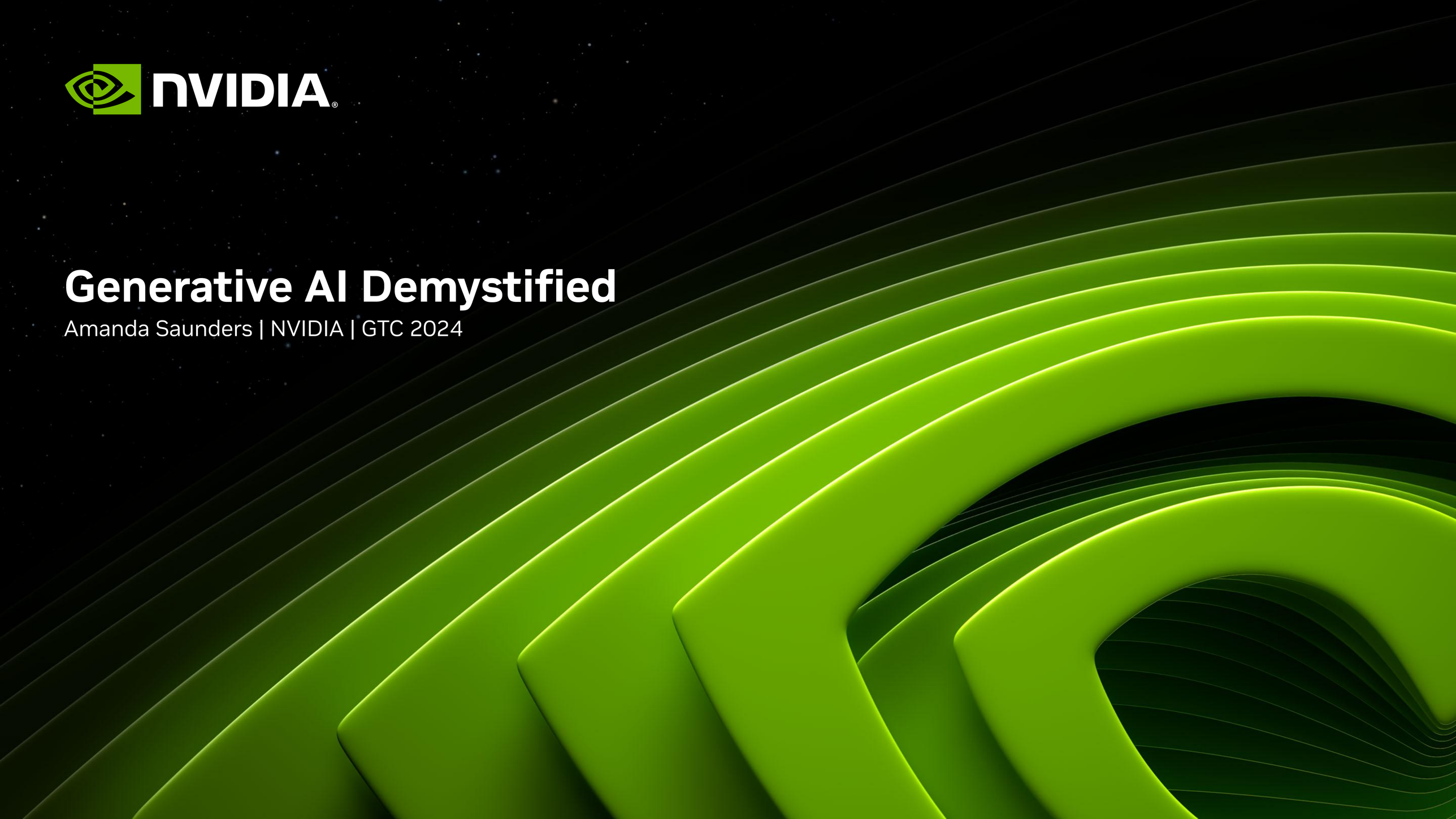
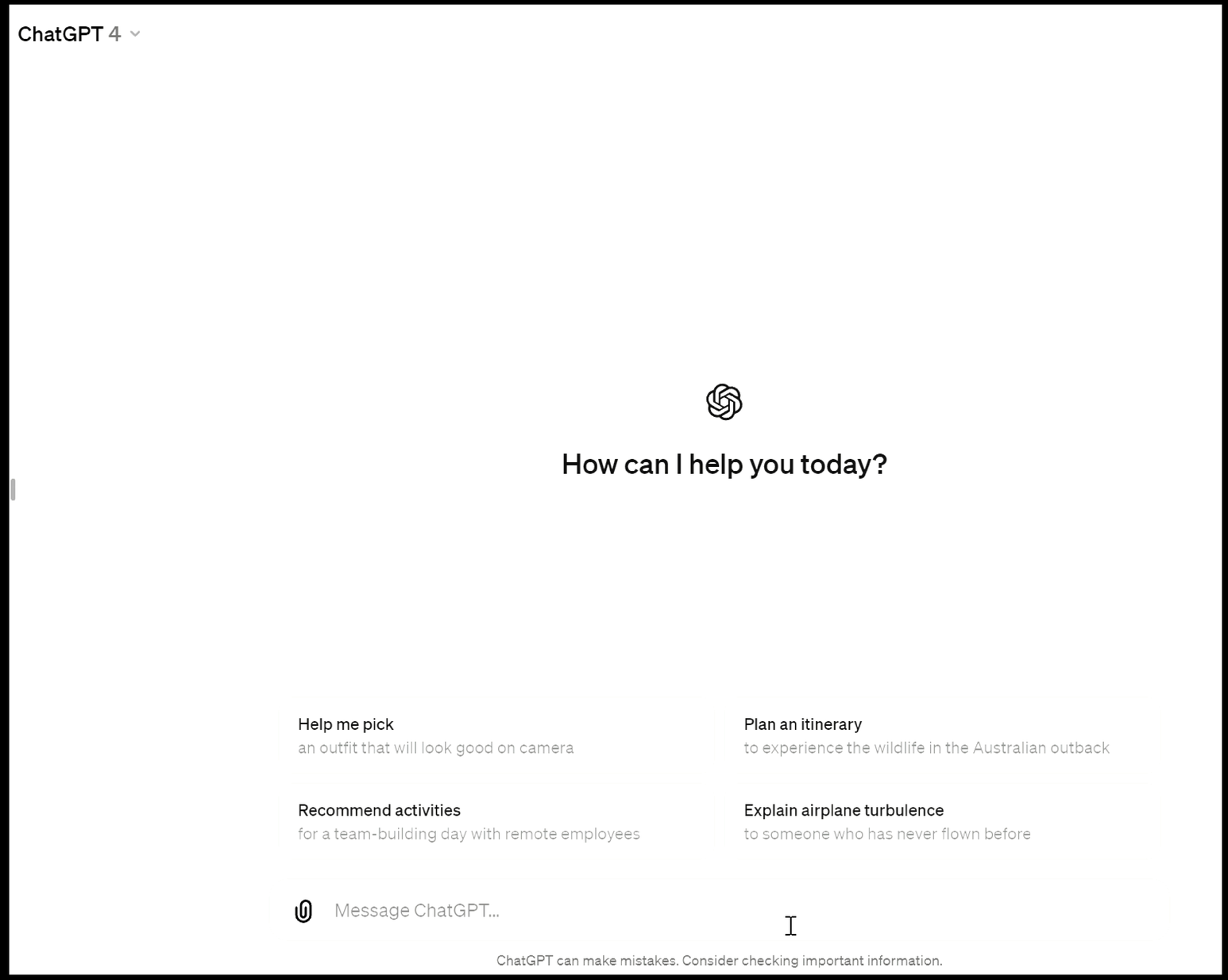


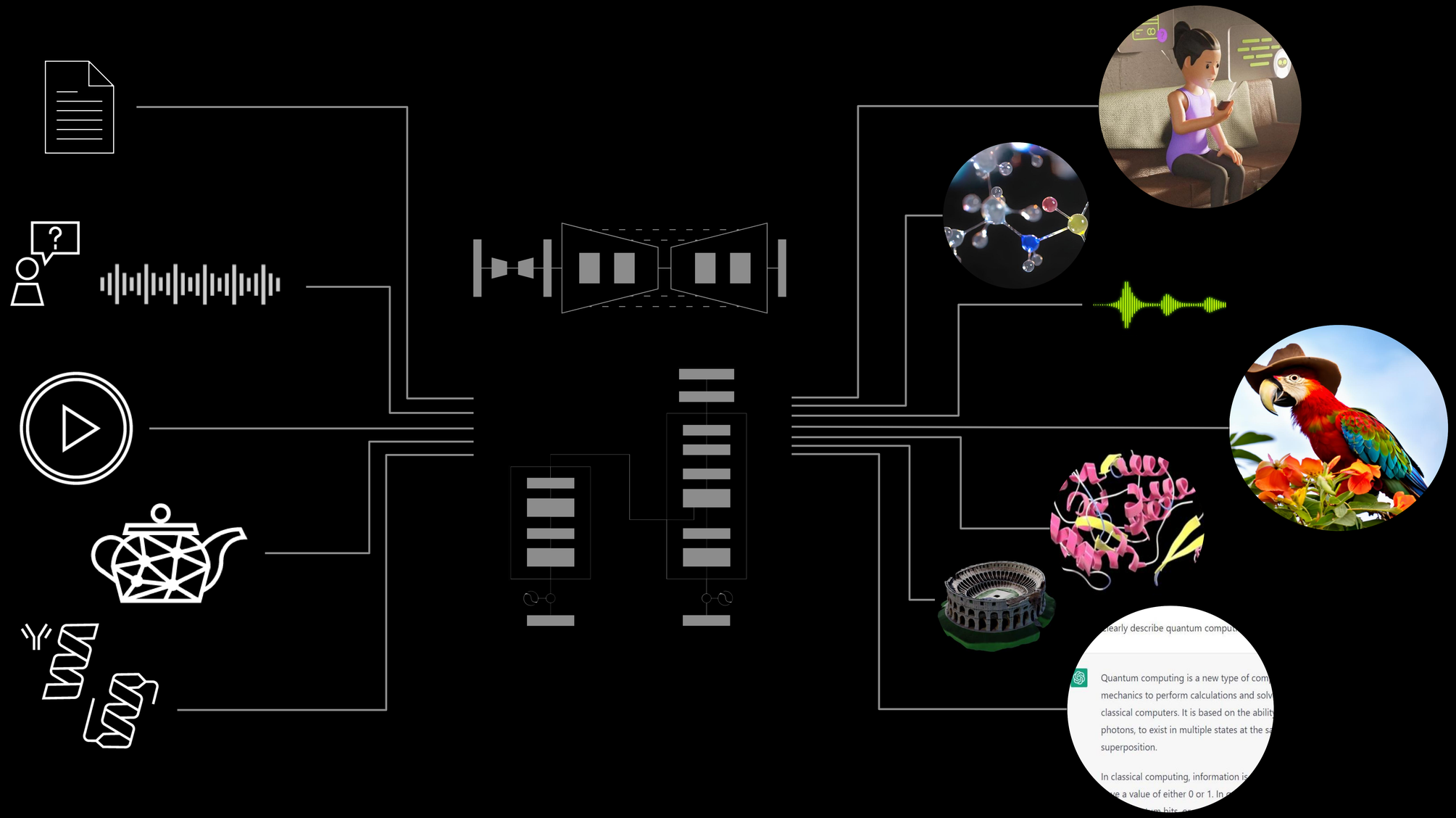
Generative AI Demystified

Amanda Saunders | NVIDIA | GTC 2024

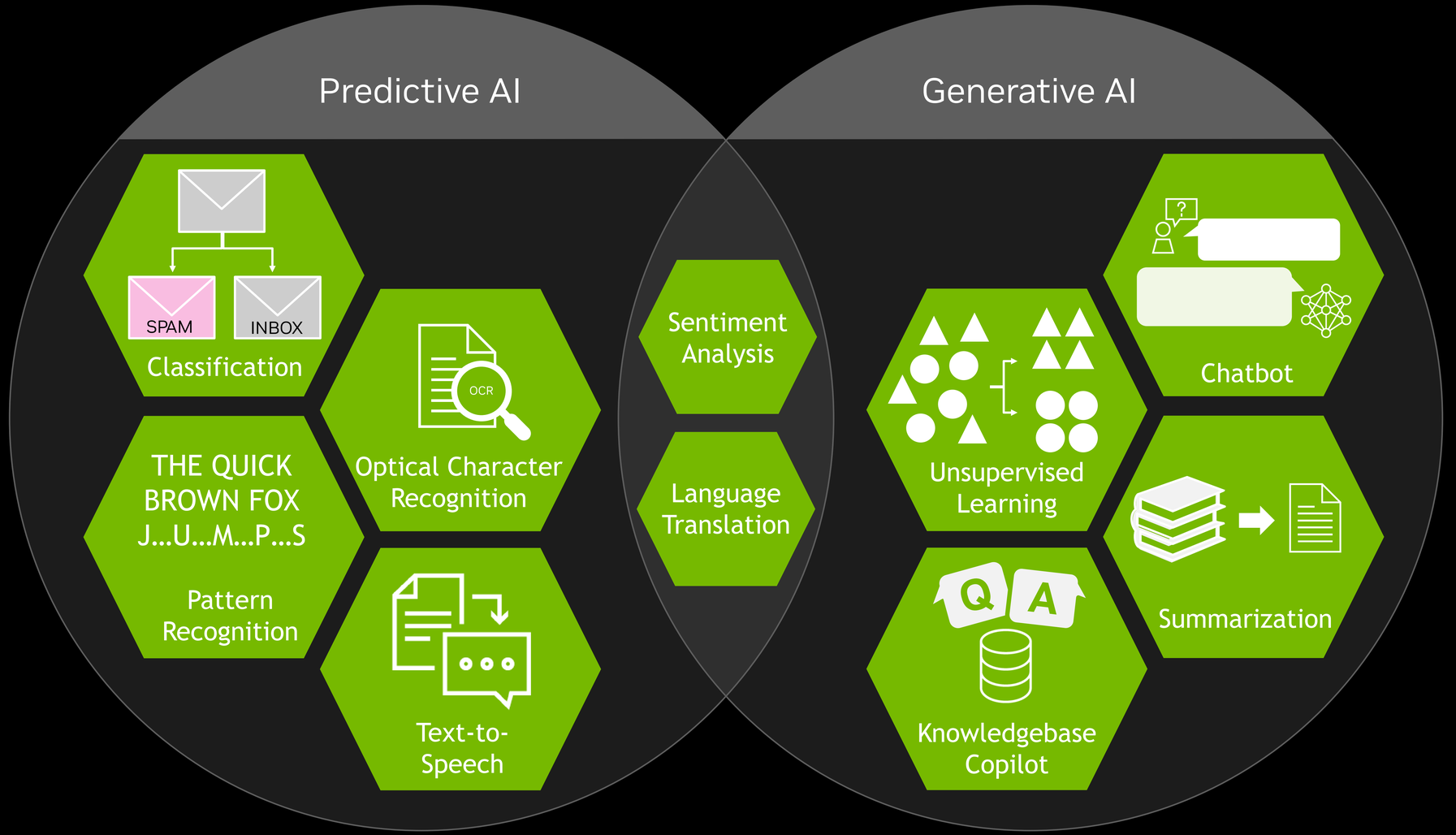
Where do I begin?



What is Generative AI?



When to Use Generative AI to Solve Challenges



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

Generative AI Journey

Explosion

ChatGPT gets announced late in 2022, gaining over 100 million users in just two months. Users of all levels can experience AI and feel the benefits firsthand.

Experimentation

Enterprise application developers kick off POCs for generative AI applications with API services and open models including Llama 2, Mistral, NVIDIA, and others.

Production

Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

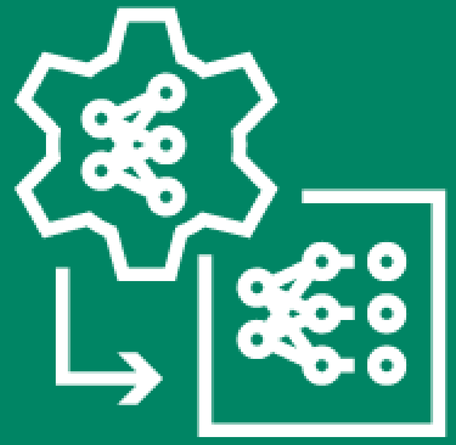
Generative AI Is a Spectrum

Less Customization

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

Moderate Customization

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

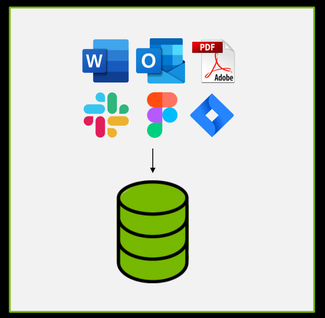


Extensive Customization

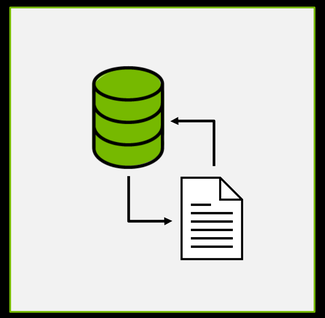
Custom foundation models or extensive finetuning
$10M+ for infrastructure and resources
6+ months for development

Building Generative AI Applications

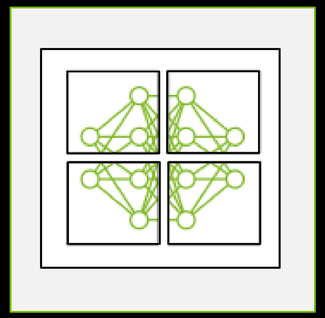
Data Acquisition



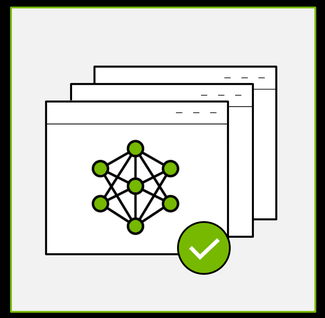
Data Curation



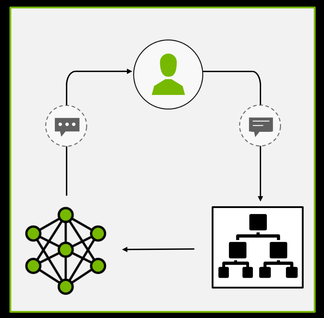
Pre-Training



Model Customization



Information Retrieval



Inference

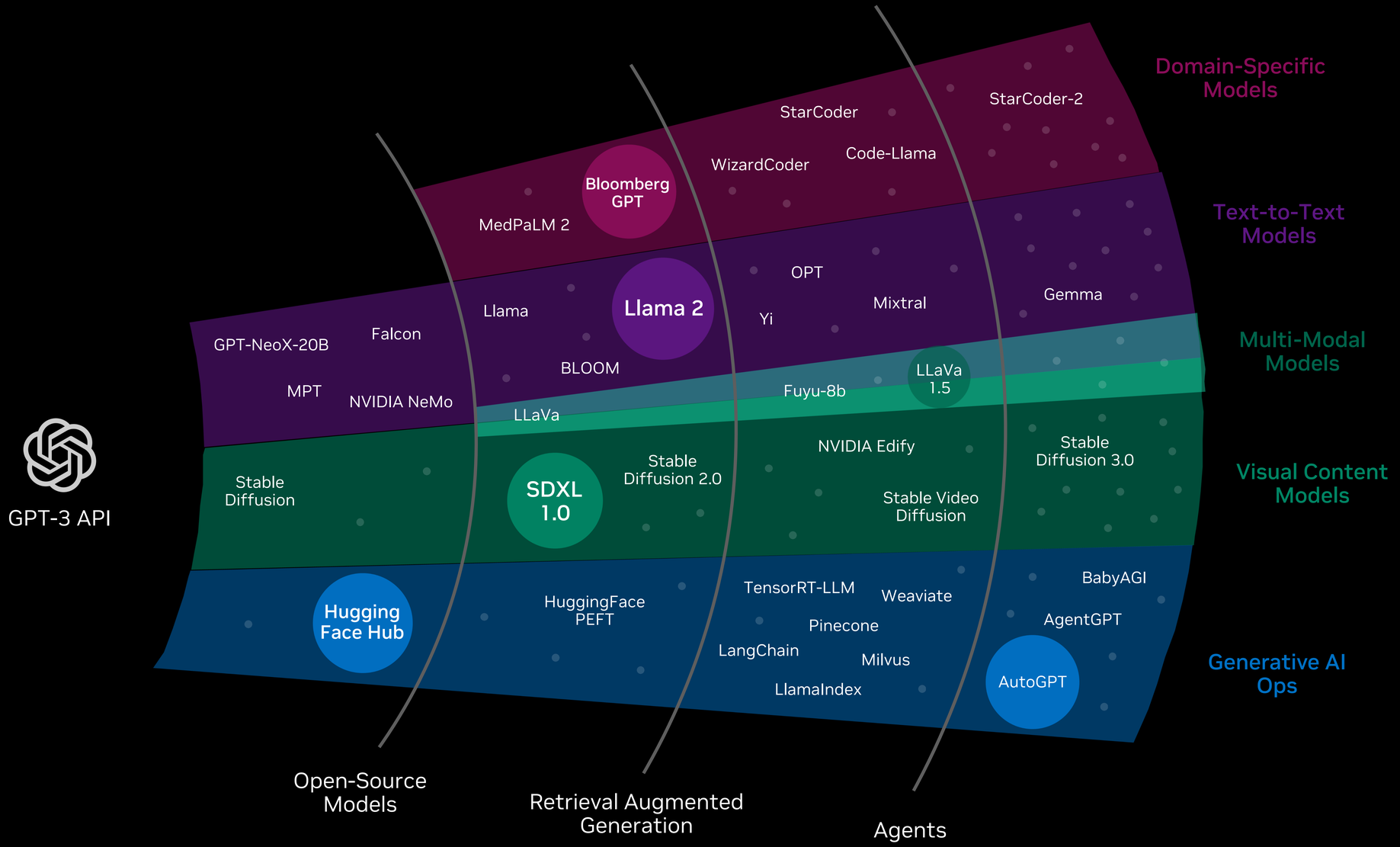
Guardrails

Data Prep

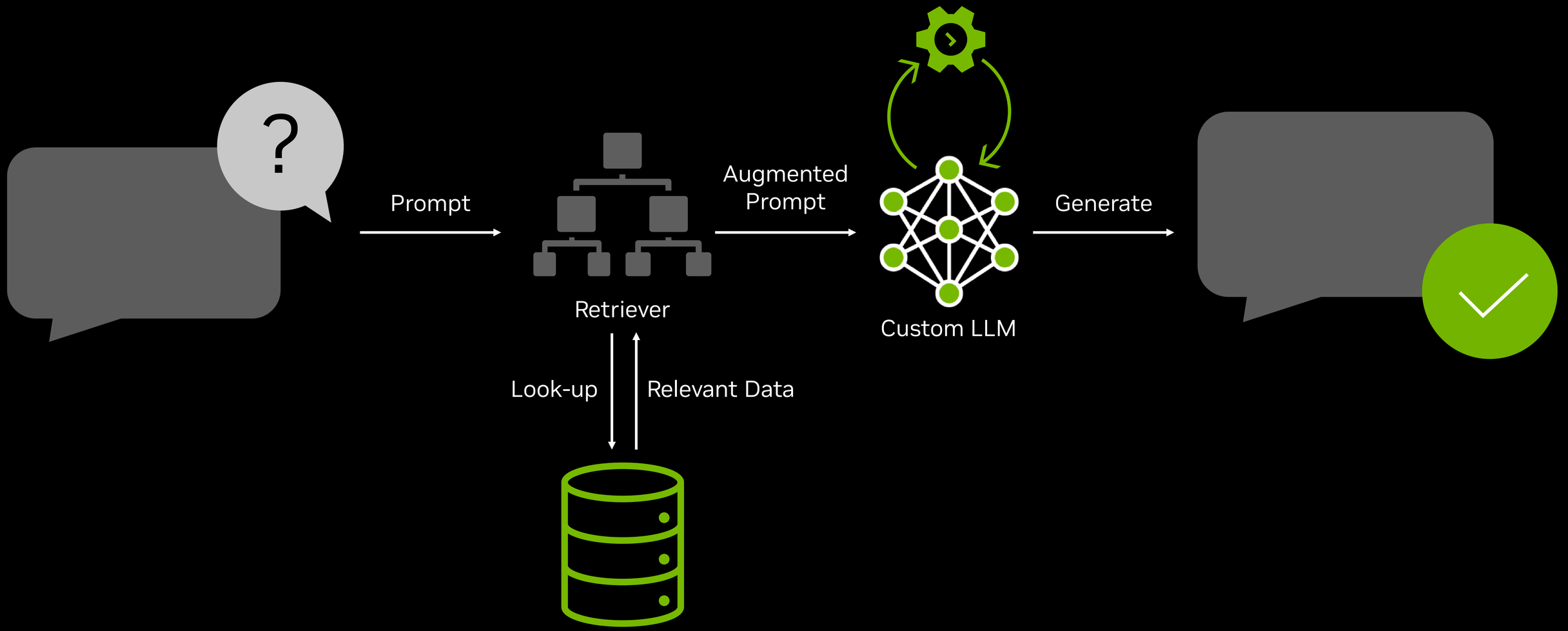
Training and Customization

Deployment

Explosion of Generative AI Models



Connecting LLMs to Enterprise Data with Retrieval Augmented Generation



Generative AI Adoption Across Industries

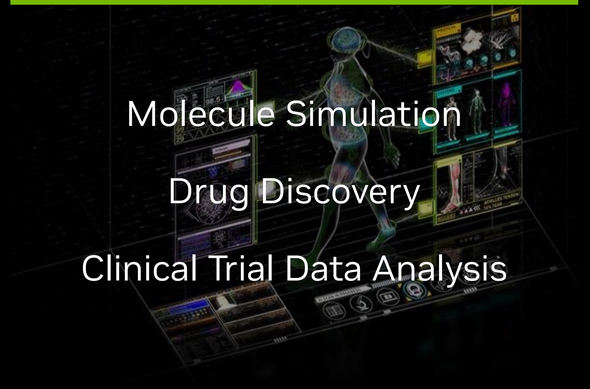
Finance

Fraud Detection

Personalized Banking

Investment Insights

Healthcare

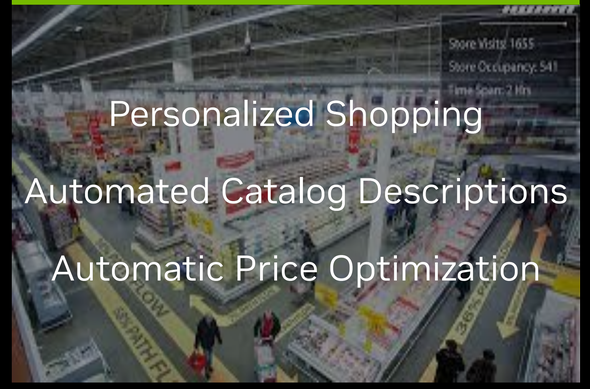


Molecule Simulation

Drug Discovery

Clinical Trial Data Analysis

Retail



Personalized Shopping

Automated Catalog Descriptions

Automatic Price Optimization

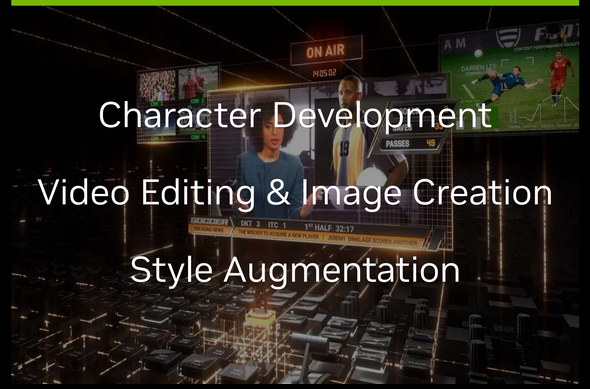
Telecommunications

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

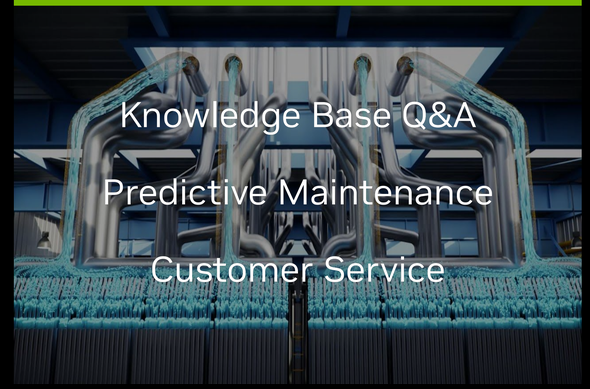
Federal

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.

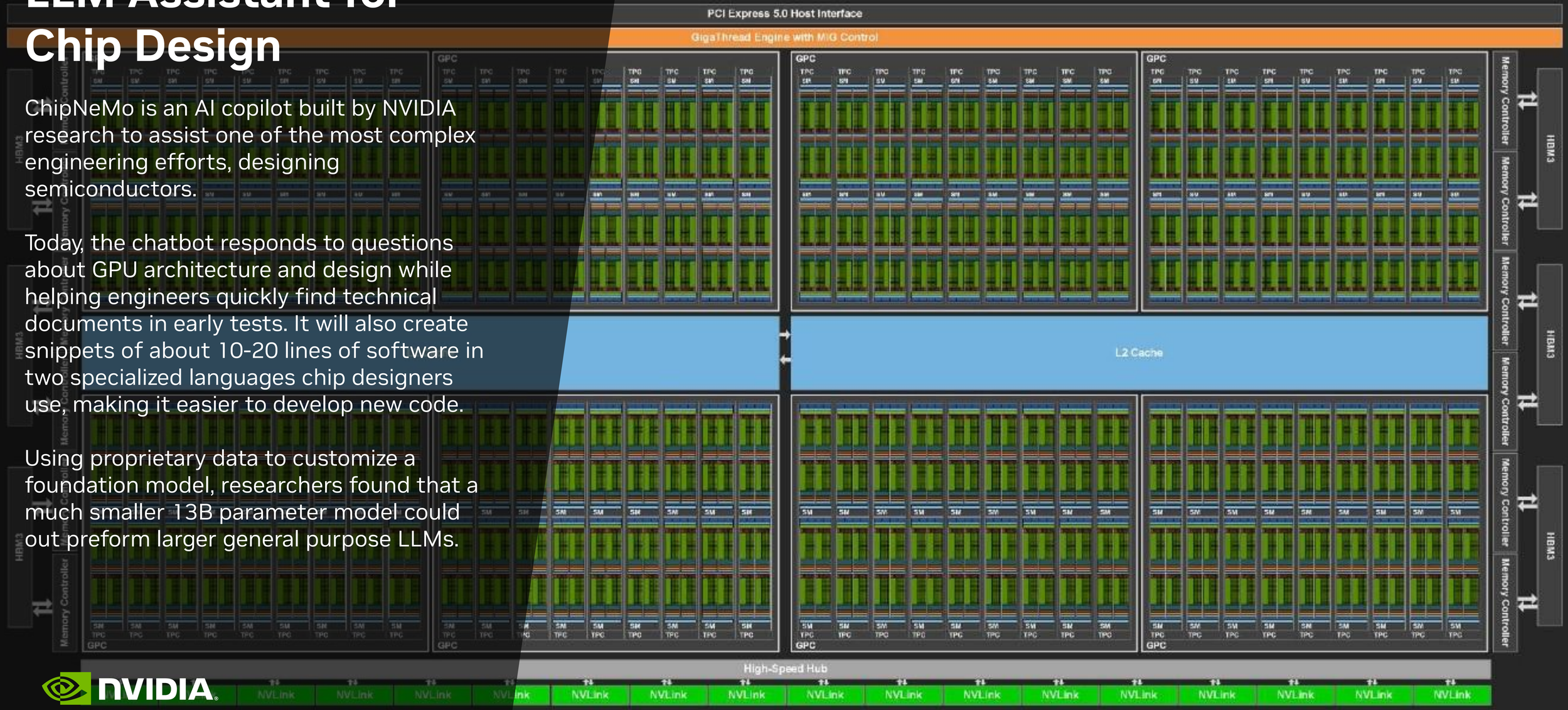
servicenow NVIDIA

LLM Assistant for Chip Design

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

Today, the chatbot 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 out preform larger general purpose LLMs.



NVIDIA


Hippocratic AI
Do No Harm

Key Things to Remember About Generative AI

Key Things to Remember About Generative AI

1. Generative AI is a tool, it’s not the full solution

Key Things to Remember About Generative AI

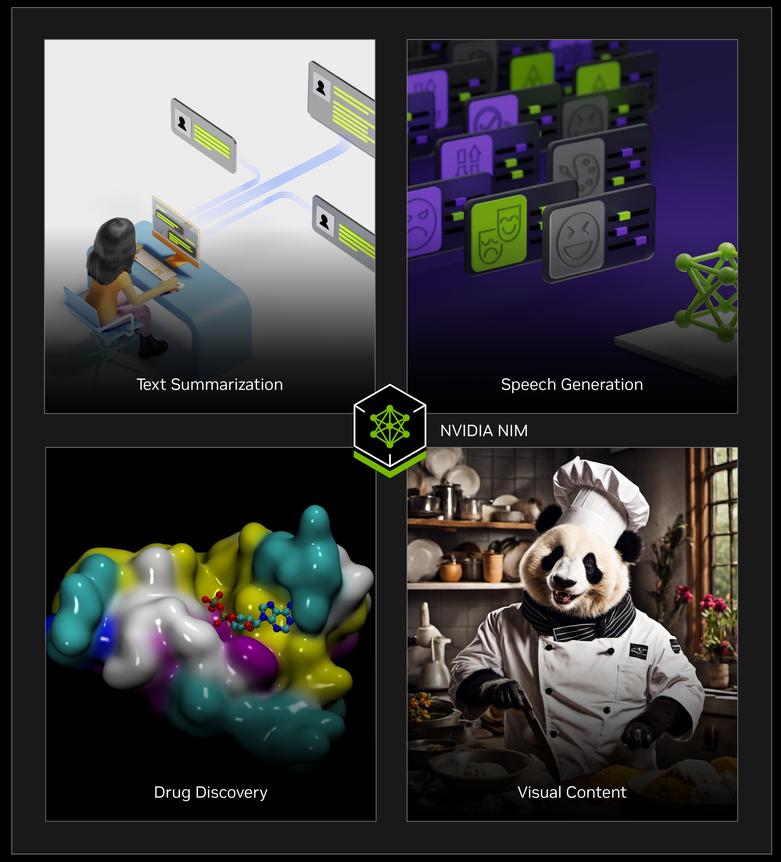
1. Generative AI is a tool, it's not the full solution

2. The space is moving quickly, build applications that have room for flexibility and adaptability

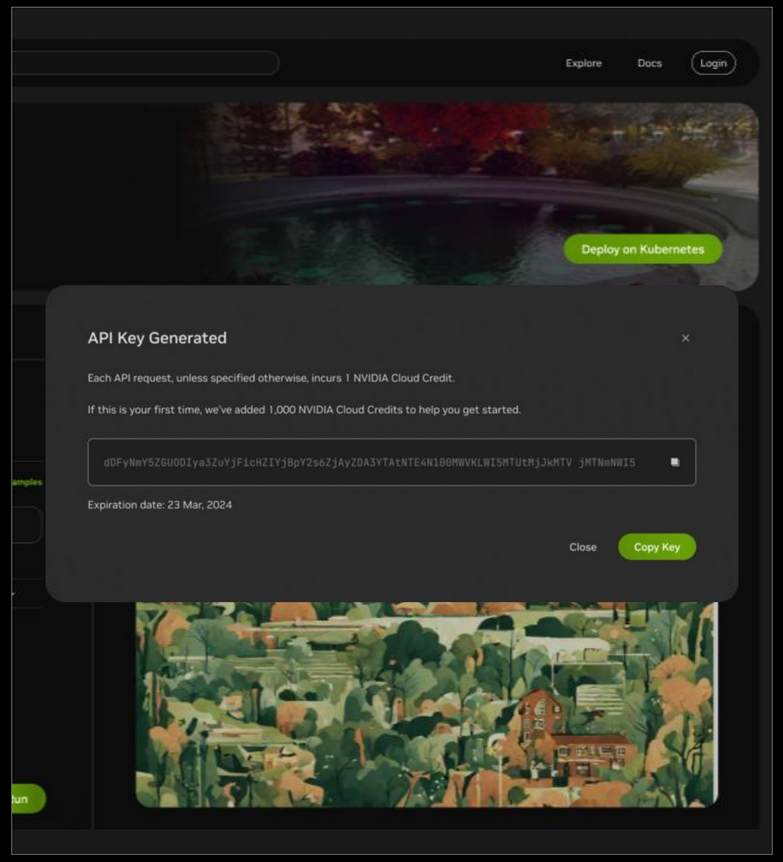
Key Things to Remember About Generative AI

1. Generative AI is a tool, it's not the full solution
2. The space is moving quickly, build applications that have room for flexibility and adaptability
3. The future is bright, buy some shades

ai.nvidia.com



Experience Models



Prototype with APIs

Deploy with NIMs


NVIDIA