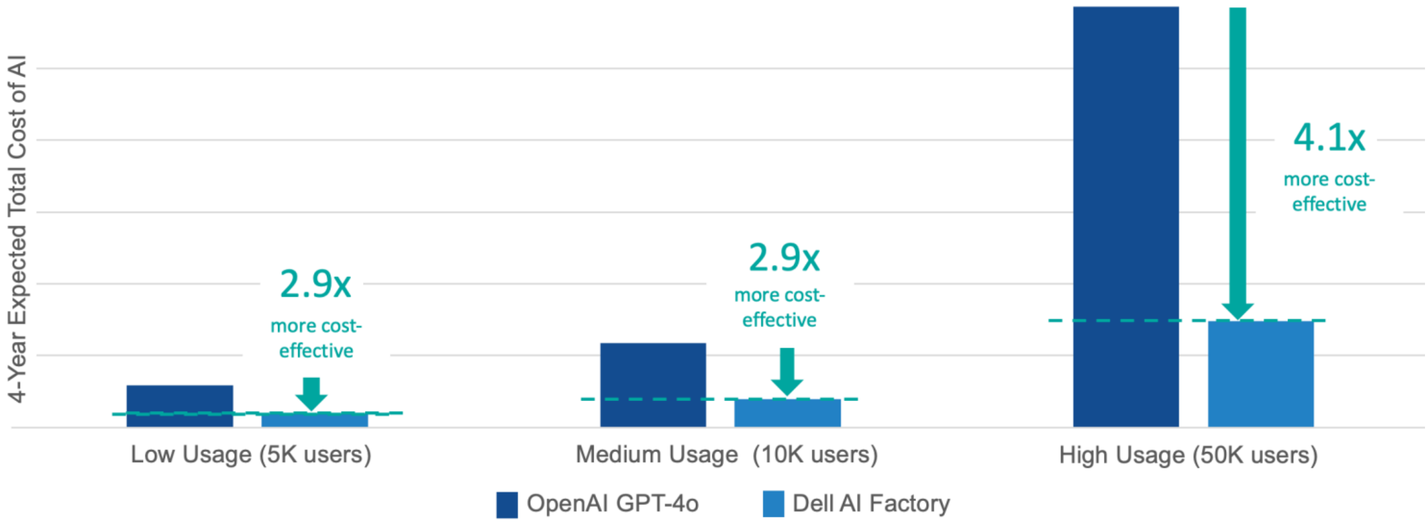
Source: Enterprise Strategy Group, now part of Omdia

Dell On-premises Infrastructure Versus API-based AI Service

We then compared the expected costs for a large organization to provide an equivalent 70 billion parameter model to 5,000 to 50,000 users using the established OpenAI API-based AI service GPT-4o, which is priced per input and output token.

For the 70 billion parameter model, our GPT-4o calculations predicted a cost of about $12.19/user/month, which compares favorably to suite-based AI assistance tools that currently cost $20-$30/user/month. With these assumptions, we found that Dell Technologies on-premises infrastructure could provide inferencing 2.9x to 4.1x (65% to 75%) more cost-effectively than using an API-based service, delivering AI capabilities for only about $3.00/user/month to $4.28/user/month for the 70B model.

Figure 3. Expected 4-year Cost to Deliver Inferencing for 70B Parameter LLM



Source: Enterprise Strategy Group, now part of Omdia

Issues to Consider

While Enterprise Strategy Group's models are built in good faith upon conservative, credible, and validated assumptions, no single modeled scenario will ever represent every potential environment. Customer savings will depend on each organization's particular use case, the nature of its data, its level of expertise, its model, and its infrastructure requirements. Following are a few more important factors to consider when modeling and comparing the expected costs of the Dell AI Factory against IaaS and API-based services:

- **Deployment location.** We expect that the Dell AI Factory deployed at colocation facilities would provide similar total cost of ownership advantages as modeled for on premises, with a slightly different cost structure.
- **Scalable hardware.** While the analysis of this paper focused on solutions leveraging NVIDIA's H100 GPUs, future solutions are expected to be even more performant. At the time of analysis, Dell had various H200 GPU-enabled solutions available, but the public cloud had limited availability and public cloud configurations continued to be more expensive than Dell's on-premises solutions. We expect that future technologies will likely continue to be more cost-effective at scale on-premises, an important consideration with an expected increase in adoption of agentic AI and the anticipated increase in computing needs going forward—even with hardware and models expected to become more efficient over time.
- **Smaller LLMs.** For workloads that can leverage smaller LLM models (e.g. 7 billion parameter models), AI PCs offer a powerful, cost-effective alternative by delivering high-performance AI capabilities directly on device. With localized processing, businesses can leverage low latency, minimized data transfer costs and maintain localized control over sensitive information.
- **Model tuning and optimization.** With IaaS and on-premise configurations, organizations have control over quality of service (QoS), model tuning and optimization, and the ability to grow resources as needed to improve model performance or reduce latency. API services have only rudimentary QoS and provide no guarantee on inferencing response times. User experience can vary.

Enterprise Strategy Group recommends that organizations perform their own analysis of available products and consult with Dell Technologies to understand and discuss the differences between the solutions proven through their own proof-of-concept testing.

Dell AI Factory for LLM Inferencing

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. Key capabilities of the Dell AI Factory approach include:

- **End-to-end integrated solutions.** The approach combines Dell PowerEdge servers, GPUs, software, and services into a stack for easier deployment and optimized AI performance.
- **Flexible AI infrastructure.** It offers scalable, workload-specific compute, storage, data protection and networking options. For compute, Dell PowerEdge servers and Dell Pro Max AI PCs span diverse capabilities and cost while meeting the needs of AI training and inferencing, using modern and energy-efficient hardware, to power diverse workloads, such as content generation, digital assistants, data analytics, computer vision, and digital twins.
- **Modern data management.** The Dell AI Data Platform helps organizations discover, prepare, and govern data efficiently using Dell PowerScale storage, Dell Data Lakehouse integrated tools, and professional services.
- **Robust security and compliance.** Built-in features support data protection, IP security, and regulatory compliance, making the platform suitable for sensitive environments.
- **Expert services and support.** Dell provides expert guidance, implementation, and support of AI platforms, as well as training and certifications to help organizations move confidently from strategy alignment and pilot projects to full-scale production.

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

Conclusion

The expanded use of AI across nearly every area of the business is a crucial factor for ensuring improved operations and future success. Enterprise Strategy Group research found that the top business drivers for implementing AI include operational efficiency, customer experience, innovation, risk management, cost reduction, and decision-making. Organizations can achieve more impactful and meaningful results by training and inferencing against their own customized version of an LLM. 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 other future AI technologies.

Several deployment methods can be used for inferencing LLMs, and each provides advantages for particular use cases and requirements. For larger organizations with thousands of users ready to take advantage of the capabilities contained in a customized LLM, the Dell AI Factory can make planning, sizing, deploying, and managing much simpler, and Dell can provide high-performance LLM inferencing up to 2.6x more cost-effectively than IaaS and up to 4.1x more cost-effectively than with API-based services. And an investment in the Dell AI Factory can provide even better economics against increased usage and higher inferencing intensity. 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.

Appendix

Table A1. Costs and Assumptions Modeled for Dell Technologies Versus Cloud IaaS Comparison

Cost Category	Dell AI Factory (On Premises)	Public Cloud IaaS (Amazon EC2)	API-based AI Service (OpenAI GPT-4o)
Initial cost of acquisition (hardware / software / services)	Price provided by Dell: PowerEdge R660 (login/header node), PowerEdge R760xa (for 5k user scenario), PowerEdge XE9680 (for 10k and 50k user scenarios), NVIDIA Spectrum X networking and support, and Dell Services (consult, deploy, and support)	N/A	N/A
Power and cooling cost	Calculated based on system specifications ($0.162/kWh), with 5% increase annually	N/A	N/A
Monthly cloud spending	N/A	p5.48xlarge EC2 instance costs calculated based on 3-year reservation discounts	Expected costs based on 3,000 tokens per query and 50 queries per day per user.
NVIDIA AI Enterprise License/GPU	Based on 5-year license (prorated for 4-years)	Per instance/h, based on 18 h/day to limit costs	N/A
Infrastructure/instance administration	Included in cost of services (Dell Services)	Modeled (7%-67% of systems engineer based on model complexity)	N/A
ML platform administration	Modeled Included in cost of services (Dell Services)	Modeled (10%-50% of ML engineer based on model complexity)	N/A
ML model administration	Modeled (10%-50% of ML engineer based on model complexity)	Modeled (10%-50% of ML engineer based on model complexity)	Modeled (2%-10% of ML engineer based on model complexity)

Source: Enterprise Strategy Group, now part of Omdia.

Table A2. Modeled Costs for 70B Parameter Model With 5K Users (Low Usage) Over 4 Years

Cost Category	Dell AI Factory (On Premises)	Public Cloud IaaS (Amazon EC2)	API-based GenAI Service (OpenAI GPT-4o)
Hardware / Software / Services (Inc. Cloud Costs, NVIDIA licenses, and Dell Services)	$917K	$2.027M	$2.925M
Power and cooling cost	$59K	$0	$0
Infrastructure/instance and ML platform administration	$0 (included in Dell Managed Services cost)	$63K	$0
ML model administration	$0 (included in Dell Managed Services cost)	$52K	$10K
Total	**$976K**	**$2.142M**	**$2.935M**

Source: Enterprise Strategy Group, now part of Omdia.

Table A3. Modeled Costs for 70B Parameter Model With 10K Users (Medium Usage) Over 4 Years

Cost Category	Dell AI Factory (On Premises)	Public Cloud IaaS (Amazon EC2)	API-based GenAI Service (OpenAI GPT-4o)
Hardware / Software / Services (Inc. Cloud Costs, NVIDIA licenses, and Dell Services)	$1.651M	$4.054M	$5.850M
Power and cooling cost	$307K	$0	$0
Infrastructure/instance and ML platform administration	$0 (included in Dell Managed Services cost)	$100K	$0
ML model administration	$0 (included in Dell Managed Services cost)	$77K	$15K
Total	**$1.958M**	**$4.231M**	**$5.865M**

Source: Enterprise Strategy Group, now part of Omdia.

Table A4. Modeled Costs for 70B Parameter Model With 50K Users (High Usage) Over 4 Years

Cost Category	Dell AI Factory (On Premises)	Public Cloud IaaS (Amazon EC2)	API-based GenAI Service (OpenAI GPT-4o)
Hardware / Software / Services (Inc. Cloud Costs, NVIDIA licenses, and Dell Services)	$5.622M	$18.242M	$29.250M
Power and cooling cost	$1.326M	$0	$0
Infrastructure/instance and ML platform administration	$0 (included in Dell Managed Services cost)	$373K	$0
ML model administration	$0 (included in Dell Managed Services cost)	$258K	$52K
Total	**$6.948M**	**$18.873M**	**$29.302M**

Source: Enterprise Strategy Group, now part of Omdia.

©2025 TechTarget, Inc. All rights reserved. The Informa TechTarget name and logo are subject to license. All other logos are trademarks of their respective owners. Informa TechTarget reserves the right to make changes in specifications and other information contained in this document without prior notice.

Information contained in this publication has been obtained by sources Informa TechTarget considers to be reliable but is not warranted by Informa TechTarget. This publication may contain opinions of Informa TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent Informa TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, Informa TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of Informa TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at cr@esg-global.com.

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