

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC



Unlock the Future:

Introduction to GenAI and Retrieval Augmented Generation (RAG) techniques for your networks

Patrick LeMaistre

Solutions Engineer

Prapanch Ramamoorthy

Principal Engineer

Dave Zacks

Distinguished Engineer

BRKAPP-2562

CISCO *Live!*

#CiscoLiveAPJC



Unlock the Future:

Introduction to GenAI and Retrieval Augmented Generation (RAG) techniques for your networks

Patrick LeMaistre

Solutions Engineer

Prapanch Ramamoorthy

Principal Engineer

Dave Zacks

Distinguished Engineer

BRKAPP-2562

CISCO *Live!*

#CiscoLiveAPJC

#HighBitRate



Cisco Webex App

Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live 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 November 15, 2024.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKAPP-2562>



Agenda

- **Introduction**

- AI/ML, Generative AI, Neural Networks, Langchain ...



- **RAG – why and what?**

- Hallucinations and context
- Retrieval, Augmentation, Generation
- Demo Time!



- **Advanced RAG**

- RAG Fusion, RAPTOR

- **On-Prem AI Solutions**

- Open Source LLMs w/ UCS and Nexus
- Demo Time!



- **Wrap-up**

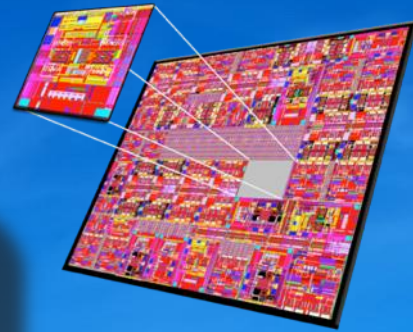
by way of Introduction ...

I am a **Distinguished Engineer** in the Network Transformation team, and have been with Cisco for 25 years.

I work primarily with large, high-performance Enterprise network architectures, designs, and systems. I have over 39 years of experience with designing, implementing, and supporting solutions with many diverse network technologies.

I have a strong background in, and focus on, customer requirements, and integrating these into the products and solutions Cisco builds.

I have a special interest in **Flexible Hardware, Fabrics, Assurance and ML/AI**.



Dave Zacks

Distinguished Engineer
dzacks@cisco.com

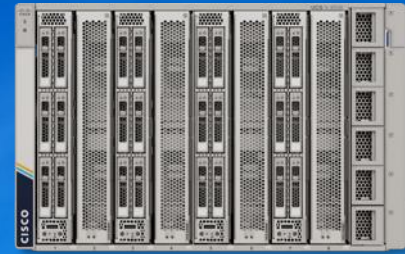


by way of Introduction ...

I am a **Compute Solutions Engineer** in Cisco's Cloud Infrastructure and Software 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

by way of Introduction ...

I am a **Principal Engineer** in Cisco TAC. I am 35 years old, with **13.5 years experience in Cybersecurity** and a strong focus in the security and networking spaces.

I am a **Problem Solver**, a **Serviceability** architect, a **Teacher**, and an **Innovator**, in addition to being a husband and father (2 boys).

Fun facts:

Love eating and trying out new things

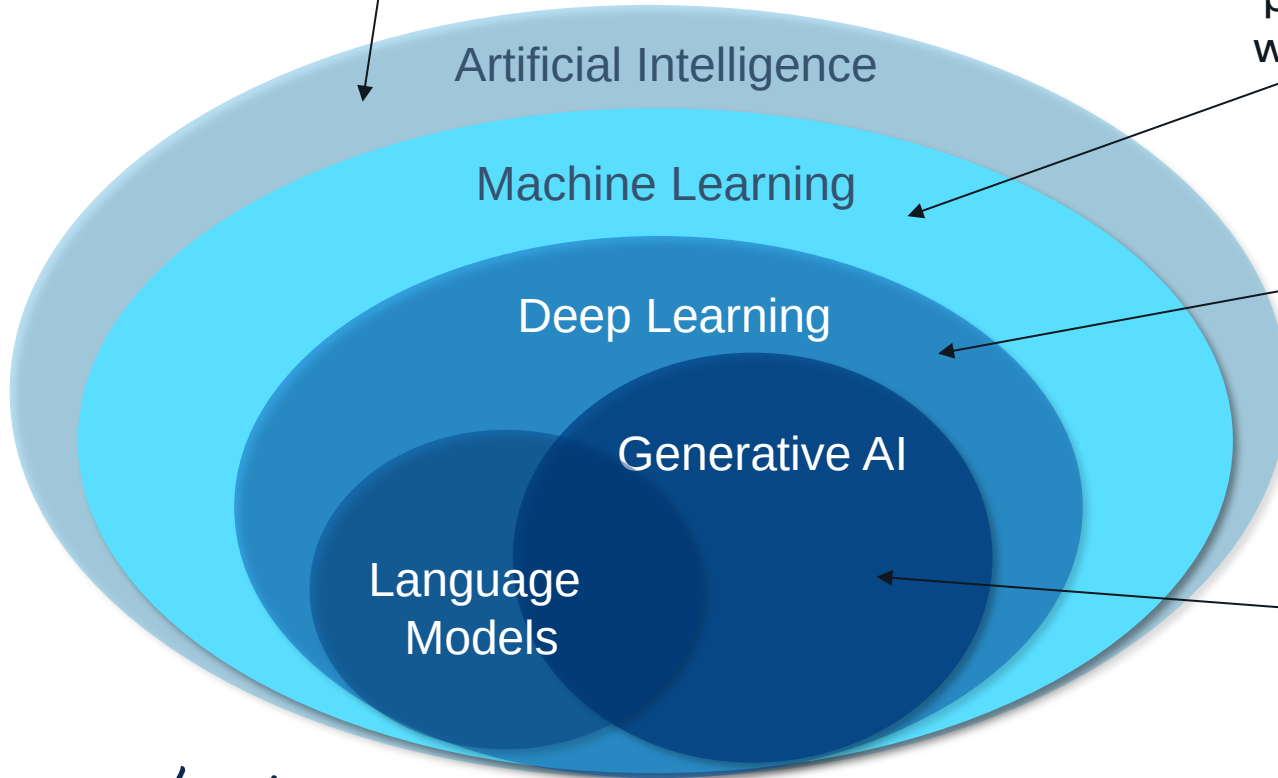
Social media phantom

Prapanch Ramamoorthy
Principal Engineer
praprama@cisco.com



AI encompasses the whole world
of ML and Deep Learning

An AI technology where the
rules are not set in the
program, but are **learned**
while the program is used



A form of ML that uses
**Neural Learning
Networks**

A powerful mechanism
that allow neural
networks to
**learn language and
generate content**

A Snapshot of GenAI Capabilities ...

Casual / Fun

R

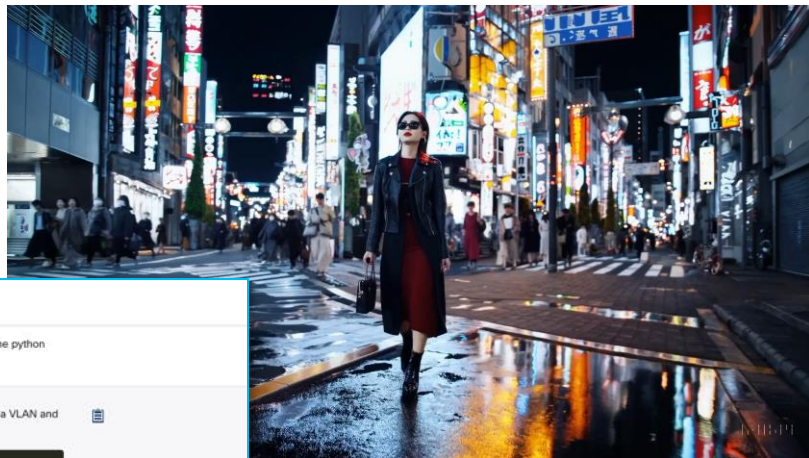
tell me a joke about the summer olympics in the style of Albert Einstein

G

"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!"

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.



Cisco Enterprise Chat AI - powered by Azure OpenAI API.

R

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

G

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 Management
entire network infrastructure. This unified approach overall efficiency.
2. Intent-Based Network
Networking, enabling than manually configuring

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?

G

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'
```

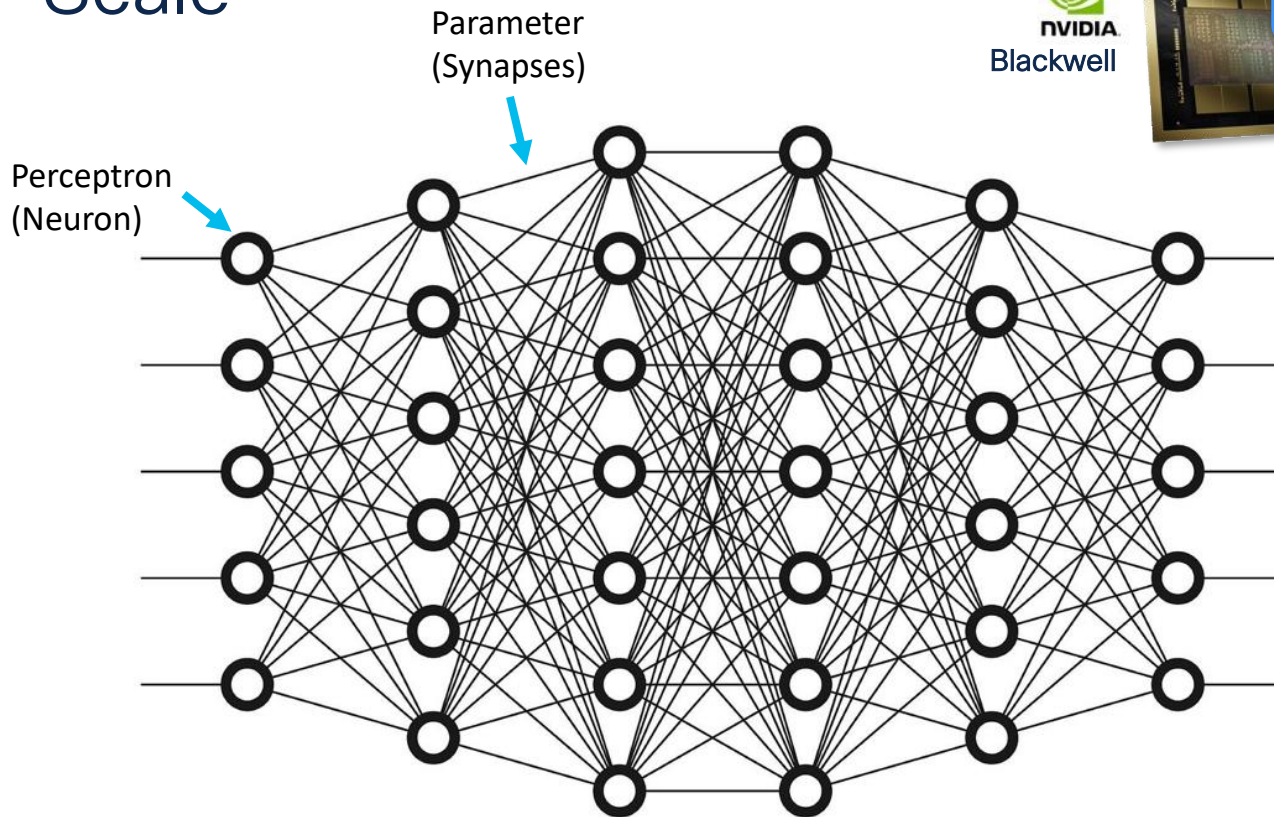
BUSINESS -
Answering RFPs,
Writing code ...

MULTIMEDIA -
Text, Pictures, Video, ...

cisco Live!

Why is this happening now?

Scale



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

cisco Live!

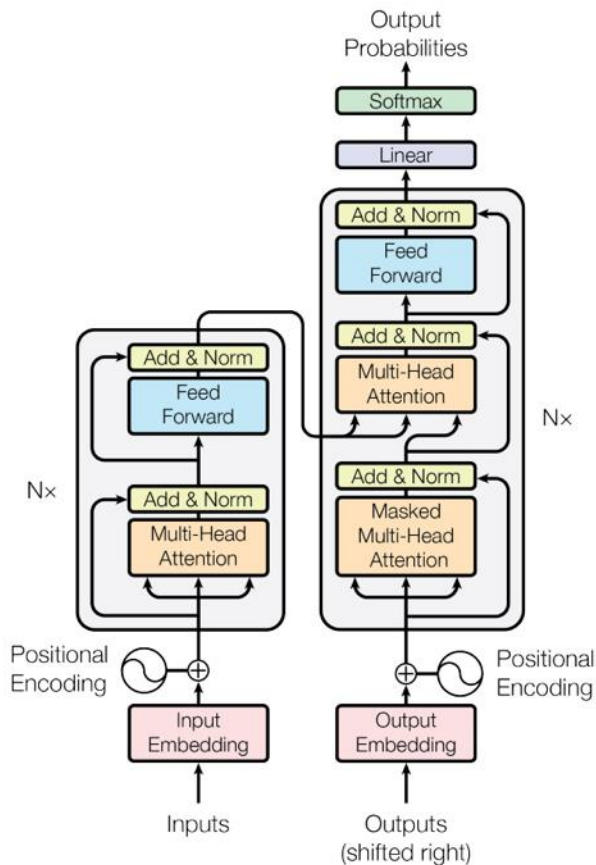


Figure 1: The Transformer - model architecture.

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](https://arxiv.org/abs/1706.03762) [cs.CL]

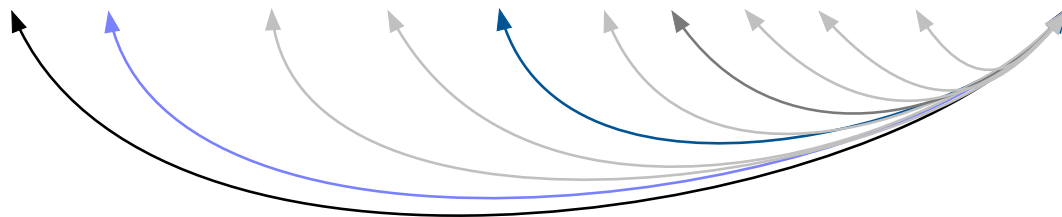
Attention Mechanism – Overview

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

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

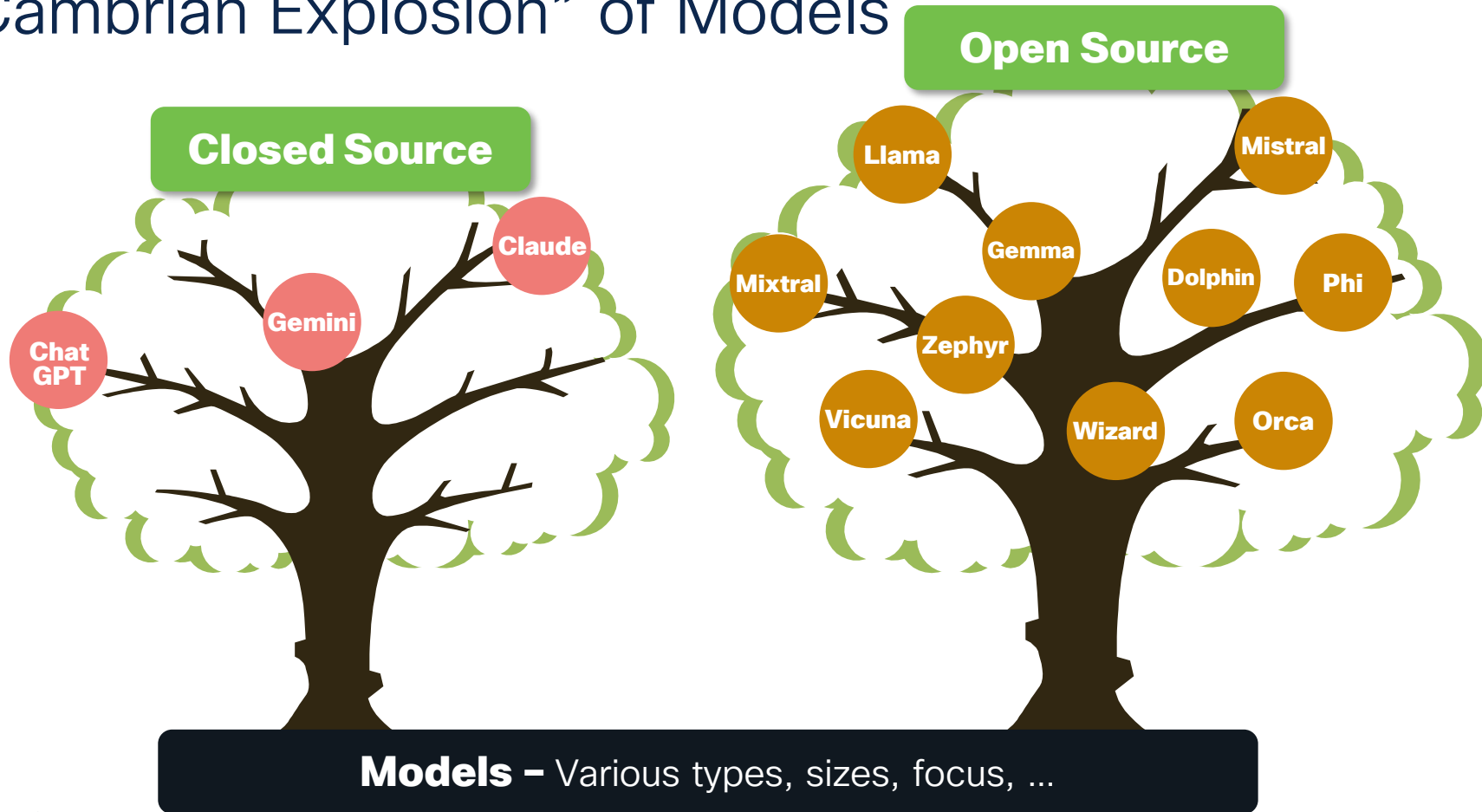
A machine needs some help...

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

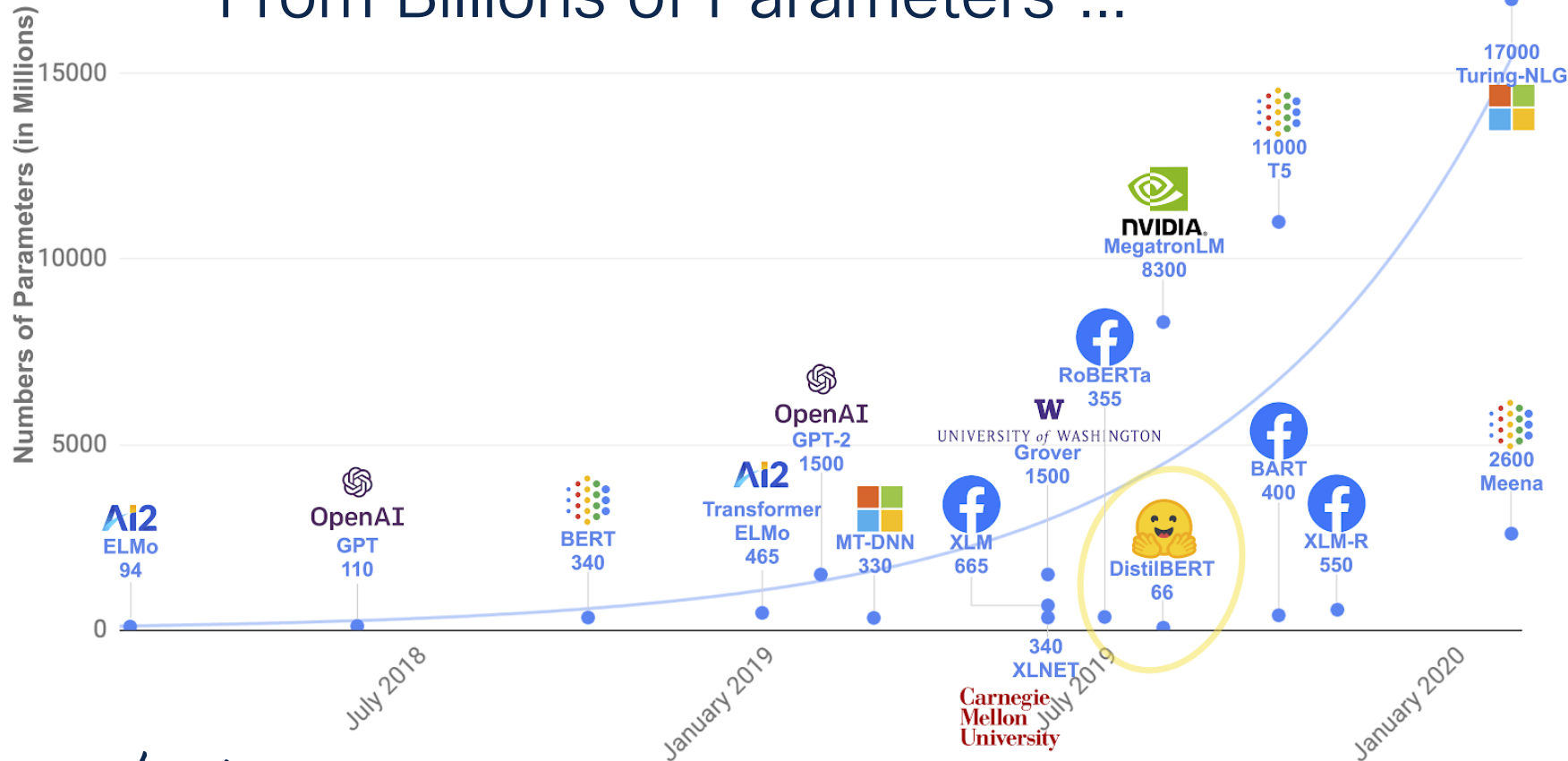


The goal of the attention mechanism is to add **contextual information** to words in a sentence.

“Cambrian Explosion” of Models



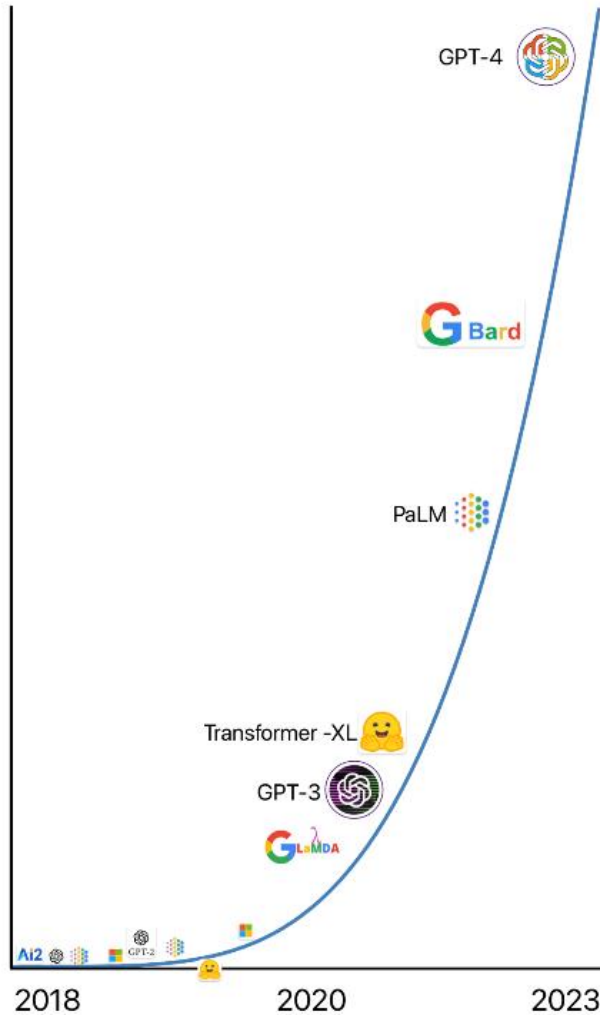
From Billions of Parameters ...



To Trillions ...

1,000,000,000,000

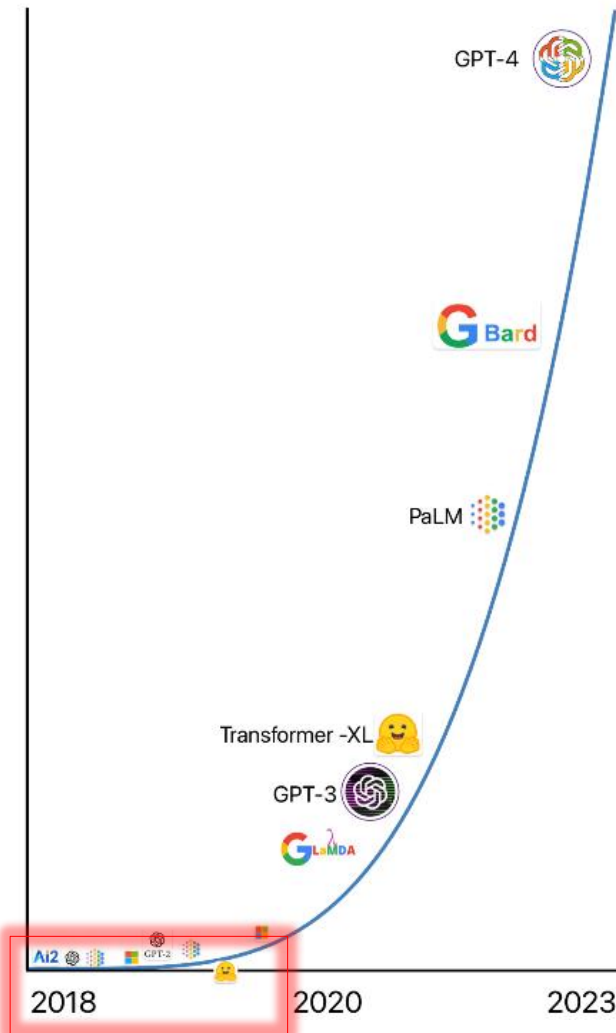
500,000,000,000



To Trillions ...

1,000,000,000,000

500,000,000,000



FUN FACT!

The human brain contains **86 billion neurons**, and over **100 trillion synaptic connections**



AI Compute

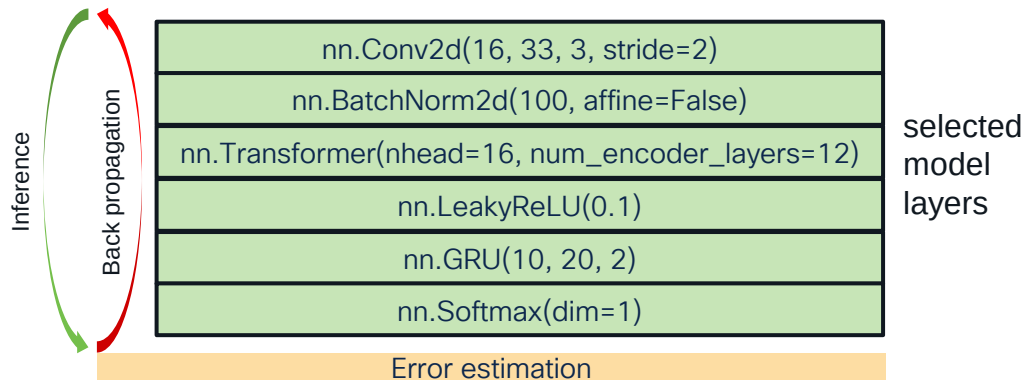
AI is a new programming paradigm – not just an optimality-vs-efficiency trade-off

Key characteristics:

- High parallelism
- Algorithm described by data-set, not procedure
- Coarse-grained selection of standard “layers” to compose network



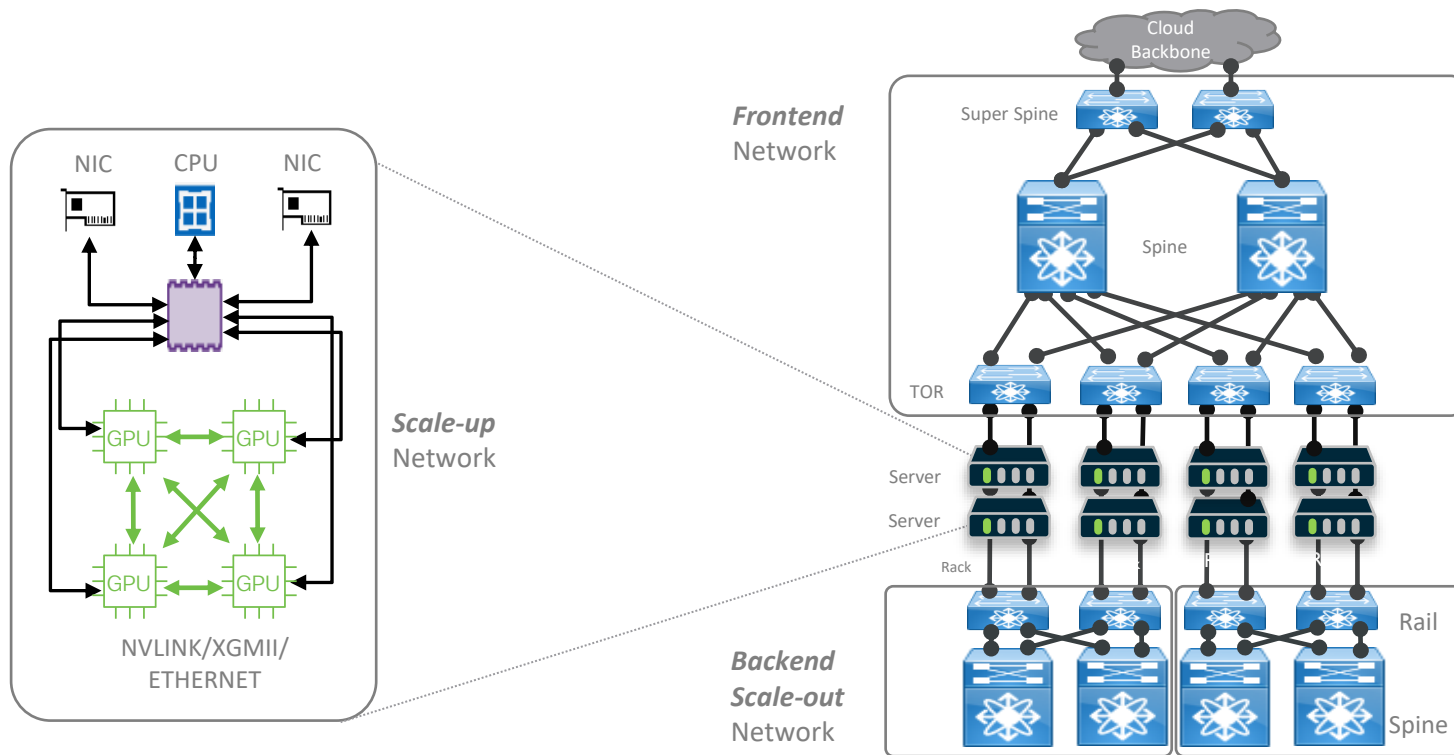
labeled
data
set



```
nn.Conv2d(W,H,Kw,Kh):  
for (int i = 0; i < outH; i++) {  
  for (int j = 0; j < outW; j++) {  
    int sum = 0;  
    for (int m = 0; m < Kh; m++) {  
      for (int n = 0; n < Kw; n++) {  
        sum += input[i + m][j + n] * kernel[m][n];  
      }  
    }  
    output[i][j] = sum;  
  }  
}
```

per ChatGPT 4/9/24

AI – Network Type Fundamentals



Why Networking is Relevant to AI Deployments

LLMs are orders of magnitude more intensive than DLRM



Deep Learning Recommendation Models

Search, Feed ranking. Ads & content recommendation

Inference needs a few Gigaflops for 100ms TTFT

Narrower scope, domain specific

Training: ~100 Gigaflop/ sentence



Inference needs 10s of Petaflops for 1 sec TTFT

Generate intelligent, creative responses

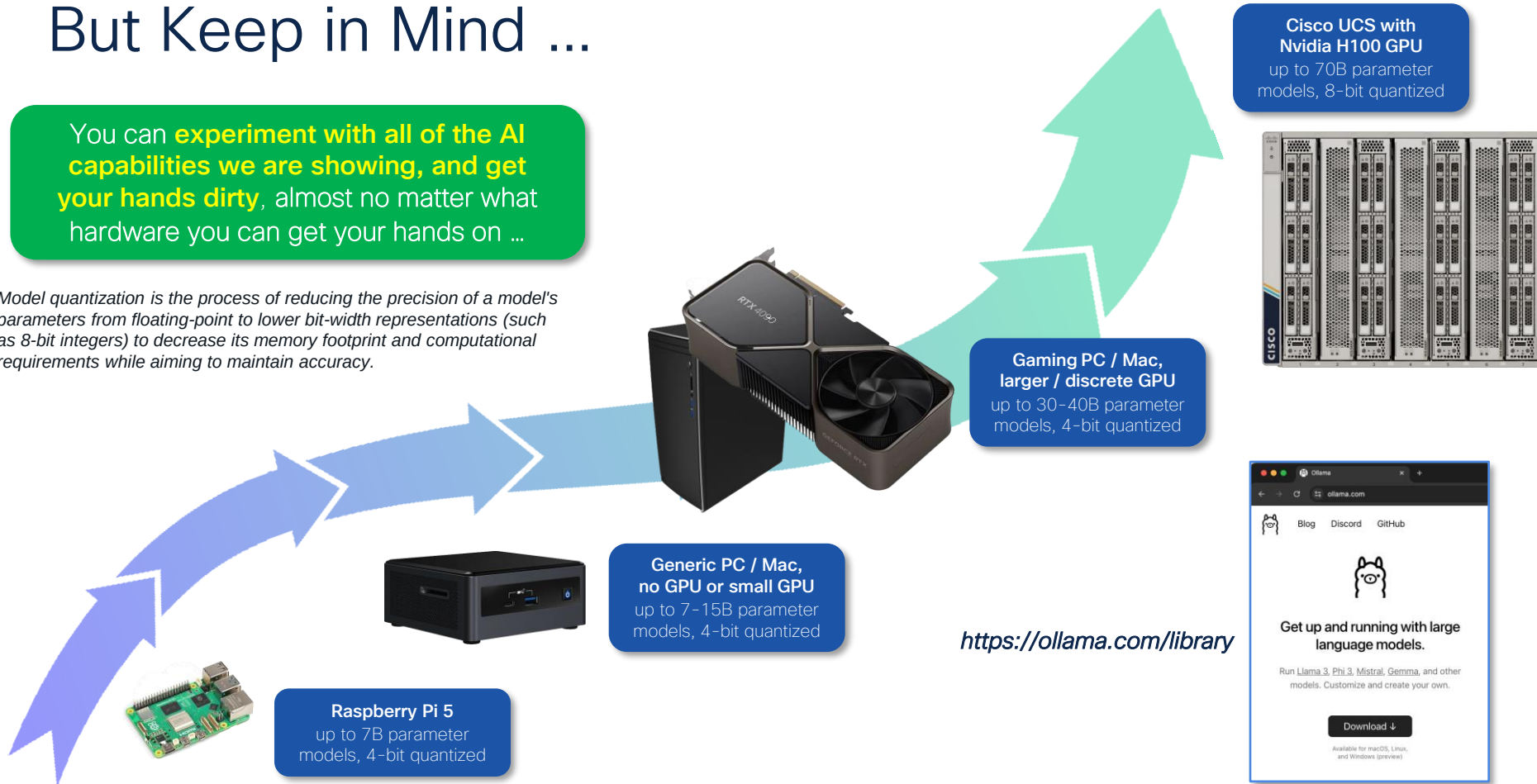
Training : ~1 Petaflop/ sentence

An Improved user experience means a *faster time to first token*, making *distributed inference an imperative*

But Keep in Mind ...

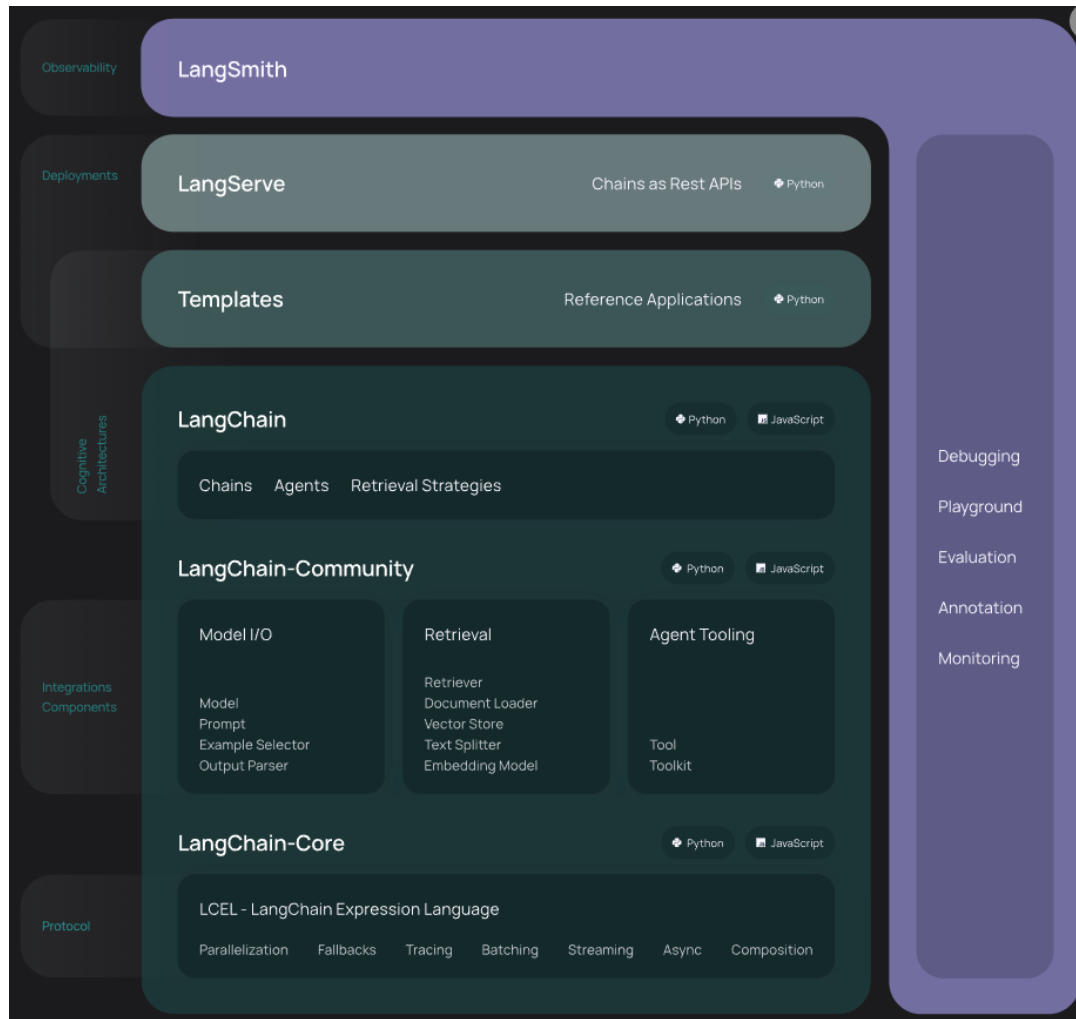
You can **experiment with all of the AI capabilities we are showing, and get your hands dirty**, almost no matter what hardware you can get your hands on ...

Model quantization is the process of reducing the precision of a model's parameters from floating-point to lower bit-width representations (such as 8-bit integers) to decrease its memory footprint and computational requirements while aiming to maintain accuracy.



CISCO Live!

LangChain



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)

RAG – Why and What?



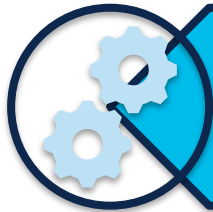
Limitations of Large Language Models



Knowledge limited to “cut-off date” of training data set



Closed-book examination – inability to access external sources of information when needed



Skills constrained by training parameters (weights, biases, training data set, etc.)

These limitations result in hallucinations

Incorrect answers

```
(base) praprama@PRAPRAMA-M-16WG ~ % ollama run phi:latest
>>> what is the world record for crossing the english channel entirely on foot?
  As of 2021, the official world record for crossing the English Channel on foot is held by Tom Hiddleston, who completed the journey in just 11 hours and 34 minutes. However, there have been many attempts to break this record over the years, with some individuals taking longer due to adverse weather conditions or unexpected obstacles.

>>>  Send a message (/? for help)
```

Out of date information



who is the current CEO of microsoft?



As of my last update in 2021, the CEO of Microsoft is Satya Nadella. He has been serving as the CEO since February 2014. However, since you asked for the "current" CEO and it's now 2024 according to your context, I would recommend checking the most recent and accurate information by visiting Microsoft's official website or consulting a reliable news source.

010110
110010
001011

Retrieval

Gather data from a data source (often a vector store) based on the user query



Augmented

Append the prompt with retrieved *context* and pass it on to the model

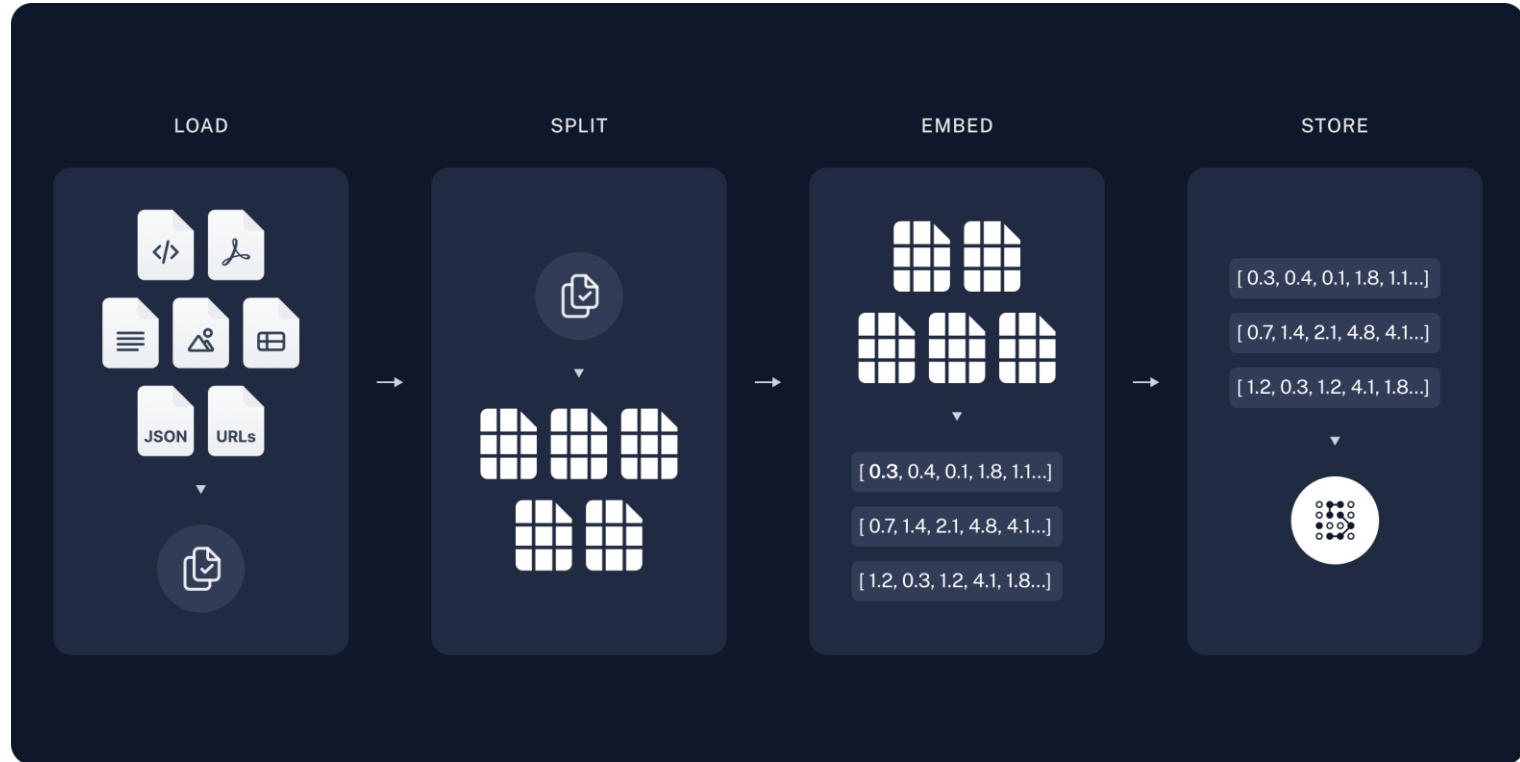


Generation

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

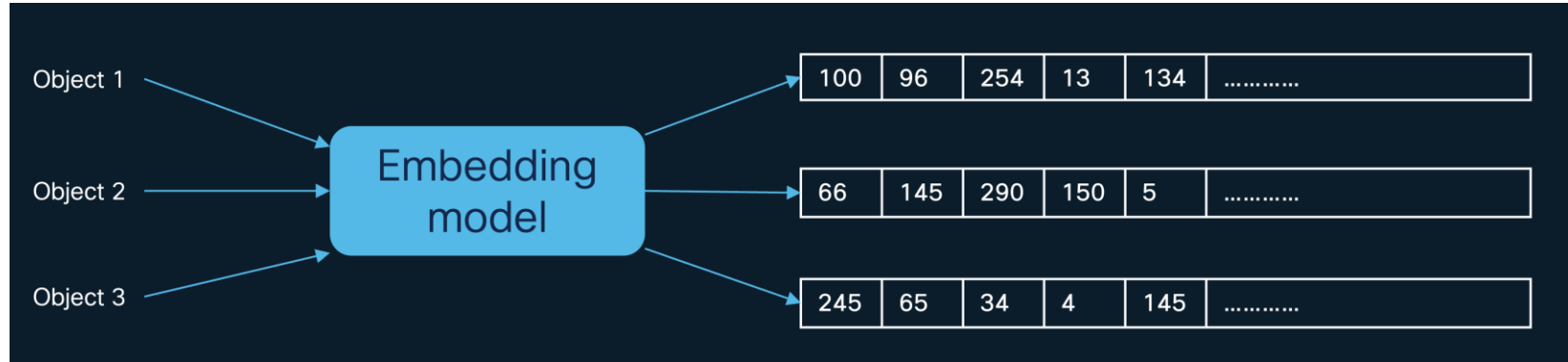
RAG Pipeline

Part 1 – Load and Store the Data



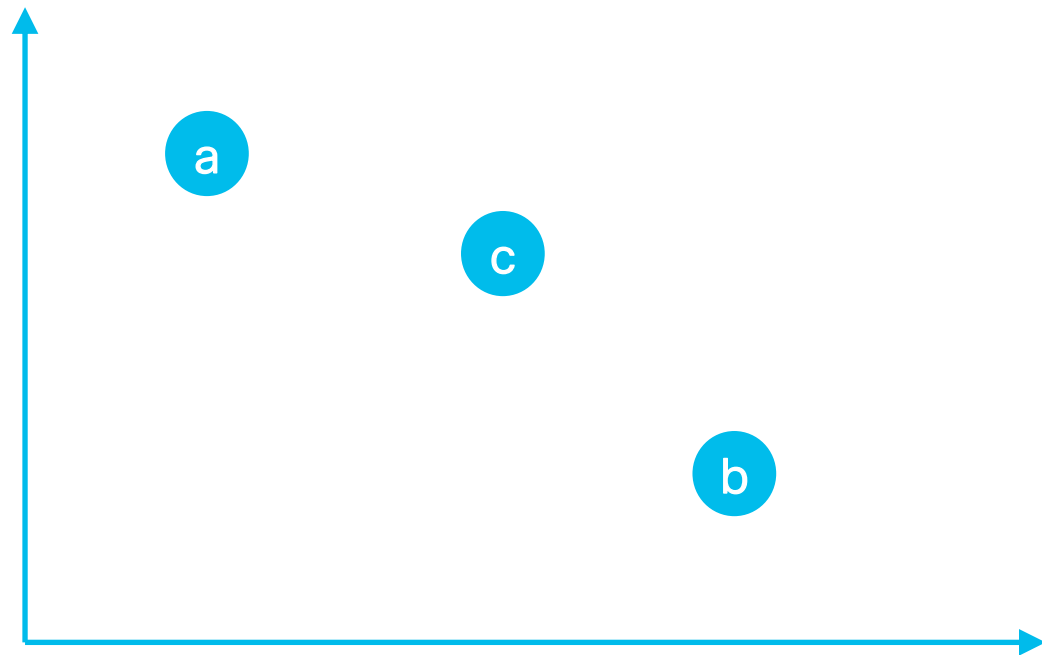
Embeddings and Semantic Search

- **Embeddings** – “*Vectors*” representing objects like words, images, audio, video, etc. in a machine consumable format
- **Embedding models** – Specially trained ML models to represent objects in a multi-dimensional vector space.
- **Vector stores** – A database specializing in handling vector data, both storing and retrieving them



Visualizing Embeddings and Semantic Search

Object ID	Sentence	Representation (2-D)
a	A bat is a flying mammal	13,145
b	The batsman's bat broke in 2	256,75
c	Bat is a noun	134,100



Embedding Models and Dimensions

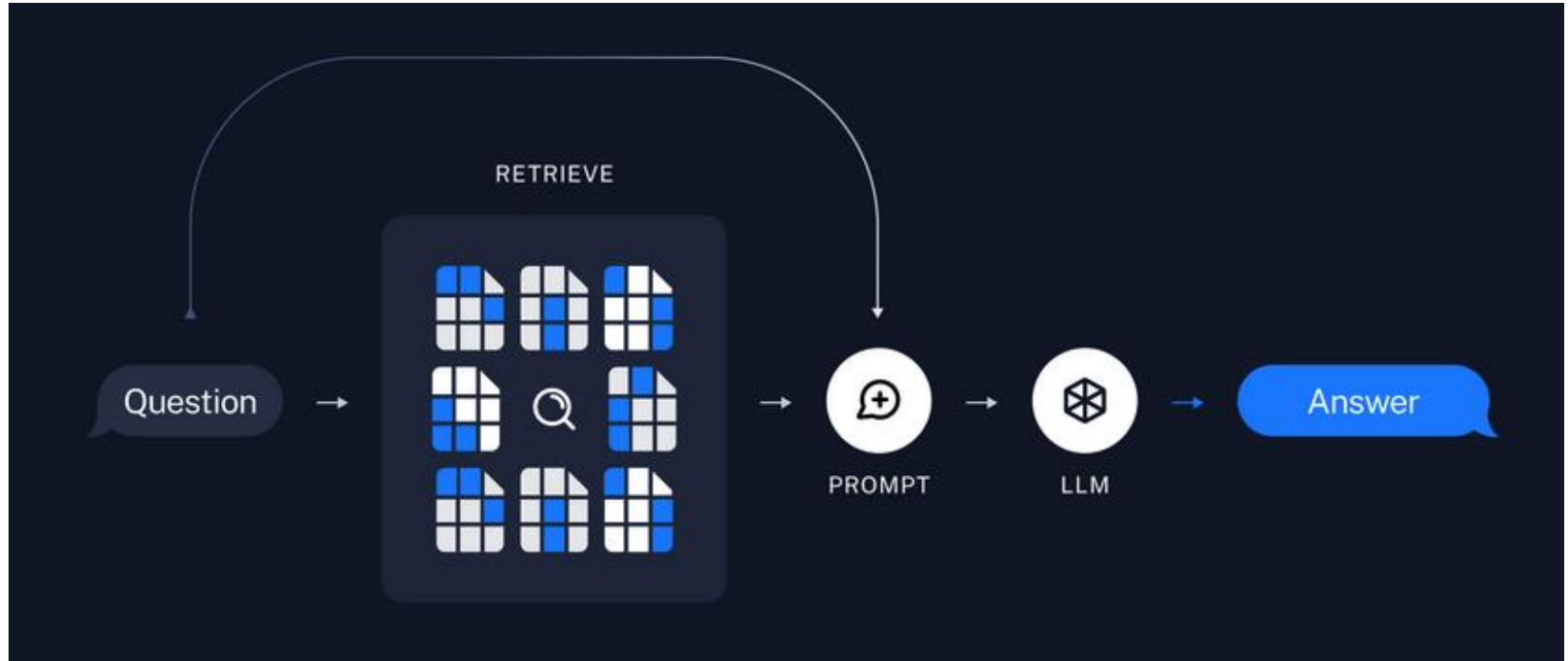
Model	Embedding dimensions
bge-base-en-v1.5	768
LLM2Vec-Mistral-supervised	4096
bge-large-en-v1.5	1024
instructor-xl	768
nomic-embed-text-v1.5-512	512
text-embedding-ada-002	1536
jina-embeddings-v2-base-en	768

Source and Massive Text Embedding Benchmark data:

<https://huggingface.co/spaces/mteb/leaderboard>

RAG Pipeline

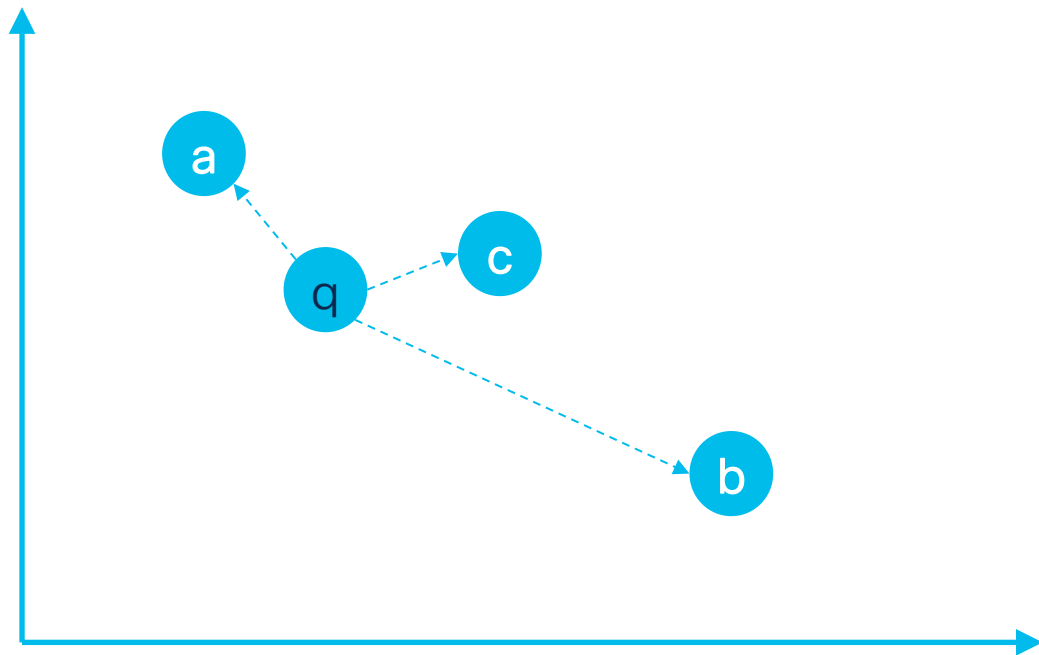
Part 2 – Answer Questions with Context



Visualizing Embeddings and Semantic Search

Object ID	Sentence	Representation (2-D)
a	A bat is a flying mammal	13,145
b	The batsman's bat broke in 2	256,75
c	Bat is a noun	134,100

q: What is a bat?



Basic RAG DIY

```
#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,
)
```

```
# 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()

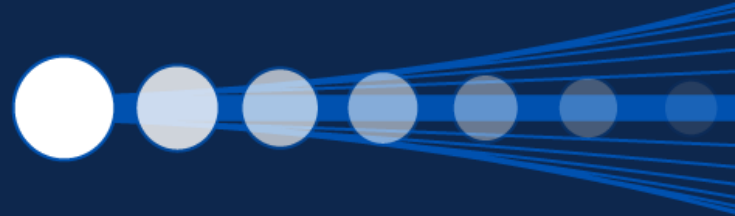
# setup memory so it remembers previous questions and answers
memory = ConversationBufferMemory(memory_key="chat_history", return_messages=True)
```

```
# perform the Conversational Retrieval Chain
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!



Simple RAG

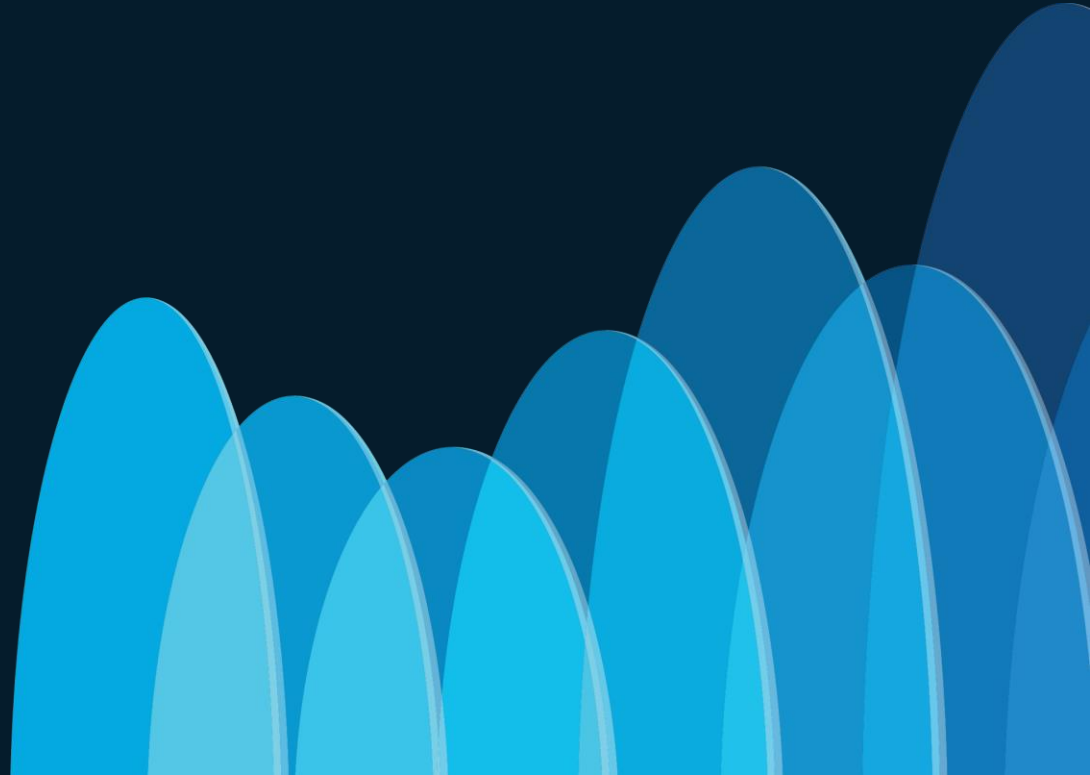
Benefits

- Ability to plug holes in a model's knowledge
- Simple implementation – easy to replace loaders, splitters, models
- Depending on use case, can run with very small local models

Drawbacks

- Necessary to tweak parameters (chunking, retrieving, prompting) for maximum results
- Does not suit itself well for dynamic data – need to keep database updated

Advanced RAG



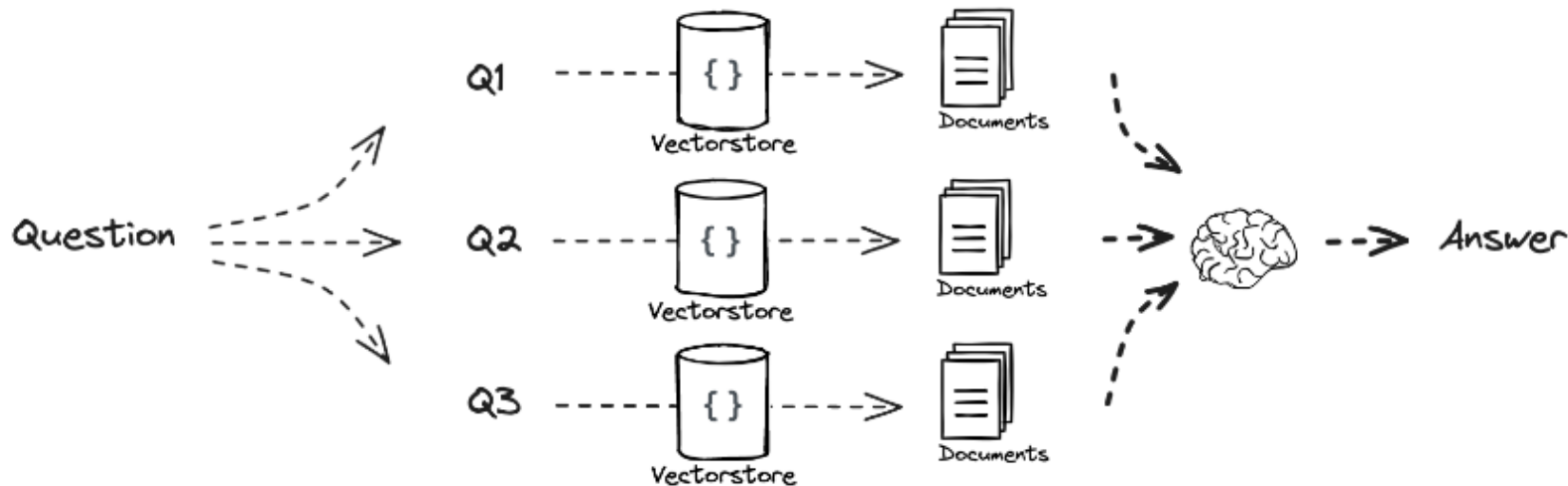
Techniques to Optimize Basic RAG

- How can we optimize the user's query?
- How can we optimize how data is embedded?
- How can we optimize the responses generated by the models?

Advanced RAG – Multi-Query

Part 5: Multi Query

Flow:



[rag-from-scratch/rag_from_scratch_5_to_9.ipynb](#) at main · langchain-ai/rag-from-scratch · GitHub

Advanced RAG – Multi-Query

Ask a question about the document data:

Please provide a SWOT analysis of this Cisco Annual report with highlight and insights

Send



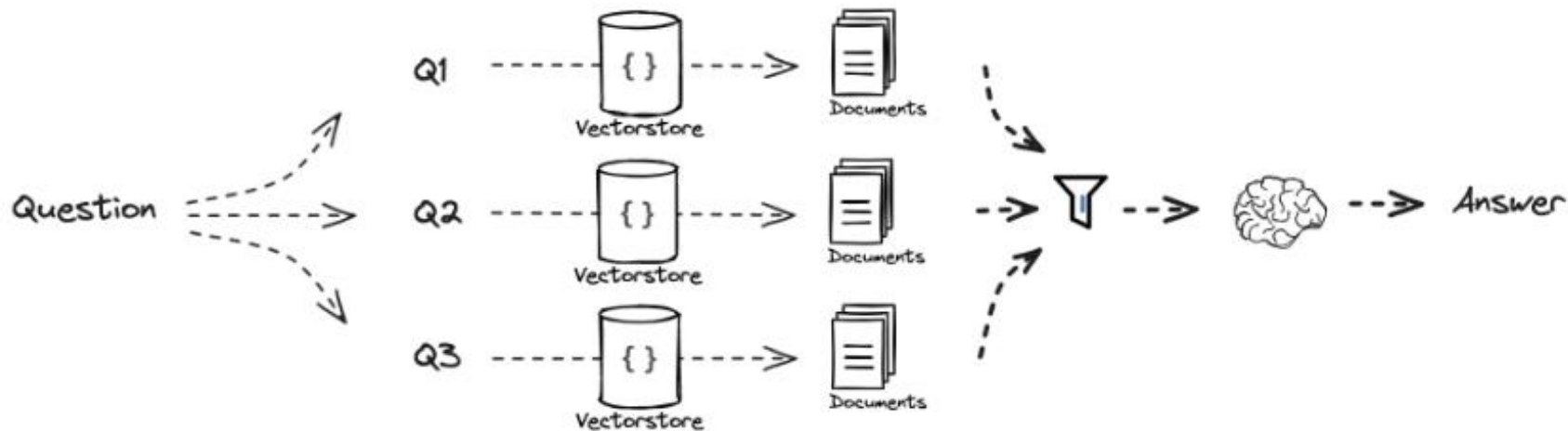
Related Queries:

```
[
  {
    "query": "Conduct an in-depth SWOT analysis of Cisco's latest annual report, focusing on their competitive positioning in the technology industry. Identify their key strengths, weaknesses, opportunities, and threats in terms of their market share, product offerings, and strategic partnerships."
  },
  {
    "query": "Explore the financial aspect of Cisco's annual report through a SWOT analysis lens. Evaluate their revenue streams, expenses, and profitability trends to uncover potential strengths, weaknesses, opportunities, and threats that could impact their financial performance."
  },
  {
    "query": "Analyze the societal and environmental implications of Cisco's operations and initiatives as presented in their annual report using a SWOT framework. Determine how their strategies address stakeholder concerns related to sustainability, ethics, and social responsibility."
  },
  {
    "query": "Employ a SWOT analysis to assess the technological innovations and trends featured in Cisco's annual report. Identify how these advancements contribute to their strengths, weaknesses, opportunities, and threats, shaping their future growth prospects in the rapidly evolving technology landscape."
  }
]
```

Advanced RAG – RAG-Fusion

Part 6: RAG-Fusion

Flow:



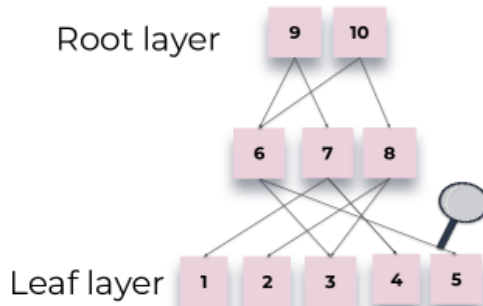
[rag-from-scratch/rag_from_scratch_5_to_9.ipynb](#) at main · langchain-ai/rag-from-scratch · GitHub

RAPTOR

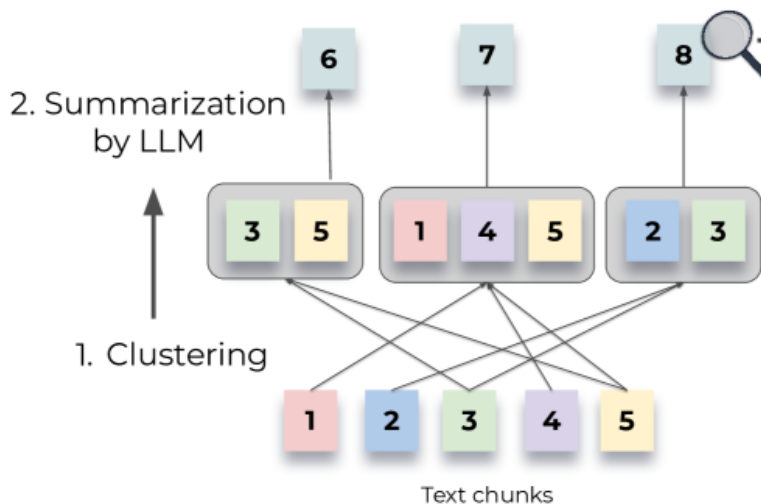


Look familiar? Like a Well Designed Data Center!

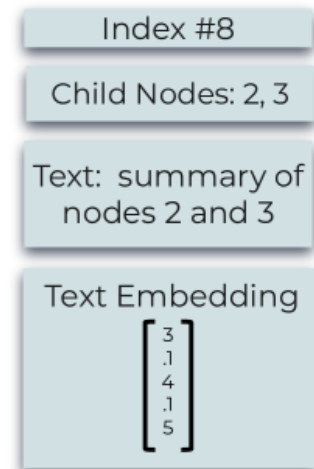
RAPTOR Tree



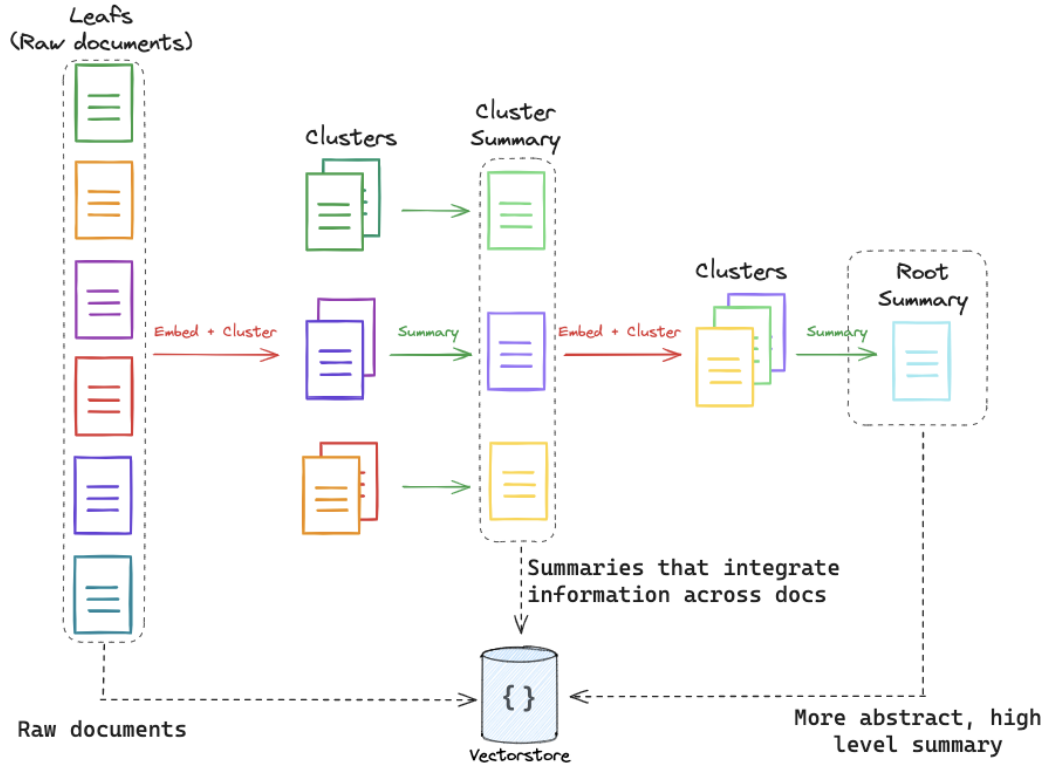
Formation of one tree layer



Contents of a node



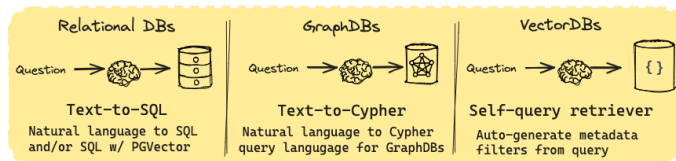
RAPTOR (Document tree at varying levels of abstraction)



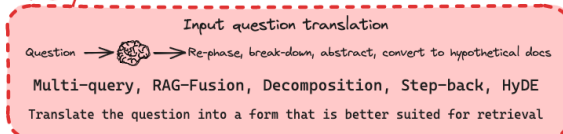
[langchain/cookbook/RAPTOR.ipynb at master · langchain-ai/langchain · GitHub](https://github.com/langchain-ai/langchain/blob/master/cookbook/RAPTOR.ipynb)

One picture to RAG them all!

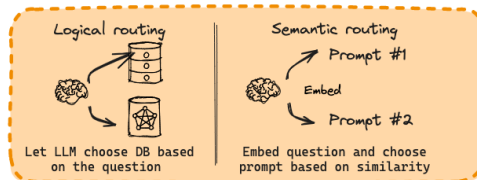
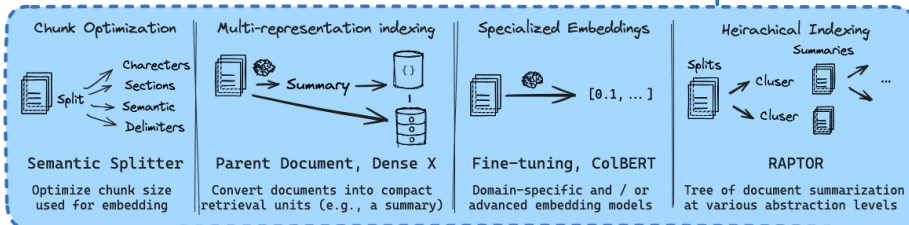
Query Construction



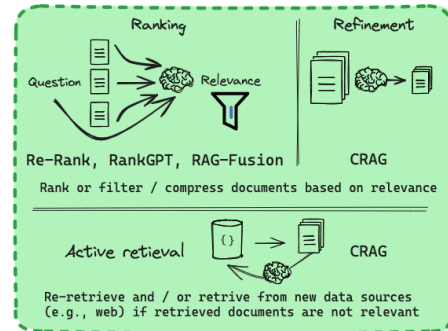
Query Translation



Routing

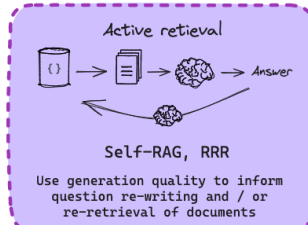


Retrieval



Indexing

Generation



On-Prem AI Solutions

Open Source LLMs with Cisco
UCS and Nexus

cisco *Live!*



Patrick

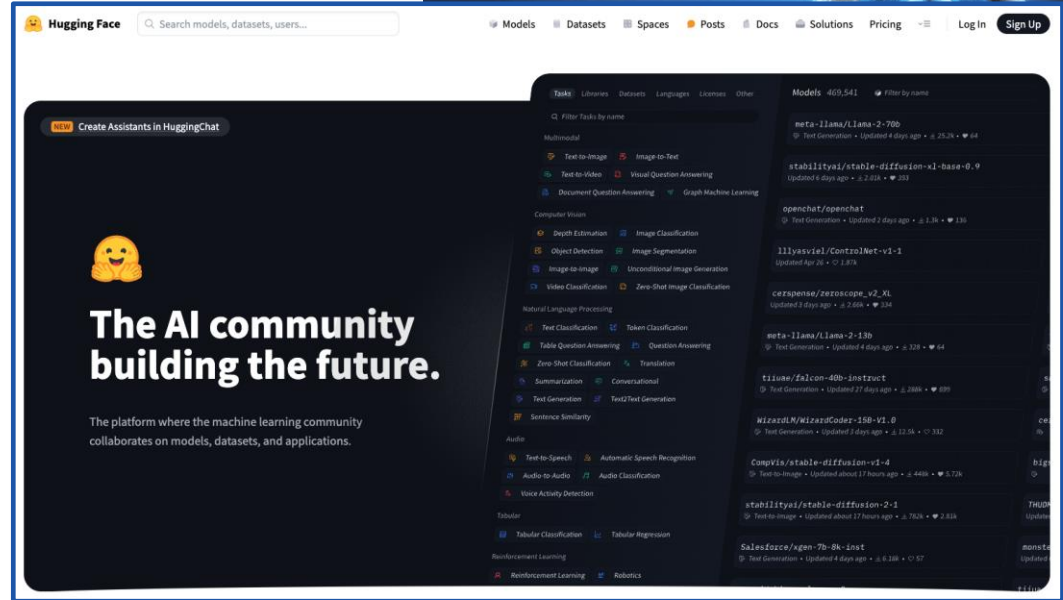
On-Prem AI Solutions:

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
- Mar 3, 2023 (a week after it's release) LLaMA's model was leaked to the open source community
- Unlike other LLMs, Meta LLaMA is now available to the Open Source community



Open Source AI and Responsible AI Use



Cisco believes **data privacy** is important



Cisco has a **Responsible AI (RAI) framework** and **controls** around transparency, fairness, accountability, privacy, security and reliability



With the emergence of Generative AI, we are **evolving our RAI framework and controls** to surface new variants of issues, such as (but not limited to):

- **False content**
- **Hallucinations**
- **Bias**
- **Unanticipated output**
- **Harmful output**
- **Deep fakes**

Please keep in mind that **open source AI tools and capabilities** may or may not have been checked against issues such as these (or others).

As such, **please use any such tools at your own risk, and in accordance with any applicable AI policies** from your organizations.

Differing
Open Source
licenses and
use

CAUTION

“Working closely with Cisco, we’re making
it easier than ever for enterprises to
obtain the infrastructure they need to
benefit from AI.”

Jensen Huang
Founder and CEO, NVIDIA

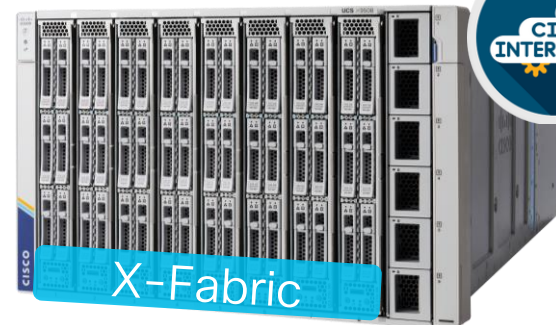


Nexus AI Networking

cisco *Live!*



UCS C885A – 8 x GPU Server

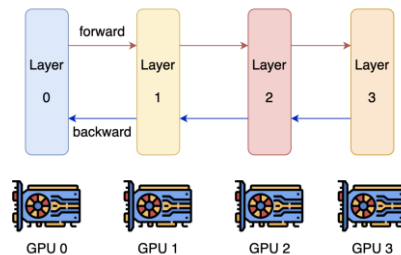


UCS-X Series – Modular Chassis

Splitting AI Model Across GPUs

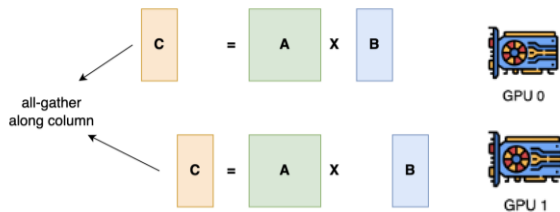
Pipeline Parallelism

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!



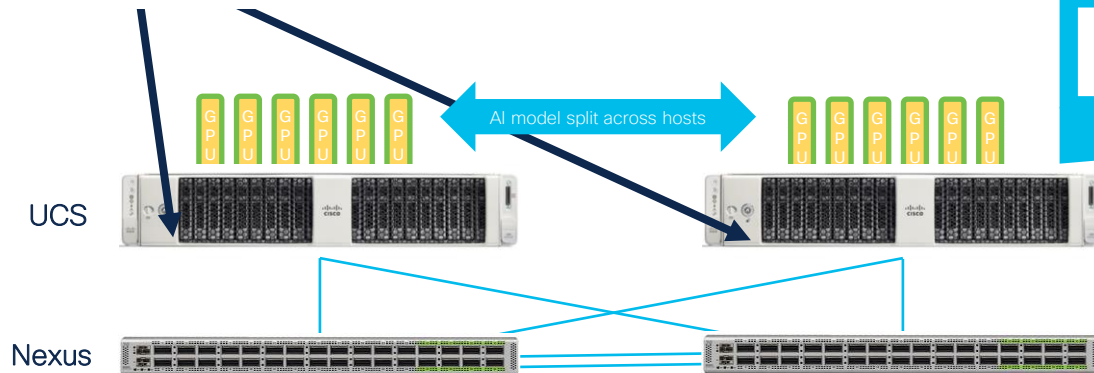
www.github.com/pl247/ai-toolkit

Demo Multi-GPU Multi-Host



Intersight

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



cisco *Live!*

OI

Open WebUI

LLM

LLM Server



Hugging Face

Models



Meta



RAY

Clustering Framework



AI
Enterprise

GPU Driver & CUDA



Ubuntu

Operating System

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