



# GIỚI THIỆU & LÀM QUEN



## Mr Lộc Đặng

- Founder nền tảng Chatbot AI Mindmaid - nền tảng Chatbot AI duy nhất vào Vòng chung khảo Nhân tài đất Việt 2023
- Founder nền tảng thu thập và phân tích dữ liệu báo chí, mạng xã hội VnAlert - Giải thưởng CĐS Việt Nam 2020
- Admin cộng đồng Build bot kiếm tiền
- Đối tác chính thức & Chuyên gia triển khai Larksuite tại Việt Nam
- Tham khảo:
  - Facebook: <https://www.facebook.com/locdh90/>
  - Báo chí:
    - Mindmaid giúp Chatbot AI phổ biến hơn
    - Các startup, trợ lý ảo AI sẽ làm gì trong 2024?



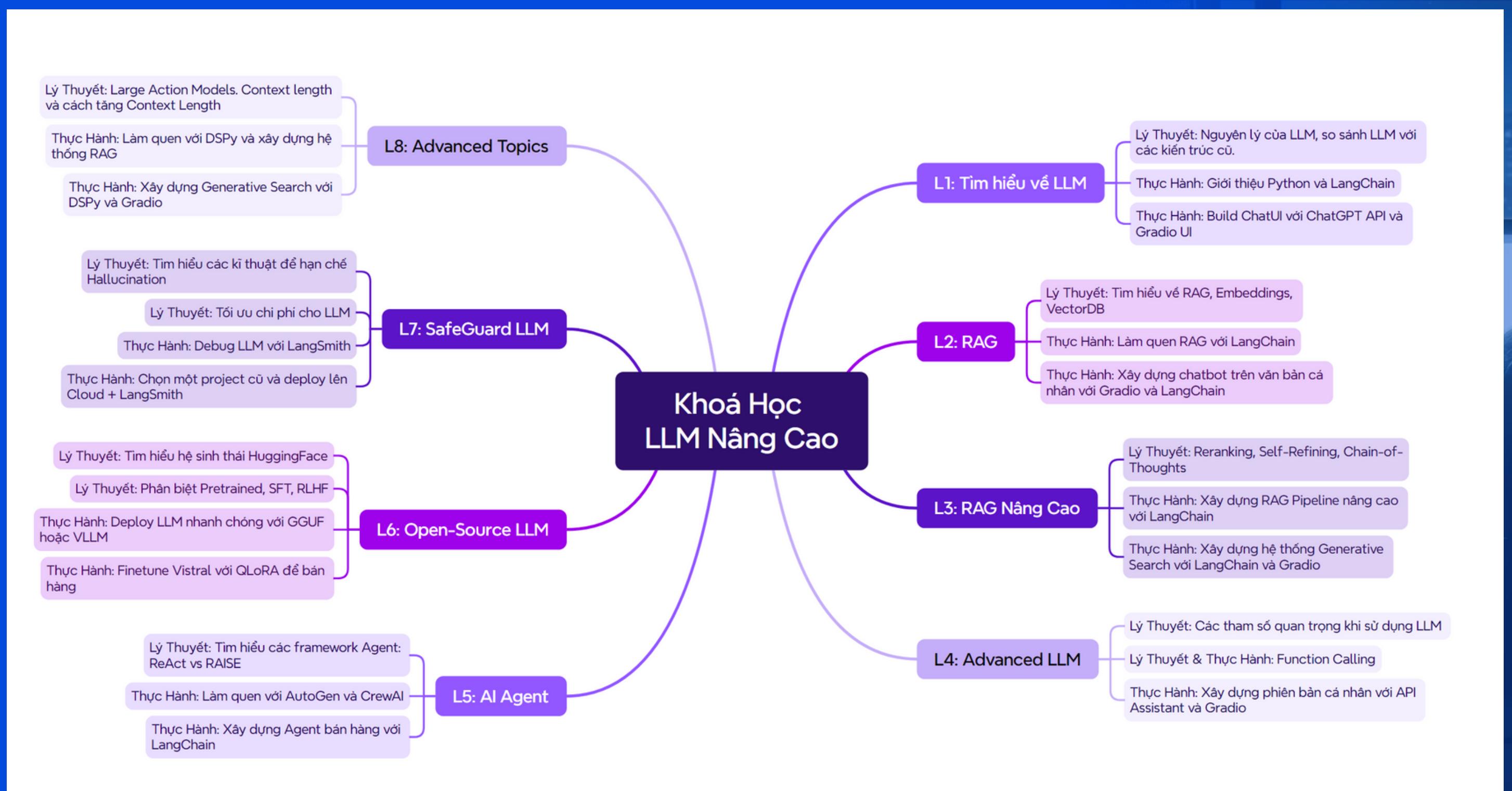
## Mr Quân Nguyễn

- Cựu kĩ sư OpenAI. Tham gia phát triển các sản phẩm như ChatGPT, Bing Chat
- Trưởng bộ phận nghiên cứu, Ontocord.ai
- Creator và tham gia huấn luyện VinaLlama và Vistral
- Tham gia vào nhiều các hoạt động Open Source trên thế giới
- Tham khảo:
  - Facebook: <https://www.facebook.com/hqmpd/>
  - Báo chí:
    - Nhóm Kỹ Sư GenZ làm AI Miễn Phí Cho Người Việt

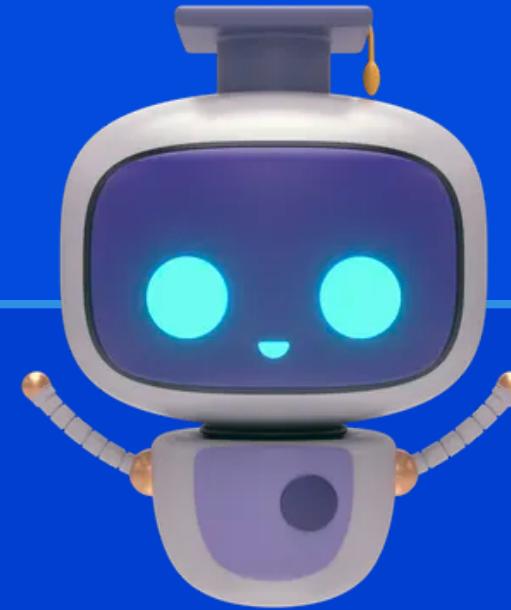


• • •  
• • •  
• • •  
• • •

# Lộ trình trở thành LLM Application Developer



Click để  
xem lớn



# Buổi 1: Tổng quan về LLM Application & phát triển Chatbot AI nâng cao

Khóa 1-2024

Giảng viên chính: Mr Quân Nguyễn  
Trưởng bộ phận nghiên cứu, Ontocord.ai

# Nội dung chính

- 1 Những cơ hội khi trở thành người LLM Application Developer chuyên nghiệp**
- 2 Lộ trình & Phương pháp học**
- 3 Thực hành: xây dựng chatbot Phiên bản số của bản thân bằng langchain & gradio**

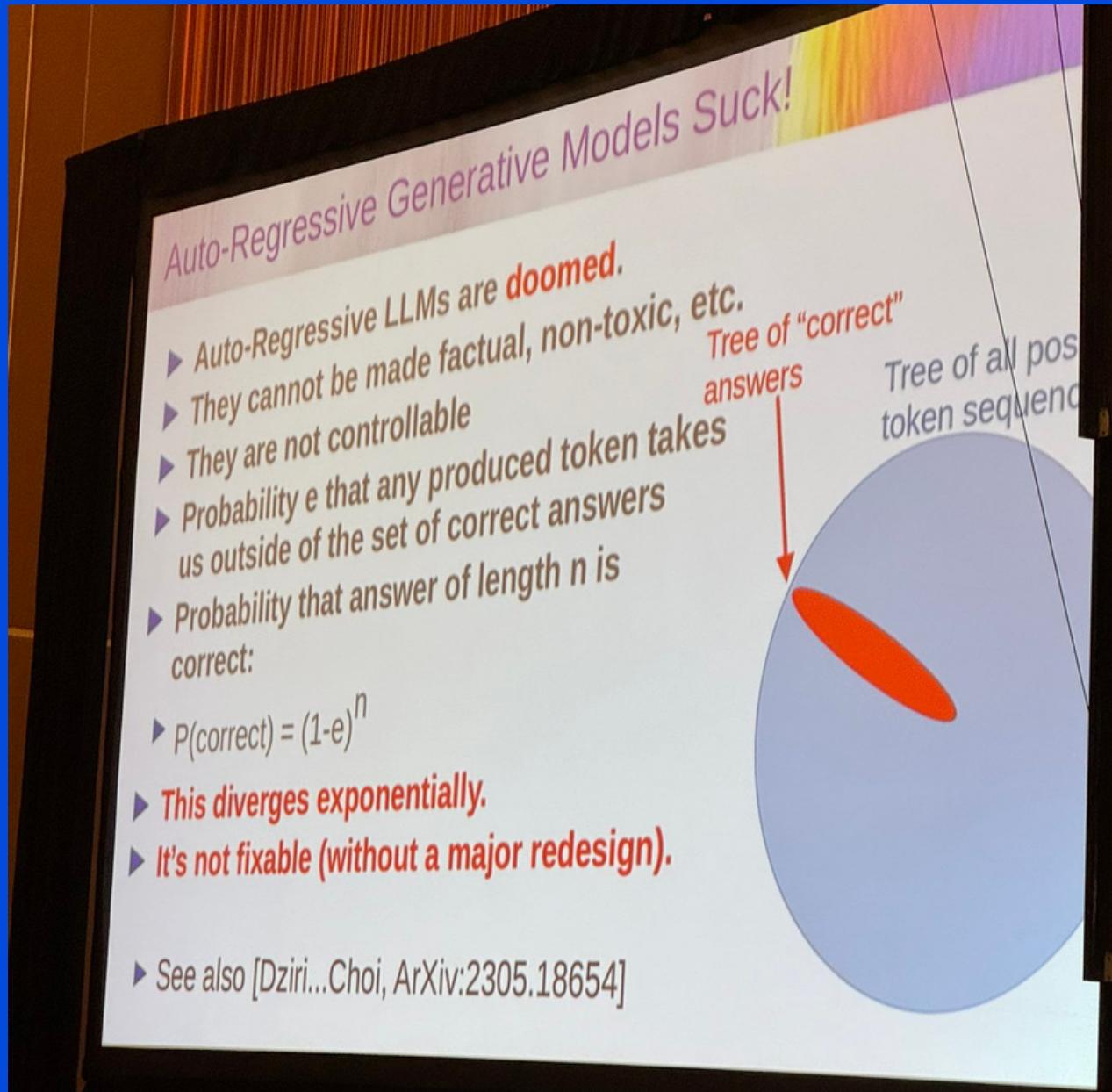




# XU HƯỚNG PHÁT TRIỂN ỨNG DỤNG LLM & BUILD CHATBOT AI



# LLM Rất Tốt, nhưng ... chúng là những cỗ máy mộng mơ



LLM ≠ Giao Diện Chat

Andrej Karpathy  
@karpathy

# On the "hallucination problem"

I always struggle a bit with I'm asked about the "hallucination problem" in LLMs. Because, in some sense, hallucination is all LLMs do. They are dream machines.

We direct their dreams with prompts. The prompts start the dream, and based on the LLM's hazy recollection of its training documents, most of the time the result goes someplace useful.

It's only when the dreams go into deemed factually incorrect territory that we label it a "hallucination". It looks like a bug, but it's just the LLM doing what it always does.

At the other end of the extreme consider a search engine. It takes the prompt and just returns one of the most similar "training documents" it has in its database, verbatim. You could say that this search engine has a "creativity problem" - it will never respond with something new. An LLM is 100% dreaming and has the hallucination problem. A search engine is 0% dreaming and has the creativity problem.

All that said, I realize that what people \*actually\* mean is they don't want an LLM Assistant (a product like ChatGPT etc.) to hallucinate. An LLM Assistant is a lot more complex system than just the LLM itself, even if one is at the heart of it. There are many ways to mitigate hallucinations in these systems - using Retrieval Augmented Generation (RAG) to more strongly anchor the dreams in real data through in-context learning is maybe the most common one. Disagreements between multiple samples, reflection, verification chains. Decoding uncertainty from activations. Tool use. All an active and very interesting areas of research.

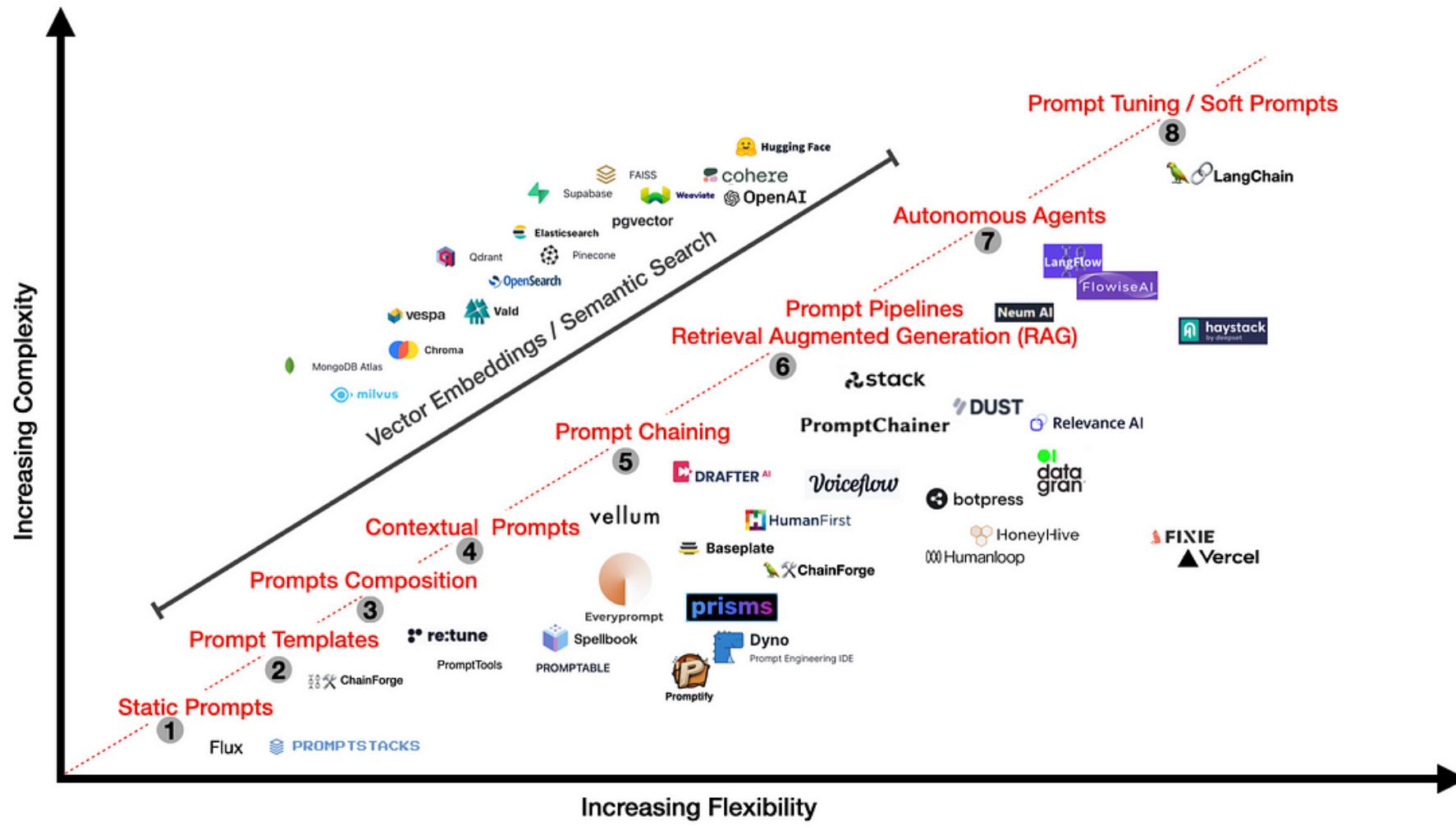
TLDL I know I'm being super pedantic but the LLM has no "hallucination problem". Hallucination is not a bug, it is LLM's greatest feature. The LLM Assistant has a hallucination problem, and we should fix it.

</rant> Okay I feel much better now :)



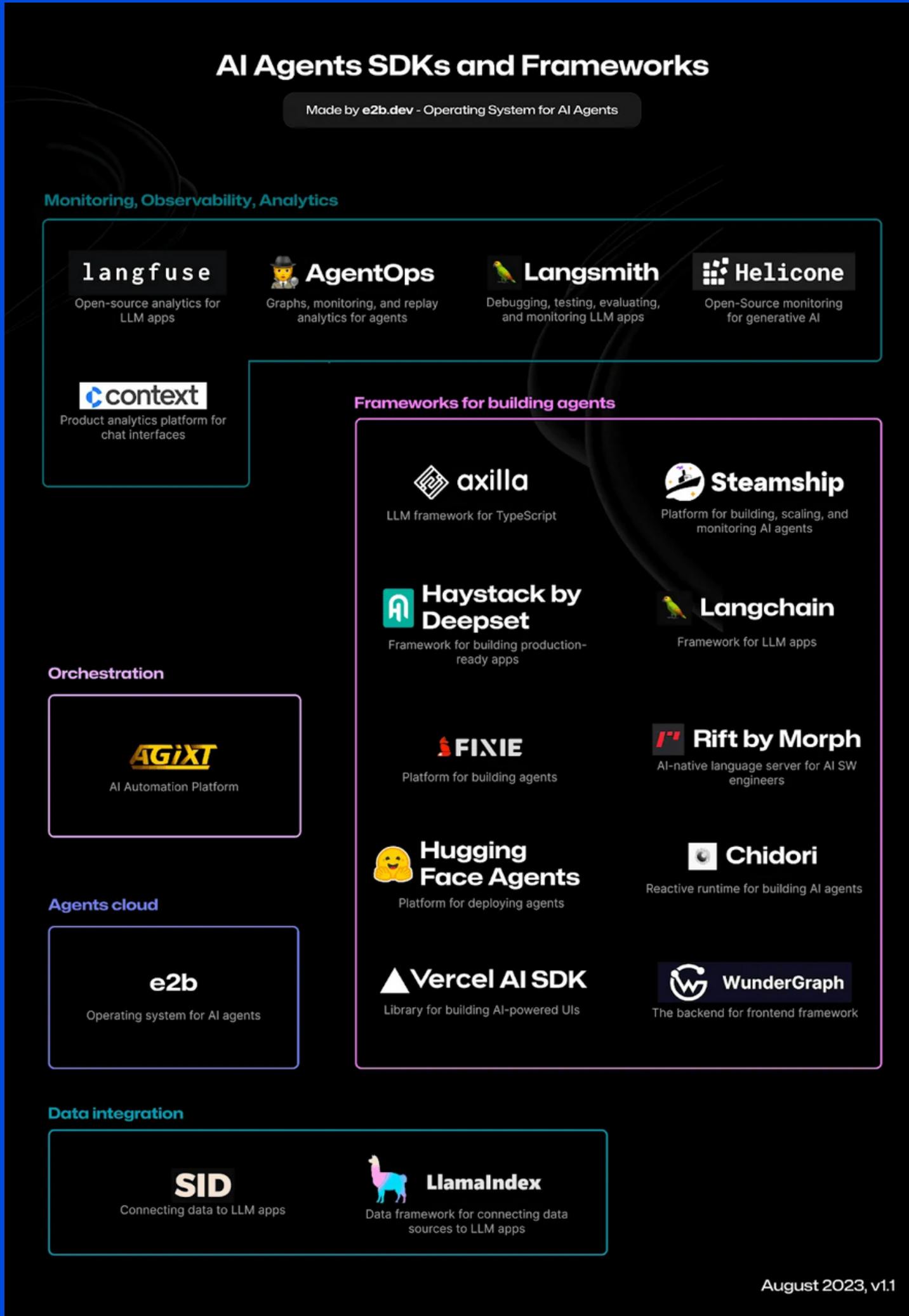
# Xây ứng dụng LLM cần nhiều sự tinh chỉnh

## Emerging LLM Application Architecture



Hiện tại, Prompting là phương pháp phổ biến nhất. Tuy nhiên để đạt được hiệu quả mong muốn, một số kỹ thuật advanced cần được áp dụng:

- **Prompt Chaining**
- **RAG**
- **Autonomous Agent**



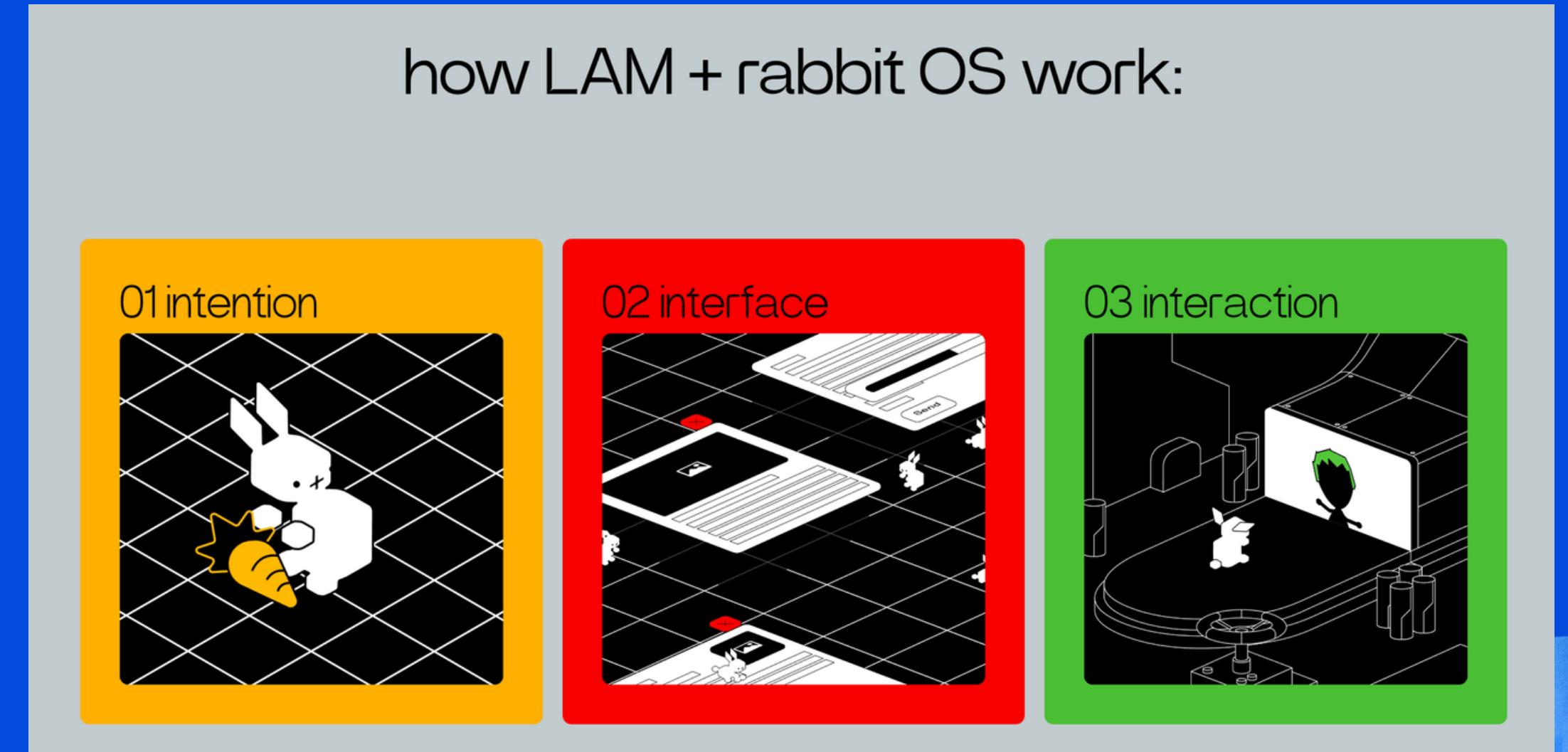
# LLM Applications & Chatbot AI đang thay thế ứng dụng kiểu cũ!!!

**Ưu điểm của LLM Application & Chatbot AI so với ứng dụng truyền thống**

- UX thân thiện hơn nhờ giao tiếp bằng ngôn ngữ, giọng nói tự nhiên
- Tự suy luận, lập kế hoạch dựa trên **Kiến thức**, thậm chí tự **viết code** thể thực hiện nhiệm vụ mà user mong muốn
- Có các **siêu năng lực** tương tự như ChatGPT, Gemini Pro...nhờ hoạt động dựa trên các công nghệ này

⇒ **Hệ sinh thái phát triển LLM Application & Chatbot AI đã phát triển cực nhanh chóng**

# Những thiết bị phần cứng trong tương lai sẽ có AI là lõi



- Rabbit R1 có thay thế được chiếc điện thoại trong túi bạn?





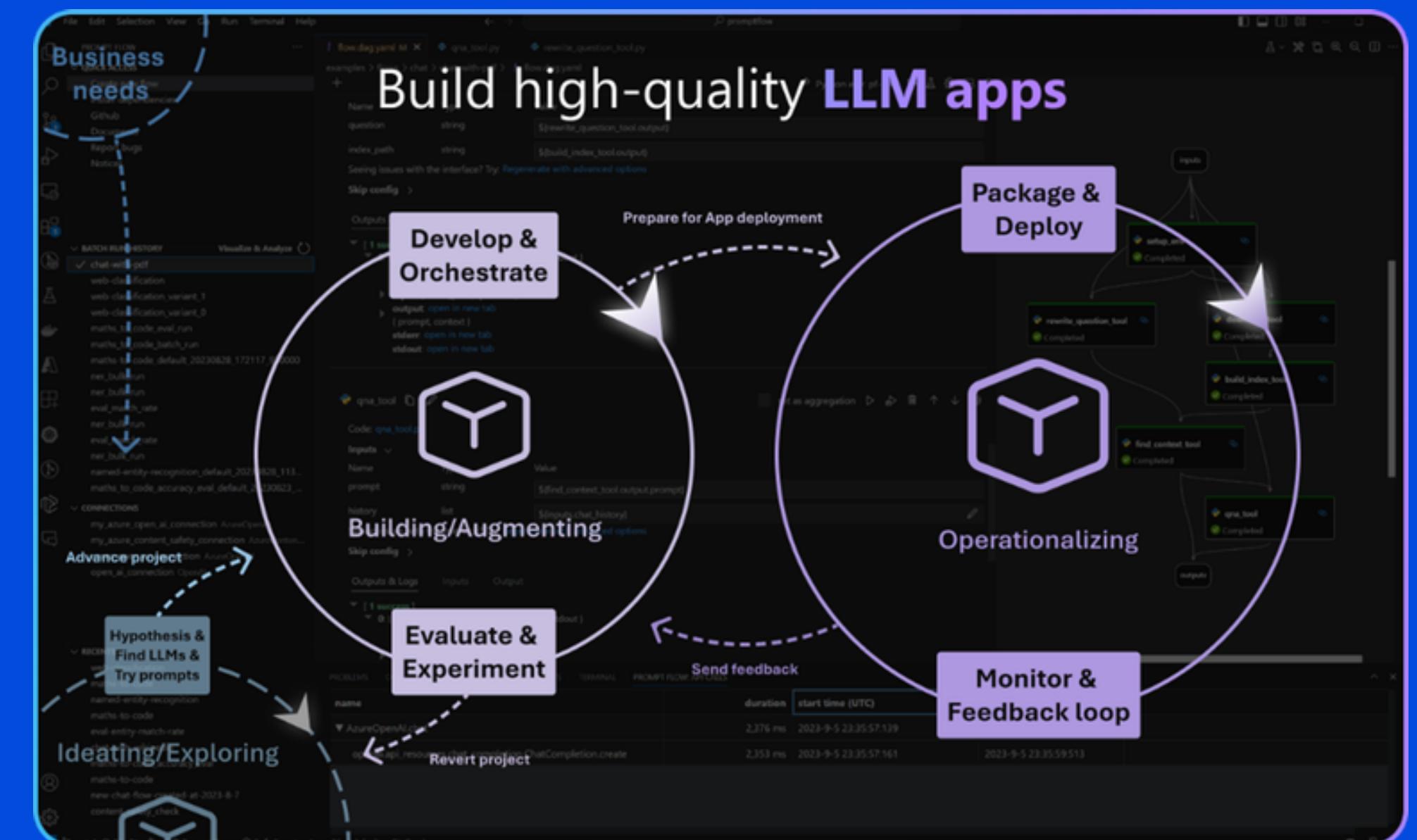
1. Năm bắt cơ hội mới trong 2024 nhờ

# XÂY DỰNG HỆ THỐNG LLM CHUYÊN NGHIỆP



# Vì sao cần LLM Developer chuyên nghiệp?

- Các hệ thống Chatbot AI no-code/low-code thường đường kèm với **rất ít tùy chỉnh (customization)**, gây ra nhiều hạn chế khi cần một luồng riêng.



- Các bài toán riêng cần **có cách giải quyết** khác nhau (vd: Hallucination trong Toán Học khác với Hallucination trong CSKH).

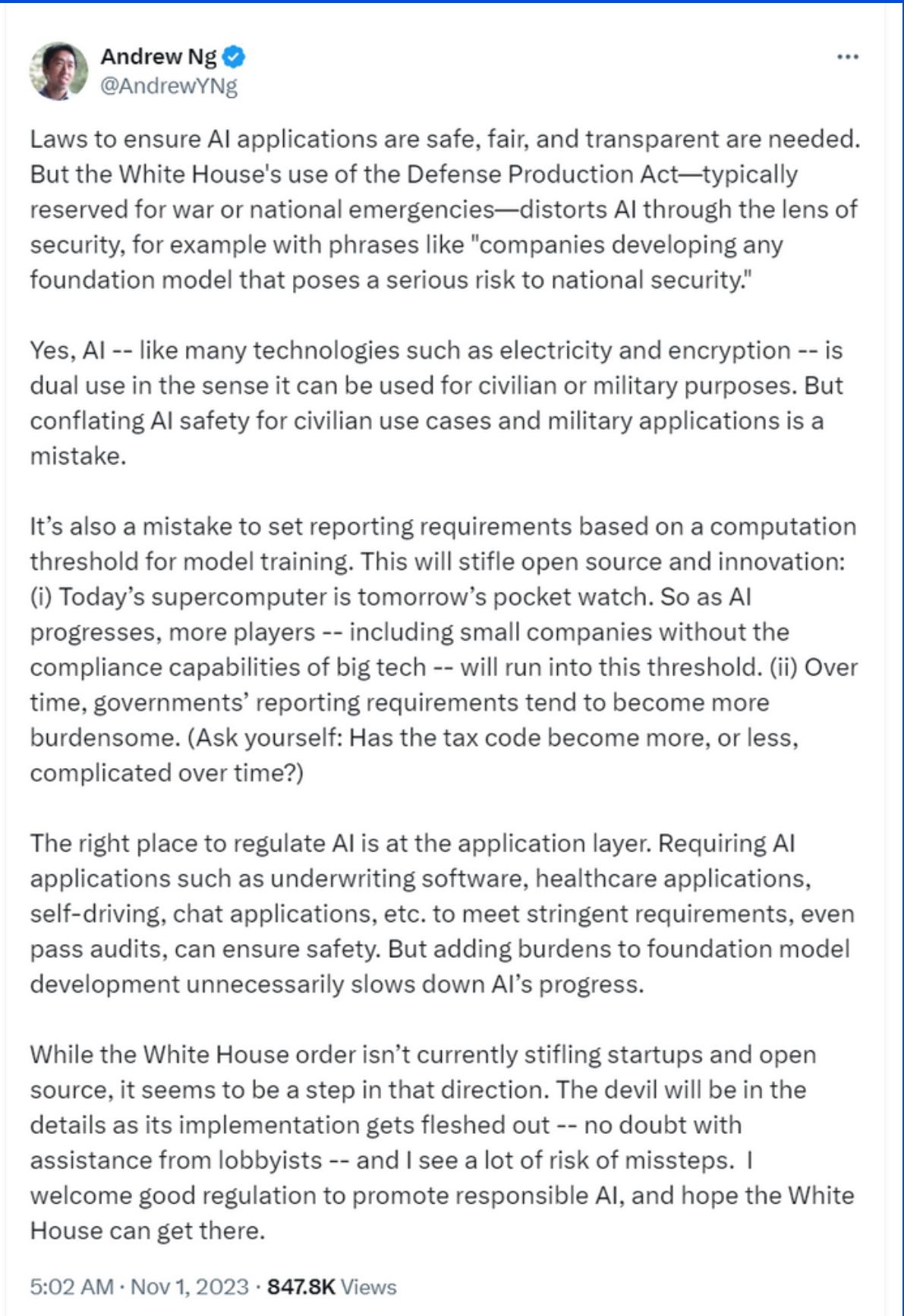
How many 'm's are in the word 'Weather'?

There is one 'm' in the word 'Weather'.

Are you sure?

Apologies for the oversight. You're right; there are no 'm's in the word 'Weather'. Thank you for pointing it out.

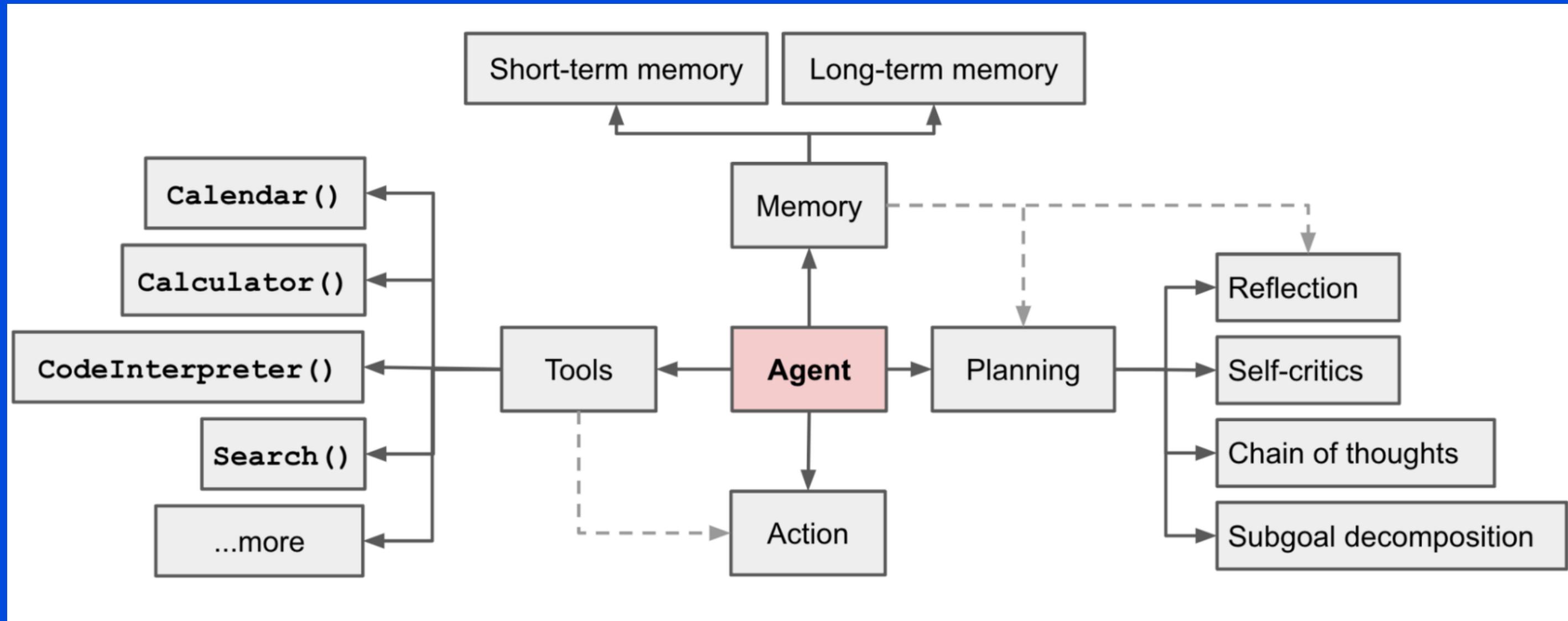
# *“Ứng dụng AI sẽ chỉ khác nhau ở tầng ứng dụng”*



Andrew Ng  @AndrewYNg ...  
Laws to ensure AI applications are safe, fair, and transparent are needed. But the White House's use of the Defense Production Act—typically reserved for war or national emergencies—distorts AI through the lens of security, for example with phrases like "companies developing any foundation model that poses a serious risk to national security."  
Yes, AI -- like many technologies such as electricity and encryption -- is dual use in the sense it can be used for civilian or military purposes. But conflating AI safety for civilian use cases and military applications is a mistake.  
It's also a mistake to set reporting requirements based on a computation threshold for model training. This will stifle open source and innovation:  
(i) Today's supercomputer is tomorrow's pocket watch. So as AI progresses, more players -- including small companies without the compliance capabilities of big tech -- will run into this threshold.  
(ii) Over time, governments' reporting requirements tend to become more burdensome. (Ask yourself: Has the tax code become more, or less, complicated over time?)  
The right place to regulate AI is at the application layer. Requiring AI applications such as underwriting software, healthcare applications, self-driving, chat applications, etc. to meet stringent requirements, even pass audits, can ensure safety. But adding burdens to foundation model development unnecessarily slows down AI's progress.  
While the White House order isn't currently stifling startups and open source, it seems to be a step in that direction. The devil will be in the details as its implementation gets fleshed out -- no doubt with assistance from lobbyists -- and I see a lot of risk of missteps. I welcome good regulation to promote responsible AI, and hope the White House can get there.

5:02 AM · Nov 1, 2023 · 847.8K Views

- Mở ra rất nhiều ứng dụng trong việc tự động hóa các công việc cá nhân.



# Tình hình Developer AI trong 2024?

- Nhu cầu tuyển dụng tăng mạnh do nhu cầu xây dựng Chatbot cho các doanh nghiệp tăng cao.
- Các AI Dev ở Việt Nam thường bắt đầu với Computer Vision, ít kinh nghiệm trong mảng xử lý ngôn ngữ tự nhiên (NLP)
- Các ứng dụng sử dụng LLM đang tăng cao, tuy nhiên rất ít tính đột phá do không có sự khác biệt trong kĩ thuật
- Các Developers với kĩ năng xây dựng ứng dụng LLM advance sẽ có lợi thế khi tham gia tuyển dụng.



# Nội dung khoá học

Trong 4.5 tuần, học viên có cơ hội tìm hiểu cũng như xây dựng các ứng dụng như:

- Xử dụng các thư viện phổ biến: **LangChain, Gradio, DSPy, ...**
- Xây dựng các **hệ thống RAG** để giảm hallucination và tăng độ chính xác cho Chatbot
- Xây dựng **Agents** để **tự động hóa** công việc
- Sử dụng **LLM mã nguồn mở** (Vistral/VinaLlama)
- **Finetune mô hình ngôn ngữ** theo dữ liệu riêng của bạn
- **Cách đánh giá và giảm thiểu các rủi ro** của hệ thống LLM-based



## 2. Lộ trình và phương pháp học

# TRỞ THÀNH LLM DEVELOPER CHUYÊN NGHIỆP



# Phương pháp học hiệu quả?

1. Tích cực tham gia các buổi học & **THỰC HÀNH** đầy đủ sau mỗi buổi học
2. Xác định một **ngách thị trường/loại hình dịch vụ** mà mình muốn **kiếm tiền**
3. Tích cực **đặt câu hỏi** trong nhóm lớp & tìm kiếm các **mảnh ghép** kết nối phù hợp
4. Liên tục **cập nhật kiến thức mới** trong ngách thị trường mình cung cấp qua:
  - follow chuyên gia Facebook
  - follow founder các platform AI trên Twitter
  - đọc blog OpenAI & tham gia community forum
  - đọc blog, newsletter của thư viện langchain và llmindex
  - đọc medium



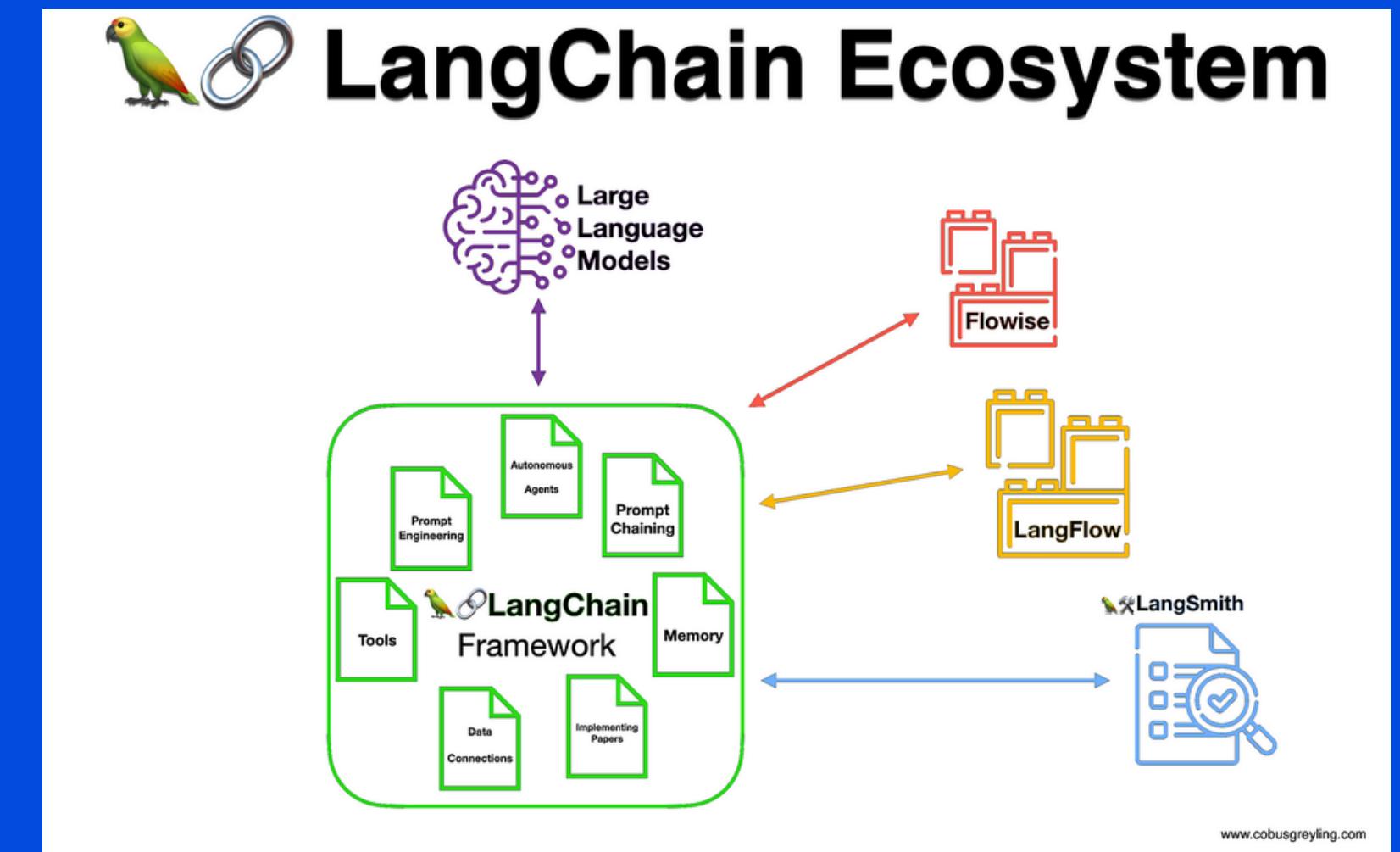
# Thư viện sẽ sử dụng trong khóa học

## LLM:

- OpenAI API

## Công cụ:

- Numpy: Thư viện dùng để tính toán trên Python
- LangChain & LangSmith: Nhanh chóng sử dụng LLM để áp dụng vào công việc
- Gradio: Thư viện UI đơn giản bằng Python
- Optional: LM Studio/Ollama – có thể sử dụng nếu muốn sử dụng LLM mã nguồn mở
- Anaconda: Dùng để quản lý các thư viện trên Python
- DSPy: Thư viện nâng cao giống LangChain



## Giao tiếp:

- Zalo Group

# Tổng hợp nguồn nên theo dõi

Danh sách các nguồn nên theo dõi





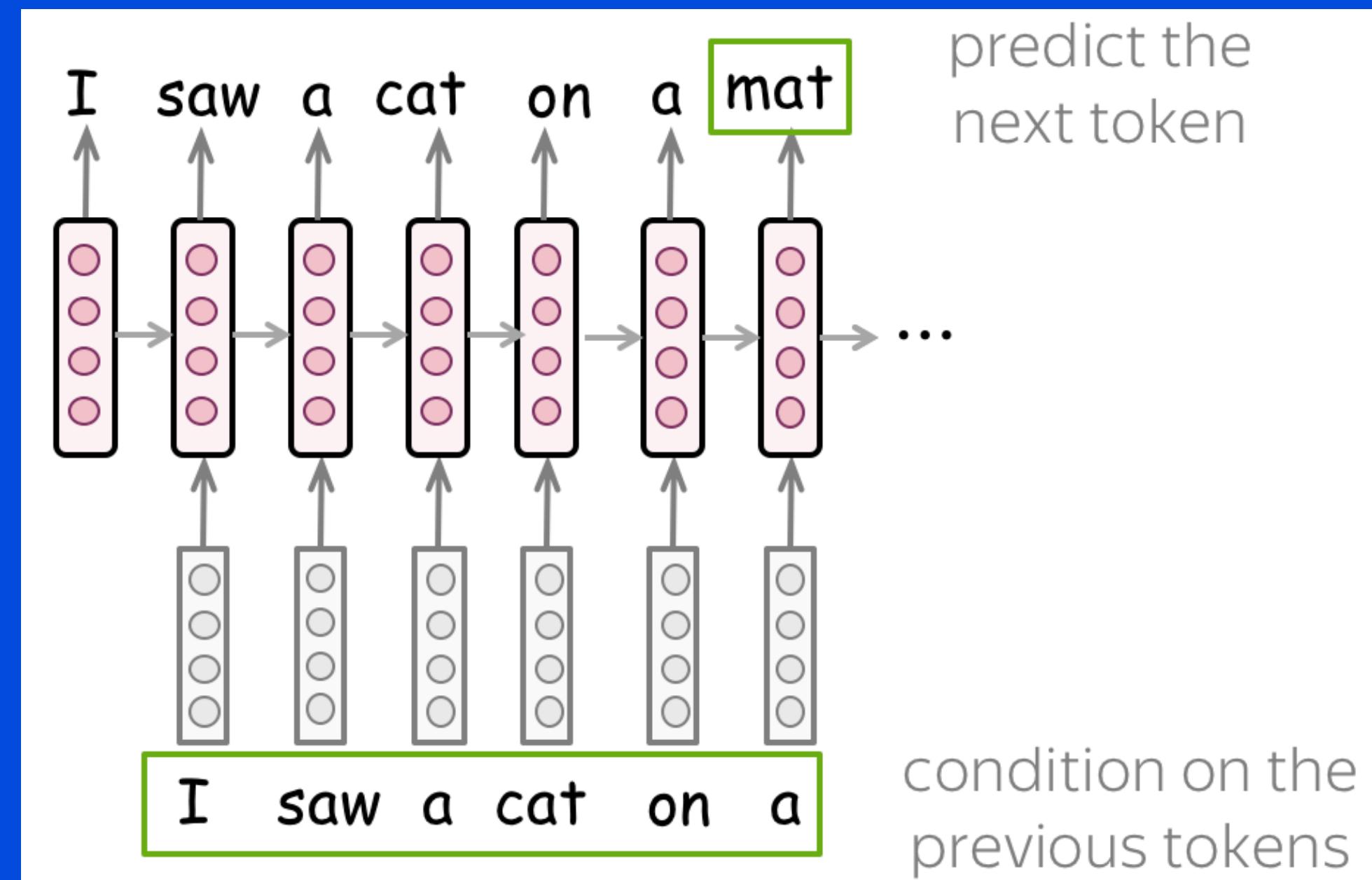
4. Lý Thuyết

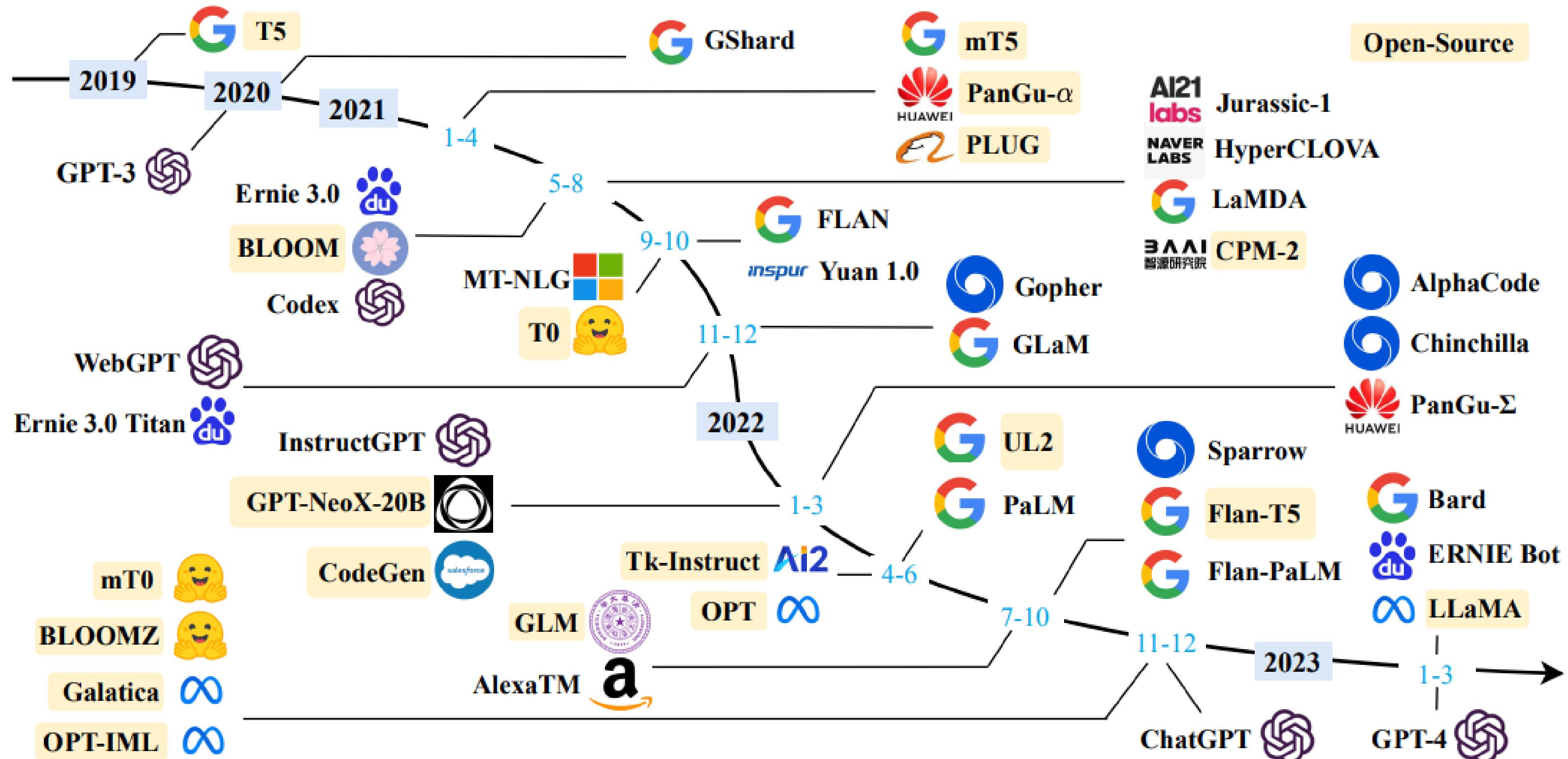
# NGUYÊN LÍ HOẠT ĐỘNG CỦA LLM



# Language Model là gì?

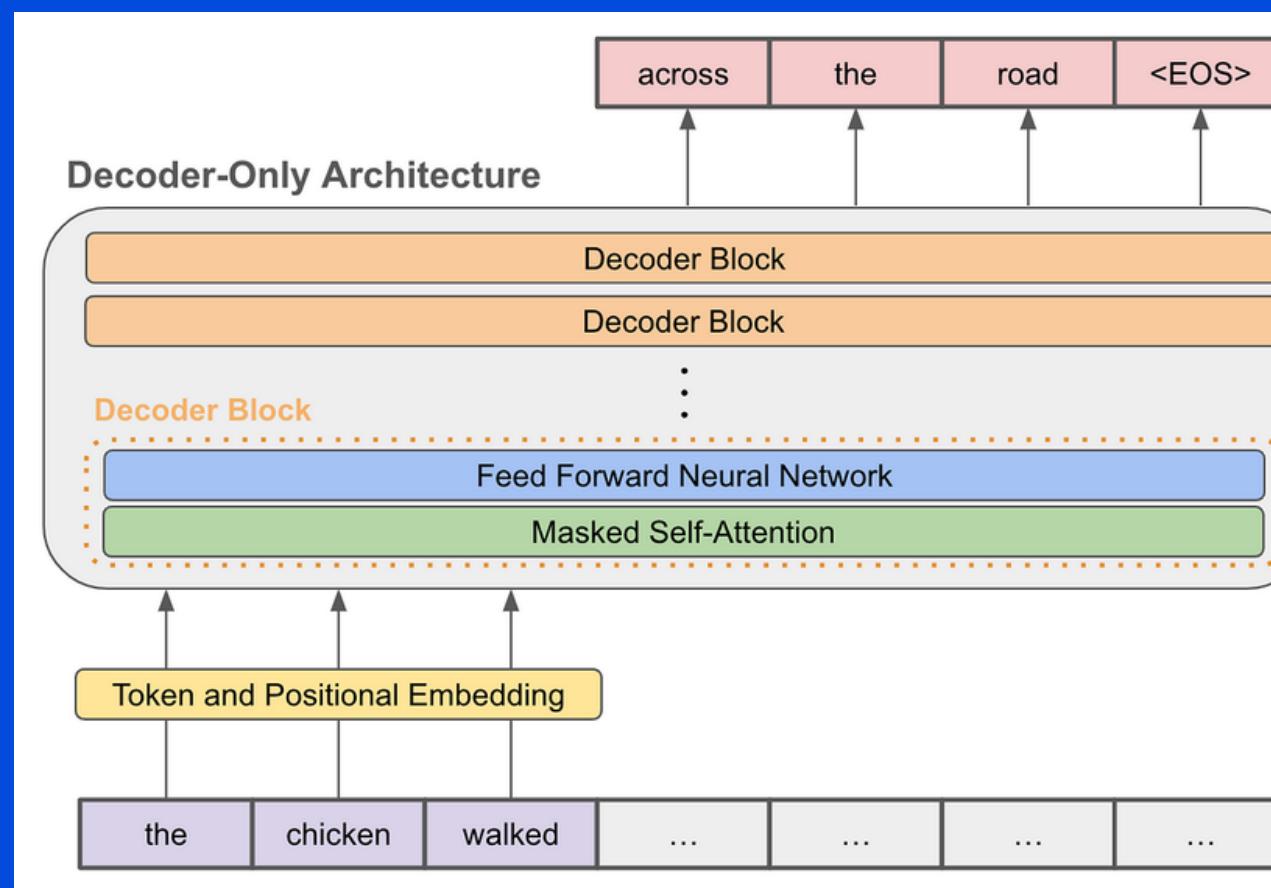
- Language Model được huấn luyện bằng việc che các từ đằng sau của một văn bản, sau đó bắt mô hình AI phải đoán các từ tiếp theo dựa trên các từ có sẵn
- Đây chính là task “Điền vào ô trống” của con người.



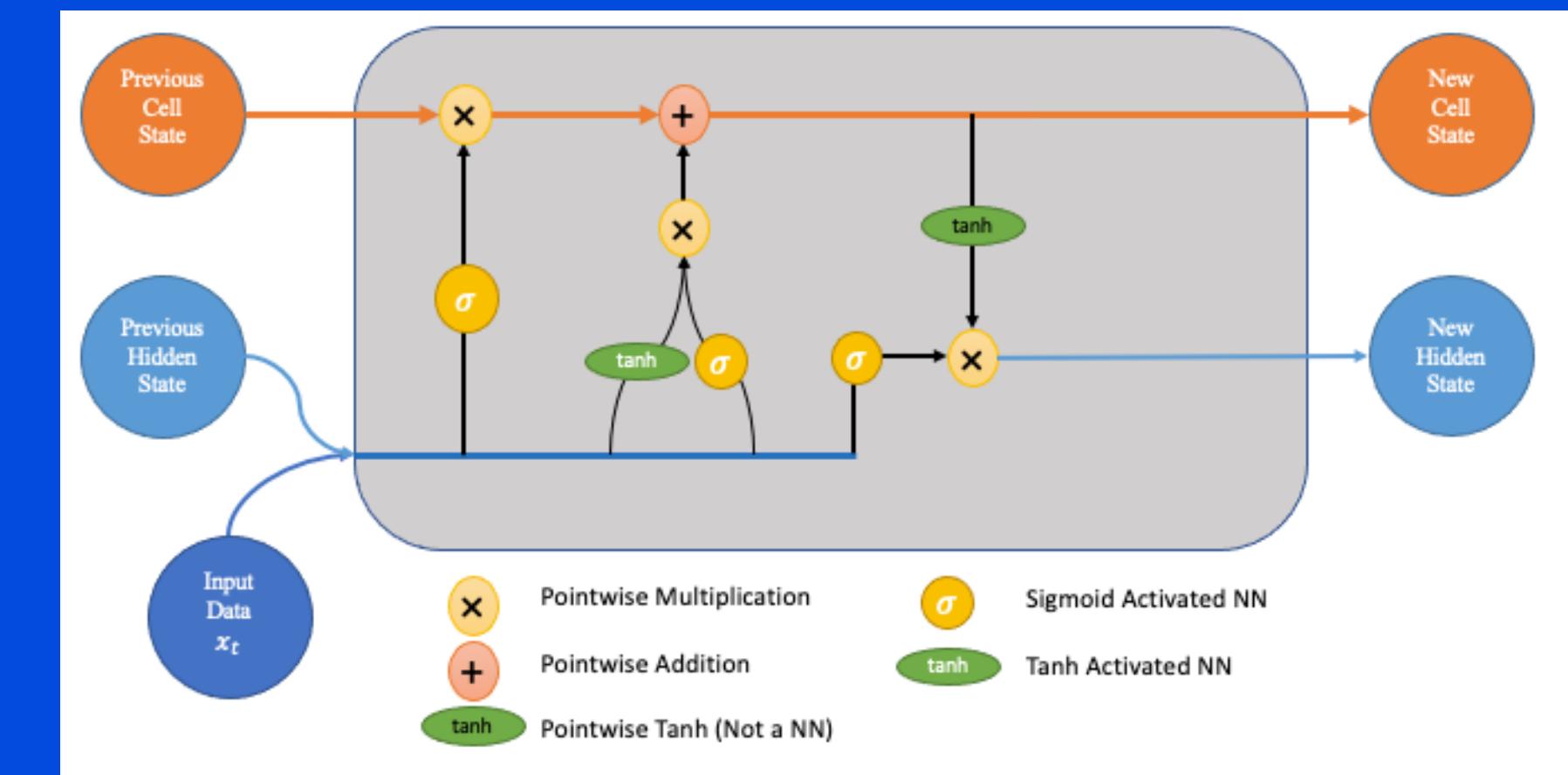


# LLM, Transformers và Nguyên Lý Hoạt Động

- LLM là viết tắt của "Large Language Model", hay còn gọi là **Mô hình ngôn ngữ lớn**.
- LLM thường được huấn luyện trên một tập dữ liệu **văn bản rất lớn (từ 2-5TB)**, nhờ vào việc được đọc rất nhiều văn bản, LLM dần học được cấu trúc câu từ cũng như ngữ pháp của ngôn ngữ con người.



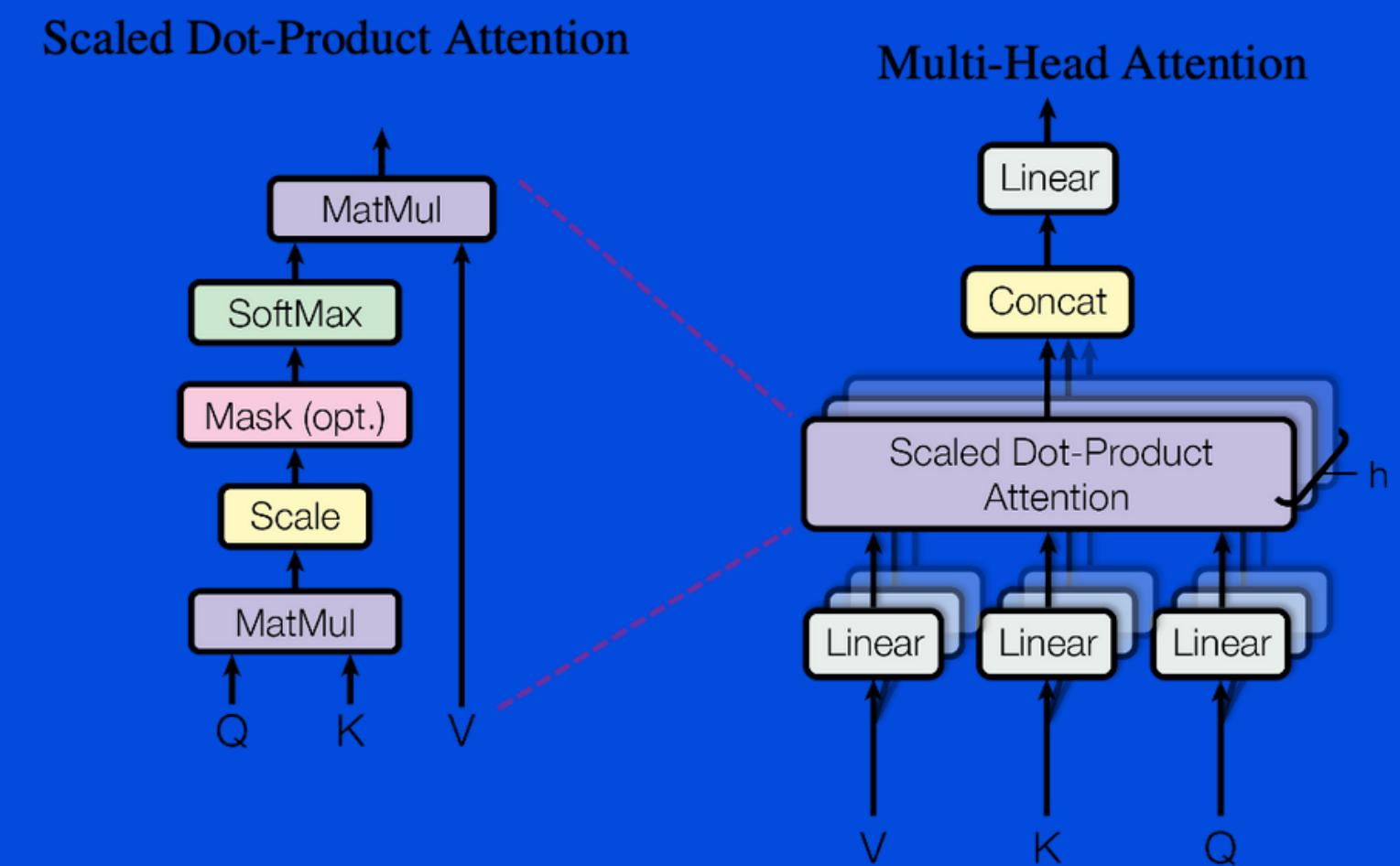
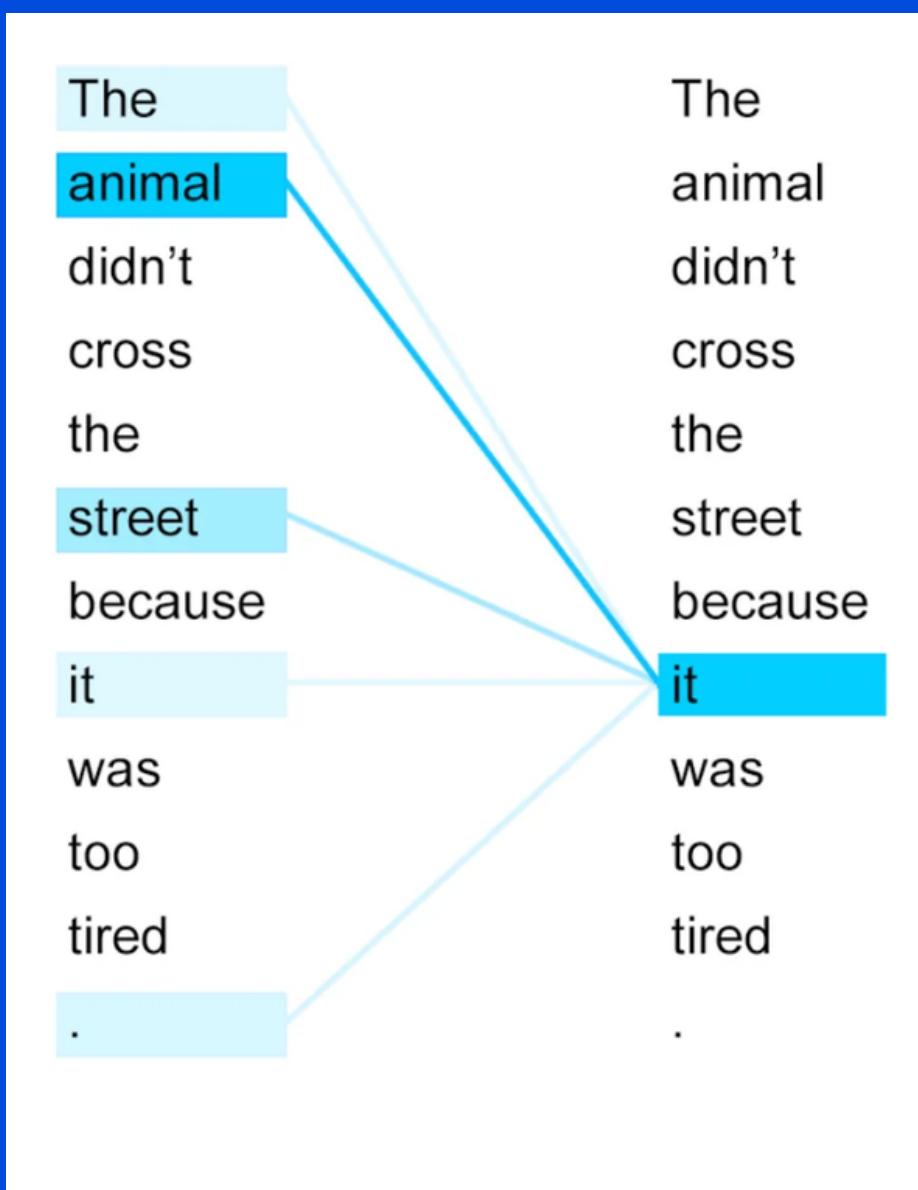
Cấu trúc Decoder của Transformers



Cấu trúc của LSTM

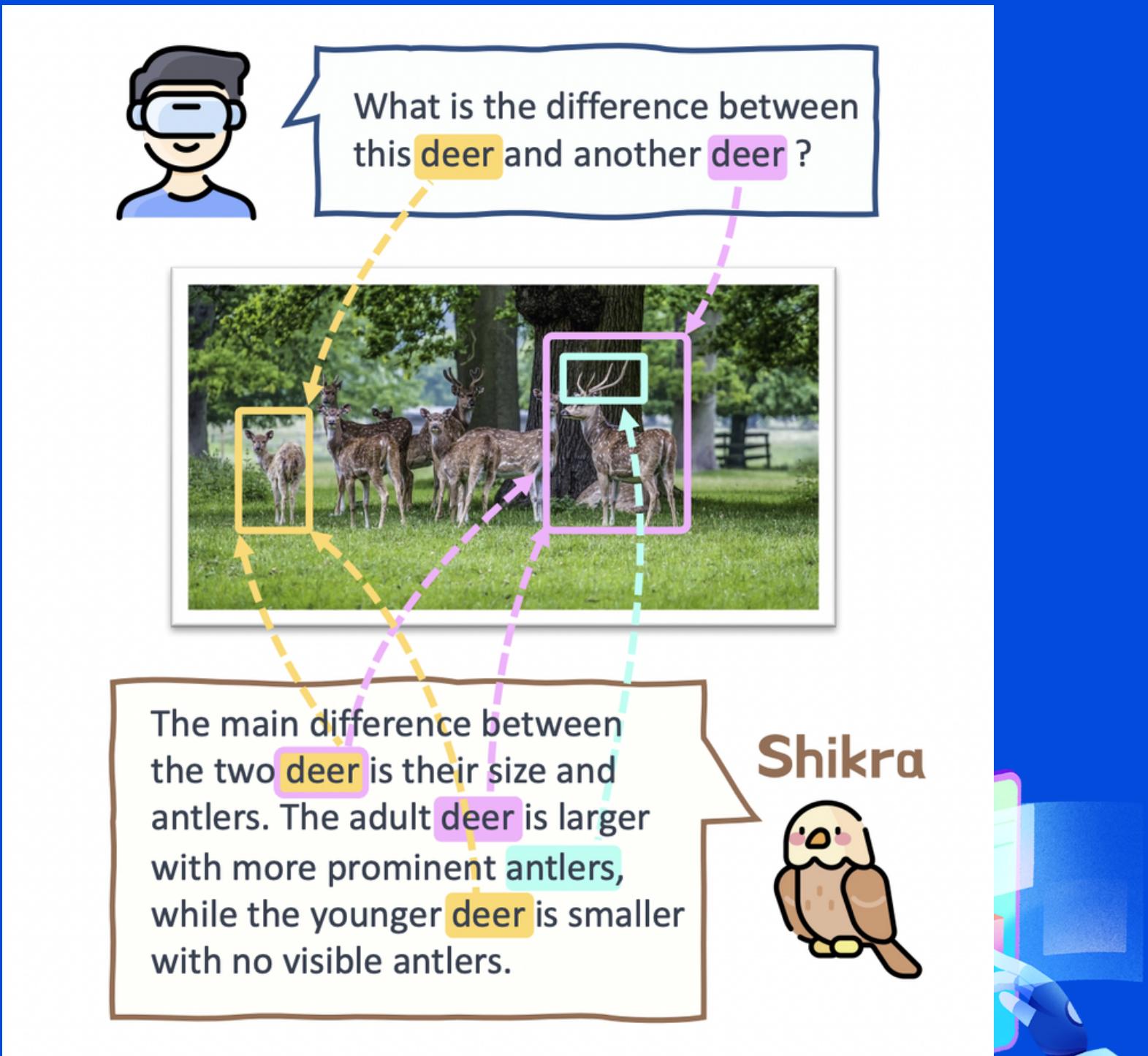
# Self-Attention: Vũ Khí Bí Mật của Transformers

- **Self-attention** là một kỹ thuật giúp mô hình học cách **tập trung vào những phần quan trọng nhất của dữ liệu** đầu vào để xử lý.
- Giống như con người khi đọc sách, ta sẽ tập trung vào những **từ khóa và câu quan trọng** để hiểu nội dung, self-attention cũng hoạt động tương tự.



# Các biến thể cao cấp của LLM

- **Multimodal:** LLM có thể tiếp nhận và xử lý các loại dữ liệu ngoài văn bản như Hình ảnh, âm thanh.
- **Large Action Model:** Các LLM có thể đưa ra các quyết định và hành động dựa trên các nội dung sẵn có để tự động hóa.
- **Worlds Model:** Mô hình AI có thể hiểu được các bản chất của cuộc sống, bao gồm vật thể, vật lí và nguyên nhân-hậu quả.





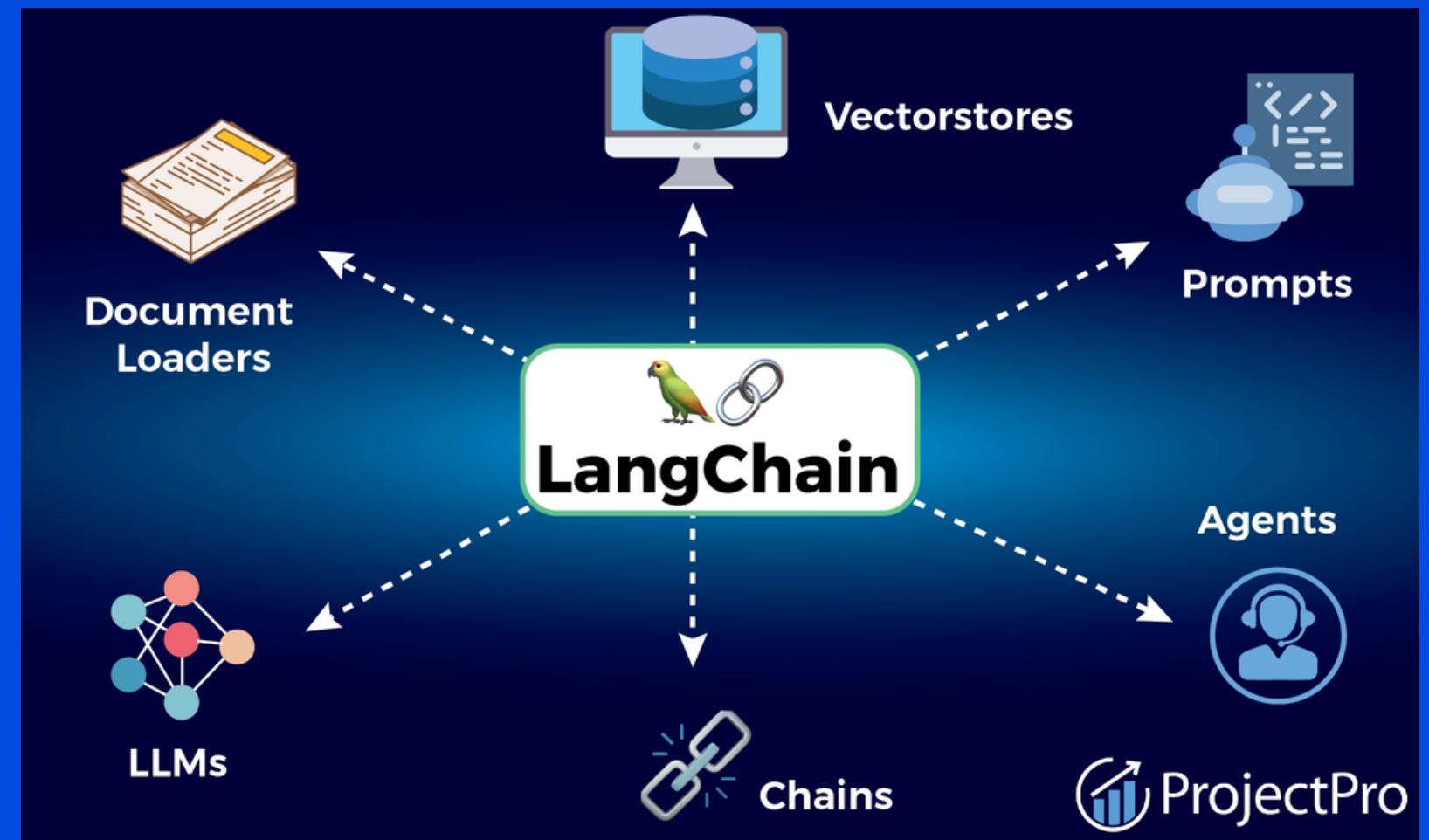
5. Lý Thuyết

# HỆ SINH THÁI LLM & LangChain



# LangChain là gì?

- LangChain giúp bạn dễ dàng tích hợp LLM như GPT, Gemini vào ứng dụng của bạn, mở ra khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ.
- LangChain cung cấp công cụ để tinh chỉnh LLM theo nhu cầu cụ thể của bạn, nâng cao độ chính xác và hiệu quả cho từng ứng dụng.
- LangChain đơn giản hóa các tác vụ phức tạp, giúp bạn xây dựng ứng dụng AI nhanh chóng và dễ dàng hơn, tập trung vào logic và chức năng cốt lõi.



# LangChain Module - LLM

```
from langchain_openai import ChatOpenAI  
from langchain_openai import OpenAI  
  
llm = OpenAI()  
chat_model = ChatOpenAI(model="gpt-3.5-turbo-0125")
```

- Module LLM cho phép người dùng đổi LLM chỉ với một dòng code

Model
AI21
AlephAlpha
AmazonAPIGateway
Anthropic
Anyscale
Aphrodite
Arcee
Aviary
AzureMLOnlineEndpoint
AzureOpenAI

# LangChain Module – Text Splitters, Document Loaders

```
pip install pypdf

from langchain_community.document_loaders import PyPDFLoader

loader = PyPDFLoader("example_data/layout-parser-paper.pdf")
pages = loader.load_and_split()

pages[0]
```

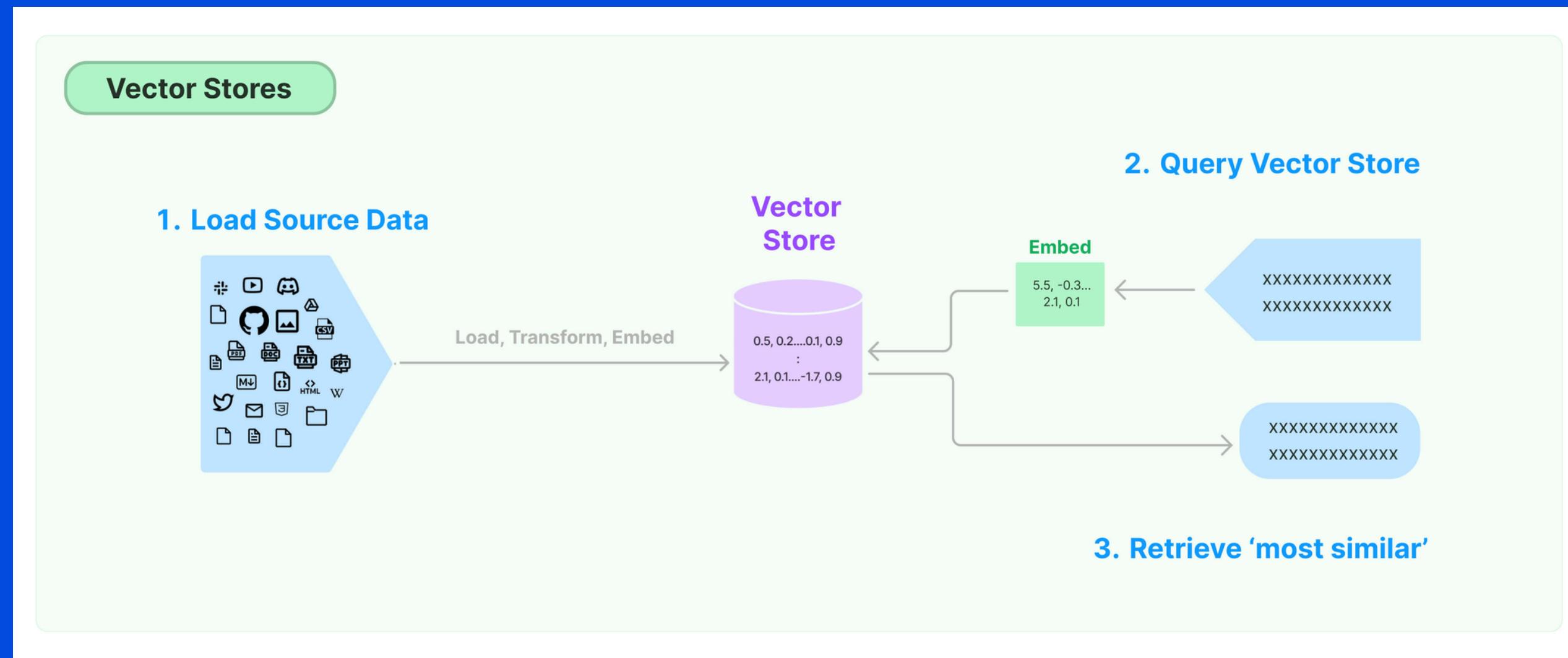
```
from langchain_text_splitters import CharacterTextSplitter

text_splitter = CharacterTextSplitter(
    separator="\n\n",
    chunk_size=1000,
    chunk_overlap=200,
    length_function=len,
    is_separator_regex=False,
)
```

- Module Document Loaders & Text Splitters giúp tiền xử lý các văn bản trong các loại file PDF, HTML, Doc(x), csv, ... trở nên đơn giản

# LangChain Module – Vectorstore

- LangChain tích hợp hầu hết các Vectorstore phổ biến hiện tại, cho phép người dùng dễ dàng tích hợp vào hệ thống RAG



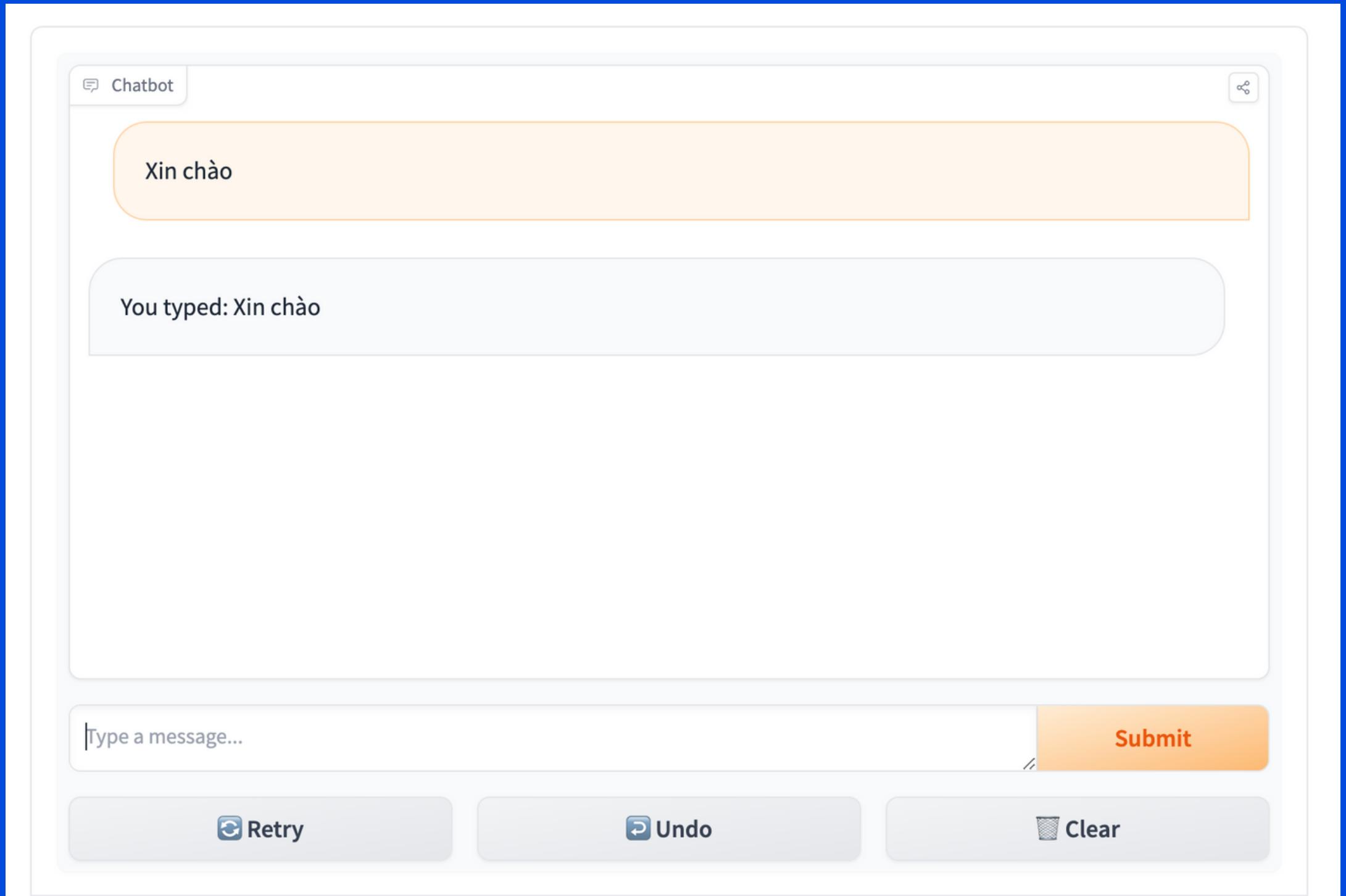
# LangChain Module - Chains

- Một số các chuỗi LLM (Chains) đã được thêm sẵn bởi LangChain.  
Giúp triển khai nhanh một số bài toán đơn giản

QAGenerationChain			Creates both questions and answers from documents. Can be used to generate question/answer pairs for evaluation of retrieval projects.
RetrievalQAWithSourcesChain		Retriever	Does question answering over retrieved documents, and cites its sources. Use this when you want the answer response to have sources in the text response. Use this over <code>load_qa_with_sources_chain</code> when you want to use a retriever to fetch the relevant document as part of the chain (rather than pass them in).
load_qa_with_sources_chain		Retriever	Does question answering over documents you pass in, and cites its sources. Use this when you want the answer response to have sources in the text response. Use this over RetrievalQAWithSources when you want to pass in the documents directly (rather than rely on a retriever to get them).
RetrievalQA		Retriever	This chain first does a retrieval step to fetch relevant documents, then passes those documents into an LLM to generate a response.
MultiPromptChain			This chain routes input between multiple prompts. Use this when you have multiple potential prompts you could use to respond and want to route to just one.
MultiRetrievalQAChain		Retriever	This chain routes input between multiple retrievers. Use this when you have multiple potential retrievers you could fetch relevant documents from and want to route to just one.

# Gradio

- Gradio là thư viện UI dựa trên Python với các element build sẵn. Giúp người dùng nhanh chóng tạo giao diện Chat.





## 6. Thực hành đầu tiên

# XÂY DỰNG PHIÊN BẢN SỐ CÁ NHÂN



# Sản phẩm

1. Tạo phiên bản cá nhân của chính bạn với OpenAI API
2. Có thể show giao diện qua Gradio UI.



# Làm Quen Với Notebook, Python và Numpy

Truy Cập Notebook

# Làm Quen Với OpenAI API, LangChain và Xây Dựng Phiên Bản Số

Truy Cập Notebook



# Tổng kết & BTVN



## Một số nội dung chính đã học trong buổi

1. Thế giới công nghệ đang chuyển dịch sang thời đại mới, nơi mà AI là một trong những công nghệ lõi.
2. Cơ hội việc làm cho Developer AI LLM ngày càng cao -> cạnh tranh cao
3. Tạo ra sự khác biệt bằng việc các well-designed LLM flow. Việc build custom chatbot sẽ mang lại nhiều lợi thế.
4. Tạo phiên bản số của bạn qua LangChain và Gradio.



# Bài tập thực hành tại nhà

1. Đọc [bài viết này](#) để hiểu thêm về LangChain
2. Đọc [bài viết này](#) để tìm hiểu trước về RAG, chuẩn bị tốt hơn cho bài học sau
3. Chọn một người nổi tiếng mà bạn yêu thích, tạo phiên bản số của họ. Sau đó đưa code này lên Notebook hoặc GitHub.
4. Thử inbox và comment vào group của lớp học về Notebook hoặc GitHub của bạn.

