Unlock high-value insights faster with GenAI

Rapidly deploy a full-stack solution for Generative Artificial Intelligence (GenAI) large language model inferencing.

Generate higher quality, faster time-to-value predictions and outputs, while accelerating decision-making with a powerful GenAI solution from Dell Technologies and NVIDIA. This jointly engineered solution addresses Inferencing challenges such as latency, responsiveness, and computational demands helping transform enterprise data into high value, smarter outcomes.

With innovative technologies, comprehensive professional services, and a broad partner ecosystem, your organization can accelerate GenAI at an enterprise-wide scale. Now IT organizations, data scientists, and AI devOps can easily deliver a modular and scalable platform for GenAI and LLM inferencing.

Increase productivity and insights

This joint architecture delivers a modular and flexible design supporting a multitude of use cases and computational requirements. Components can be mixed and matched, and independently scaled depending on your application needs.

Some notable examples of inferencing use cases supported:

**Natural Language Generation:**
Gen Models can be used for text generation tasks such as document writing, dialogue generation, summarization, or content creation.

**Chatbots and Virtual Assistants:**
GenAI powers conversational agents, chatbots, and virtual assistants by generating natural language responses based on user queries or instructions.

**Code Development:**
Get assistance in software development with features like code completion, ability to generate unit tests, or a chat function for explain code.

Reduce time-to-results with a proven solution

Quickly build on-premises infrastructure for your application needs with a validated design and reference architecture made to simplify adoption. By reducing the complexity of every step of the way, you can now drive more insights and faster decisions, while boosting productivity.
Dell Technologies and NVIDIA work together to enable and accelerate Generative AI workloads, deliver engineering-validated hardware and software to accelerate AI, ML and DL workloads to meet customer needs across all businesses and verticals. With this Validated Design for LLM inferencing, you can accelerate your digital transformation with real-time data that improves key decision-making at-scale, with solutions optimized for fastest time to value from your AI initiatives.