Unlock high-value insights faster with GenAI

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

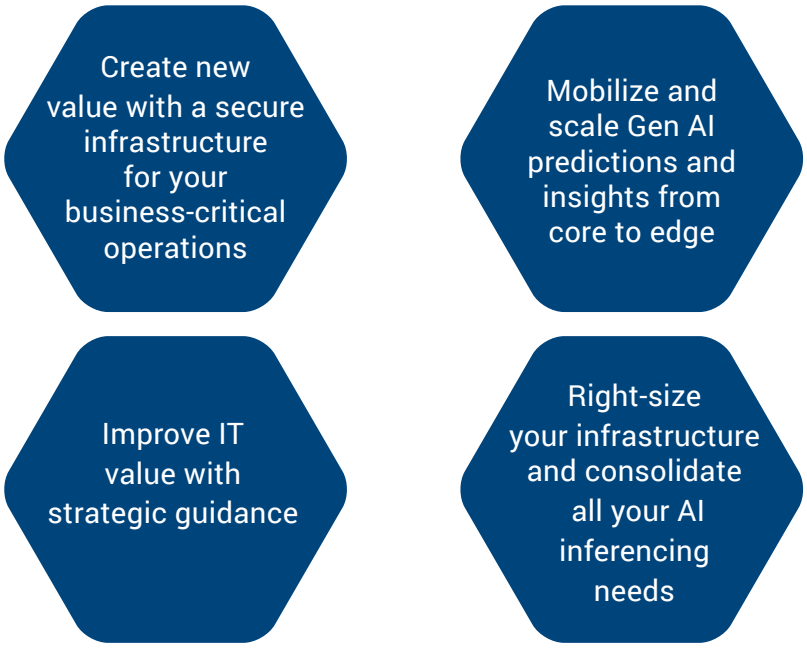
Natural Language Generation: Gen Models can be used for text generation tasks such as document writing, dialogue generation, summarization, or content creation

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.

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.



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.

Learn more

- See Design Guide
- AI InfoHub
- delltechnologies.com/ai
- Dell Technologies and NVIDIA

What is Inferencing?

Inferencing in AI refers to the process of using a trained model to generate predictions, make decisions, or produce outputs based on input data. It involves applying the learned knowledge and patterns acquired during the model's training phase to new, unseen data.

During inferencing, the trained model takes input data and processes it through its computational algorithms or neural network architecture to produce an output or prediction. The model applies its learned parameters, weights, or rules to transform the input data into meaningful information or actions.

Inferencing is a crucial stage in the lifecycle of an AI system. After training a model on labeled or unlabeled data to learn patterns and correlations, inferencing allows the model to generalize its knowledge and make predictions or generate responses on real-world or unseen data.

Deliver outcomes faster with our help

Dell Services experts help you realize the value of GenAI for your data more quickly with a portfolio of services 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** - establish your platform, leveraging Dell Validated Designs to implement GenAI inferencing hardware and software
- **Adopt** - accelerate the value of your GenAI use cases by implementing a pre-trained inferencing model
- **Scale** - Manage your GenAI innovation portfolio with resident technical experts and training offers to develop the skills of your team

Technical Specifications

The Validated Design configurations are based on the newest, AI-acceleration-optimized Dell PowerEdge XE and rack servers, leveraging the latest NVIDIA GPUs and NVIDIA AI Enterprise, with Triton Inference Server and the NeMo framework. Fast, ample data lake storage for Generative AI and large language models is provided by Dell PowerScale all-flash or hybrid storage arrays.

Compute	Accelerators	Networking	Software	Storage
Dell PowerEdge R760xa servers	NVIDIA A100 or H100 GPUs	NVIDIA Networking, Dell PowerSwitch S5232F-ON or S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise with Nemo Framework for LLMs and Triton Inference Server; NVIDIA Base Command Manager Essentials	Supported by Dell PowerScale, ECS, and ObjectScale

Dell Technologies and NVIDIA

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.

Learn more about Dell solutions

Contact a Dell Technologies Expert

View more resources

Join the conversation with #HashTag