NVIDIA's Agentic AI Tools:
NeMo, Nemotron, and NIM Microservices

Nathan Stephens
ASU Computing Expo 2025

Evolution of AI

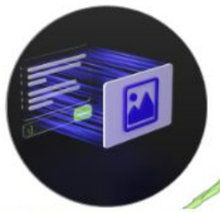
Agentic AI Enables More Powerful AI Applications

PHYSICAL AI

SELF-DRIVING CARS
GENERAL ROBOTIC

AGENTIC AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE



GENERATIVE AI

DIGITAL MARKETING
CONTENT CREATION

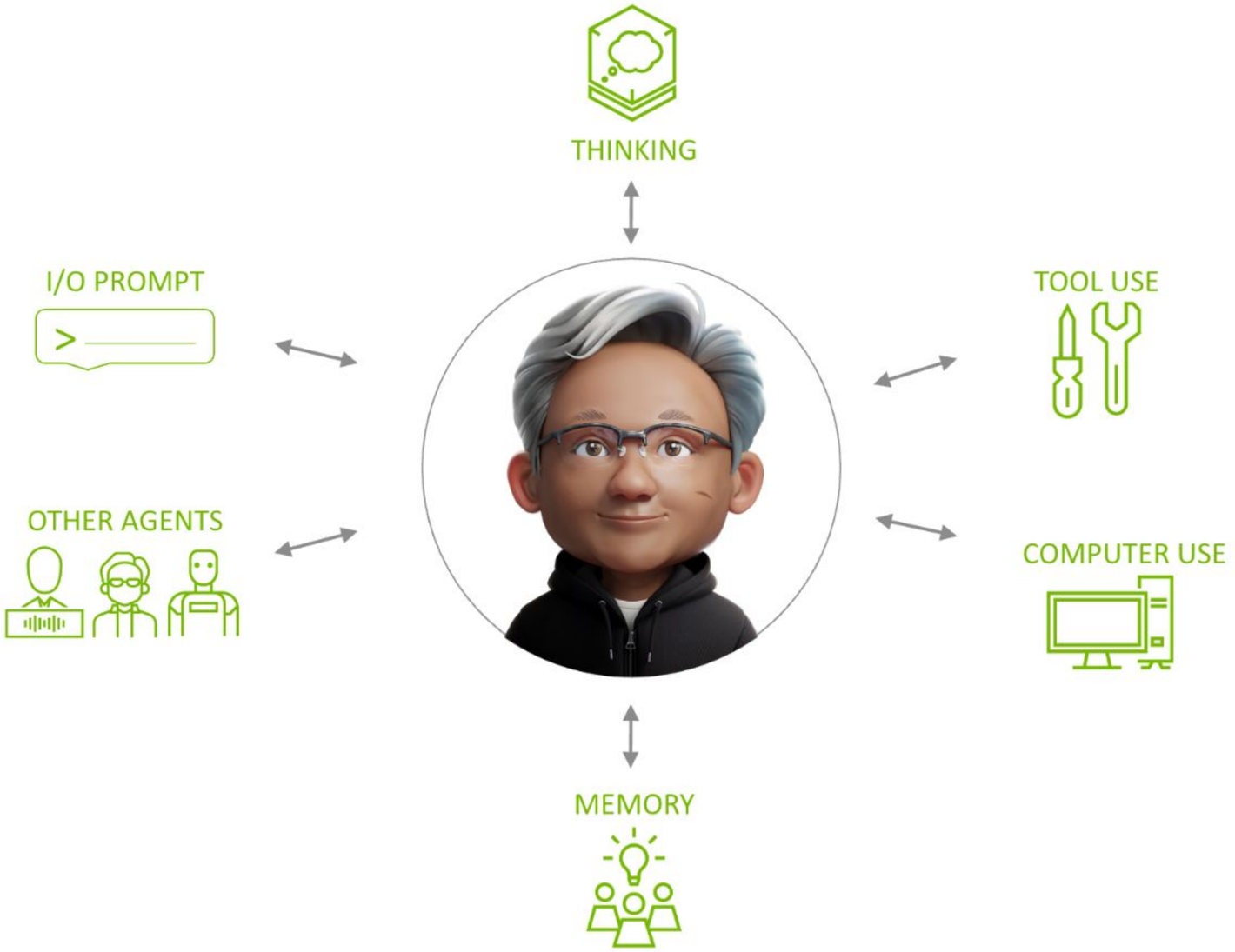
PERCEPTION AI

SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

2012 ALEXNET

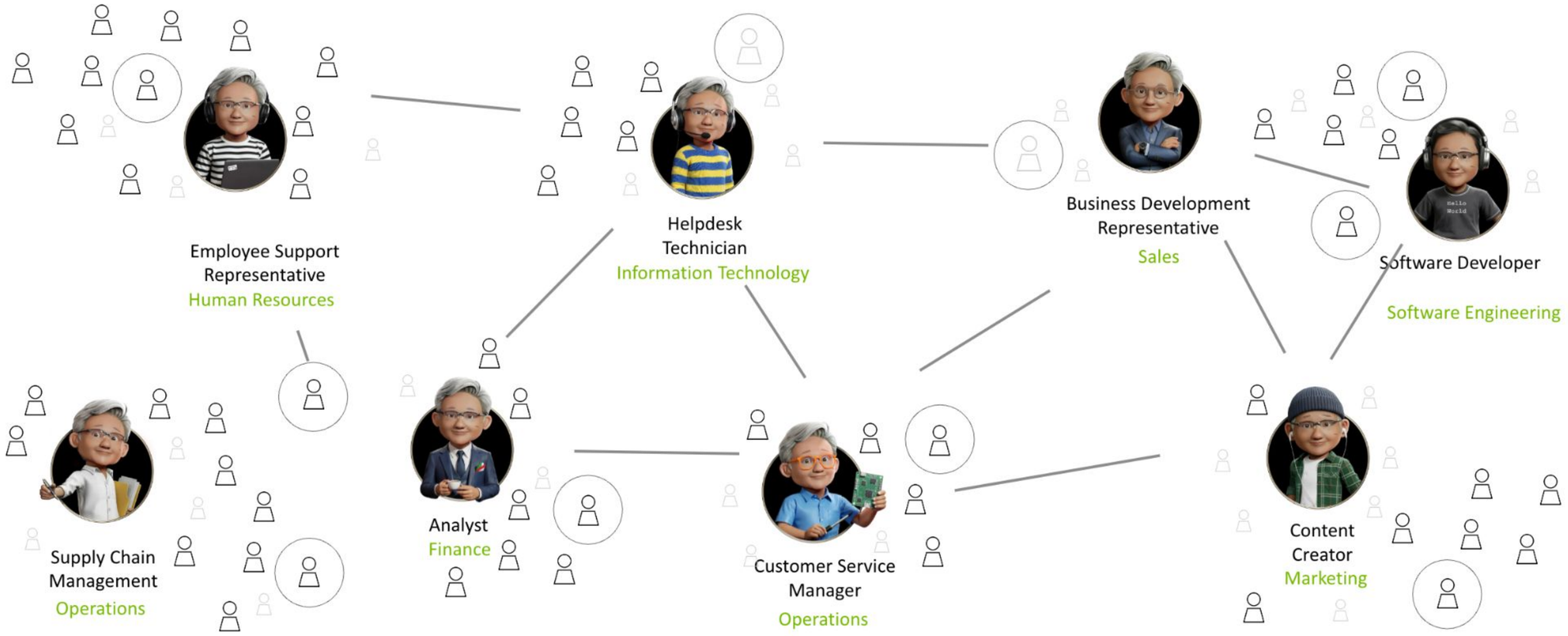
How AI Agents Work

AI Agents use Tools and Collaborate to Complete Complex Projects



Agents work together to solve complex problems

AI Agents will drive performance gains, better problem solving, and faster time to action



Building Agentic Systems is Complex

Agentic applications should integrate with existing tools, heterogeneous data, and perform reliably

Architectural Complexity
Many tools exist—but don't work together

Repeatability
Challenging to guarantee consistent results

Code Reuse
Fragmented solutions result in duplicate work

Performance
System-level accelerations require knowledge of the entire system

Learning & Adaptation
Continuous feedback and improvements

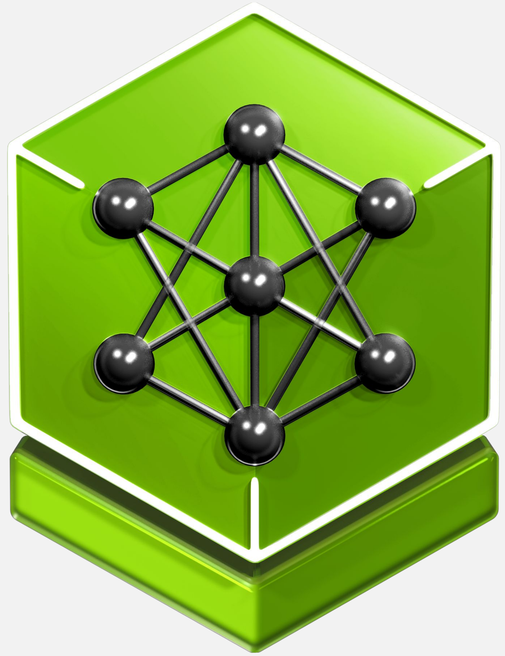
Safety, Security & Privacy
Preventing unwanted behaviors and protecting proprietary data

Enterprise Readiness
Evaluation system and telemetry

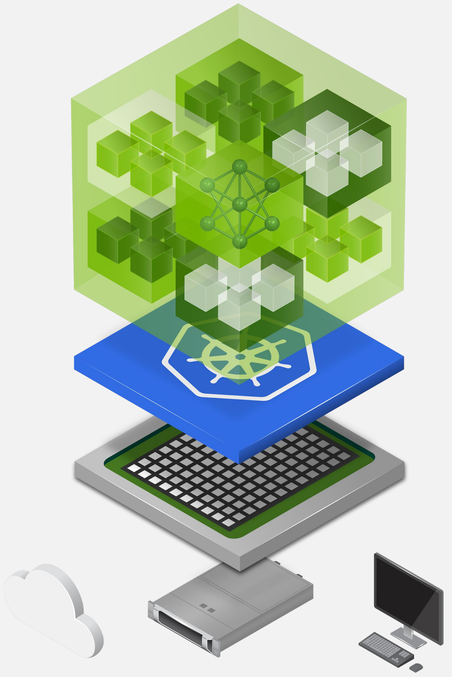
User-AI Interaction
Natural, effective communications and understanding

NVIDIA’s Agent Platform

AI Agent Lifecycle Management
NVIDIA NeMo



Accelerate Leading Open Models
NVIDIA Nemotron

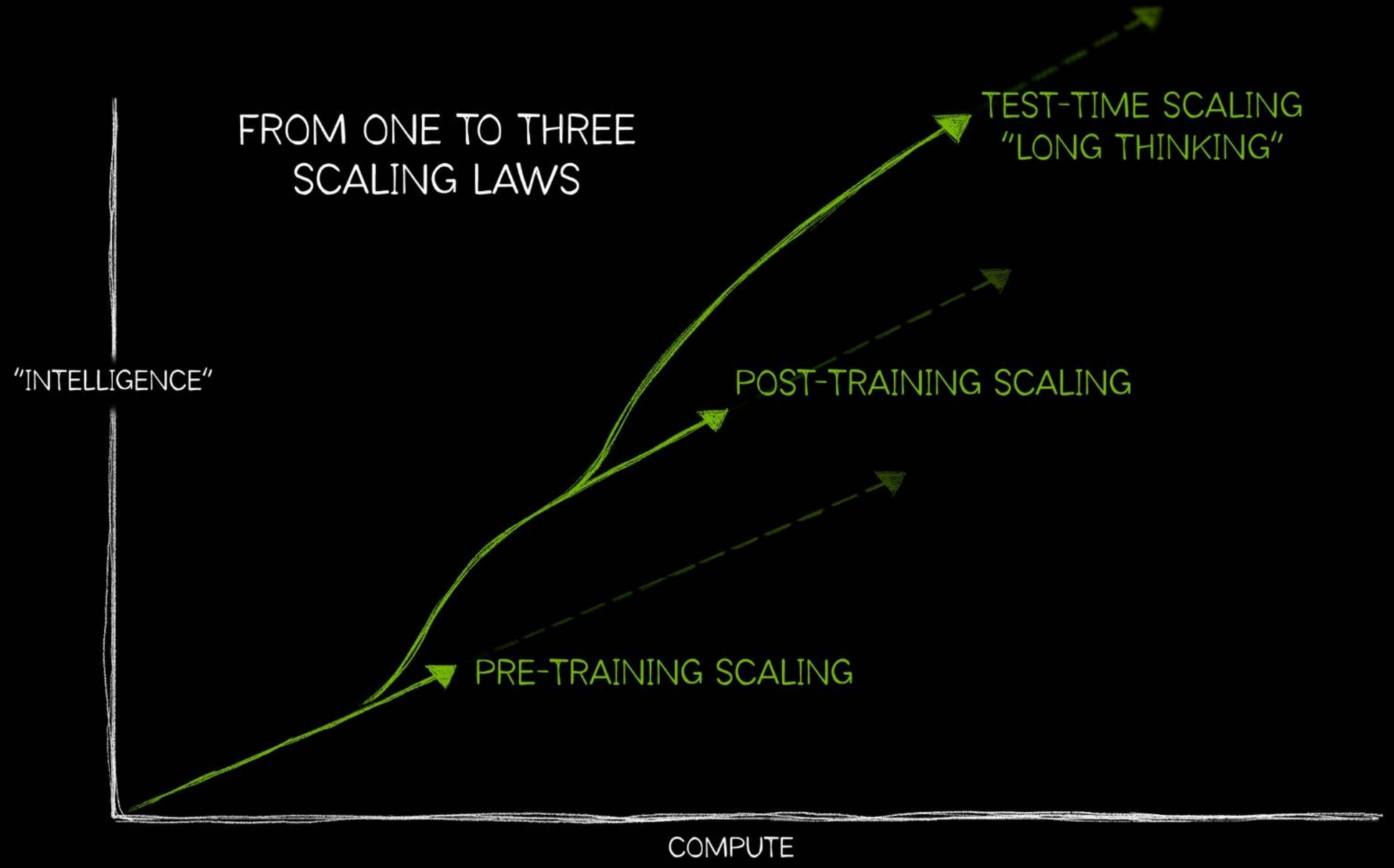


Package Open Models for Production
NVIDIA NIM

NeMo

AI Agent Lifecycle Management
NVIDIA NeMo

How Scaling Laws Drive Smarter, More Powerful AI



NVIDIA NeMo for Managing the AI Agent Lifecycle

Creating Specialized Agents

NeMo Data Designer
Generate

NeMo Curator
Process

NeMo RL
Adapt

NeMo Evaluator
Evaluate

NeMo Guardrails
Guardrail

NeMo Retriever
Learn

NeMo Agent Toolkit
Connect, Evaluate, Optimize

Build

NeMo Data Designer

Design high-quality, domain-specific data from scratch

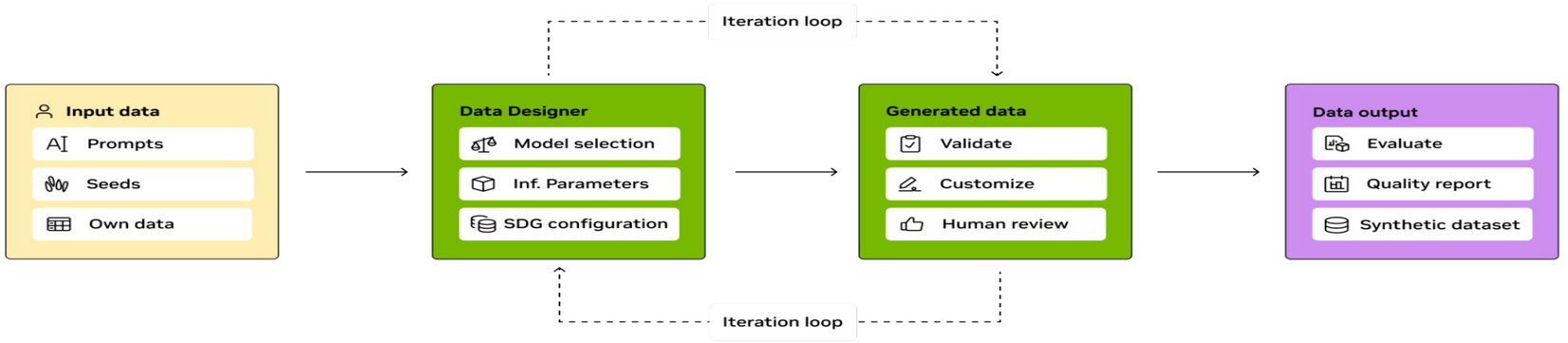
NeMo Data Designer is a compound AI system that combines large language models, pre-trained components, and user-defined configurations to produce accurate and realistic synthetic datasets.

High-Quality
Accurate, realistic synthetic datasets built for production-grade AI.

Customizable
Easily configurable to integrate any LLM such as Nemotron, GPT-5, etc.

Scalable
Optimized for high-throughput synthetic data generation at scale.

Secure
Avoids sensitive data exposure with safe, synthetic alternatives.



NVIDIA

Build

NVIDIA NeMo Curator

Python SDK for scalable, configurable pipelines to curate text, image and video datasets for higher model accuracy

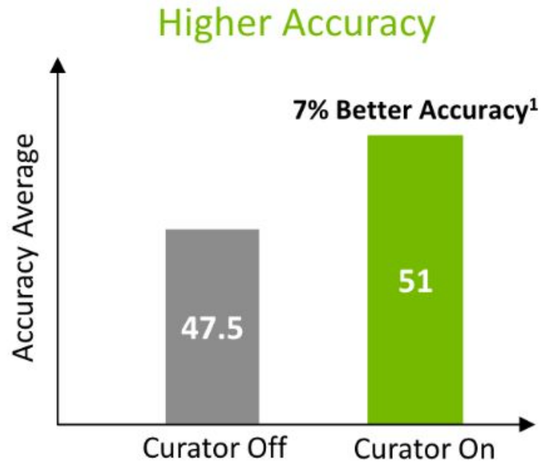
Efficient: Improve accuracy with less data and compute

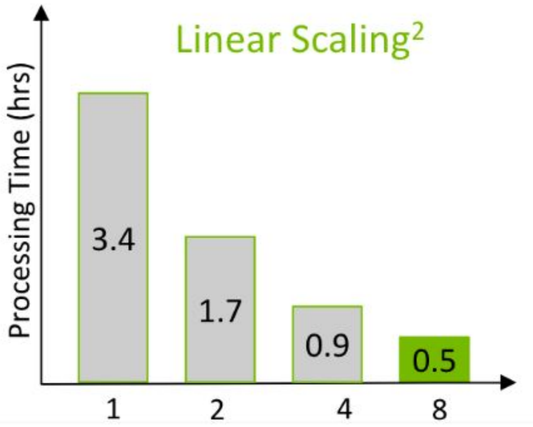
Fast: Accelerated data processing

Scalable: Easily scale to process large datasets

Robust: SOTA classifier models for safety, content, and diversity

Secure: Run anywhere - Python APIs in a customizable and modular OSS library



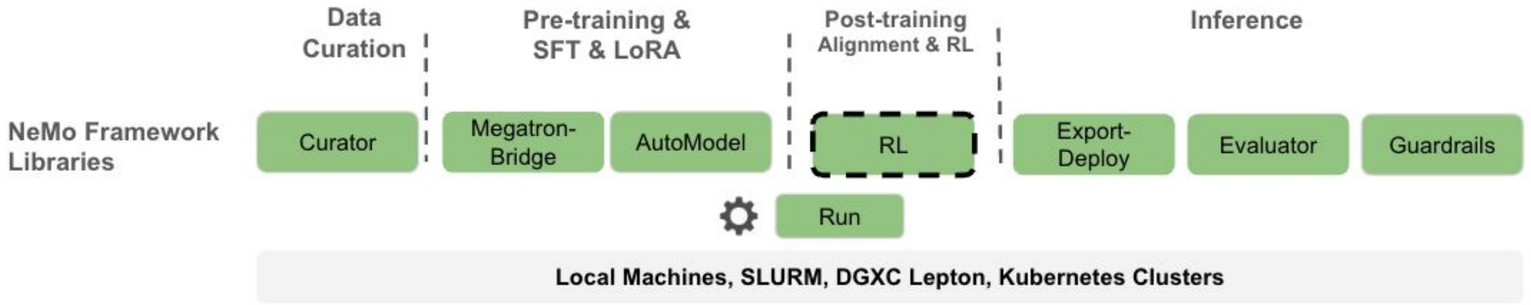


Build

Nemo-RL

Efficient Reinforcement Learning from single to thousands of GPUs

Enables developers and researchers to efficiently post-train, and align language models and multimodal models at scale.



PAIN POINT	NeMo RL BENEFIT
Reinforcement learning on large models is slow and resource-intensive	Ray-based distributed orchestration ensures scalable, efficient, and fault-tolerant training
Limited flexibility for experimentation and research	Modular, PyTorch-native design enables quick iteration and easy algorithm customization
Fragmented toolchains across training, rollout, and inference	Unified, end-to-end framework for training, rollout, evaluation, and optimization
Inconsistent performance across backends	Automatic selection between DTensor and Megatron Core for optimal efficiency
Lack of integration with existing ecosystems	Hugging Face integration provides access to a wide range of pre-trained models. PyTorch native for training and vLLM for generation backend
Scaling across hardware and environments is complex	Optimized training across multi-node, multi-GPU setup with Megatron-core backend

Build

NVIDIA NeMo Evaluator

1 API call to evaluate 20+ academic benchmarks & custom metrics

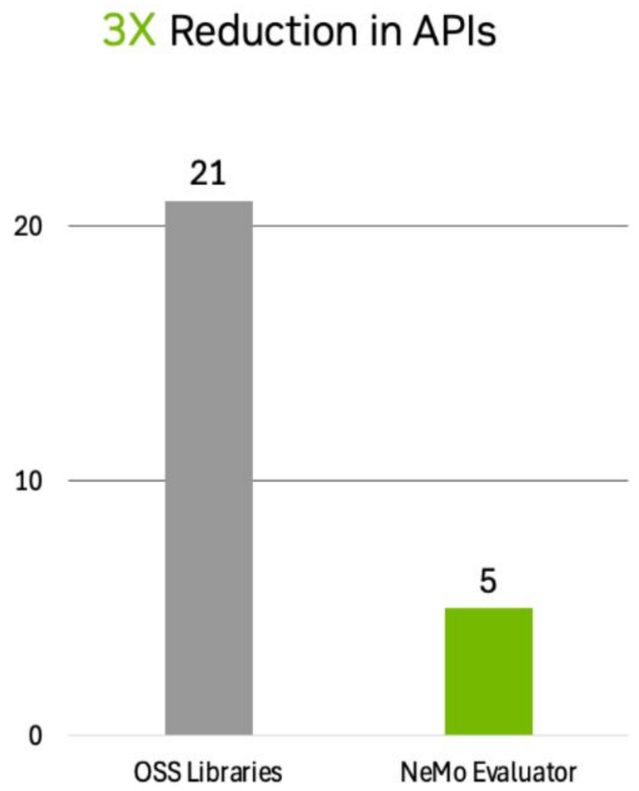
Flexible: Full control of LLMs and AI pipelines evals for custom and standard benchmarks

Simple: Run evals at scale with a single API call anywhere with full data control and maintain CI/CD

Repeatable: Run consistent evals across teams with varying benchmark versioning using config files

Top-of-tree: Enterprise-grade, supports latest benchmarks with security vulnerability patching

Secure: Runs anywhere - on-prem and in the cloud with Kubernetes

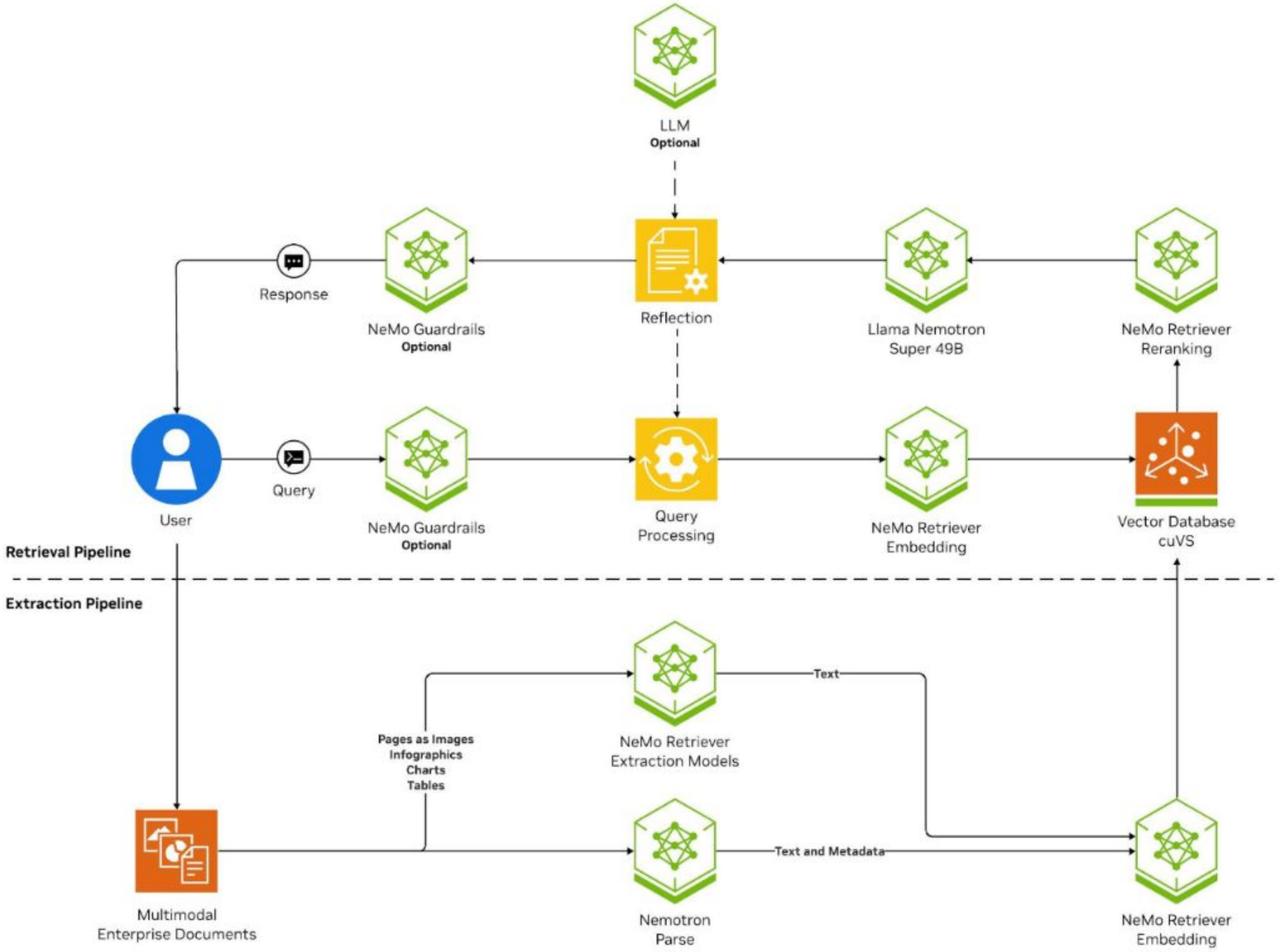


Deploy

NVIDIA NeMo Retriever Accelerates RAG Applications

NVIDIA open, commercial microservices power enterprise RAG pipelines turning data into knowledge

- Optimized Inference Engines, Built on NVIDIA NIM
- State-of-the-art, customizable models, fine-tuned for accuracy
- Flexible and modular deployment
- Accelerated vector search
- Production Ready



Try RAG Blueprint at build.nvidia.com

NVIDIA

Deploy

NVIDIA NeMo Guardrails

Open-source, enterprise grade microservice to meet safety & security for your Gen AI solution

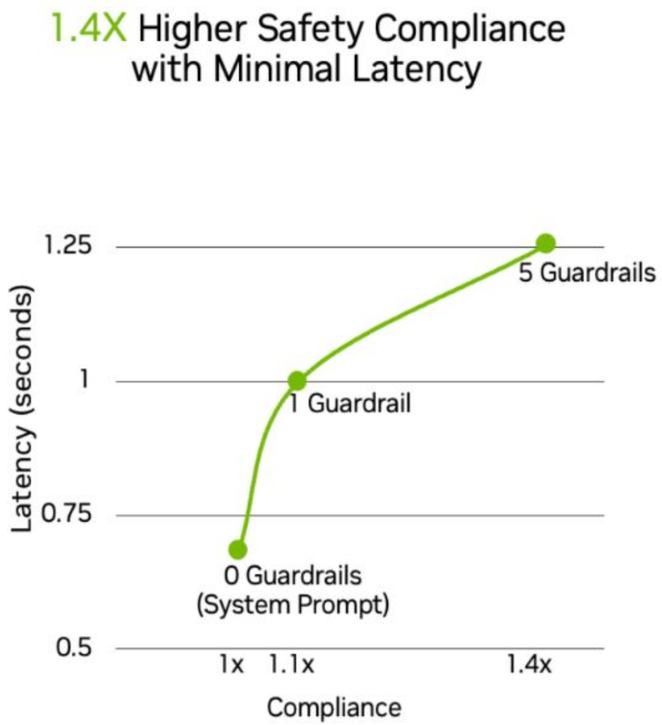
Industry Leading Guardrail: Content moderation, reduce hallucinations, toxicity, off-topic dialogs, PII-breach, and security jailbreaks

Customizable: Build rails targeted for your specific use-cases while adhering to own enterprise policies

Rail Orchestration: Rails helps to improve protection rate while reducing latency, increasing throughput and lowering costs

Easy Integration: Compatible with 3rd party APIs, such as ActiveFence, Fiddler, DataRobot etc.

Native Integration: Popular Gen AI app developer tools integration (LangChain & LlamaIndex)



Optimize

NVIDIA NeMo Agent Toolkit

An open-source library for building enterprise-ready agentic systems

Profiling & Optimizations

- Fine-grained AI workflow telemetry collected can be used to implement agentic system accelerations.

Evaluation & Observability

- Evaluate system level accuracy
- Understand and debug inputs and outputs for each component in the AI workflow

Agent Interconnect

- Universal descriptors for agents, tools, and workflows across frameworks
- Reusable Agent/Tool registry
- Workflow Configuration/Builder

Nemo Framework

Full Stack, End-to-End LLM Building Framwork

Performance & Scalability

- More than 800 TFLOPs/sec/GPU FP8 Hopper
- Trained over 16k+ cluster size
- Supports 1M+ sequence length
- 4D parallelism
- GPU-accelerated data curation

Model Coverage

- Broad support for HF models
- SOL accelerated support for most popular model families - Incl LLM, SSMs, MOEs, SD, VLMs, VFMs, VLAs

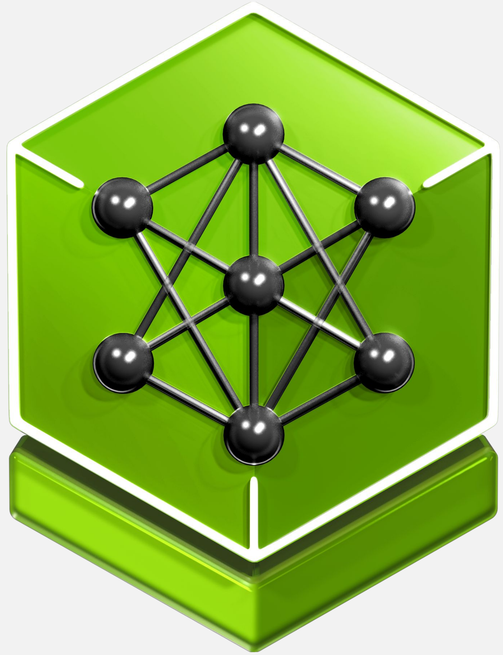
SOTA Algorithms

- PEFT: LoRA, p-tuning, IA3, QLoRA, Adapters (Canonical)
- Reinforcement learning & Model alignment: GRPO, RLHF PPO, DPO, KTO, IPO, RLAIF, SteerLM, Rejection Sampling

Usability & Compatibility

- Hugging-face like pythonic APIs
- Fault tolerance and Resiliency to ensure smooth training via NVRx

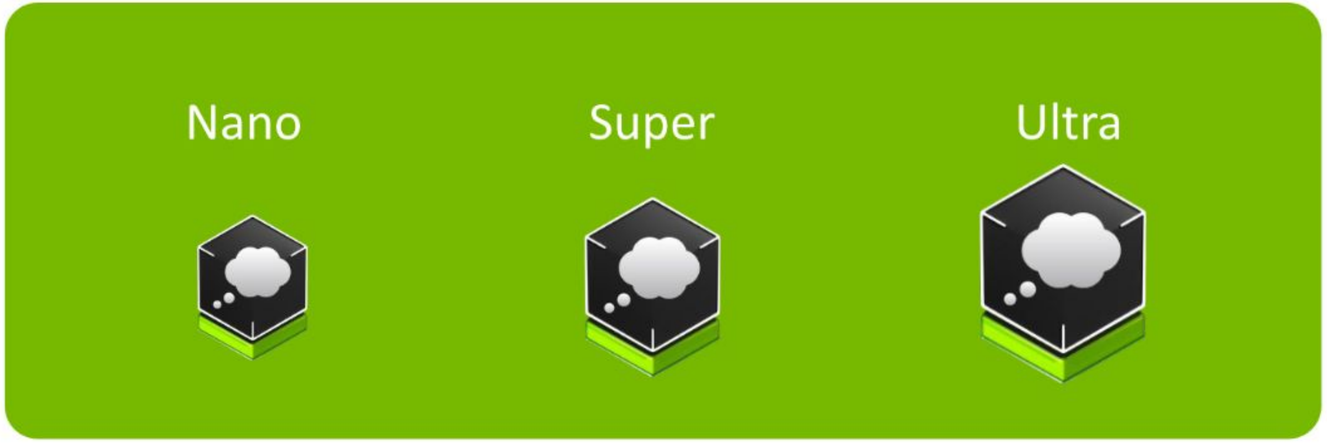
Nemotron



Accelerate Leading Open Models
NVIDIA Nemotron

Announcing Llama Nemotron Reasoning Model Family

Leading Open Reasoning NIM Microservices for Agentic AI



NVIDIA NeMo Framework

Post-Training

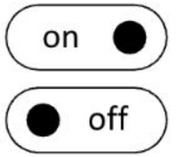
Open NVIDIA Datasets

Model-Building Recipe

60B Tokens, 360K H100 Hours, 45K Annotation Hours

Leading Accuracy

Highest Efficiency

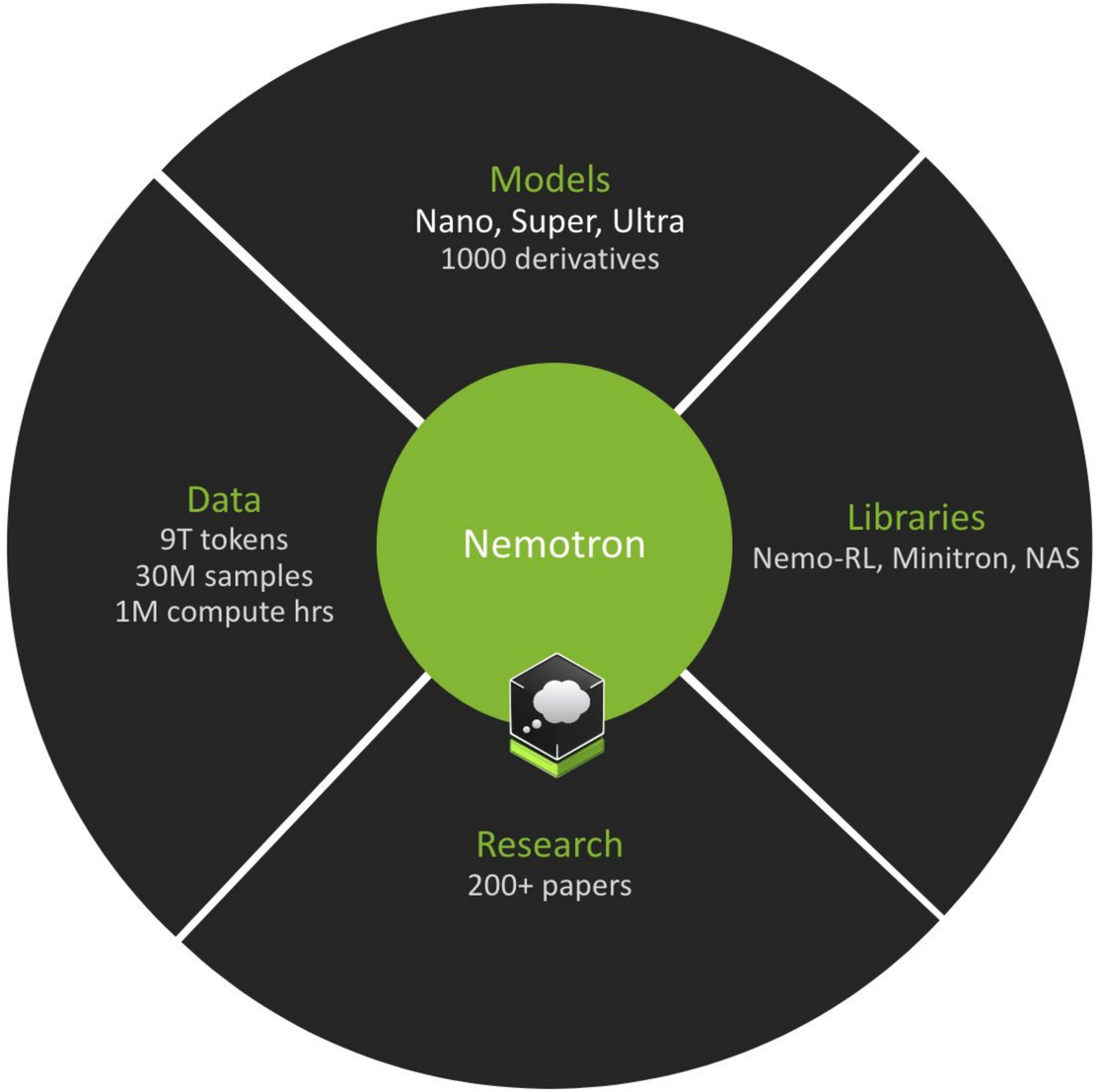


Reasoning ON/OFF

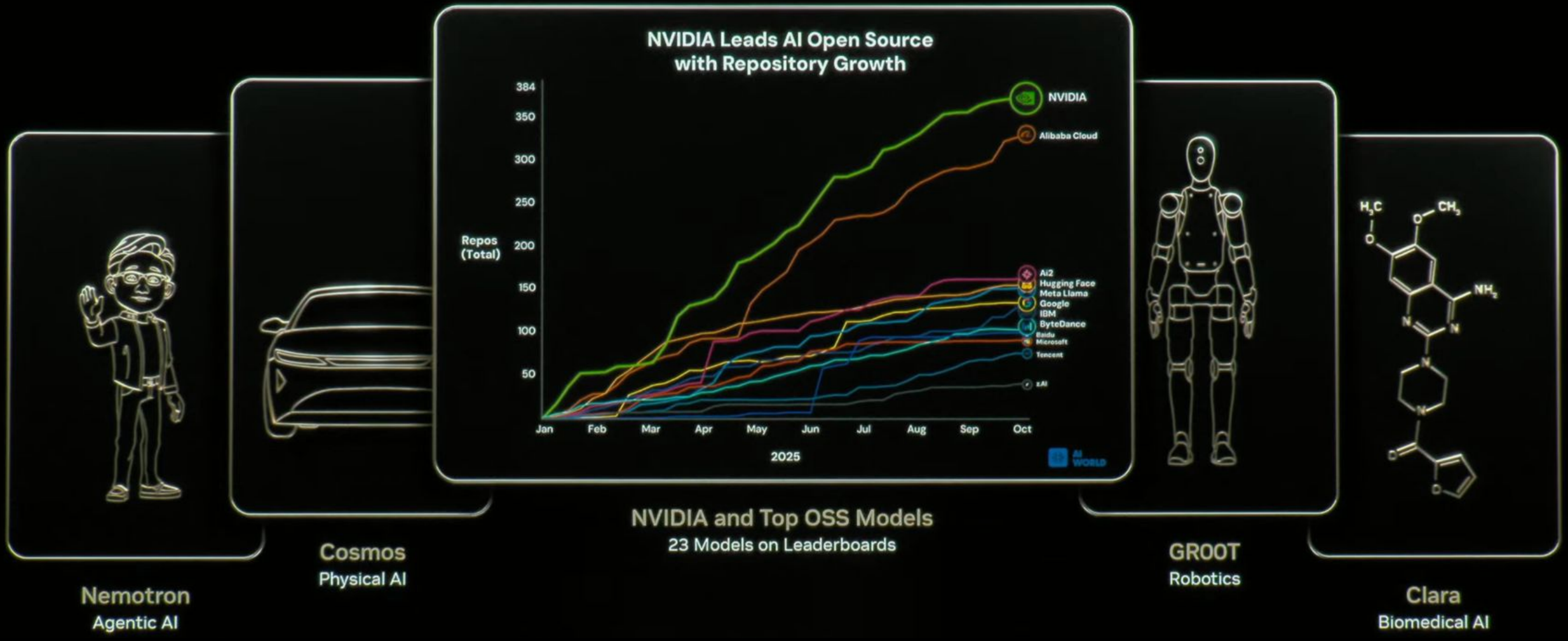
Enterprise Ready

Open

NVIDIA Nemotron – Open Family of AI Models, Datasets and Techniques



NVIDIA Open Models, Data, Libraries Top Leaderboards



Nemotron
Agentic AI

Cosmos
Physical AI

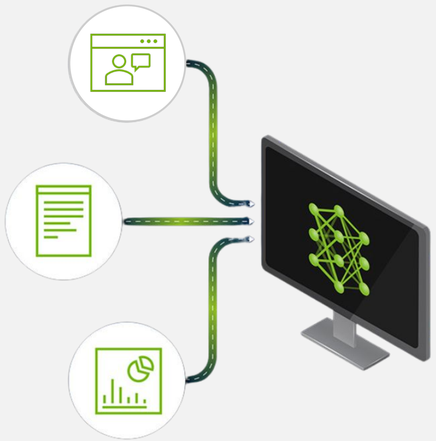
NVIDIA and Top OSS Models
23 Models on Leaderboards

GR00T
Robotics

Clara
Biomedical AI

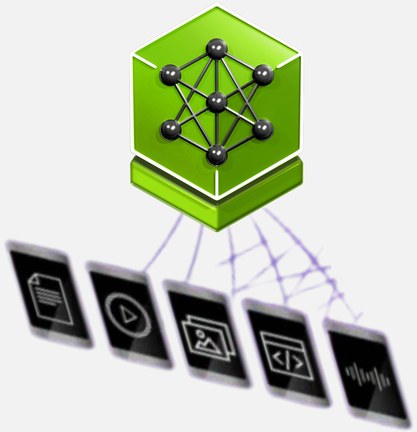
New Nemotron Models

Highly Efficient Models with Open Datasets and Recipes for Agentic AI



Nemotron Nano 2 VL

Video understanding and Image reasoning

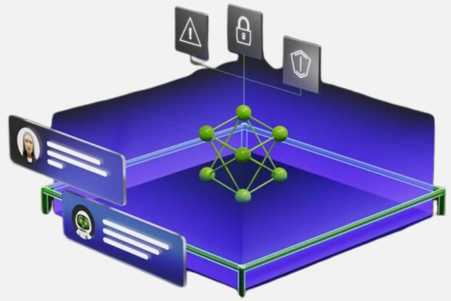


Nemotron Parse 1.1

Document extraction

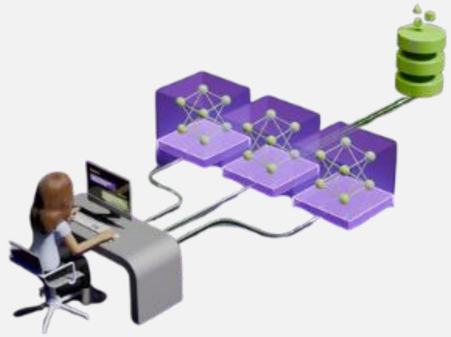
Nemotron Nano 3

Advanced reasoning for agentic AI



Llama Nemotron Safety Guard

Multilingual content safety

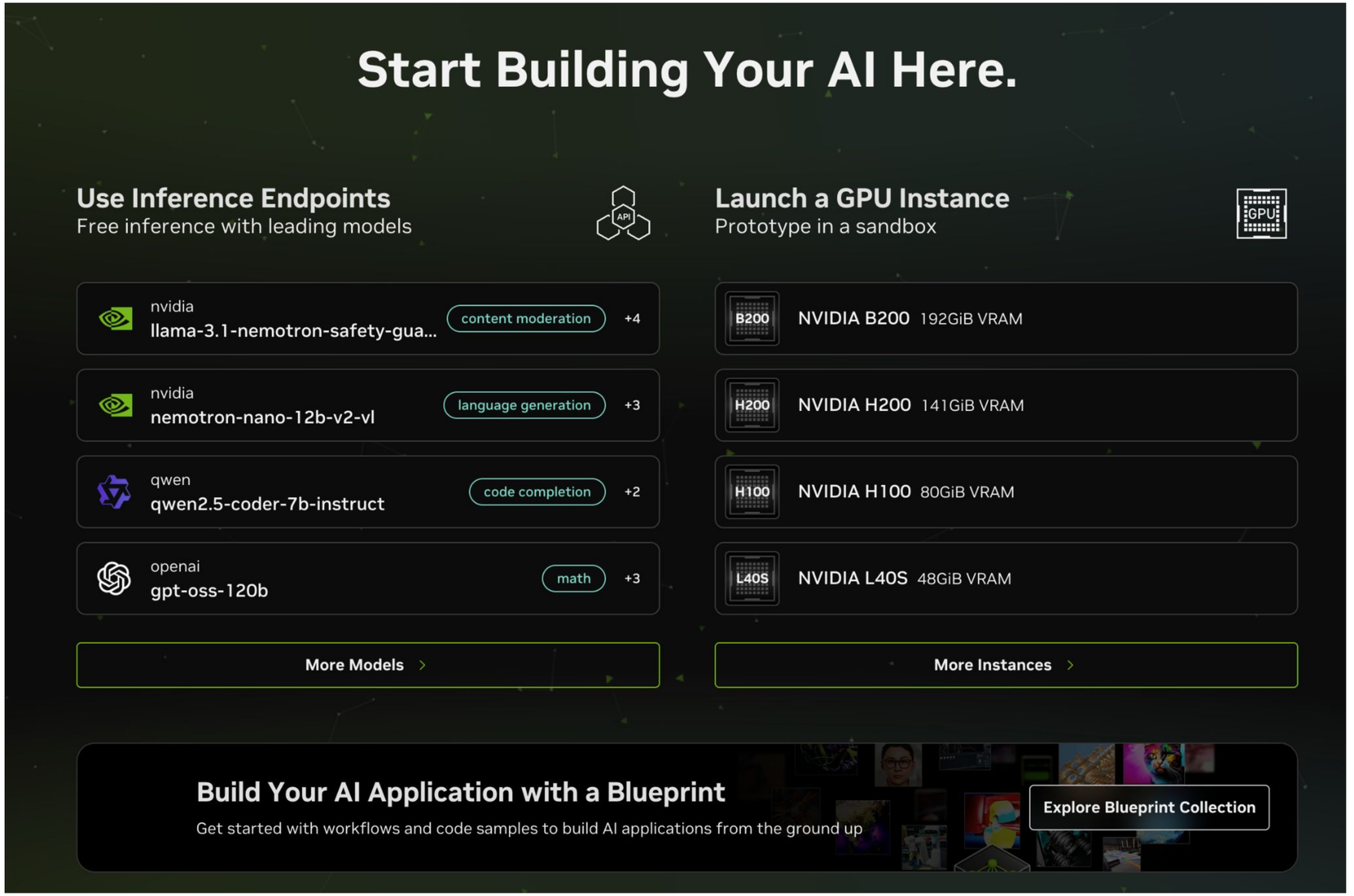


Nemotron RAG

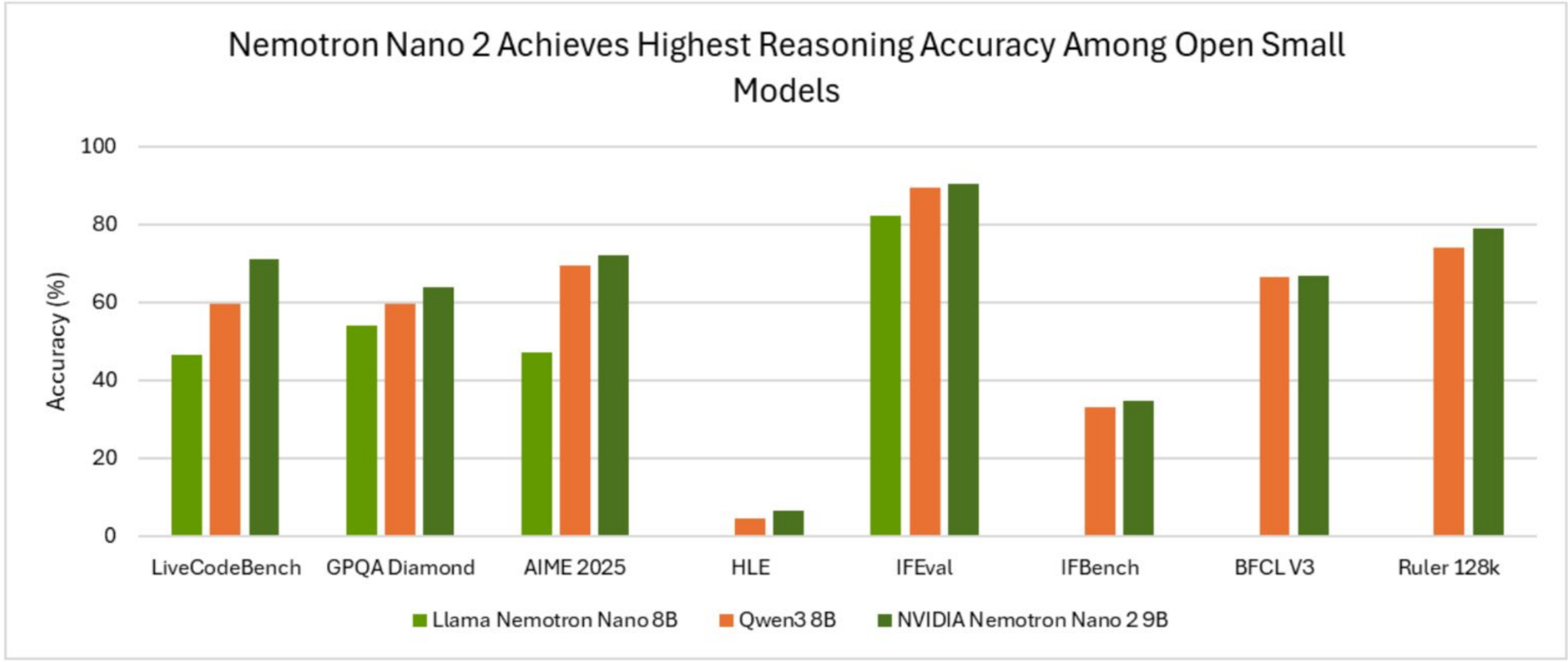
Extraction, Embedding, and Reranking

Try Nemotron on build.nvidia.com

Free inference with leading models; Prototype in a sandbox



Power AI Agents with Advanced Reasoning Using NVIDIA Nemotron Nano 2



Coding
LiveCodeBench

Scientific Reasoning
GPQA Diamond

Complex Math
AIME 2025

Problem Solving
Humanity's Last Exam

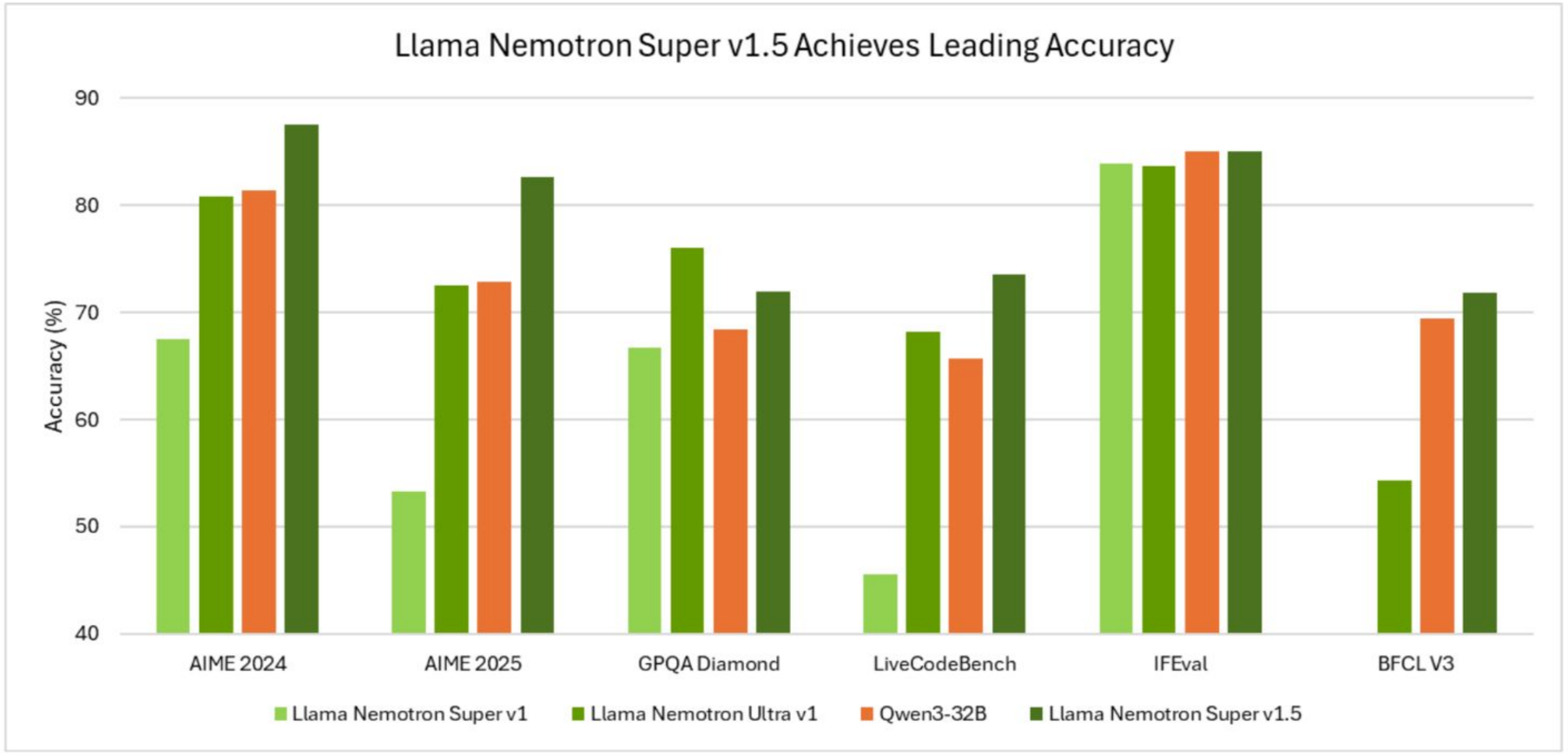
Instruction Following
IFEval, IFBench

Function Calling
BFCL

Long-Context
Ruler

Highest Accuracy for Enterprise Agents Making Complex Reasoning Decisions

Leads in Scientific Reasoning, Coding, and Agentic Tasks



Complex Math
AIME 2024/25

Tool Calling
BFCL

Coding
LiveCodeBench

Instruction Following
IFEval

1x NVIDIA H100. BF16 precision

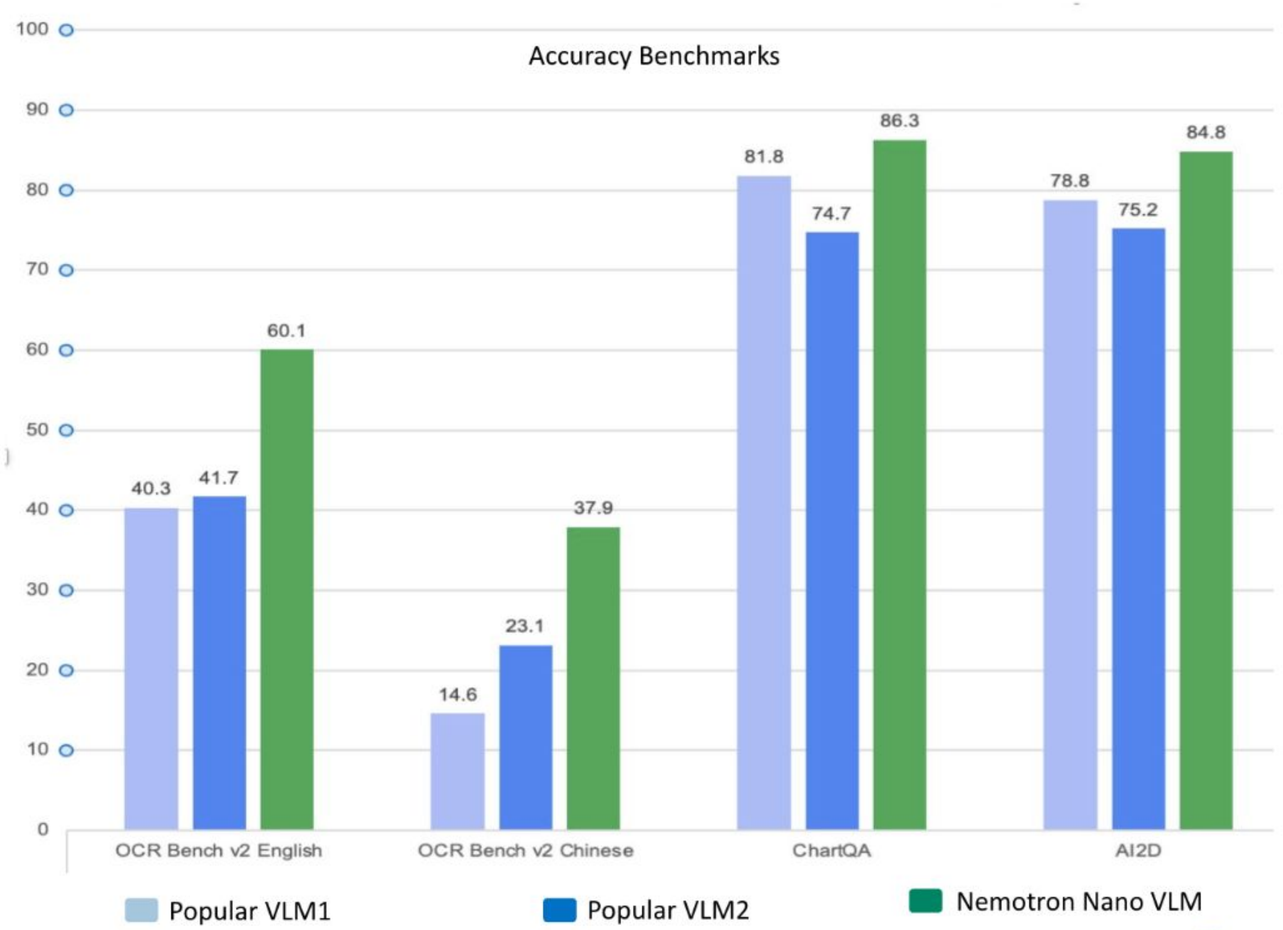
Llama Nemotron Nano VL

8B VLM with Doc Intel Available Today!

Advance Doc AI agent with Nano VLM NIM

- Top of OCRBench V2 leaderboard
- Downloadable VLM NIM.
  Customize VLM using NeMo
- API support for easy use in legacy apps and services
- High performance:
  - TRT-LLM Optimization
  - FP8 quantization
- Create AI agent with RAG using VLM NIM

Best-in-class tables, text, charts understanding



NVIDIA Nemotron Open Datasets

9T Tokens, 30M Samples, 1M GPU hours

Train specialized AI models with high-quality, multimodal synthesized data

Nemotron-PII
Sensitive Data Detection

Nemotron-Safety-Guard
LLM Safety

Nemotron-AIQ-Agentic Safety
Agentic System Safety

Nemotron-Personas
Sovereign AI Development (US, JP, IN)

Nemotron-Pretraining-Code
Coding

Nemotron-Post-Training
Reasoning

NVIDIA Nemotron RAG

Collection of extraction, embedding, reranking models

Extraction

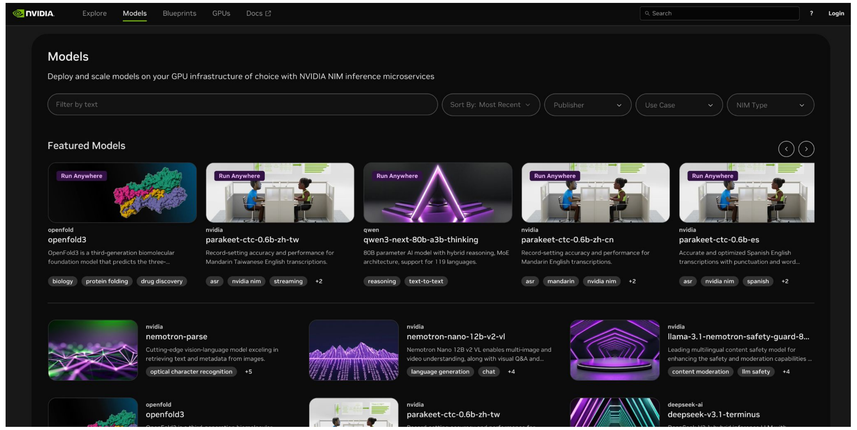
- Nemotron Page Elements: An object detection model for detecting and classifying structured images in documents.
- Nemotron Table Structure: An object detection model for detecting rows and columns and preserving the structure of tables in Markdown format.
- Nemotron Graphic Elements: An object detection model for detecting components of charts such as titles, legends, axes, etc.
- Nemotron OCR: An optical character recognition (OCR) model for extracting text from tables, charts, and infographics.

Embedding

- Nemotron Embedding RAG: multilingual and cross-lingual text question-answering retrieval with long context support and optimized data storage

Reranking

- Nemotron Reranking RAG: fine-tuned reranking model for multilingual, cross-lingual text question-answering retrieval, with long context support



Try at: build.nvidia.com/explore/retrieval

NVIDIA NIM Microservices

Package Open Models for Production
NVIDIA NIM

NVIDIA NIM

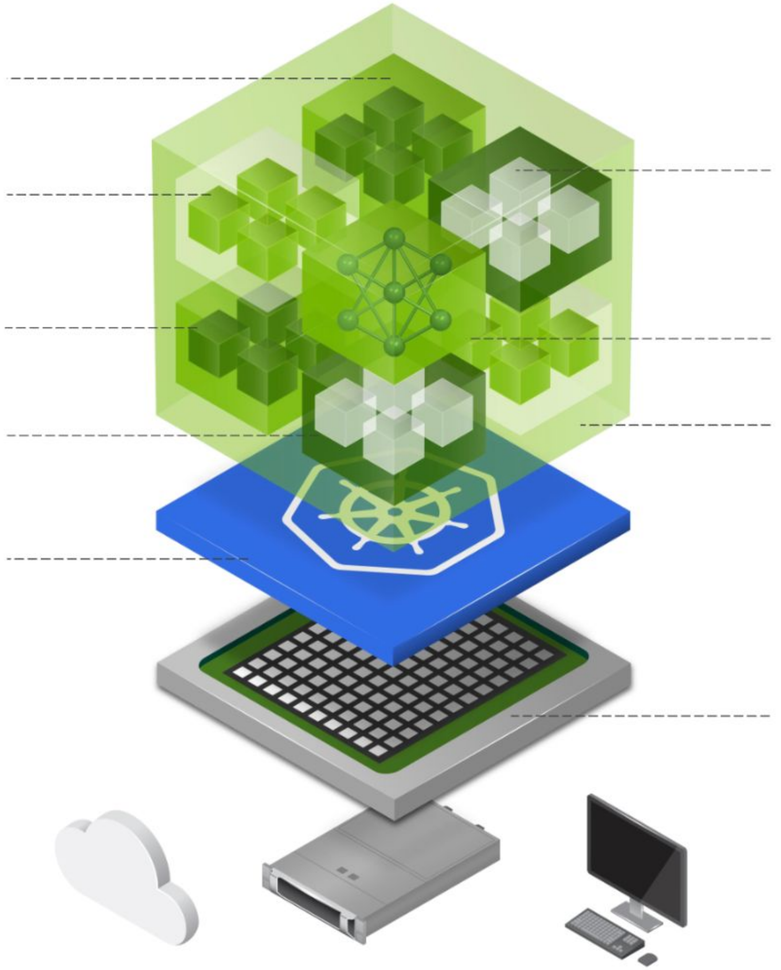
Standard APIs
Text, Speech, Image,
Video, 3D, Biology

NVIDIA Dynamo-Triton
cuDF, CV-CUDA, DALI, NCCL,
Postprocessing Decoder

Cloud-Native Stack
GPU Operator, Network Operator

Enterprise Management
Health Check, Identity, Metrics,
Monitoring, Secrets Management

Kubernetes



NVIDIA TensorRT, TensorRT-LLM, vLLM, SGLang
cuBLAS, cuDNN, In-Flight Batching,
Memory Optimization, FP8 Quantization

Optimized Model
Single GPU, Multi-GPU, Multi-Node

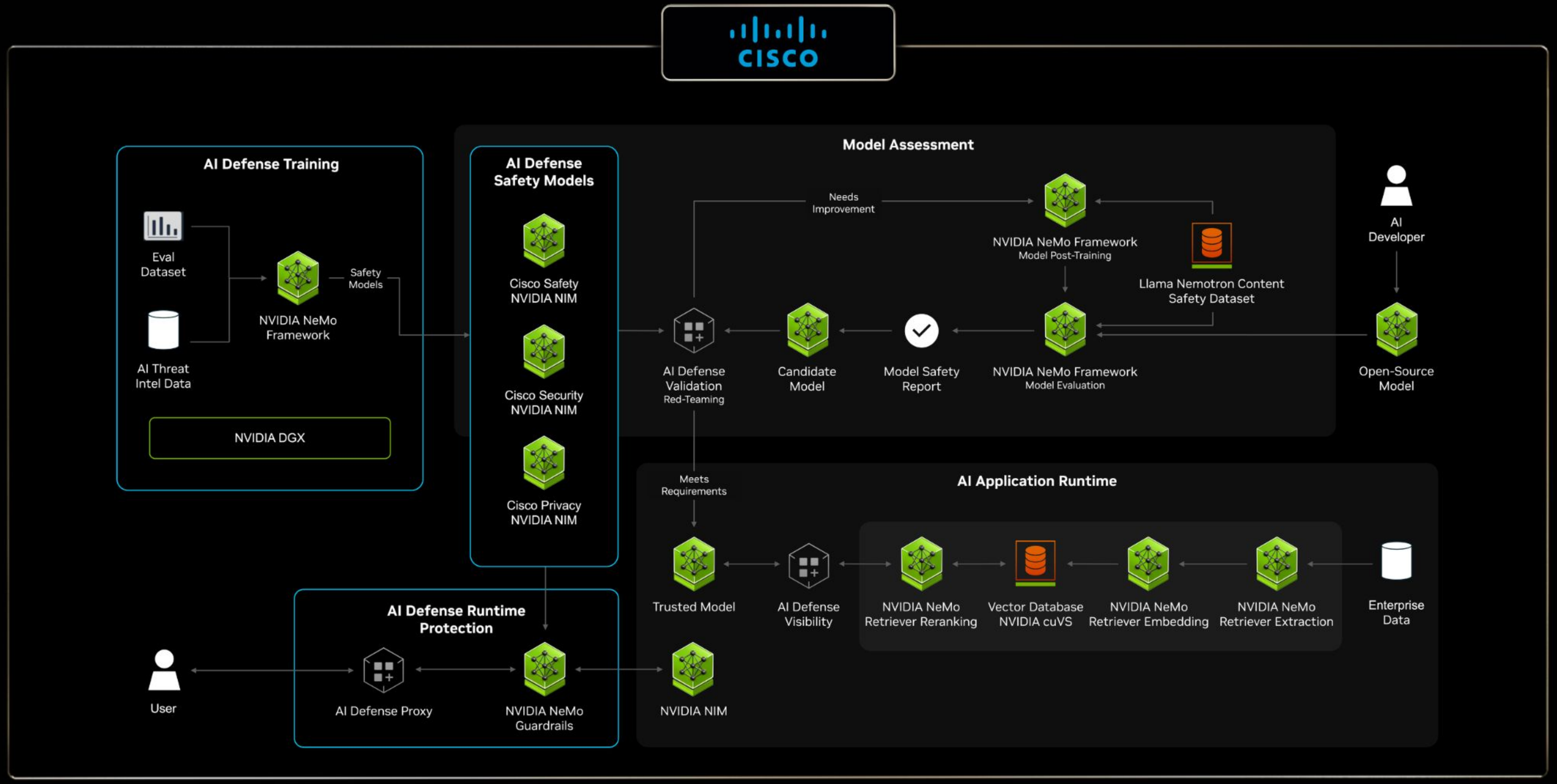
Customization Cache
P-Tuning, LoRA, Model Weights

CUDA

NVIDIA NIM is the Fastest Path to AI Inference

Reduces engineering resources required to deploy optimized, accelerated models

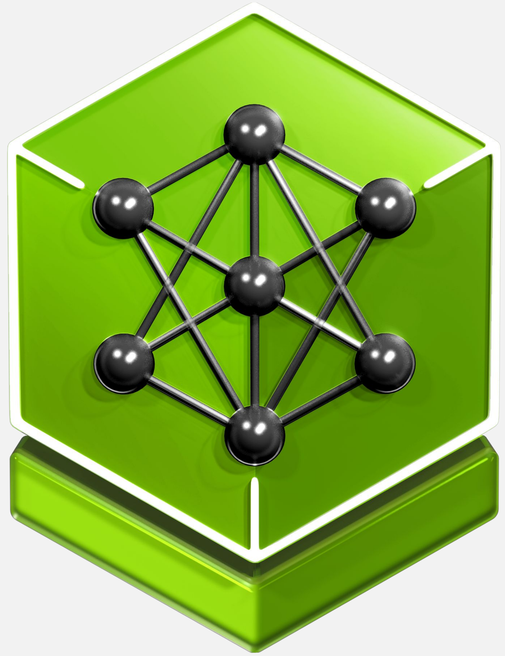
	NVIDIA NIM	Do It Yourself
Deployment Time	5 minutes	1 week +
API Standardization	Industry standard protocol OpenAI for LLMs, Google Translate for Speech	Implement the API layer for each domain and model family according to industry standard specifications
Optimized Engines	Pre-built engines for NVIDIA and community models MISTRAL AI_ Meta starcoder NVIDIA Nemotron	Build your own engine and manually customize for workload and hardware specific requirements
Pre and Post Processing Pipelines	Pre-built with optimized pipeline engines to handle pre/post processing (tokenization)	Implement custom logic
Model Server Deployment	Automated	Manual setup and configuration
Customization	LoRA is supported, more planned	Create custom logic
Container Validation	Extensive workload specific QA support matrix validation	No validation
Enterprise Support	Delivered with NVIDIA AI Enterprise Security and CVE scanning/patching and tech support	Self supported


cisco
AI Defense Training
Eval Dataset
AI Threat Intel Data
NVIDIA NeMo Framework
Safety Models
NVIDIA DGX
AI Defense Safety Models
Cisco Safety NVIDIA NIM
Cisco Security NVIDIA NIM
Cisco Privacy NVIDIA NIM
Model Assessment
Needs Improvement
NVIDIA NeMo Framework
Model Post-Training
Llama Nemotron Content Safety Dataset
AI Developer
AI Defense Validation
Red-Teaming
Candidate Model
Model Safety Report
NVIDIA NeMo Framework
Model Evaluation
Open-Source Model
Meets Requirements
AI Application Runtime
Trusted Model
AI Defense Visibility
NVIDIA NeMo Retriever Reranking
Vector Database NVIDIA cuVS
NVIDIA NeMo Retriever Embedding
NVIDIA NeMo Retriever Extraction
Enterprise Data
AI Defense Runtime Protection
User
AI Defense Proxy
NVIDIA NeMo Guardrails
NVIDIA NIM

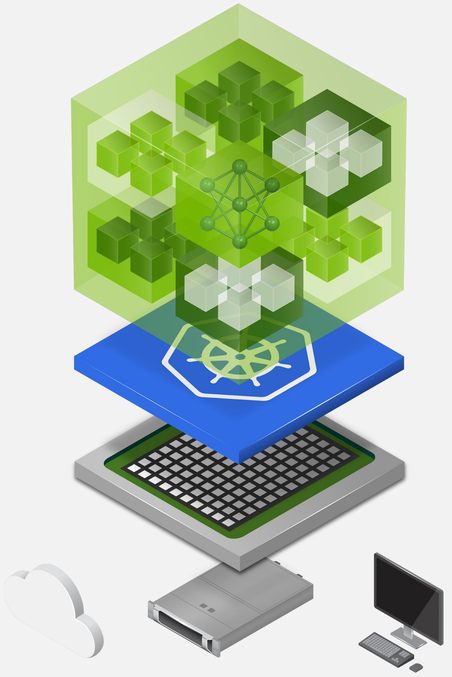
Conclusion

NVIDIA’s Agent Platform

AI Agent Lifecycle Management
NVIDIA NeMo



Accelerate Leading Open Models
NVIDIA Nemotron



Package Open Models for Production
NVIDIA NIM

Get Started Integrating New Nemotron and NeMo Tools Today

Nemotron Nano 2 VL
EA API/weights – available today
API access – Oct 28 | NIM GA – Oct 21

Nemotron Parse
EA container/weights – available today
API access – Oct 28 | NIM GA – Oct 17

Nemotron RAG
EA container/weights – available today
API access – available | NIM GA – Oct 21

Llama Nemotron Safety Guard
EA container/weights – available today
API access – Oct 28 | NIM GA – Oct 17

Nemotron Nano 3
Coming soon

Nemotron-Personas Datasets
EA – USA, Japan-versions available today
GA – India-version Oct 13

Nemotron-PII Dataset
GA – Oct 28

NeMo Data Designer
EA – Available today
GA– Oct 27

NeMo RL
Available today on GitHub

NeMo Agent Toolkit
Available today on GitHub

NeMo Evaluator OSS
Available today on GitHub

NeMo Curator
GA: Audio & Video Processing
Available today on GitHub