aws marketplace

GENERATIVE AI SOLUTION SPOTLIGHT

Scale and optimize your generative AI development

Deliver accurate responses for AI applications with NVIDIA on AWS

Generative artificial intelligence (AI) poses a significant opportunity for enterprises. Large language models (LLMs) can help increase efficiency across time-consuming enterprise tasks, such as copyediting, programming, and more. However, these models often struggle keeping up with real-time events and specific knowledge domains, which can lead to inaccuracies. Fine-tuning these models can enhance their knowledge, but it can be costly, labor-intensive, and require ample technical expertise.

NVIDIA AI Enterprise on Amazon Web Services (AWS) provides a secure, end-to-end software platform which includes NeMo, a cloud-native framework that allows organizations to quickly train, customize, and deploy LLMs at scale leveraging existing code and pretrained models.

NVIDIA NeMo Retriever Microservice, a service within NVIDIA AI Enterprise, helps enterprises enhance generative AI applications with retrieval-augmented generation (RAG) capabilities. With this service, enterprises can connect custom LLMs to enterprise data to deliver highly accurate responses for their AI applications.

How it works

NeMo Retriever, part of NVIDIA AI Enterprise, on AWS uses RAG to combine information retrieval with LLMs for open-domain question-answering applications. RAG provides LLMs with vast, updatable knowledge, effectively addressing time-consuming limitations from fine-tuning models.

In collaboration with:

NVIDIA

Partnering with AWS

The collaboration between AWS and NVIDIA extends across infrastructure, software, and services, offering a full-stack solution for customers to accelerate time to solution and reduce total cost of ownership (TCO) for deploying AI into production.

AWS provides scalability, reliability, global availability, and ease of use to make it seamless to get started with generative AI development using NVIDIA. Available in AWS Marketplace, NVIDIA AI Enterprise provides monthly releases to give you access to the latest features and performance improvements.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works (continued)

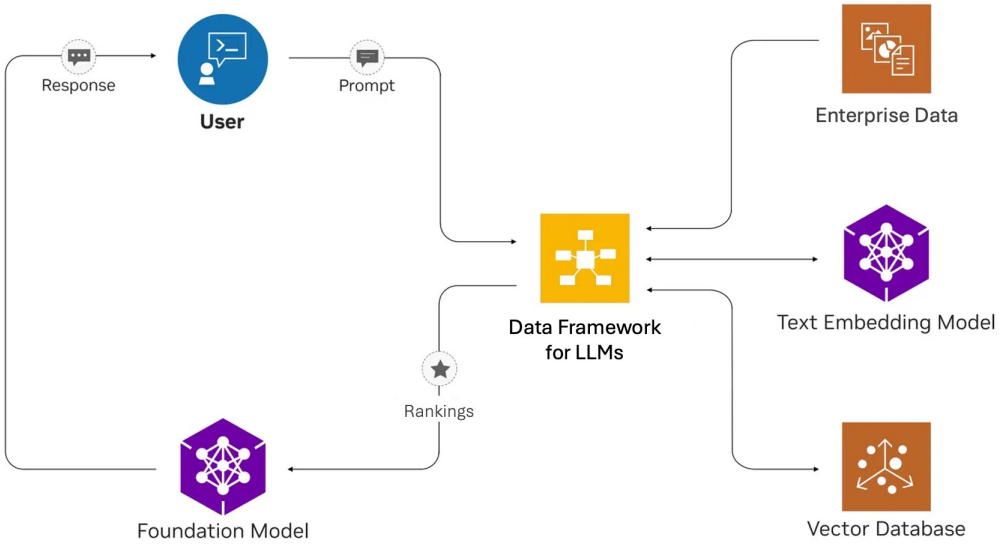
Enterprises can use this service to optimize the embedding and retrieval process of RAG to deliver higher accuracy and more efficient responses. Using NeMo Retriever, enterprises can connect their LLMs to multiple data sources and knowledge bases so that users can easily interact with data and receive accurate, up-to-date answers using simple, conversational prompts.

Businesses using Retriever-powered applications can allow users to securely gain access to information across data modalities, such as text, PDFs, images, and videos.

Leveraging NVIDIA for generative AI development

NeMo Retriever offers state-of-the-art, commercially ready models and microservices, optimized for the lowest latency and highest throughput. The service supports production-ready generative AI with API stability, security patches, and enterprise support.

With multiple pretrained models as starting points, developers can quickly customize models for their domain-specific use cases, such as IT or human resources (HR) help assistants, and research and development (R&D) assistants.



Featured AWS products

Amazon SageMaker

Amazon SageMaker is a fully managed service that lets you build, train, and deploy machine learning (ML) models for any use case with fully managed infrastructure, tools, and workflows.

Amazon Elastic Container Service (Amazon ECS)

Amazon ECS is a fully managed container orchestration service that helps you easily deploy, manage, and scale containerized applications.

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2 offers a broad, deep compute platform, with over 750 instances and choice of the latest processor, storage, networking, operating system, and purchase model to best match the needs of your workload.

Amazon Elastic Kubernetes Service (Amazon EKS)

Amazon EKS is a managed service that makes it easy for you to run Kubernetes on AWS without needing to install and operate your own Kubernetes clusters.

Amazon Simple Storage Service (Amazon S3)

Amazon S3 is an object storage service offering industry-leading scalability, data availability, security, and performance.

Learn more ›

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits

- **Time to solution**
  With NVIDIA NeMo Retriever, businesses can quickly build custom enterprise-grade models with their data and domain expertise. Enterprises can harness the powerful insights of generative AI, instead of maintaining and tuning their AI development platform.
- **Ease of use**
  Companies can streamline generative AI development with NVIDIA's suite of model-making services, pretrained models, frameworks, and APIs within NVIDIA AI Enterprise. They can simplify management with end-to-end software, including cluster management across cloud and data center environments, automated model deployment, and cloud-native orchestration.
- **Production-ready**
  Organizations can create enterprise-grade models that protect privacy, data security, and intellectual property. They can streamline AI projects with NVIDIA Enterprise Support, assurance of API stability, continuous monitoring, and regular security patches for common vulnerabilities and exposures (CVEs).
- **Unprecedented performance**
  Enterprises can leverage the powerful accelerators for generative AI, optimized for training and deploying LLMs. They can run their solutions on AWS to accelerate generative AI development.

Get started with training and deploying LLMs at scale.

Accelerate generative AI development with NVIDIA AI Enterprise on AWS.

About NVIDIA

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI, and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing infrastructure company with data-center-scale offerings that are reshaping industry.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.