NVIDIA Generative AI solutions

Meriem Bendris, Senior Deep Learning Data Scientist

About Me

Meriem Bendris

- Senior Deep Learning Data Scientist at NVIDIA
- Focus on Conversational AI, Natural Language Processing, Large-scale Training
- PhD in Signal and Image Processing – Telecom Paris and Orange Labs

Pioneering Accelerated Computing

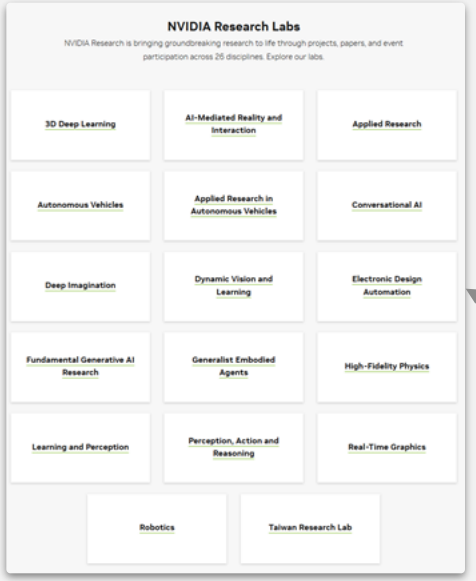
Accelerated computing requires full-stack optimization, from chip architecture, systems, and acceleration libraries, to refactoring the applications. The global NVIDIA ecosystem spans 4 million developers, 40,000 companies, and over 3,000 applications.

NVIDIA AI Technology Center Program

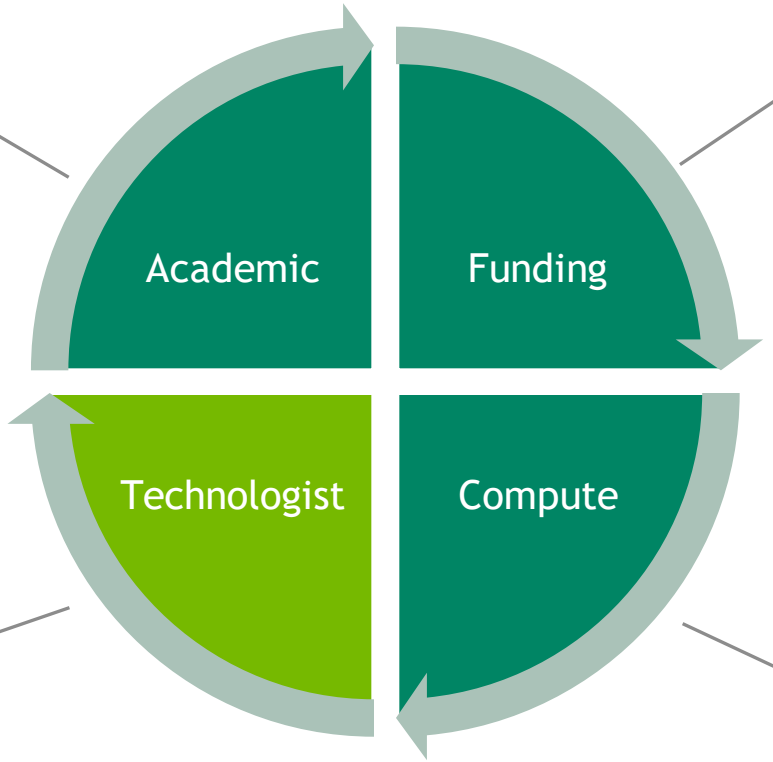
EMEA

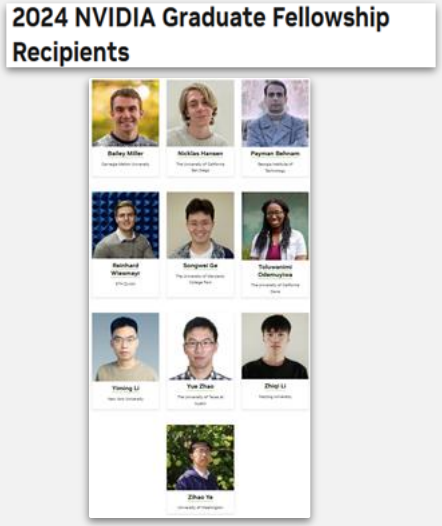
AI Research Projects

Enabling AI research through efficient GPU Computing



NVIDIA Research Labs | Research





NVIDIA Graduate Fellowship Program

NVIDIA AI TECHNOLOGY CENTER (NVAITC) Program

Academic Grant Program for Researchers

NVIDIA AI TECHNOLOGY CENTER (NVAITC)

Catalyse AI transformation through research-centric integrated engagements

NVAITC EMEA

Operating since 2020

74 collaborators from 48 institutions

Out of 426 PI approached at 150+ institutions

81 peer-reviewed publications

Out of 149 past submissions + 8 new papers in the works

4650+ academics trained live

From technology lectures to scientific workshops

NVAITC Scientific Workshops

Last Accepted at European Conference on Computer Vision (ECCV) 2024 Milano

- **International Workshop on Computational Aspects of Deep Learning** (CADL)
  - previously at ICPR'20, ECCV'22
  - 3rd Workshop at BMVC 23
  - 4th Workshop at ECCV 2024
  - 5th Workshop at ISC 2025

- **International Workshop on Uncertainty Quantification for Computer Vision** (UNCV)
  - 1st Workshop at ECCV 2022
  - 2nd Workshop at ICCV 2023
  - 3rd Workshop at ECCV 2024
  - 4th Workshop at CVPR 2025

NVAITC

Playbooks | Webinars

README.md

NVIDIA AI Technology Center

The goal of the NVIDIA AI Technology Center (NVAITC) is to enable and accelerate AI research, education and adoption, using supercomputing resources based on NVIDIA technology, as well as to foster academic collaborations across the NVAITC network. A central objective of NVAITC collaborations is to provide support to specific research projects in AI and Applied AI, governed by an approved statement of work. The goal of projects is to foster technological development based on research, and to publish and otherwise disseminate results of the project's work. NVAITC enables researchers to benefit from NVIDIA's expertise in utilizing GPU and A I Computing. NVAITC projects enable academics at all levels to do their research more efficiently.

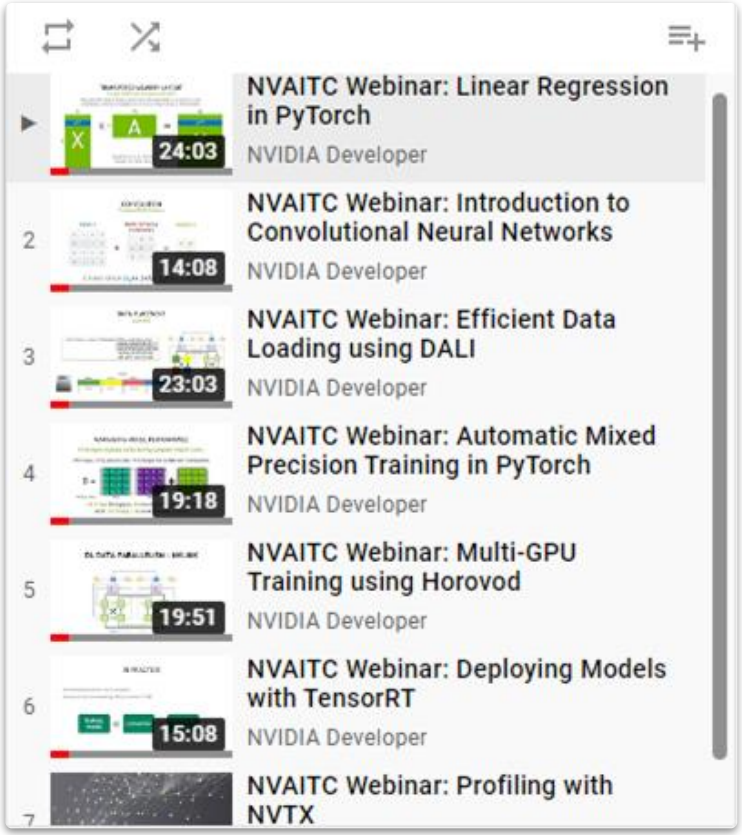
Playbooks

As researchers invest a lot of effort to unlock the full potential of AI, the need for scalable and efficient tools has become funamental, particularly when dealing with large models, diverse data formats, and high-performance systems. As NVAITC, we are working to facilitate this process by developing comprehensive tools, materials, and recipes that simplify AI adoption at scale.

The **NVAITC Playbooks** provide NVIDIA-based scalable reference implementations for common AI use-cases in research.

- multi-scale-agentic-rag-playbook: A playbook showcasing how to create a RAG pipeline working at different scales.
- synthetic-data-generation-and-sft-playbook: A playbook showcasing a scalable pipeline to finetune an LLM on synthetically enriched data through using the NeMo Framework.

- **Multi-scale RAG pipeline**
- **Synthetic Data Generation**
- Scalable Data Curation and Model Finetuning using NeMo
- Collaborative data aggregation and 3D visualization of digital twins
- Online Training of deep learning for HPC Apps
- NSIGHT Systems profiling on GH
- Hyper-parameter Optimization for NeMo
- Fortran to Python Code Modernization
- Accelerated Video Processing and Model Training



NVIDIA AI Technology Center

Developer Program

Resources to accelerate building

Learn more: developer.nvidia.com/join

Learn

News
Industry and technical

Training
Hands-on self-paced courses and instructor-led workshops

Certification
Industry-recognized credentials

Learning
Tutorials, guides, blogs, research, docs, code samples, reference apps

Best Practices
Setup, optimization, reference architecture

Ecosystem
GTC, NVIDIA Partner Network, Accelerated App Catalog

Build

Software
100s of APIs, models, SDKs, microservices, early access to NVIDIA tech

Cloud APIs
Evaluation access to NIM microservices and optimized APIs

Sample Apps
GPU accelerated software: notebooks, sample apps, frameworks

Connect

Community
Technical forums, Discord, user groups, Slack

GTC
Networking sessions, Connect With Experts sessions

Events
Meetups, hackathons, bootcamps

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Salamandra Technical Report

Language Technologies Unit

Barcelona Supercomputing Center

Abstract

This work introduces Salamandra, a suite of open-source decoder-only large language models available in three different sizes: 2, 7, and 40 billion parameters. The models were trained from scratch on highly multilingual data that comprises text in 35 European languages and code. Our carefully curated corpus is made exclusively from open-access data compiled from a wide variety of sources. Along with the base models, supplementary checkpoints that were fine-tuned on public-domain instruction data are also released for chat applications. Additionally, we also share our preliminary experiments on multimodality, which serve as proof-of-concept to showcase potential applications for the Salamandra family. Our extensive evaluations on multilingual benchmarks reveal that Salamandra has strong capabilities, achieving competitive performance when compared to similarly sized open-source models. We provide comprehensive evaluation results both on standard downstream tasks as well as key aspects related to bias and safety. With this technical report, we intend to promote open science by sharing all the details behind our design choices, data curation strategy and evaluation methodology. In addition to that, we deviate from the usual practice by making our training and evaluation scripts publicly accessib
license in order to foste
contributing to the open-

Models
Code

B Acknowledgements

This project has benefited from the contributions of numerous teams and institutions, mainly through data contributions, knowledge transfer or technical support.

We are grateful to our ILENIA project partners: CENID, HiTZ and CiTIUS for their collaboration. We also extend our genuine gratitude to the Spanish Senate and Congress, Fundación Dialnet, and the University of Las Palmas de Gran Canaria. Many other institutions have been involved in the project. Our thanks to Òmnium Cultural, Parlament de Catalunya, Institut d'Estudis Aranesos, Racó Català, Vilaweb, ACN, Nació Digital, El món and Aquí Berguedà. We thank the Welsh government, DFKI, Occiglot project, especially Malte Ostendorff, and The Common Crawl Foundation, especially Pedro Ortiz, for their collaboration.

We would also like to give special thanks to the NVIDIA team, with whom we have met regularly, specially to: Ignacio Sarasua, Adam Henryk Grzywaczewski, Oleg Sudakov, Sergio Perez, Miguel Martinez, Felipes Soares and Meriem Bendris. Their constant support has been especially appreciated throughout the entire process.

We truly appreciate the support provided by BSC's operations team, specially to its leader David Vicente for his patience and help in HPC-related issues. Their valuable efforts have been instrumental in the development of this work.

Finally, we are deeply grateful to the Spanish and Catalan governments for their financial support, which has made this entire endeavor possible. This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project Modelos del Lenguaje and the ILENIA Project with reference 2022/TL22/00215337, and by the Government of Catalonia through the Aina Project.

[33] * equal contributions.

Salamandra Technical Report

NVAITC EMEA

NVAITC
Project Submission Template

- **Contact:** nvaitc-project-emea@nvidia.com

Project Header

Date	*Date of submission of this form*
Title	*Title for this research project*
Principal Investigator(s)	*Name, affiliation, contact information*
Contributors	*List all researchers (faculty, grad students, PhD, etc) involved in the project with their affiliation*
References	*List 5 peer-reviewed publications in past 3 years by the PI and main contributors above*
NVIDIA Mutual NDA in Place	*Yes|No* *Does the PI's institution have a Mutual NDA in place with NVIDIA? Refer to NVIDIA standard model.*
Target Venue for Publication	*Research collaboration must lead to a scientific publication where NVAITC is acknowledged according to its contribution (eg co-author or acknowledgment section)*
NVIDIA Technology	*List the NVIDIA hardware and software technologies targeted to be used throughout the project*

Description of Research Project

Give a concise description of the state-of-the-art and of the project objectives, workplan, risks, timeline and committed resources (all kind). Provide links/access to source code, datasets, draft manuscripts, forums, etc for further analysis.

Discussion of Computing

Discuss in more details the computing dimension in the project.

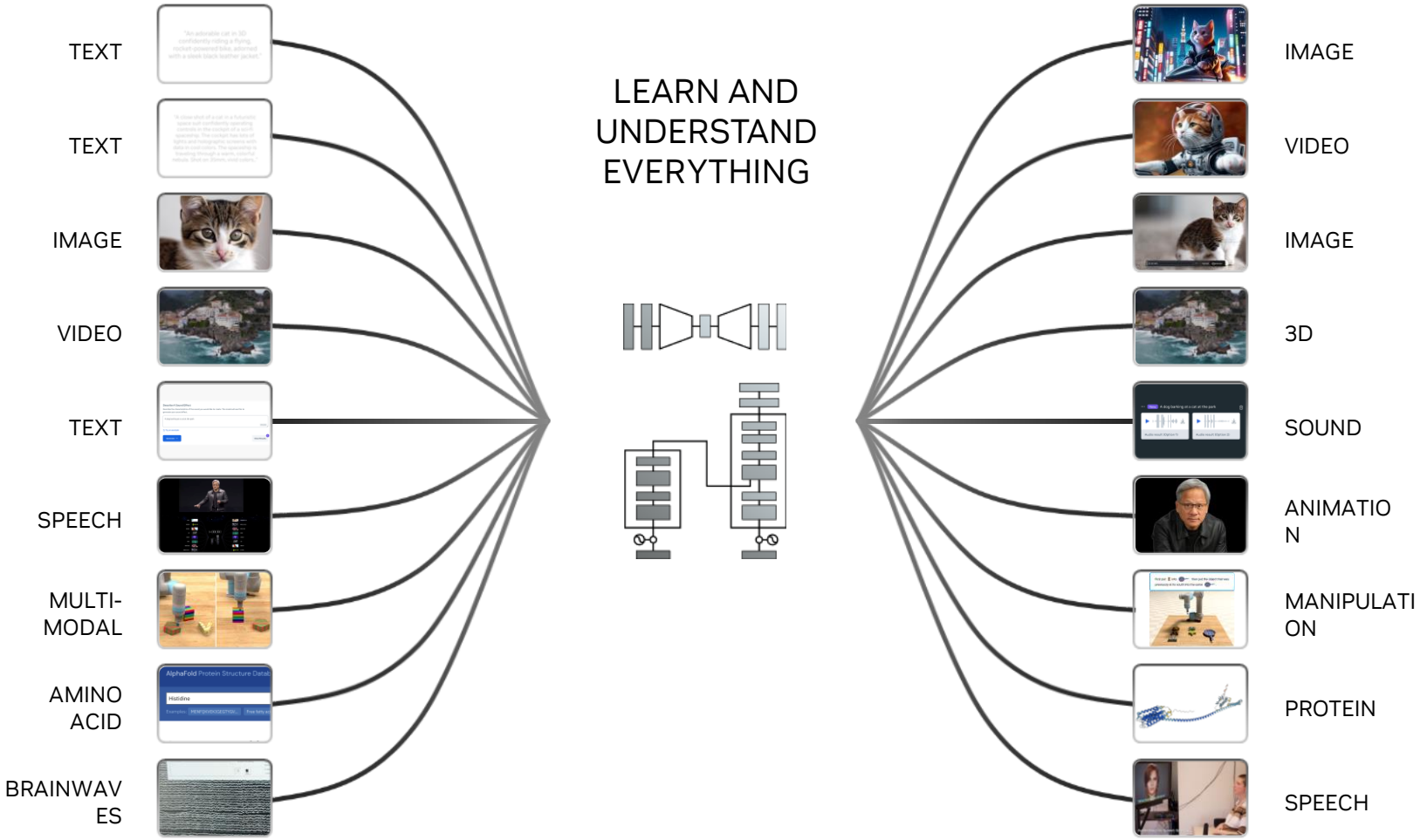
What scale of computing are you targeting for this project?
How many GPU can your software/model leverage in parallel?
Describe the volume and type of data/datasets.
What HPC system provider (eg CINECA, CSC, University) do you have in mind for this project?
Etc.

Discussion of NVAITC Participation

Discuss in more details the help requested. Where in the project and how much time of a NVAITC engineer do you need? What specific skills does he/she need to bring? Etc.

Frederic Pariente

Generative AI



Generative AI For Language Processing

Before Generative AI

Sentiment Analysis

Document Classification

...

Summarization

Translation

...

With Generative AI

How has NVIDIA contributed to acceleration of AI?

LLM

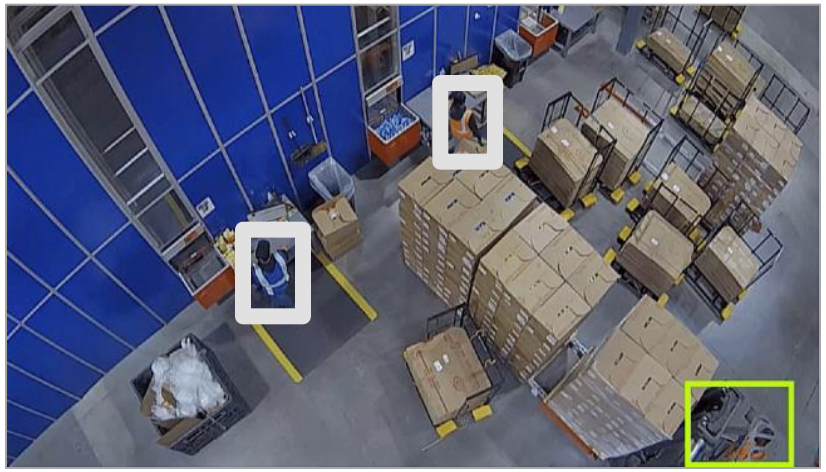
NVIDIA has been a pioneer in the field of AI since the very beginning. Our GPU platform has enabled the rapid development of AI – from the training of neural networks, to inference in the data center, on-device AI in the car and in the cloud, and the deployment of AI to tackle challenging problems like conversational AI and translation.

NVIDIA's GPU-accelerated computing platform is the engine of AI – it is the most important computing platform of our time.

***Generated using NVIDIA NeMo service*

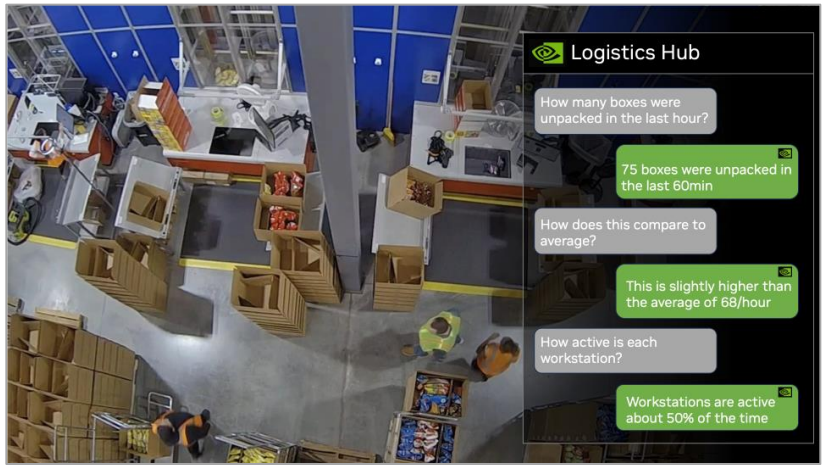
Generative AI For Computer Vision

Legacy CNNs



Specialized, Rigid / Rule-based
Requires tons of labeled data
Slow Development Cycle

Generative AI



Zero-shot Learning, Generalizable
Faster Development Cycle
Natural Language Prompts

Visual Language Models

Open Vocabulary Object Detection | Optical Character Detection and Recognition

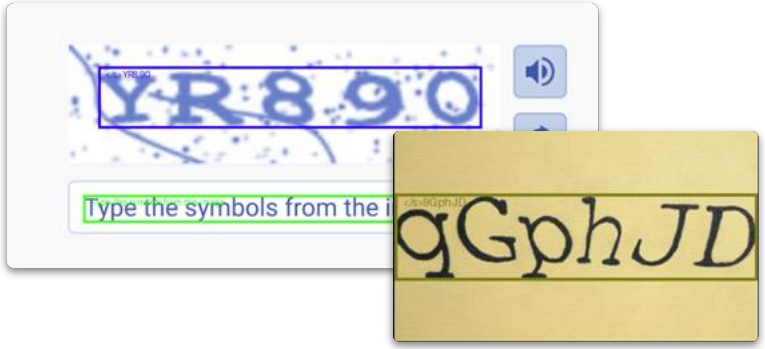
Prompt: person, robotic arm

nv-grounding-dino

Ocdrnet

Prompt: Where is the coach?

microsoft-kosmos-2

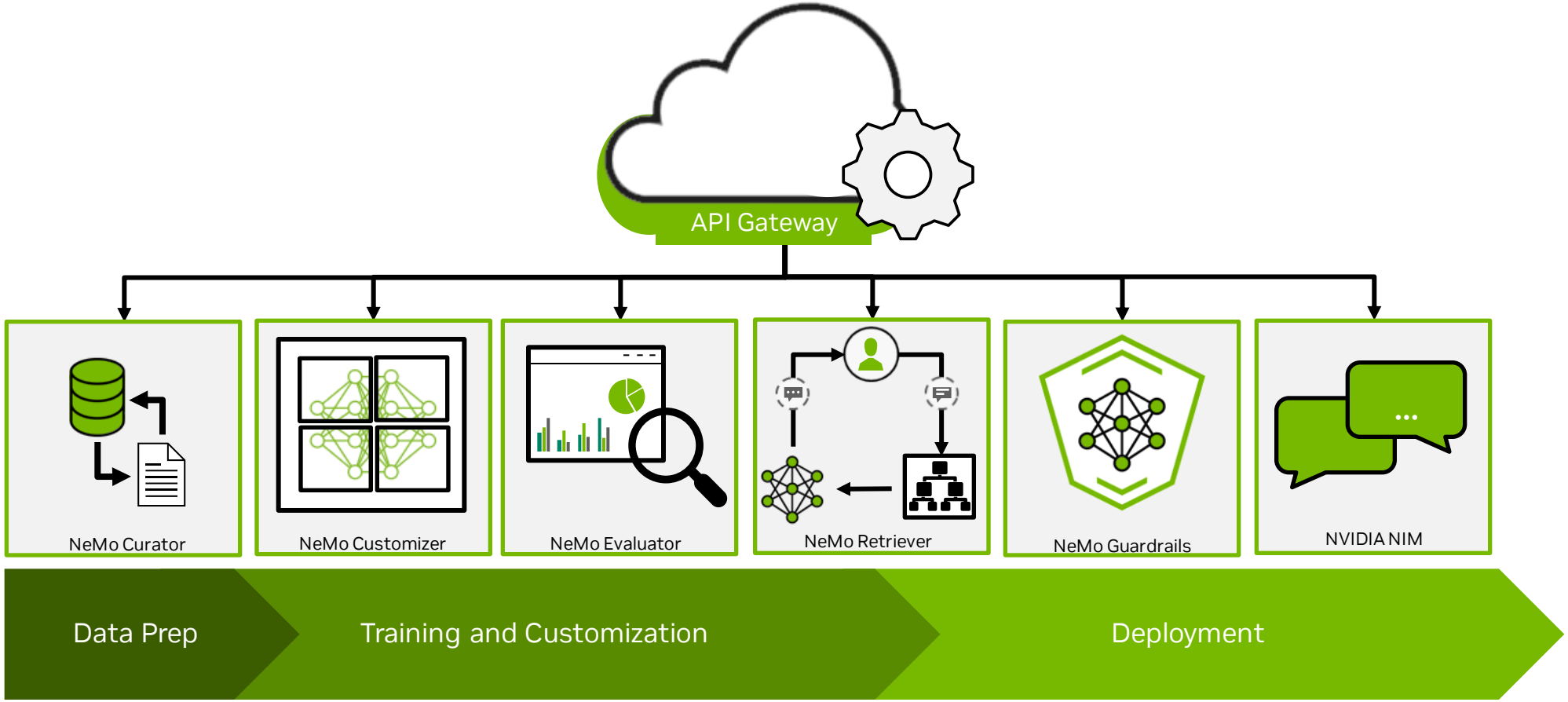


microsoft-florence-2

NVIDIA Nemo

Building Generative AI Applications for the Enterprise

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



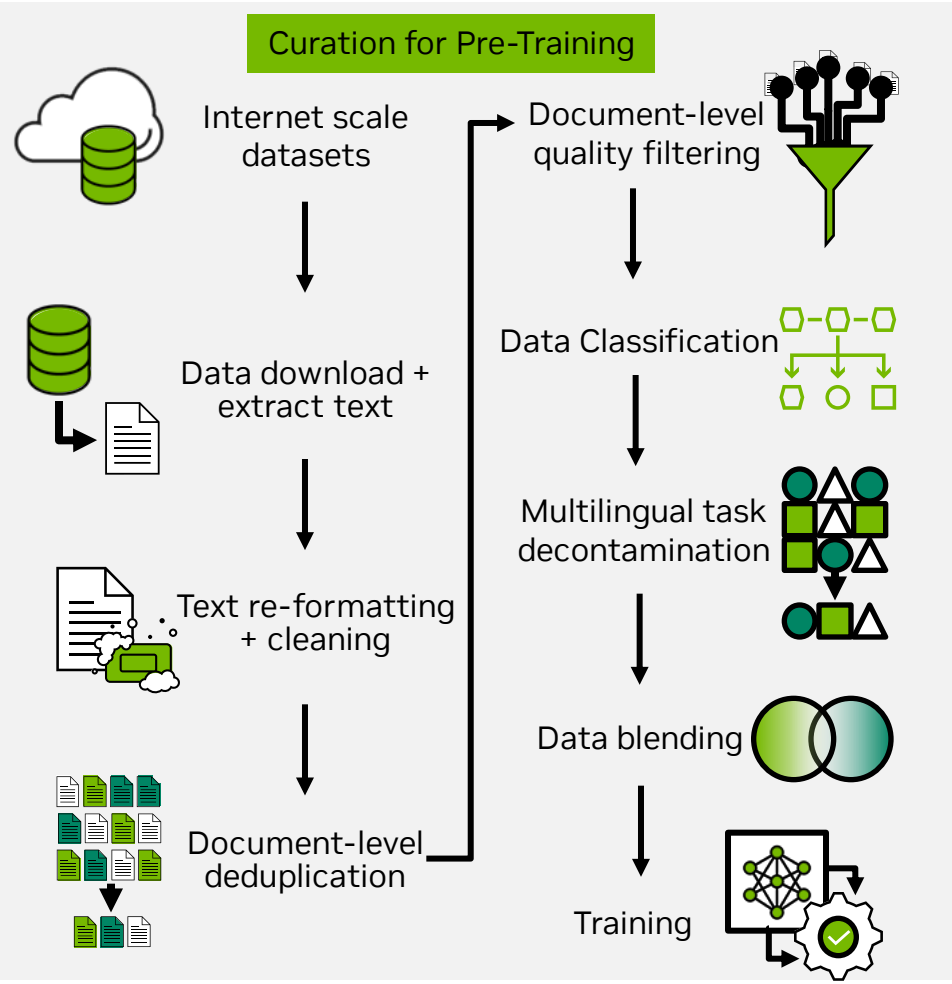
NeMo Curator

Enabling Large-scale high-quality datasets for Pre-training

- GPU-accelerated data curation at scale
- Best practices for data preparation

NeMo Curator steps:

1. Data download and text extraction
   - Download from Common Crawl, Wikidumps, and ArXiv
   - Flexibility for users to customize and extend to other datasets
2. **Text re-formatting and cleaning** - Bad Unicode, newline, repetition
3. GPU-accelerated Document Level **Deduplication**
   - Fuzzy deduplication
   - Exact deduplication
   - Semantic deduplication
4. Document-level **Filtering**
   - Classifier filtering
   - Quality filtering
   - Heuristic-based filtering
5. **Data Classification:** PII removal/redaction filter, domain classifier, toxicity classifier, task classifier, complexity classifier
6. Downstream-task decontamination



NeMo Curator

Increased Accuracy With a Variety of GPU-accelerated Features

Synthetic Data Generation

- **Pre-built pipelines** - for tasks like prompt generation, dialogue generation, and entity classification
- **Modular** - Easily integrate NeMo Curator's features into your existing pipelines
- **OpenAI API compatible** - Integrate custom Instruct and Reward models

Deduplication & Classification

- **Lexical Deduplication** – Identical (Exact) or near identical (Fuzzy)
- **Semantic Deduplication** – focuses on the meaning rather than the exact text
- **Classifier Models** - State-of-the-art open models to either enrich or filter your data.

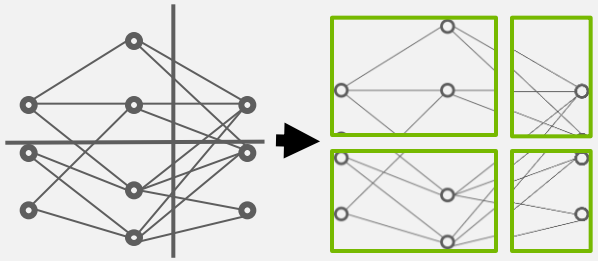
GPU Acceleration with RAPIDS

- **cuDF** - for deduplication & classifer models
- **cuML** - for K-means clustering in semantic deduplication
- **cuGraph** – for fuzzy deduplication

Building Generative AI Foundation Models

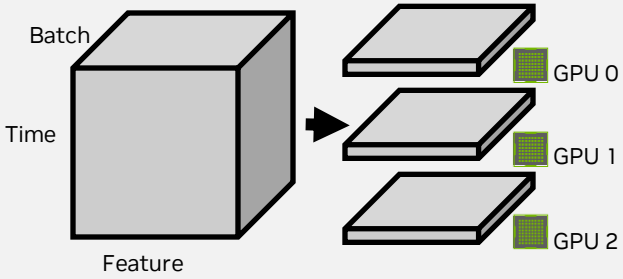
Efficiently and quickly training models using NVIDIA NeMo

Tensor & Pipeline Parallelism



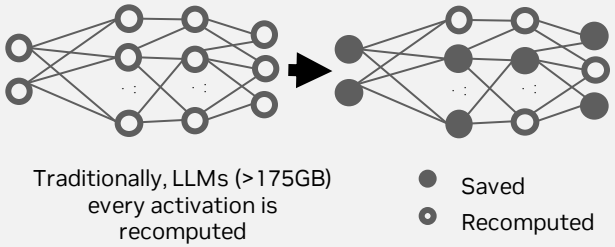
Reduced memory footprint and allows for large-scale training of LLMs across accelerated infrastructure

Sequence Parallelism



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

Selective Activation Recomputation

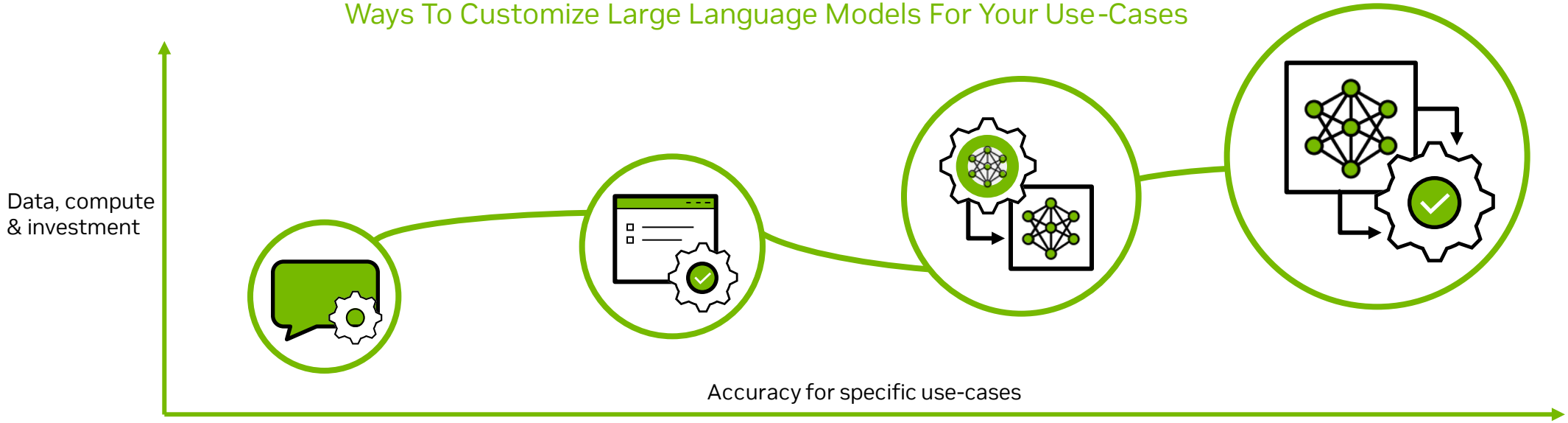


Smart activation checkpointing provides greatest trade-off between memory and recomputation

https://github.com/NVIDIA/NeMo

Suite of Model Customization Tools in NeMo

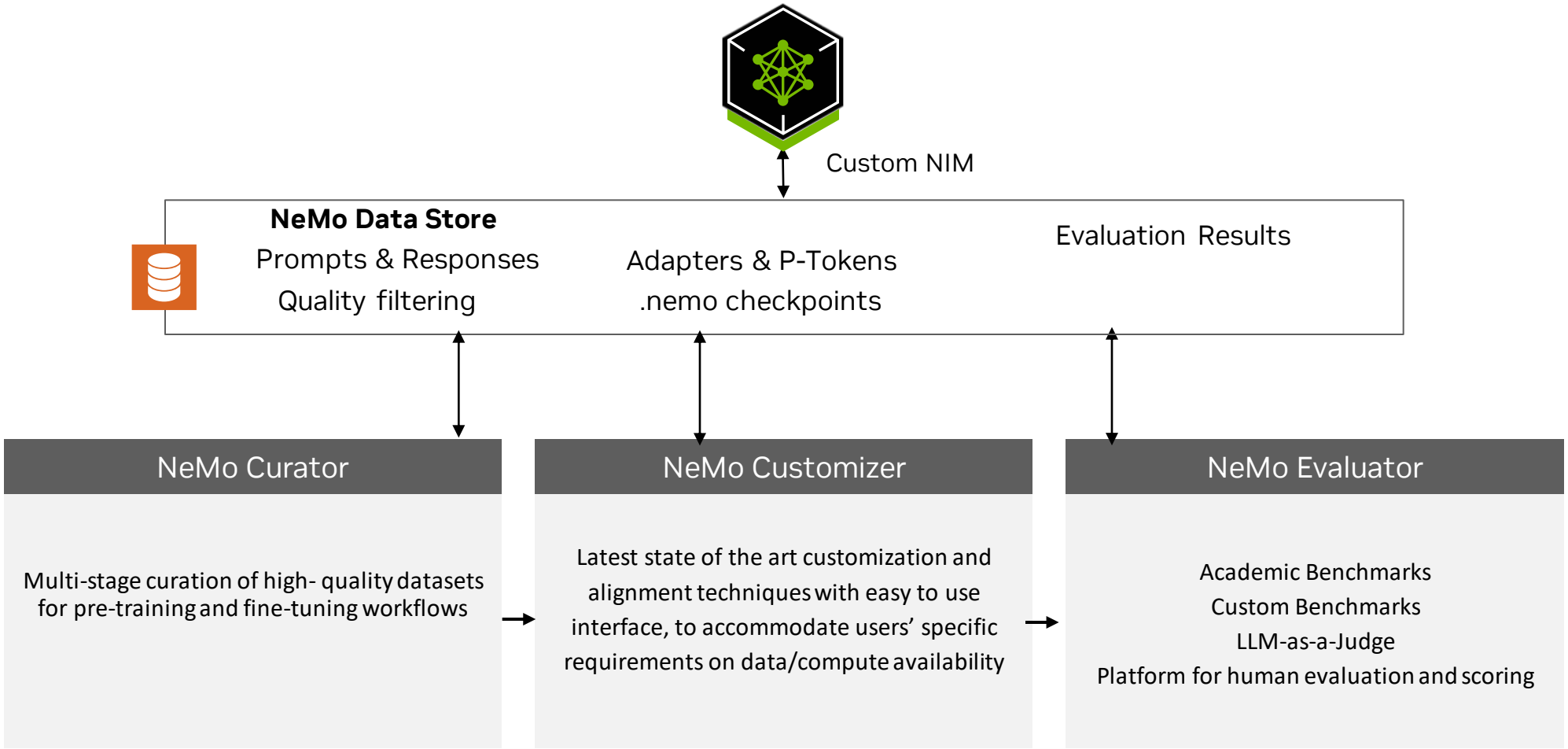
Ways To Customize Large Language Models For Your Use-Cases



	PROMPT ENGINEERING	PROMPT LEARNING	PARAMETER EFFICIENT FINE-TUNING	FINE TUNING
Techniques	• Few-shot learning • Chain-of-thought reasoning • System prompting	• Prompt tuning • P-tuning	• Adapters • LoRA • IA3	• SFT • RLHF
Benefits	• Good results leveraging pre-trained LLMs • Lowest investment • Least expertise	• Better results leveraging pre-trained LLMs • Lower investment • Will not forget old skills	• Best results leveraging pre-trained LLMs • Will not forget old skills	• Best results leveraging pre-trained LLMs • Change all model parameters
Challenges	• Cannot add as many skills or domain specific data to pre-trained LLM	• Less comprehensive ability to change all model parameters	• Medium investment • Takes longer to train • More expertise needed	• May forget old skills • Large investment • Most expertise needed

https://github.com/NVIDIA/NeMo-Aligner

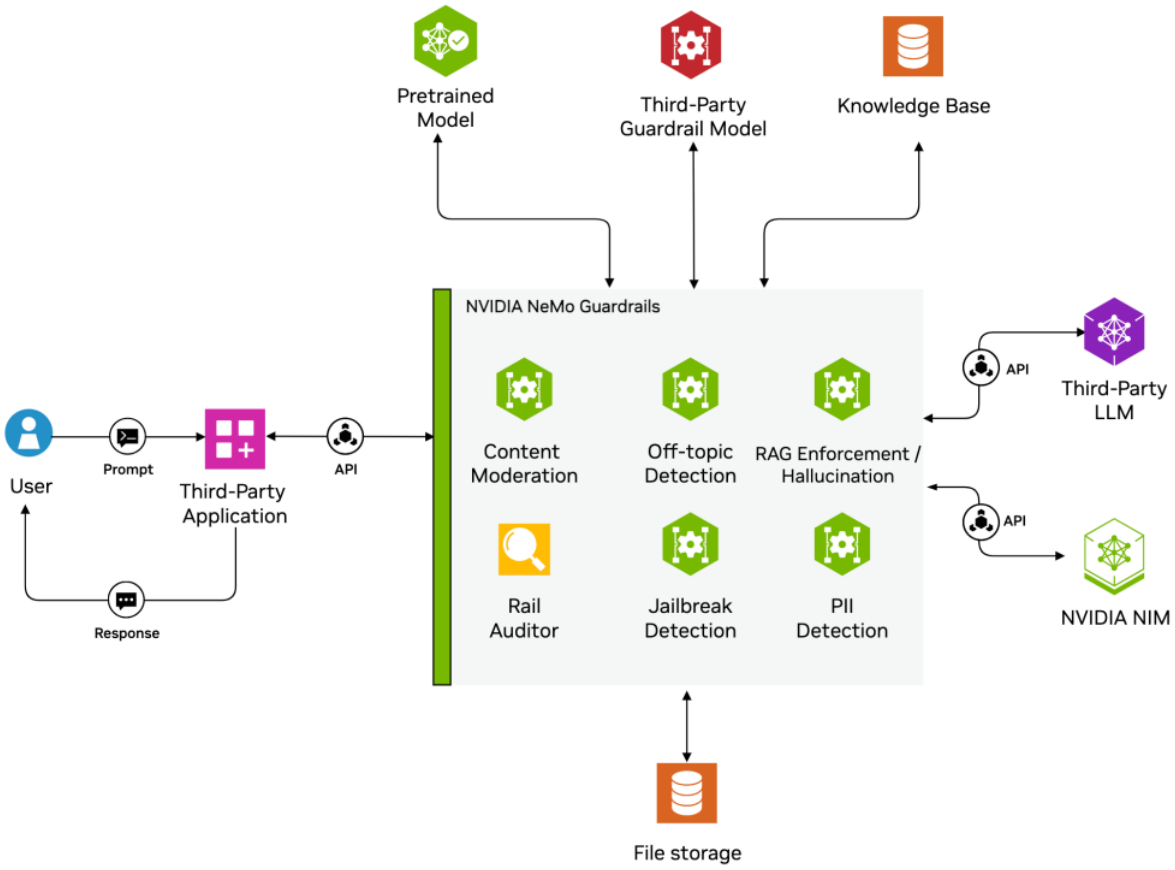
AI Model Evaluation with NeMo Evaluator



NVIDIA NeMo Guardrails

Scalable rail orchestration for safeguarding enterprise generative AI

- Efficiently orchestrate multiple rails across applications with a modular framework
- Use smart defaults or customize and extend rails leveraging a robust 3rd party ecosystem
- Continuously improve rail and application effectiveness with built-in auditing and analytics
- Leverage open-source and portable, enterprise grade microservices ecosystem



NVIDIA NIM

NVIDIA NIM

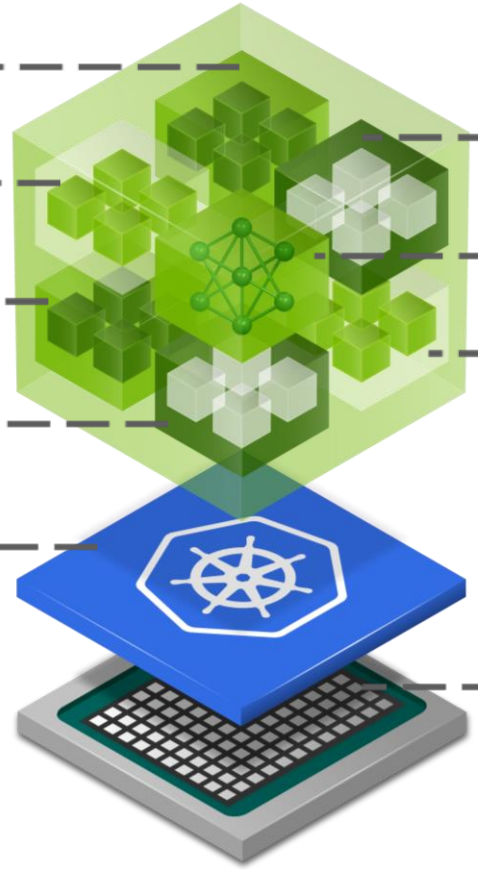
Industry Standard APIs
Text, Speech, Image, Video, 3D, Biology

Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL, Post Processing Decoder

Cloud Native Stack
GPU Operator, Network Operator

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

Kubernetes



TensorRT and TensorRT-LLM
cuBLAS , cuDNN, In-Flight Batching, Memory Optimization, FP8 Quantization

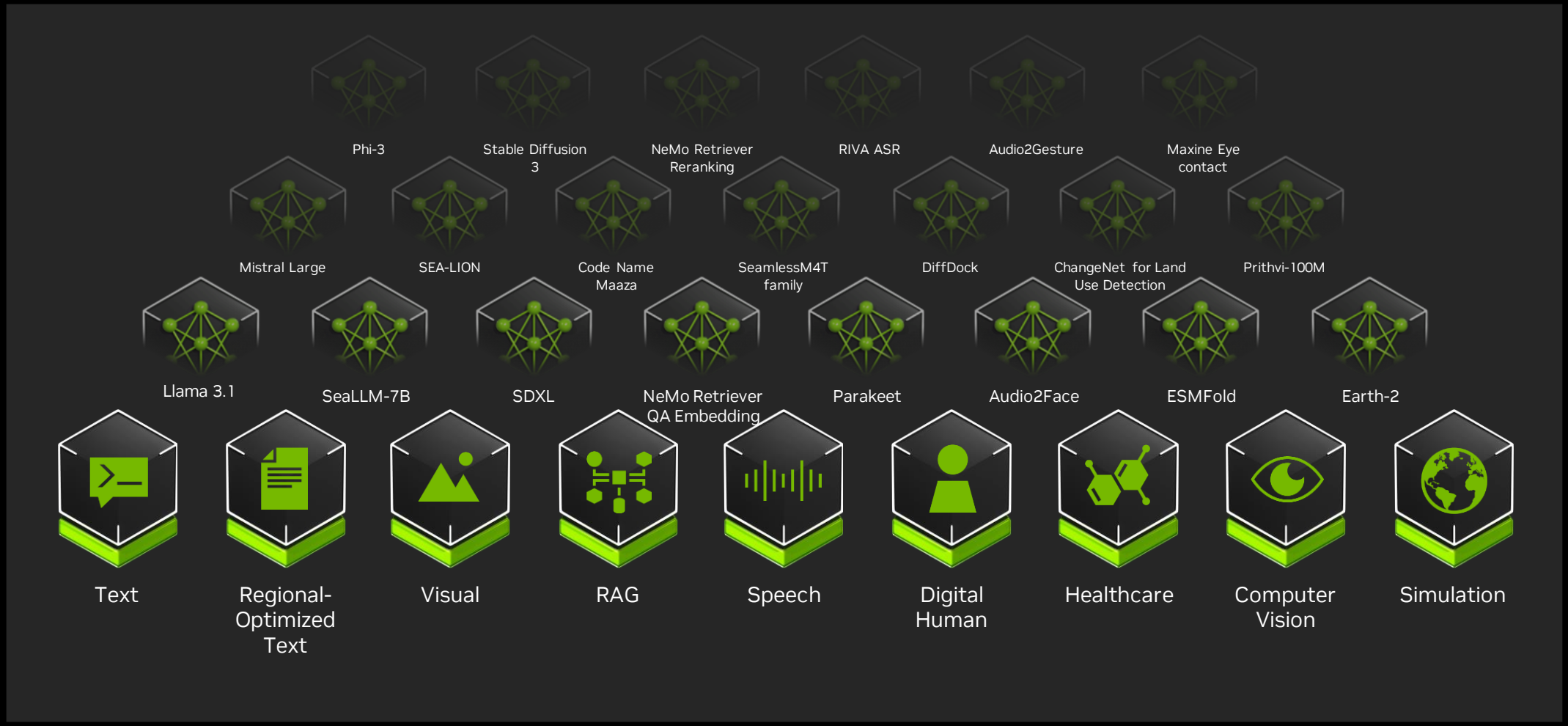
Optimized Model
Single GPU, Multi-GPU, Multi-Node

Customization Cache
P-Tuning, LORA, Model Weights

NVIDIA CUDA

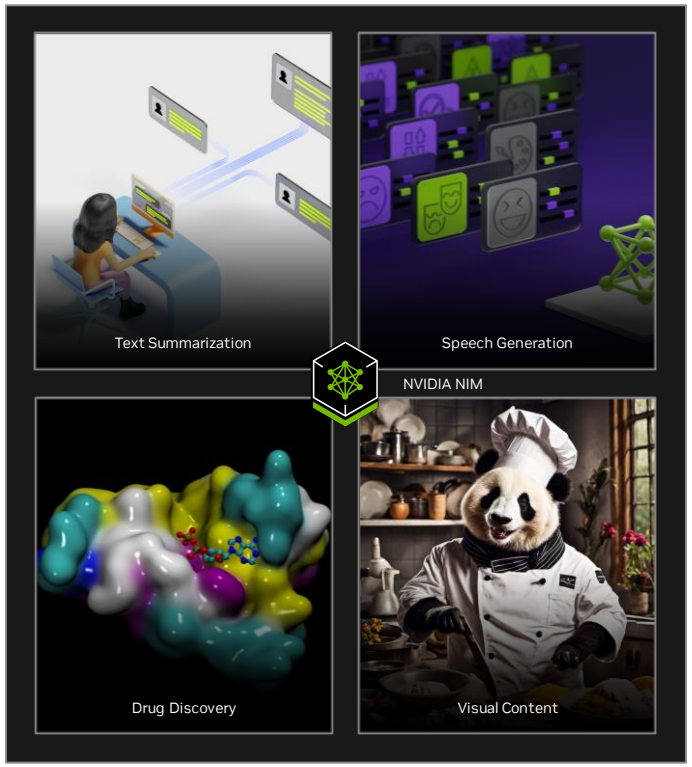
100's of Millions of CUDA GPUs Installed Base

NVIDIA NIM For Every Domain

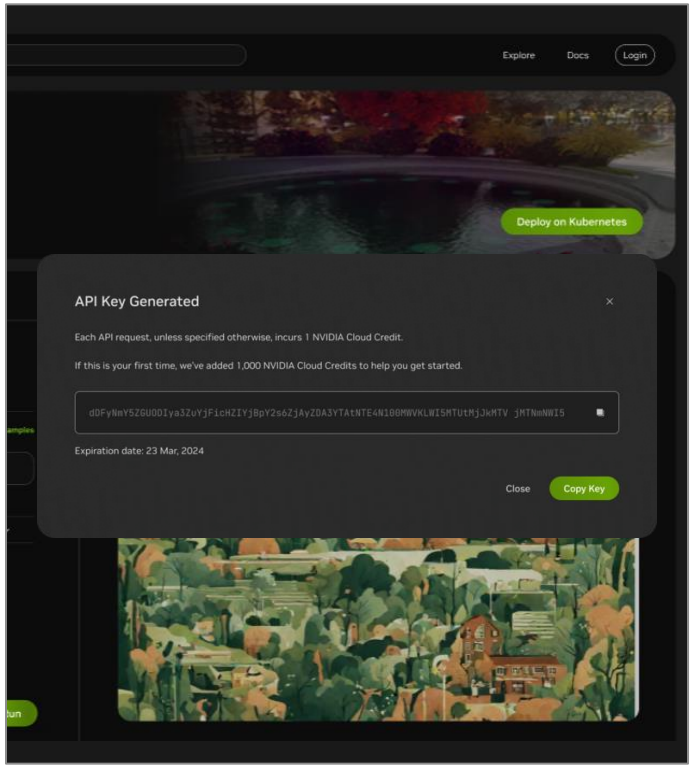


Experience and Run Enterprise Generative AI Models Anywhere

Seamlessly integrate AI in business applications with NVIDIA AI APIs



Experience Models



Prototype with APIs

Deploy with NIMs

Blueprints

Reference Pipelines

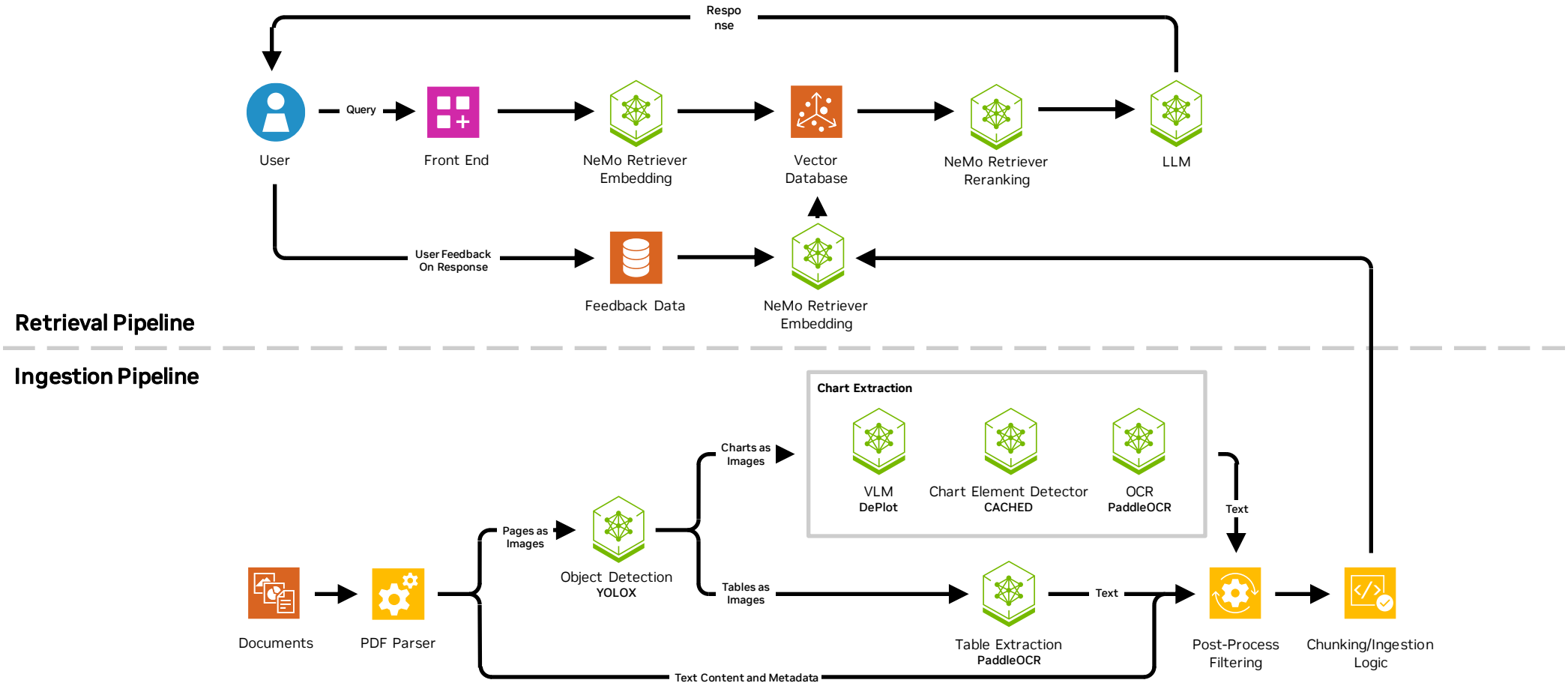
NVIDIA Blueprints for Building AI Agents

An Easy Starting Point for Building Fast, Smart, Enterprise-Grade AI Agents

NVIDIA AI Blueprints	PDF to Podcast	AI Assistant for Customer Service	Vulnerability Analysis for Container Security	Generative Virtual Screening for Drug Discovery	Video Search and Summarization
Specialized AI Agents	 Research Assistant Agent	 Customer Service Agent	 Software Security Agent	 Virtual Lab Agent	 Video Analytics Agent

Multimodal PDF Data Extraction for Enterprise RAG

Unlocks knowledge from trillions of PDFs



NVIDIA GTC

What's Next in AI Starts Here

Workshops March 16-20
Keynote March 18
Conference and Expo March 17–21