To maximize the value of pre-trained language models, they need to be combined with your own data for more accurate and reliable outputs. When using only a pre-trained model, performance is limited to the specific use cases against which the model was trained. However, using a retrieval-based approach (RAG) with your own data, you can achieve the innovation you need.

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

- Provides a scalable and reliable framework for building GenAI models
- Integrates with existing frameworks for GenAI
- Enables better performance via the use of your own data

Leverage the potential of GenAI for key use cases

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

Why use a retrieval-based approach?

- Improves 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

- A main challenge is when a model makes a wrong guess, called a hallucination. End users expect accurate and true answers from your GenAI implementations.

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

- By using your own data, you can further improve the accuracy and reliability of your models. Dell Generative AI Solutions with NVIDIA leverage the power of retrieval-based generation (RAG) to provide trustworthy answers.

Dell Generative AI Frameworks and Services

- Dell and NVIDIA work together to deliver outcomes faster with Dell Professional Services
- Dell and NVIDIA work together to accelerate AI, ML and DL workloads to meet customer needs

Start with an enterprise RAG framework for AI models today

- Get up and running quickly with joint architecture from Dell Technologies and NVIDIA
- Using only a pre-trained model + your own data (with RAG)

Deliver outcomes faster with Dell Professional Services

- Dell and NVIDIA work together to deliver outcomes faster with Dell Professional Services
- Dell and NVIDIA work together to accelerate AI, ML and DL workloads to meet customer needs

¹ Based on internal analysis, October 2023