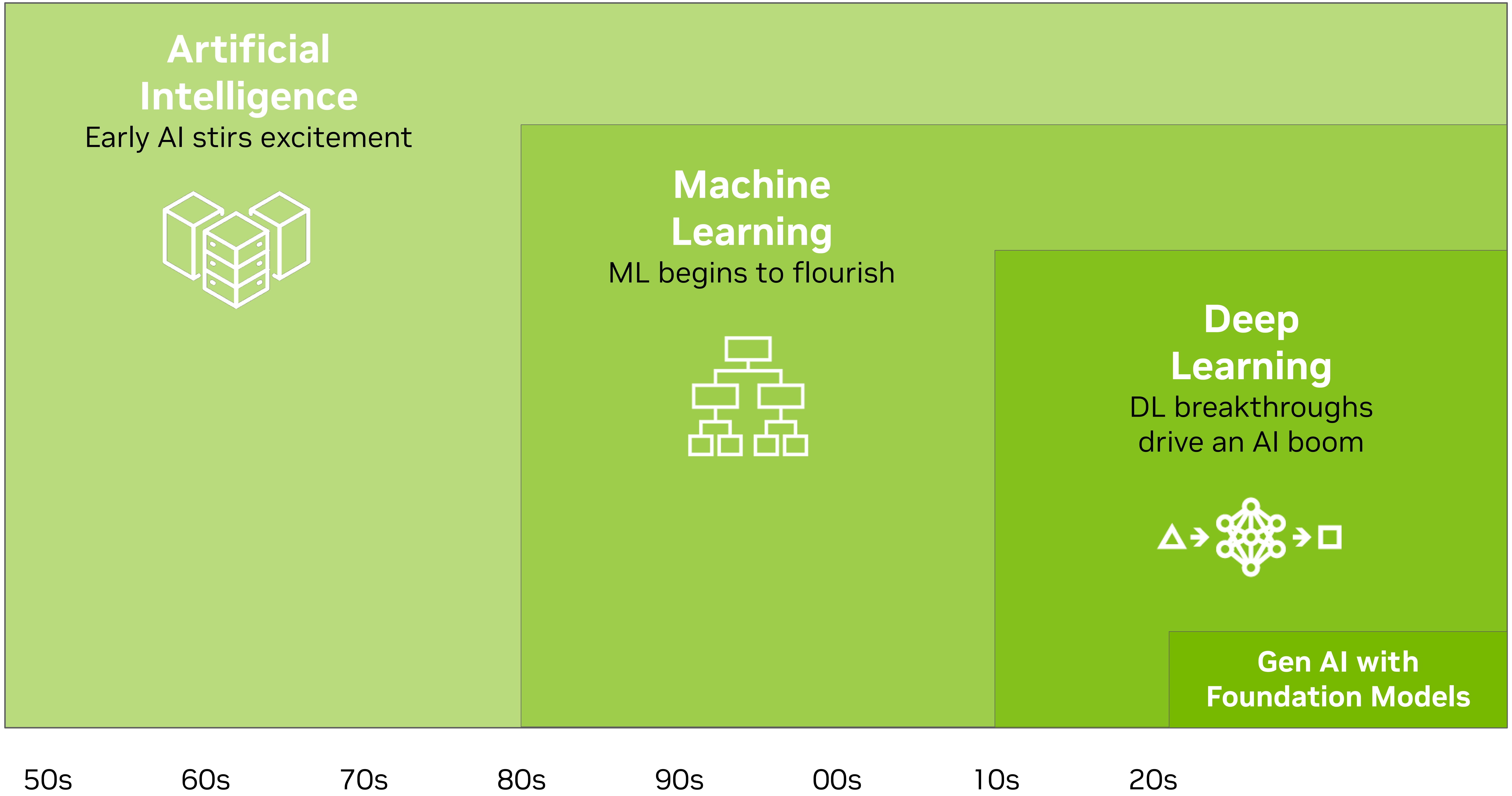
Unlocking the Power of LLMs with NVIDIA NeMo

Zahra Ronaghi, PhD – Manager, Solution Architecture at NVIDIA

Agenda

- The Evolution of AI
- Generative AI Adoption Across Industries
- The process and implications of creating LLM Models
- Pre-training foundation models
- Model alignment (SFT, PEFTs)
- BioNeMo – Example workflow
- Important Takeaways

The Evolution of AI



Artificial Intelligence

Early AI stirs excitement

Machine Learning

ML begins to flourish

Deep Learning

DL breakthroughs drive an AI boom

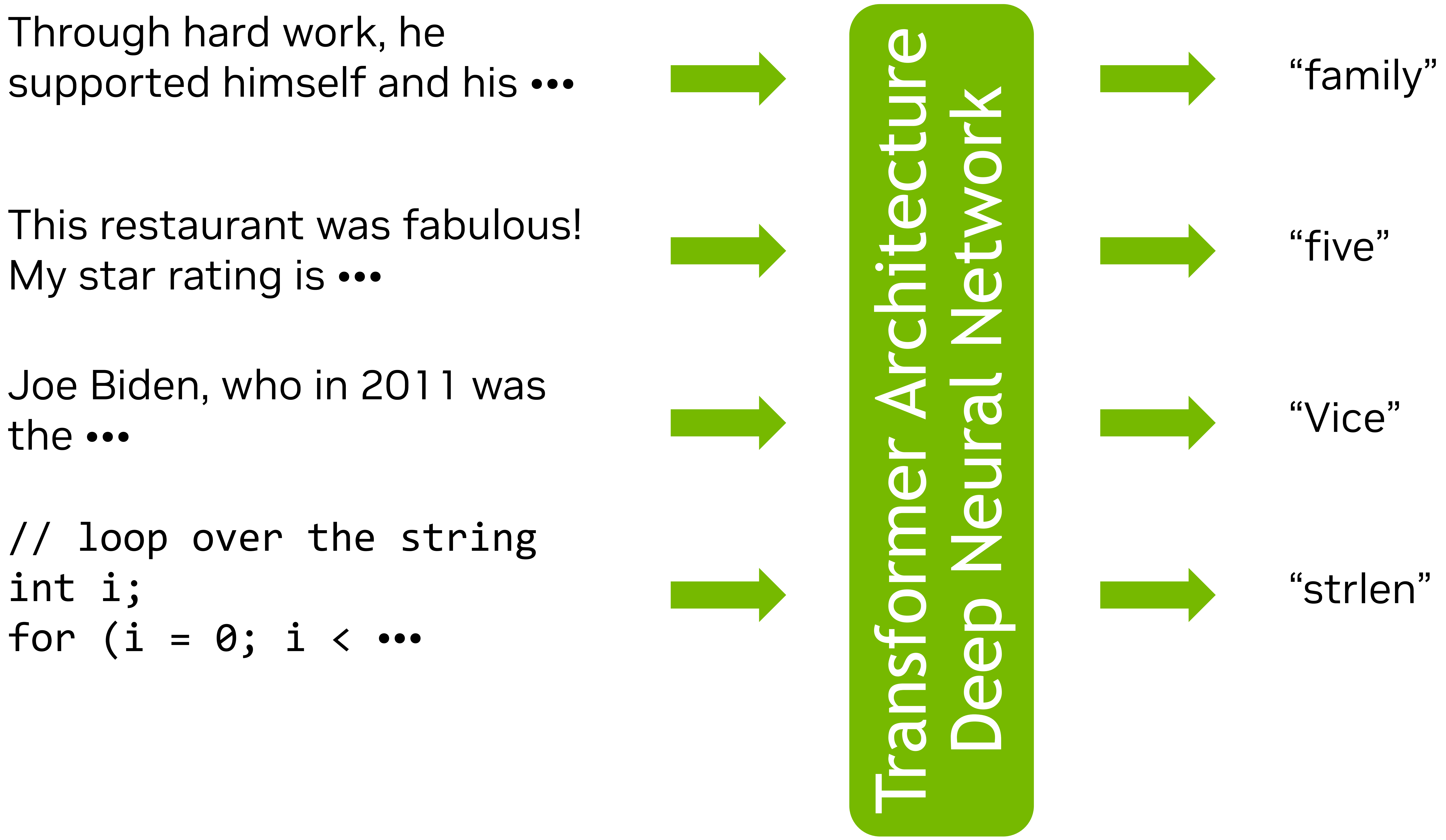
Gen AI with Foundation Models

50s 60s 70s 80s 90s 00s 10s 20s

An LLM is a Deep Neural Network

Map from “all previous words” to “next word”

Input: A few thousand previous words for context
Output: predict the next word or group of words



Through hard work, he supported himself and his •••

This restaurant was fabulous! My star rating is •••

Joe Biden, who in 2011 was the •••

```
// loop over the string
int i;
for (i = 0; i < •••
```

Transformer Architecture
Deep Neural Network

“family”

“five”

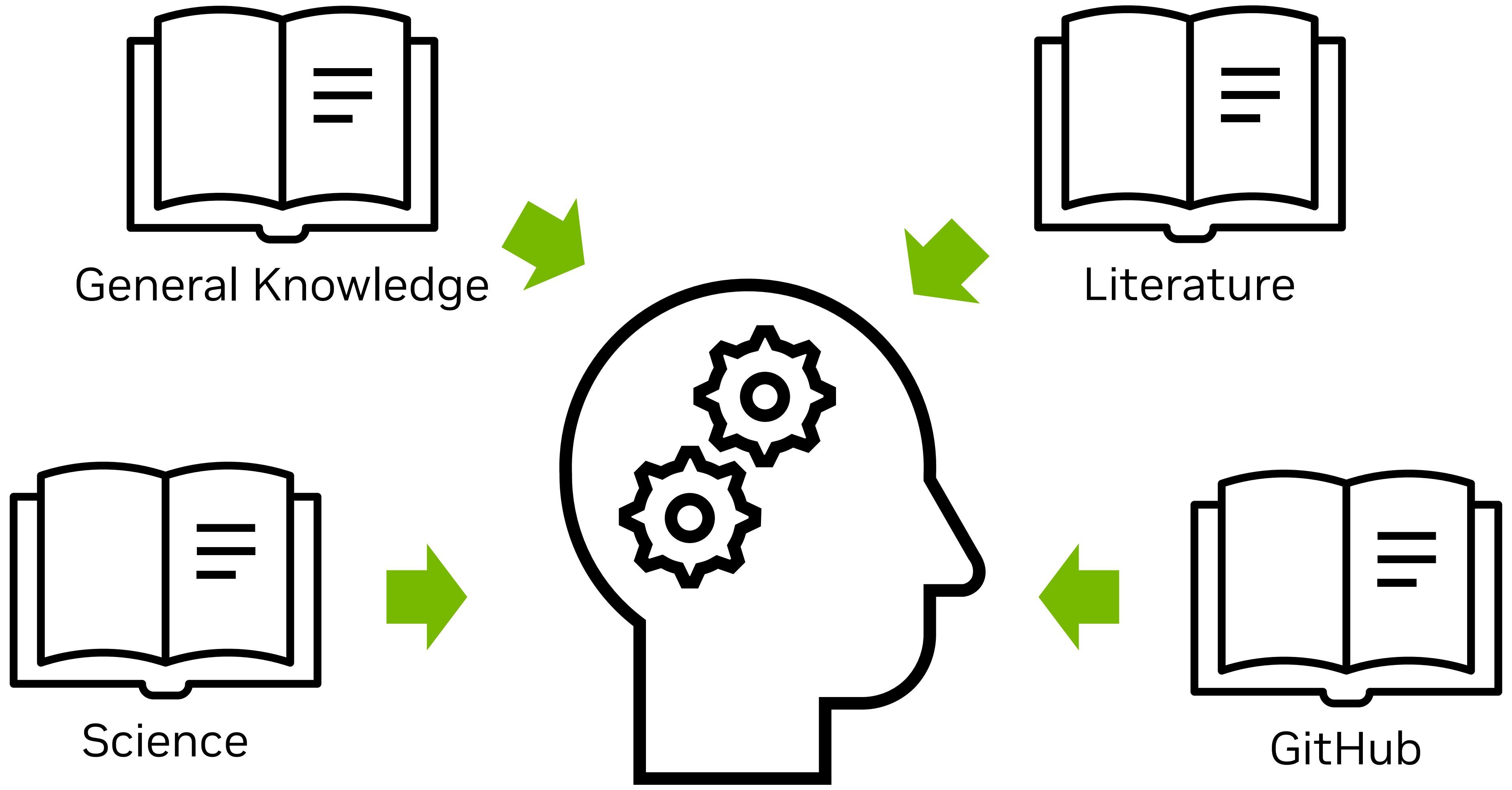
“Vice”

“strlen”

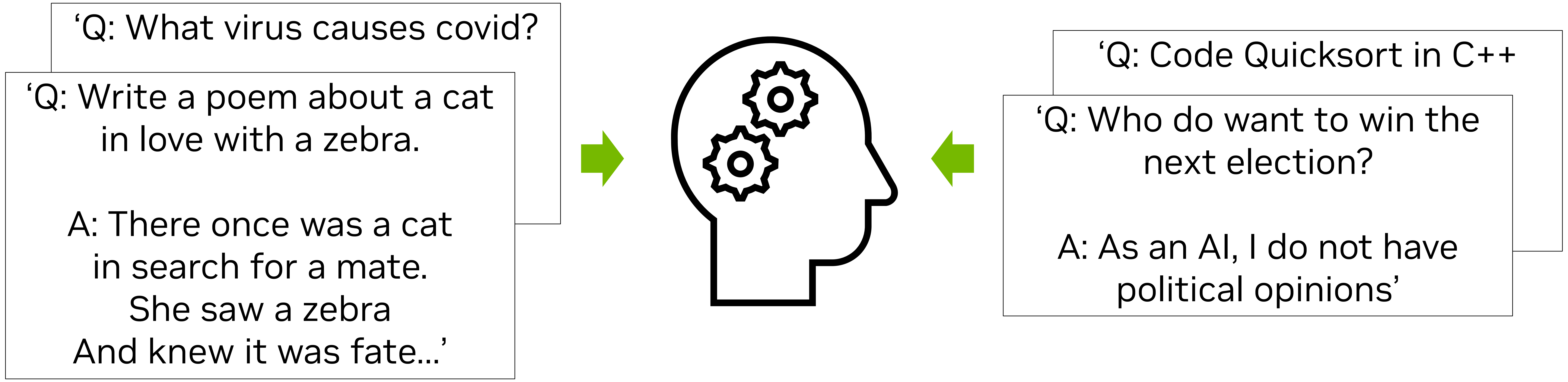
How to train an LLM

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++

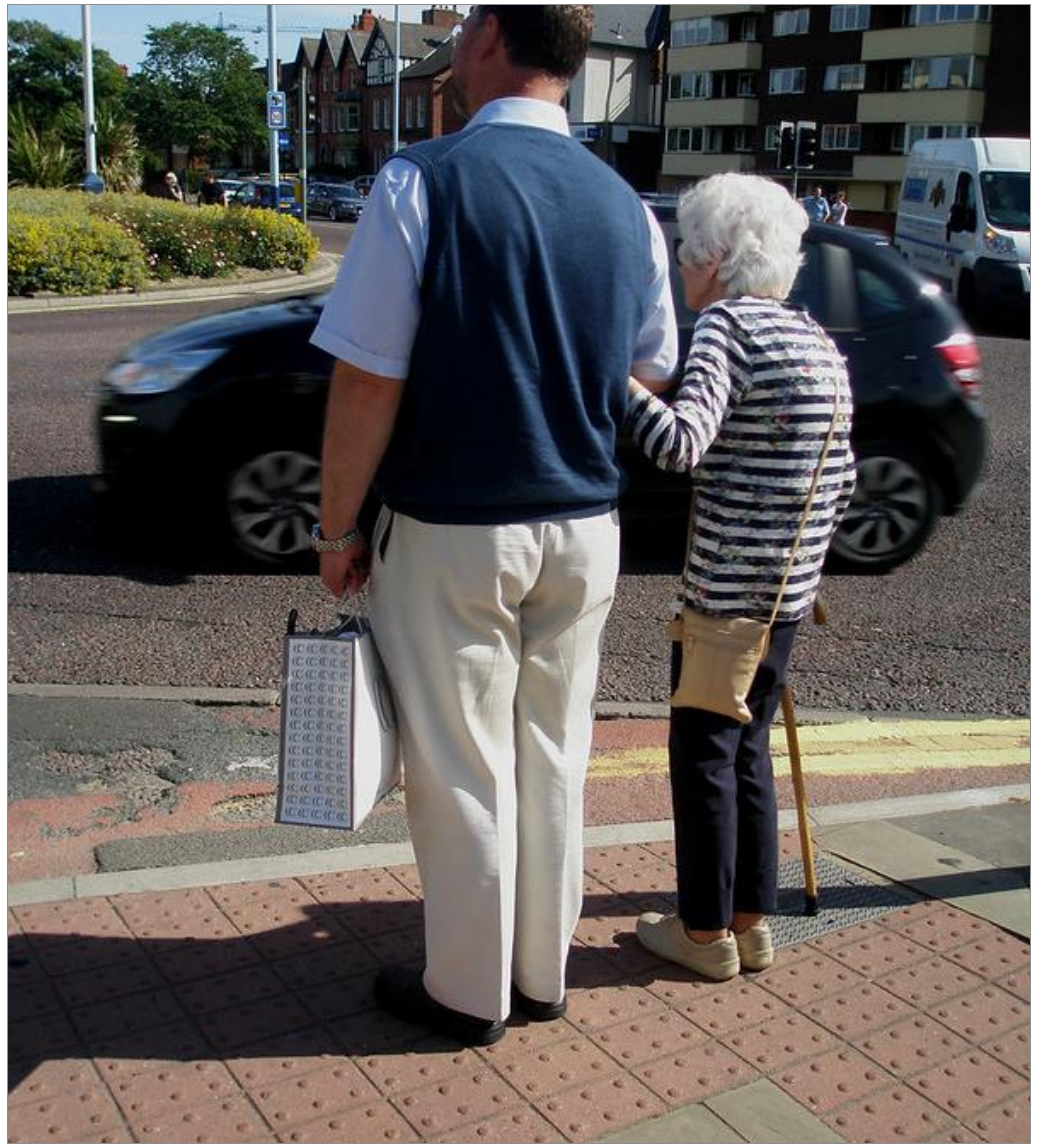
‘Q: Who do want to win the
next election?

A: As an AI, I do not have
political opinions’

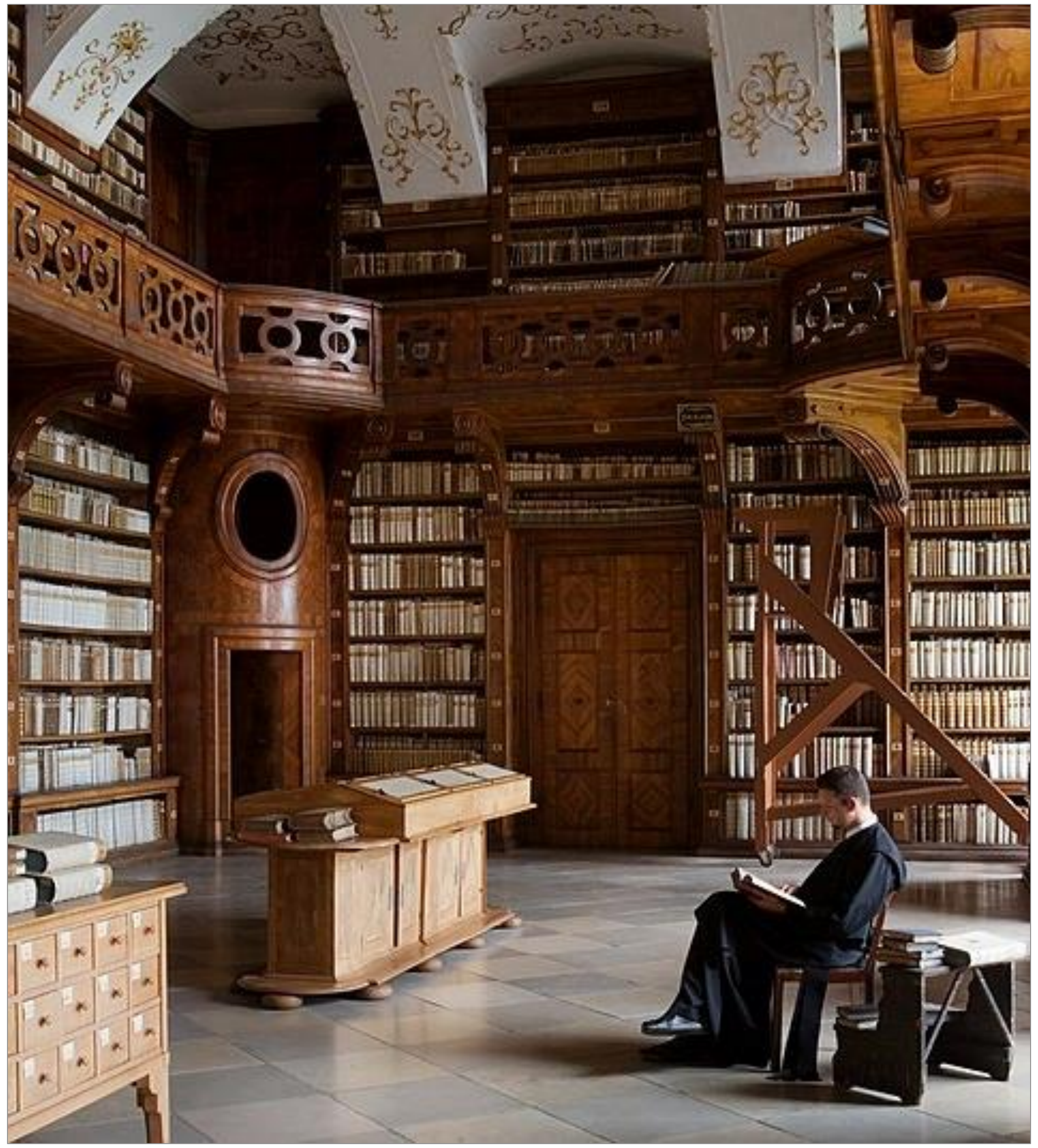
Custom AIs

Turn foundation model into a domain-specific AI
(p-tuning, LoRA, SFT, RLHF, SteerLM, …)

Train it on a skill
Learn to 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

Generative AI is Transforming Business

TEXT GENERATION

Summarization

GPT-3

Marketing Copy

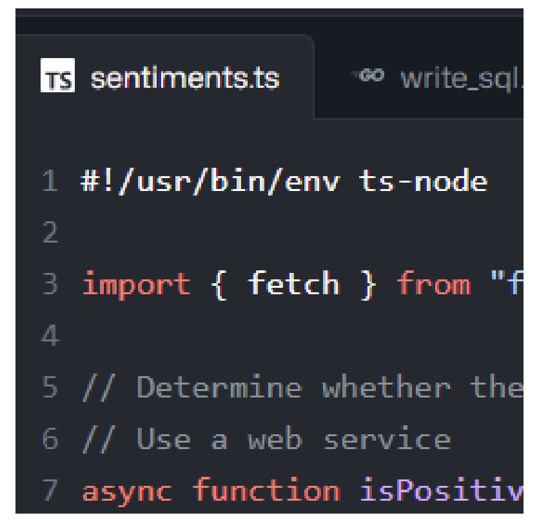
TRANSLATION

Translating Wikipedia

NLLB-200

Real-Time Translation

CODING



```
#!/usr/bin/env ts-node

import { fetch } from "f

// Determine whether the
// Use a web service
async function isPositiv
```

Dynamic Code Commenting

CODEX

Function Generation

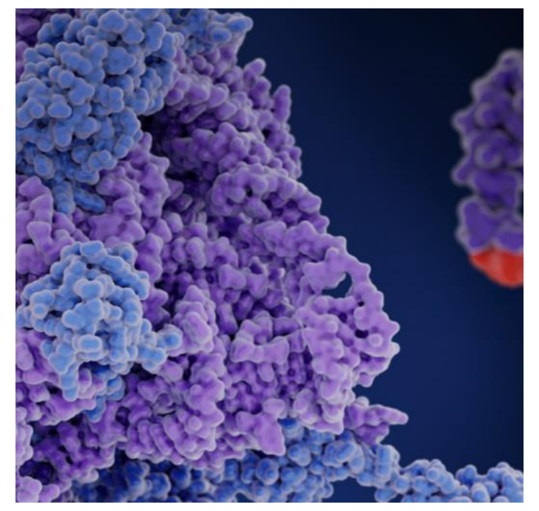
VISUAL CONTENT

Brand Creation

e-Diffi

Gaming Characters

LIFE SCIENCE



Molecular Representations

MegaMolBART

Drug Discovery

Enterprises that adopt next-generation AI like LLMs and Generative AI are **2.6X more likely to increase revenue by 10% or more** but must invest in their AI infrastructure to fully reap the benefits.

-Accenture Research. Breakthrough Innovation: Is your organization equipped for breakthrough innovation? WEF 2023.

How Enterprises are Using Generative AI

Fastest
Fastest Time to Adoption
Slowest
Least Difficult
Level of Difficulty to Develop & Deploy
Most Difficult

Some Customization

Generative AI as a Service - ChatGPT, Google Bard, Amazon Bedrock, Existing Services
Consumption model, $ per inference
Fastest time to market
Langchain and Inferencing

Moderate Customization

P-tuning and fine tuning of pre-trained model
$M+ for infrastructure and resources
Weeks to months for development

Extensive Customization

Custom foundation models or extensive finetuning
$10M+ for infrastructure and resources
6+ months for development

Requirements for Building Custom LLMs

Training Data

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

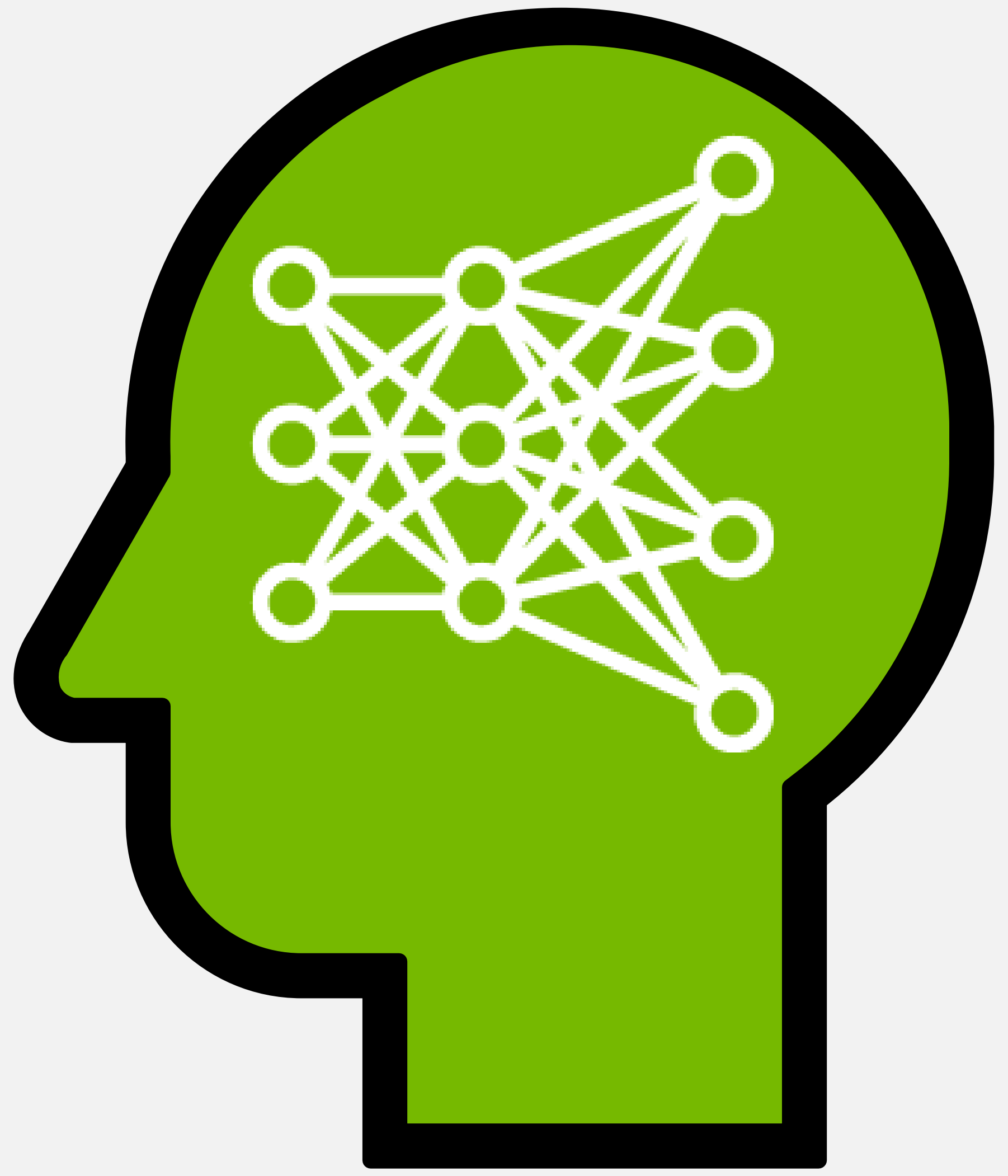
Foundations Models

Training & Customization

Accelerated Inference

AI Expertise

Internal Expertise



Solution Delivery Partners

Generative AI Adoption Across Industries

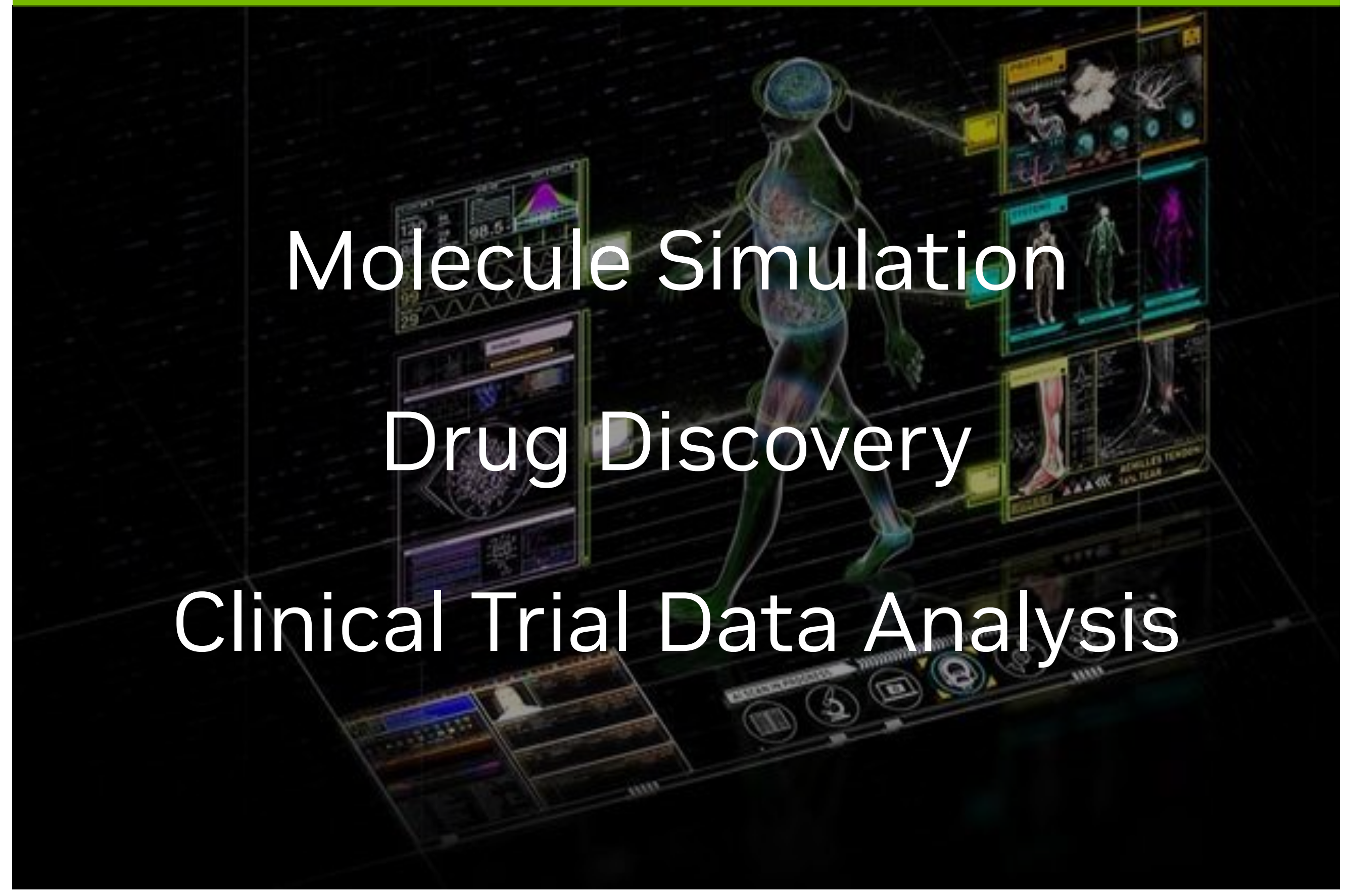
Finance

Fraud Detection

Personalized Banking

Investment Insights

Healthcare

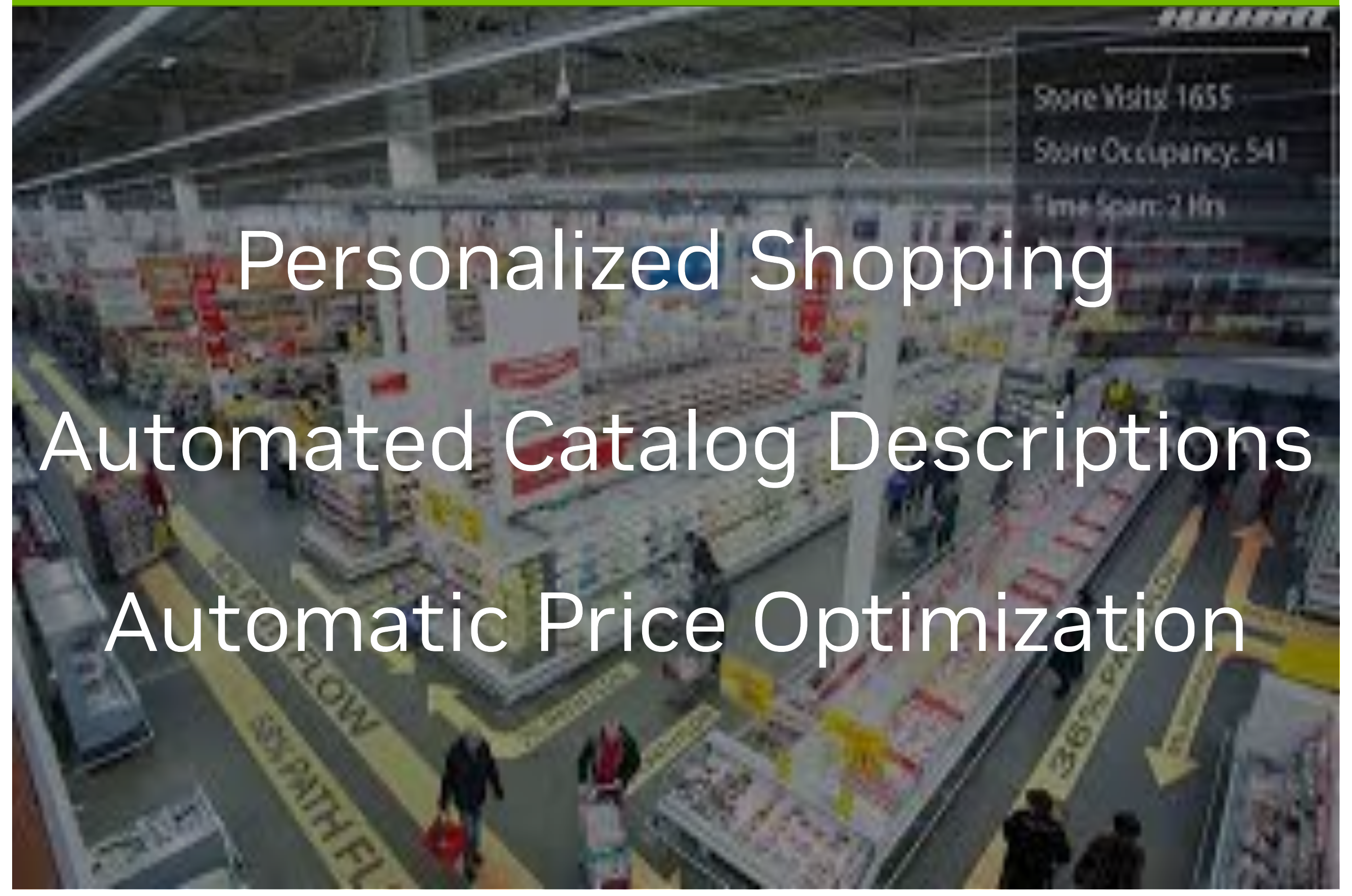


Molecule Simulation

Drug Discovery

Clinical Trial Data Analysis

Retail



Personalized Shopping

Automated Catalog Descriptions

Automatic Price Optimization

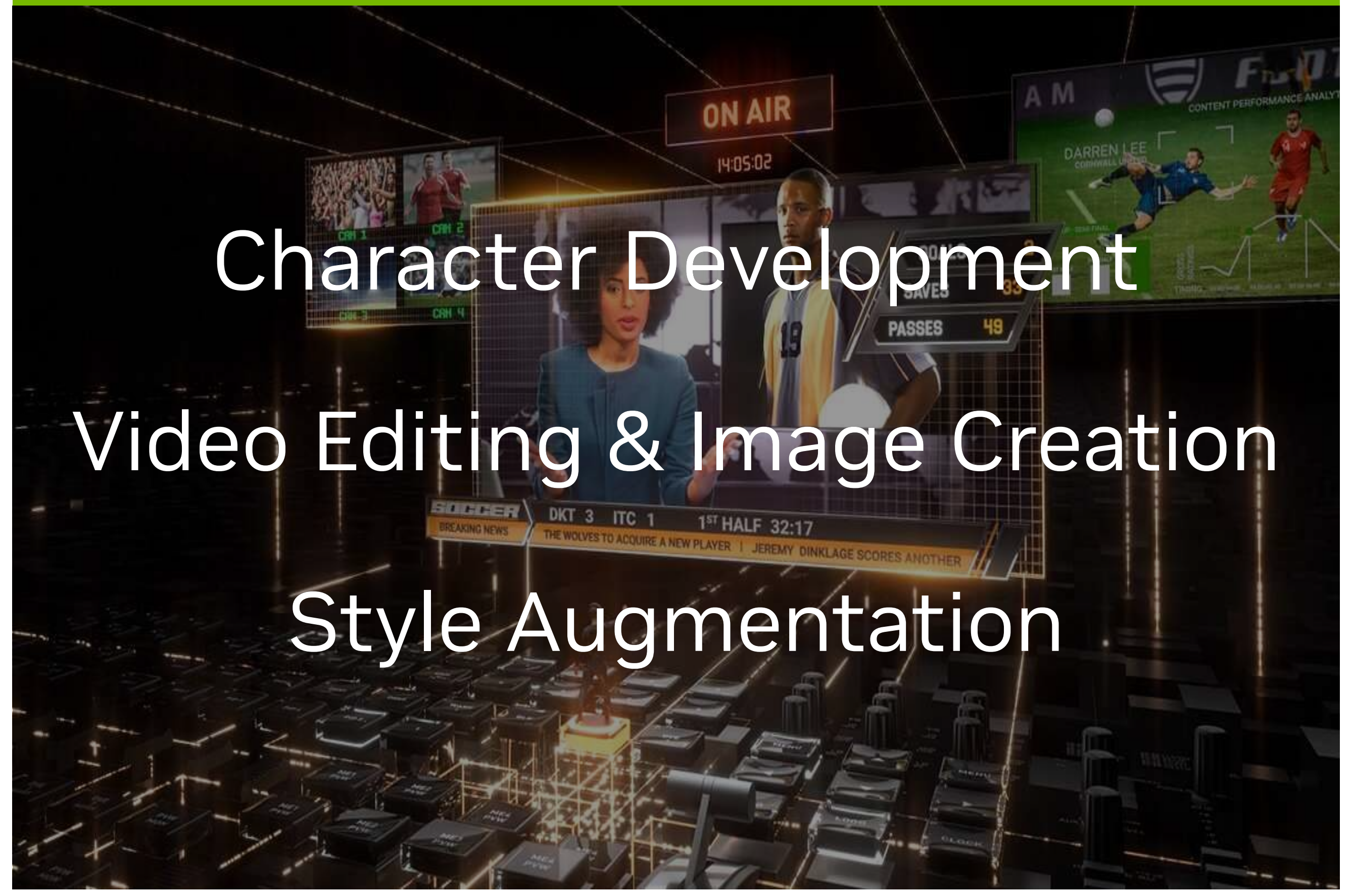
Telecommunications

AI Virtual Assistants

Network Performance Tuning

Remote Support Capabilities

Media & Entertainment



Character Development

Video Editing & Image Creation

Style Augmentation

Manufacturing

Factory Simulation

Product Design

Predictive Maintenance

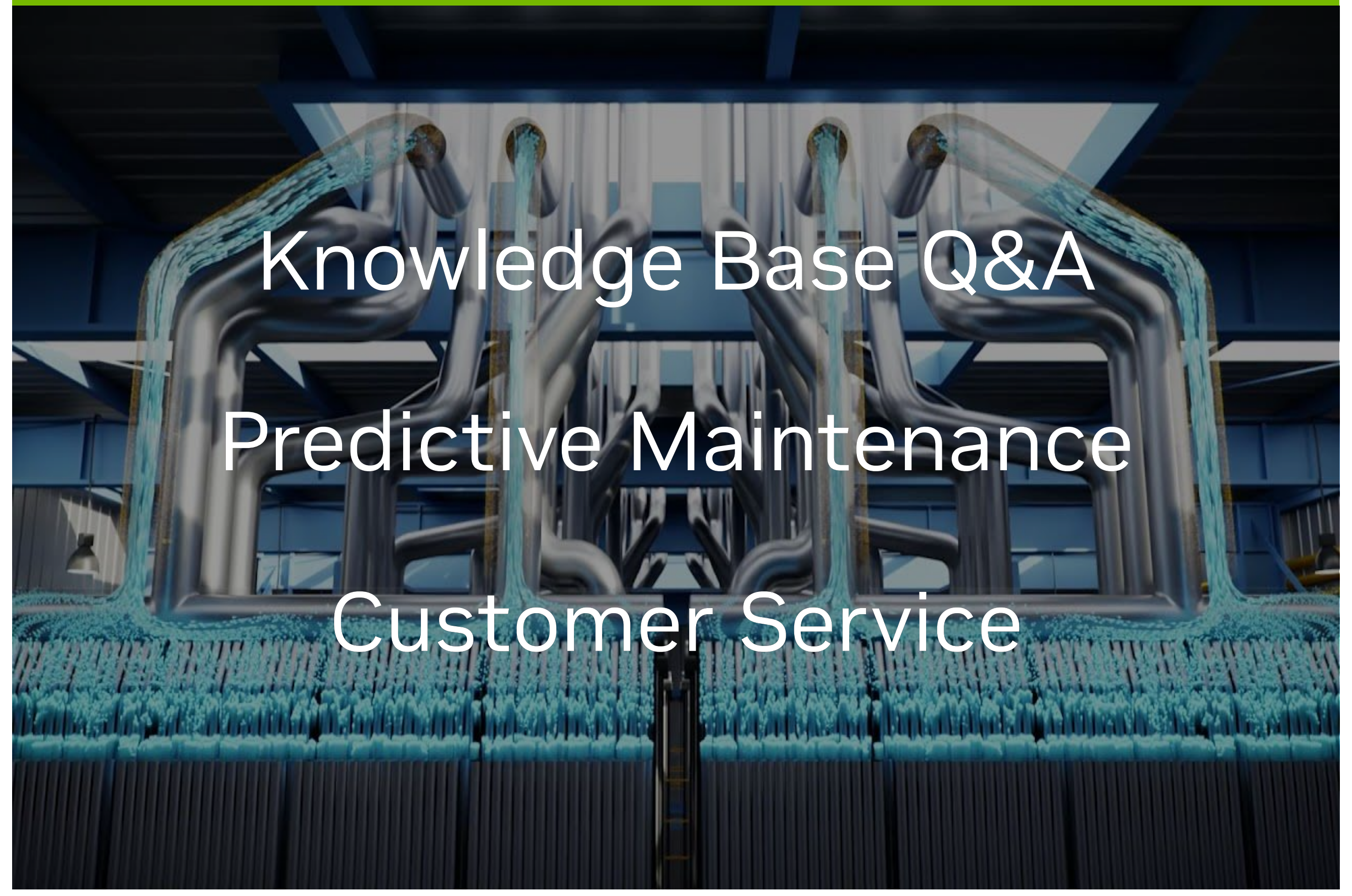
Federal

Document Summarization

Audit Compliance

AI Virtual Assistants

Energy



Knowledge Base Q&A

Predictive Maintenance

Customer Service

Custom Generative AI for Enterprise IT

ServiceNow and NVIDIA have partnered to develop generative AI capabilities aimed at enhancing workflow automation across various business processes.

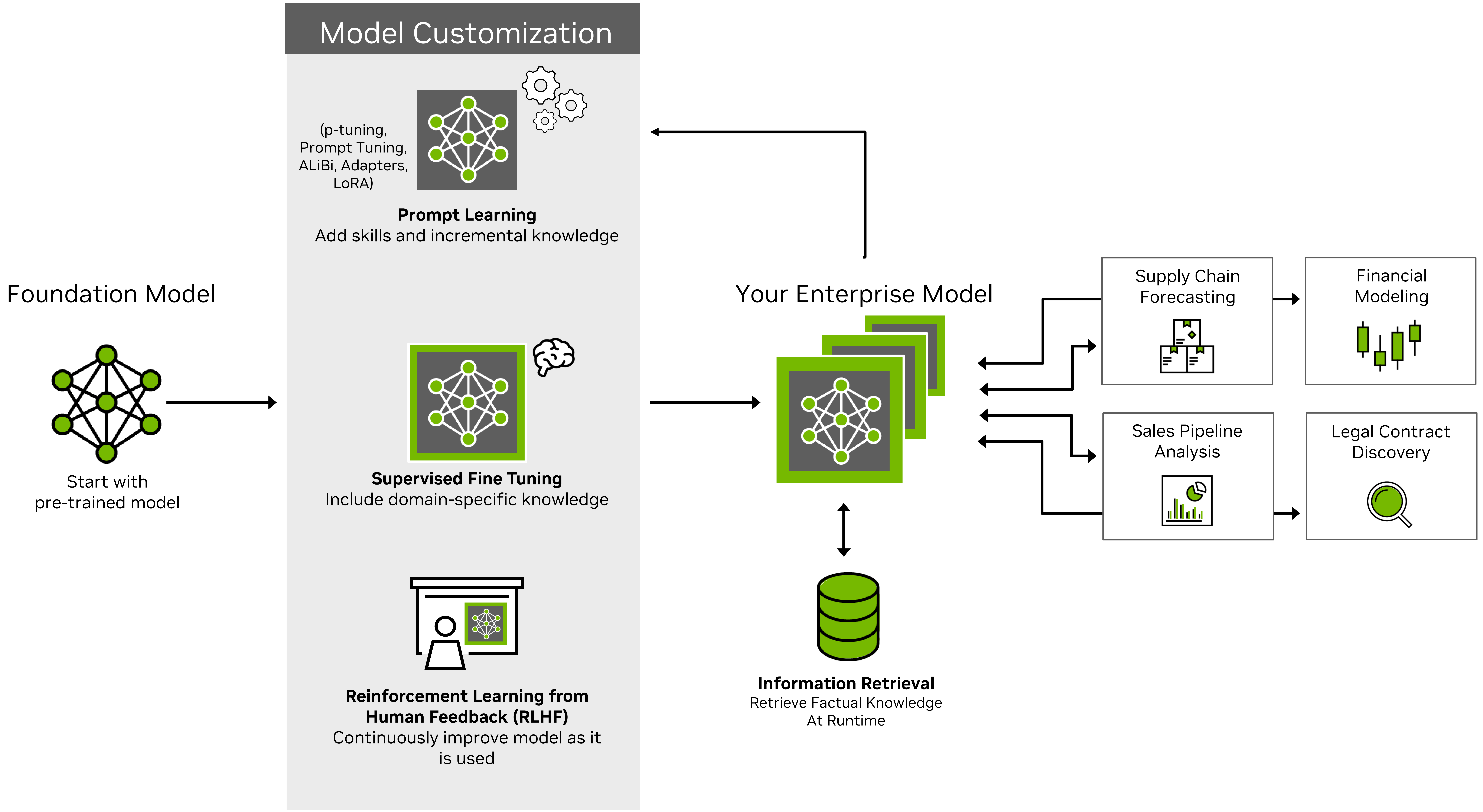
Leveraging NVIDIA's technology, ServiceNow is creating large language models trained on its specific data. This will enhance ServiceNow's existing AI functionality, enabling new applications of generative AI across the enterprise, including IT, customer service, and developers, to bolster workflow automation and boost productivity.

This innovative AI solution will provide higher accuracy and value in IT tasks, reshape customer service, and improve the employee experience.

servicenow NVIDIA

Model Customization for Enterprise Ready LLMs

Customization techniques to overcome the challenges of using foundation models



Foundation Model

Start with
pre-trained model

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

Information Retrieval
Retrieve Factual Knowledge
At Runtime

Supply Chain
Forecasting

Financial
Modeling

Sales Pipeline
Analysis

Legal Contract
Discovery

How it all fits together

Training from the left, Inference from the right

Data Science Team
Model Development
Foundation Model Building
Data Curation
Distributed Training
Pre-Trained Foundation Models
GPT-8B
GPT-22B
GPT-43B
Community Models (Llama2, StarCoder, ...)
Model Customization
Deploy
Inference
Vector Database
Guardrails
Queries
Proprietary Knowledge
Associates
Vendors
Customers
NeMo Framework
NVIDIA AI Enterprise

Multi-Modality

Build language, image, generative AI models

Data Curation at 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

Run optimized inference at-scale anywhere

Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA AI Enterprise and experts by your side to keep projects on track

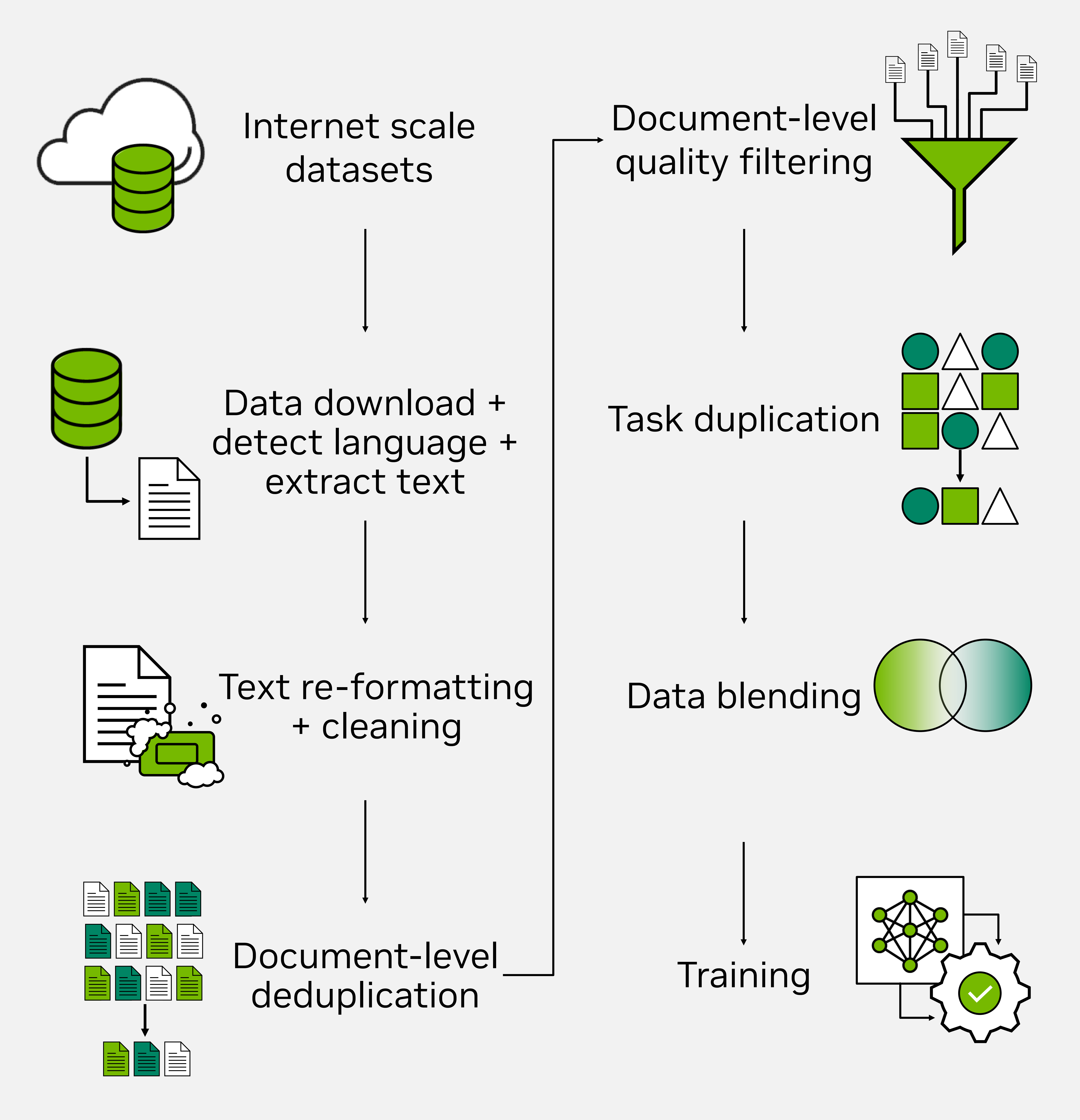
Data Curation Improves Model Performance

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


Internet scale datasets
Data download + detect language + extract text
Text re-formatting + cleaning
Document-level deduplication
Document-level quality filtering
Task duplication
Data blending
Training

NVIDIA NeMo Works with Powerful Generative Foundation Models

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

Fastest Responses

GPT-8

GPT-8B w/ 3.5T tokens. +SFT, SteerLM.
53 Languages I/O: 4K tokens

Balance of Accuracy - Latency

GPT-22

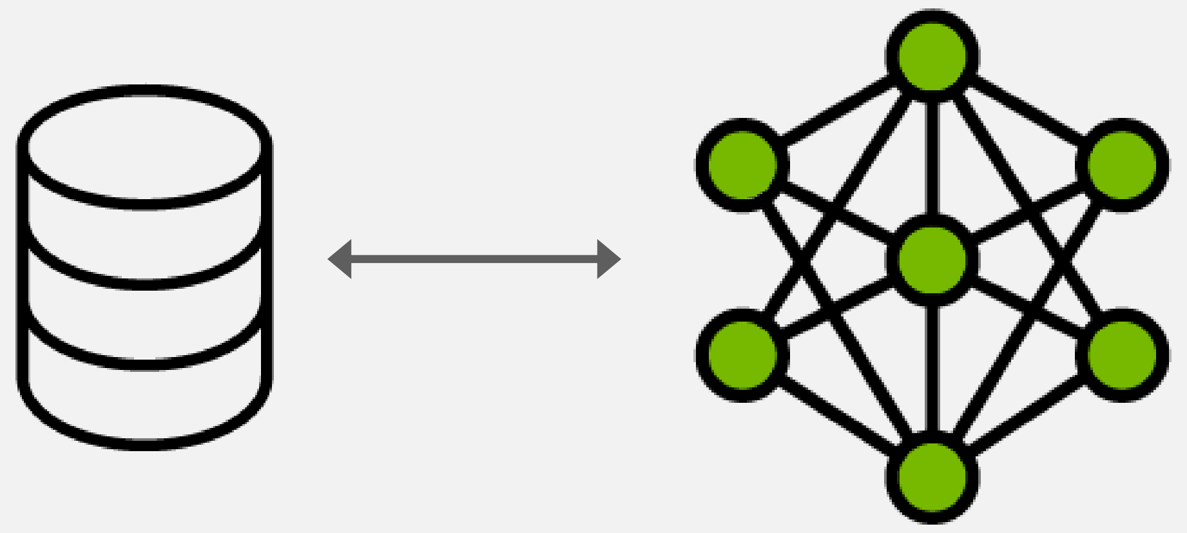
GPT-22B w/ 1.1T tokens. + SFT private mix.
50 Languages. I/O: 4K tokens

For Complex Tasks

GPT-43

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

Information Retrieval



Community-Built Models

Code Llama
Meta

Falcon LLM
Falcon

Llama 2
Meta

MPT
Mosaic ML

StarCoder
ServiceNow & Hugging Face

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	INSTRUCTION TUNING
Techniques	· Few-shot learning · Chain-of-thought reasoning · System prompting	· Prompt tuning · P-tuning	· Adapters · LoRA · IA3	· SFT · RLHF
Pros	· Good results leveraging pre-trained LLMs · Lowest investment · Least expertise	· Better results leveraging pre-trained LLMs · Lower investment · Will not forget old skills	· Best results leveraging pre-trained LLMs · Will not forget old skills	· Best results leveraging pre-trained LLMs · Change all model parameters
Cons	· Cannot add as many skills or domain specific data to pre-trained LLM	· Less comprehensive ability to change all model parameters	· Medium investment · Takes longer to train · More expertise needed	· May forget old skills · Large investment · Most expertise needed

“Using hyperparameter optimization tools in NeMo allowed us to train LLMs 2x faster than with other frameworks.”

Hwijung Ryu, LLM Development Team Lead
Korea Telecom

Auto-Configurator Tool

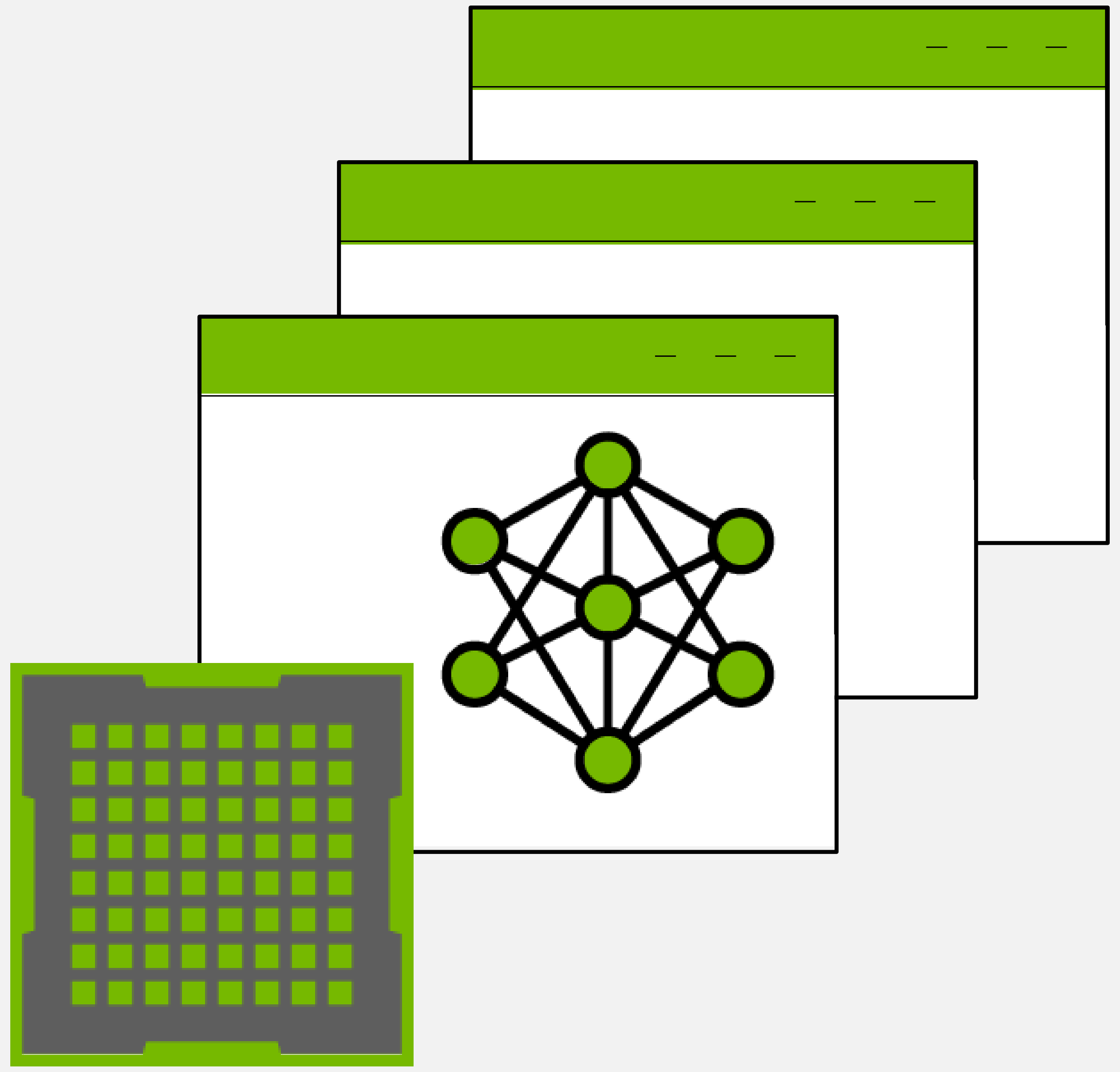
Automatically search and optimize model configurations on any given compute or time constraints

- 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)

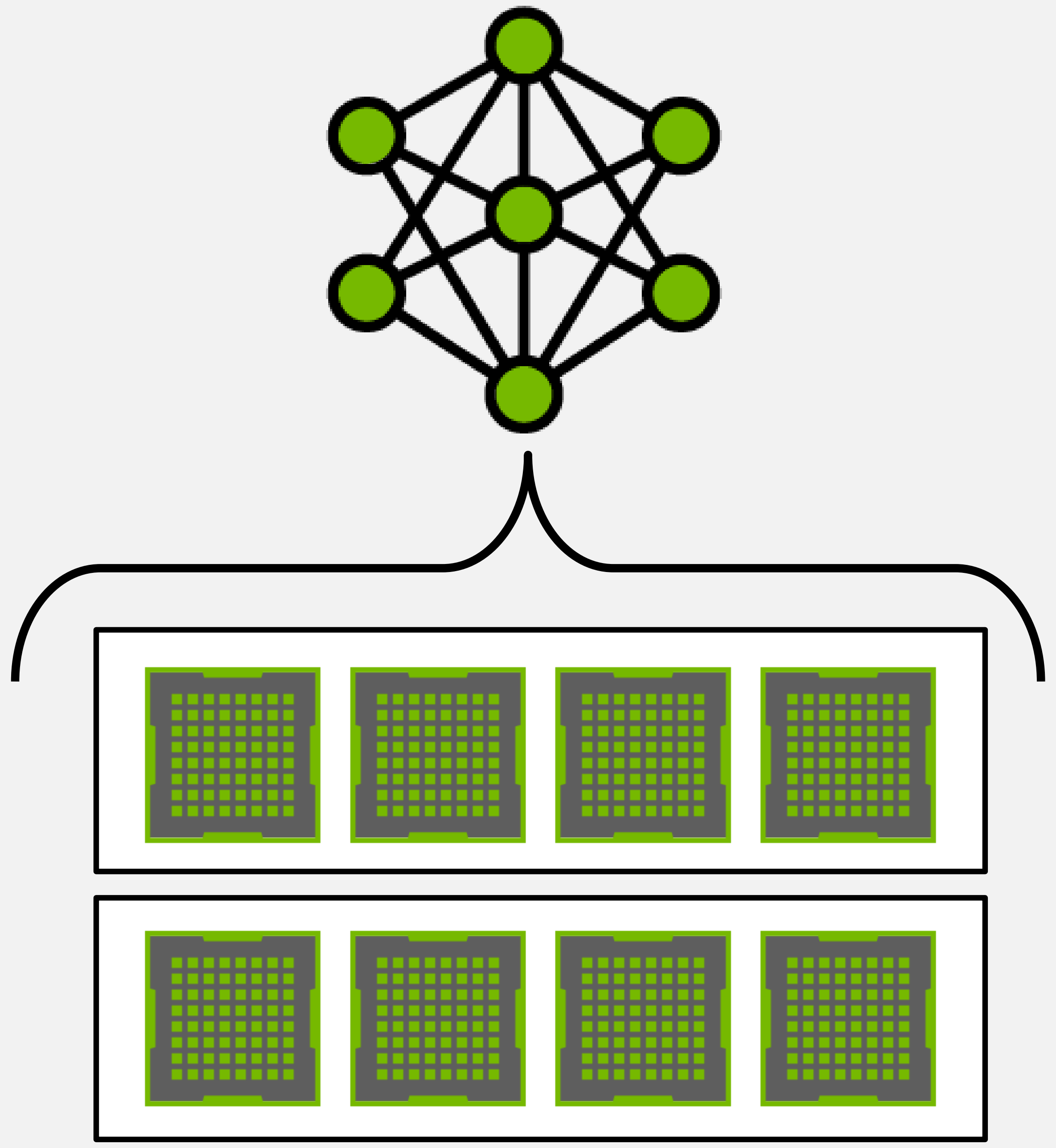
Deploying Large Scale Inference for Generative AI

Efficiently Deploy Generative AI Models At-scale With NeMo

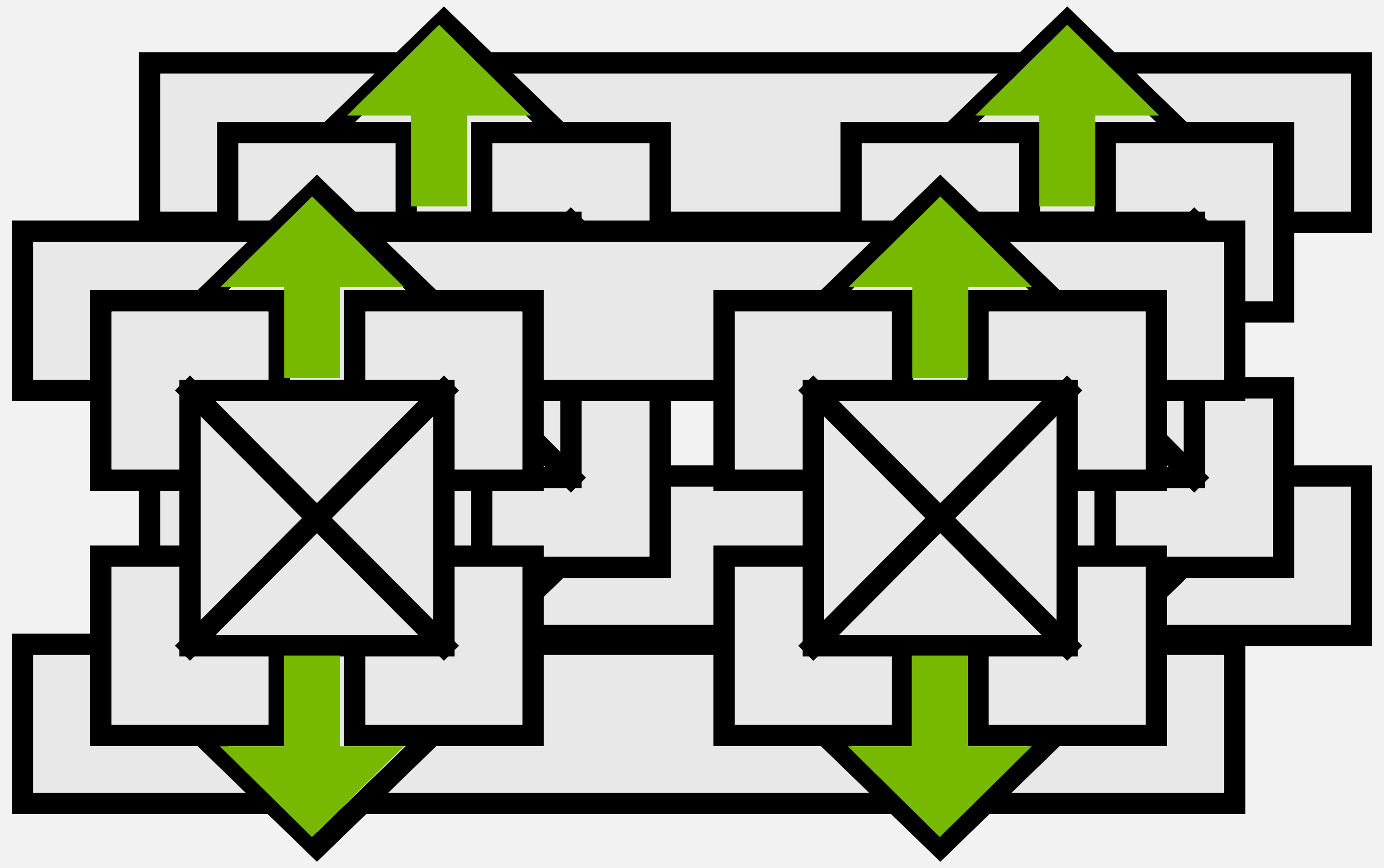
Optimized Kernels for Accelerated Performance



Multi-GPU and Multi-Node Inferencing



Intra/Inter-Node Communication



Summary

LLMs build on long history of AI and Deep Learning

Innovation in AI continues to accelerate *exponentially*

Two simultaneous revolutions: Rise of LLM and Rise of Accelerated Computing

"Zero shot" foundational models generalize to solve new problems *without* training data – this is their value!

But with proprietary data, they get even better!

LLMs will transform business in every industry

Get Started with NeMo

Web Pages

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

Blogs

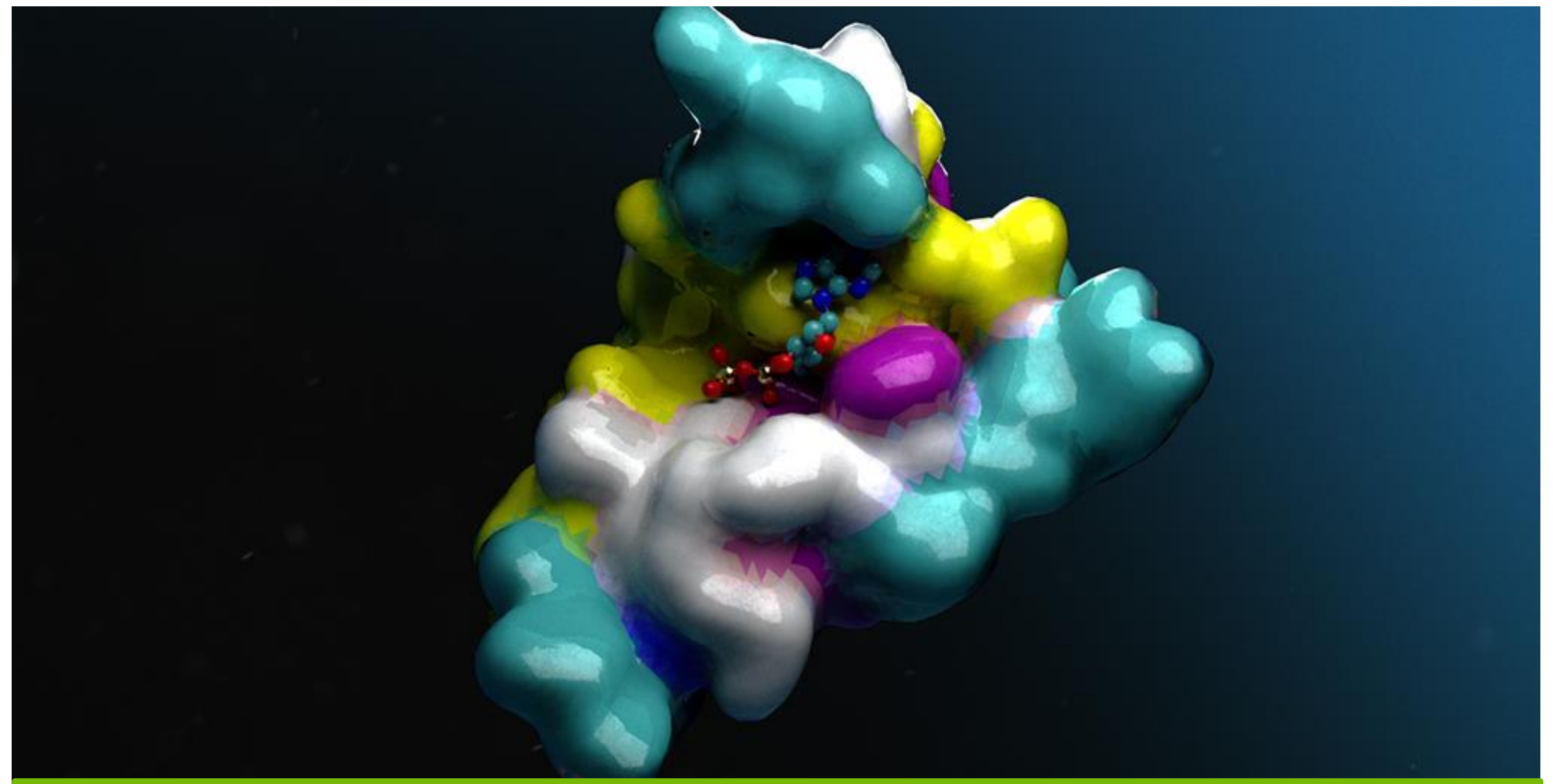
- What are Large Language Models?
- What Are Large Language Models Used For?
- What are Foundation Models?
- How To Create A Custom Language Model?
- Adapting P-Tuning to Solve Non-English Downstream Tasks
- NVIDIA AI Platform Delivers Big Gains for Large Language Models
- The King's Swedish: AI Rewrites 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

NVIDIA Generative AI Platform

NeMo
Language



BioNeMo
Life Sciences

Picasso
Visual Content

NVIDIA AI Enterprise

DGX & DGX Cloud

aws · Google Cloud · Microsoft Azure · ORACLE Cloud Infrastructure

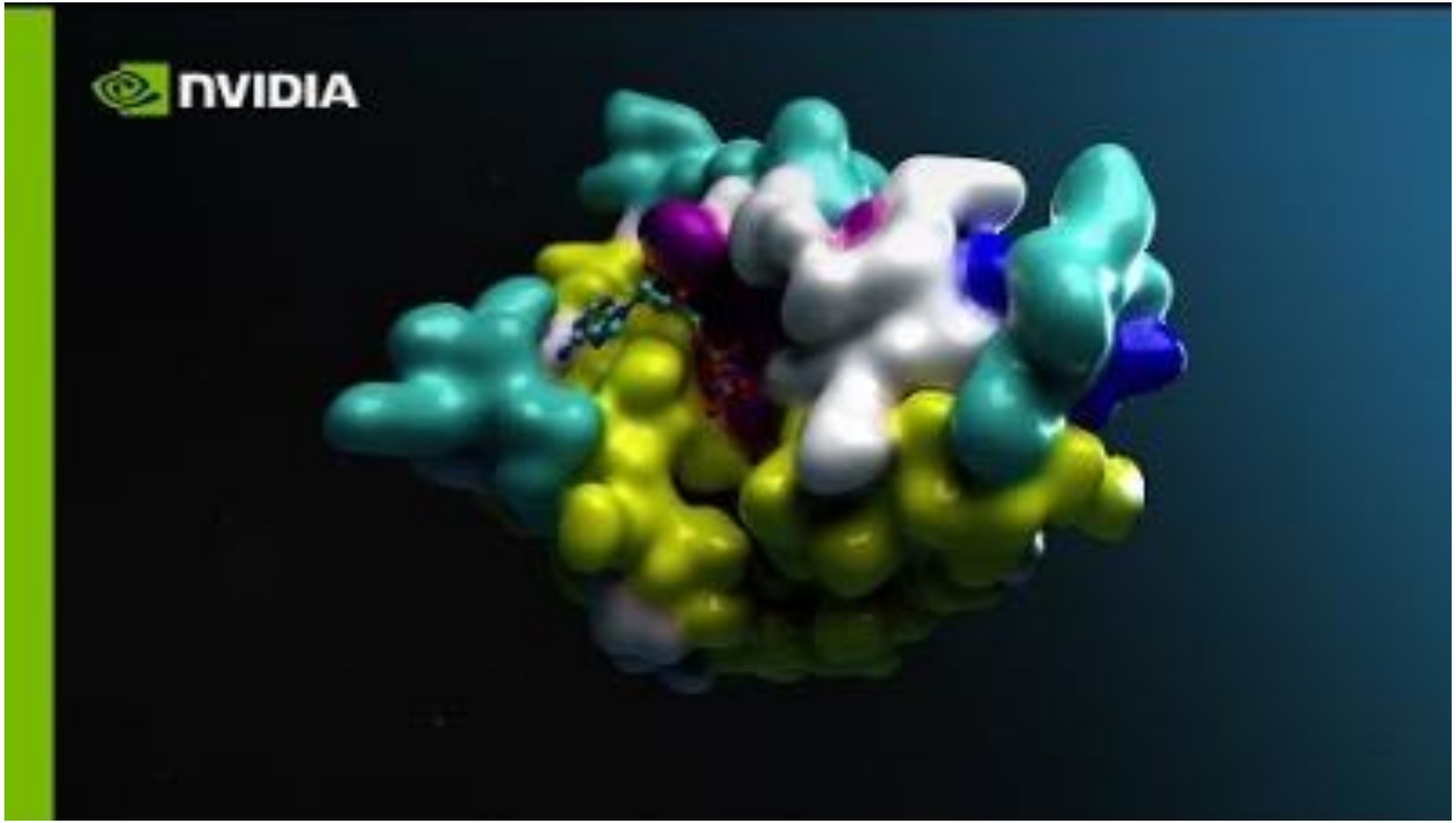
Cloud

DELL Technologies · Hewlett Packard Enterprise · Lenovo

On-Prem

Accelerated Compute Infrastructure

BioNeMo Demo



NVIDIA®