ZERTIFIZIERTE KI: Workshop Foundation Models

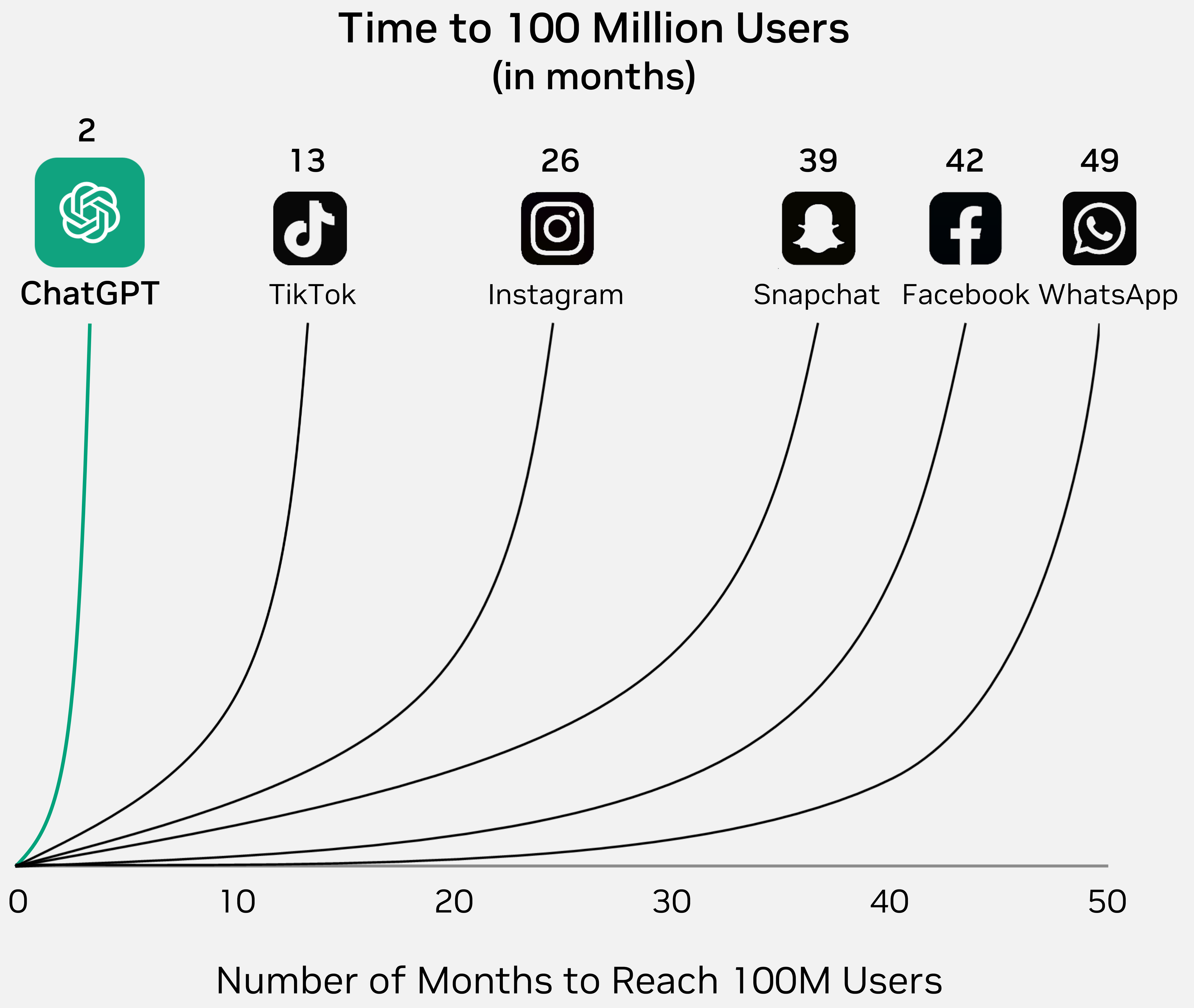
Dr. Jochen Papenbrock, Head of Financial Technology EMEA

June 2023

jpapenbrock@nvidia.com

ChatGPT — 2022 "The AI Heard Around the World"

From AlexNet to ChatGPT in 10 years

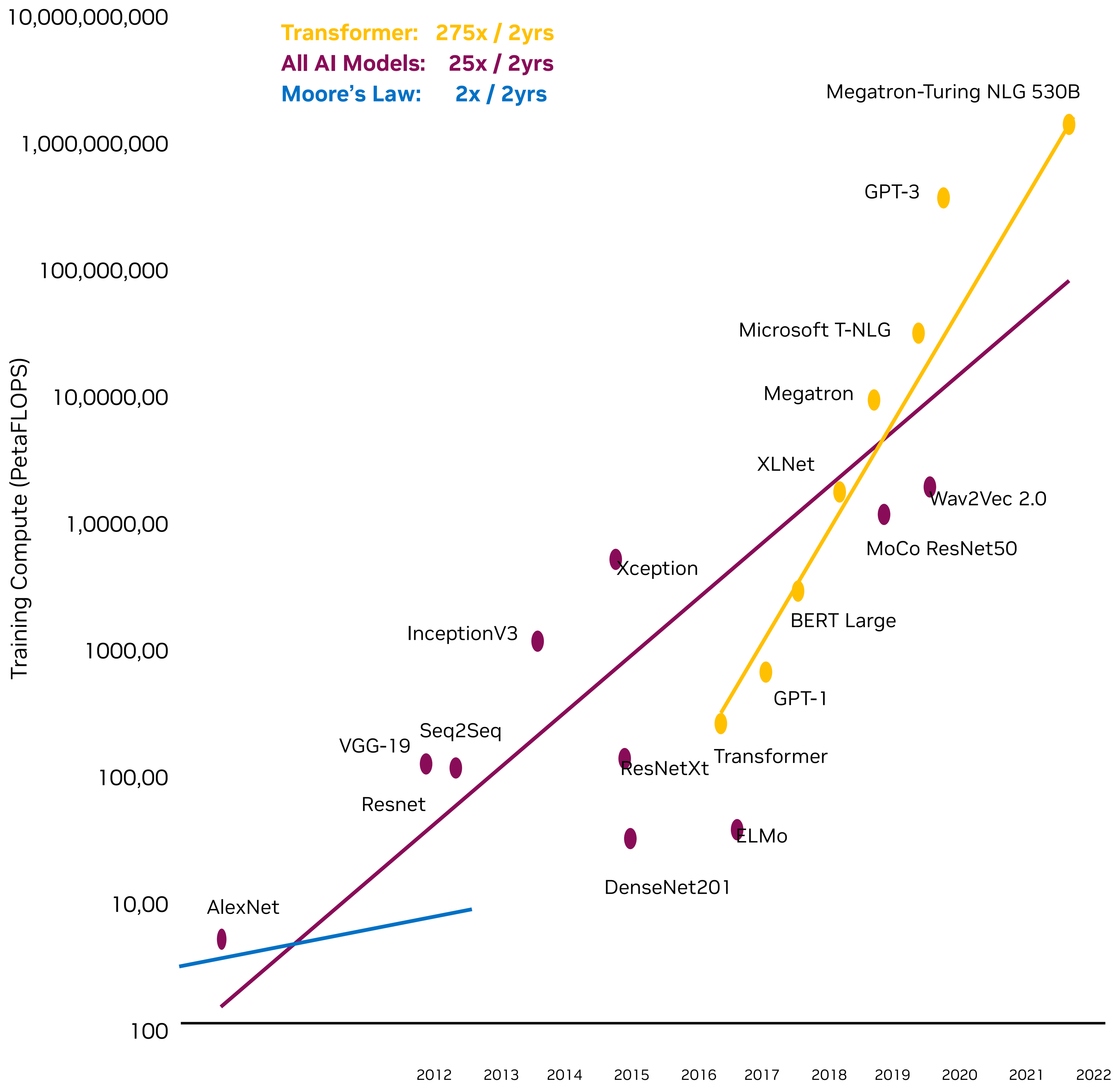


Massive AI Models Drive New Use Cases

LLMs and Gen AI Driving an Inflection Point

Moore's Law is Ending

AI has an enormous appetite for compute — we are only seeing the tip of the iceberg!

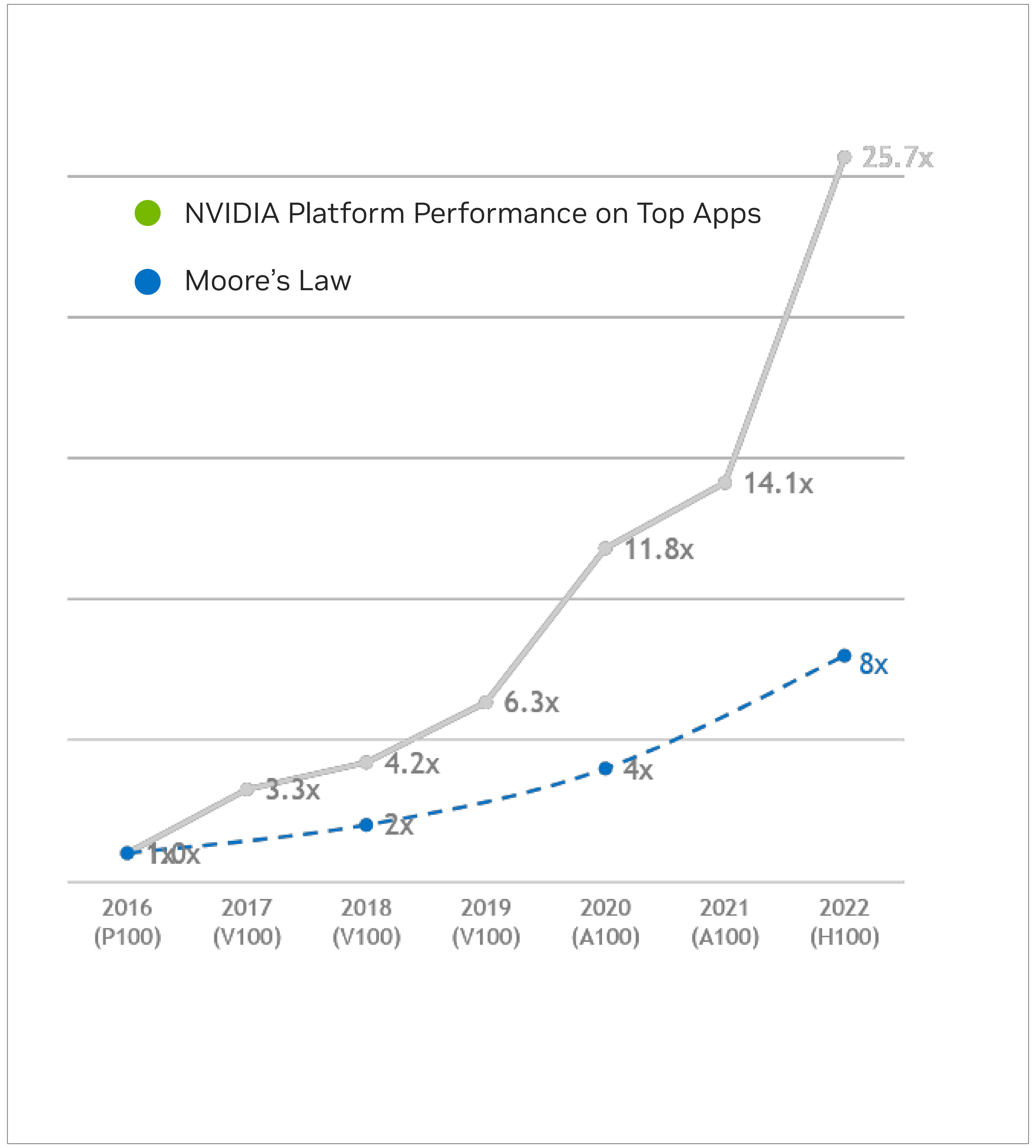


Models Growing Exponentially

Forecasted Share of Energy Usage	2%
Share of Global Energy Usage	5% by 2030
Data Center through to Electricity Usage	>200 TWh/year

Data Centers are Power Limited

Need to Become More Efficient



26x Performance In 6 Years

NVIDIA

Accelerated Compute — Essential for AI

"The shift to accelerated compute away from conventional CPU is structural"
— Vivek Arya, Bank of America, Analyst

Top500
Platform of Choice

71%
Use NVIDIA

MLPerf HPC & Training
AI Transforming HPC

#1
Leadership Perf

Accelerated Network
Driving Modern Workloads

#1
Most adopted in Top 500

Cloud HPC
Democratizing Access

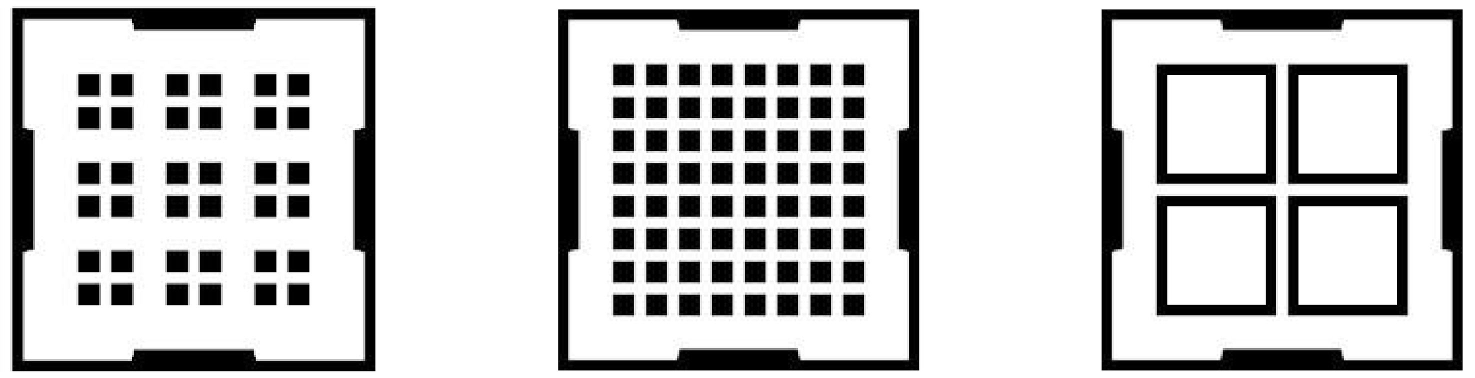
100%
8 instances in Top 50

Powering World's Fastest Supercomputers

APPLICATION

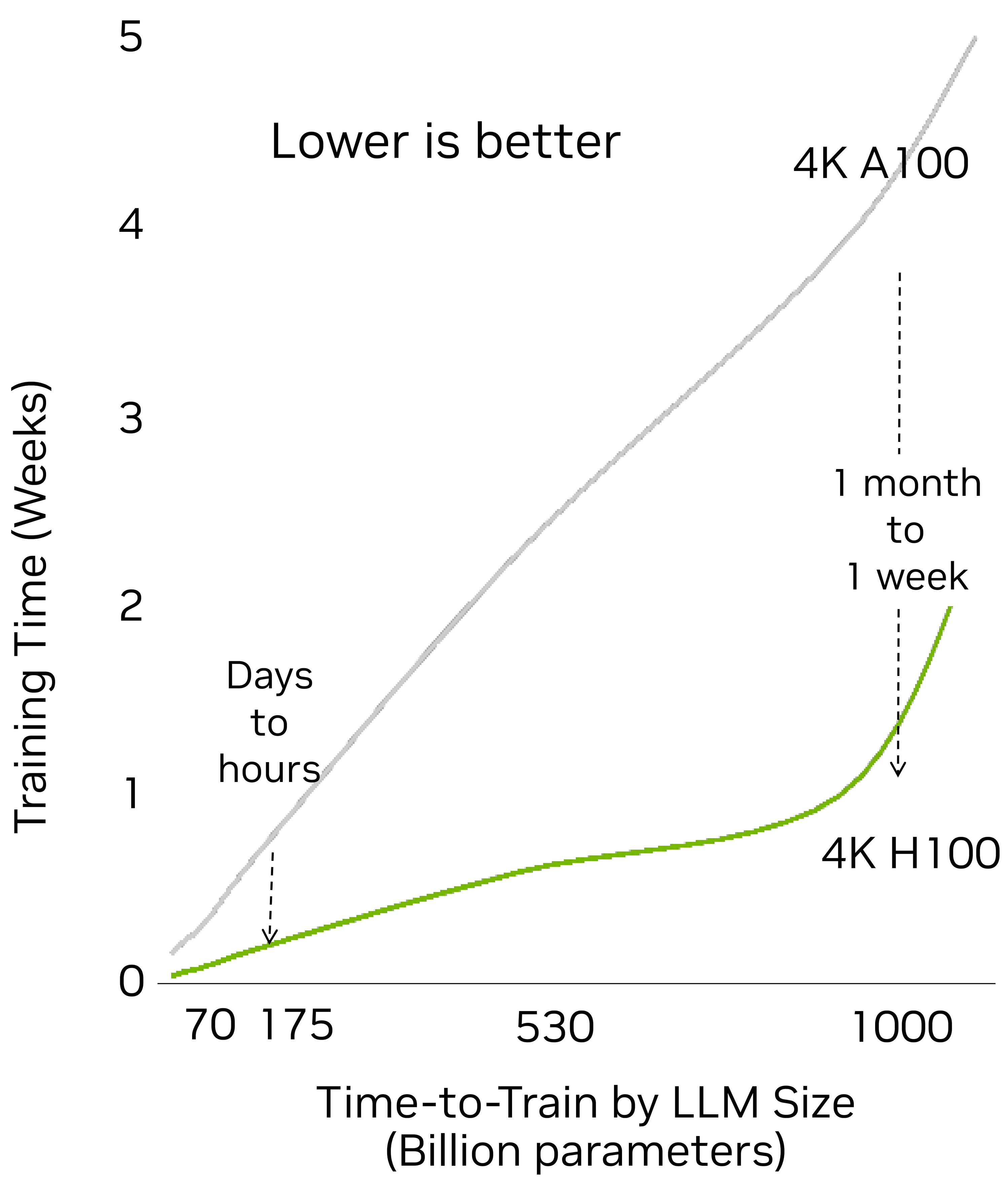
ACCELERATION LIBRARY

SYSTEM



DPU GPU CPU

Acceleration Takes a Full Stack



H100 Supercharges AI

Jensen Huang, Co-founder, President and CEO, NVIDIA – GTC 2023 Keynote

"

In 2012, Alex Kerchevsky, Ilya Suskever, and Geoff Hinton needed an insanely fast computer to train the AlexNet computer vision model.

The researchers trained AlexNet with 14 million images on nVidia GeForce GTX 580, processing 262 quadrillion floating-point operations, and the trained model won the ImageNet challenge by a wide margin and ignited the Big Bang of AI.

A decade later, the transformer model was invented.

And Ilya, now at OpenAI, trained the GPT-3 large language model to predict the next word.

323 sextillion floating-point operations were required to train GPT-3.

One million times more floating-point operations than to train AlexNet.

The result this time – ChatGPT, the AI heard around the world.

"

Potential of AI/GenAI

Generative AI Unlocks New Opportunities

How has NVIDIA contributed to acceleration of AI?

NVIDIA has been a pioneer in the field of AI since the very beginning. Our GPU platform has enabled the rapid development of AI – from the training of neural networks, to inference in the data center, on-device AI in the car and in the cloud, and the deployment of AI to tackle challenging problems like conversational AI and translation.

NVIDIA's GPU-accelerated computing platform is the engine of AI – it is the most important computing platform of our time.

***Generated using NVIDIA NeMo service*

530B

TEXT GENERATION

Summarization

GPT-3

Marketing Copy

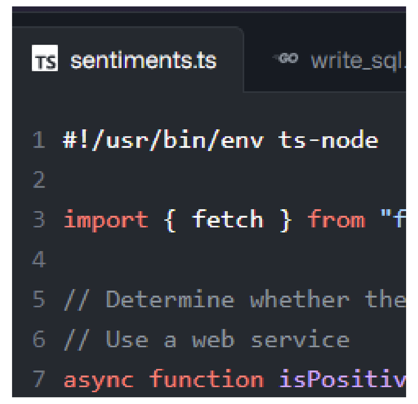
TRANSLATION

Translating Wikipedia

NLLB-200

Real-Time Metaverse Translation

CODING



Dynamic Code Commenting

CODEX

Function Generation

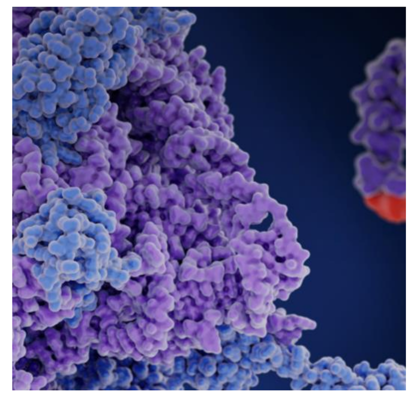
IMAGE GENERATION

Brand Creation

e-Diffi

Gaming Characters

LIFE SCIENCE

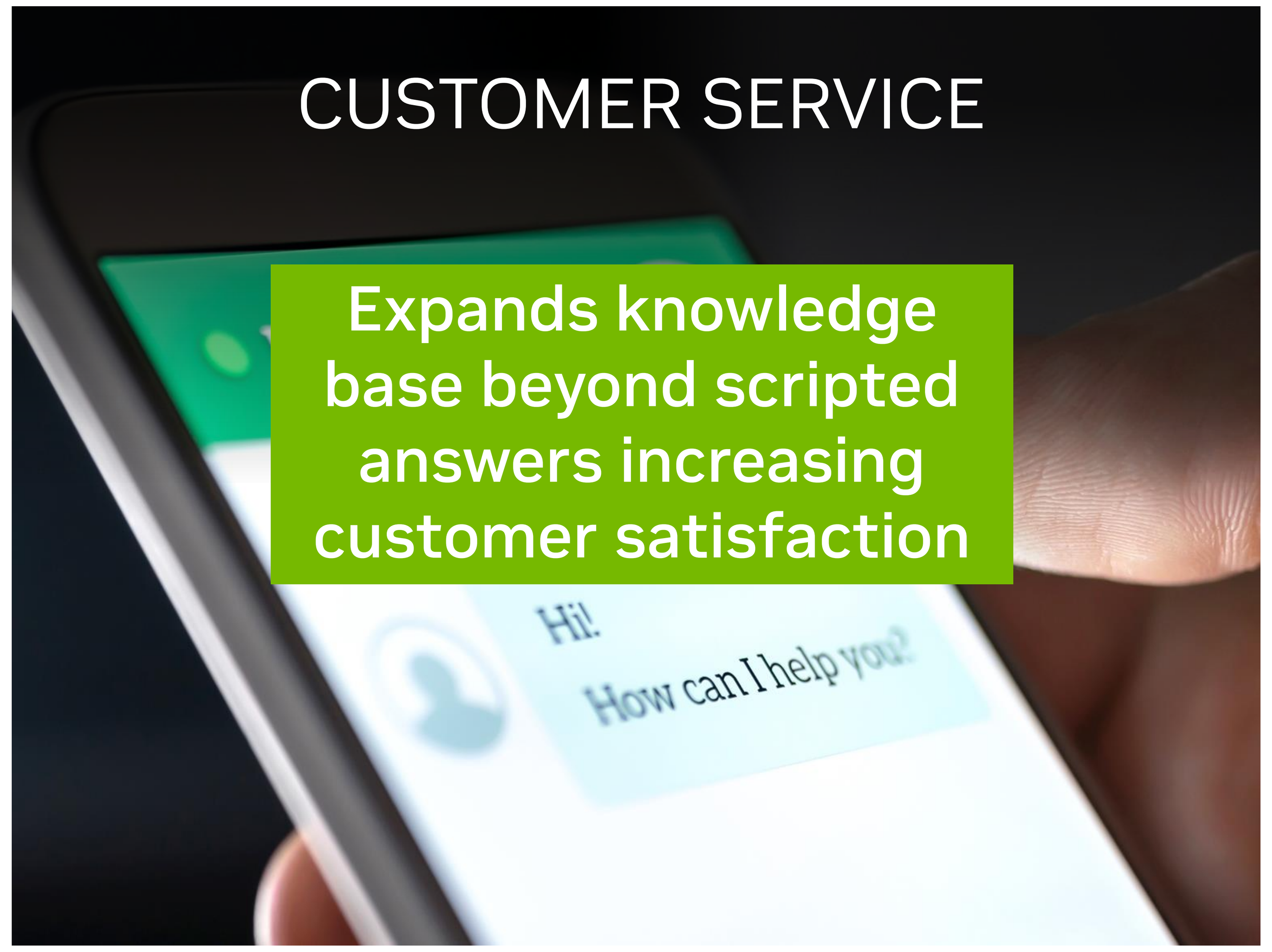


Molecular Representations

MegaMolBART

Drug Discovery

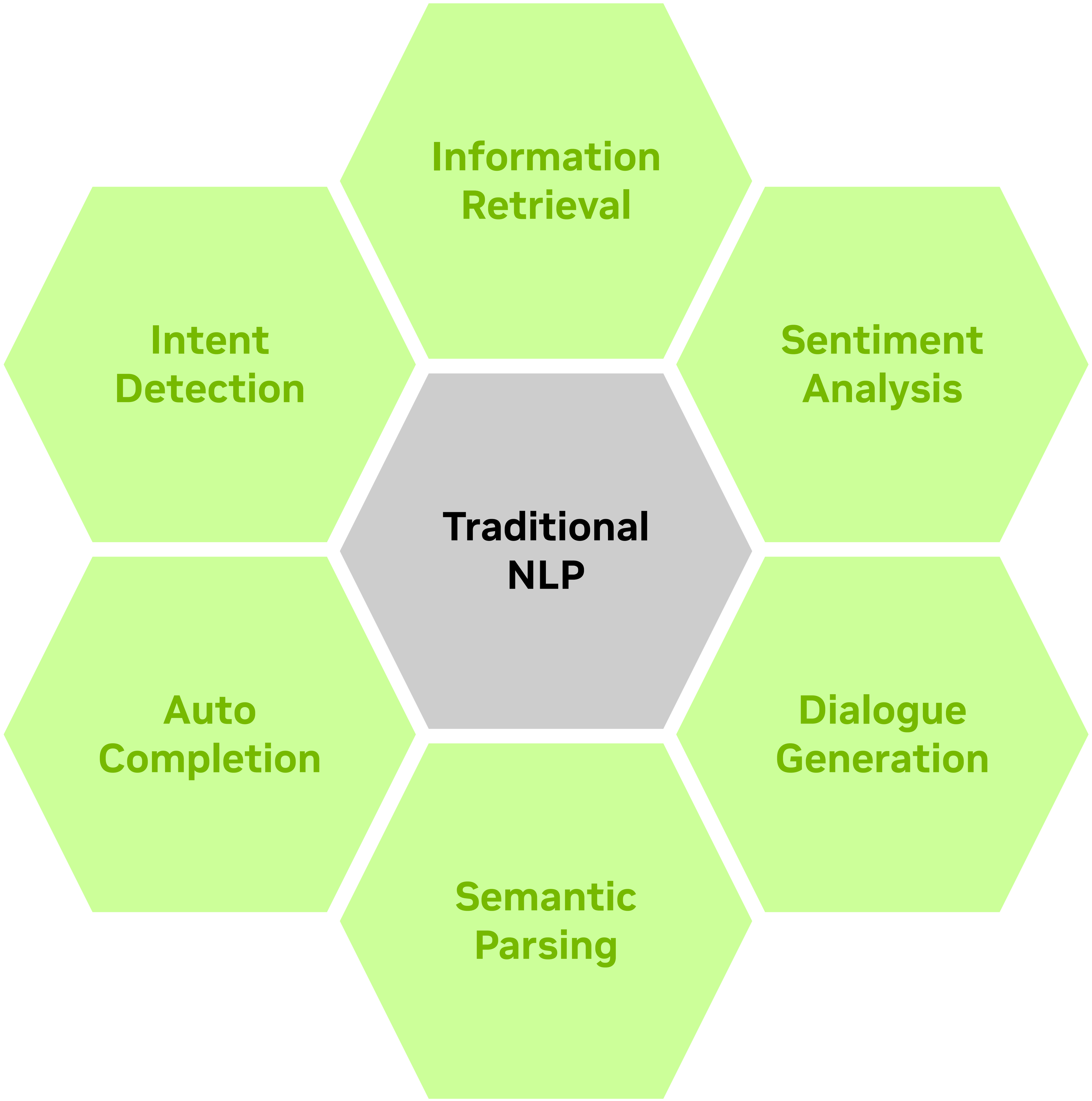
Generative AI Impacts Every Function in a Bank



NVIDIA

Benefits and Challenges of LLMs in production

Traditional NLP Use-Cases



Current State of Traditional NLP

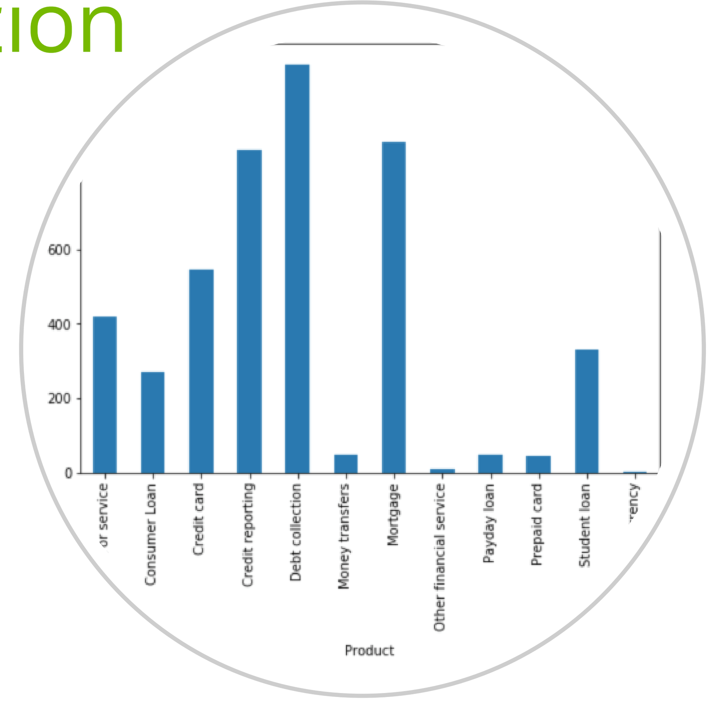
Challenges of Traditional NLP

- Need to build models for domain specific use-cases
- Requires extensive data inputs and data labeling
  - Intents and Entities
- Frequent (re)training for out of vocabulary use cases, and for new data. Model drifts over time
- Challenges in deciphering meaning of complex/dual inputs

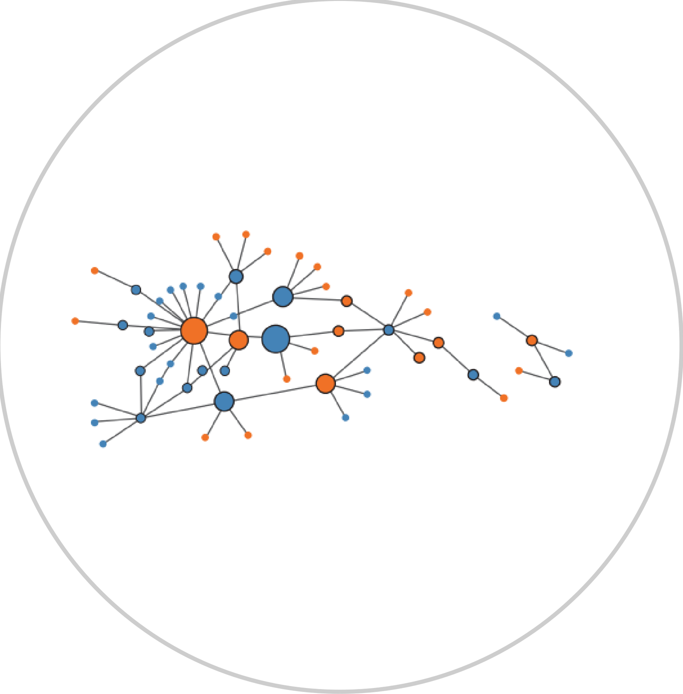
	Traditional NLP Approach	Large Language Models
Requires labelled data	Yes	No
Parameters	100s of millions	Billions to trillions
Desired model capability	Specific (one model per task)	General (model can do many tasks, also new ones)
Training frequency	Retrain frequently with task-specific training data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

One to rule them all

Domain Classification



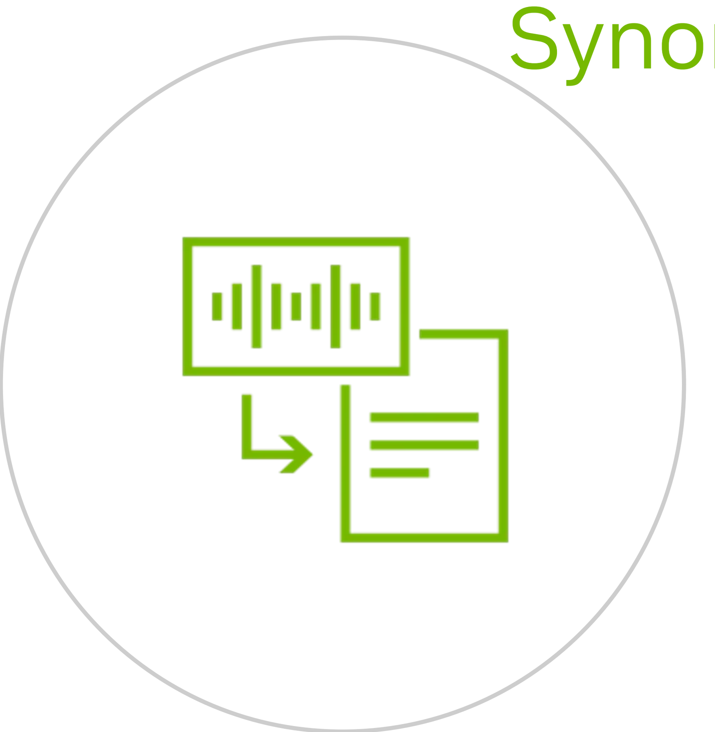
Relation extraction



Intent Tagging

Topic Modeling

Synonym and Semantic Search



NLP, ASR, TTS

Sentiment Analysis

Entity Recognition

TL; DR

Text Summarization

Question Answering

Translation

LLMs are more efficient

They can leverage prompts as in-context information | Prompt Engineering and Prompt Tuning approaches

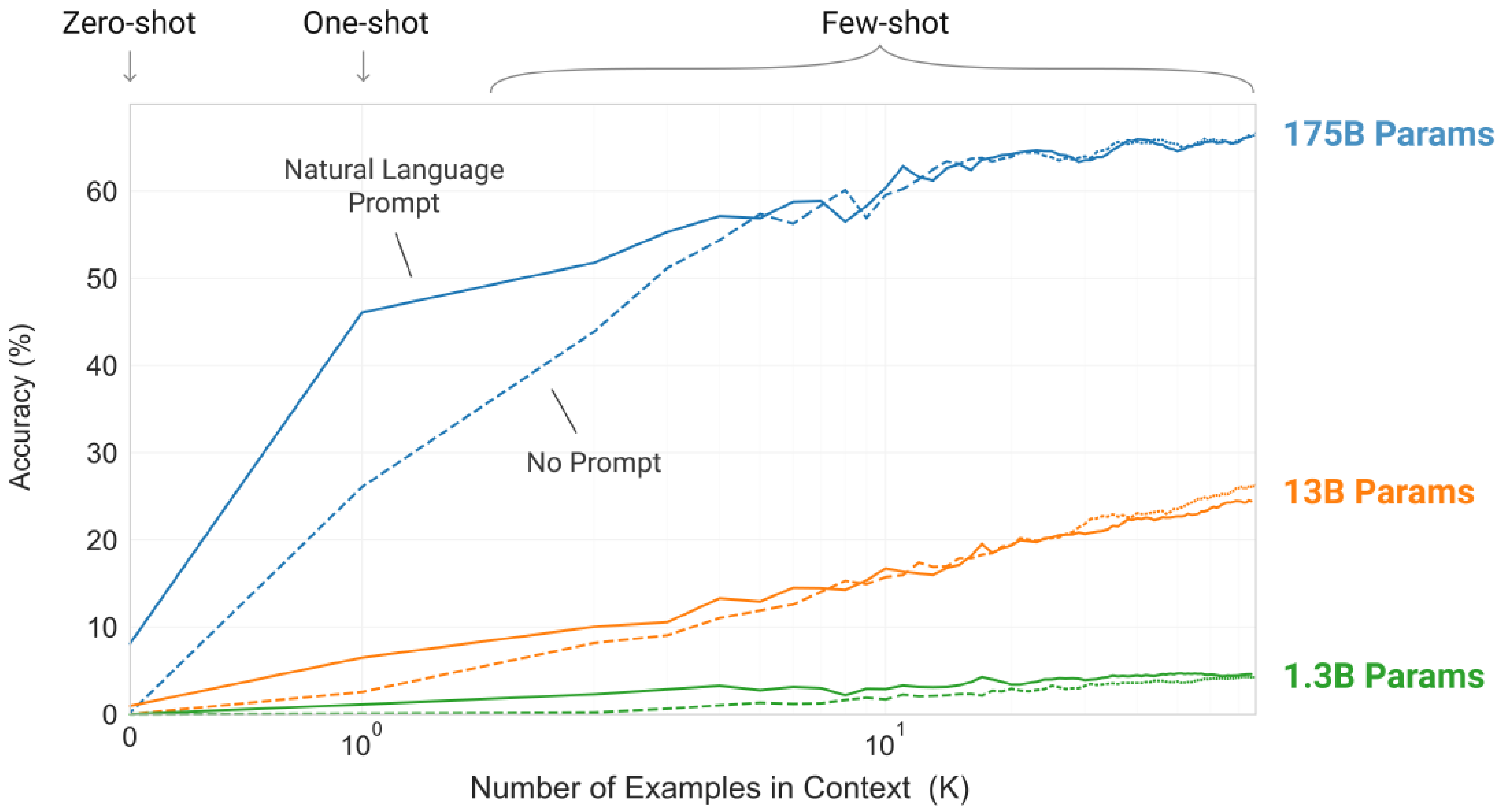


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Source: T. Brown et al (2020), "Language Models are Few-Shot Learners", arXiv:2005.14165

Enterprise Challenges Of Developing Generative AI

Challenges of Building Foundation Models

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

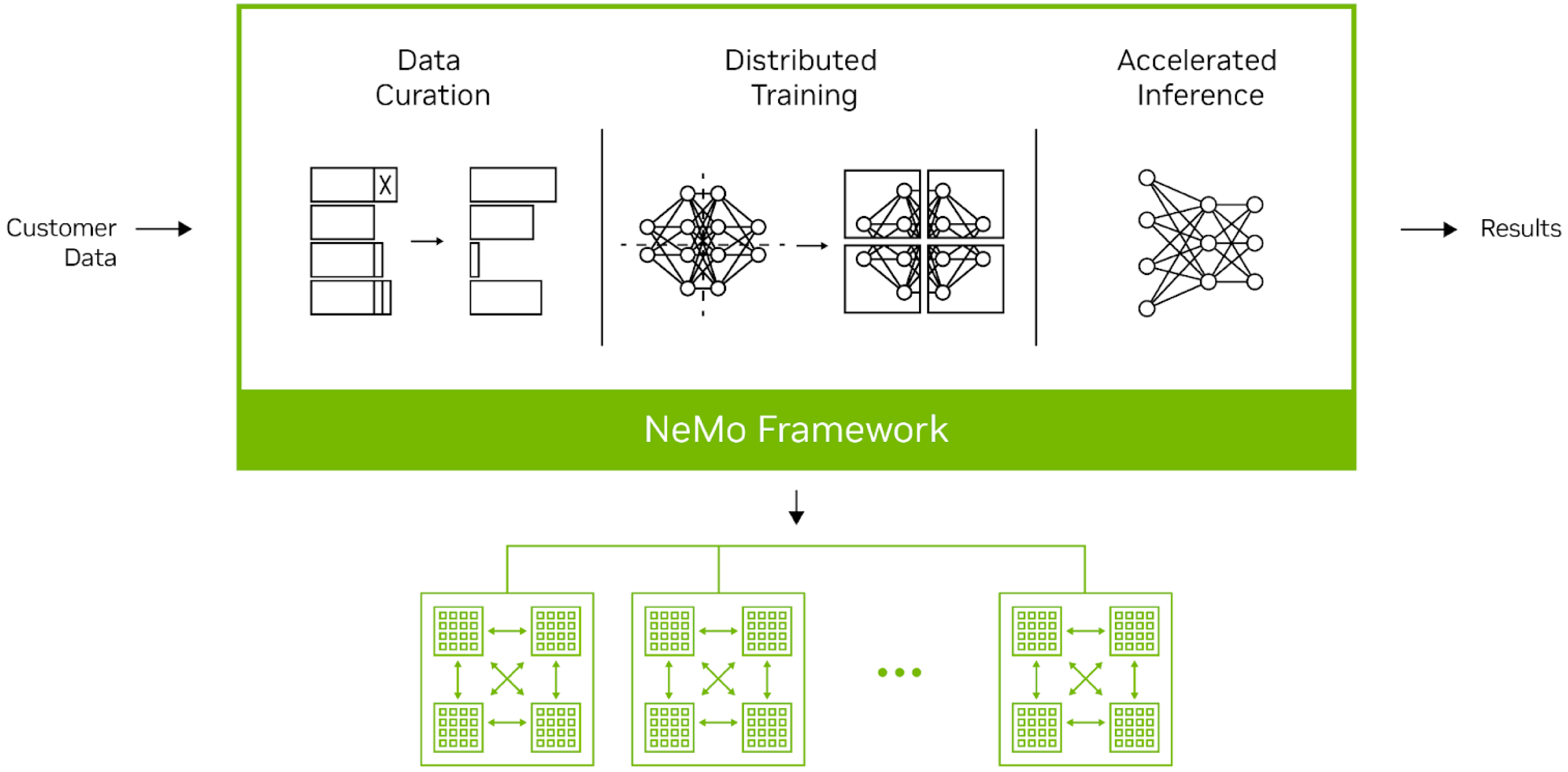
Challenges of Using Foundation Models

- Lack domain / enterprise specific knowledge
- Frozen in Time
- Hallucinate and provide undesired information
- Bias & Toxic Information

Enterprise Scaling LLMs with Trust

NeMo Framework

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



Microsoft Azure

aws

ORACLE

- open-source framework LLMs with billions and trillions of parameters
- hyper-personalization and at-scale deployment (e.g. Inference)
- simplifying and accelerating the path to build and deploy LLMs
- guardrails: reduce bias and toxicity, to align to human intentions
- finding optimal hyperparameters, convergence of models
- retrieval augmented models based on-prem data

Deploy anywhere

DGX SuperPODs

DGX Cloud

DGX Systems

Multi-modality support

Build language, image, generative AI models

Accelerated Workflow

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

Data Curation

Mine and curate high-quality training data @ scale

Customize Foundation Models

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

Support

NVIDIA AI Enterprise keep projects on track

Deploy Anywhere

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

Overcoming Challenges Of Using Foundation Models

Generalized AI Will Not Work; Enterprises Need Their Own AI

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

Customization is Required to Address Business-specific Tasks

Zero-Shot Response

"What is the yellow part in an egg?"

LLM

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

P-Tuned Response

"What is the yellow part in an egg?

+

Trained Prompts (Context)

LLM

Nutrition Chatbot

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

Prenatal Chatbot

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

Culinary Chatbot

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

Enterprises Require Responses Based on Current Information

Foundation/Custom Model Response

"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 "*

Information Retrieval 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 new opportunities for greater intelligence

Less Frequent Re-Training

Significant cost and time savings to maintain LLMs

NeMo Generative Foundation Models

Suite Of Generative Foundation Language Models Built For Enterprise Hyper-personalization

Fastest Responses

GPT-8

GPT-8B w/ 1.1T tokens. SFT w/ FLAN. I/O: 4K tokens

Optimal balance of accuracy - latency

GPT-43

GPT-43B w/ 1.1T tokens. SFT w/FLAN. 50 Languages. I/O: 4K tokens

For Complex Tasks

GPT-530

GPT-530B w/ 340B tokens. SFT w/FLAN. I/O: 2K tokens

Information Retrieval

Inform

Community-built model

BigScience

mT0-XXL

mT0-XXL 13B w/ 340B tokens. 101 Languages. I/O: 2K tokens. Encoder-only - T5 model

All models available through NeMo service

Hyper-personalizing Foundation Models for Enterprises

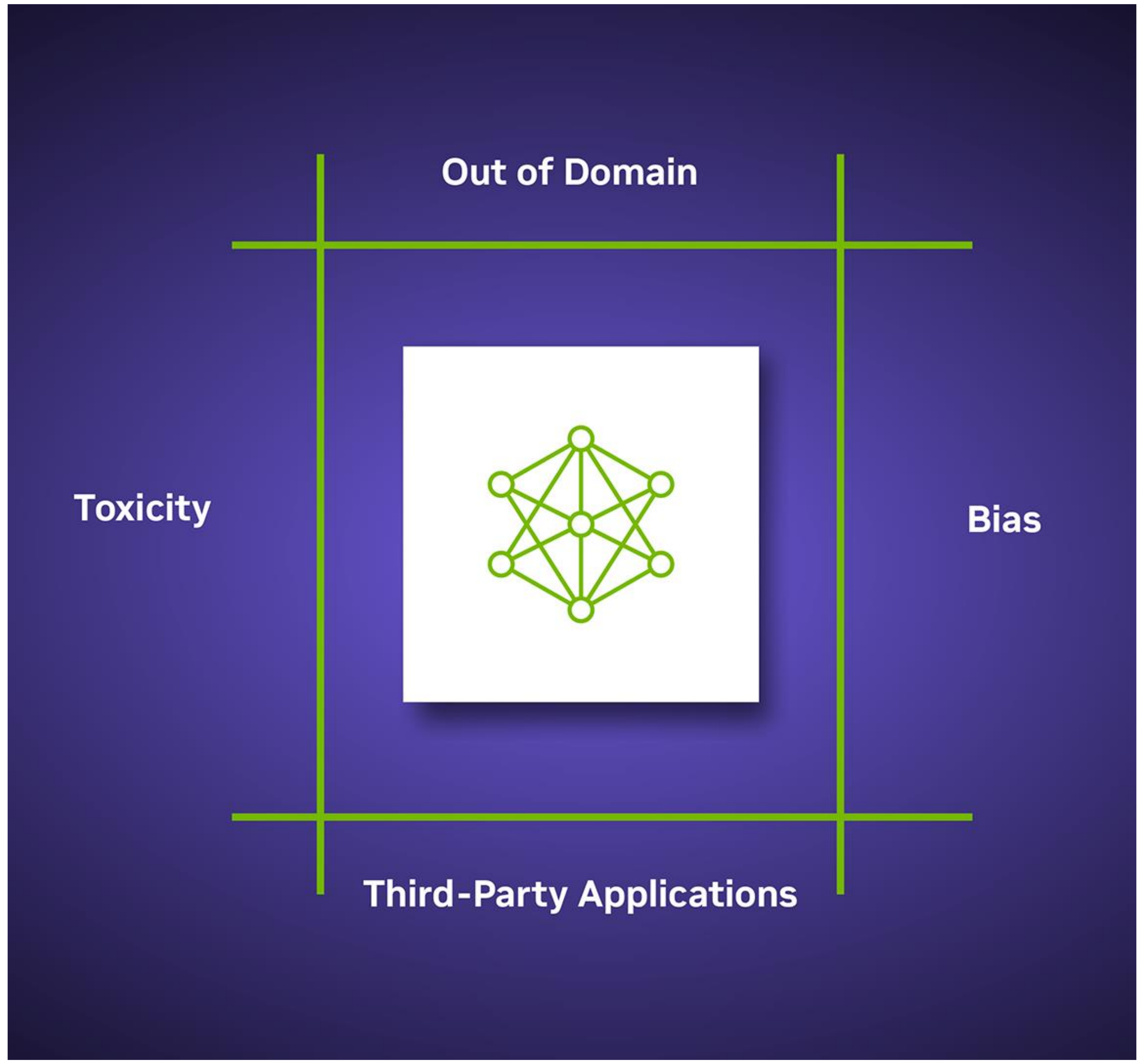
Methods To Build And Hyper-personalize Foundation Models For Specific Use-cases

Personalization / Customization		Methods & Techniques
Learn New Knowledge (New domain)	→	Foundation Model Training / Fine-Tuning – *Via NeMo framework*
Learn a skill (ex. Article summarization)	→	Incremental Knowledge - Prompt Learning Techniques (p-tuning, adapters)
Filter Bias & Inappropriate Content	→	Toxicity Classifiers to indicate toxicity score for both Inputs & Outputs
Include Proprietary and Topical Knowledge Base	→	Runtime Knowledge - Inform (Information Retrieval Models)
Learn to Implement Guardrails	→	Tune Model Parameters to Stay within Specified Domain - Supervised Fine Tuning
Continuously Improve Models Over-Time	→	Continuous Knowledge – Reinforcement Learning with Human Feedback

Legend: Available Today | On the Roadmap

Enterprise Use-Cases Require Guardrails

Exclude everything outside functional domain, eliminate bias and toxicity, to align to Enterprise goals



- **Topical guardrails** prevent apps from veering off into undesired areas. For example, they keep customer service assistants from answering questions about the weather.
- **Safety guardrails** ensure apps respond with accurate, appropriate information. They can filter out unwanted language and enforce that references are made only to credible sources.
- **Security guardrails** restrict apps to making connections only to external third-party applications known to be safe.

- Programmable rails for LLMs: Steering the LLM towards producing outputs that accurately and effectively meet user intent
- Align the LLMs with the business goals of the Enterprise
- Prevent the model from generating undesirable, biased, or harmful content
- Toxicity classifier (BERT based classifier) assigns a toxicity score for every input and output
- Developer can use the toxicity score to filter inappropriate responses for their use-case

Designed for Enterprise Adoption

Trustworthy and Responsible AI Development

Privacy

- Training data not shared
- Control logging of prompts & outputs

Safety & Security

- Toxicity Classification
- Deploy Anywhere using NeMo Framework
- Soc-2 compliance
- Implement guardrails

Transparency & Explainability

- Foundation models trained on licensed data
- Connect to proprietary knowledge base
- Cite sources for model answers

Nondiscrimination

- Bias classification

Legend: Available Today | On the Roadmap

NVIDIA

AVATARS & DIGITAL HUMANS USE MANY AI/ML MODELS

[BOT MAKER - EARLY ACCESS PROGRAM | NVIDIA DEVELOPER]

[OMNIVERSE AVATAR CLOUD ENGINE (ACE) | NVIDIA DEVELOPER]

[SOFTSERVE EXAMPLE]

HOW CAN I HELP?

Vision

Conversation

Speech

Planning & Action

Face Animation

Gesture

Realistic Graphics

Body Animation

Accelerated Computing for AI Quality/Certification projects

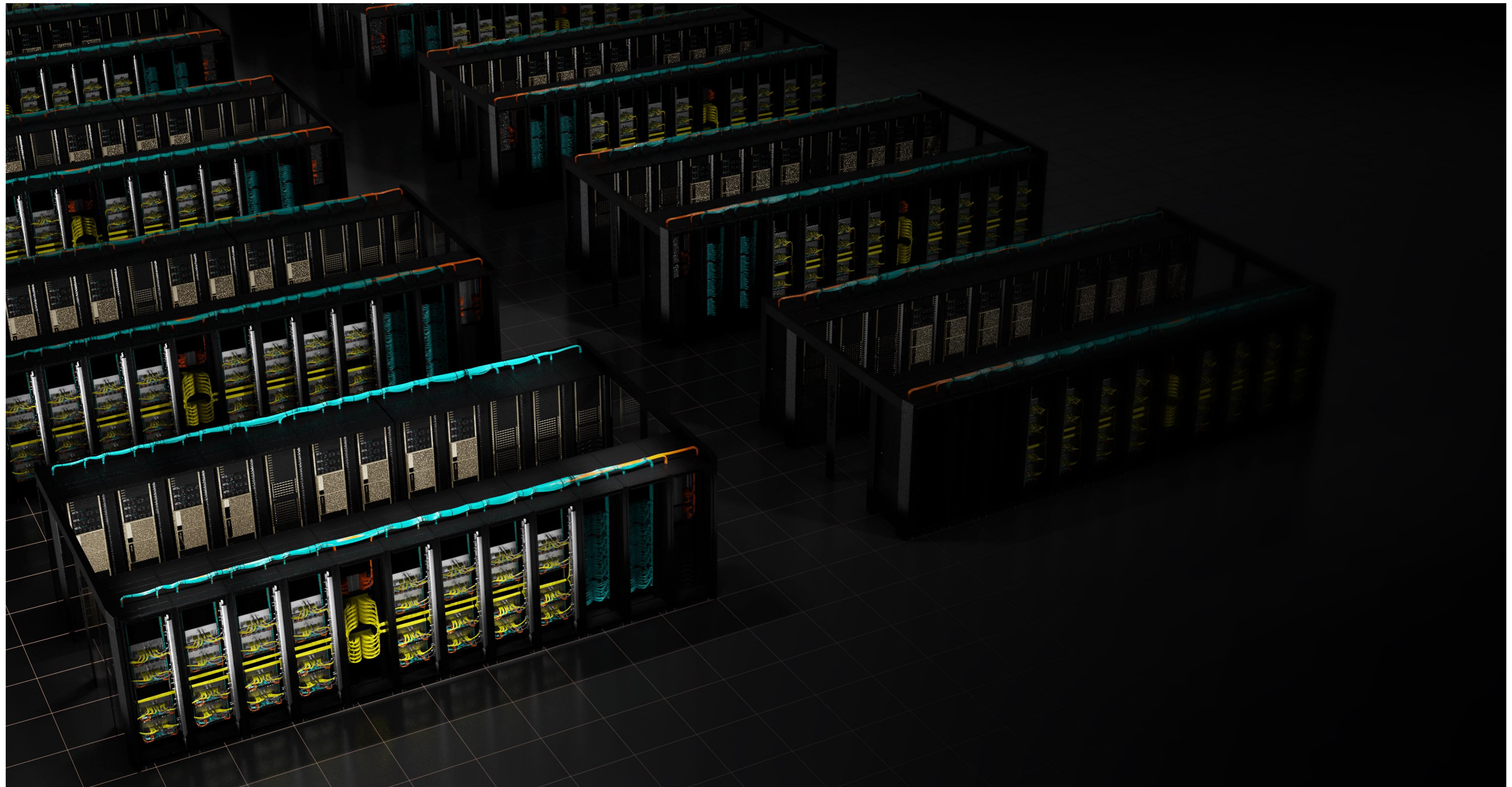
Trustworthy and Responsible AI Development

The NVIDIA accelerated computing platform can be leveraged to develop, deploy and validate AI Quality

Scaling and (semi-)automation is the next level - establishing a stack with tools/technologies for effective execution and cost reduction is critical

Examples:

- Post-hoc Explainable AI
- Robustness tests and simulation (synthetic data)
- MLOps workflows at scale
- Selection of fittest model from large model pool
- …



Leadership in AI Quality involves Accelerated Computing and ‚AI & Simulation Factories‘

Contact:

jpapenbrock@nvidia.com