Generative AI for Science

Unlocking the power of LLMs with NVIDIA NeMo

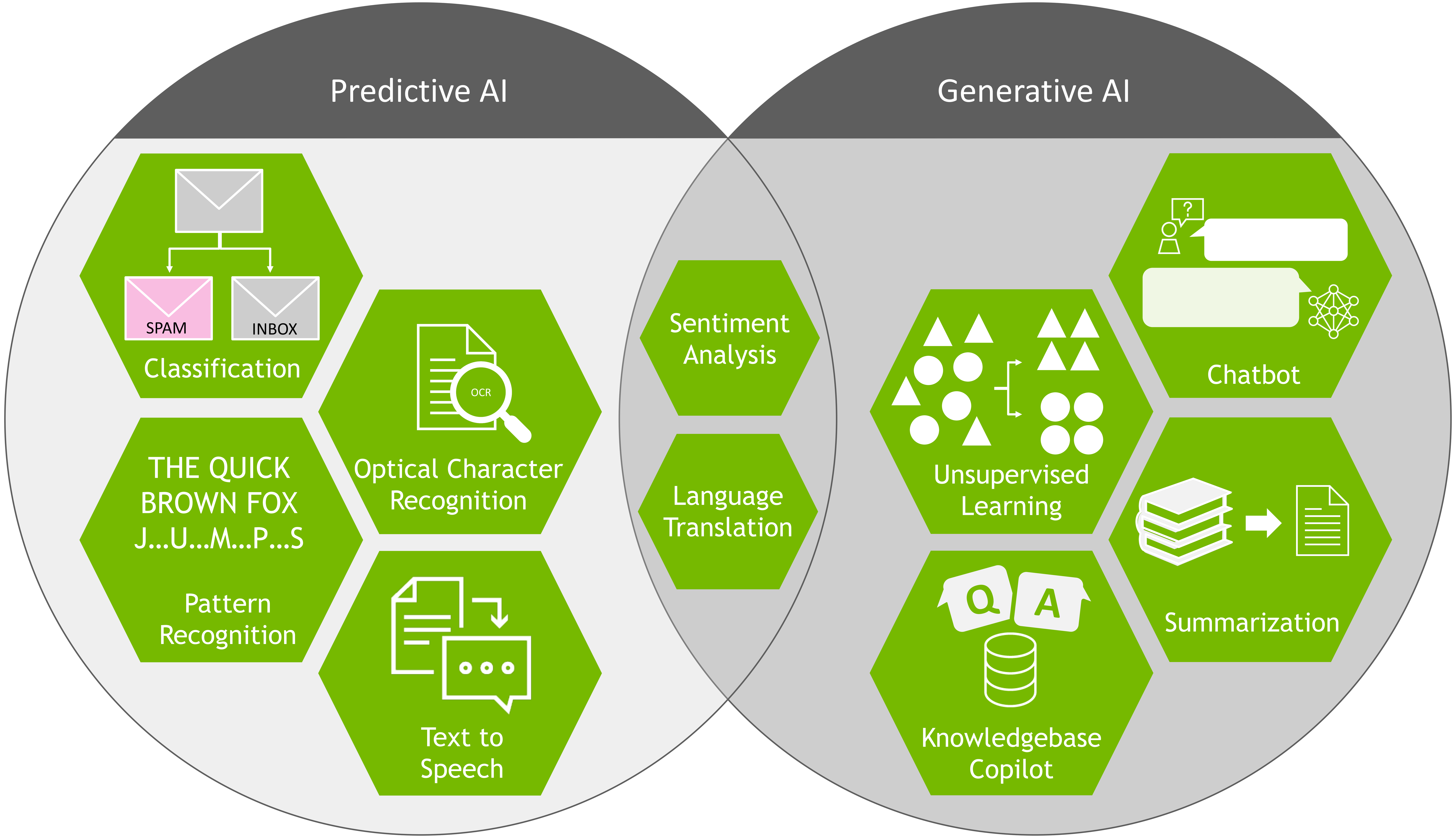
Janaki Vamaraju, Senior Solution Architect, NVIDIA

Zahra Ronaghi, Manager Solution Architect, NVIDIA

Agenda

- Generative AI and Large Language Models (LLMs)
- NVIDIA NeMo Framework
- Retrieval Augmented Generation (RAG)
- Domain Adapted LLMs

When to use Generative AI?



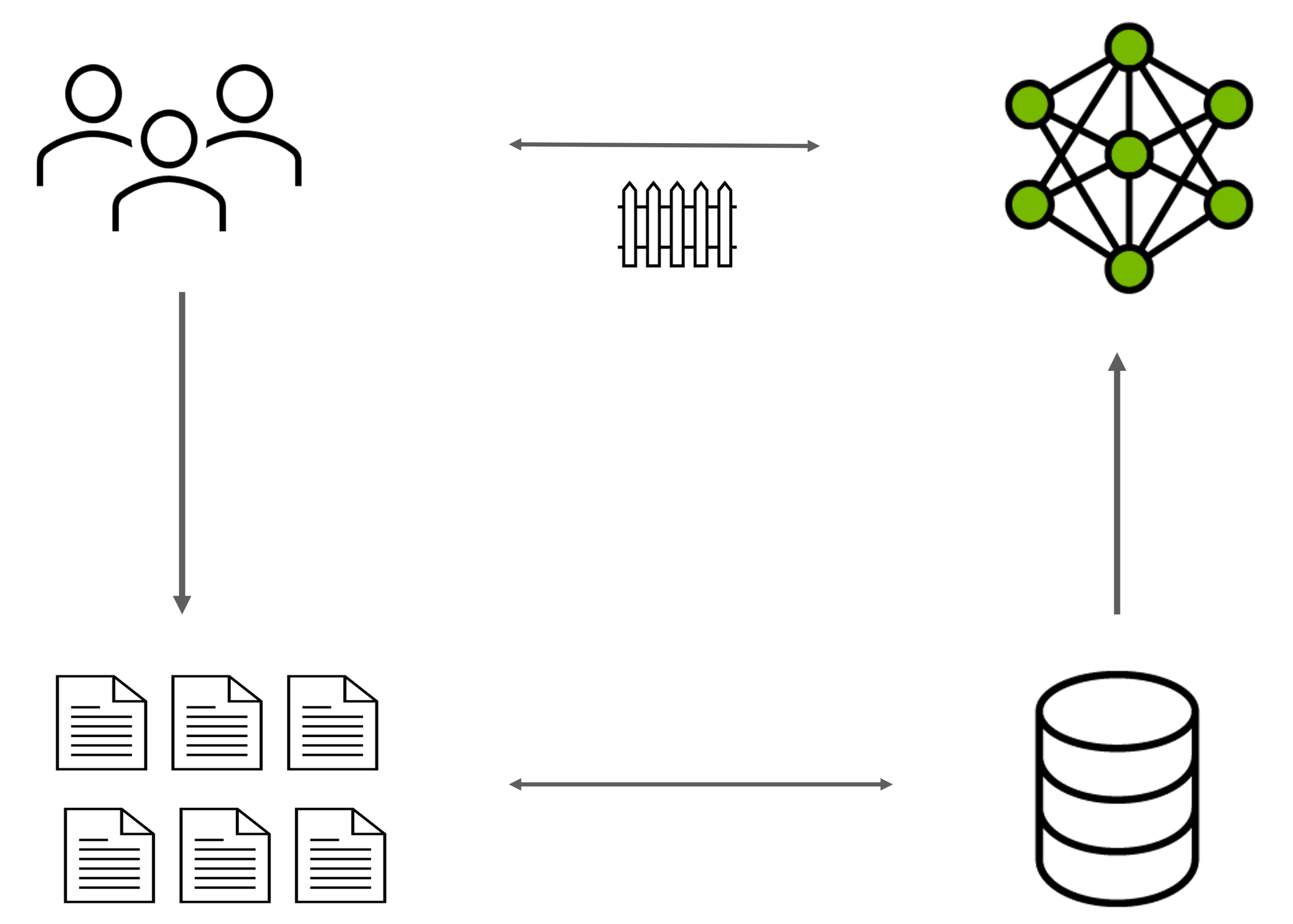
Predictive AI focuses on understanding historical data and making accurate predictions

Generative AI creates new data based on patterns and trends learned from training data

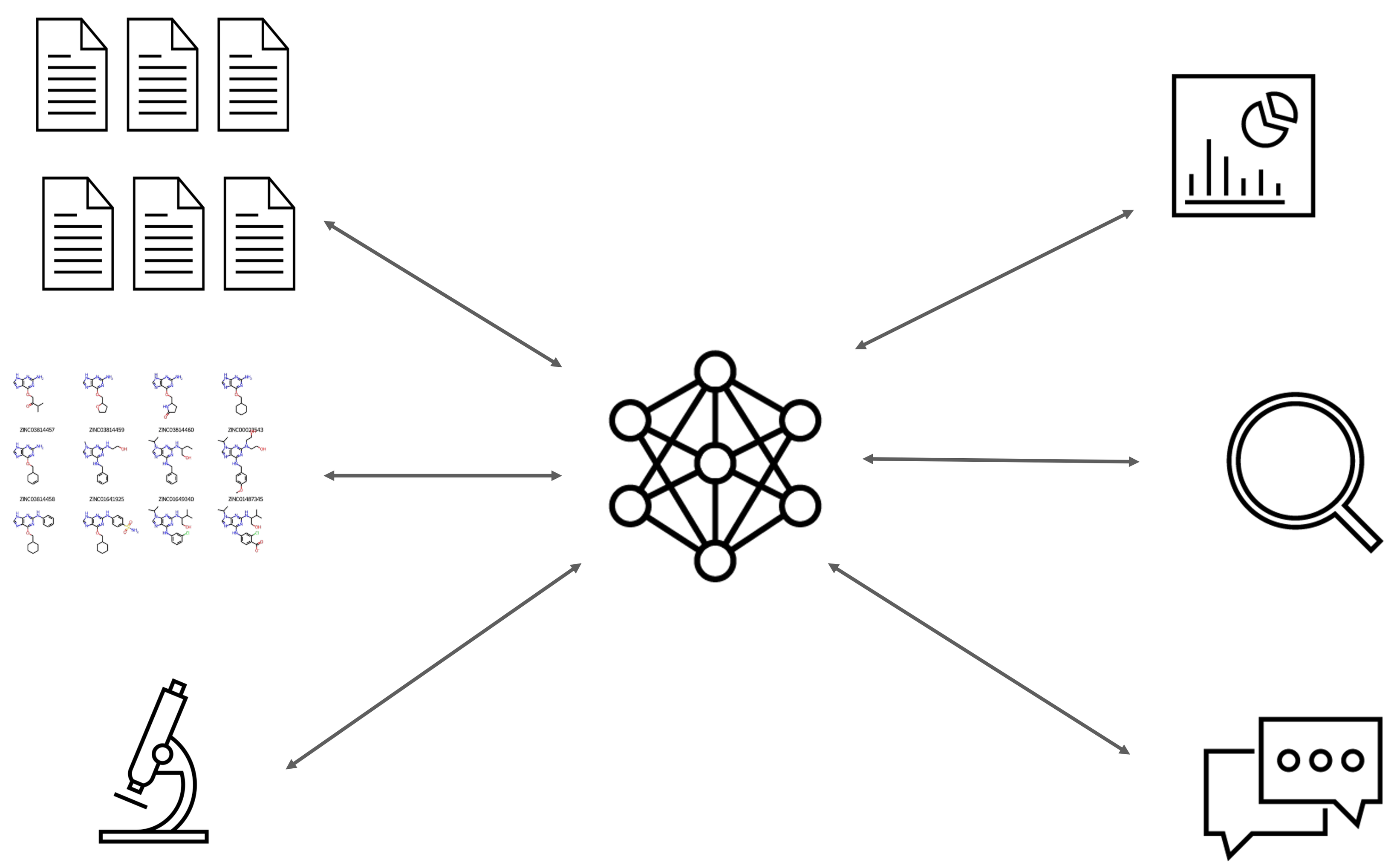
Intersection of Gen AI and Science

Building Foundation Models for Science Research and Discovery

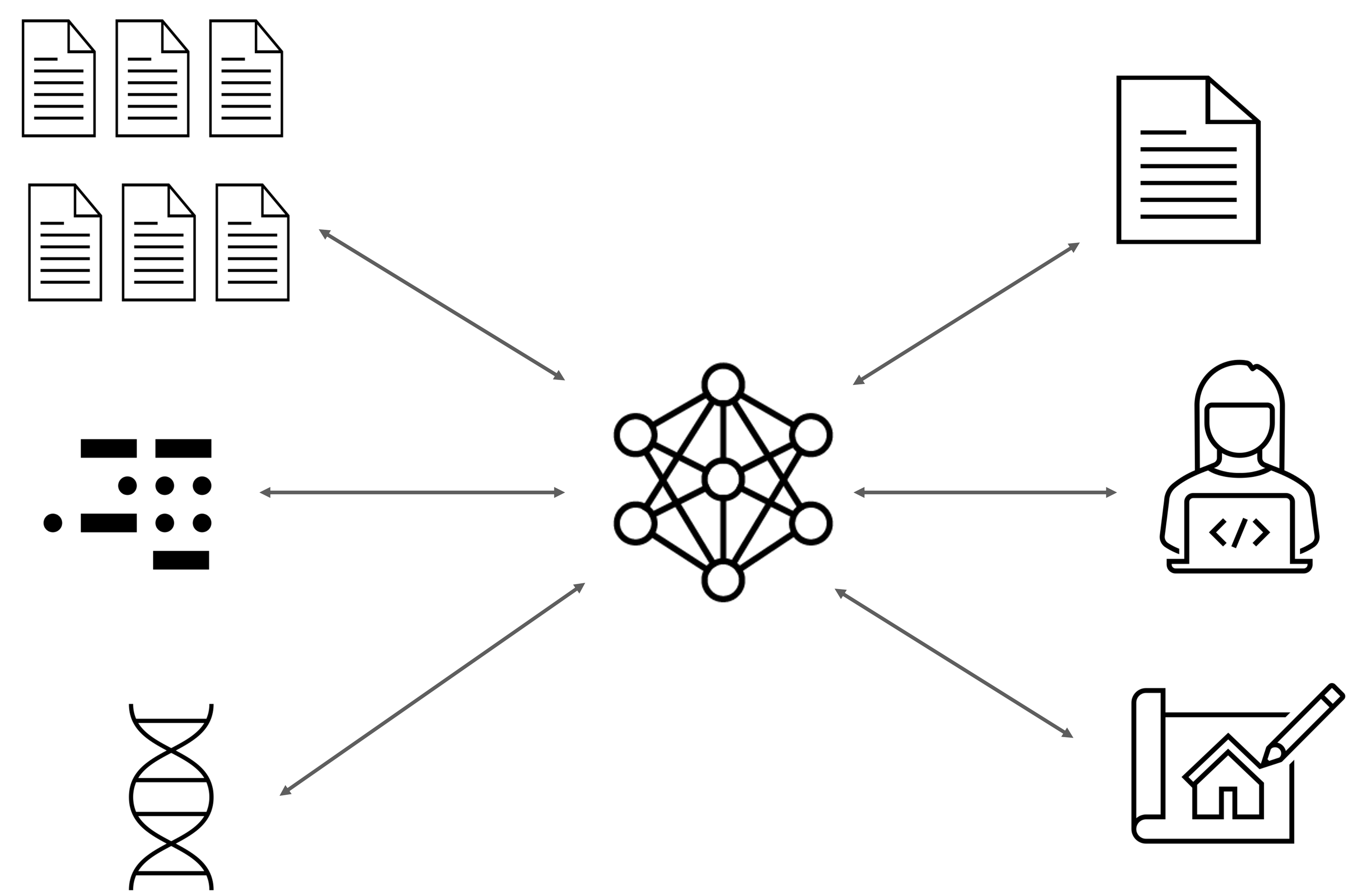
Summarize



Synthesize



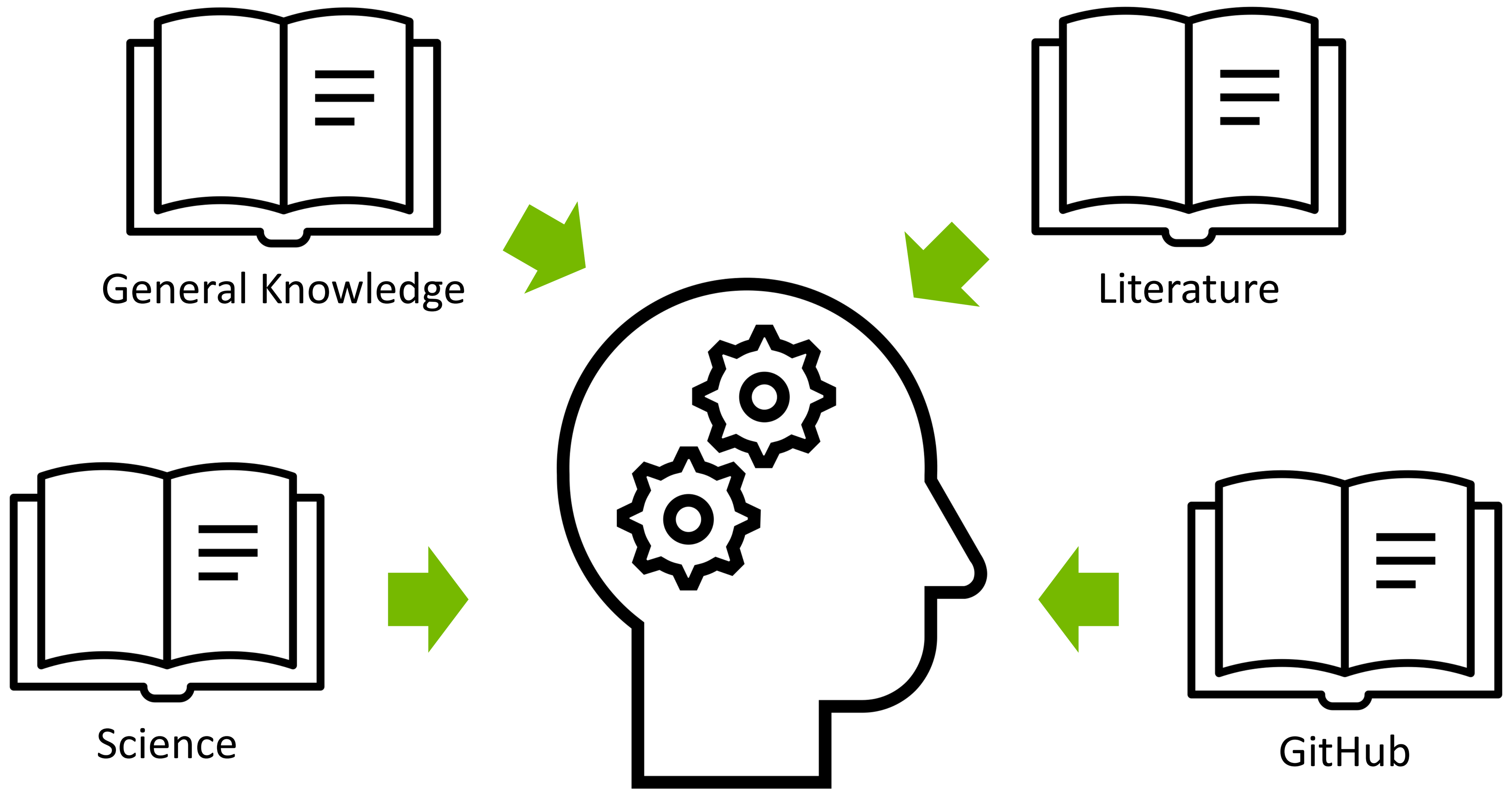
Generate



How to train an LLM

Creating a “Foundation Model”

- Step 1 - **Pretraining**. Feed it an enormous corpus to learn from.

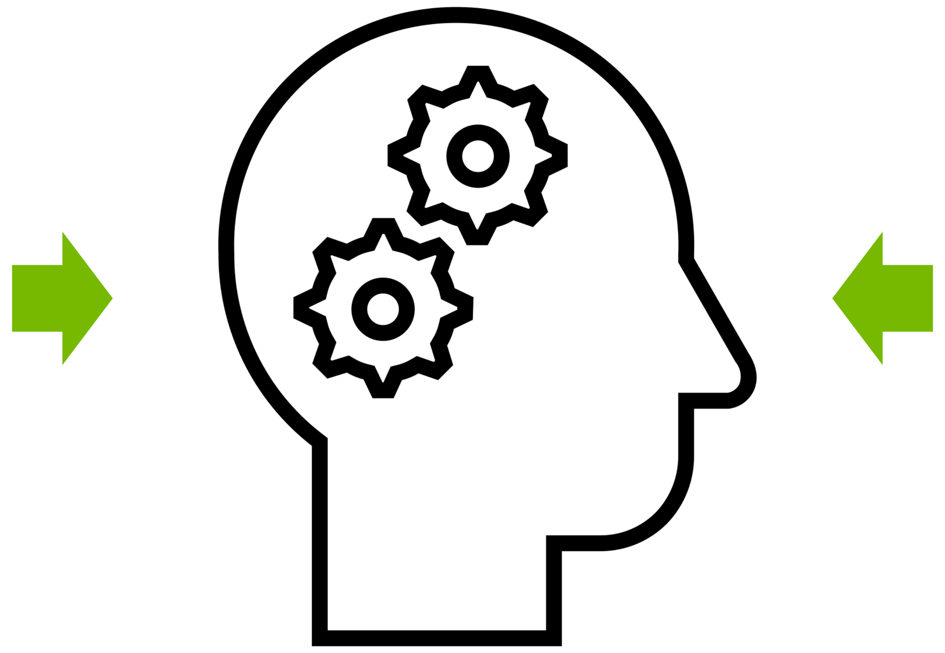


- Step 2 – **Fine tuning**. Provide demonstrations of how you want it to answer questions

‘Q: What virus causes covid?

‘Q: Write a poem about a cat in love with a zebra.

A: There once was a cat
in search for a mate.
She saw a zebra
And knew it was fate…’



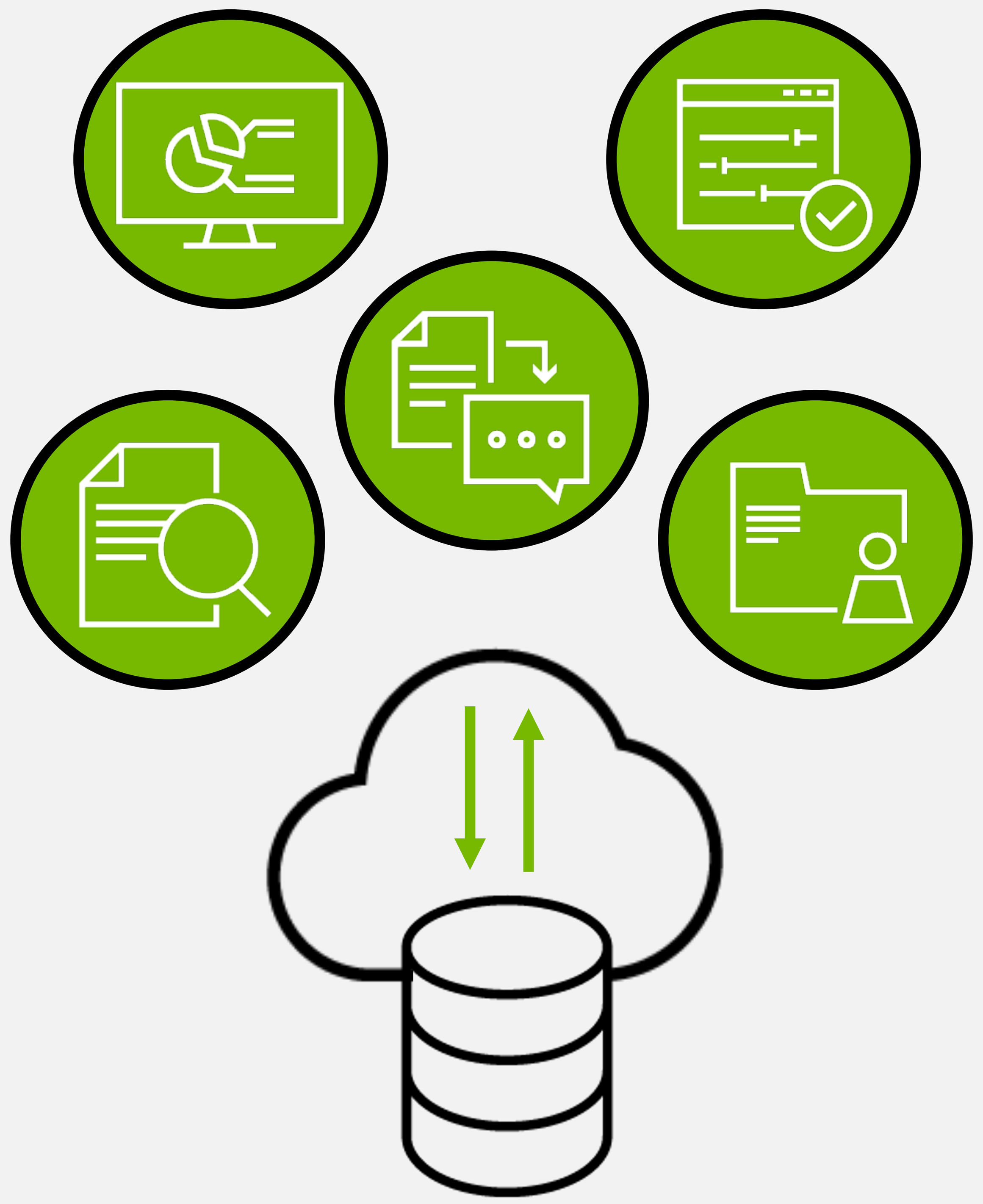
‘Q: Code Quicksort in C++

‘Q: Who do want to win the next election?

A: As an AI, I do not have political opinions’

Requirements for Building Custom LLMs

Training Data



Accelerated Computing

DGX &
DGX Cloud

aws
Google Cloud
Microsoft Azure
ORACLE Cloud Infrastructure

DELL Technologies
Hewlett Packard Enterprise
Lenovo
SUPERMICRO

Training and Inference Tools

Data Curation

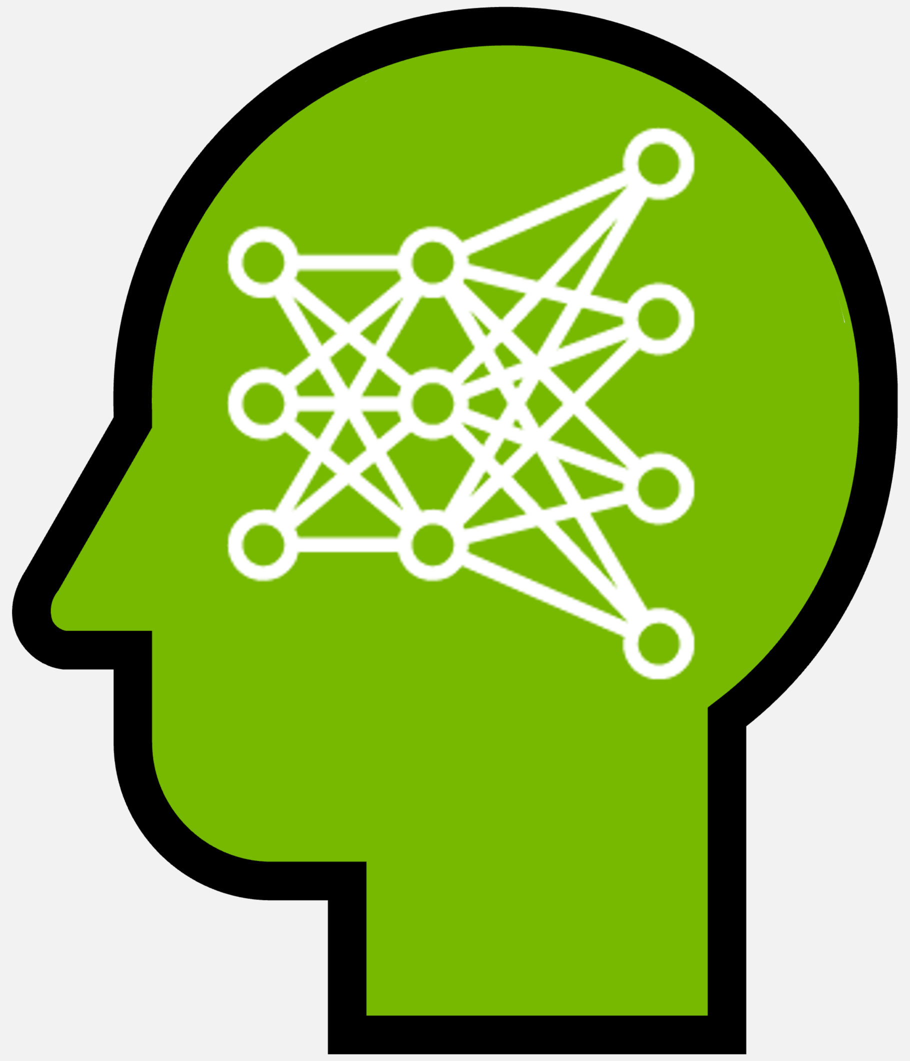
Foundations Models

Training & Customization

Accelerated Inference

AI Expertise

Internal Expertise



Solution Delivery Partners

Building Generative AI Applications

Build, customize and deploy generative AI models with NVIDIA NeMo

https://github.com/NVIDIA/NeMo

Multi-Modality

Build language, image, generative AI models

Data Curation at Scale

Extract, deduplicate, filter info from large unstructured data @ scale

Optimized Training

Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.

Model Customization

Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi

Deploy at Scale

Run optimized inference at-scale anywhere

Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

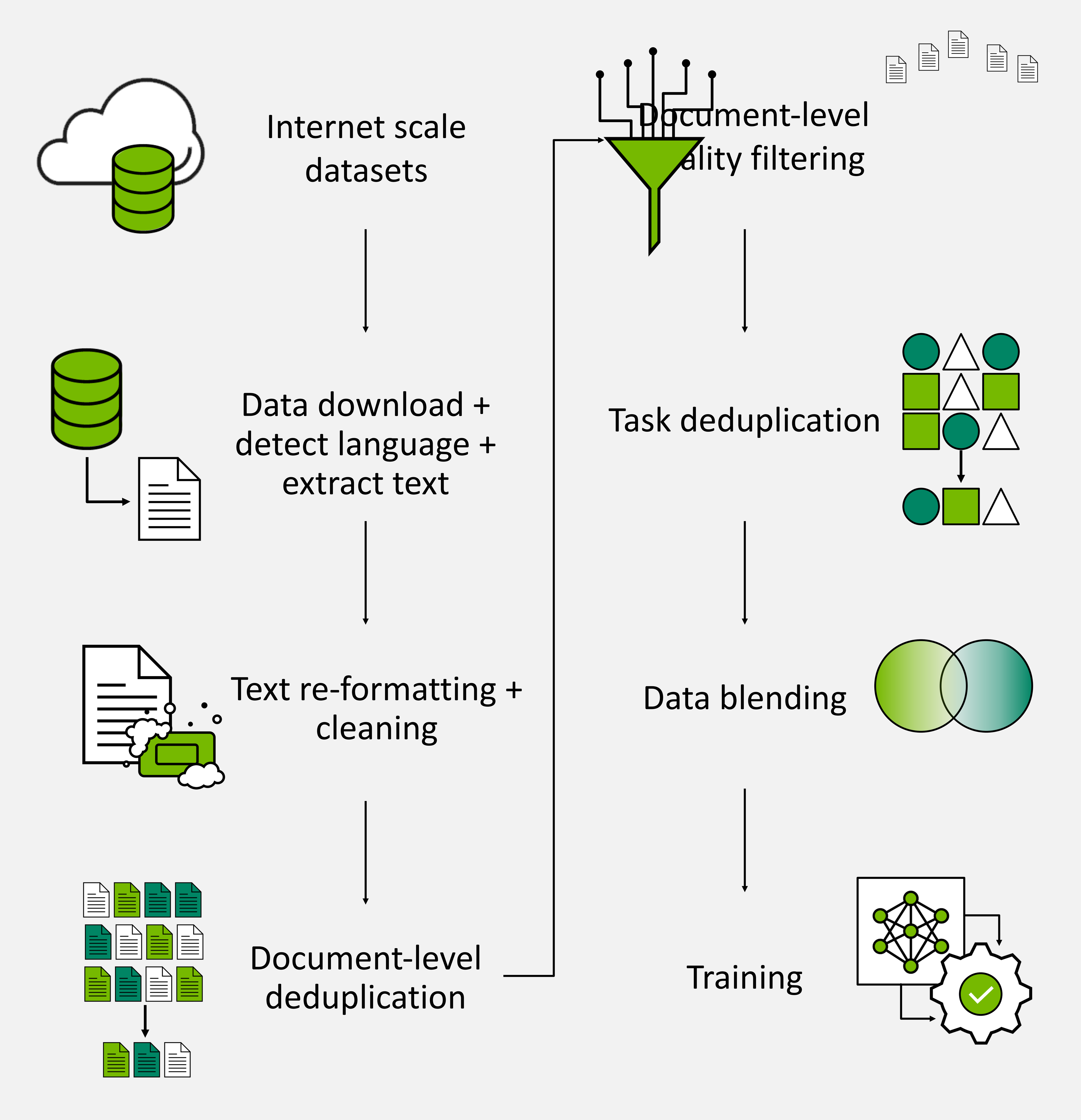
Data Curation Improves Model Perfomance

NeMo Data Curator enabling large-scale high-quality datasets for LLMs

- Reduce the burden of combing through unstructured data sources
- Download data and extract, clean, deduplicate, and filter documents at scale

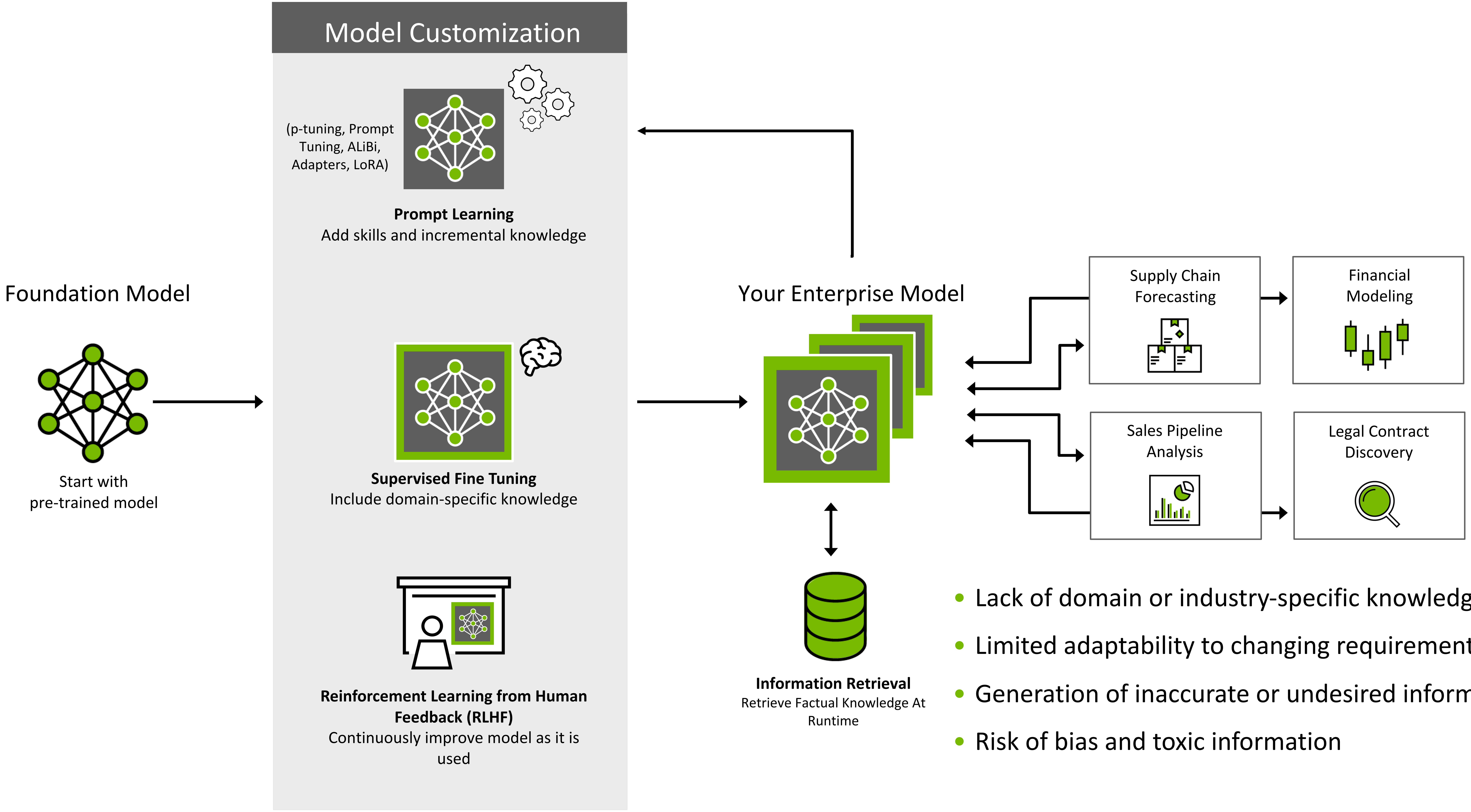
NeMo Data Curator steps:

1. Data download, language detection and text extraction - HTML and LaTeX files
2. Text re-formatting and cleaning - Bad Unicode, newline, repetition
3. GPU accelerated Document Level Deduplication
   - Fuzzy Deduplication
   - Exact Deduplication
4. Document-level quality Filtering
   - Classifier-based filtering
   - Multilingual Heuristic-based filtering
5. Task Deduplication - Performs intra-document deduplication



Model Customization for LLMs

Customization techniques to overcome the challenges of using foundation models



- Lack of domain or industry-specific knowledge
- Limited adaptability to changing requirements
- Generation of inaccurate or undesired information
- Risk of bias and toxic information

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 • SteerLM
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

NVIDIA NeMo Works with Powerful Generative Foundation Models

Suite of generative foundation language models built for enterprise hyper-personalization

Fastest Responses

Nemotron-3 8B

GPT-8B w/ 3.5T tokens. +SFT, SteerLM.
53 Languages I/O: 4K tokens

Balance of Accuracy - Latency

Nemotron-3 22B

GPT-22B w/ 1.1T tokens. + SFT private mix.
50 Languages. I/O: 4K tokens

For Complex Tasks

Nemotron-3 43B

GPT-43B w/ 1.1T tokens. + SFT private mix.
50 Languages. I/O: 4K tokens

Information Retrieval

NeMo Retriever

Community-Built Models

Code Llama
Meta

Falcon LLM
Falcon

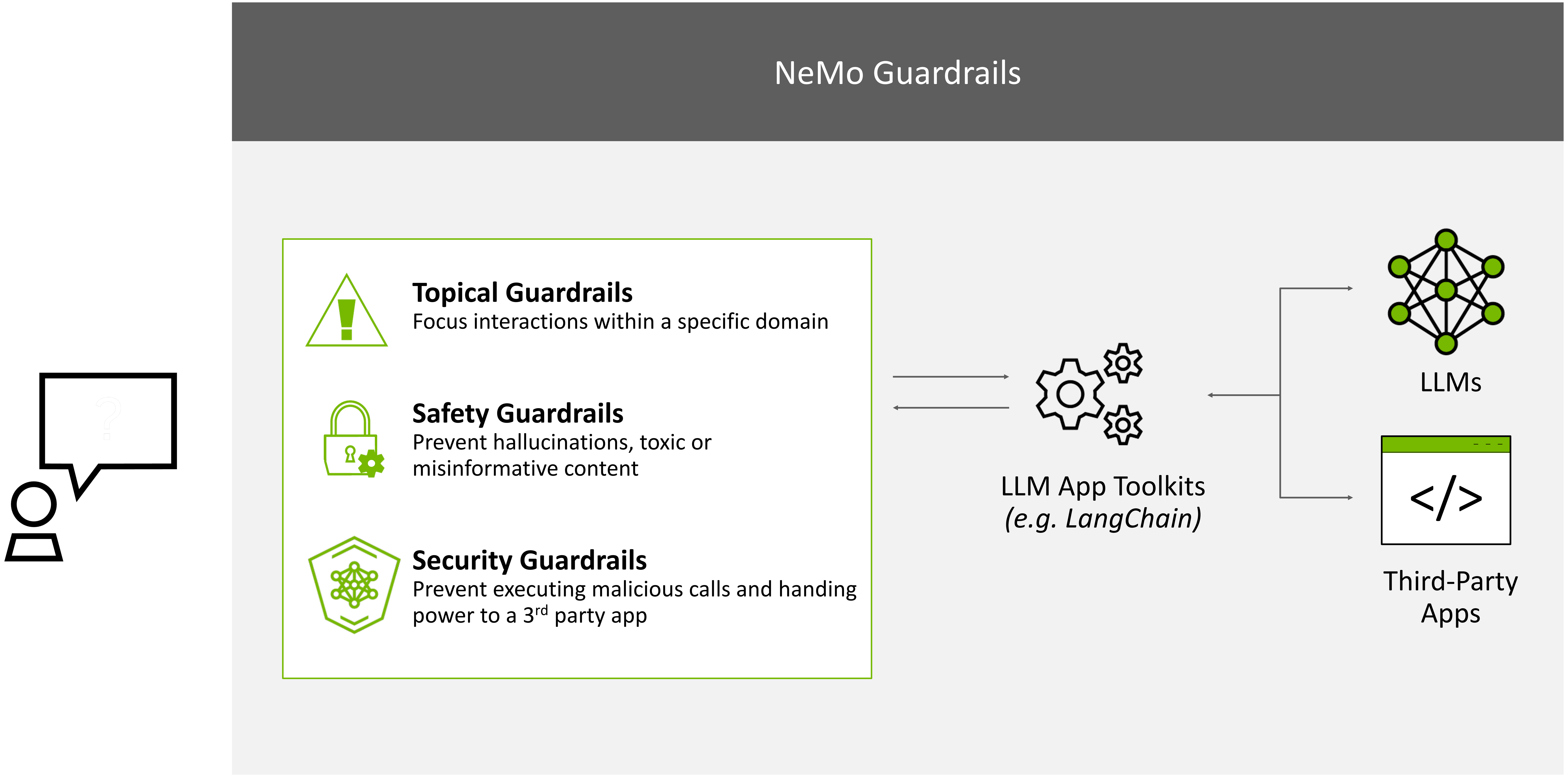
Llama 2
Meta

MPT
Mosaic ML

StarCoder
ServiceNow & Hugging Face

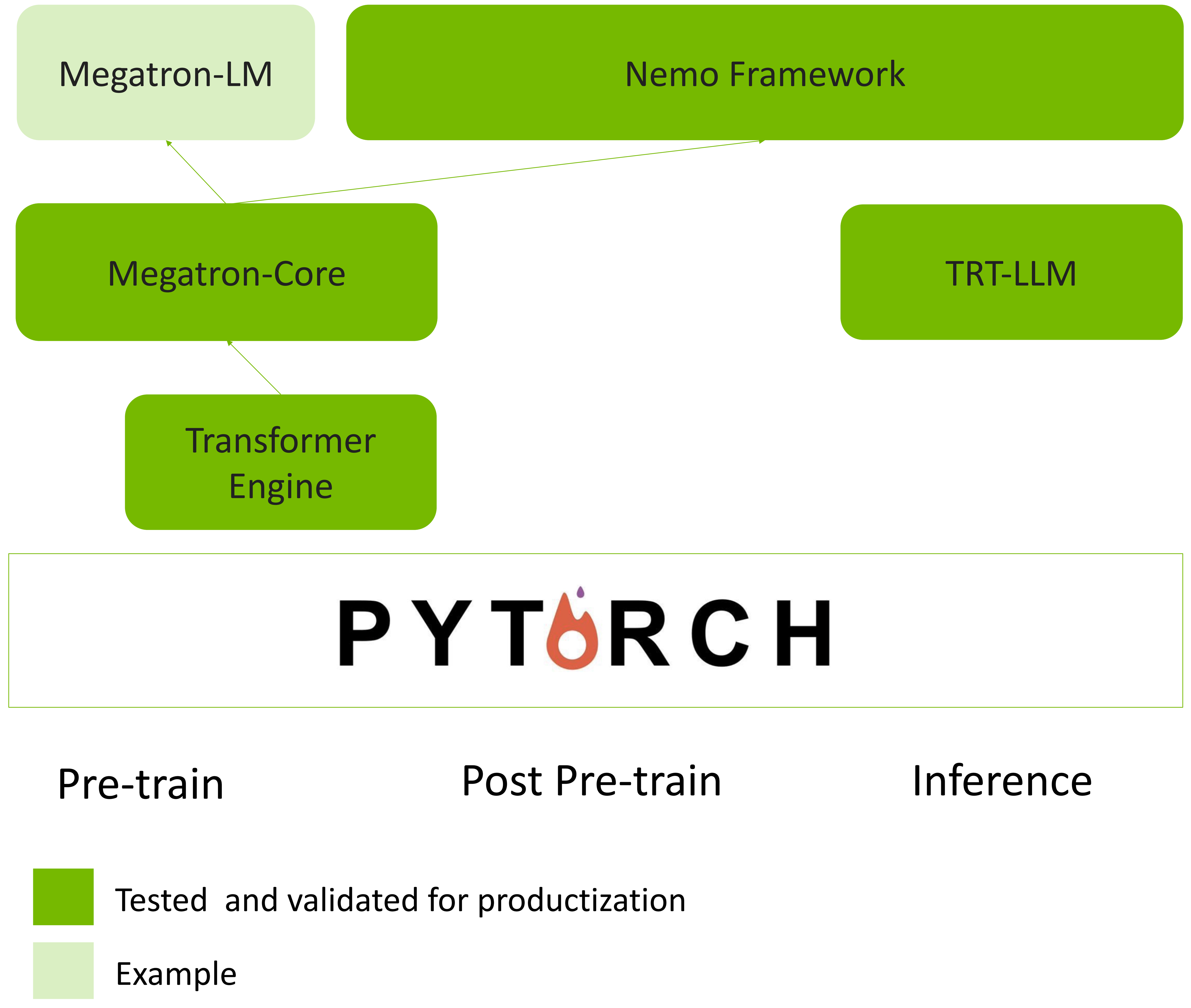
Guardrails Can Keep Generative AI On Track

Ensure accuracy, appropriateness, and security in LLMs



NVIDIA's LLM offerings for Training And Inference

All Are available on Github and NGC



Nemo Framework: An OOTB FW for experimenting, building, training, tuning and deploying LLM models.
https://github.com/NVIDIA/NeMo

Megatron-LM: A lightweight framework reference for using Megatron-Core to build your own LLM framework.
https://github.com/NVIDIA/Megatron-LM

Megatron-Core: A library for GPU optimized techniques for LLM training. Can be used to build custom LLM frameworks.
https://github.com/NVIDIA/Megatron-LM/tree/main/megatron/core

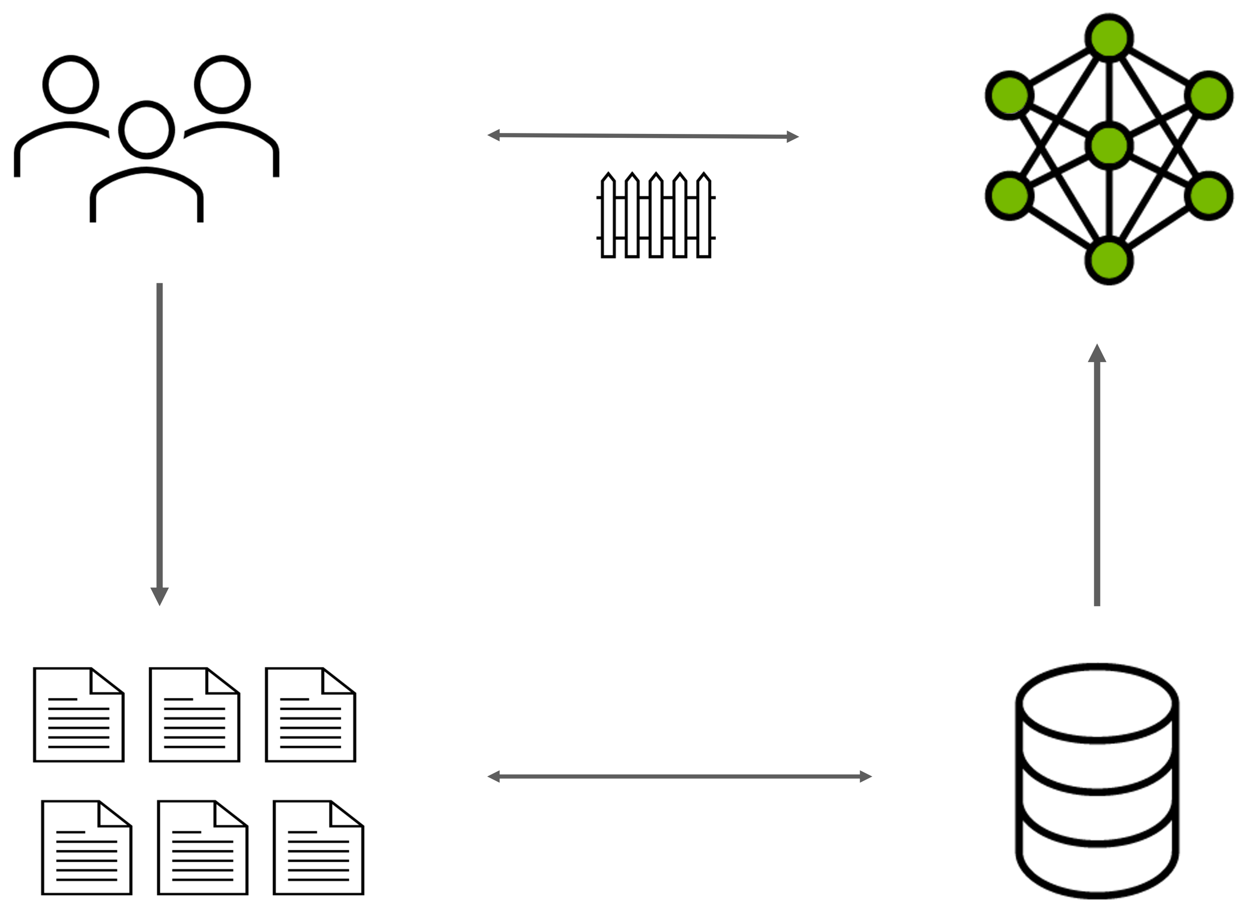
Transformer Engine: Hopper accelerated Transformer models. Specific acceleration library, including FP8 on Hopper.

TRT-LLM: an open-source library for optimal performance on the latest LLMs for inference on NV GPUs.
https://github.com/NVIDIA/TensorRT-LLM

Decades of Scientific Research Intersecting with GenAI

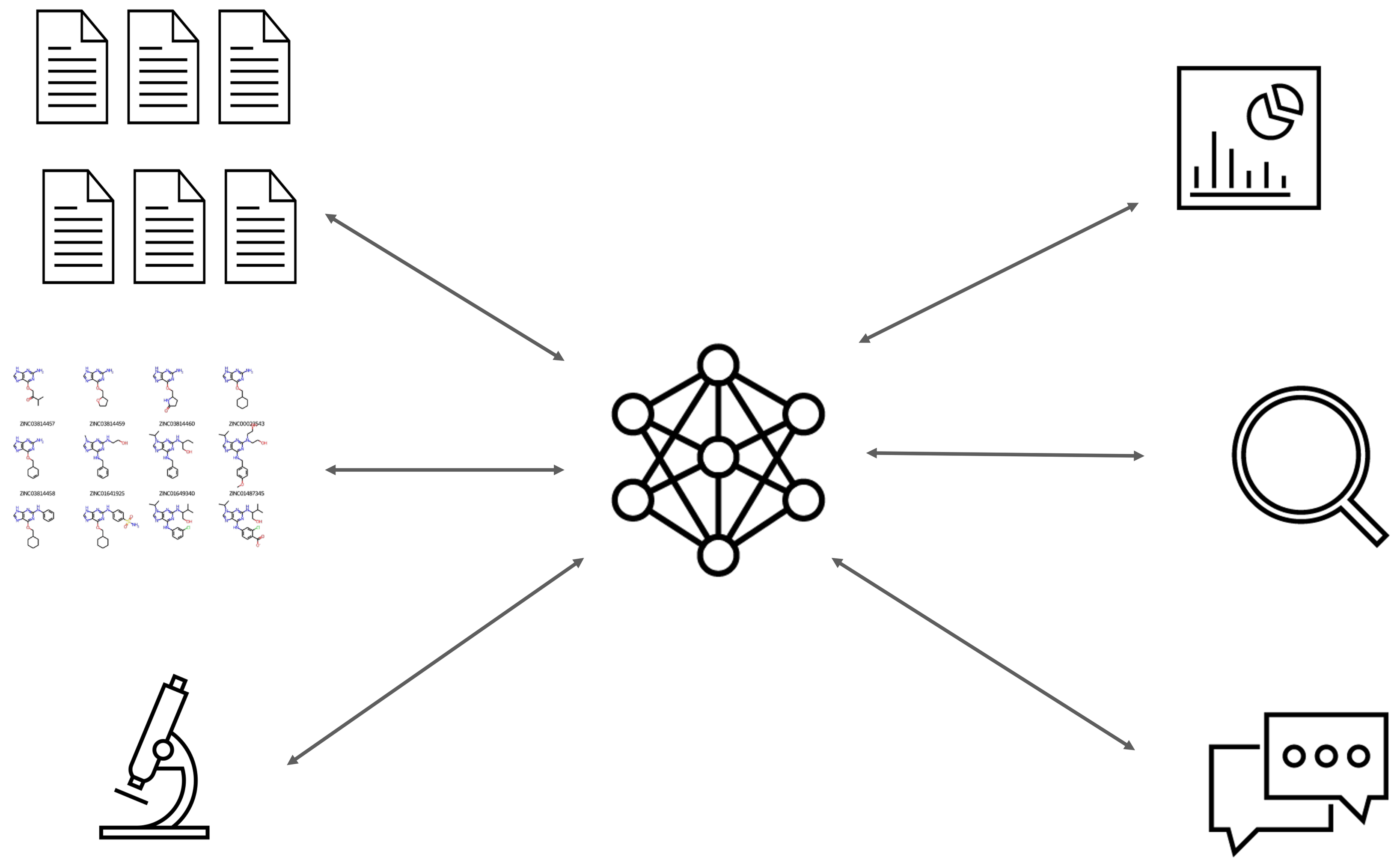
3 Distinct Categories

Summarize



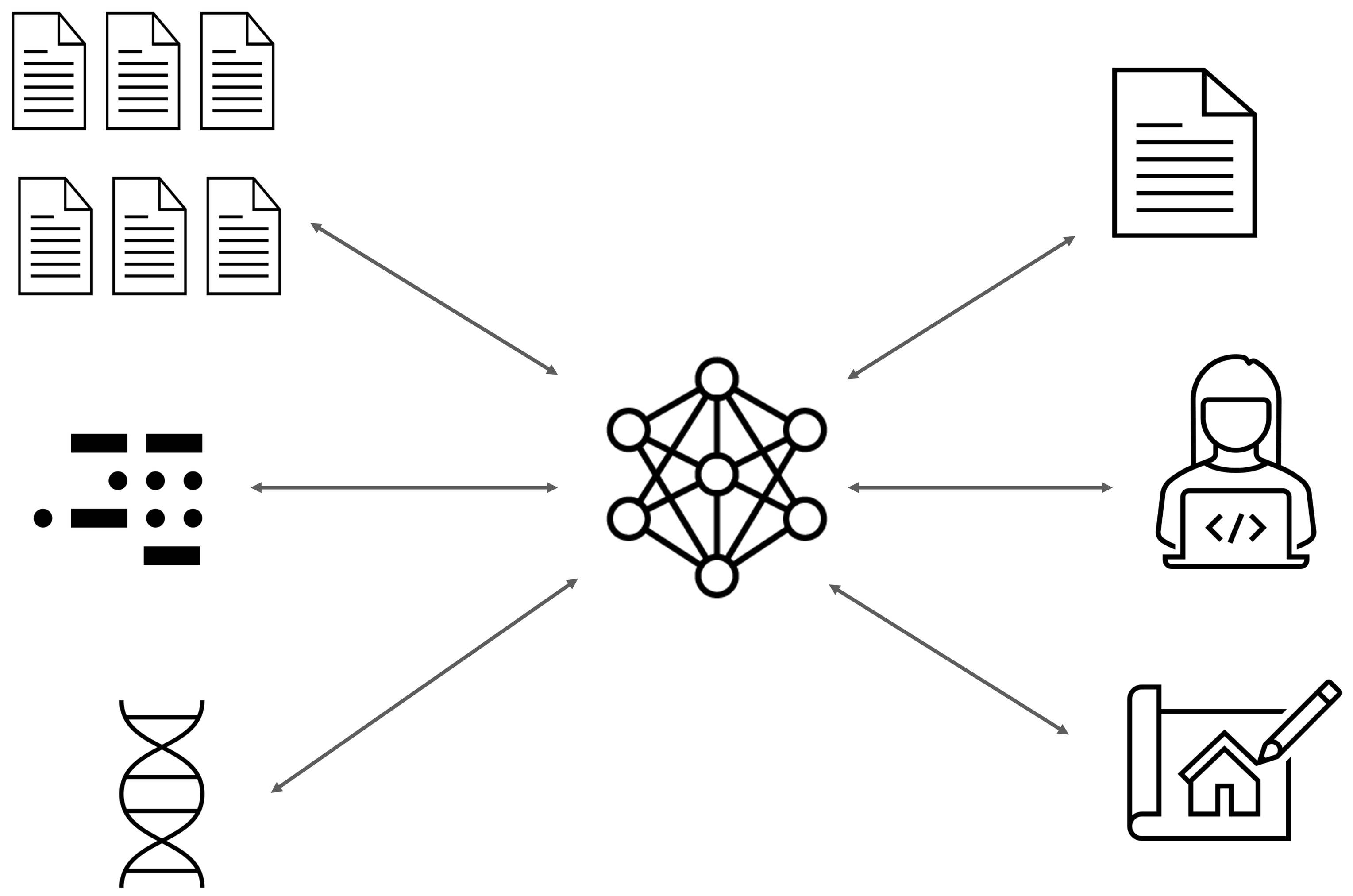
OTS LLMs
RAG
Guardrails

Synthesize



OTS LLM
Multiple Data Sources
Customization/Tuning
Guardrails
RAG

Generate



LLM from Scratch
Multiple Data Sources,
Customization/Tuning
Guardrails
RAG

NIM, NeMo Models, NeMo Retriever, Guardrails | Nemo FW, TRT-LLM | MegatronCore

NVIDIA NIM Streamlines the Path to Production

Easiest and most performant way to deploy generative AI and LLM models coupled with industry-standard APIs

Prebuilt container and helm chart tested and validated across infrastructure

Industry standard APIs with NVIDIA Unified Cloud Standards

Domain specific code for each NIM domain category, including LLMs, Images, VLMs, video, healthcare, biology, genomics, and more

Optimized inference engines for each model and hardware SKU

Support for custom models built by users targeted use cases

NVIDIA AI Enterprise approved base container

NVIDIA NIM

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 state-of-the-art 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

NVIDIA NIM is the Fastest Path to AI Inference

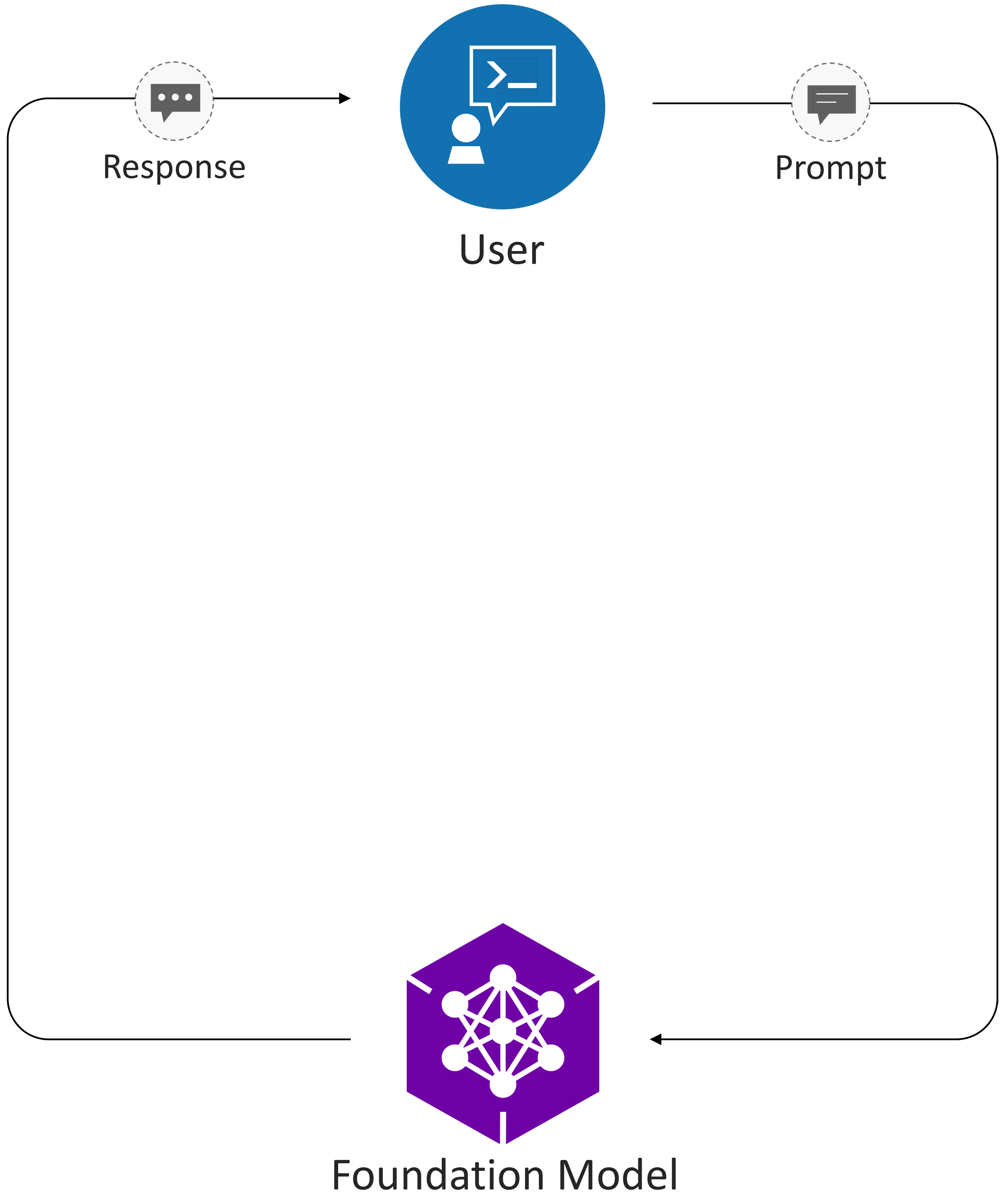
Reduces engineering resources required to deploy optimized, accelerated models

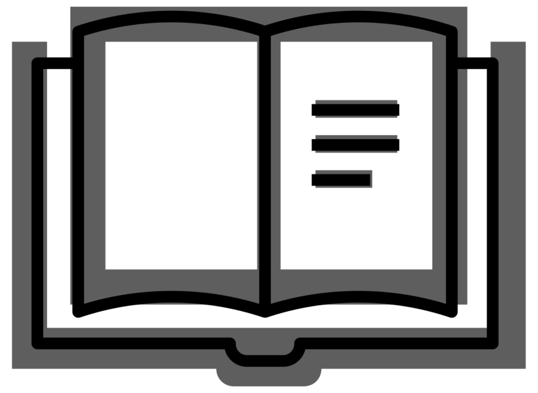
	NVIDIA NIM	Triton + TRT-LLM Opensource
Deployment Time	5 minutes	~1 week
API Standardization	Industry standard protocol OpenAI for LLMs, Google Translate Speech	User creates a shim layer (reducing performance) or modify Triton to generate custom endpoints
Pre-Built Engine	Pre-built TRT-LLM engines for NV and community models MISTRAL AI_ Llama 2 starcoder NVIDIA Nemotron	User converts checkpoint to TRTLLM 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

Retrieval Augmented Generation (RAG)

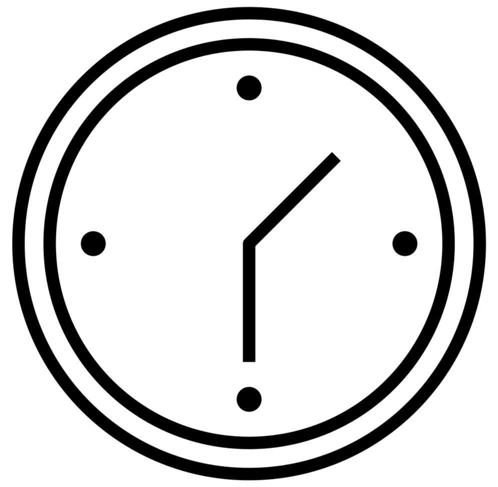
LLMs are Powerful Tools but Not Accurate Enough

Without a connection to enterprise data sources, LLMs cannot provide accurate information





Lacking proprietary knowledge

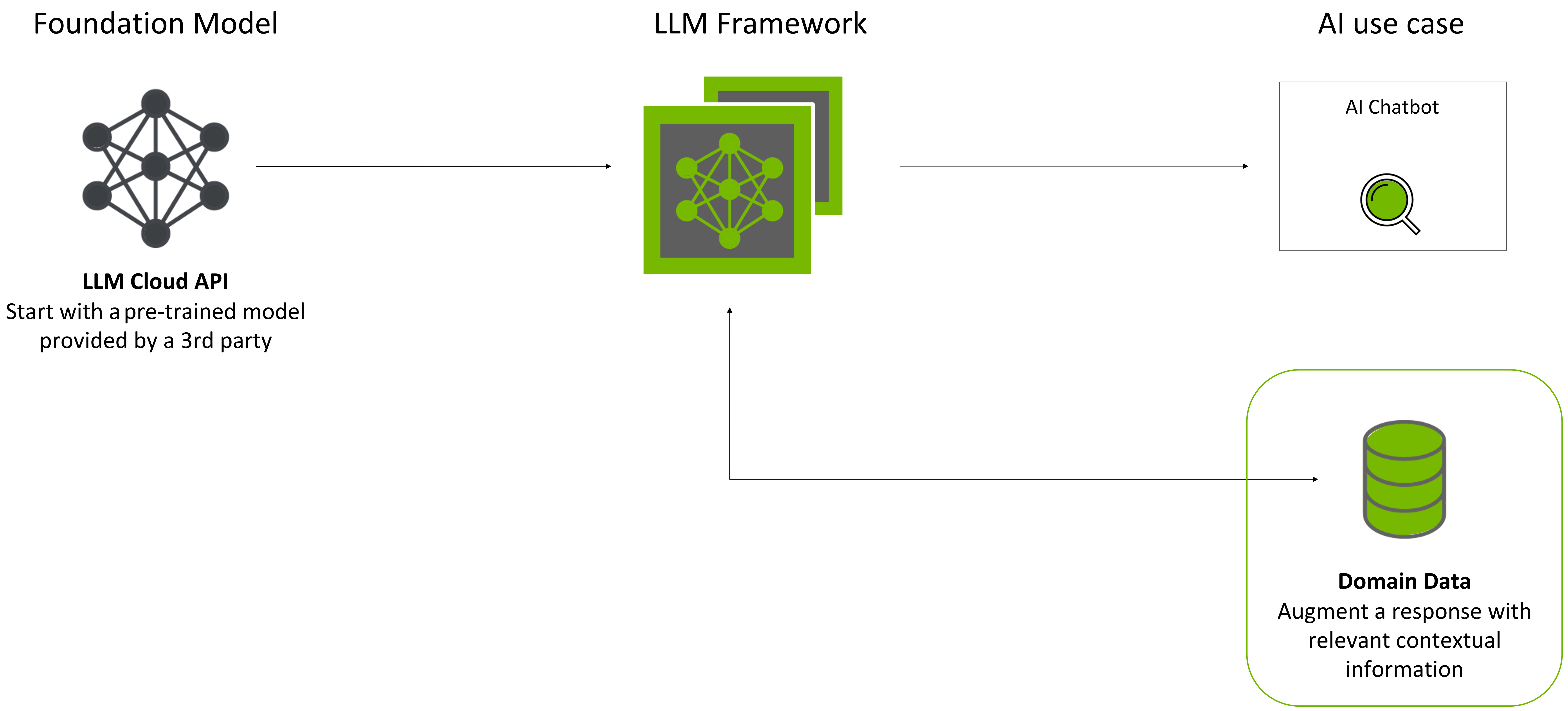


Risk of outdated information

Hallucinations

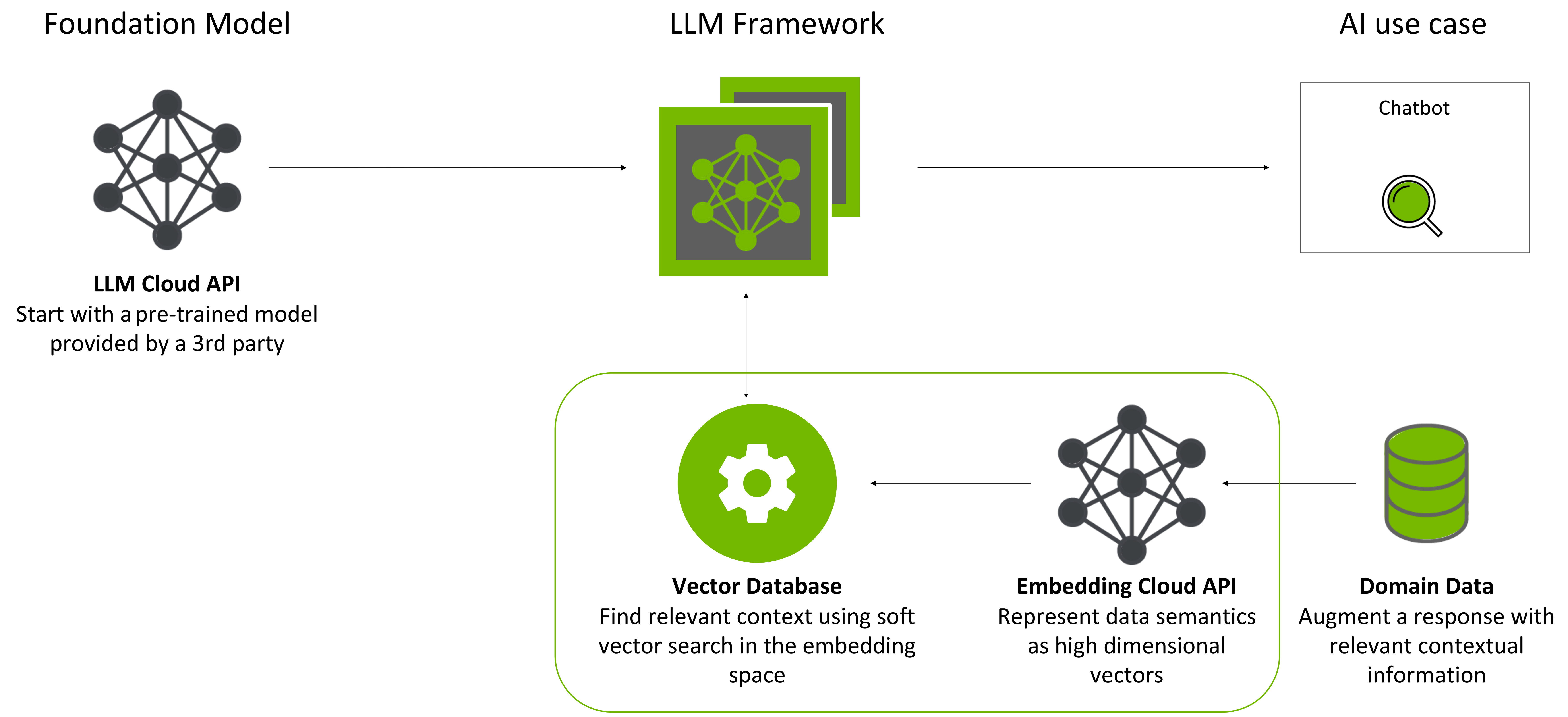
Use Retrieval-Augmented Generation (RAG)

Provide context at a query time to minimize hallucinations and keep LLM answers fresh



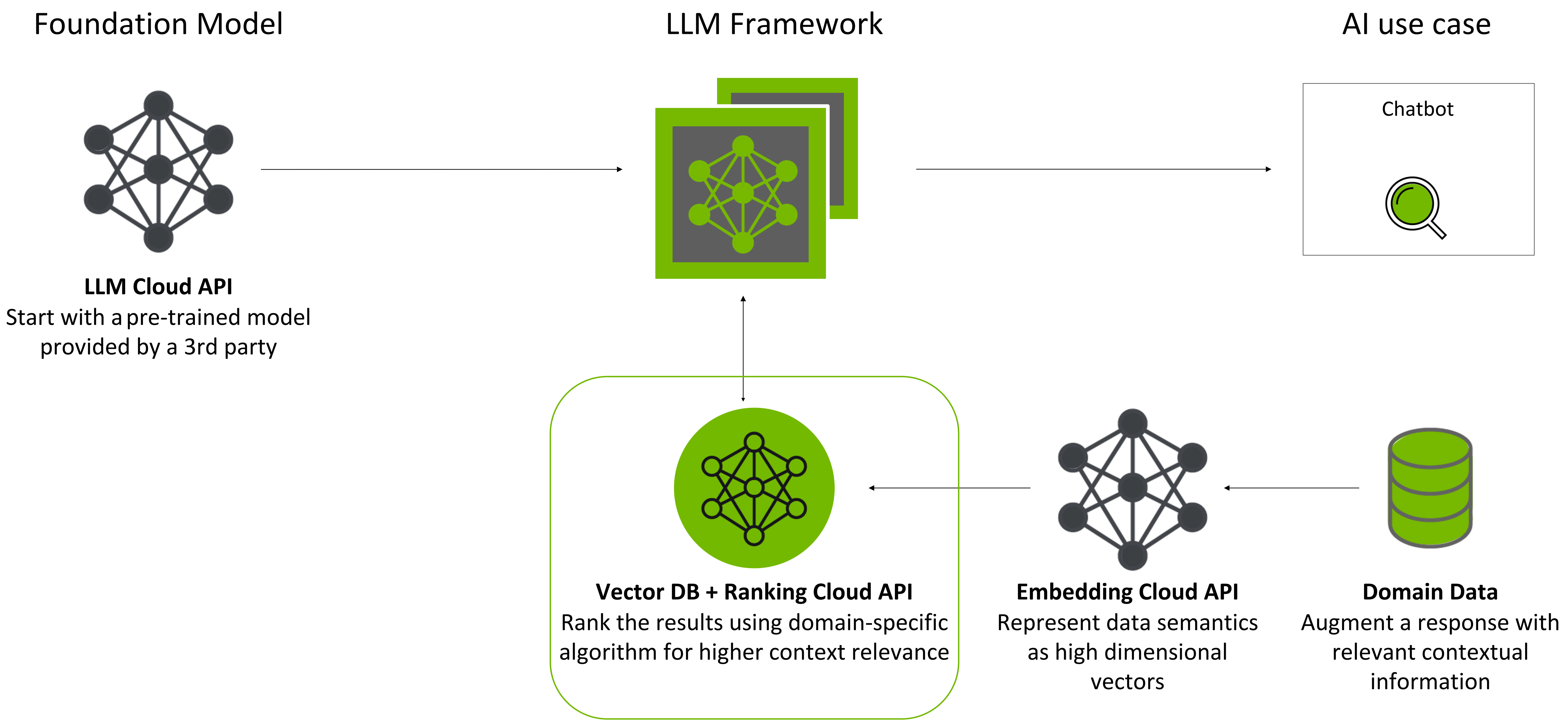
Use Retrieval-Augmented Generation (RAG)

Represent data as embeddings to support “soft” vector similarity search



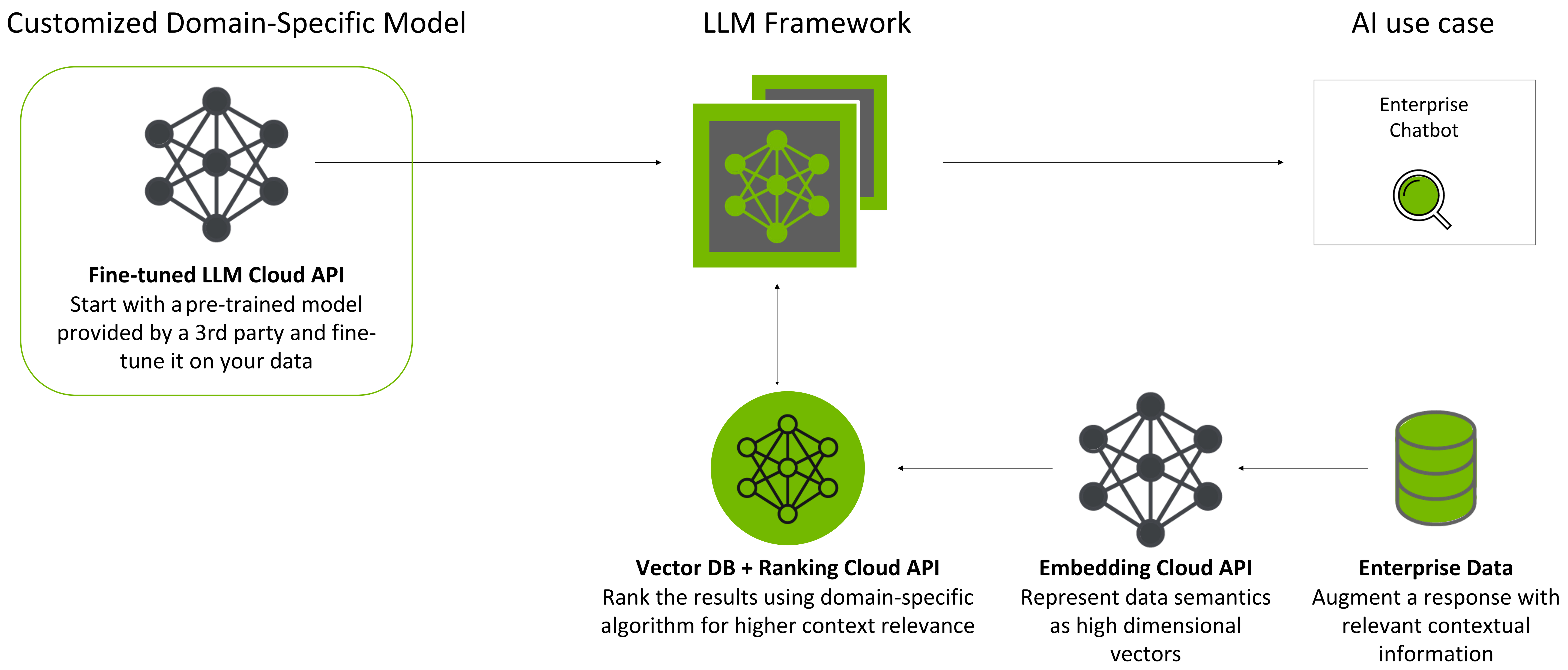
Use Retrieval-Augmented Generation (RAG)

Increase context relevance using domain-specific (re)ranking algorithm



Fine-tune Your Model to Understand Domain Semantics

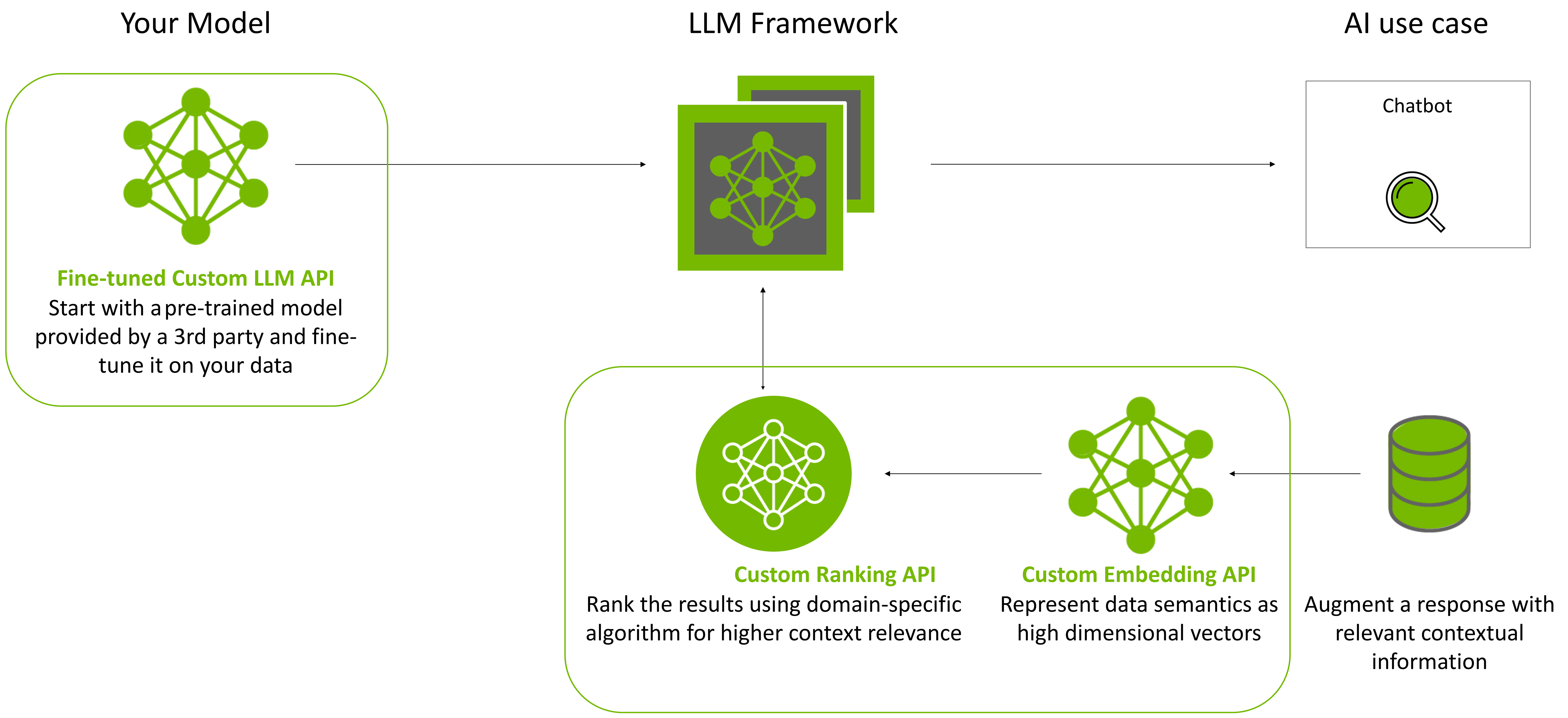
Increase LLM accuracy by customizing for your enterprise use case



Adopt Open Source Models to Gain Flexibility and Control

Open source models (LLM, embedding, ranking) help protect enterprise data and IP

- Falcon 40B
- Gemma 2B
- Gemma 7B
- Llama-2 7B
- Llama-2 13B
- Llama-2 70B
- Code Llama 34B
- Mistral 7B
- Mixtral 8x7B
- Nemotron 8B
- Nemotron 43B
- GPT3 175B
- MPT 30B

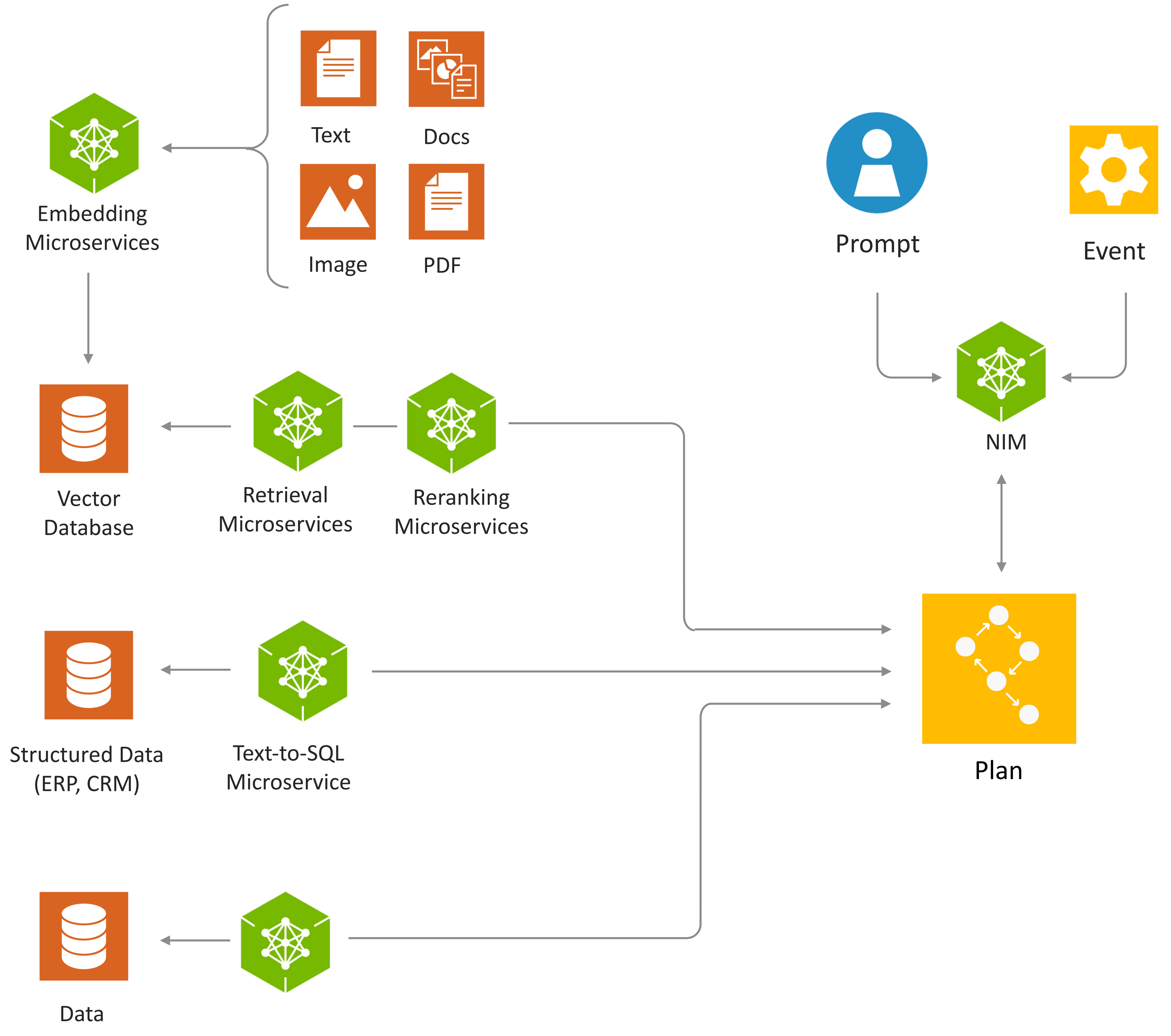


NVIDIA Provides Optimized Retrieval Augmented Generation

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

NeMo Retriever Supercharges RAG Applications

World Class Accuracy and Throughput



2X World-class accuracy with nearly 2x fewer incorrect answers

7X Faster embedding inference throughput

Optimized Inference Engines

World class models and community model support

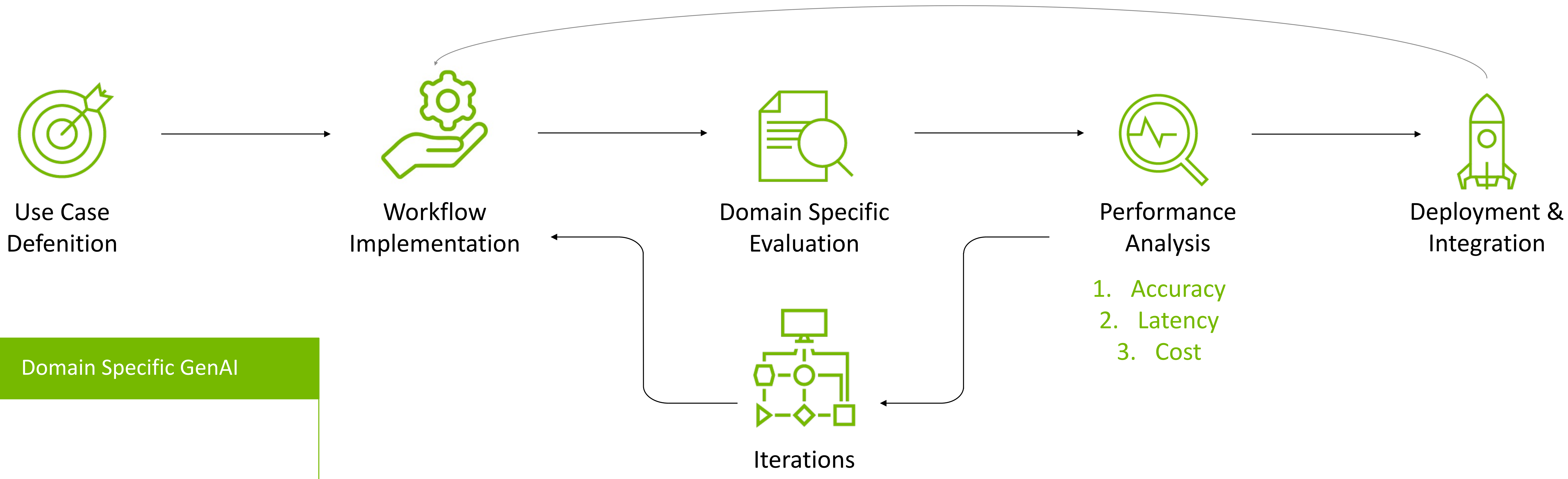
Flexible and modular deployment

Customizable models and pipelines

Production Ready

Domain Adapted LLMs

Building a Domain Specific Gen AI model is a Multistage Process



Domain Specific GenAI

Train it on a skill – perform a task in a certain way

Give it ethics and personality – align its response based on human preferences and values

Teach it a set of facts – connect to a knowledge base

Pipeline Structure

Models Selection

RAG

Context Learning

Dataset

Domain Data

Data Blends

Open-Source Dataset

Model Domain Adaptation

LLM Model

Embedding

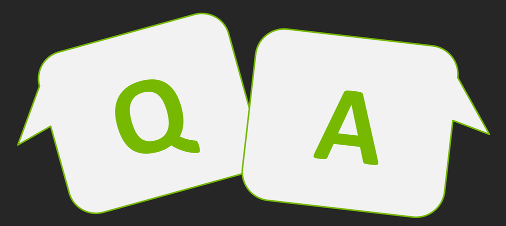
Re-Ranker

LLM Assistant for Chip Design - ChipNeMo

AI **copilot** built by NVIDIA research to assist one of the most complex engineering efforts, designing semiconductors.

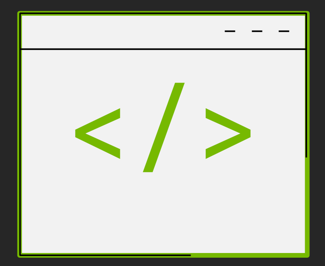
Responds to **questions about GPU architecture and design** while helping engineers quickly find technical documents in early tests. It will also **create snippets of about 10-20 lines of software** in two specialized languages chip designers use, making it easier to develop new code.

Using proprietary data to customize a foundation model, researchers found that a much **smaller 13B parameter model could out preform larger general purpose LLMs.**



Q&A for GPU ASIC Architecture

Bug Analysis & Reports



Code Generation for VLSI Tools

ChipNeMo LLM Assistant

Three chip design use cases: EDA Code Generation, Bug Summarization, Design-assist Chatbot

Accuracy

Correctness on wide range of **domain-specific tasks**

Avoid security risks with third party APIs

Model groundedness in the chip domain (e.g. retrieval hit-rate)

Latency

Fast batch evaluation on domain-specific benchmarks

Real-time responses for NVIDIA engineers

Cost

Development (GPU training time, number of data samples and pre-training tokens)

Operations (reduced inference cost at scale)

End-2-End ChipNeMo Customization Workflow

Domain-specific models lead to higher accuracy and lower cost

Pretraining	Foundation Model	Domain Adaptive Pretraining	ChipNeMo Foundation Model	Supervised Fine-Tuning	ChipNeMo Chat Model
Trillions of tokens of internet data	Llama 2	24B tokens of chip design docs/code	(13B, 7B)	128K chat instructions + 1.1K task instructions	(13B, 7B)
10^5-10^6 GPU hrs	(70B, 13B, 7B)	~5000 GPU hrs		~100 GPU hrs	

Pre-trained Models → Training and Customization → Deployment

https://arxiv.org/abs/2311.00176

ChipNeMo Data Curation

Balanced datasets combining NVIDIA-proprietary chip design specific data and publicly available datasets

Data Source Type	Data Percentage (%)	Data Tokens (B)	Training Percentage (%)	Training Tokens (B)
Bug Summary	9.5%	2.4	10.0%	2.4
Design Source	47.0%	11.9	24.5%	5.9
Documentation	17.8%	4.5	34.0%	8.2
Verification	9.1%	2.3	10.4%	2.5
Other	7.9%	2.0	12.0%	2.9
Wikipedia	5.9%	1.5	6.2%	1.5
Github	2.8%	0.7	3.0%	0.7
Total	100.0%	25.3	100.0%	24.1

Breakdown of DAPT data for ChipNeMo after filtering (**24.1 billion tokens**)

Domain Source	Number of Samples
Design Knowledge	280
EDA Script Generation	480
Bug summarization and analysis	392
Total	1152

Breakdown of Domain SFT data (**128000 samples**)

Data relevance and quality > quantity

Data anonymization and privacy should be considered in dataset compilation

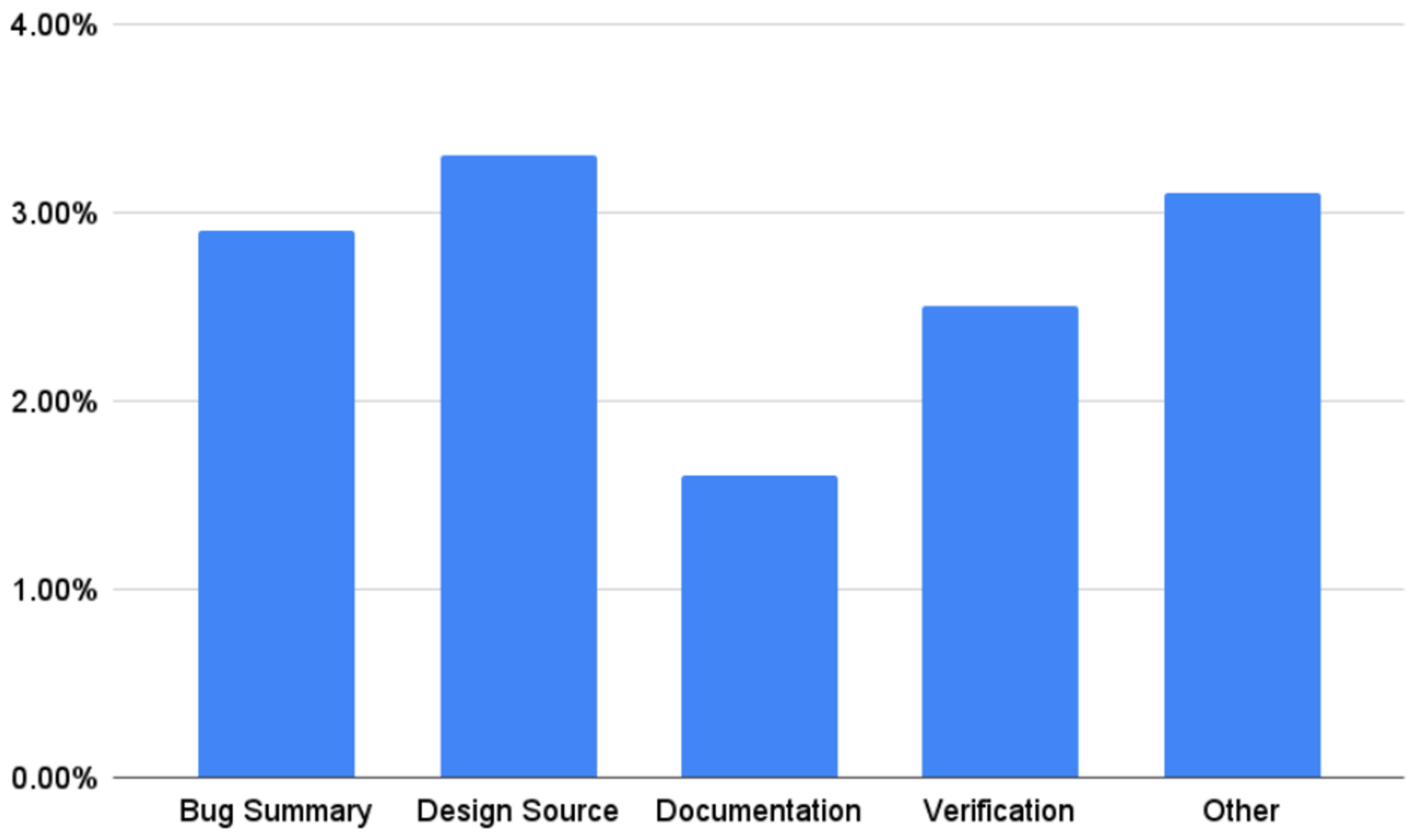
Continuous data updating process critical to keep the training set relevant

Data curation & management play important role

Domain-adaptive Foundation Model Pretraining

Custom Tokenization

ChipNeMo's tokenizer enhancements (**9k new tokens**) improved tokenization efficiency (**1.6% to 3.3% improvement**) across various design datasets without significant accuracy decline on public benchmarks

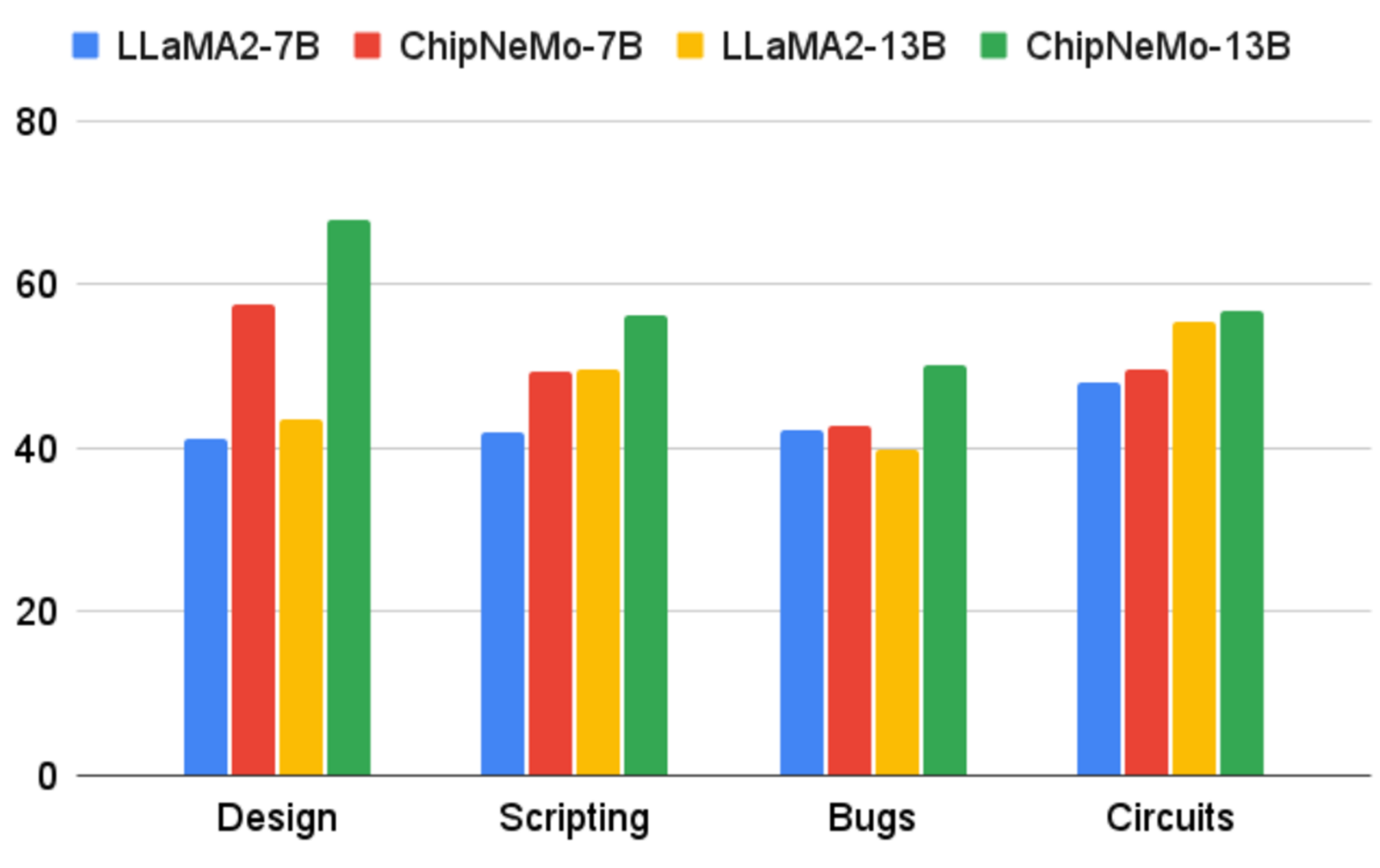


ChipNeMo Tokenizer Augmentation Improvements

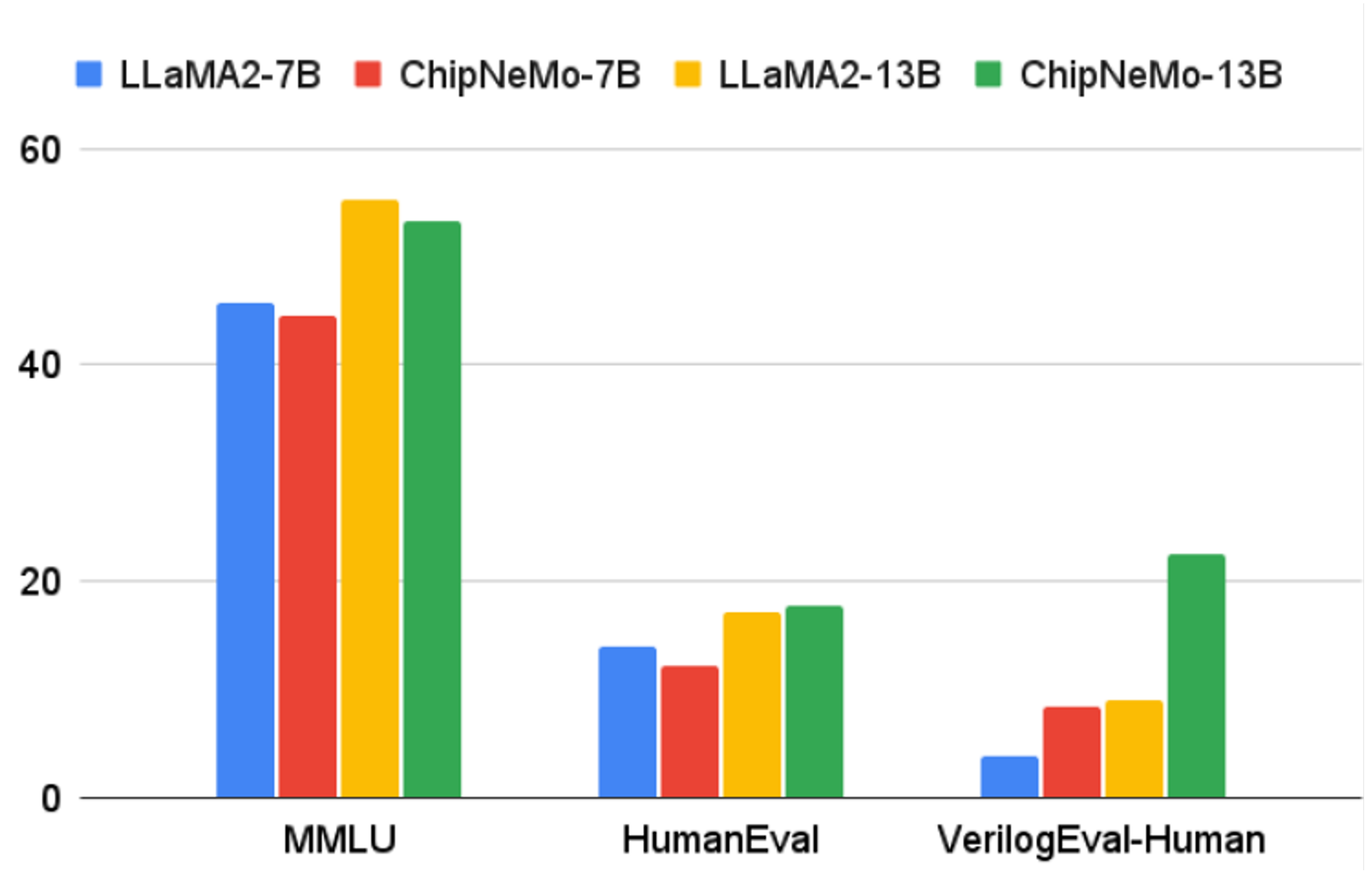
- Balance between generic language understanding and domain-specific nuances
- Iterative refinement of tokenizer based on feedback and model performance
- Collaboration between domain experts and ML engineers to identify critical tokens

Domain-adaptive Foundation Model Pretraining

ChipNeMo uses domain-adaptive pre-training to better understand chip design contexts



Chip Design Domain Benchmarks



Academic Benchmarks

Balancing between continuing pre-training and overfitting risks. Smaller learning rate plays a dual role.

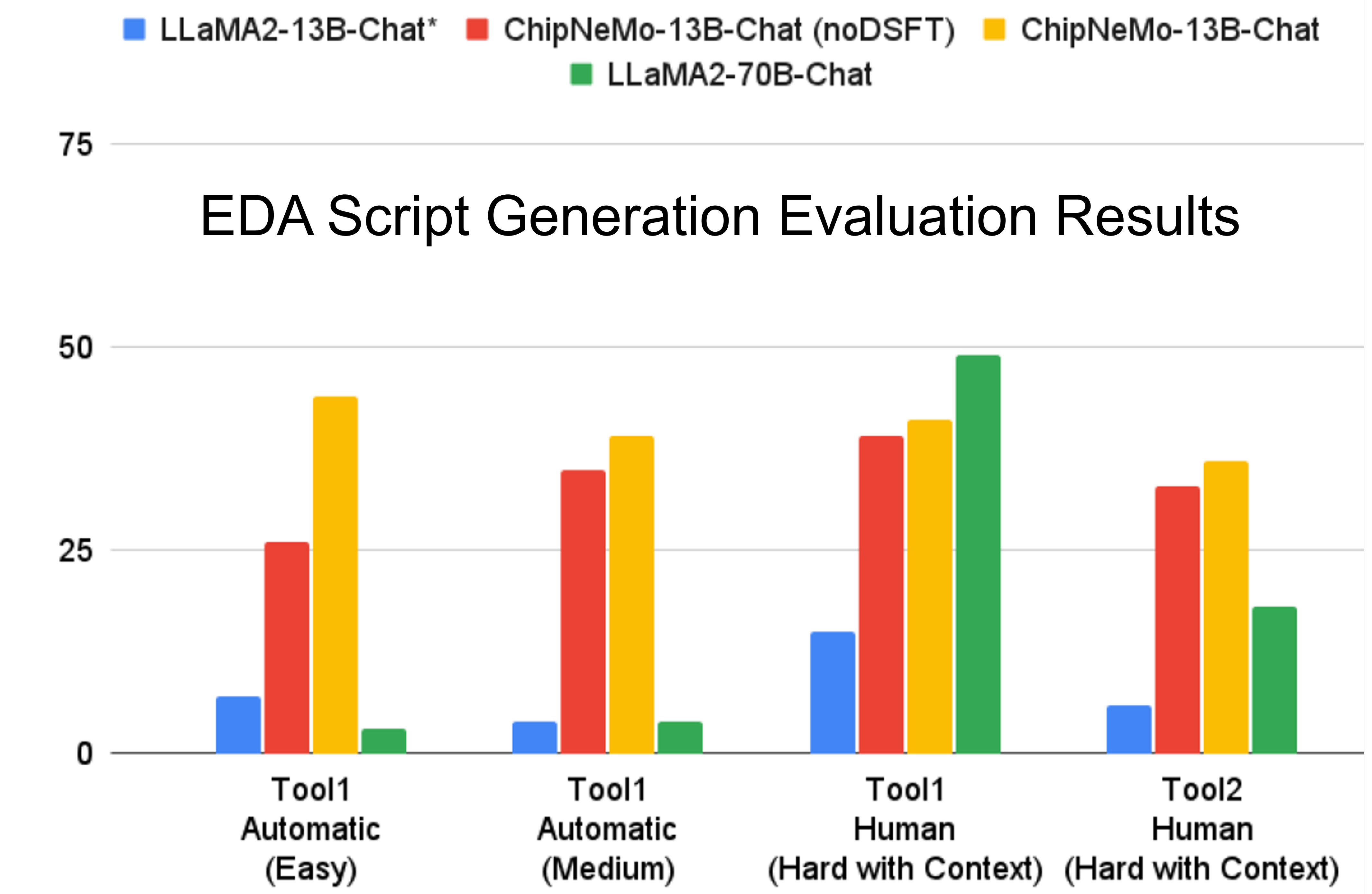
Larger and more performant foundational models yielded better zero shot results on domain-specific tasks.

Supervised Fine-Tuning

Customization of model behaviour for high performance on specific tasks

Model Size	Pretraining	DAPT	SFT
7B	184,320	2,620	90
13B	368,640	4,940	160
70B	1,720,320	-	-

Training cost of models in GPU hours.



ChipNeMo-Chat: Models fine-tuned with both domain and general chat data
ChipNeMo-Chat (noDSFT): Models fine-tuned with general chat data exclusively.

Importance of quality and relevance of labeled data for fine-tuning

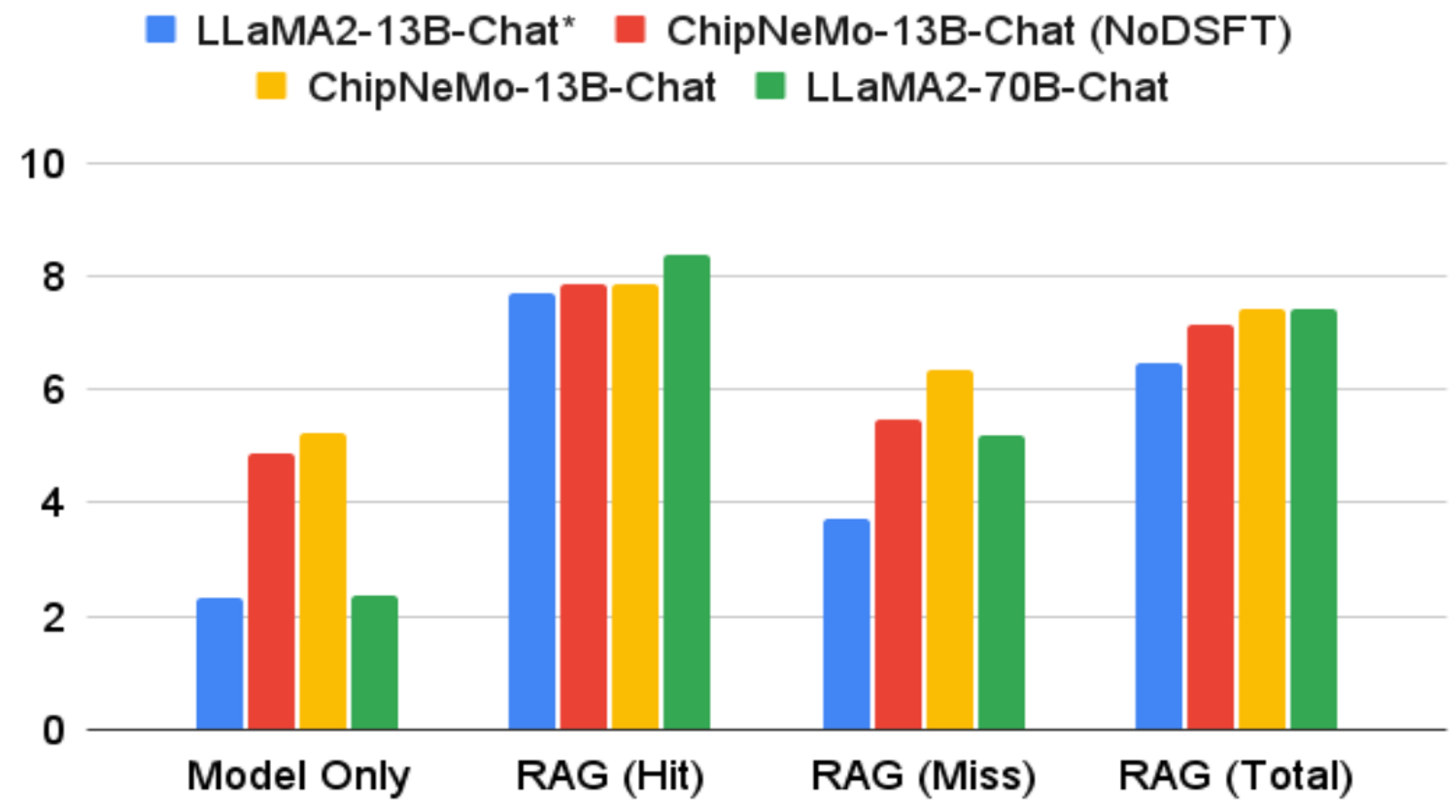
Adopt techniques for efficient fine-tuning without compromising model generalizability

Evaluation metrics tailored to specific tasks to gauge SFT success

⬆️
⬆️
⬆️ 0.79 out of 7 point scale

Retrieval-Augmented Generation (RAG)

Fine-tuning ChipNeMo retrieval model + domain-specific data improves the hit rate by 30% leading to better RAG



Human Evaluation of Different Models

Addition of in-domain context through **RAG** significantly boosts human scores

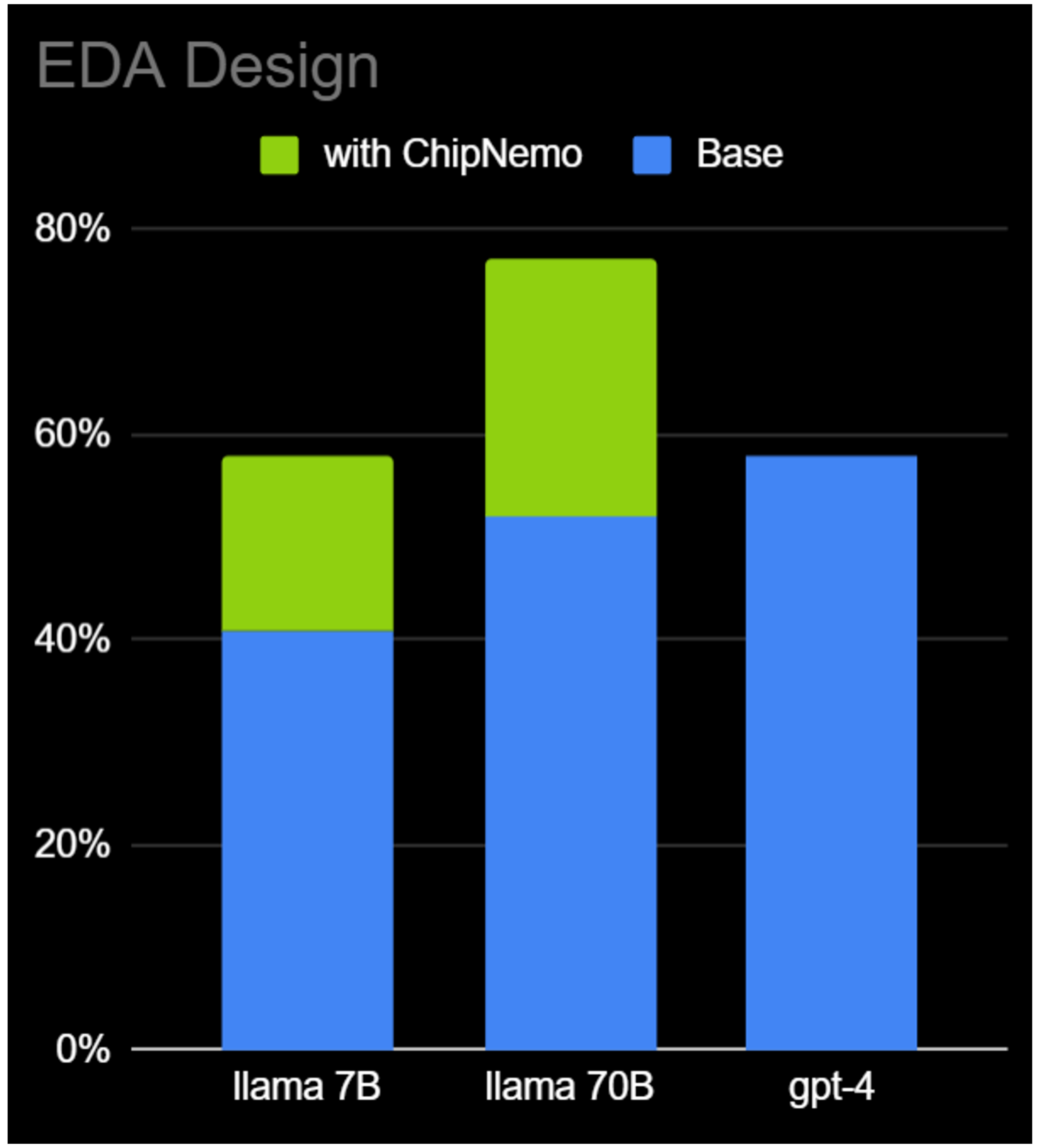
ChipNeMo (DAPT+SFT) models outperforms fine-tuned same size LLaMa chat model.

ChipNeMo-13b-Chat with RAG achieves same score as the 5X larger model LLaMA2-70B-Chat with RAG. **Domain adaptation however makes up for the misses.**

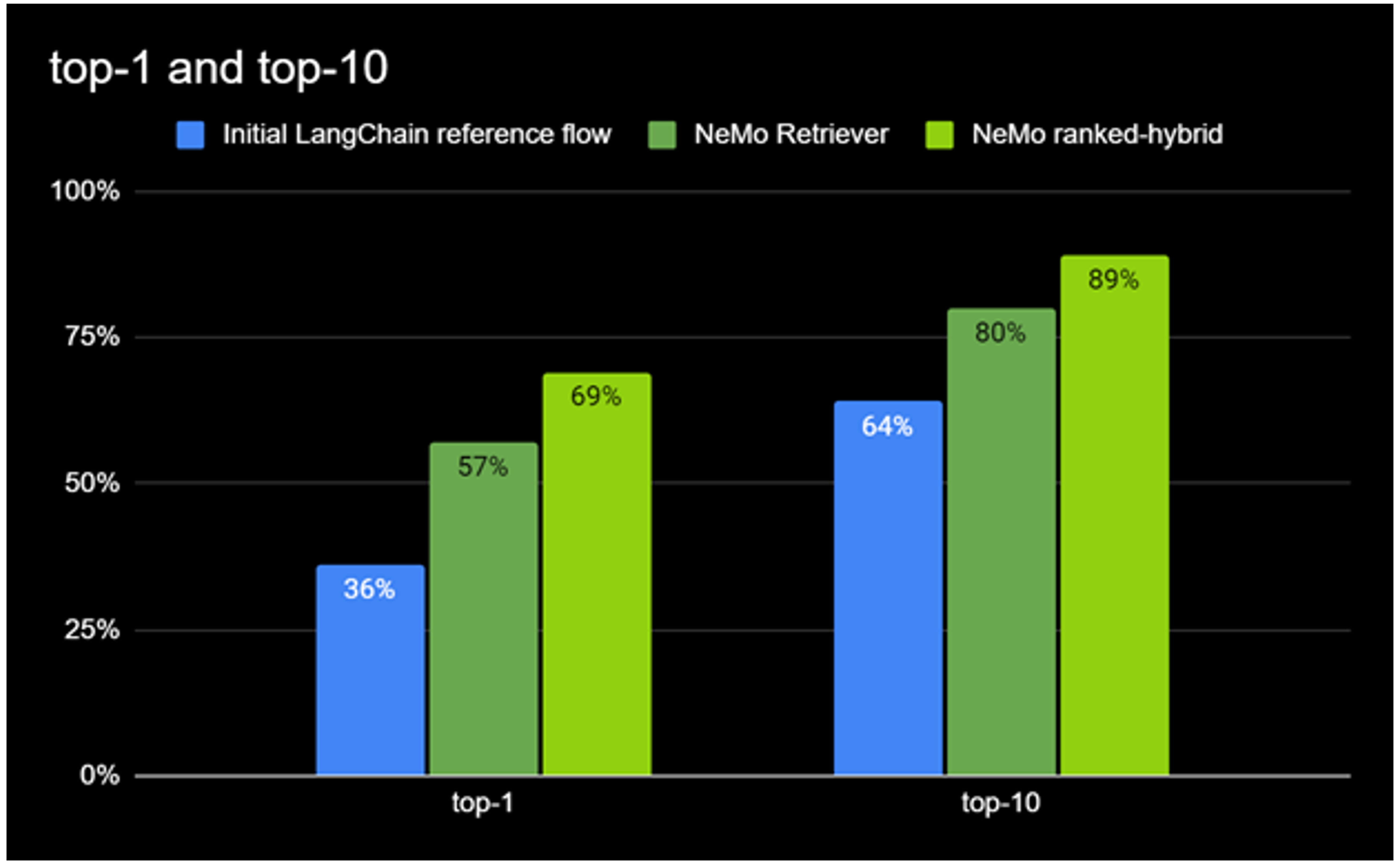
Domain SFT improves performance of ChipNeMo-13B-Chat with/without RAG.

Customization Lead to Large Performance Improvement

Domain-adapted ChipNeMo significantly outperforms OOTB solutions



Customized Llama-2 7B achieves GPT-4 accuracy, while Llama-2 70B demonstrates state-of-the-art results



Domain-specific embedding, ranking, and re-ranking models lead to higher context relevance resulting in more downstream customer value

Unified Stack to Accelerate Generative AI Adoption for Enterprises

Enabling end-2-end generative AI journey from data curation to model customization, optimization, evaluation, inference

NVIDIA NIM

Pre-trained & custom LLMs

NeMo Data Store

Prompts, responses, PII redaction, quality filtering

Adapters, P-tokens as .nemo checkpoints

Custom datasets, evaluation results

NeMo Curator

Scalable multi-stage curation of high-quality training and evaluation datasets for pre-training and fine-tuning data pipelines

RAPIDS

NeMo Customizer

State-of-the-art customization techniques with easy-to-use API for diverse data / compute scenarios balancing accuracy, latency, cost, skill level

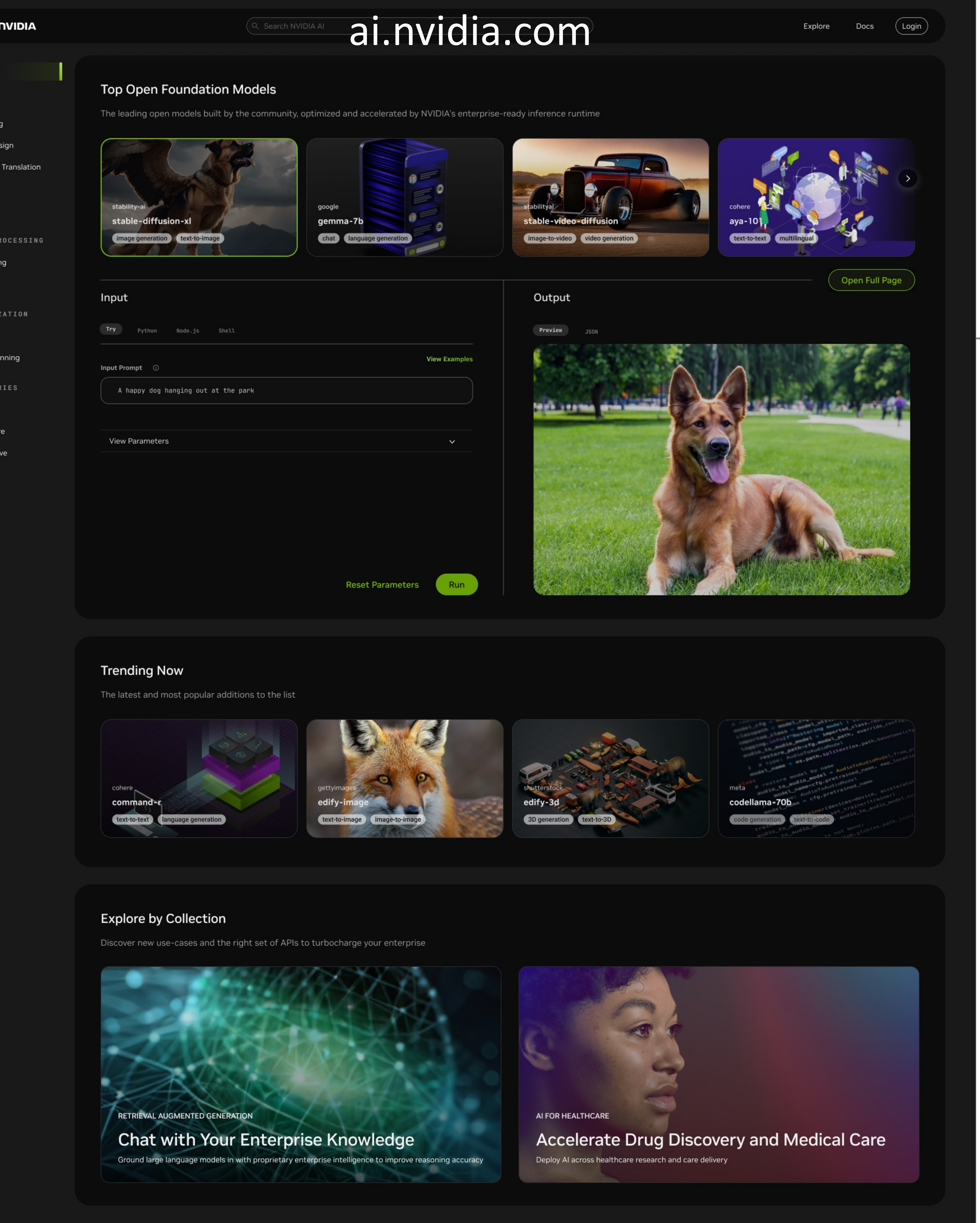
NeMo Framework

NeMo Evaluator

Automated evaluation of foundation models and fine-tuned LLMs on academic benchmarks and custom datasets using LLM-as-a-judge and pre-defined metrics

NVIDIA NIM

Microsoft Azure · aws · Google Cloud · ORACLE · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO



Resources to Get Started

- Explore **NVIDIA API Catalog**: https://ai.nvidia.com/
- **NVIDIA RAG:**
  - https://build.nvidia.com/explore/retrieval
  - https://github.com/NVIDIA/GenerativeAIExamples
- **NeMo Microservices:**
  - Apply for Early Access: developer.nvidia.com/nemo-microservices-early-access
  - https://developer.nvidia.com/docs/nemo-microservices/index.html

Get Started with NeMo

Download Now - Language

Apply Now - Multimodal

Web Pages

- NVIDIA Generative AI Solutions
- NVIDIA NeMo Framework
- NeMo Guardrails TechBlog

Blogs

- What are Large Language Models?
- What Are Large Language Models Used For?
- What are Foundation Models?
- How To Create A Custom Language Model?
- Adapting P-Tuning to Solve Non-English Downstream Tasks
- NVIDIA AI Platform Delivers Big Gains for Large Language Models
- The King's Swedish: AI Rewrites the Book in Scandinavia
- eBook Asset
- No Hang Ups With Hangul: KT Trains Smart Speakers, Customer Call Centers With NVIDIA AI

Webinars

- Learn more about LLM Application Development
- How to Build Generative AI for Enterprise Use-cases
- Leveraging Large Language Models for Generating Content
- Power Of Large Language Models: The Current State and Future Potential
- Generative AI Demystified
- Efficient At-Scale Training and Deployment of Large Language Models – GTC Session
- Hyperparameter Tool GTC Session