

CISCO *Live!*

GO BEYOND

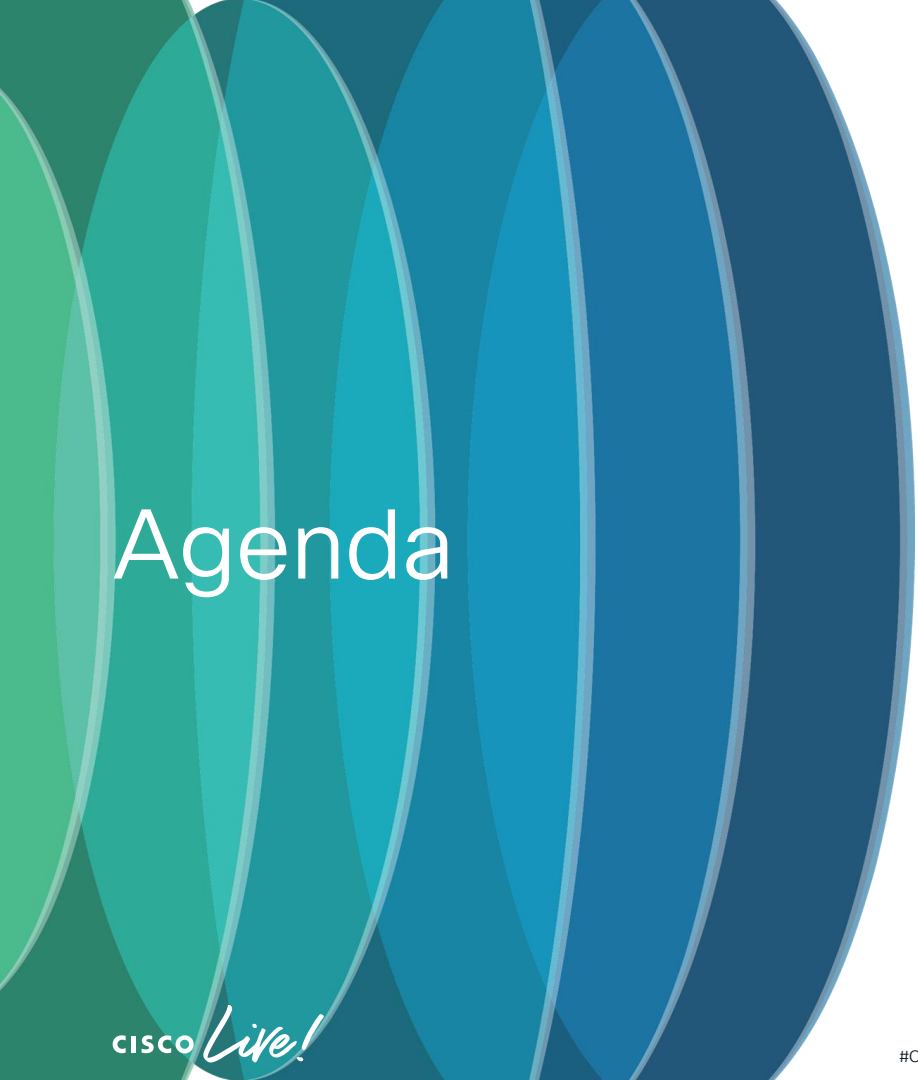
#CiscoLiveAPJC



Getting Started with Large Language Models (LLMs)

From the Start to Today

Flo Pachinger, Developer Advocate
@flopachinger
DEVNET-2161



Agenda

- Unpacking AI
- Predictive vs Generative AI
- How to get started: 4 main GenAI Architectures
- Toolset



@flopachinger
flopach

developer.cisco.com

cisco *Live!*

Flo Pachinger

Developer Advocate

- Focusing on Data & AI
- Based in Vienna, Austria
- DevNet Projects:
 - API+Code Assistant for Catalyst Center
 - Creator of johann (Network Monitoring Tool)
 - Po Robot Arm Challenge @ CiscoLive 2022
 - Play Minecraft on Catalyst 9300 Switch

WSJ WSJ

How Generative AI Will Change the Way You Use the Web, From Search to Shopping

People seeking information online will increasingly go first to TikTok, ChatGPT and other applications powered by generative artificial...



MIT Sloan

Sam Altman believes AI will change the world (and everything else)

The OpenAI CEO talks artificial intelligence and regulation, AI and bias, AI and daily life, and more. For many reasons, he's optimistic.



F Forbes

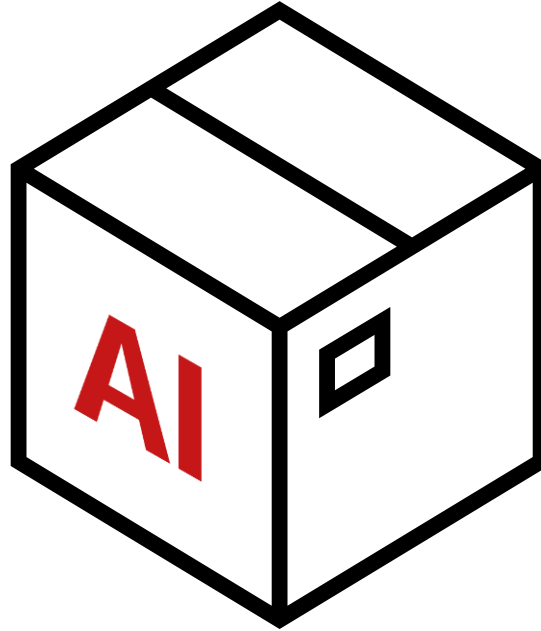
How Generative AI Will Change All Of Our Jobs In 2024

As we step into 2024, the landscape of work is undergoing a transformation unlike any before, primarily driven by the advancements in...



“The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings”

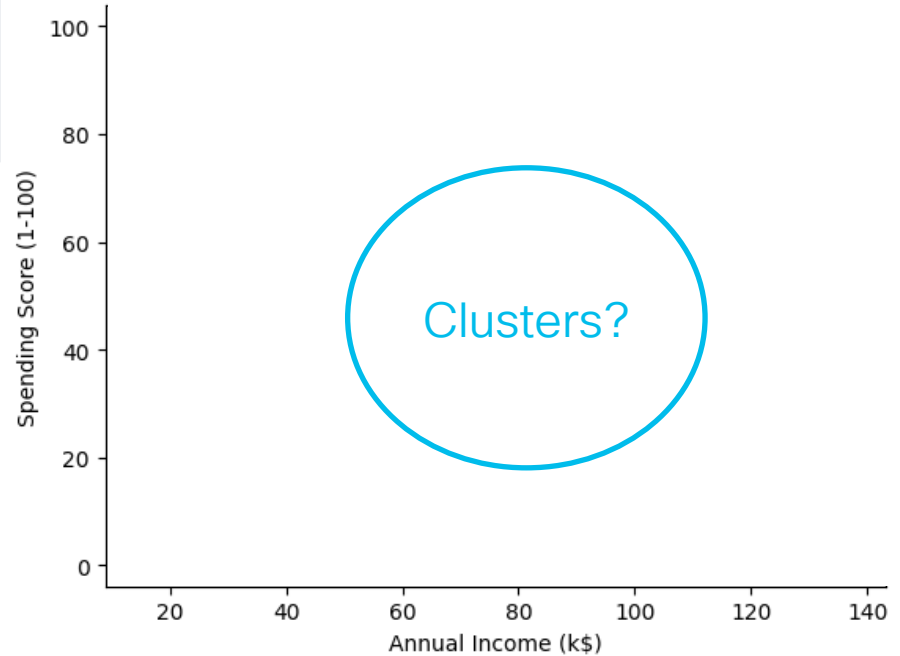
AI definition of Encyclopædia Britannica



Unmasking AI example: clustering customers

	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

Magic?



“The Magic” = Kmeans algorithm

1. **Initialize:** $\mu_1, \mu_2, \dots, \mu_K$

2. **Iterate:**

- Assign $\mathbf{x}_j$ to the nearest μ_i :

$$c_j = \arg \min_i \|\mathbf{x}_j - \mu_i\|^2$$

- Update μ_i :

$$\mu_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$$

3. **Converge** when centroids μ_i stabilize.



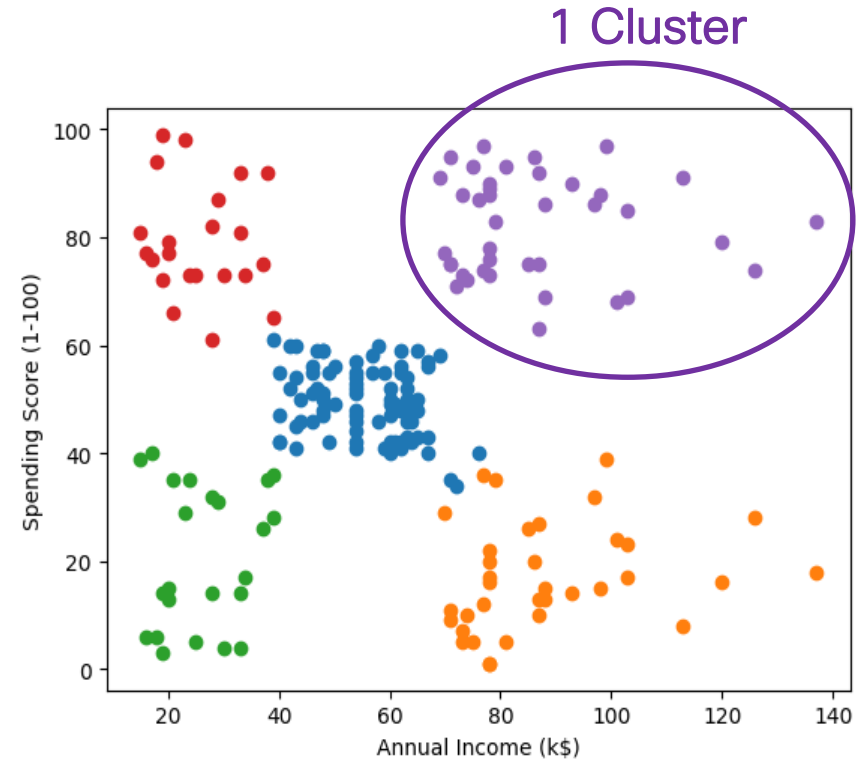
```
from sklearn.cluster import KMeans
km = KMeans(n_clusters=5, random_state=42)
km.fit(df[['Annual Income (k$)', 'Spending Score (1-100)']])
```

Unmasking AI: clustering customers

	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40



Kmeans
algorithm



1. **Initialize:** $\mu_1, \mu_2, \dots, \mu_K$
2. **Iterate:**
 - Assign $\mathbf{x}_j$ to the nearest μ_i :
$$c_j = \arg \min_i \|\mathbf{x}_j - \mu_i\|^2$$
 - Update μ_i :
$$\mu_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$$
3. **Converge** when centroids μ_i stabilize.



Understanding the Math

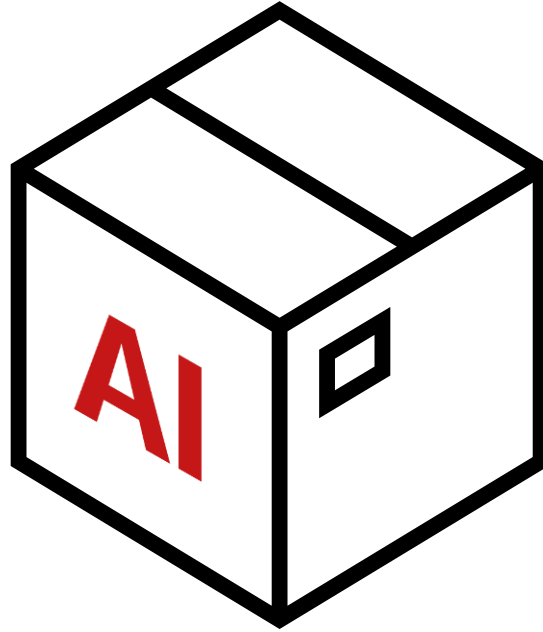
Math & Statistics

```
from sklearn.cluster import KMeans
km = KMeans(n_clusters=5, random_state=42)
km.fit(df[['Annual Income (k$)', 'Spending Score (1-100)']])
```

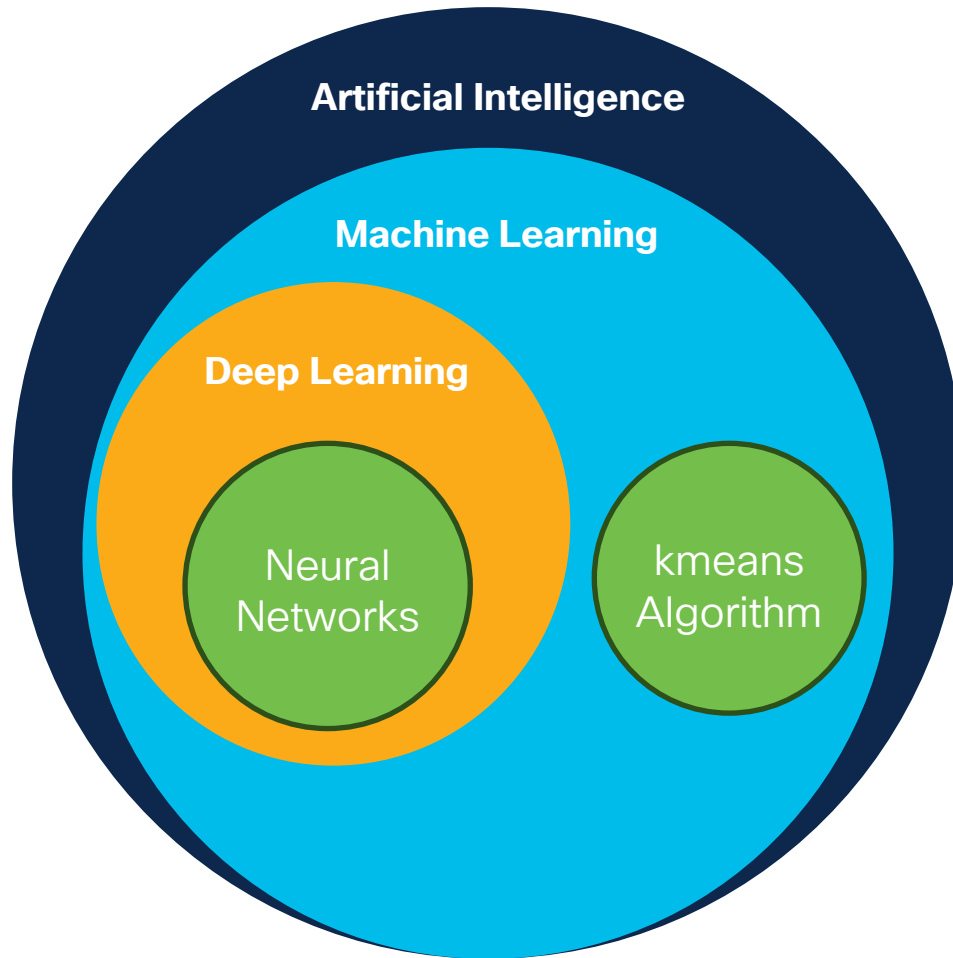
Programming Code

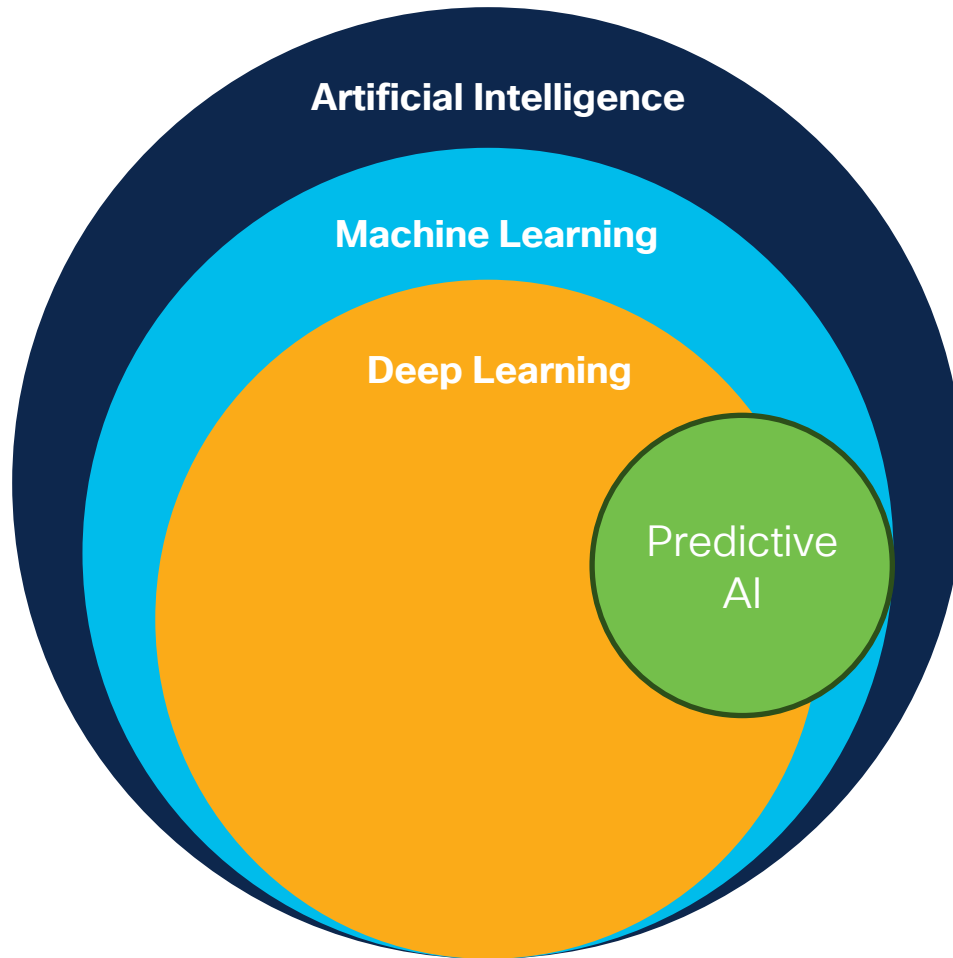


Understanding
how to apply the math and
what algorithm/method to
choose
for which problem





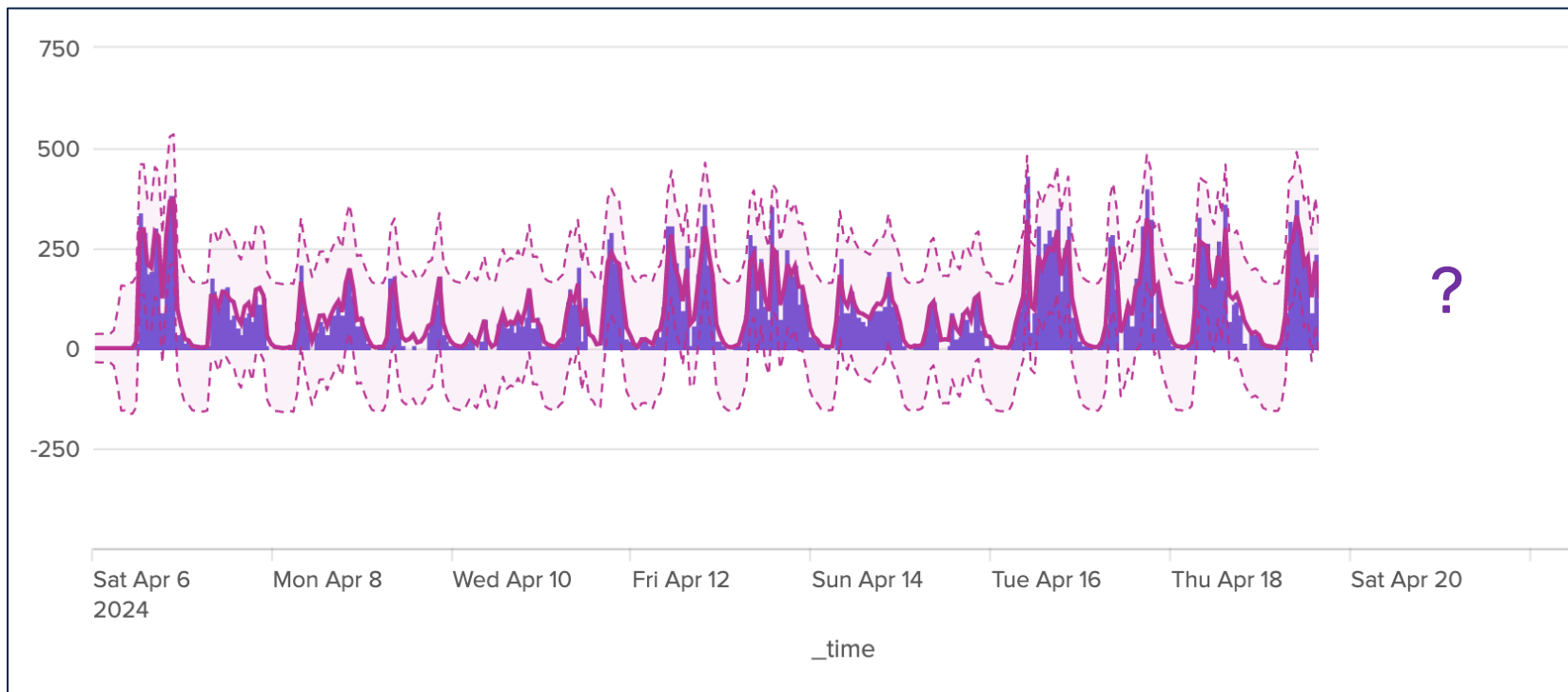






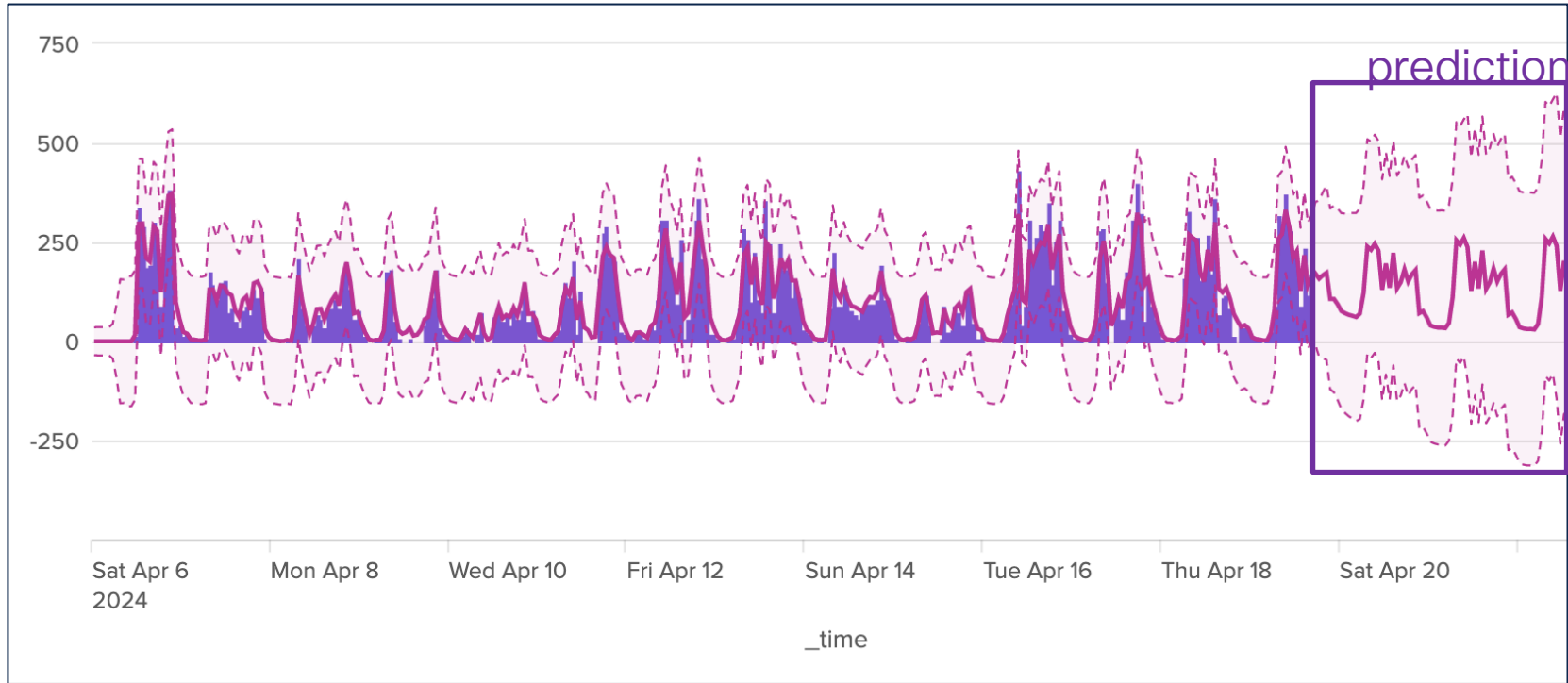
Muffin or
Chihuahua?

Predicted Facebook usage in the network



Splunk Dashboard

Predictive AI: Predicted Facebook usage in the network

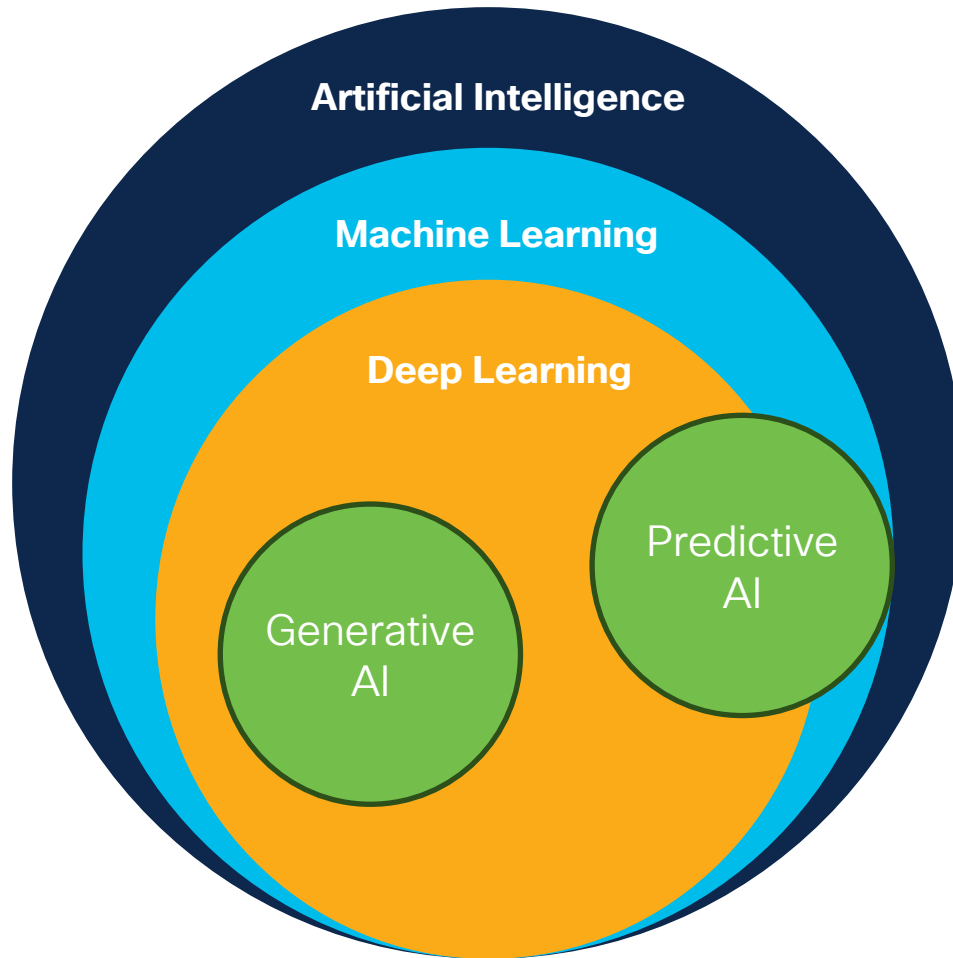


Splunk Dashboard

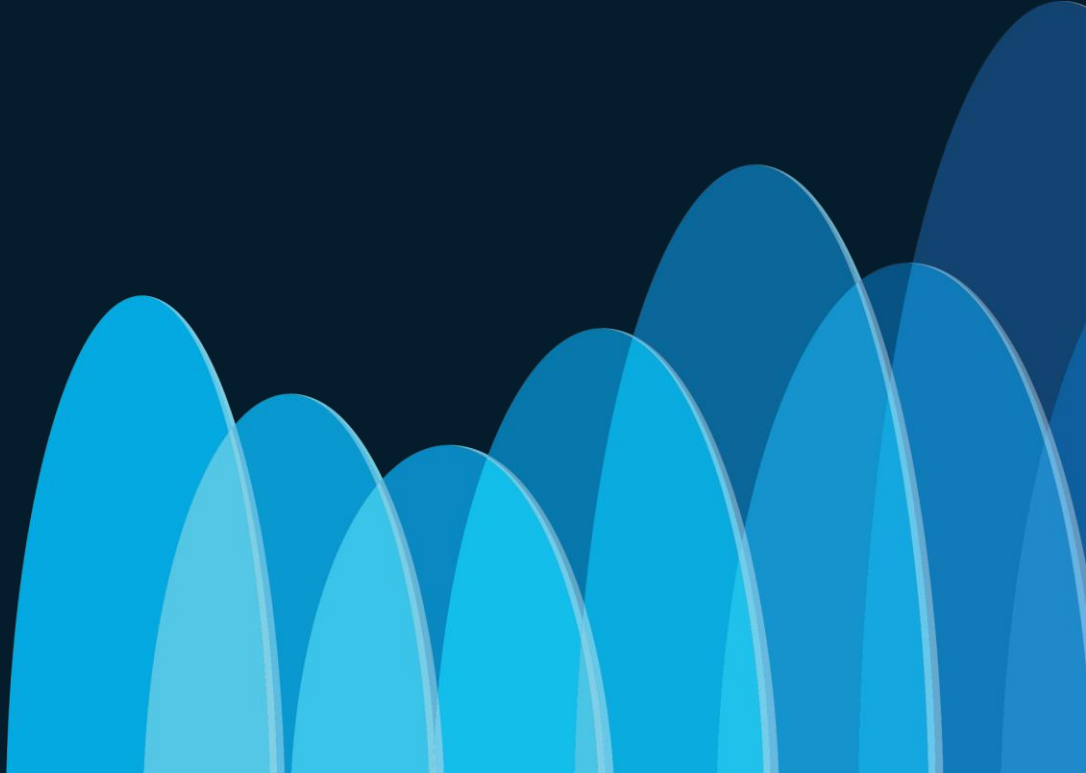
Predictive vs. Generative: test scenario

- **Goal:** Predict the revenue for 2025
- Revenue data from 2010 – 2024
- **Generative AI:** LLMs
- **Predictive AI:** Linear Regression

	year	revenues
0	2010	10000
1	2011	18000
2	2012	25000
3	2013	50000
4	2014	56000
5	2015	70000
6	2016	80000
7	2017	90000
8	2018	100000
9	2019	95000
10	2020	110000
11	2021	130000
12	2022	110000
13	2023	140000
14	2024	180000



How did GenAI happened?



AI key milestones (Selected)

Foundation

Foundational statistical methods for machine learning

Further Research

Feedforward neural networks (FNNs)

Nearest Neighbor

Visible Success

Data-driven

Unsupervised machine learning

Long Short-Term Memory (LSTM) neural network

GenAI Boom

Large language models

GPT-4, BERT, Llama

Pre 1940s

1940s/50s

1960s

1980s

1990s/2000s

2010s

2020s

Birth of AI

Turing test

Neural network & artificial neuron

Perceptron

New Hope

Convolutional neural networks (CNNs)

Recurrent neural network (RNN)

Rediscovery of Backpropagation

Reinforcement Learning

Deep Learning Revolution

Image & object recognition

Generative Adversarial Network (GAN) 2014

Transformer Architecture 2017

Attention is all you need (2017)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

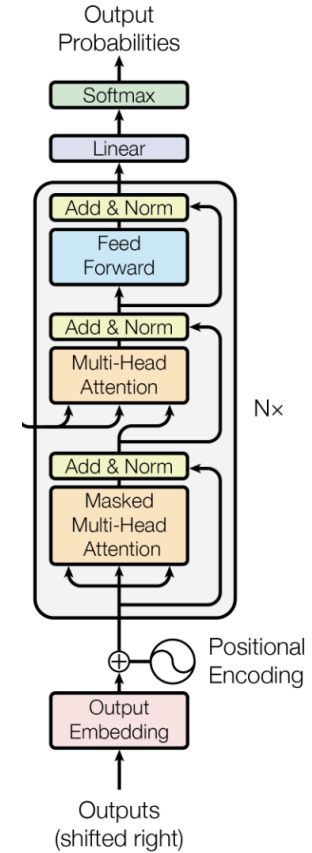
Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Transformer architecture



© Ashish Vaswani et al. (Attention Is All You Need)

Predictive AI

- Analyzing existing data for prediction & classification
- Regained popularity in 2010s

Generative AI

- Creating text, images, audio & other data
- Mainstreamed in 2022

GenAI: opportunities and limitations

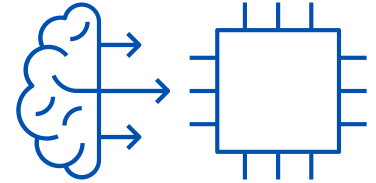
- ☑ Creation of code, text, audio & video
- ☑ Conversational user interfaces
- ☑ Knowledge discovery
- ☑ Classification of specific data

- ✗ GenAI is hardly useful for:
 - ✗ Prediction e.g. network utilization prediction
 - ✗ Specific classification tasks e.g. customer clustering
- ✗ Can produce inaccurate and fabricated answers

“So how do I start now with generative AI?”

GenAI architecture approaches

**Pre-train from
scratch**



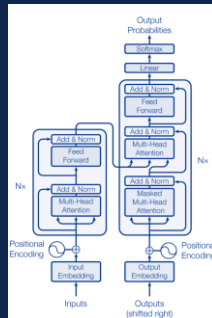
**Training the LLM from
scratch**



LLM Training



Millions to trillions
of tokens of text
data



Training using
transformer
architecture



Trained LLM-Model
(weights, biases, instructions)

LLM Inferencing

When was Cisco founded?

User Input



Trained LLM-Model



Cisco
Cisco Systems
Cisco Systems was
Cisco Systems was founded
Cisco Systems was founded on
Cisco Systems was founded on December
Cisco Systems was founded on December 10
Cisco Systems was founded on December 10, 1984

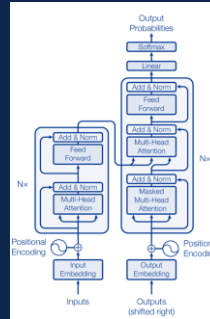
Model Output

LLM training example: Llama3 70B

LLM Training



15 trillion tokens
of text data



Training on
48.000x H100-80GB
GPUs

After 17 days



Llama3
70B parameters

LLM creation data overview

Model Name + Parameters	Training Tokens	GPU Time (hours)	GPU Hardware	Training Time
Llama2 7B	2T	184.320	A100-80GB	n/a
Llama2 70B	2T	368.640	A100-80GB	n/a
Llama3 8B	15T	1.3M	2x 24k H100-80GB	~3 days
BioMedLM 2.7B	34.6B	n/a	128x A100-40GB	6.25 days
phi-3-mini 3.8B	3.3T	n/a	512x H100-80GB	7 days
Llama3 70B	15T	6.4M	2x 24k H100-80GB	~17 days
BloombergGPT 50B	700B	1.3M	512x A100-40GB	53 days
Bloom 176B	350B	1M	384x A100 80GB	4 months

H100 is superior to the A100

Example: GPU infrastructure costs LLM training

- 400 GPUs: A100-80GB → 18.000 \$
- Depends on: # of model parameters, training time, input data

On-prem Infrastructure:

Item	Single price	Quantity	Power consumption (kW/h)	Total Costs	Remarks
A100-80GB	\$18,000.00	400	0.3	\$7,200,000.00	List price
Electricity costs per day for total GPU quantity*	\$ 480.38	50		\$ 24,019.20	for 50 days
*average USD price in USA (kW/h):	\$ 0.1668			\$7,224,019.20	

Cloud:

Item	Single price	Quantity	Total Costs
A100-80GB GPU per hour (Cloud)	\$3.00	1,000,000.00	\$3,000,000.00
			\$3,000,000.00

Building from scratch

- ☑ Privacy
- ☑ 100% control
- ☑ Performance
- ☑ Accuracy

- ✗ High costs & resources (data, infrastructure, employees)
- ✗ Takes time

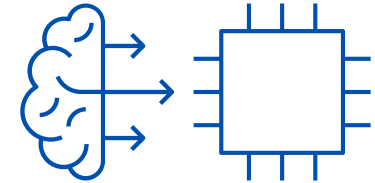
GenAI architecture approaches

Prompt engineering



Natural language text
describes the tasks
for the model

Pre-train from scratch

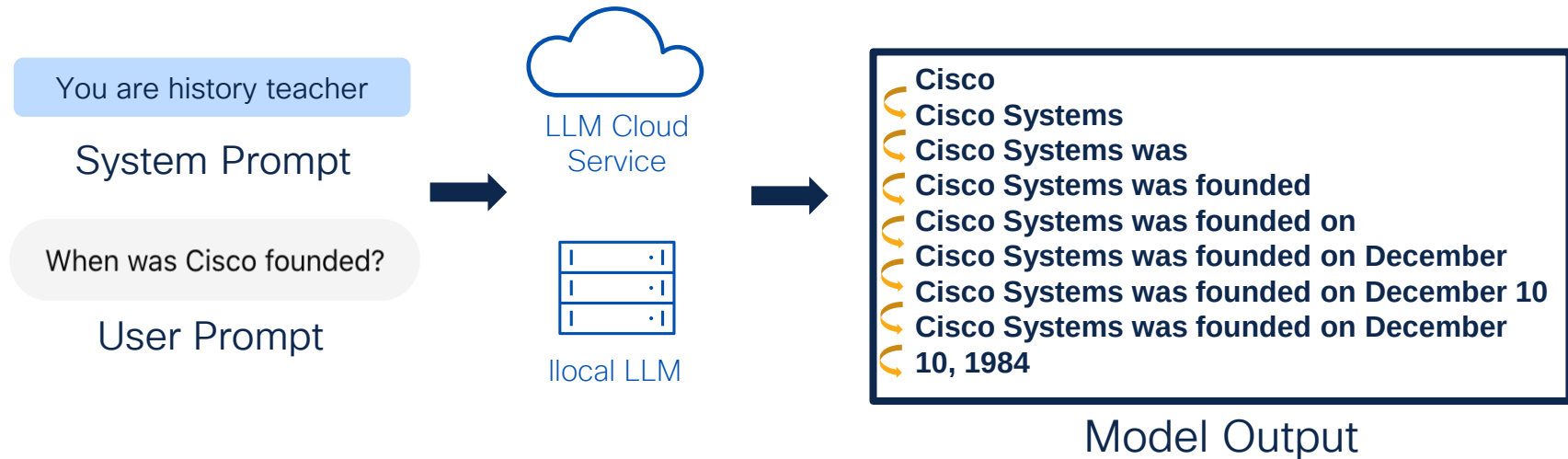


Training the LLM from
scratch



Prompt engineering = optimizing prompts

LLM Inferencing



Prompting example: create Python code

```
response = ask_llm(messages=[
    {
        'role': 'system',
        'content': f'''You are an assistant to a Cisco networking engineer.
The user wants to get information from a Cisco networking device.
The user-request is the problem which needs to be solved using the programming language Python.
Break down in a numbered list what tasks in the Python code are needed to solve this problem. Begin with:
1. Use the most suitable Python libraries listed below:
- netmiko
- ncclient with NETCONF. Use this for XML data format.
- requests with RESTCONF. Use this for JSON data format.
- NAPALM
2. The device type is defined as "{profile["type"]}" by the user. Change it accordingly.
3. Establish a connection to the device.
4. Choose the correct method to get data from the device.
Do not use regular expressions for parsing or filtering.
Add more steps as needed.
Always print the answer to the user.
Only output the numbered list.
'''
    }, {
        'role': 'user',
        'content': f'user-request: {userinput}',
    }
])
```

Prompt engineering best practices

- Give clear and specific instructions. Simplify them.
 - Use the 'requests' Python library. If loops or advanced code is needed, provide it.
- Ask the LLM not to hallucinate / make things up
 - Tell the user if you do not know the answer
- Give examples & use prompt separators (e.g. using XML tags)
 - `<examples> {example 1} {example 2} </examples>`
- Adopt a persona
 - You are the Cisco Catalyst Center REST API and Python code assistant

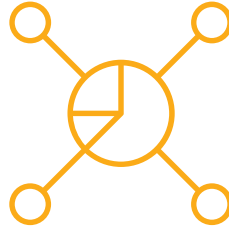
GenAI architecture approaches

Prompt engineering



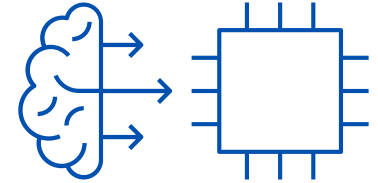
Natural language text describes the tasks for the model

Retrieval augmented generation (RAG)



Retrieving external knowledge for LLM-input

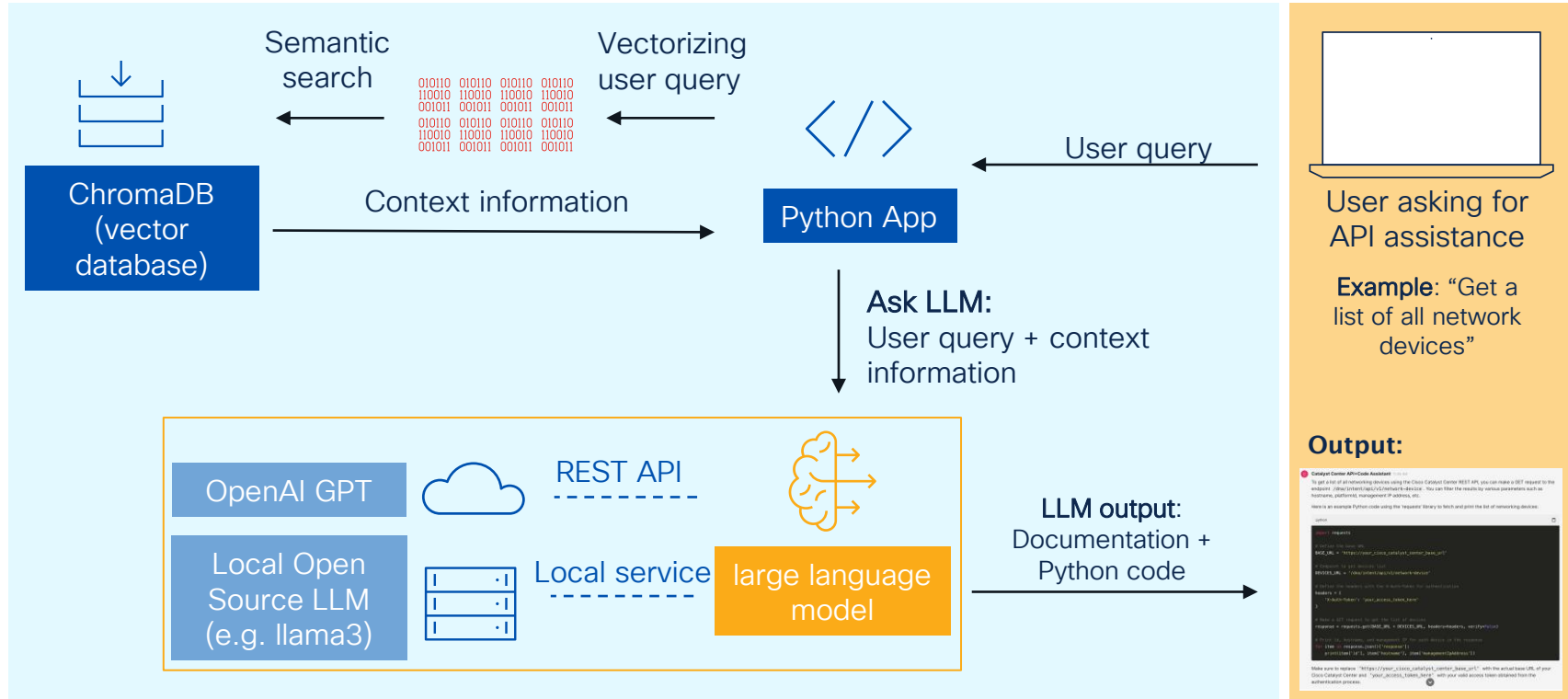
Pre-train from scratch



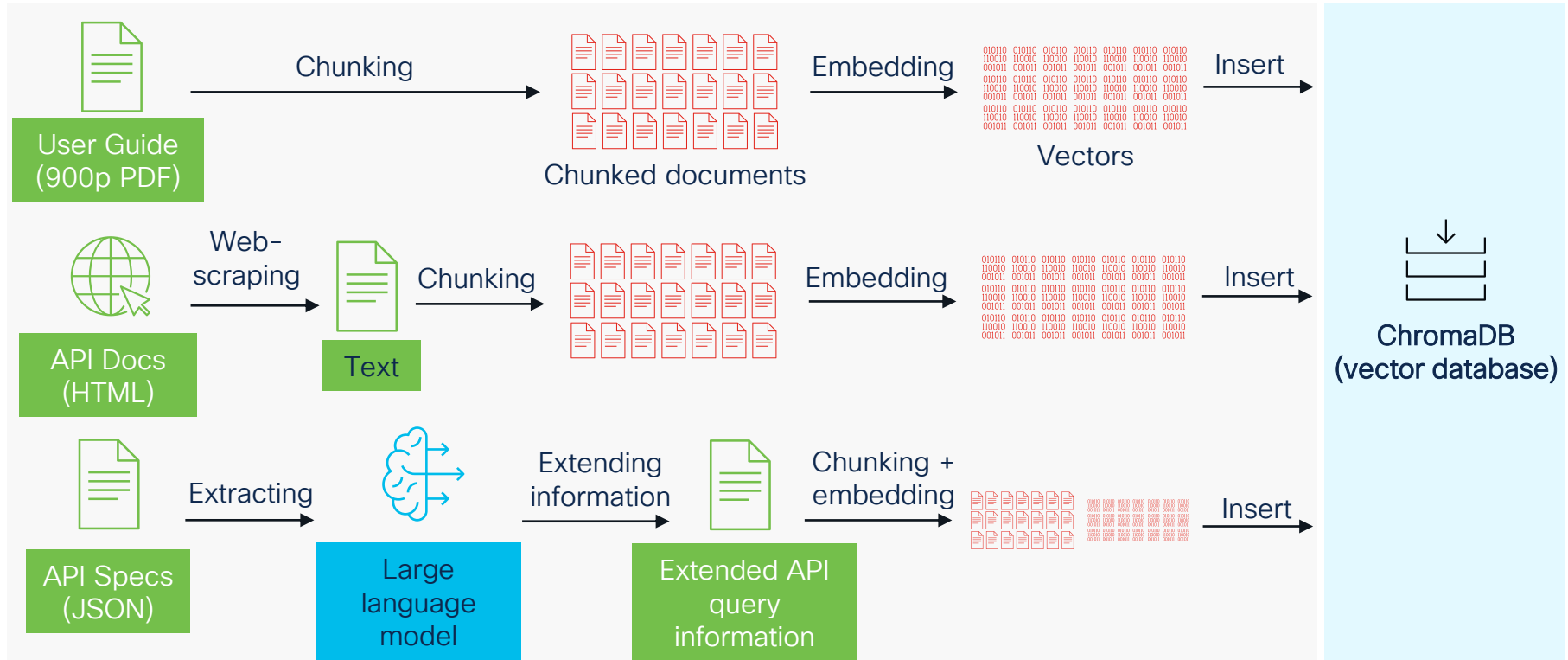
Training the LLM from scratch

← lower *Costs & compute-intensiveness* higher →

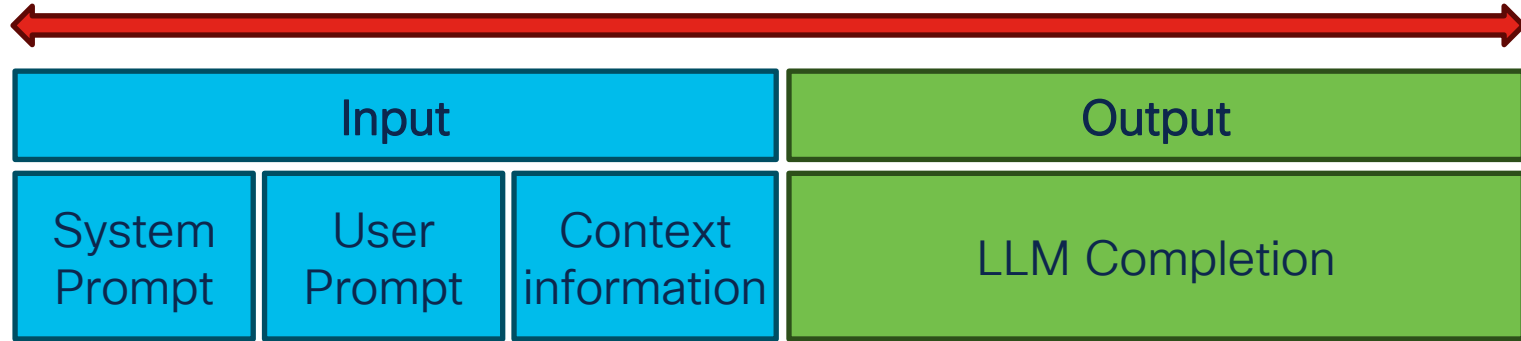
RAG architecture overview



RAG: 3 importing data examples



LLM Context Window

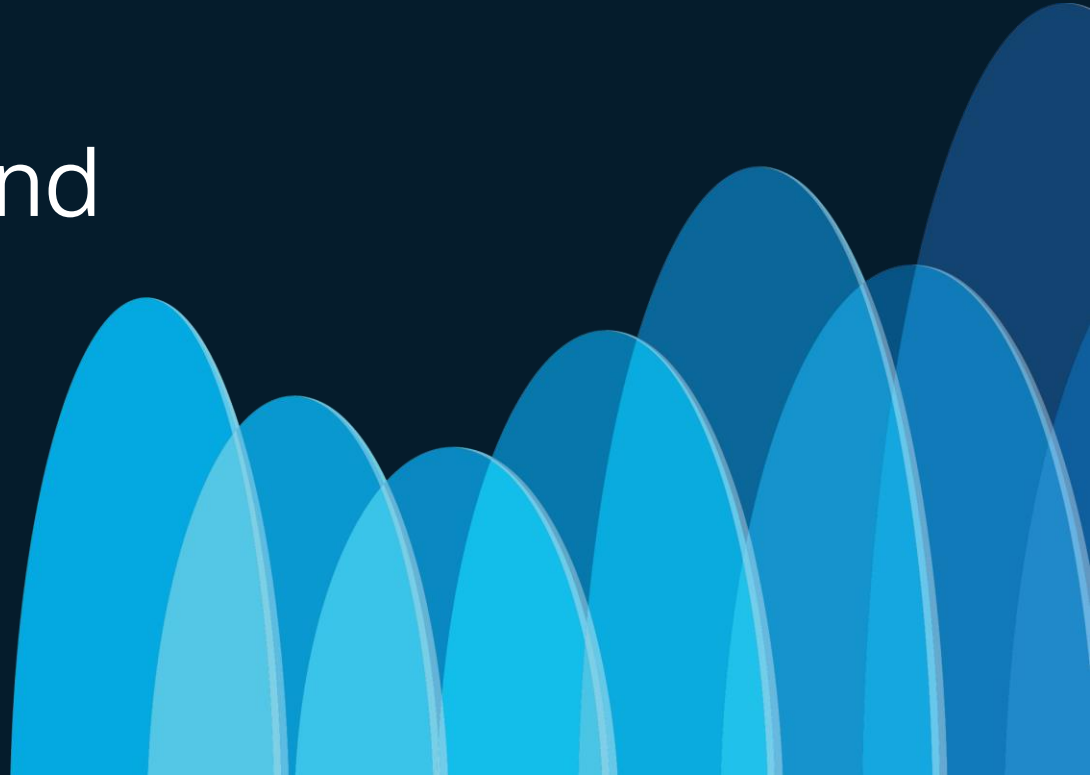


Context window on:

- Llama3.1: 128k tokens
- GPT-4o: 128k Tokens
- Claude: 200k Tokens

1k tokens $\pm$ 750 words

Why Chunking and Embedding?



Text to Import

Cisco Catalyst Center is at the heart of Cisco's intent-based network architecture. Catalyst Center supports the expression of business intent for network use cases, such as base automation capabilities in the enterprise network. The Analytics and Assurance features of Catalyst Center provide end-to-end visibility into the network with full context through data and insights.

Cisco customers and partners can use the Catalyst Center platform to create value-added applications that leverage the native capabilities of Catalyst Center. Use Catalyst Center Intent APIs, Integration Flows, Events and Notification Services to enhance the overall network experience by optimizing end-to-end IT processes, reducing Total Cost of Ownership (TCO), and creating value-added networks.

779 Characters = 129 Tokens (Llama3)

Tokens visualized

Cisco Catalyst Center is at the heart of Cisco's intent-based network architecture. Catalyst Center supports the expression of business intent for network use cases, such as base automation capabilities in the enterprise network. The Analytics and Assurance features of Catalyst Center provide end-to-end visibility into the network with full context through data and insights.\n\nCisco customers and partners can use the Catalyst Center platform to create value-added applications that leverage the native capabilities of Catalyst Center. Use Catalyst Center Intent APIs, Integration Flows, Events and Notification Services to enhance the overall network experience by optimizing end-to-end IT processes, reducing Total Cost of Ownership (TCO), and creating value-added networks.

779 Characters = 129 Tokens (Llama3)

Large language models process the text using **tokens**.

Tokens visualized

Cisco Catalyst Center is at the heart of Cisco's intent-based network architecture. Catalyst Center supports the expression of business intent for network use cases, such as base automation capabilities in the enterprise network. The Analytics and Assurance features of Catalyst Center provide end-to-end visibility into the network with full context through data and insights.\n\nCisco customers and partners can use the Catalyst Center platform to create value-added applications that leverage the native capabilities of Catalyst Center. Use Catalyst Center Intent APIs, Integration Flows, Events and Notification Services to enhance the overall network experience by optimizing end-to-end IT processes, reducing Total Cost of Ownership (TCO), and creating value-added networks.

Token IDs

[92137, 88957, 5955, 374, 520, 279, 4851, 315, 43234, 753, 7537, 6108, 4009, 18112, 13, 88957, 5955, 11315, 279, 7645, 315, 2626, 7537, 369, 4009, 1005, 5157, 11, 1778, 439, 2385, 33762, 17357, 304, 279, 20790, 4009, 13, 578, 33527, 323, 84882, 4519, 315, 88957, 5955, 3493, 842, 4791, 13368, 24035, 1139, 279, 4009, 449, 2539, 2317, 1555, 828, 323, 26793, 382, 92137, 6444, 323, 8717, 649, 1005, 279, 88957, 5955, 5452, 311, 1893, 907, 87442, 8522, 430, 33164, 279, 10068, 17357, 315, 88957, 5955, 13, 5560, 88957, 5955, 9005, 34456, 11, 41169, 3061, 4336, 11, 18093, 323, 16980, 8471, 311, 18885, 279, 8244, 4009, 3217, 555, 74142, 842, 4791, 13368, 8871, 11618, 11, 18189, 10884, 11443, 315, 86036, 320, 7905, 46, 705, 323, 6968, 907, 87442, 14488, 13,]

Tokens visualized

Cisco Catalyst Center is at the heart of Cisco's intent-based network architecture. Catalyst Center supports the expression of business intent for network use cases, such as base automation capabilities in the enterprise network. The Analytics and Assurance features of Catalyst Center provide end-to-end visibility into the network with full context through data and insights.\n\nCisco customers and partners can use the Catalyst Center platform to create value-added applications that leverage the native capabilities of Catalyst Center. Use Catalyst Center Intent APIs, Integration Flows, Events and Notification Services to enhance the overall network experience by optimizing end-to-end IT processes, reducing Total Cost of Ownership (TCO), and creating value-added networks.

Token IDs

[92137, 88957, 5955, 374, 520, 279, 4851, 315, 43234, 753, 7537, 6108, 4009, 18112, 13, 88957, 5955, 11815, 279, 7645, 315, 2626, 7537, 369, 4009, 1005, 5157, 11, 1778, 439, 2385, 33762, 17357, 304, 279, 20790, 4009, 13, 578, 33527, 323, 84882, 4519, 315, 88957, 5955, 3493, 842, 4791, 13368, 24035, 1139, 279, 4009, 449, 2539, 2317, 1555, 828, 323, 26793, 382, 92137, 6444, 323, 8717, 649, 1005, 279, 88957, 5955, 5452, 311, 1893, 907, 87442, 8522, 430, 33164, 279, 10068, 17357, 315, 88957, 5955, 13, 5560, 88957, 5955, 9005, 34456, 11, 41169, 3061, 4336, 11, 18093, 323, 16980, 8471, 311, 18885, 279, 8244, 4009, 3217, 555, 74142, 842, 4791, 13368, 8871, 11618, 11, 18189, 10884, 11443, 315, 86036, 320, 7905, 46, 705, 323, 6968, 907, 87442, 14488, 13,]

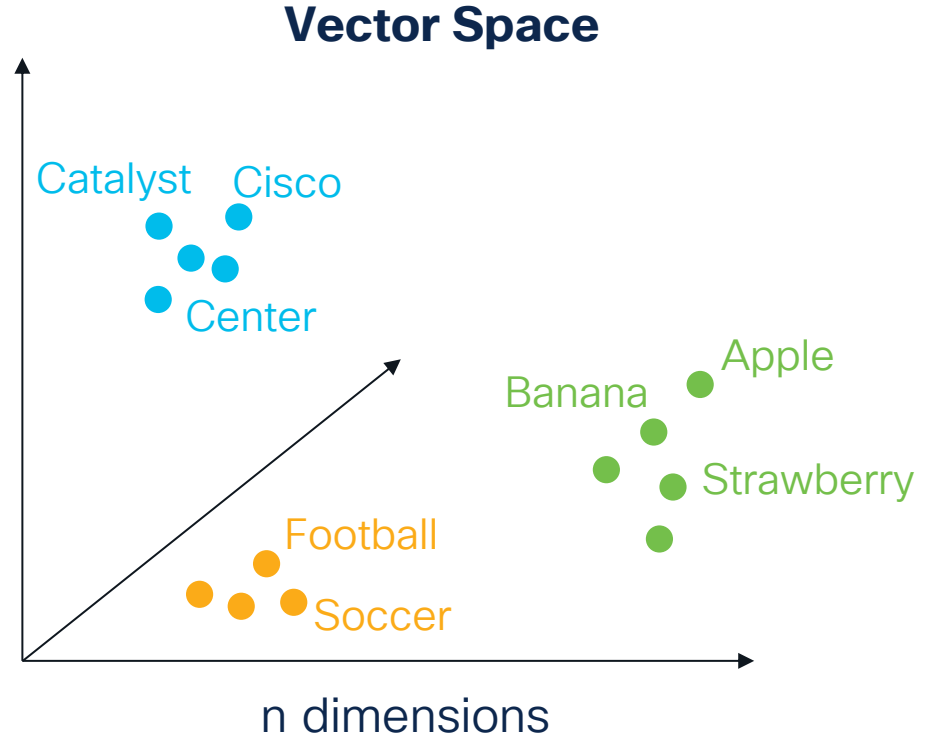
Vectorized Text

[-1.53614916e-02 1.05633754e-02 -5.66645004e-02 -6.19251188e-03 6.68779016e-02 2.10219994e-02 6.05151691e-02 7.23722875e-02 -4.16289344e-02 -2.71013863e-02 -1.01383314e-01 1.10760117e-02 -1.22514507e-02 6.36807233e-02 4.60679084e-02 1.85244009e-02 6.16239570e-02 -1.15199238e-01 -4.89140686e-04 -3.99950482e-02 3.34187062e-03 -7.70600662e-02 -5.15669882e-02 5.33867255e-03 -4.83192466e-02 6.42568767e-02 1.69622842e-02 -1.93460640e-02 -5.83380135e-03 -1.94340385e-02 -1.17660025e-02 7.47456923e-02 8.54562372e-02 2.81155687e-02 -6.18175901e-02 -5.65470643e-02 -2.26882417e-02 -4.52071428e-02 9.00947154e-02 -1.29134297e-01 3.32980417e-02 -1.89650655e-02 -5.85596003e-02 4.69980463e-02 -1.27475997e-02 -7.96920434e-02 -1.51178036e-02 -3.54162715e-02 -1.90531474e-03 2.22081002e-02 -1.67971440e-02 -4.96574230e-02 -4.80264649e-02 -1.39647434e-02 5.32995425e-02 9.77718979e-02 1.37339961e-02 3.94268101e-03 1.56628415e-02 -4.21853140e-02 6.75667301e-02 -6.72868863e-02 1.33521128e-02 6.92976639e-02 -6.00161292e-02 6.74986392e-02 4.00929861e-02 2.61038914e-02 -2.69671436e-02 -8.52481276e-02 -5.16354330e-02 -2.49810647e-02 -2.24567745e-02 2.44225748e-02 4.42139395e-02 3.41943502e-02 6.41698018e-02 -1.89136956e-02 -8.31962097e-03 -5.94300590e-02 1.74471382e-02 8.28464404e-02 -8.98656249e-02 5.64863719e-02 -1.93580631e-02 -2.60147806e-02 -3.21673192e-02 4.23833318e-02 3.81893367e-02 -3.74885425e-02 -4.21928205e-02 -4.86377999e-02 -2.01516785e-02 -8.35248902e-02 2.34446917e-02 -1.16874534e-03 -7.59318545e-02 -1.34604037e-01 -6.11120388e-02 3.29650603e-02 9.74610597e-02 3.29474136e-02 -2.93879658e-02 -1.68610681e-02 3.06183547e-02 5.11360317e-02 5.54264262e-02 -4.92100790e-02 -2.49358863e-02 -1.26581052e-33 -1.23462908e-01 4.33101431e-02 ... 7.64974654e-02 2.84792837e-02 7.45524839e-02 2.85661742e-02 7.72002265e-02 -1.83822196e-02]

Embedding function: **all-MiniLM-L6-v2** (384 dimensions)

Semantic search

- **Embeddings:** represent items (words, documents, images) in a continuous vector space
- Capturing the semantic and syntactic relationships between words

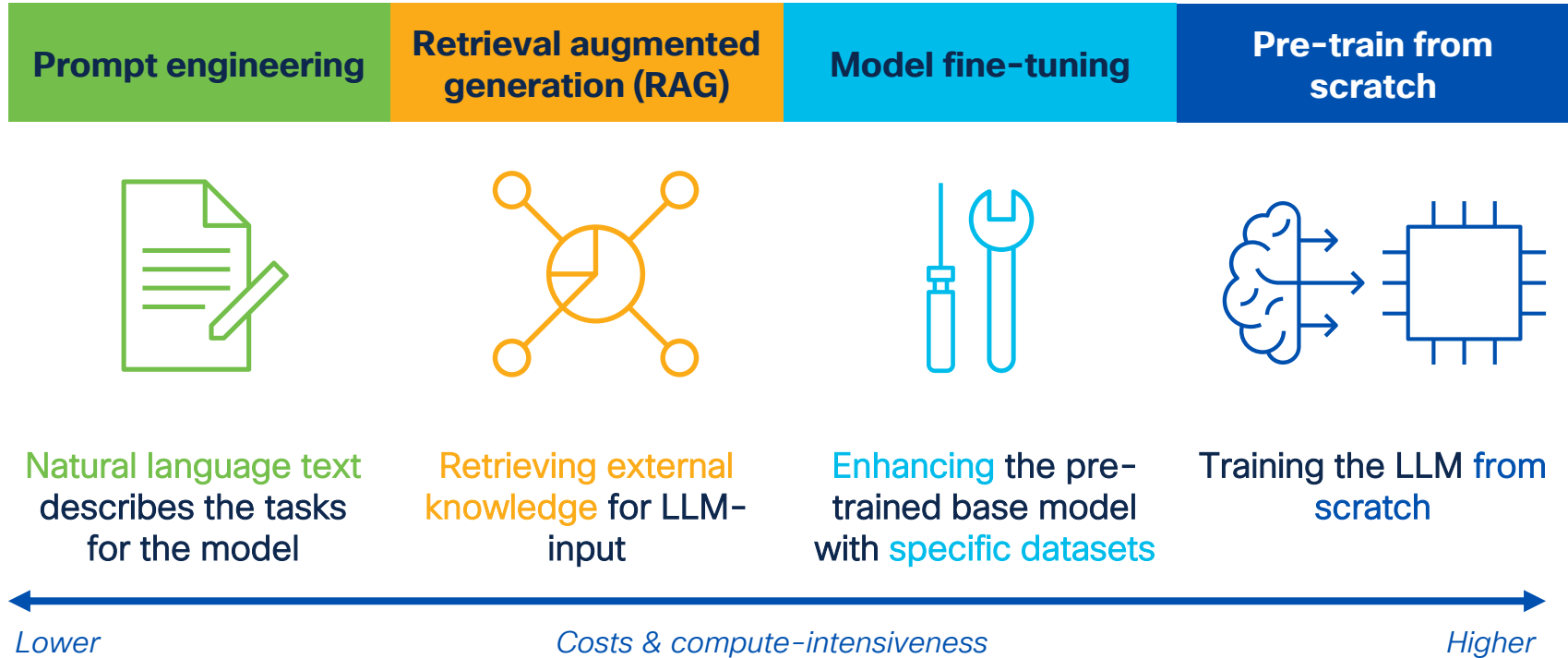


Retrieval Augmented Generation (RAG)

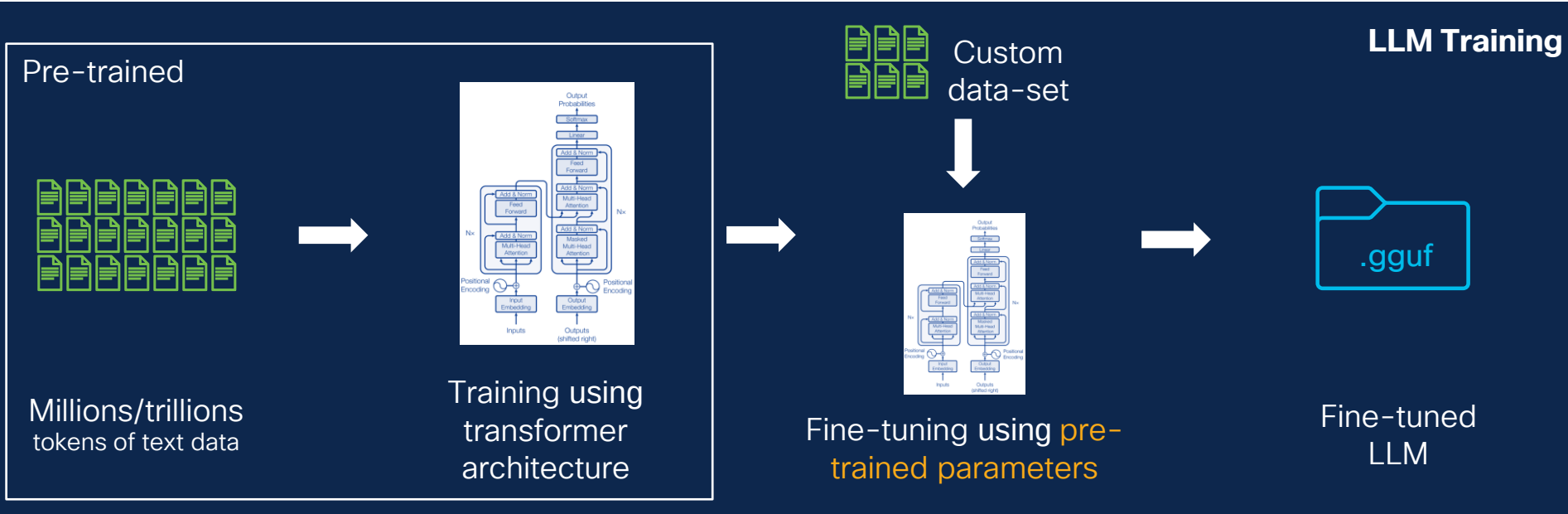
- ☑ Cost-efficient
- ☑ Easy to implement
- ☑ VectorDB Data control

- ✗ Dependence on external resources
- ✗ Requires additional infrastructure

GenAI architecture approaches



LLM fine-tuning example



Adapting a pre-trained *LLM* to a specific task by training it on domain-specific data.

Example: supervised fine-tuning

- **Goal:** Creating Ansible configuration files for specific use-cases
- Training on labeled dataset:

Collection	Description	Use-case	Playbook Code
Cisco.ios	Configuring access control lists (ACLs)	Block malicious traffic	--- - name: new ACLs hosts: cisco_ios gather_facts: no connection: network_cli tasks: - name: ...
Cisco.Meraki	Create or Update SSID on Meraki network	Renaming the wifi name	--- - name: Configure Meraki SSID hosts: localhost gather_facts: no connection: local tasks: - cisco.meraki.meraki_ssid:
...			

Model fine-tuning

- ☑ Cost-efficient on domain specific tasks
- ☑ A smaller fine-tuned LLM can outperform a larger LLM

- ✗ Data preparation is a time-consuming task
- ✗ Exposing the model with private data → privacy issues
- ✗ Whole model needs to be retrained when adding new data

Popular libraries/software to use (selection)

Name	Type	Description
Langchain, Llamaindex	Python Framework	Framework for developing LLM applications
ChromaDB, Pinecone, Pgvector	Vector Database	Storing data for a RAG architecture
Chainlit, Streamlit	Chat serving Frameworks	Front-end to chat with the LLM application
Ollama	Local LLM serving framework	Run LLMs locally
OpenAI APIs, Gemini, Groq, Mistral	Cloud LLM serving	LLM inferencing, fine-tuning
Hugging Face	GenAI/ML Community Platform	Access to models, datasets, peers etc.

Key learning points

- Mathematics & Statistics are **fundamental** for AI
- **GenAI is here to stay**. But it is **not the solution to everything**.
 - Generative AI: Generating data
 - Predictive AI: Analyzing existing data
- LLMs are dealing with **tokens & embeddings**
- 4 main GenAI architecture approaches
 - **Prompt engineering, RAG, fine-tuning, train from scratch**

Keynote Deep Dives

Wednesday
10:30am – 11:30am



Experiences Amplified:
How AI Can Fuel Better Employee and Customer Experiences

Level 1
Room 106



Smart, Secure, Seamless:
Transforming Experiences with Next-Generation Networking

Level 2
Room 204



Harness a Bold New Era:
Transform Data Centre and Service Provider Connectivity

Level 2
Room 203



Securing User to Application and Everything in Between

Level 2
Melbourne Room 2



Unlocking Digital Resilience through Unified Observability

The HUB
Centre Stage

Complete Your Session Evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a **Cisco Live T-Shirt**.



Complete your surveys in the **Cisco Live mobile app**.



Continue your education



- Visit the Cisco Stand for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand



Thank you

CISCO *Live!*

#CiscoLiveAPJC

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC