NVIDIA Brings Business Intelligence to Chatbots, Copilots and Summarization Tools With Enterprise-Grade Generative AI Microservice

Cadence, Dropbox, SAP, ServiceNow First to Access NVIDIA NeMo Retriever to Optimize Semantic Retrieval for Accurate AI Inference

AWS re:Invent—NVIDIA today announced a generative AI microservice that lets enterprises connect custom large language models to enterprise data to deliver highly accurate responses for their AI applications.

NVIDIA NeMo™ Retriever — a new offering in the NVIDIA NeMo family of frameworks and tools for building, customizing and deploying generative AI models — helps organizations enhance their generative AI applications with enterprise-grade retrieval-augmented generation (RAG) capabilities.

As a semantic-retrieval microservice, NeMo Retriever helps generative AI applications provide more accurate responses through NVIDIA-optimized algorithms. Developers using the microservice can connect their AI applications to business data wherever it resides across clouds and data centers. It adds NVIDIA-optimized RAG capabilities to AI foundries and is part of the NVIDIA AI Enterprise software platform, available in AWS Marketplace.

Cadence, Dropbox, SAP and ServiceNow are among the pioneers working with NVIDIA to build production-ready RAG capabilities into their custom generative AI applications and services.

“Generative AI applications with RAG capabilities are the next killer app of the enterprise,” said Jensen Huang, founder and CEO of NVIDIA. “With NVIDIA NeMo Retriever, developers can create customized generative AI chatbots, copilots and summarization tools that can access their business data to transform productivity with accurate and valuable generative AI intelligence.”

Global Leaders Enhance LLM Accuracy With NeMo Retriever
Electronic systems design leader Cadence serves companies across hyperscale computing, 5G communications, automotive, mobile, aerospace, consumer and healthcare markets. It is working with NVIDIA to develop RAG features for generative AI applications in industrial electronics design.

“Generative AI introduces innovative approaches to address customer needs, such as tools to uncover potential flaws early in the design process,” said Anirudh Devgan, president and CEO of Cadence. “Our researchers are working with NVIDIA to use NeMo Retriever to further boost the accuracy and relevance of generative AI applications to reveal issues and help customers get high-quality products to market faster.”

Cracking the Code for Accurate Generative AI Applications
Unlike open-source RAG toolkits, NeMo Retriever supports production-ready generative AI with commercially viable models, API stability, security patches and enterprise support.

NVIDIA-optimized algorithms power the highest accuracy results in Retriever’s embedding models. The optimized embedding models capture relationships between words, enabling LLMs to process and analyze textual data.

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 spanning numerous data modalities, such as text, PDFs, images and videos.

Enterprises can use NeMo Retriever to achieve more accurate results with less training, speeding time to market and supporting energy efficiency in the development of generative AI applications.

Reliable, Simple, Secure Deployment With NVIDIA AI Enterprise
Companies can deploy NeMo Retriever-powered applications to run during inference on NVIDIA-accelerated computing on virtually any data center or cloud. NVIDIA AI Enterprise supports accelerated, high-performance inference with NVIDIA NeMo, NVIDIA Triton Inference Server™, NVIDIA TensorRT™, NVIDIA TensorRT-LLM and other NVIDIA AI software.

To maximize inference performance, developers can run their models on NVIDIA GH200 Grace Hopper Superchips with TensorRT-LLM software.

Availability
Developers can sign up for early access to NVIDIA NeMo Retriever.

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 company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, and availability of our products, services, and technologies, including NVIDIA NeMo Retriever, NVIDIA NeMo, NVIDIA AI Enterprise, and NVIDIA GH200; pioneers working with NVIDIA to build production-ready RAG capabilities into their custom generative AI applications and services; generative AI applications with RAG capabilities being the next killer app of the enterprise; businesses having hundreds of customized generative AI chatbots, copilots and summarization tools that can access their data to deliver accurate and valuable intelligence; and global leaders enhancing LLM accuracy with NeMo Retriever, including the benefits and impact thereof are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

© 2023 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA NeMo, NVIDIA Triton Inference Server, NVIDIA TensorRT and NVIDIA Grace Hopper are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Anna Kiachian
Senior PR Manager
NVIDIA Corporation
+1-650-224-9820
akiachian@nvidia.com