ECONOMIC WHITE PAPER

Understanding the Total Cost of Inferencing Large Language Models

How Leveraging Dell Technologies On-premises Solutions Can Be 38% to 88% More Cost-effective for Inferencing LLMs With RAG Compared to the Public Cloud or Token-based APIs

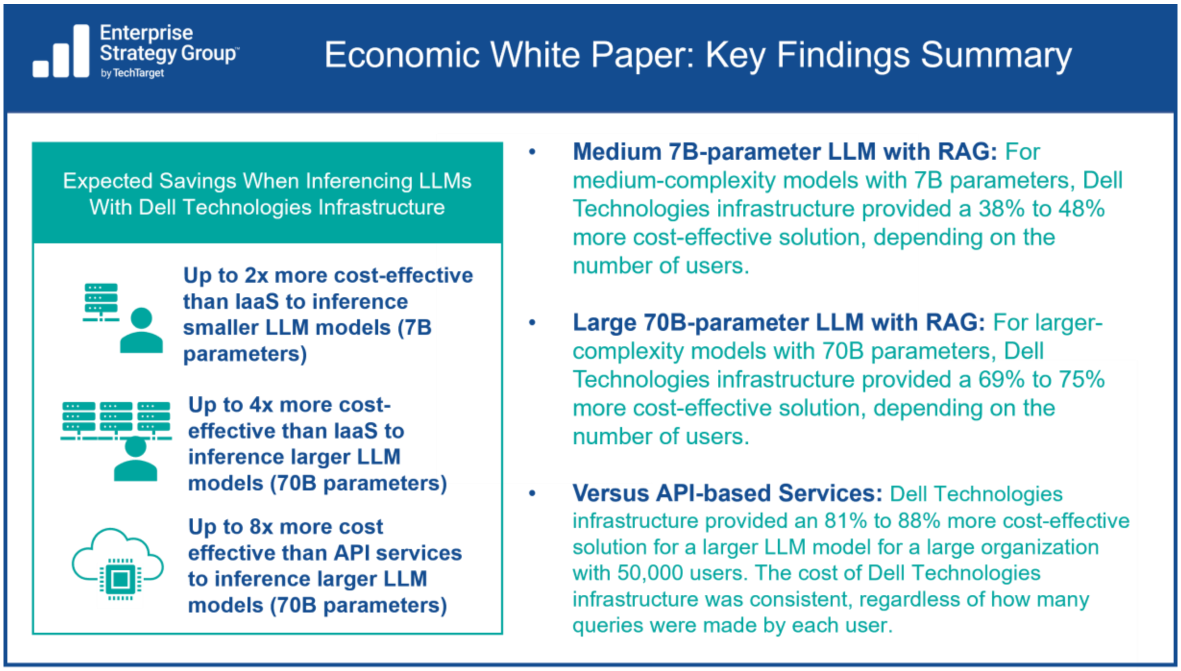
By Aviv Kaufmann, Practice Director and Principal Validation Analyst
Enterprise Strategy Group

April 2024

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

Contents

Introduction **3**
Challenges 3
Key Considerations of LLM Inferencing 4
Enterprise Strategy Group Economic Analysis **5**
Dell Technologies On-premises Infrastructure Versus Public Cloud IaaS 5
Smaller-sized Model: 7B Parameter Mistral 7B LLM 6
Larger-sized Model: 70B Parameter Llama 2 LLM 7
Dell Technologies On-premises Infrastructure Versus API-based GenAI Service 8
Issues to Consider **8**
Dell Technologies for LLM Inferencing **9**
Conclusion **9**



Enterprise Strategy Group by TechTarget

Economic White Paper: Key Findings Summary

Expected Savings When Inferencing LLMs With Dell Technologies Infrastructure

Up to 2x more cost-effective than IaaS to inference smaller LLM models (7B parameters)

Up to 4x more cost-effective than IaaS to inference larger LLM models (70B parameters)

Up to 8x more cost effective than API services to inference larger LLM models (70B parameters)

- **Medium 7B-parameter LLM with RAG:** For medium-complexity models with 7B parameters, Dell Technologies infrastructure provided a 38% to 48% more cost-effective solution, depending on the number of users.
- **Large 70B-parameter LLM with RAG:** For larger-complexity models with 70B parameters, Dell Technologies infrastructure provided a 69% to 75% more cost-effective solution, depending on the number of users.
- **Versus API-based Services:** Dell Technologies infrastructure provided an 81% to 88% more cost-effective solution for a larger LLM model for a large organization with 50,000 users. The cost of Dell Technologies infrastructure was consistent, regardless of how many queries were made by each user.

Introduction

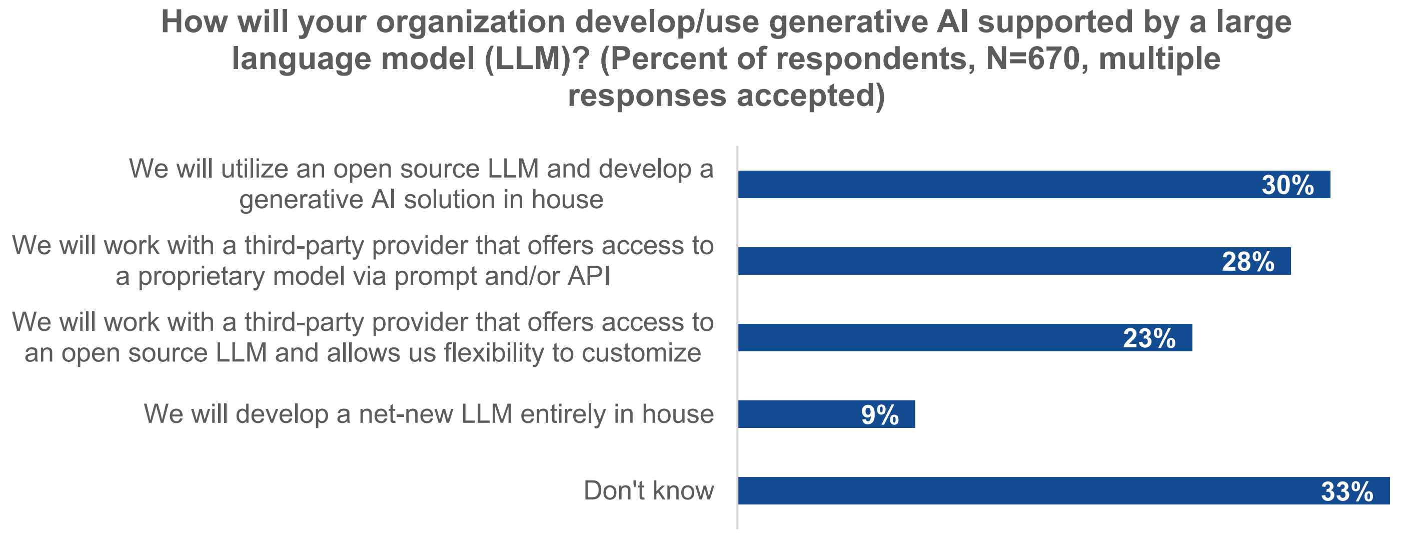
This Economic White Paper presents some of the options and considerations for delivering text-based generative AI (GenAI) capabilities to organizations. TechTarget's Enterprise Strategy Group modeled and compared the expected costs to inference large language models (LLMs) utilizing retrieval-augmented generation (RAG) on on-premises Dell Technologies infrastructure versus using native public cloud infrastructure as a service (IaaS) or the OpenAI GPT-4 Turbo LLM model service through an API. We found that Dell Technologies could provide LLM inferencing up to 4x more cost-effectively than IaaS and up to 8x more cost-effectively than with GPT-4 Turbo API.

Challenges

Organizations are embracing the power of GenAI and LLMs that leverage company-specific data and other intellectual property to automate content generation, answer questions, and make insights readily available to decision-makers. Along with many other benefits, respondents to an Enterprise Strategy Group research study reported that the primary benefits of using GenAI in their organization include improving and/or automating processes and workflows, supporting data analytics and business intelligence, increasing employee productivity, and improving operational efficiency.[1]

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. Ready-made API-based services such as OpenAI GPT offer a simpler solution, but inferencing (i.e., querying) costs can quickly add up, especially for larger organizations and more complex LLMs. Alternatively, organizations can build and control their own LLM inferencing solution on powerful GPU-enabled enterprise servers or equivalent GPU-enabled cloud instances and a machine learning platform, like NVIDIA's AI Enterprise, running open source LLMs. Not surprisingly, Enterprise Strategy Group found that the most popular strategy for organizations to develop and use GenAI supported by an LLM was to utilize an open source LLM and develop a GenAI solution in-house.[2]

Figure 1. Most Organizations Plan to Develop Their Own Generative AI Solution In-house



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

[1] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.
[2] Ibid.

Key Considerations of LLM Inferencing

Text-based LLMs focus on learning, understanding, and producing text-based content, answers, summaries, and questions that can be tailored to a particular industry, use case, and organization. RAG augments the results of GenAI 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. When delivering 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 infrastructure required to facilitate inferencing workloads. There are several deployment methods that can be used for inferencing LLMs, including:

- **Traditional infrastructure.** Purchased or leased traditional infrastructure consisting of compute, memory, GPUs, and storage can be deployed and managed along with a commercial or open source AI platform, giving the organization control over all aspects of the deployment. This method may be the most cost-effective for larger and predictable workloads.
- **Public cloud IaaS.** Similarly, organizations could 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 platform, with agility and easy integration with existing tools. This method may be the most cost-effective for small deployments and those with unpredictable or seasonal requirements.
- **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 may be the best for exploring and getting started, smaller deployments, and those that do not require a large degree of customization or control.

Before deciding on an LLM platform, organizations should invest time to understand their requirements and capabilities, as well as discuss some of the following considerations around choosing a platform for LLM inferencing, such as:

- **Cost/ROI.** Organizations should consider the cost and benefits of implementing and using every technology investment. According to an Enterprise Strategy Group research study, cost savings and ROI were the most common metrics that organizations said they use to measure the effectiveness of their AI initiatives.[3]
- **Performance and scalability.** Sizing the infrastructure with enough resources in processors, GPUs, memory, and storage 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 of the LLM will happen on the same platform or on a higher-performance dedicated training platform before being moved to the inferencing platform.
- **Management simplicity.** 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 operating the infrastructure and platforms (e.g., administration, support and maintenance, and power/cooling). Colocation options can also 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.
- **Expected user workloads.** 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 may suffice, but as an organization supports more users and inferences, building a proprietary platform will become more cost-effective. It is important that organizations consider expected growth in adoption and usage frequency over time to ensure that infrastructure is sized appropriately and can grow with the needs of the business.

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

- **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 or is easily retrievable where it is needed, while the public cloud can make collection and centralization of data easier in some cases. On-premises instances also enable organizations to better control security and ensure compliance of sensitive data. 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.

Enterprise Strategy Group Economic Analysis

Enterprise Strategy Group created an economic analysis that compared the expected costs of delivering inferencing for several open source LLMs utilizing RAG of various complexities (with the number of parameters including 7B and 70B) and for different sized organizations (with the number of users between 5K and 50K). We assumed that the model was providing an internal text-based Q&A and that inferencing occurred where the data was located, so there was no high cost of data migration. The analysis looked at all the costs associated with running and inferencing the models over a three-year period, including providing and running the infrastructure, administering the systems, and paying for cloud services if required.

Dell Technologies 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 configurations 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 described in Table 1 for both the Dell Technologies infrastructure and the equivalent EC2 configuration.

Table 1. Costs and Assumptions Modeled for Each LLM Inferencing Workload Requirement

Cost Category	Dell Technologies (on-premises)	Public Cloud IaaS (Amazon EC2)
Initial cost of acquisition (hardware and software)	Price provided by Dell Technologies for Dell PowerEdge R760xa and R660 with ProDeploy and ProSupport	N/A
Additional cost of capital (interest) and depreciation (benefit)	Factored into model (8% WACC, 6% annual depreciation benefit)	N/A
Power and cooling cost	Calculated based on system specifications ($0.173/kWh)	N/A
Monthly cloud spending	N/A	p5.48xlarge EC2 instance costs calculated based on 3-year reservation discounts
NVIDIA AI Enterprise license/GPU	Based on 5-year license (prorated)	Per instance/h, based on 16 h/day, 5 days per week to limit costs
Infrastructure/instance administration	Modeled (10%-100% of system admin based on number of nodes)	66% lower than on-premises model
ML model and platform administration	Modeled (20%-100% of ML engineer based on qty of model instances)	Same as on-premises model

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Smaller-sized Model: 7B Parameter Mistral 7B LLM

For the first comparison, we modeled the costs to deliver a smaller model containing about 7 billion parameters, similar to the open source Mistral 7B LLM. To size the requirements, we used a sizing tool based on the results of testing that predicted the server and GPU configurations that would be capable of providing an average latency per request of around 0.4 seconds and an estimated throughput of 2.29 to 6.86 inferences per second. The high-level assumptions for instance and GPU counts are shown in Table 2.

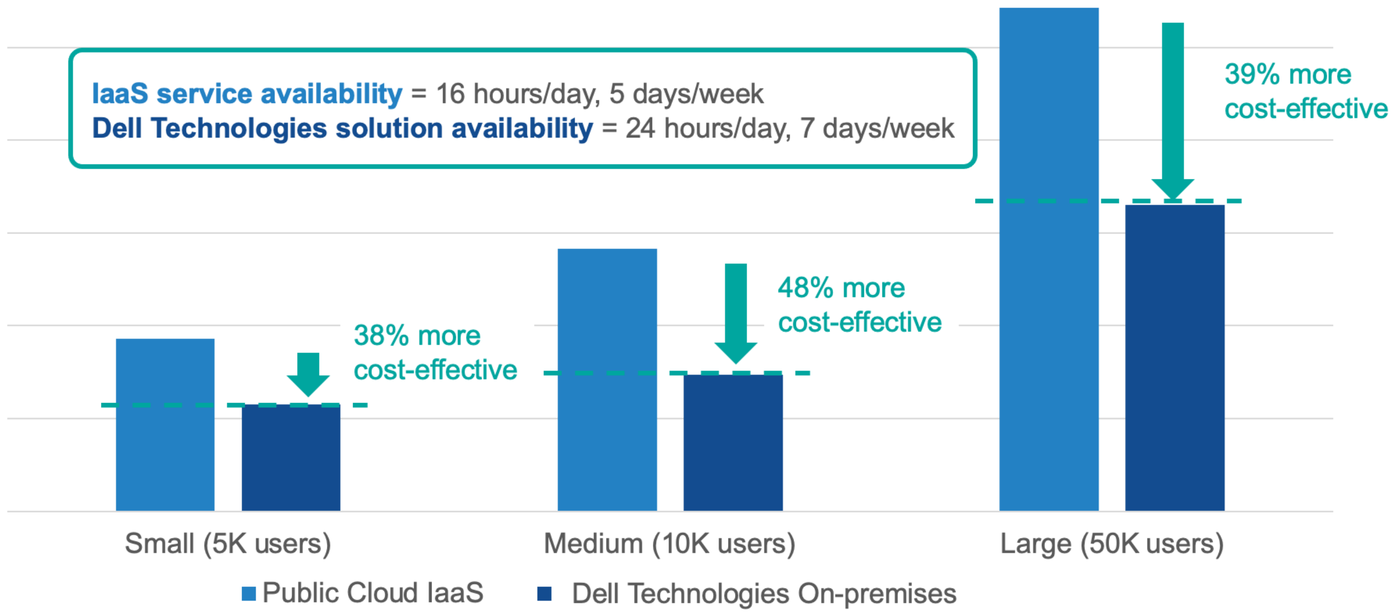
Table 2. Configuration Assumptions for the Mistral 7B Parameter Model Inferencing

LLM Model (Number of parameters)	Number of Users	Number of Inferencing Nodes/instances	Number of H100 GPUs
Mistral (7B)	5,000	1	1
	10,000	1	2
	50,000	1	4

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

We then modeled all the costs summarized in Table 1 for each configuration. As shown in Figure 3, Dell Technologies infrastructure was 1.6x to 1.9x (38% to 48%) more cost-effective at delivering inferencing for the organization. while also being made available to the organization 24/7.

Figure 2. Expected Cost to Deliver Inferencing for 7B Parameter Mistral LLM Using RAG



IaaS service availability = 16 hours/day, 5 days/week
Dell Technologies solution availability = 24 hours/day, 7 days/week

38% more cost-effective

48% more cost-effective

39% more cost-effective

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Larger-sized Model: 70B Parameter Llama 2 LLM

We then modeled the expected costs to deliver a larger model with 70 billion parameters, similar to the open source Llama 2 70B LLM. We again sized the requirements with the same sizing tool to predict server and GPU configurations that would be capable of providing a slightly higher average latency per request of around 1.8 seconds and an estimated throughput of 2.29 to 22.86 inferences per second. The high-level assumptions for instance and GPU counts are shown in Table 3.

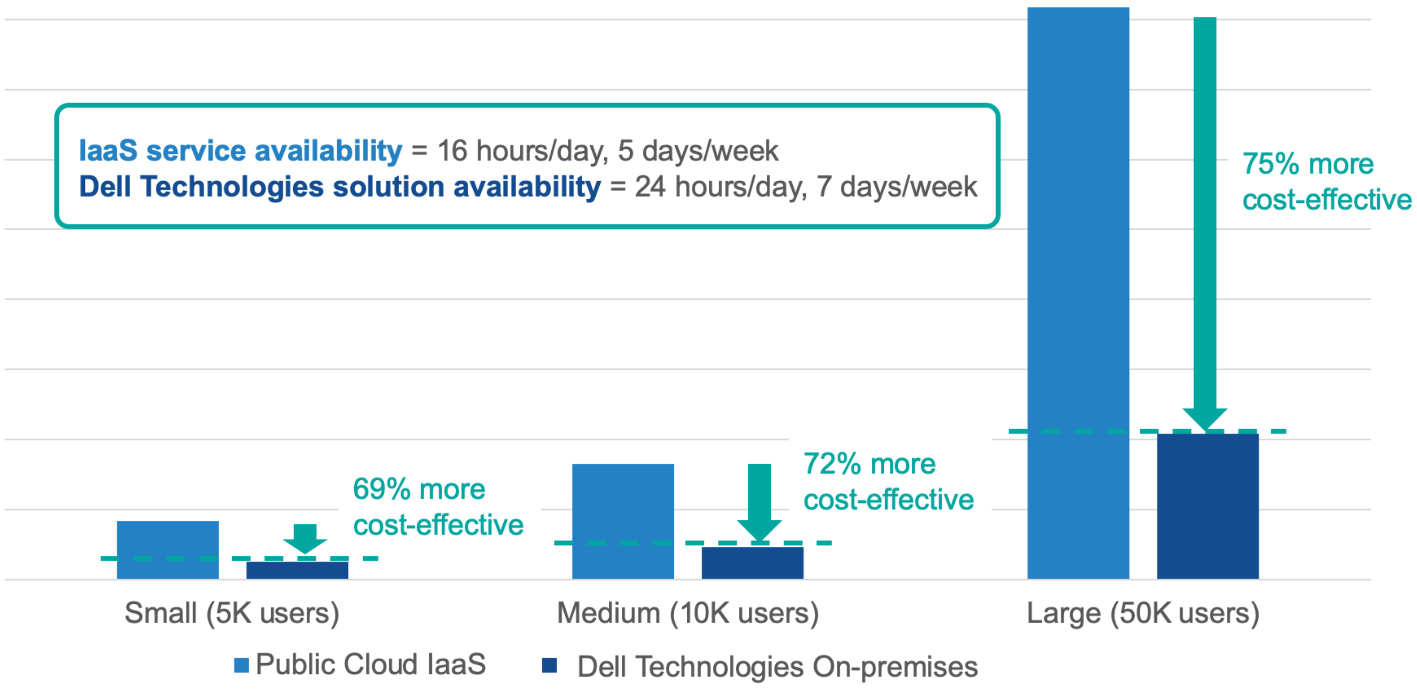
Table 3. Configuration Assumptions for the Llama 2 70B Parameter Model Inferencing

LLM Model (Number of parameters)	Number of Users	Number of Inferencing Nodes/instances	Number of H100 GPUs
Llama 2 (70B)	5,000	2	8
	10,000	4	16
	50,000	20	80

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

After again modeling all the costs summarized in Table 1 for each configuration shown above, we found that Dell Technologies infrastructure was 3.3x to 4x (69% to 75%) more cost-effective at delivering inferencing for the organization, while also being made available to the organization 24/7.

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

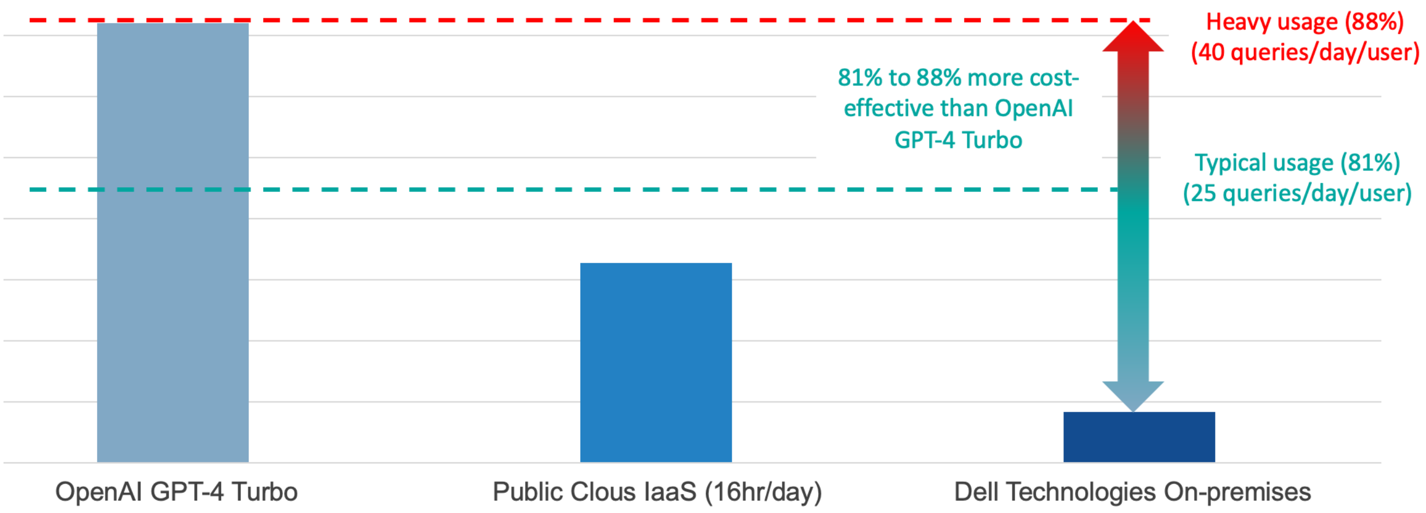


Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Dell Technologies On-premises Infrastructure Versus API-based GenAI Service

We then compared the expected costs for a large organization to provide an equivalent 70B parameter model to its 50,000 users using the established OpenAI API-based GenAI service GPT-4 Turbo, which is priced cost-effectively per input and output "token." Text-based Q&A 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. We assumed 1,500 total (input plus output) tokens per query, with an average of about 25 queries per day, per user, for both the on-premises and API-based solutions. 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 per user and more established organizations averaging as many as 40 queries per user, per day. Our GPT-4 Turbo calculations predicted a cost of about $12.50/user/month, which compares favorably to the suite-based AI assistance tools that can cost roughly $30/user/month. With these assumptions, we found that Dell Technologies on-premises infrastructure could provide inferencing 5.4x to 8.6x (81% to 88%) more cost-effectively than using an API-based service, delivering GenAI capabilities for only about $2.31/user/month.

Figure 4. Expected Cost to Deliver Inferencing for 70B Parameter Llama 2 LLM to 50K Users



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

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 their particular use case, the nature of their data, their level of expertise, and their model and infrastructure requirements. Enterprise Strategy Group recommends that you perform your own analysis of available products and consult with Dell Technologies to understand and discuss the differences between the solutions proven through your own proof-of-concept testing.

Dell Technologies for LLM Inferencing

Dell Technologies helps organizations easily bring AI to their data, no matter where it resides. This means offering the broadest portfolio of AI services—from desktop, to data center, to cloud—so organizations can right-size their investments and leverage data to build their AI factories and bring AI use cases to life efficiently, securely, and sustainably. Dell does this by providing access to a comprehensive services portfolio and a broad, open ecosystem of partners to assist organizations no matter where they are in their AI journeys—whether they're developing AI strategies or accelerating and scaling their GenAI investments.

For organizations challenged by data security threats, compliance concerns, data silos, and unvalidated data sets, Dell Professional Services for Generative AI can help create consensus among business and IT leaders around prioritized use cases, provide an actionable roadmap to achieve objectives, prepare enterprise data for LLM integration, advance cybersecurity maturity, and establish an AI platform aligned to specific business needs. In addition, with Dell APEX, organizations can subscribe to AI solutions and optimize them for multicloud use cases.

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

Conclusion

The expanded use of GenAI across nearly every area of the business is a crucial factor to ensuring improved operations and future success. Enterprise Strategy Group research finds that the top areas in which organizations are currently applying GenAI today include research, marketing, software development, product development, and IT operations, and the potential for usage across every area is expected to increase.[4] Organizations can achieve more impactful and meaningful results by training and inferencing against their own customized version of an LLM.

There are several deployment methods that 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, Dell Technologies infrastructure can provide high-performance LLM inferencing up to 4x more cost-effectively than IaaS and up to 8x more cost-effectively than with OpenAI GPT-4 Turbo. Enterprise Strategy Group strongly recommends that companies implementing LLMs to power their organizations consider taking advantage of the cost-effective technologies and knowledgeable services that Dell Technologies provides to ensure a successful outcome, accelerate their GenAI initiatives, and reduce the time to achieve these expected savings.

[4] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

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

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group 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