aws marketplace

NVIDIA AI Enterprise

Reviews, tips, and advice from real users






Powered by PeerSpot

Contents

Product Recap

NVIDIA AI Enterprise

NVIDIA AI Enterprise Recap

NVIDIA AI Enterprise provides a comprehensive suite of AI tools designed for deployment across diverse industries, enabling businesses to harness the power of AI for scalable, efficient operations.

NVIDIA AI Enterprise offers a robust set of AI technologies tailored for advanced data analytics, machine learning, and neural networks. It streamlines AI deployment, optimizing workload management and facilitating rapid model training and deployment. With support for a range of frameworks and environments, it integrates seamlessly into existing IT infrastructures, reducing complexity and enhancing scalability.

What are the most important features of NVIDIA AI Enterprise?

- Pre-trained Models: Enables accelerated AI implementation with a library of models designed for common use cases.
- Containerized Environment: Simplifies deployment and management through isolated environments.
- Security Enhancements: Incorporates advanced security protocols ensuring data protection.
- Scalability: Allows for seamless scaling of resources to meet changing demand.

What benefits or ROI should users look for in reviews?

- Improved Efficiency: Users highlight reduced training times and increased workflow productivity.
- Cost Savings: Gain insights into significant reductions in overhead through streamlined processes.
- Enhanced Innovation: Leverage insights into faster deployment cycles leading to quicker time-to-market.
- Integration: Reviews often note easy alignment with existing IT structures.

Industries like healthcare, retail, and finance leverage NVIDIA AI Enterprise to implement predictive analytics, optimize supply chains, and enhance customer interactions. In healthcare, it's used for advanced diagnostics, while retailers utilize its capabilities for personalized marketing. Finance sectors harness its power for risk assessment and fraud detection.

Valuable Features

Excerpts from real customer reviews on PeerSpot:

"Overall, it increased productivity for both development teams and end-users by making AI solutions faster, scalable, and easier to maintain."

Verified user

Operations Analyst at a non-profit with 51-200 employees

"Now I am able to easily handle more than 60 high-resolution camera streams simultaneously on a single H100 GPU with excellent throughput and very low latency, and the development time for new vision pipelines has dramatically dropped from three to four weeks to only four to six days because of DeepStream."

Nandhavignesh Ramalingam

NVIDIA AI Architect at Lenovo

"For that kind of use case, NVIDIA AI Enterprise is ideal when it compares to other AMD or Dell, because AMD may not provide a complete solution the way NVIDIA AI Enterprise is providing for the enterprise."

Bharath _Kumar

Solution Architect at Velocis Systems

What users had to say about valuable features:

"Regarding the integration with AI framework on your project development, the impact of NVIDIA AI Enterprise is easily consumable. The license has an enterprise license and all of those components. It is easy to adopt. How it impacts is very helpful in terms of choosing the options.

I do see that it helps to minimize downtime for AI applications because it has a lot of valuable features. I do see a benefit from it. Mainly at the time of doing any kind of opportunity where precision computing and all those things will come, the Tensor Cores bring a certain kind of value. It is mainly helping me to speed up the training of the AI models. That is where in most of the AI factories, the Tensor Cores make a difference when you have mixed-precision computing. Mostly the HPC is part of the HPC. They recently launched the Blackwell fifth-generation Tensor Cores.

In terms of the price of the license, I would say NVIDIA AI Enterprise is expensive.."

Bharath _Kumar
Solution Architect at Velocis Systems

Read full review

"The best features of NVIDIA AI Enterprise are GPU-accelerated AI and GenAI workloads, NVIDIA NIM microservices for fast LLM deployment, and enterprise-grade security and support. Another strong feature is support for a hybrid environment so workloads can run across clouds, data center, and edge systems. It also includes orchestration and infrastructure tools for better GPU resource management, which is very useful for large-scale AI workloads.

In my day-to-day work, I rely most on the NVIDIA NIM microservices and the GPU-optimized inference because they make LLM deployment faster, reduce latency, and simplify scalable production deployment. I also value the pre-validated enterprise stacks because they save time on compatibility issues between drivers, frameworks, and libraries. Instead of spending efforts on environment setup, I can focus more on building and improving the AI solution using NVIDIA AI Enterprise.

Another important advantage is seamless integration with enterprise infrastructure in Kubernetes, VMware, and cloud platforms, which makes production deployment and scaling much easier.."

Verified user

Operations Analyst at a non-profit with 51-200 employees

Read full review

"The best features NVIDIA AI Enterprise offers are high-performance multi-stream processing, end-to-end GPU accelerations for full pipelines, seamless Kubernetes integration for easy deployment of NVIDIA GPU operators, stability, support, and advanced tracking with multi-view tracking capabilities.

"The Kubernetes integration helps my team by simplifying deployment, as I previously had to manually manage Docker containers, GPU allocations, and scaling for new vision pipelines, but now I define my pipelines in YAML manifest and let Kubernetes handle scheduling, GPU resource allocations, and autoscaling, enabling me to automatically scale up DeepStream pods during high workloads and down during low traffic, optimizing GPU cost.

"NVIDIA AI Enterprise has positively impacted my organization by significantly reducing processing time as I'm now handling more than 60 high-resolution cameras instead of two to three weeks before, achieving operational efficiencies, reducing processing costs by approximately 45%, and enabling me to handle 5x more camera streams.

"On the manufacturing side, the product quality has improved with real-time defect detection that reduced faulty products reaching customers by 38%, leading to increased customer satisfaction scores along with fewer returns and warranty claims.."

Nandhavignesh Ramalingam
NVIDIA AI Architect at Lenovo

Read full review

Other Solutions Considered

"There is no such competition for NVIDIA AI Enterprise, as they are addressing the complete AI-related space. Even if AMD has GPUs, Dell has that, and all of this, NVIDIA AI Enterprise is leading because they are addressing each and every component in the AI infrastructure.."

Bharath _Kumar
Solution Architect at Velocis Systems

Read full review

"Before adopting NVIDIA AI Enterprise, we were primarily using a combination of open-source tools and custom-built ML infrastructure. This included a standard Python-based ML stack, Docker-based deployment, and manual management of GPU environments on cloud providers like AWS.."

Verified user
Operations Analyst at a non-profit with 51-200 employees

Read full review

"Before choosing NVIDIA AI Enterprise, we evaluated a few other options to compare performance, cost, and ease of deployment, including AWS SageMaker, Google Vertex AI, and standard open-source MLOps stacks. NVIDIA AI Enterprise was preferred for better GPU performance optimization, lower inference latency, and tighter integration with on-premises hybrid GPU infrastructure.."

Verified user

Operations Analyst at a non-profit with 51-200 employees

Read full review

ROI

Real user quotes about their ROI:

"NVIDIA AI Enterprise has positively impacted our organization by improving the speed and efficiency of deploying AI solutions. It helped reduce the setup time of GPU environments, streamline model deployment, and improve performance for inference workloads. It also enabled us to build more reliable production-grade AI applications such as an internal knowledge assistant and a document automation system. Overall, it increased productivity for both development teams and end-users by making AI solutions faster, scalable, and easier to maintain.

We saw around a 30 to 40% inference performance improvement, reduced deployment time using pre-built NVIDIA AI Enterprise tools, and better GPU resource utilization for large-scale GenAI workloads.

We saw around a 25 to 30% reduction in infrastructure cost due to better GPU utilization and approximately 40% reduction in model deployment time, which improved overall delivery speed and reduced the engineering efforts needed for production release.."

Verified user

Operations Analyst at a non-profit with 51-200 employees

Read full review

Use Case

“My main use case is building and deploying GenAI applications like RAG pipelines, LLM inference service, and GPU-accelerated AI workloads with a scalable enterprise deployment.

I use NVIDIA AI Enterprise to deploy a RAG-based chatbot using NVIDIA NIM microservices and GPU acceleration for faster LLM inference, document retrieval, and scalable enterprise deployment on Kubernetes..”

Verified user

Operations Analyst at a non-profit with 51-200 employees

Read full review

"Regarding use cases, mainly if you want to do anything on AI workloads, you have an option to choose because NVIDIA has the full stack. They have the software, they have their GPUs, and all of those components. Based on the solution, suppose some customers might be asking for some kind of computer vision models they want to adopt in order to have a quality of inspections and all of those in their factory or in their healthcare. For one of the customers where we worked, we wanted to implement a computer vision model where they want to identify some kind of artifacts in the health reports. It means in terms of identifying the quality and inspecting the particular lab X-rays and whatever is health-related. At that time, we need to work from the infrastructure level to the model and also have a software; the full stack has to be there. For that kind of use case, NVIDIA AI Enterprise is ideal when it compares to other AMD or Dell, because AMD may not provide a complete solution the way NVIDIA AI Enterprise is providing for the enterprise. In those cases, it is very ideal.."

Bharath _Kumar
Solution Architect at Velocis Systems

Read full review

"I have been using NVIDIA AI Enterprise for the past 10 months in a production environment, primarily for learning large language model inference, RAG pipelines, and some computer vision workloads on H100s and GPUs.

"There are many use cases for NVIDIA AI Enterprise, mostly on different verticals, but most of them are on vision workloads.

"A quick specific example of a vision workload I'm running with NVIDIA AI Enterprise is using DeepStream SDK, which delivers high-performance, multi-stream video processing with low latency, and TAO Toolkit makes transfer learning and model optimization straightforward for me, while TensorRT optimizations provide a huge inferencing speedup.

"DeepStream and the TAO Toolkit are game-changers for me, as I was struggling with traditional OpenCV plus PyTorch setups and could only process 8 to 12 camera streams reliably for one of our customers on our hardware, with frequent frame drops and high latencies. Now I am able to easily handle more than 60 high-resolution camera streams simultaneously on a single H100 GPU with excellent throughput and very low latency, and the development time for new vision pipelines has dramatically dropped from three to four weeks to only four to six days because of DeepStream.

"NVIDIA AI Enterprise does a lot for my workflow because model development and operational reliability have all started on that platform, fitting perfectly into my framework since it is not the single solution I am working on with customers, and I am processing camera pipelines, reducing them, and changing focus from business outcomes with orchestration layer, model integration layer, data flow layer, monitoring layer, and security compliances across various frameworks.

"Additionally, I have started exploring the BioNeMoTron framework with NVIDIA AI Enterprise, and I'm looking forward to advancements in the Triton Inferencing servers, as well as enhanced analytics and metadata integrations. Improvements in debugging tools and flexible pricing are important for mid-market customers, particularly in terms of enhanced documentation for edge deployments.."

Nandhavignesh Ramalingam
NVIDIA AI Architect at Lenovo

Read full review

Setup

The setup process involves configuring and preparing the product or service for use, which may include tasks such as installation, account creation, initial configuration, and troubleshooting any issues that may arise. Below you can find real user quotes about the setup process.

"As for the installation part, to be honest, I have not installed NVIDIA AI Enterprise right now. We had done only an eight-GPU deployment in our CoE. Eight built servers with eight GPUs were deployed for our lab setup. For the customers, I think there is another team who generally takes care of that.."

Bharath _Kumar

Solution Architect at Velocis Systems

Read full review

"Before choosing NVIDIA AI Enterprise, we evaluated a few other options to compare performance, cost, and ease of deployment, including AWS SageMaker, Google Vertex AI, and standard open-source MLOps stacks. NVIDIA AI Enterprise was preferred for better GPU performance optimization, lower inference latency, and tighter integration with on-premises hybrid GPU infrastructure.."

Verified user

Operations Analyst at a non-profit with 51-200 employees

Read full review

Customer Service and Support

"Customer support for NVIDIA AI Enterprise has been generally good, especially for enterprise-level issues. We have had 24/7 enterprise support for fast response times through NVIDIA Enterprise support portals and access to dedicated technical accounts and managers for critical issues. Most production issues are resolved quickly with clear guidance and regular updates.."

Verified user
Operations Analyst at a non-profit with 51-200 employees

Read full review

"I think I did not deal with the TAC and all of this, but the way the solution team provides design-level queries and answers questions about sizing is valuable. If you have any challenges in terms of sizing and you reach out to them, that kind of proactive support is always there. That means I can say that it is good. Based on my observations and experience with support, I can give it eight points from zero to ten, where ten is the best.."

Bharath _Kumar
Solution Architect at Velocis Systems

Read full review

Other Advice

“My advice to others considering NVIDIA AI Enterprise would be to first clearly define their workloads, requirements, and infrastructure setup before adoption. It works best for teams that are already using or planning to use GPU-accelerated AI workloads, especially in production environments. Understanding your use case, whether it is training, inference, or RAG pipelines, is important before investing. I would rate this product an 8 out of 10..”

Verified user
Operations Analyst at a non-profit with 51-200 employees

Read full review

"In terms of measuring the effectiveness of the project, I mostly work only in terms of the sizing of the infra piece for AI workloads. What exactly, what type of AI workloads the customer is having? And whether the primary workload is training-heavy or inferencing, what AI models they have? And in terms of performance, we just mainly ask in terms of what is the target for that token latencies. When you talk about AI, it is all about tokens. What are the expected average and peak tokens? That is the kind of sizing I understand.

Regarding whether my clients have NVIDIA AI Enterprise on cloud or on-premise, I can say it is a mix. It is mixed because it depends on the usage of your AI workload. If it is frequent, where people are trying to access, upload, and download, then definitely on-prem will be ideal, where they will go with NVIDIA AI Enterprise. And if it is not that much, then they will go with NVIDIA AI Enterprise from AWS or any cloud where you are able to spin the GPUs of NVIDIA in the cloud. I am not much into AWS on the cloud part.

My overall rating for NVIDIA AI Enterprise is eight out of ten.."

Bharath _Kumar
Solution Architect at Velocis Systems

Read full review

"I choose to rate NVIDIA AI Enterprise a 9 out of 10 because there are different frameworks I am working with customers on very customized pipelines, and I am unable to utilize 100 percent of NVIDIA AI Enterprise in those use cases, although it has the best features like superior performance optimization, DeepStream SDKs, and enterprise-grade stability. Better flexibility and affordable pricing options, particularly around interactions with the latest open-source models, could be improved.

"Regarding NVIDIA AI Enterprise's governance and security, I find it to be one of the strongest aspects I have utilized, including STIG hardening containers,

Distroless images, and compliance with regulatory environments, along with AI-specific governance features like NeMo Guardrails for prompt protections and output filtering.

“In terms of accuracy and reliability of output, I maintain 98 to 99 percent of the original model accuracy with my internal RAG models, achieving 3 to 5x higher output throughput with FP16 and int8 quantization options, resulting in overall system reliability of more than 95 to 98 percent.

“I would advise others considering NVIDIA AI Enterprise to definitely use it due to its superior performance on the inferencing side, seamless Kubernetes integration, strong governance, and high accuracy and reliability. My overall rating for NVIDIA AI Enterprise is 9 out of 10..”

Nandhavignesh Ramalingam
NVIDIA AI Architect at Lenovo

Read full review

About this buyer's guide

Thanks for downloading this PeerSpot report.

The summaries, overviews and recaps in this report are all based on real user feedback and reviews collected by PeerSpot's team. Every reviewer on PeerSpot has been authenticated with our triple authentication process. This is done to ensure that every review provided is an unbiased review from a real user.

Get a custom version of this report... Personalized for you!

Please note that this is a generic report based on reviews and opinions from the collective PeerSpot community. We offer a customized report of solutions recommended for you based on:

- Your industry
- Company size
- Which solutions you're already considering

The customized report will include recommendations for you based on what other people like you are using and researching.

Answer a few questions in our short wizard to get your customized report.
Get your personalized report here

About PeerSpot

PeerSpot is the leading review site for cloud, AI, and business software. We created PeerSpot to provide a trusted platform to share information about software, applications, and services. Since 2012, over 22 million people have used PeerSpot to choose the right software for their business.

PeerSpot helps tech professionals by providing:

- A list of products recommended by real users
- In-depth reviews, including pros and cons
- Specific information to help you choose the best vendor for your needs

Use PeerSpot to:

- Read and post reviews of products
- Access over 30,000 buyer's guides and comparison reports
- Request or share information about functionality, quality, and pricing

Join PeerSpot to connect with peers to help you:

- Get immediate answers to questions
- Validate vendor claims
- Exchange tips for getting the best deals with vendor

Visit PeerSpot: www.peerspot.com

PeerSpot

244 5th Avenue, Suite R-230 • New York, NY 10001

reports@peerspot.com

+1 646.328.1944