

# Dự án Chatbot AI

7/2/2025:

**Giai đoạn 1:** Tìm hiểu tổng quan, trải nghiệm thử các chatbot AI thương mại, xác định rõ ưu và nhược điểm.

- Tổng quan về chatbot và phân loại.
- Mô hình tích hợp chatbot
- Các chatbot thương mại hiện nay. (đang update)
- Thêm bảng xếp hạng chatbot AI LLM cập nhật liên tục

**Giao đoạn 2:** Dựng LLM + RAG thử nghiệm

- Sử dụng model sẵn có (ko cần training lại, rất tốn thời gian).
- Ưu tiên dùng model của LLMA3.2 (đã có bản 3.3 nhưng 70 tỉ tham số -> nặng)

<https://www.llama.com/>

- Model LLMA 3.3 , hướng dẫn:
- <https://blog.dailydoseofds.com/p/building-a-rag-app-using-llama-33?ref=dailydev>

1 vài key mới: DeepSeek, AI Agent

Dự án Chatbot AI	1
1. Chatbot và phân loại.	2
a. Định nghĩa:	2
b. Các trường hợp sử dụng	2
c. Phân loại chatbot	2
2. Các tổ chức xây dựng mô hình AI thế nào ?	3
3. LLM là gì ?	4
4. Bảng xếp hạng chatAI LLM	5
5. Các sản phẩm chatbot AI đã thương mại hoá	6
6. Các LLM mã nguồn mở.	8
7. LLM + RAG	9
8. Ollama vs vLLM	9
9. Fine-tuning vs RAG	10

## 1. Chatbot và phân loại.

### a. Định nghĩa:

**Chatbot** là chương trình hoặc ứng dụng có thể trò chuyện với người dùng bằng giọng nói hoặc văn bản.

**Chatbot truyền thống** sử dụng *quy tắc định sẵn* để trò chuyện với người dùng và cung cấp câu trả lời theo kịch bản.

**Chatbot hiện nay** xử lý ngôn ngữ tự nhiên (NLP) để hiểu người dùng và có thể trả lời các câu hỏi phức tạp với độ chính xác cao, sử dụng mô hình ngôn ngữ lớn như LLM.

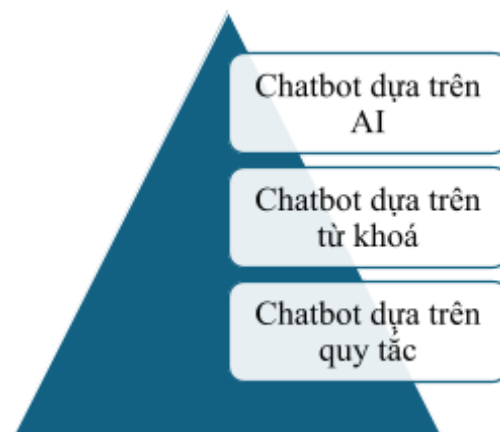
### b. Ứng dụng:

Các tổ chức trong các ngành nghề sử dụng chatbot để hợp lý hóa trải nghiệm của khách hàng, tăng hiệu quả vận hành và giảm chi phí.

- **Nâng cao năng suất doanh nghiệp:** truy xuất tài liệu, tối ưu hoá các quy trình-nghệ vụ lặp lại.
- **Trợ lý cá nhân:** lịch, ghi chú, hỏi đáp với dữ liệu được cho phép bởi người dùng.
- **Trung tâm cuộc gọi:** trả lời tự động với tin nhắn dạng chữ, hội thoại.

### c. Phân loại chatbot:

Chatbot có thể chia thành 3 loại dựa theo các đặc điểm:



STT	Loại chatbot	Đặc điểm
1	Chatbot dựa trên quy tắc	Đơn giản, người dùng chỉ hỏi hoặc chọn các câu hỏi cụ thể, trả lời là giống nhau cho tất cả người dùng. Chatbot có một từ điển tích hợp sẵn để ánh xạ câu trả lời cụ thể cho mọi câu hỏi. (dạng như từ điển).
2	Chatbot dựa trên từ khoá	Trích xuất từ khoá và đưa ra câu trả lời tương ứng.

		VD: "Làm cách nào để kích hoạt tài khoản của tôi?", chatbot phát hiện “kích hoạt tài khoản của tôi” là từ khóa và phản hồi bằng hướng dẫn từng bước.
3	<b>Chatbot dựa trên AI</b> 😊	Hiểu và sinh được ngôn ngữ tự nhiên. <b>Từ khoá:</b> NLP, NLU, NLG, LLM.

## 2. Các tổ chức xây dựng mô hình AI thế nào ?

**Cách 1: Tự xây dựng mô hình:** => tốn kém chi phí và thời gian.

**Cách 2: Tuỳ chỉnh LLM hiện có** (như LLaMA...)

Chatbot có được kiến thức cơ sở và **kiến thức nội bộ** (dữ liệu riêng của doanh nghiệp).

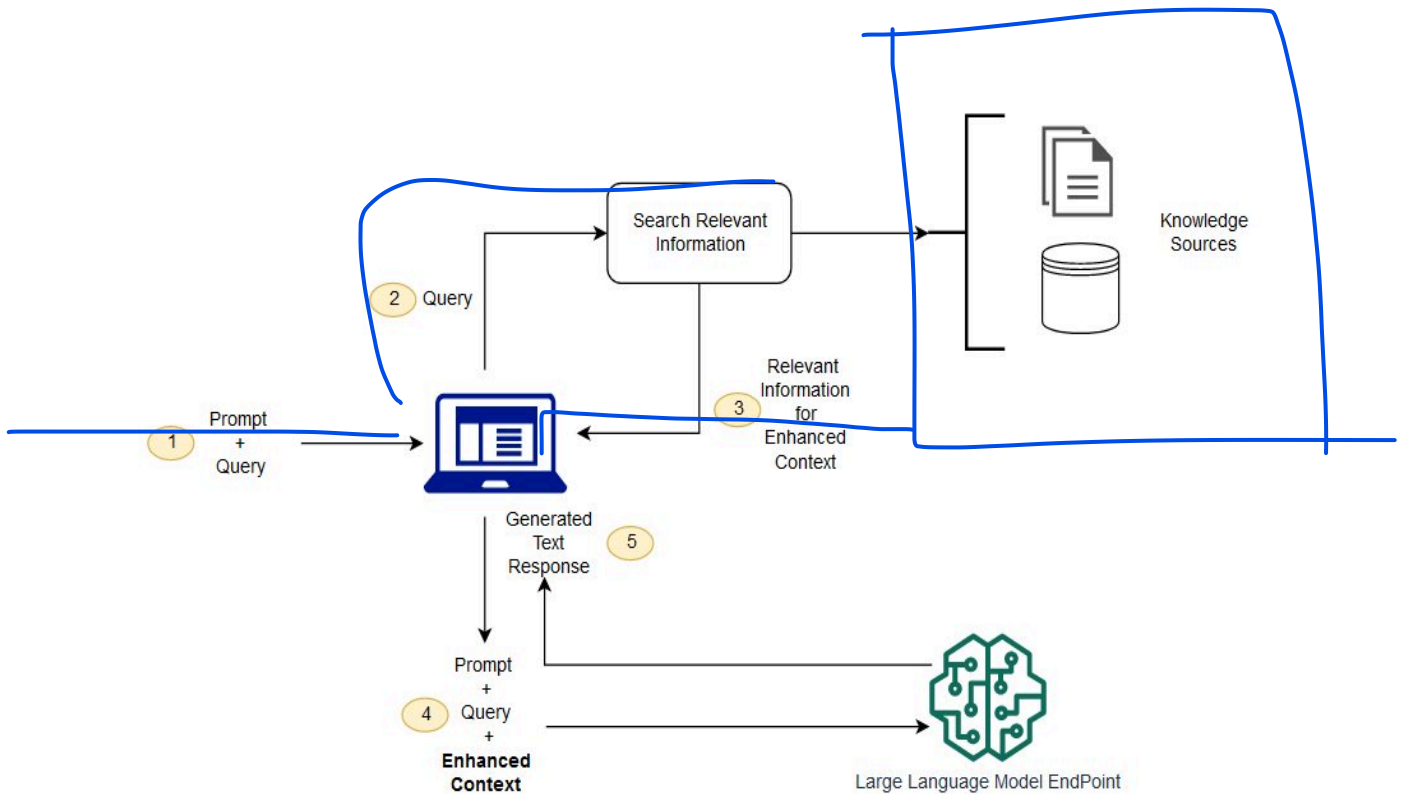
Để chatbot có thể sử dụng kiến thức nội bộ có thể sử dụng bằng **fine-tuning** hoặc sử dụng **RAG**, mỗi cách tiếp cận sẽ có ưu và nhược điểm riêng.

Cần tìm hiểu về các cách khác nhau đó, mới rút ra dc ưu nhược và chọn cách tiếp cận phù hợp.

Theo AWS: Tao có kết hợp truy xuất thông tin ngoài (RAG) là kỹ thuật chính để nâng cao LLM.

RAG mở rộng các khả năng vốn đã mạnh mẽ của **LLM sang các miền cụ thể hoặc cơ sở kiến thức nội bộ của tổ chức mà không cần phải đào tạo lại mô hình**. Đây là một cách tiếp cận hiệu quả về mặt chi phí để cải thiện đầu ra LLM để nó vẫn phù hợp, chính xác và hữu ích trong nhiều bối cảnh khác nhau.

**Mô hình khi sử dụng RAG:**



### 3. LLM là gì ?

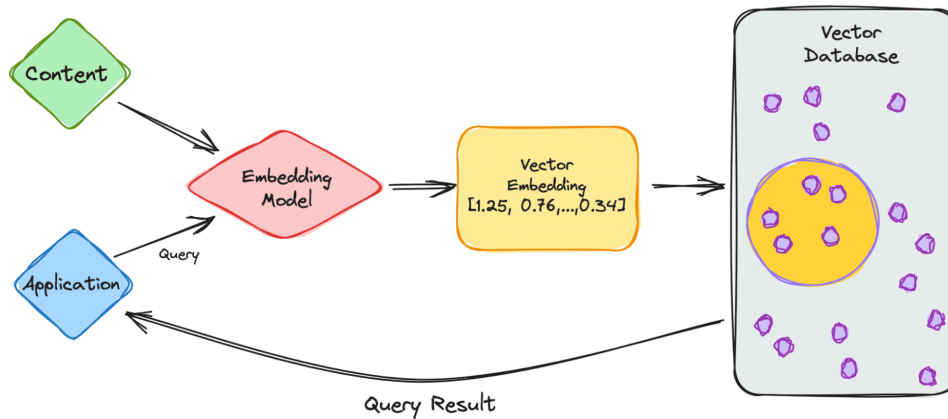
#### a. Định nghĩa:

LLM là viết tắt của **Large Language Model** có nghĩa là mô hình ngôn ngữ lớn, là một mô hình DeepLearning (học sâu) được đào tạo với lượng dữ liệu khổng lồ.

Dữ liệu có thể từ hàng tỉ trang web, từ điển wikipedia...ví dụ như [Common Crawl](#) với hơn 50 tỷ trang web, và Wikipedia với khoảng 57 triệu trang...

#### b. LLM hoạt động thế nào ?

Một yếu tố quan trọng trong cách thức hoạt động của LLM là cách chúng biểu diễn các từ. Các hình thức **máy học** trước đây sử dụng một bảng số để biểu diễn từng từ. Tuy nhiên, hình thức biểu diễn này không thể nhận ra mối quan hệ giữa các từ, chẳng hạn như các từ có nghĩa tương tự. Hạn chế này đã được khắc phục bằng cách **sử dụng các véc-tơ đa chiều**, thường được gọi là nhúng từ, để biểu diễn các từ sao cho các từ có nghĩa theo ngữ cảnh tương tự nhau hoặc các mối quan hệ khác sẽ gần nhau trong không gian véc-tơ.



Tìm hiểu thêm: <https://www.analyticsvidhya.com/articles/llm-vs-agents/>

### c. LLM được đào tạo thế nào ?

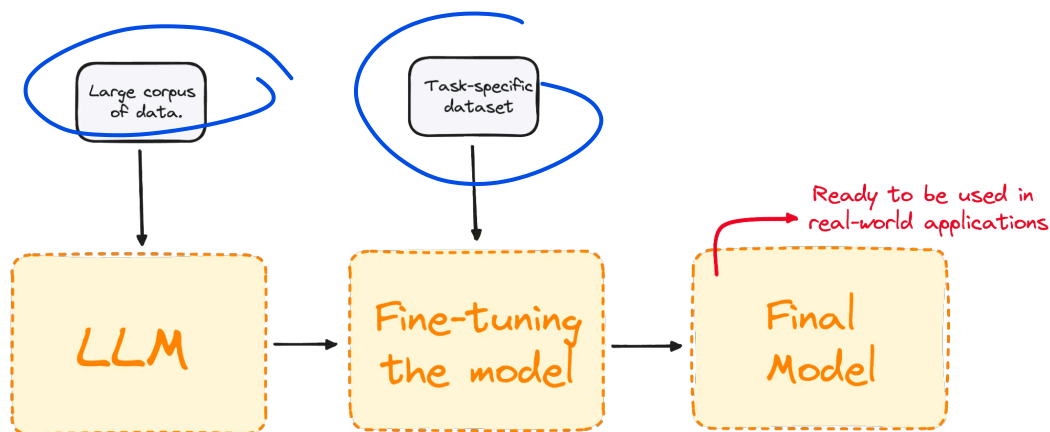
Quá trình đào tạo sử dụng *nguồn dữ liệu chất lượng cao*, liên tục điều chỉnh các tham số để có kết quả tốt.

Sau khi đào tạo LLM có thể được tinh chỉnh bằng tập dữ liệu có giám sát (nhỏ).

- **Zero-shot learning:** Học bằng dữ liệu chưa từng gặp; LLM cơ sở có thể phản hồi một loạt các yêu cầu mà không cần đào tạo rõ ràng, thường thông qua câu lệnh văn bản, mặc dù độ chính xác của câu trả lời có thể khác nhau.
- **Few-shot learning:** Học với ít dữ liệu đào tạo: Bằng cách cung cấp một số mẫu đào tạo liên quan, hiệu năng của mô hình cơ sở cải thiện đáng kể trong lĩnh vực cụ thể đó.
- **Fine-tuning:** Đây là một phần mở rộng của mô hình học với ít dữ liệu đào tạo, trong đó các nhà khoa học dữ liệu đào tạo một mô hình cơ sở để điều chỉnh các tham số của nó với dữ liệu bổ sung liên quan đến ứng dụng cụ thể.

Tham khảo AWS: <https://aws.amazon.com/vi/what-is/large-language-model/>

### Fine-Tuning LLM thế nào ?



Các loại fine-tuning:

- Supervised fine-tuning
- Few-shot learning
- Transfer learning
- Domain-specific fine-tuning

**Todo:**

<https://www.datacamp.com/tutorial/fine-tuning-large-language-models>

#### 4. Bảng xếp hạng chatAI LLM

Link: <https://lmarena.ai/?leaderboard>

Tính đến 17/1/2025:

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	<a href="#">Gemini-Exp-1206</a>	1374	+4/-5	20227	Google	Proprietary
1	1	<a href="#">ChatGPT-4o-latest (2024-11-20)</a>	1365	+4/-3	33383	OpenAI	Proprietary
1	4	<a href="#">Gemini-2.0-Flash-Thinking-Exp-1219</a>	1364	+5/-6	15728	Google	Proprietary
2	4	<a href="#">Gemini-2.0-Flash-Exp</a>	1357	+6/-4	19030	Google	Proprietary
3	1	<a href="#">o1-2024-12-17</a>	1351	+7/-7	7289	OpenAI	Proprietary
6	4	<a href="#">o1-preview</a>	1335	+4/-4	33194	OpenAI	Proprietary
7	7	<a href="#">DeepSeek-V3</a>	1319	+6/-6	10510	DeepSeek	DeepSeek
7	10	<a href="#">Step-2-16K-Exp</a>	1305	+8/-9	3374	StepFun	Proprietary
8	11	<a href="#">o1-mini</a>	1306	+3/-3	47977	OpenAI	Proprietary
8	8	<a href="#">Gemini-1.5-Pro-002</a>	1303	+4/-4	44485	Google	Proprietary
11	13	<a href="#">Grok-2-08-13</a>	1288	+3/-3	65752	xAI	Proprietary
11	14	<a href="#">Yi-Lightning</a>	1287	+4/-5	28961	01 AI	Proprietary
11	9	<a href="#">GPT-4o-2024-05-13</a>	1285	+3/-2	117751	OpenAI	Proprietary
11	7	<a href="#">Claude 3.5 Sonnet (20241022)</a>	1284	+3/-3	46758	Anthropic	Proprietary
11	21	<a href="#">Qwen2.5-plus-1127</a>	1282	+6/-6	7009	Alibaba	Proprietary
11	18	<a href="#">Deepseek-v2.5-1210</a>	1279	+7/-6	7259	DeepSeek	DeepSeek
14	21	<a href="#">Athena-v2-Chat-72B</a>	1277	+4/-4	20322	NexusFlow	NexusFlow
15	20	<a href="#">GLM-4-Plus</a>	1274	+4/-3	27774	Zhipu AI	Proprietary
15	21	<a href="#">GPT-4o-mini-2024-07-18</a>	1273	+4/-3	60183	OpenAI	Proprietary
15	35	<a href="#">Llama-3.1-Nemotron-70B-Instruct</a>	1269	+8/-6	7596	Nvidia	Llama 3.1

## 5. Các sản phẩm chatbot AI đã thương mại hoá

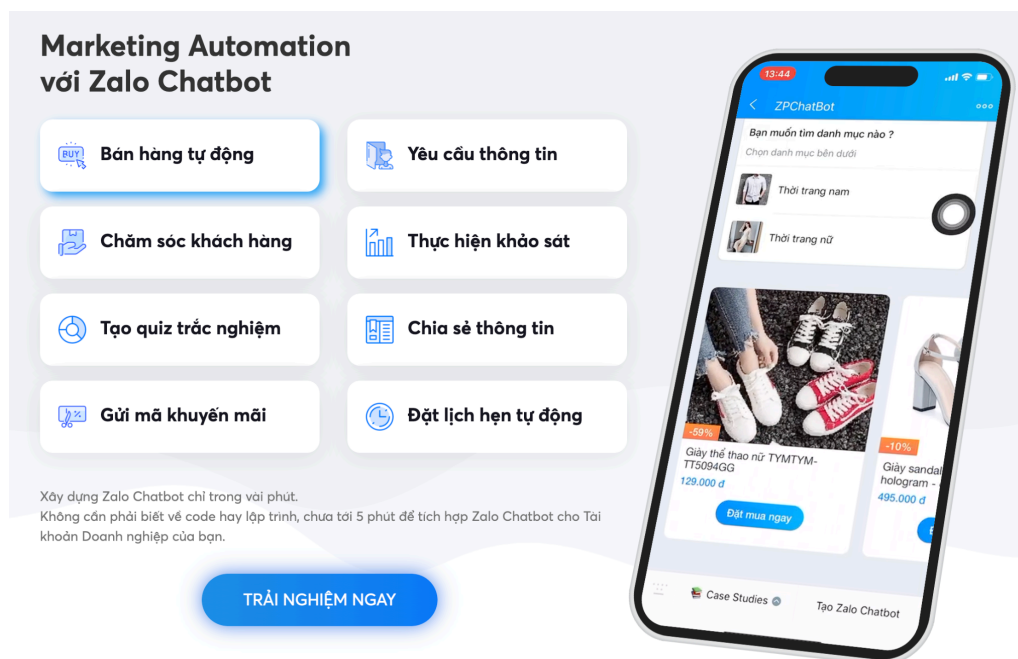
- Thực hiện các khảo sát về chi phí, giá, ưu nhược điểm.
- Cách các giải pháp tích hợp của họ vào hệ thống hiện tại tại doanh nghiệp.
- Bao gồm trong nước và ngoài nước

1

### Zalo Chatbot

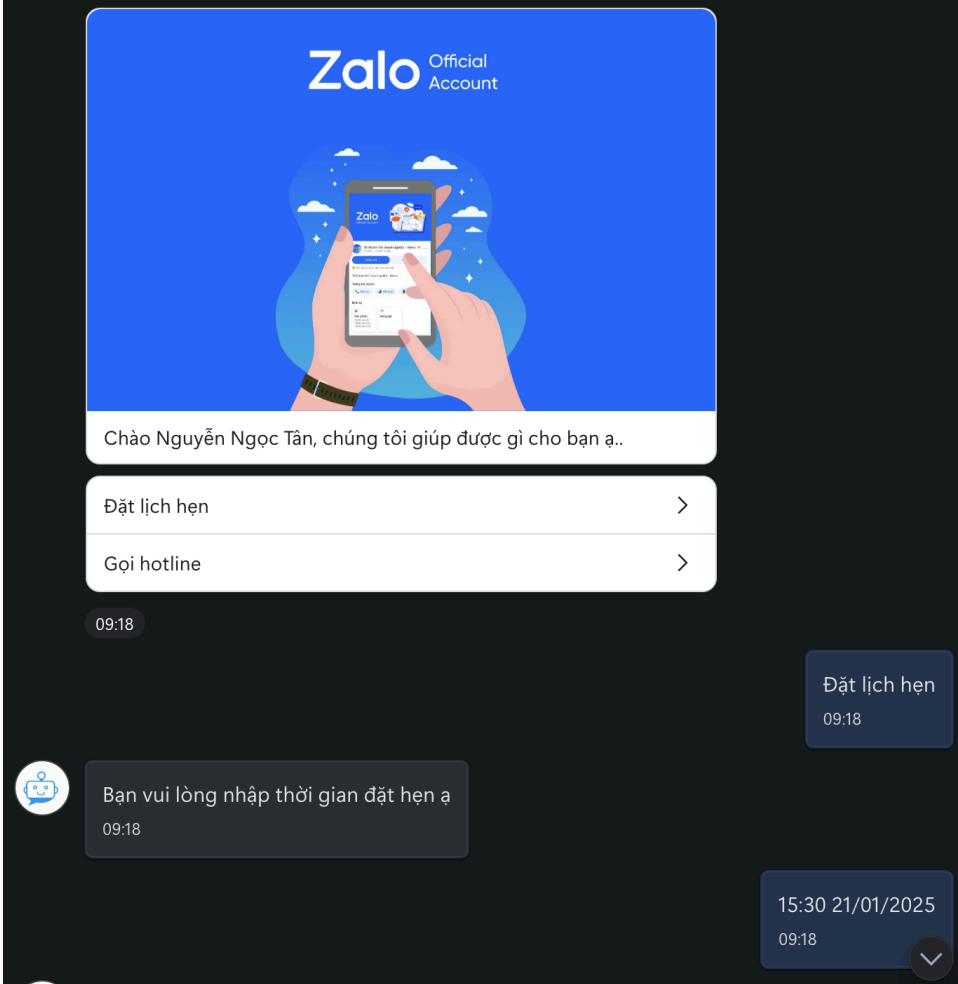
Link: <https://chatbot.zalo.me/>

Có tính phí, hiện tại đang dùng luôn nền tảng zalo, chưa xem kỹ có cách tích hợp khác hay không.



### Đánh giá:

Có thể xây dựng được kịch bản tự động, tự đưa ra gợi ý tiếp theo. Đánh giá là chưa tốt hơi giống **chatbot loại 1, loại 2** vì người dùng vẫn phải chọn menu đúng theo từ khoá.

	 <p>Người bán hàng phải xây dựng nhiều kịch bản khác nhau =&gt; chưa tốt.</p>

**Todo...khảo thêm các nguồn bên ngoài.**

## **6. Các LLM mã nguồn mở.**

Tham khảo:

<https://github.com/eugeneyan/open-llms>

<https://github.com/Hannibal046/Awesome-LLM>

**Cần thử nghiệm và lựa chọn model phù hợp.**

**Ưu tiên dùng model dc dc phép dùng cho thương mại.**



## 7. LLM + RAG

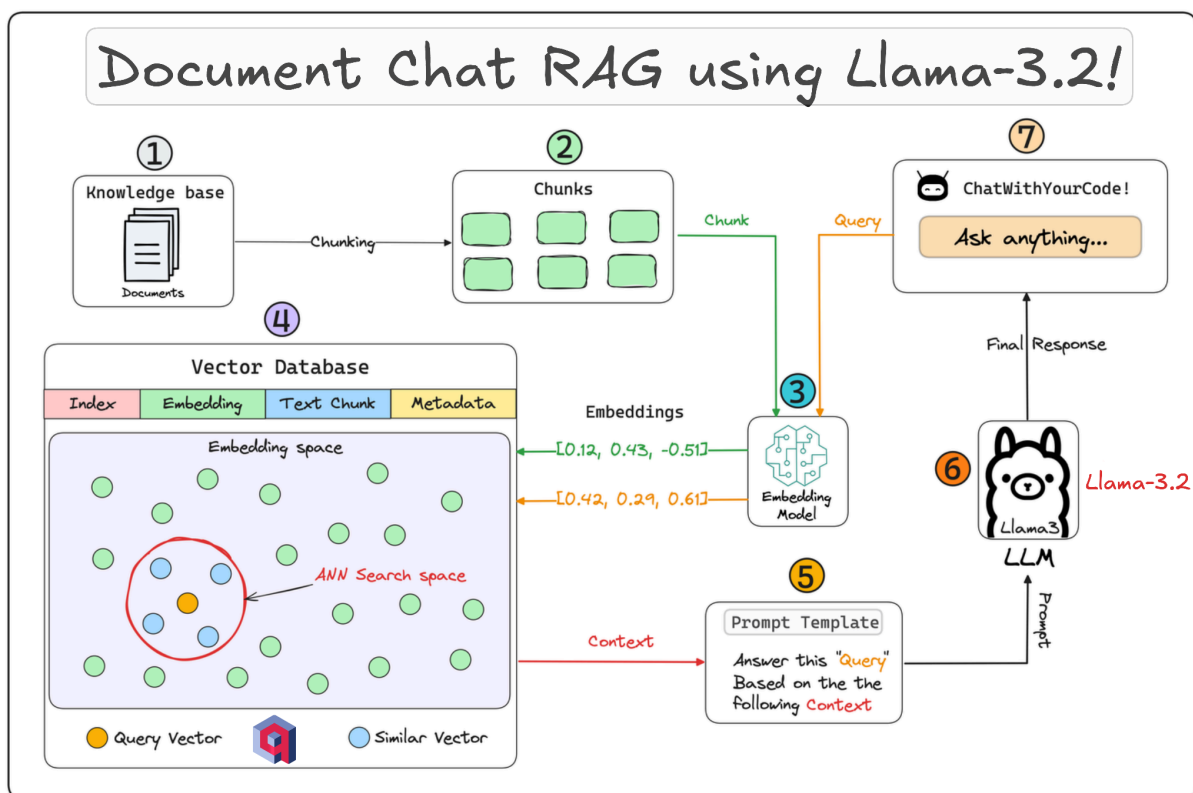
Tham khảo thêm:

<https://viblo.asia/p/chatgpt-series-5-tim-hieu-ve-retrieval-augmented-generation-rag-Ny0VG Rd7LPA>

Todo:

<https://lightning.ai/lightning-ai/studios/rag-using-llama-3-2-by-meta-ai?section=featured>

📺 Full RAG using Open-Source [Ollama + Llama 3.2] 🚀



### Các bước thực hiện:

Bước 1: Cài đặt Ollma

Bước 2:

Bước 3:

## 8. Ollama vs vLLM

**Ollama và vLLM về cơ bản cung cấp môi trường để chạy các mô hình ngôn ngữ lớn.**

**Đối tượng sử dụng:** Ollama hướng đến các nhà phát triển cá nhân và nhóm nhỏ, trong khi vLLM phục vụ cho các tổ chức lớn với quy mô phức tạp hơn.

**Hiệu suất:** vLLM thường có lợi thế hơn về hiệu suất khi xử lý các mô hình lớn trong môi trường phân tán.

**Độ phức tạp:** Ollama đơn giản hơn trong việc thiết lập và sử dụng, phù hợp cho những người mới bắt đầu hoặc cần thử nghiệm nhanh.

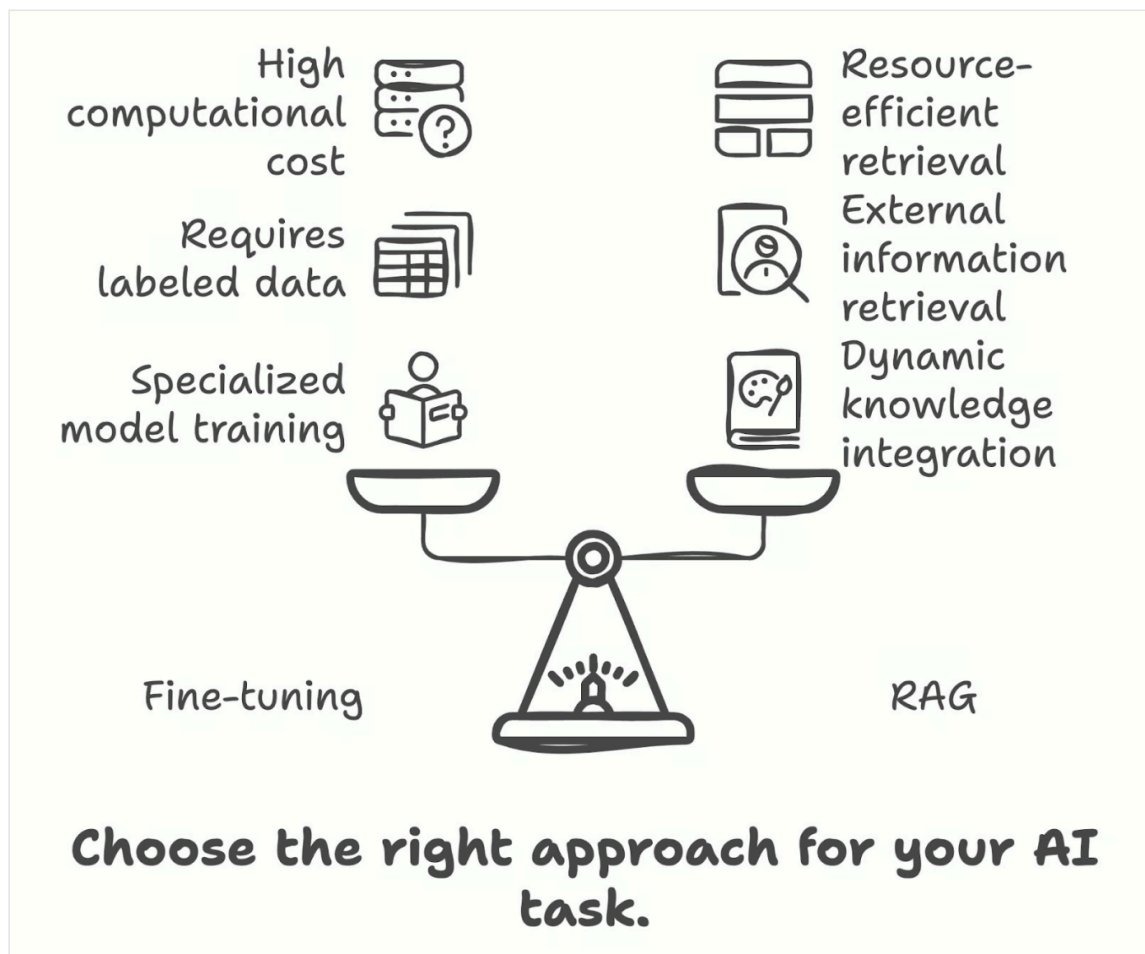
Tham khảo:

<https://ollama.com/>

<https://docs.vllm.ai/en/stable/>

<https://collabnix.com/ollama-vs-vllm-choosing-the-best-tool-for-ai-model-workflows/>

## 9. Fine-tuning vs RAG



<https://learn.microsoft.com/en-us/azure/developer/ai/augment-llm-rag-fine-tuning>