



# NVIDIA NeMo Framework

Deep Dive

22<sup>nd</sup> May 2023

Ettikan Kandasamy Karuppiah (Ph.D)

Director/Technologist, Asia Pacific South Region





# Agenda

- The Opportunity of LLMs

---
- NeMo Framework Overview

---
- Latest advancements in NeMo Framework

---
- NeMo Framework Availability

---
- Customer Using NeMo Framework Today

---



The background features a dark field with numerous thin, bright green lines that curve and flow across the frame, creating a sense of motion and digital connectivity. On the far left, there is a solid, vertical green bar. Overlaid on this background is the text 'The Opportunity of Generative AI' in a white, bold, sans-serif font.

# The Opportunity of Generative AI




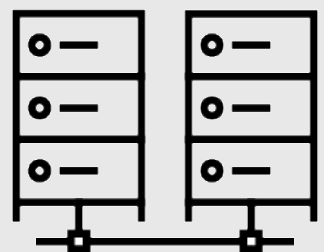

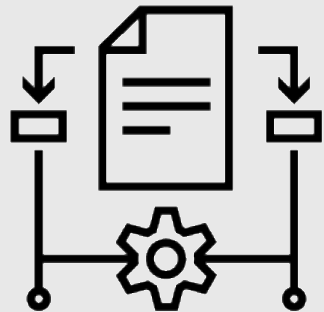


# Challenges Building Generative AI for the Enterprise

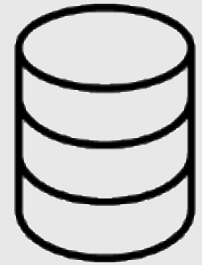
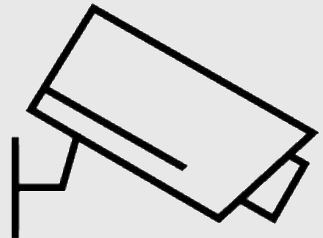
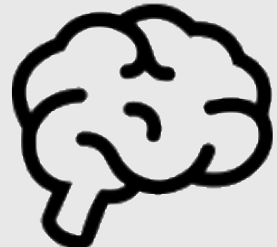

## LLM Enterprise Use Cases and Goals:

MAINTAIN & UTILIZE COMPANY KNOWLEDGE 	QUESTION & ANSWER 	CUSTOMER SERVICE ASSISTANT 	SUPPLY CHAIN FORECASTING 	SALES PIPELINE ANALYSIS 	FINANCIAL MODELING 	LEGAL CONTRACT DISCOVERY 
---	--	---	---	--	---	---

## Challenges of Building Foundation Models

	Massive Training Datasets
	Large-scale compute infrastructure for training & inferencing, costing upwards of \$XX M
	Deep technical expertise
	Algorithm Selection and complex experimentation to achieve convergence

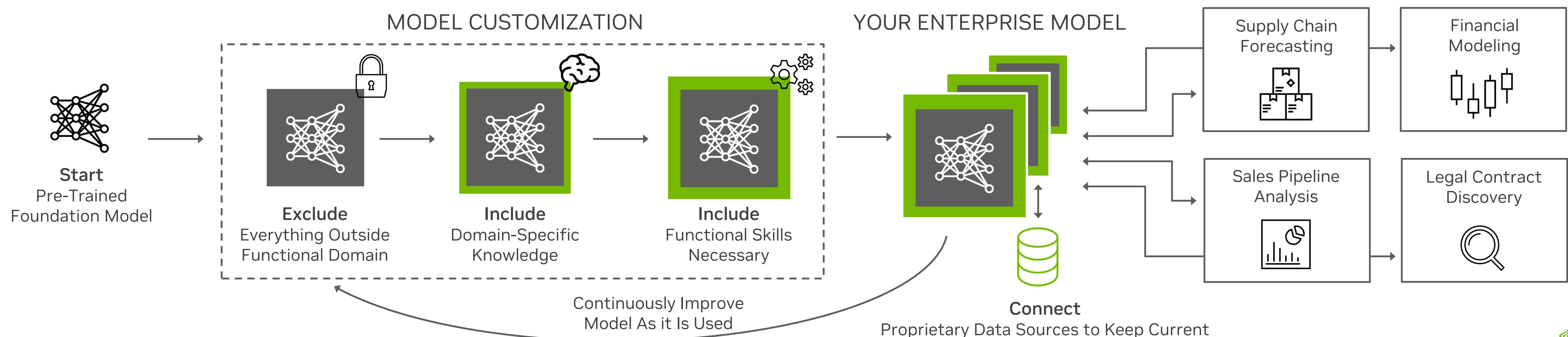
## Challenges of Using Foundation Models

	Trained on Publicly available information and datasets
	Outdated Information, as Models are Frozen in Time.
	Hallucination
	Bias & Toxic Information

# NVIDIA Provides the Tools to Overcome LLM Challenges

## Using Foundation Models

- **Generalized AI does not achieve Enterprise needs as it lacks domain knowledge and can have non-factual responses.**
- Model customization is key to enable **inclusion** of domain specific knowledge & proprietary information, and **exclusion** of unwanted information or responses.
- NVIDIA NeMo LLM enables:
  - **Functional Skills:** Specialized skills to solve customer and business problems.
  - **Focus with Guardrails:** Exclude everything outside functional domain, eliminate bias and toxicity, align to human intentions.
  - **Domain Specific Knowledge:** Encode and embed your enterprise's real-time information to provide the latest responses.
  - **Continuous Improvement:** Reinforcement Learning with Human Feedback techniques allow for your enterprise model to get smarter over time, aligned to your specific enterprise domain



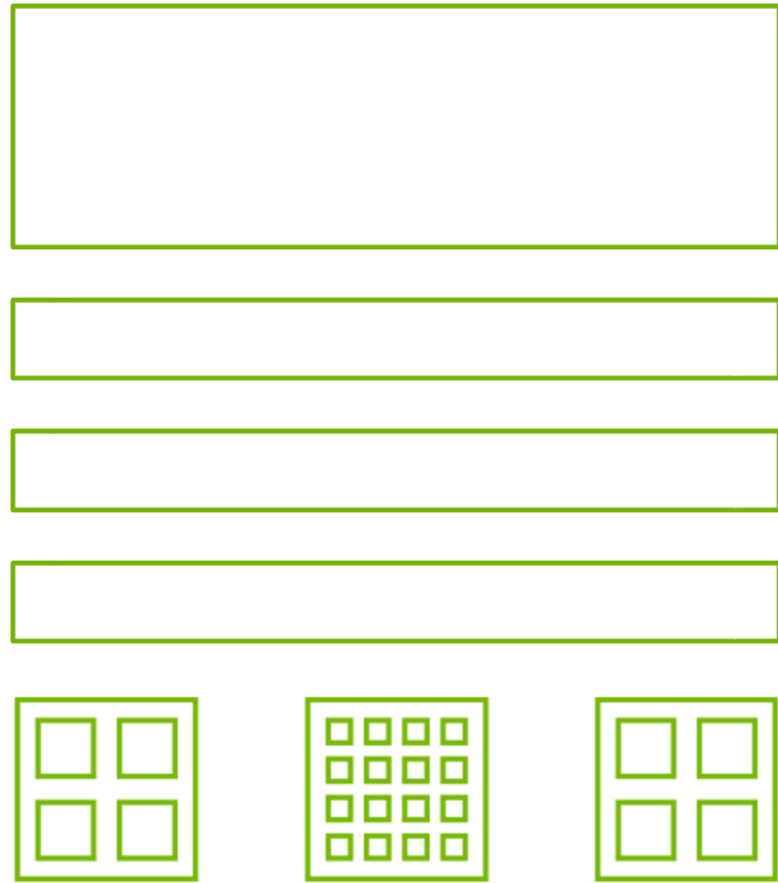


Full Stack, 3 Chips, Data Center Scale

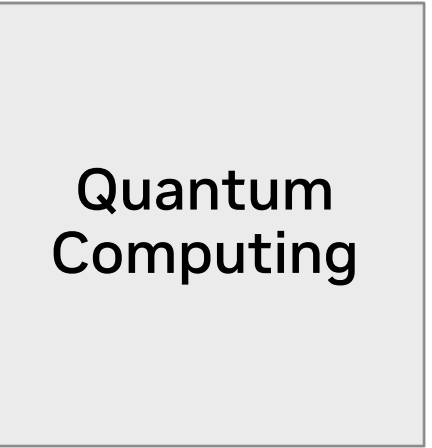
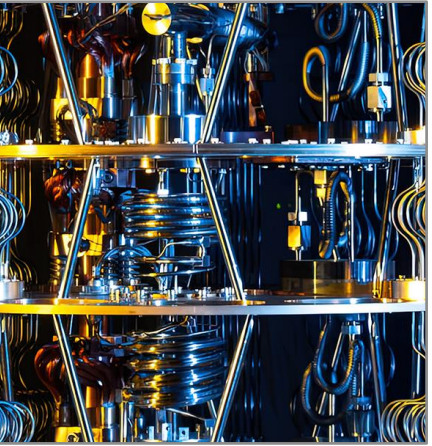
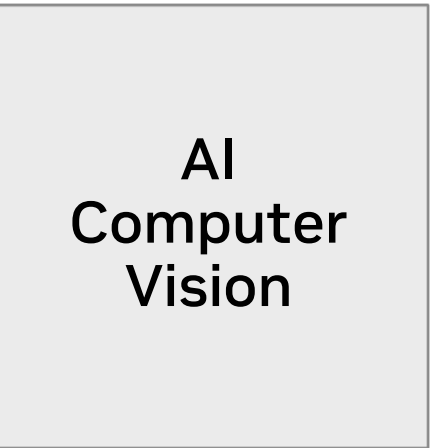
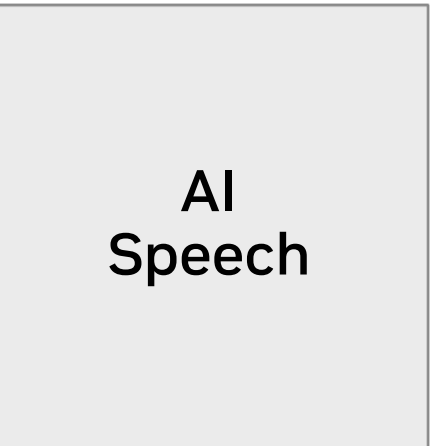
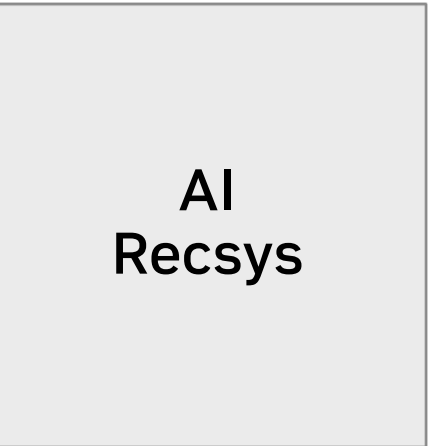
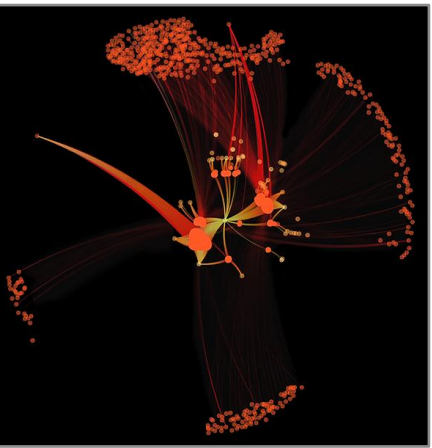
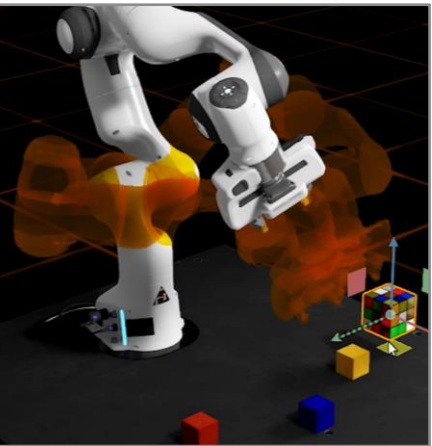
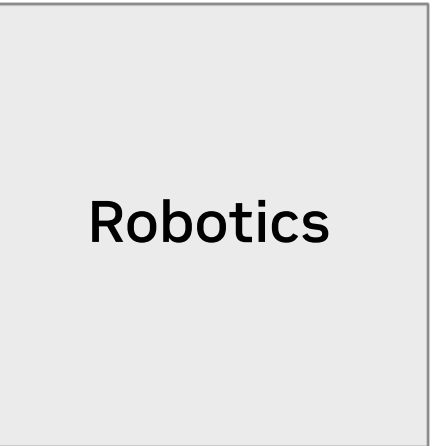
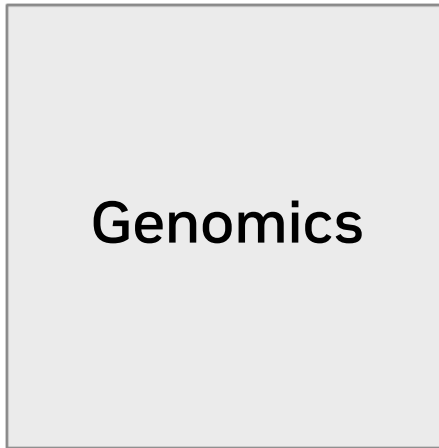
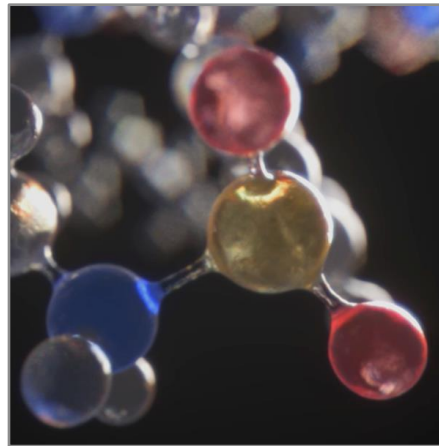
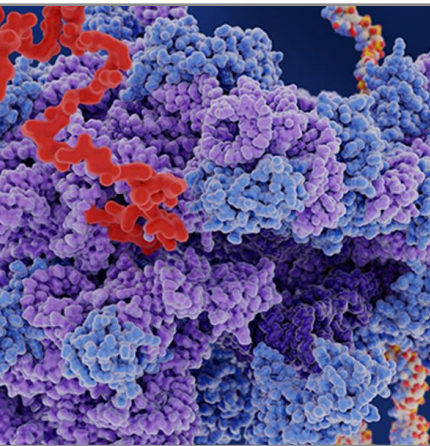
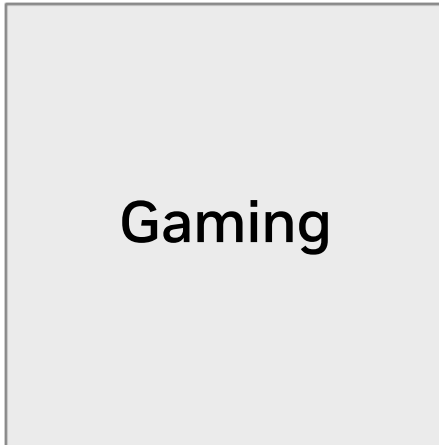
33 Million CUDA Downloads

150 SDKs

\$100 Trillion Industry Served

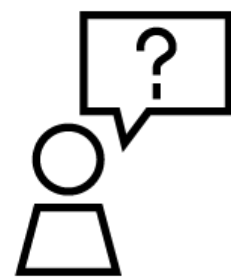


NVIDIA Accelerated Computing





# Generative AI Unlocks New Opportunities



How has NVIDIA contributed to acceleration of AI?

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.



530B

*\*\*Generated using NVIDIA NeMo service*

## TEXT GENERATION



Summarization

GPT-3

Marketing Copy

## TRANSLATION

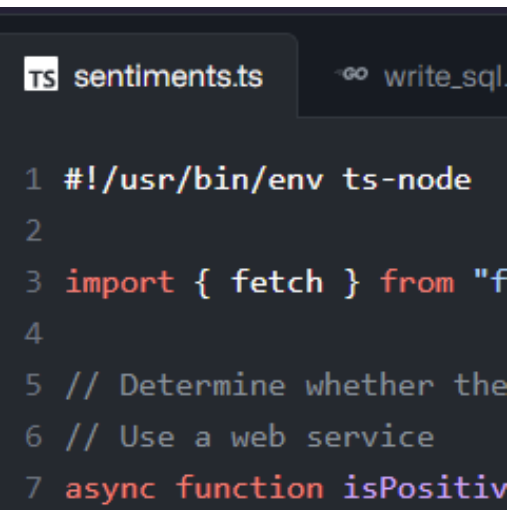


Translating Wikipedia

NLLB-200

Real-Time Metaverse Translation

## CODING



Dynamic Code Commenting

CODEX

Function Generation

## IMAGE GENERATION

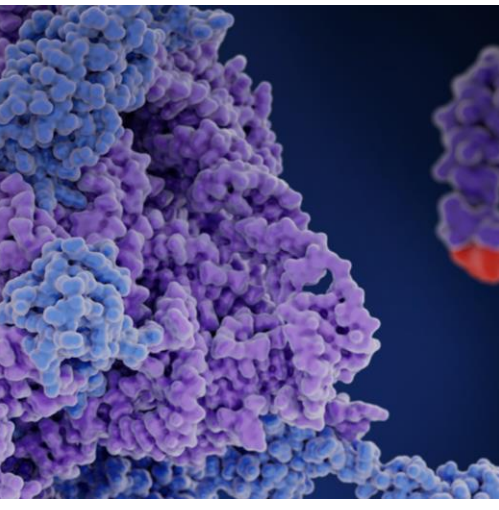


Brand Creation

e-Diffi

Gaming Characters

## LIFE SCIENCE

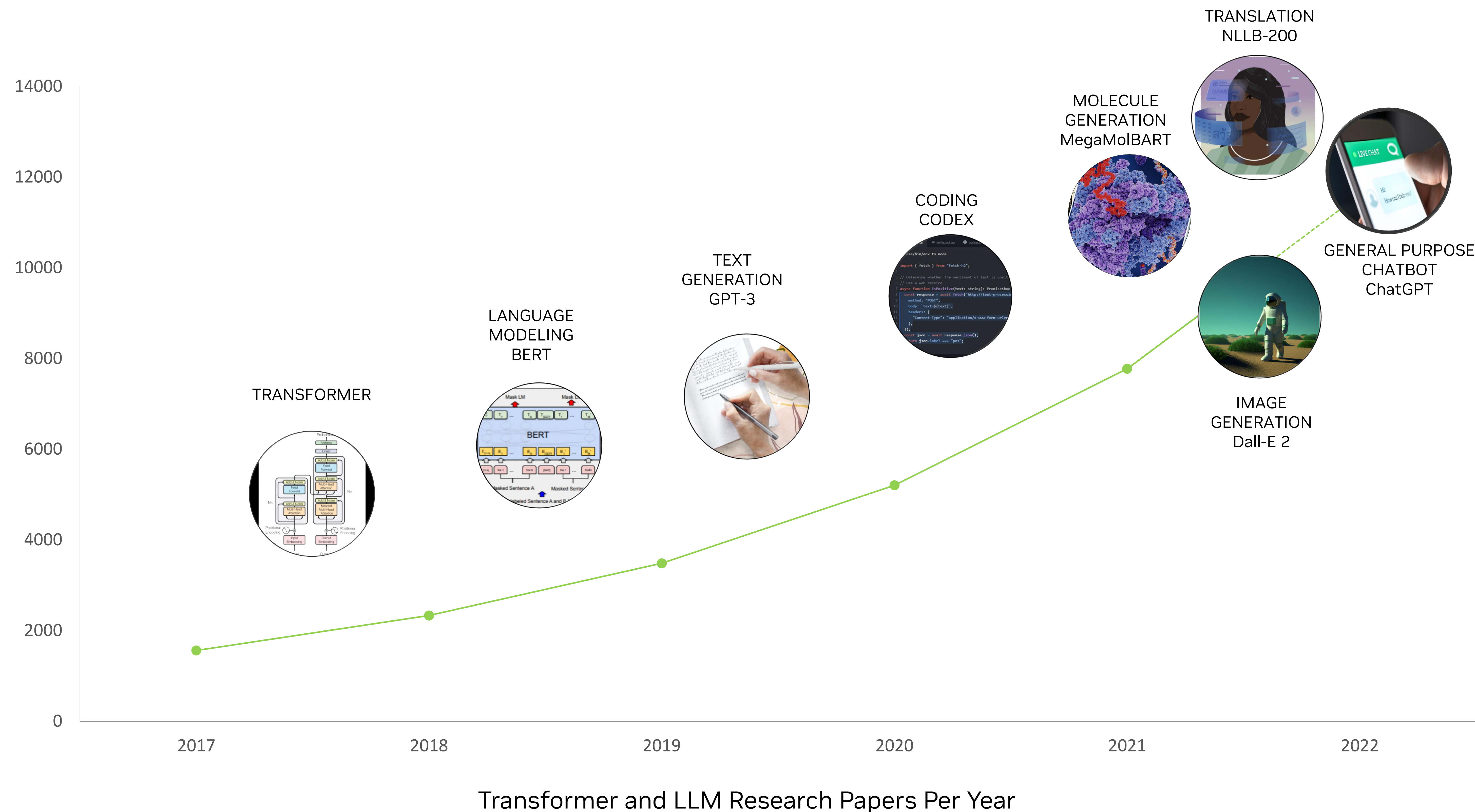


Molecular Representations

MegaMolBART

Drug Discovery





# Transformer Models Making Waves

Few of the most significant milestones in LLMs shaping industries



The background features a complex pattern of glowing green lines and shapes on a black field. On the left, numerous thin, parallel green lines radiate outwards. On the right, there are larger, more intricate structures resembling stylized leaves or petals, composed of many fine, overlapping green lines that create a sense of depth and movement.

# NeMo Framework Overview



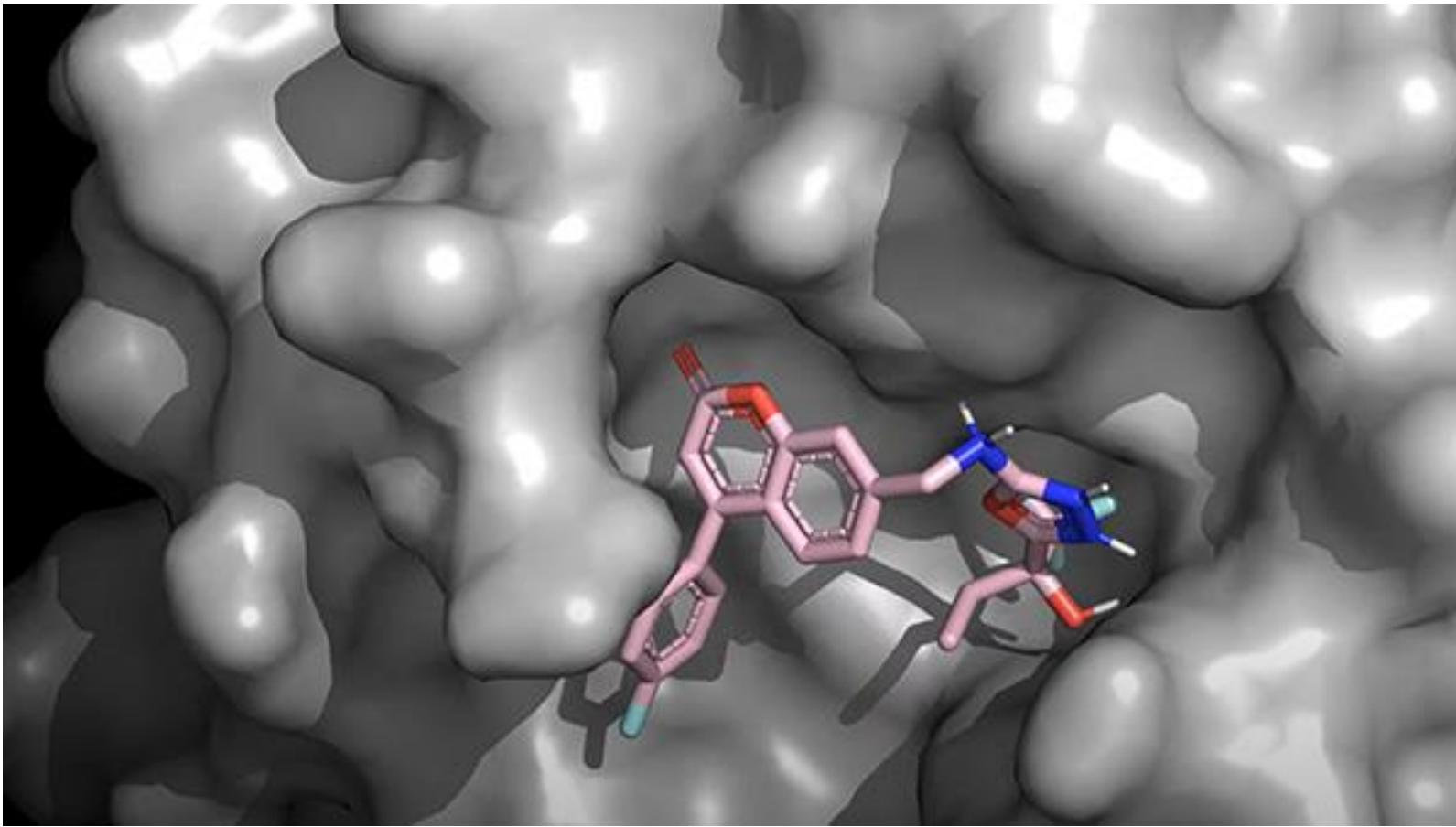
# NVIDIA's Generative AI Solutions

Foundations to Build and Run Your Generative AI

## NVIDIA NeMo service



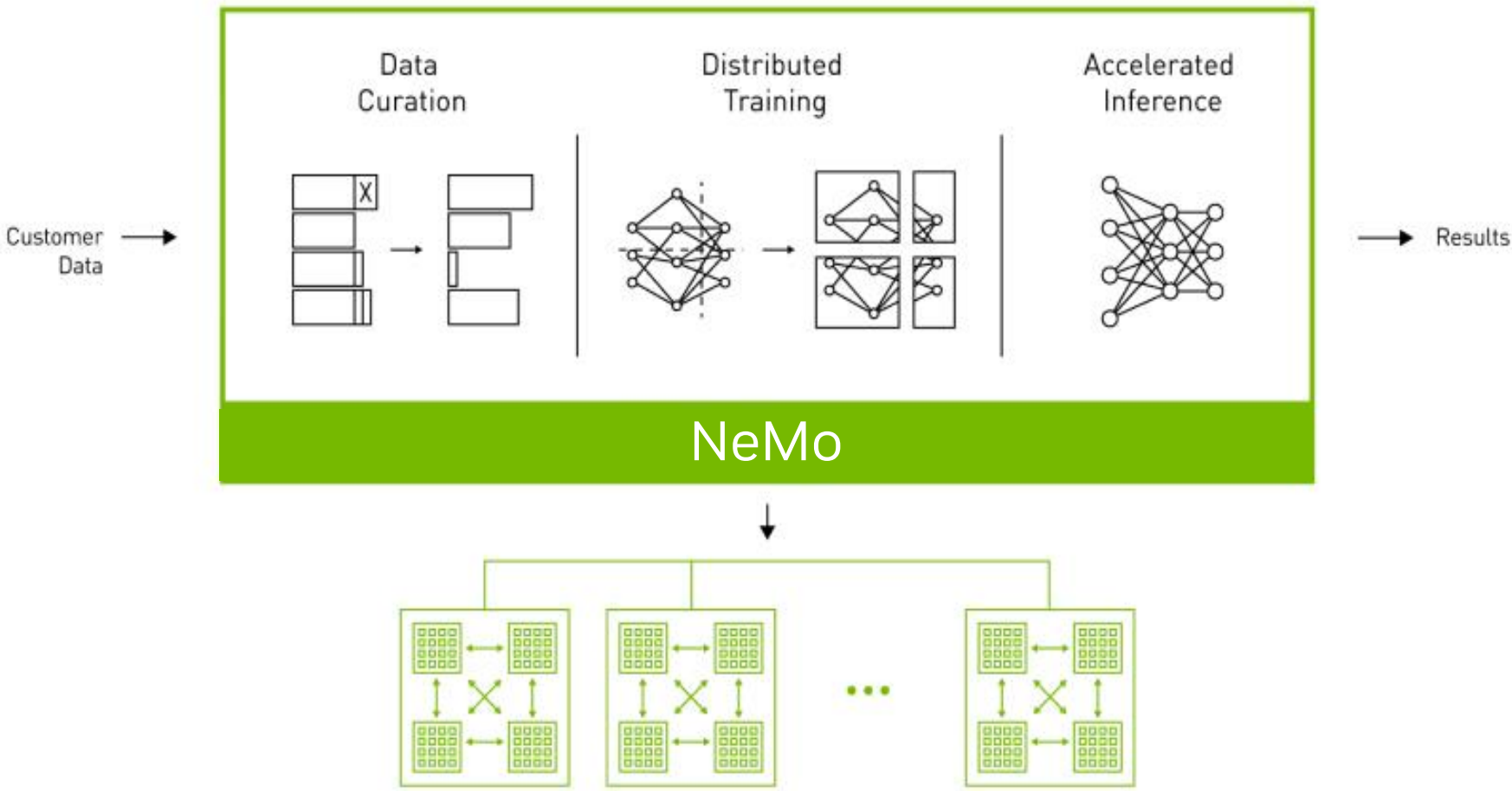
## NVIDIA BioNeMo service



## NVIDIA Picasso service



## NVIDIA NeMo framework



NVIDIA AI Foundations

NVIDIA AI Enterprise

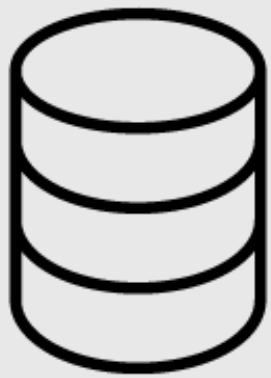
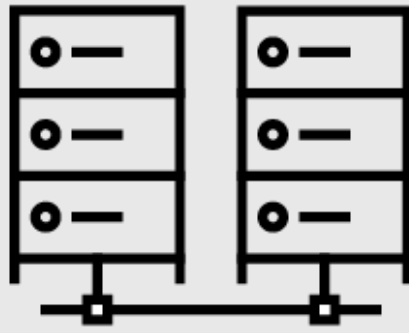

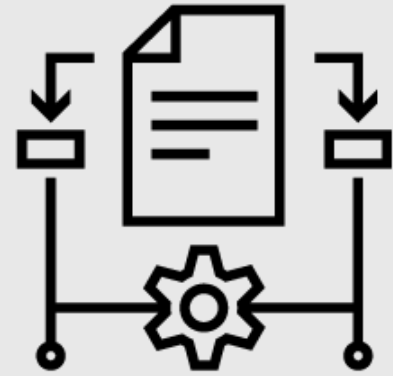
NVIDIA DGX Cloud



# Enterprise Challenges Of Developing Generative AI

Foundation models are AI neural networks trained on massive unlabeled datasets to handle a wide variety of tasks

## Challenges of Building Foundation Models

	Mountains of Training Data
	Large-scale compute infrastructure for training & inferencing, costing \$10 M+ in just cloud costs
	Deep technical expertise
	Complex algorithms to build on large-scale infrastructure

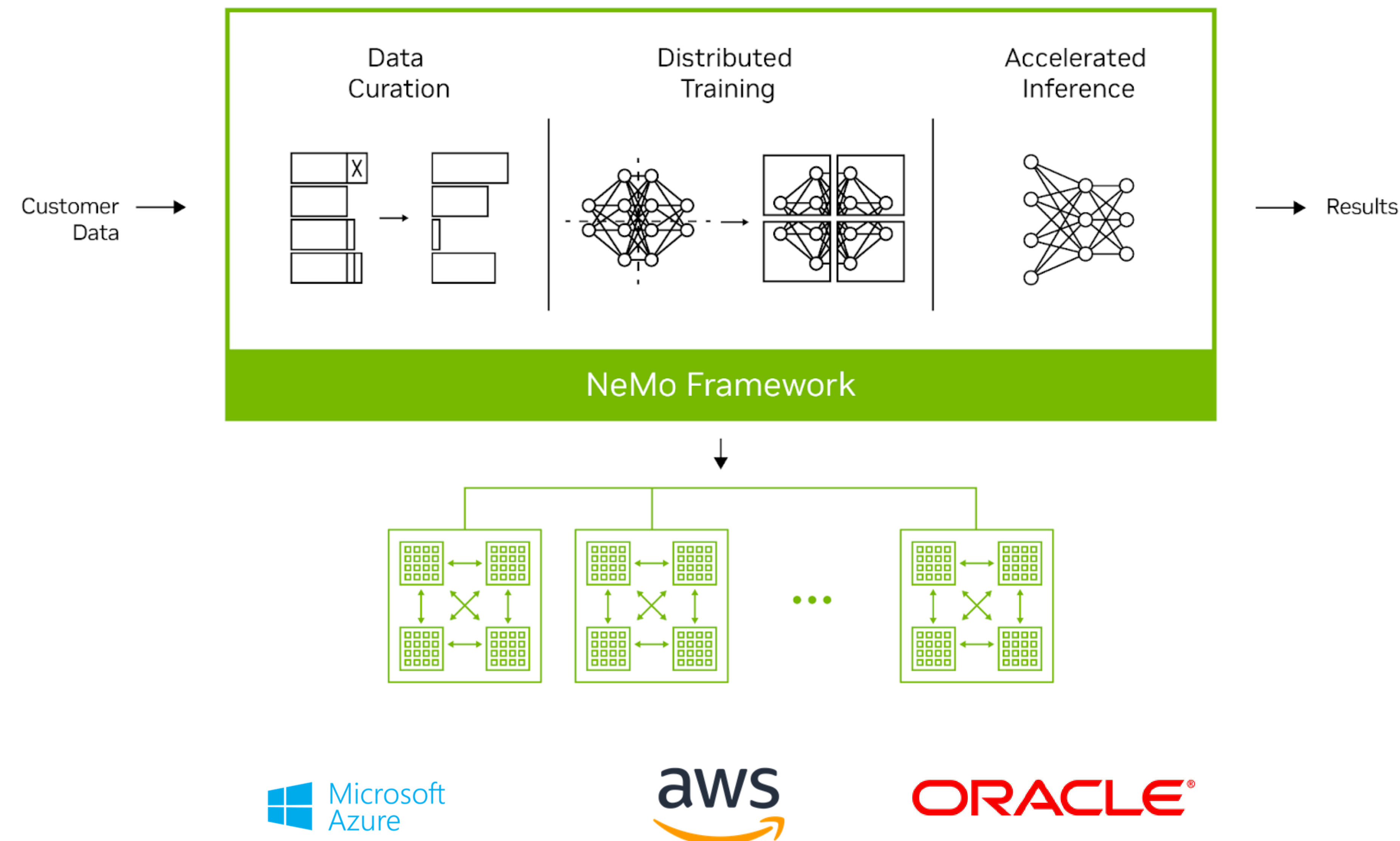
## Challenges of Using Foundation models

	Don't contain domain / enterprise specific knowledge
	Frozen in Time
	Hallucinate and provide undesired information
	Bias & Toxic Information



# NeMo Framework

An end-to-end, cloud-native enterprise framework to build, customize and deploy generative AI models



## Multi-modality support

Build language, image, generative AI models

## Accelerated Workflow

Speed up workflows with 3D parallelism & distributed training and inference techniques

## Data Curation

Mine and curate high-quality training data @ scale

## Customize Foundation Models

State of the art customization techniques for LLMs including Adapters, RLHF, AliBi, SFT

## Support

NVIDIA AI Enterprise keep projects on track

## Deploy Anywhere

On any NVIDIA accelerated system: NVIDIA DGX Cloud, major CSPs (Azure, AWS, OCI), or on-prem



### Unmet Needs

Large-Scale Data Processing

Multilingual data processing & training

Finding optimal hyperparameters

Convergence of Models

Scaling on Clouds

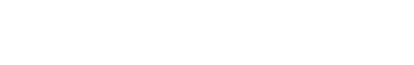
Deploying for inference

Deployment at-scale

Evaluating models in industry standard benchmarks

Differing infrastructure setups

Lack of Expertise



### NeMo addressing needs...

Data Curation & Preprocessing Tools

Relative Positional Embedding (RPE) – Multilingual Support

Hyperparameter Tool

Verified recipes for large GPT & T5-style models

Scripts/configs to run on Azure, OCI, and AWS

Model navigator + export to FT functionalities

Quantization to accelerate inferencing

Productization evaluation harness

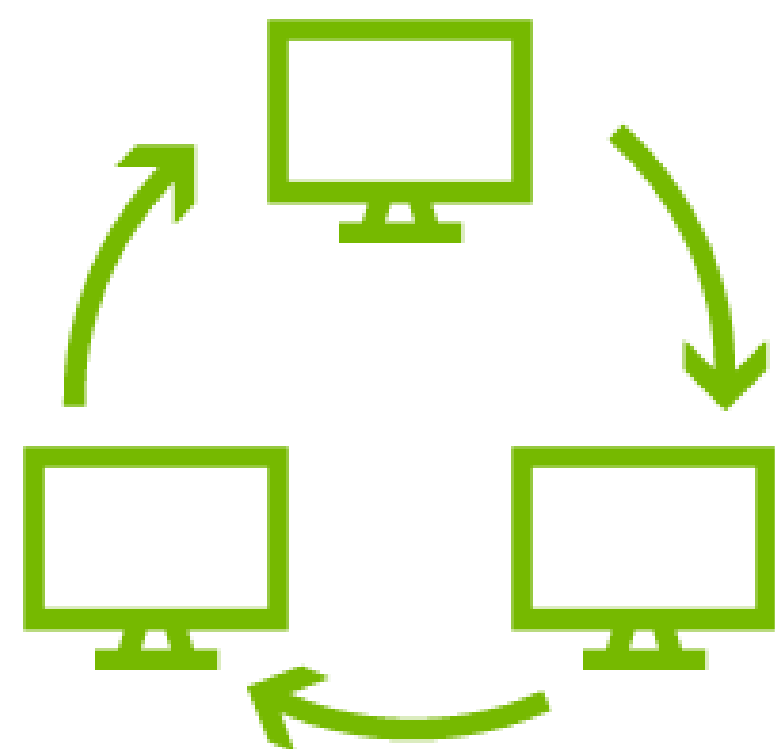
Full-Stack support with FP8 & Hopper Support

Documentation

**Solving pain-points across the stack**



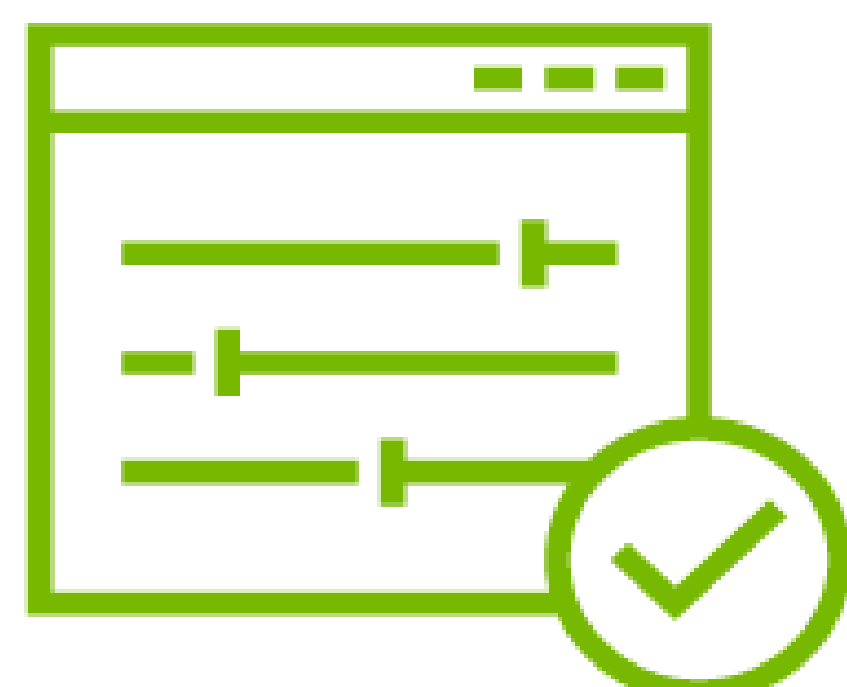
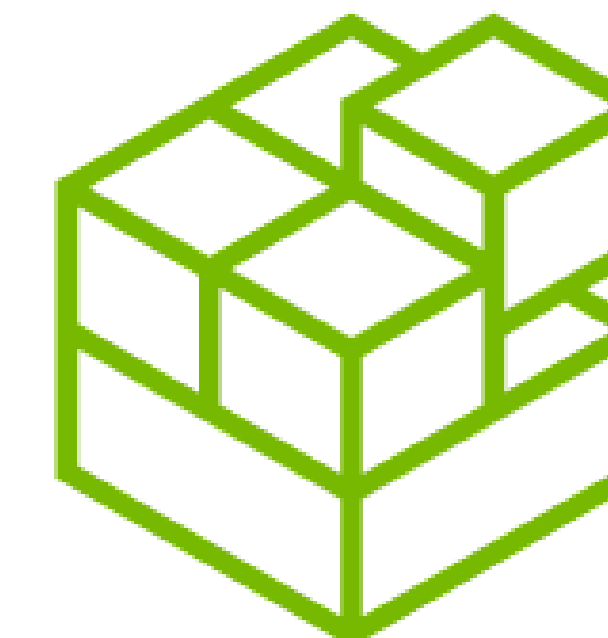
**End-to-End**  
Bring your own data, train & deploy LLM



**Fastest Performance at-Scale**  
SOTA training techniques and tools



**Easy-to-Use**  
Containerized framework



**Fully Flexible**  
Open-source approach



**Run Anywhere**  
Train & deploy on your choice of infrastructure



**Battle-Hardened**  
Verified recipes to work OOTB

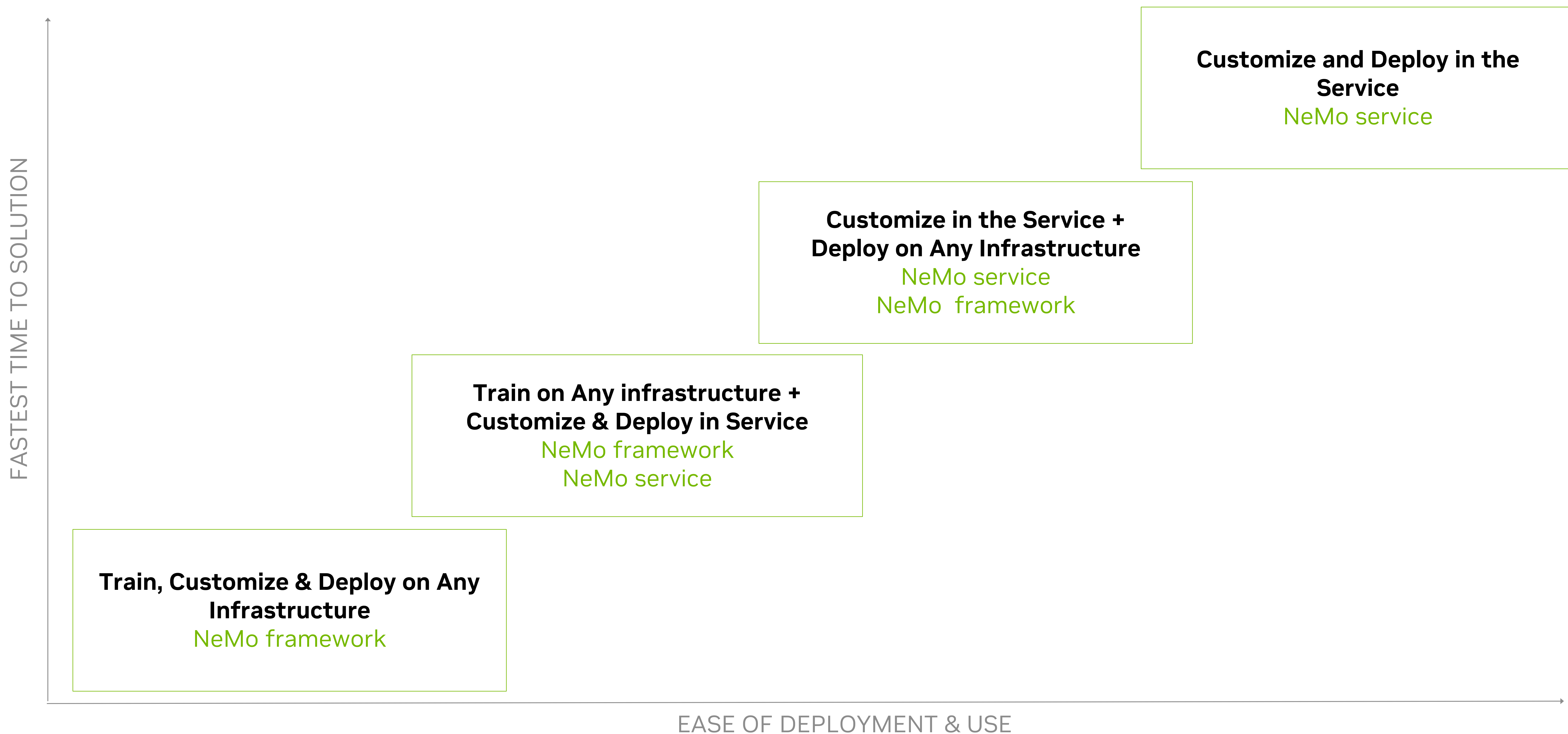
# NeMo Framework

Simplifying and accelerating the path to build and deploy large-scale generative AI models



# Enabling Enterprises Create Their Own Generative AI Models

Abstracting Complexities of Deep Learning Expertise, MLOps & Large-Scale Infrastructure





The background features a complex pattern of glowing green lines and shapes against a solid black field. On the left, numerous thin, parallel green lines radiate outwards. On the right, there are larger, more intricate structures resembling stylized leaves or overlapping rectangular frames, composed of many fine green lines that create a sense of depth and movement.

# NeMo Framework – Deep Dive



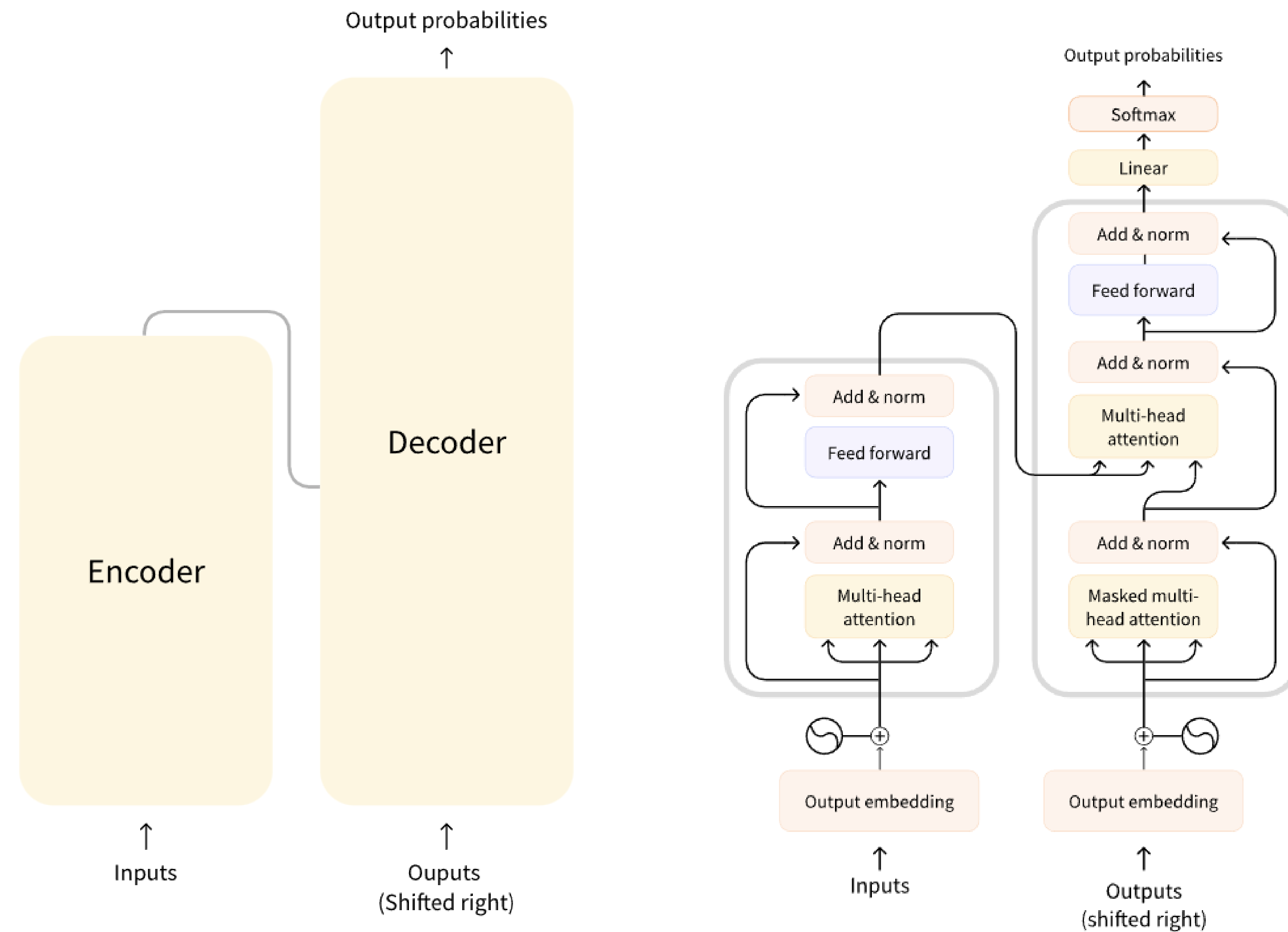
# When Large-Language-Models Make Sense

	Traditional NLP Approach	Large Language Models
Requires labelled data	Yes	No
Parameters	100s of millions	Billions to trillions
Desired model capability	Specific (one model per task)	General (model can do many tasks)
Training frequency	Retrain frequently with task-specific training data	Never retrain, or retrain minimally

- Zero-Shot (or Few Shot Learning)
  - Painful & Impractical to get a large corpus of labelled data
- Models can learn new tasks
  - If you want models with “common sense” and can generalize well to new tasks
- A single model can serve all use-cases
  - At-scale you avoid costs and complexity of many models, saving cost in data curation, training, and managing deployment



## ARCHITECTURE



- A **transformer** is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.
- Introduced in Attention Is All You Need
- Based on Encoder-Decoder Architecture, wherein encoder understands language, whilst decoder generates language

# Transformers

The Next Wave of AI



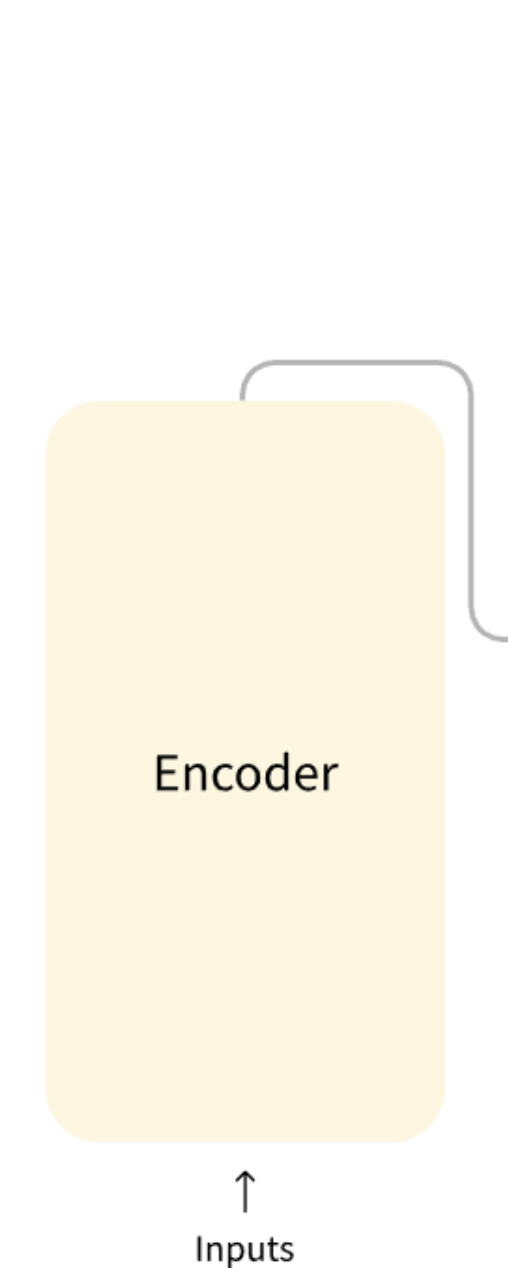
## Encoders

For Understanding Language

Suited for task requiring an understanding of the full sentence, such as sentence classification, named entity recognition, and extractive question answering.

Supported Models:

- BERT



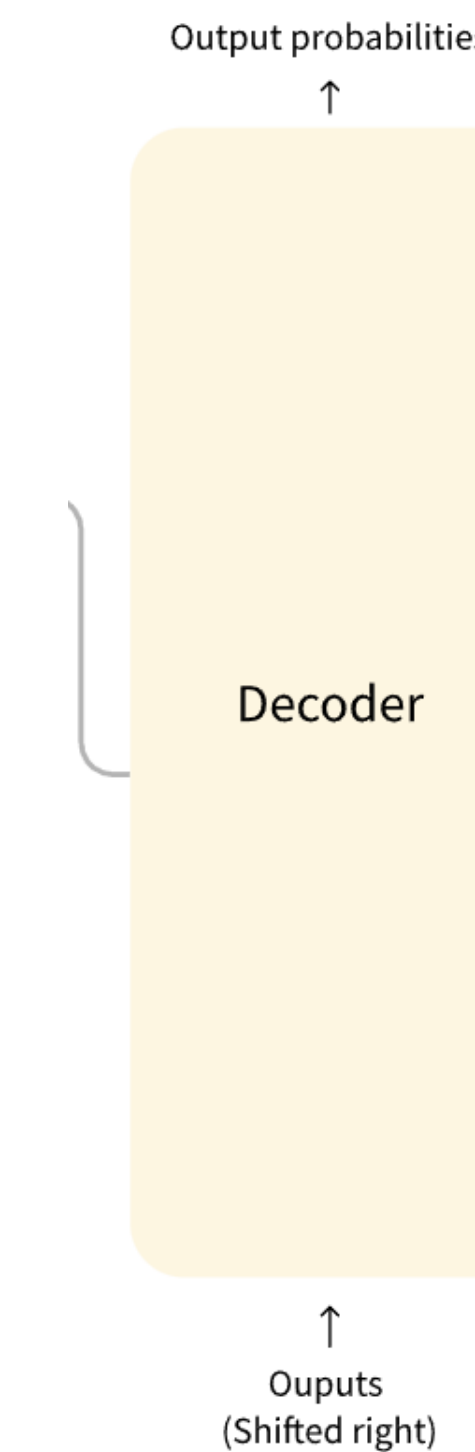
## Decoders

For Generative Models

Suited for tasks involving Text Generation

Supported Models:

- GPT-3



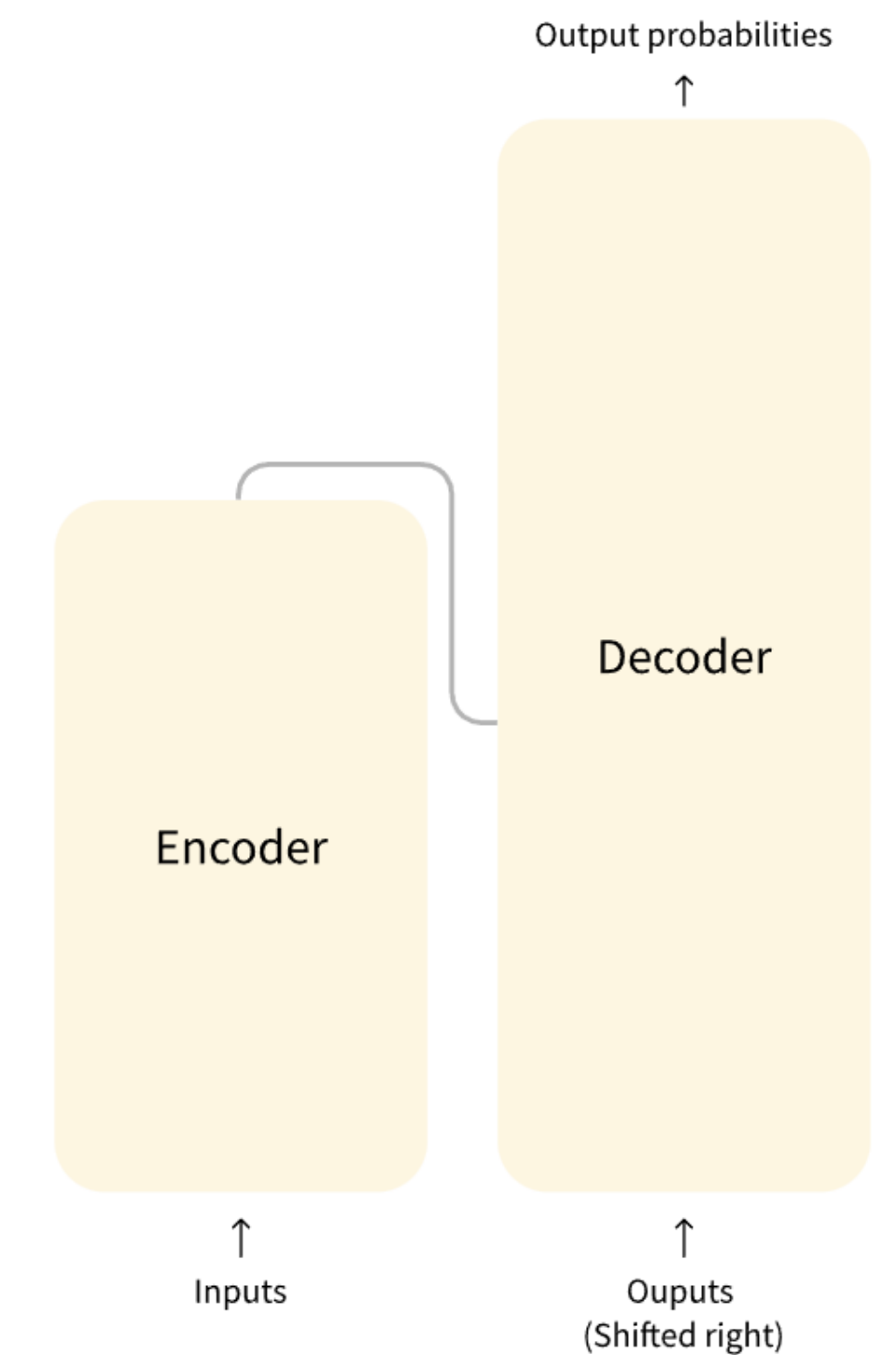
## Encoder-Decoders

Sequence-to-Sequence

Suited for tasks around generating new sentences depending on a given input, such as summarization, translation, or generative question answering.

Supported Models:

- T5
- Multilingual – mT5

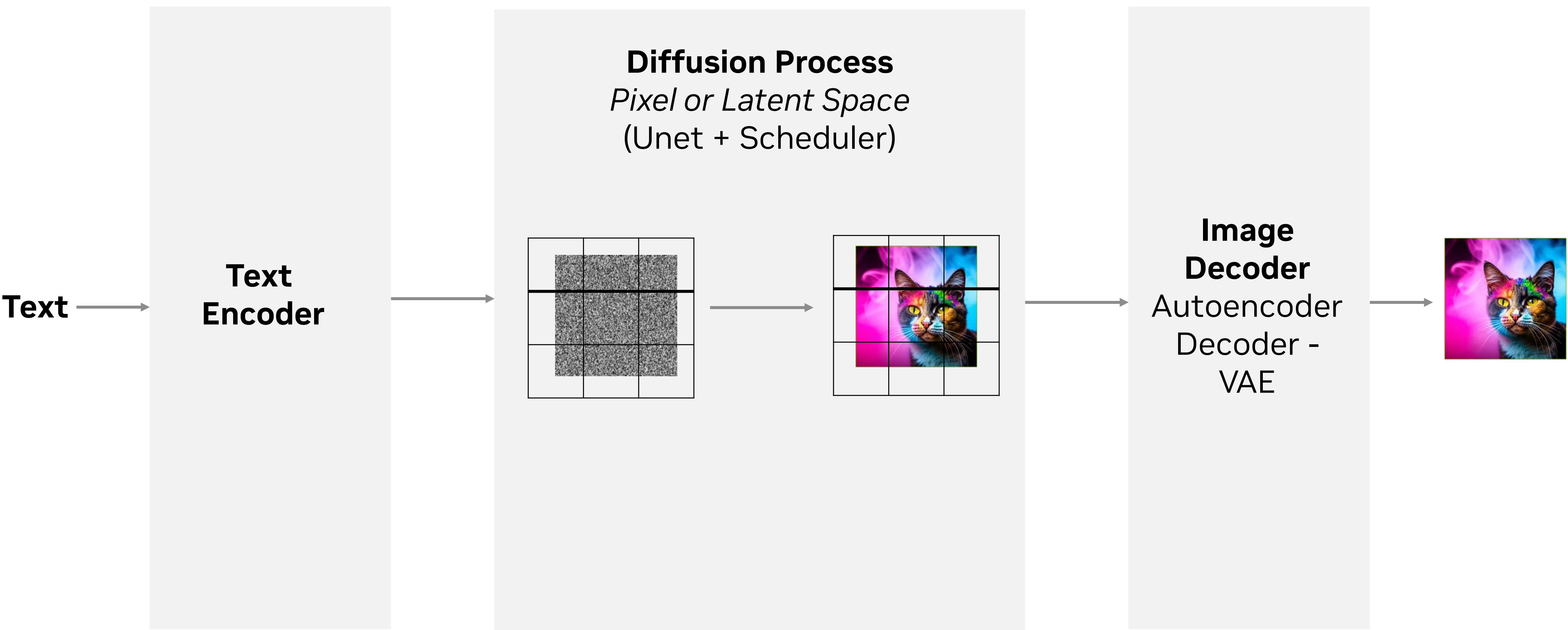


# Supported Language Models



# Generative Image Models

Text to Image Generative Models



## Supported Models In NeMo framework:

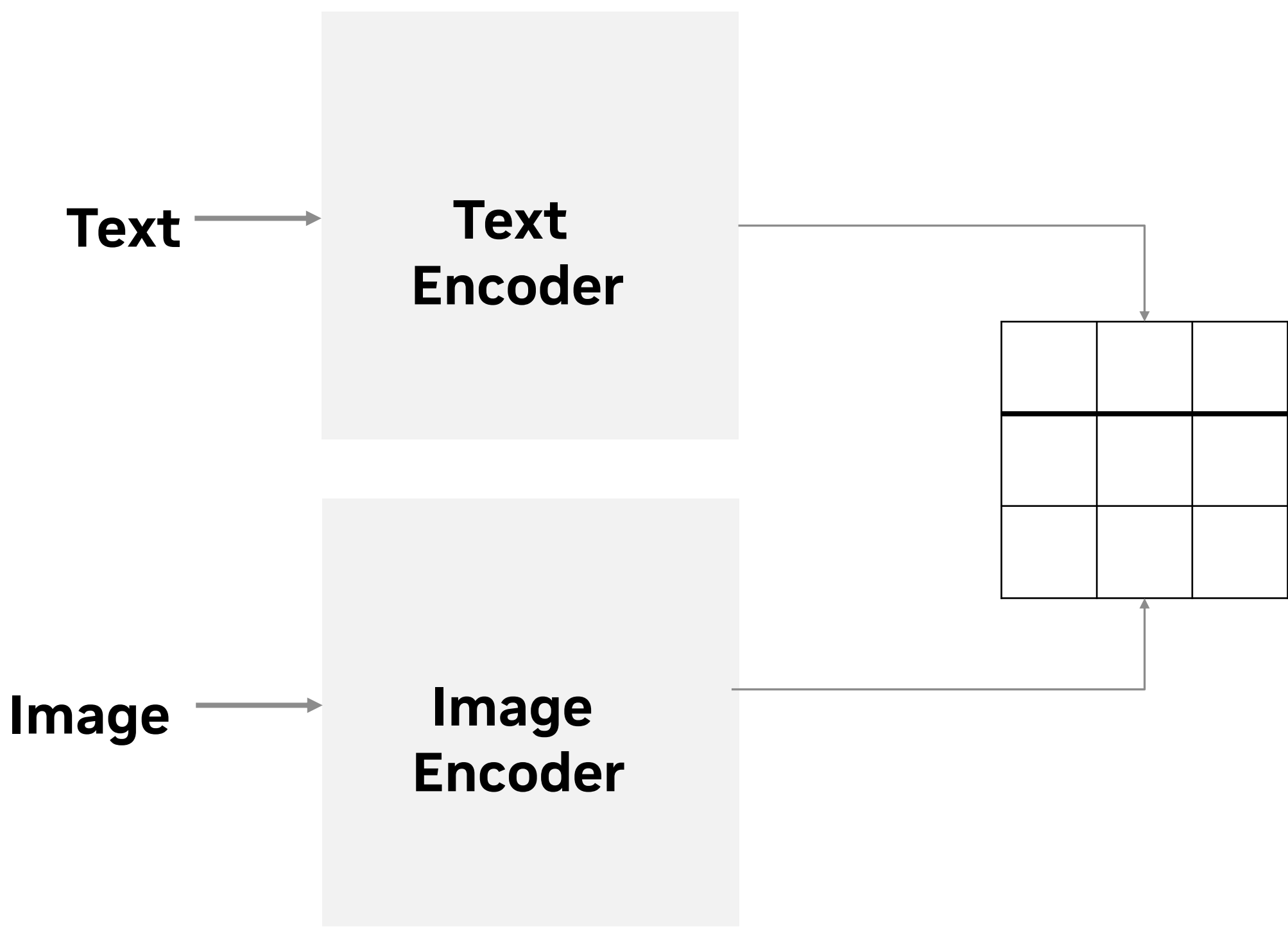
*Diffusion in Latent Space:* Stable Diffusion v1.5

*Diffusion in Pixel Space:* Imagen

*Image-to-Image Models:* Instruct-Pix2Pix (For editing images – No text encoder)

# Discriminative

Suitable for Tasks Like Image Classification, Object Detection



## Supported Models In NeMo framework:

*Text-Encoder:* Vision-Transformer

*Multi-Modal:* CLIP

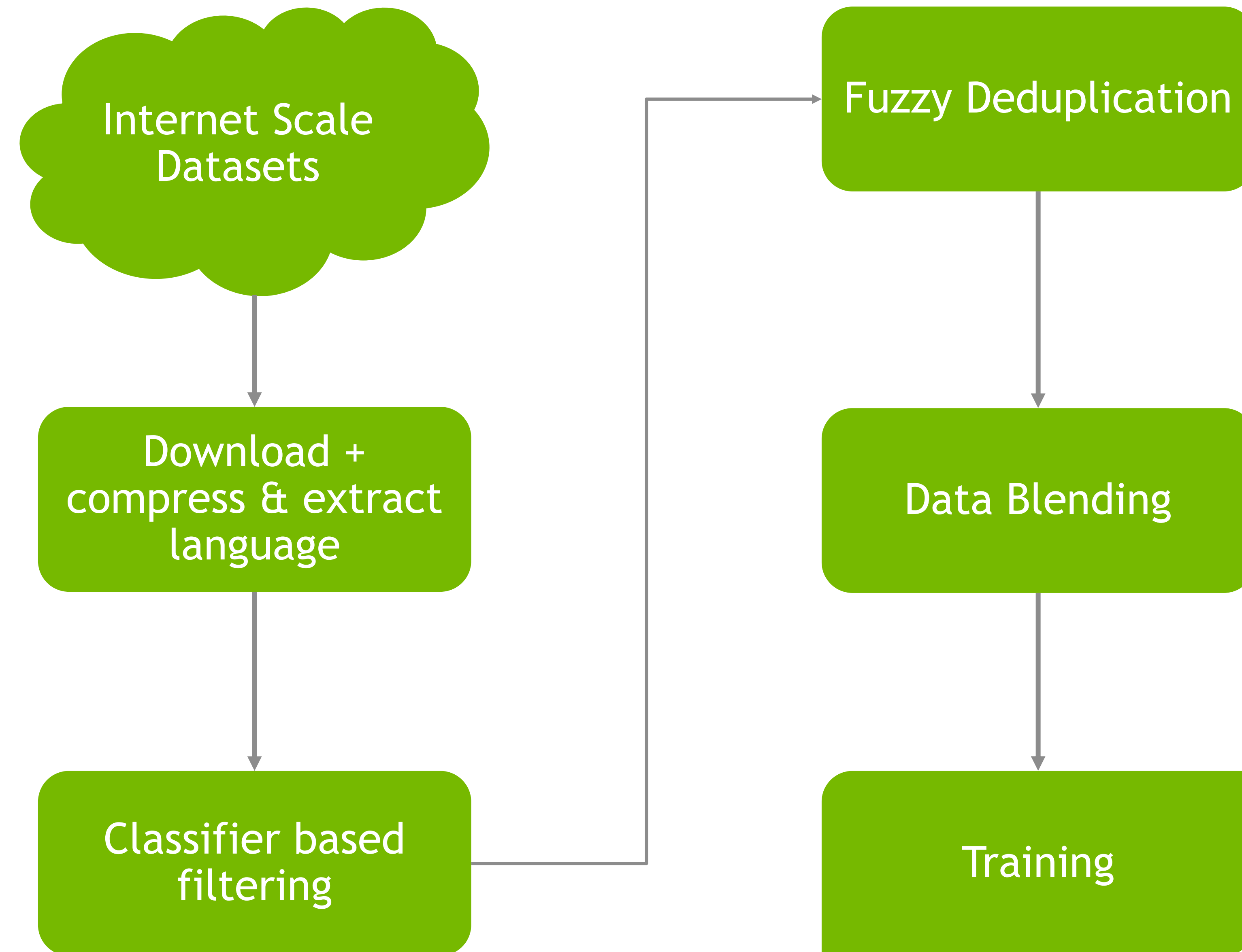
*Overall Model:* ViT-CLIP

# Support for Multi-Modal Models



## Bring your own dataset to train LLMs

Framework Agnostic Distributed Data Curation Tools for Filtering, Deduplication, and Blending



- Distributed processing leveraging DASK
  - DASK enabled auto load balancing for distributed processing
- De-duplication
- Data Cleaning-Bad Unicode, newline, repetition
- Extraction- HTML files and JavaScript

## Data Curation & Preprocessing

Enabling Large-Scale High-Quality Datasets for LLMs



*“Using hyperparameter optimization tools in NeMo allowed us to train LLMs 2x faster than with other frameworks.”*

Hwijung Ryu, LLM Development Team Lead  
Korea Telecom

## Auto-Configurator Tool

Automatically search and optimize model configurations on any given compute or time constraints

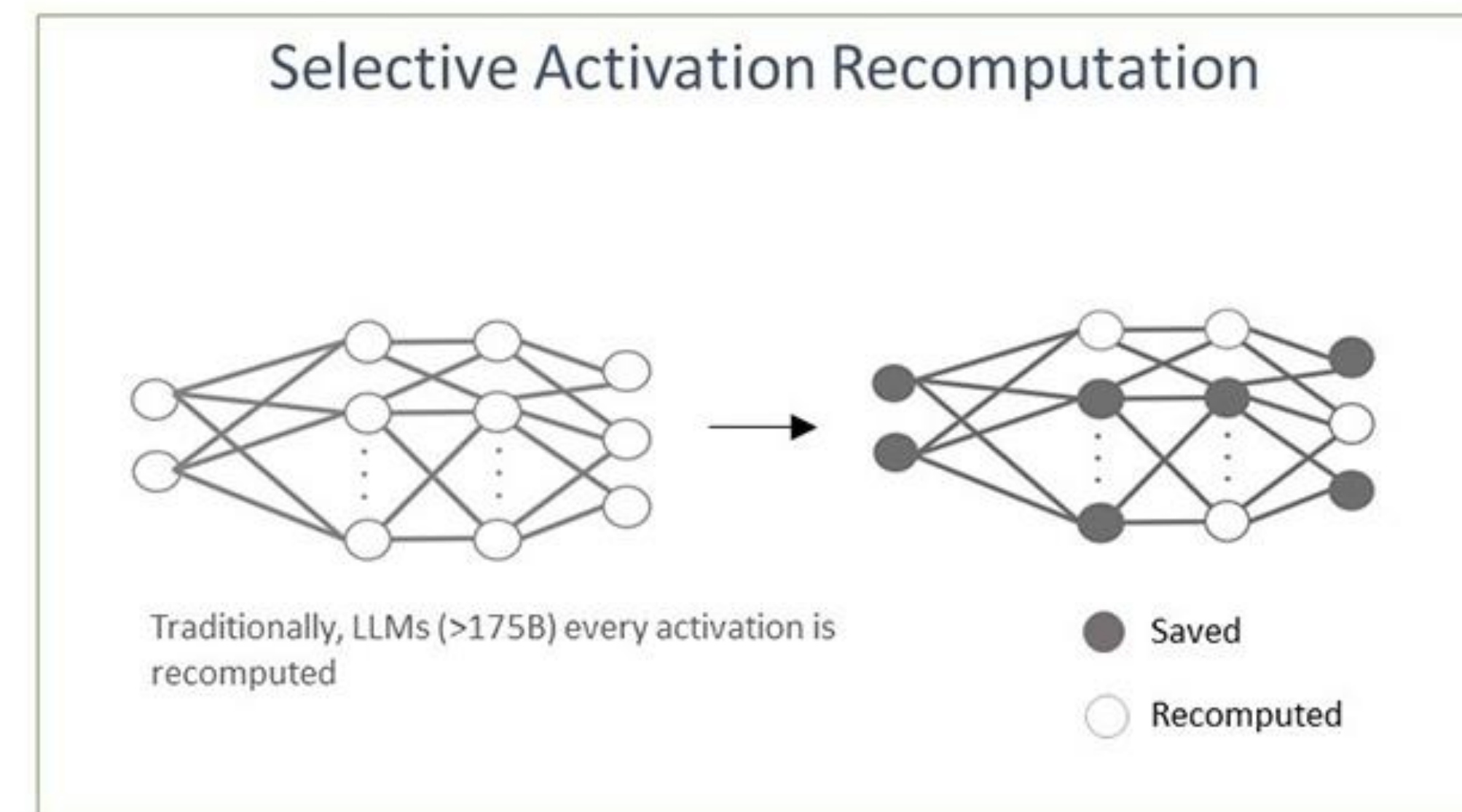
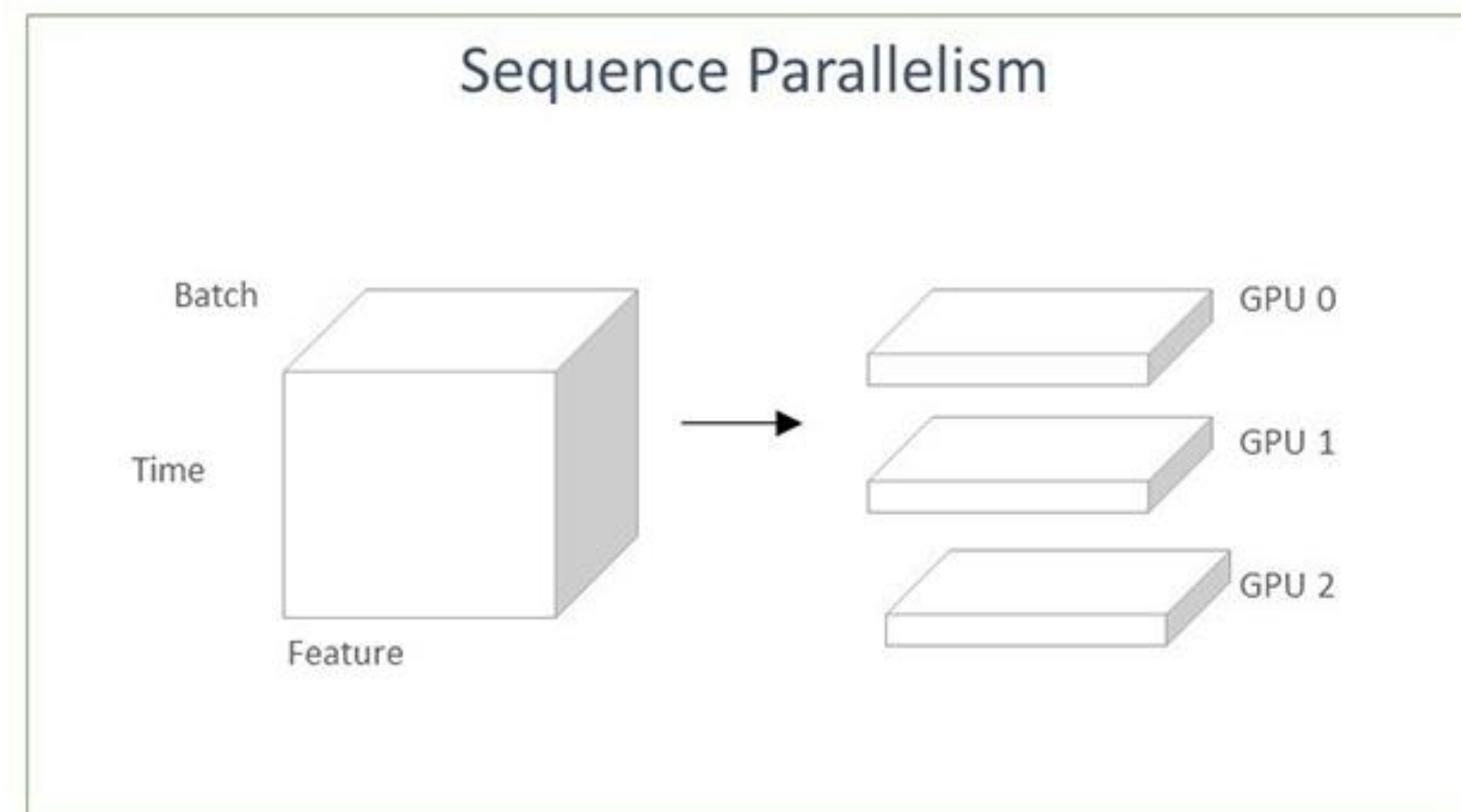
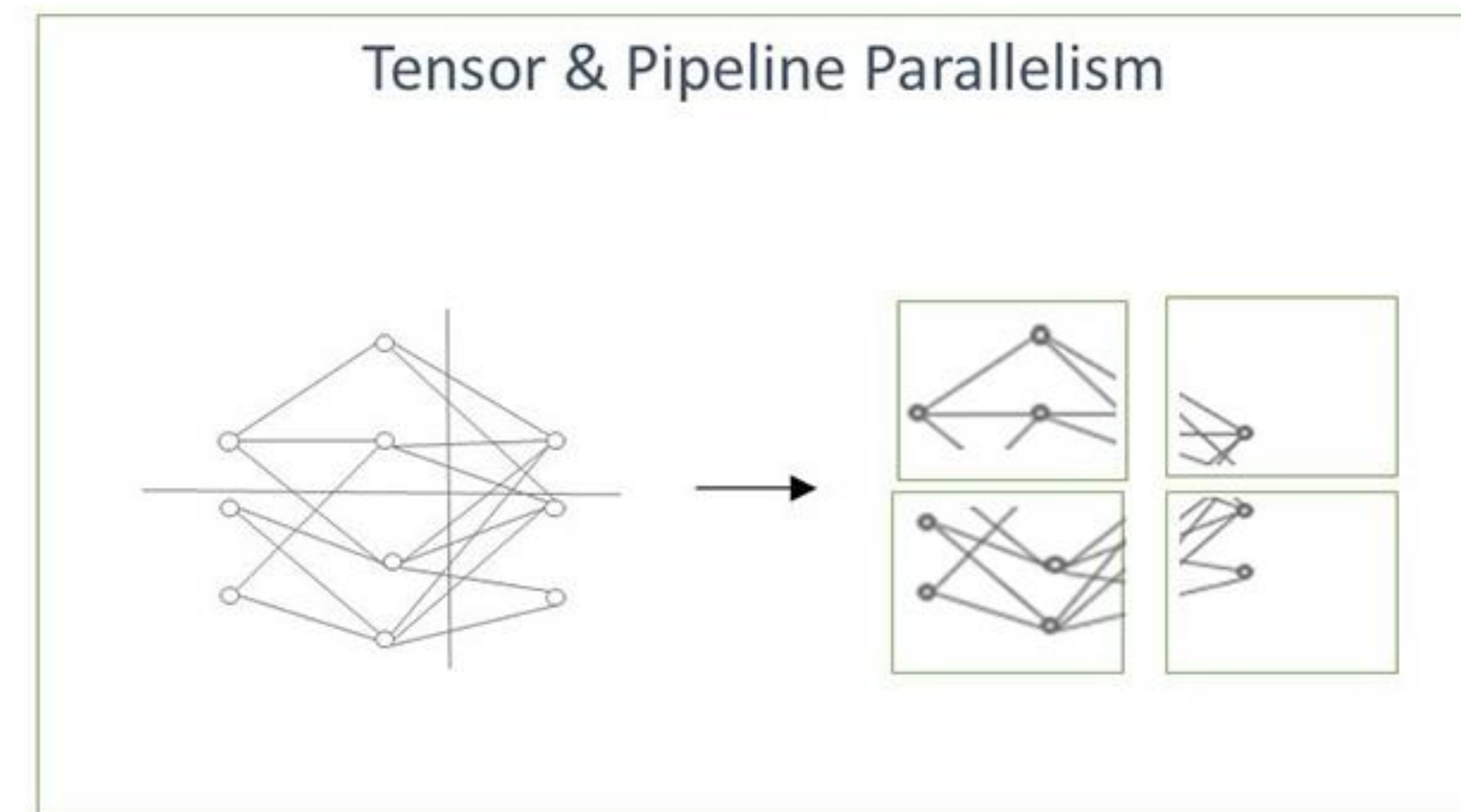
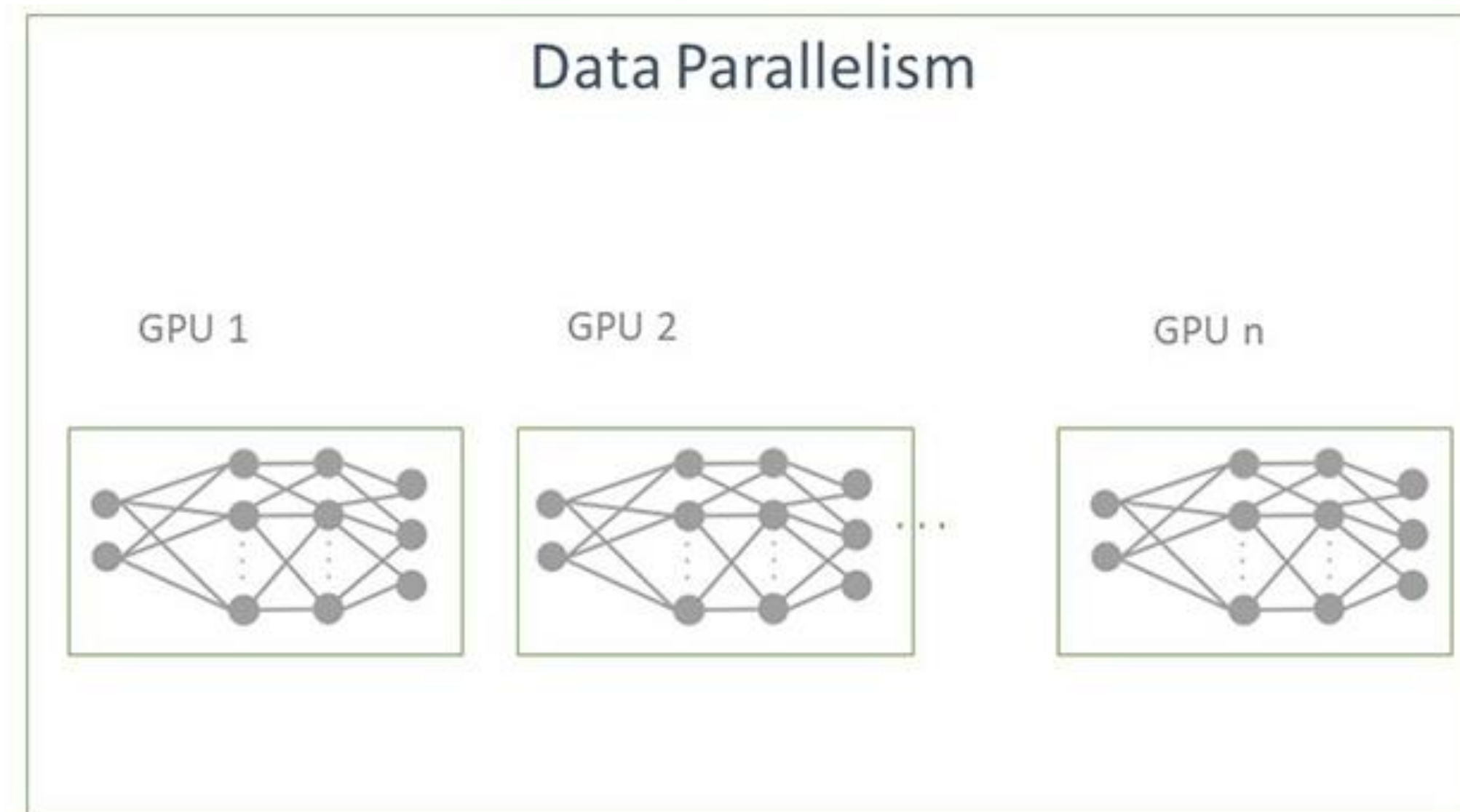
---

- Decides the model size based on your hardware constraints, inference or time constraints
- Best training and inference configurations can be found in minutes (for small models) or a few hours (for large models)



# 3D Parallelism Techniques To Build Foundation Model

NeMo framework offers efficient algorithms to train large-scale models



- Requires extensive experimentation to configure hyperparameters
- Needs state-of-the-art algorithms to process internet-scale data across an entire datacenter
- **Maximize GPU Utilization over InfiniBand and Minimum Latency within a Single Node**

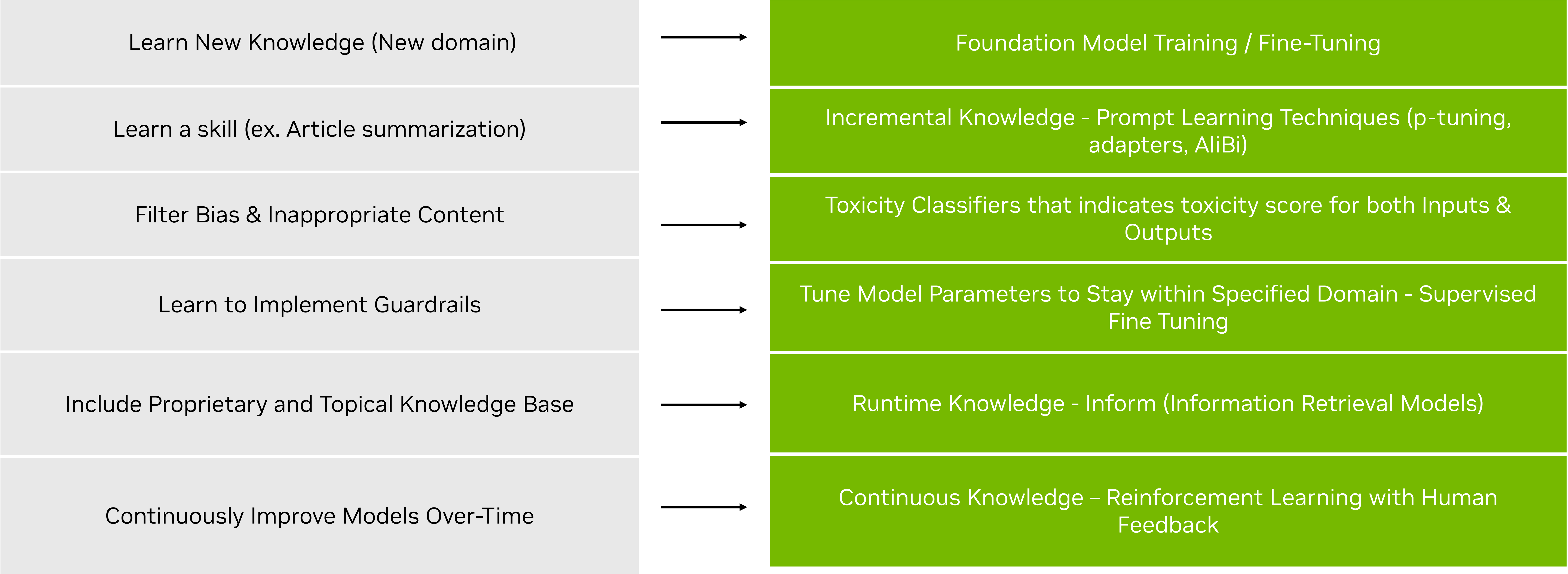


# Hyper-personalizing Foundation Models for Enterprises

Methods to build and hyper-personalize foundation models for specific use-cases

## Personalization / Customization

## Methods & Techniques

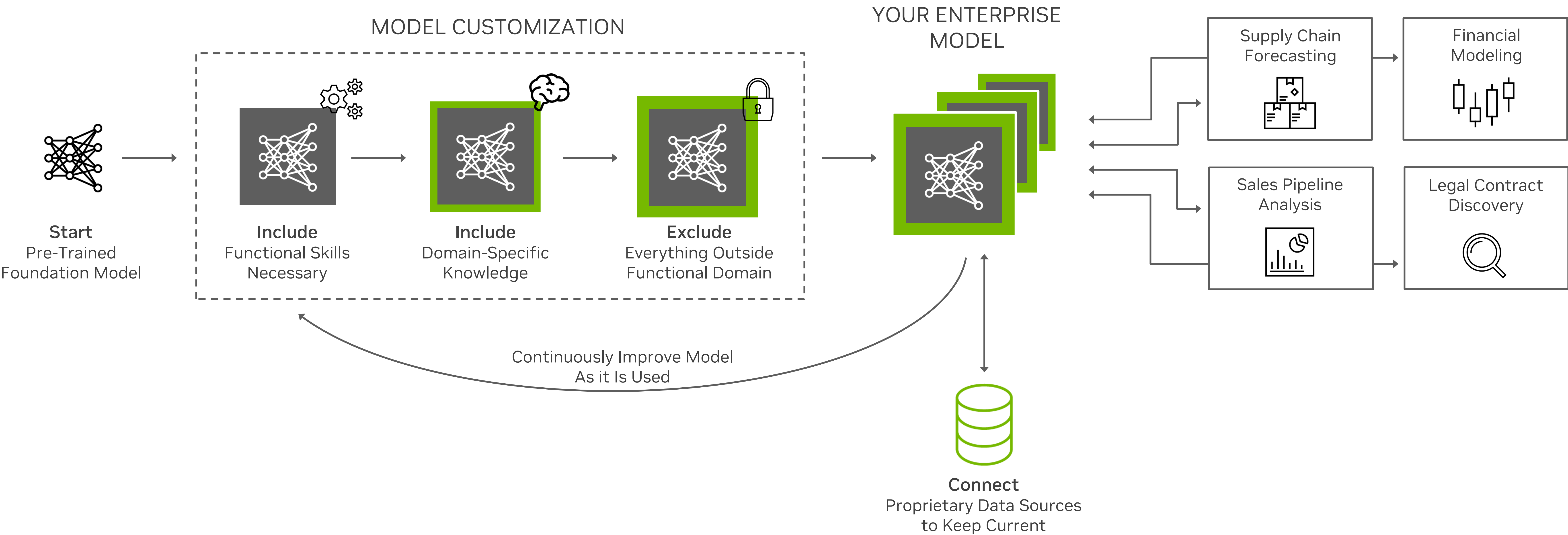
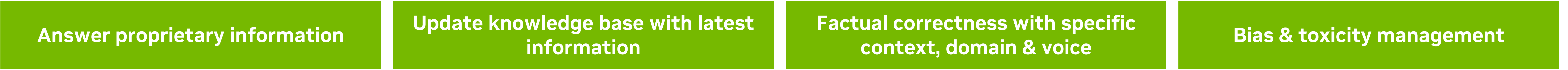


Legend: Available Today



# Overcoming Challenges Of Using Foundation Model

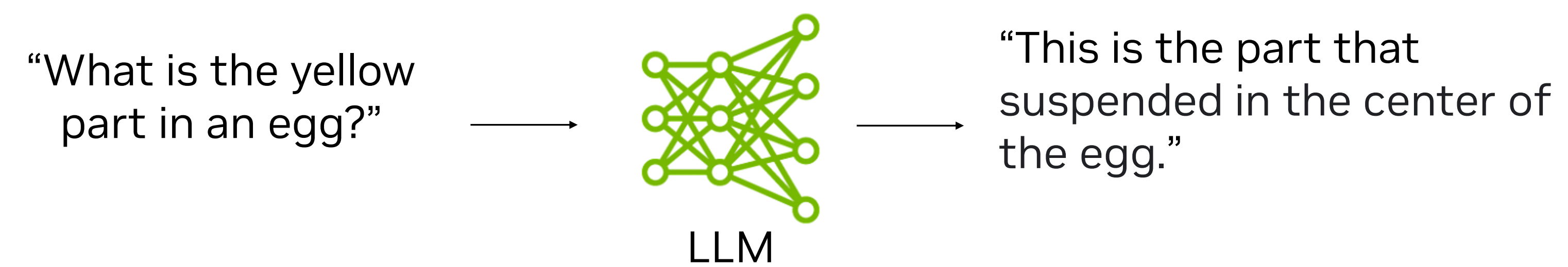
Generalized AI will not work – Enterprise need their own AI





# LLMs Are Knowledgeable To Answer General Questions

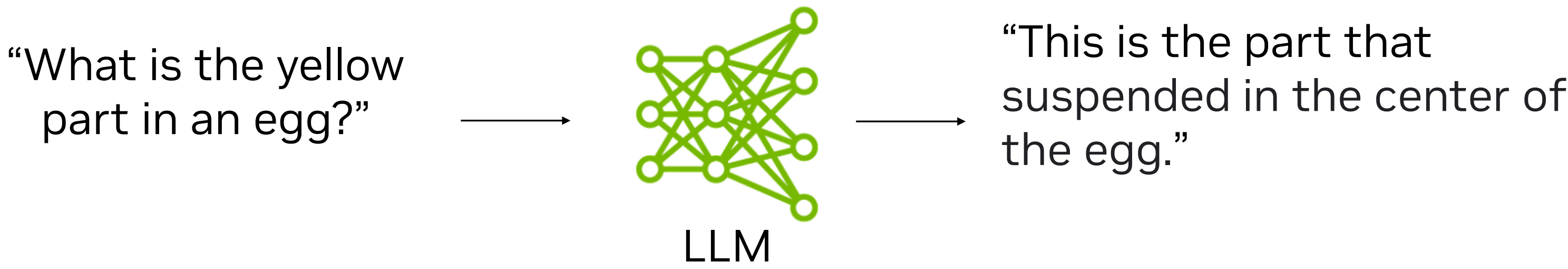
## Zero-Shot Response



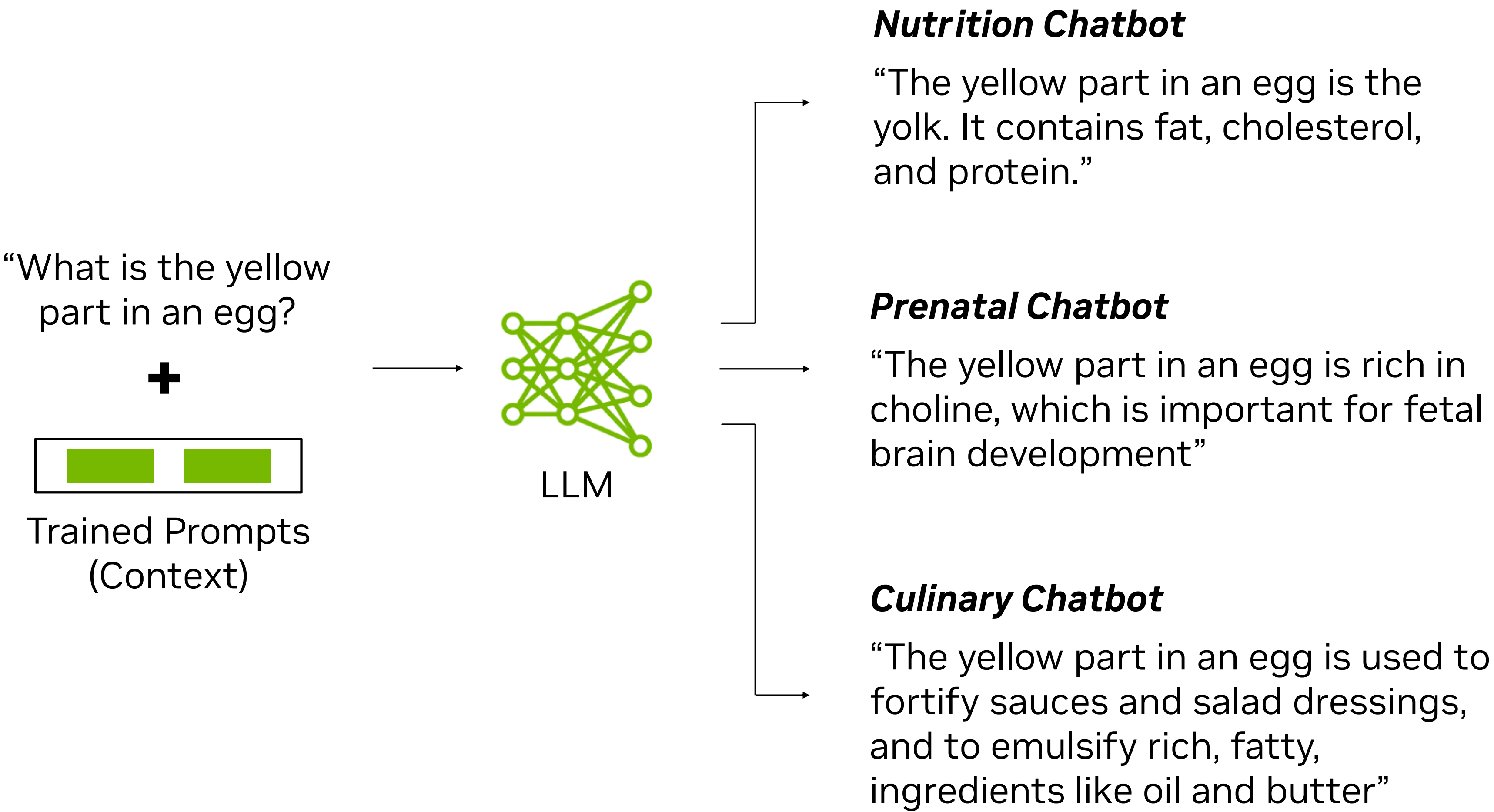


# Customization is Required to Address Business-specific Tasks

## Zero-Shot Response



## P-Tuned Response





# Enterprises Require Responses Based on Current Information



70%

**Of Enterprise Data is Untapped**

Unlock new opportunities for greater intelligence



**Less Frequent Re-Training**

Significant cost and time savings to maintain LLMs

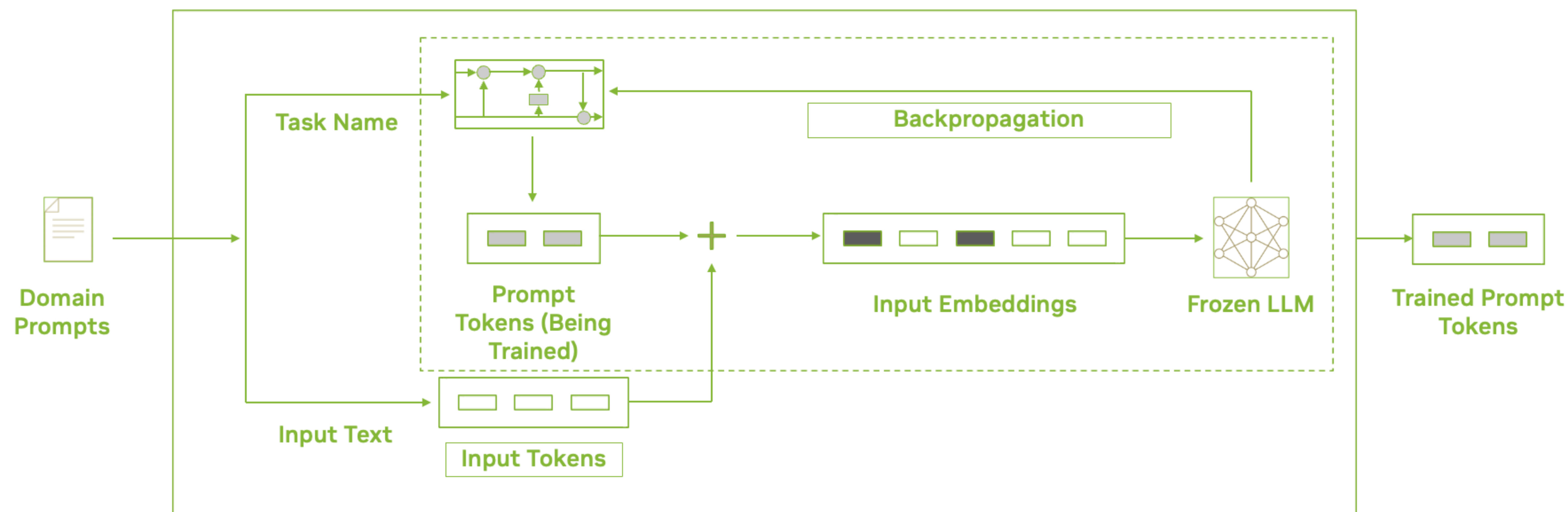


# Enterprise Use-Cases Require Functional Skills

Specialized skills to solve customer and business problems

## Provide Context to Models

Parameter efficient ways to customize LLMs for specific use-cases



- Freeze foundational model, and learn the prompt tokens using a supervised learning approach
- Can achieve high accuracy for specific use-cases with just 100s of samples



# Enterprise Use-Cases Require Domain Specific Knowledge

Encode and embed your AI with your enterprise's real-time information to provide the latest responses



70%

**Of enterprise data is untapped**

Unlock many new opportunities for greater intelligence

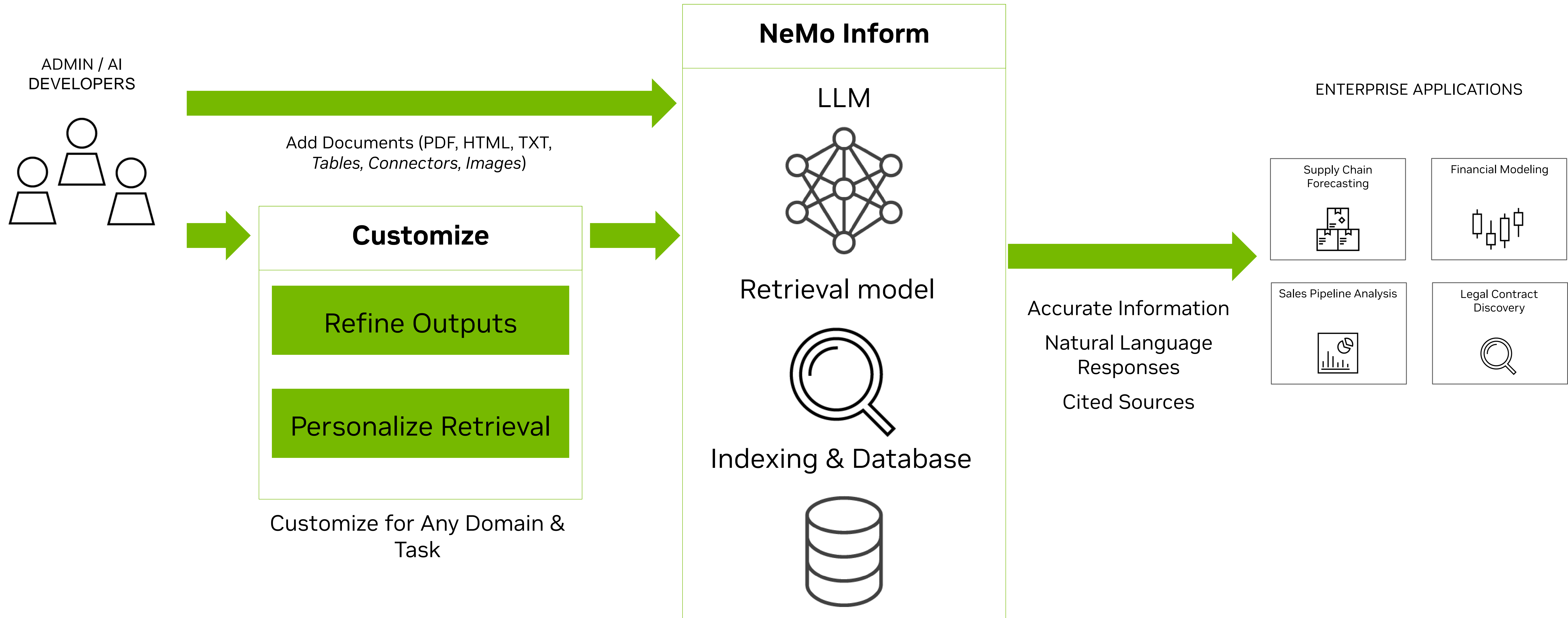


**Less Frequent Re-Training**

Significant cost and time savings in long-run to maintain LLMs



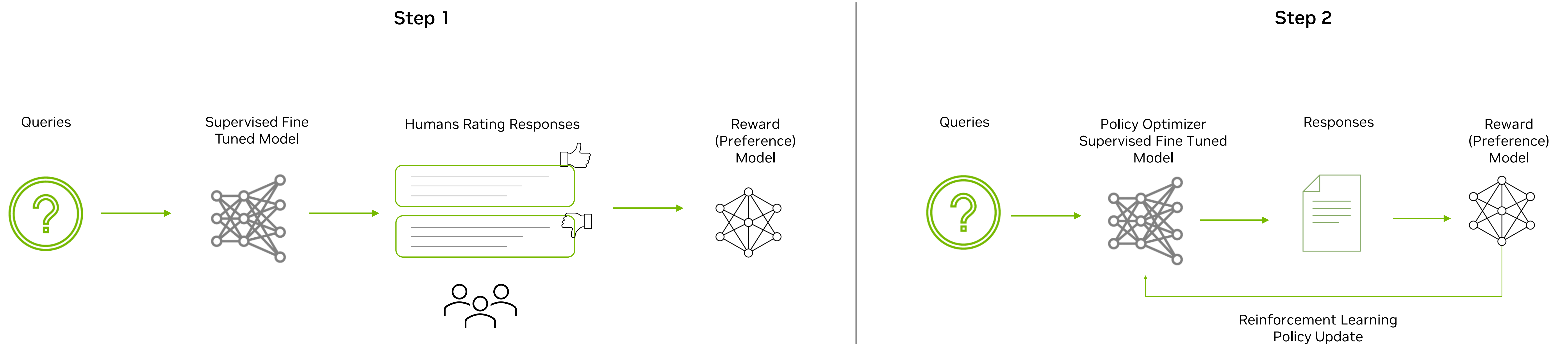
# Retrieval Augmented Models





# Enterprise Use-Cases Require Continuous Improvement

Reinforcement Learning with Human Feedback techniques allow for your enterprise model to get smarter over time, aligned to your specific enterprise domain

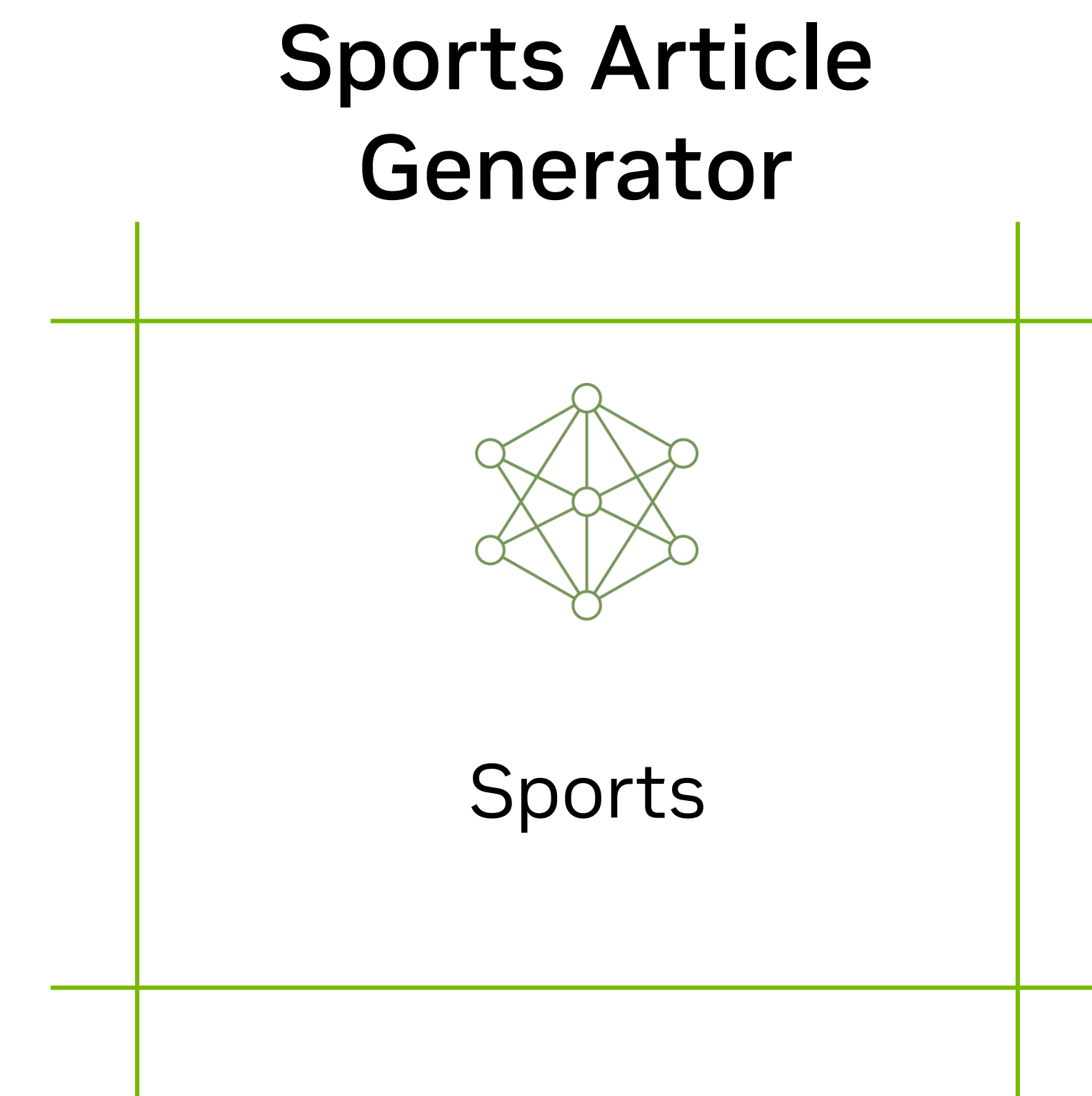
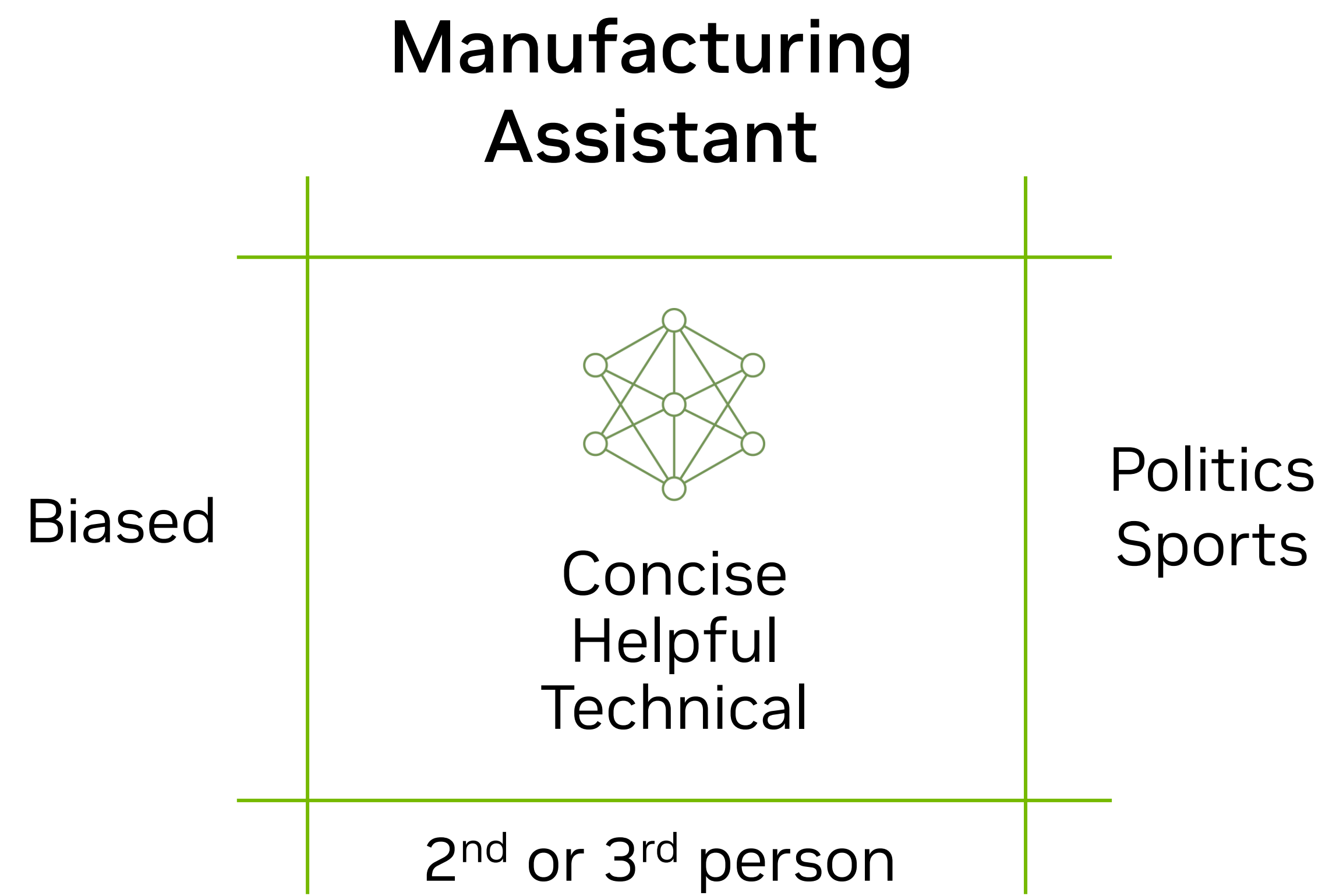


*On roadmap – not available today*



# Enterprise Use-Cases Require Guardrails

Exclude everything outside functional domain, eliminate bias and toxicity, to align to human intentions



- Toxicity classifier (BERT based classifier) assigns a toxicity score for every input and output
- Developer can use the toxicity score to filter inappropriate responses for their use-case



### Unmet Needs

### NeMo Framework addressing needs...

Large-Scale Data Processing	→	Data Curation & Preprocessing Tools
Multilingual data processing & training	→	Relative Positional Embedding (RPE) – Multilingual Support
Finding optimal hyperparameters	→	Hyperparameter Tool
Convergence of Models	→	Verified recipes for large GPT & T5-style models
Scaling on Clouds	→	Scripts/configs to run on Azure, OCI, and AWS
Deploying for inference	→	Model navigator + export to FT functionalities
Deployment at-scale	→	Quantization to accelerate inferencing
Evaluating models in industry standard benchmarks	→	Productization evaluation harness
Differing infrastructure setups	→	Full-Stack support with FP8 & Hopper Support
Lack of Expertise	→	Documentation

**Solving pain-points across the stack**



# NeMo Framework Performance - Training

	Time to train 300B tokens in days (A100) – BF16			
	800 GPUs (5x DGX SuperPod)	480 GPUs (3x DGX SuperPod)	160 GPUs (1x DGX SuperPod)	64 GPUs (8x DGX A100)
GPT-3: 126M	0.07	0.12	0.37	0.92
GPT-3: 5B	0.8	1.3	3.9	9.8
GPT-3: 20B	3.6	6	18.1	45.3
GPT-3: 40B	6.6	10.9	32.8	82
GPT-3: 175B	28	46.7	140	349.9



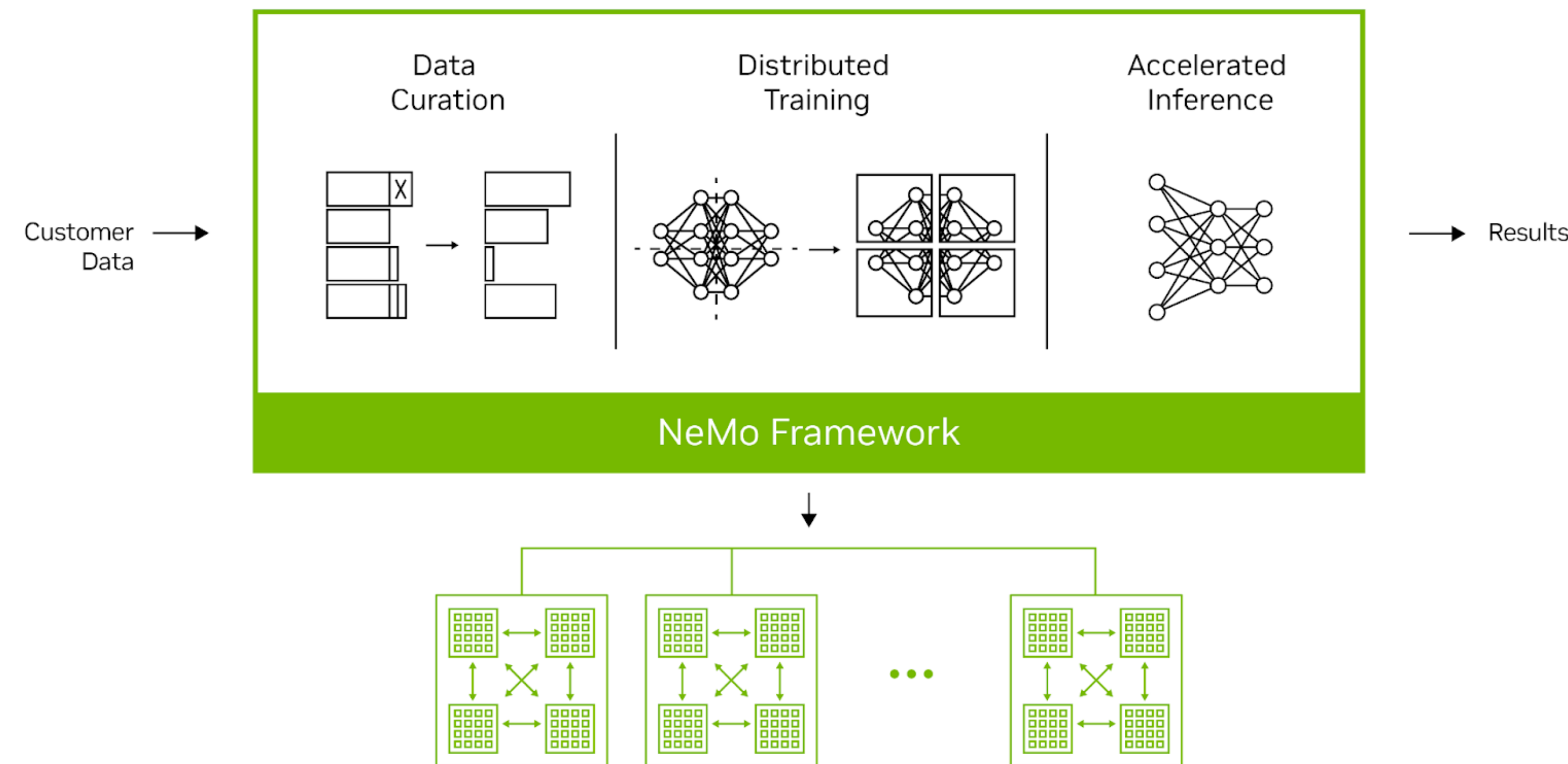
The background features a complex pattern of glowing green lines and shapes against a solid black field. On the left, numerous thin, parallel green lines radiate outwards. On the right, there are larger, more intricate structures resembling stylized leaves or overlapping rectangular frames, composed of many fine green lines that create a sense of depth and movement.

# NeMo Framework Availability



# NeMo Framework

An end-to-end, cloud-native enterprise framework to build, customize and deploy generative AI models



**NVIDIA DGX SuperPODs**

**NVIDIA DGX Cloud**

**NVIDIA DGX Systems**

## Training



## Inference



- ✓ Expanding support across modalities
  - ✓ **Multi-modal:** SD, ViT, ViT-CLIP, Instruct-Pix2Pix, Imagen
  - ✓ **LLMs:** BERT > 100B, T5-MoE, T5, GPT-3, Inform
- ✓ 3D parallelism techniques: Data, Tensor & Pipeline, Sequence Parallelisms, Selective Activation Recomputation
- ✓ Customization techniques for LLMs: Adapters, RLHF, AliBi, SFT
- ✓ Auto-configurator tool to find optimal hyperparameters
- ✓ Orchestration support: SLURM, Nephele, Kubernetes – K8s (Inference)

**General Availability with NVIDIA AI Enterprise in Q2'2023  
(LLMs Only)**

**Multi-modal functionality available via early access now**



## Now Available on Public Clouds



- NeMo framework is now available in open-beta
- [Sign-up and download containers from NVIDIA GPU Cloud](#)
- Source-code and training examples are available for [use through our NeMo GitHub Repo](#)
- Available for use on:
  - NVIDIA DGX SuperPODs
  - Cloud Platforms

## Access NeMo Framework

Now Available for developers around the world through open-beta



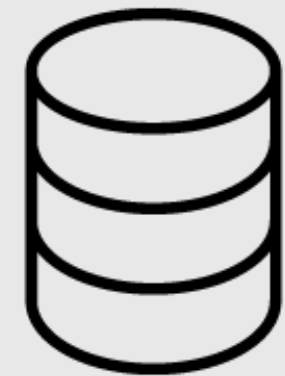
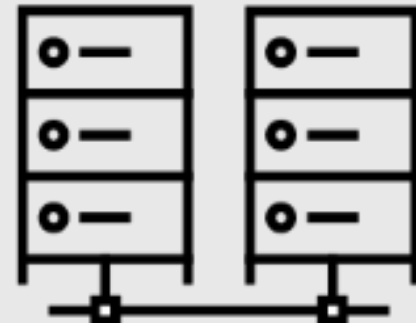




# **Distributed Training, Inferencing and Customization of Large Language Models**



# Training & Deploying of Foundation Models are Challenging

Foundation models are neural networks trained on massive unlabeled datasets to handle a wide variety of tasks

	Mountains of Training Data
	Large-scale compute infrastructure for training & inferencing, costing \$10 M+ in just cloud costs
	Complex techniques to train and deploy on large-scale infrastructure
	Deep technical expertise



# Training & Deploying of GPT-3

## Training

Train 300B tokens in days (A100) - BF16			
	800 GPUs (5x DGX SuperPod)	3x DGX SuperPod	1x DGX SuperPod
GPT-3: 126M	0.07	0.12	0.37
GPT-3: 5B	0.8	1.3	3.9
GPT-3: 20B	3.6	6	18.1
GPT-3: 40B	6.6	10.9	32.8
GPT-3: 175B	28	46.7	140

## Inference

Estimated Inference Capacity					
GPT-3 Model Parameter Count	Precision	Input/Output Length (Tokens)	Batch Size	Estimated GPU Memory Size	Estimated # of A100 80GB
100M - 3B	FP16	60/20 200/200	1-256	200MB - 6GB	1
5B - 20B	FP16	60/20 200/200	1-256	10GB - 600GB	1-8
100B - 300B	FP16	60/20 200/200	1-256	200GB - 2TB	8-32 GPUs 1-4 Nodes
500B - 1T	FP16	60/20 200/200	1-256	1TB - 5TB	16-64 GPUs 2-8 Nodes



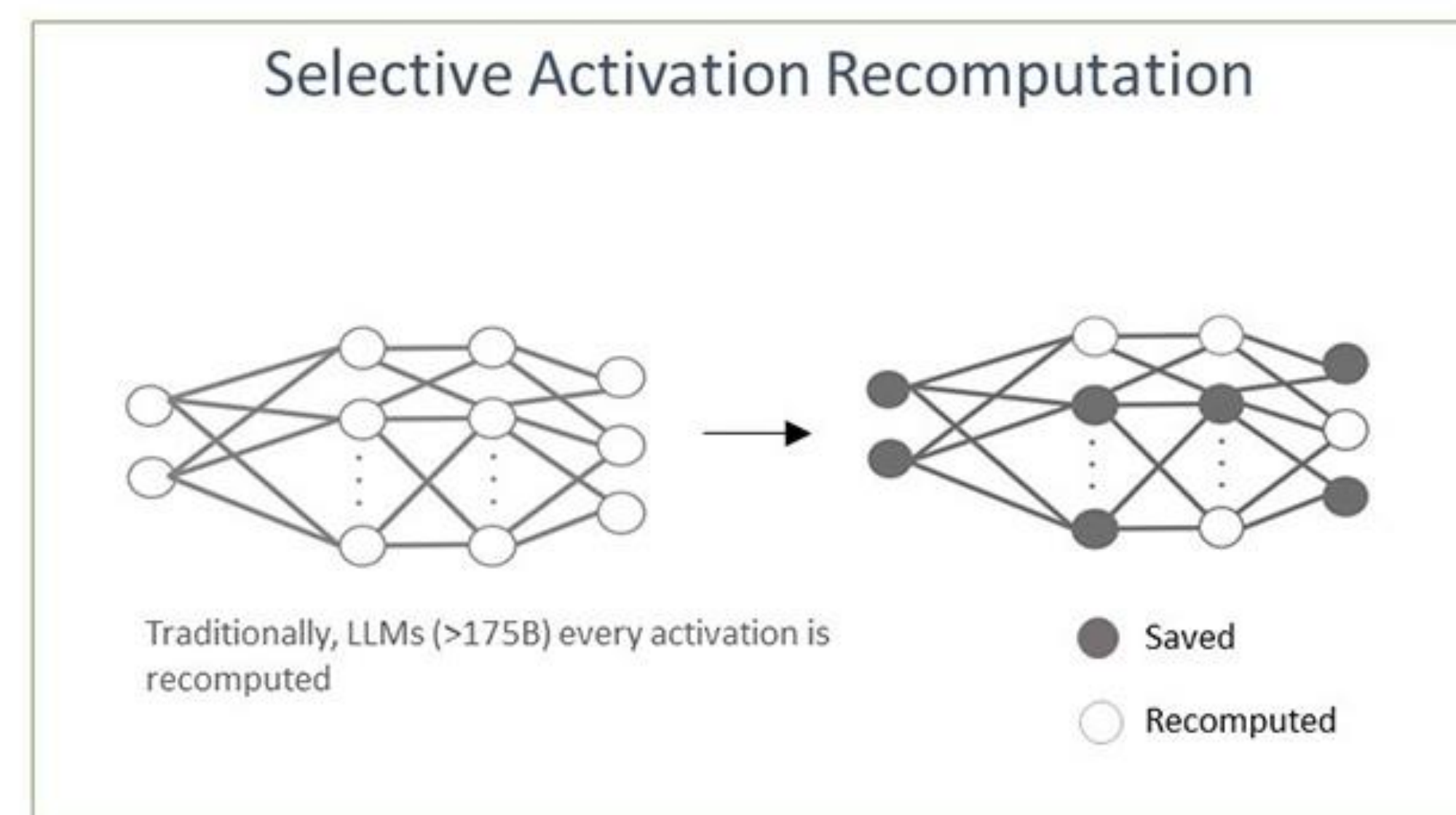
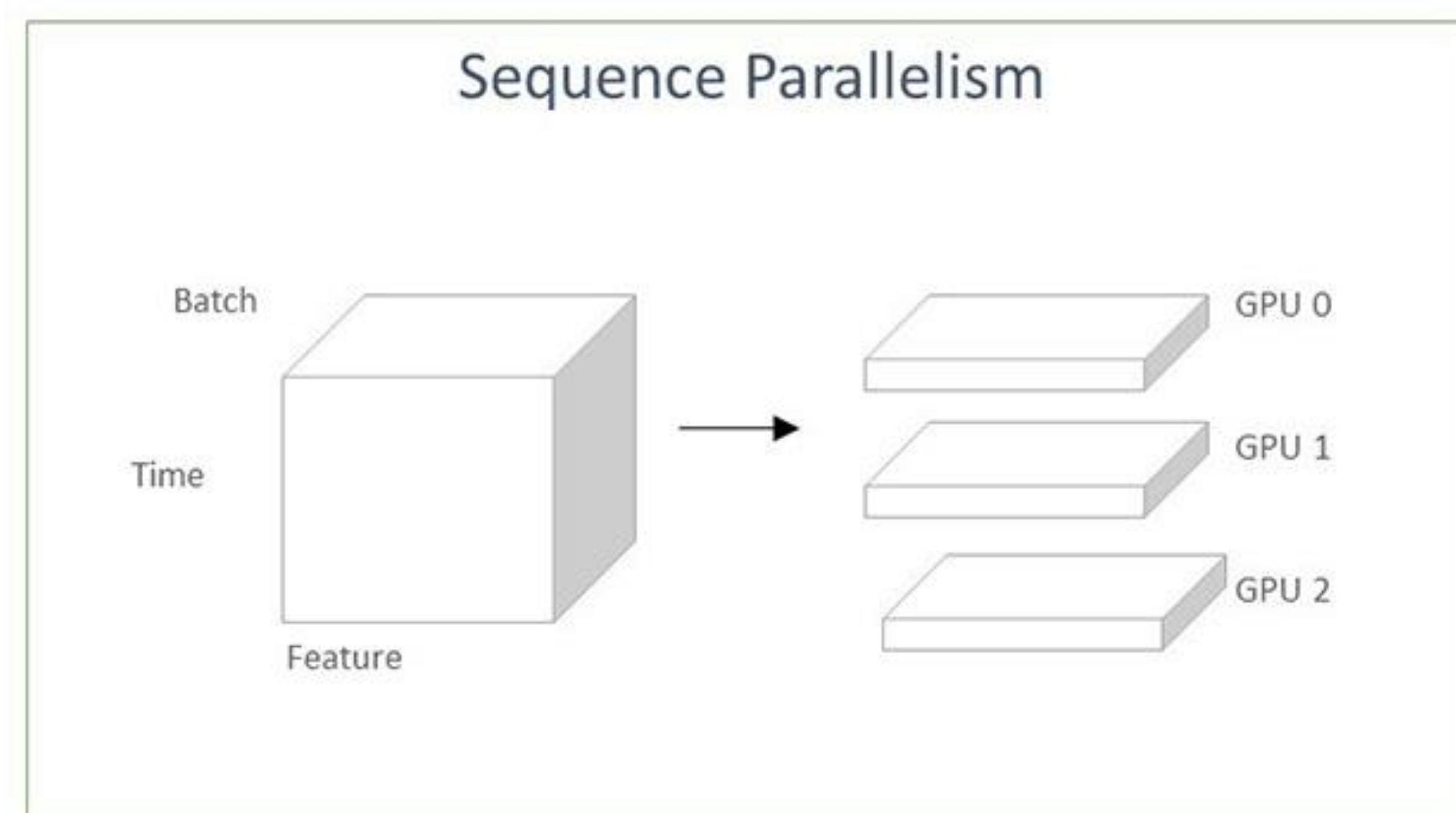
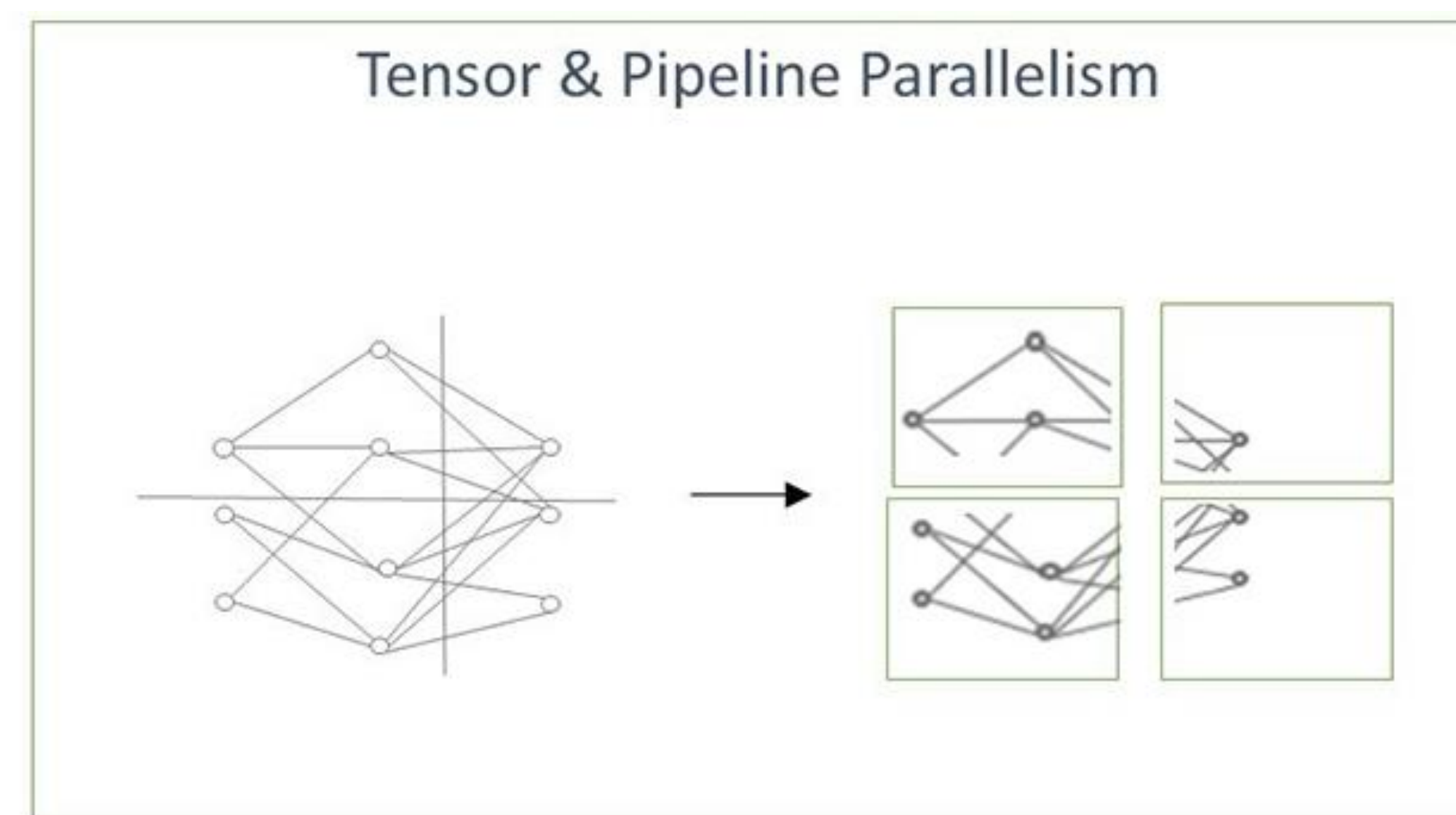
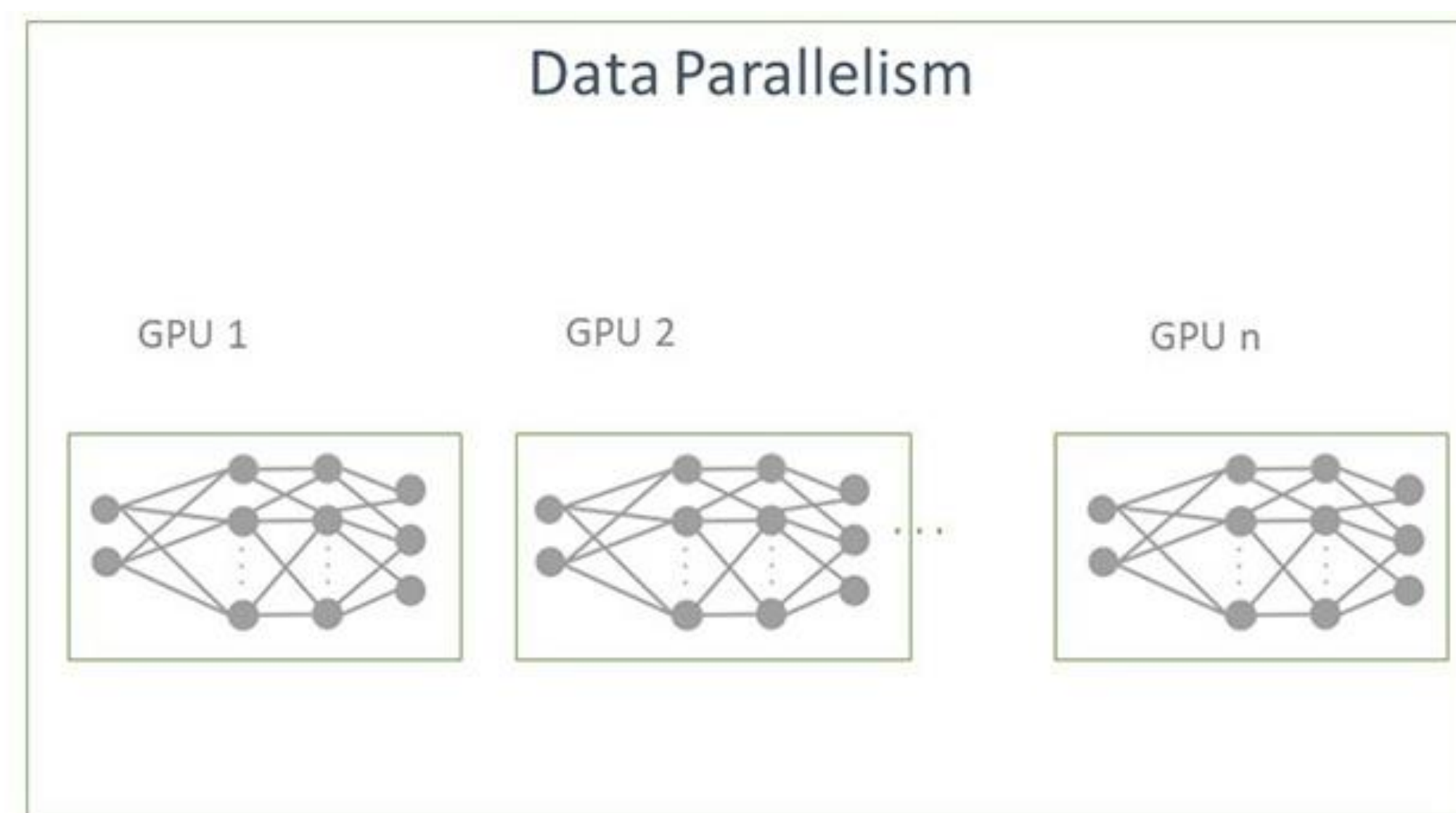
The background of the slide is a black field filled with numerous thin, curved, and slightly blurred lines in shades of green and yellow. These lines appear to be moving or flowing across the frame, creating a sense of dynamic energy and data flow. The lines are more concentrated in the lower right quadrant, where they form a more complex, almost geometric pattern, and become sparser towards the top left.

# Distributed Training



# Overcoming Challenges of Training Foundation Model

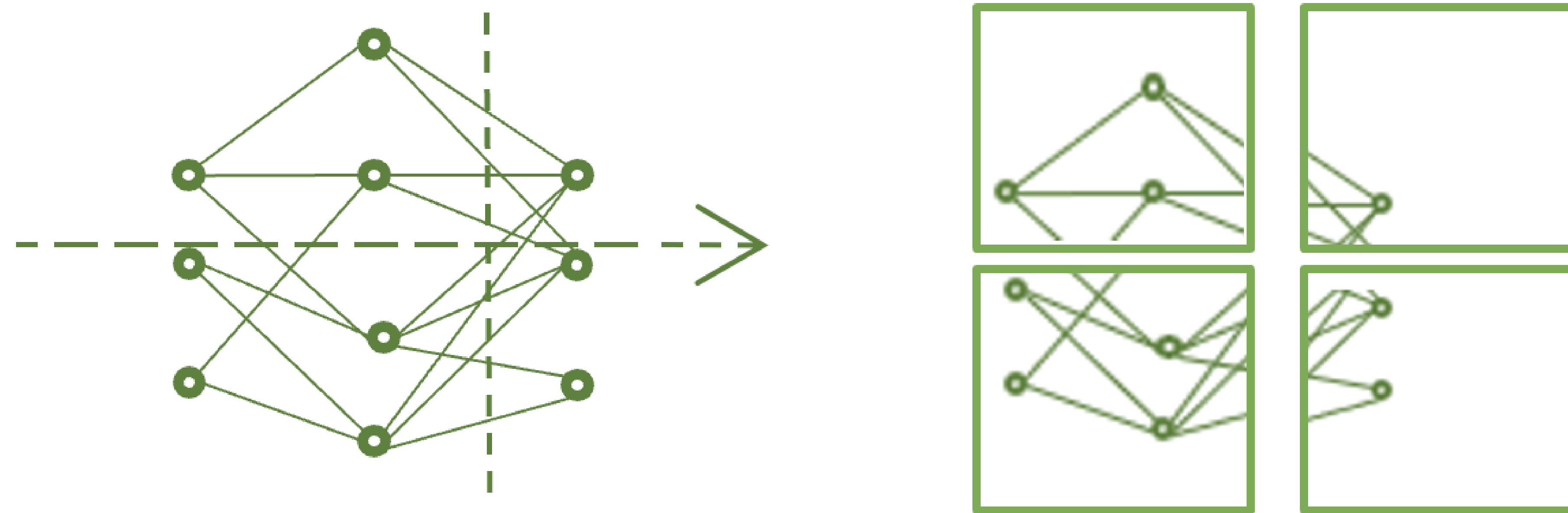
NeMo framework offers efficient algorithms to train large-scale models



- Requires extensive experimentation to configure hyperparameters
- Needs state-of-the-art algorithms to process internet-scale data across an entire datacenter



## Maximize GPU Utilization over InfiniBand and Minimum Latency within a Single Node



### Pipeline (Inter-Layer) Parallelism

- Split contiguous sets of layers across multiple GPUs
- Layers 0,1,2 and layers 3,4,5 are on different GPUs

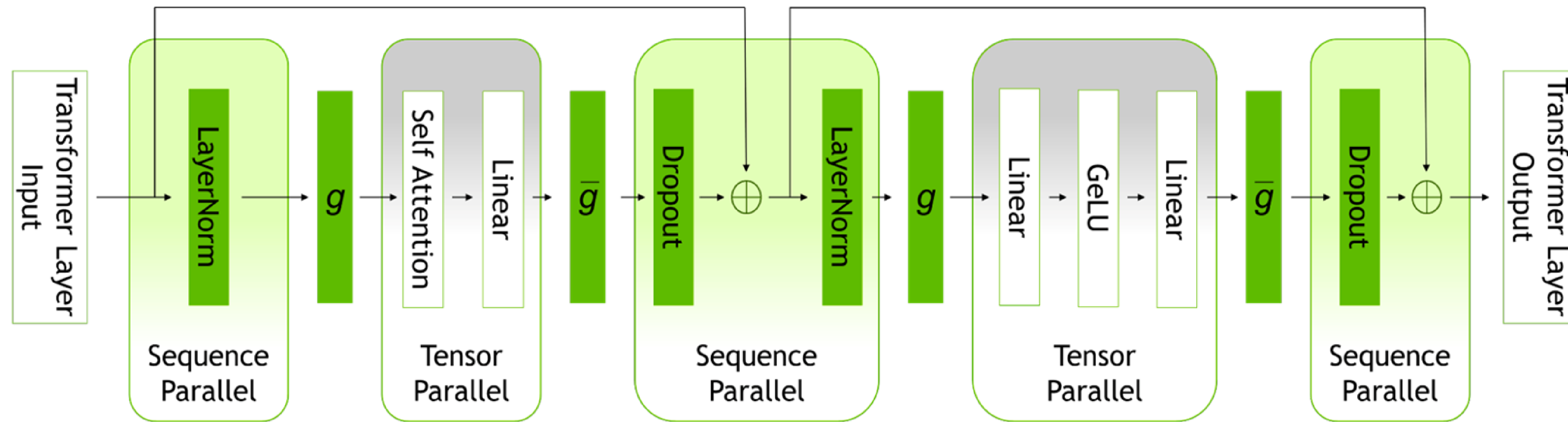
### Tensor (Intra-Layer) Parallelism

- Split individual layers across multiple GPUs
- Devices compute different parts of Layers 0,1,2,3,4,5

# Pipeline & Tensor Parallelism for Training

Training Models at Scale





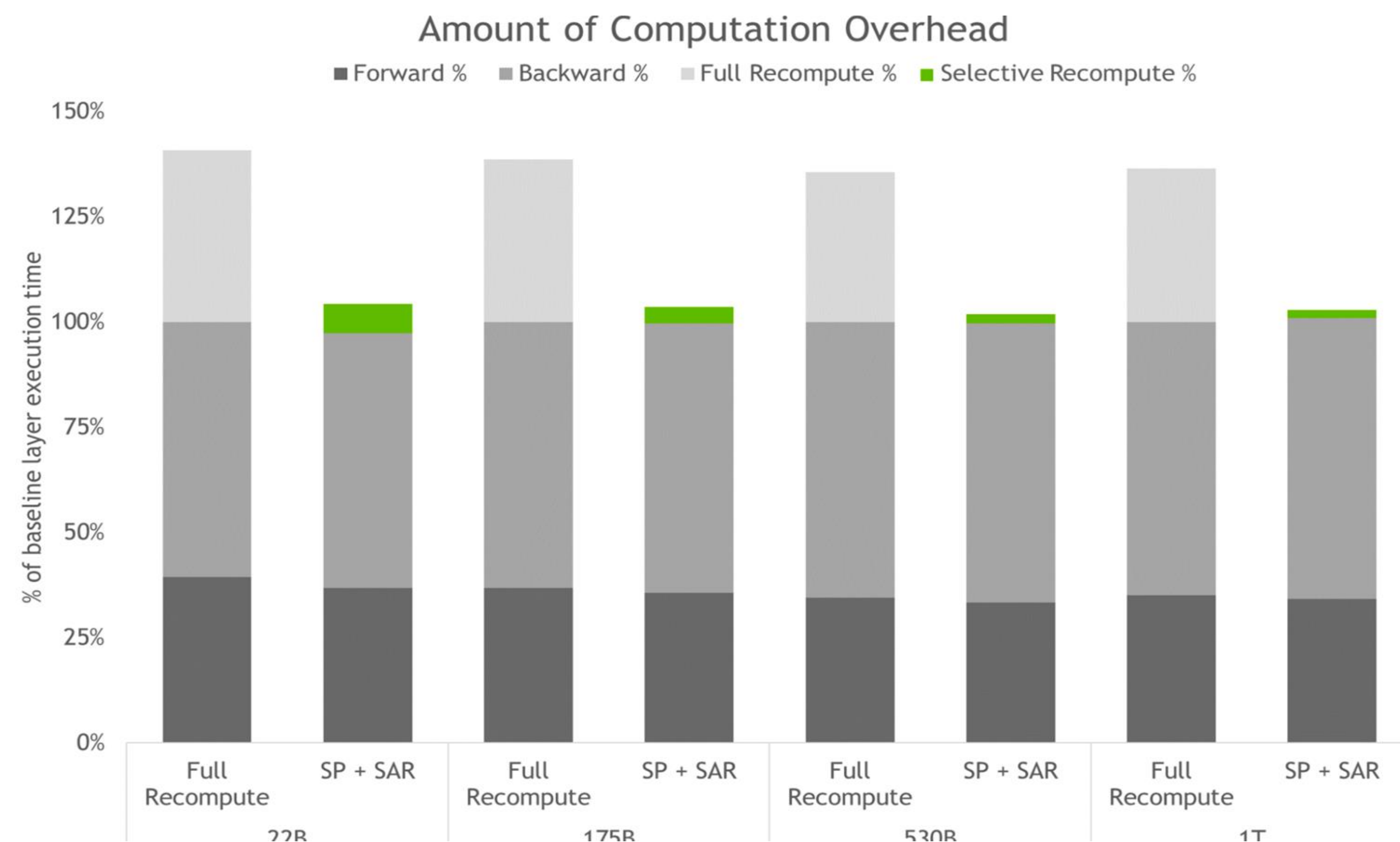
- Splits tensors across sequence dimension
- Reduce memory consumption of activation to reduce re-computation of activations during back-prop

## Sequence Parallelism for Training

Increase throughput during back-propagation



## Selective Activation Recomputation



- Choose activations to calculate based on compute-memory tradeoff
- Lower memory footprint of activations and increase throughput of network

## Selective Activation Recomputation for Training



# Distributed Training with Nemo

## Example of Config

**model:**

.....

**tensor\_model\_parallel\_size: 8**

**pipeline\_model\_parallel\_size: 16**

.....

**## Activation Checkpointing**

**activations\_checkpoint\_granularity: selective # 'selective' or 'full'**

.....

**## Sequence Parallelism**

**sequence\_parallel: True**



The background features a solid green vertical bar on the left side. The rest of the image is a dark field filled with numerous thin, bright green lines that appear to be moving or flowing, creating a sense of dynamic energy. Some lines are straight, while others are curved or bundled together.

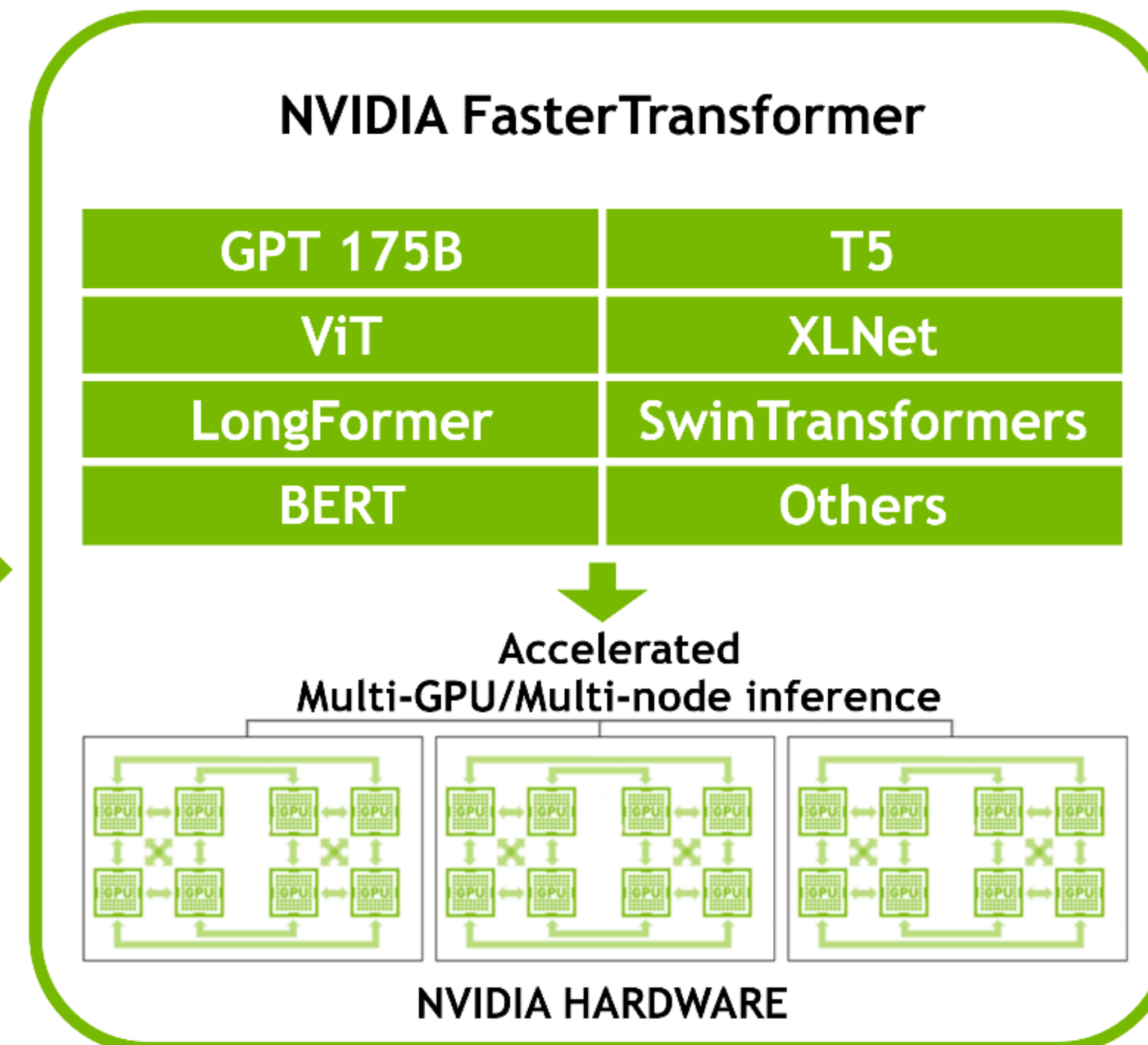
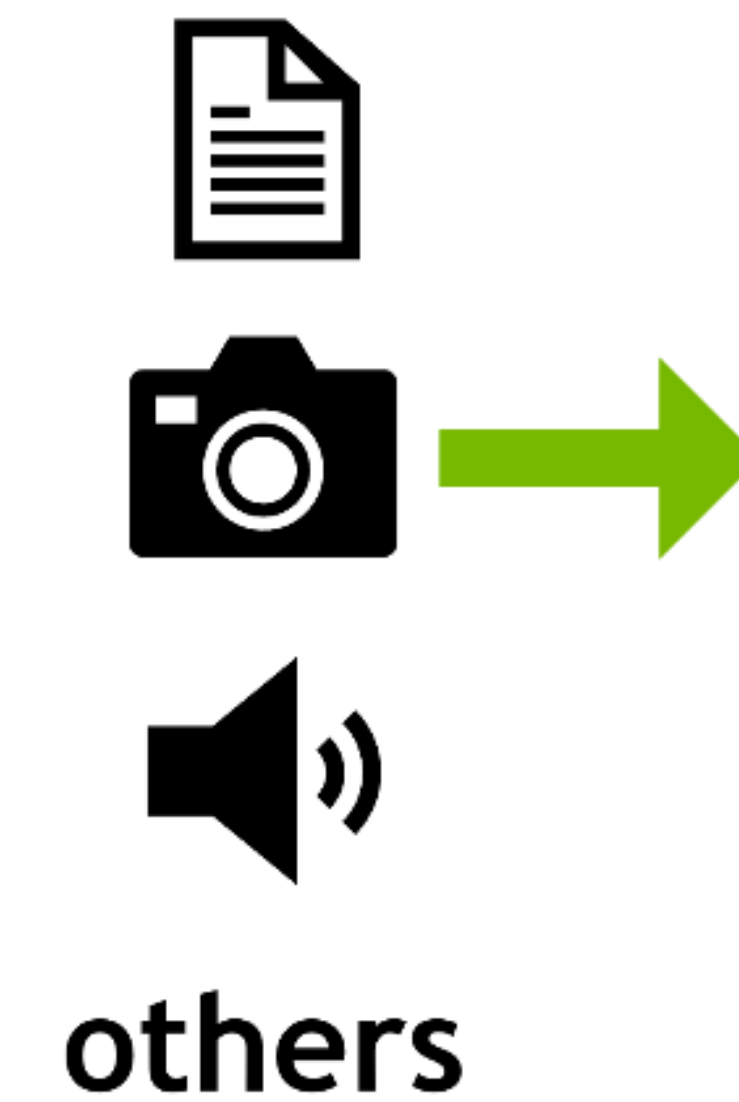
# Distributed inference



# Distributed inference with fastertransformer

- Accelerated engine for the inference of transformer-based models
- Leverage highly optimized cuBLAS, cuBLASLt, and cuSPARSELt libraries.
- Highly optimize transformer blocks.
  - Layer fusion
  - GEMM autotuning
  - Quantization
- Distributed inference with MNMG.
  - Usage of MPI and NCCL

Inputs



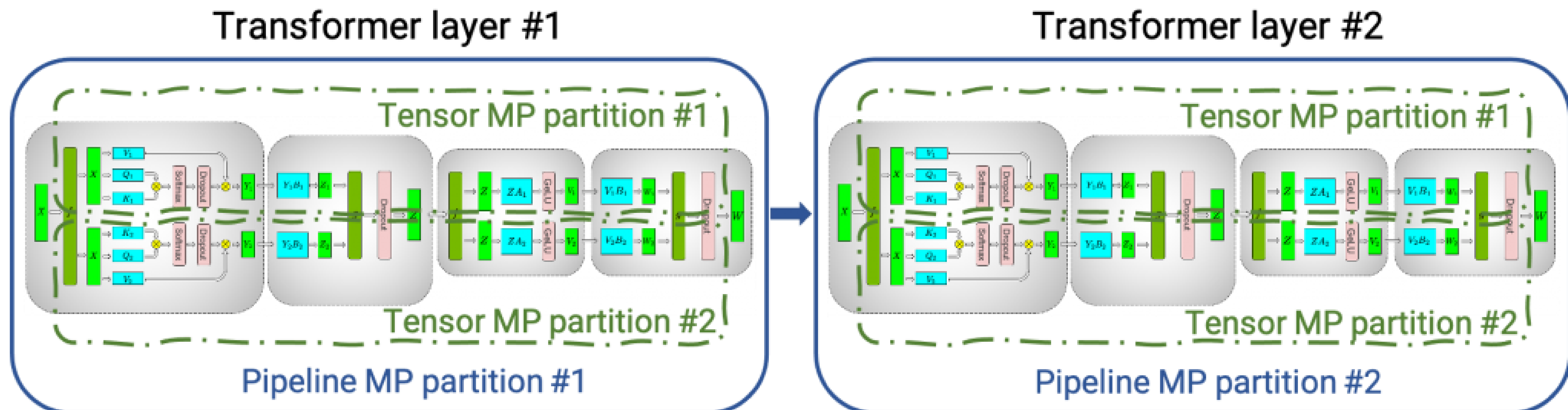
Tasks/  
Outputs

- Classification
- Generation
- Summarization
- Representation (Embeddings)
- Others



# DISTRIBUTED INFERENCE WITH FASTERTRANSFORMER

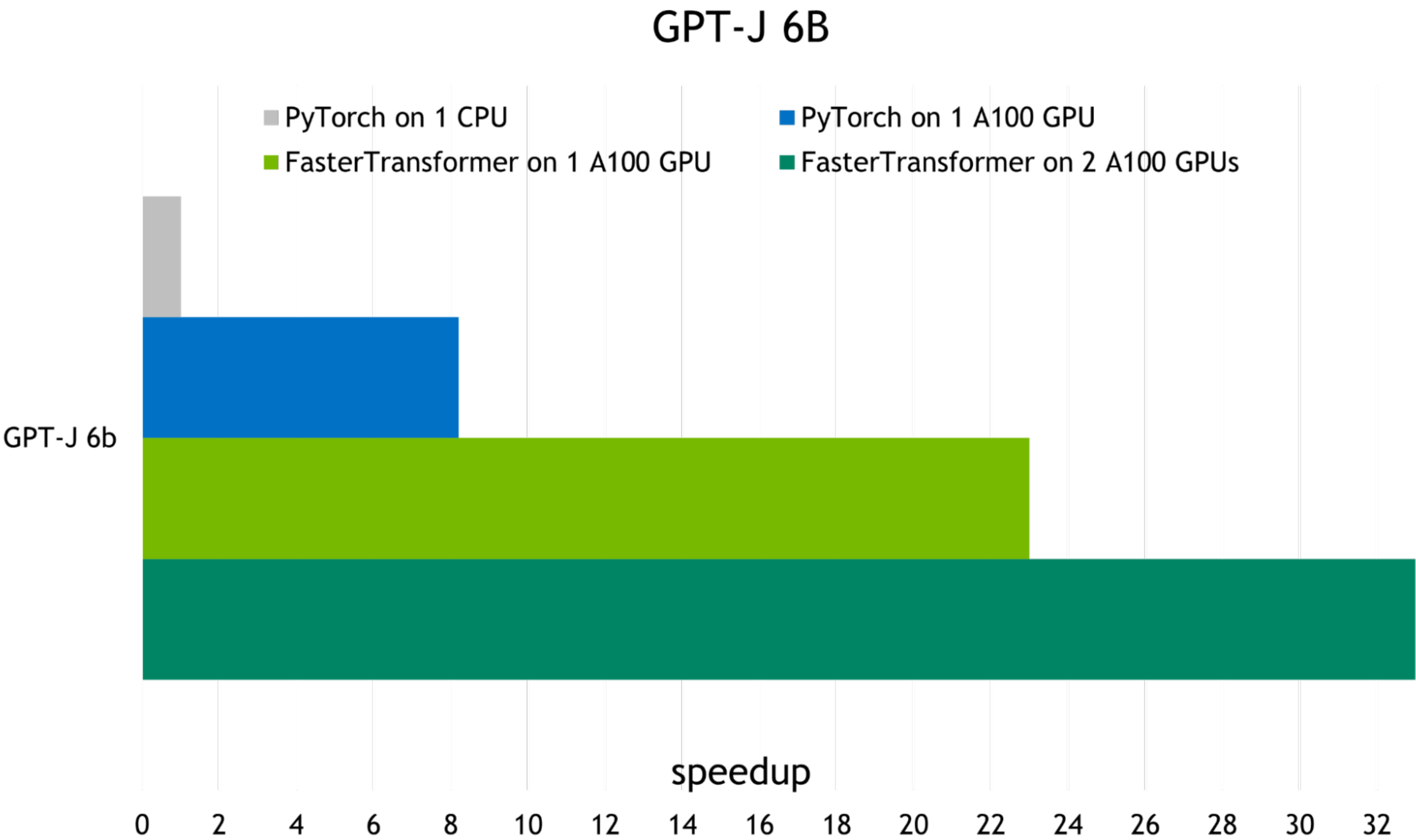
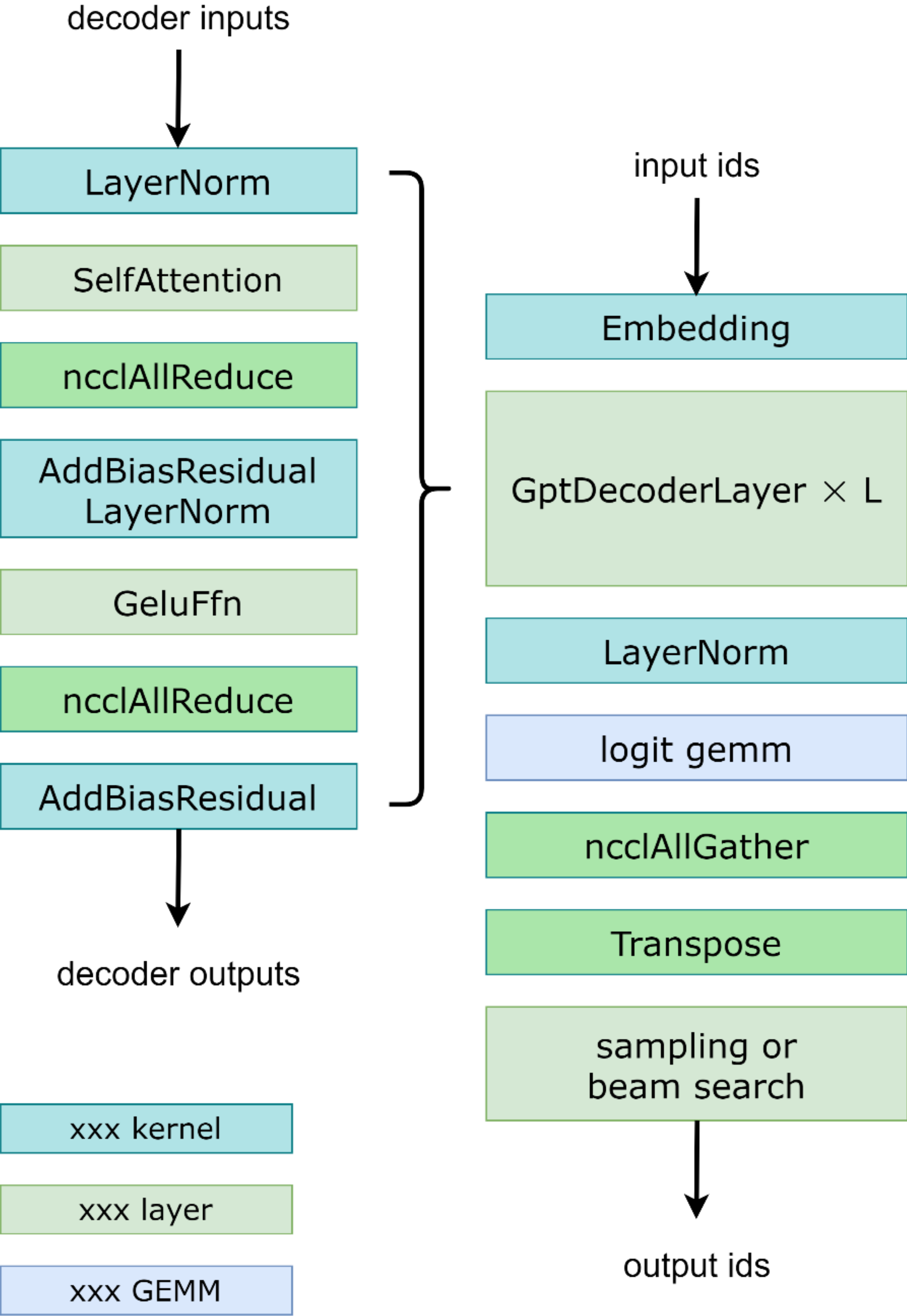
MNMG





# Distributed inference with fastertransformer

GPT with optimize transformer blocks

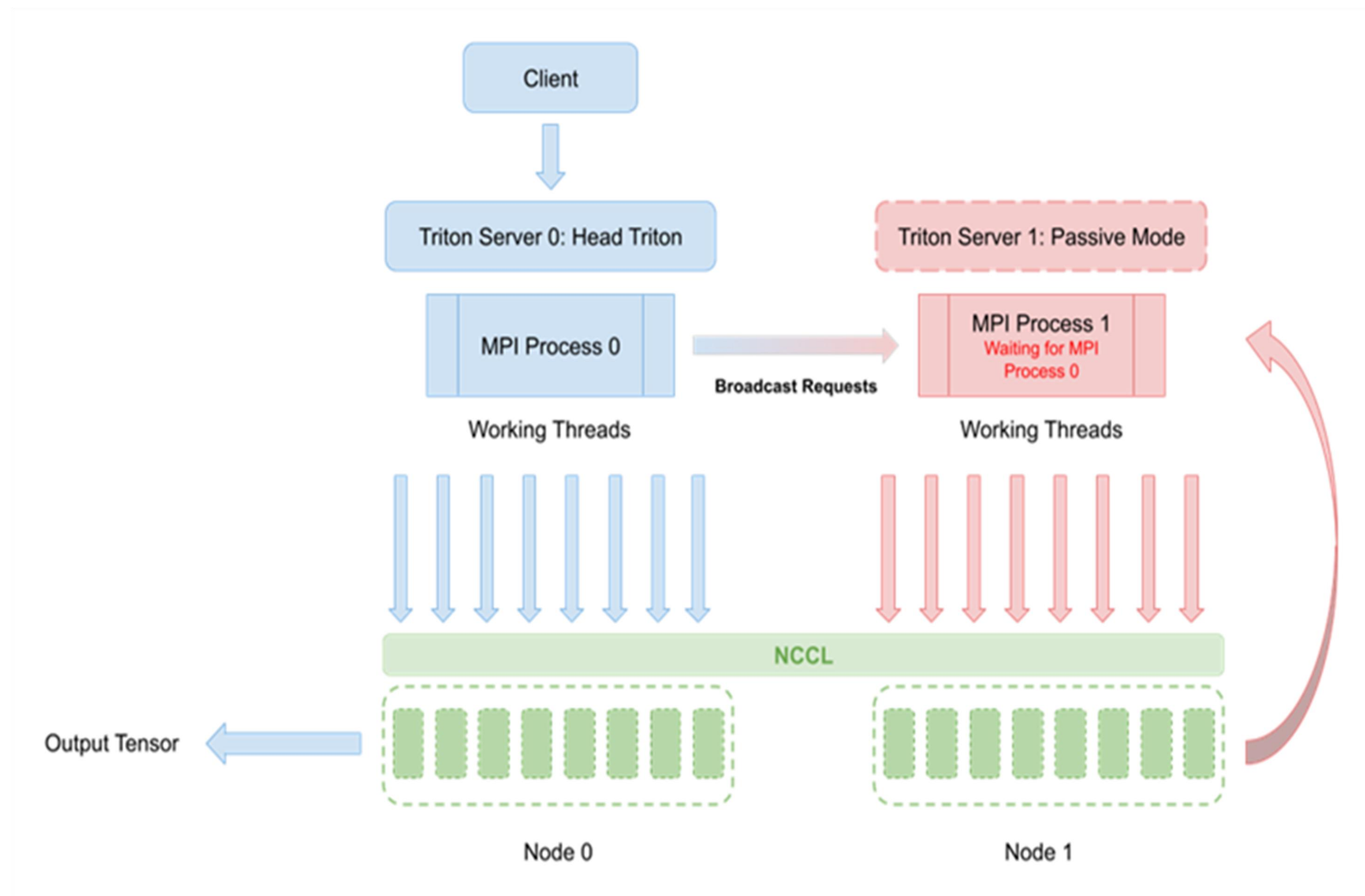




# Distributed inference with fastertransformer

Serve giant transformer models and accelerate inference

- Optimize kernels to accelerate inference for encoder/decoder layers of transformer models
- Integrated as a backend in Triton Inference Server
- Uses tensor/pipeline parallelism for multi-GPU, multi-node inference
- Uses MPI and NCCL to enable inter/intra node communication
- Supports BERT, GPT, T5, ViT and Swin-T style models
- Megatron, HuggingFace and ONNX converters provided





# Distributed inference with nemo

```
python3 FasterTransformer/examples/pytorch/gpt/utils/nemo_ckpt_convert.py \  
  --in-file /checkpoints/nemo_gpt1.3B_fp16.nemo \  
  --infer-gpu-num 1 \  
  --saved-dir /model_repository/gpt3_1.3b \  
  --weight-data-type fp16 \  
  --load-checkpoints-to-cpu 0
```

.....

```
python3 /export_scripts/prepare_triton_model_config.py \  
  --model-train-name gpt3_1.3b \  
  --template-path  
/opt/bignlp/fastertransformer_backend/all_models/gpt/fastertransformer/config.pbtxt \  
  --ft-checkpoint /model_repository/gpt3_1.3b/1-gpu \  
  --config-path /model_repository/gpt3_1.3b/config.pbtxt \  
  --max-batch-size 256 \  
  --pipeline-model-parallel-size 1 \  
  --tensor-model-parallel-size 1 \  
  --data-type bf16'
```



The background of the slide is a dynamic, abstract composition. It features a solid black field overlaid with numerous thin, bright green and yellow lines. These lines are mostly oriented diagonally, creating a sense of rapid movement or data flow. Some lines are straight, while others are curved or bundled together, resembling a complex network or a high-speed data stream. The overall effect is futuristic and technological.

# Customers Using NeMo Framework Today





## *Korean Language Models Powering:*

1. AI Contact Center - Cloud-based solution handling 100K calls/day without human intervention, reducing consultation times by 15 seconds.
2. Providing home assistant functions through IPTV, serving 8 Million families





*Accelerated NLP industry applications in Sweden by making the power of a 100-billion-parameter model for Nordic languages easily accessible to the Nordic ecosystem.*





**JD.COM**

*Improved downstream NLP tasks,  
like sentiment analysis, dialogue,  
and translation, by training custom  
Large Language Models using  
NeMo framework.*



- [Register here](#)
- [Find out more here](#)
- [NVIDIA Brings Large Language AI Models to Enterprises Worldwide | NVIDIA Newsroom](#)

#### DEVBLOGS and VIDEOS:

- [Adapting P-Tuning to Solve Non-English Downstream Tasks](#)
- [NVIDIA AI Platform Delivers Big Gains for Large Language Models](#)
- [Efficient At-Scale Training and Deployment of Large Language Models – GTC Session](#)
- [Hyperparameter Tool GTC Session](#)
- [Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model | NVIDIA Developer Blog](#)

#### CUSTOMER STORIES:

[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](#)

---

## Resources

Get Started



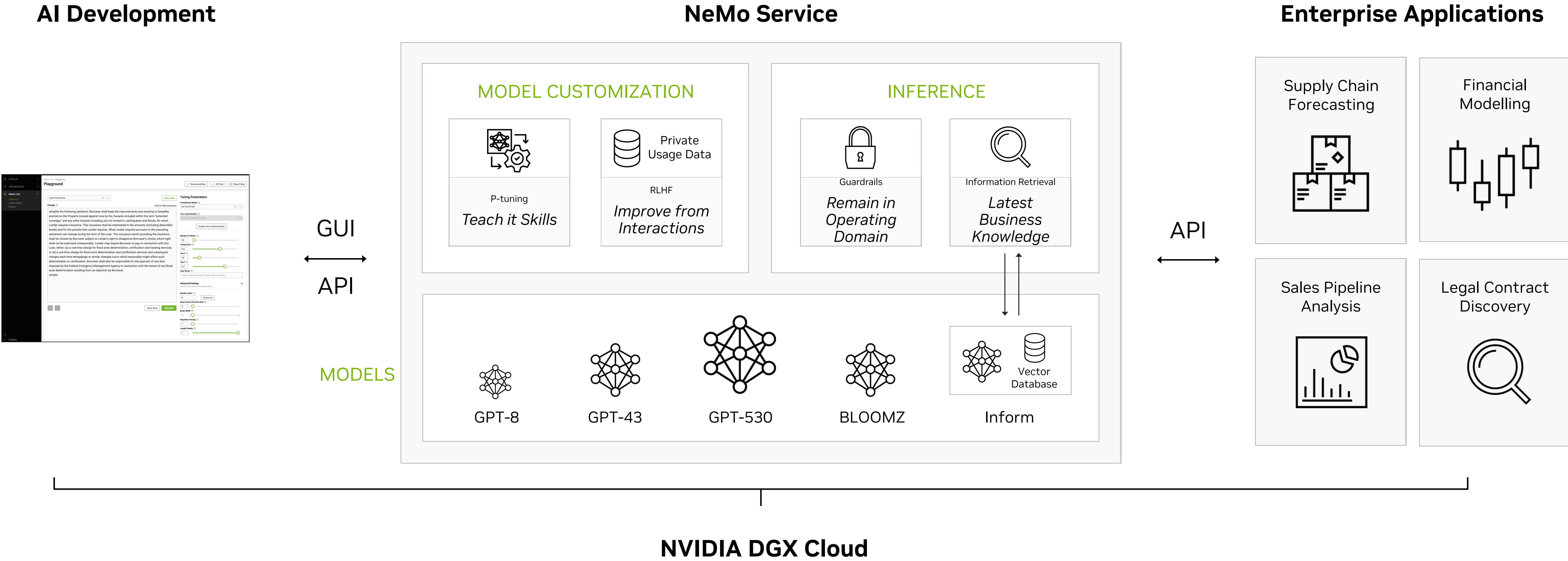


# Introduction to NVIDIA NeMo Service



# NVIDIA NeMo Service

Enterprise Hyper-Personalization and At-Scale Deployment of Intelligent Large Language Models



**Your Enterprise AI**  
Customize state-of-the-art pre-trained language models

**Easily Develop & Connect Applications**  
GUI-based Playground and Scalable Cloud API

**Deploy Anywhere**  
In the Service, Across Public Clouds, or On-Premises

**Enterprise Support**  
Fully supported by NVIDIA AI Experts from Customization to Deployment At-Scale



# NeMo Service – Key Features

## Customize Easily



---

Curated training techniques for enterprise hyper-personalization.

## Achieve Higher Accuracy



---

Best-in-class suite of AI foundation models design for customization, trained with up to 1T tokens.

## Run Anywhere



---

Run inference of large-scale custom models in the service or deploy across clouds or private data centers with [NVIDIA AI Enterprise software](#).

## Fastest Performance at Scale



---

State-of-the-art training techniques, tools, and inference—powered by [NVIDIA DGX™ Cloud](#).

## Ease of Use



---

Easily access the capabilities of your custom enterprise LLM through just a few lines of code or an intuitive GUI-based playground.

## Enterprise Support



---

Fully supported by NVIDIA AI experts every step of the way.



# Get Started with NeMo Service

**Apply Now**



## Web Pages

- [NVIDIA Generative AI Solutions](#)
- [NVIDIA NeMo Service](#)



## 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](#)



## GTC Sessions

- [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](#)



