NVIDIA

Introduction, Demystifying LLMs and Practical Considerations

26th May 2023
Ettikan Kandasamy Karuppiah (Ph.D)
Director/Technologist, Asia Pacific South Region

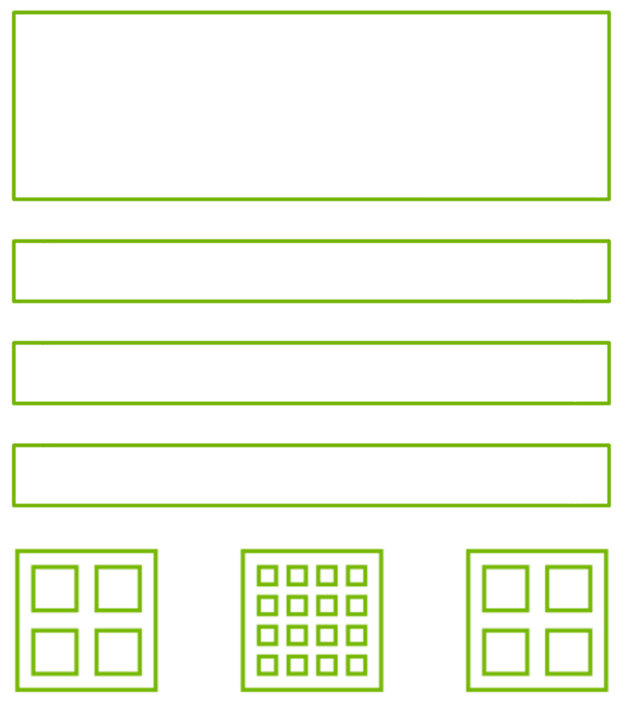
MODULUS
MONAI
RIVA
MAXINE
NEMO
MERLIN
CUOPT
MORPHEUS
TOKKIO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
AI APPLICATION FRAMEWORK
PLATFORMS
NVIDIA HPC
NVIDIA AI
NVIDIA Omniverse
ACCELERATION LIBRARIES
cuNumeric
CV-CUDA
cuQuantum
Parabricks
Sionna
JetPack
RAPIDS
Spark
cuDNN
cuGraph
TensorRT
Triton
DeepStream
Flare
DOCA
Mag IO
Aerial
CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS
RTX
DGX
HGX
EGX
OVX
Super POD
AGX
3 CHIPS
GPU
CPU
DPU

Full Stack, 3 Chips, Data Center Scale

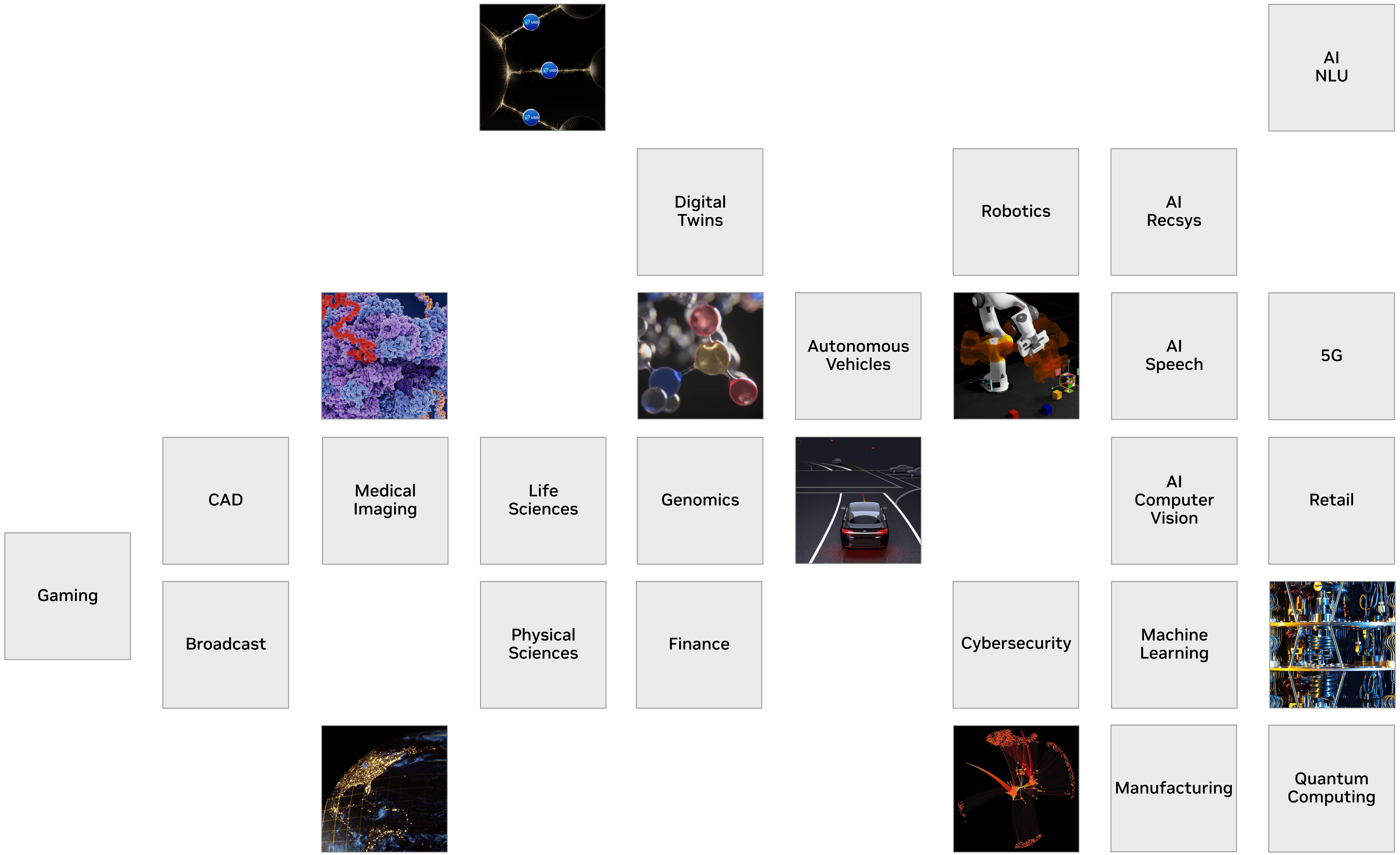
33 Million CUDA Downloads

150 SDKs

$100 Trillion Industry Served



NVIDIA
Accelerated Computing



The Opportunity of Generative AI

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

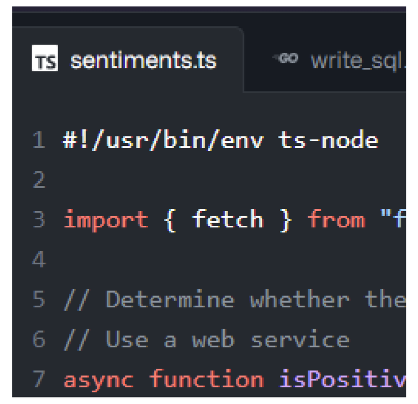
TRANSLATION

Translating Wikipedia

NLLB-200

Real-Time Metaverse Translation

CODING



Dynamic Code Commenting

CODEX

Function Generation

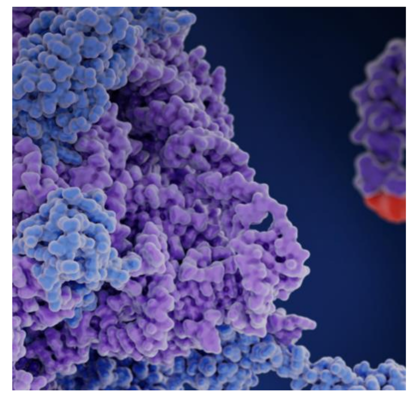
IMAGE GENERATION

Brand Creation

e-Diffi

Gaming Characters

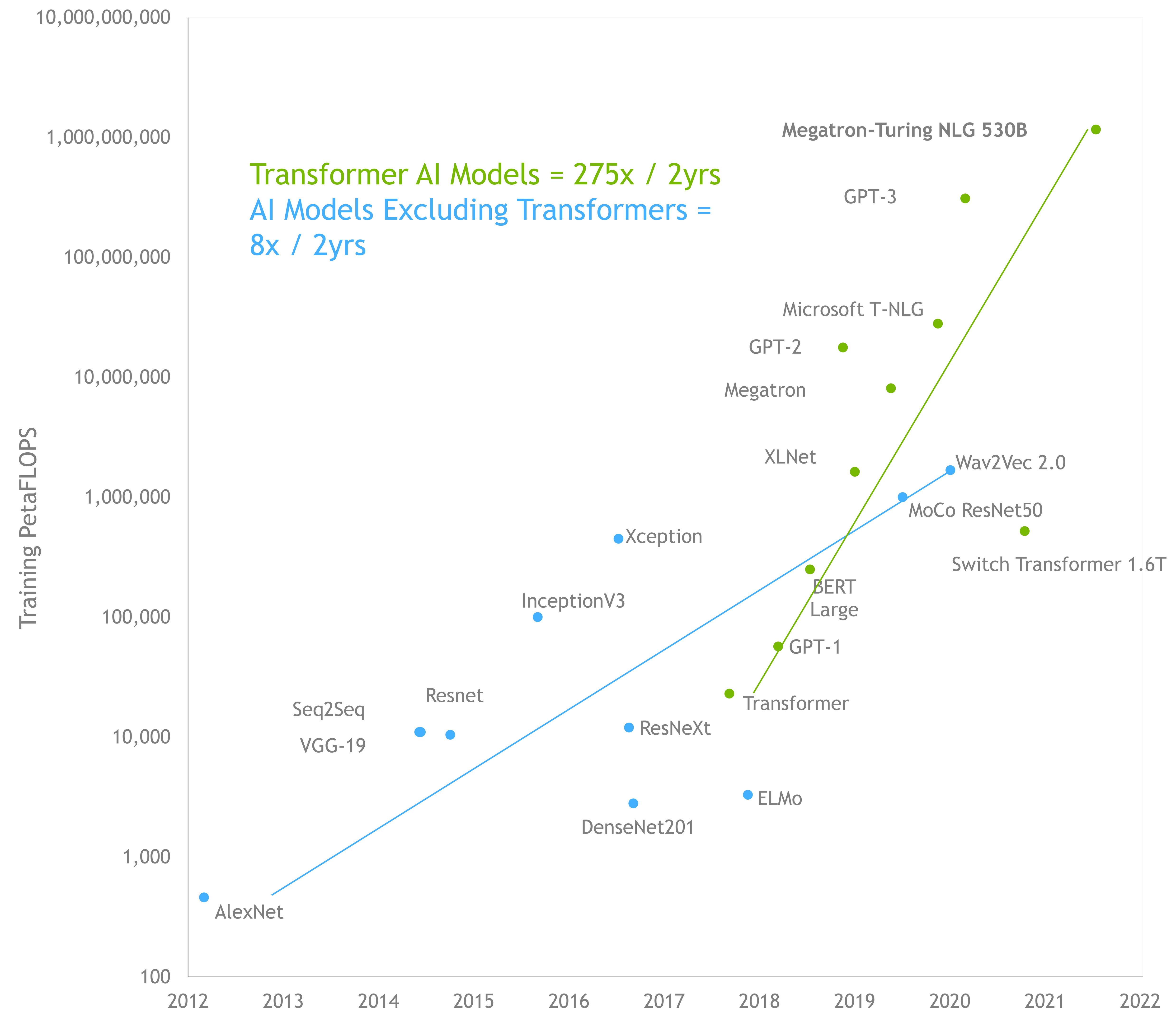
LIFE SCIENCE



Molecular Representation s

MegaMolBART

Drug Discovery



Next Wave of AI Requires Performance and Scalability

Exploding computation requirements

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

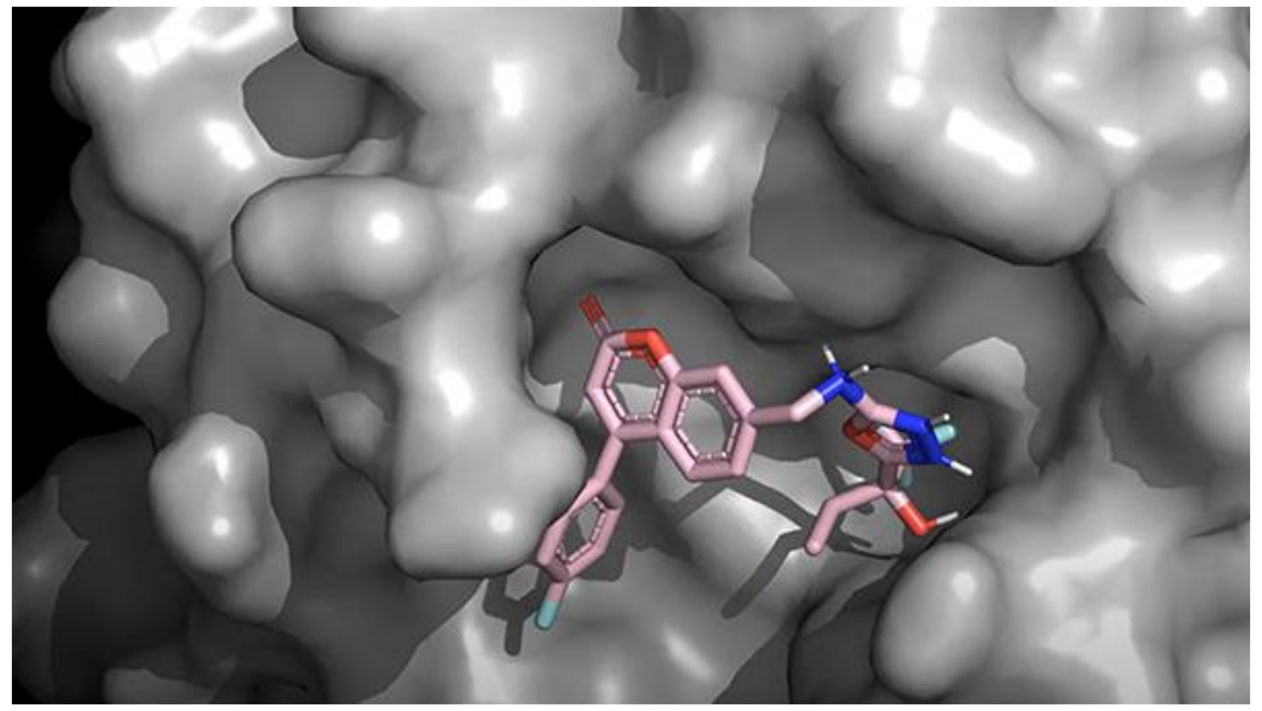
NVIDIA Generative AI Solutions

NVIDIA's Generative AI Solutions

Foundations to Build and Run Your Generative AI

NVIDIA NeMo service

NVIDIA BioNeMo service

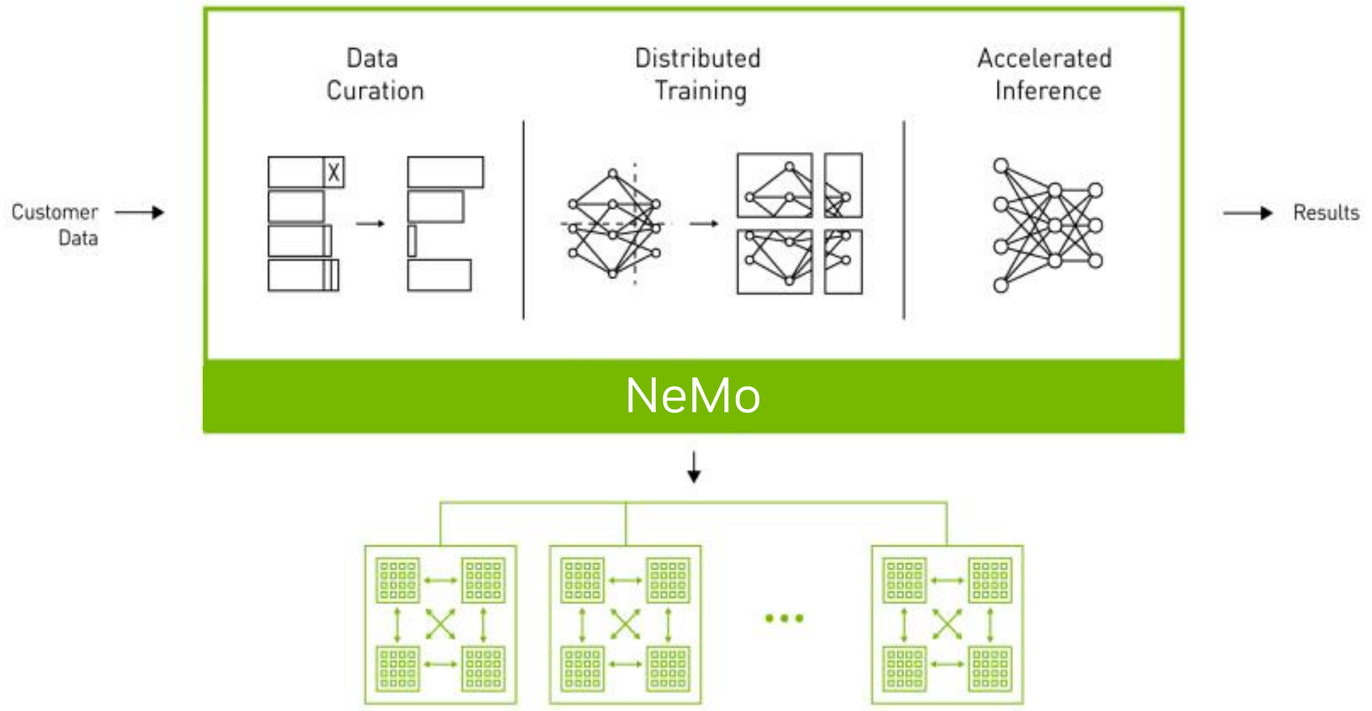


NVIDIA Picasso service

A photo of a golden retriever puppy wearing a green shirt. The shirt has text that says 'NVIDIA rocks". Background office. 4k dslr.

NVIDIA AI Foundations

NVIDIA NeMo framework

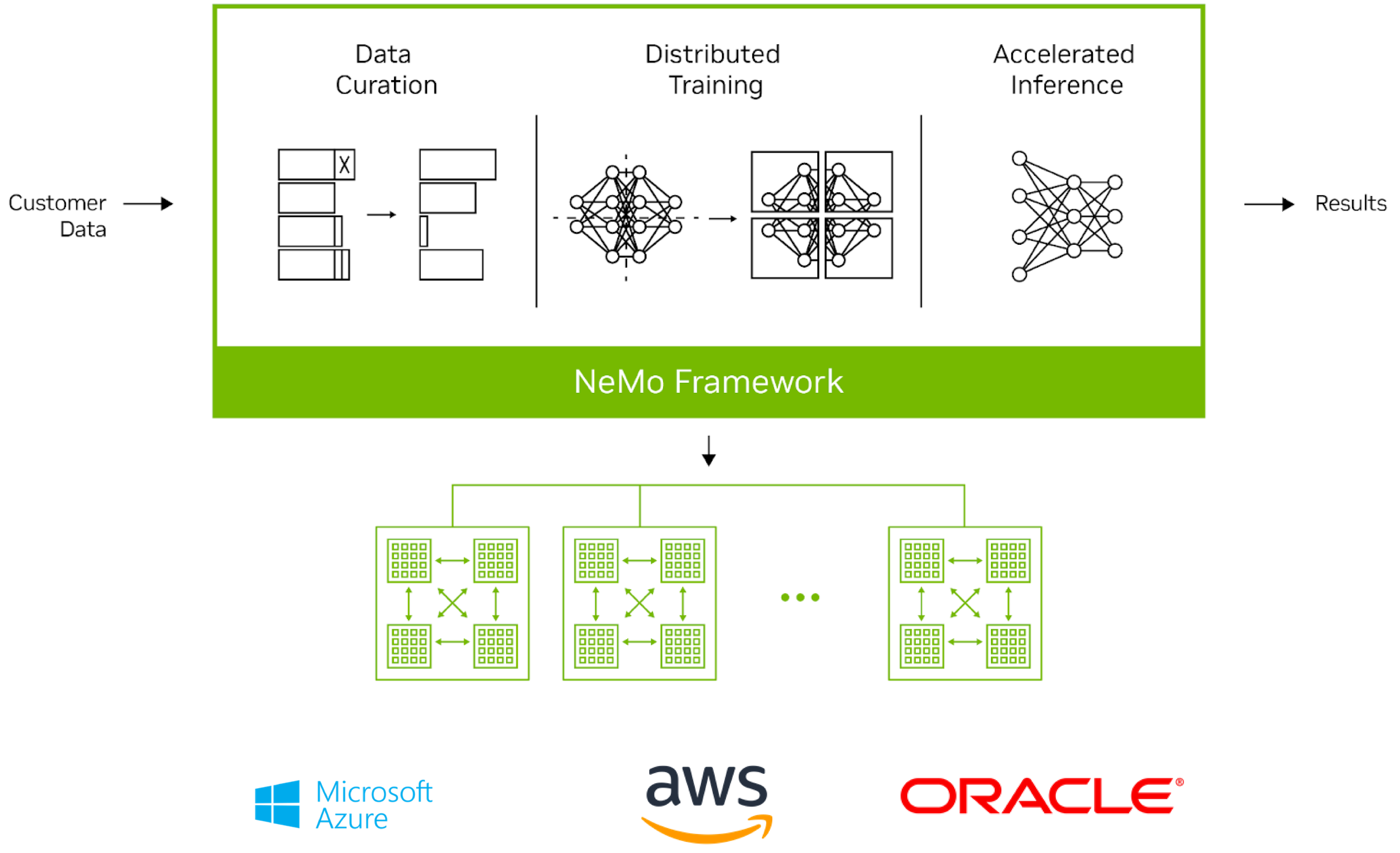


NVIDIA AI Enterprise

NVIDIA DGX Cloud

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

Solving pain-points across the stack

Unmet Needs		NeMo Megatron 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

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

Practical Considerations when working with LLMs

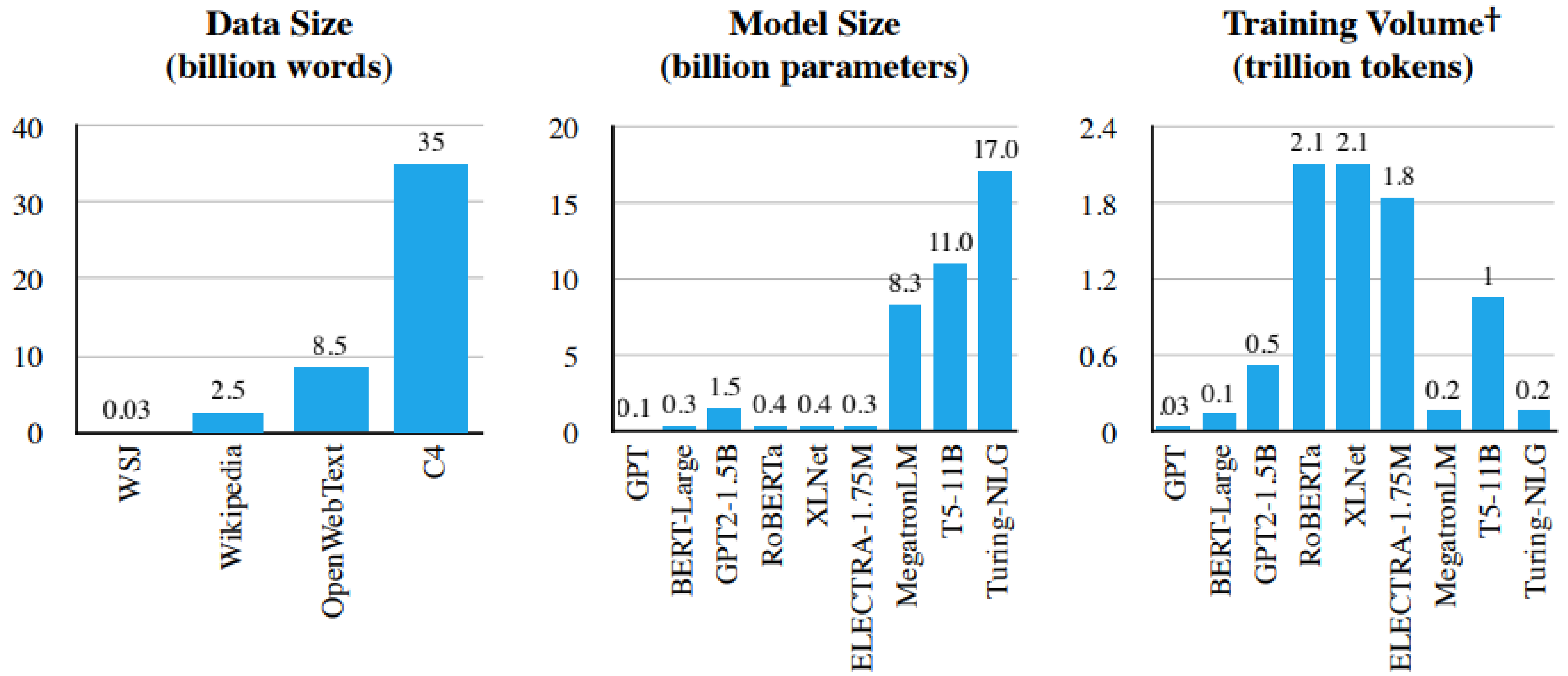
Optimise Training

WHAT DRIVES THE COST OF TRAINING LLMs

Directly Impacting The Cost Of Training

Direct Cost of Training

- **Size of the Dataset** (approximated by number of words)
- **Model Size** (approximated by the number of parameters)
- **Training Volume** (approximated by the number of tokens processed during pre-training)



The Cost Of Training NLP Models Or Sharir, Barak Peleg, Yoav Shoham AI2I Labs

Scheme	Number of parameters (billion)	Model-parallel size	Batch size	Number of GPUs	Microbatch size	Achieved teraFlOP/s per GPU	Training time for 300B tokens (days)
PTD Parallelism	174.6	96	1536	384	1	153	84
				768	1	149	43
				1536	1	141	23
	529.6	280	2240	560	1	171	156
				1120	1	167	80
				2240	1	159	42

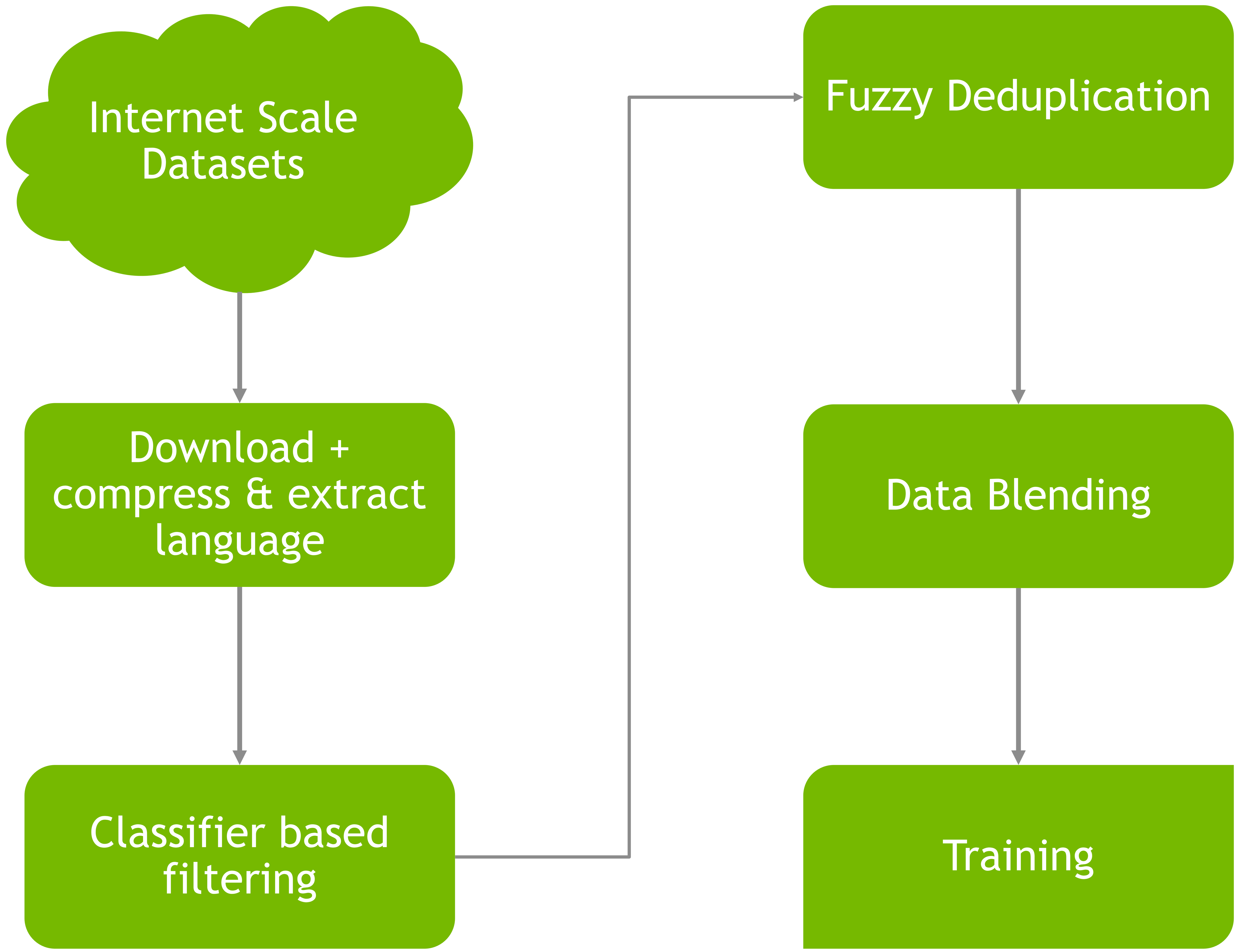
*Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM (arxiv.org)
Shows performance on NVIDIA A100 GPUs

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

Bring your own dataset to train LLMs

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



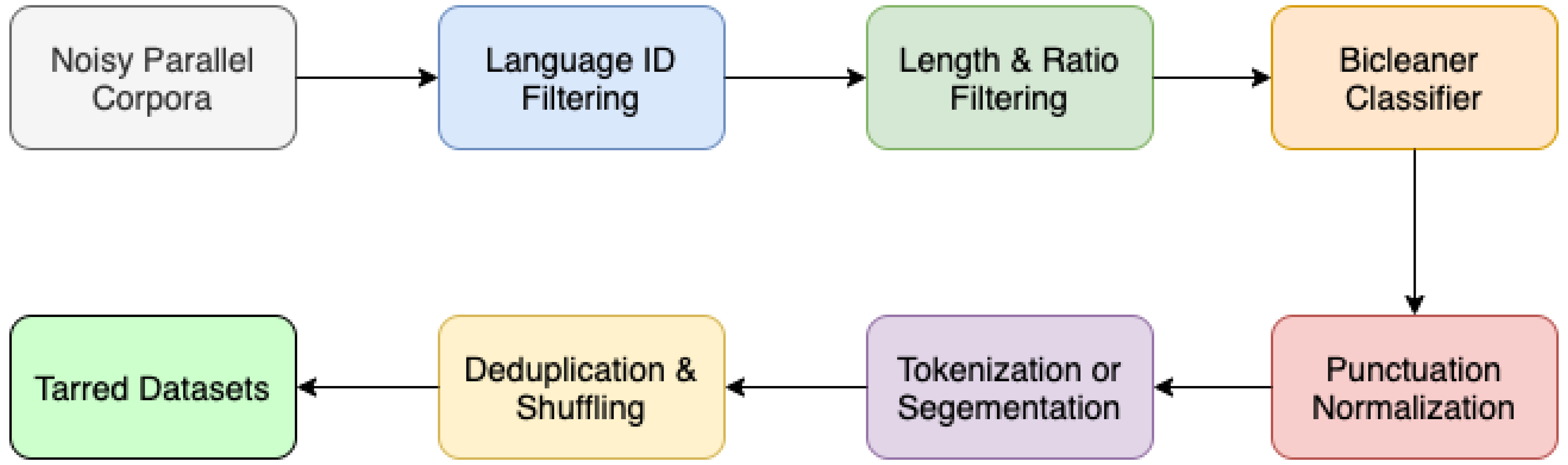
- Distributed processing leveraging DASK
  - DASK enabled auto load balancing for distributed processing
- De-duplication
- Data Cleaning-Bad Unicode, newline, repetition
- Extraction- HTML files and JavaScript

Data Curation & Preprocessing

Enabling Large-Scale High-Quality Datasets for LLMs

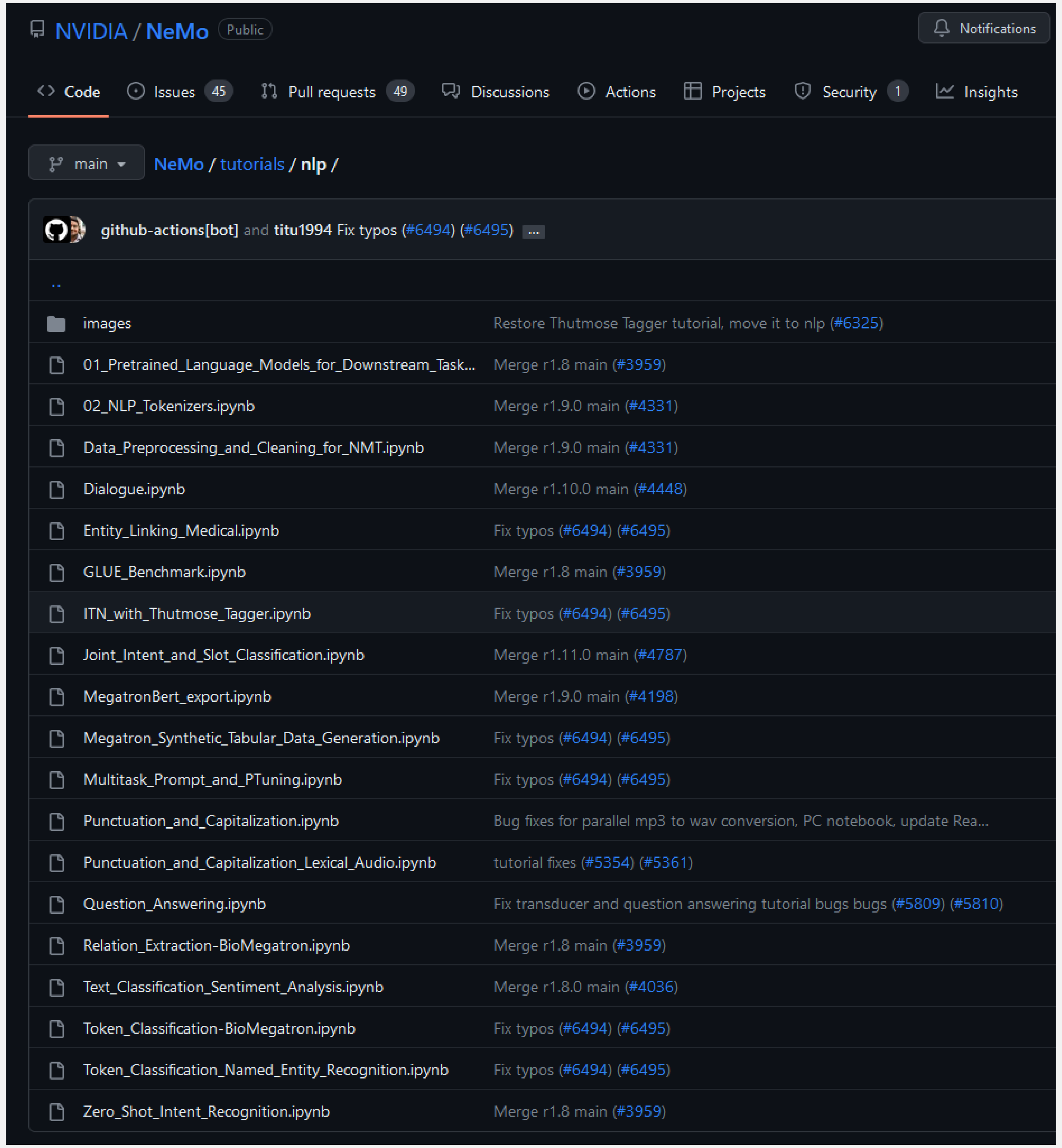
Sample Data Processing Workflow

Available on the NeMo Framework github



Scan the QR Code to visit:

[NeMo/tutorials/nlp at main · NVIDIA/NeMo · GitHub]



Reducing THE COST OF TRAINING LLMs

Existing Solutions Focus on Direct Cost Drivers

- **Improvement in Training Techniques**
  - Algorithmic Improvements such as DeepSpeed-MoE for NLG, Primer, etc.
  - Better Optimizers such as ZeRO, ZeRO-Infinity
  - Better Memory Utilization such as Gradient, Activation Checkpointing, etc.
  - Better Parallelization and Distributed Training such as 3D parallelism Megatron-Turing NLG 530B
  - …
- **Innovations in Hardware**
  - Tensor Cores with NVLink and NVSwitch, HDR InfiniBand Networking
  - Targeting Structured Sparsity present in the networks
  - …
- **Framework level optimizations**
  - PyTorch DistrubtedDataParallel and Best Practices
- ***… and so on***

Current methods do not address inefficiencies in the training process arising from **Slow and Manual Hyperparameter Search** and **Multiple exploratory runs to meet the target performance and scaling efficiency** that can be completely avoided with limited search scope and initial heuristics

WHAT DRIVES THE COST OF TRAINING LLMs

Extremely High Cost Of Experimentation

Hidden Cost of Training LLMs

- Besides the direct cost drivers such as Dataset Size, Model Size and Training Volume, there are inefficiencies in how we approach experiments that escalate the overall cost of experimentation

Multiple Runs To train Meaningful Models

- Multiple runs to arrive at stable hyperparameter configuration
- Multiple runs to achieve high scaling efficiency during training
- Multiple inference tests to achieve high throughput and low latency
- Repeat the above steps multiple times for different model sizes
- …
- This reduces or eliminates the scope of making corrections or changes to an already trained model

- What is the right model size for my Hardware?
- Which hyperparameters are the most sensitive to convergence?
- How should I improve the throughput of my training runs?
- Which hyperparameters should I change when we add more GPUs
- How can I use my existing compute optimally
- … *and so on*

REDUCING THE COST OF TRAINING LLMs

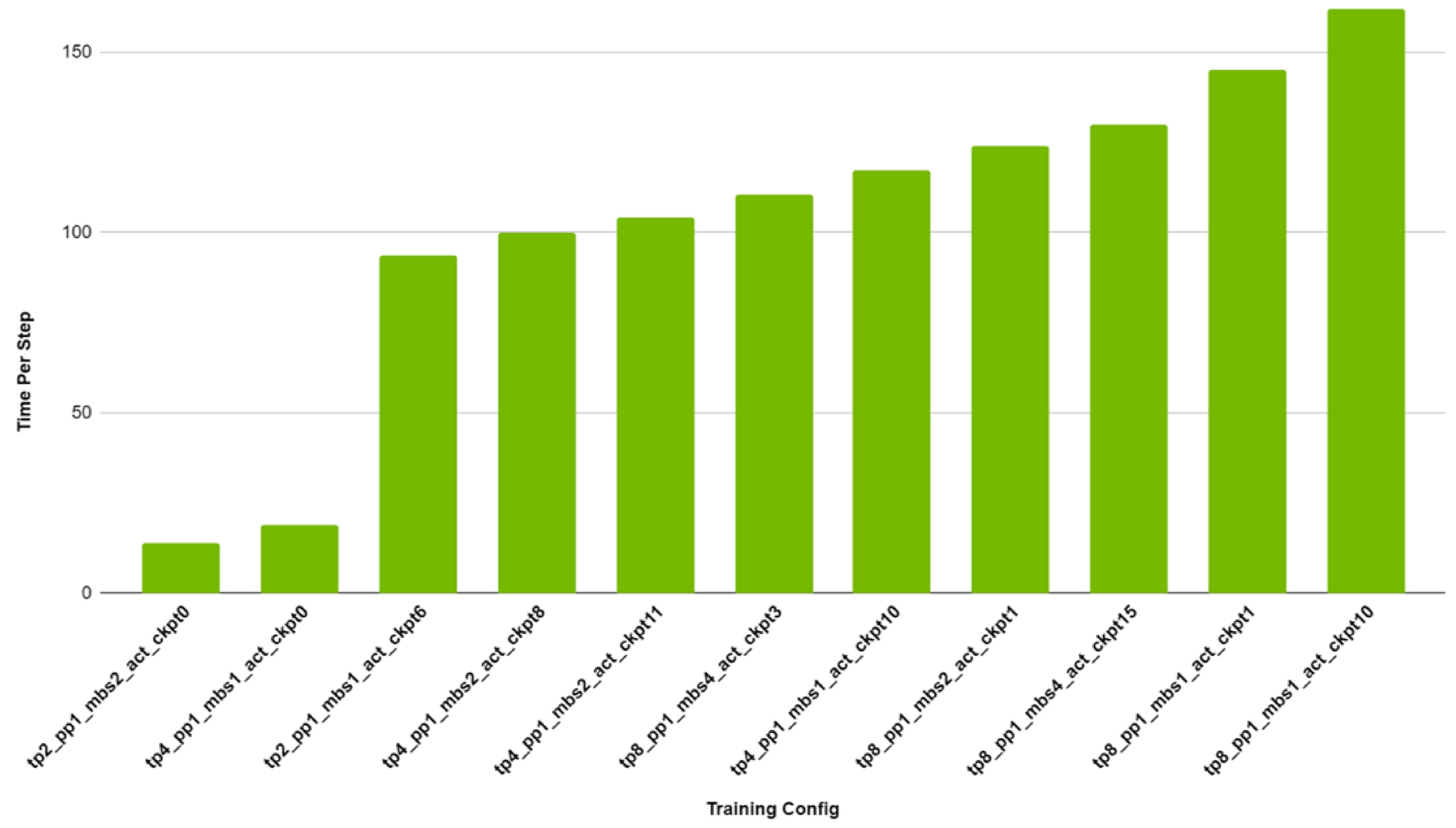
Targeting the hidden cost of training Large Language Models

Performance Speedups achieved using the NeMo-Megatron HP selection utility for the 5B parameter GPT-3 model

The chart below shows an 11.61x speedup achieved between the best (tp2_pp1_mbs2_act_ckpt0) i.e. Tensor Parallel = 2, Pipeline Parallel = 1, Micro Batch Size = 2 and number of activation checkpoint layers = 0 and the worst model configuration (tp8_pp1_mbs1_act_ckpt10) i.e. Tensor Parallel = 8, Pipeline Parallel = 1, Micro Batch Size = 1 and number of activation checkpoint layers = 10.

Selecting the correct parallelism values, micro batch size and activation checkpoint layers can have a huge impact on the training speed.

5B GPT-3 Model: 11.61x training speedup



Performance speedups achieved using NeMo-Megatron Hyperparameter Search tool for a 5B GPT3 model trained on DGX A100 GPUs

A poor hyperparameter config can slow down the model training by several days.
This is very expensive on a shared cluster with multiple users, as well as considering the expense of each run of model training

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

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

Tuning

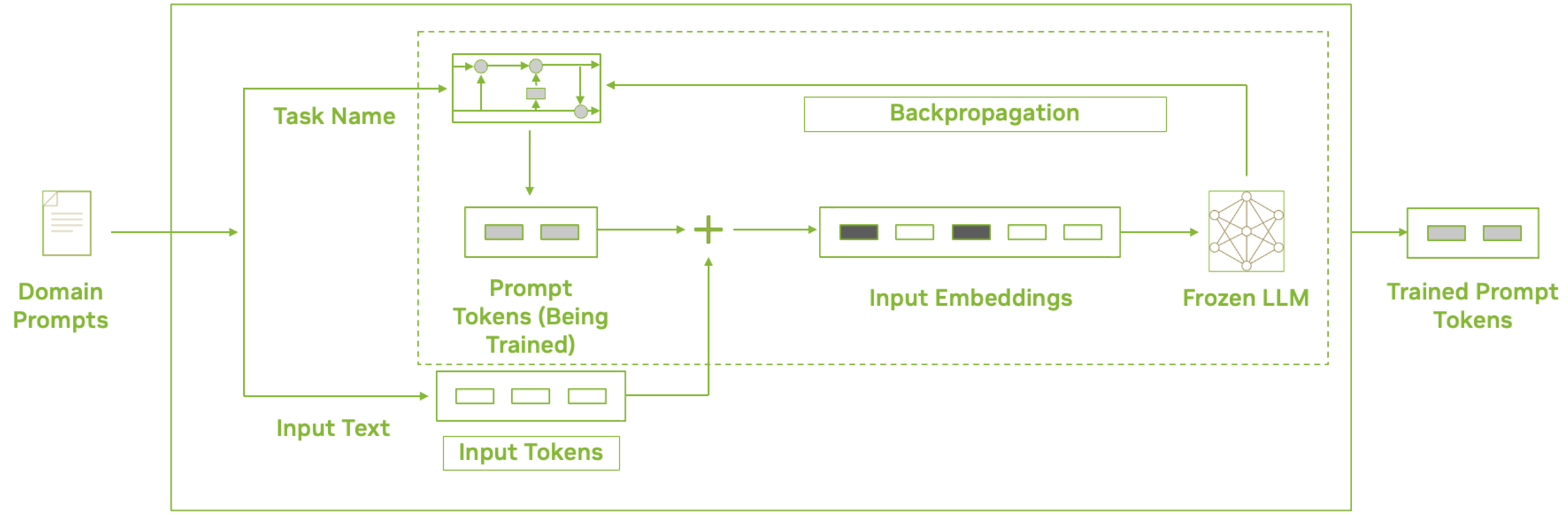
Overcoming Challenges Of Using Foundation Model

Generalized AI will not work – Enterprise need their own AI

Answer proprietary information	Update knowledge base with latest information	Factual correctness with specific context, domain & voice	Bias & toxicity management

Provide Context to Models

Parameter efficient ways to customize LLMs for specific use-cases



- Prompt learning includes both p-tuning, and prompt tuning
- 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
- Domain prompts include task name, “prompt”, and desired output
  - Ex: Q&A, “What are the rental options?”, Answer: *“We offer Economy, Compact, & Full-Size vehicles for rental”*

Prompt Learning Capabilities

Customize Models using SOTA prompt learning techniques

Customization is Required to Address Business-specific Tasks

Zero-Shot Response

"What is the yellow part in an egg?"

LLM

"This is the part that suspended in the center of the egg."

P-Tuned Response

"What is the yellow part in an egg?

+

Trained Prompts (Context)

LLM

Nutrition Chatbot

"The yellow part in an egg is the yolk. It contains fat, cholesterol, and protein."

Prenatal Chatbot

"The yellow part in an egg is rich in choline, which is important for fetal brain development"

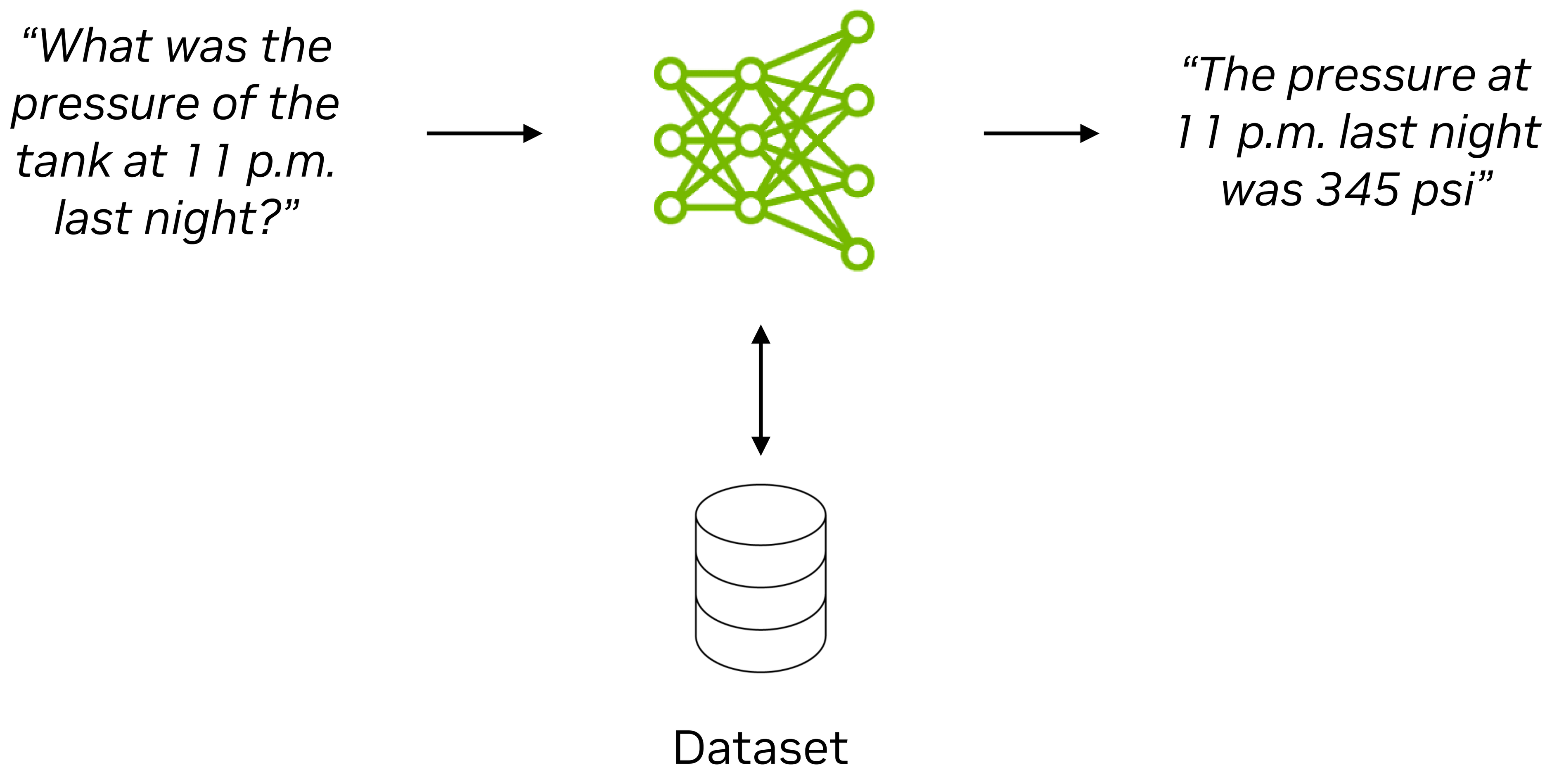
Culinary Chatbot

"The yellow part in an egg is used to fortify sauces and salad dressings, and to emulsify rich, fatty, ingredients like oil and butter"

Enterprises Require Responses Based on Current Information

Foundation/Custom Model Response

Information Retrieval Model Response



70%

Of Enterprise Data is Untapped

Unlock new opportunities for greater intelligence

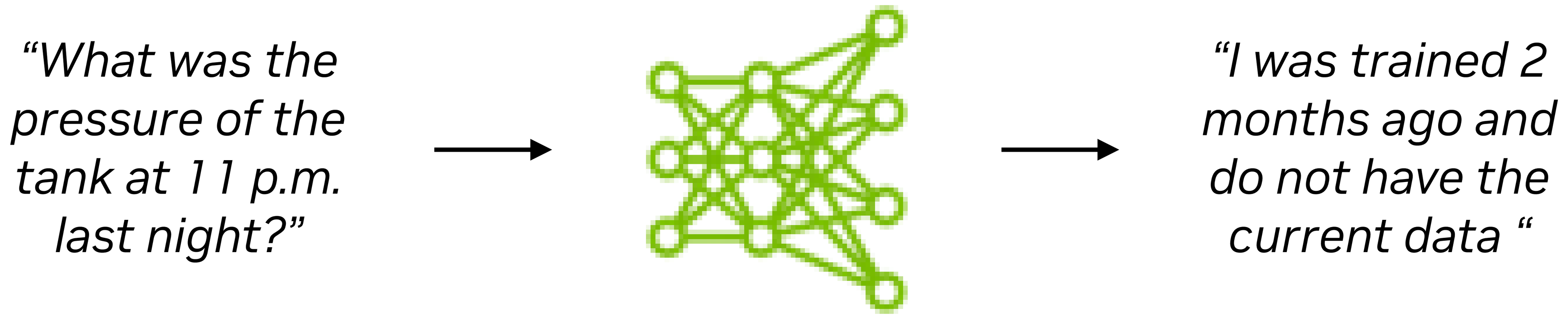
Less Frequent Re-Training

Significant cost and time savings to maintain LLMs

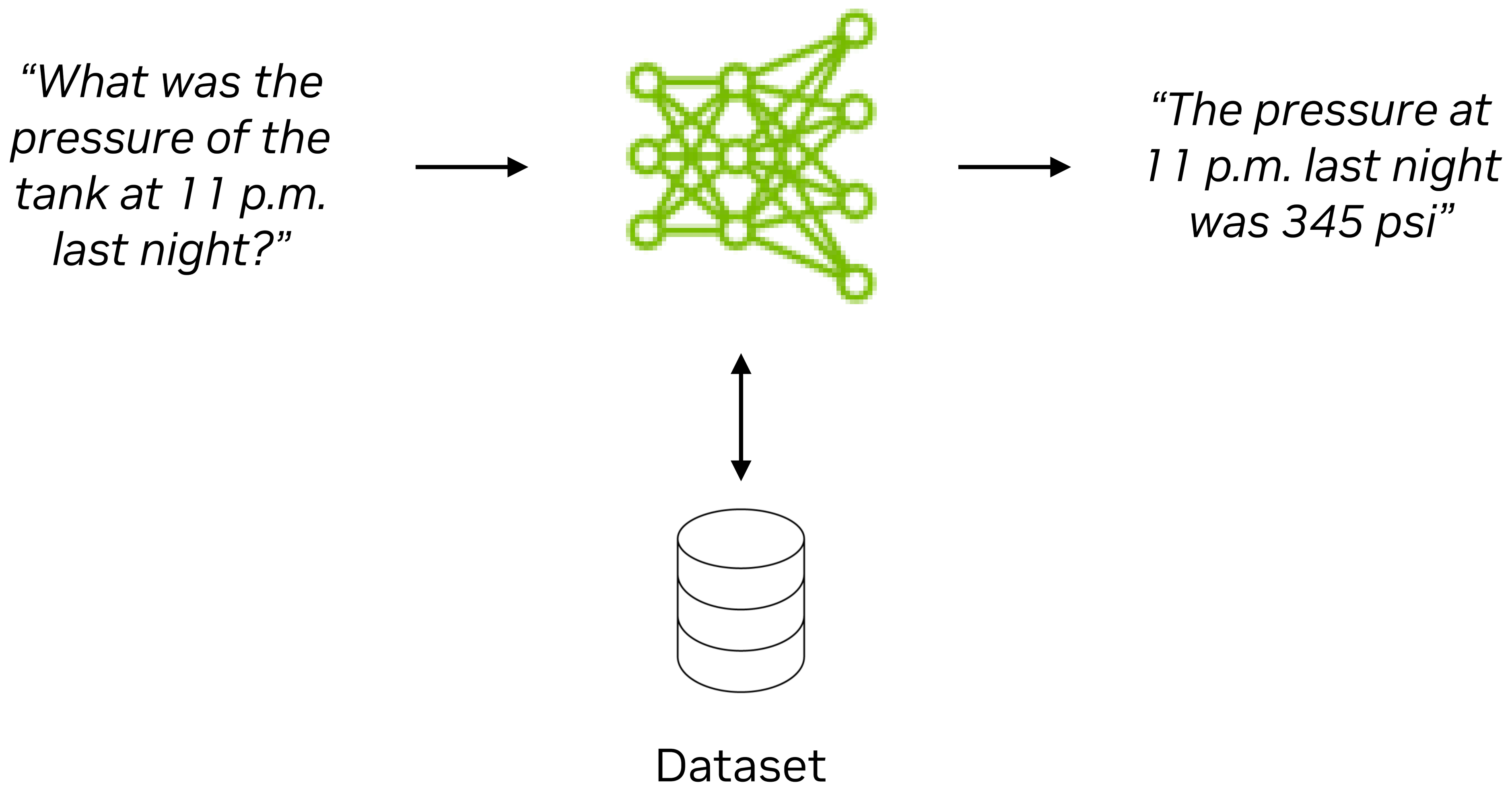
Enterprise Use-Cases Require Domain Specific Knowledge

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

Foundation/Custom Model Response



Inform Model Response



70%

Of enterprise data is untapped

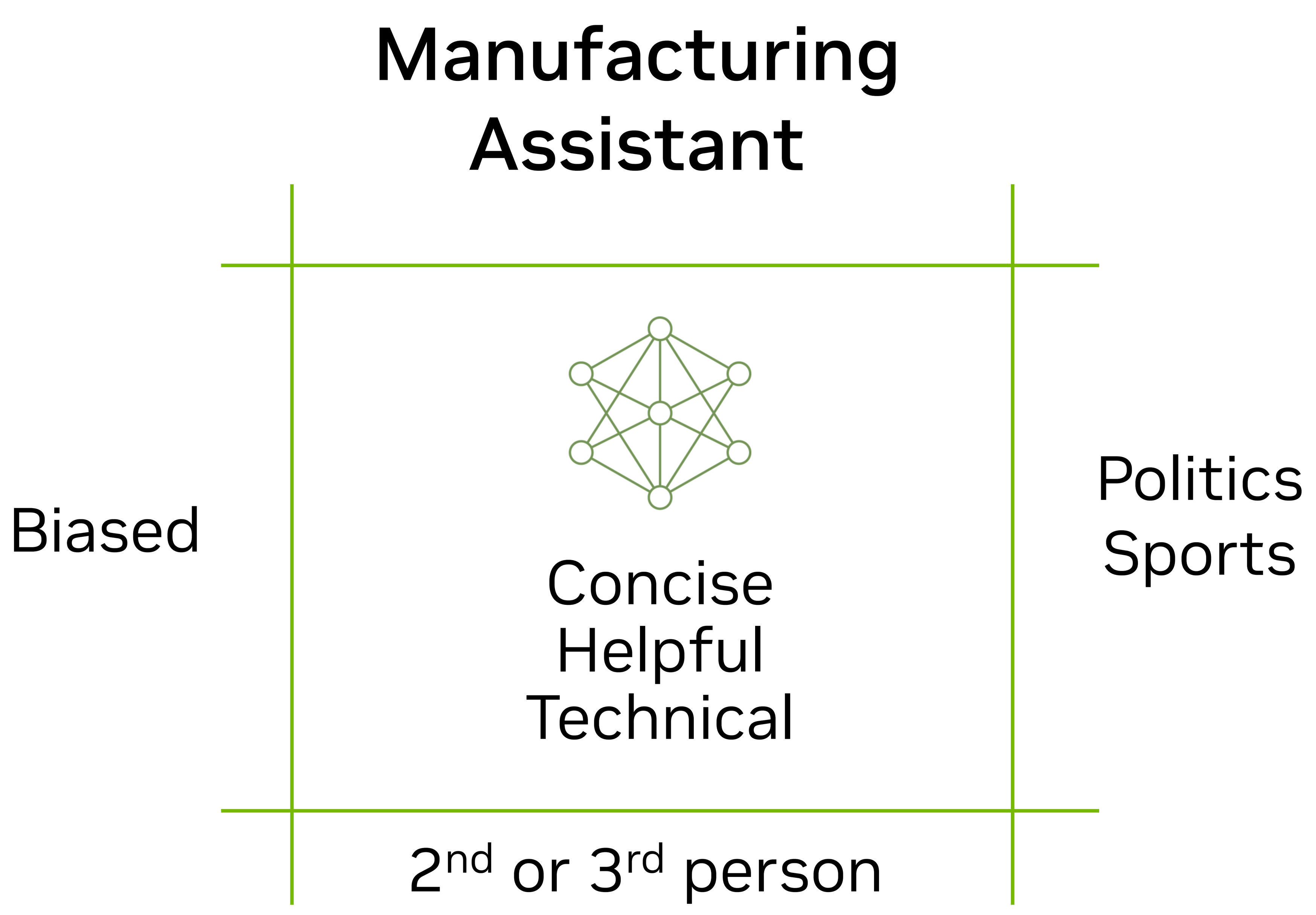
Unlock many new opportunities for greater intelligence

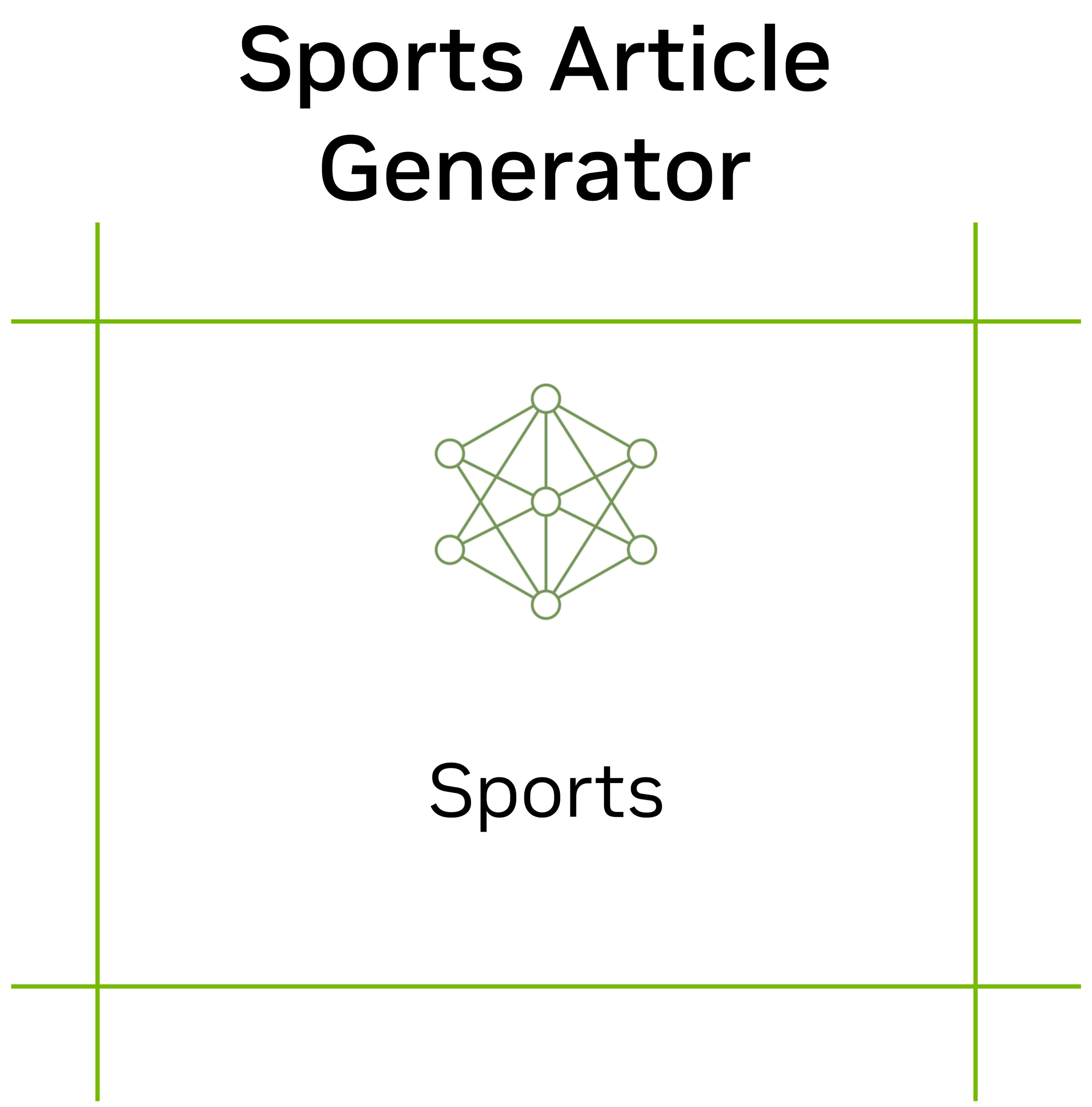
Less Frequent Re-Training

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

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

Customization Definitions and Examples

Zero Shot

Asking the foundation model to perform a task with no previous example or knowledge

Few Shot

Providing a couple examples to the foundation model before giving it a task

P-Tuning

Training a "prompt-model" with 100s to 1000s of examples to foundation model at inference time

Fine-Tuning

Re-training layers of the foundation model with specific datasets

Customization Definitions and Examples

Zero Shot

Asking the foundation model to perform a task with no previous example or knowledge

Few Shot

P-Tuning

Fine-Tuning

530B with no examples

Summarize the following article.

Article:Building, maintaining, and deploying these models require expansive amounts of training data, AI expertise, and compute.

Enterprises can fast-track their generative AI adoption with NVIDIA NeMo service running on NVIDIA DGX Cloud. The quickest path is starting with one of NVIDIA's state-of-the-art pre-trained foundation models.

With the NeMo service, organizations can easily customize a model with p-tuning to teach it specialized skills, like summarizing financial documents, creating brand-specific content, and composing emails with personalized writing styles.

Connecting the model to a proprietary knowledge base ensures that responses are accurate, current, and cited for their business.

And NeMo's playground is available for rapid prototyping before moving to the cloud API for larger-scale evaluation and application integration.

Summary:

The article talks about the importance of AI in the business world.

Customization Definitions and Examples

Zero Shot

Asking the foundation model to perform a task with no previous example or knowledge

Few Shot

Providing a couple examples to the foundation model before giving it a task

P-Tuning

Fine-Tuning

530B with 1 example provided

English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.
Summary: The Transformer architecture based solely on the attention mechanism deliver superior quality on several translation tasks while being more parallelizable and requiring significantly less time to train compared to recurrence and convolution alternatives.
===
Summarize the following article:
Article: Building, maintaining, and deploying these models require expansive amounts of training data, AI expertise, and compute. Enterprises can fast-track their generative AI adoption with NVIDIA NeMo service running on NVIDIA DGX Cloud. The quickest path is starting with one of NVIDIA's state-of-the-art pre-trained foundation models.With the NeMo service, organizations can easily customize a model with p-tuning to teach it specialized skills, like summarizing financial documents, creating brand-specific content, and composing emails with personalized writing styles. Connecting the model to a proprietary knowledge base ensures that responses are accurate, current, and cited for their business. NeMo's playground is available for rapid prototyping before moving to the cloud API for larger-scale evaluation and application integration.

Summary: NVIDIA's NeMo service allows enterprises to build, customize, and deploy generative AI models with minimal AI expertise and compute.

Customization Definitions and Examples

Zero Shot

Asking the foundation model to perform a task with no previous example or knowledge

Few Shot

Providing a couple examples to the foundation model before giving it a task

P-Tuning

Training a "prompt-model" with 100s to 1000s of examples to foundation model at inference time

Fine-Tuning

20B P-tuned on summarization

Summarize the following article.
Article:Building, maintaining, and deploying these models require expansive amounts of training data, AI expertise, and compute.
Enterprises can fast-track their generative AI adoption with NVIDIA NeMo service running on NVIDIA DGX Cloud. The quickest path is starting with one of NVIDIA's state-of-the-art pre-trained foundation models.
With the NeMo service, organizations can easily customize a model with p-tuning to teach it specialized skills, like summarizing financial documents, creating brand-specific content, and composing emails with personalized writing styles.
Connecting the model to a proprietary knowledge base ensures that responses are accurate, current, and cited for their business.
And NeMo's playground is available for rapid prototyping before moving to the cloud API for larger-scale evaluation and application integration.
Summary:

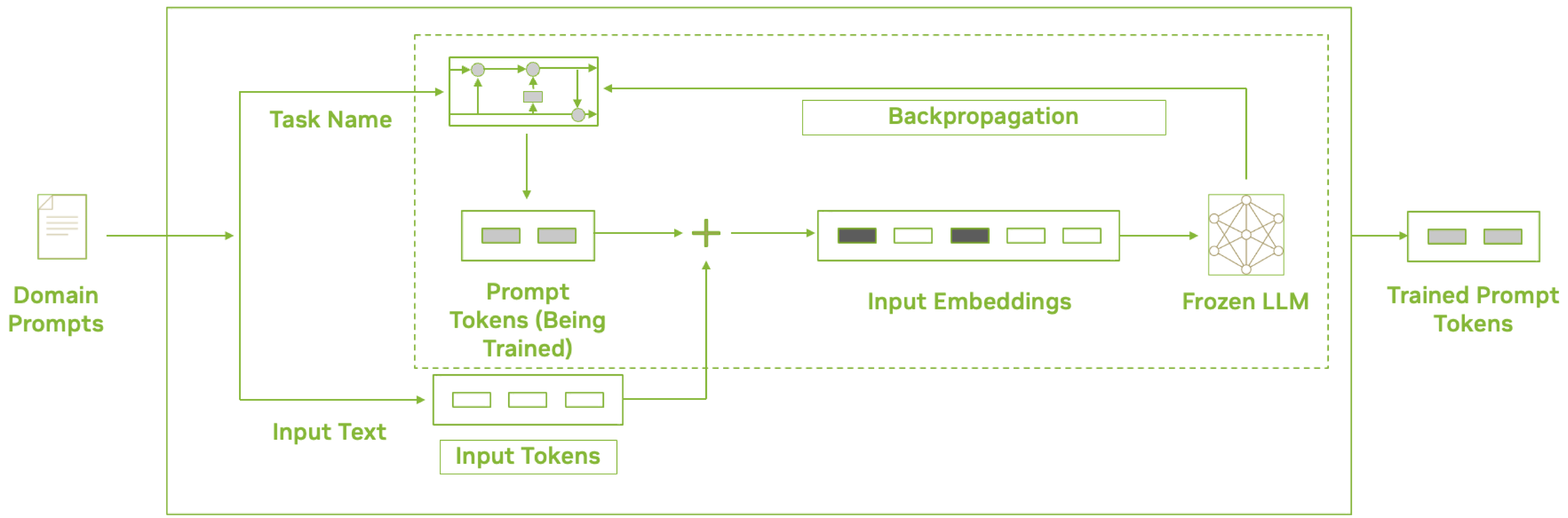
Enterprises can fast-track their generative AI adoption with NVIDIA NeMo service running on NVIDIA DGX Cloud.

Customization Definitions and Examples

Zero Shot

Few Shot

P-Tuning



Fine-Tuning

Re-training layers of the foundation model with specific datasets

An Example Reference Architecture

NeMo Inform

LLM-based Information Retrieval Service for the Enterprise

Thank you!

Q&A