

Dell Generative AI Solutions with NVIDIA

Deliver more accurate and reliable outputs for Generative AI workloads



A scalable architecture for Retrieval-Augmented Generation (RAG) with NVIDIA Microservices

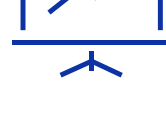
End users expect accurate and true answers from your GenAI implementations.

- Out-of-the-box use of Large Language Models (LLMs) offer limited value since they were trained on publicly available data
- A main challenge is when a model makes a wrong guess, called a hallucination

To maximize the value of pre-trained language models, they need to be combined and trained on your own data.



Quickly deploy scalable information retrieval frameworks for GenAI models



Enable more precise and trustworthy answers



Get better model performance via a retrieval-based approach

Leverage the potential of GenAI for key use cases



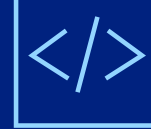
Content Creation

Across marketing, sales, back-office operations, and more



Digital Assistants

A tailored self-service experience in almost any language



Code Development

Assistance in generating initial code drafts for example

Why use a retrieval-based approach?

Improve accuracy and reliability of your GenAI models by using your own data for contextual retrieval. Build trust by enabling end users to check any claims against a referenceable source (like footnotes in a research paper).

Using only a pre-trained model



Show me all the product specifications for ACME Parachute



Generative AI Language Model



Sorry, I don't have access to any specific information about Acme Parachute

Using a pre-trained model + your own data (with RAG)



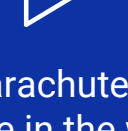
Show me all the product specifications for ACME Parachute



Generative AI Language Model



Your data; specific knowledge base



ACME Parachute is the safest parachute in the world:

- Crafted from cutting-edge, bulletproof materials for maximum safety
- Equipped with specially designed handles for precise control in the air
- Reflective interior baffles make you virtually invisible during descent

Start with an enterprise RAG framework for AI models today

Get up and running quickly with joint architecture from Dell Technologies and NVIDIA that reduces risk by helping you avoid design, planning, and adoption pitfalls.

Over 340k engineering hours spent on design, development and validation on GenAI solutions¹

GenAI Frameworks and Services



NVIDIA NeMo RAG Microservices



Community Source Models

AI Optimized Platforms and Libraries

NVIDIA Triton Inference Server

NVIDIA RAPIDS RAFT

Embedding Microservice + Vector Database

Infrastructure Management

Kubernetes

PowerScale Kubernetes CSI Driver

Enterprise Linux

Infrastructure

The world's broadest GenAI infrastructure portfolio that spans from cloud to client devices, all in one place²



PowerEdge + NVIDIA GPUs



PowerScale



PowerSwitch

Deliver outcomes faster with Dell Professional Services

Dell experts are here to assist you at every stage of your GenAI journey:

Strategize

Build your roadmap to achieve the innovation objectives of your IT and business stakeholders

Implement

Prepare your data and selected platform, leveraging documented and engineering-validated designs to implement the required hardware and software

Adopt

With your unique use cases in mind, Dell experts deploy and configure infrastructure to meet your needs

Scale

Manage your innovation portfolio with resident technical experts and training offers to develop the GenAI skills of your team

Dell Technologies and NVIDIA

Dell Technologies and NVIDIA work together to enable and accelerate Generative AI adoption, deliver engineering-validated hardware and software to accelerate AI, ML and DL workloads to meet customer needs across all businesses and verticals. With Dell Technologies and NVIDIA, you can deploy AI solutions to accelerate your digital transformation through real-time data that improves key decision-making, with solutions optimized for fastest time to value from your AI initiatives.



[Learn more](#) about Dell solutions



[Contact](#) a Dell Technologies Expert



[View more](#) resources



[Join the conversation](#) with [#PowerEdge](#) [@DellTech](#)

¹ Based on internal analysis, October 2023

² Based on Dell analysis, August 2023. Dell Technologies offers hardware solutions engineered to support AI workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services.