ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

BK TP.HCM

AI, GenAI and AI Agents: A case-study at HCMUT

Quan Thanh Tho

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

About speaker

Assoc. Prof. Dr. Quan Thanh Tho

Dean, Faculty of Computer Science and Engineering

University of Technology – Vietnam National University,

Ho Chi Minh City.

NỘI DUNG

01 Artificial Intelligence (AI): Introduction

02 AI Agents

03 Agents, LLM and RAG

04 A Case-study at HCMUT

Introduction to Artificial Intelligence

- What is Artificial Intelligence?
- Artificial Intelligence: How does it work?
- Artificial Intelligence: Key Milestones
- Artificial Intelligence: Current Achievements
- Generative AI

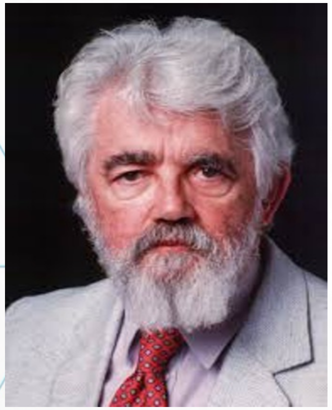
What is Artificial Intelligence?

Artificial Intelligence (**A**rtificial **I**ntelligence, **AI**)

A field of research that enables machines to possess human-like abilities.

Proposed at: Dartmouth Conference (1956)

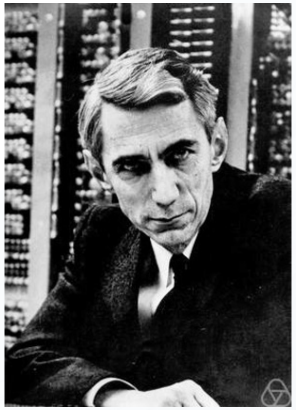
Summer Research Project on Artificial Intelligence



John McCarthy

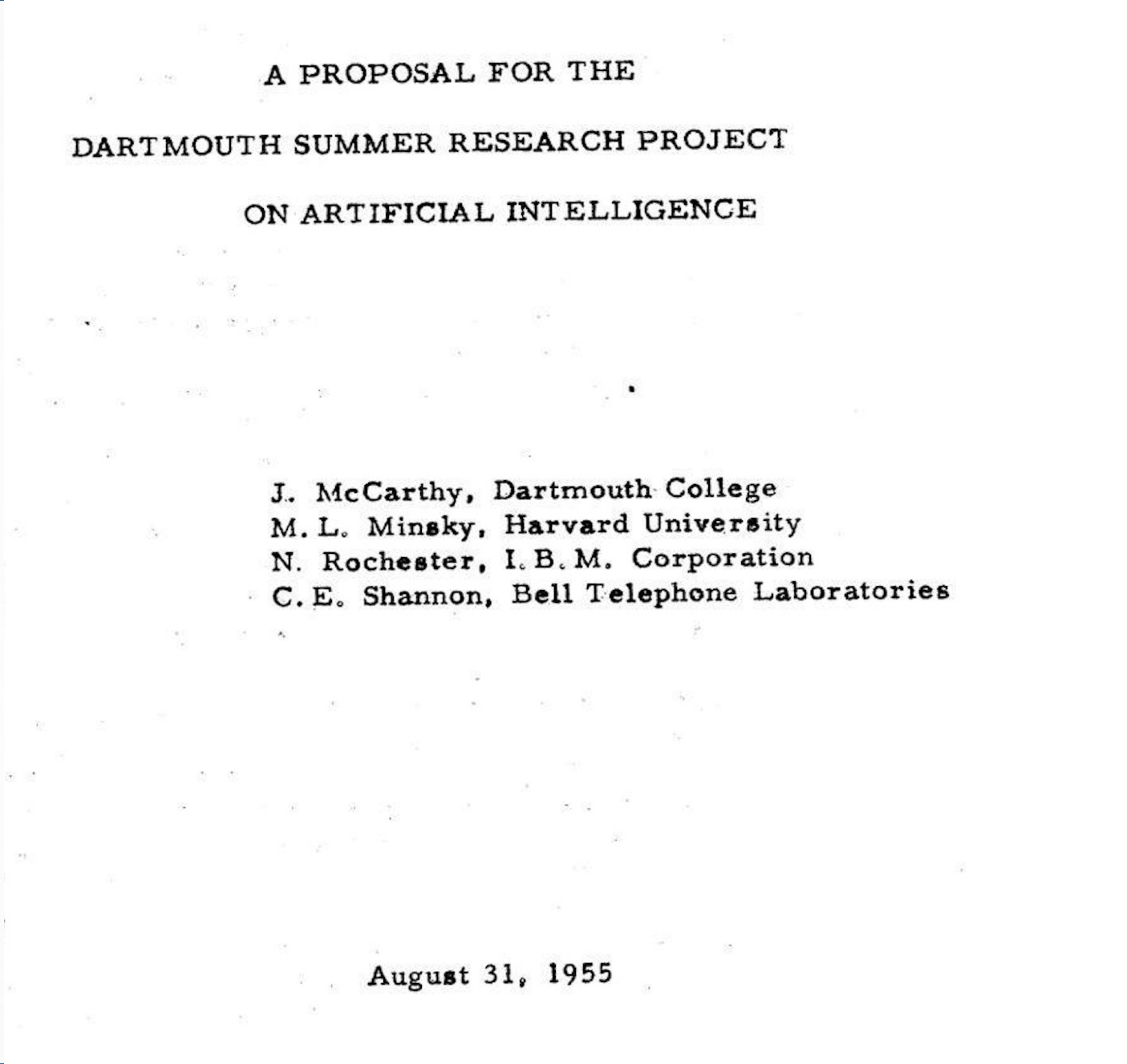
Marvin Minsky

…



Claude Shannon

Allen Newell



A PROPOSAL FOR THE

DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

August 31, 1955

WHAT SPECIAL ABILITIES DO HUMANS HAVE?

What is Artificial Intelligence?

Vision from around ~1950

01 **Language ability**

Understanding text, summarizing documents, writing

02 **Vision and perception**

Understanding images and videos

03 **Communication through speech**

Speech recognition and generation

What is Artificial Intelligence?

Vision from around ~1950

04 **Knowledge Representation and Reasoning**

Performing a transformation chain:

Data → Information → Knowledge → Intelligence

Reasoning based on knowledge

Reasoning is the weakest capability and is currently receiving significant attention from the research community.

What is Artificial Intelligence?

Vision from around ~1950

05 **Language ability**

Carrying out intelligent actions:
expressing emotions, walking, standing, running, jumping, coordination, controlling vehicles and flying devices, etc.

06 **Learning Capability**

Learning from data

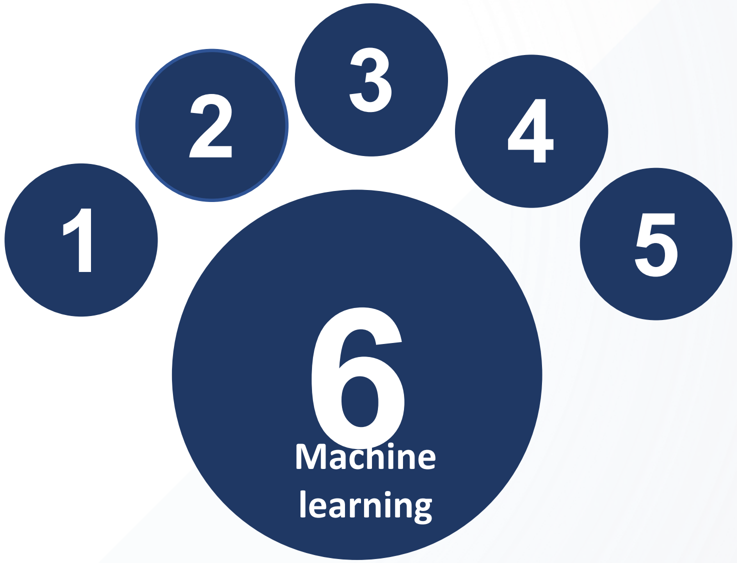
What is Artificial Intelligence?

Artificial Intelligence vs Human Intelligence

Human	Artificial Intelligence

Emphasis on learning

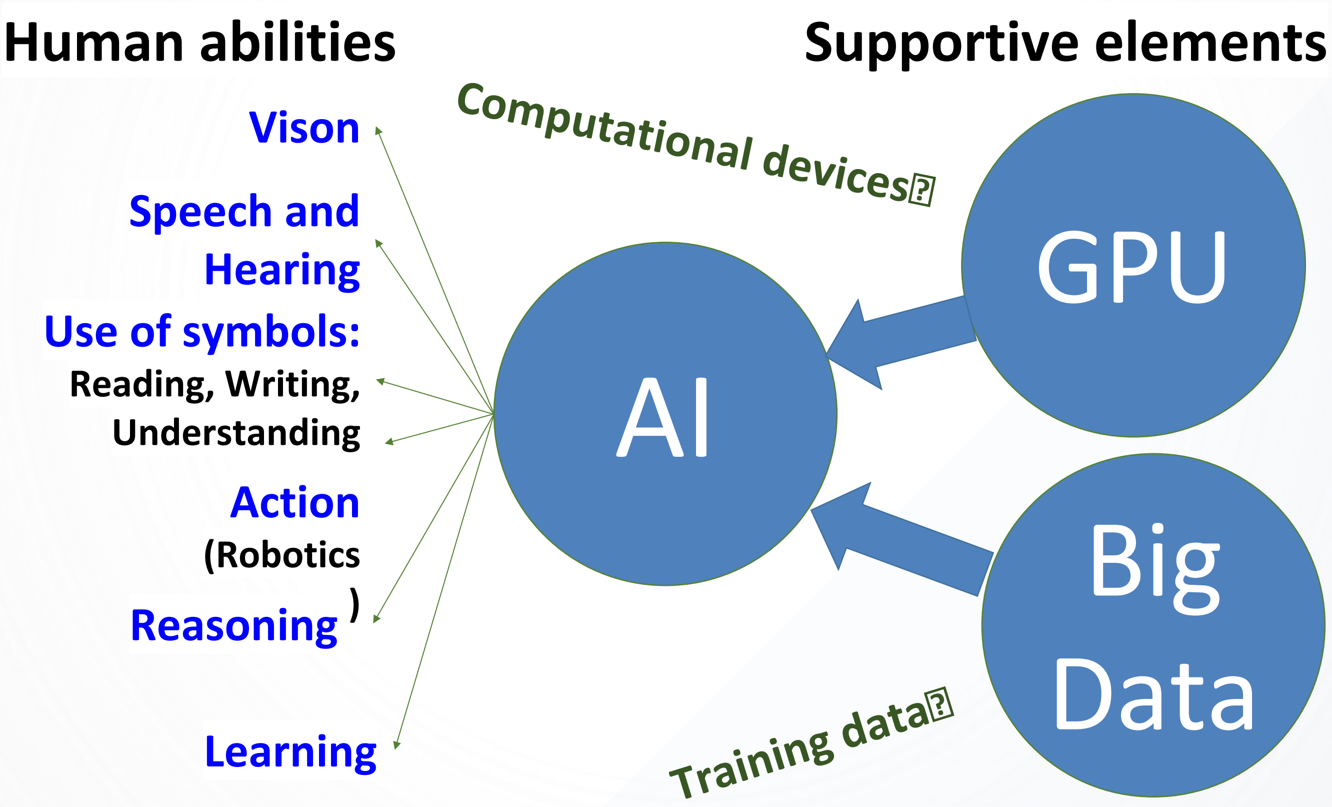
- **Vietnam:** *Learn, Learn more, Keep learning*
- **USA (ABET Standard):** *Lifelong learning*



Learning (item 6) is a foundational skill that enables the development of other skills (items 1–5)

Artificial Intelligence?

How? (supporting elements are needed)



Human abilities

Supportive elements

Vison

Speech and Hearing

Use of symbols: Reading, Writing, Understanding

Action (Robotics)

Reasoning

Learning

Computational devices

Training data

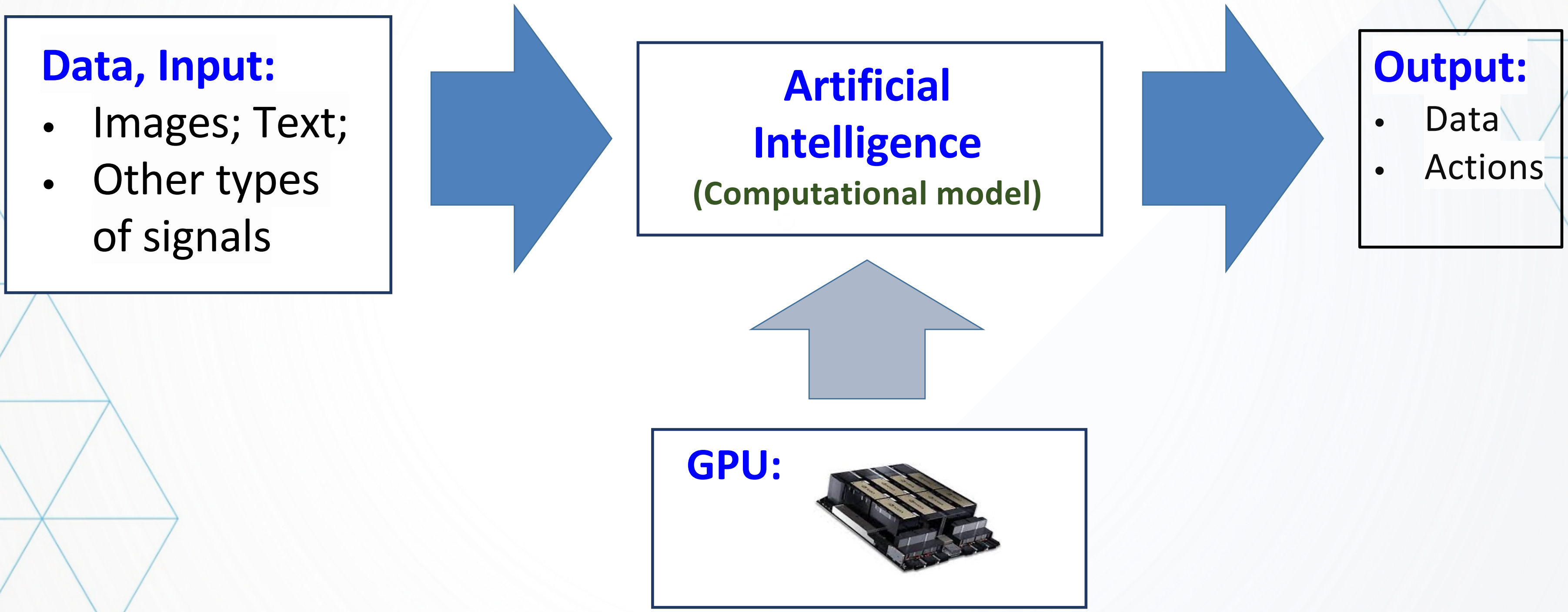
AI

GPU

Big Data

Artificial Intelligence?

How? (computational model)



Data, Input:

- Images; Text;
- Other types of signals

Artificial Intelligence

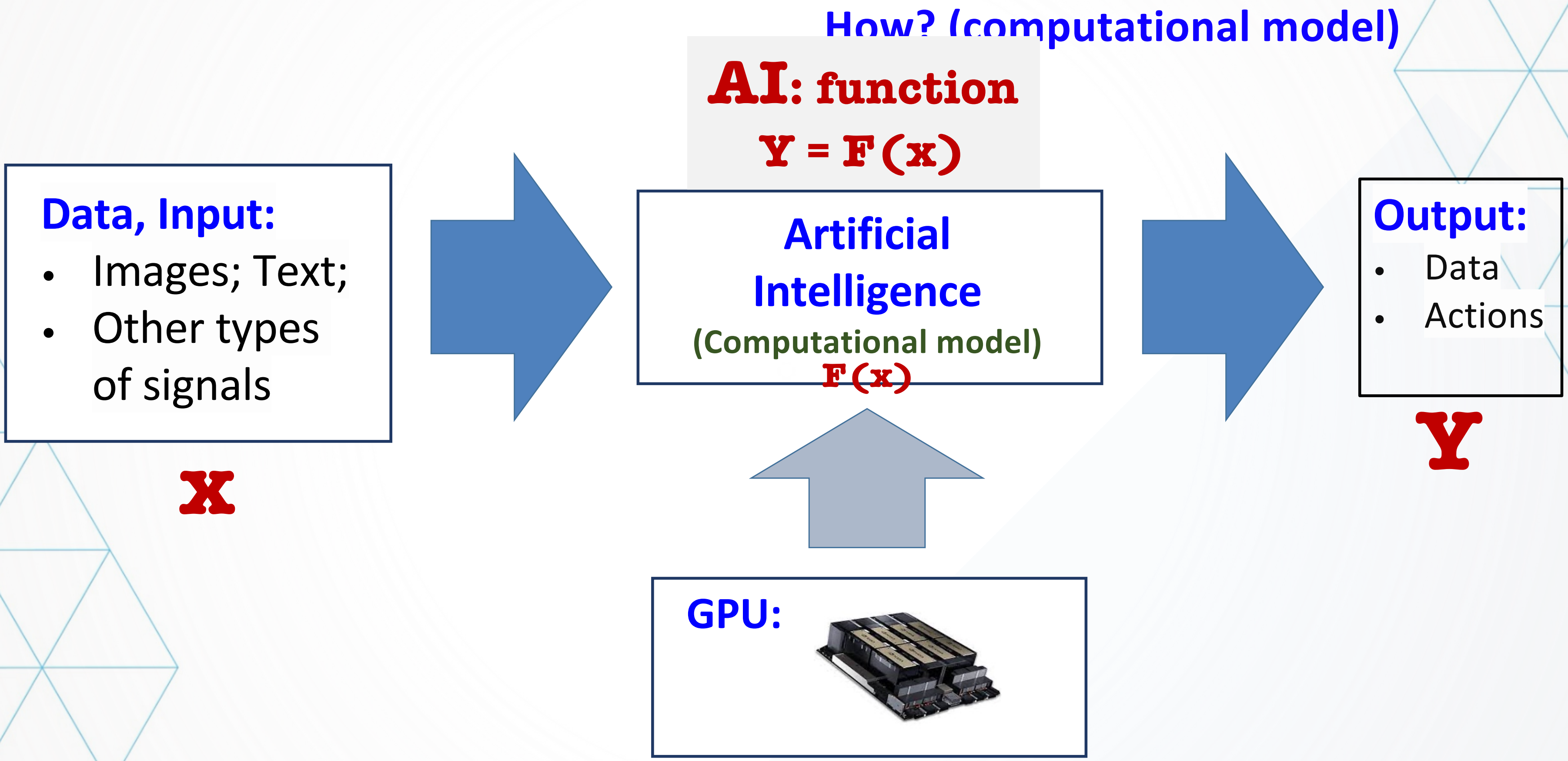
(Computational model)

Output:

- Data
- Actions

GPU:

Artificial Intelligence?



How? (computational model)

AI: function

Y = F(x)

Data, Input:

- Images; Text;
- Other types of signals

x

Artificial Intelligence

(Computational model)

F(x)

Output:

- Data
- Actions

Y

GPU:

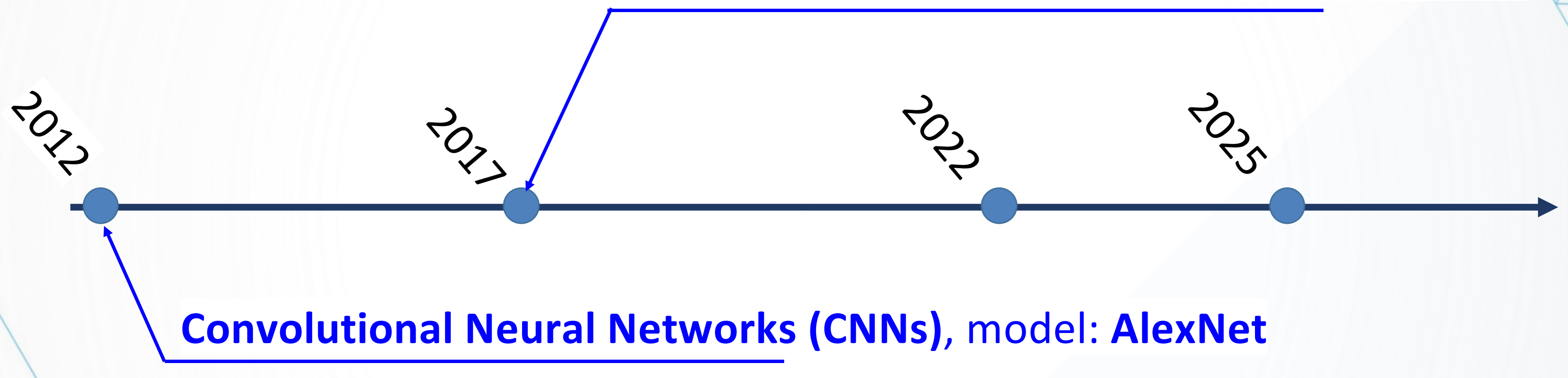
Milestones

2012
2017
2022
2025

Beginning of a Trend:

- Unification of data structure formats (text, images, time-series, etc.)
- Emphasis on self-supervised learning to address the challenges of manual labeling

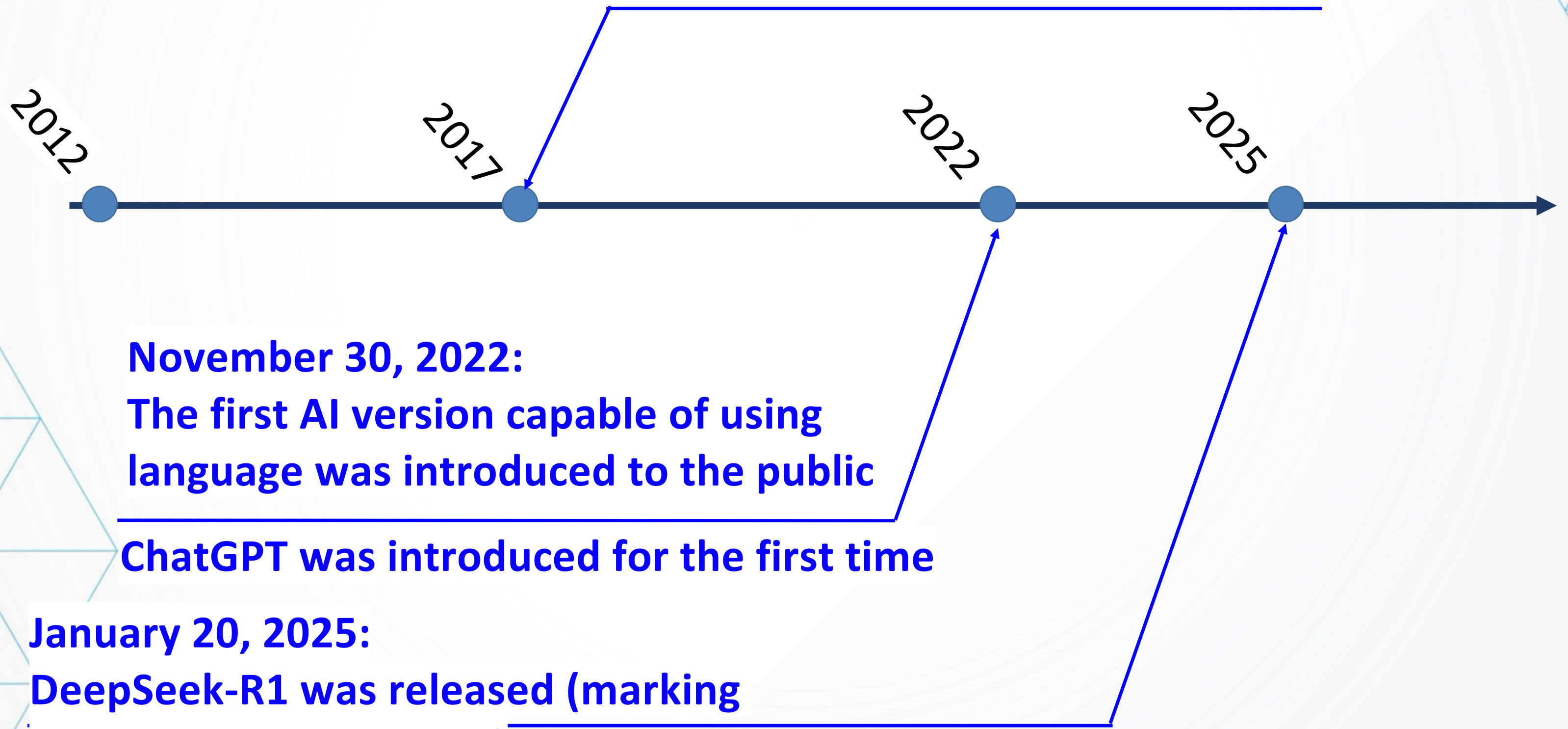
Transformer computation block, attention mechanism



Convolutional Neural Networks (CNNs), model: **AlexNet**

The Start of a Paradigm Shift:

- Unification of computational techniques: AI models represented as computational graphs (replacing many disparate traditional methods)
- Use of 1D, 2D, and 3D convolutions for processing various signal types



Beginning of a Trend:

- Unification of data structure formats (text, images, time-series, etc.)
- Emphasis on self-supervised learning to address the challenges of manual labeling

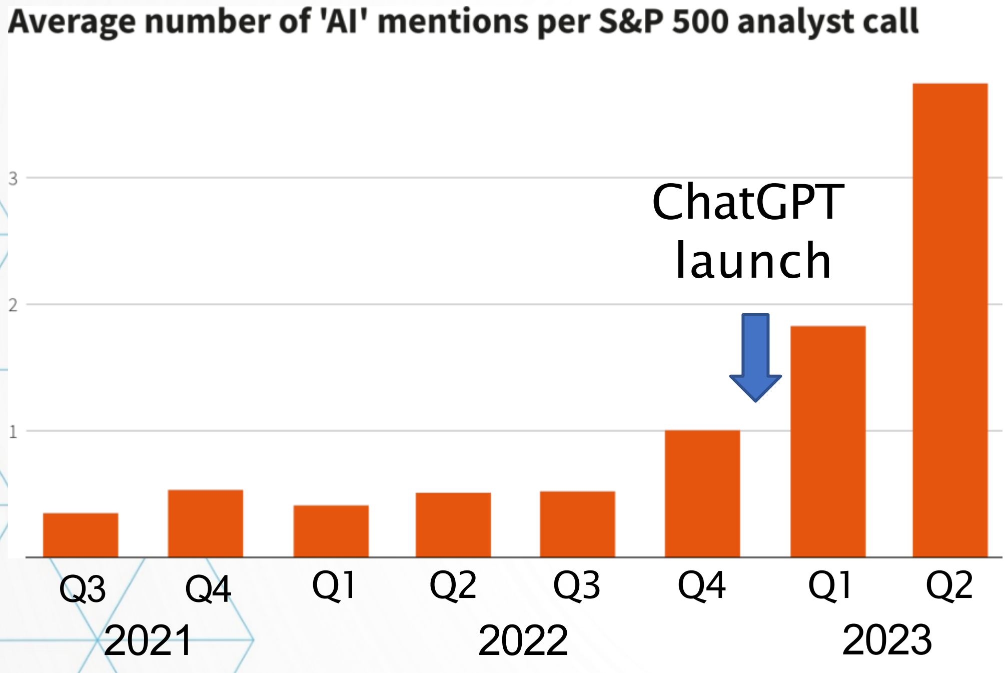
Transformer computation block, attention mechanism

November 30, 2022:
The first AI version capable of using language was introduced to the public

ChatGPT was introduced for the first time

January 20, 2025:
DeepSeek-R1 was released (marking the trend of low-cost AI)

The rise of generative AI



Source: Reuters

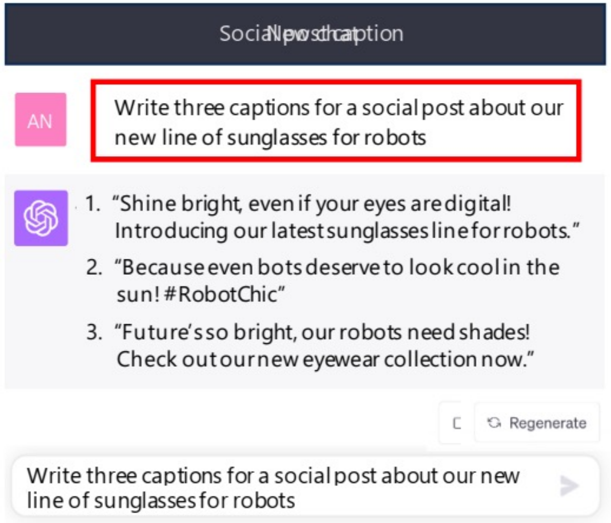
Generative AI could

- Add \$2.6-\$4.4 trillion annually to the economy[1]
- Raise global GDP by 7% in the next 10 years[2]
- Impact 10% of the tasks carried out daily by 80% of workers[3]

What is generative AI

Artificial intelligence systems that can produce high quality content, specifically text, images, and audio.

Prompts

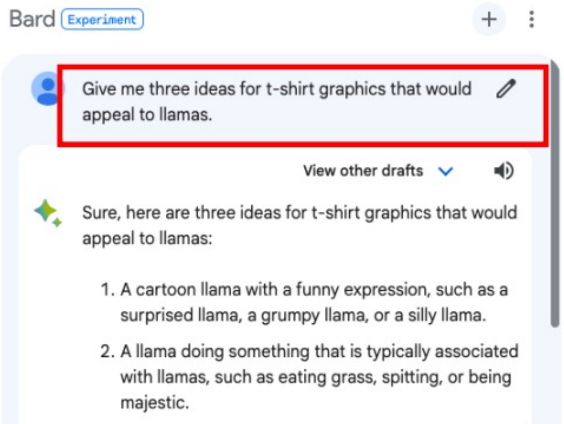


Write three captions for a social post about our new line of sunglasses for robots

1. "Shine bright, even if your eyes are digital! Introducing our latest sunglasses line for robots."
2. "Because even bots deserve to look cool in the sun! #RobotChic"
3. "Future's so bright, our robots need shades! Check out our new eyewear collection now."

Write three captions for a social post about our new line of sunglasses for robots

ChatGPT/OpenAI

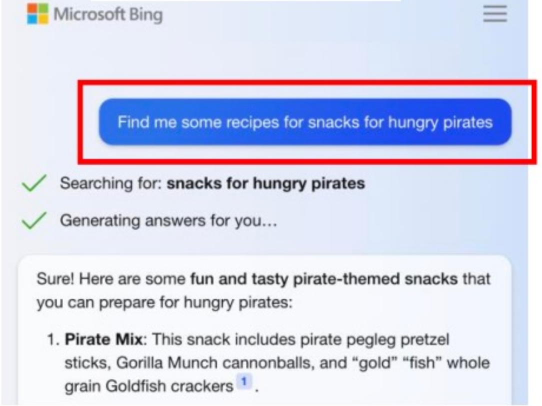


Give me three ideas for t-shirt graphics that would appeal to llamas.

Sure, here are three ideas for t-shirt graphics that would appeal to llamas:

1. A cartoon llama with a funny expression, such as a surprised llama, a grumpy llama, or a silly llama.
2. A llama doing something that is typically associated with llamas, such as eating grass, spitting, or being majestic.

Bard/Google



Find me some recipes for snacks for hungry pirates

✓ Searching for: **snacks for hungry pirates**

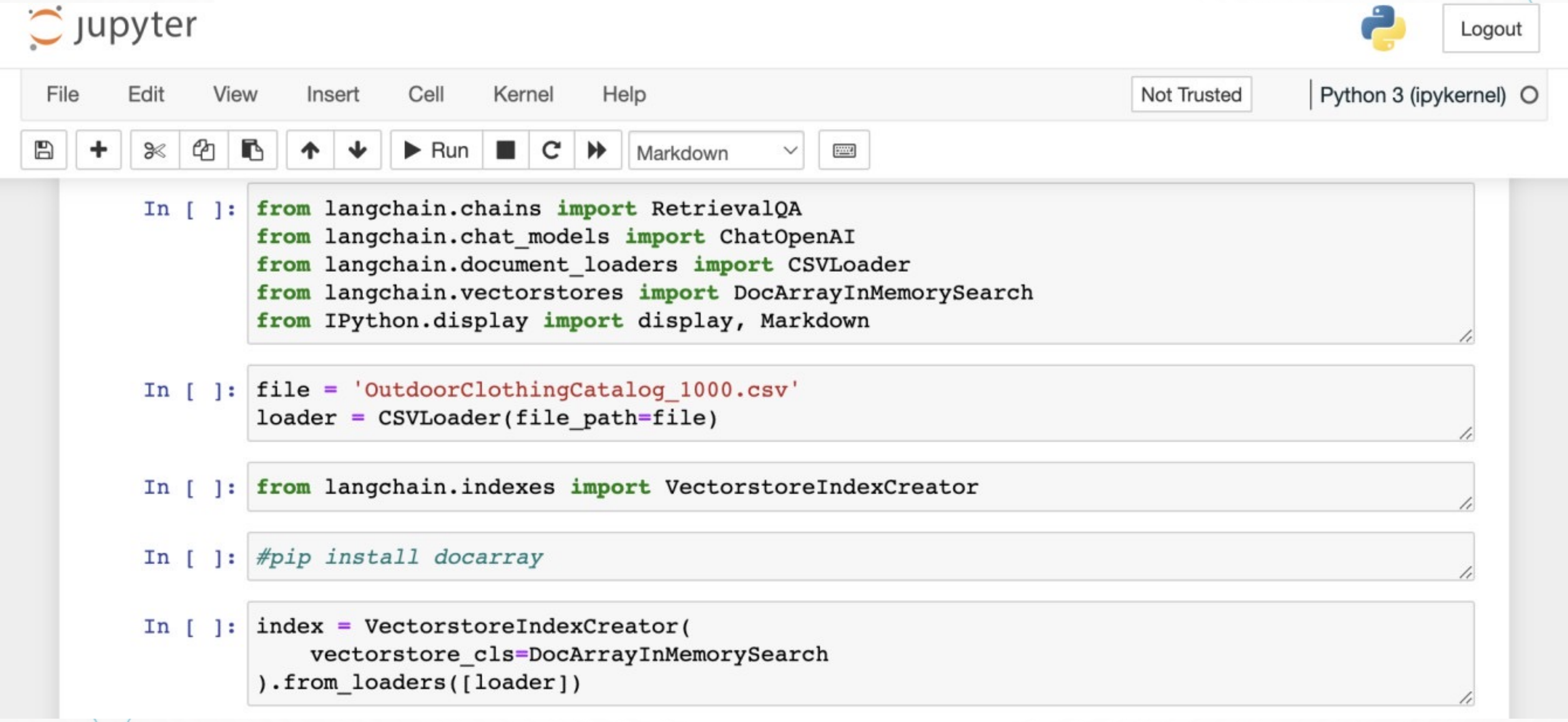
✓ Generating answers for you…

Sure! Here are some **fun and tasty pirate-themed snacks** that you can prepare for hungry pirates:

1. **Pirate Mix**: This snack includes pirate pegleg pretzel sticks, Gorilla Munch cannonballs, and "gold" "fish" whole grain Goldfish crackers [1].

Bing Chat/Microsoft

Generative AI can also be a tool



```
In [ ]: from langchain.chains import RetrievalQA
        from langchain.chat_models import ChatOpenAI
        from langchain.document_loaders import CSVLoader
        from langchain.vectorstores import DocArrayInMemorySearch
        from IPython.display import display, Markdown

In [ ]: file = 'OutdoorClothingCatalog_1000.csv'
        loader = CSVLoader(file_path=file)

In [ ]: from langchain.indexes import VectorstoreIndexCreator

In [ ]: #pip install docarray

In [ ]: index = VectorstoreIndexCreator(
            vectorstore_cls=DocArrayInMemorySearch
        ).from_loaders([loader])
```

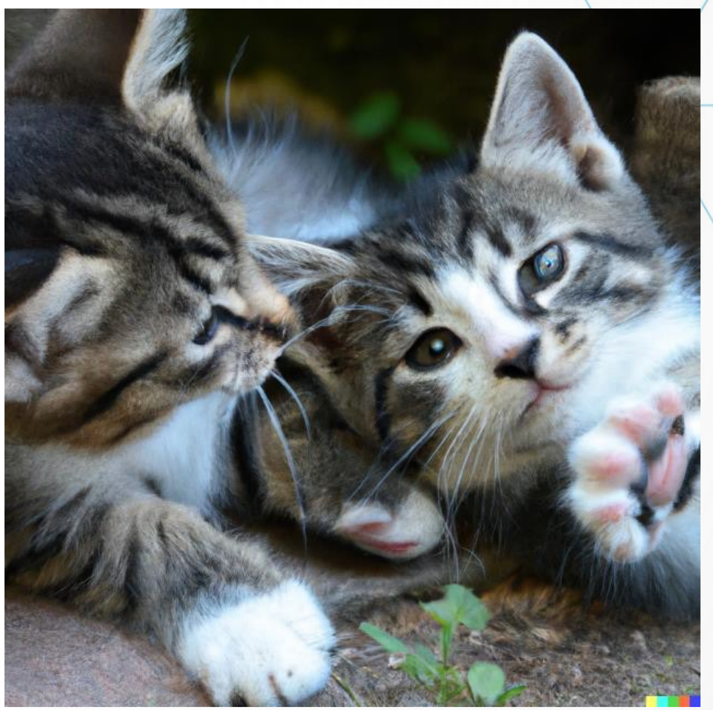
AI is every where

AI technology	Example
Web search s	Google, Bing
Fraud detection	Credit card payments
Recommender system	Amazon, Netflix

Generative Image and Video

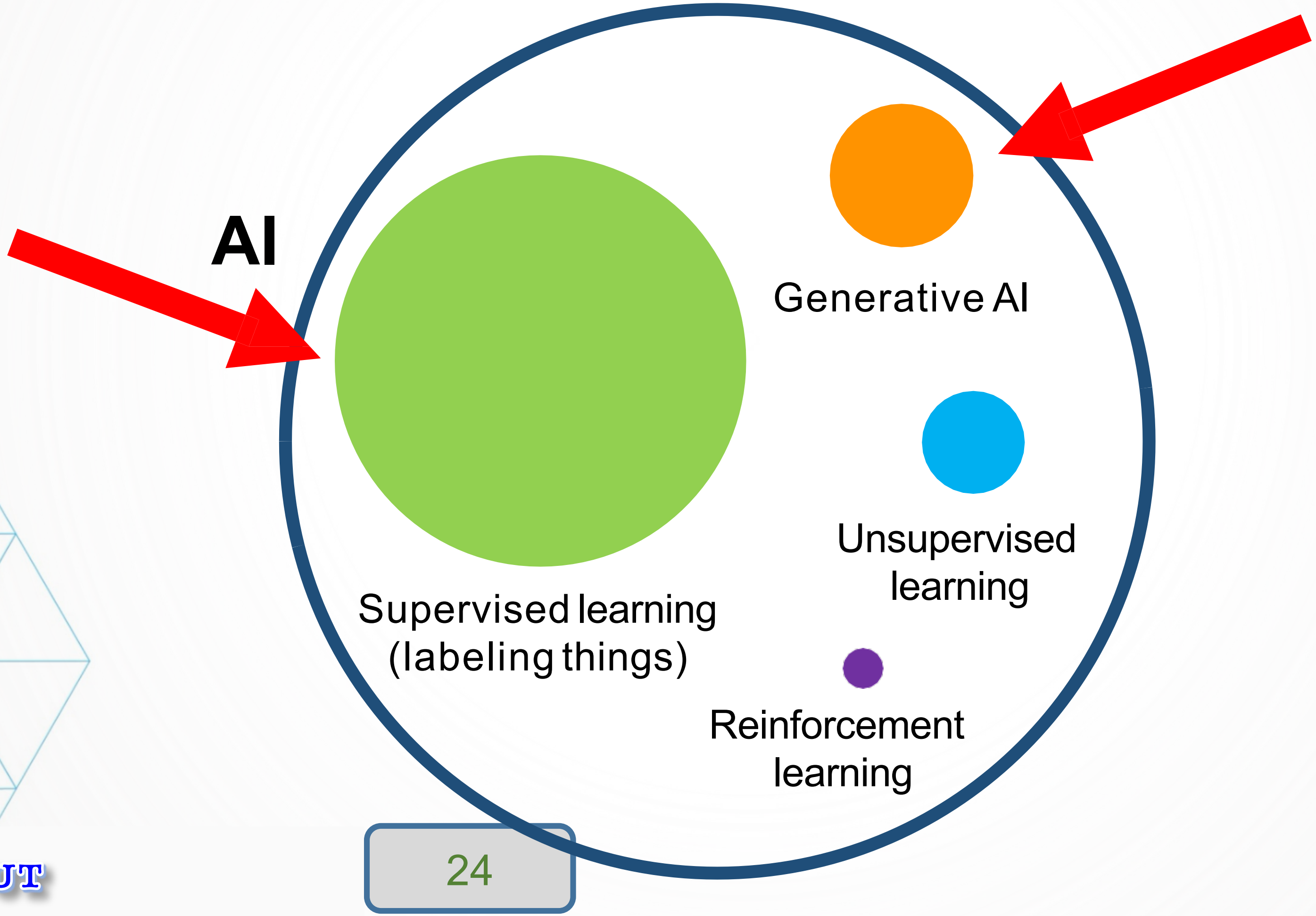
A beautiful, pastoral mountain scene.
E) Landscape painting style (Midjourney)

Two cute kittens playing (DALL-



HOW GEN AI WORKS

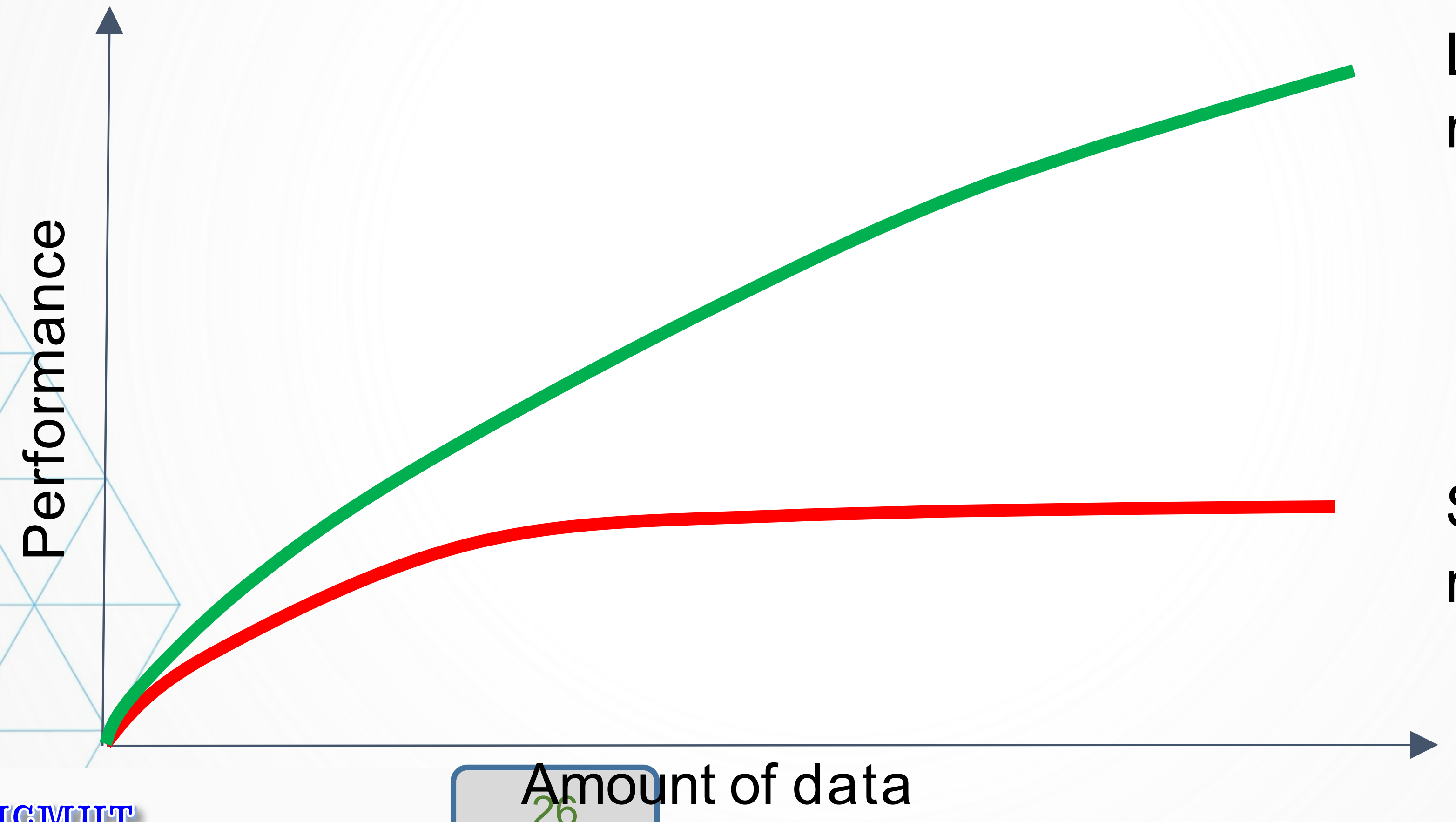
AI Techniques



Supervised learning

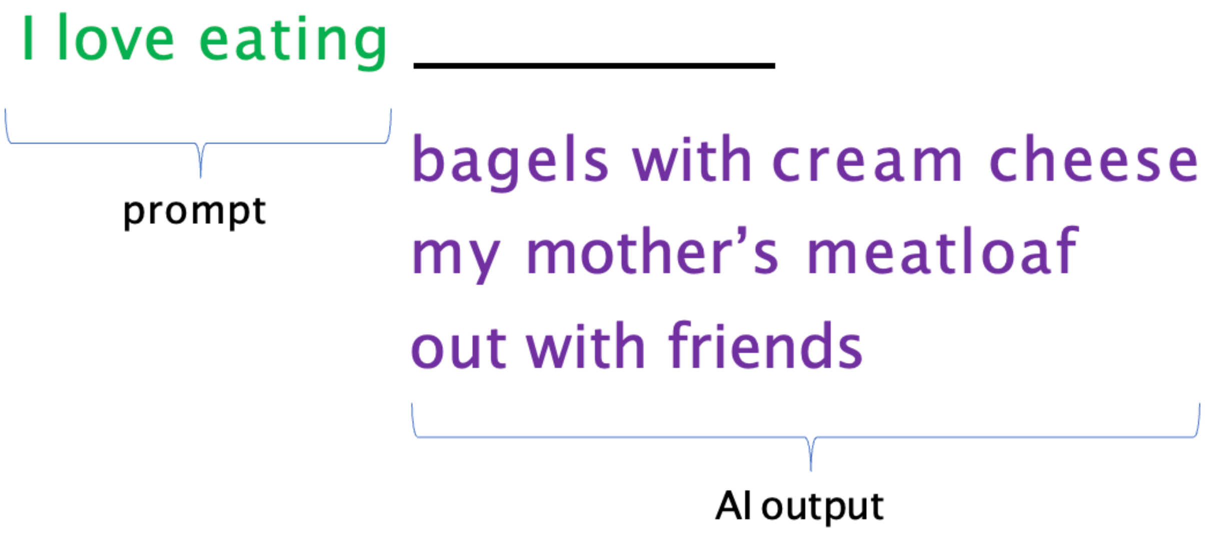
Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Ad, user info	Click? (0/1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
X-ray image	Diagnosis	Healthcare
Image of phone	Defect? (0/1)	Visual inspection
Audio recording	Text transcript	Speech recognition
Restaurant reviews	Sentiment (pos/neg)	Reputation monitoring

The trend of supervised learning



Large Language Models

Text generation process



I love eating ____________

prompt

bagels with cream cheese

my mother's meatloaf

out with friends

AI output

Training an LLM

LLMs are built by using supervised learning (A→B) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese

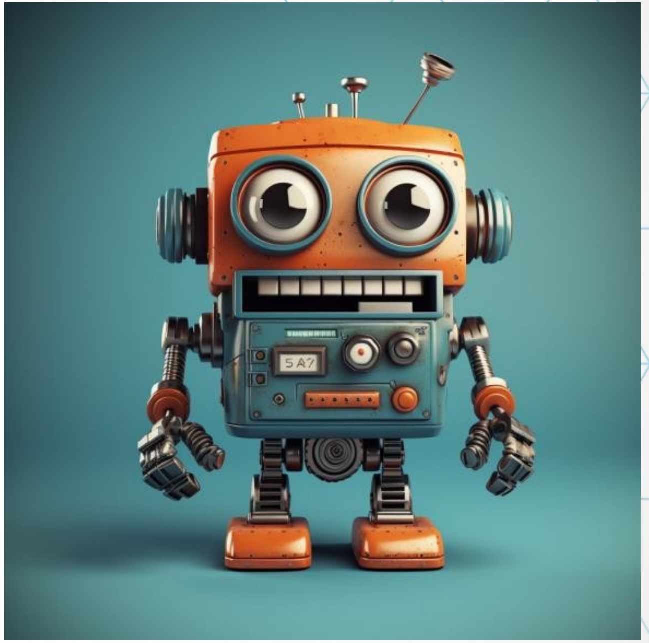
Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

When we train a very large AI system on a lot of data (hundreds of billions of words), we get a Large Language Model like ChatGPT.

Image Generation

A picture of a woman smiling

A futuristic city scene



A cool, happy robot

Image Processing

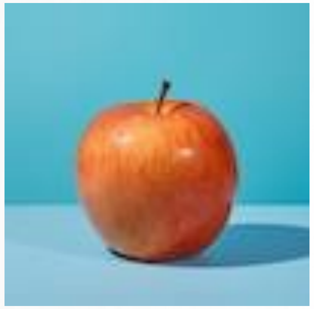


Image 1

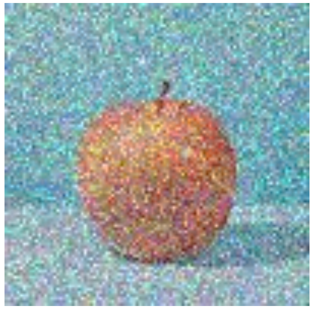


Image 2

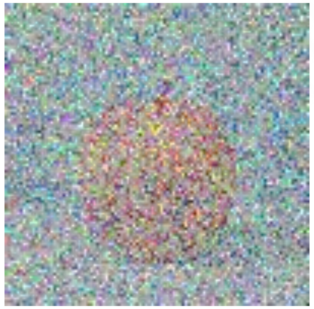


Image 3

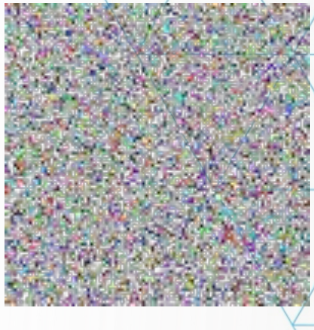


Image 4

Input (A)

Noisy image

➡

Output (B)

Slightly less noisy image

Input (A)	Output (B)
Image 2	Image 1
Image 3	Image 2
Image 4	Image 3

Text and Image

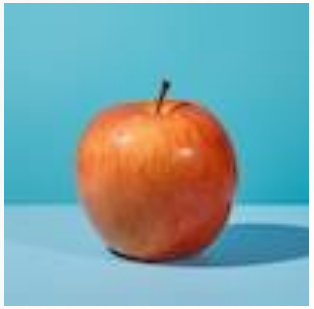


Image 1
"red apple"

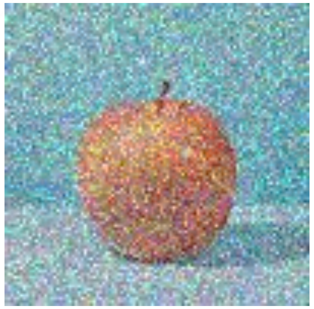


Image 2

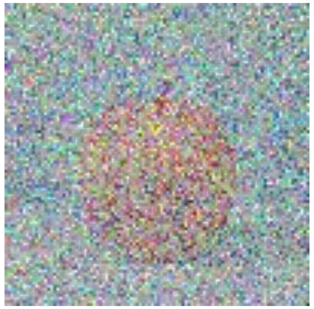


Image 3

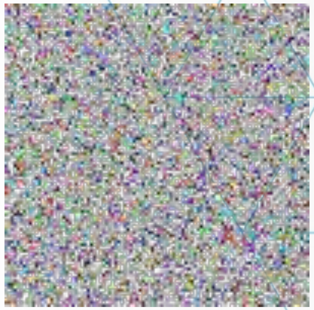


Image 4

Input (A)

Noisy image and caption

→

Output (B)

Slightly less noisy image

Input (A)	Output (B)
Image 2, "red apple"	Image 1
Image 3, "red apple"	Image 2
Image 4, "red apple"	Image 3

What LLM can do

Writing

Suggest three names for our new analytics platform

Sure, here are three ideas:
1. NexaInsight
2. Quantivue
3. AnalytiSphere

web

Is there parking for staff?

Yes, employees can park on levels 1and 2 of the office lot. You can get a sticker at [..

app

Reading

I love my new llama T-shirt! The fabric is so soft.

Complaint: NO
Department: Apparel

app

I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show.

Complaint: YES
Department: Apparel

app

Chatting

Welcome to BettaBurgers!

I'd like a cheeseburger for delivery!

Awesome. Anything else?

That's it thanks!

Ok, it'll be there in 20 minutes. Enjoy!

app

LLM disadvantages

- **Knowledge Cutoff**
- **Hallucination**
- **Structured Data**

Cutoff Knowledge

An LLM's knowledge of the world is frozen at the time of its training

- A model trained on data scraped from internet in January 2022 has no information about more recent events

What was the highest grossing film of 2022?

As of January 2022, I don't have data on the highest- grossing movie for that year. ✖

Avatar: The Way of Water

Hallucination

Give me three quotes that Shakespeare wrote about Beyonce

1. Her vocals shine like the sun.
2. All hail the queen, she is most worthy of love.
3. Such a voice, so electric and alive, none else can compare!

✖ Hallucination

Generative AI does not work well with structured (tabular) data

Home prices

size (square feet)	price (1000$)
523	100
645	150
708	200
1034	300
2290	350
2545	440

A B

Use supervised learning (A → B) to estimate price.

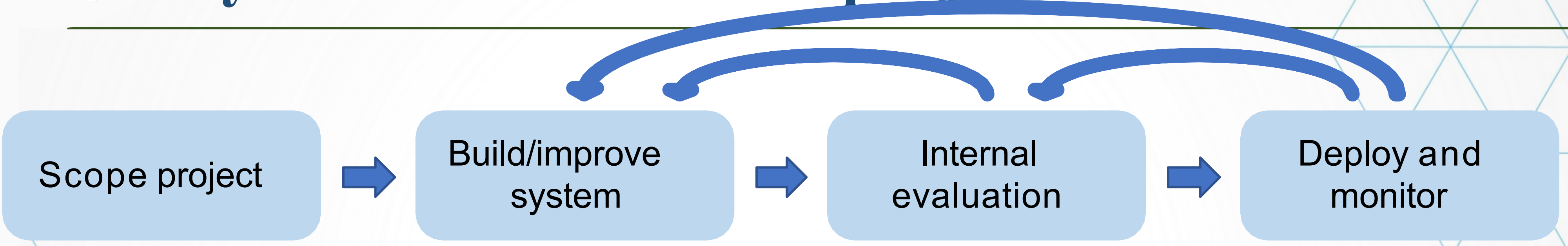
Purchases on website

user ID	time	price ($)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:13:.5	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes

A B

GEN AI PROJECTS

Lifecycle of Generative AI projects

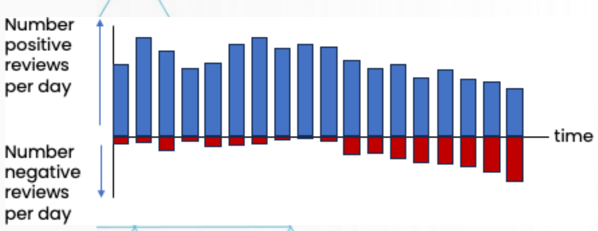


Scope project

Build/improve system

Internal evaluation

Deploy and monitor



Initially a prototype, that we will improve over time

Classify the sentiment of the following review as either positive or negative:

My miso ramen tasted like tonkotsu ramen.

Positive

GenAI applications

Writing

Developing sales strategy

Writing a press release

Translation

Reading

Proofreading

Summarizing an article

Summarizing call center conversations

Customer email analysis

Reputation monitoring

Chatbots

How can I vacation in Paris inexpensively?

Here are some ideas to save money in Paris:

1. Eat at bakeries
2. Take metro, not taxis
3. Visit free attractions […]

Trip planner

I'm nervous about my big presentation at work…

It's natural to feel nervous. What worries you most?

That I'll forget what to say…

You aren't alone! Here are some tips that may help:

1. Use index cards.
2. Picture a friend in the room to present to […]

Career coach

What can I make with the following ingredients?
Pasta, eggs, lemons, ham

Here's a recipe you can make with those ingredients:

Ham and Lemon Carbonara

Instructions:

1. Cook pasta
2. Whisk eggs, lemon juice and zest in bowl […]

Recipe ideas

Technology Options

Retrieval Augmented Generation

Give LLM access to external data sources

Fine-tuning LLM

Adapt LLM to your task

Pre-training LLM

Train LLM from scratch

Retrieval Augemented Generation

General Chatbot

Is there parking for employees?

I need more specific information about your workplace to answer that question.

Chatbot with RAG

Is there parking for employees?

Yes, employees can park on levels 1 and 2 of the office lot. You can get a sticker at […]

Pretraining and Finetuning

Pretraining

My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

Learns from 100Bs of words

Fine-tuning

What a wonderful chocolate cake
The novel was thrilling

Input (A)	Output (B)
What	a
What a	wonderful
What a wonderful	chocolate
What a wonderful chocolate	cake

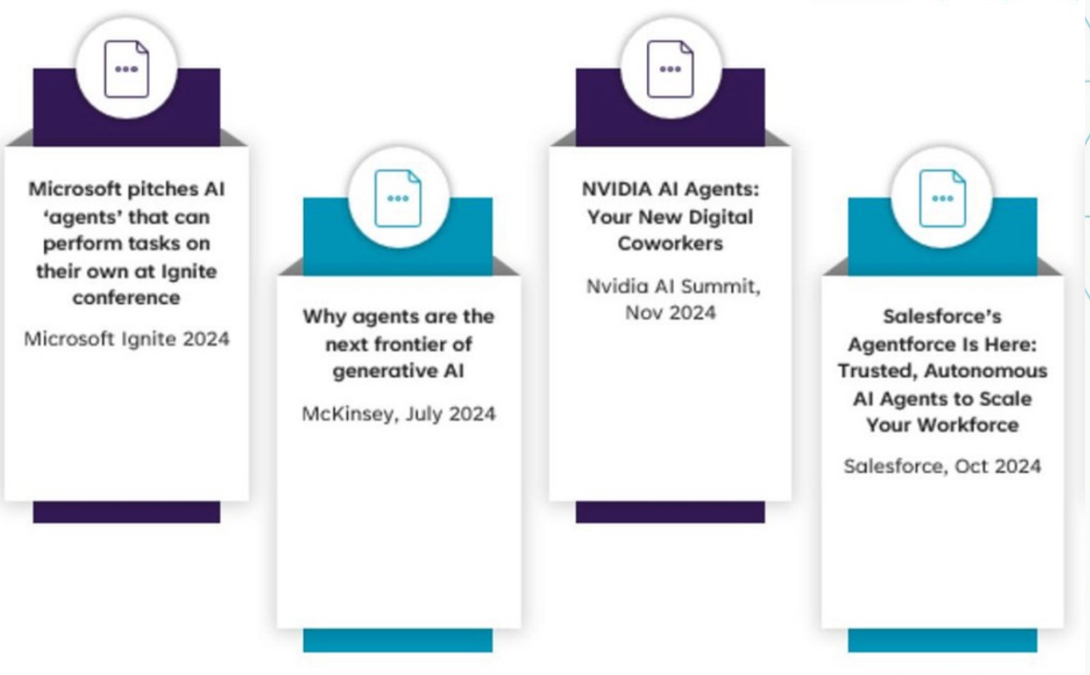
Learns from 1000s to 10,000s of words

AI Agents

AI Agents

In the Generative AI context, Agents are representative of an **Autonomous Agent** that can execute complex tasks, e.g.,

- -make a sale, -plan a trip, -
- make a flight booking, -
- book a contractor to do a
- house job, -order a pizza.
-



Microsoft pitches AI 'agents' that can perform tasks on their own at Ignite conference

Microsoft Ignite 2024

Why agents are the next frontier of generative AI

McKinsey, July 2024

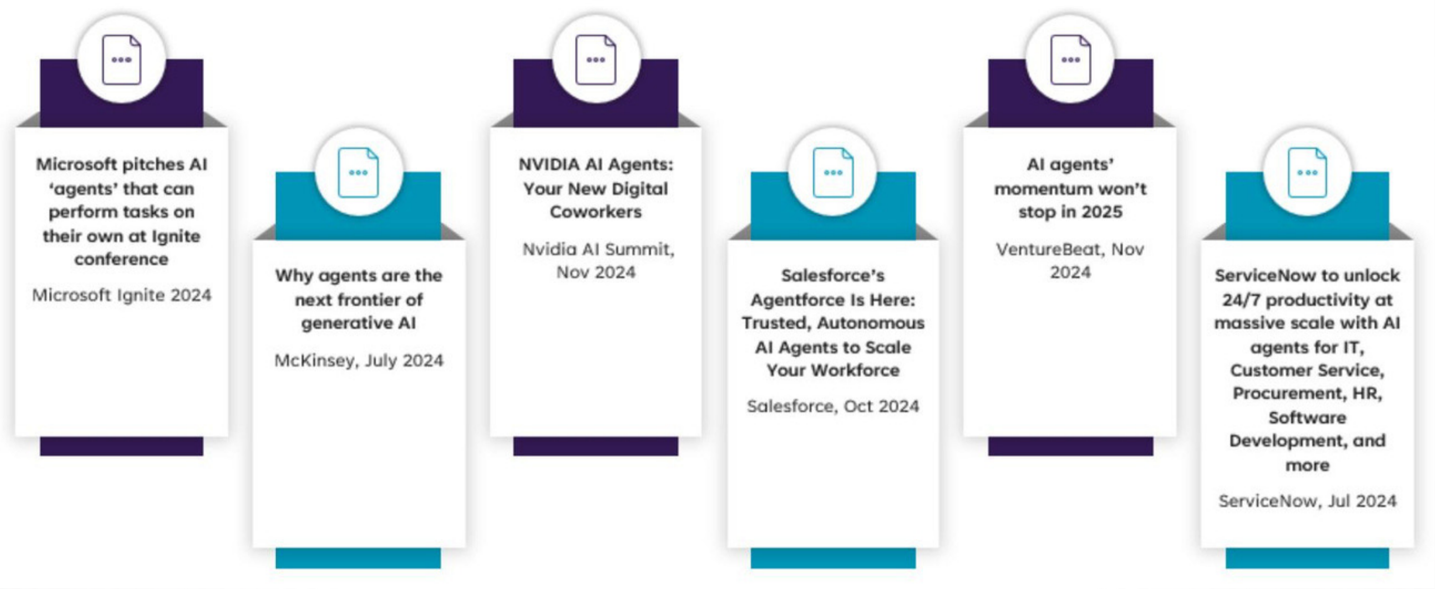
NVIDIA AI Agents: Your New Digital Coworkers

Nvidia AI Summit, Nov 2024

Salesforce's Agentforce Is Here: Trusted, Autonomous AI Agents to Scale Your Workforce

Salesforce, Oct 2024

Agentic AI in the News



Microsoft pitches AI 'agents' that can perform tasks on their own at Ignite conference

Microsoft Ignite 2024

Why agents are the next frontier of generative AI

McKinsey, July 2024

NVIDIA AI Agents: Your New Digital Coworkers

Nvidia AI Summit, Nov 2024

Salesforce's Agentforce Is Here: Trusted, Autonomous AI Agents to Scale Your Workforce

Salesforce, Oct 2024

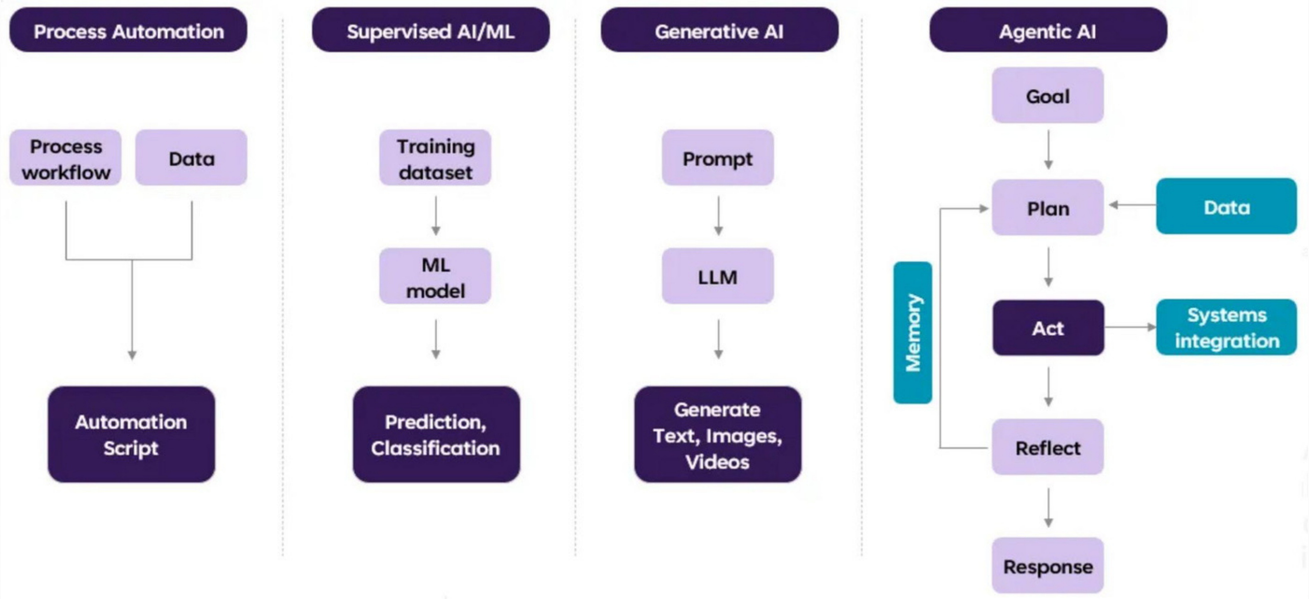
AI agents' momentum won't stop in 2025

VentureBeat, Nov 2024

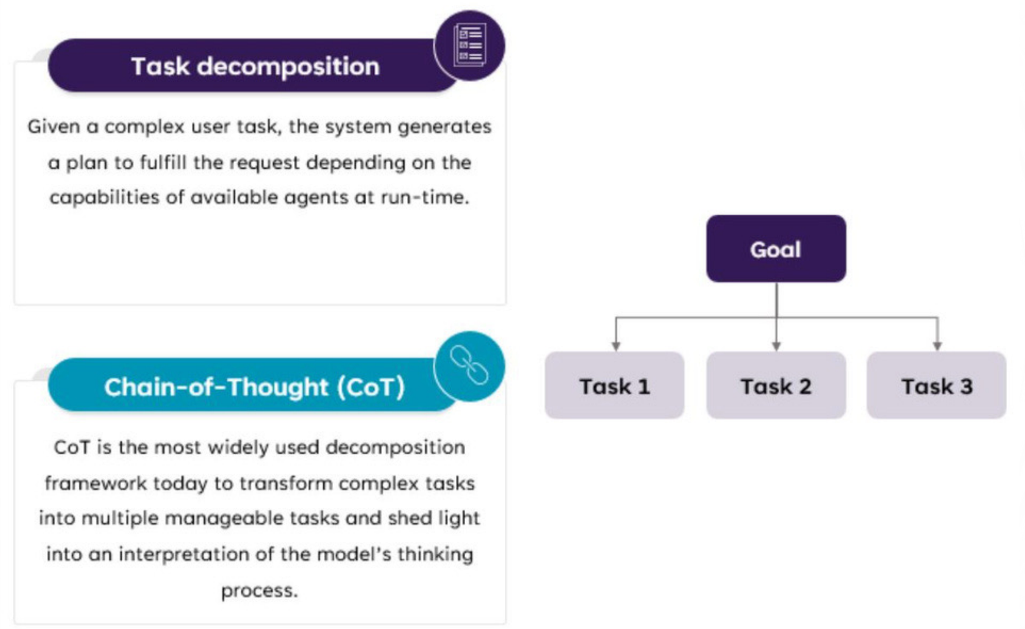
ServiceNow to unlock 24/7 productivity at massive scale with AI agents for IT, Customer Service, Procurement, HR, Software Development, and more

ServiceNow, Jul 2024

Agentic AI Evolution



Agentic AI capabilities –Task Decomposition



Task decomposition

Given a complex user task, the system generates a plan to fulfill the request depending on the capabilities of available agents at run-time.

Chain-of-Thought (CoT)

CoT is the most widely used decomposition framework today to transform complex tasks into multiple manageable tasks and shed light into an interpretation of the model's thinking process.

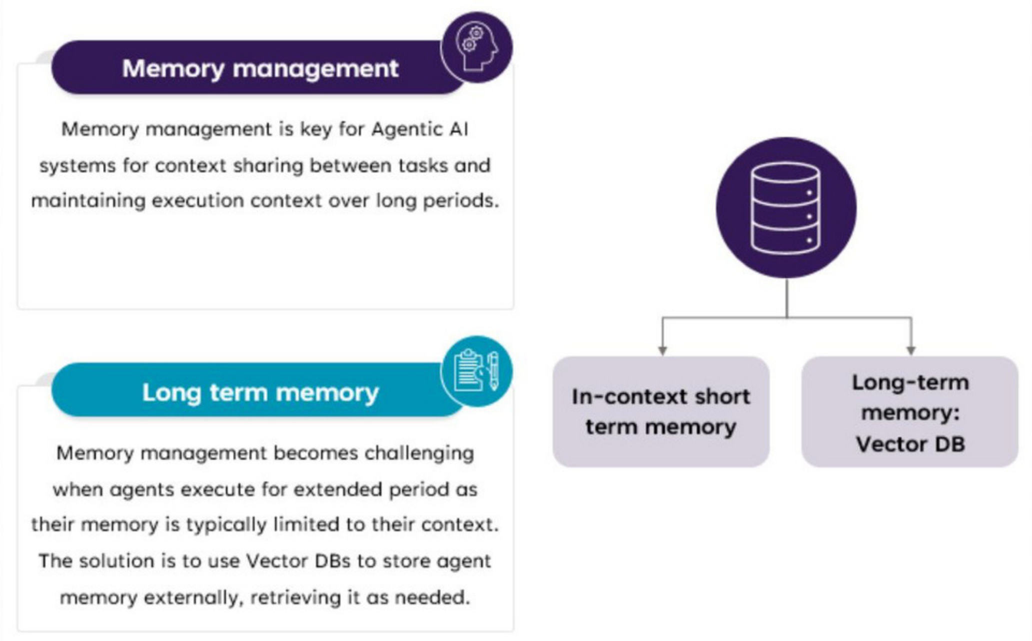
Goal

Task 1

Task 2

Task 3

Agentic AI capabilities –Memory Management



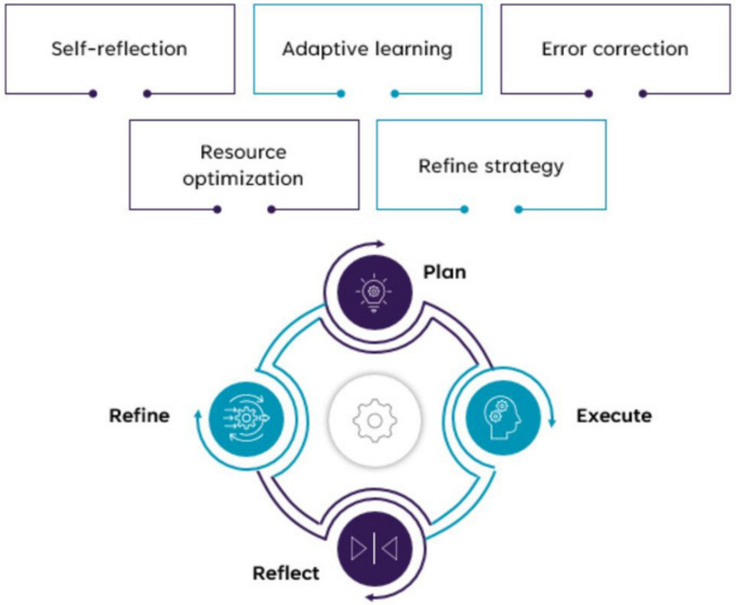
Memory management

Memory management is key for Agentic AI systems for context sharing between tasks and maintaining execution context over long periods.

Long term memory

Memory management becomes challenging when agents execute for extended period as their memory is typically limited to their context. The solution is to use Vector DBs to store agent memory externally, retrieving it as needed.

Agentic AI capabilities –Reflect & Adapt



Self-reflection

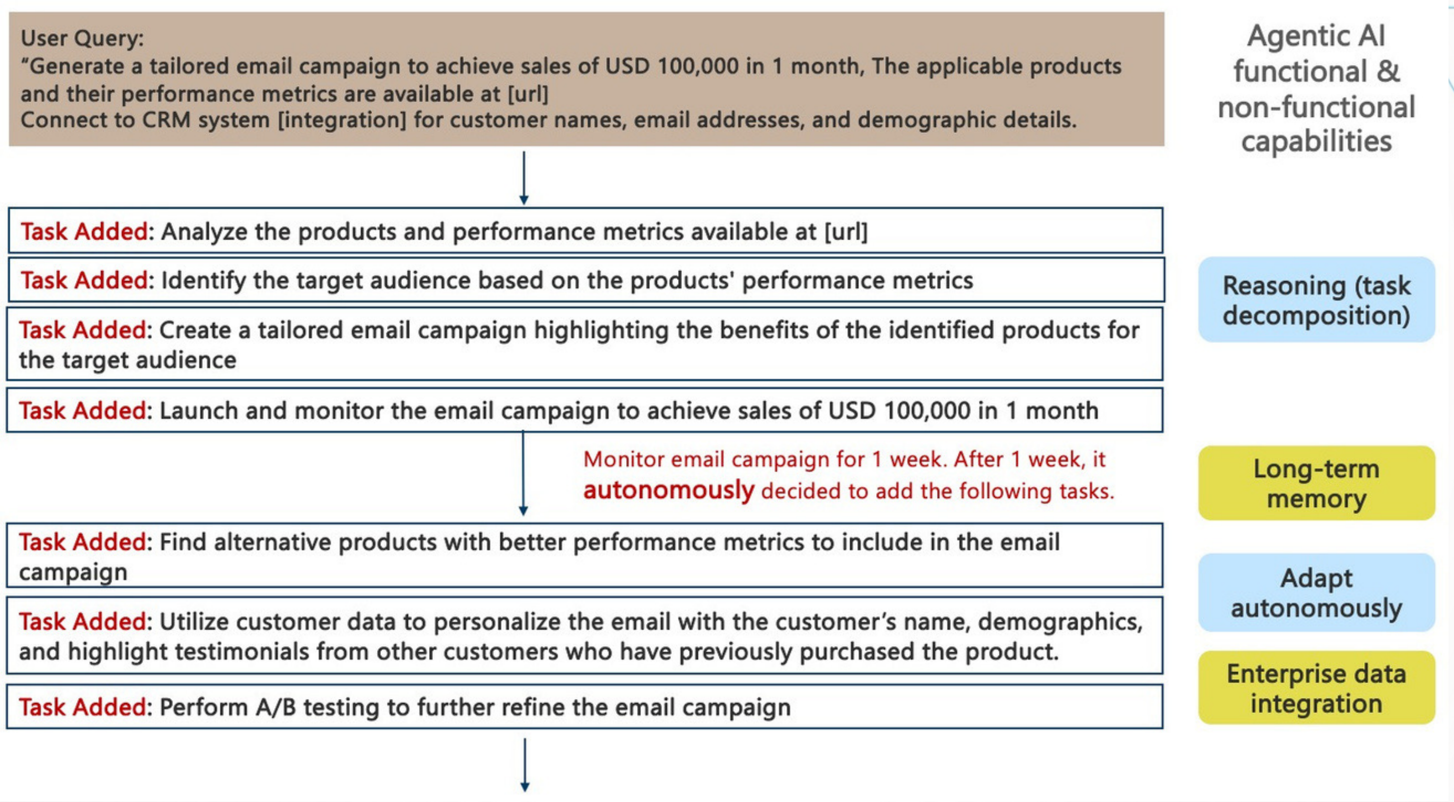
Adaptive learning

Error correction

Resource optimization

Refine strategy

Agentic AI Use-case: Funds Email Marketing Campaign



User Query:
"Generate a tailored email campaign to achieve sales of USD 100,000 in 1 month, The applicable products and their performance metrics are available at [url]
Connect to CRM system [integration] for customer names, email addresses, and demographic details.

Task Added: Analyze the products and performance metrics available at [url]

Task Added: Identify the target audience based on the products' performance metrics

Task Added: Create a tailored email campaign highlighting the benefits of the identified products for the target audience

Task Added: Launch and monitor the email campaign to achieve sales of USD 100,000 in 1 month

Monitor email campaign for 1 week. After 1 week, it **autonomously** decided to add the following tasks.

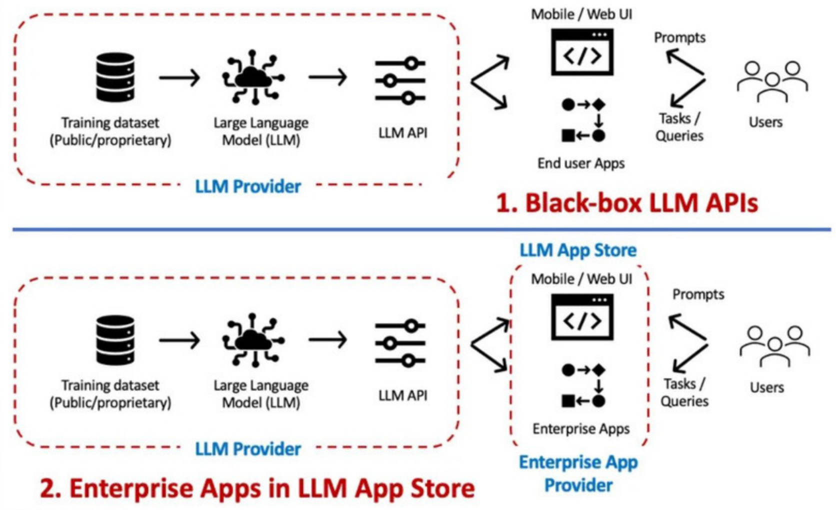
Task Added: Find alternative products with better performance metrics to include in the email campaign

Task Added: Utilize customer data to personalize the email with the customer's name, demographics, and highlight testimonials from other customers who have previously purchased the product.

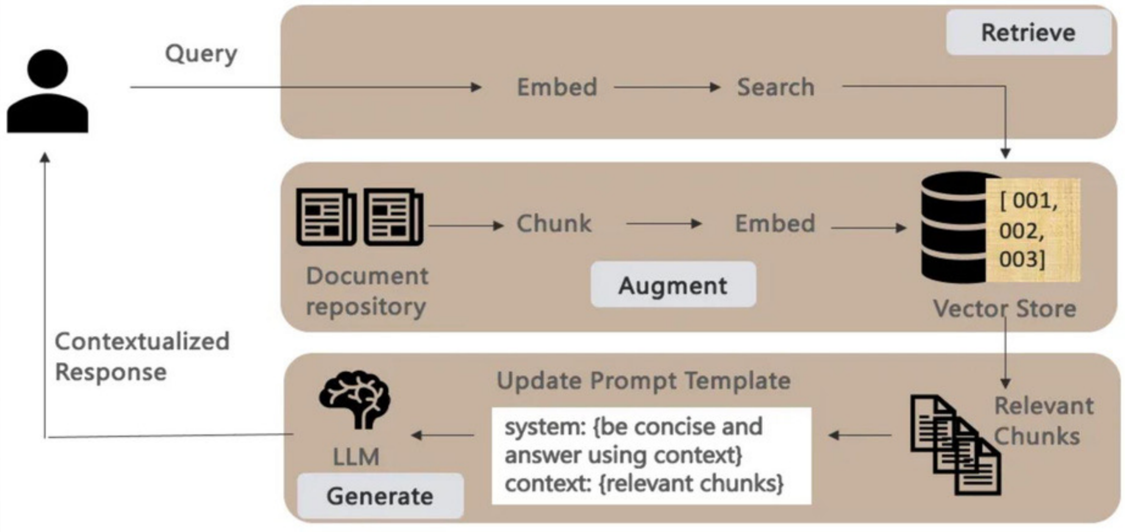
Task Added: Perform A/B testing to further refine the email campaign

Agentic AI functional & non-functional capabilities
Reasoning (task decomposition)
Long-term memory
Adapt autonomously
Enterprise data integration

Gen AI Architecture Patterns –APIs & Embedded Gen AI

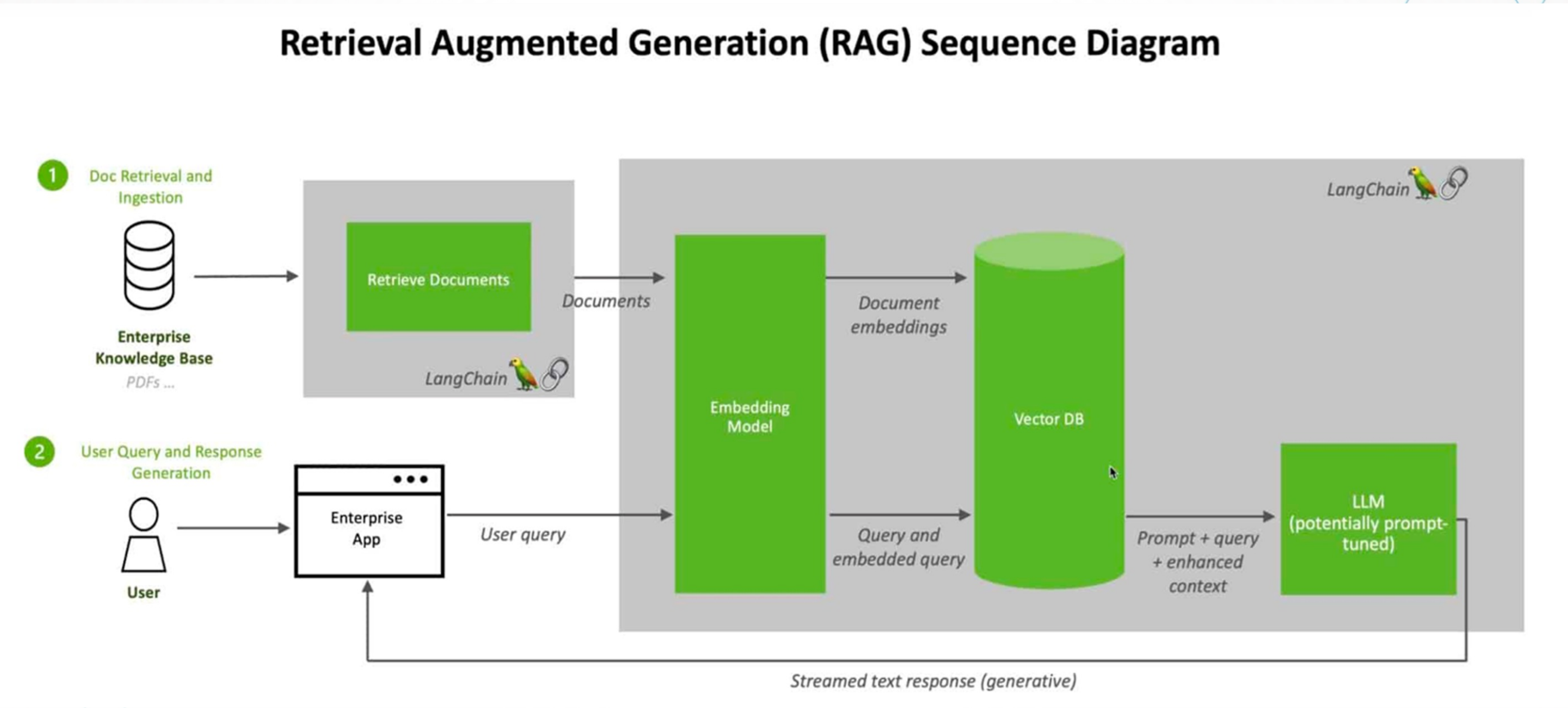


Gen AI Architecture Patterns –Retrieval-Augmented-Generation (RAG)



system: {be concise and answer using context}
context: {relevant chunks}

RAG Sequence diagram

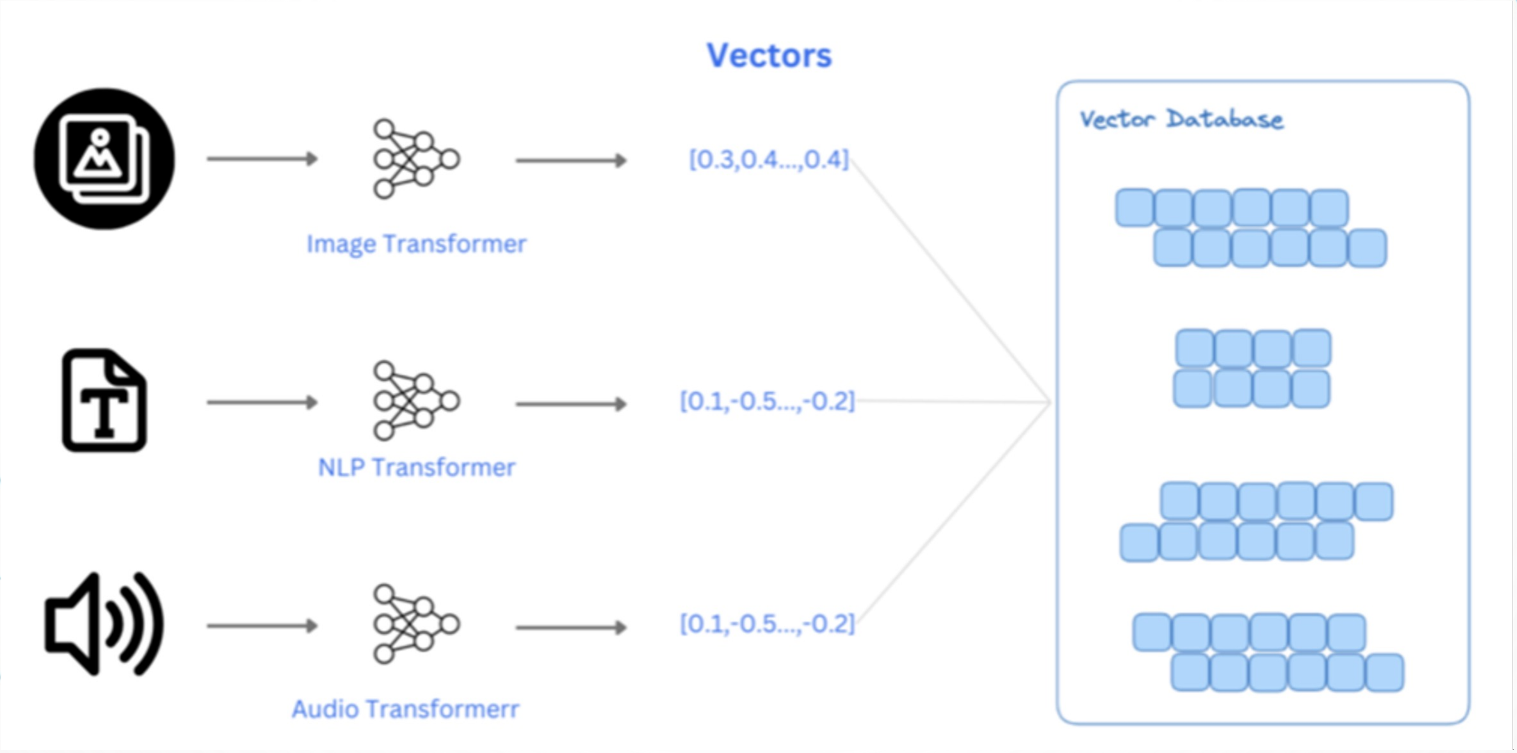


Retrieval Augmented Generation (RAG) Sequence Diagram

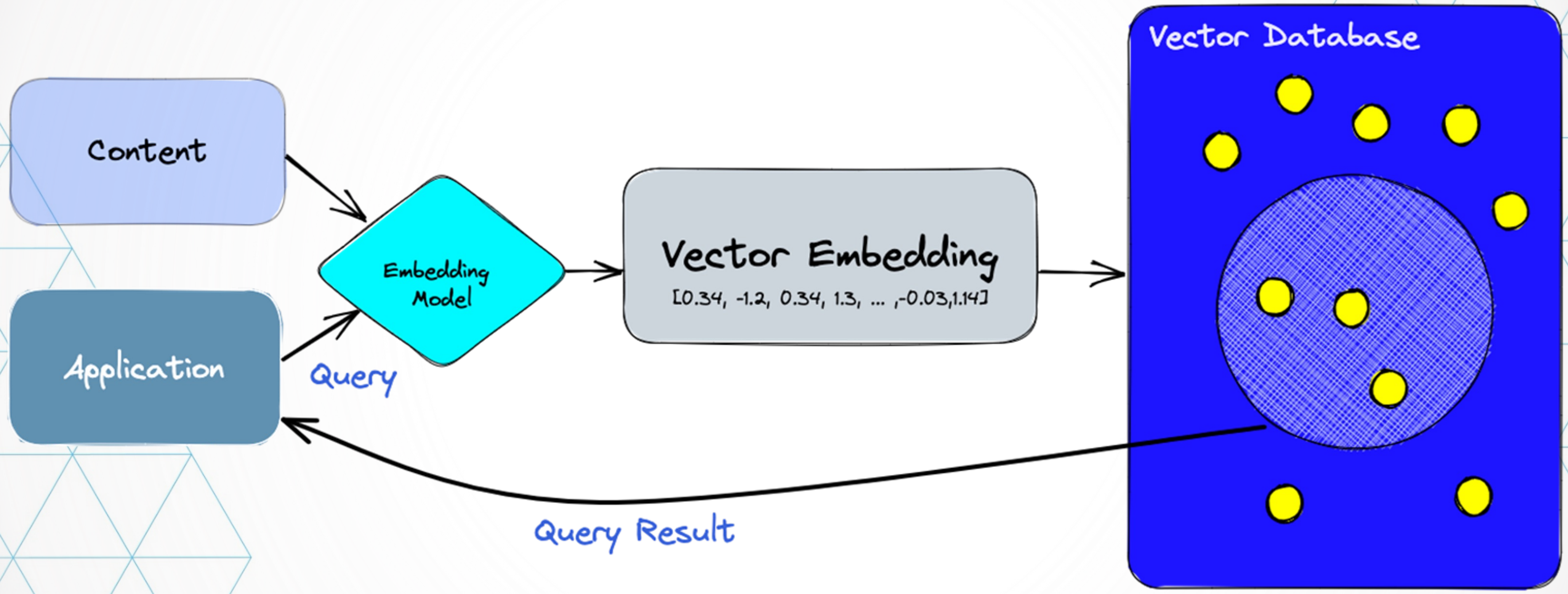
1
Doc Retrieval and Ingestion
Enterprise Knowledge Base
PDFs ...
Retrieve Documents
LangChain
Documents
Embedding Model
Document embeddings
Vector DB
LangChain
2
User Query and Response Generation
User
Enterprise App
User query
Query and embedded query
Prompt + query + enhanced context
LLM (potentially prompt-tuned)
Streamed text response (generative)

Vector Database

A vector database is a type of database that indexes and stores vector embeddings for fast retrieval and similarity search, with capabilities like CRUD operations, metadata filtering, and horizontal scaling.



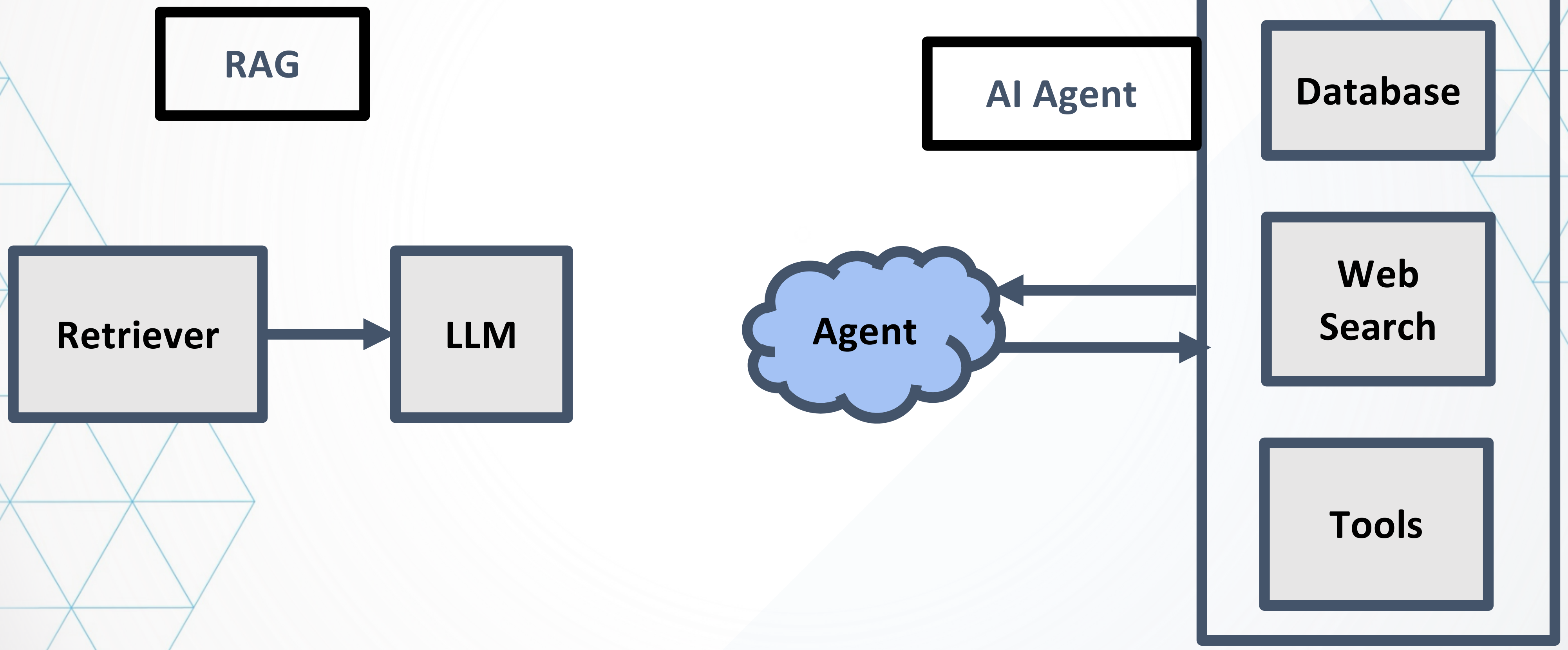
Vector Database



Beyond RAGs

The era of AI Agents

RAG vs AI Agent



RAG

Retriever

LLM

AI Agent

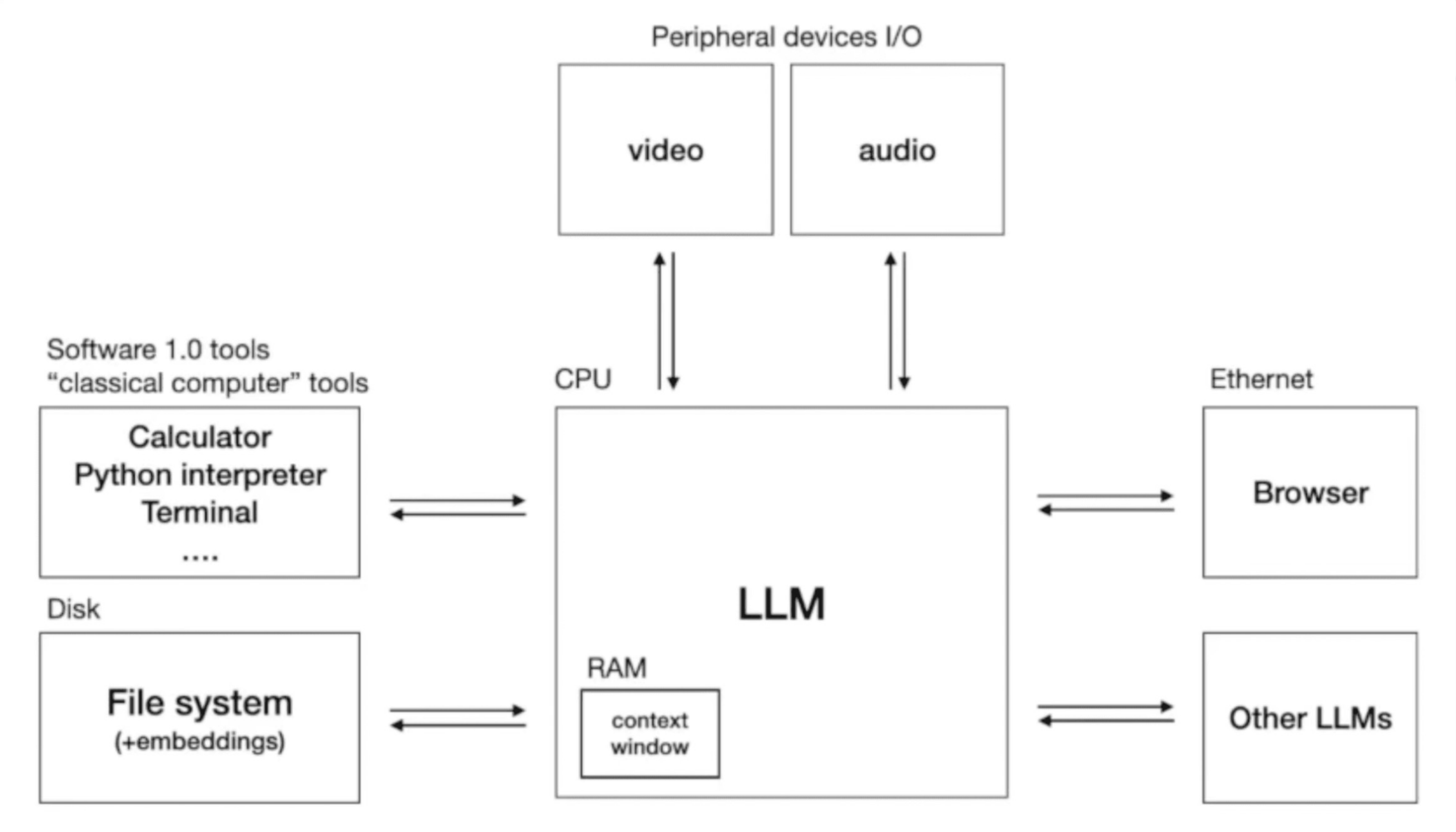
Agent

Database

Web Search

Tools

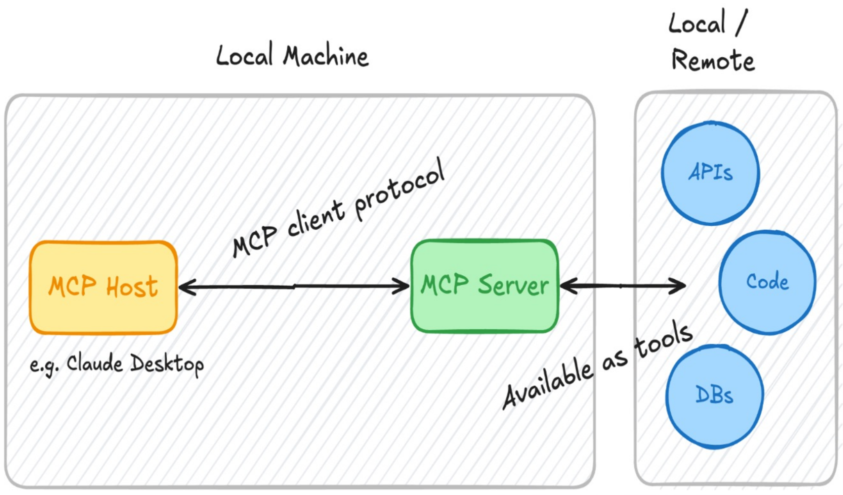
LLM OS


Peripheral devices I/O
video
audio
Software 1.0 tools
"classical computer" tools
Calculator
Python interpreter
Terminal
....
CPU
Ethernet
Browser
LLM
Disk
File system
(+embeddings)
RAM
context window
Other LLMs

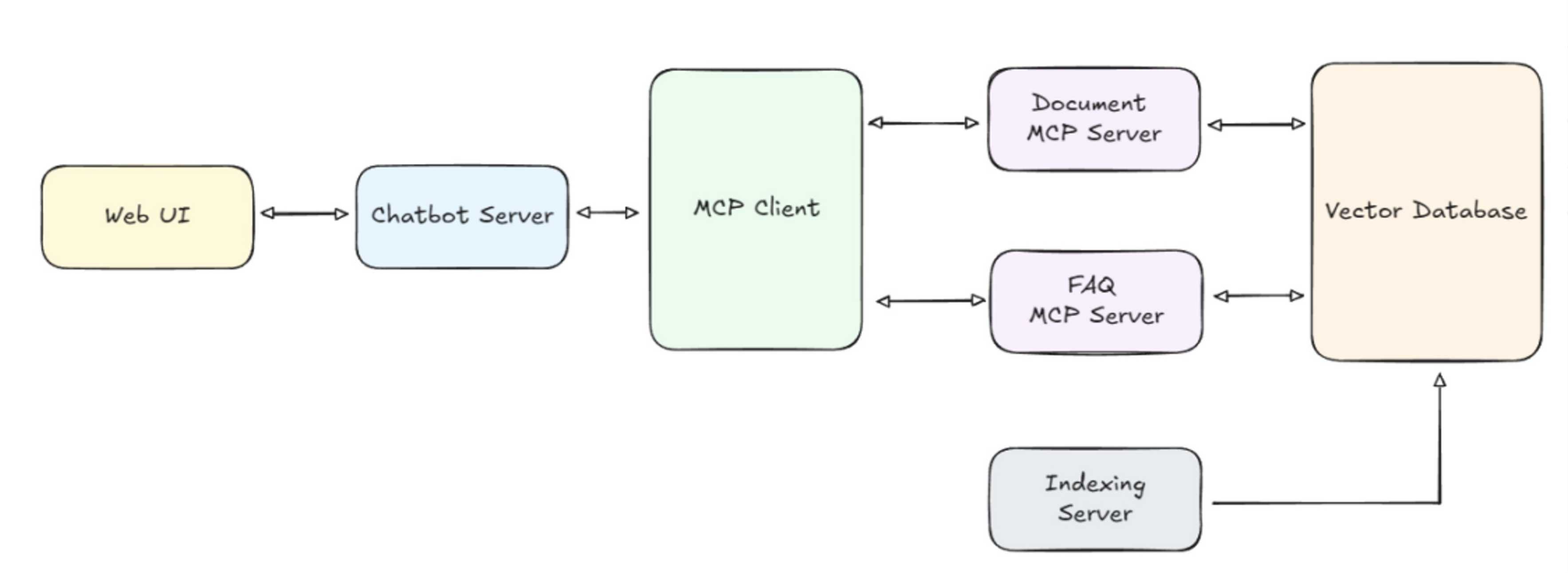
What is MCP?

Model Context Protocol (MCP) is an open protocol that standardizes how applications provide context to LLMs. Think of MCP like a USB-C port for AI applications. MCP provides a standardized way to connect AI models to different data sources and tools.

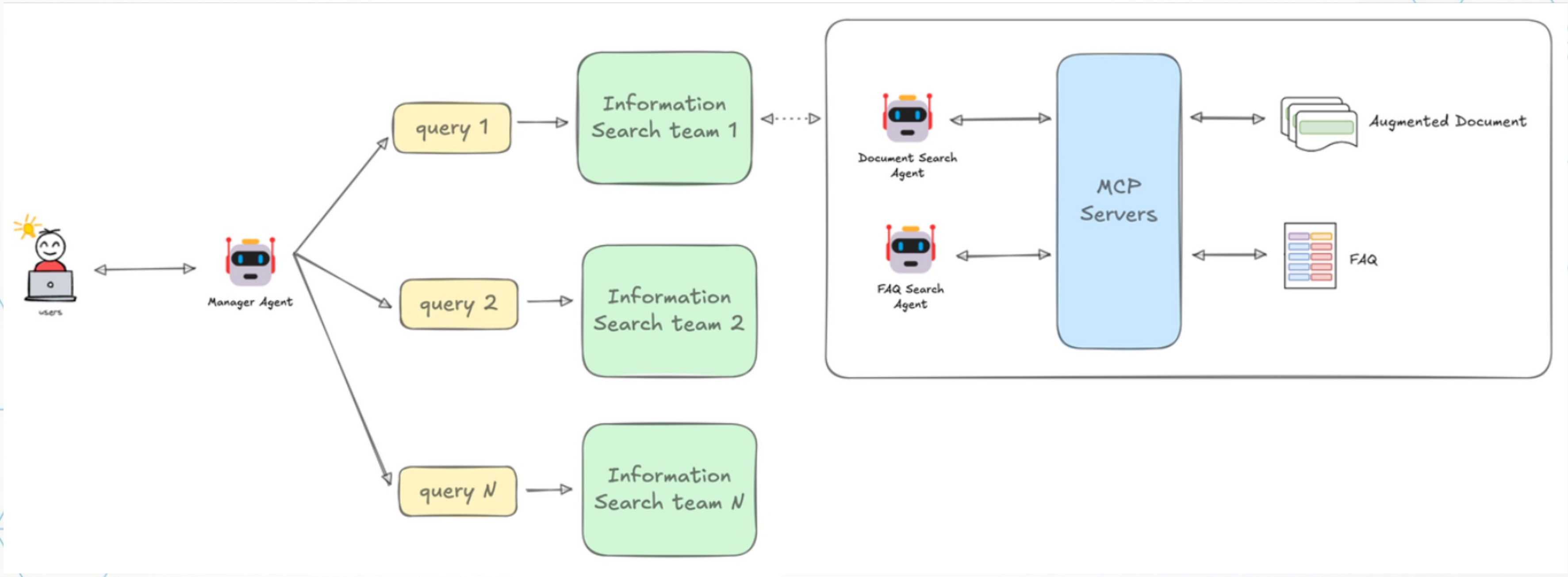
MCP follows a client-server architecture where a host application can connect to multiple servers:



Case Study: System overview

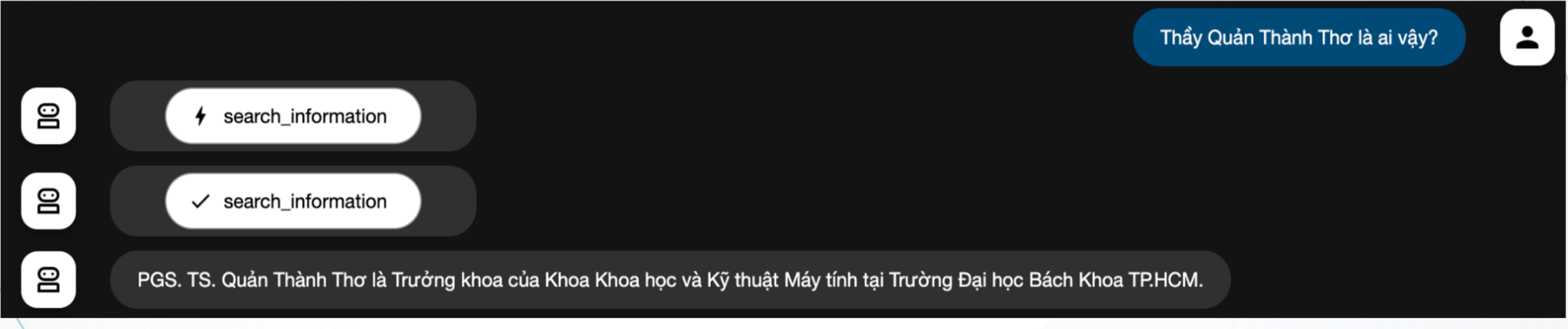


Inference pipeline



Use cases

Case 1 – Clear & Specific Query



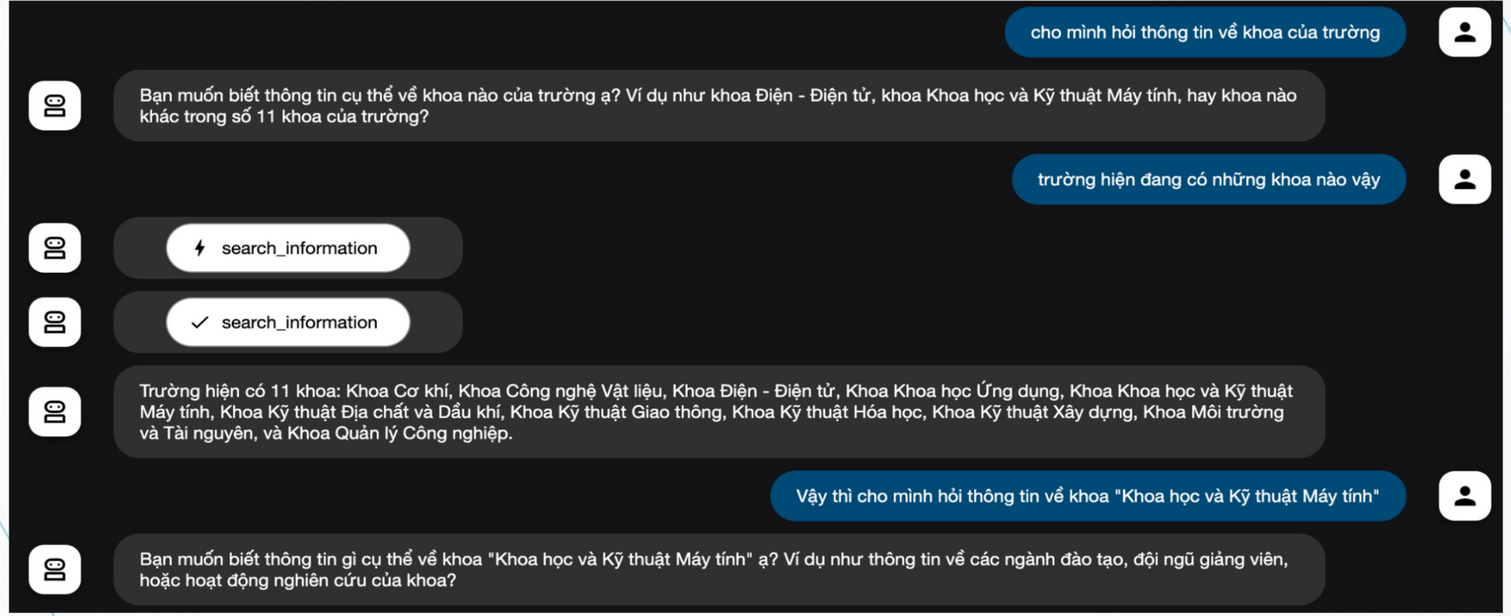
Thầy Quản Thành Thơ là ai vậy?

PGS. TS. Quản Thành Thơ là Trưởng khoa của Khoa Khoa học và Kỹ thuật Máy tính tại Trường Đại học Bách Khoa TP.HCM.

Detects that **required info slot is already complete** → triggers a single retrieval call and answers immediately.

Use cases

Case 2 – Broad or Vague Query



cho mình hỏi thông tin về khoa của trường

Bạn muốn biết thông tin cụ thể về khoa nào của trường ạ? Ví dụ như khoa Điện - Điện tử, khoa Khoa học và Kỹ thuật Máy tính, hay khoa nào khác trong số 11 khoa của trường?

trường hiện đang có những khoa nào vậy

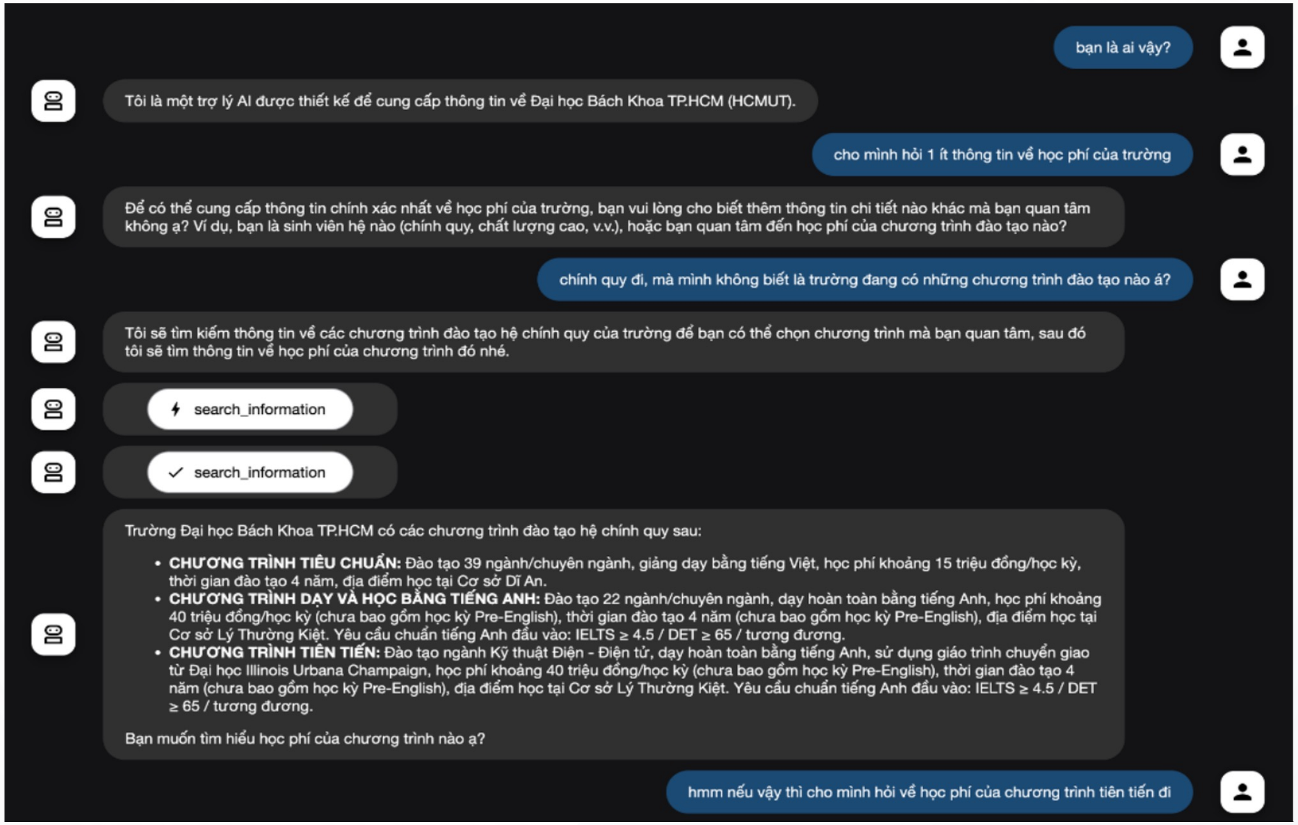
Trường hiện có 11 khoa: Khoa Cơ khí, Khoa Công nghệ Vật liệu, Khoa Điện - Điện tử, Khoa Khoa học Ứng dụng, Khoa Khoa học và Kỹ thuật Máy tính, Khoa Kỹ thuật Địa chất và Dầu khí, Khoa Kỹ thuật Giao thông, Khoa Kỹ thuật Hóa học, Khoa Kỹ thuật Xây dựng, Khoa Môi trường và Tài nguyên, và Khoa Quản lý Công nghiệp.

Vậy thì cho mình hỏi thông tin về khoa "Khoa học và Kỹ thuật Máy tính"

Bạn muốn biết thông tin gì cụ thể về khoa "Khoa học và Kỹ thuật Máy tính" ạ? Ví dụ như thông tin về các ngành đào tạo, đội ngũ giảng viên, hoặc hoạt động nghiên cứu của khoa?

Sees missing detail, does a quick exploratory search, then asks user a clarifying question to narrow scope. Iterates clarify → targeted search → answer until scope is specific enough, then delivers the response.

DEMO



bạn là ai vậy?

Tôi là một trợ lý AI được thiết kế để cung cấp thông tin về Đại học Bách Khoa TP.HCM (HCMUT).

cho mình hỏi 1 ít thông tin về học phí của trường

Để có thể cung cấp thông tin chính xác nhất về học phí của trường, bạn vui lòng cho biết thêm thông tin chi tiết nào khác mà bạn quan tâm không ạ? Ví dụ, bạn là sinh viên hệ nào (chính quy, chất lượng cao, v.v.), hoặc bạn quan tâm đến học phí của chương trình đào tạo nào?

chính quy đi, mà mình không biết là trường đang có những chương trình đào tạo nào á?

Tôi sẽ tìm kiếm thông tin về các chương trình đào tạo hệ chính quy của trường để bạn có thể chọn chương trình mà bạn quan tâm, sau đó tôi sẽ tìm thông tin về học phí của chương trình đó nhé.

Trường Đại học Bách Khoa TP.HCM có các chương trình đào tạo hệ chính quy sau:

- **CHƯƠNG TRÌNH TIÊU CHUẨN:** Đào tạo 39 ngành/chuyên ngành, giảng dạy bằng tiếng Việt, học phí khoảng 15 triệu đồng/học kỳ, thời gian đào tạo 4 năm, địa điểm học tại Cơ sở Dĩ An.
- **CHƯƠNG TRÌNH DẠY VÀ HỌC BẰNG TIẾNG ANH:** Đào tạo 22 ngành/chuyên ngành, dạy hoàn toàn bằng tiếng Anh, học phí khoảng 40 triệu đồng/học kỳ (chưa bao gồm học kỳ Pre-English), thời gian đào tạo 4 năm (chưa bao gồm học kỳ Pre-English), địa điểm học tại Cơ sở Lý Thường Kiệt. Yêu cầu chuẩn tiếng Anh đầu vào: IELTS ≥ 4.5 / DET ≥ 65 / tương đương.
- **CHƯƠNG TRÌNH TIÊN TIẾN:** Đào tạo ngành Kỹ thuật Điện - Điện tử, dạy hoàn toàn bằng tiếng Anh, sử dụng giáo trình chuyển giao từ Đại học Illinois Urbana Champaign, học phí khoảng 40 triệu đồng/học kỳ (chưa bao gồm học kỳ Pre-English), thời gian đào tạo 4 năm (chưa bao gồm học kỳ Pre-English), địa điểm học tại Cơ sở Lý Thường Kiệt. Yêu cầu chuẩn tiếng Anh đầu vào: IELTS ≥ 4.5 / DET ≥ 65 / tương đương.

Bạn muốn tìm hiểu học phí của chương trình nào ạ?

hmm nếu vậy thì cho mình hỏi về học phí của chương trình tiên tiến đi

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

BK TP.HCM

Thank you

https://www.cse.hcmut.edu.vn/

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA