Ứng dụng GenAI trong Tổ chức & Doanh nghiệp

Triển khai Generative AI dựa trên dữ liệu riêng của Doanh nghiệp

MYGPT

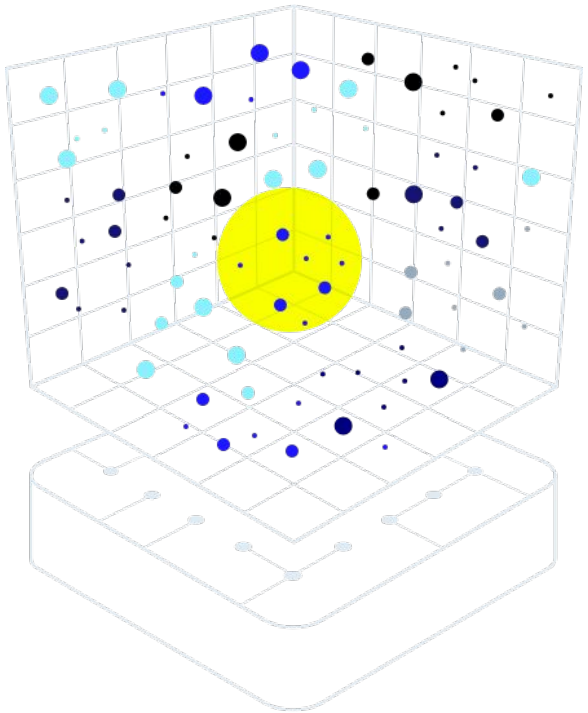
Tạ Quang Thái
Tech advisor - MyGPT
mygpt.vn

21/01/2025

Thuật ngữ

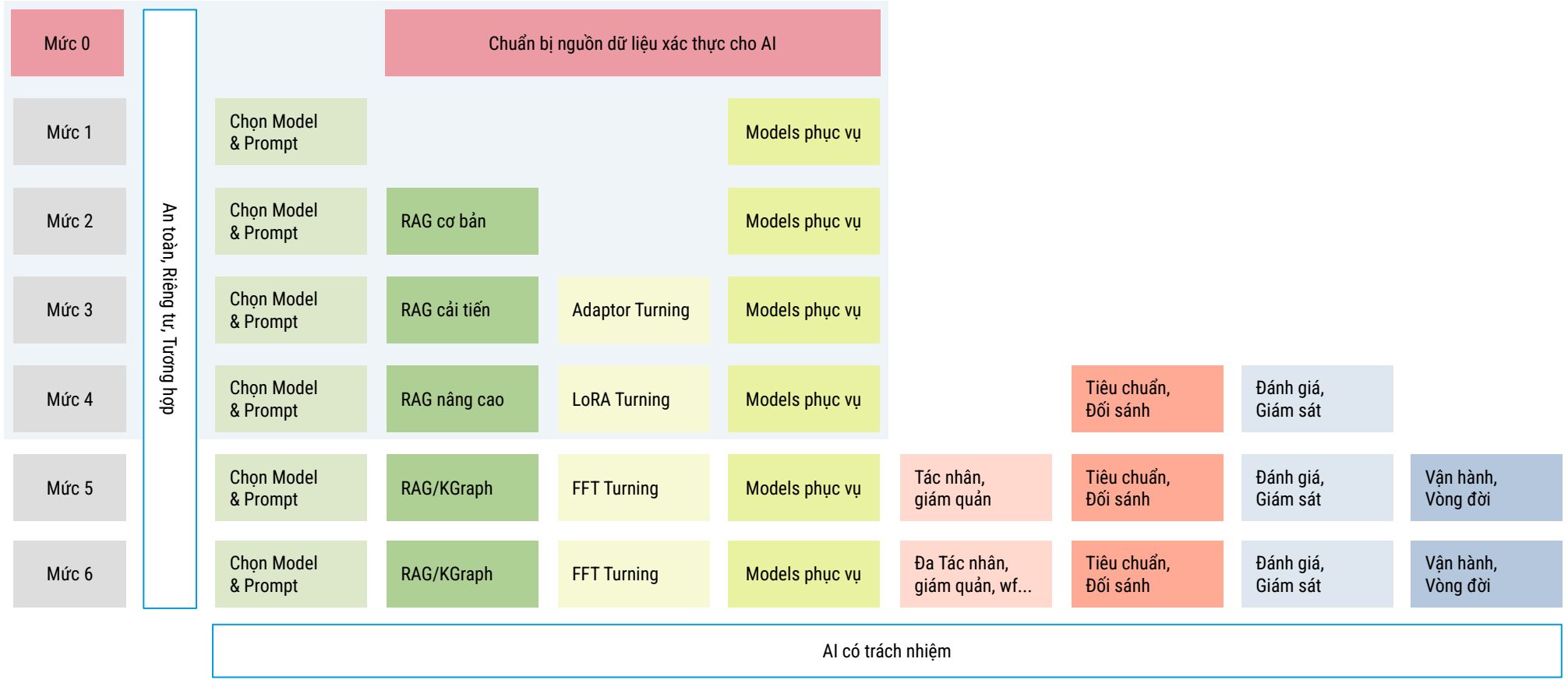
- **LLMs:** Mô hình ngôn ngữ lớn
- **Foundation Model:** Mô hình ngôn ngữ gốc
- **OpenAI, Meta, Anthropic, Nvidia...:** Một số tổ chức giới thiệu và huấn luyện các mô hình ngôn ngữ gốc
- **NLP:** Xử lý ngôn ngữ tự nhiên
- **GPT:** Thuật toán sản sinh chữ (tạo sinh) dựa trên Mô hình ngôn ngữ lớn được huấn luyện sẵn
- **GPT x, Llama x, Claude x, Gemini, Qwen, Mistral, Gemma, DeepSeek...: M**ột số mô hình ngôn ngữ được huấn luyện sẵn với nhiều đóng gói có số lượng tham số khác nhau từ 3tỉ, 13 tỷ, 65 tỷ… đến 175 tỉ tham số, 1000 tỷ, 2xxx tỷ. (tham số: trọng số (weights) & độ chệch (biases) trong mạng nơ-ron)
- **ChatGPT, Claude 2, Gemini, Bing, Co-pilot:** Là một ứng dụng để tương tác với các LLMs...
- **MyGPT**: Cung cấp dịch vụ huấn luyện AI bao gồm **XỬ LÝ DỮ LIỆU & PHÁT TRIỂN PHẦN MỀM** với dữ liệu riêng của các tổ chức, doanh nghiệp.

Technology stack

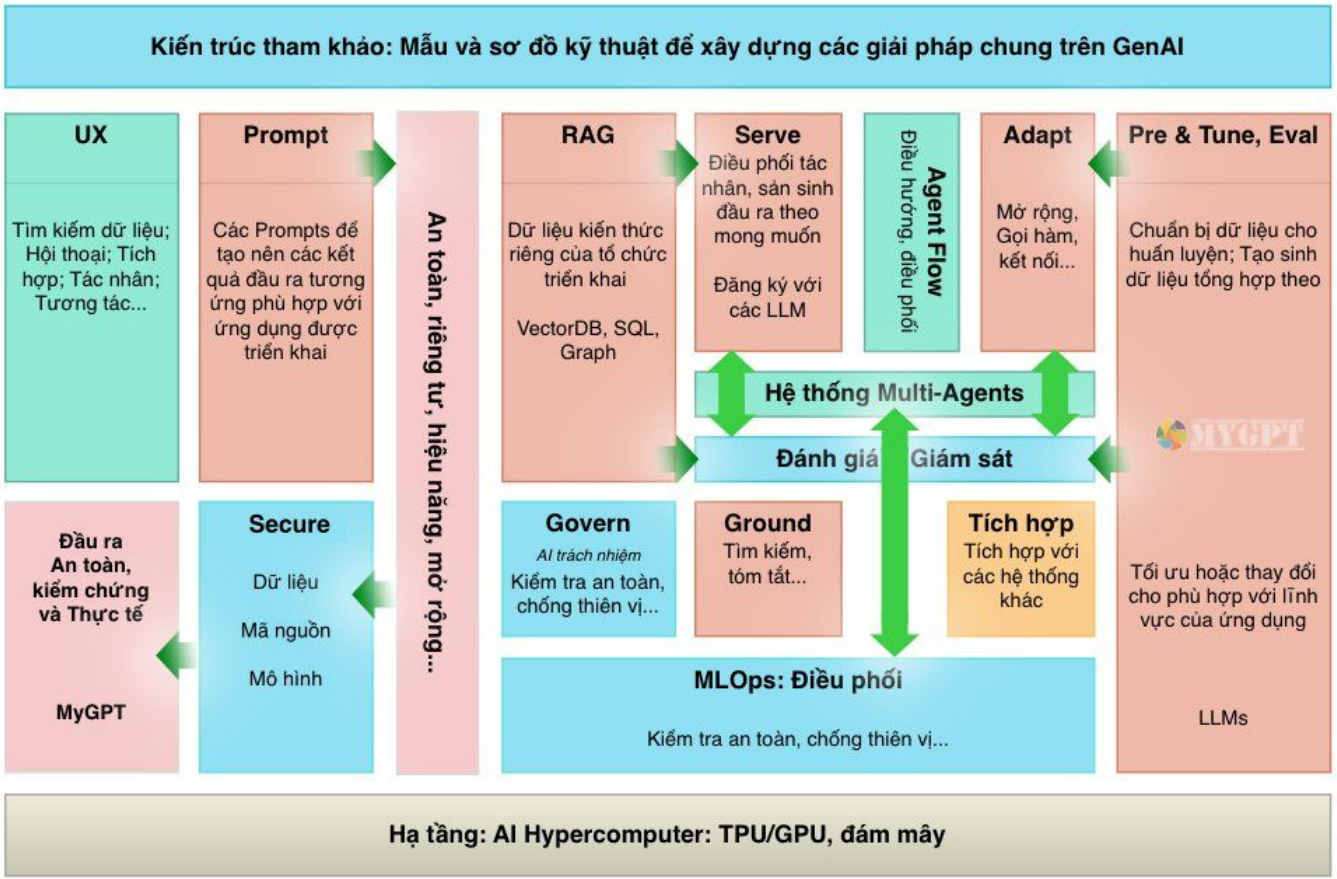


- **Data processing & back-end application:** R, Python, Node.js, Next.js...
- **NLP Framework:** NTTK, Bert, OpenNLP, TensorFlow, Keras, Theano, spaCy…
- **LLMs framework:** Langchain/LangGraph, Awesome-LLM, CodiumAI, Llama Index, Auto-GPT, Ollama...
- **Database:** Pinecone, Supabase, Neo4J, Faiss, Chroma, Milvus…
- **Front application:** Next.js, React.js
- **Server:** Debian 10 & up
- **Type:** OpenSource

7 mức độ ứng dụng GenAI



Kiến trúc tham chiếu ứng dụng GenAI



Lưu ý:

- Triển khai từng phần theo yêu cầu của ứng dụng trong tổ chức
- Áp dụng với các dữ liệu công khai trước khi hướng đến các dữ liệu riêng tư.
- Tập trung vào dữ liệu tĩnh trước khi làm đến dữ liệu động
- Đặt mục tiêu giảm nhẹ phiền phức trước khi tính đến chuyện thay thế
- Tránh locked in (all in) vào một nhà cung cấp hoặc 1 tech stack cụ thể
- Giám sát và tái huấn luyện liên tục nhằm tối ưu kết quả

Tổng quát triển khai ứng dụng GenAI

Data Analysis

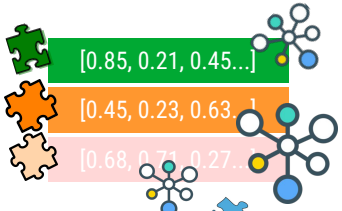
Processing, modeling

Objects, Dataset

Embedding/Pipeline

Embedding/Extraction Model
GPT, BERT, Doc2Vec, Llama...

fine-tuning



NLU

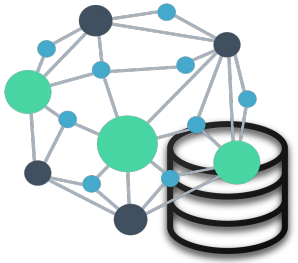
GPUs system

Foundation Model
GPT3.5 turbo, GPT4, Germini hoặc local model (PhoGPT, Mistral, Llama, Gema...)

Prompting

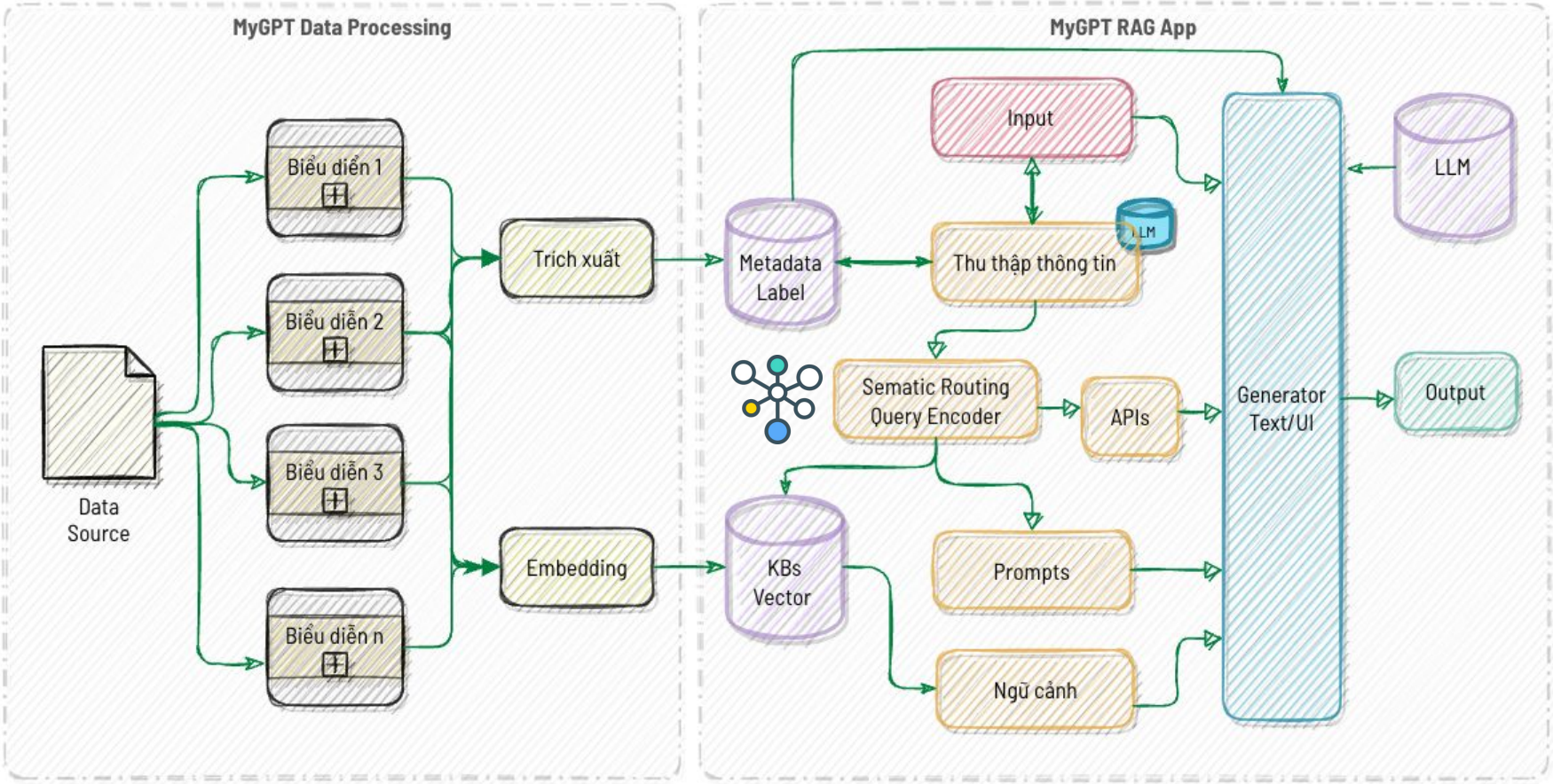
Application/Agents

Vectors Search/Prompt



Local Vectors/DB/KGraph or LLM

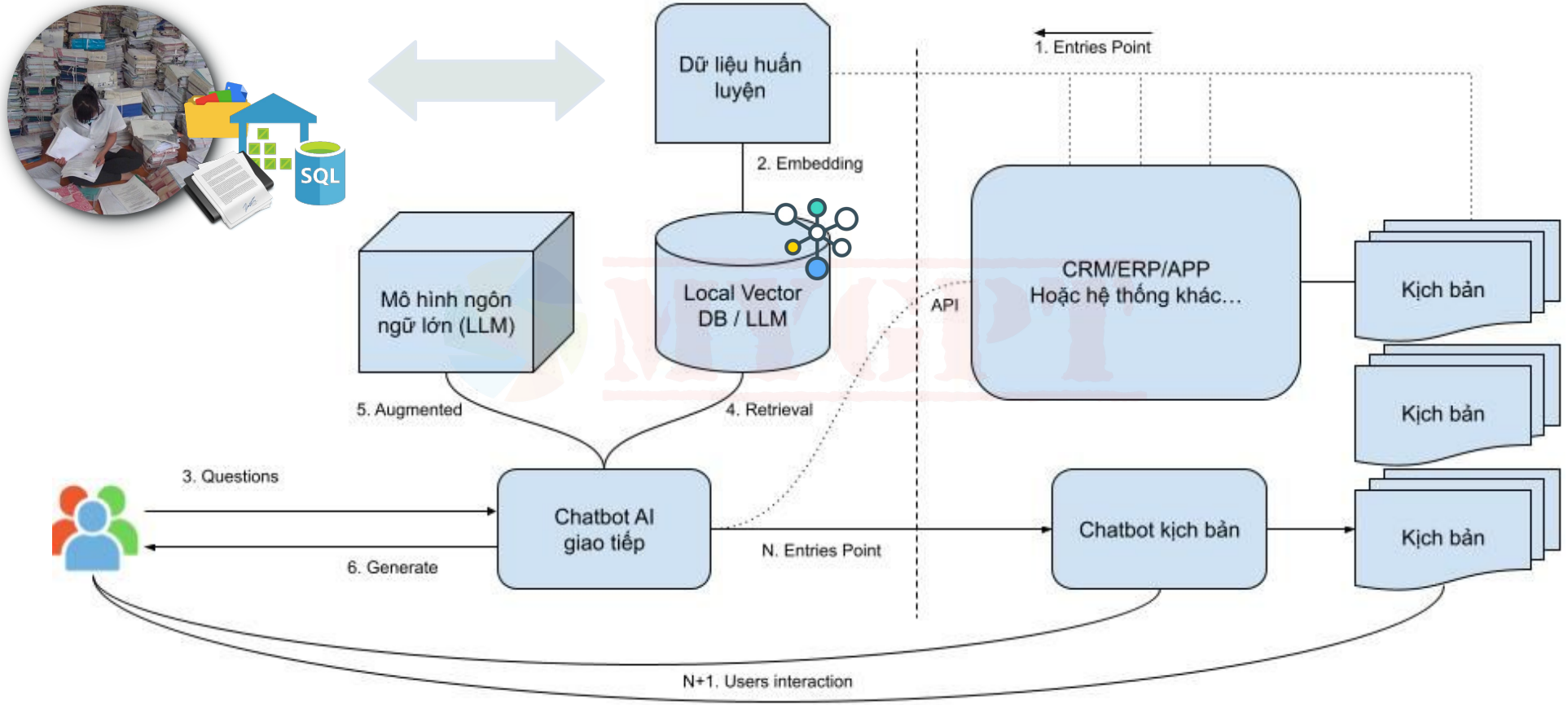
Truy xuất tăng cường



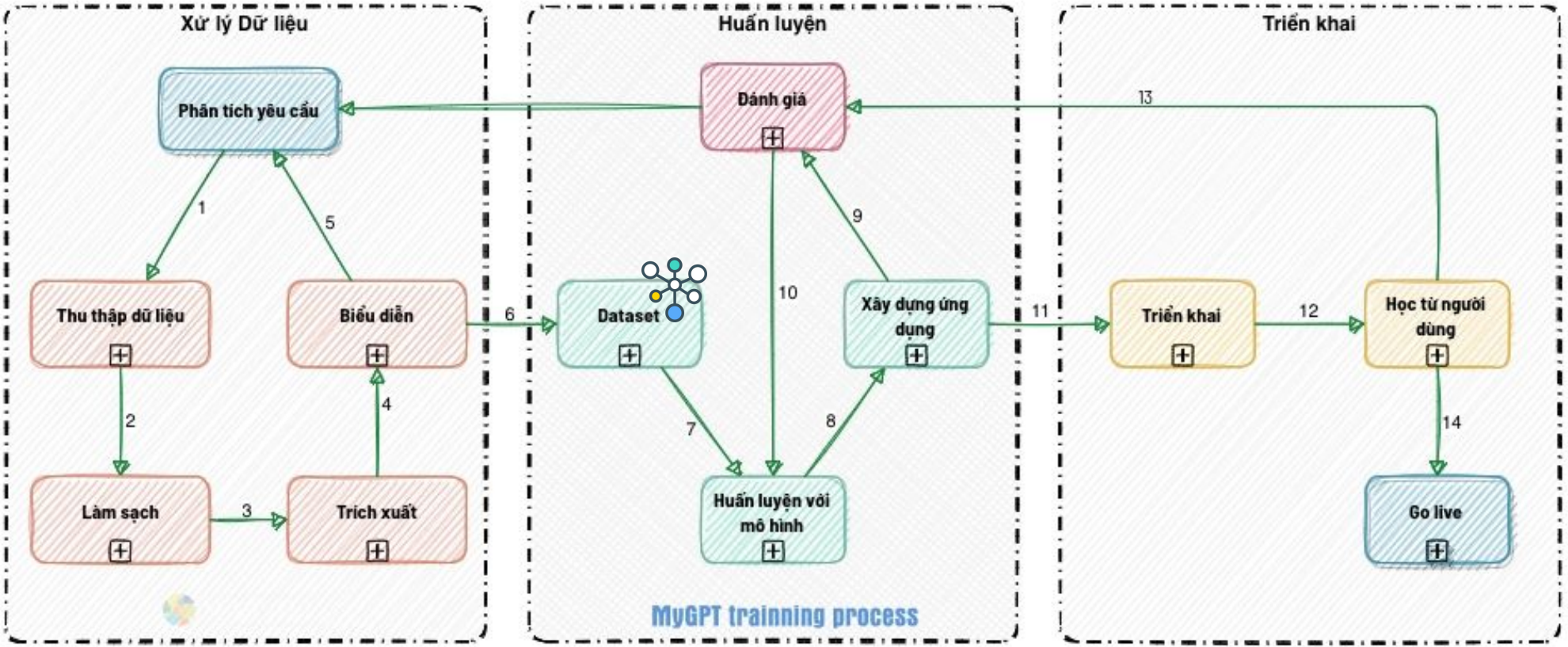
Highlight:

- Biểu diễn
- Thu thập thông tin
- Metadata
- Semantic Routing
- Multiple Prompts

Ứng dụng GenAI với hệ thống khác



Chu trình triển khai chi tiết



Đặc điểm kỹ thuật của ứng dụng GenAI

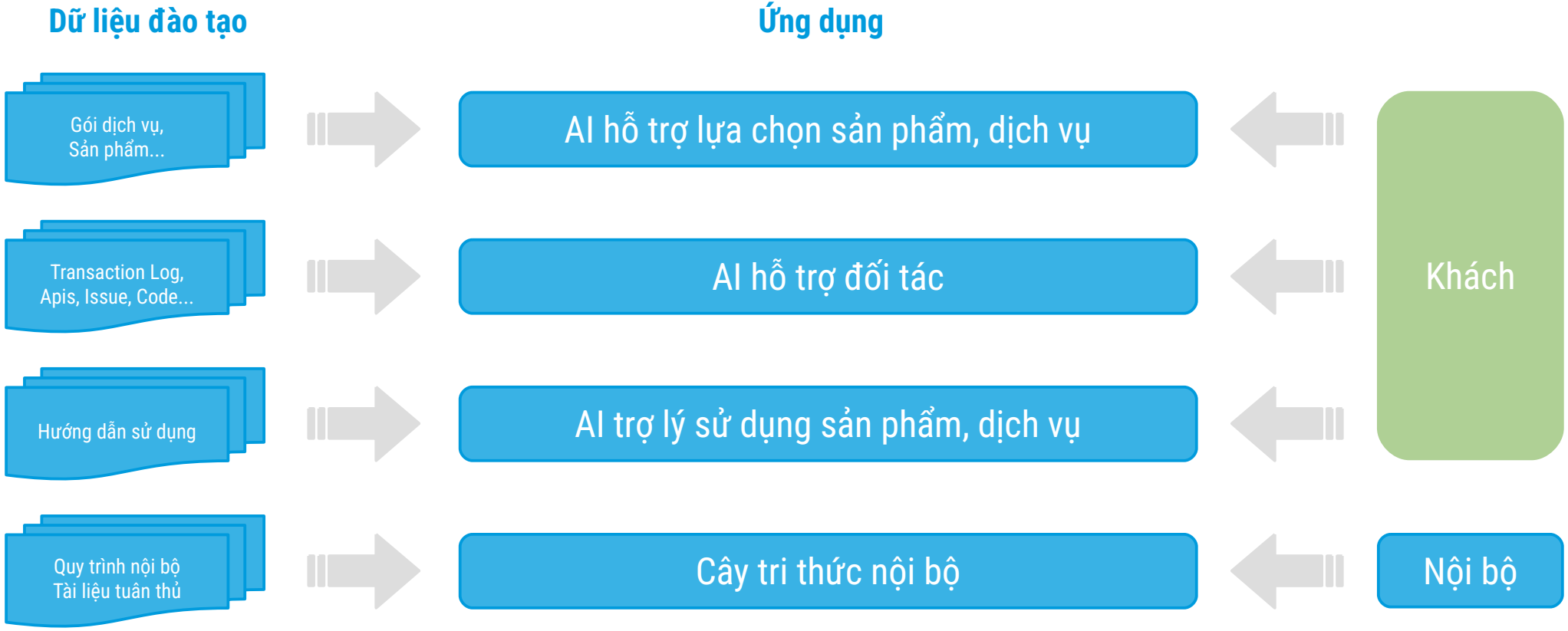
1. Hiểu câu hỏi bằng ngôn ngữ tự nhiên
2. Chủ động hỏi để thu thập thông tin
3. Tìm kiếm kiến thức nhanh trên DB, VectorDB hay KGraph
4. Sử dụng mô hình ngôn ngữ phù hợp với mục tiêu
5. Dữ liệu lưu trữ riêng
6. Đáp ứng nhiều cách thức tổ chức dữ liệu khác nhau
7. Trả lời bằng ngôn ngữ tự nhiên, dễ hiểu
8. Có khả năng tích hợp vào nhiều ứng dụng khác nhau

GenAI trong Nhà máy sản xuất

GenAI trong Giáo dục

GenAI trong Ngân hàng

GenAI trong Thương mại, dịch vụ, mkt



GenAI trong Nhà nước

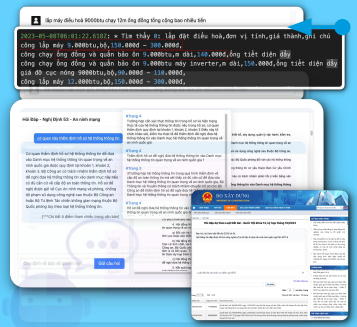
Hiệu quả khi ứng dụng GenAI

Tăng năng suất đến 50% của tổ chức

Hiểu câu hỏi, ngữ cảnh bằng ngôn ngữ tự nhiên

Không cảm xúc, luôn lịch sự

Hoạt động 24/7/365



Chấp nhận lỗi chính tả

Cung cấp chính xác câu trả lời từ dữ liệu

Suy đoán logic trên dữ liệu

Trả lời liên tục những câu hỏi nhàm chán

Giúp tổ chức tối ưu nguồn lực nội bộ

Tăng thời gian phải hồi khách hàng

Khẳng định thương hiệu của tổ chức

Tăng doanh thu, mở thêm nhiều cơ hội mới

Cân nhắc phương án

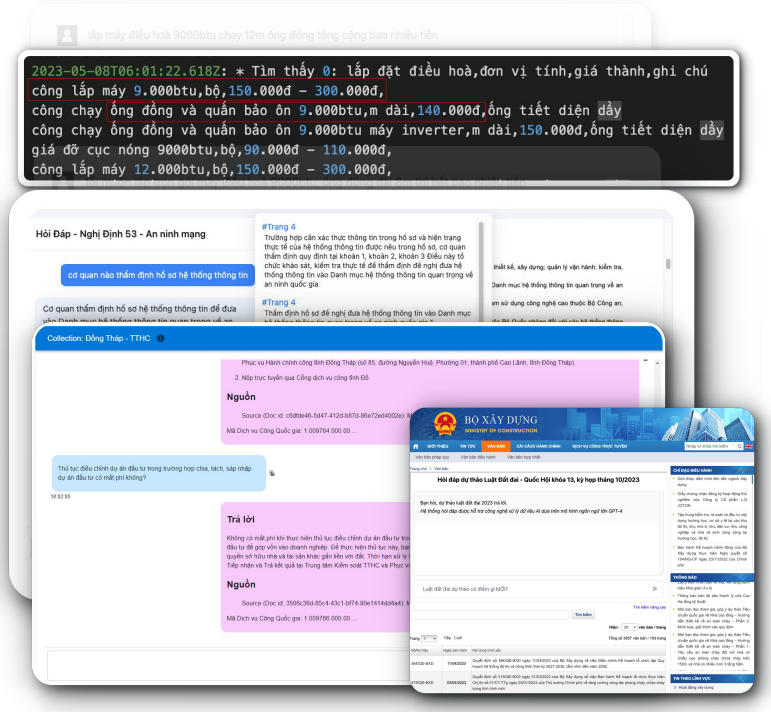
Giá thành - Độ lớn - Thời gian - Tính riêng rẽ

Build model | Fine tuning | RAG

Tham khảo với huấn luyện mô hình gốc Llama 2:

- Năng lượng điện tiêu thụ: 13.244.246 KWh
- Phát thải carbon: 539 tấn
- Phần cứng: 1 cụm siêu máy tính phục vụ nghiên cứu của Meta *(không có thông số)* và 1 cụm máy tính hội tụ dùng riêng. Cụm máy tính dùng riêng dùng Card đồ hoạ A100 của nVidia với khoảng 2000 GPU *(giá GPU A100 khoảng 20-30k $US/cái)*.
- Nhân sự: Không có công bố chính thức tuy nhiên Meta sử dụng rất nhiều team khác nhau và trong đó có team xác định rủi ro với quy mô là 350 người.

Các tổ chức đã triển khai



- Tổ chức giáo dụng trực tuyến Funix: https://funix.edu.vn
- Công ty CP bóng đèn phích nước Rạng Đông: https://rangdongstore.vn
- Hệ thống Dauthau.Info: https://dauthau.asia
- Công ty CP Rada: https://apprada.vn
- Công ty HomeID: https://homeid.asia
- Công ty Vijases: https://globicare.com
- Nhà máy bình phích thuỷ tinh Rạng Đông: Nội bộ
- Mẫu Chat Ngân hàng: http://115.146.127.40:9999/vpbank
- Mẫu Chat tuyển sinh: http://115.146.127.40:9999

Giá thành & Thời gian

Phân tích dữ liệu	Xử lý	Mã hoá	Finetune	Ứng dụng	Tích hợp	Vận hành
		2 tuần - 3 tháng				
Đọc, hiểu, lên phương án tổ chức mô hình dữ liệu đáp ứng yêu cầu thực tế	Làm sạch dữ liệu, chuẩn hóa dữ liệu, phân hoạch, đánh nhãn và băm dữ liệu thành các đơn vị lưu trữ	Vector hoá các đơn vị dữ liệu, lưu trữ lên mạng neuron nội bộ	Kiểm tra, hiệu chỉnh, tối ưu hóa phản hồi, tiến hành xử lý và vector lại dữ liệu đến mức độ tối ưu theo yêu cầu	Hiệu chỉnh ứng dụng, phân tích & nhận dạng câu hỏi, căn chỉnh kết quả search vector và phản hồi người dùng	Tích hợp vào các nền tảng đích (app, web hoặc các ứng dụng đặc thù)	Logs tracking
Tổ chức nhà nước **300đ/15 ký tự hoặc tùy thuộc độ phức tạp/vênh dữ liệu với mô hình**		**APIs pay**		**355tr-3tỉ**	**Theo thực tế**	**APIs pay & host**
Doanh nghiệp *(dữ liệu thuộc nhóm MyGPT đã trained)*		5-30 ngày		**90-200tr**	**Embedding code**	**APIs pay & host**

Demo - Q&A

- Công ty & Sản phẩm
- Dataset, Cây tri thức
- Văn bản pháp quy
- Thủ tục hành chính