Generative AI for Science

Unlocking the power of LLMs with NVIDIA NeMo

Janaki Vamaraju, Senior Solution Architect, NVIDIA
Zahra Ronaghi, Manager Solution Architect, NVIDIA
Generative AI and Large Language Models (LLMs)

NVIDIA NeMo Framework

Retrieval Augmented Generation (RAG)

Domain Adapted LLMs
When to use Generative AI?

Predictive AI focuses on understanding historical data and making accurate predictions.

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

- **Classification**
- **Optical Character Recognition**
- **Pattern Recognition**
- **Text to Speech**
- **Sentiment Analysis**
- **Language Translation**
- **Unsupervised Learning**
- **Knowledgebase Copilot**
- **Summarization**
- **Chatbot**

**Examples of Predictive AI:**
- SPAM and INBOX
- THE QUICK BROWN FOX J...U...M...P...S

**Examples of Generative AI:**
- OCR
- QA
- Optical Character Recognition
- Language Translation
- Unsupervised Learning
- Knowledgebase Copilot
- Summarization
- Chatbot

**Technologies:**
- Text to Speech
- OCR
- Q&A
- Language Translation
- Unsupervised Learning
- Knowledgebase Copilot
- Summarization
- Chatbot
Intersection of Gen AI and Science
Building Foundation Models for Science Research and Discovery

Summarize

Synthesize

Generate
How to train an LLM
Creating a “Foundation Model”

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

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

  - ‘Q: What virus causes covid?
  - ‘Q: Write a poem about a cat in love with a zebra.
    - A: There once was a cat
      in search for a mate.
      She saw a zebra
      And knew it was fate…’

  - ‘Q: Code Quicksort in C++
  - ‘Q: Who do want to win the next election?
    - A: As an AI, I do not have political opinions’
Requirements for Building Custom LLMs

**Training Data**
- Data collection and preparation
- Cloud storage

**Accelerated Computing**
- DGX & DGX Cloud
- AWS, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure, Dell Technologies, Hewlett Packard Enterprise, Lenovo, Supermicro

**Training and Inference Tools**
- Data Curation
- Foundations Models
- Training & Customization
- Accelerated Inference

**AI Expertise**
- Internal Expertise
- Solution Delivery Partners
Building Generative AI Applications
Build, customize and deploy generative AI models with NVIDIA NeMo
https://github.com/NVIDIA/NeMo

Data Curation
Megatron Core
NeMo Curator

Distributed Training
NeMo Aligner

Model Customization
In-domain, secure, cited responses
NeMo Aligner

Model Evaluation
In-domain queries
NeMo Evaluator

Accelerated Inference
Triton & TensorRT-LLM
NeMo Retriever

Retrieval Augmented Generation
NeMo Guardrails

Multi-Modality
Build language, image, generative AI models

Data Curation at Scale
Extract, deduplicate, filter info from large unstructured data at scale

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

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

Deploy at Scale
Run optimized inference at-scale anywhere

Guardrails
Keep applications aligned with safety and security requirements using NeMo Guardrails

NVIDIA NeMo
Data Curation Improves Model Performance
NeMo Data Curator enabling large-scale high-quality datasets for LLMs

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

**NeMo Data Curator steps:**

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

![Diagram of data curation process]

Internet scale datasets

Data download + detect language + extract text

Text re-formatting + cleaning

Document-level deduplication

Training

Document-level quality filtering

Task deduplication

Data blending
Model Customization for LLMs

Customization techniques to overcome the challenges of using foundation models

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

**Model Customization**

- (p-tuning, Prompt Tuning, ALibi, Adapters, LoRA)
- **Prompt Learning**
  Add skills and incremental knowledge
- **Supervised Fine Tuning**
  Include domain-specific knowledge
- **Reinforcement Learning from Human Feedback (RLHF)**
  Continuously improve model as it is used

**Your Enterprise Model**

- **Supply Chain Forecasting**
- **Financial Modeling**
- **Sales Pipeline Analysis**
- **Legal Contract Discovery**

**Foundation Model**

- Start with pre-trained model
# Suite of Model Customization Tools in NeMo

**Ways To Customize Large Language Models For Your Use-Cases**

## Techniques

<table>
<thead>
<tr>
<th>PROMPT ENGINEERING</th>
<th>PROMPT LEARNING</th>
<th>PARAMETER EFFICIENT FINE-TUNING</th>
<th>FINE TUNING</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Few-shot learning</td>
<td>- Prompt tuning</td>
<td>- Adapters</td>
<td>- SFT</td>
</tr>
<tr>
<td>- Chain-of-thought reasoning</td>
<td>- P-tuning</td>
<td>- LoRA</td>
<td>- RLHF</td>
</tr>
<tr>
<td>- System prompting</td>
<td></td>
<td>- IA3</td>
<td>- SteerLM</td>
</tr>
</tbody>
</table>

## Benefits

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>Accuracy for specific use-cases</th>
<th>Data, compute &amp; investment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few-shot learning</td>
<td>Good results leveraging pre-trained LLMs</td>
<td>Lowest investment</td>
</tr>
<tr>
<td>Chain-of-thought reasoning</td>
<td>Better results leveraging pre-trained LLMs</td>
<td>Lower investment</td>
</tr>
<tr>
<td>System prompting</td>
<td>Best results leveraging pre-trained LLMs</td>
<td>Will not forget old skills</td>
</tr>
</tbody>
</table>

## Challenges

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>Accuracy for specific use-cases</th>
<th>Data, compute &amp; investment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannot add as many skills or domain specific data to pre-trained LLM</td>
<td>May forget old skills</td>
<td>May forget old skills</td>
</tr>
<tr>
<td>Less comprehensive ability to change all model parameters</td>
<td>Medium investment</td>
<td>Medium investment</td>
</tr>
<tr>
<td>More expertise needed</td>
<td>Large investment</td>
<td>Large investment</td>
</tr>
<tr>
<td>Will not forget old skills</td>
<td>Most expertise needed</td>
<td>Most expertise needed</td>
</tr>
</tbody>
</table>

---

https://github.com/NVIDIA/NeMo-Aligner
NVIDIA NeMo Works with Powerful Generative Foundation Models

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

**Fastest Responses**
- Nemotron-3 8B
  - GPT-8B w/ 3.5T tokens + SFT, SteerLM
  - 53 Languages I/O: 4K tokens

**Balance of Accuracy - Latency**
- Nemotron-3 22B
  - GPT-22B w/ 1.1T tokens + SFT private mix
  - 50 Languages I/O: 4K tokens

**For Complex Tasks**
- Nemotron-3 43B
  - GPT-43B w/ 1.1T tokens + SFT private mix
  - 50 Languages I/O: 4K tokens

**Information Retrieval**
- NeMo Retriever

**Community-Built Models**
- Code Llama
  - Meta
- Falcon LLM
  - Falcon
- Llama 2
  - Meta
- MPT
  - Mosaic ML
- StarCoder
  - ServiceNow & Hugging Face
Guardrails Can Keep Generative AI On Track

Ensure accuracy, appropriateness, and security in LLMs

Topical Guardrails
Focus interactions within a specific domain

Safety Guardrails
Prevent hallucinations, toxic or misinformative content

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

NeMo Guardrails

LLM App Toolkits (e.g. LangChain)

LLMs

Third-Party Apps
NVIDIA’s LLM offerings for Training And Inference
All Are available on Github and NGC

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

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

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

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

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

---

**Pre-train**  
**Post Pre-train**  
**Inference**

- **Tested and validated for productization**
- **Example**
Decades of Scientific Research Intersecting with GenAI

3 Distinct Categories

**Summarize**
- OTS LLMs
- RAG
- Guardrails

**Synthesize**
- OTS LLM
- Multiple Data Sources
- Customization/Tuning
- Guardrails
- RAG

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

NIM, NeMo Models, NeMo Retriever, Guardrails

Nemo FW, TRT-LLM

MegatronCore
NVIDIA NIM Streamlines the Path to Production
Easiest and most performant way to deploy generative AI and LLM models coupled with industry-standard APIs

- **Prebuilt container and helm chart** tested and validated across infrastructure
- **Industry standard APIs** with NVIDIA Unified Cloud Standards
- **Domain specific code** for each NIM domain category, including LLMs, Images, VLMs, video, healthcare, biology, genomics, and more
- **Optimized inference engines** for each model and hardware SKU
- **Support for custom models** built by users targeted use cases
- **NVIDIA AI Enterprise** approved base container
- **Deploy anywhere and maintain control** of generative AI applications and data
- **Simplified development** of AI application that can run in enterprise environments
- **Day 0 support** for state-of-the-art generative AI models providing choice across the ecosystem
- **Improved TCO** with best latency and throughput running on accelerated infrastructure
- **Best accuracy** for enterprise by enabling tuning with proprietary data sources
- **Enterprise software** with feature branches, validation and support
NVIDIA NIM is the Fastest Path to AI Inference
Reduces engineering resources required to deploy optimized, accelerated models

<table>
<thead>
<tr>
<th>NVIDIA NIM</th>
<th>Triton + TRT-LLM Opensource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployment Time</td>
<td>5 minutes</td>
</tr>
<tr>
<td>API Standardization</td>
<td>Industry standard protocol, OpenAI for LLMs, Google Translate Speech</td>
</tr>
<tr>
<td>Pre-Built Engine</td>
<td>Pre-built TRT-LLM engines for NV and community models</td>
</tr>
<tr>
<td>Triton Ensemble/ BLS Backend</td>
<td>Pre-built with TRT-LLM to handle pre/post (tokenization) processing</td>
</tr>
<tr>
<td>Triton Deployment</td>
<td>Automated</td>
</tr>
<tr>
<td>Customization</td>
<td>Supported – P-tuning and LORA, more planned</td>
</tr>
<tr>
<td>Container Validation</td>
<td>Pre-validated with QA testing</td>
</tr>
<tr>
<td>Support</td>
<td>NVIDIA AI Enterprise - Security and CVE support, scanning/patching and tech support</td>
</tr>
</tbody>
</table>
Retrieval Augmented Generation (RAG)
LLMs are Powerful Tools but Not Accurate Enough

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

- Lacking proprietary knowledge
- Risk of outdated information
- Hallucinations
Use Retrieval-Augmented Generation (RAG)

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

Foundation Model

LLM Cloud API
Start with a pre-trained model provided by a 3rd party

LLM Framework

AI use case

AI Chatbot

Domain Data
Augment a response with relevant contextual information
Use Retrieval-Augmented Generation (RAG)
Represent data as embeddings to support “soft” vector similarity search

Foundation Model

LLM Cloud API
Start with a pre-trained model provided by a 3rd party

LLM Framework

Vector Database
Find relevant context using soft vector search in the embedding space

Embedding Cloud API
Represent data semantics as high dimensional vectors

Domain Data
Augment a response with relevant contextual information

Al use case
Chatbot
Use Retrieval-Augmented Generation (RAG)
Increase context relevance using domain-specific (re)ranking algorithm

Foundation Model

LLM Cloud API
Start with a pre-trained model provided by a 3rd party

LLM Framework

Vector DB + Ranking Cloud API
Rank the results using domain-specific algorithm for higher context relevance

Embedding Cloud API
Represent data semantics as high dimensional vectors

Domain Data
Augment a response with relevant contextual information

AI use case
Chatbot
Fine-tune Your Model to Understand Domain Semantics

Increase LLM accuracy by customizing for your enterprise use case

Customized Domain-Specific Model

Fine-tuned LLM Cloud API
- Start with a pre-trained model provided by a 3rd party and fine-tune it on your data

LLM Framework

Embedding Cloud API
- Represent data semantics as high dimensional vectors

Vector DB + Ranking Cloud API
- Rank the results using domain-specific algorithm for higher context relevance

AI use case

Enterprise Chatbot

Enterprise Data
- Augment a response with relevant contextual information
Adopt Open Source Models to Gain Flexibility and Control
Open source models (LLM, embedding, ranking) help protect enterprise data and IP

Your Model

Fine-tuned Custom LLM API
Start with a pre-trained model provided by a 3rd party and fine-tune it on your data

LLM Framework

Custom Ranking API
Rank the results using domain-specific algorithm for higher context relevance

Custom Embedding API
Represent data semantics as high dimensional vectors

AI use case

Chatbot

Augment a response with relevant contextual information

Falcon 40B
Gemma 2B
Gemma 7B
Llama-2 7B
Llama-2 13B
Llama-2 70B
Code Llama 34B
Mistral 7B
Mixtral 8x7B
Nemotron 8B
Nemotron 43B
GPT3 175B
MPT 30B
NVIDIA Provides Optimized Retrieval Augmented Generation

Commercially viable, optimized embedding, reranking, and personalization to deliver highest accuracy and performance
NeMo Retriever Supercharges RAG Applications

World Class Accuracy and Throughput

- **2X** World-class accuracy with nearly 2x fewer incorrect answers
- **7X** Faster embedding inference throughput
- Optimized Inference Engines
- World class models and community model support
- Flexible and modular deployment
- Customizable models and pipelines
- Production Ready

Diagram:

- Embedding Microservices
- Vector Database
- Structured Data (ERP, CRM)
- Data
- Retrieval Microservices
- Text-to-SQL Microservice
- Reranking Microservices
- Prompt
- Event
- NIM
- Plan
Domain Adapted LLMs
Building a Domain Specific Gen AI model is a Multistage Process

1. Accuracy
2. Latency
3. Cost

Domain Specific GenAI
- Train it on a skill – perform a task in a certain way
- Give it ethics and personality – align its response based on human preferences and values
- Teach it a set of facts – connect to a knowledge base

Pipeline Structure
- Models Selection
- RAG
- Context Learning

Dataset
- Domain Data
- Data Blends
- Open-Source Dataset

Model Domain Adaptation
- LLM Model
- Embedding
- Re-Ranker
LLM Assistant for Chip Design - ChipNeMo

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

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

Using proprietary data to customize a foundation model, researchers found that a much smaller 13B parameter model could outperform larger general purpose LLMs.
ChipNeMo LLM Assistant

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

**Accuracy**
- Correctness on wide range of **domain-specific tasks**
- Avoid **security risks** with third party APIs
- Model **groundedness** in the chip domain (e.g. retrieval hit-rate)

**Latency**
- Fast batch evaluation on domain-specific benchmarks
- **Real-time responses** for NVIDIA engineers

**Cost**
- **Development** (GPU training time, number of data samples and pre-training tokens)
- **Operations** (reduced inference cost at scale)
End-2-End ChipNeMo Customization Workflow

Domain-specific models lead to higher accuracy and lower cost

Pretraining
- Trillions of tokens of internet data
  - $10^5$-$10^6$ GPU hrs

Foundation Model
- Llama 2 (70B, 13B, 7B)

Domain Adaptive Pretraining
- 24B tokens of chip design docs/code
  - ~5000 GPU hrs

Supervised Fine-Tuning
- 128K chat instructions + 1.1K task instructions
  - ~100 GPU hrs

ChipNeMo Foundation Model (13B, 7B)

ChipNeMo Chat Model (13B, 7B)

Pre-trained Models
Training and Customization
Deployment

https://arxiv.org/abs/2311.00176
ChipNeMo Data Curation

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

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

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

Breakdown of Domain SFT data (128000 samples)

<table>
<thead>
<tr>
<th>Domain Source</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Knowledge</td>
<td>280</td>
</tr>
<tr>
<td>EDA Script Generation</td>
<td>480</td>
</tr>
<tr>
<td>Bug summarization and analysis</td>
<td>392</td>
</tr>
<tr>
<td>Total</td>
<td>1152</td>
</tr>
</tbody>
</table>

Data relevance and quality > quantity
Data anonymization and privacy should be considered in dataset compilation
Continuous data updating process critical to keep the training set relevant
Data curation & management play important role
Domain-adaptive Foundation Model Pretraining

Custom Tokenization

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

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

ChipNeMo Tokenizer Augmentation Improvements
Domain-adaptive Foundation Model Pretraining

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

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

Larger and more performant foundational models yielded better zero shot results on domain-specific tasks.
Supervised Fine-Tuning
Customization of model behaviour for high performance on specific tasks

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Pretraining</th>
<th>DAPT</th>
<th>SFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>7B</td>
<td>184,320</td>
<td>2,620</td>
<td>90</td>
</tr>
<tr>
<td>13B</td>
<td>368,640</td>
<td>4,940</td>
<td>160</td>
</tr>
<tr>
<td>70B</td>
<td>1,720,320</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Training cost of models in GPU hours.

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

Importance of quality and relevance of labeled data for fine-tuning
Adopt techniques for efficient fine-tuning without compromising model generalizability
Evaluation metrics tailored to specific tasks to gauge SFT success

---

EDA Script Generation Evaluation Results

<table>
<thead>
<tr>
<th>Tool1</th>
<th>Tool1</th>
<th>Tool1</th>
<th>Tool2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic (Easy)</td>
<td>Automatic (Medium)</td>
<td>Human (Hard with Context)</td>
<td>Human (Hard with Context)</td>
</tr>
</tbody>
</table>

EDA Script Generation Evaluation Results:

- Tool1 Automatic (Easy): 0.33 out of 10 point scale
- EDA script generation: 18% correctness
- Bug summarization: 0.79 out of 7 point scale

0.79 out of 7 point scale
Retrieval-Augmented Generation (RAG)

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

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

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

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

Domain SFT improves performance of ChipNeMo-13B-Chat with/without RAG.
Customization Lead to Large Performance Improvement

Domain-adapted ChipNeMo significantly outperforms OOTB solutions

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

Domain-specific embedding, ranking, and re-ranking models lead to higher context relevance resulting in more downstream customer value.
Unified Stack to Accelerate Generative AI Adoption for Enterprises
Enabling end-2-end generative AI journey from data curation to model customization, optimization, evaluation, inference

**NVIDIA NIM**
Pre-trained & custom LLMs

**NeMo Data Store**
Prompts, responses, PII redaction, quality filtering
Adapters, P-tokens as .nemo checkpoints
Custom datasets, evaluation results

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

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

**NeMo Evaluator**
Automated evaluation of foundation models and fine-tuned LLMs on academic benchmarks and custom datasets using LLM-as-a-judge and pre-defined metrics
Resources to Get Started

- Explore NVIDIA API Catalog: [https://ai.nvidia.com/](https://ai.nvidia.com/)
- NVIDIA RAG:
  - [https://build.nvidia.com/explore/retrieval](https://build.nvidia.com/explore/retrieval)
  - [https://github.com/NVIDIA/GenerativeAIExamples](https://github.com/NVIDIA/GenerativeAIExamples)
- NeMo Microservices:
Get Started with NeMo

Download Now - Language

Apply Now - Multimodal

Web Pages

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

Blogs

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

Webinars

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