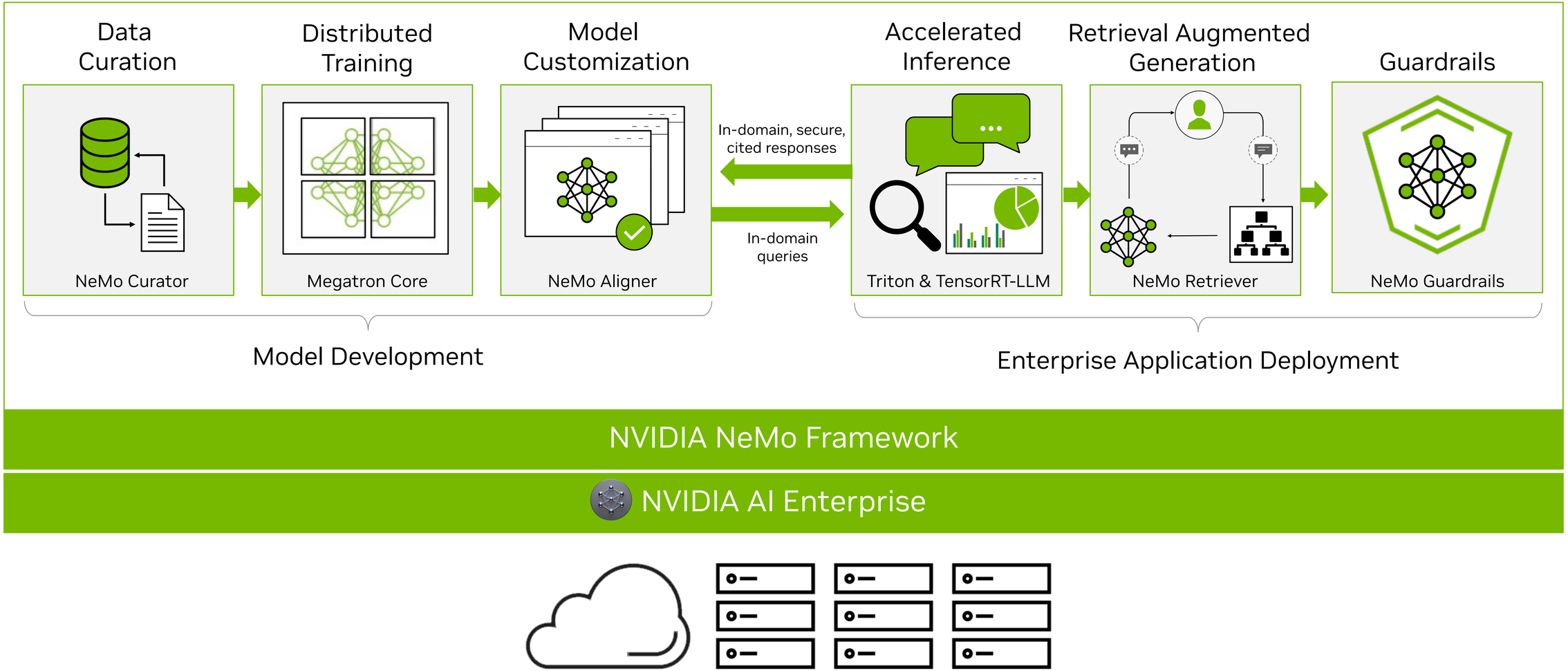
NVIDIA NeMo Framework

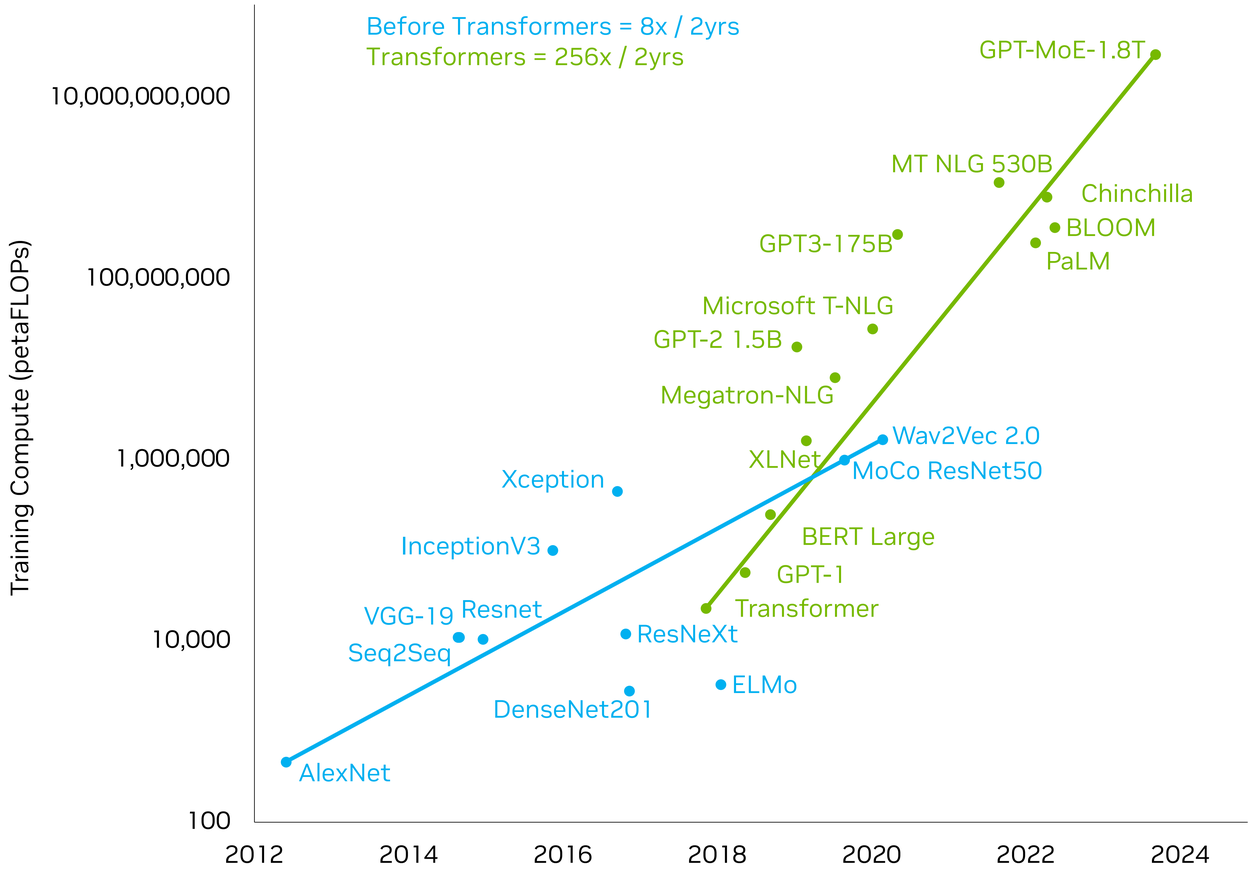
Accelerate Enterprise AI Development

Building Generative AI Applications for the Enterprise

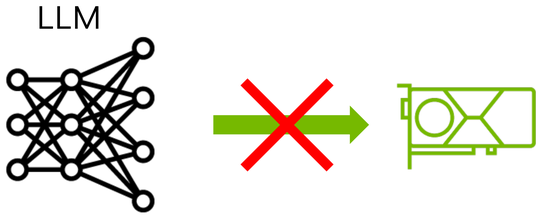
Build, customize and deploy generative AI models with NVIDIA NeMo



Explosive Growth in AI Computational Requirements



	Time (GPU hours)	Power Consumption (W)
Llama 3 8B	1.3M	700
Llama 3 70B	6.4M	700
Total	7.7M	

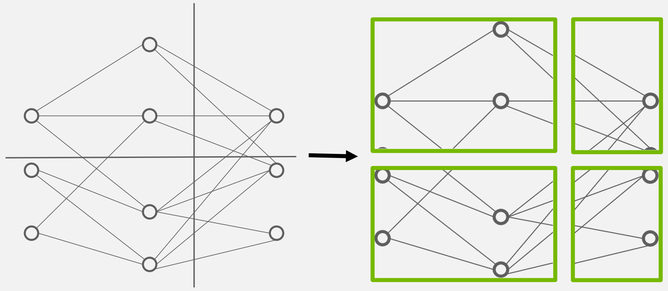


Can’t fit into single GPU

Building Generative AI Foundation Models

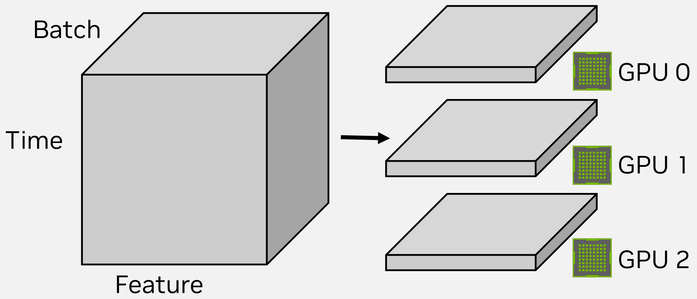
Efficiently and quickly training models using NVIDIA NeMo

Tensor & Pipeline Parallelism



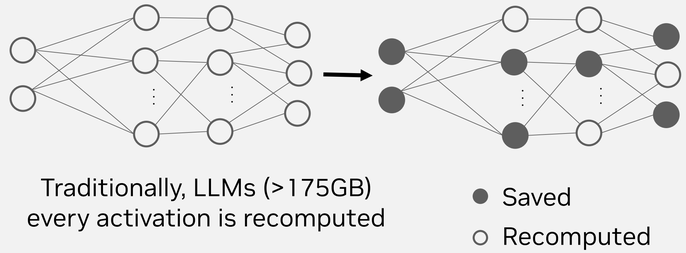
Reduced memory footprint and allows for large-scale training of LLMs across accelerated infrastructure

Sequence Parallelism



Working with tensor processing to increase the batch size that can be support for training

Selective Activation Recomputation



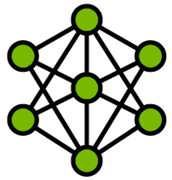
Smart activation checkpointing provides greatest trade-off between memory and recomputation

Model Customization for Enterprise Ready LLMs

Customization techniques to overcome the challenges of using foundation models

Stage1: Pre-training

Foundation Model

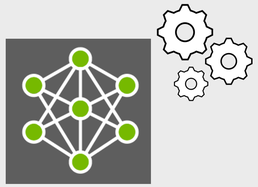


Start with
pre-trained model

- NVIDIA / Nemotron-3
- Meta / Llama-3
- Mistral / Mixtral
- Google / Gemma
- ServiceNow/StarCoder

Stage2: Fine-tuning

(P-tuning,
Prompt Tuning,
Adapters, LoRA)

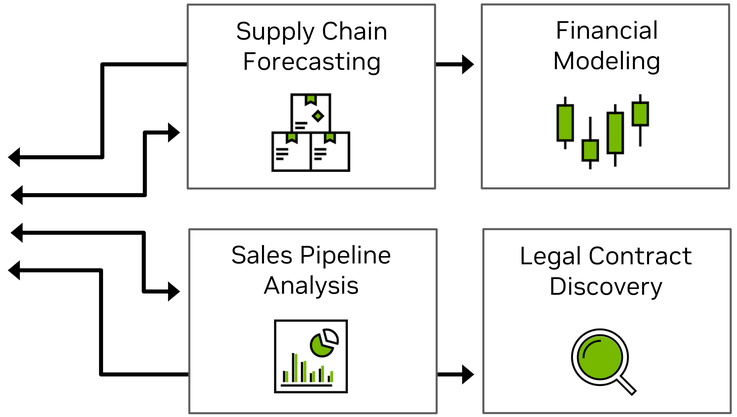


Parameter Efficient Fine-tuning
Add skills and incremental knowledge

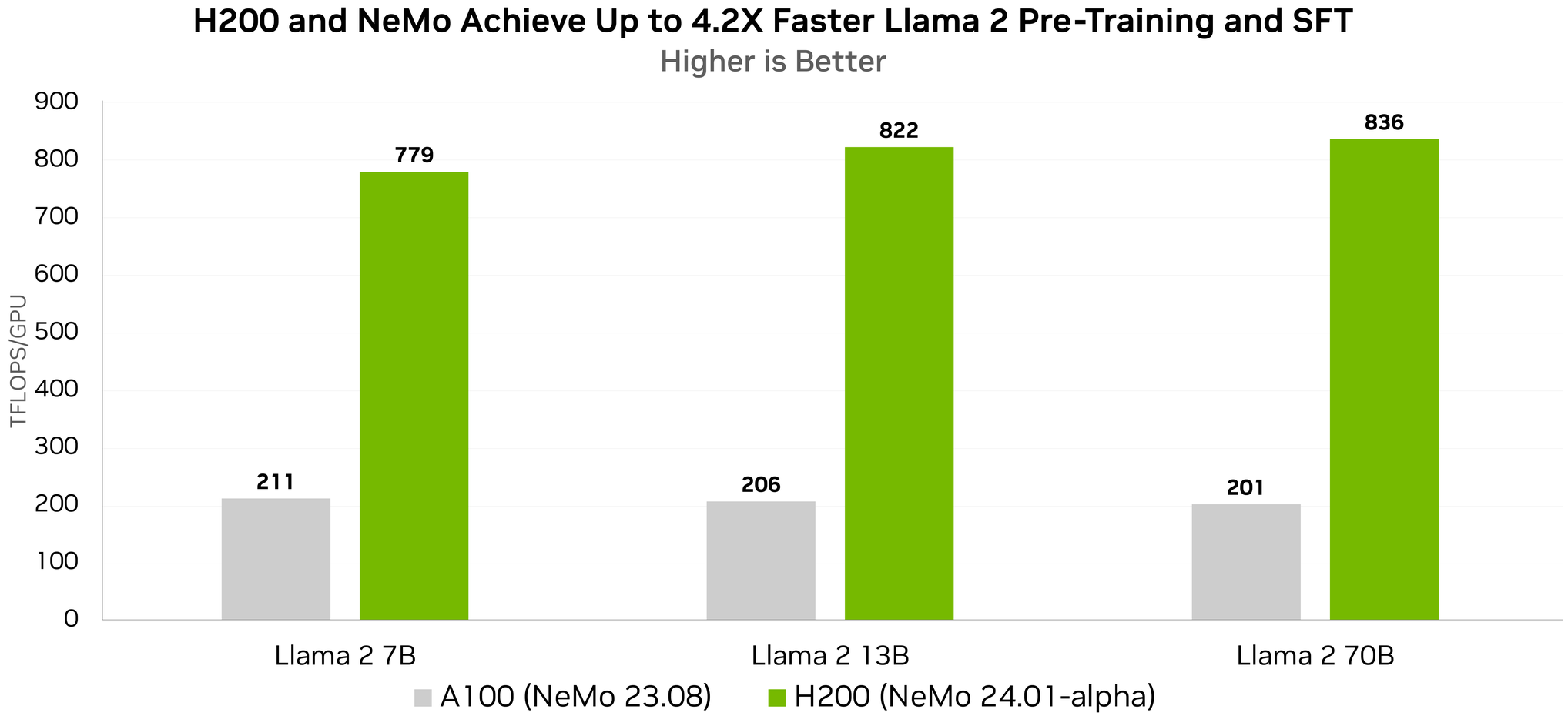
Supervised Fine Tuning
Include domain-specific knowledge

Reinforcement Learning from Human Feedback (RLHF)
Continuously improve model as it is used

Enterprise Model



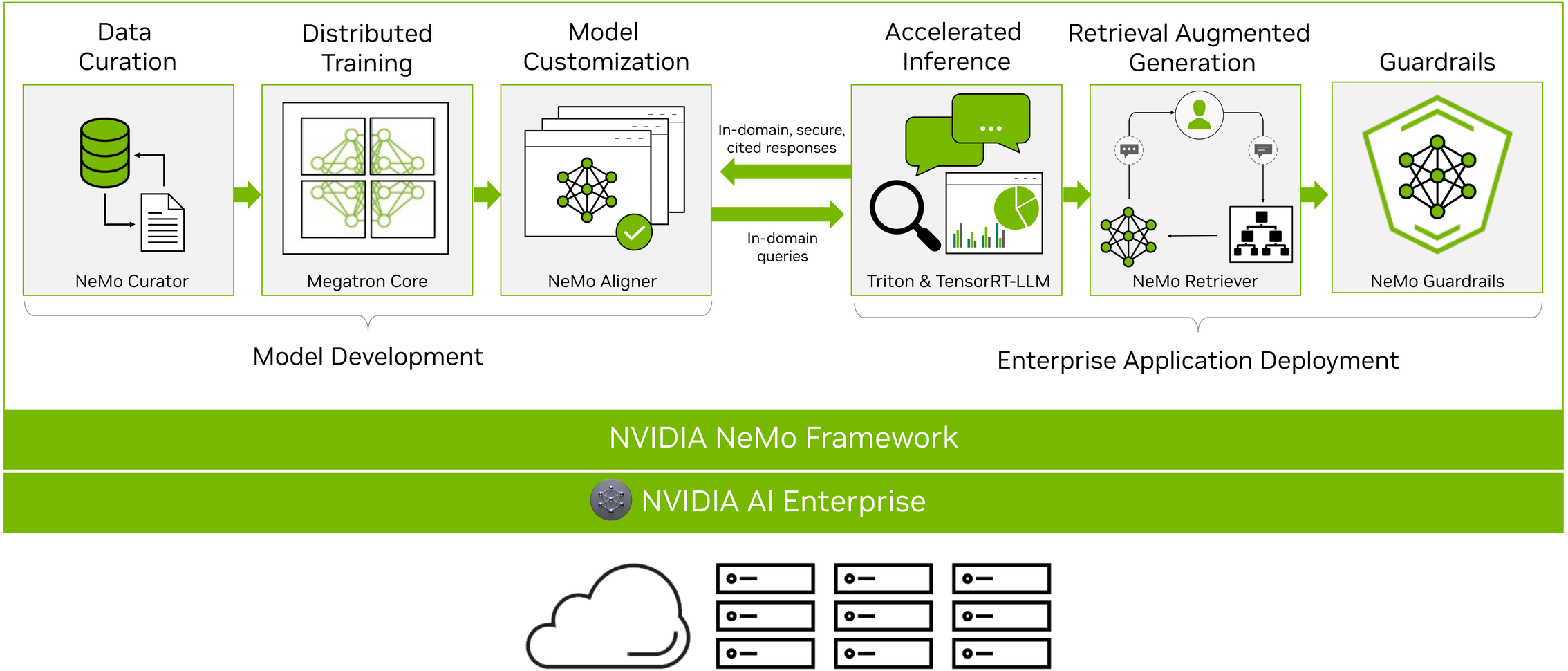
NVIDIA H200 with NeMo Delivers Stunning Performance



Measured performance per GPU. Global Batch Size = 128.
Llama 2 7B: Sequence Length 4096 | A100 8x GPU, NeMo 23.08 | H200 8x GPU, NeMo 24.01-alpha
Llama 2 13B: Sequence Length 4096 | A100 8x GPU, NeMo 23.08 | H200 8x GPU, NeMo 24.01-alpha
Llama 2 70B: Sequence Length 4096 | A100 32x GPU, NeMo 23.08 | H200 8x GPU, NeMo 24.01-alpha

Building Generative AI Applications for the Enterprise

Build, customize and deploy generative AI models with NVIDIA NeMo



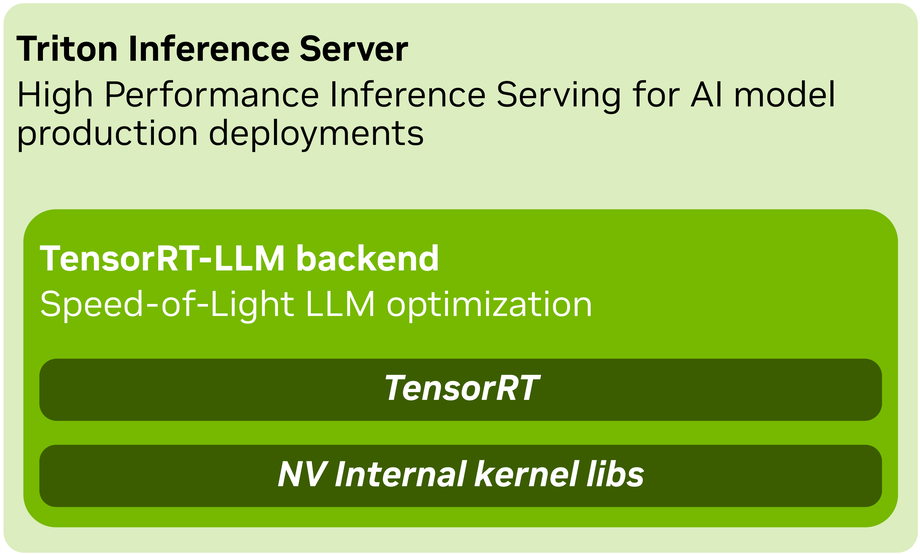
Triton & TensorRT-LLM

NVIDIA AI Inference Platform

- **Triton Inference Server**
  - Triton Inference Server deploy AI model from multiple deep learning and machine learning frameworks. It supports inference across cloud, data center, edge and embedded devices on NVIDIA GPUs, x86 and ARM CPU, or AWS Inferentia. Triton Inference Server delivers optimized performance for many query types, including real time, batched, ensembles and audio/video streaming.

- **TensorRT-LLM**
  - For deployments with high amounts of customization or control required
    - New or customized model architectures
    - Fine tuned control over optimizations, quantization, & performance
  - Model developers & optimizers desiring deep control of optimization
  - Requires higher layers for serving (Triton) and functionality (Guardrails, RAG, etc.) for final deployments

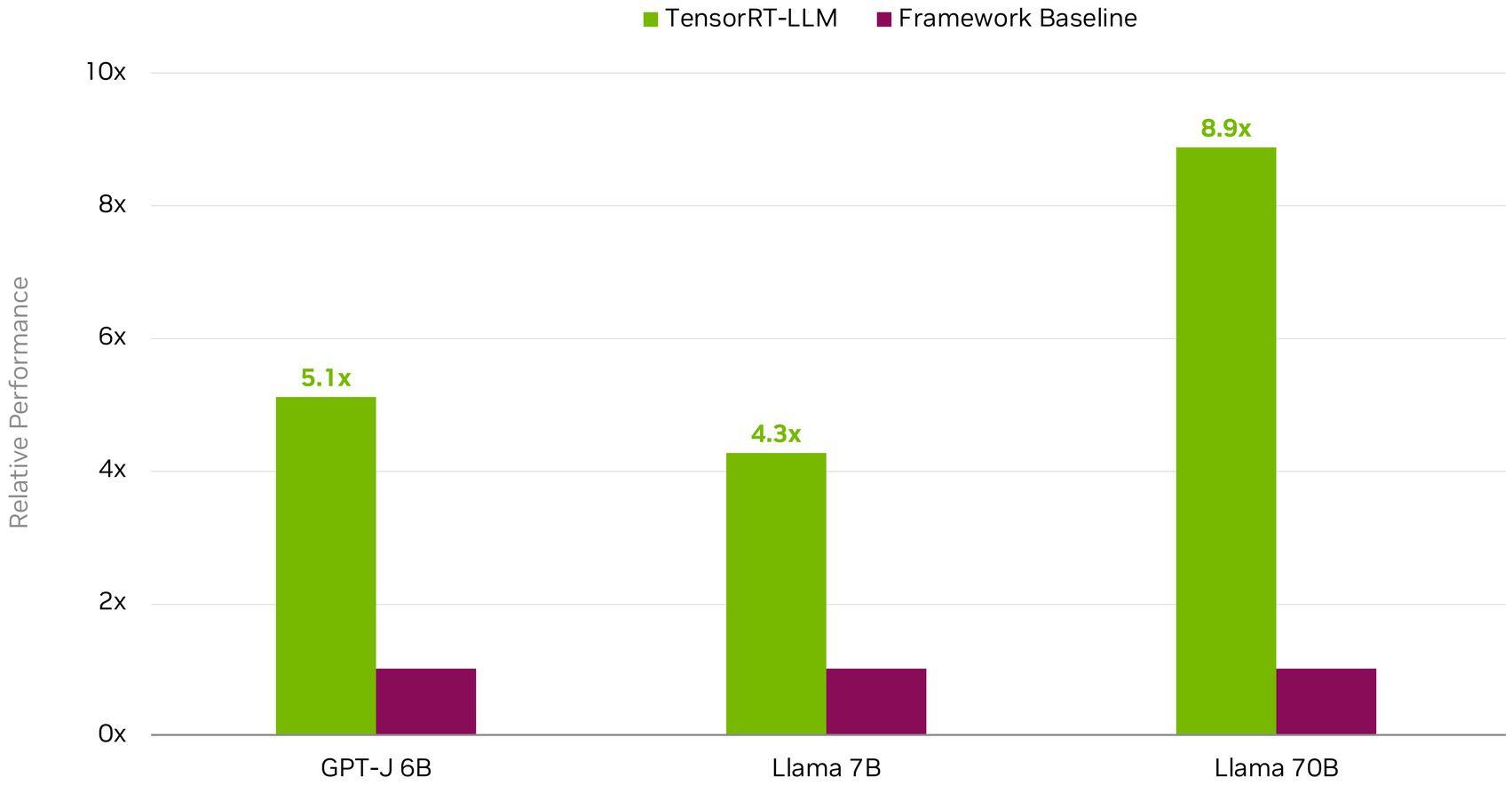
Recommended that most developers start with NeMo Framework



NVIDIA LLM Inference Stack

TensorRT-LLM Performance Improvement

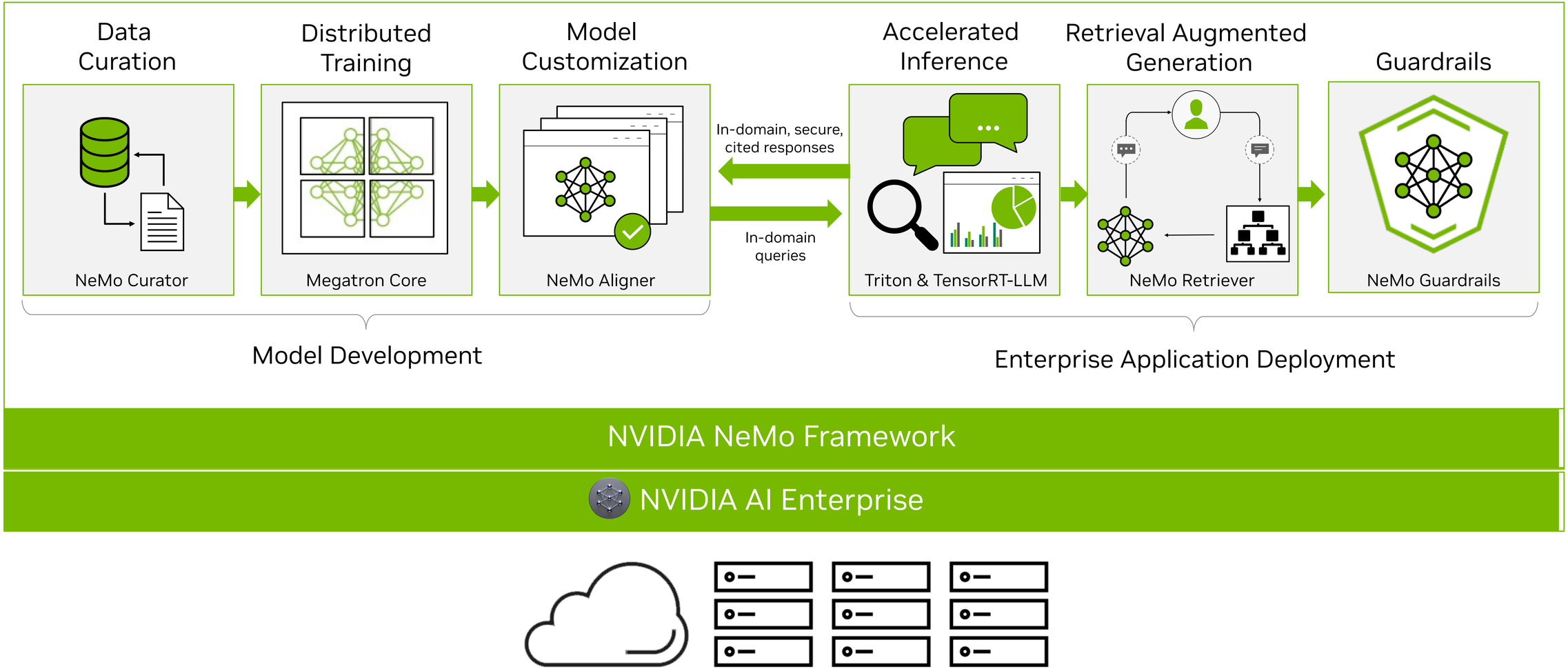
Up to 9x better throughput than implementations in DL frameworks



TensorRT-LLM v0.5.0 internal build. HF Accelerate. Tokens/s/GPU relative improvement
DGX H100. TensorRT-LLM FP8, HF Accelerate FP16.
Max batchsize up to 64. Input & output sequence length 128:128
TensorRT-LLM Llama 70B TP2, HF Accelerate PP4

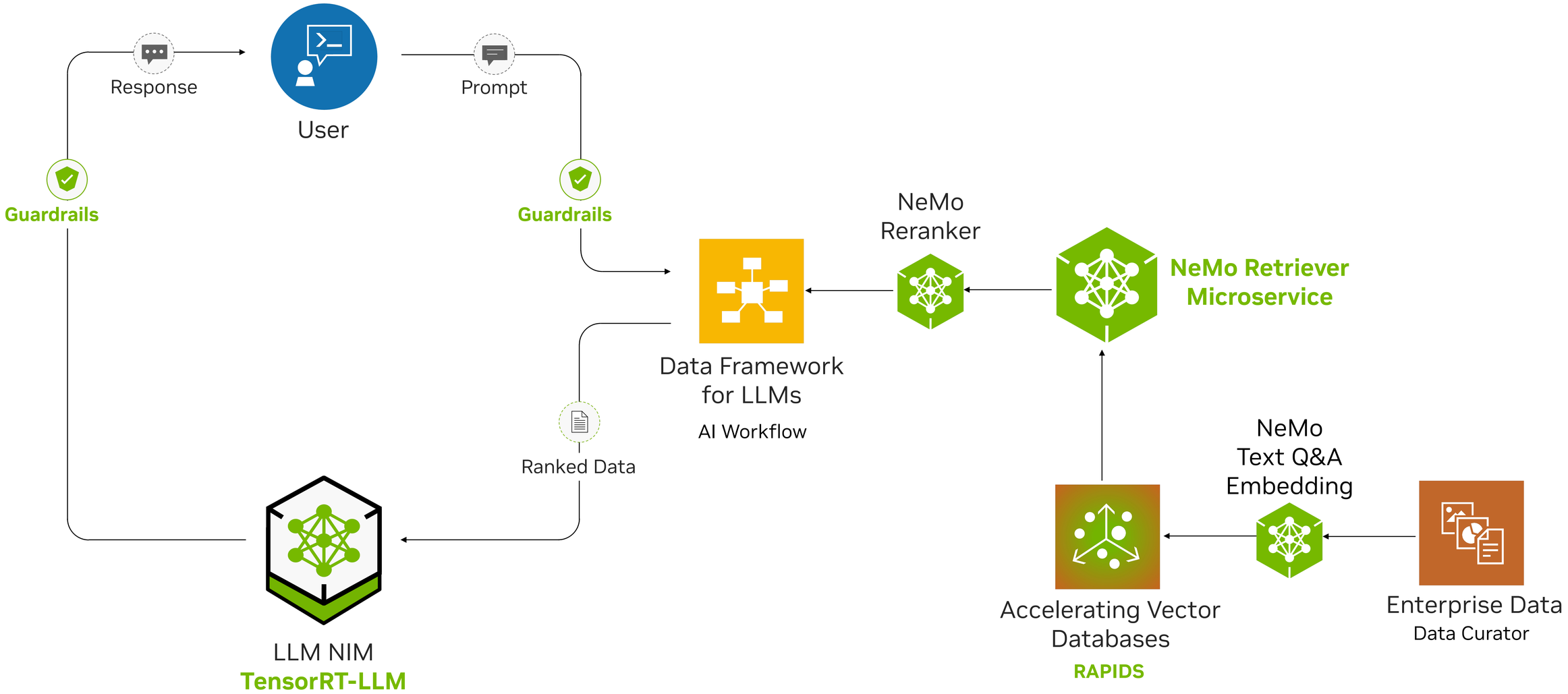
Building Generative AI Applications for the Enterprise

Build, customize and deploy generative AI models with NVIDIA NeMo



NVIDIA Provides Optimized Retrieval Augmented Generation

Commercially viable, optimized embedding, reranking, and personalization to deliver highest accuracy and performance

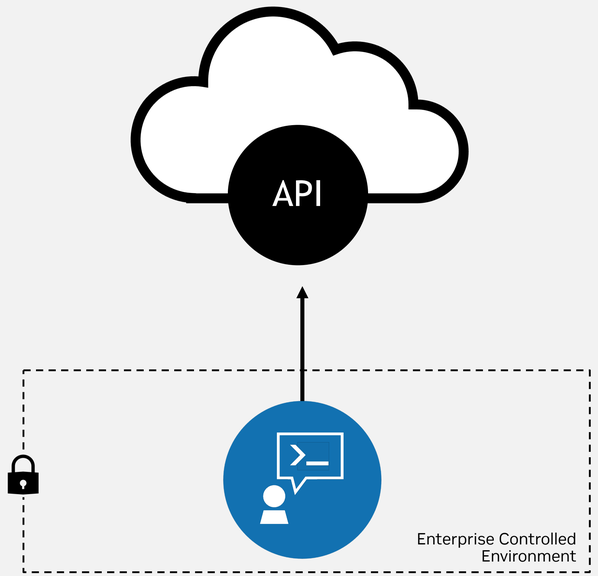


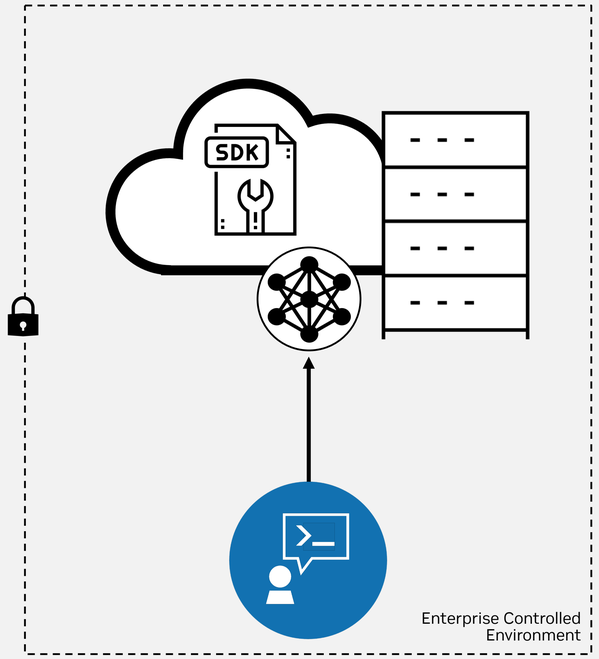
NeMo Microservice

Enterprises Face Challenges Experimenting with Generative AI

Organizations must choose between ease of use and control

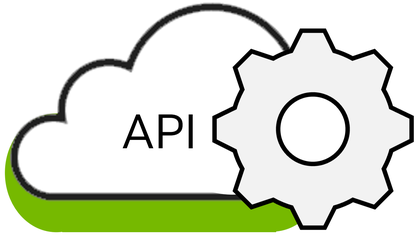
	Managed Generative AI Services	Open-Source Deployment	
Easy to use APIs for development			**Run anywhere** across data center and cloud
Fast path to getting started with AI			**Securely manage** data in self hosted environment
Infrastructure limited to managed environment			**Tuning required** for different infrastructure
Data and prompts are **shared externally**			**Custom code** for APIs and fine-tuned models
Limited control for overall generative AI strategy			**Ongoing maintenance** and updates



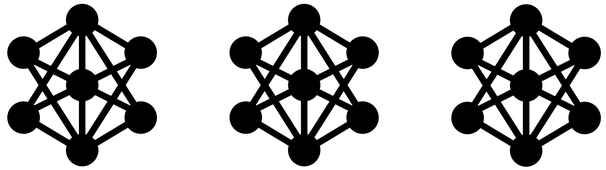


NVIDIA Inference Microservice

Easy-to-use microservice to run generative AI in production at scale anywhere

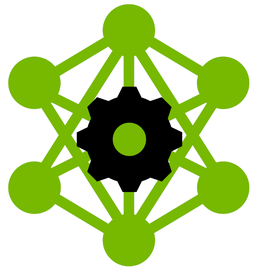


Community AI Models



NVIDIA AI Foundation Models

Custom AI Models



NVIDIA Inference Microservice

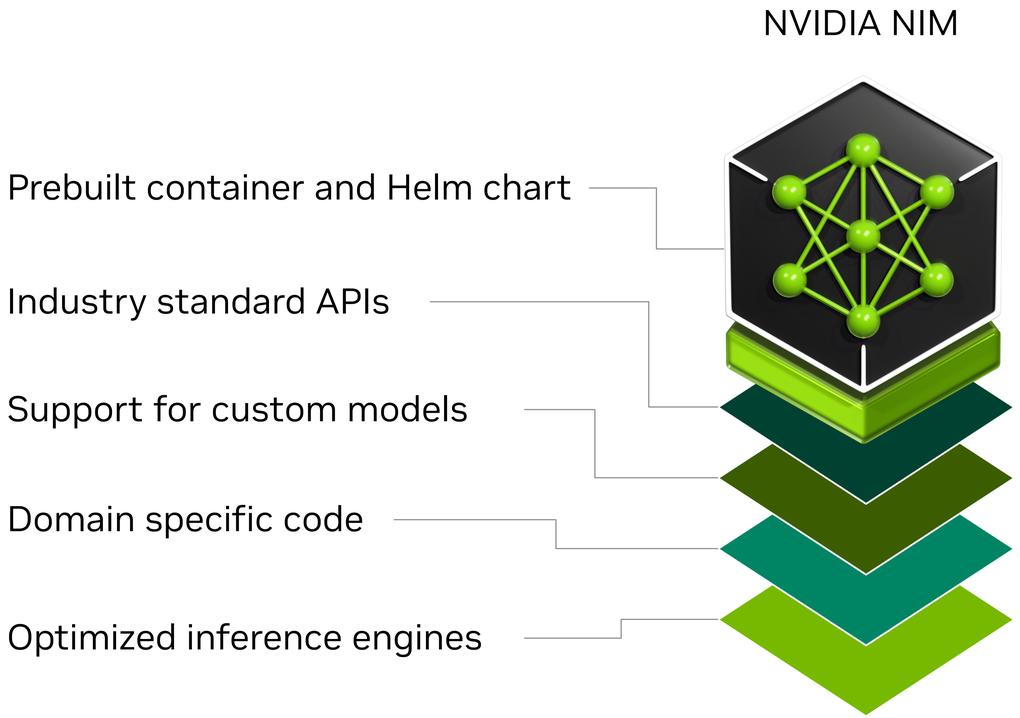
Triton Inference Server

TensorRT-LLM

PyTorch

NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI



Deploy anywhere and maintain control of generative AI applications and data

Simplified development of AI application that can run in enterprise environments

Day 0 support for all generative AI models providing choice across the ecosystem

Improved TCO with best latency and throughput running on accelerated infrastructure

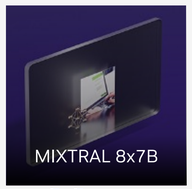
Best accuracy for enterprise by enabling tuning with proprietary data sources

Enterprise software with feature branches, validation and support

Inference Microservices for Generative AI

NVIDIA NIM is the fastest way to deploy AI models on accelerated infrastructure across cloud, data center, and PC

NVIDIA API Catalog

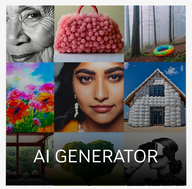


MISTRAL AI_

Google

ADEPT

NVIDIA



gettyimages

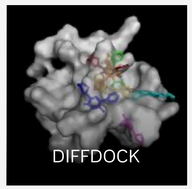
Microsoft

shutterstock

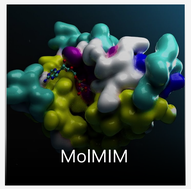
NVIDIA

Meta

NVIDIA



MIT



NVIDIA

NVIDIA NIM for Every Domain

LANGUAGE NIMs


Code Llama 70B


Cohere 35B


Gemma 7B


Jamba


Llama 3 70B


Mistral 7B


Mixtral 8x7B


Nemotron-3 22B Persona


Phi-2

VISUAL / MULTIMODAL NIMs


Adept 110B


Deplot


Edify. Getty


Edify. Shutterstock


FuYu 8B, 55B


Kosmos-2


NeVA


SDXL 1.0


SDXL Turbo

DIGITAL HUMAN NIMs


Audio2Face


Riva ASR

OPTIMIZATION / SIMULATION NIMs


cuOpt


Earth-2

APPLICATION NIMs


Llama Guard


Retrieval Embedding


Retrieval Reranking

DIGITAL BIOLOGY NIMs


DeepVariant


DiffDock


ESMFold


MolMIM


Vista 3D

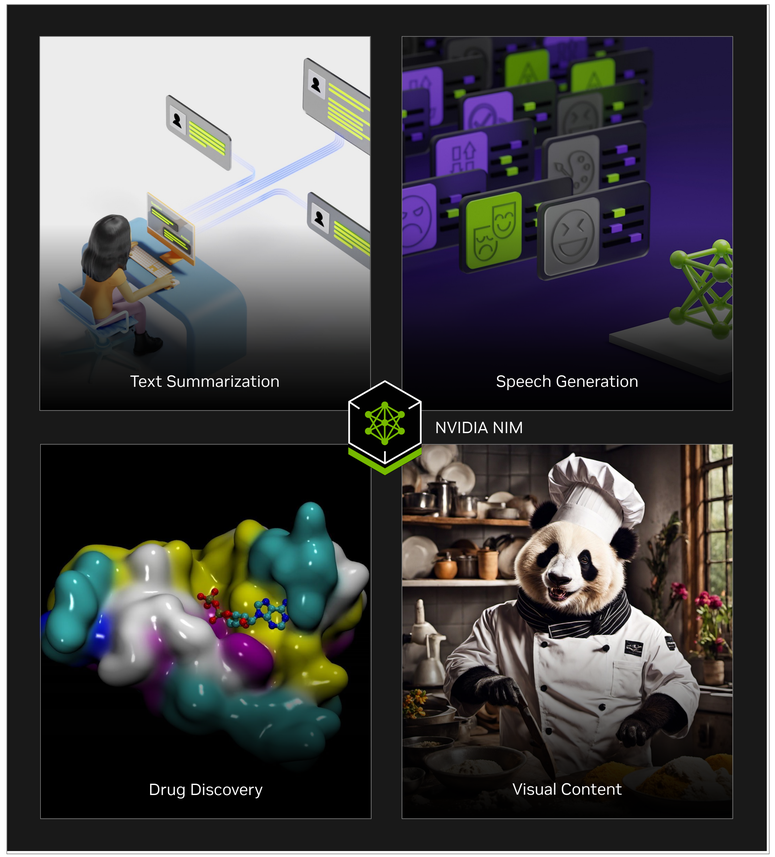
NVIDIA NIM is the Fastest Path to AI Inference

Reduces engineering resources required to deploy optimized, accelerated models

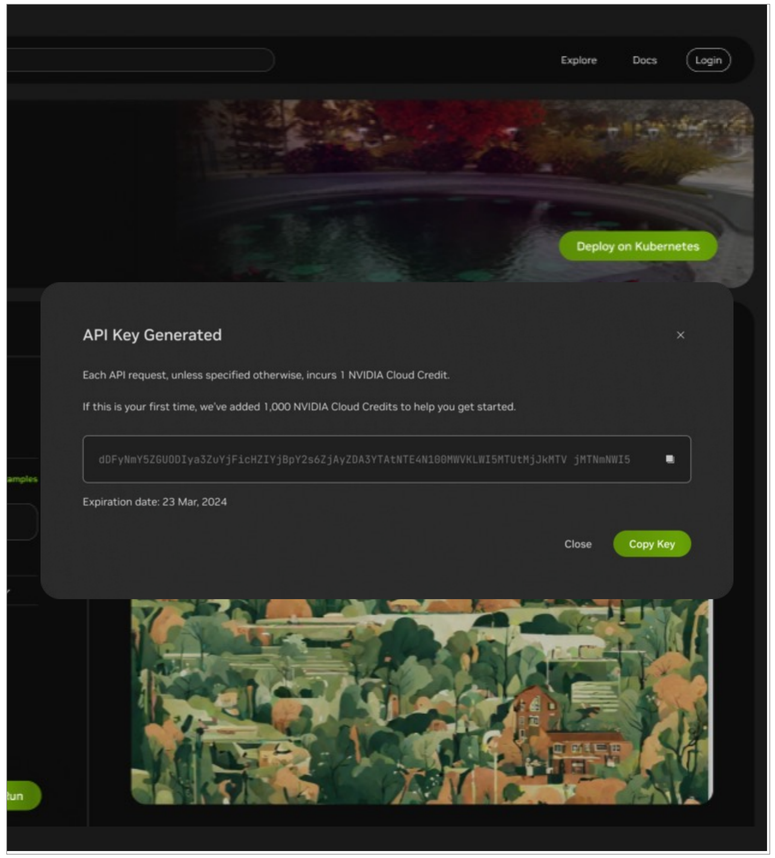
	NVIDIA NIM	Triton + TRT-LLM Opensource
Deployment Time	5 minutes	~1 week
API Standardization	Industry standard protocol OpenAI for LLMs, Google Translate Speech	User creates a shim layer (reducing performance) or modify Triton to generate custom endpoints
Pre-Built Engine	Pre-built TRT-LLM engines for NV and community models MISTRAL AI_ Llama 2 starcoder NVIDIA Nemotron	User converts checkpoint to TRT-LLM format and creates and runs sweeps through different parameters to find the optimal config
Triton Ensemble/ BLS Backend	Pre-built with TRT-LLM to handle pre/post processing (tokenization)	User manually sets up + configures
Triton Deployment	Automated	User manually sets up + configures
Customization	Supported – P-tuning and LORA, more planned	User needs to create custom logic
Container Validation	Pre-validated with QA testing	No pre-validation
Support	NVIDIA AI Enterprise - Security and CVE scanning/patching and tech support	No enterprise support

Experience and Run Enterprise Generative AI Models Anywhere

Seamlessly integrate AI in business applications with NVIDIA AI APIs



Experience Models



Prototype with APIs

Deploy with NIMs

Building Generative AI Applications for the Enterprise

Build, customize, and deploy generative AI models with NVIDIA NeMo.