Exploring the Future of Multi-Modal AI: Integrating Text, Audio, Images, and Video

Prapanch Ramamoorthy - Principal Engineer
Patrick LeMaistre - Solutions Engineer
AIHUB-1970

Agenda

- Introduction
- Multi Modality – Text, Audio, Image and Video
  - Large Language models
  - Diffusion models
- Demo time
- Serving AI models on Cisco UCS and Nexus
- Multi modal embeddings and RAG
- Demo time

by way of **Introduction ...**

I am a **Principal Engineer** in Cisco TAC. I am 37 years old, with **15 years experience in Cybersecurity** and a strong focus in the security and networking spaces. Presently focused on developing Agentic workflows to help improve customer interactions.

I am a **Problem Solver**, a **Serviceability architect**, a **Teacher**, and an **Innovator**, in addition to being a husband and father (2 boys). I spend my spare time volunteering with organizations in the cybersecurity space.

Prapanch Ramamoorthy
Principal Engineer
praprama@cisco.com

by way of **Introduction ...**

I am a Compute **Solutions Engineer** in Cisco's Cloud and AI Infrastructure Group. I have been with Cisco for **25 years** focused on the areas of **AI, high performance computing, cloud, mass-scale infrastructure and mobility**.

I am the author of the **AI Toolkit** for UCS X-Series and C-Series which can be found on GitHub at: www.github.com/pl247/ai-toolkit

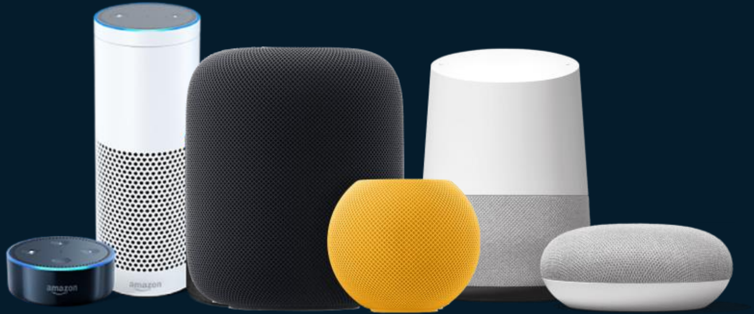
Prior to Cisco, I was co-founder and Vice President of Network Operations for Escape Communications, and hold a patent pertaining to Data Center Ethernet.

Patrick LeMaistre
Solutions Engineer
plemaist@cisco.com

Artificial Intelligence is Everywhere

Facial Recognition

Deep Learning



Voice Activated Assistants

Natural Language Processing

Chat Bots

Generative AI

A Snapshot of GenAI Capabilities …

Casual / Fun

R tell me a joke about the summer olympics in the style of Albert Einstein

"Why did the photon refuse to compete in the Summer Olympics?

Because it didn't want to be weighed down by mass and preferred to move at the speed of light, where time dilation makes every event feel like a lifetime!"

Cisco Enterprise Chat AI - powered by Azure OpenAI API.

R Help me write an RFP response why a customer would choose DNA Center

Dear [Customer],

Thank you for the opportunity to respond to your Request for Proposal (RFP) regarding the selection of a network management solution. After carefully evaluating your requirements and considering various options, we believe that Cisco DNA Center is the ideal choice for your organization. Allow

1. Unified Network Mana
entire network infrastr
This unified approach
overall efficiency.

2. Intent-Based Network
Networking, enabling
than manually configu

New Chat Ask me anyt

Cisco Enterprise Chat AI - powered by Azure OpenAI API.

R can you write a python script to add a VLAN and SVI to a catalyst switch using the python library paramiko?

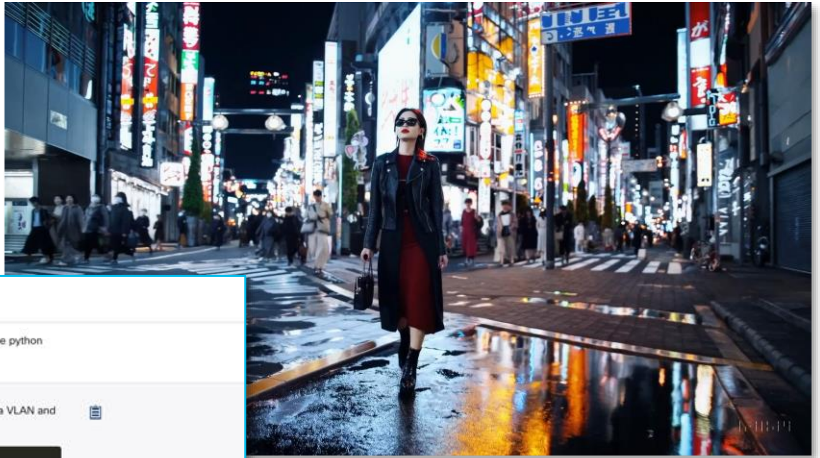
Certainly! Here's an example Python script that uses the Paramiko library to add a VLAN and SVI (Switched Virtual Interface) to a Cisco Catalyst switch:

```
import paramiko

# Define the switch connection details
hostname = 'your_switch_hostname'
username = 'your_username'
password = 'your_password'

# Define the VLAN and SVI details
vlan_id = '10'
vlan_name = 'VLAN10'
svi_ip = '192.168.1.1'
```

New Chat Ask me anything...

BUSINESS –
Answering RFPs,
Writing code …

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.



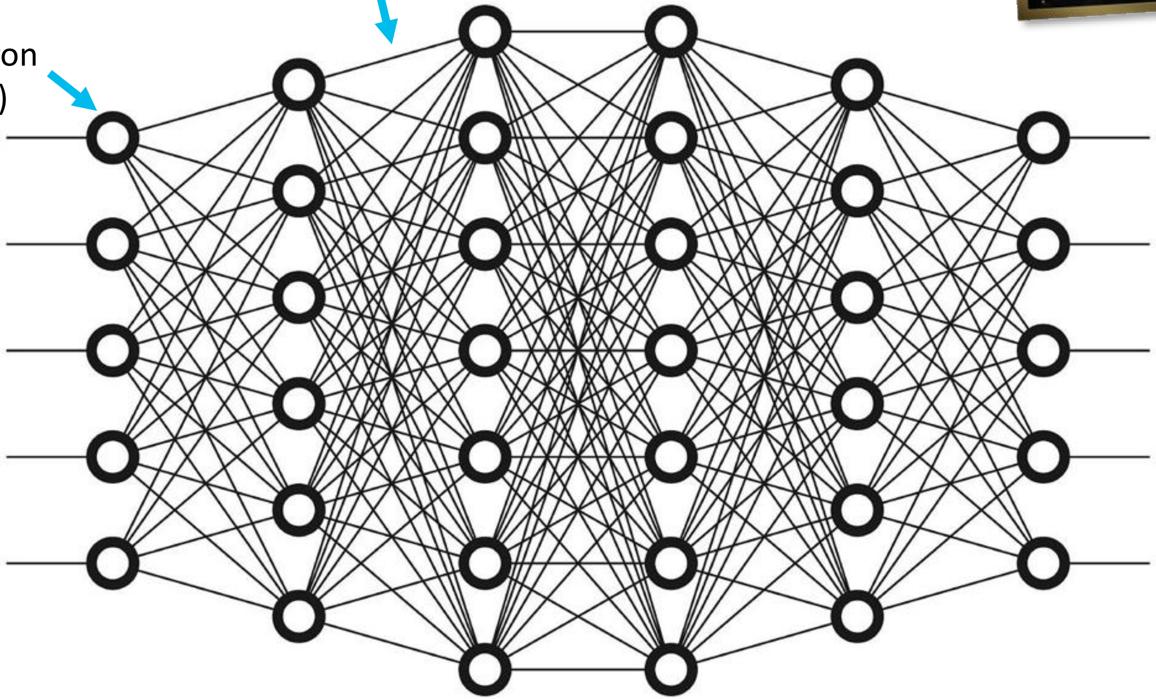
MULTIMEDIA –
Text, Pictures, Video, …

Why is this happening **now?**

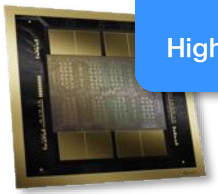
Scale

Perceptron
(Neuron)

Parameter
(Synapses)



NVIDIA
Blackwell



Advances in Silicon –
High-density, High-performance GPUs

NVIDIA Flagship Accelerator Specification Comparison

	B200	H100	A100 (80GB)
FP32 CUDA Cores	A Whole Lot	16896	6912
Tensor Cores	As Many As Possible	528	432
Boost Clock	To The Moon	1.98GHz	1.41GHz
Memory Clock	8Gbps HBM3E	5.23Gbps HBM3	3.2Gbps HBM2e
Memory Bus Width	2x 4096-bit	5120-bit	5120-bit
Memory Bandwidth	8TB/sec	3.35TB/sec	2TB/sec
VRAM	192GB (2x 96GB)	80GB	80GB

Geoffrey Hinton – the "Godfather" of Deep Learning

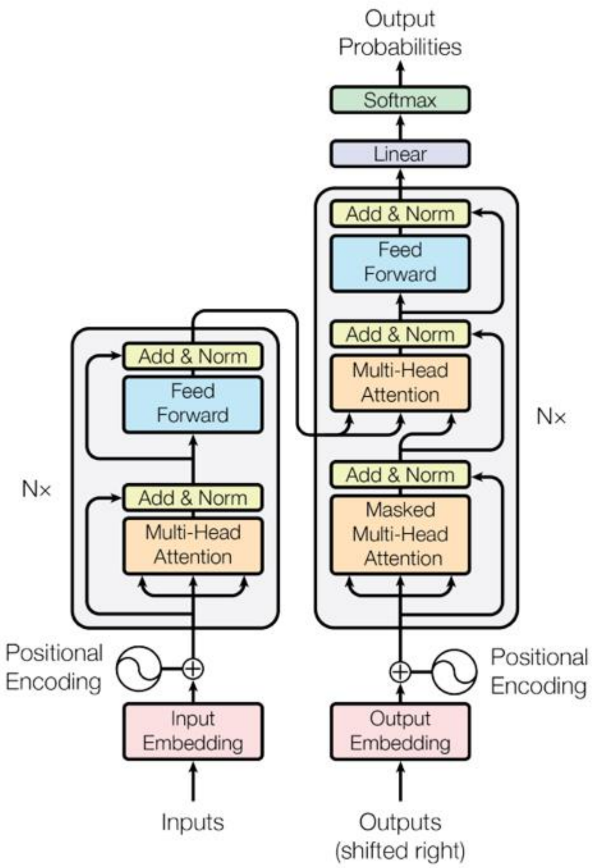


Figure 1: The Transformer - model architecture.

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

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

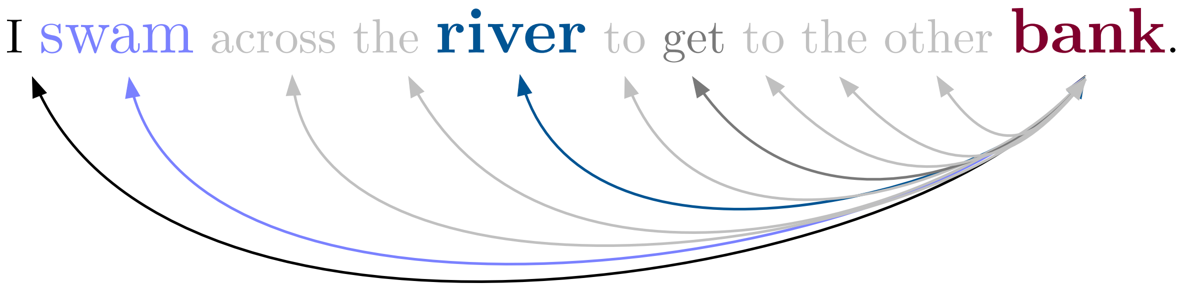
arXiv:1706.03762 **[cs.CL]**

Attention Mechanism – Overview

You have no problem interpreting “bank” in the following sentence:

“I swam across the river to get to the other bank.”

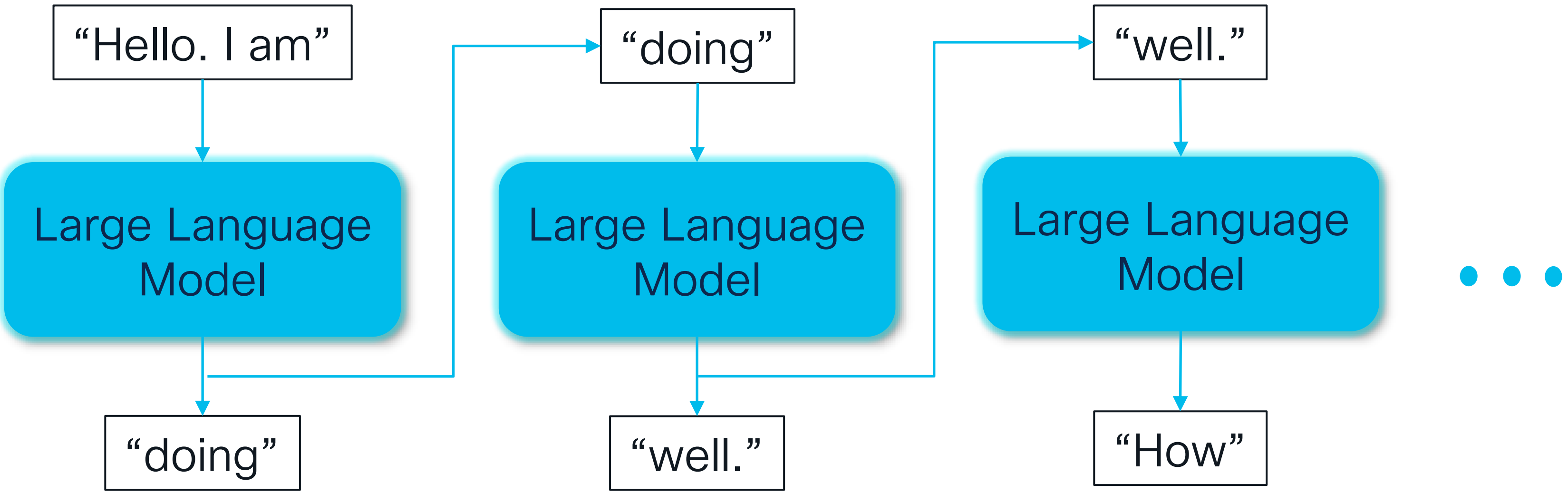
A machine needs some help…



The goal of the attention mechanism is to add **contextual information** to words in a sentence.

Text generation models - LLMs

Given a sequence of “tokens”, generate the subsequent token



Audio, Image and Video Generation

How are Diffusion Models Trained?

Step 1: Data Collection
(Feeding Knowledge)

Step 2: Noise Addition
(Disrupting the Image)

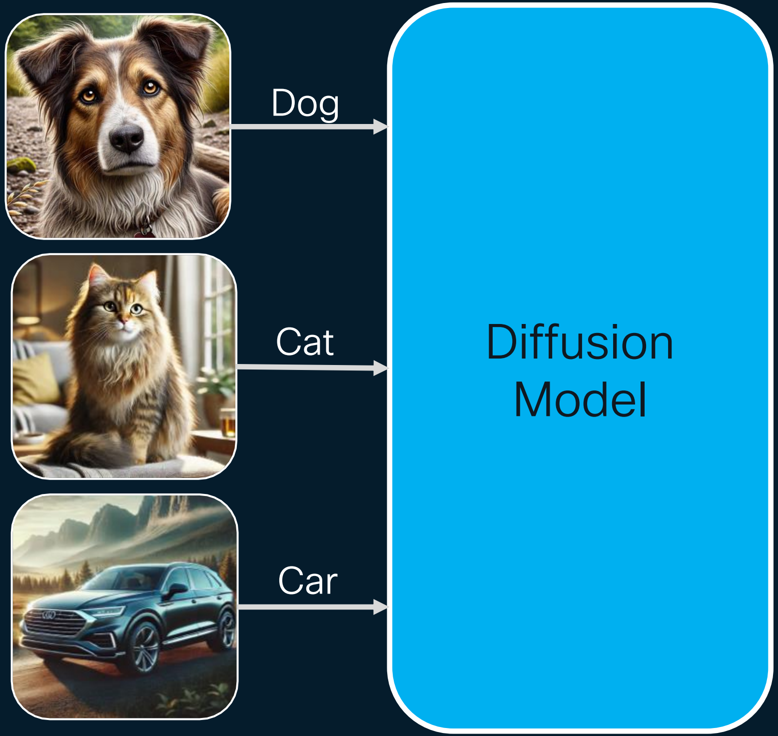
Step 3: Denoising
(Restoring the Image)

Step 1: Data Collection (Feeding Knowledge)

What Happens?
- Feel model large datasets of images paired with descriptive text to learn.

Analogy:
- Given a person photos with captions they learn to associate certain text with specific visual patterns.

Fun Fact: DALL·E is trained on billions of images with text descriptions.

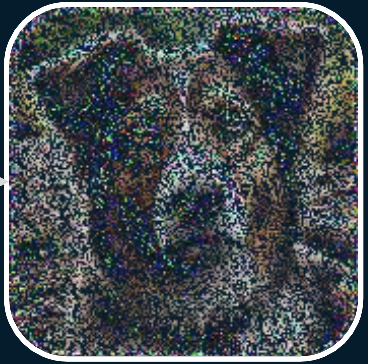


Step 2: **Add Noise** (Disrupting the Image)

What Happens?
- The model adds noise to the image until it's unrecognizable.

Analogy:
- Crumpling a piece of so it's harder to recognize the original content.

Original



Noise Added

Fun Fact: The noise used in diffusion models isn't random—it's carefully structured to ensure that when the model learns to denoise, it's learning how to generate fine details.

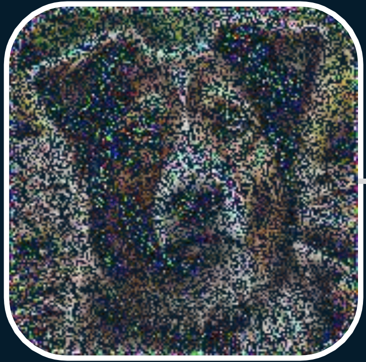
Step 3: **Denoising** (Restoring the Image)

What Happens?

- The model attempts to reconstruct the noisy image to its original form while learning in the process.

Analogy:

- Learning to draw the artwork from that crumpled paper.



Noise Added

Denoised

How Diffusion Model Converts Text to Image

Step 1: Understanding the Prompt

Step 2: Noise Initialization

Step 3: Denoising with Text

Step 4: Image Refinement

Step 1: Understanding the Prompt

Prompt: Generate an image of a dog on a flying carpet!

Tokenized Text

[“Generate", “an”, “image”, “of”, “a”, “dog”, “on”, …]

Vectorized Tokens

[0.23, -0.56, 1.04, 0.78, 0.99, 1.21, 0.89, 0.32, ...]

Neural Network Layers

Analyzes and understands the meaning and context

Step 2: Noise Initialization

Prompt: Generate an image of a dog on a flying carpet!

How It Works

- Start with a random noisy image.
- Text encoding guides the process using learned text-image pairs.

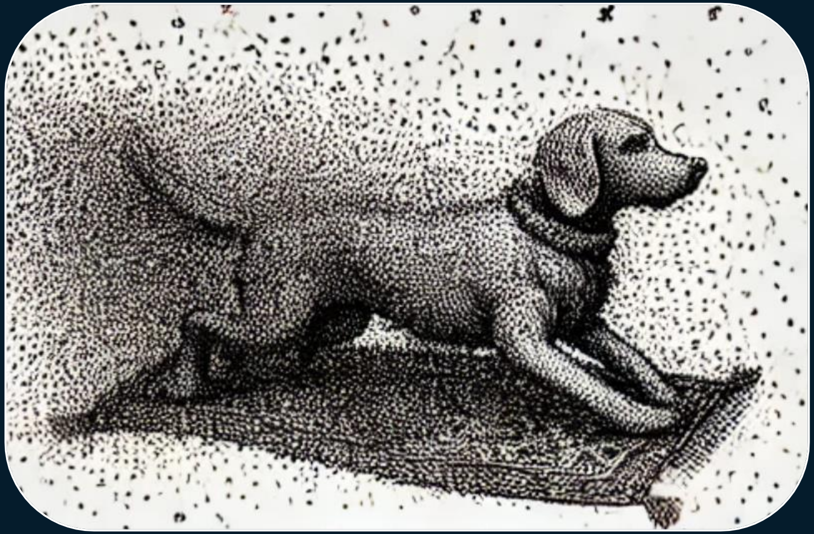
Step 3: Denoising with Text Guidance

Prompt: Generate an image of a dog on a flying carpet!

How It Works

- The model removes noise using patterns learned from training.

- The shapes of the dog and carpet become clearer.



Step 4: Refining the Image

Prompt: Generate an image of a dog on a flying carpet!

How It Works

- Noise is entirely removed.
- Final image of the dog flying on the magic carpet is complete.

Demo Time!

Serving AI Models
on Cisco UCS and Nexus

cisco Live!

On-Prem AI Solutions:
Business Challenges of Proprietary GenAI Systems

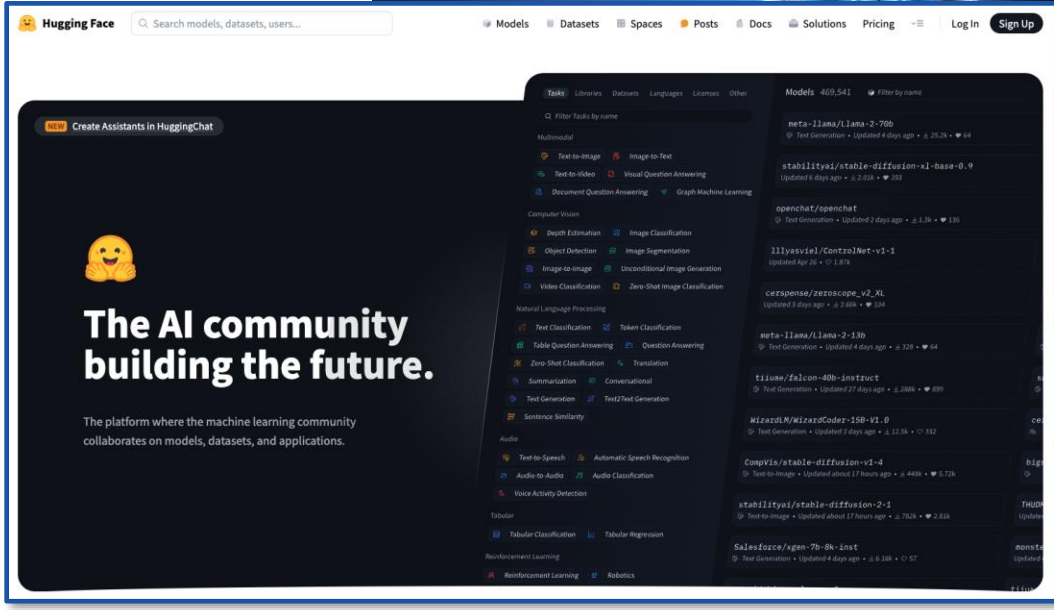
1. **Cost:** Cloud-based GenAI solutions are becoming a lucrative source of recurring revenue for OpenAI, Google, Anthropic, etc.
2. **Privacy and Security:** Commercially available LLMs are something of a “black box” for users. The users have no control over how they were trained, bias, etc.
3. **Training Gap:** Models are usually months or years out of date. Requires Retrieval Augmented Generation (RAG)
4. **Fine Tuning:** Cloud-based LLMs are “foundational models” and lack fine tuning for certain use cases

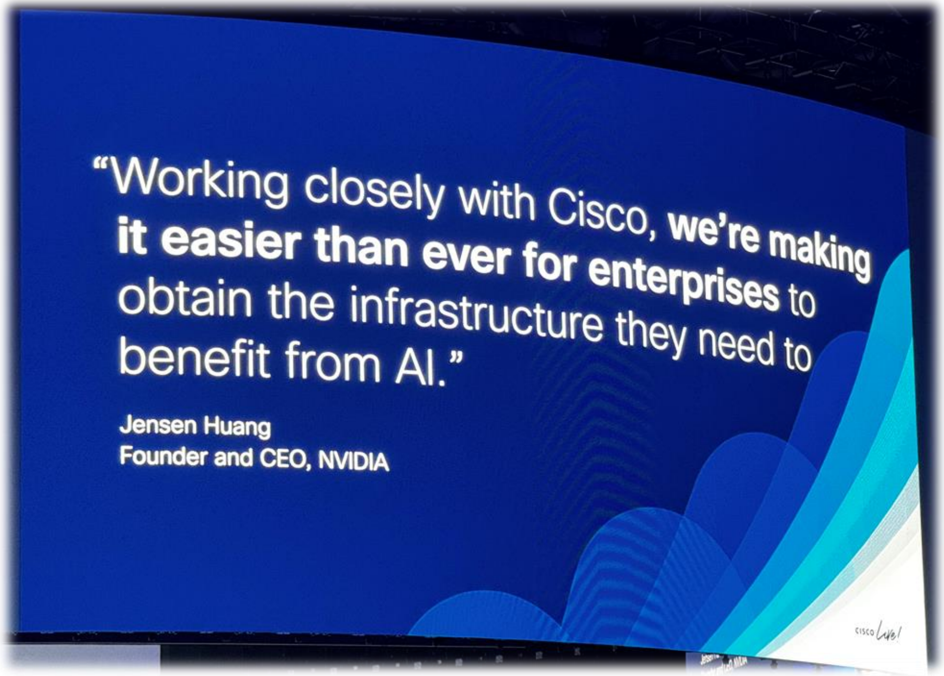
Meta has a different Approach – open source AI Llama (Large Language Model Meta AI)

- Llama (Large Language Model Meta AI) is an LLM released by Meta AI in February 2023
- Multimodal LLaMA 3.2 released September 2024
    - Medium sized vision LLMs (11B and 90B)
    - Image reasoning (graph or map reading)

Meta has a different Approach – open source AI Llama (Large Language Model Meta AI)

- Llama (Large Language Model Meta AI) is an LLM released by Meta AI in February 2023
- Multimodal LLaMA 3.2 released September 2024
  - Medium sized vision LLMs (11B and 90B)
  - Image reasoning (graph or map reading)





UCS C885A – 8 x GPU Server

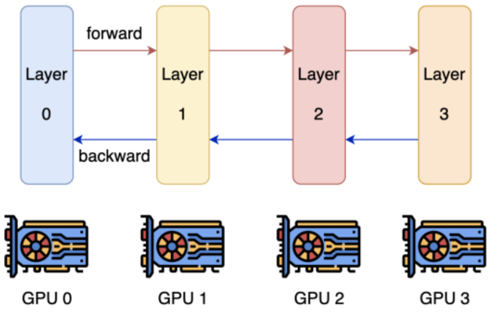
Nexus AI Networking

UCS-X Series – Modular Chassis

Splitting AI Model Across GPUs

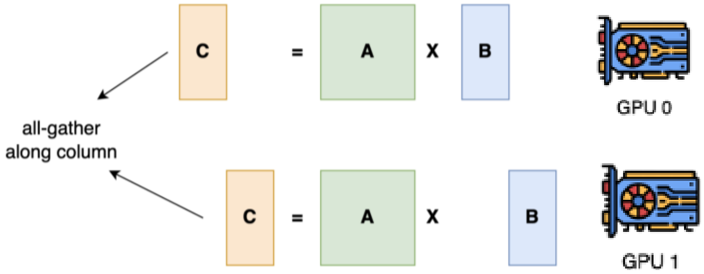
Pipeline Parallelism

Different layers to different GPUs. Usually on GPU on different hosts
(lower bit rate)



Tensor Parallelism

Layer computations sent to different GPUs. Best for multi-GPU single host
(higher bit rate)



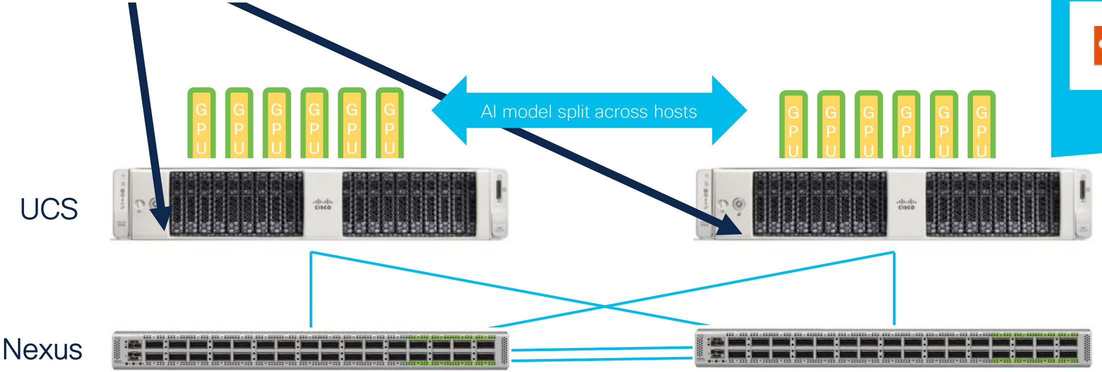
Demo Time!

Demo
Multi-GPU Multi-Host

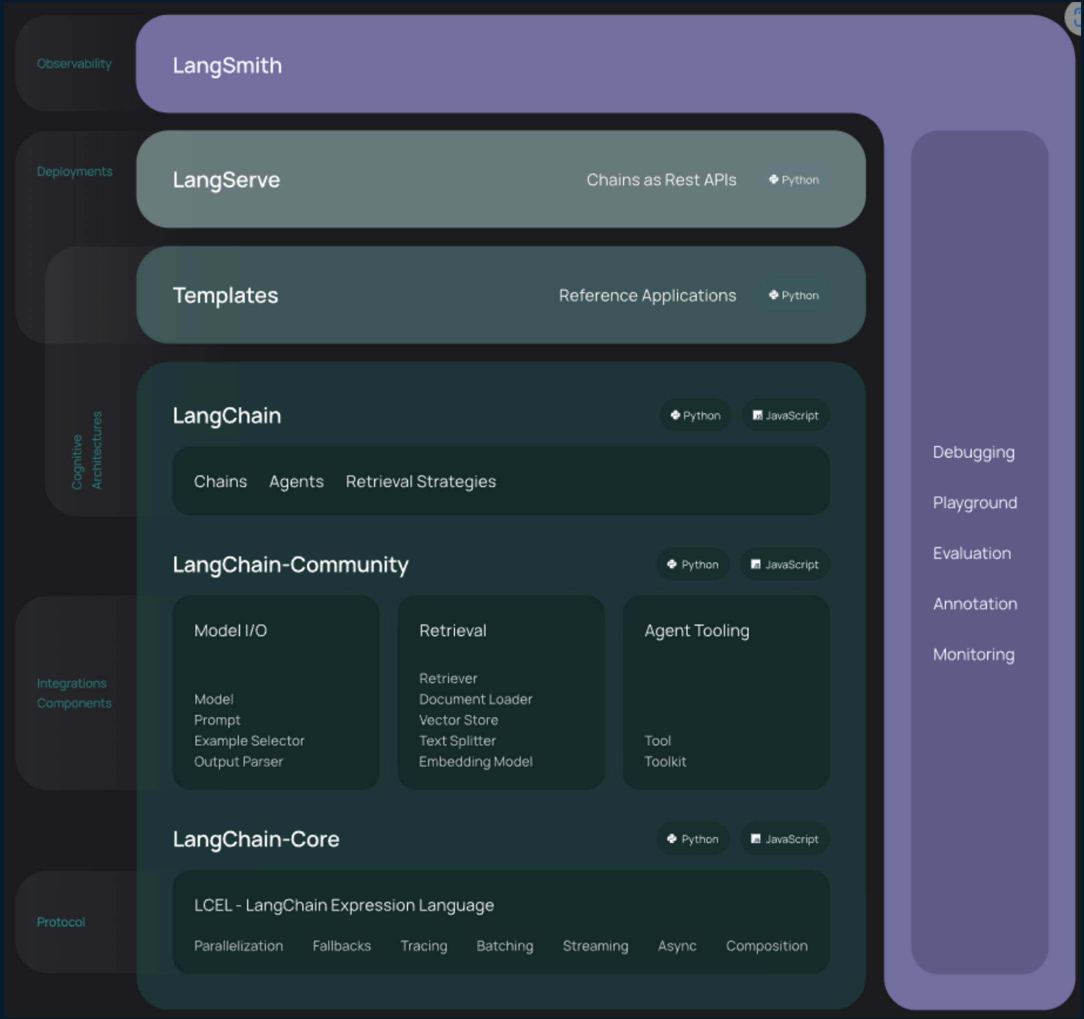
OI	Open WebUI
vLLM	LLM Server (OAPI)
Hugging Face	Models → Meta
RAY	Clustering Framework
NVIDIA AI Enterprise	GPU Driver & CUDA
Ubuntu	Operating System

Intersight

- Deploy Server Profile
- OS (ubuntu) & AI Tools



Multi-model Embeddings and RAG


Observability
LangSmith
Deployments
LangServe
Chains as Rest APIs
Python
Templates
Reference Applications
Python
Cognitive Architectures
LangChain
Python
JavaScript
Chains
Agents
Retrieval Strategies
LangChain-Community
Python
JavaScript
Integrations Components
Model I/O
Model
Prompt
Example Selector
Output Parser
Retrieval
Retriever
Document Loader
Vector Store
Text Splitter
Embedding Model
Agent Tooling
Tool
Toolkit
LangChain-Core
Python
JavaScript
Protocol
LCEL - LangChain Expression Language
Parallelization
Fallbacks
Tracing
Batching
Streaming
Async
Composition
Debugging
Playground
Evaluation
Annotation
Monitoring

An open-source framework for AI development

- Python or JavaScript
- Highly abstracted
- Easy
- Accessible

Integrated openAI

Options for ***private LLMs***

Evolving quickly and daily

“Like developing on wet cement”

LCEL (Langchain Expression Language)

Basic RAG

Retrieval

- We will retrieve data from a vector store with relevant records from, for example, a semantic search (top result returned)

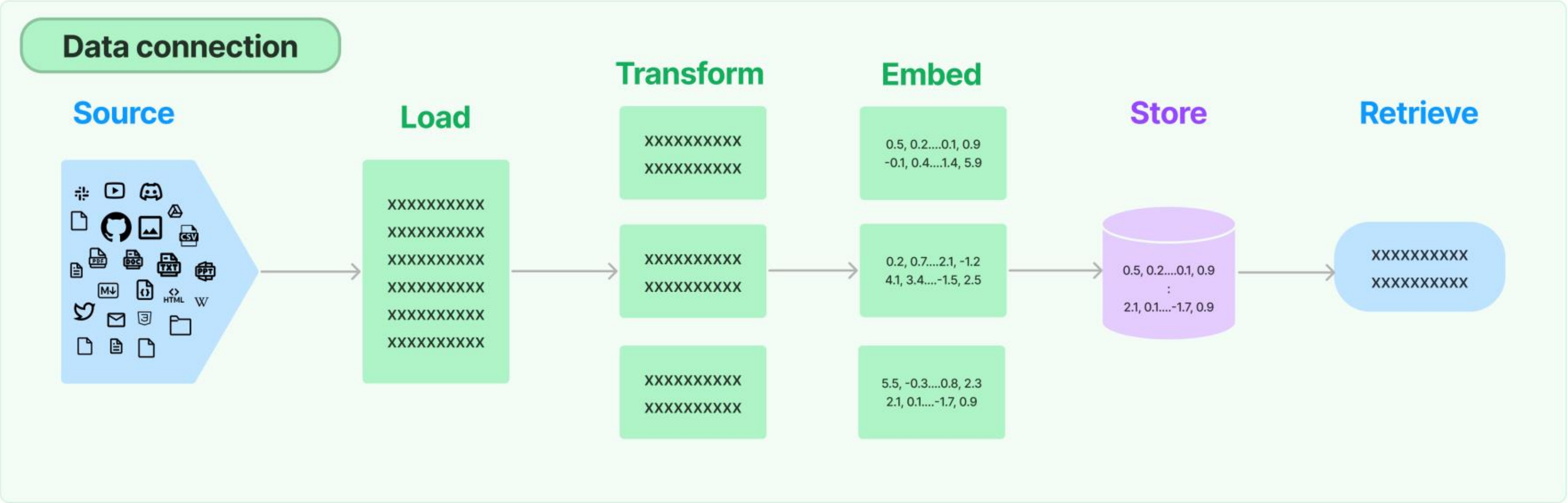
Augmented

- We will *augment* our **prompt** by supplementing the relevant information
- Often in the form of JSON

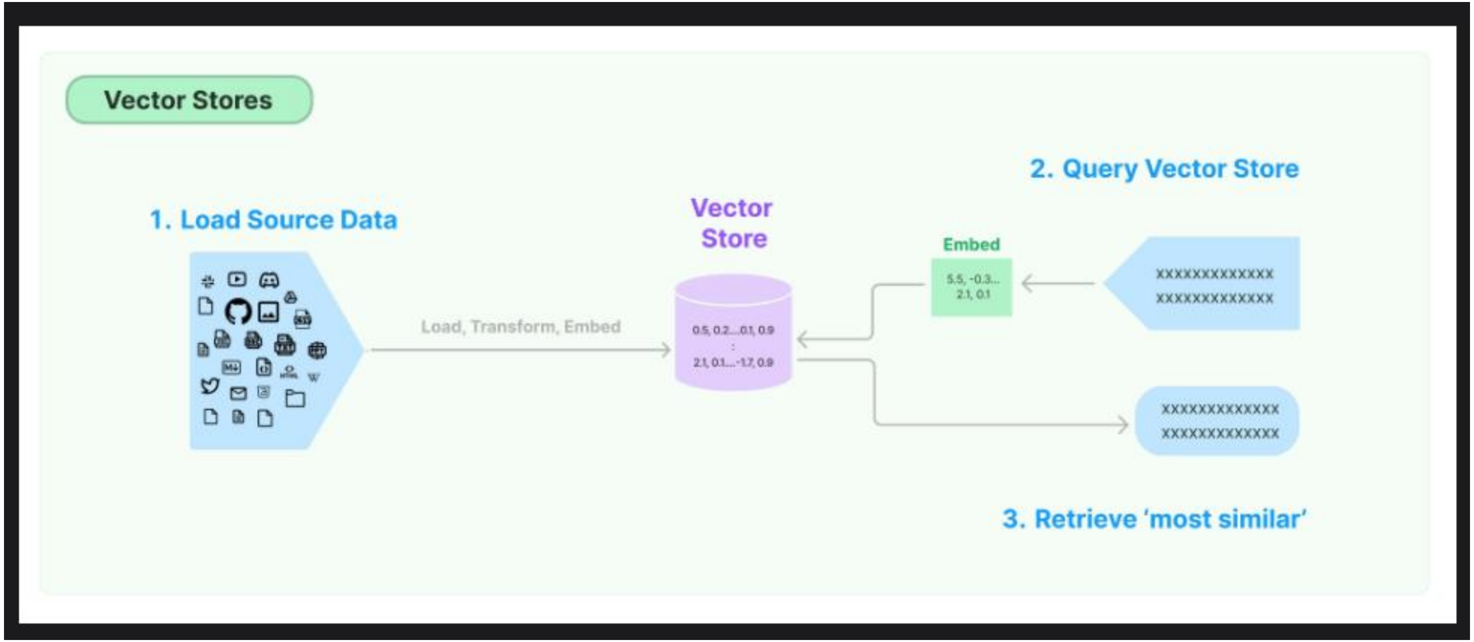
Generation

- Generate the response taking both the prompt and the contextual data into consideration

A high-level LangChain pipeline abstraction



Vector Store



```python
#Load OpenAI API Key
os.environ["OPENAI_API_KEY"] = openai.api_key

#Setup Loader - in this case a PDF Loader
loader = PyMuPDFLoader("cisco-annual-report-2022.pdf")

# Load and split the pdf into pages
pages = loader.load_and_split()

# setup a text splitter
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=250,
    chunk_overlap=20,
    length_function=len,
)
```

```python
# split the pages into smaller chunks called docs
docs = text_splitter.split_documents(pages)

# transform to embeddings
embeddings = OpenAIEmbeddings()

# setup and store docs and embeddings into ChromaDB
vectordb = Chroma.from_documents(docs, embedding=embeddings,
                                 persist_directory=".")

#Make the database persisten
vectordb.persist()

# setup memory so it remembers previous questions and answers
memory = ConversationBufferMemory(memory_key="chat_history", return_messages=True)
```

```python
# perform the Conversational Retreival Chain
pdf_qa = ConversationalRetrievalChain.from_llm(OpenAI(temperature=0.5) , vectordb.as_retriever(), memory=memory)

#Run the question
question = "Who is the senior leadership team at Cisco?"
result = qa.run(question)

#Print the results to the screen
print(result)
```

Demo Time!

Webex App

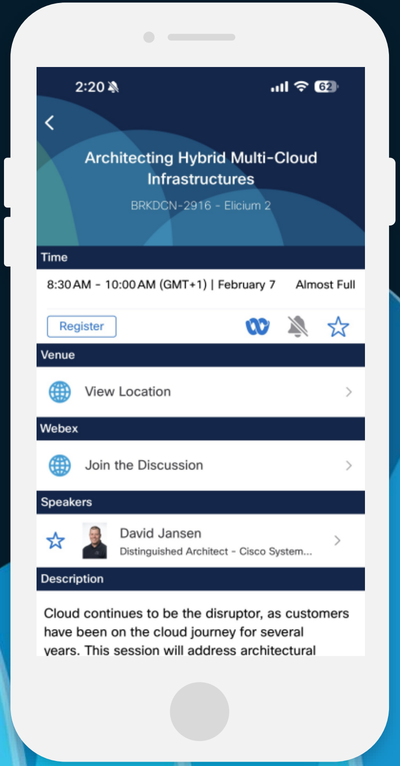
Questions?

Use the Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events mobile app
2. Click “Join the Discussion”
3. Install the Webex app or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until February 28, 2025.



Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

All surveys can be taken in the Cisco Events mobile app or by logging in to the Session Catalog and clicking the 'Participant Dashboard'

Content Catalog

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at ciscolive.com/on-demand. Sessions from this event will be available from March 3.

Thank you

GO **BEYOND**