- -Các nhánh chính của AI:
- +Machine Learning: day cho máy tính học từ dữ liệu để đưa ra dự đoán hoặc quyết định

Ví dụ : dự đoán giá nhà, giá cổ phiếu, hệ thống gợi ý

+Deep learning (thuộc ML) : sử dụng mạng noron nhiều lớp để đưa ra kết quả

Ví dụ : nhận dạng khuôn mặt, xử lý ảnh

+GenAI: tạo ra nội dung mới như văn bản, hình ảnh,âm thanh, ... bằng các kỹ thuật như fine-tuning, prompt, RAG, ..

Ví dụ : chatbot, content marketing, sáng tác nghệ thuật

+NLP : giúp máy tính hiểu và sinh ngôn ngữ tự nhiên như con người bằng các kỹ thuật như tokenization, ...

Ví dụ: Dịch tự động, trợ lý ảo

+ Computer Vision : giúp máy hiểu nội dung từ ảnh

Ví dụ: Nhận diện khuôn mặt, ...

+Robotics: Kết hợp AI và phần cứng (robot công nghiệp,..)

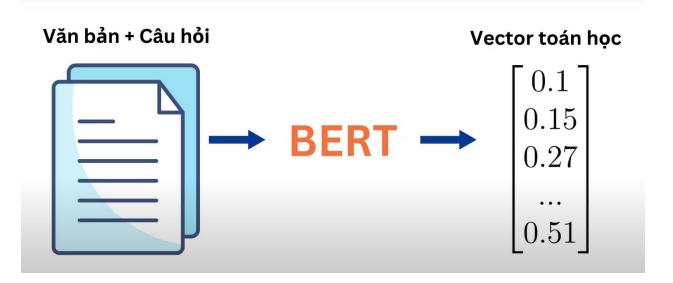
+Expert Systems : hệ thống chẩn đoán bệnh,...

+Planning: giải các bài toán lên lịch, đường đi ngắn nhất,...

## -Lô trình làm chatbot :

https://chatgpt.com/share/687e1689-8130-800a-91d5-5c67ef07ad3d

- -Tìm hiểu về RAG và dify (tự động hóa workflow, có 2 loại là chatflow và workflow):
- +Các mô hình ngôn ngữ lớn thường có limited knowledge, việc re-training dữ liệu mới cũng khiến tốn nhiều chi phí
  - Khi gắng hỏi thì hay bị hallucination (ảo giác kiến thức)
- +RAG gồm hai thành phần:
  - + Retrieve : thu thập các văn bản có liên quan đến câu hỏi của người dùng



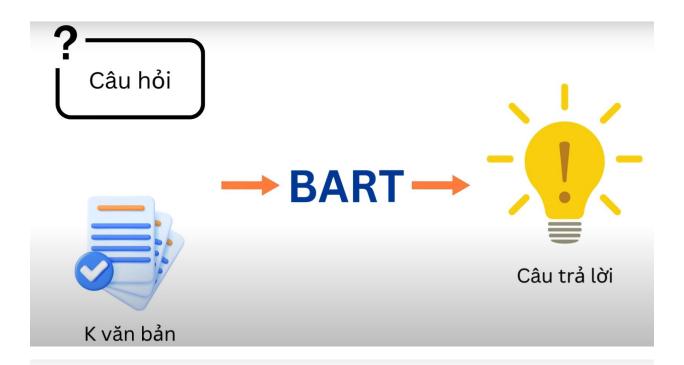
## Tìm K văn bản sao cho

$$d^Tq\,$$
 lớn nhất

d: vector mã hoá văn bản

q: vector mã hoá câu hỏi

+ Generate : Tạo sinh ngôn ngữ sử dụng các văn bản liên quan



## Tìm y sao cho

p(y|x,z) lớn nhất

y: chuỗi câu trả lời

x: chuỗi câu hỏi

z: chuỗi các văn bản liên quan

+ Ví du: