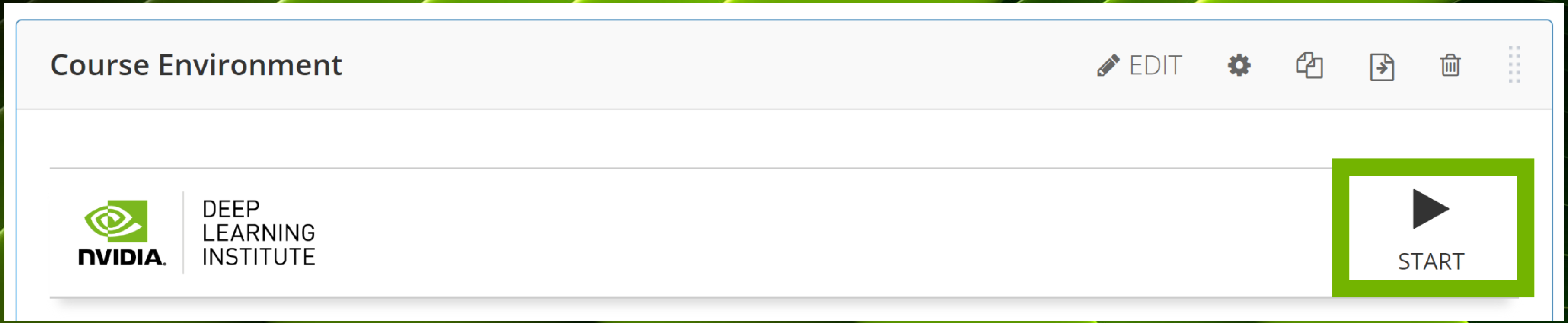
Building LLM Applications with Prompt Engineering

Introduction

Building LLM Applications with Prompt Engineering

Introduction



Large Language Models

Artificial Intelligence system designed to understand, generate, and process human language

2022

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.

2023

Experimentation

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

TODAY

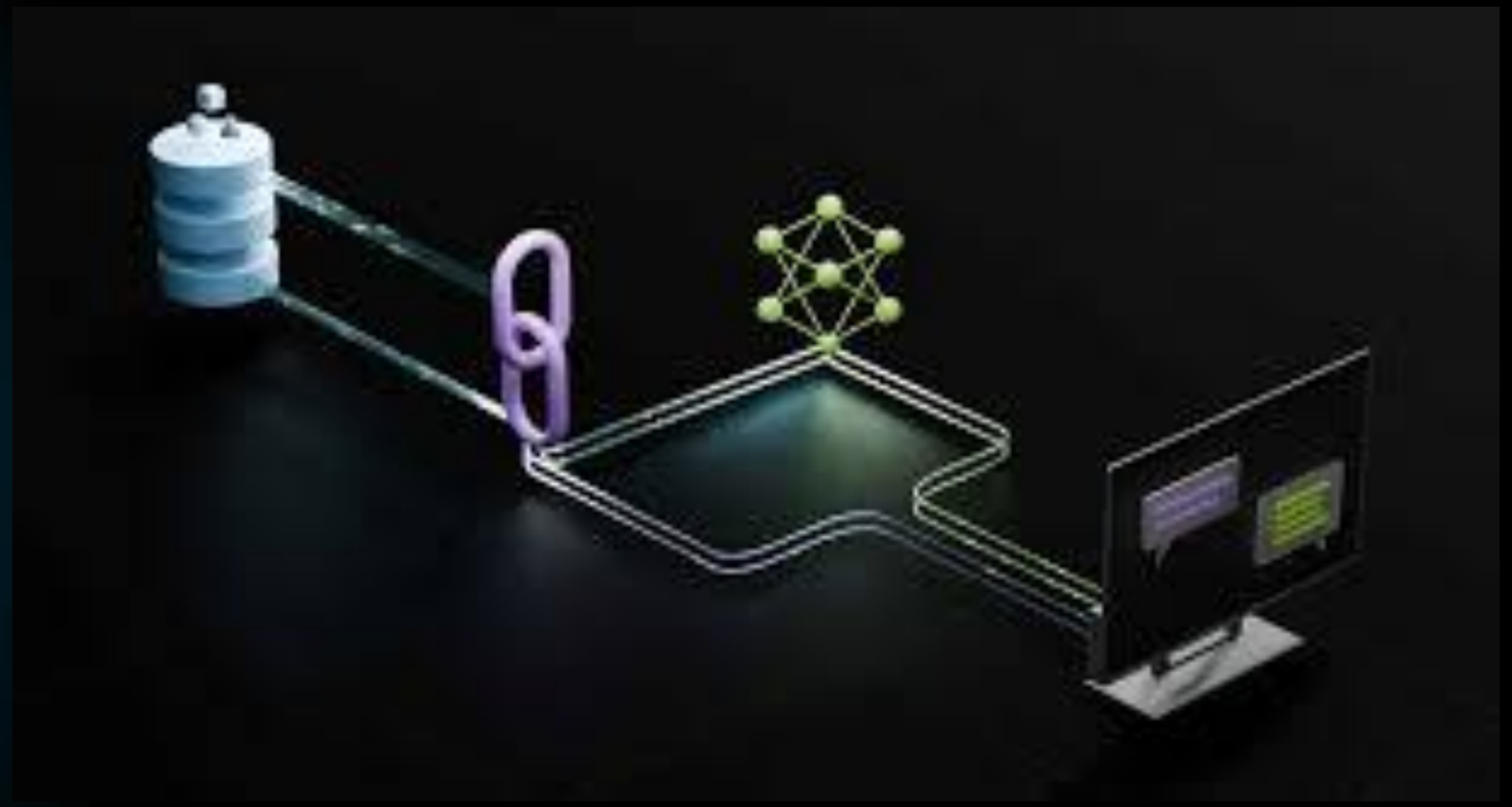
Production

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

3 Methods for Interacting with LLMs

Prompt Engineering

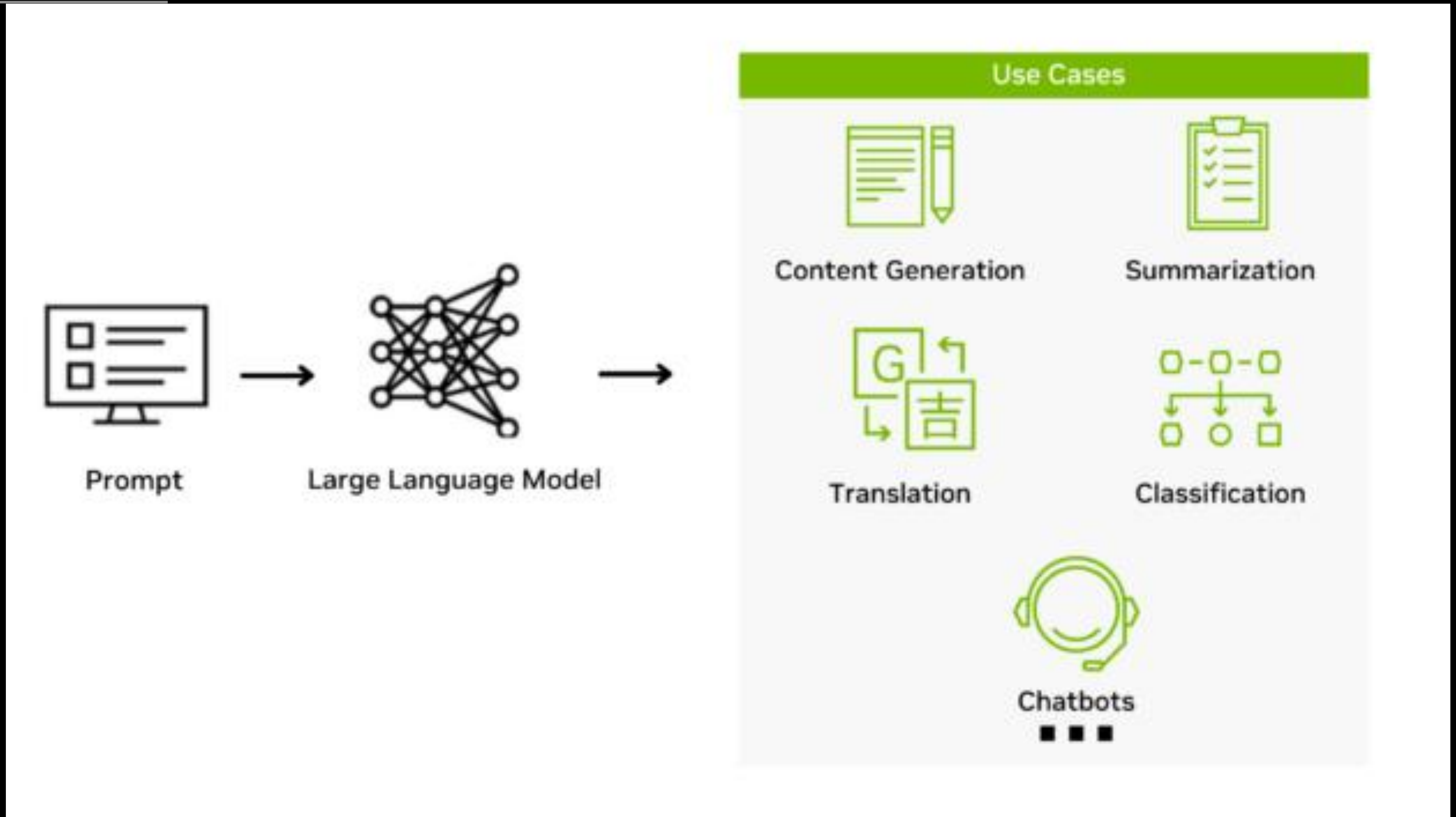
Retrieval Augmented Generation



(Parameter Efficient) Fine-tuning

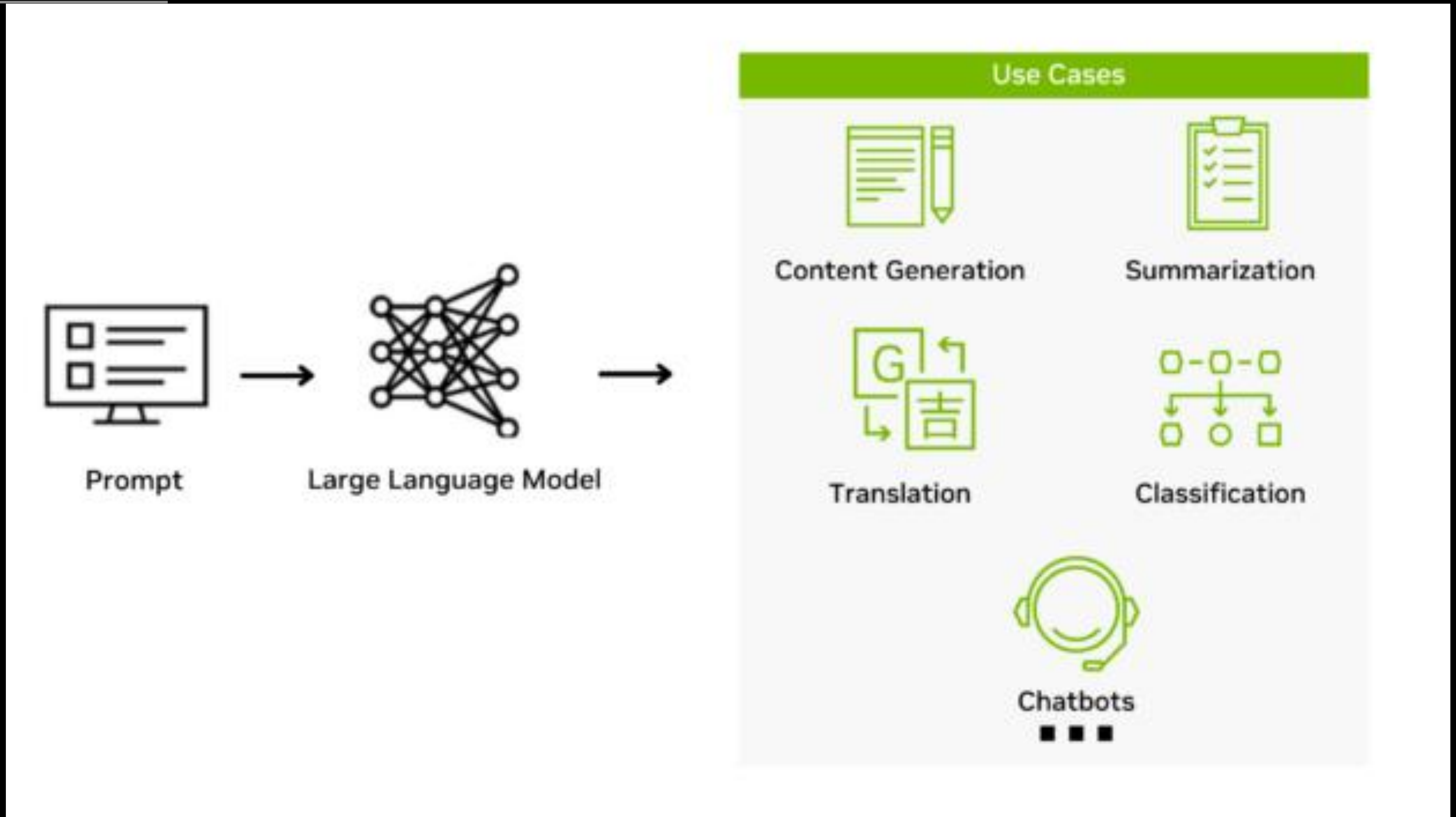
Goal of this Course

- Learn to interact programmatically with Chat-variant LLMs
- Be capable of using LLMs for a wide variety of application use cases
- Become fluent with fundamental LangChain techniques
- Develop good habits around iterative prompt engineering



Goal of this Course

- Introduction to Large Language Models (slides)
- NVIDIA NIM (slides)
- Intro to Prompting (interactive)
- LCEL Chains (interactive)
- PE techniques w/ messages (interactive)
- Chatbots (interactive)
- Structured data and document tagging (interactive)
- Tools and Agents (interactive)



Course Content

- Introduction to Large Language Models (slides)
- NVIDIA NIM (slides)
- Intro to Prompting (interactive)
- LCEL Chains (interactive)
- PE techniques w/ messages (interactive)
- Chatbots (interactive)
- Structured data and document tagging (interactive)
- Tools and Agents (interactive)

Prerequisites

Building LLM Applications with Prompt Engineering

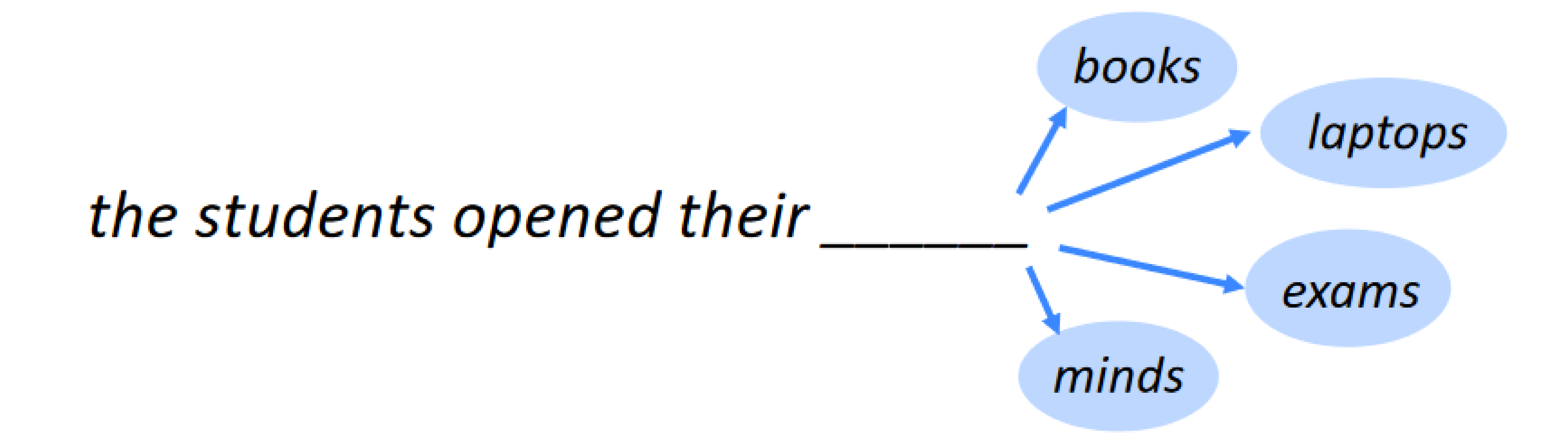
- Basic Python Proficiency
- Prior LLM Exposure

- Plus: Experience with LangChain or Knowledge in Natural Language Processing

Building LLM Applications with Prompt Engineering

Introduction to Large Language Models

What is Language Modeling?



the students opened their ________

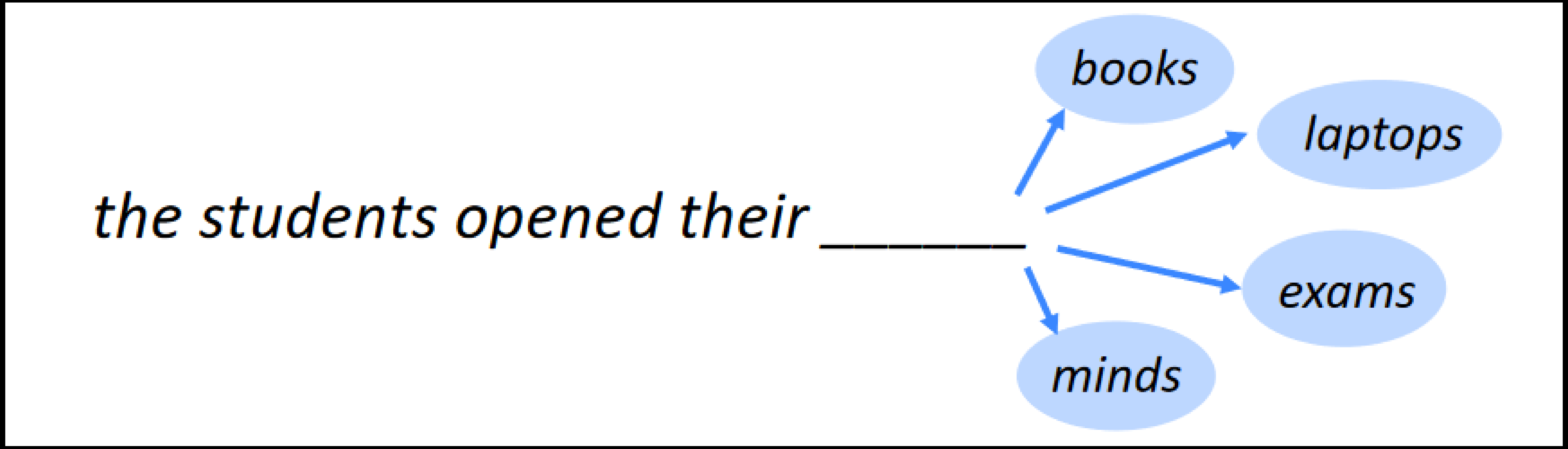
books

laptops

exams

minds

Large Language Models Pipeline



the students opened their _______

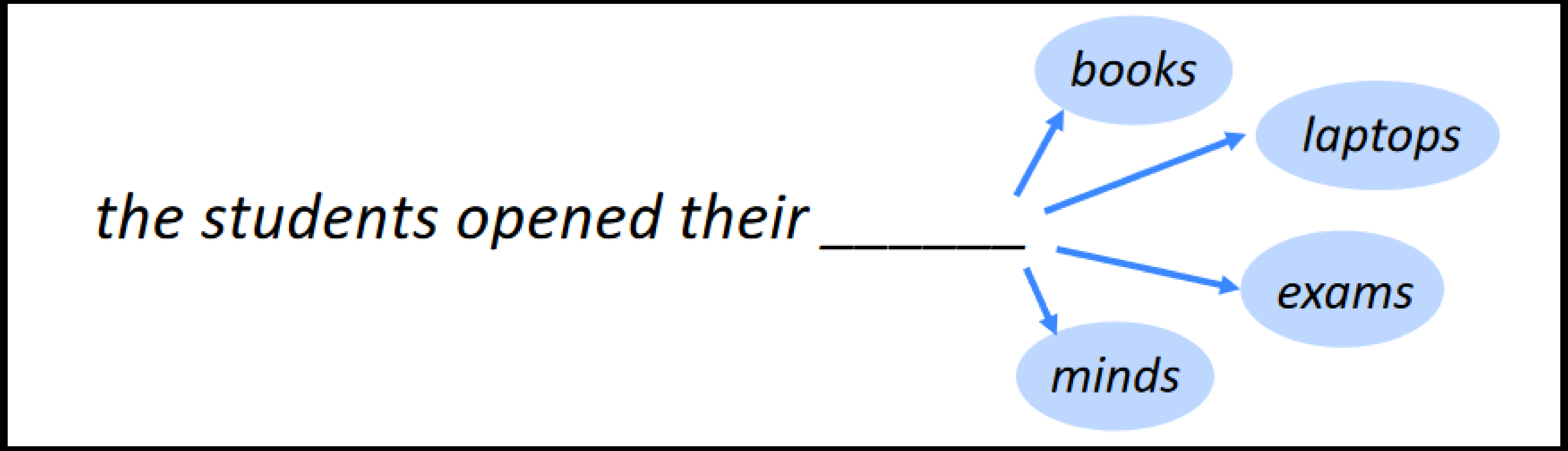
books

laptops

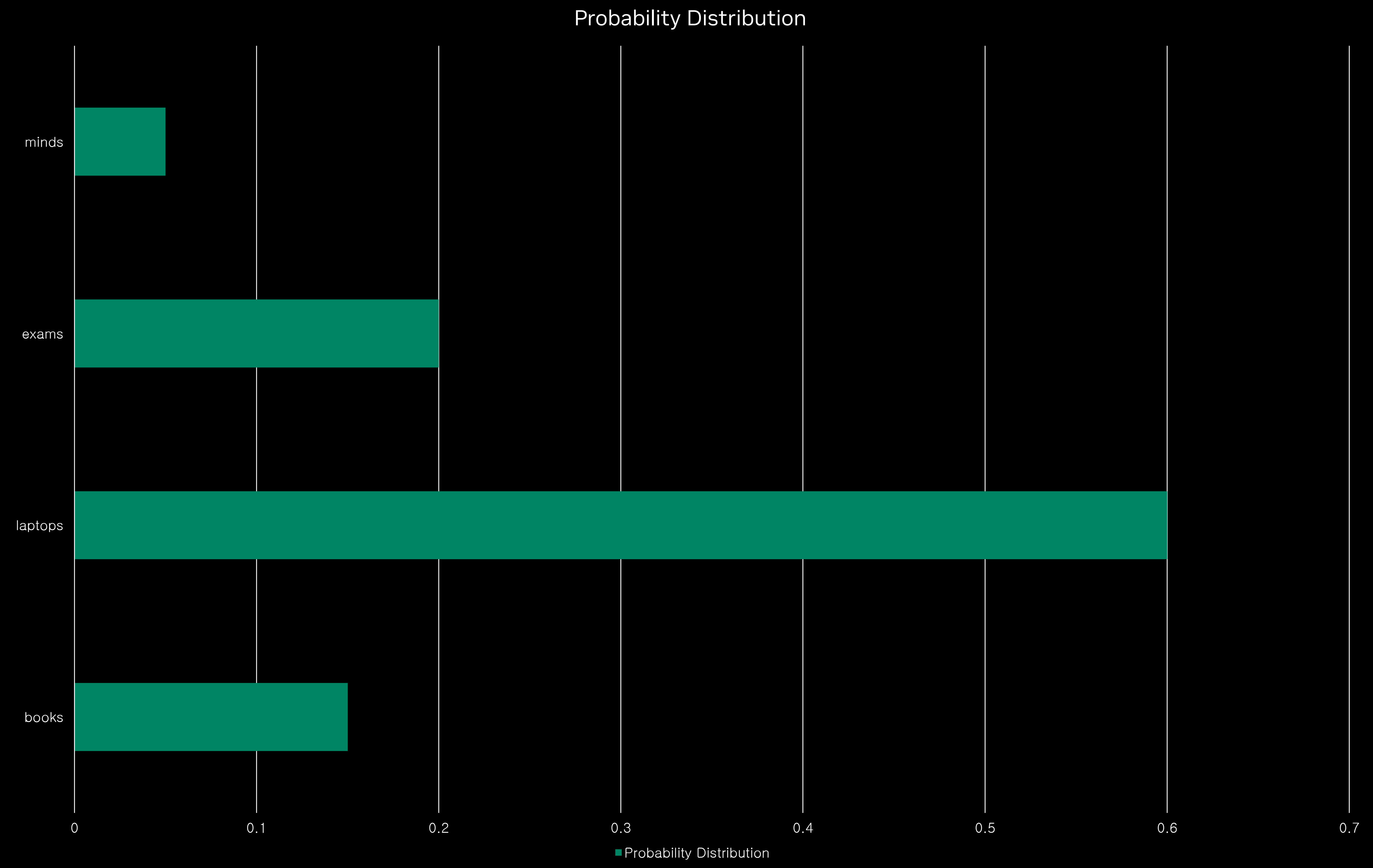
exams

minds

Large Language Models



the students opened their _______



Training LLMs: Tokenization

Hello! How is it going?!

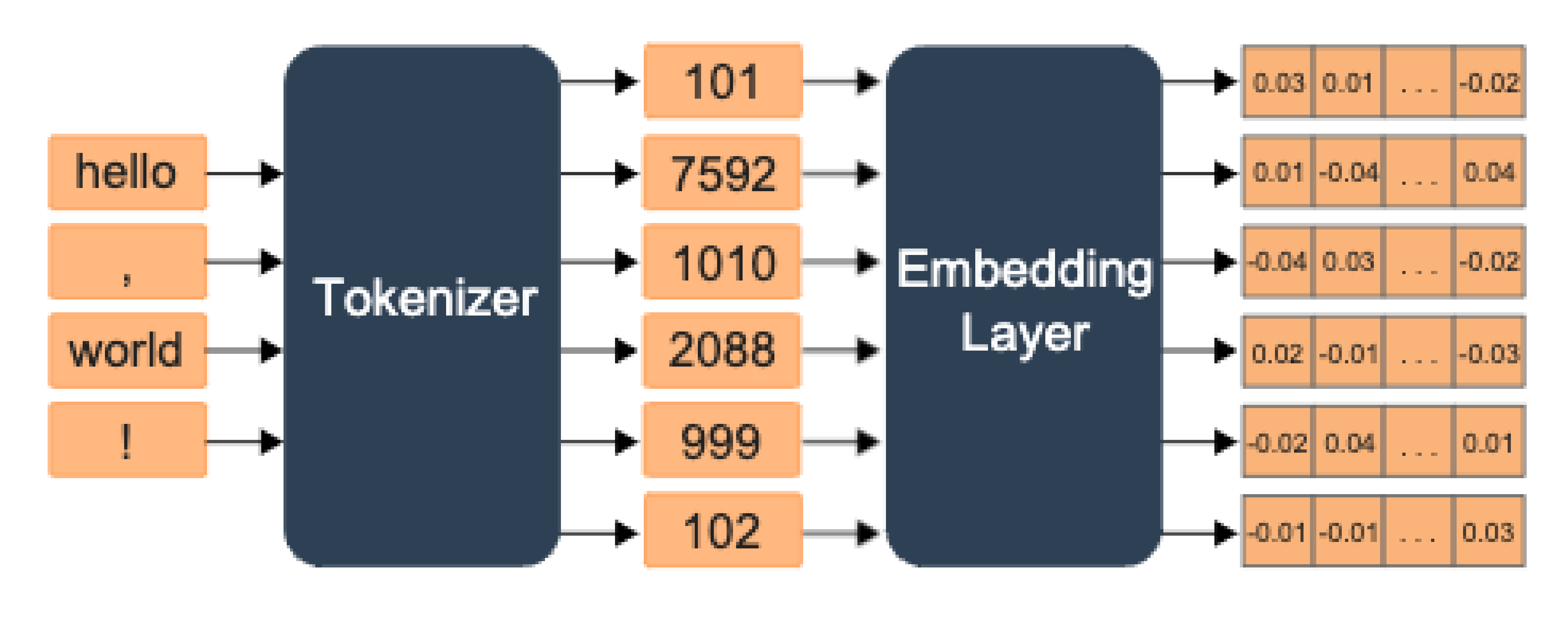
Sequence of Letters

Tokenization

Hello! How is it going?!

Ordered Sequence of Classes

Training LLMs: Embedding



hello

,

world

!

Tokenizer

101

7592

1010

2088

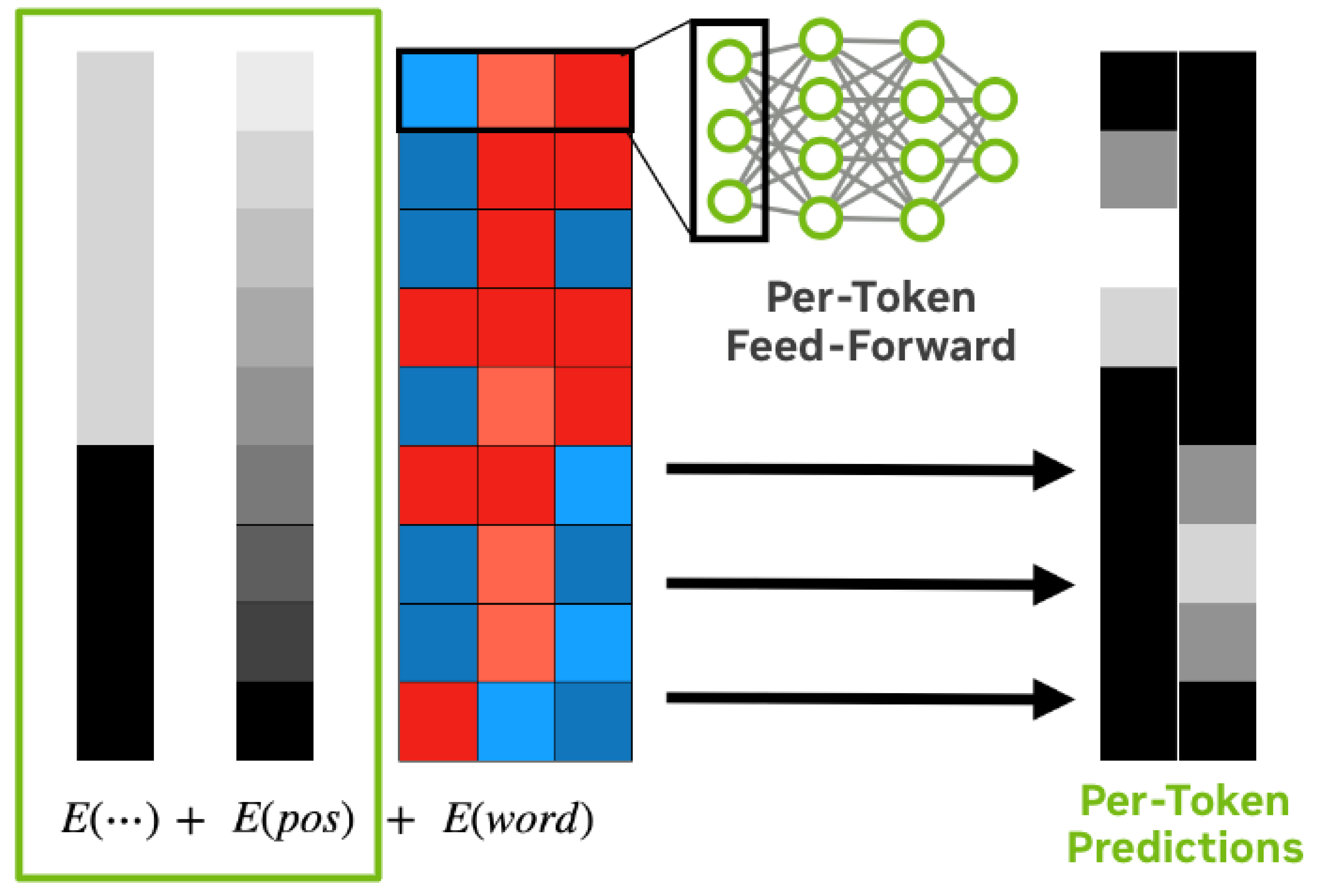
999

102

Embedding Layer

0.03	0.01	. . .	-0.02
0.01	-0.04	. . .	0.04
-0.04	0.03	. . .	-0.02
0.02	-0.01	. . .	-0.03
-0.02	0.04	. . .	0.01
-0.01	-0.01	. . .	0.03

Training LLMs: Embedding



Large Language Models

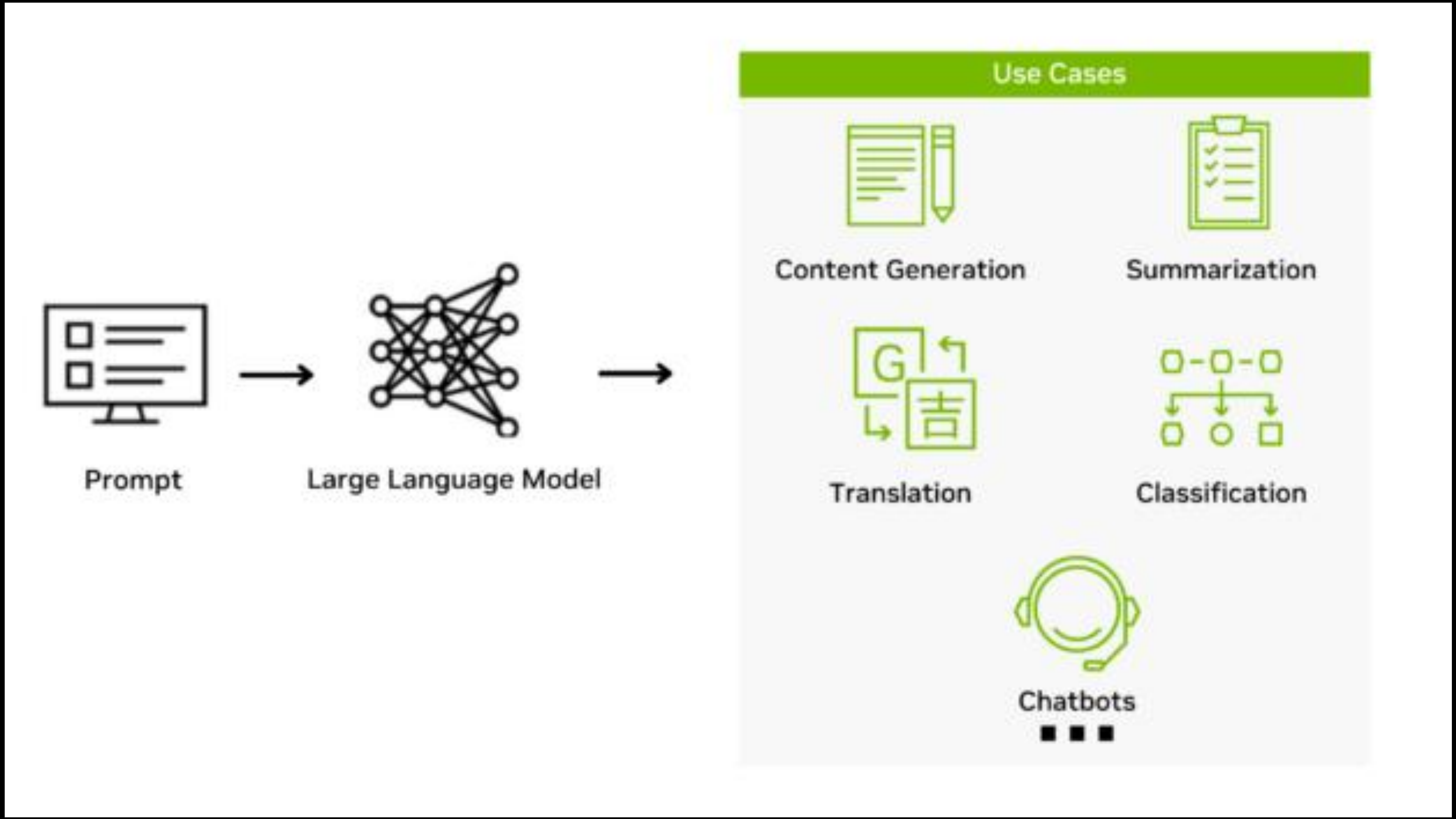
Next-token prediction task

- Massively Multi-task learner capable of doing all sorts of tasks

Hard to control the text generation

Task	Example sentence in pre-training that would teach that task
Grammar	In my free time, I like to {code, apples}
Lexical Semantic	I went to the store to buy papaya, dragon fruit, and {pineapples, croissant}
World Knowledge	The capital of South Korea is {Seoul, Paris}
Sentiment Analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {bad, good}
Math Question	Arithmetic exam answer key: 3 + 8 + 4 = {15, 384}
Translation	The word for pretty in Spanish is {bonita, hola}

Prompt Engineering



Prompt Engineering is asking the right question to the LLM depending on your use case.

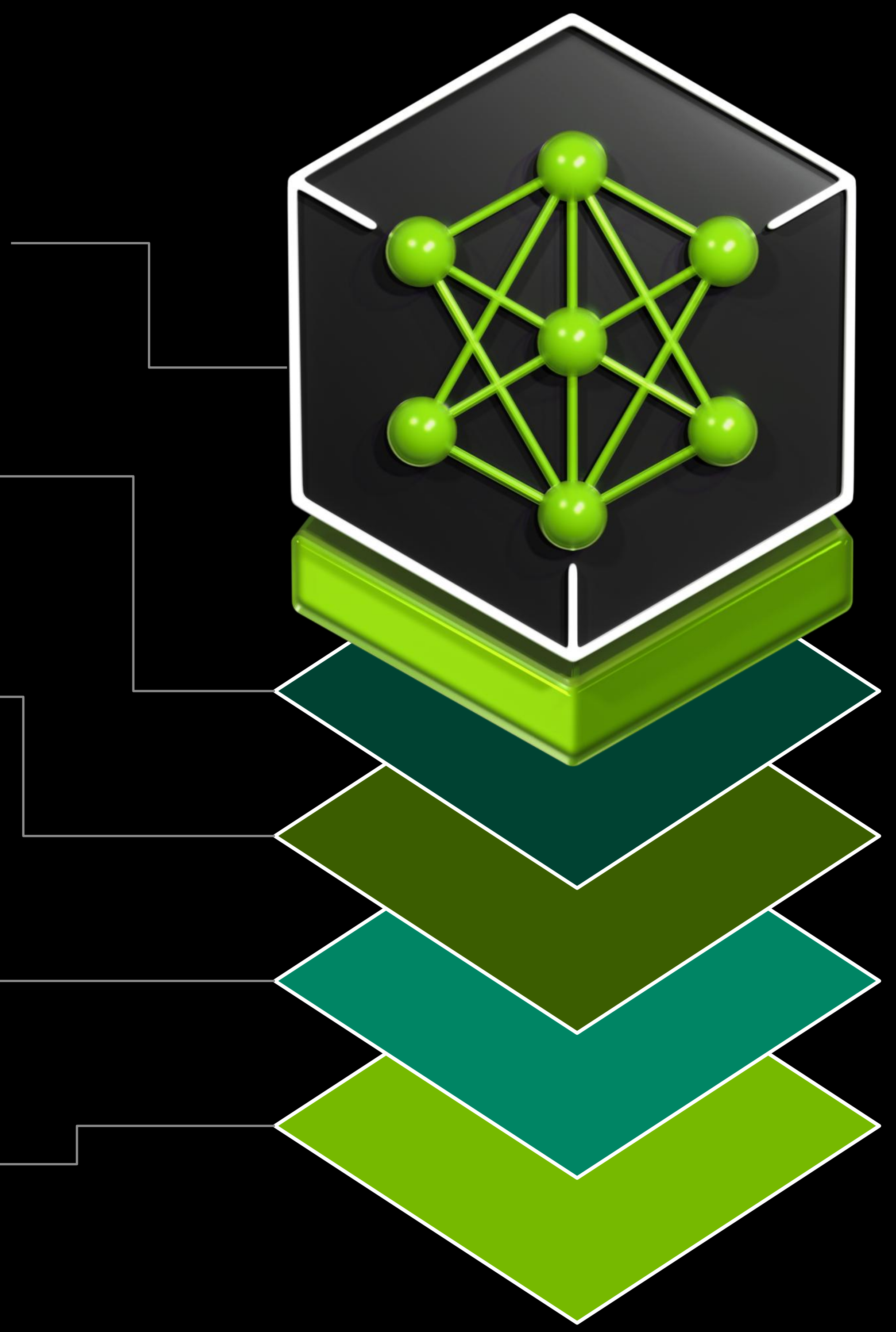
Building LLM Applications with Prompt Engineering

NVIDIA NIM

NVIDIA NIM Optimized Inference Microservices

NVIDIA NIM is the fastest way to deploy AI models on accelerated infrastructure across cloud, data center, and PC

NVIDIA NIM



Prebuilt container and Helm chart

Industry standard APIs

Support for custom models

Domain specific code

Optimized inference engines

Deploy anywhere with security and control

Speed time to market

Empower developers

Optimize throughput

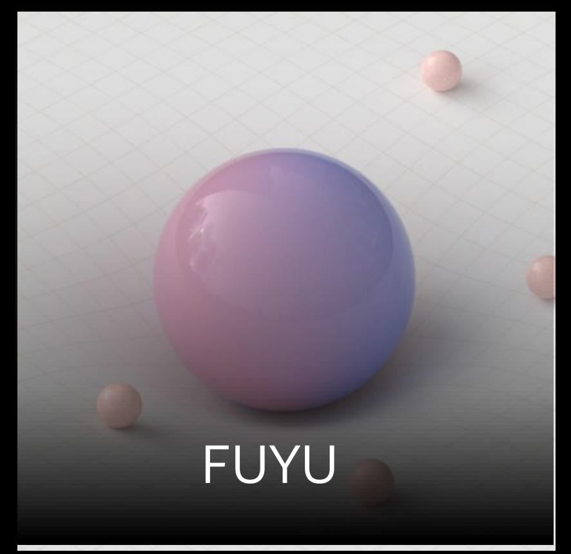
Boost accuracy

Deploy in production

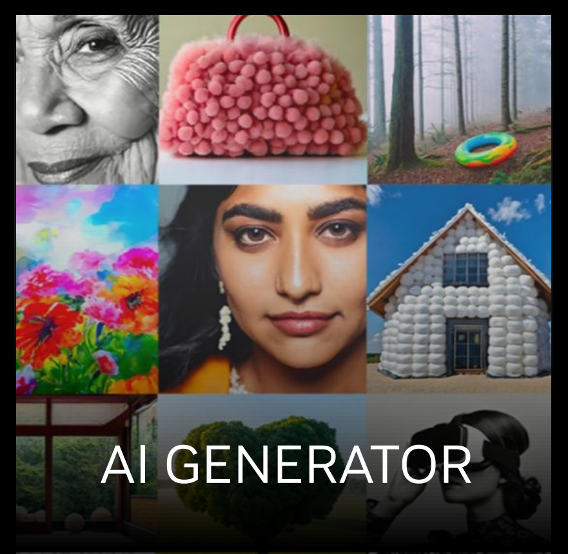
LLAMA 3.1

MIXTRAL 8x7B

GEMMA 7B

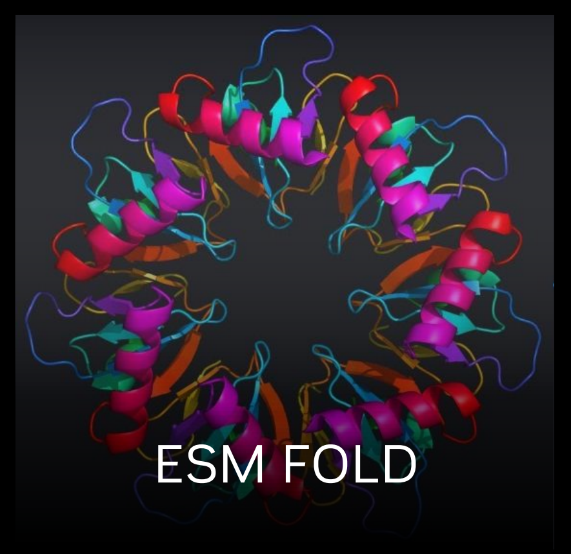

FUYU

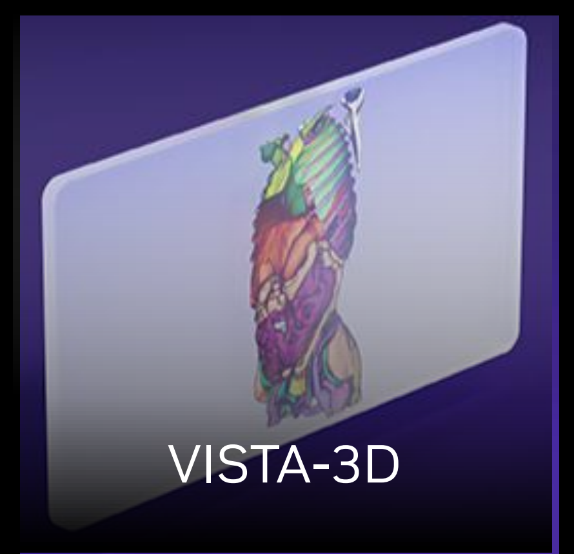
NEMO RETRIEVER

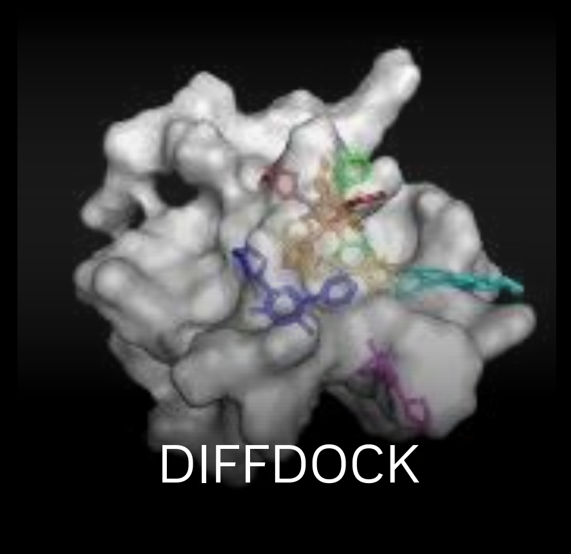

AI GENERATOR

3D GENERATOR

AUDIO2FACE


ESM FOLD

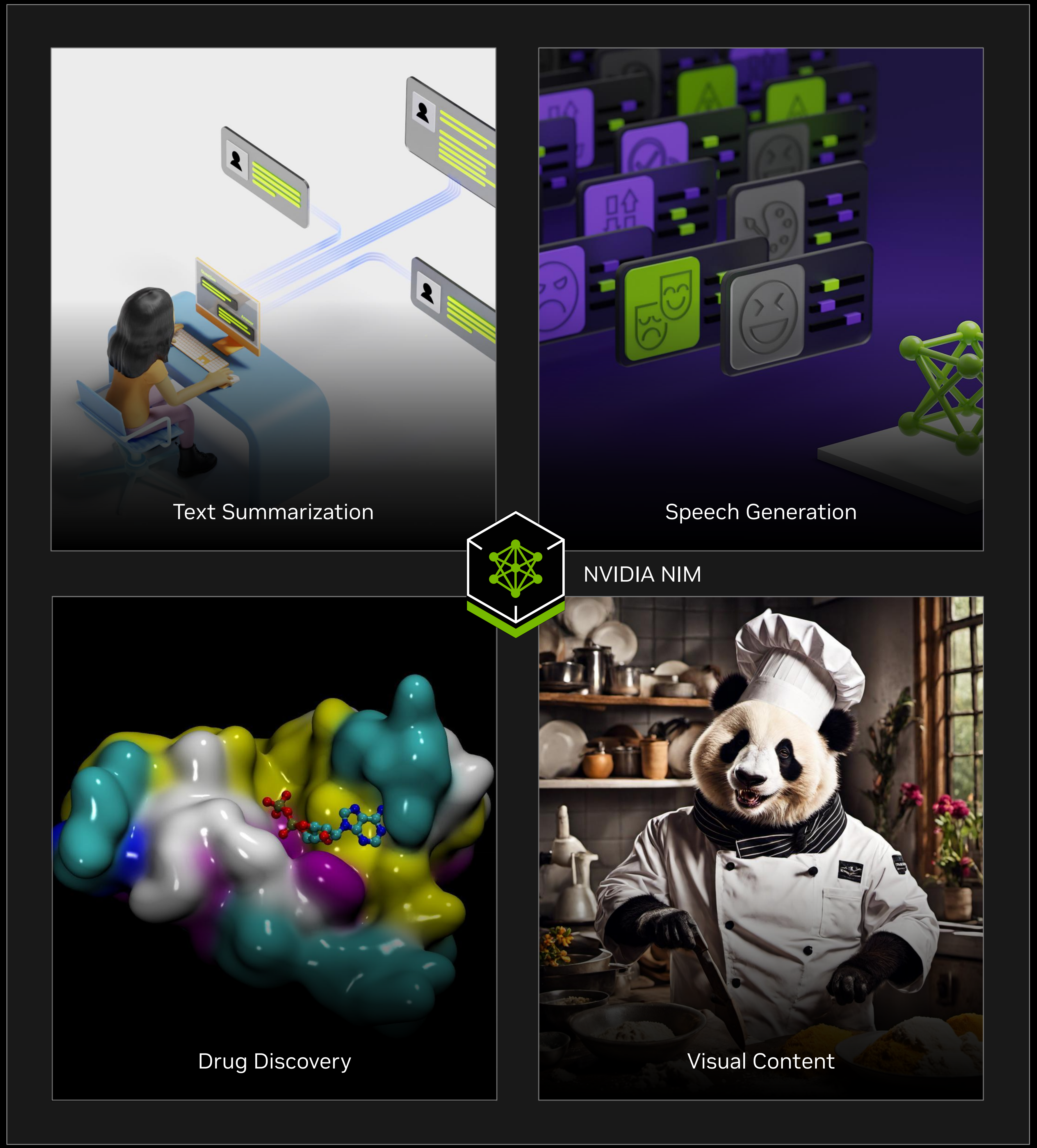

VISTA-3D

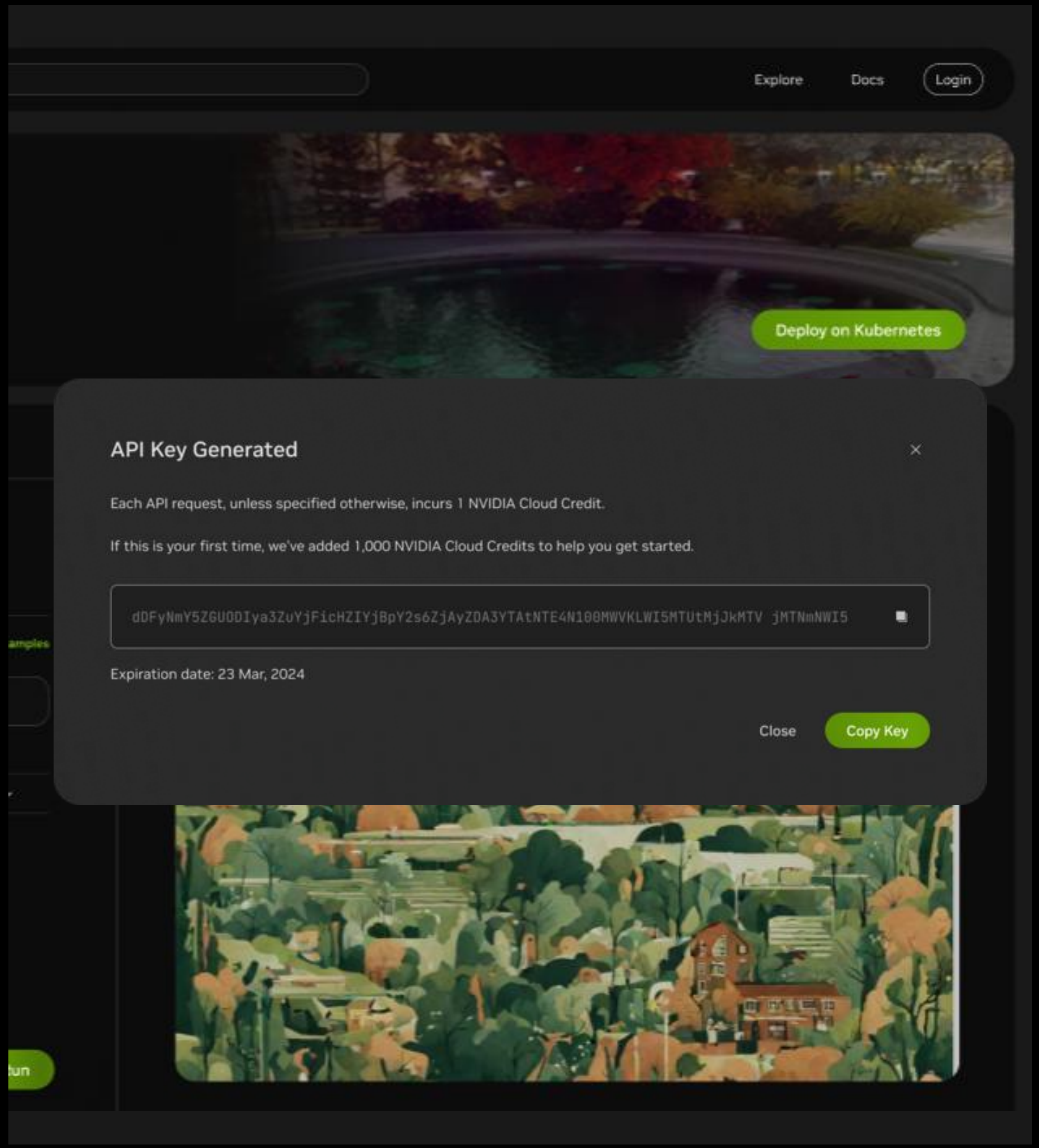

DIFFDOCK

MolMIM

Experience and Run Enterprise Generative AI Models Anywhere

Seamlessly integrate AI in business applications with NVIDIA AI APIs





API Key Generated

Each API request, unless specified otherwise, incurs 1 NVIDIA Cloud Credit.

If this is your first time, we've added 1,000 NVIDIA Cloud Credits to help you get started.

```
dDFyNmY5ZGUODIya3ZuYjFicHZIYjBpY2s6ZjAyZDA3YTAtNTE4N100MWVKLWI5MTUtMjJkMTV jMTNmNWI5
```

Expiration date: 23 Mar, 2024

- Experience Models
- Prototype with APIs
- Deploy with NIMs

NIM Deployment Cycle

Locally-hosted LLMs reduce latency as opposed to models hosted on a public server

Scalable Deployment

Ownership of your intellectual property

Building LLM Applications with Prompt Engineering