Australia LLM Bootcamp

September 2023

Agenda

- Connecting to the Cluster
- NeMo Framework – a brief recap
- Fundamentals of NeMo – Lab
- Question Answering – Lab
- Multitask Prompt and P-tuning - Lab
- Megatron GPT 1.3B Text Generation Server Inferencing - Lab
- Challenge

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NLP-LLM Bootcamp User Guide

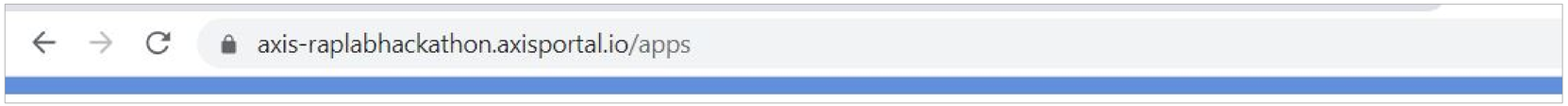
Last update June 23rd, 2023

Cluster Login Instructions

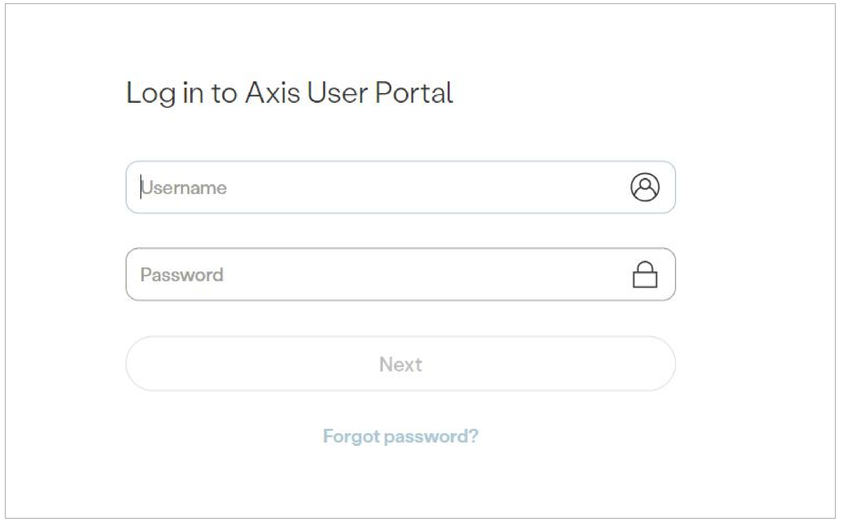
Login Instructions

Login Credentials

- Step 1:
  - Open the browser and enter the login URL: **https://axis-raplabhackathon.axisportal.io/apps**



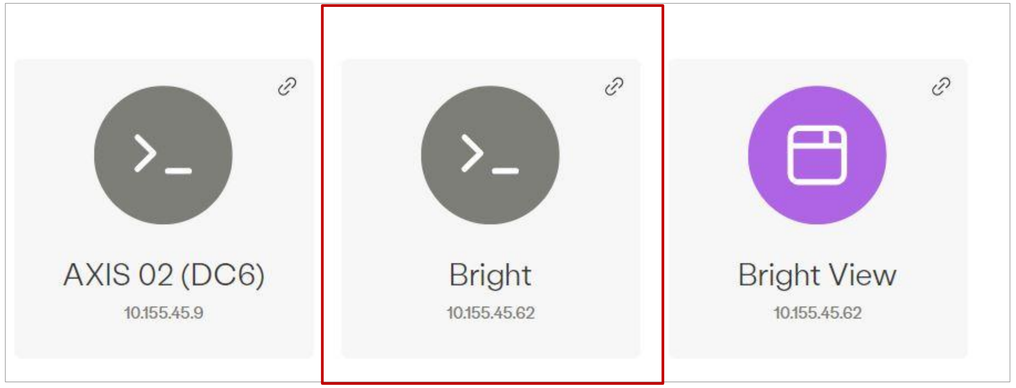
- Step 2:
  - Enter your cluster login credentials



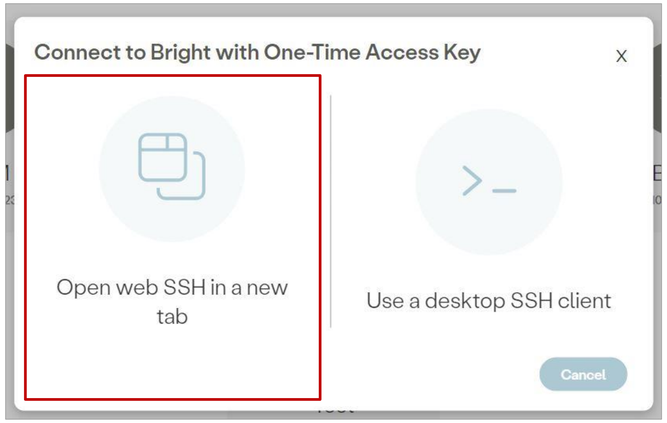
Login Instructions

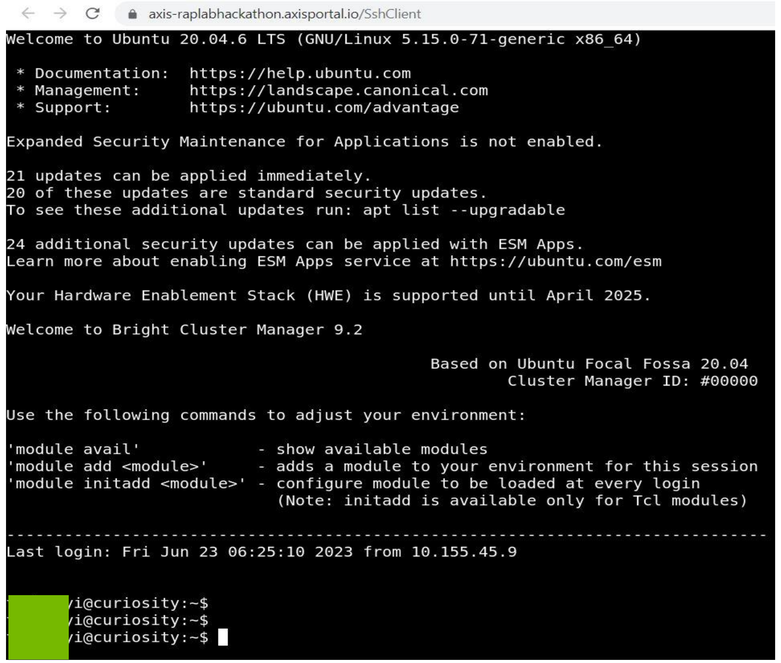
Select Bright option

- Step 3:
  - Click on **Bright**



- Step 4:
  - Click on **Open web SSH in a new tab**

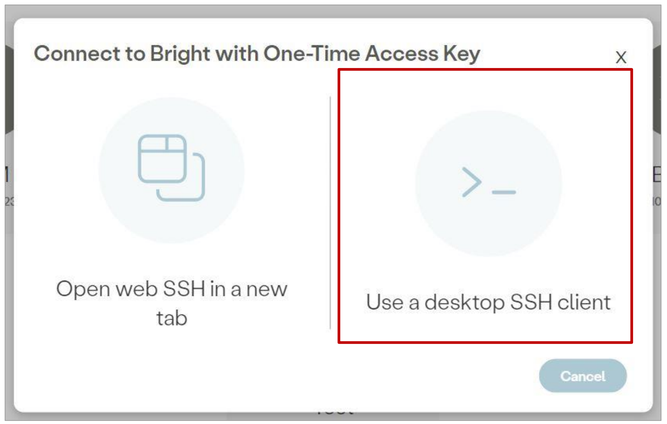


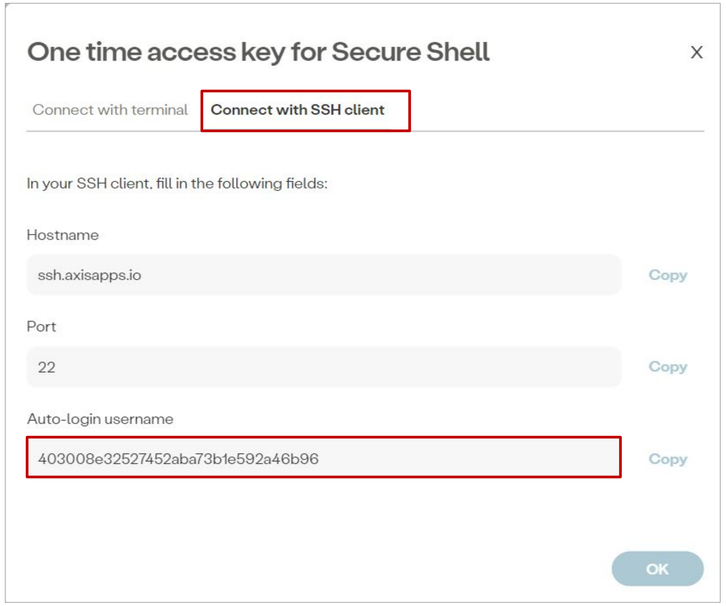


Login Instructions

Copy the Hashcode

- Step 5:
  - Click on the option **Use a desktop SSH client.** Click on **Connect with SHH client** and copy the hashcode





Running the NeMo Megatron Lab

NeMo Megatron-GPT Lab

Running the Lab Script and Port-forwarding (1/2)

- The Nemo Megatron lab contains the following hands-on notebooks:
  1) Nemo Fundamentals
  2) Question Answering
  3) Prompt-Tuning/P-Tuning
  4) The Megatron-GPT Language Model Inferencing
  5) Challenge

Steps:

- From the browser terminal, run the sbatch script and press on the Enter key to submit job:

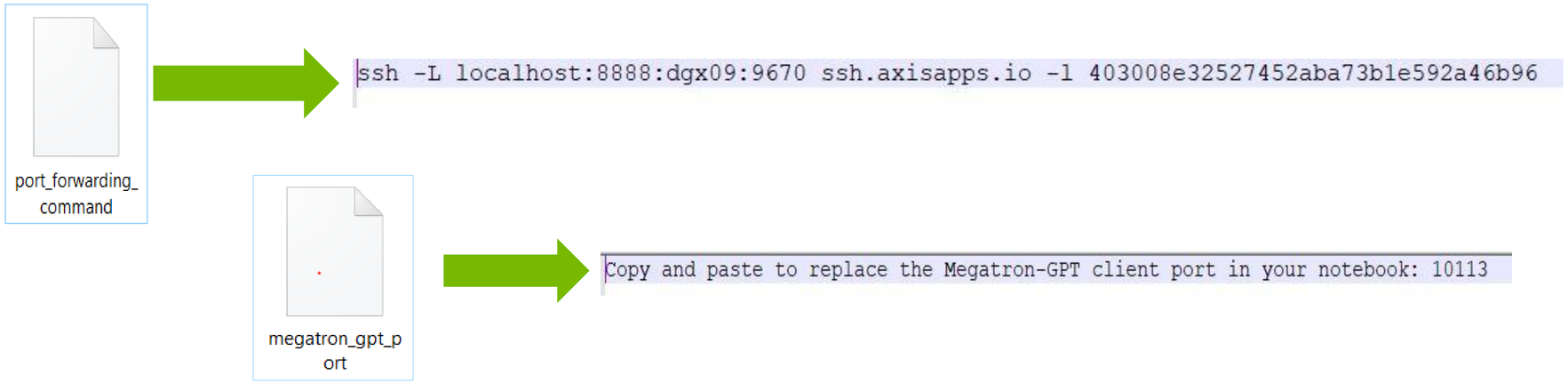
sbatch /bootcamp_scripts/NLP/**sbatch-llm** <hashcode>

```
yi@curiosity:~$ sbatch /bootcamp_scripts/NLP/sbatch-llm cad857f39ec14668b098892bef85cc3
```

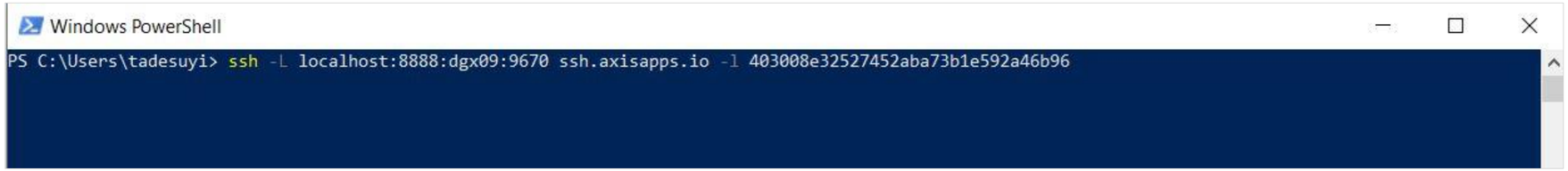
NeMo Megatron-GPT Lab

Running the Lab Script and Port-forwarding (2/2)

- In approximately **8mins** a port forwarding (port_forwarding_command) and a megatron-gpt inferencing port (megatron_gpt_port) files will be created in your home directory.
  - To see the files created, type the command this into the terminal: ls
  - To display the **port_forwarding_command** file content, type the command: cat port_forwarding_command
  - To display the content of **megatron_gpt_port** file, type the command: cat megatron_gpt_port



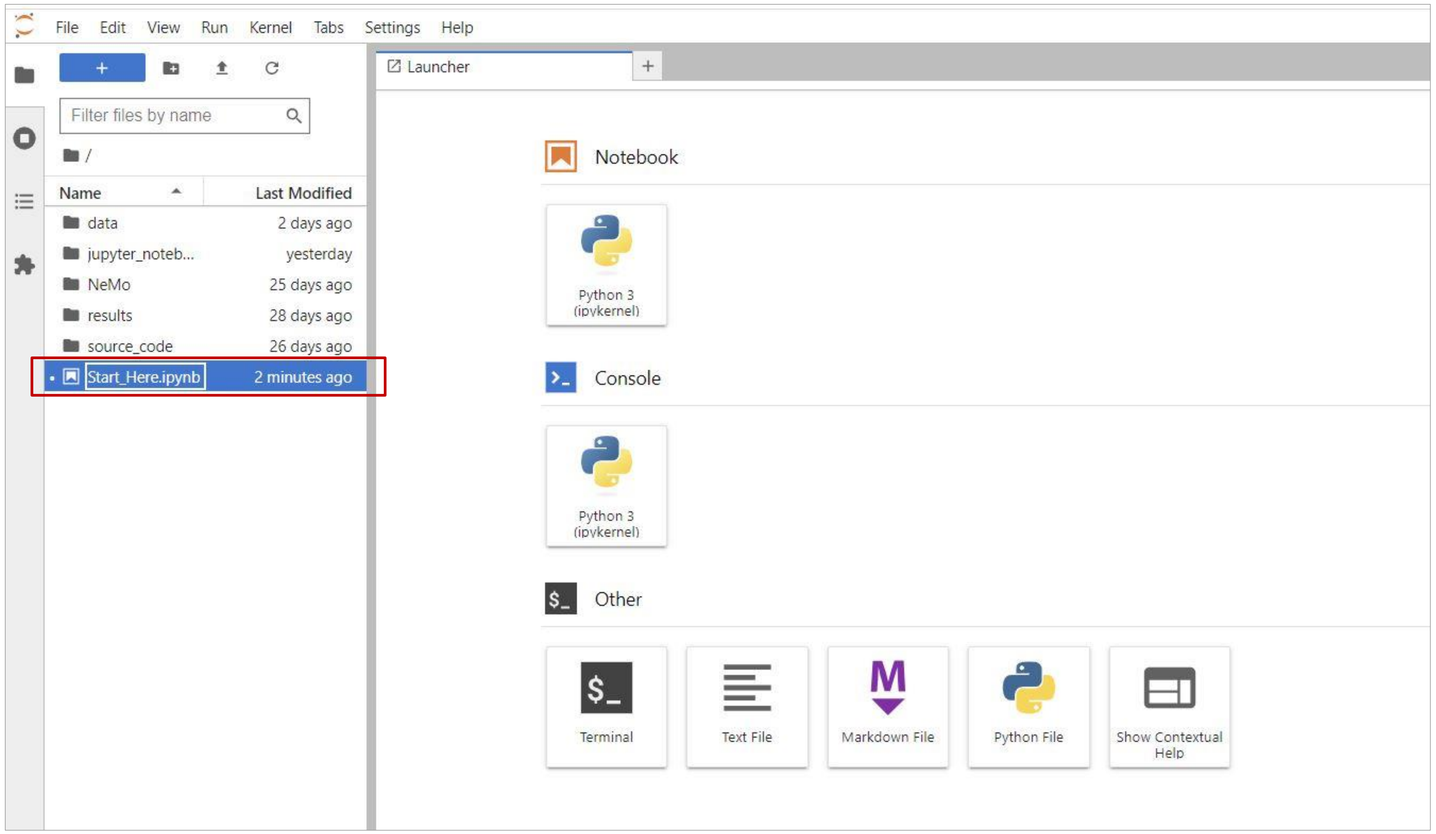
- Copy and paste the content of the port-forwarding file into your local machine or laptop terminal (e.g. CMD prompt, PowerShell, Terminal, etc.). Press on the Enter key.



NeMo Megatron Lab

Running the Lab

- Run the notebooks via the browser :
  - Open a new tab on your browser at http://localhost:8888
  - Click on the Start_Here.ipynb notebook



NeMo Megatron Lab

Start with Nemo Fundamentals

- Under the “NeMo Megatron,” Click on Nemo Fundamentals.

Table of Content

The following contents will be covered:

1. Data preprocessing
   - A. Overview of QA Dataset
     - a. Introduction to QA
     - b. Brief on QA Dataset
   - B. Common Preprocessing Techniques for Raw Text Data
   - C. QA Text Data preprocessing
     - a. SQuAD Dataset Structural Format
     - b. Text Data Source
     - c. Mannual QA Extraction
     - d. Automatic QA Generation with T5 model
   - D. Exercise
   - E. Summary
2. Deployment
   - A. RIVA Deployment
   - B. Challenge
3. NeMo Megatron
   - A. Nemo Fundamentals
   - B. Question Answering
   - C. Prompt Tuning/P-Tuning
   - D. Megatron-GPT 1.3B: Language Model Inferencing
   - E. Challenge

Troubleshooting

- Everyone can only run 1 job at a time.
- The jobs will run for 12 hours so you do not need to re-submit the sbatch script. All you need is the content of the "port_forwarding_command" to reconnect to the cluster.
- Each person will get access to 1 GPU only.
- The Hash key is valid for 24 hours.
- Do not use VPN or eduroam
- The access will be revoked by the end of the day.

If you get the following error in the final Step :

```
bind [127.0.0.1]:8888: Address already in use
channel_setup_fwd_listener_tcpip: cannot listen to port: 8888
Could not request local forwarding.
```

You will need to change the port 8888 in your ssh command to some other port, examples would be 8890, 8900 etc.

For example: `ssh -L localhost:8890:dgx05:9538 ssh.axisapps.io -l 7aab57129c95450cb80c4df5c13836bd`

Troubleshooting

- Useful commands:
  - Check the number of jobs submitted via squeue --me
  - Cancel the submitted job via scancel {job id}
  - Check the slurm output via cat slurm-{job id}.out
- To check if the new "`port_forwarding_command`" has been created, run `ls port_forwarding_command`
- Type the commands instead of copy-pasting to avoid errors.
- Please direct the questions to the "**cluster-support**" channel on the Slack.

Introduction to NeMo Framework

Language Model

An adult man walks into a bar

Task of LM is next word (token) prediction

An adult man walks into a ?

Language Model

$\underbrace{\text{An}}_{w_1}\ \underbrace{\text{adult}}_{w_2}\ \underbrace{\text{man}}_{w_3}\ \underbrace{\text{walks}}_{w_4}\ \underbrace{\text{into}}_{w_5}\ \underbrace{\text{a}}_{w_6}\ \underbrace{\overset{?}{____}}_{w_7}$

Estimate $P(w_7 | w_6)$

1) a library
2) a shop
3) a book
4) a door
5) a LLM
6) a beer
7) a knob
8) ....
9) ....
10) ....
1000) ....
100k) ....

Language Model

$$\underbrace{\text{An}}_{w_1}\ \underbrace{\text{adult}}_{w_2}\ \underbrace{\text{man}}_{w_3}\ \underbrace{\text{walks}}_{w_4}\ \underbrace{\text{into}}_{w_5}\ \underbrace{\text{a}}_{w_6}\ \underbrace{\overset{?}{____}}_{w_7}$$

Estimate $P(w_7 \mid w_5, w_6)$

1) into a library
2) into a shop
3) ~~into a book~~
4) into a door

5) ~~into a LLM~~
6) ....
7) ....
8) ....

9) ....
10) ....
1000) ....
100k) ....

Language Model

$$\underbrace{\text{An}}_{w_1}\ \underbrace{\text{adult}}_{w_2}\ \underbrace{\text{man}}_{w_3}\ \underbrace{\text{walks}}_{w_4}\ \underbrace{\text{into}}_{w_5}\ \underbrace{\text{a}}_{w_6}\ \underbrace{\overset{?}{\underline{\qquad}}}_{w_7}$$

Estimate $P(w_7 \mid w_2, w_3, w_4, w_5, w_6)$

1) ~~adult man walks into a library~~
2) adult man walks into a shop
3) adult man walks into a bar

Large Language Model

LLM is basically LM with billions parameters trained on internet-scale datasets

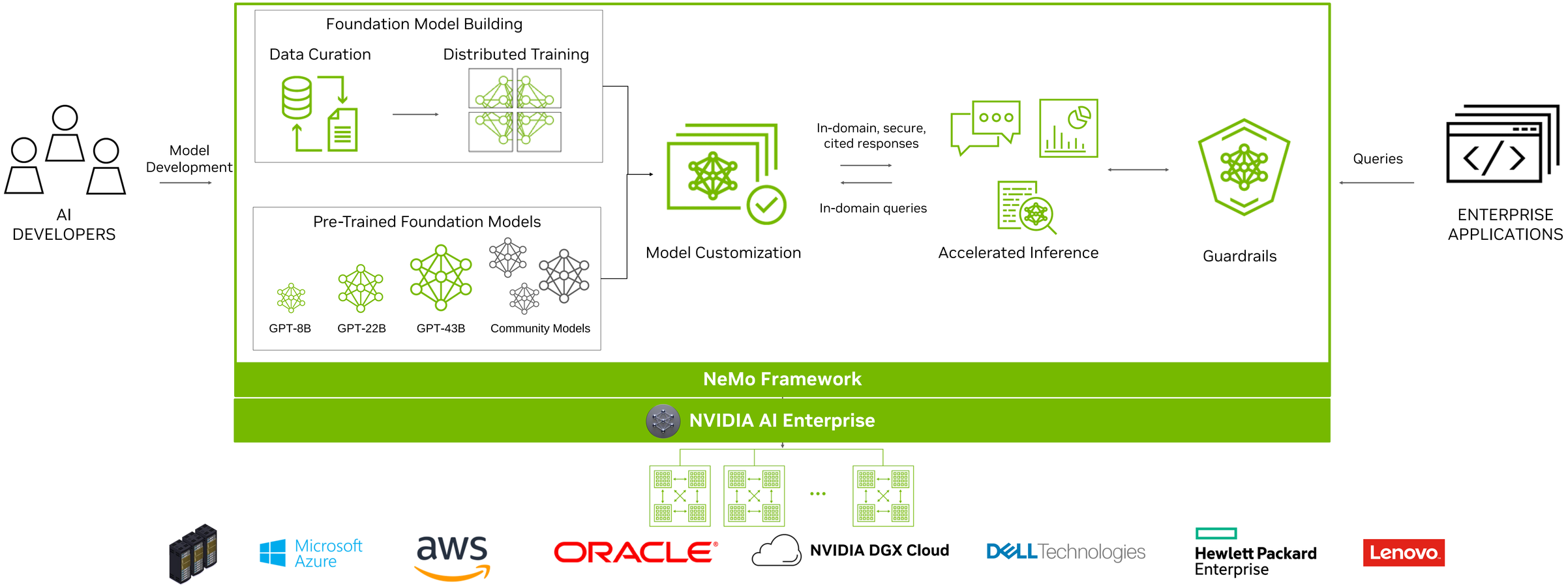
GPUs

Large-scale optimization

Science

NVIDIA NeMo for Custom LLMs

End-to-end, open and interoperable platform to build, customize and deploy generative AI models anywhere



Multi-modality support

Build language, image, generative AI models

Data Curation @ 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 Anywhere

Run optimized inference at-scale anywhere

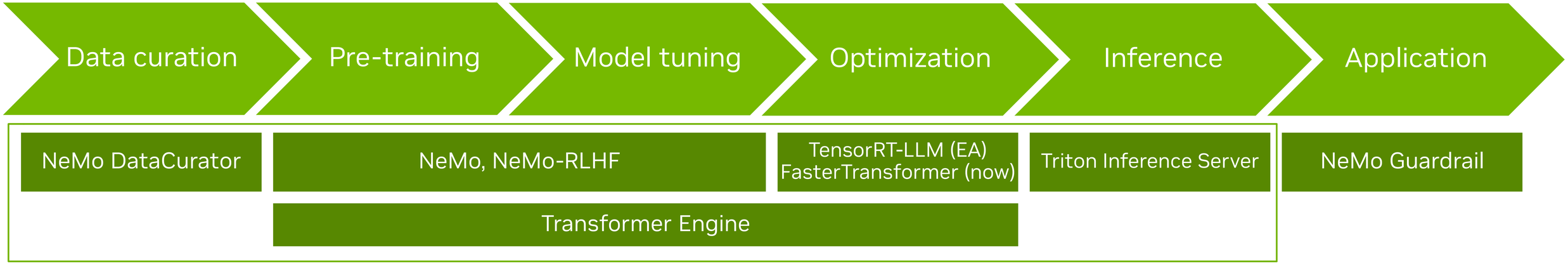
Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA experts by your side to keep projects on track

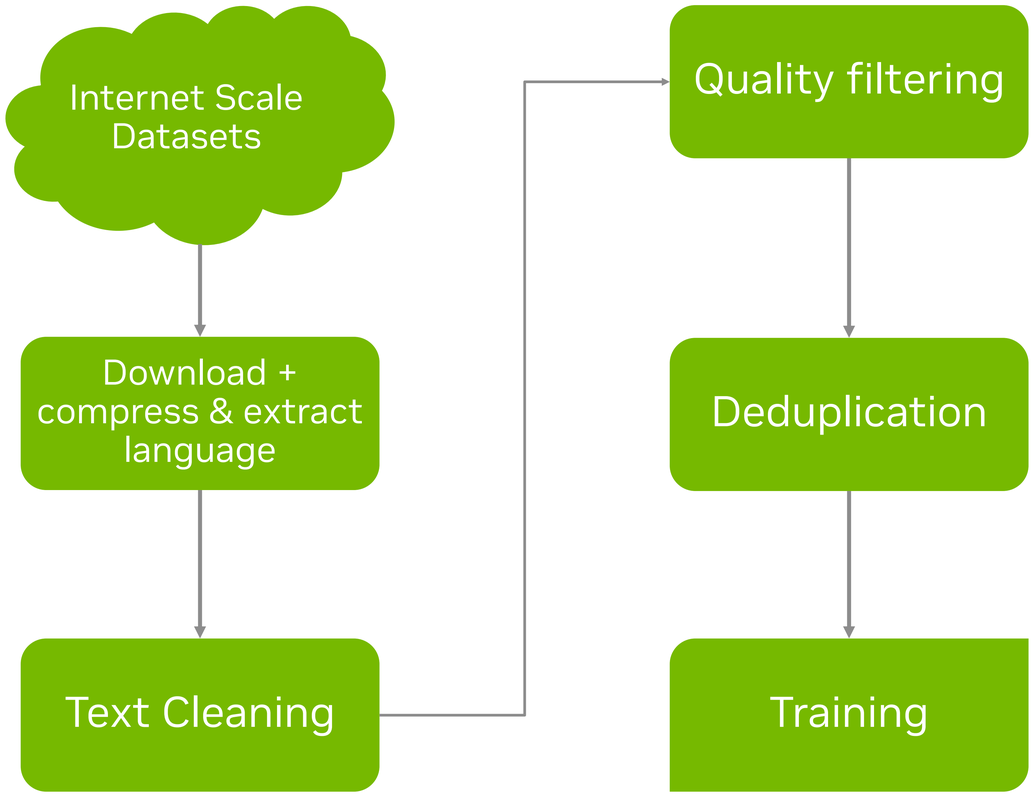
Build Large Language Model Pipeline



NeMo Framework

NeMo Data Curator

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



Text Cleaning

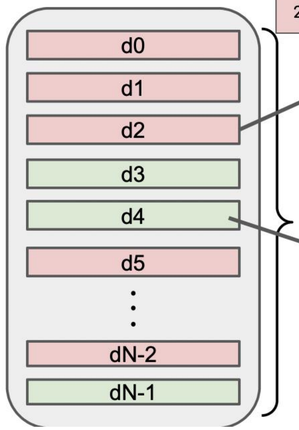
"The Mona Lisa doesnÃƒ ¢Ã¢â€š ¬Ã¢â€ž ¢t have eyebrows."
"âœ" No problems"

↓ Text reformatting

"The Mona Lisa doesn't have eyebrows."
"✓ No problems"

Quality Filtering

Cleaned documents in dataset




```
BEGIN:VTIMEZONE\nTZID:America/Chicago\nX-LIC-LOCATION:America/Chicago\nBEGIN:STANDAR
D\nDTSTART:20181104T020000\nRRULE:FREQ=YEARLY;BYDAY=1SU;BYMONTH=11\nTZNAME:CST\nTZOF
FSETFROM:-0500\nTZOFFSETTO:-0600\nEND:STANDARD\nBEGIN:DAYLIGHT\nDTSTART:20190310T020
000\nRRULE:FREQ=YEARLY;BYDAY=2SU;BYMONTH=3\nTZNAME:CDT\nTZOFFSETFROM:-0600\nTZOFFSET
TO:-0500\nEND:DAYLIGHT\nEND:VTIMEZONE\nBEGIN:VEVENT\nCREATED;TZID=America/Chicago:20
201123T111003\nDESCRIPTION:Renewable energy: the future is now! \\n
```

Both high and low quality documents

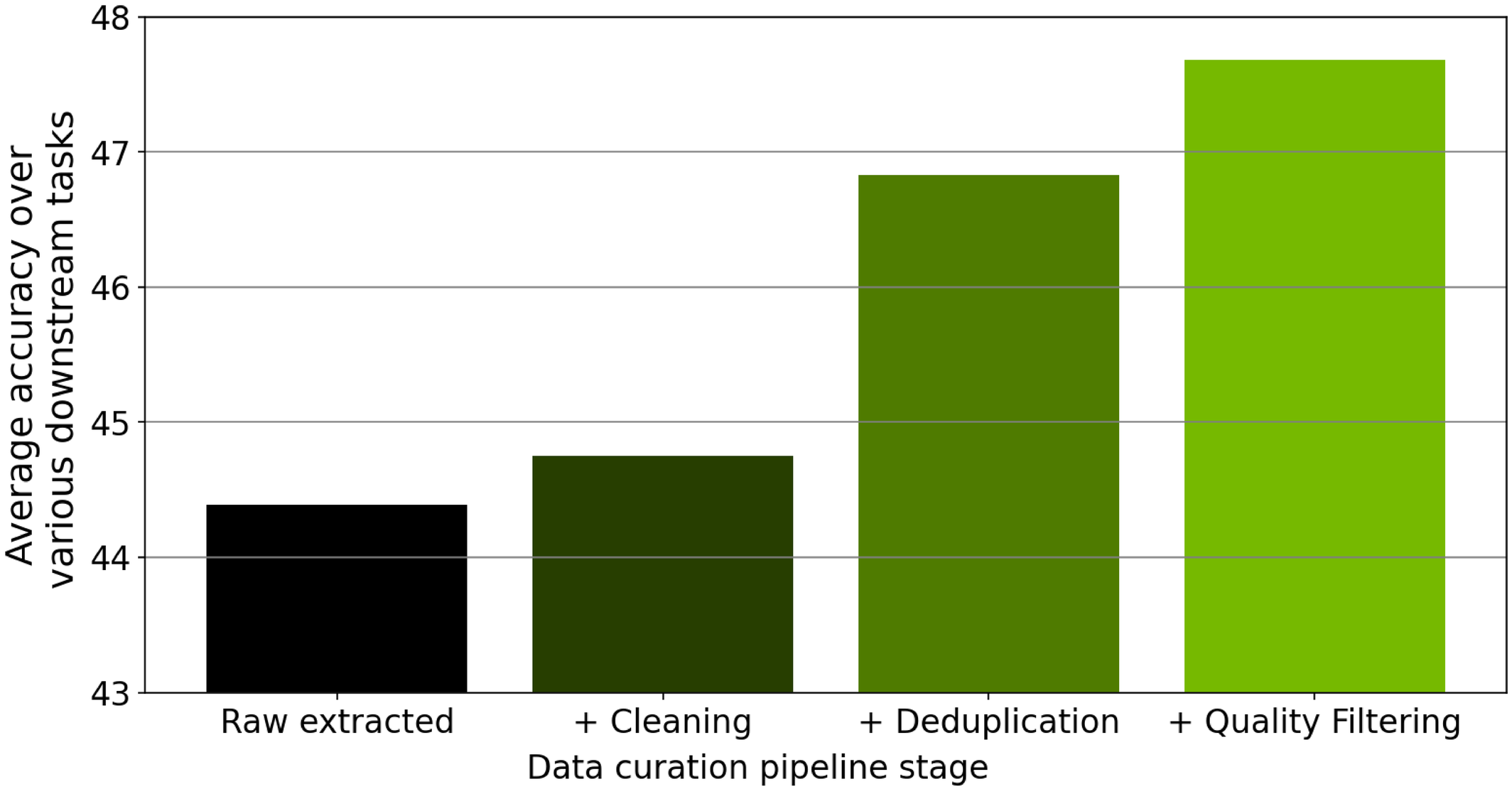
How to filter out the low quality?

```
Philippine police are ready to resume President Rodrigo Duterte's war on
the drugs trade which had returned to the streets, the police chief said
on Monday (27/02), a month after Duterte halted police operations,
labeling the force 'rotten to the core.' (Reuters Photo/Janis Alano)
Regional Tensions, Paris Attacks Hang Over APEC Summit BY :ROSEMARIE
FRANCISCO & MANUEL MOGATO NOVEMBER 16, 2015 Manila. Regional tensions
over the South China Sea and security concerns after the attack by
Islamist militants on Paris could eclipse efforts by Pacific-rim leaders
this week to boost trade and growth across a region of around three
billion people.
```

Data Curation & Preprocessing

Enabling Large-Scale High-Quality Datasets for LLMs

Performance after Data Curation



Data Blending

Learnings from Bloomberg GPT

Dataset	Docs 1e4	C/D	Chars 1e8	C/T	Toks 1e8	T%
FINPILE	*175,886*	*1,017*	*17,883*	*4.92*	*3,635*	*51.27%*
Web	158,250	933	14,768	4.96	2,978	42.01%
News	10,040	1,665	1,672	4.44	376	5.31%
Filings	3,335	2,340	780	5.39	145	2.04%
Press	1,265	3,443	435	5.06	86	1.21%
Bloomberg	2,996	758	227	4.60	49	0.70%
PUBLIC	*50,744*	*3,314*	*16,818*	*4.87*	*3,454*	*48.73%*

Table 1: Breakdown of the full training set used to train BLOOMBERGGPT. The statistics provided are the average number of characters per document ("C/D"), the average number of characters per token ("C/T"), and the percentage of the overall tokens ("T%"). Units for each column are denoted in the header.

Even a small amount of proprietary data does a lot!

	BLOOMBERGGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$
ConvFinQA	**43.41**	30.06	27.88	36.31
FiQA SA	**75.07**	50.59	51.60	53.12
FPB	**51.07**	44.64	48.67	50.25
Headline	**82.20**	73.22	79.41	76.51
NER	60.82	**60.98**	57.49	55.56
All Tasks *(avg)*	**62.51**	51.90	53.01	54.35
All Tasks *(WR)*	**0.93**	0.27	0.33	0.47

Table 8: Results on financial domain tasks.

Shape	
Number of Layers	70
Number of Heads	40
Vocabulary Size	131,072
Hidden Dimension	7,680
Total Parameters	50.6B
Hyperparameters	
Max Learning Rate	6e-5
Final Learning Rate	6e-6
Learning Rate schedule	cosine decay
Gradient Clipping	0.3
Training	
Tokens	569B
Hardware	64 × 8 A100 40GB
Throughput	32.5 sec/step
avg. TFLOPs	102
total FLOPS	2.36e23

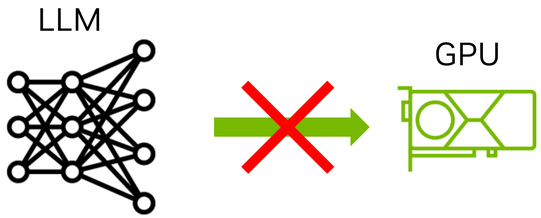
Table 4: A summary of the hyper-parameters and their values for BLOOMBERGGPT.

2303.17564.pdf (arxiv.org)

Challenges to Train Large Language Model

Memory cost of GPT3 175B

- **Challenges**
  - Large Language Model can not fit in **single GPU** or even **single GPU Server**.



- **GPT-3 175B** Memory costs
  - Model parameters: **350 GB** = 175 B * 2 byte
  - Gradients: **350 GB** = 175 B * 2 byte
  - Optimizer state: **2100 GB** = 175 B * 12 byte
  - Total: **2800 GB ++**

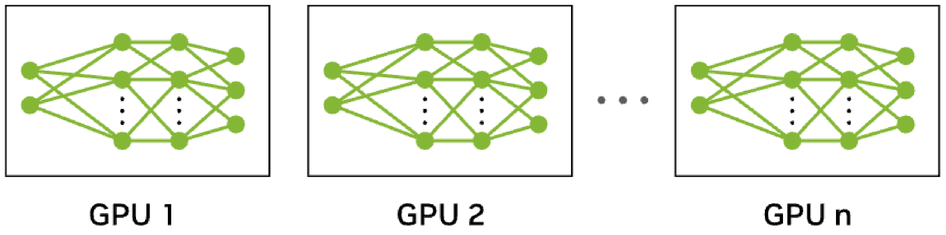
H100 80GB

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

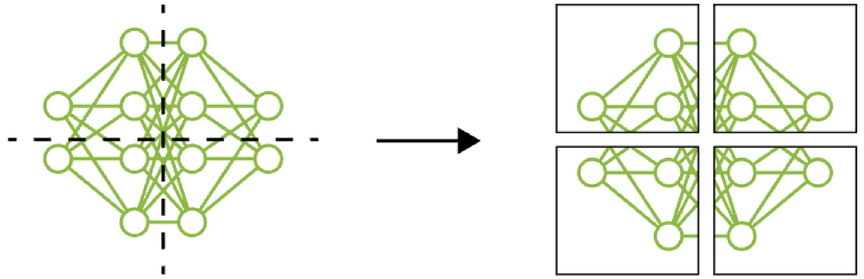
3D Parallelism technique

Save GPU memory cost

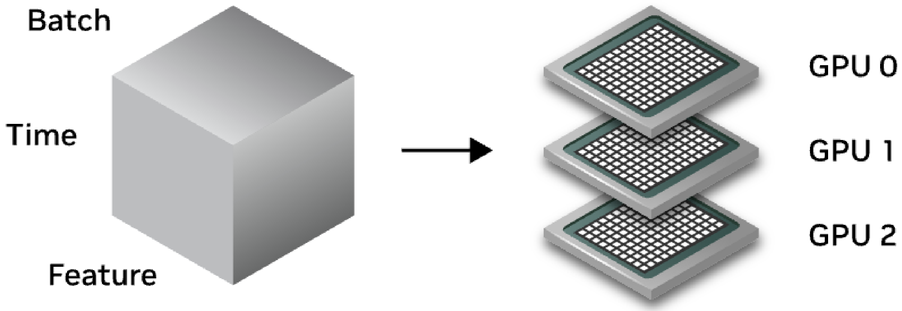
Data Parallelism



Tensor and Pipeline Parallelism

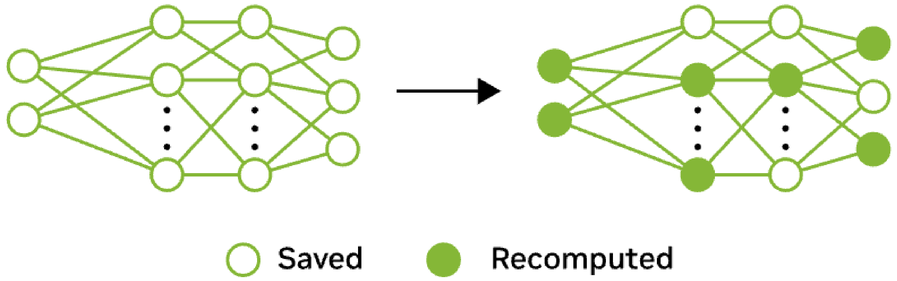


Sequence Parallelism



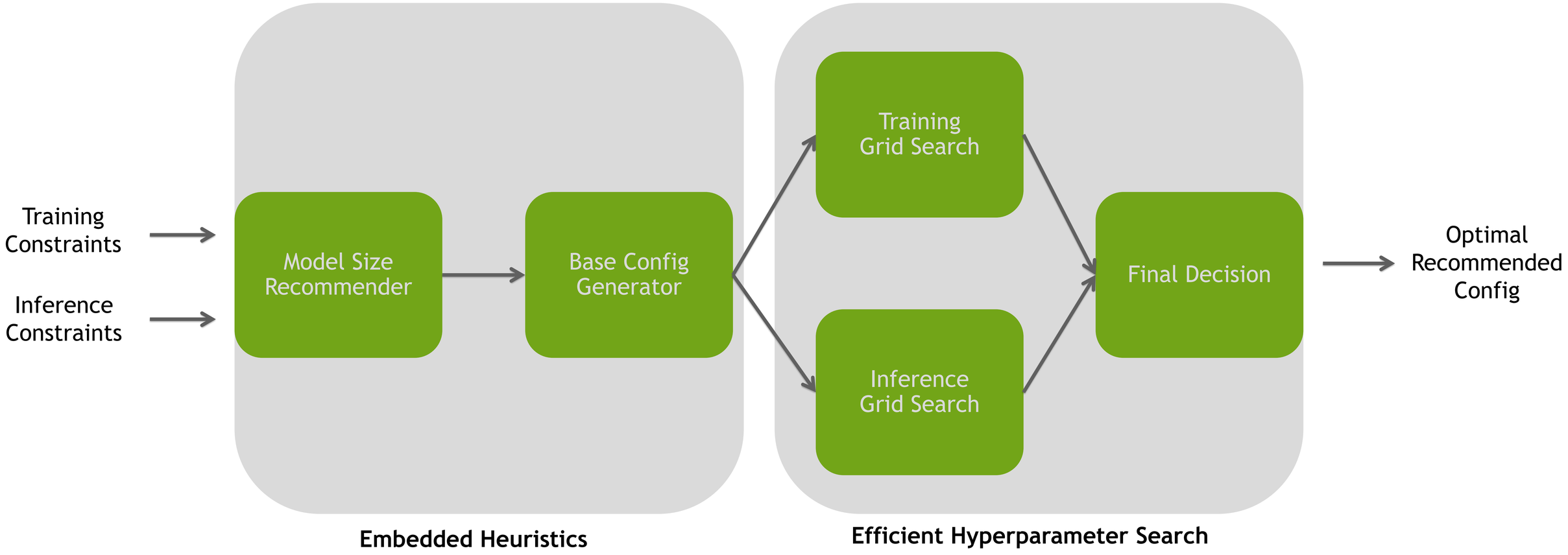
Selective Activation Recomputation

Traditionally, LLMs (>175B) every activation is recomputed



AUTO-CONFIGURATOR TOOL

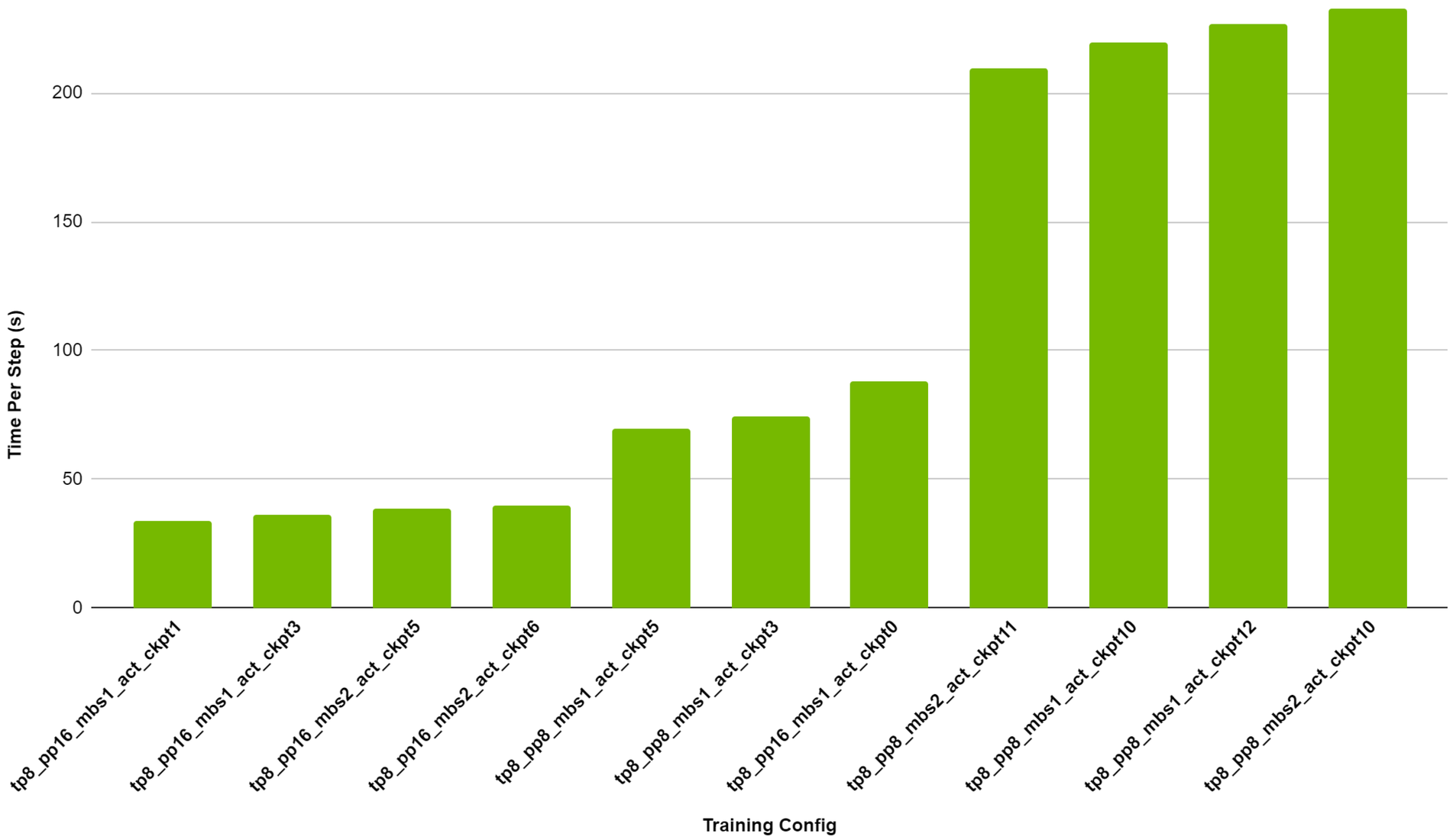
Automatically search and optimize model configurations on any given compute or time constraints



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)

175B GPT-3 MODEL: 6.85X TRAINING SPEEDUP



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

As simple as using recipes, by default a GPT3 5B model

NeMo-Megatron-Launcher / launcher_scripts / conf / config.yaml

```yaml
defaults:
  - _self_
  - cluster: bcm  # Leave it as bcm even if using bcp. It will be ignored for bcp.
  - data_preparation: gpt3/download_gpt3_pile
  - training: gpt3/5b
  - conversion: gpt3/convert_gpt3
  - fine_tuning: null
  - prompt_learning: null
  - adapter_learning: null
  - ia3_learning: null
  - evaluation: gpt3/evaluate_all
  - export: gpt3/export_gpt3
  - override hydra/job_logging: stdout

hydra:
  run:
    dir: .
  output_subdir: null

debug: False

stages:
  - data_preparation
  - training
  - conversion
  - evaluation
  - export
```

- Download the docker container

> *docker pull nvcr.io/ea-bignlp/bignlp-training:22.08.01-py3*

- Run the container, eg

> *docker run --gpus all -it --ipc=host --ulimit memlock=-1 *
> *--ulimit stack=67108864 *
> *nvcr.io/ea-bignlp/nemofw-training:23.04.1-py3 /bin/bash*

- In the training container

> *cd /opt/NeMo-Megatron-Launcher/launcher_scripts#*

> *python3 main.py*

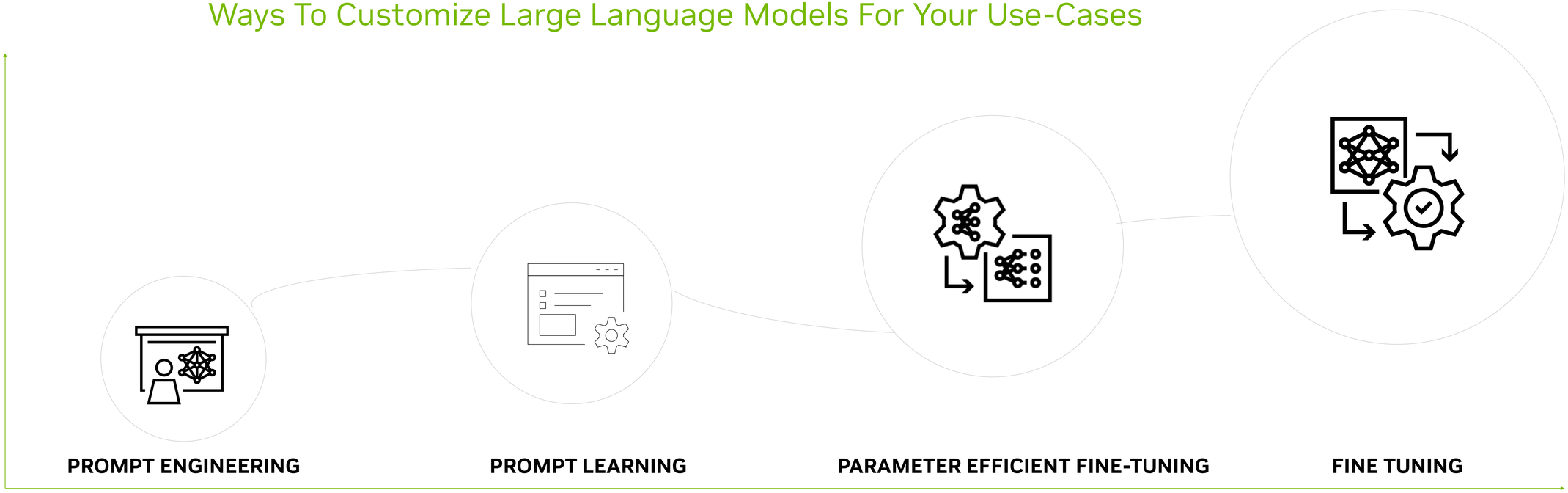
NVIDIA/NeMo-Megatron-Launcher: NeMo Megatron launcher and tools (github.com)

NeMo-Megatron-Launcher/launcher_scripts/conf/config.yaml at master · NVIDIA/NeMo-Megatron-Launcher (github.com)

Suite of Model Customization Tools in NeMo

Ways To Customize Large Language Models For Your Use-Cases

Data, compute & investment

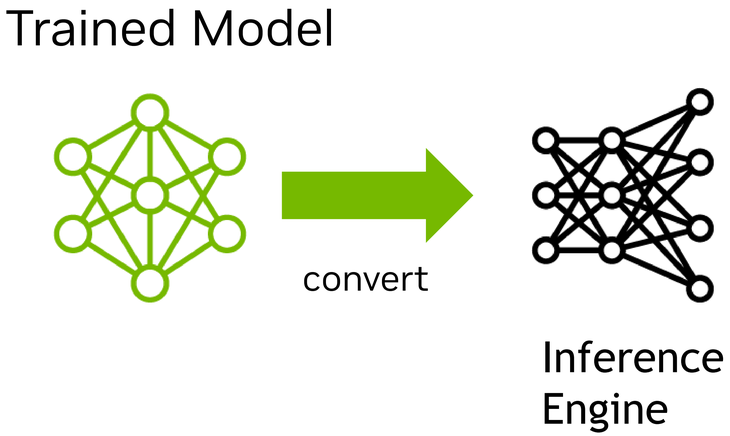


Accuracy for specific use-cases

	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
HOW	Prompt templates	Tune companion model	Add ”custom” layers to LLM	Tune LLM model weights

FasterTransformer – TRT LLM

Serve and Accelerate Large Transformer Models

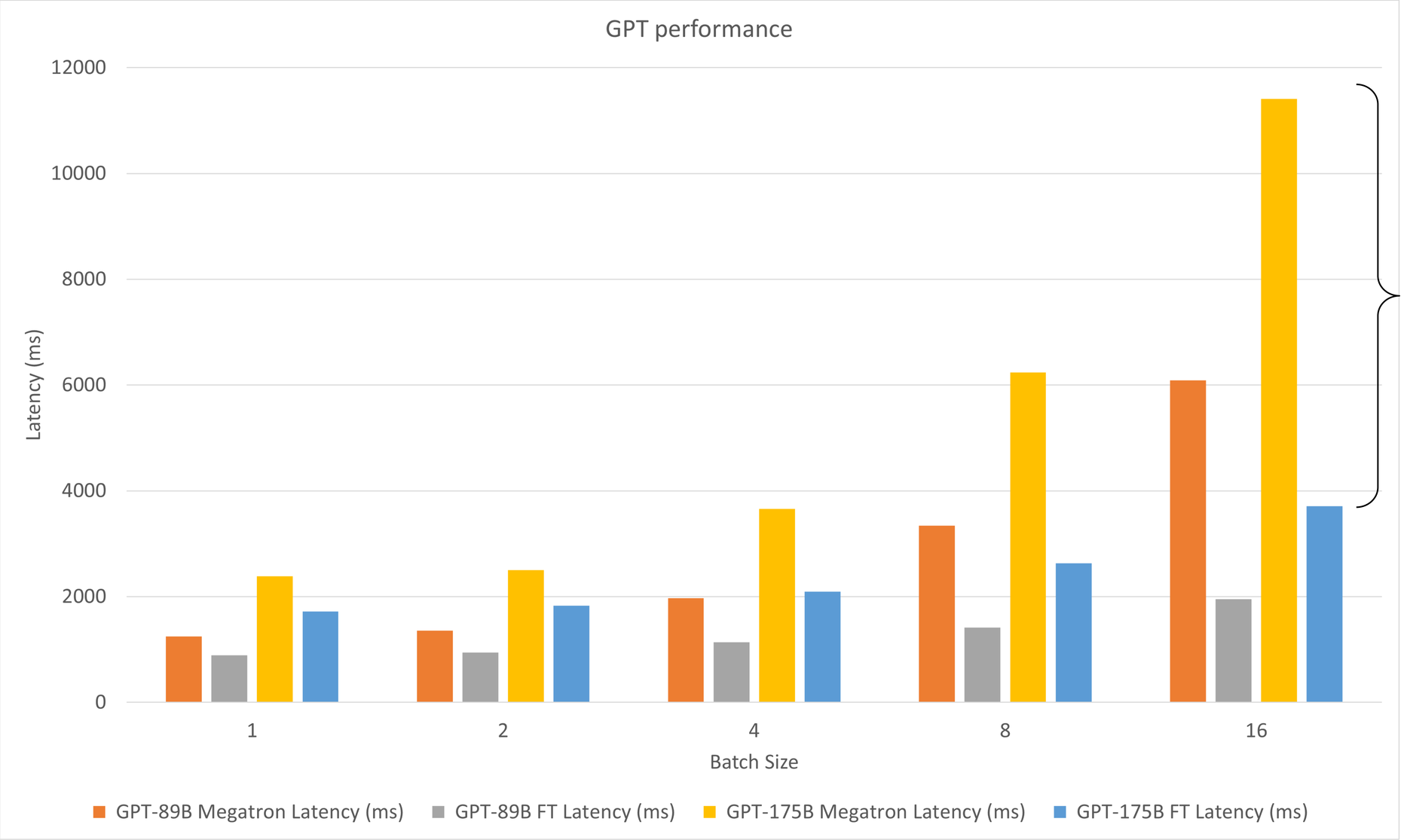


	Optimized kernels to accelerate performance for Transformers
	Tensor/Pipeline parallelism for multi-GPU and multi-node inference
	Support various models, like GPT-3, GPT-J, GPT-NeoX, T5, Meta OPT, Google UL2, BLOOM, BERT

FasterTransformer -> TensorRT LLM (EA)

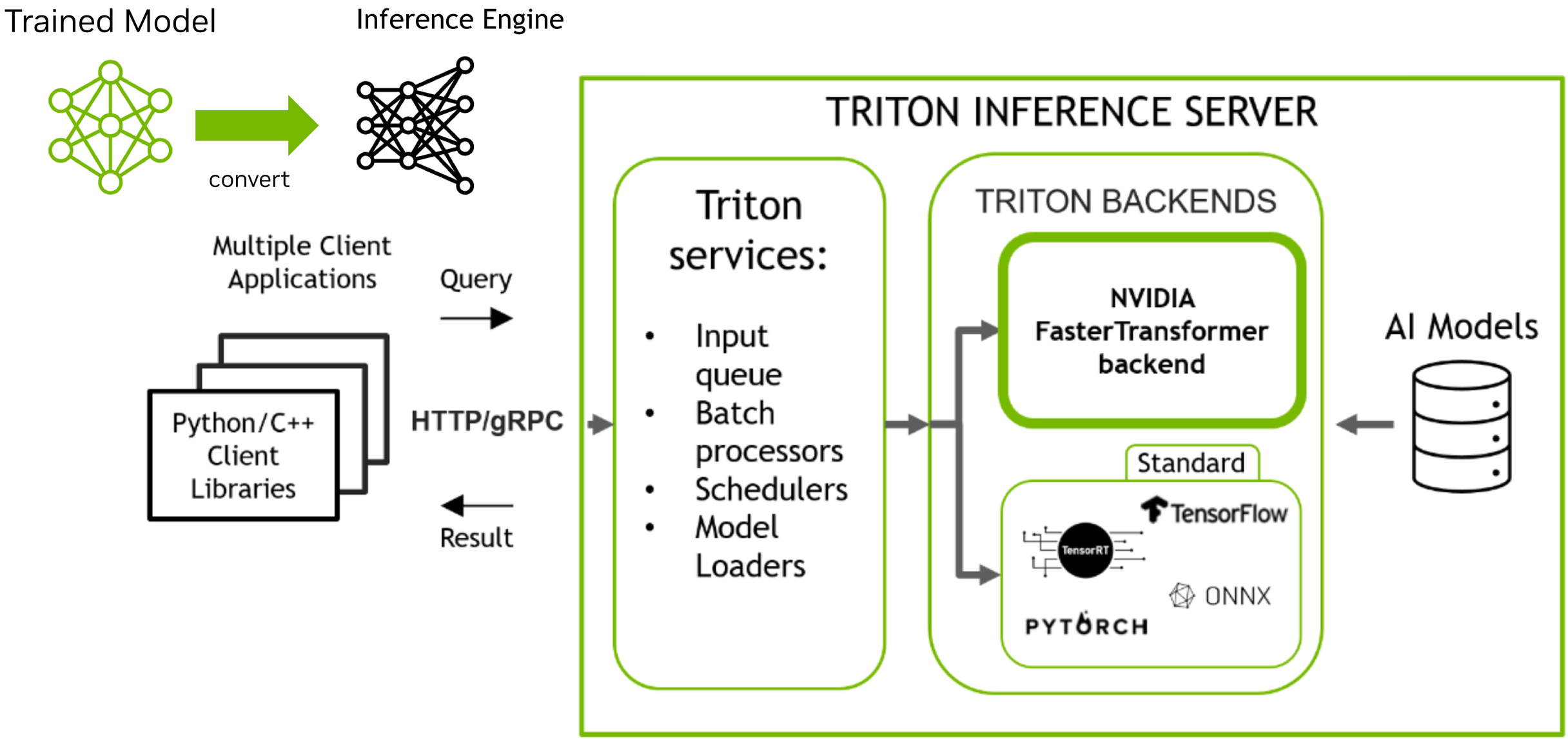
FasterTransformer – TensorRT LLM

Serve and Accelerate Large Transformer Models



NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

Triton Inference Server



NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

Inference

As simple as one command line – Based on Triton Inference Server

```
docker run --rm \
    --name triton-inference-server \
    -d \
    --gpus all \
    -p 8000-8002:8000-8002 \
    -v /path/to/checkpoints/output:/model_repository \
    nvcr.io/ea-bignlp/bignlp-inference:22.08-py3 \
    bash -c 'export CUDA_VISIBLE_DEVICES=0 && \
    tritonserver --model-repository /model_repository'
```

```
$ docker logs triton-inference-server
I0928 14:29:34.011299 1 server.cc:629]
+-----------+---------+--------+
| Model     | Version | Status |
+-----------+---------+--------+
| gpt3_1.3b | 1       | READY  |
+-----------+---------+--------+
```

Inference

Using Python Triton Client

```
$ python3 infer.py "1 2 3 4 5 6"
```

This sends the prompt “1 2 3 4 5 6” to the local inference server and should output the following to complete the sequence up to the default response token limit of 30:

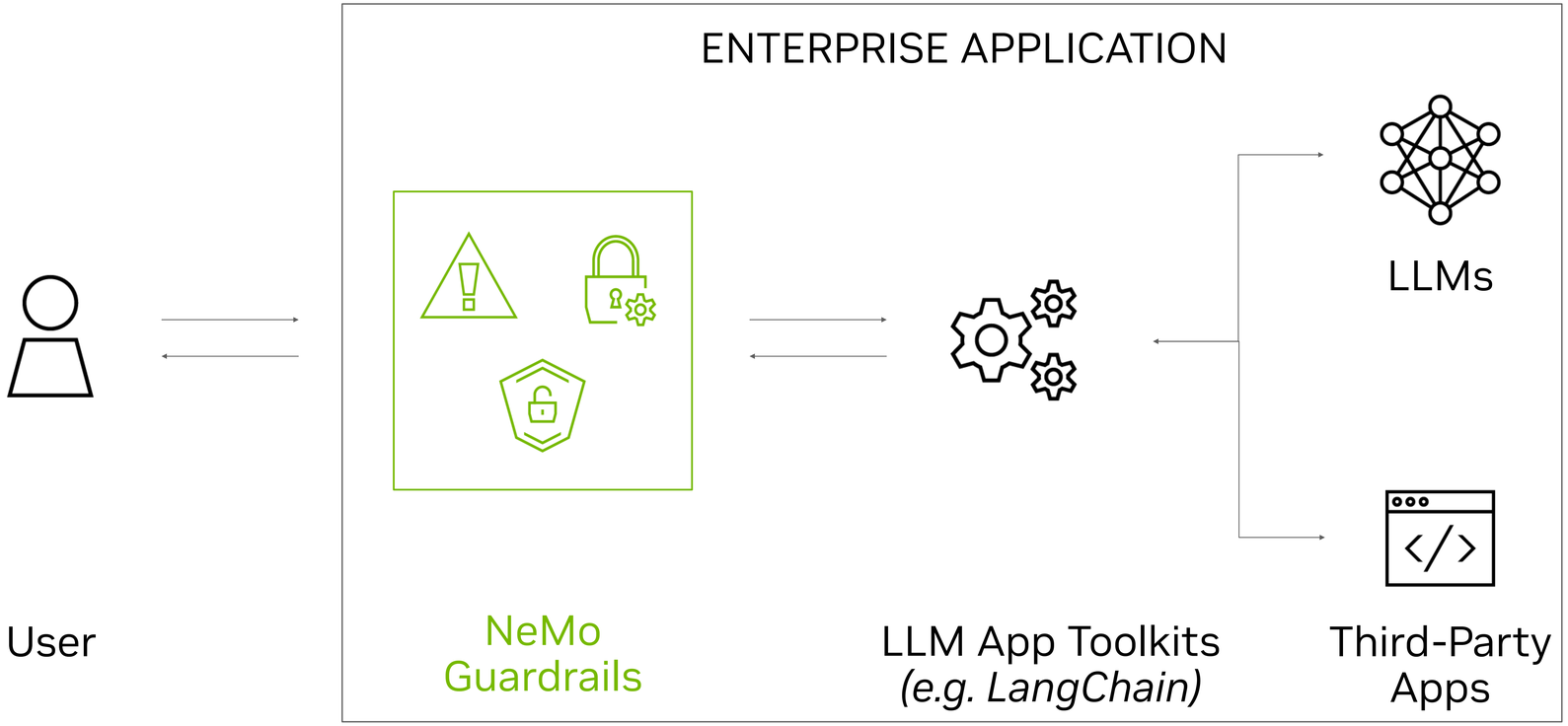
```
“1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36"
```

Deploying a 1.3B GPT-3 Model with NVIDIA NeMo Framework | NVIDIA Technical Blog

```python
from argparse import ArgumentParser
import numpy as np
import tritonclient.http as httpclient
from tritonclient.utils import np_to_triton_dtype
from transformers import GPT2Tokenizer

def fill_input(name, data):
    infer_input = httpclient.InferInput(name, data.shape, np_to_triton_dtype(data.dtype))
    infer_input.set_data_from_numpy(data)
    return infer_input

def build_request(query, host, output):
    with httpclient.InferenceServerClient(host) as client:
        request_data = []
        request = np.array([query]).astype(np.uint32)
        request_len = np.array([[len(query)]]).astype(np.uint32)
        request_output_len = np.array([[output]]).astype(np.uint32)
        top_k = np.array([[1]]).astype(np.uint32)
        top_p = np.array([[0.0]]).astype(np.float32)
        temperature = np.array([[1.0]]).astype(np.float32)

        request_data.append(fill_input('input_ids', request))
        request_data.append(fill_input('input_lengths', request_len))
        request_data.append(fill_input('request_output_len', request_output_len))
        request_data.append(fill_input('runtime_top_k', top_k))
        request_data.append(fill_input('runtime_top_p', top_p))
        request_data.append(fill_input('temperature', temperature))
        result = client.infer('gpt3_1.3b', request_data)
        output = result.as_numpy('output_ids').squeeze()
        return output

def main():
    parser = ArgumentParser('Simple Triton Inference Requestor')
    parser.add_argument('query', type=str, help='Enter a text query to send to '
                        'the Triton Inference Server in quotes.')
    parser.add_argument('--output-length', type=int, help='Specify the desired '
                        'length for output.', default=30)
    parser.add_argument('--server', type=str, help='Specify the host:port that '
                        'Triton is listening on. Defaults to localhost:8000',
                        default='localhost:8000')
    args = parser.parse_args()

    tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
    query = tokenizer(args.query).input_ids
    request = build_request(query, args.server, args.output_length)
    print(tokenizer.decode(request))

if __name__ == '__main__':
    main()
```

NeMo Guardrails

Add Boundaries Ensure Chatbots Operate According to Use Cases



Question: ask capabilities

"What can you do?"
"tell me what you can do"

Answer:

"I am an AI assistant which helps answer questions based on a given knowledge base."

Question: ask off topic

"Who is the president?"
"Can you recommend the best stocks to buy?"
"Can you tell me a joke?"

Answer:

"I cannot comment on anything which is not relevant to the job report"

TOPICAL

Focus interactions within a specific domain

SAFETY

Prevent hallucinations, toxic or misinformative content

SECURITY

Prevent executing malicious calls and handing power to a 3rd party app

Enterprise Use-Cases Require Domain Specific Knowledge

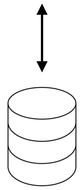
Encode and embed your AI with your enterprise's real-time information to provide the latest responses

Foundation/Custom Model Response

"What was the pressure of the tank at 11 p.m. last night?" → → *"I was trained 2 months ago and do not have the current data "*

Inform Model Response

"What was the pressure of the tank at 11 p.m. last night?" → → *"The pressure at 11 p.m. last night was 345 psi"*



Dataset

70%

Of enterprise data is untapped

Unlock many new opportunities for greater intelligence

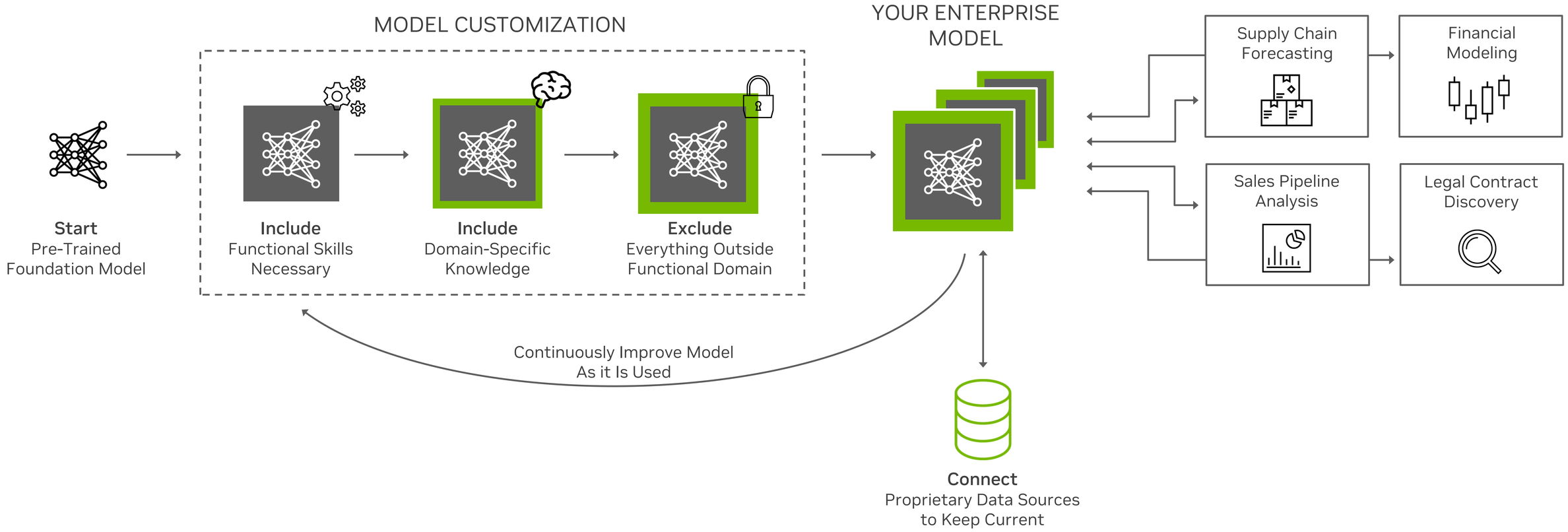
Less Frequent Re-Training

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

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



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

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
- Efficient At-Scale Training and Deployment of Large Language Models – GTC Session
- Hyperparameter Tool GTC Session

Building Generative AI Foundation Models

Efficiently and quickly training models using NeMo

Requirements for Building Foundation Models

Mountains of Training Data

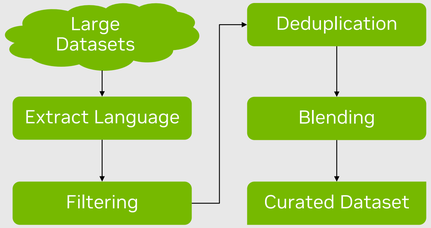
Large-scale compute infrastructure for training & inferencing, costing $10 M+

Deep technical expertise

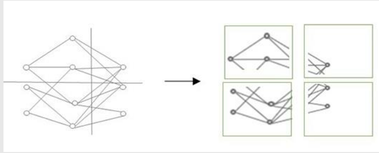
Complex algorithms to build on large-scale infrastructure

Accelerated Training With NeMo

Data Curation Tool

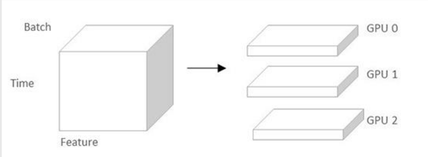


Accelerated Training



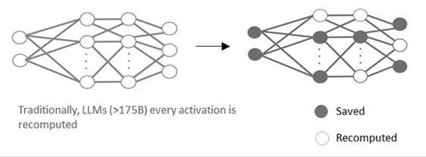
Tensor & Pipeline Parallelism

Accelerated Training



Sequence Parallelism

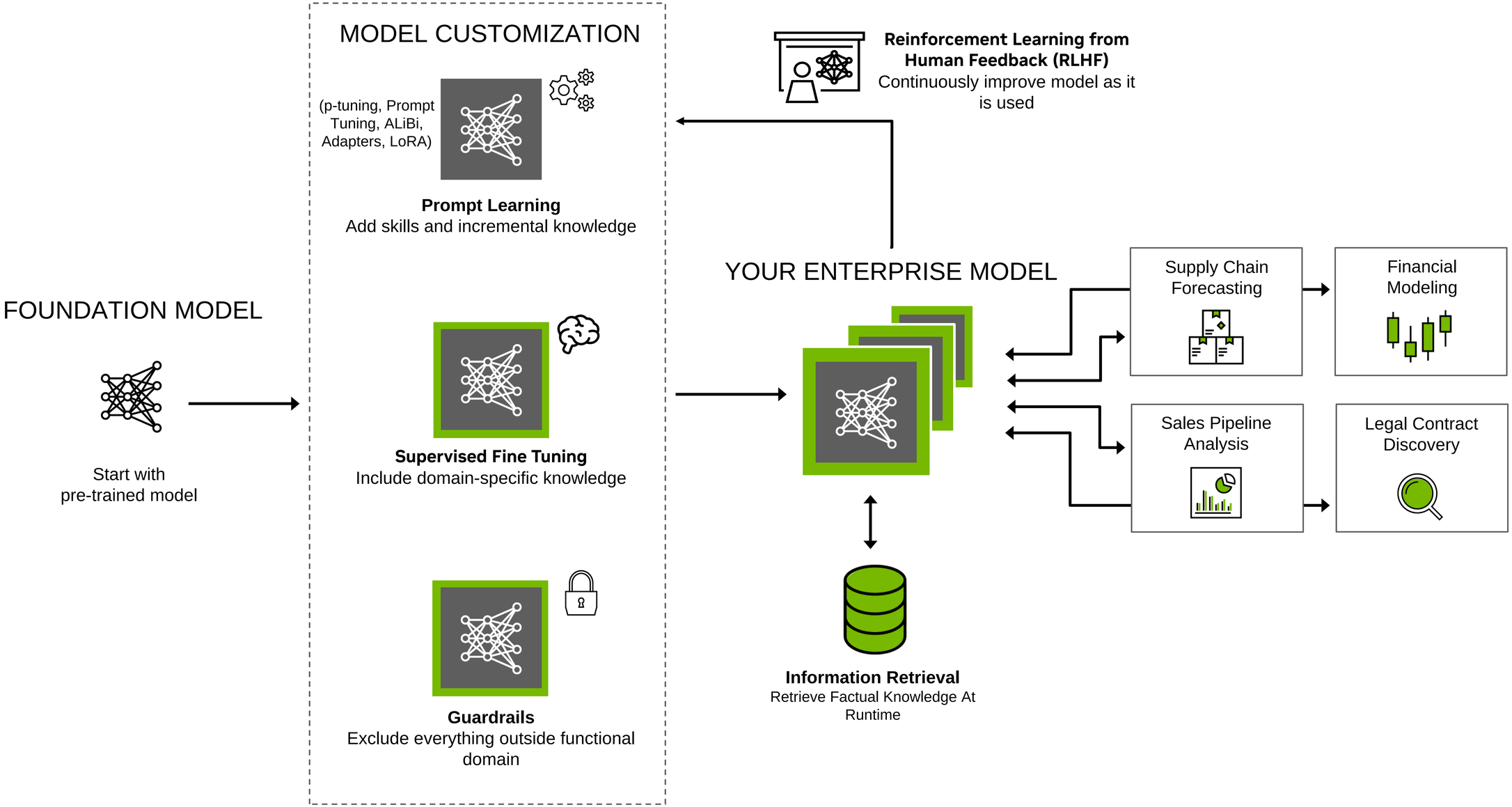
Accelerated Training



Selective Activation Recomputation

Model Customization for Enterprise Ready LLMs

Customization techniques to overcome the challenges of using foundation models

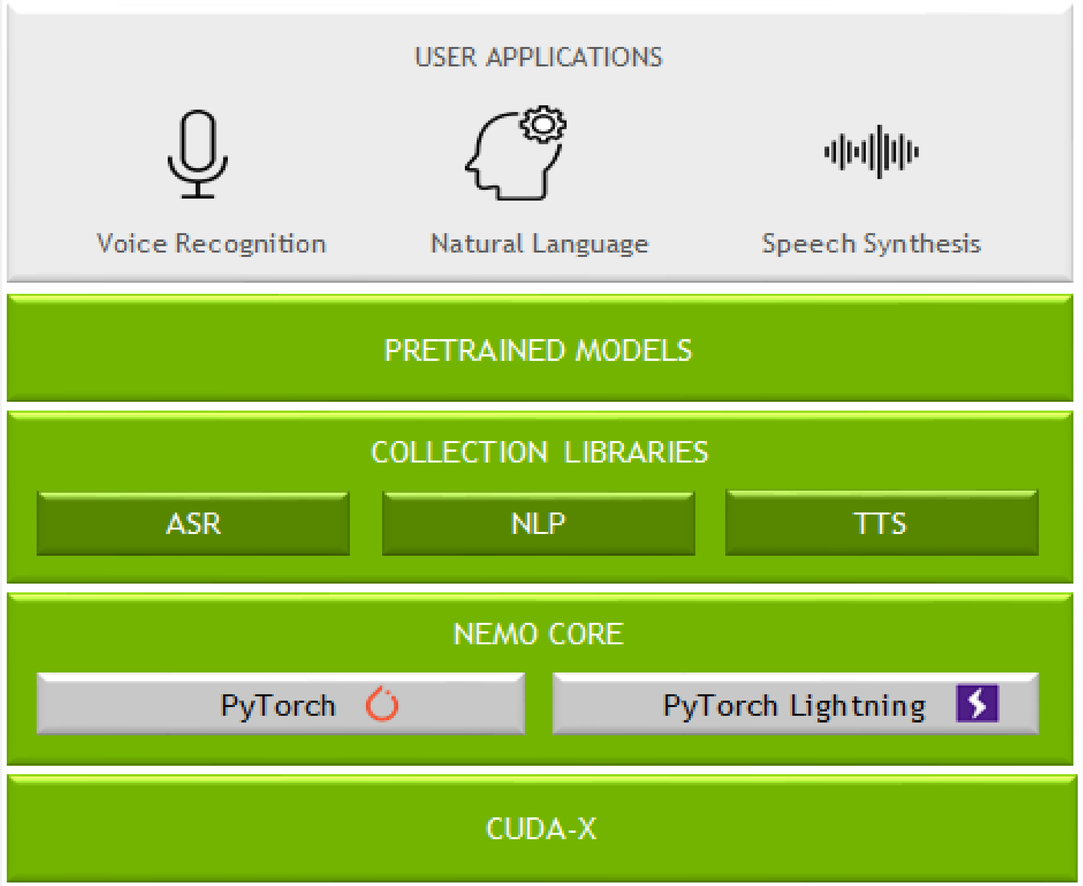


Fundamentals of NeMo

NVIDIA NEMO

Toolkit for Building SOTA Conversational Models

- DL-Based Speech & Language Understanding Models
- Include Semantic Checking for Correct-by-Construction Architectures
- Open-Sourced
- Integrated with PyTorch & PyTorch Lightning
- Easy-to-Use APIs
- Optimized Training Performance
- 100+ Pre-Trained GPU-Optimized Checkpoints
- Scale to 1000s of NVIDIA GPUs



https://ngc.nvidia.com/catalog/containers/nvidia:nemo

https://github.com/NVIDIA/NeMo

Using NeMo

Only a few simple steps

- Download NeMo from NGC and run the container
- docker run --runtime=nvidia -it
  --rm -v --shm-size=16g
  -p 8888:8888 -p 6006:6006
  --ulimit memlock=-1
  --ulimit stack=67108864
  nvcr.io/nvidia/nemo:22.08

- Start Python 3
- Import NeMo
- Import a pre-trained model
  - (4 GiB of VRAM in this case)
- And use it ☺

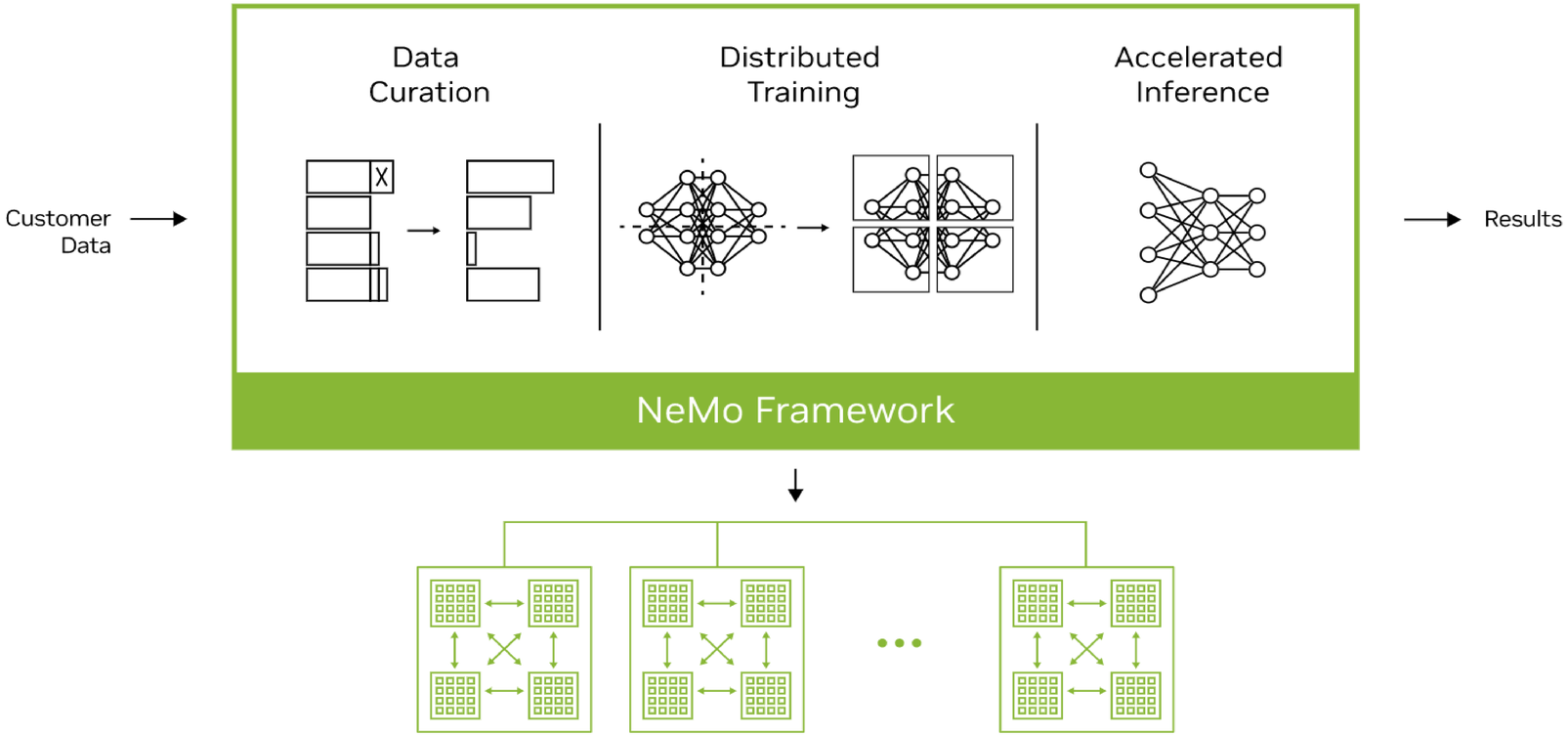
```python
from nemo.collections.nlp.models import MTEncDecModel

# To get the list of pre-trained models
MTEncDecModel.list_available_models()

# Download and load the a pre-trained to
# translate from English to Spanish
model = MTEncDecModel.from_pretrained(
                        "nmt_en_fr_transformer24x6")

# Translate a sentence or list of sentences
translations = model.translate(
       ["Hello! How are you today?"],
       source_lang="en", target_lang="fr")
```

```
>>> translations
["Bonjour ! Comment allez-vous aujourd'hui ?"]
```

NeMo Framework Fundamentals

Introduction

- NVIDIA NeMo™ framework is an end-to-end, cloud-native enterprise framework to build, customize, and deploy generative AI models with billions of parameters.



Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Framework Fundamentals

Models for different modalities

- The framework comes with extendable collections of pre-built modules and ready-to-use models for automatic speech recognition (ASR), natural language processing (NLP) and text synthesis (TTS).
- Built for speed, NeMo can utilize NVIDIA's Tensor Cores and scale out training to multiple GPUs and multiple nodes.

Language

- BERT
- GPT-3
- T5
- T5-MoE
- Inform

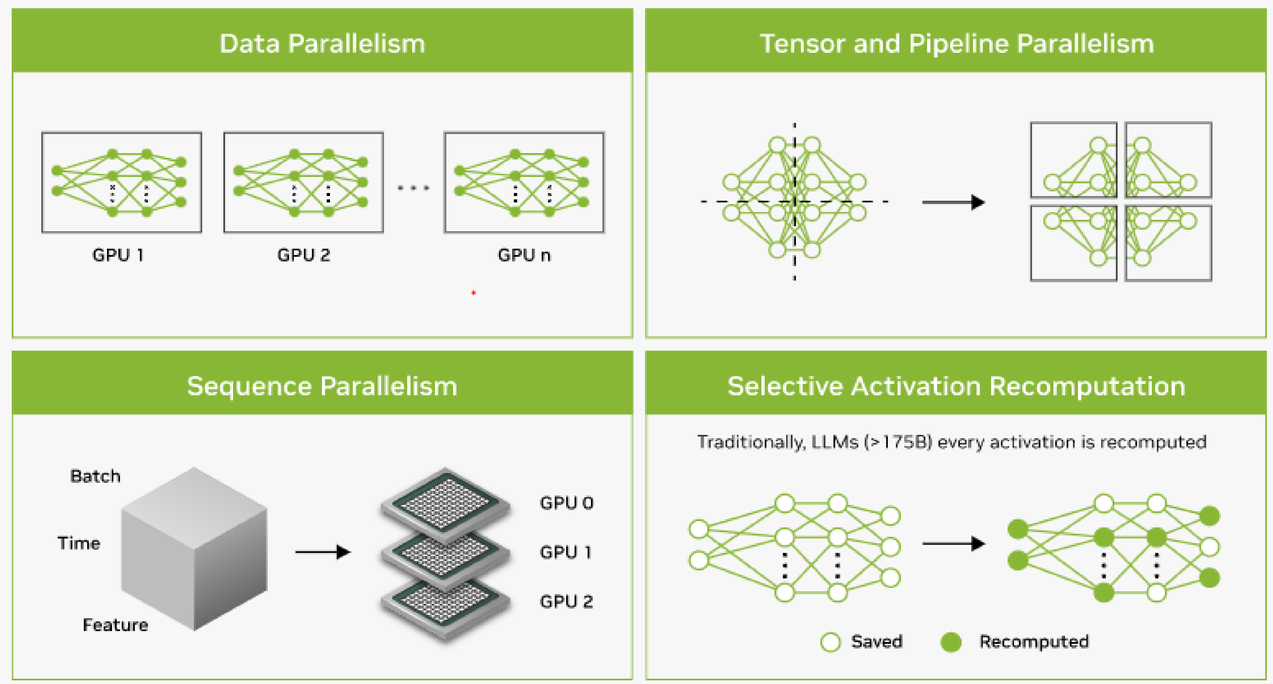
Multimodal

- Stable Diffusion v1.5
- VisionTransformers (ViT)
- CLIP
- Instruct-Pix2Pix
- Imagen

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Parallelism



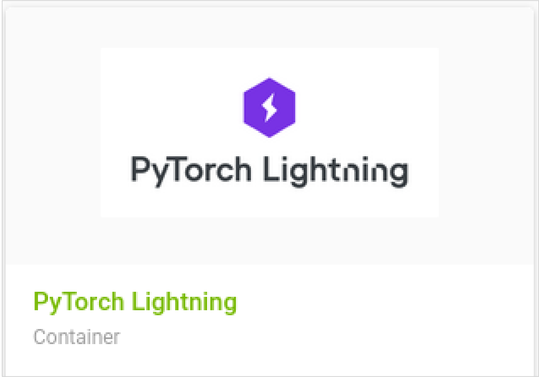
- The NeMo Megatron delivers high levels of training efficiency, making training of large-scale foundation models possible, using 3D parallelism techniques
- NeMo Megatron supports 4 types of parallelisms:
  - Distributed Data parallelism
  - Tensor Parallelism
  - Pipeline Parallelism
  - Sequence Parallelism

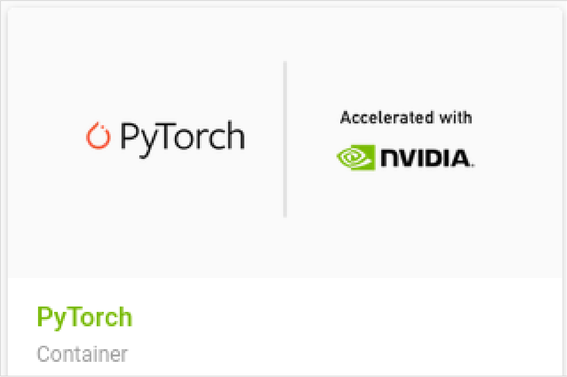
Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Framework Fundamentals

NeMo Models

- NeMo models leverage PyTorch Lightning Module and are compatible with the entire PyTorch ecosystem.





Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Framework Fundamentals

NeMo Models

- A "Model" is the neural network(s) as well as all the infrastructure supporting those network(s), wrapped into a singular, cohesive unit.
- As such, all NeMo models are constructed to contain the following:
  - Neural Network architecture
  - Dataset + Data Loaders
  - Preprocessing + Postprocessing
  - Optimizer + Schedulers
  - Any other supporting infrastructure

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Framework Fundamentals

NeMo Models in Collections

```
import nemo.collections.asr as nemo_asr
import nemo.collections.nlp as nemo_nlp
import nemo.collections.tts as nemo_tts
```

```
asr_models = [model for model in dir(nemo_asr.models) if model.endswith("Model")]
```

```
nlp_models = [model for model in dir(nemo_nlp.models) if model.endswith("Model")]
```

```
tts_models = [model for model in dir(nemo_tts.models) if model.endswith("Model")]
```

```
citrinet = nemo_asr.models.EncDecCTCModelBPE.from_pretrained('stt_en_citrinet_512')
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Model Configuration using OmegaConf

- OmegaConf is an excellent library that is used throughout NeMo to enable yaml configuration management more easily.
- All NeMo models come packaged with their model configuration inside the cfg attribute
- The cfg is an essential tool to modify the behavior of the Model after it has been constructed.

```
from omegaconf import OmegaConf
import copy
cfg = copy.deepcopy(citrinet.cfg)
print(OmegaConf.to_yaml(cfg))
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Model Config Content

```
train_ds:
  manifest_filepath: null
  sample_rate: 16000
  batch_size: 32
  trim_silence: true
  max_duration: 16.7
  shuffle: true
  is_tarred: false
  tarred_audio_filepaths: null
validation_ds:
  manifest_filepath: null
  sample_rate: 16000
  batch_size: 32
  shuffle: false
test_ds:
  manifest_filepath:
  - /home/smajumdar/PycharmProjects/nemo-eval/nemo_eval/librispeech/manifests/de
  sample_rate: 16000
  batch_size: 32
  shuffle: false
  num_workers: 12
  pin_memory: true
model_defaults:
  repeat: 5
  dropout: 0.0
  separable: true
  se: true
  se_context_size: -1
tokenizer:
  dir: /home/smajumdar/PycharmProjects/nemo-eval/nemo_beta_eval/asrset/manifests/
unigram_v1024/
  type: bpe
preprocessor:
  _target_: nemo.collections.asr.modules.AudioToMelSpectrogramPreprocessor
  sample_rate: 16000
  normalize: per_feature
  window_size: 0.025
  window_stride: 0.01
  window: hann
  features: 80
  n_fft: 512
  frame_splicing: 1
  dither: 1.0e-05
  pad_to: 16
  stft_conv: false
spec_augment:
  _target_: nemo.collections.asr.modules.SpectrogramAugmentation
  freq_masks: 2
  time_masks: 10
  freq_width: 27
```

```
encoder:
  _target_: nemo.collections.asr.modules.ConvASREncoder
  feat_in: 80
  activation: relu
  conv_mask: true
  jasper:
  - filters: 512
    repeat: 1
    kernel:
    - 5
    stride:
    - 1
    dilation:
    - 1
```

```
decoder:
  _target_: nemo.collections.asr.modules.ConvASRDecoder
  feat_in: 640
  num_classes: 1024
  vocabulary:
  - <unk>
  - s
  - _the
  - t
  - _a
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Model Config Content


```
  optim:
    name: novograd
    lr: 0.05
    betas:
    - 0.8
    - 0.25
    weight_decay: 0.001
    sched:
      name: CosineAnnealing
      warmup_steps: 1000
      warmup_ratio: null
      min_lr: 1.0e-09
      last_epoch: -1
  target: nemo.collections.asr.models.ctc_bpe_models.EncDecCTCModelBPE
  nemo_version: 1.17.0
  decoding:
    strategy: greedy
    preserve_alignments: null
    compute_timestamps: null
    word_seperator: ' '
    ctc_timestamp_type: all
    batch_dim_index: 0
    greedy:
      preserve_alignments: false
      compute_timestamps: false
      preserve_frame_confidence: false
      confidence_method_cfg: null
    beam:
      beam_size: 4
      search_type: default
      preserve_alignments: false
      compute_timestamps: false
      return_best_hypothesis: true
      beam_alpha: 1.0
      beam_beta: 0.0
      kenlm_path: null
      flashlight_cfg:
        lexicon_path: null
        beam_size_token: 16
        beam_threshold: 20.0
        unk_weight: -.inf
        sil_weight: 0.0
        unit_lm: false
      pyctcdecode_cfg:
        beam_prune_logp: -10.0
        token_min_logp: -5.0
        prune_history: false
        hotwords: null
        hotword_weight: 10.0
    confidence_cfg:
```


Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Components of the Model

- NeMo has special utilities for a few components. They are:
  - setup_training_data
  - setup_validation_data and setup_multi_validation_data
  - setup_test_data and setup_multi_test_data
  - setup_optimization

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Optimizer & Scheduler Config

- **Optim structure**

```yaml
optim:
    name: novograd
    lr: 0.01

    # optimizer arguments
    betas: [0.8, 0.25]
    weight_decay: 0.001

    # scheduler setup
    sched:
      name: CosineAnnealing

      # Optional arguments
      max_steps: -1 # computed at runtime or explicitly set here
      monitor: val_loss
      reduce_on_plateau: false

      # scheduler config override
      warmup_steps: 1000
      warmup_ratio: null
      min_lr: 1e-9
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

Saving and restoring models

save_to

```
citrinet.save_to('citrinet_512.nemo')
```

restore_from

```
temp_cn = nemo_asr.models.EncDecCTCModelBPE.restore_from('citrinet_512.nemo')
```

load_from_checkpoint

```
citrinet = nemo_asr.models.EncDecCTCModelBPE.load_from_checkpoint(<path to checkpoint>)
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

NeMo with Hydra

- Hydra is used throughout NeMo as a way to enable rapid prototyping using predefined config files.
- While all NeMo models come with at least 1 default config file, one might want to switch configs without changing code.
- This is easily achieved by the following commands:

```
--config-path:
--config-name:
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Models

NeMo with Hydra: Overriding config from the command line

- Hydra allows users to provide command line overrides to any part of the config.
- There are three cases to consider:
  - Override existing value in config
  - Add new value in config
  - Remove old value in config

Overriding existing values in config: change the optimizer from novograd to adam

```
$ python <script>.py \
  --config-path="dir to config" \
  --config-name="name of config" \
  model.optim.name="adam" \
  model.optim.betas=[0.9,0.999]
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Examples

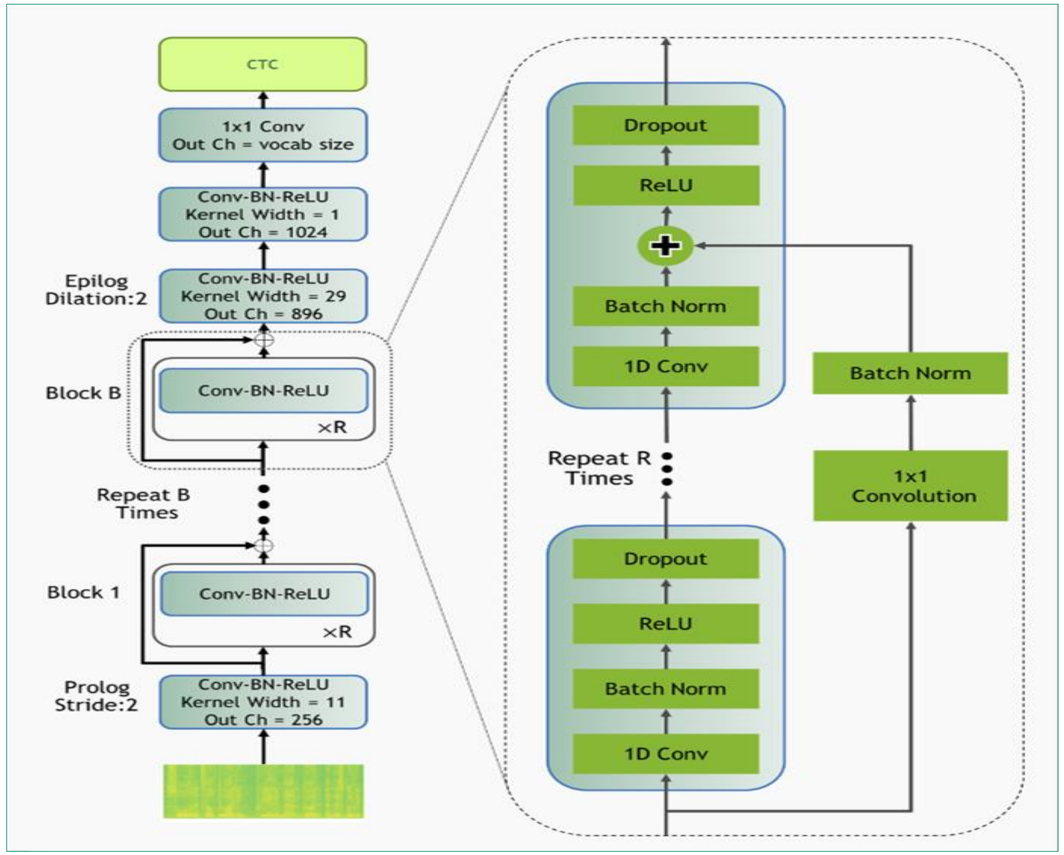
ASR Models

- NeMo supports multiple Speech Recognition models such as:
  - Jasper
  - QuartzNet
  - Citrinet
  - ContextNet
  - Conformer-CTC
  - Conformer-Transducer
  - Fast-Conformer
  - Cache-aware Streaming Conformer
  - LSTM-Transducer
  - LSTM-CTC
  - Squeezeformer-CTC
  - Hybrid-Transducer-CTC. All of these can be trained on various datasets

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

ASR Models

Jasper ("Just Another Speech Recognizer")



Jasper Architecture

- The Jasper family of models are denoted as **Jasper_[BxR]**
- Jasper models can be instantiated using the **EncDecCTCModel** class.

Source: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/asr/models.html

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Examples

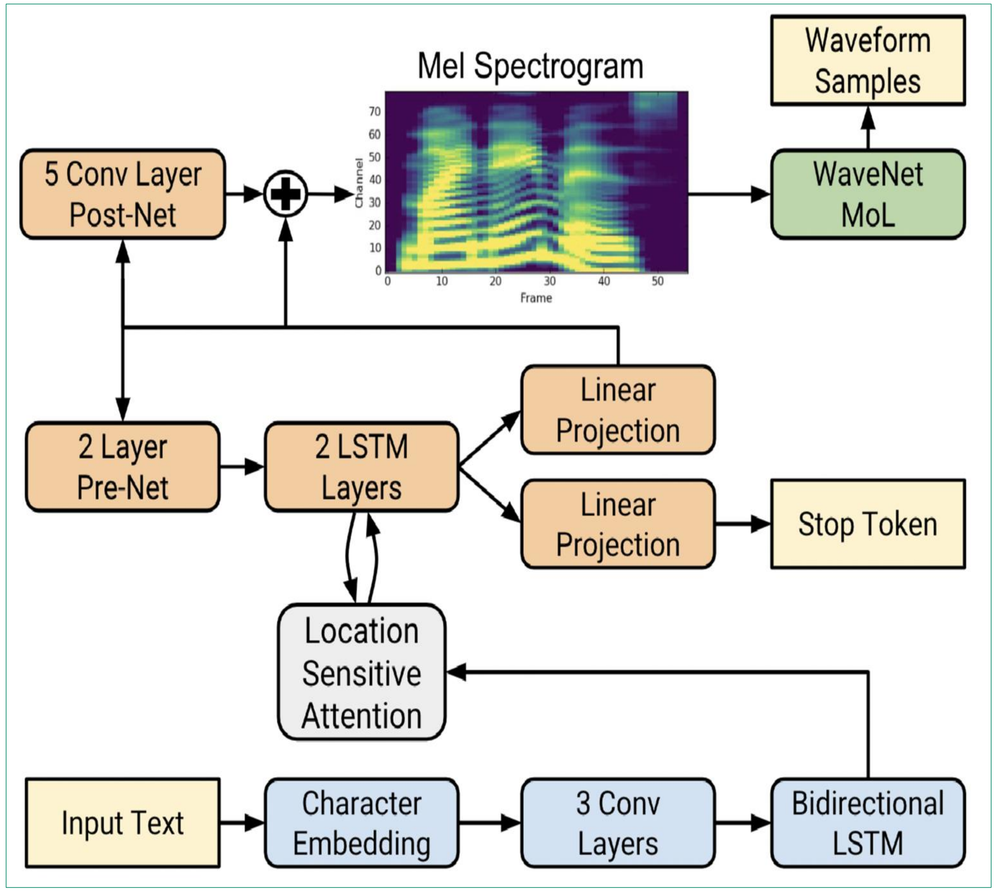
TTS Models

- NeMo supports Text To Speech (TTS, aka Speech Synthesis) via a two-step inference procedure.
- First, a model is used to generate a **mel spectrogram** from the text. Second, a model is used to generate audio from a mel spectrogram.
- Supported Models:
  - Mel Spectrogram Generators:
    - Tacotron2
    - FastPitch
    - Talknet
    - And more…
  - Audio Generators (Vocoders):
    - WaveGlow
    - HiFiGAN
    - UnivNet
    - And more…

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

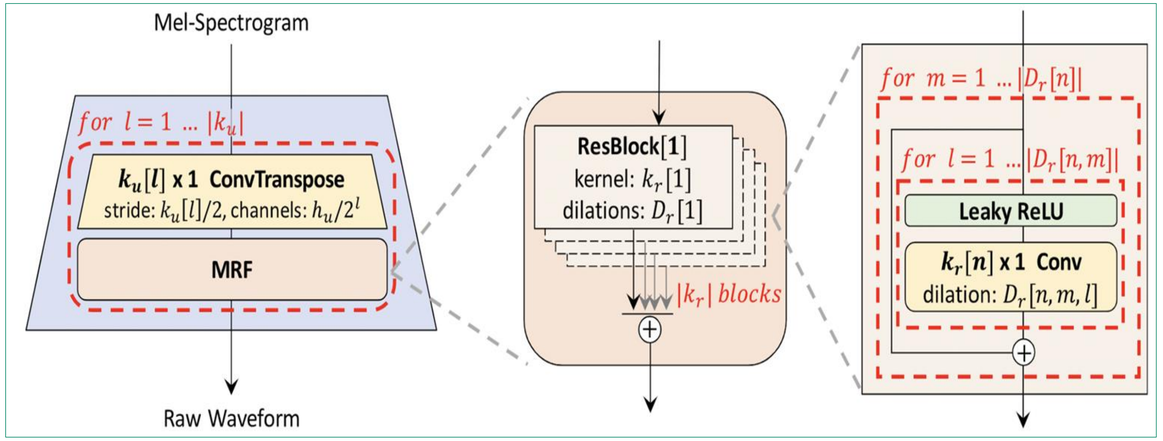
TTS Models

Tacotron 2 & HiFi-GAN Architecture



Tacotron 2

Source: https://en.wikipedia.org/wiki/GPT-2



HiFi-GAN

Source:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Examples

NLP

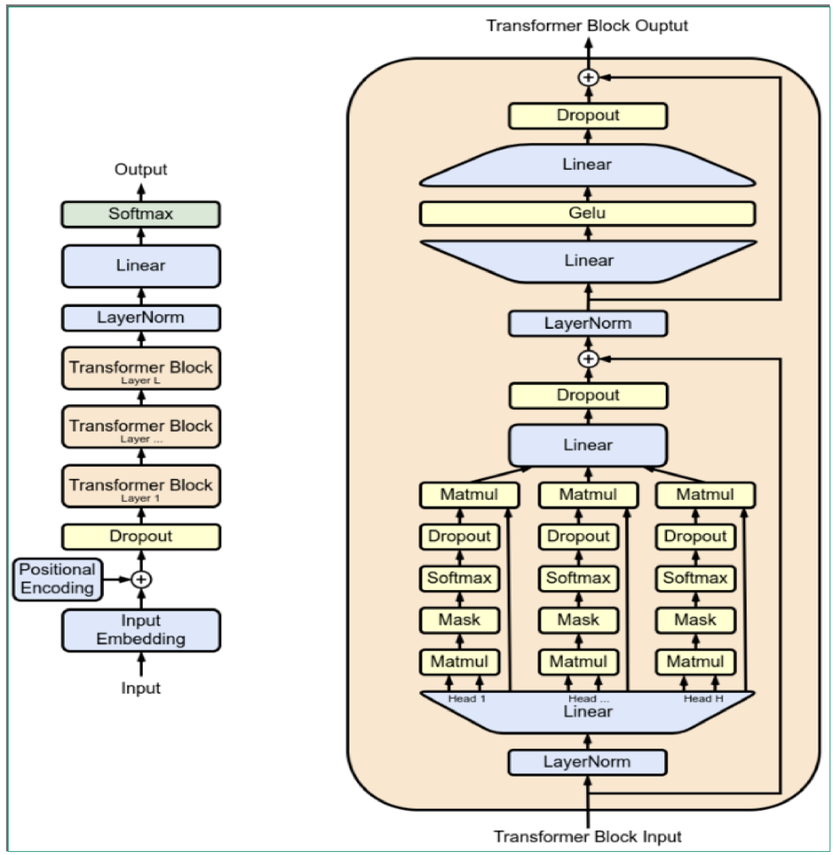
- **NeMo supports a wide variety of tasks in NLP**:
  - Punctuation And Capitalization Models
  - Token Classification (Named Entity Recognition) Model
  - Joint Intent and Slot Classification
  - Text Classification model
  - BERT
  - Language Modeling
  - Prompt Learning
  - Question Answering
  - Dialogue tasks
  - Dialogue tasks
  - GLUE Benchmark
  - Information Retrieval
  - Entity Linking
  - Model NLP
  - Machine Translation Models

- **List of tokenizers**:
  - WordPiece tokenizer,
  - SentencePiece tokenizer or
  - simple tokenizers like Word tokenizer.

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

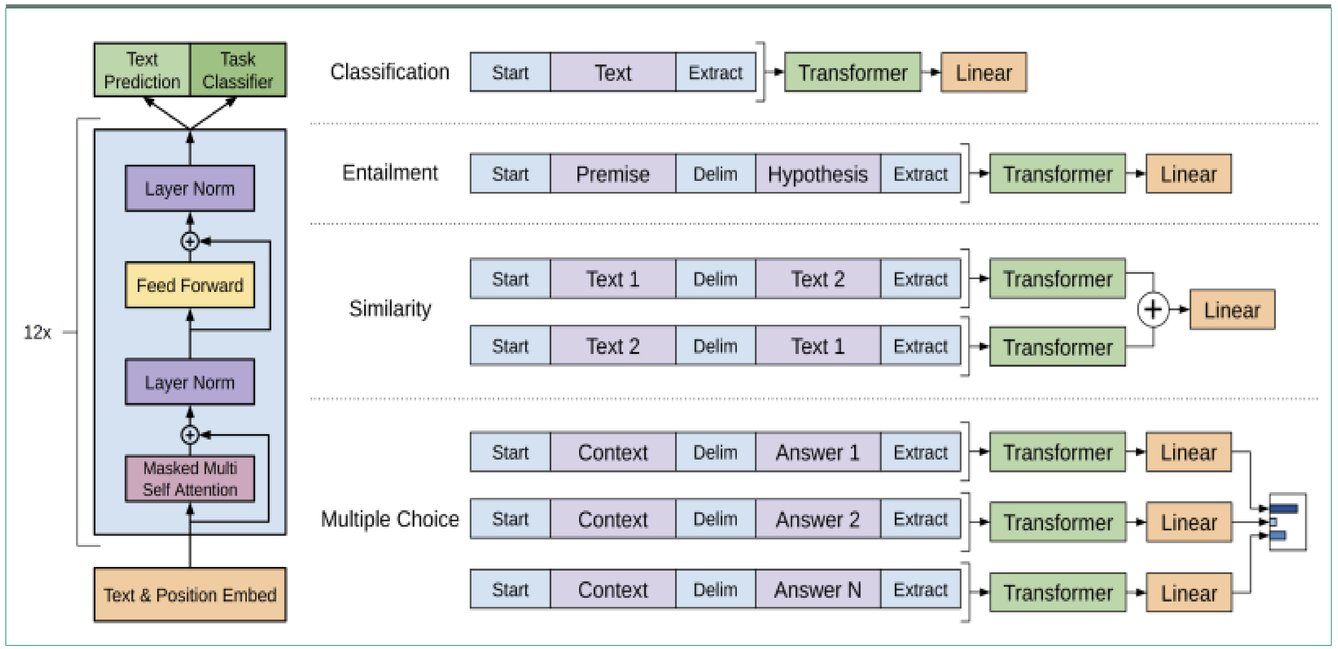
NLP Transformer

GPT (Generative Pre-trained Transformer) Architecture



Original GPT architecture

Source: https://en.wikipedia.org/wiki/GPT-2



Improved GPT architecture by OpenAI

Source:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering

Agenda

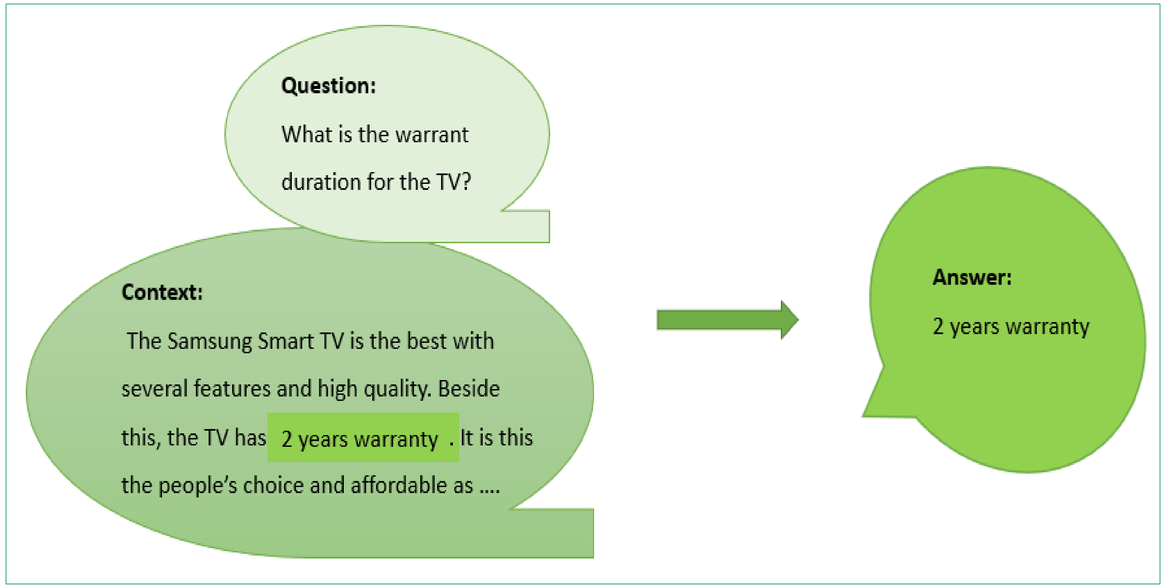
- Introduction to Question Answering (QA) Models and Architectures
- Overview of Question Answering Dataset

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Introduction to Question Answering (QA) Models and Architectures

Question Answering

Introduction



QA Definition

- Question Answering (QA) is about information retrieval whereby a question is posed to the system and a corresponding answer is replied in return.
- The QA system does this by retrieving the answer from a given context such as text or document.

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering

Different Types of QA

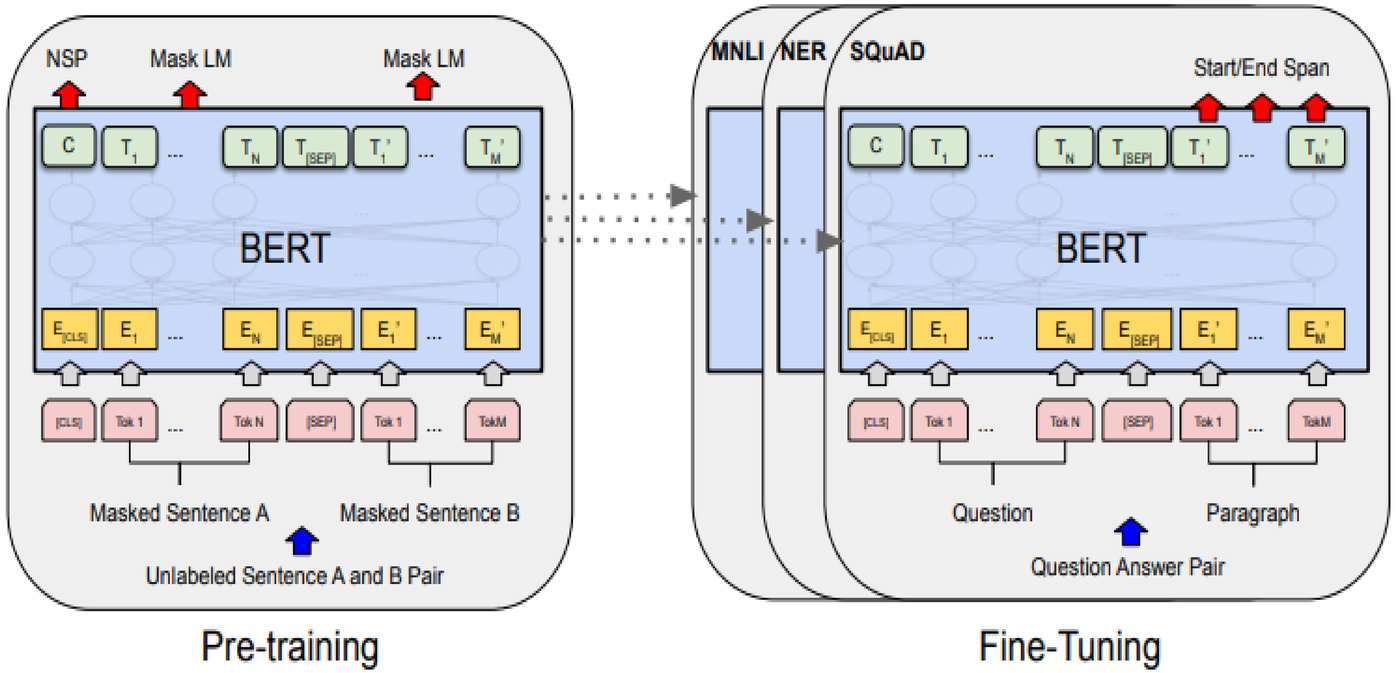
- Based on the inputs and output pattern, there are 3 different types of QA:
  - **Extractive QA** - which extracts answers from a text or document referred to as context.
  - **Open Generative QA** - that generates direct text using the context given
  - **Closed Generative QA** - generates answers without any given context
- Our focus would be on Extractive QA including examples of such datasets, and how to build custom datasets for extractive QA

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

BERT

BERT Model for Question Answering

- **BERT** (Bidirectional Encoder Representations from Transformers) model is typically trained in two phases, **pre-training** and **fine-tuning.**
- BERT is pre-trained using two unsupervised tasks: Masked LM & Next Sentence Prediction (NSP)

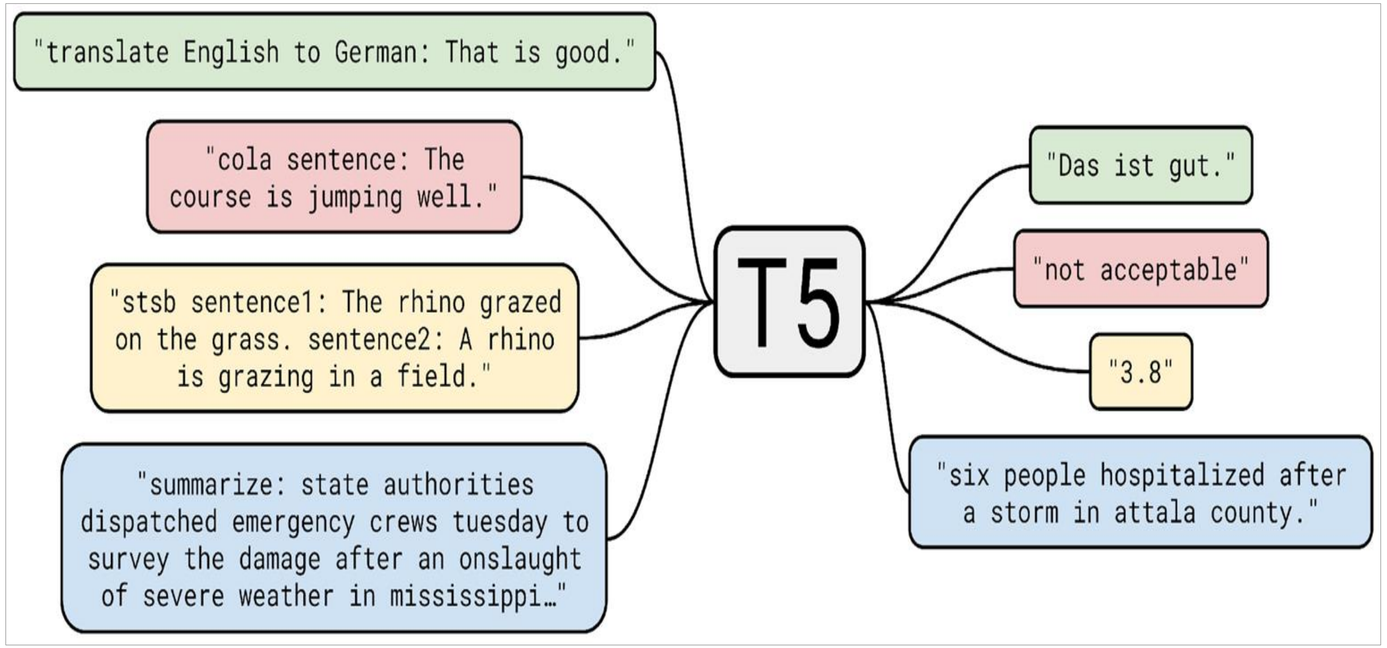


Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

T5 Model

T5 Model (Text-to-Text Transfer Transformer)



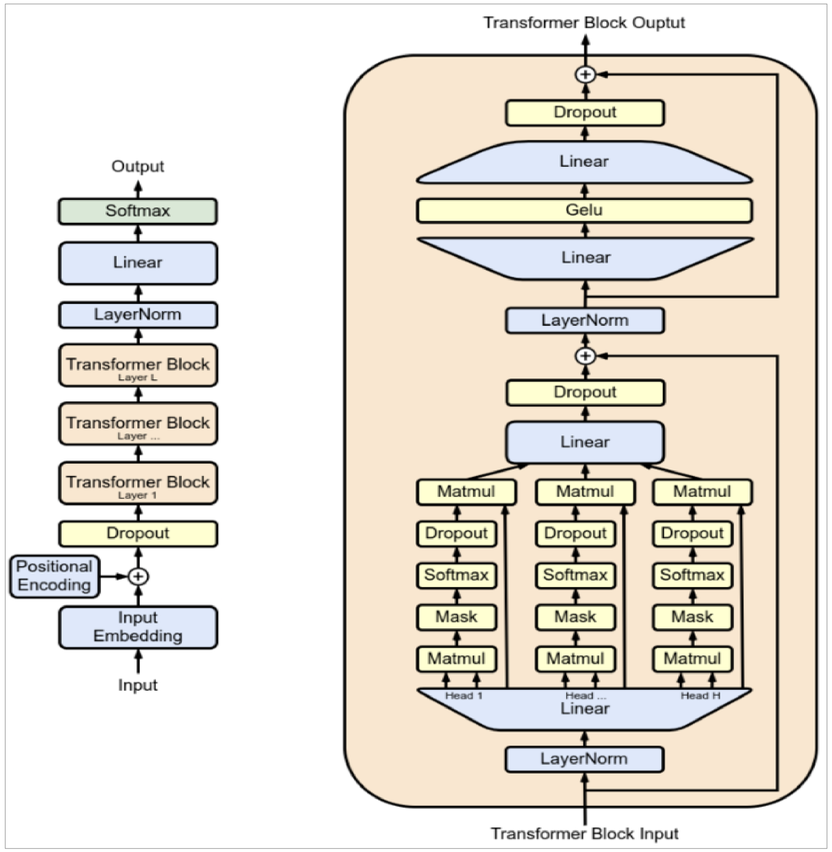
- T5 is all about reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings
- The T5 model, pre-trained on C4 dataset, achieves state-of-the-art results on many NLP benchmarks.

Source: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Raffel et al., 2020

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

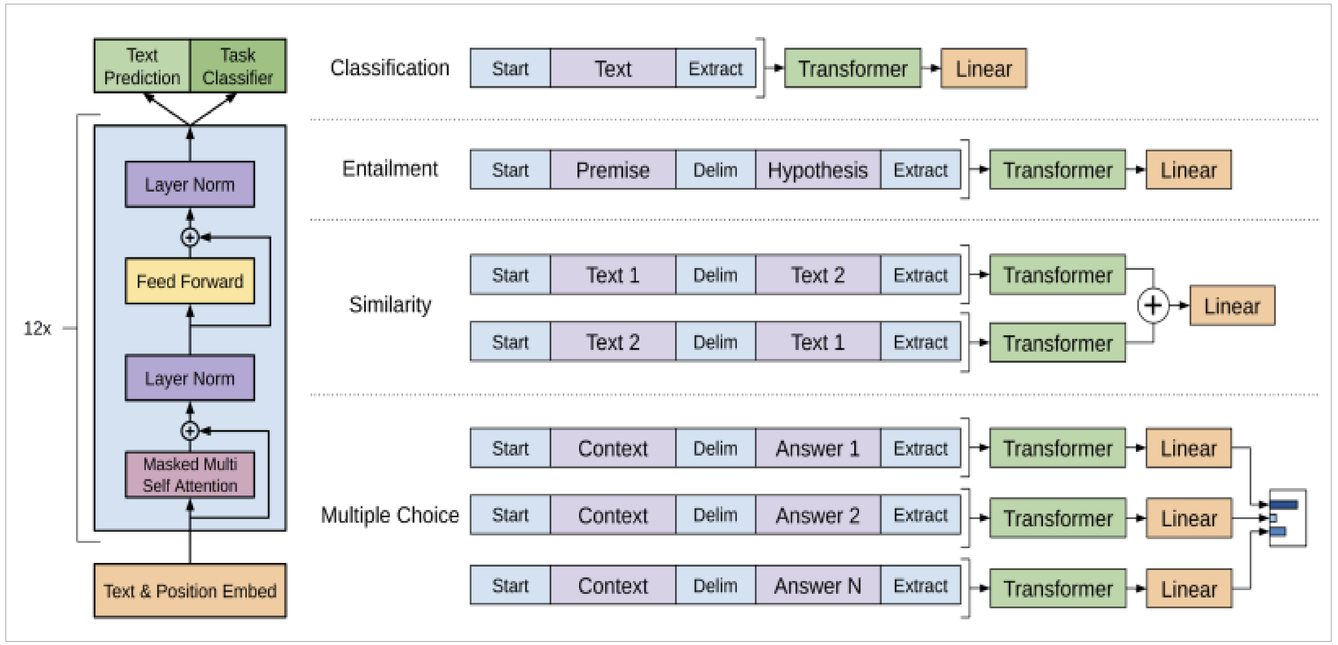
GPT Model

GPT (Generative Pre-trained Transformer) Architecture



Original GPT architecture

Source: https://en.wikipedia.org/wiki/GPT-2



Improved GPT architecture by OpenAI

Source:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Model

Other Models

Model Name	Paper link
RoBERTa	RoBERTa: A Robustly Optimized BERT Pretraining Approach Yinhan by Liu, et al., arXiv preprint, 2019.
DistilBERT	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter Victor sanh, et al., arXiv, 2019
ProQA	ProQA: Resource-efficient method for pretraining a dense corpus index for open-domain QA and IR. (2020)
GPT-4	GPT-4 Technical Report by OpenAI, 2023
DiffusionBERT	DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models by Zhengfu et al., 2022

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Overview of Question Answering Dataset

Question Answering Dataset

Stanford Question Answering Dataset (SQuAD)

Extract from SQuAD 2.0 dataset


```
{'qas': [{'question': 'What fraction of New Yorkers in the private sector are employed by foreign companies?',
   'id': '56cf4722aab44d1400b88f06',
   'answers': [{'text': 'One out of ten', 'answer_start': 113}],
   'is_impossible': False},
  {'question': 'What publication ranked New York first in the 2013 American Cities of the Future rankings?',
   'id': '56cf4722aab44d1400b88f07',
   'answers': [{'text': 'FDi Magazine', 'answer_start': 372}],
   'is_impossible': False}],
 'context': 'Many Fortune 500 corporations are headquartered in New York City, as are a large number of foreign corporations. One
out of ten private sector jobs in the city is with a foreign company. New York City has been ranked first among cities across the
globe in attracting capital, business, and tourists. This ability to attract foreign investment helped New York City top the FDi
Magazine American Cities of the Future ranking for 2013.'}
```



```
{
  'version':
    'data': [
                {
                        'title':
                 'paragraphs': [
                        { 'qas': [ { }, { }, { }, ..., { } ], 'context': }, { 'qas': [ ... ], 'context': }, { } ... { }
                        ]
                },
                { },
                { },
                .....
                {
                        'title': ,
                 'paragraphs': [
                        { 'qas': [ { }, { }, { }, ..., { } ], 'context': }, { 'qas': [ ... ], 'context': }, { } ... { }
                        ]
                }
        ]
}
```



```
{
        'question' :
                'id' : ,
        'answers' : [ {'text': , 'answer_start': } ],
    'is_impossible' :
}
```


Simplified SQUAD JSON Format

- SQuAD is a reading comprehension dataset that contains questions posed by crowd workers on a set of Wikipedia articles.
- These questions are answerable within a text paragraph known as context.
- The data format include:
  - version
  - data
  - title
  - paragraphs
  - qas
  - context
- There are 442 topics/domains and 442 paragraphs covered in the SQuAD json dataset

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Dataset

Natural Questions (NQ)

Google AI Blog

The latest from Google Research

Natural Questions: a New Corpus and Challenge for Question Answering Research

Wednesday, January 23, 2019

Posted by Tom Kwiatkowski and Michael Collins, Research Scientists, Google AI Language

Open-domain question answering (QA) is a benchmark task in natural language understanding (NLU) that aims to emulate how people look for information, finding answers to questions by reading and understanding entire documents. Given a question expressed in natural language (*"Why is the sky blue?"*), a QA system should be able to read the web (such as this Wikipedia page) and return the correct answer, even if the answer is somewhat complicated and long. However, there are currently no large, publicly available sources of *naturally occurring* questions (i.e. questions asked by a person seeking information) and answers that can be used to train and evaluate QA models. This is because assembling a high-quality dataset for question answering requires a large source of real questions and significant human effort in finding correct answers.

- The Natural Questions is a large-scale corpus dataset from Google that target open-domain question answering system.
- It contains questions issued to Google search engines and long and short answers that were annotated from Wikipedia pages.
- The full dataset is 42GB including HTML of Wikipedia pages, and contains 307k training examples, 8k examples each for testing and development respectively.
- The simplified version of NQ training dataset is 4GB

Google AI Blog Natural Questions is released under the Creative Commons Share-Alike 3.0 license

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Dataset

Natural Questions (NQ) Format


```
{'document_text': "Email marketing - Wikipedia <H1> Email marketing </H1> Jump to : navigation , search <Table> <Tr> <Td>
</Td> <Td> ( hide ) This article has multiple issues . Please help improve it or discuss these issues on the talk page .
( Learn how and when to remove these template messages ) <Table> <Tr> <Td> </Td> <Td> This article needs additional
citations for verification . Please help improve this article by adding citations to reliable sources . Unsourced
material may be challenged and removed . ( September 2014 ) ( Learn how and when to remove this template message ) </Td>
</Tr> </Table> <Table> <Tr> <Td> </Td> <Td> This article possibly contains original research . Please improve it by
verifying the claims made and adding inline citations . Statements consisting only of original research should be removed
. ( January 2015 ) ( Learn how and when to remove this template message ) </Td> </Tr> </Table> ( Learn how and when to
remove this template message ) </Td> </Tr> </Table> <Table> <Tr> <Td> Part of a series on </Td> </Tr> <Tr> <Th> Internet
marketing </Th> </Tr> <Tr> <Td> <Ul> <Li> Search engine optimization </Li> <Li> Local search engine optimisation </Li>
<Li> Social media marketing </Li>........
This email resulted in $13 million worth of sales in DEC products , and highlighted the potential of marketing through
mass emails . However , as email marketing developed as an effective means of direct communication , users began blocking
out content from emails with filters and blocking programs . In order to effectively communicate a message through email
, marketers had to develop a way of pushing content through to the end user , without being cut out by automatic filters
and spam removing software .....
</Li> <Li> </Li> <Li> </Li> <Li> </Li> <Li> </Li> <Li> </Li> </Ul> <Ul> <Li> </Li> <Li> </Li> </Ul>",

'long_answer_candidates': [{'start_token': 14, 'top_level': True, 'end_token': 170}, {'start_token': 15, 'top_level':
False, 'end_token': 169}, {'start_token': 52, 'top_level': False, 'end_token': 103}, {'start_token': 53, 'top_level':
False, 'end_token': 102}, {'start_token': 103, 'top_level': False, 'end_token': 156}, {'start_token': 104, 'top_level':
False, 'end_token': 155}, {'start_token': 170, 'top_level': True, 'end_token': 321}, {'start_token': 171, 'top_level':
False, 'end_token': 180}, {'start_token': 180, 'top_level': False, 'end_token': 186}, {'start_token': 186, 'top_level':
False, 'end_token': 224}, {'start_token': 188, 'top_level': False, 'end_token': 222}, {'start_token': 189, 'top_level':
False,.... }],

'question_text': 'which is the most common use of opt-in e-mail marketing',
'annotations': [{'yes_no_answer': 'NONE', 'long_answer': {'start_token': 1952, 'candidate_index': 54, 'end_token': 2019},
'short_answers': [{'start_token': 1960, 'end_token': 1969}], 'annotation_id': 593165450220027640}],

'document_url': 'https://en.wikipedia.org//w/index.php?title=Email_marketing&amp;oldid=814071202',

'example_id': 5655493461695504401}
```


Extract from NQ Dataset

- Each example of NQ contains:
  - a document paragraph (document_text),
  - long answer candidates (long_answer_candidates),
  - question (question_text),
  - annotations,
  - document_url , and
  - example_id.
- Training examples from the simplified version (v1.0-simplified_simplified-nq-train.jsonl.gz) are shown image on the left-side

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Dataset

Conversational Question Answering (CoQA)

CoQA: A Conversational Question Answering Challenge

Siva Reddy* **Danqi Chen*** **Christopher D. Manning**
Computer Science Department
Stanford University
{sivar,danqi,manning}@cs.stanford.edu

Abstract

Humans gather information through conversations involving a series of interconnected questions and answers. For machines to assist in information gathering, it is therefore essential to enable them to answer conversational questions. We introduce CoQA, a novel dataset for building **C**onversational **Q**uestion **A**nswering systems.[1] Our dataset contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. The questions are conversational, and the answers are free-form text with their corresponding evidence highlighted in the passage. We analyze CoQA in depth and show that conversational questions have challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning. We evaluate strong dialogue and reading comprehension models on CoQA. The best system obtains an F1 score of 65.4%, which is 23.4 points behind human performance (88.8%), indicating there is ample room for improvement. We present CoQA as a challenge to the community at https://stanfordnlp.github.io/coqa.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i)

Extract from CoQA paper

- Conversational Question Answering (CoQA) is a large-scale dataset for building conversational question-answering systems.
- The goal is to have a dataset that can measure the ability of machines to comprehend a text passage and correctly respond to a series of interconnected questions within a conversation.

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Dataset

CoQA Format


```
{'source': 'wikipedia',

 'id': '3zotghdk5ibi9cex97fepx7jetpso7',

 'filename': 'Vatican_Library.txt',

 'story': 'The Vatican Apostolic Library (), more commonly called the Vatican Library or simply the Vat, is the library of the Holy See, located
in Vatican City. Formally established in 1475, although it is much older, it is one of the oldest libraries in the world and contains one of the
most significant collections of historical texts. It has 75,000 codices from throughout history, as well as 1.1 million printed books, which
include some 8,500 incunabula. \n\nThe Vatican Library is a research library for history, law, philosophy, science and theology. The Vatican
Library is open to anyone who can document their qualifications and research needs. Photocopies for private study of pages from books published
between 1801 and 1990 can be requested in person or by mail. \n\nIn March 2014, the Vatican Library began an initial four-year project of
digitising its collection of manuscripts, to be made available online. \n\nThe Vatican Secret Archives were separated from the library at the
beginning of the 17th century; they contain another 150,000 items. \n\nScholars have traditionally divided the history of the library into five
periods, Pre-Lateran, Lateran, Avignon, Pre-Vatican and Vatican. \n\nThe Pre-Lateran period, comprising the initial days of the library, dated
from the earliest days of the Church. Only a handful of volumes survive from this period, though some are very significant.',
 'questions': [{'input_text': 'When was the Vat formally opened?',
   'turn_id': 1},
  {'input_text': 'what is the library for?', 'turn_id': 2},
  {'input_text': 'for what subjects?', 'turn_id': 3},
  {'input_text': 'and?', 'turn_id': 4},
  {'input_text': 'what was started in 2014?', 'turn_id': 5},
  {'input_text': 'how do scholars divide the library?', 'turn_id': 6},
  ..................................................................
  {'input_text': 'what will this allow?', 'turn_id': 20}],

 'answers': [{'span_start': 151,
   'span_end': 179,
   'span_text': 'Formally established in 1475',
   'input_text': 'It was formally established in 1475',
   'turn_id': 1},
  {'span_start': 454,
   'span_end': 494,
   'span_text': 'he Vatican Library is a research library',
   'input_text': 'research',
   'turn_id': 2},
  {'span_start': 457,
   'span_end': 511,
   'span_text': 'Vatican Library is a research library for history, law',
   'input_text': 'history, and law',
   'turn_id': 3},
  {'span_start': 457,
   'span_end': 545,
   'span_text': 'Vatican Library is a research library for history, law, philosophy, science and theology',
   'input_text': 'philosophy, science and theology',
   'turn_id': 4},
 {'span_start': 769,
  'span_end': 879,
  'span_text': 'March 2014, the Vatican Library began an initial four-year project of digitising its collection of manuscripts',
  'input_text': 'a  project',
  'turn_id': 5},
 {'span_start': 1048,
  'span_end': 1127,
  'span_text': 'Scholars have traditionally divided the history of the library into five period',
  'input_text': 'into periods',
  'turn_id': 6},
..............................................................................
 {'span_start': 868,
  'span_end': 910,
  'span_text': 'manuscripts, to be made available online. ',
  'input_text': 'them to be viewed online.',
  'turn_id': 20}],

'name': 'Vatican_Library.txt'}
```


- CoQA data format contains:
  - Source
  - Id
  - Filename
  - Story
  - Questions
  - Answers
  - Name

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Question Answering Dataset

Other Datasets

QA Dataset Name	Download Link	Paper link
Explain Like I'm Five (ELI5)	https://github.com/facebookresearch/ELI5	Long Form Question Answering
TriviaQA	http://nlp.cs.washington.edu/triviaqa/	TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension:
Question Answering in Context (QuAC)	https://quac.ai/	Question Answering in Context:
TWEETQA	https://aclanthology.org/P19-1496/	TWEETQA: A Social Media Focused Question Answering Dataset:

- For more on large and small Question Answering dataset, see:
  - *10 Question-Answering Datasets To Build Robust Chatbot System* by Ambika Choudhury, 2019 and
  - *University of Freiburg: Algorithms and Data Structures Group* large-qa-datasets GitHub page .

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Multitask Prompt and P-tuning

NeMo Megatron

Prompt Learning

Prompt Tuning/P-Tuning Forward Pass

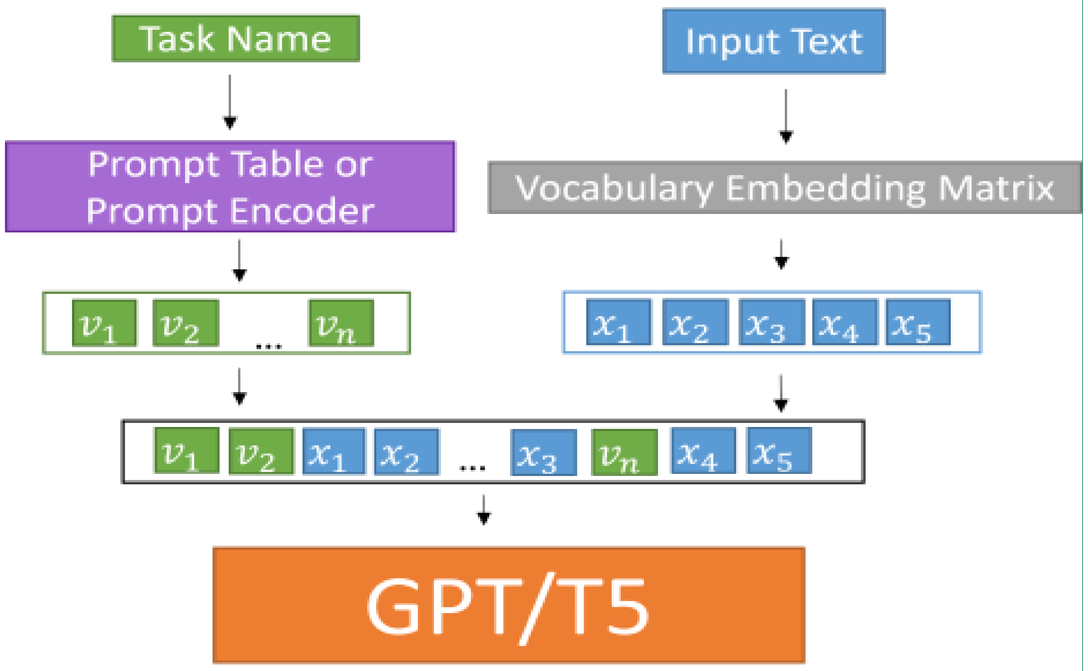
Task name and input text are passed to the model

↓

Task specific virtual tokens are retrieved based on the task name. The input text is tokenized, and token embeddings are retrieved.

↓

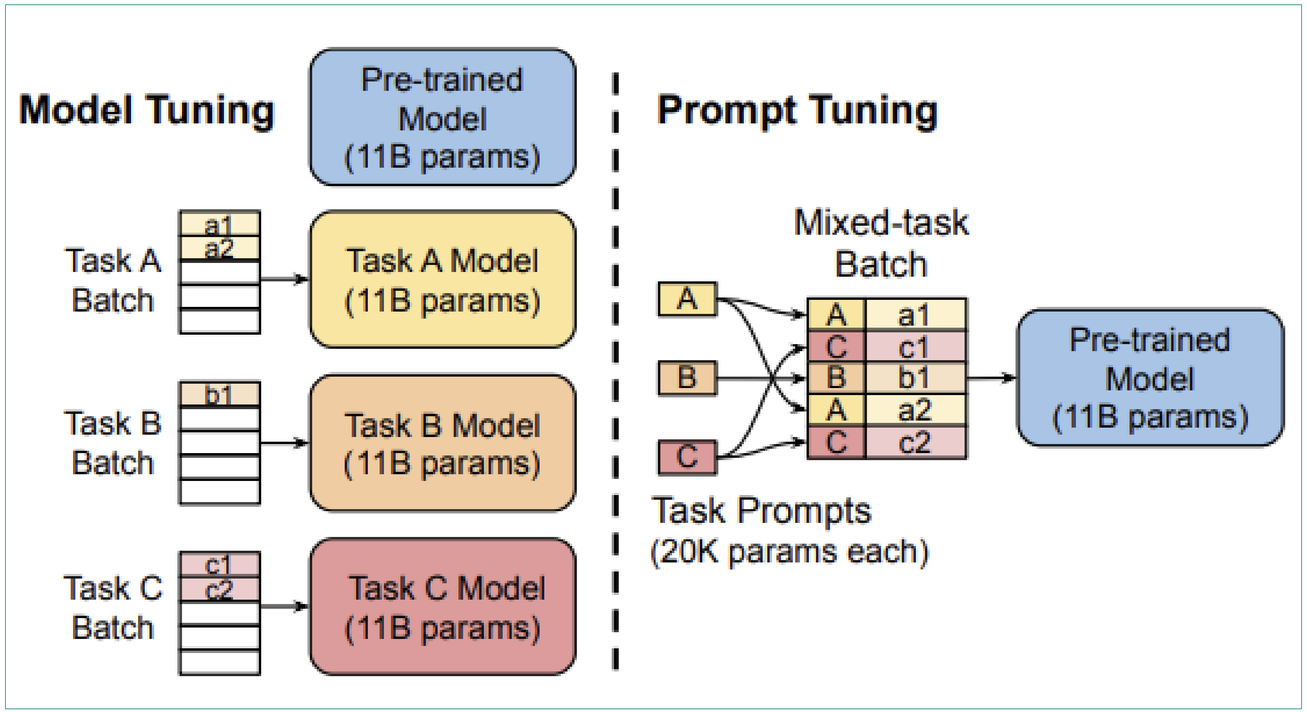
Virtual token embeddings are inserted among discrete token embeddings and passed together into the rest of the model.



Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Prompt-Tuning



Source: The Power of Scale for Parameter-Efficient Prompt Tuning by Lester et. al's EMNLP 2021 paper (https://arxiv.org/pdf/2104.08691.pdf)

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

P-Tuning

In p-tuning, an LSTM model is used to predict virtual token embeddings.

Source: GPT Understands, Too by Liu et al (https://arxiv.org/abs/2103.10385)

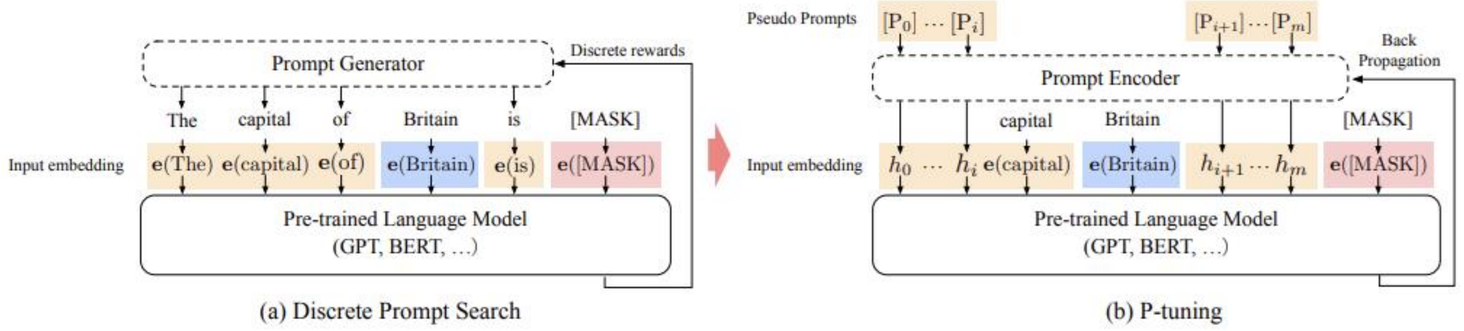


Figure 2. An example of prompt search for "The capital of Britain is [MASK]". Given the context (blue zone, "Britain") and target (red zone, "[MASK]"), the orange zone refer to the prompt tokens. In (a), the prompt generator only receives discrete rewards; on the contrary, in (b) the pseudo prompts and prompt encoder can be optimized in a differentiable way. Sometimes, adding few task-related anchor tokens (such as "capital" in (b)) will bring further improvement.

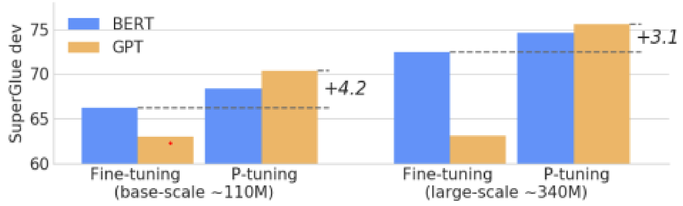
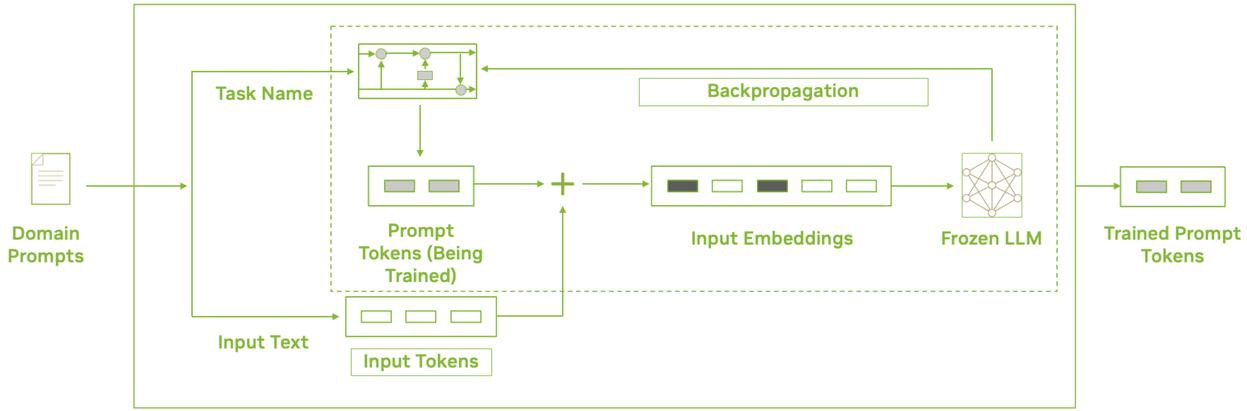


Figure 1. Average scores on 7 dev datasets of SuperGlue. GPTs can be better than similar-sized BERTs on NLU with P-tuning.



Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Prompt tuning/P-Tuning

- A single pretrained GPT model can use both p-tuning and prompt-tuning.
- While you must decide to use either p-tuning or prompt-tuning for each task you want your model to perform, you can p-tune your model on a set of tasks A, then prompt tune your same model on a different set of tasks B, then finally run inference on tasks from both A and B at the same time.
- During prompt-tuning or p-tuning, tasks tuned at the same time must use the same number of virtual tokens.
- During inference, tasks using differing amounts of virtual tokens can be run at the same time.

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Dataset Preprocessing

```
[
  {"taskname": "squad", "context": [CONTEXT_PARAGRAPH_TEXT1], "question": [QUESTION_TEXT1], "answer": [ANSWER_TEXT1]},
  {"taskname": "squad", "context": [CONTEXT_PARAGRAPH_TEXT2], "question": [QUESTION_TEXT2], "answer": [ANSWER_TEXT2]},
  {"taskname": "intent_and_slot", "utterance": [UTTERANCE_TEXT1], "label": [INTENT_TEXT1][SLOT_TEXT1]},
  {"taskname": "intent_and_slot", "utterance": [UTTERANCE_TEXT2], "label": [INTENT_TEXT2][SLOT_TEXT2]},
  {"taskname": "sentiment", "sentence": [SENTENCE_TEXT1], "label": [SENTIMENT_LABEL1]},
  {"taskname": "sentiment", "sentence": [SENTENCE_TEXT2], "label": [SENTIMENT_LABEL2]},
]
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

SQUAD Dataset Prompt Formatting

```
{"taskname": "squad", "context": "Super Bowl 50 was an American football ga... numerals 50.",
"question": "What does AFC stand for?", "answer": "American Football Conference“ }.
```

```
config.model.task_templates = [

    {
        "taskname": "squad",
        "prompt_template": "<|VIRTUAL_PROMPT_0|> Context: {context}\n\nQuestion:
{question}\n\nAnswer:{answer}",
        "total_virtual_tokens": 15,
        "virtual_token_splits": [15],
        "truncate_field": "context",
        "answer_only_loss": True,
        "answer_field": "answer",
    },

]
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Setting The Pre-Trained GPT Model

Check what GPT .nemo models we have available on NGC

```
from nemo.collections.nlp.models.language_modeling.megatron_gpt_model import MegatronGPTModel
MegatronGPTModel.list_available_models()
```

Direct use of 345M parameter GPT model

```
gpt_model = MegatronGPTModel.from_pretrained(model_name="megatron_gpt_345m", trainer=trainer).cuda()
```

Download the model from NGC and set GPT model path on prompt learning config

```
gpt_file_name = "megatron_gpt_345m.nemo"
!wget  -nc --content-disposition
https://api.ngc.nvidia.com/v2/models/nvidia/nemo/megatron_gpt_345m/versions/1/files/megatron_g
pt_345m.nemo -O {NEMO_DIR}/{gpt_file_name}
config.model.language_model_path = gpt_file_name
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Setting P-Tuning Specific Params

Set the model.virtual_prompt_style hyperparameter:

```
from nemo.collections.nlp.modules.common import VirtualPromptStyle
config.model.virtual_prompt_style = VirtualPromptStyle.P_TUNING
```

We can set the two p-tuning specific parameters:

- p_tuning.dropout
- p_tuning.num_layers

```
config.model.p_tuning.dropout = 0.0
config.model.p_tuning.num_layers = 2
config.model.global_batch_size = 2
config.model.micro_batch_size = 1
```

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

NeMo Megatron

Setting Prompt-Tuning Specific Params

- Prompt tuning specific parameters
  - **`prompt_tuning.new_prompt_init_methods`**
  - **`prompt_tuning.new_prompt_init_text`**

- Each of the above hyperparameters are a list of strings.
- `new_prompt_init_methods` would look like `["text", "random", "text", "text"]`
- `new_prompt_init_text` might look like `["some text I want to use", None, "some other text", "task text goes here"]` for those four new tasks

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

Q & A

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.

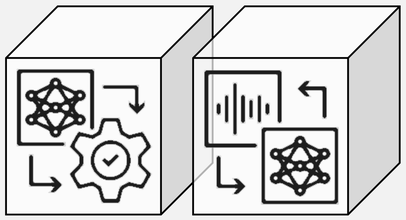
Wrapping Up

NVIDIA NeMo

Multiple Products to Help Build Custom LLMs

NeMo Framework

General availability in Sept. with **NVIDIA AI Enterprise 4.0**



NVIDIA AI Enterprise

Also available on Git Hub and NGC

NeMo Pretrained Models

Available from NGC and Hugging Face

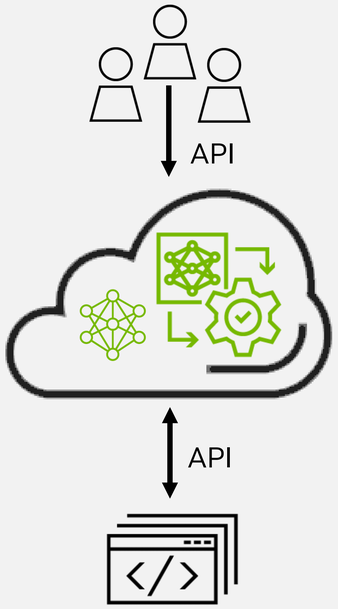
Community Models

Compatible with NeMo framework

Falcon LLM Llama 2 MPT

NeMo Service

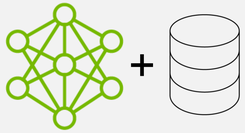
Limited early access for target accounts



NeMo Research

Early development techniques for researchers/developers

NeMo Inform



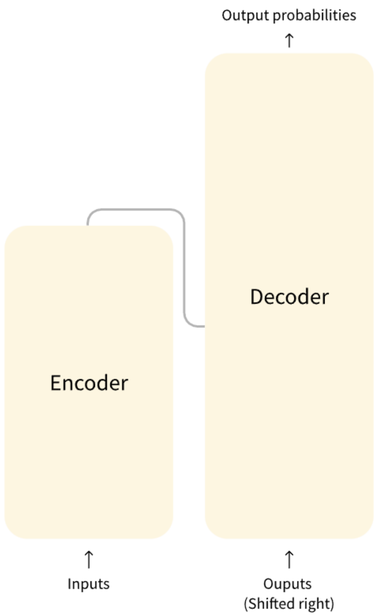
NeMo Guardrails

NeMo SteerLM

NeMo Expanding Support Across Modalities

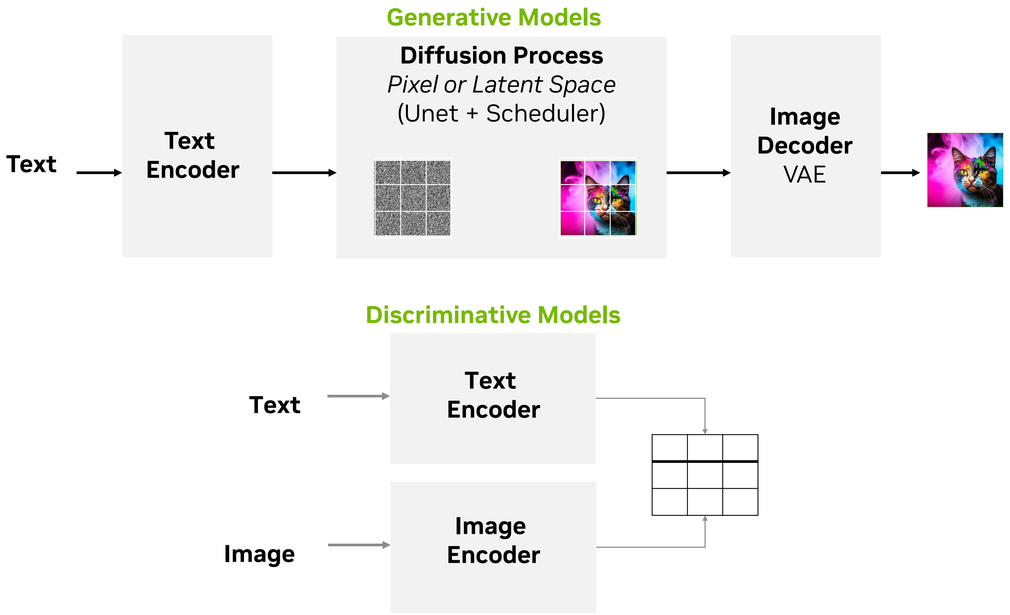
NeMo offers multi-modality support

Language Models



- **Encoder Only** *(Understanding Use-Cases)* **-** BERT
- **Decoder Only** *(Generating Use-Cases)* **-** GPT
- **Encoder-Decoder** *(Translation and Q&A Use-Cases)* **–** T5, mT5, T5-MoE

Image Models



Generative Models

- *Diffusion in Latent Space:* Stable Diffusion v1.5
- *Diffusion in Pixel Space:* Imagen
- *Image-to-Image Models:* Instruct-Pix2Pix, DreamBooth

Discriminative Models

- Vision-Transformers (ViT)
- *Multi-Modal:* CLIP

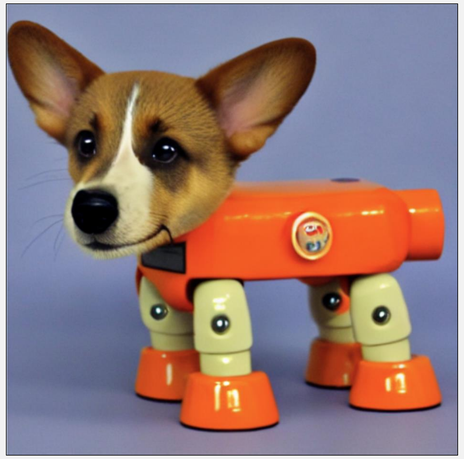
Model Architecture Support Across Modalities

Nemo Framework Now Supports Training And Deployment of Popular Model Architectures

Language Models

GPT
T5, mT5, T5-MoE
BERT

Text-to-Image Models



Prompt: A 'sks' dog mecha robot.

Stable Diffusion v1.5
Imagen
Vision Transformers
CLIP

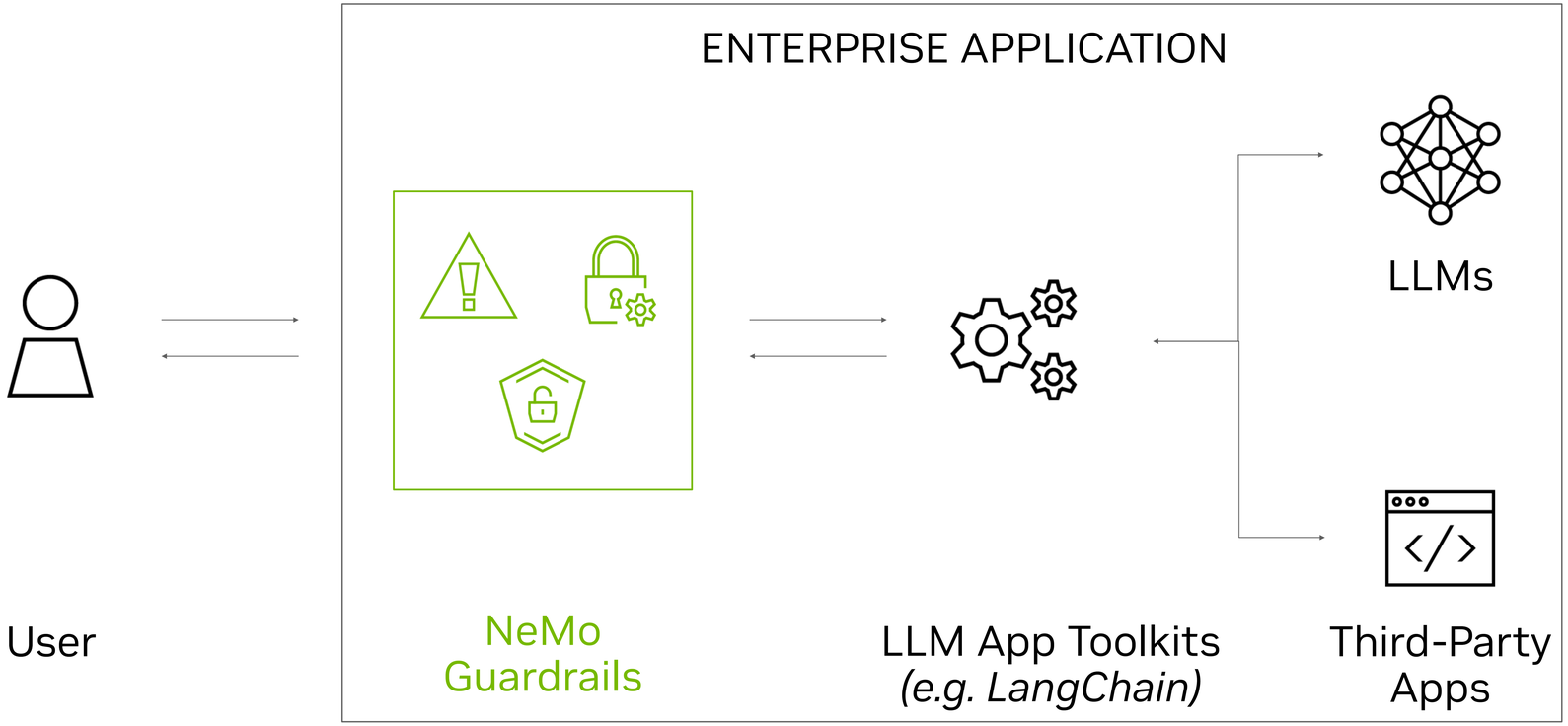
Image-to-Image Models

Instruction: Make it on a beach

Dreambooth
InstructPix2Pix

NeMo Guardrails

Add Boundaries Ensure Chatbots Operate According to Use Cases



Question: ask capabilities

"What can you do?"
"tell me what you can do"

Answer:

"I am an AI assistant which helps answer questions based on a given knowledge base."

Question: ask off topic

"Who is the president?"
"Can you recommend the best stocks to buy?"
"Can you tell me a joke?"

Answer:

"I cannot comment on anything which is not relevant to the job report"

TOPICAL

Focus interactions within a specific domain

SAFETY

Prevent hallucinations, toxic or misinformative content

SECURITY

Prevent executing malicious calls and handing power to a 3rd party app

Announcing NeMo SteerLM

Model customization technique to create various personalities at inference with just a single training run

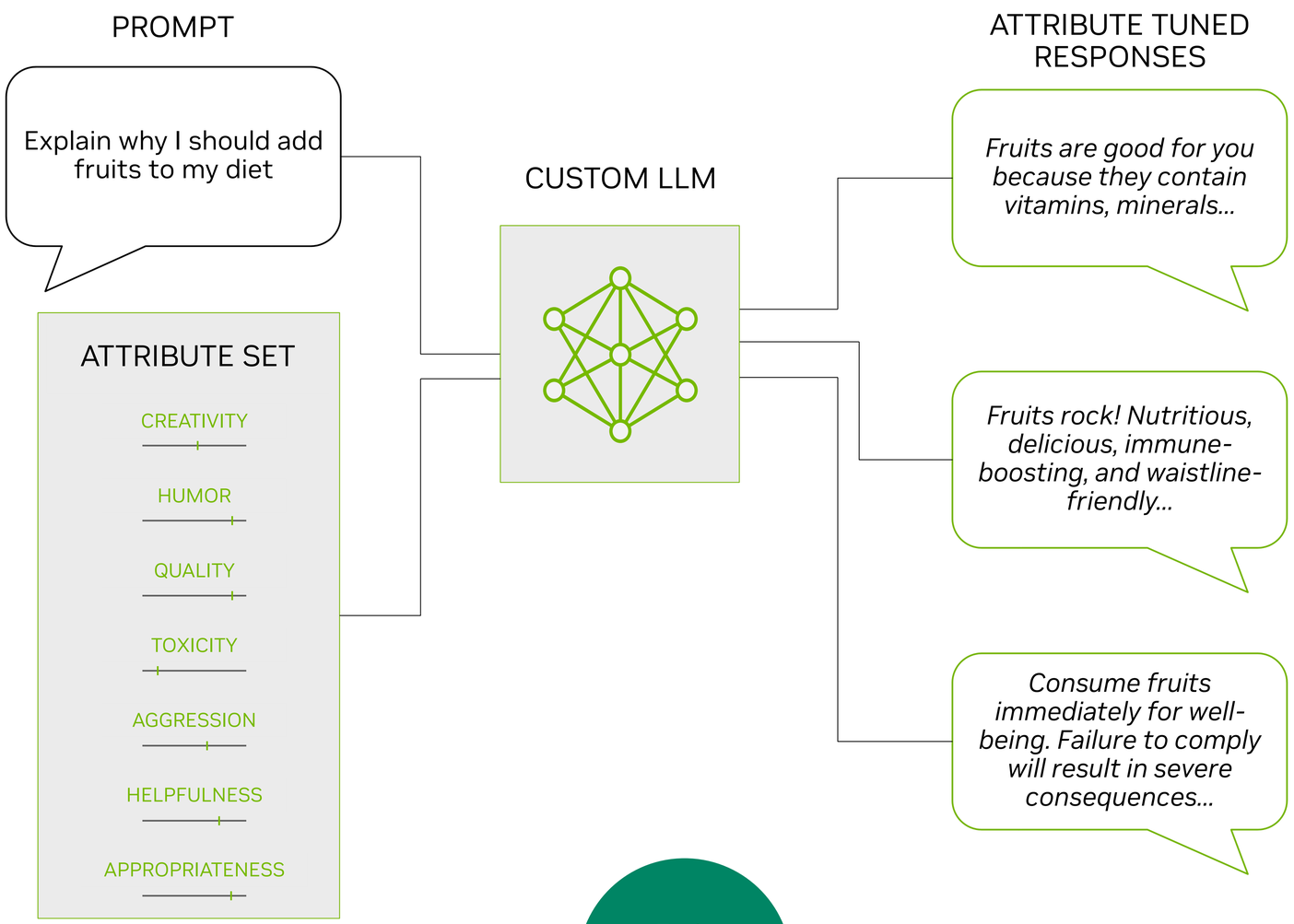
New training and inferencing technique from NVIDIA Research

Define Tunable Attributes During Training

Simple SFT-Like Training Process

Tune LLM responses across multiple attributes during inference

Now Available For Experimentation In Open-Source From GitHub



Available for experimentation in open-source from GitHub

Early Access in August

NVIDIA Developer Program

The Community that Builds

Program Benefits:

Tools

- 550+ exclusive SDKs and models
- GPU-optimized software, model scripts, and containerized apps
- Early access programs

Training

- Research papers, technical documentation, webinars, blogs, and news
- Technical training and certification opportunities
- 1,000s of technical sessions from industry events On-Demand

Community

- NVIDIA developer forums
- Exclusive meetups, hackathons, and events

Join the Community

** The Hardware Grant Program is available to qualified researchers and educators.*

Challenge

NVIDIA
THANK YOU

NeMo Megatron

Hands-On

- How to build a PyTorch Lightning Trainer
- Setting up a NeMo Experiment
- Sample P-Tuning and inferencing session

Details on these are covered in the Hands-On Jupyter Notebooks

Copyright © 2023 OpenACC-Standard.org. This material is released by OpenACC-Standard.org, in collaboration with NVIDIA Corporation, under the Creative Commons Attribution 4.0 International (CC BY 4.0). These materials include references to hardware and software developed by other entities; all applicable licensing and copyrights apply.