Slot17 – Giới thiệu về AI
- Tự build 1 model AI hỗ trợ các công ty
=> mỗi 1 công ty sẽ có 1 model riêng
========
=> offline chạy trên server công ty
=> Cấu trúc
+ Giao diện: Viết bằng Nodejs (Express), React Native
+ AI Backend: python (FastAPI, LangChain, Ollama)
+Model: phi3 hoặc llama3 (cài qua Ollama)
=== Dự kiến tối thiểu 20GB===
React AI Demo
    web
        server.js
        package.json
        public
            index.html
    ai
        main.py # FastAPI AI Backend
        requirements.txt
=======
Thư viện
B1 = Cài Ollama và Model AI
brew install ollama
ollama pull phi3 #llama3
Cài xong thì khởi động Ollama:
ollama serve
=====
Mô hình 2:

Sử dụng Gemini API Key => train để Gemini hiểu các tài liệu mà chúng ta cung cấp
=> hiện nay Gemini chưa cho fine-tine trực tiếp như GPT, Claude
Nhưng: cho phép xây dựng 1 lớp “kiến thức riêng” bằng kỹ thuật RAG (Retrivel-Augmented Generation)
===
Cấu trúc
React Native => Backend (Nodejs) => Python Service
+ Trích xuất dữ liệu (PyMuPDF)
+Tạo embendding (Google Text Embedding API)
+Lưu dữ liệu (FAIS hoặc ChromeDB)
=> Gemini API:
+Nhận promt (câu hỏi, tài liệu trích dẫn)
+Sinh câu trả lời theo ngữ cảnh
====
Các bước triển khai với Gemini:
B1- Cài
npm install express body-parser axios
pip install PyMuPDF chromadb google-generativeai
========