Practical AI: Definitions, Concepts, & Techniques

Lee Solomonson | December 2024
Director of Advanced Technologies
lee.solomonson@nth.com

WELCOME!

The webinar will begin shortly.
We are waiting for everyone to load.

Before We Begin...

- Microphones have been muted – mics will open at the end.
- Please use the Q&A function for questions you would like to see addressed.
- Slides will be provided.
- This is an educational event focused on the technical concepts and techniques around AI. Specific solution offerings will be introduced, with in-depth explanations offered at a later date.

Course Outline

Section 1
- AI/ML/DL
- Deep Learning Phases (Explore/Deploy/Apply)
- Predictive AI vs. Generative AI
- Functions and Use Cases

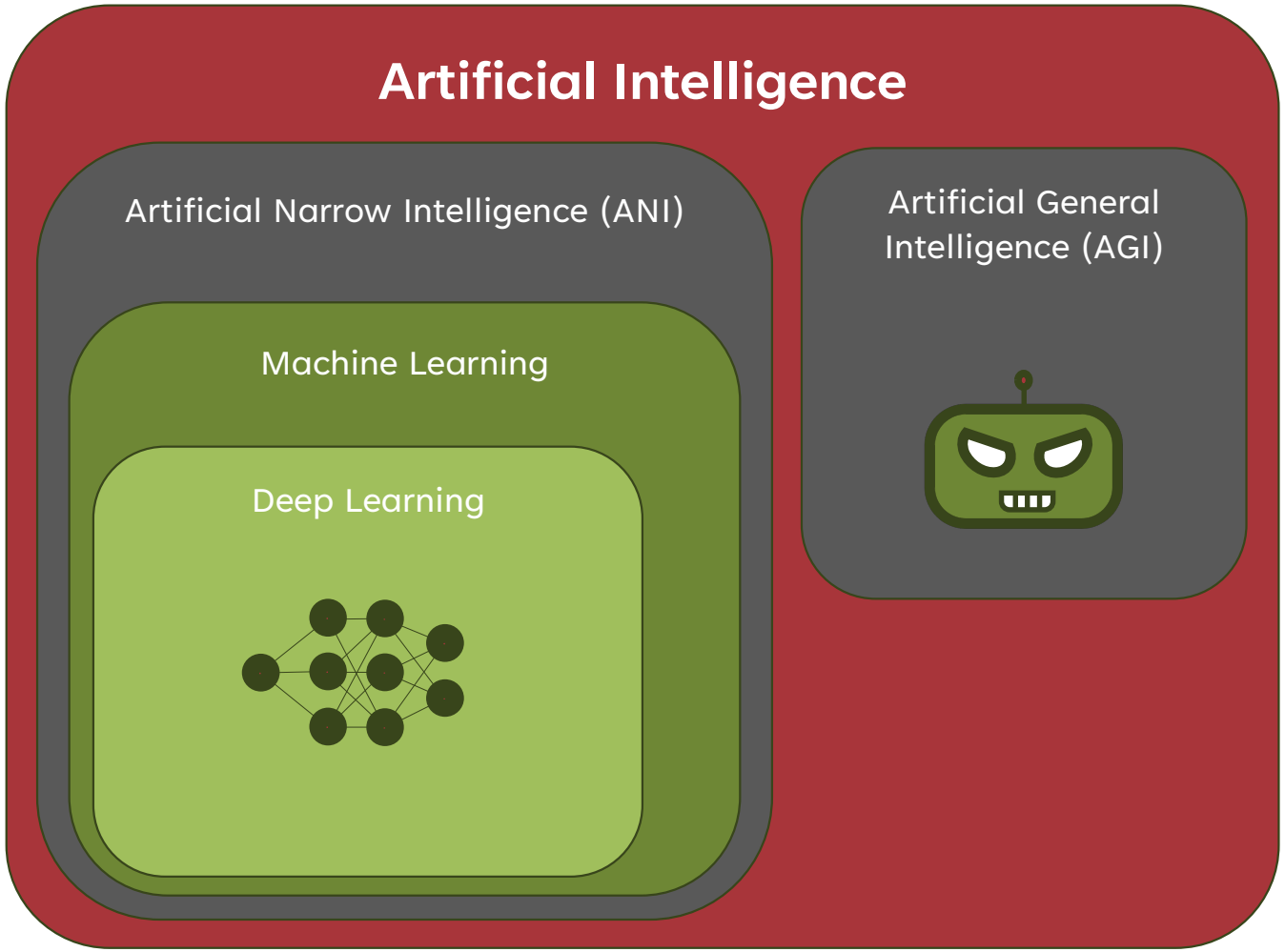
Section 2
- What is a Neural Network?
  - Networks and Architectures
  - Neurons, Weights, Bias, Activation
  - Matrix Multiplication
- Model Training Overview

Section 3
- Model Selection & Sources
  - Quantization

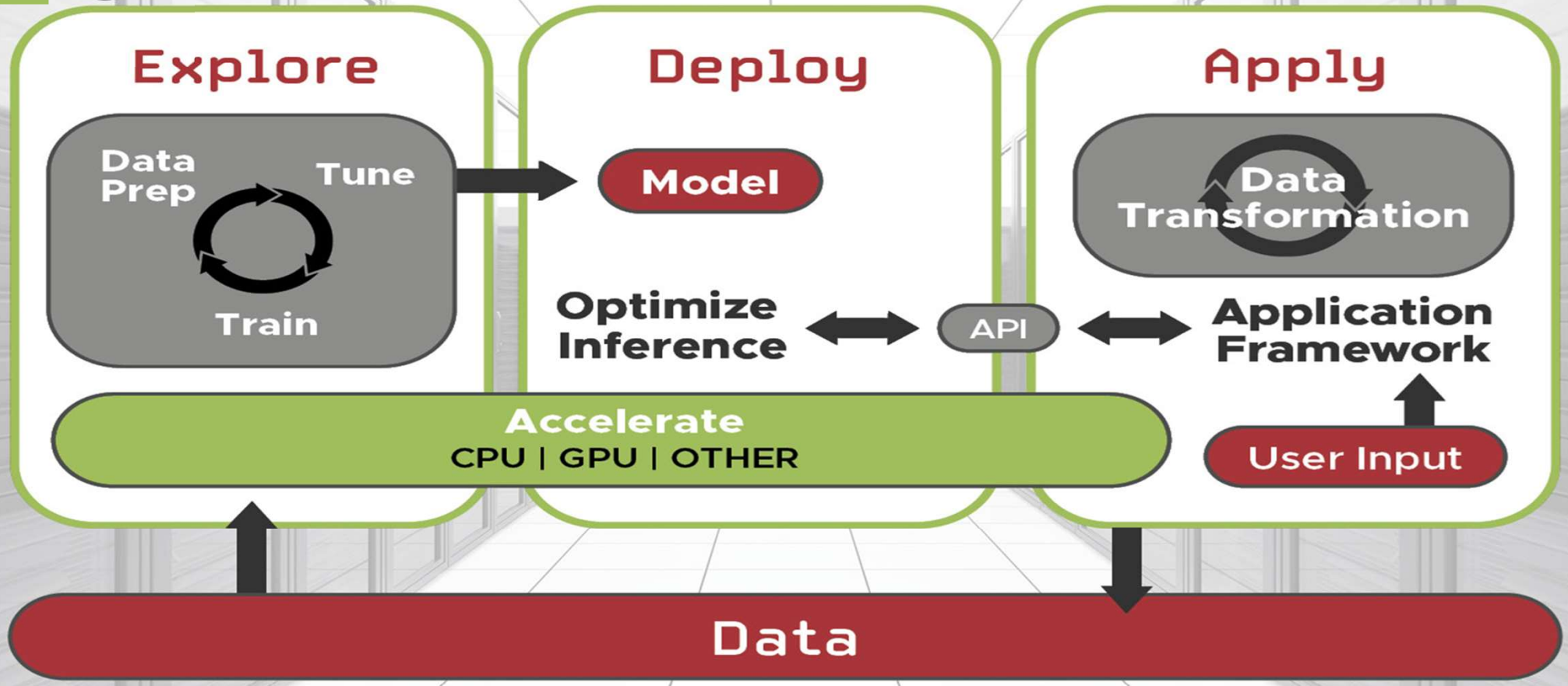
Section 4
- Model Deployment
- Applying Model
  - Prompt Engineering
  - Retrieval Augmented Generation (RAG)
  - Foundational Models
- Getting Started

Poll: What is your experience level?

Taxonomy

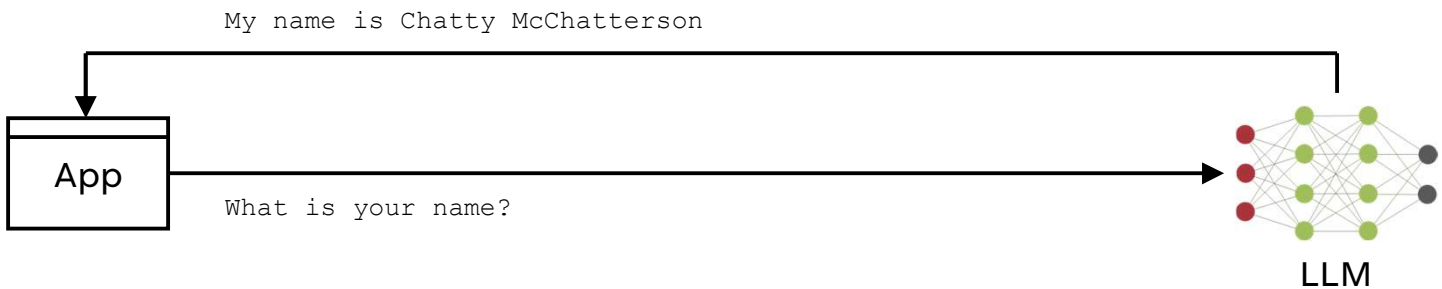

Artificial Intelligence
Artificial Narrow Intelligence (ANI)
Machine Learning
Deep Learning
Artificial General Intelligence (AGI)

AI Blueprint



Predictive AI

Generative AI



"You shall know a word by the company it keeps."
--J.R. Firth, 1957

It is windy outside.
The road is windy outside of town.

```
What is your name?

My Name is Chatty McChatterson

what is 397652 X 1.5342?
```

```
[3923, 374, 701, 836, 1980, 5159, 4076, 374, 921, 23758, 45
84, 1163, 1617, 1293, 271, 12840, 374, 220, 20698, 23181, 1
630, 220, 16, 13, 22467, 17, 30]
```

https://tiktokenizer.vercel.app

AI Functions

Ask an interesting business question	Identify the appropriate DL task
Is 'it' present or not?	Detection
What type of thing is 'it'?	Classification
To what extent is 'it' present?	Segmentation
What is the likely outcome?	Prediction
What will likely satisfy the objective?	Recommendation

INPUT DATA

Text

Images

Video

Audio

EXAMPLES OF OUTPUT BY INDUSTRY

Healthcare	Retail	Finance
Cancer Detection	Targeted Ads	Cybersecurity
Image Classification	Basket Analysis	Credit Scoring
Tumor Size / Shape Analysis	Build 360° Customer View	Credit Risk Analysis
Survivability Prediction	Sentiment & Behavior Recognition	Fraud Detection
Therapy Recommendation	Recommendation Engine	Algorithmic Trading

Image provided courtesy of NVIDIA

Generative AI Use Cases

Intelligent Chatbot

Focus is on question-and-answer tasks.

Ex. Customer Service Agent, Brand Ambassador, Help Desk

Knowledge Base Copilot

Connects to knowledge bases performs tasks such as writing, coding, generating images, etc.

Ex. Documentation Copilot, IT Bugs Assistant, Field Agent Copilot

Code Generation

Help develop or troubleshoot code based on natural language. Can work across common languages or be proprietary languages.

Ex. GitHub Copilot, ChatUSD, Software Development Assistant

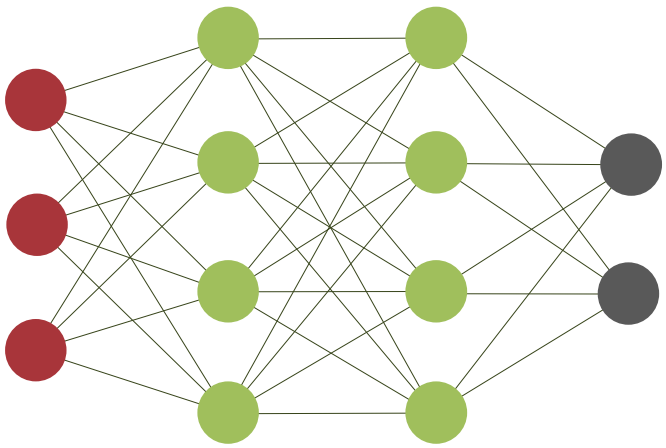
Image provided courtesy of NVIDIA

Multimodal Conversion

Demonstration

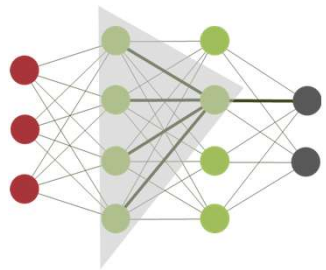
Up Next: Anatomy of a Neural Network

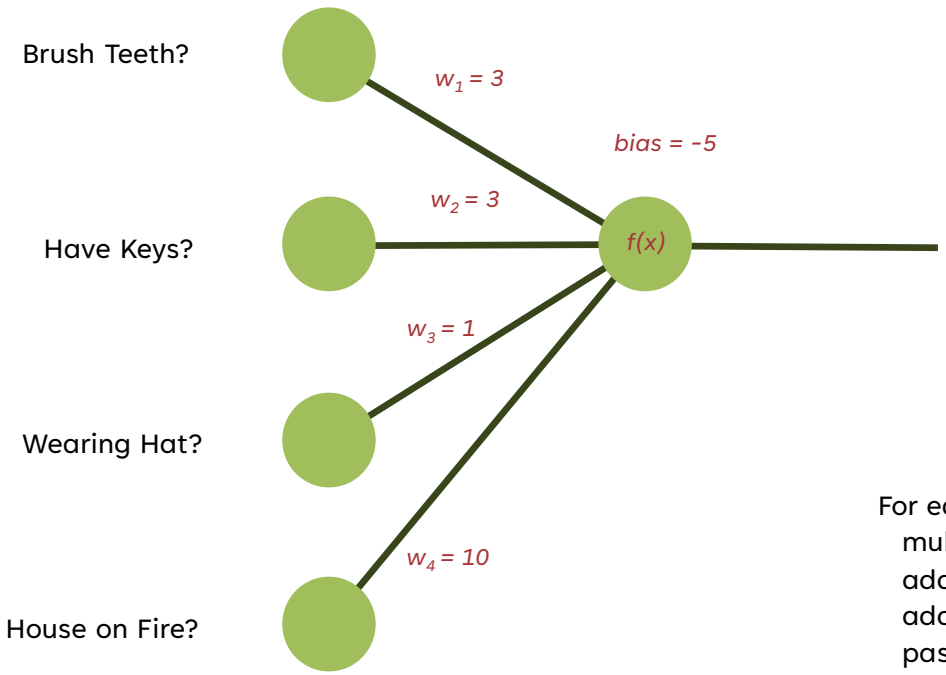
Artificial Neural Network



Artificial Neural Network	Description/Purpose	Typical LT
Convolutional Neural Networks (CNNs)	Classification & Recognition	Supervised
Recurrent Neural Networks (RNNs)	Sequential Patterns	Supervised
Transformer Networks (LLM)	Comprehension of Context	Semi-Supervised
Generative Adversarial Networks (GANs)	Competing ANNs (Generator & Discriminator)	Unsupervised & Semi-Supervised
Diffusion Model	Forward Diffusion & Reverse De-Noise	Semi-Supervised

Neurons, Weights, and Biases





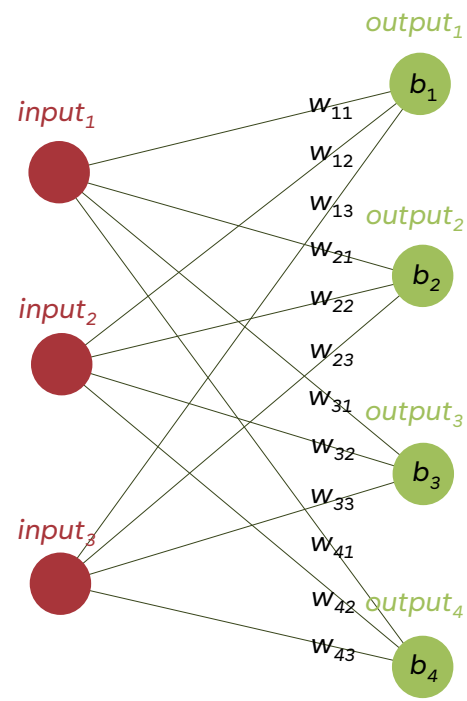
$w_1 * I$	$w_2 * I$	$w_3 * I$	$w_4 * I$	bias	value	result
Teeth	Keys	Hat	Fire			
3	3	1	0	-5	**2**	active
3	0	1	0	-5	**-1**	inactive
0	0	0	10	-5	**5**	active
3	3	1	10	-5	**12**	active

For each output node:
multiply weight and input value,
add the weighted inputs together,
add bias to sum
pass results through activation function

Matrix Multiplication

$$\text{activation}\left(\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \end{bmatrix} \times \begin{bmatrix} input_1 \\ input_2 \\ input_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}\right) = \begin{bmatrix} output_1 \\ output_2 \\ output_3 \\ output_4 \end{bmatrix}$$

$$\begin{bmatrix} w_{11} \times input_1 + w_{12} \times input_2 + w_{13} \times input_3 \\ w_{21} \times input_1 + w_{22} \times input_2 + w_{23} \times input_3 \\ w_{31} \times input_1 + w_{32} \times input_2 + w_{33} \times input_3 \\ w_{41} \times input_1 + w_{42} \times input_2 + w_{43} \times input_3 \end{bmatrix}$$

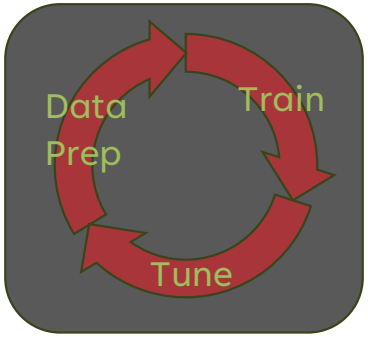


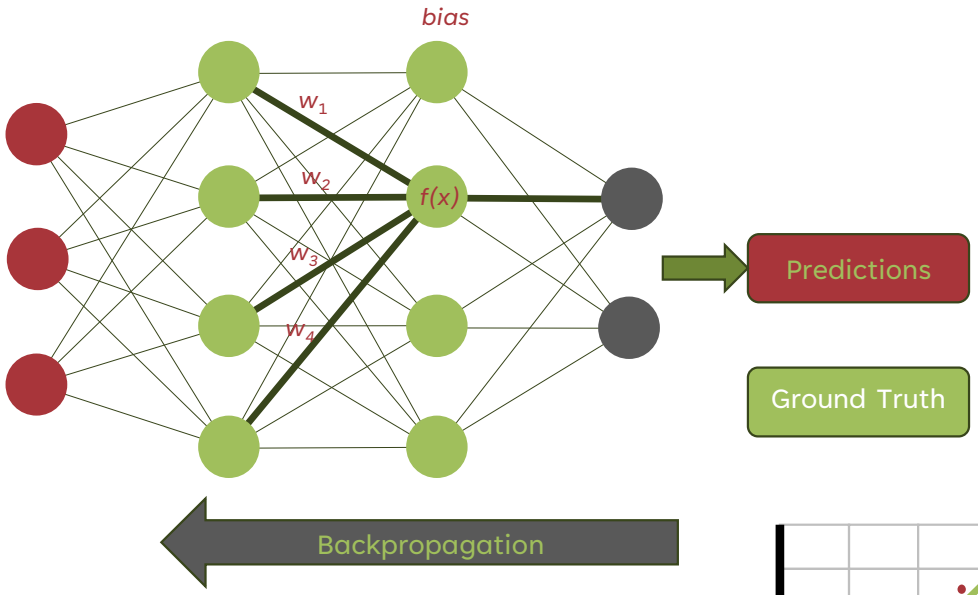
For each output node:
- multiply weight and input value,
- add the weighted inputs together,
- add bias to sum
- pass results through activation function

Preparing the Model

- Labeling
- Missing/Incorrect Data
- Normalization/Standardization
- Augmentation
- Split

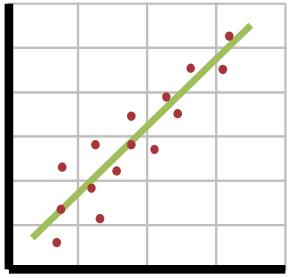
  Train / Validate / Test





- Fine Tune vs. Retraining
- Hyper Parameter Adjustments
- Pre-Trained Model Selection

$y = mx + b$
$\boldsymbol{x}$ = input(s)
$\boldsymbol{y}$ = output



Customizing LLMs

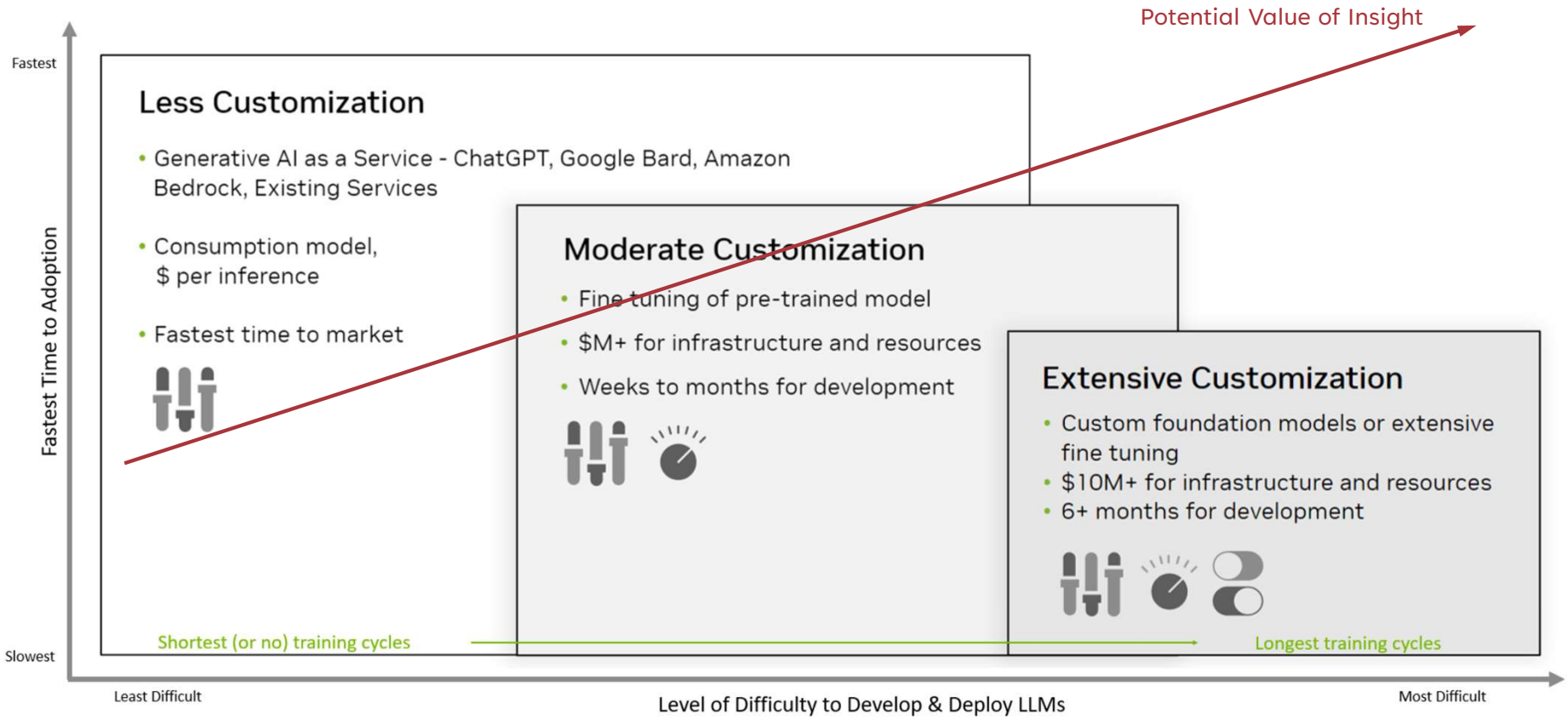
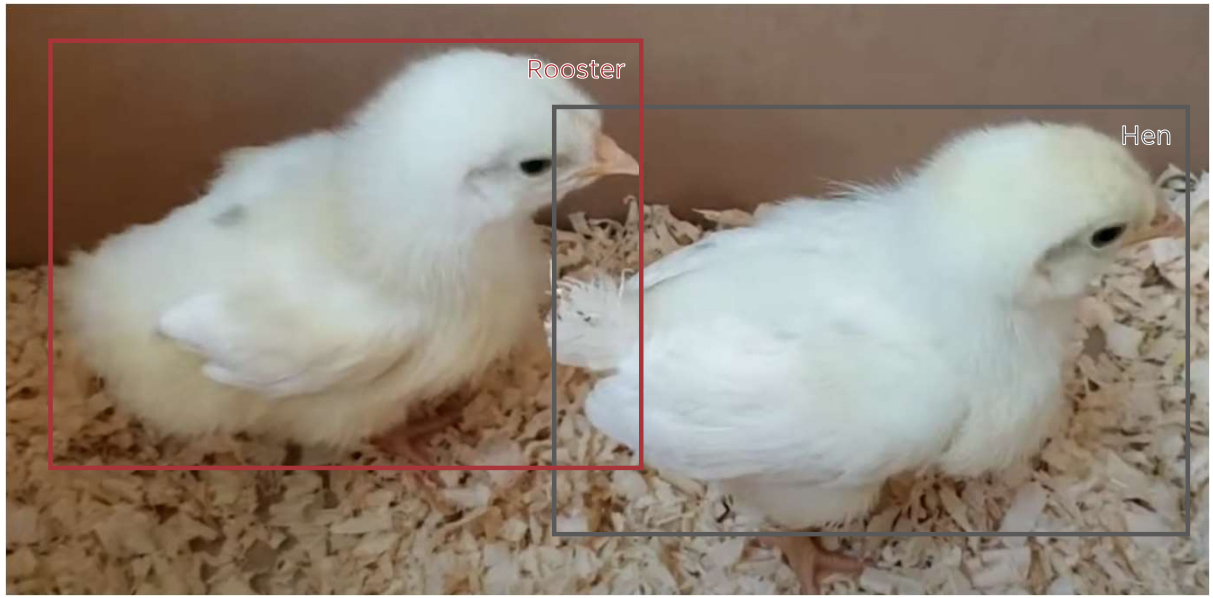


Image provided courtesy of NVIDIA

Improving Insights



Up Next: Sourcing Foundational Models

Model Selection

Hugging Face

NVIDIA NGC Catalog

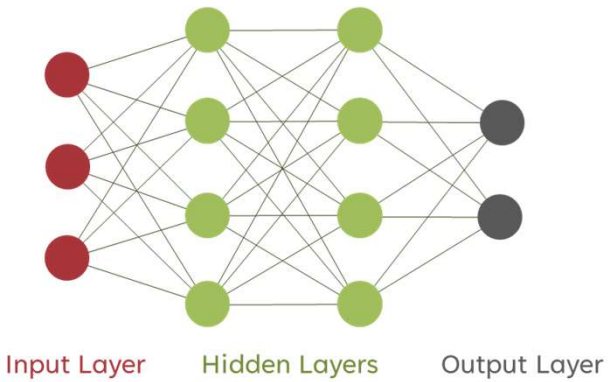
Token count
19

This is an example of token usage.

Prompts and responses all account for the context window.

2028, 374, 459, 3187, 315, 4037, 10648, 382, 36286, 13044, 323, 14847, 682, 2759, 369, 279, 2317, 3321, 4286

https://tiktokenizer.vercel.app



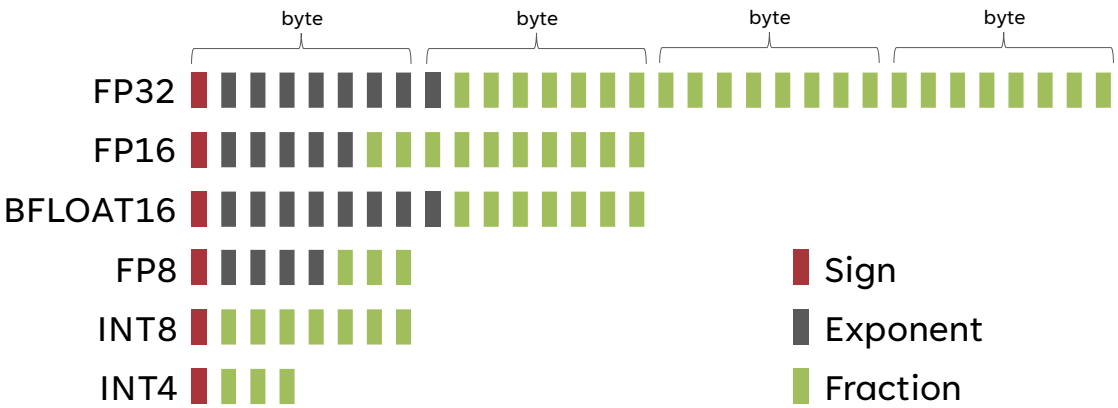
Model Repository Features:

- Model Card
- Live Interface
- Code Examples

Key Terminology:

- Context Window / Length
- Parameter
- Quantization
- Licensing

Quantization



Floating Points:

$$\text{value} = (-1)^{\text{sign}} \times 2^{(E-127)} \times \left(1 + \sum_{i=1}^{23} b_{23-i} 2^{-i}\right)$$

Equation source: Wikipedia

FPvalue = $(-1)^{sign}$ * 1.[*fraction*] * $2^{[exponent]-(highBit - 1)}$

3.1416 = $(-1)^{0}$ * 1.[*5708*] * 2^{1}

More Exponent Bits = Larger Range
More Fraction Bits = Greater Precision

NVIDIA NEMO

Quanto

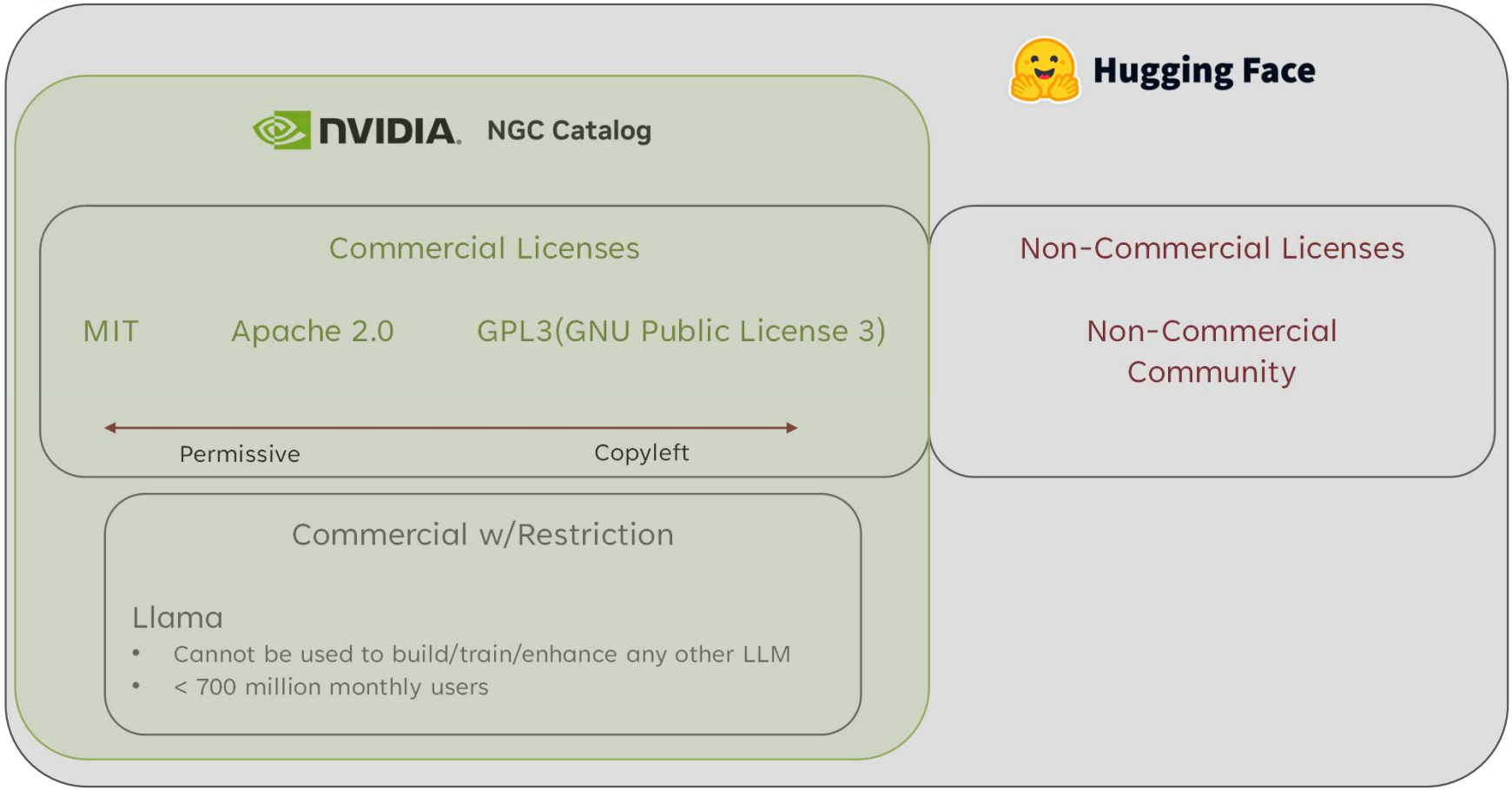
ONNX

GGUF

Acroyms:

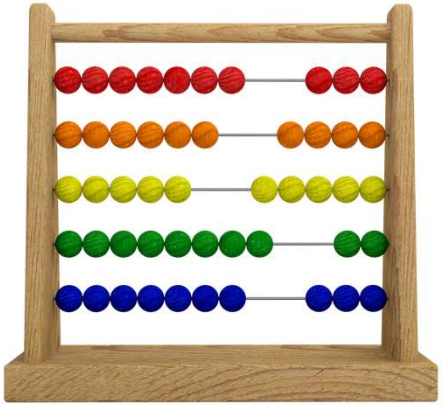
NeMo (Neural Modules)
ONNX (Open Neural Net Exchange)
GGUF (GPT-Generated Unified Format)
PTQ (Post Training Quantization)
QAT (Quantization-Aware Training)

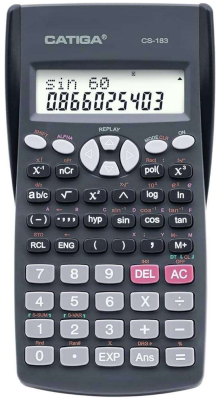
Licensing



This is not legal advice

Weighing Options





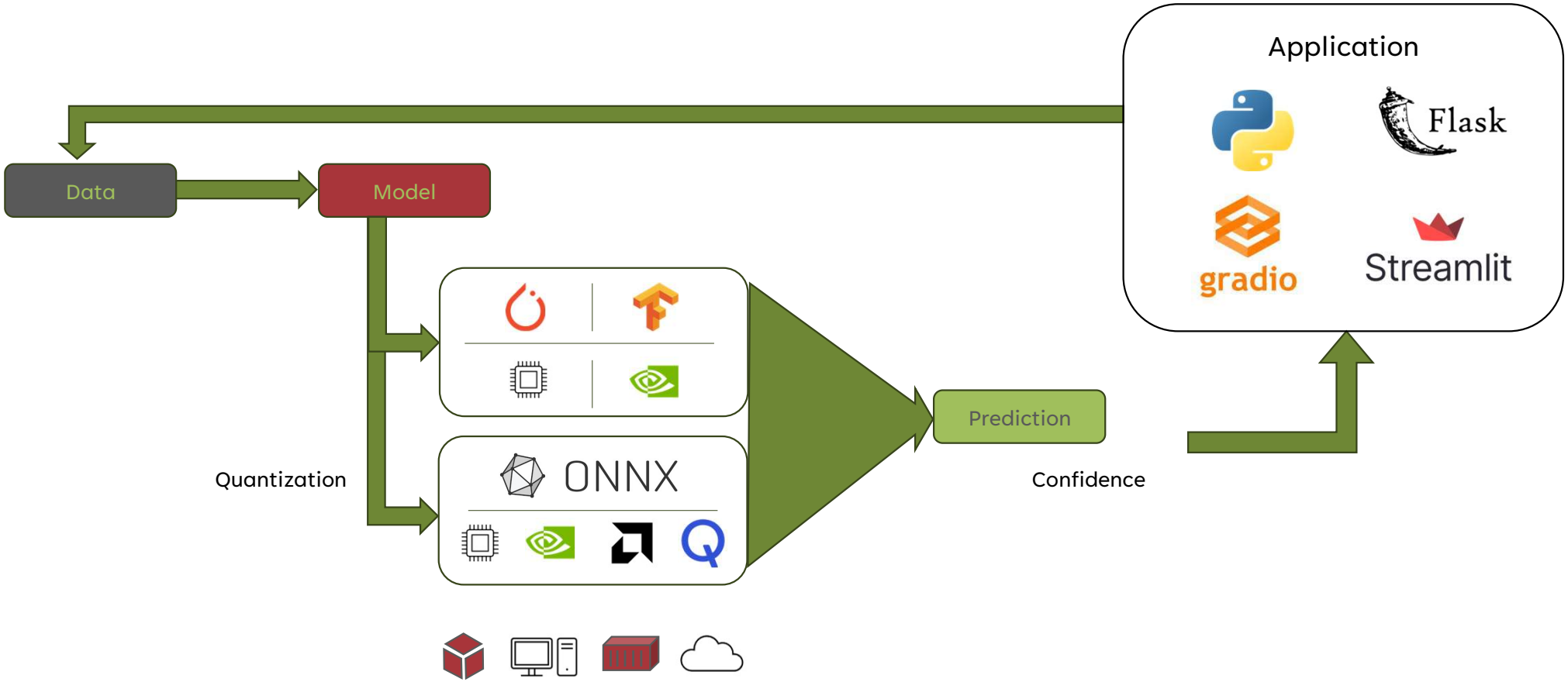
Cost
Precision
Speed
Size & Scalability
Ease of Use

Up Next: Putting AI Models to Work

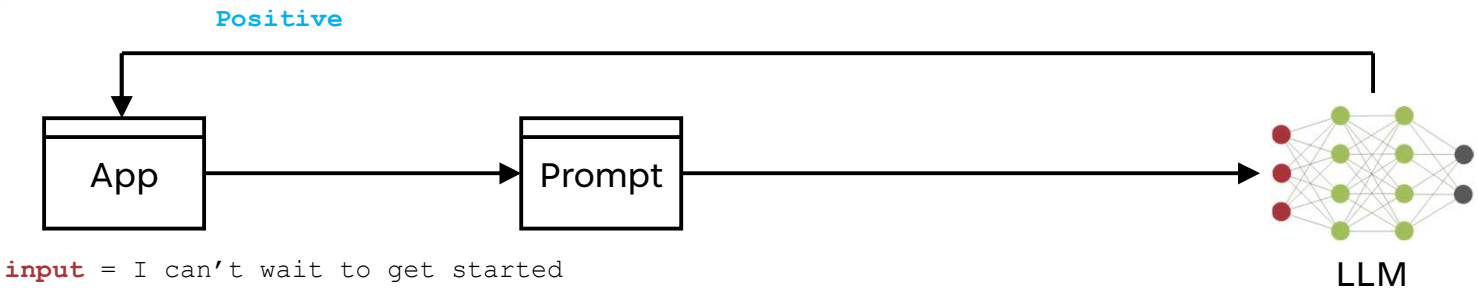
IT + CYBERSECURITY

EST. 1991

Deploy & Apply



Prompt Engineering



Prompt Template

```
prompt = f"""
{role}
{context}
{instruct}
"""
```

Few-Shot

```
instruct = """
Classify this comment:
Awesome presentation!
Sentiment: Positive

Classify this comment:
What a waste of time!
Sentiment: Negative

Classify this comment:
{input}
Sentiment:"""
```

Role Prompt

```
role = """You are an event planner
reviewing audience feedback from a
survey taken after a technical
presentation"""
```

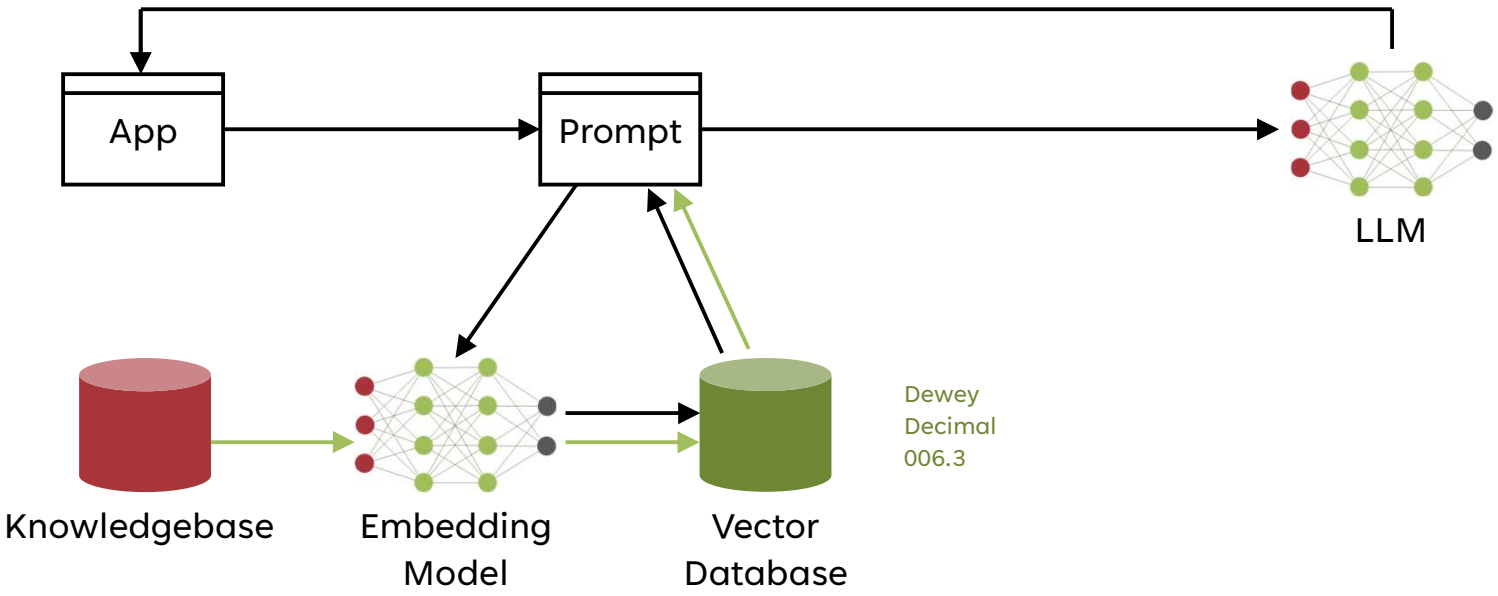
Context

```
context = """Other relevant info"""
```

Prompt

```
You are an event planner
reviewing audience feedback
from a survey taken after a
technical presentation

Other relevant info

Classify this comment:
Awesome presentation!
Sentiment: Positive

Classify this comment:
What a waste of time!
Sentiment: Negative

Classify this comment:
I can't wait to get started
Sentiment:
```

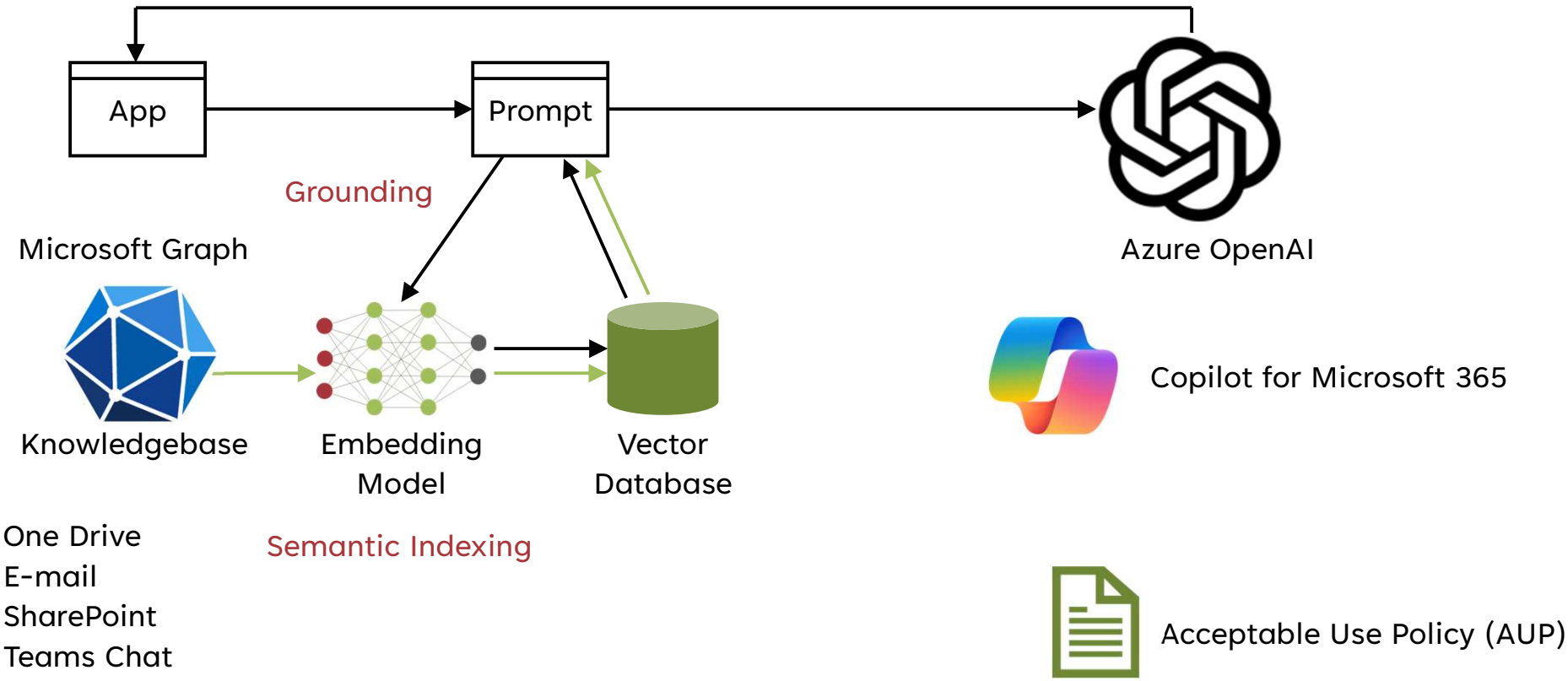
Retrieval Augmented Generation (RAG)



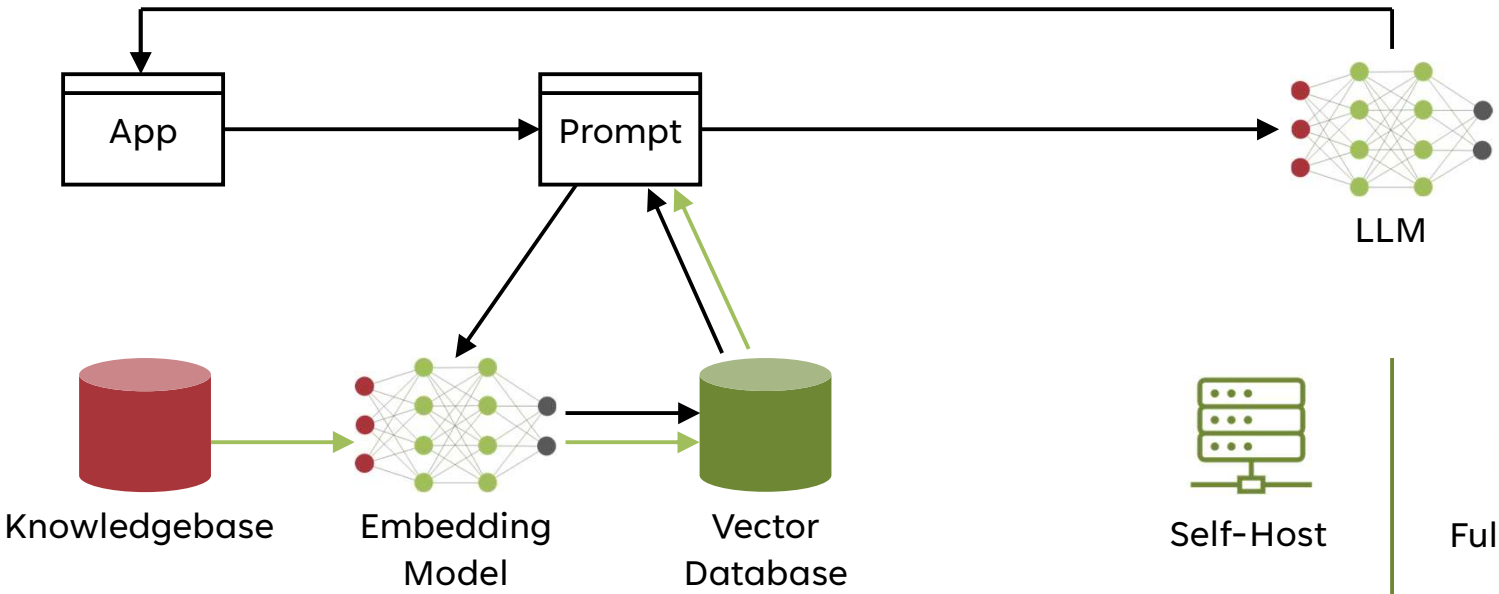
Key Benefits:

- Minimize Hallucinations
- Current Data
- Citing Sources

Copilot for Microsoft 365



LLM Considerations – On-Prem vs. External



Self-Host

Open Source
On-Prem / Cloud

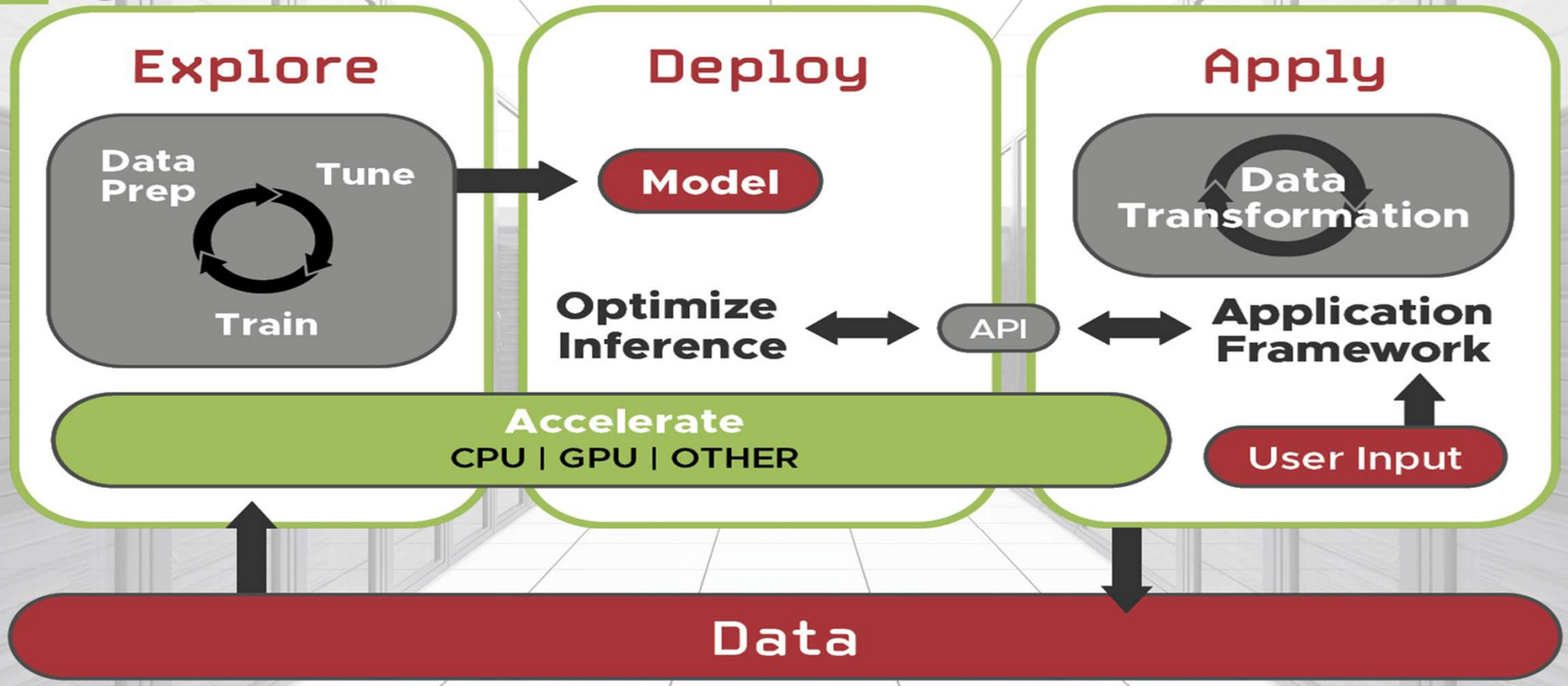
OpenELM Mistral Gemma Phi-3 Llama 3

Fully Managed

Token Based Billing
Cloud Only

OpenAI API

AI Blueprint



Getting Started

Prerequisits:

- Code Editor (VS Code)
- Access to a GPU (Eventually)
- Python Proficiency
- Project Ideas
- Commitment of Time

Python Module	Purpose
Jupyter Notebooks	Experimenting
NumPy	Number Matrix
Pandas	Tabular Data
OpenCV	Visual Data
Flask / Gradio / Streamlit	User Interface

Web Resources:

General	Data Science	Model Deploy	LLM Frameworks
Google	colab.google	ai.nvidia.com	LangChain
YouTube	Kaggle	huggingface.co	LlamaIndex
Coursera	Fast.ai	Ollama	

Nth University: AI Curriculum

Practical AI: Definitions, Concepts, and Techniques

Applied AI: Putting AI to Work

Deployed AI: Run & Maintain Inferencing Services

Practical AI:
AI Taxonomy
Predictive AI / Generative AI
Neural Network Overview
Model Sourcing & Sizing
- NVIDIA
- Hugging Face

LLM Inferencing Techniques
Getting Started

Applied AI:
Development Environment
- VENV / Jupyter

Model Serving
- NVIDIA NIM
- vLLM

User Interface
- Gradio / Streamlit

LLM Tools
- LangChain / LlamaIndex

Data Retrieval (RAG)
- Data Chunking
- Embedding Model
- Vector Database

Deployed AI:
GPU Drivers / Prep
Bare Metal
Virtual Machine
- VMware
- KVM

Docker
- Install
- GPU Toolkit
- Basic Commands

Kubernetes
Ecosystem Solutions

Review of Use Cases

Video Analytics

- Perimeter Surveillance
- Intrusion Detection
- Vehicle Identification / License Plate Reader
- Manufacturing Quality Control

Large Language Models (LLMs)

- ChatBot Assistant
- Code Generation
- Content & Image Creation
- Sentiment Analysis
- Text Translation and Summarization

Multi-Modal

- Image to Text
- Text to Image
- Speech to Text
- Text to Speech
- Video to Text
- Text to Video

Poll: How Can Nth Help?

Conclusion Q & A

Thank You

Nth GENERATION

Nth.com | Security.Nth.com | 800.548.1883 | info@nth.com