



Large Language Model (LLM) Technical Workshop, Introduction

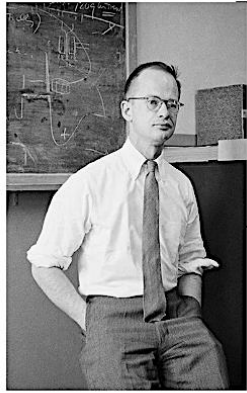
26th May 2023

Ettikan Kandasamy Karupiah (Ph.D)
Director/Technologist, Asia Pacific South Region

WORKSHOP AGENDA

TIME / DURATION	TOPIC
5min	Sponsored Networking Lunch
15min	Openings
40min	Title: Introduction, Demystifying LLM and Data Curation
20min	Break
40min	Title: LLM Training and Inference at Scale. Customized LLM with Prompt-Learning.
20min	NVIDIA LLM Service Demo

NUTSHELL DEEP LEARNING STORY



Computer model based on neural networks of human brain: 'Threshold Logic'

Walter Pitts & Warren McCulloch

1943

First Convolutional Neural Network: 'Neocognitron'



Kunihiko Fukushima

1965

1979

The 80s

Start of the games from TD-Gammon & IBM's Deep Blue to IBM's Watson



The 90s & Beyond

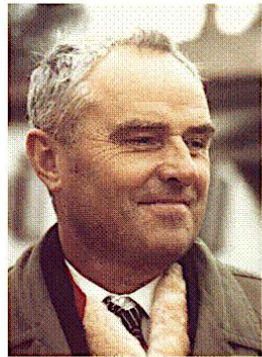
1999

2006

Deep Learning gets a name from Geoffrey Hinton



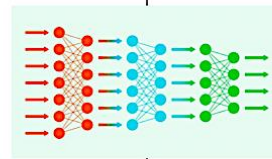
Geoffrey Hinton, Godfather of Deep Learning



О.Г. Ивахненко (1967 г.)

Alexey Ivakhnenko & V.G. Lapa

Models of complicated equations that were statistically analyzed



Recurrent NN, Backpropagation Technique, Handwritten digit recognition



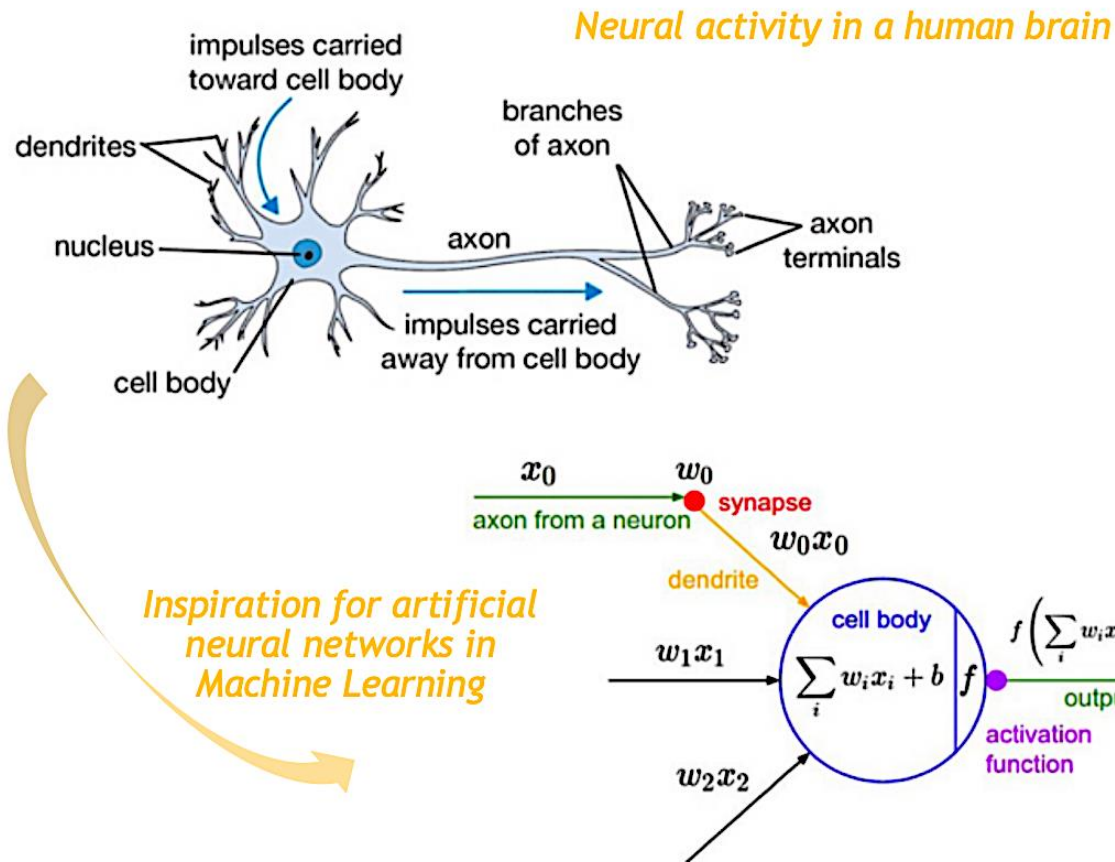
GPUs were developed & data processing capabilities exploded

World's First GPU

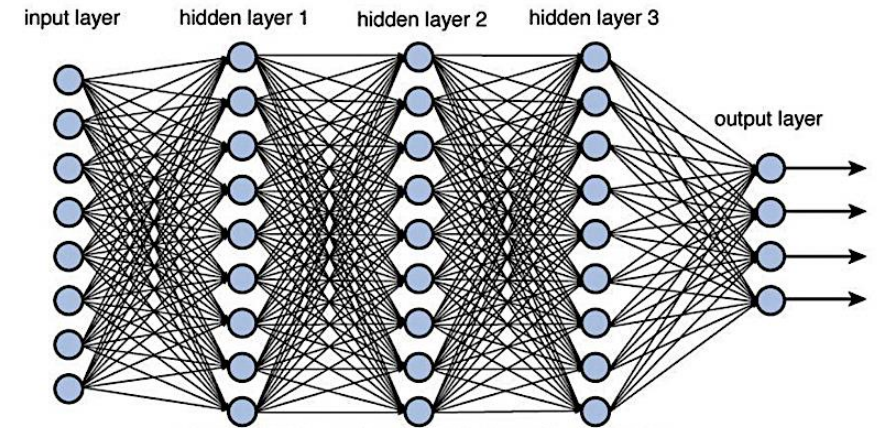


ARTIFICIAL NEURAL NETWORKS (ANN)

How Deep Learning Mimics Brain Activity



Deep Network Architecture with Several Layers



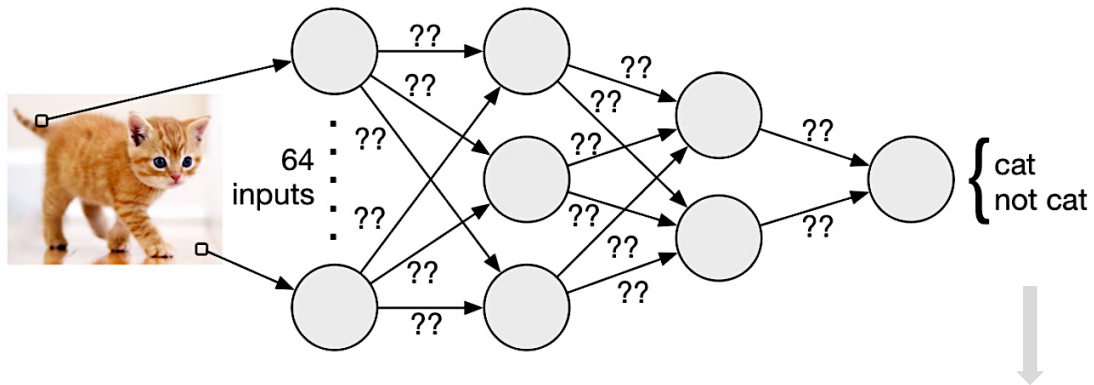
"I have always been convinced that the only way to get Artificial Intelligence to work is to do the computation in a way similar to the human brain. That is the goal I have been pursuing. We are making progress, though we still have lots to learn about how the brain actually works."

Geoffrey Hinton, Godfather of Deep Learning

TRAINED ARTIFICIAL NEURAL NETWORKS

Inferring incoming data

Training an ANN with lots of data



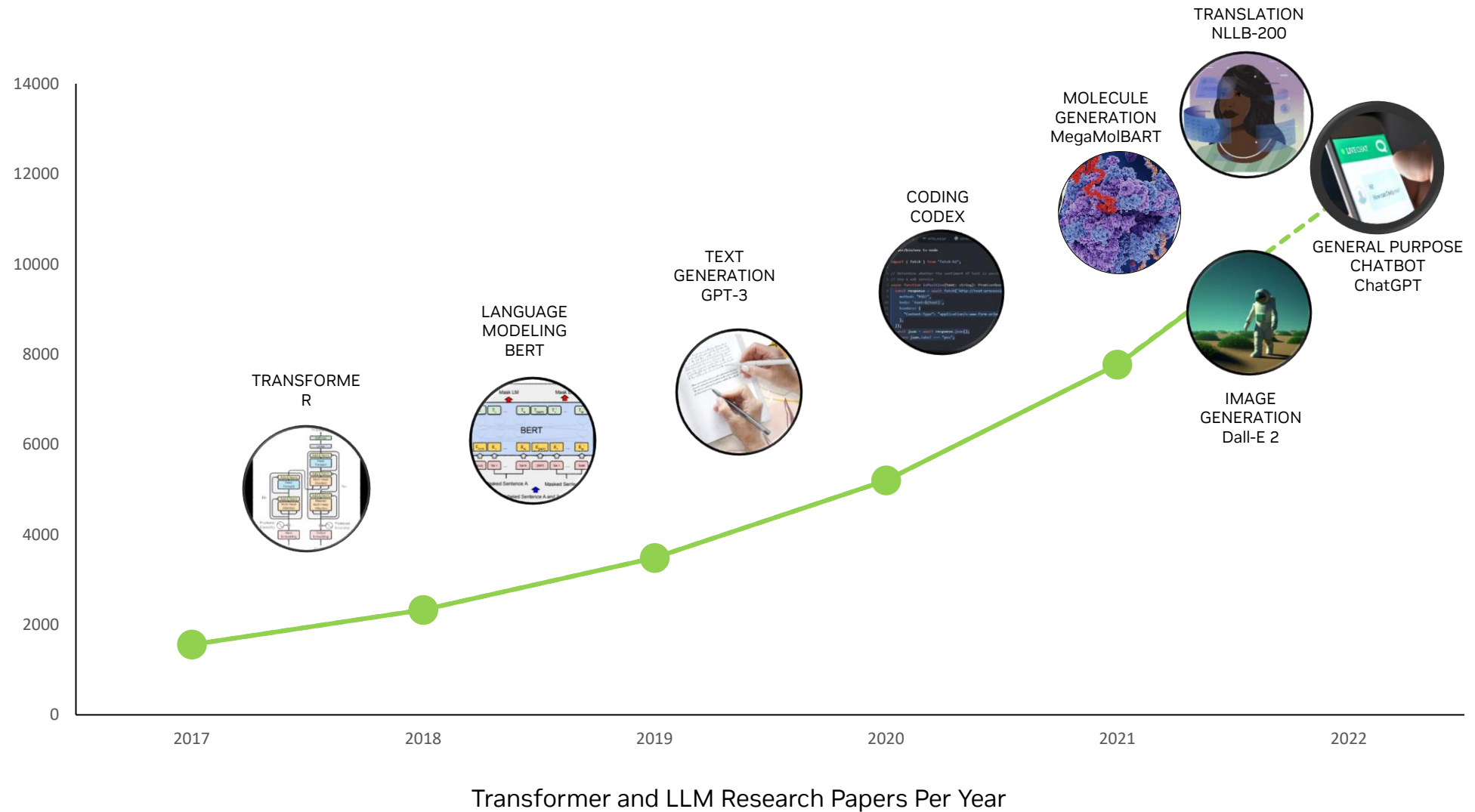
Trained ANN predicting from new incoming data

Characteristics of a trained ANN:

- Becomes an expert in an area
- Makes its own decisions and judgements
- Has hard-coded knowledge that is permanent
- Produces repeatable and accurate results

*“Early AI was mainly based on logic. You're trying to make computers that reason like people.
The second route is from biology: You're trying to make computers that can perceive and act and adapt like animals.”*

Geoffrey Hinton

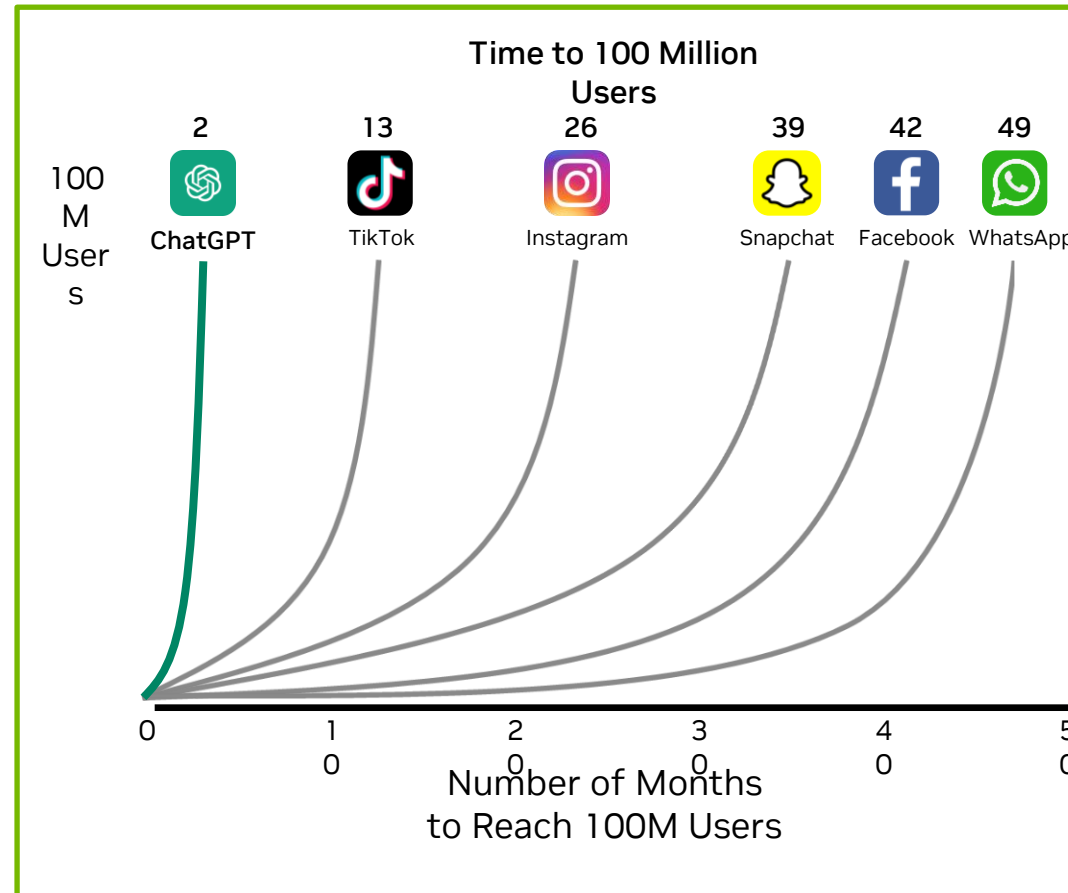


Generative AI From Research to Production in 5 Years

Few of the most significant milestones in LLMs shaping industries

Massive AI Models Drive New Use Cases

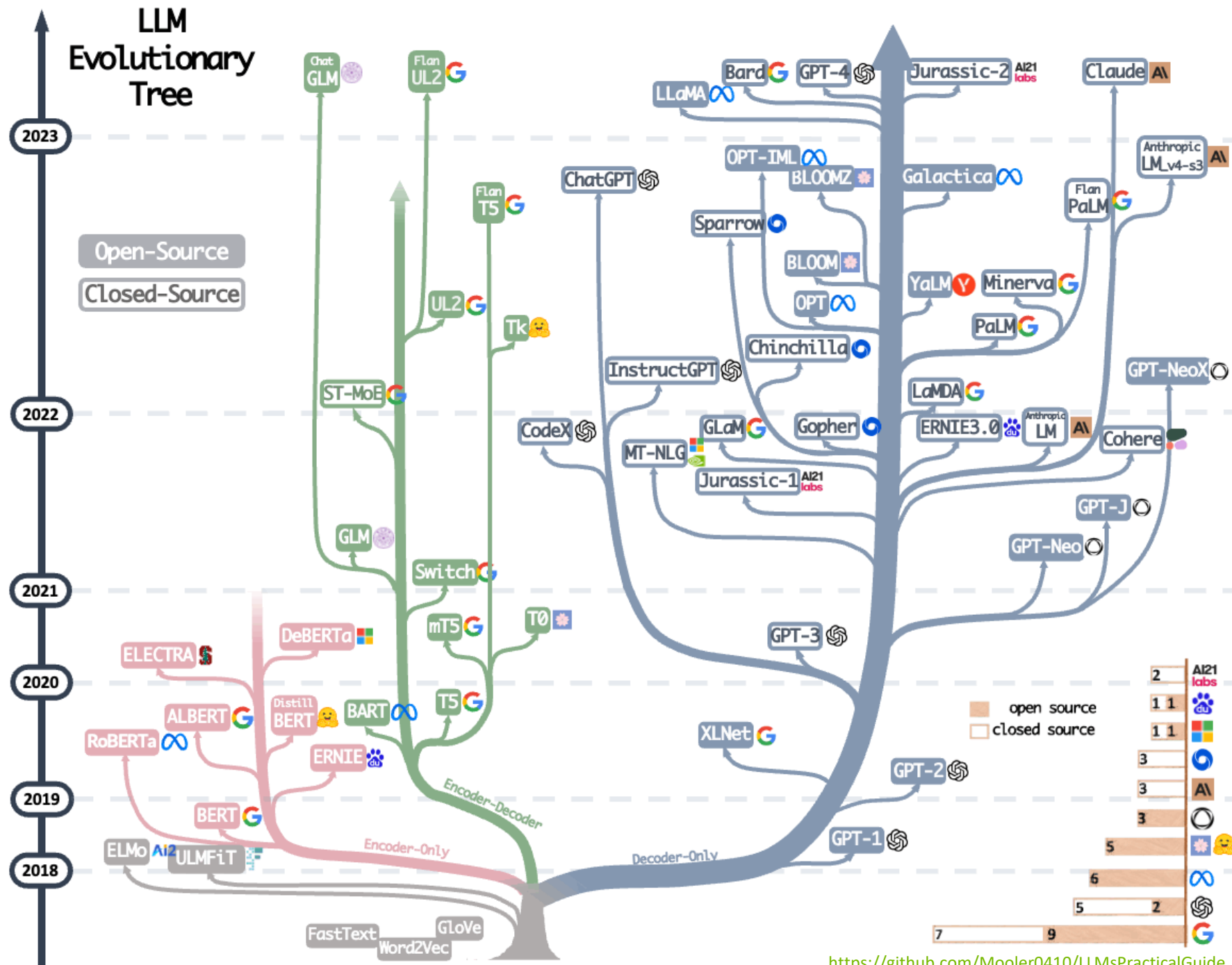
LLMs and GenAI Driving an Inflection Point



Animating 2D Avatar and Portraits

Making them sing!









<https://github.com/Mooler0410/LLMsPracticalGuide>

Challenges Building Generative AI for the Enterprise





LLM Enterprise Use Cases and Goals:

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

Challenges of Building Foundation Models

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

Challenges of Using Foundation Models

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

Consider Which Path to Take in LLM Adoption

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

Personalization / Customization

Learn Knowledge



Include Task Specific Knowledge



Include Proprietary and Topical Knowledge Base



Continuously Improve Models Over-Time



Methods & Techniques

Foundation Model Training / Fine-Tuning

Incremental Knowledge - Prompt Learning Techniques (p-tuning, adapters)

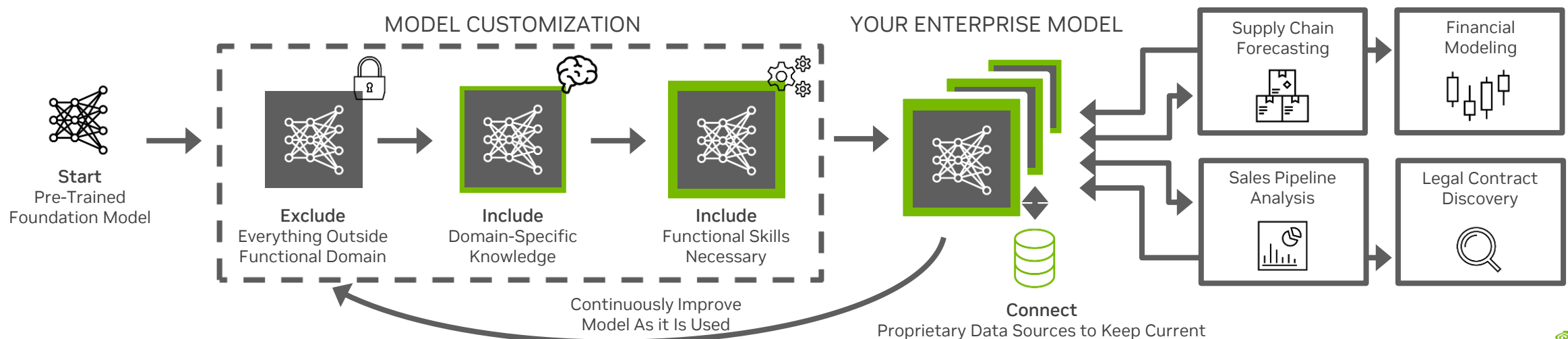
Runtime Knowledge - Information Retrieval Models

Continuous Knowledge - Reinforcement Learning with Human Feedback

NVIDIA Provides the Tools to Overcome LLM Challenges

Using Foundation Models

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



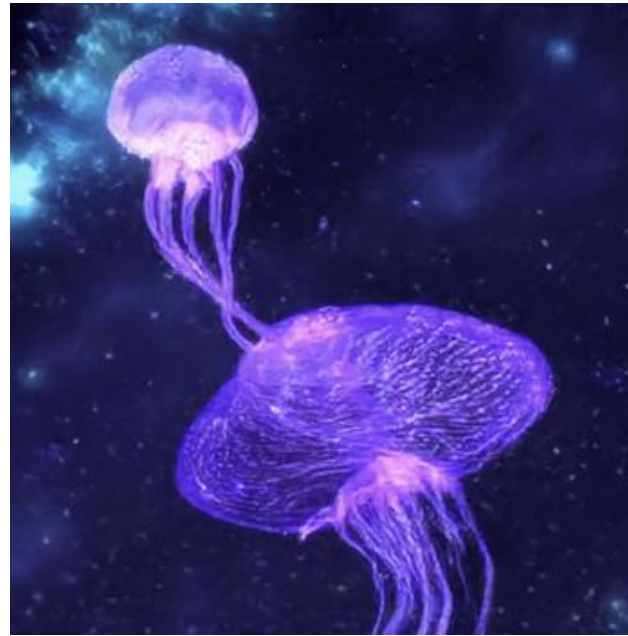
NVIDIA Picasso

Cloud Service For AI-Powered Image, Video & 3D Applications



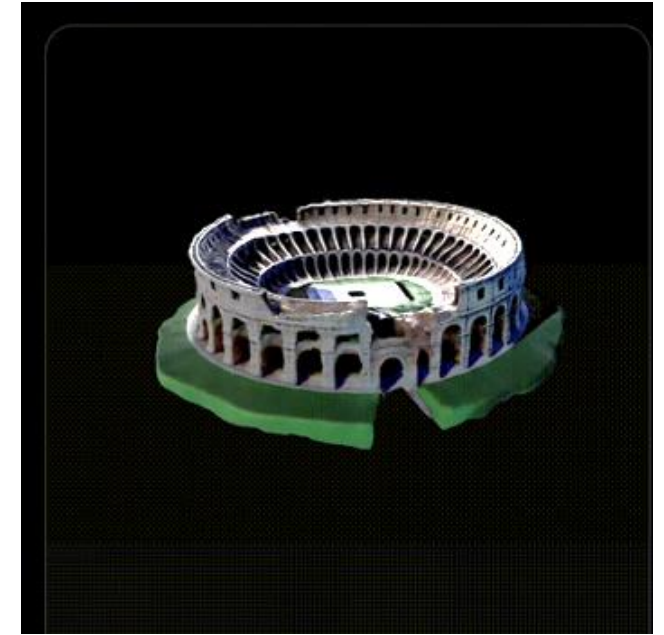
Text-to-Image

A photo of a cute cat with lots of Holi colors



Text-to-Video

Purple bioluminescent jellyfish swimming in space



Text-to-3D

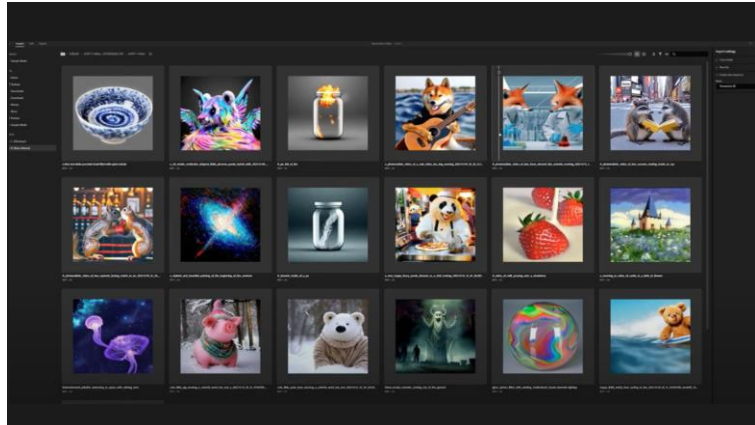
A DSLR photo of a 3D model of the colosseum



NVIDIA DGX Cloud

NVIDIA Picasso

Add Generative AI To Your Application



Add Generative AI To Your Design Process



Optimize Your Model For Inference



Leverage NVIDIA Omniverse Ecosystem



Sign Up To Get Notified Of Updates & Availability

www.nvidia.com/picasso

KEY TAKEAWAYS FOR TODAY'S SESSION

Session 1 : Demystifying LLM and Data Curation

1. Introduction and Demystifying LLM
2. Introduction to NVIDIA NEMO Framework Toolkit
3. Data Curation for Fine-Tuning and P-Tuning

Session 2: LLM Training and Inference at Scale. Customized LLM with Prompt-Learning

1. Basics of Training Foundation Models at Scale (Pre-Training)
2. Prompt-Learning Techniques for Pre-Trained Models
3. Inferencing LLMs at Scale



3-DAYS HANDS-ON BOOTCAMP (BY-INVITATION)

SAVE THE DATE: 11TH TO 13TH JULY

Prerequisites: Developers. Experience with Python. Ideally Well Versed with NLP Domain. No GPU Programming Knowledge is Required.

Duration: 3 Days, 6 Hours Daily

DAY 1

- Introduction to Q&A Models and Architectures
- Pre-Processing Raw Text Data
- Q&A Dataset Generation
- GPT Tokenizer

Day 2

- Model Optimization with TensorRT
- Model Deployment Pipeline
- Introduction to Nemo Framework with Lab

Day 3

- Prompt-Tuning Lab with Nemo Framework