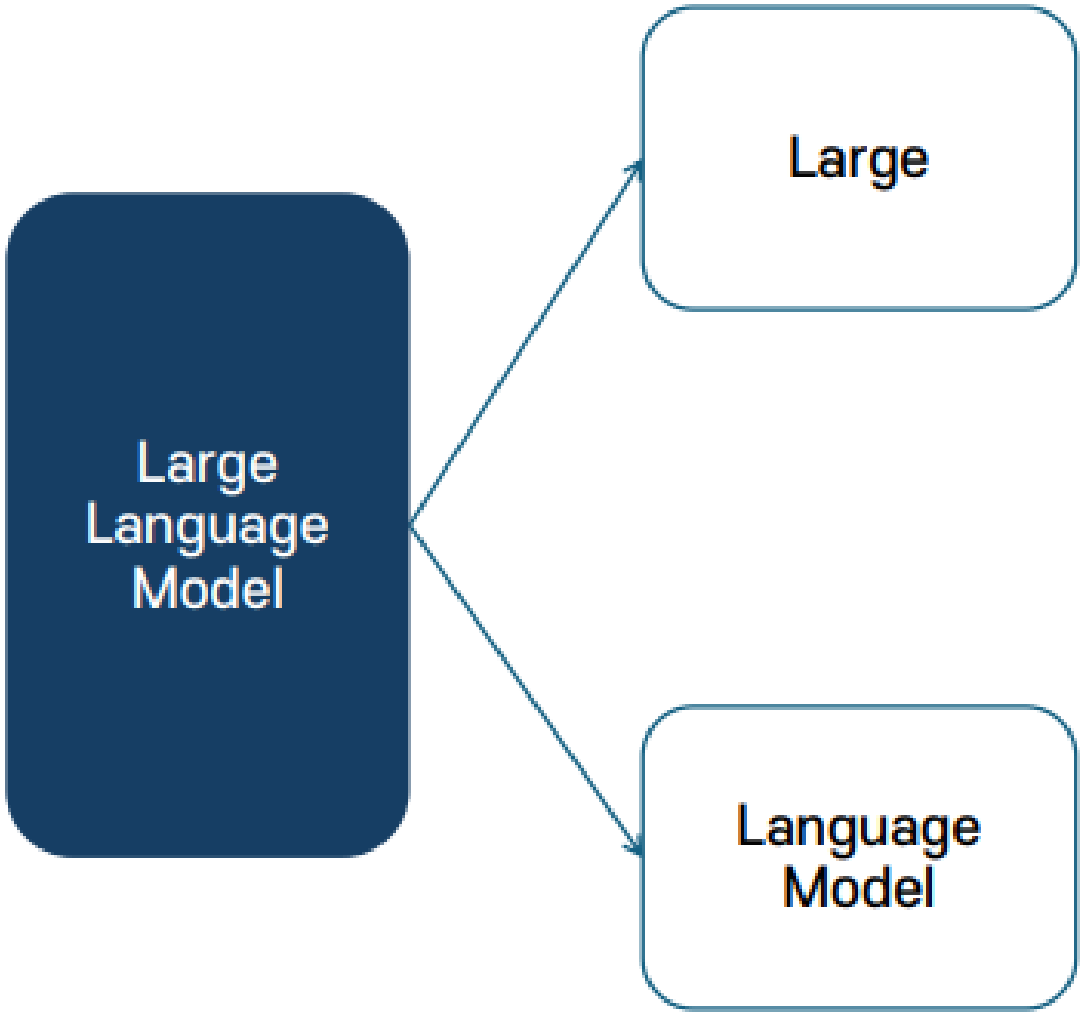
LangChain으로 LLM 서비스 개발하기

류영표 강사
ryp1662@gmail.com

Langchain으로 LLM 서비스 개발하기

LLM이란?



모델이 학습하는데 사용된 데이터의 양,
모델이 가지고 있는 파라미터의 수,
그리고 처리할 수 있는 언어의 범위와 능력

단어(또는 문자) 시퀀스에
확률을 할당(assign)하는 모델

Language Model

Language Model

언어 모델은 단어 시퀀스에 확률을 할당(assign) 하는 일을 하는 모델입니다.

이를 조금 풀어서 쓰면, 언어 모델은 가장 자연스러운 단어 시퀀스를 찾아내는 모델입니다.

https://wikidocs.net/21668

Language Model을 어떻게 만들까?

1. 어떤 문제를 풀게 할 것인가?

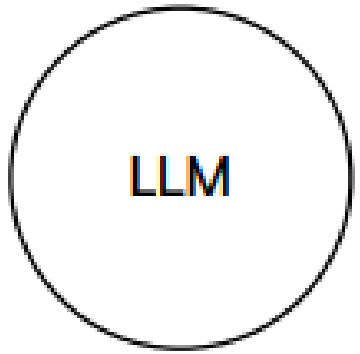
가장 보편적으로 사용되는 방법은
언어 모델이 이전 단어들이 주어졌을 때 다음 단어를 예측하도록 하는 것입니다.

이 외에도 주어진 양쪽의 단어들로부터 가운데 비어있는 단어를 예측하는 언어 모델 등
다양한 방법들이 활용되었습니다.

2. 어떤 방법론을 쓸 것인가?

Ngram, 확률

RNN, LSTM, Transformer



GPT-3

Large Language Model

Large Language Model

** TB
(약 45TB 추정)

1750억개

GPT-3를 기준으로 보는 LLM의 크기

Large Language Model

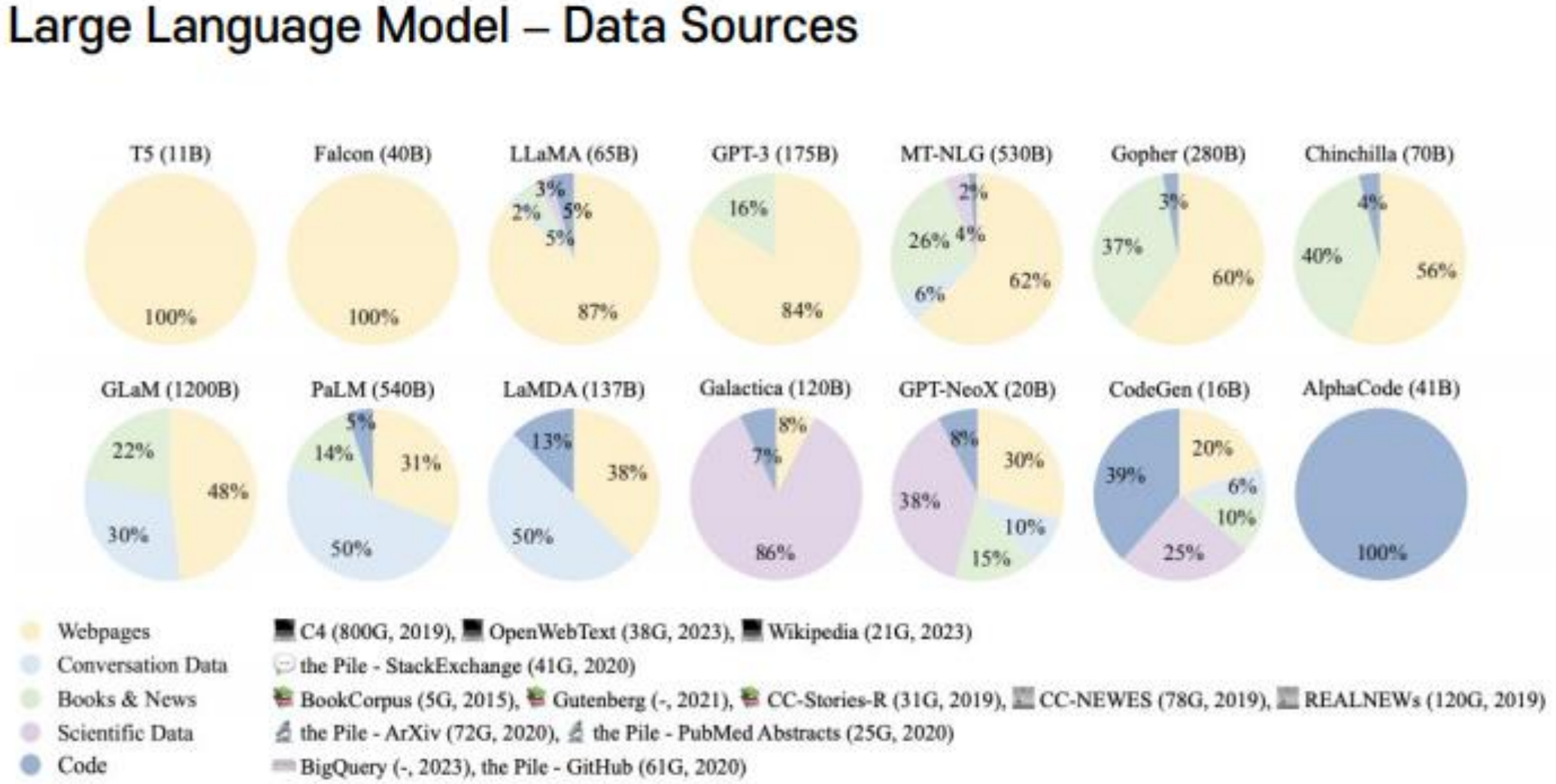
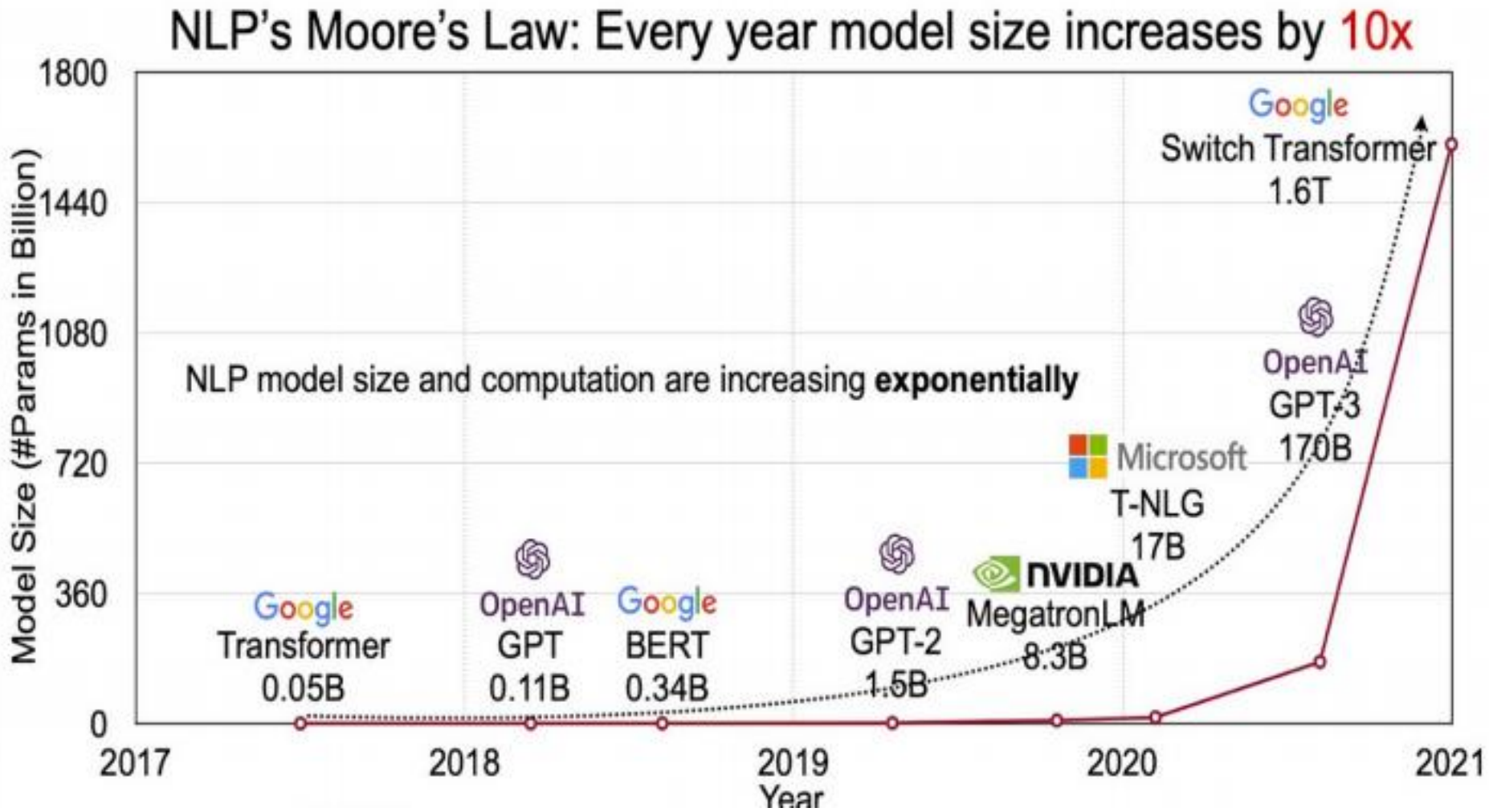


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.

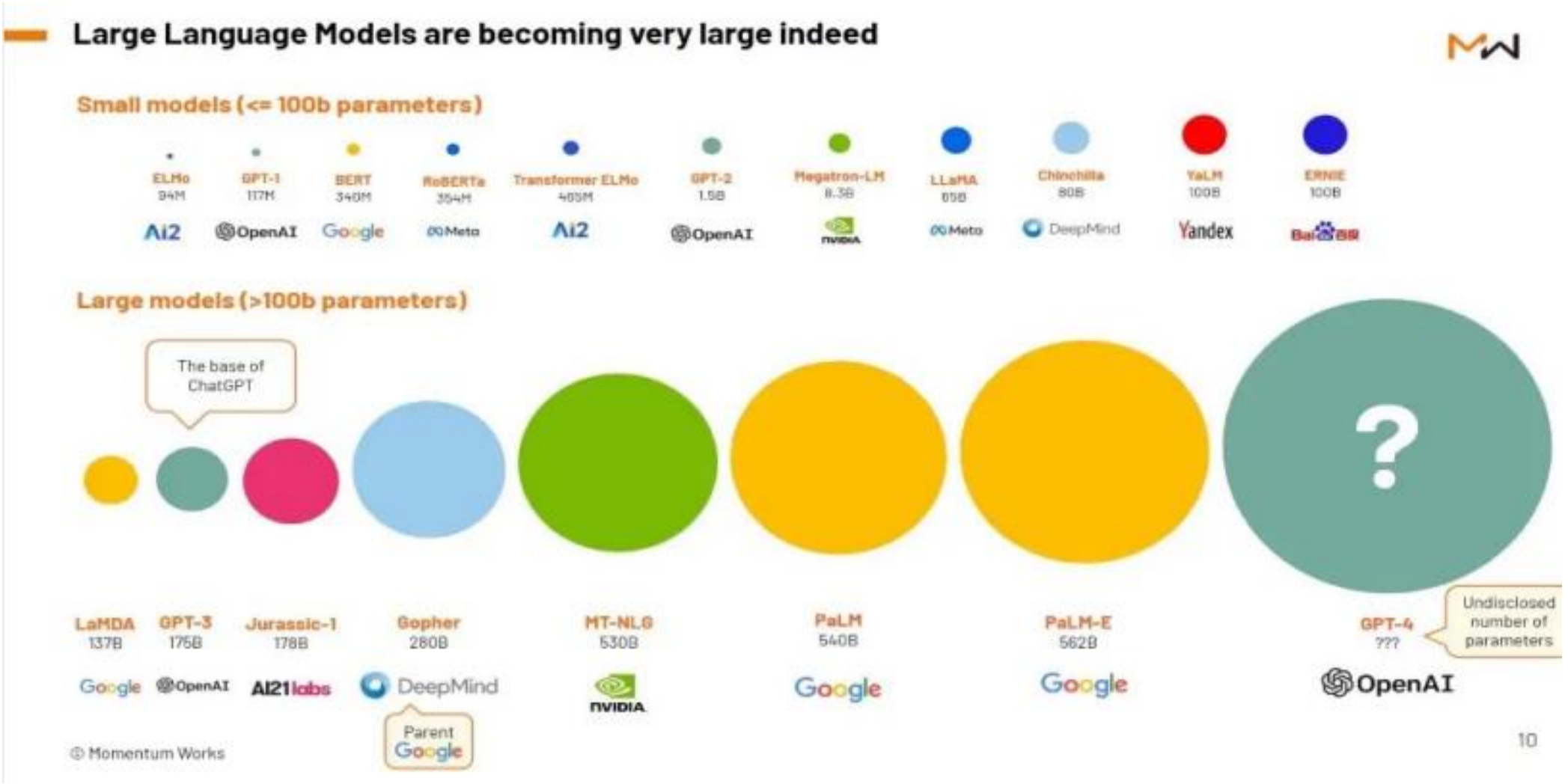
https://arxiv.org/pdf/2303.18223

Large Language Model



https://medium.com/@harishdatalab/unveiling-the-power-of-large-language-models-llms-e235c4eba8a9

Large Language Model

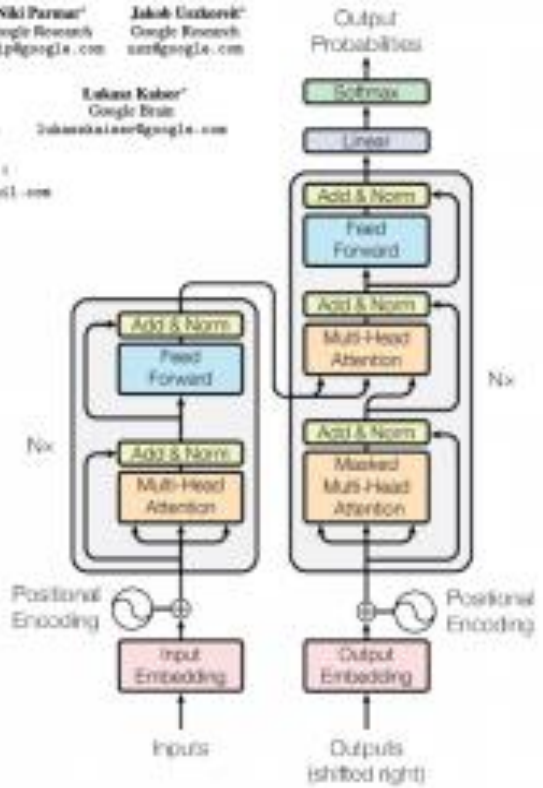


https://www.bureauworks.com/blog/what-is-large-language-models-llm

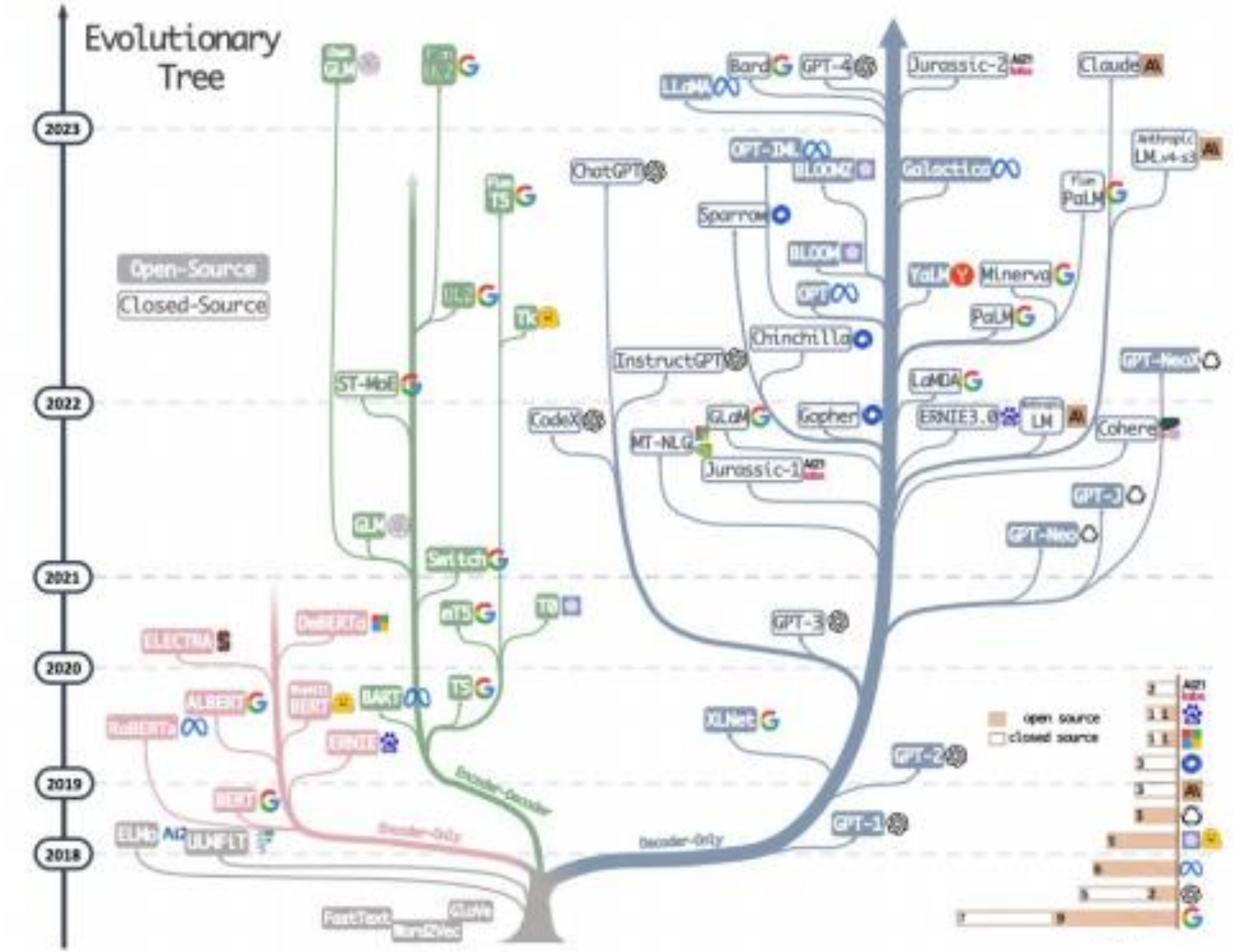
LLM의 핵심 구조 - 트랜스포머(Transformer)

Attention Is All You Need

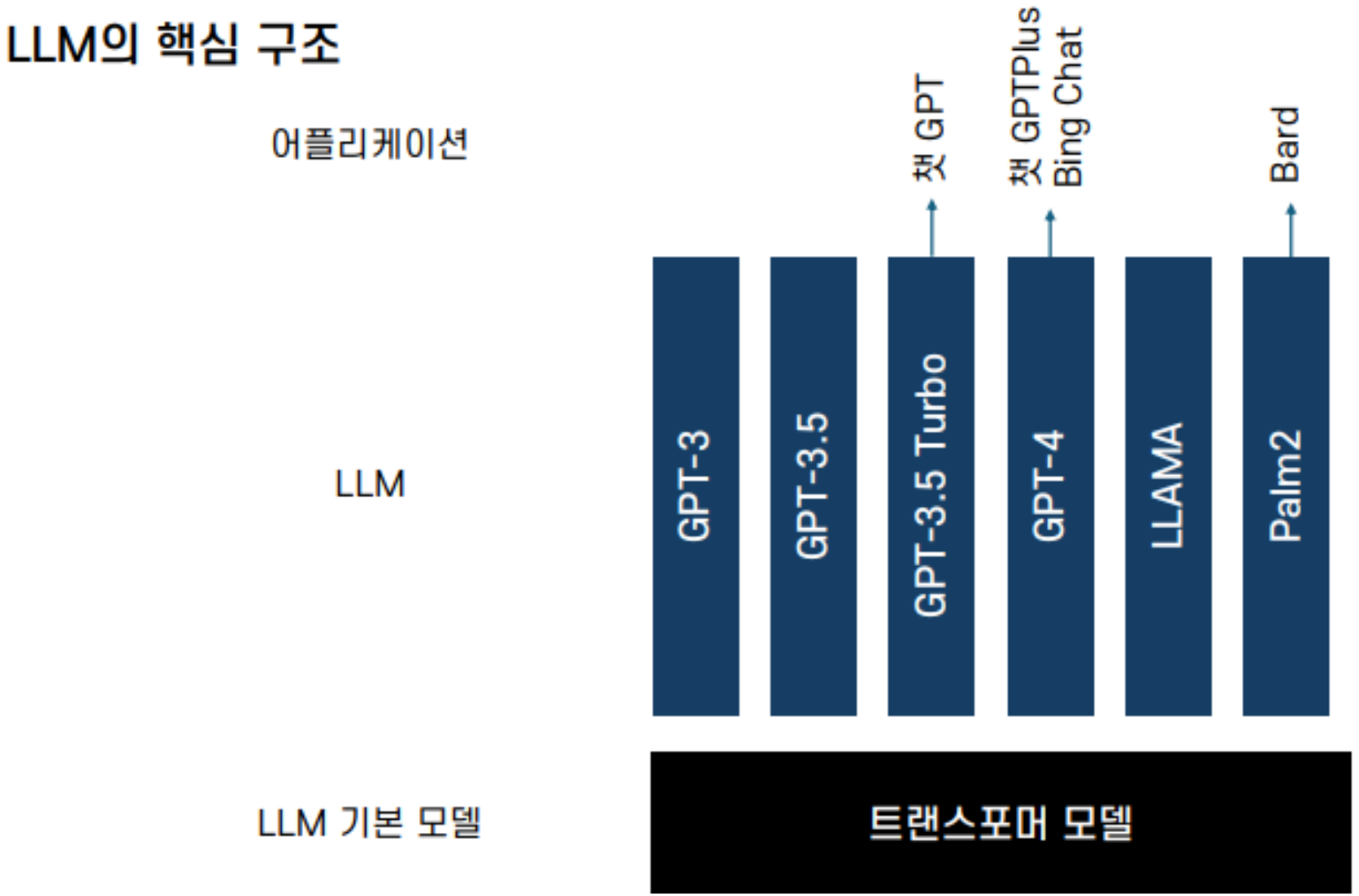
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* University of Toronto aidan@cs.toronto.edu
Lukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* illia.polosukhin@gmail.com



https://arxiv.org/abs/1706.03762
https://arxiv.org/pdf/2304.13712.pdf



LLM의 핵심 구조



책 : [랭체인으로 LLM 기반의 AI 서비스 개발하기] - 서지영

LLM

LangChain의 구조

LLM

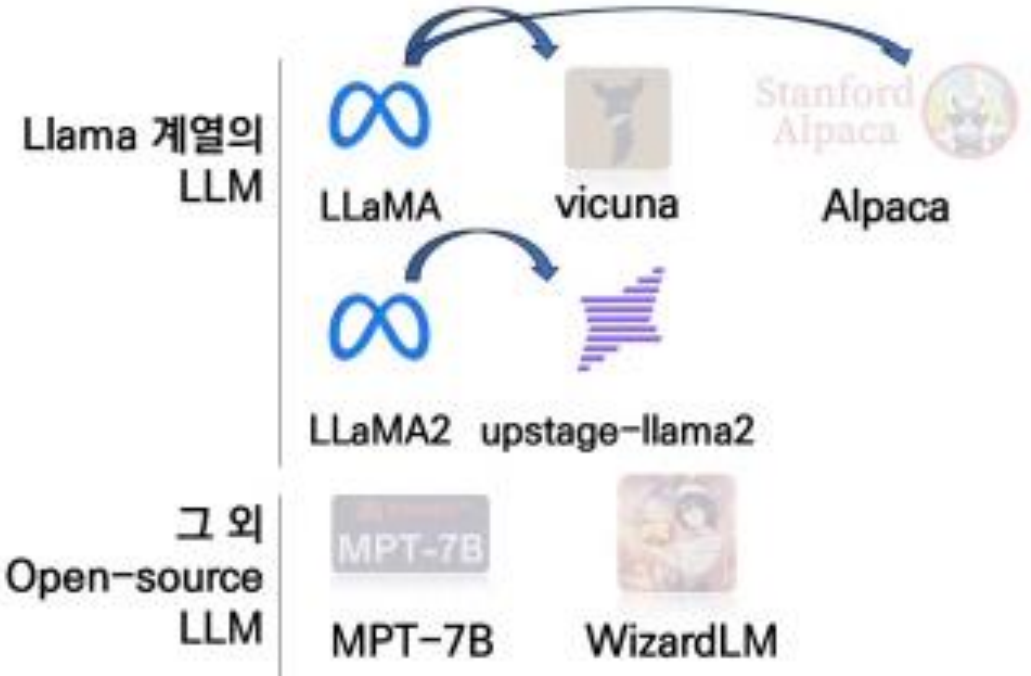
Closed Source

	OpenAI	Google
개발사		
개발 모델	GPT-3	PALM
	GPT-3.5	LaMDA
	GPT-4	Bard

Good! 뛰어난 성능, API 방식의 편리한 사용성

Bad 보장할 수 없는 보안, API 호출 비용

Open Source



Good! Closed source 못지 않은 성능, 높은 보안성, 낮은 비용

Bad 개발 난이도 높음, 사용 위한 GPU 서버 필요

https://www.youtube.com/watch?v=XwlLeVhWCCc / [모두의 AI] LANGCHAIN 편

LLM

오늘의 주제로 돌아와서

Langchain으로 LLM 서비스 개발하기

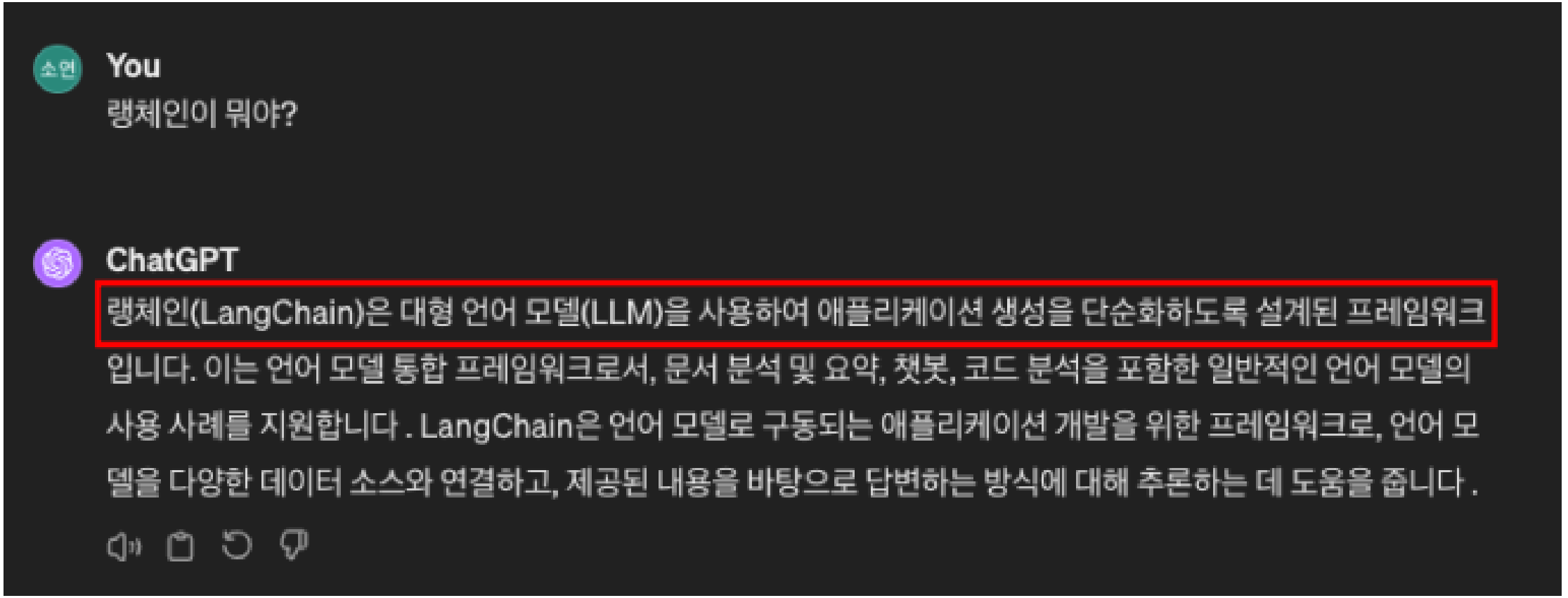
LLM 만들기 ❌

LLM 활용하기 🟢

https://www.youtube.com/watch?v=XwlLeVhWCCc / [모두의 AI] LANGCHAIN 편

Langchain으로 LLM 서비스 개발하기

Langchain이란?



Langchain이란?

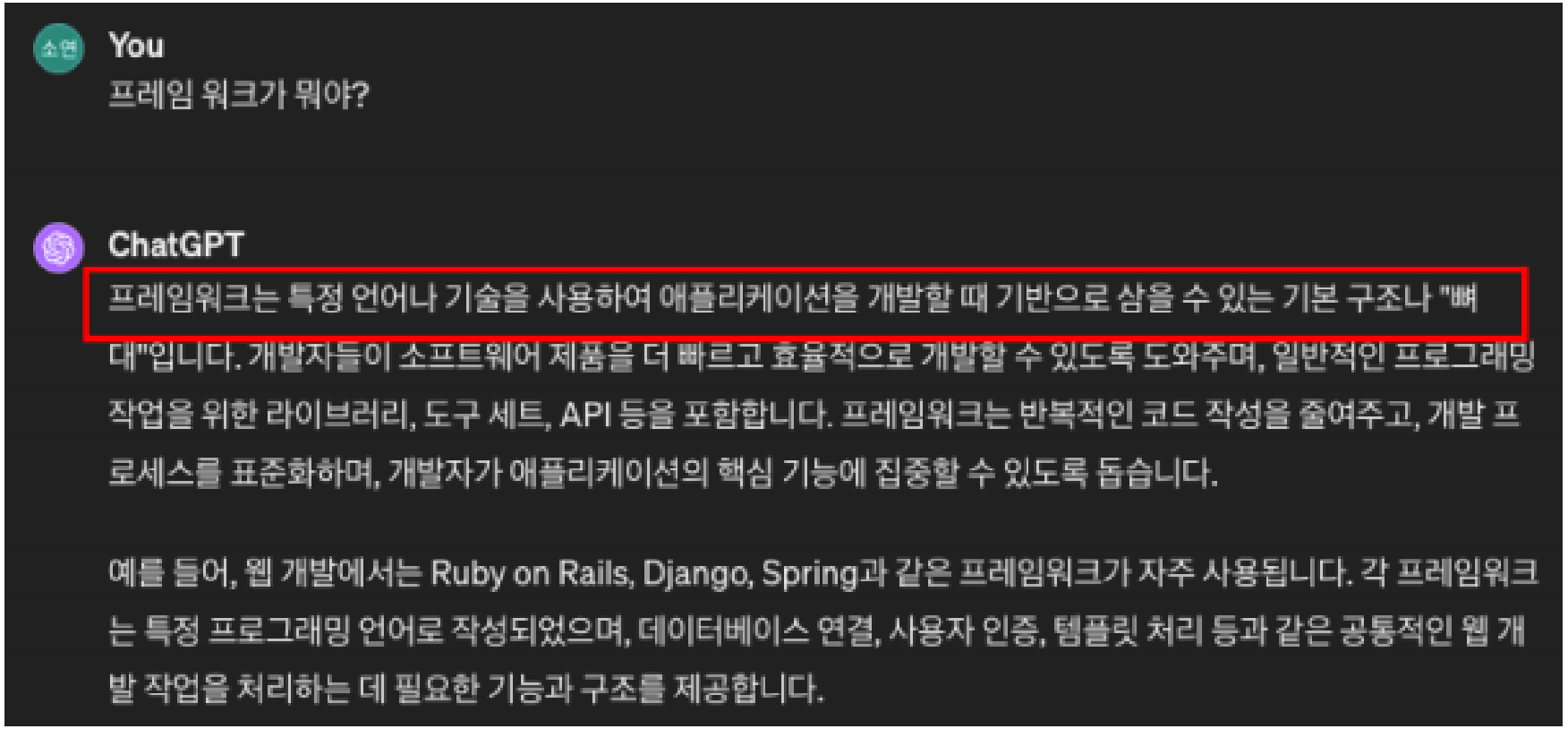
대형 언어모델(LLM) 어플리케이션 개발 프레임워크

LangChain

프레임워크

프레임워크란?

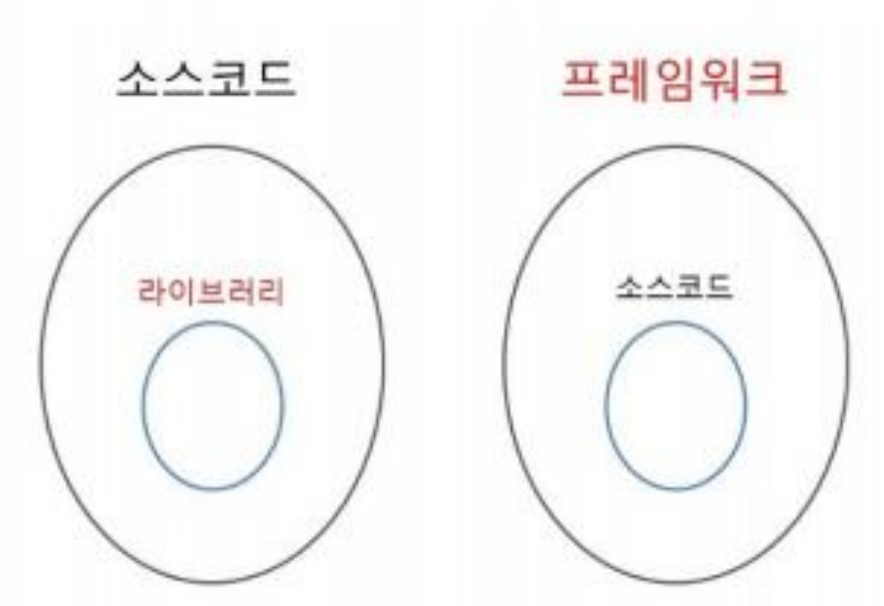
= 관련 도구 모음



프레임워크

프레임워크란?

= 관련 도구 모음



특징	프레임워크	라이브러리
유저코드의 작성	프레임워크 클래스를 서브 클래싱 해서 작성	독립적으로 작성
호출흐름	프레임워크코드가 유저코드를 호출	유저코드가 라이브러리를 호출
실행흐름	프레임워크가 제어	유저코드가 제어
객체의 연동	구조프레임워크가 정의	독자적으로 정의

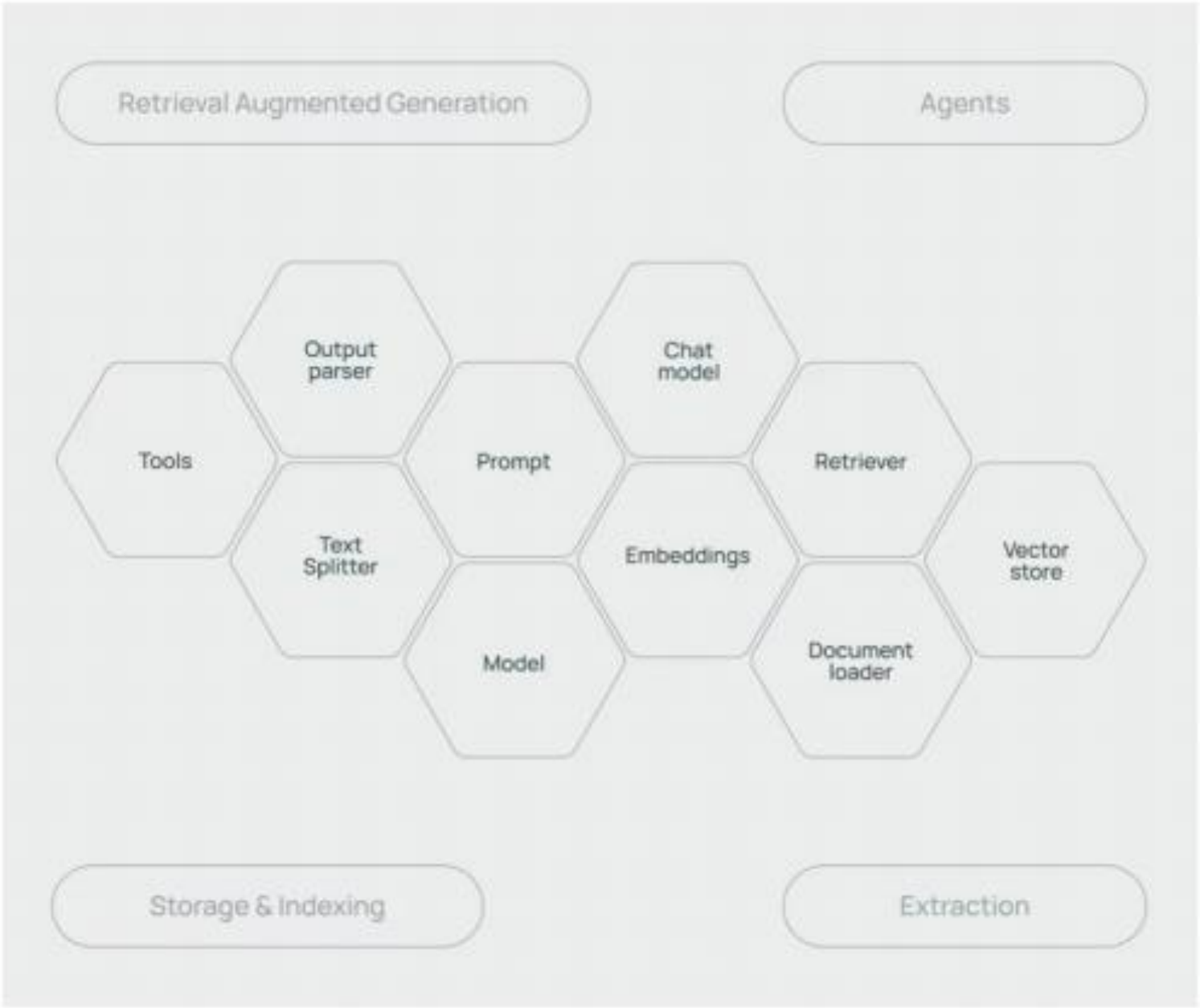
〈표 1〉 프레임워크와 라이브러리의 비교

Langchain이란?

LLM으로 어플리케이션 만들 때
많이 사용하는 도구들의 모음

LangChain

Langchain이란?



LangChain을 사용하는 이유

LangChain 사용하는 이유

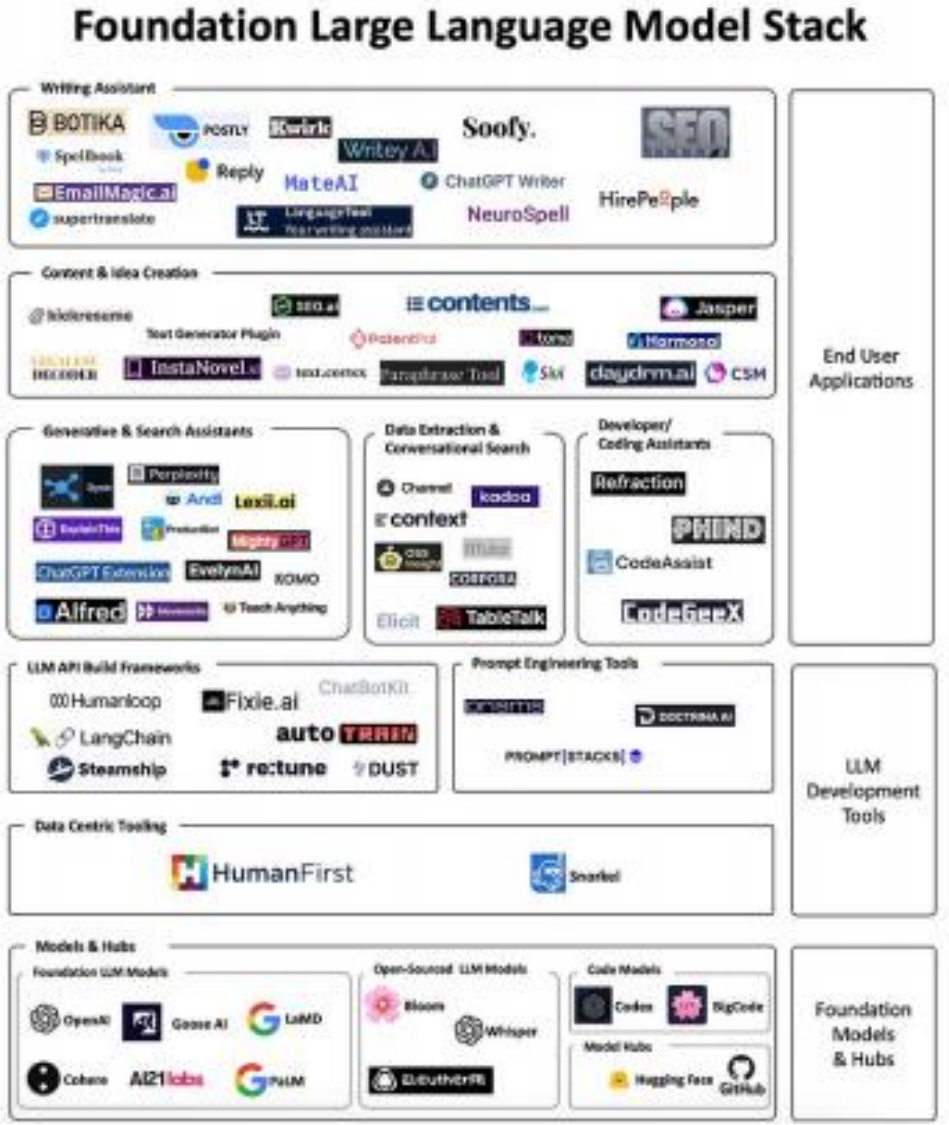
LLM 프레임워크를 사용하는 이유

1. 여러 도구들을 연결하고 관리하기 편리
2. 교체 편리
3. 추상화로 코드가 짧고 이해가 쉬움

LangChain을 사용하는 이유

커뮤니티가 가장 큼 (가장 많이 씀)

https://github.com/langchain-ai/langchain



https://cobusgreyling.medium.com/the-foundation-large-language-model-llm-tooling-landscape-8a849ebc7228

LangChain을 사용하는 이유

오늘의 주제로 돌아와서

Langchain으로 LLM 서비스 개발하기

1. OPEN AI API 살펴보기
2. LangChain으로 OpenAI API 활용하기
3. RAG 기반 챗봇 만들기

LLM 활용 방법

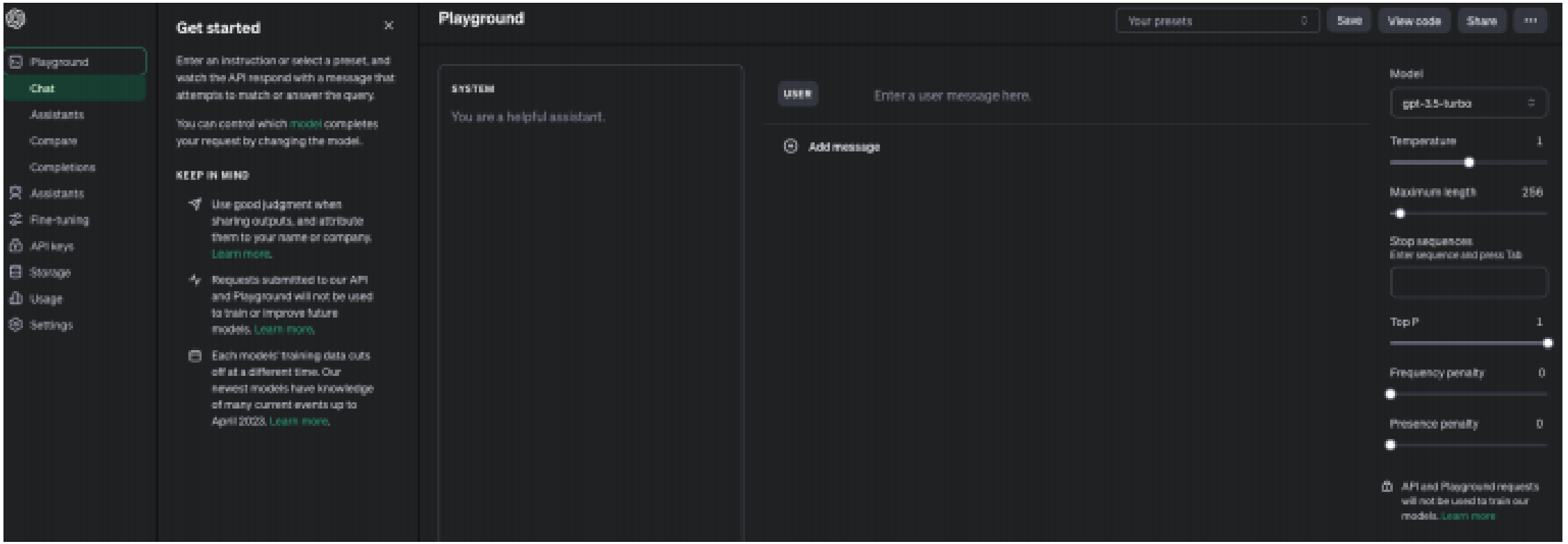
우리 회사 내규에 관한 챗봇을 만들려면?

1. 그냥 챗GPT를 사용하면 어떻게 될까?

LLM 활용 방법

1. OPEN AI API 살펴보기

OpenAI playground : https://platform.openai.com/playground/chat

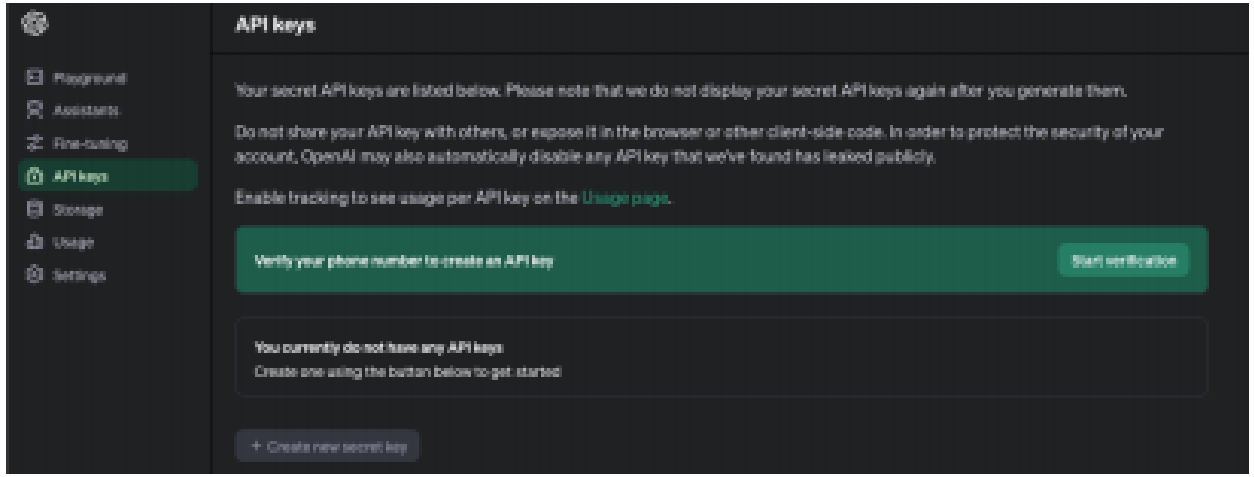


LLM 활용 방법

1. OPEN AI API 살펴보기

OpenAI : https://openai.com/product

24년도 API Key발급받기 : https://teddylee777.github.io/openai/openai-api-key/



무료 토큰이 만료되었고, 새로 결제하고 싶지 않다면? HuggingFace를 이용합시다.

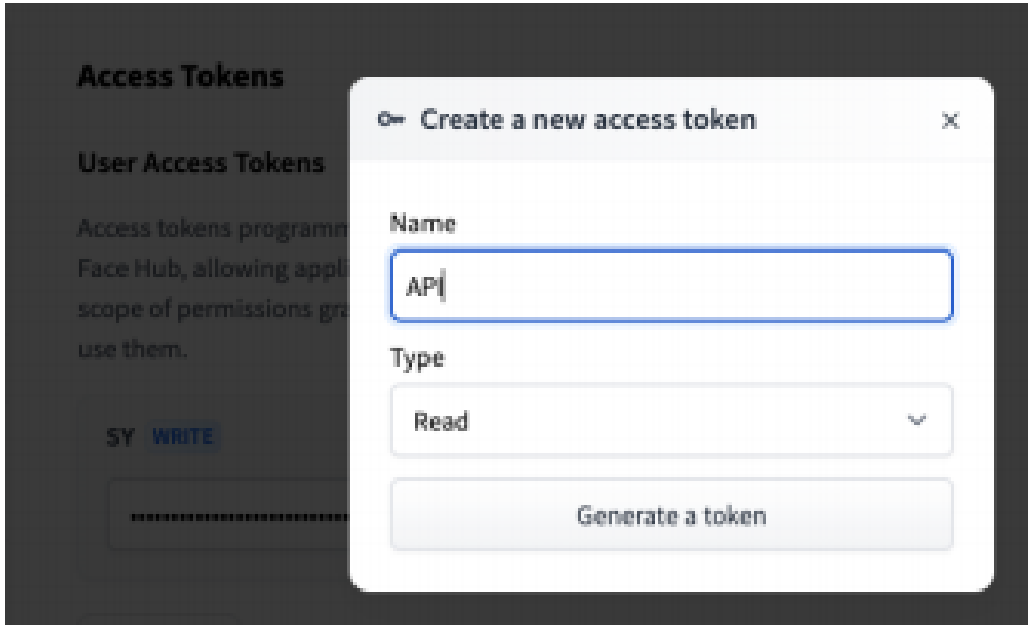
다만, 코드가 상이해지므로 원활한 수업을 위해 OpenAI API로 진행하는 것을 추천합니다!

LLM 활용 방법

1. HuggingFace API

HuggingFace : https://huggingface.co/docs/hub/security-tokenxs

24년도 API Key발급받기 : https://teddylee777.github.io/langchain/langchain-tutorial-02/



LLM 활용 방법

1. 정보 접근 제한
   ChatGPT(GPT3.5)는 2021년까지의 데이터를 학습한 LLM
   2022년부터의 정보에 대해서는 답변을 하지 못하거나, 거짓된 답변을 제공한다

2. 토큰 제한
   ChatGPT에서 제공하는 모델인 GPT3.5와 GPT4는 각각 4096, 8192토큰이라는
   입력 토큰 제한이 존재한다.

3. 환각현상(Hallucination)
   엉뚱한 대답을 하거나 거짓말을 하는 경우가 많다.

4. 보안이슈

https://www.youtube.com/watch?v=WWRCLzXxUgs&list=PLQIgLu3Wf-q_Ne8vv-ZXuJ4mztHJaQb_v

LLM 활용 방법

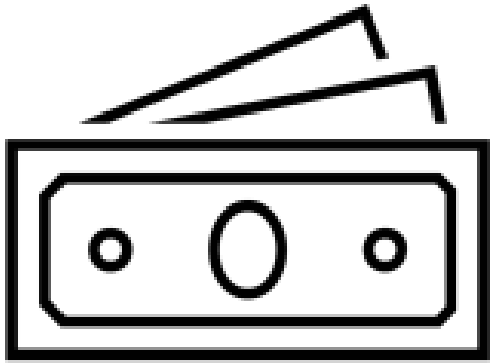
우리 회사 내규에 관한 챗봇을 만들려면?

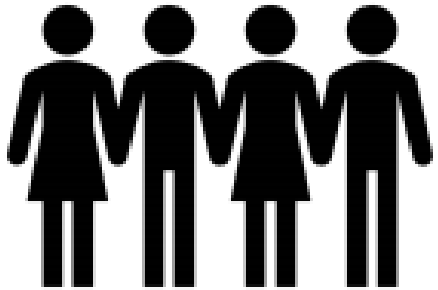
2. 자체 LLM을 만들어볼까?

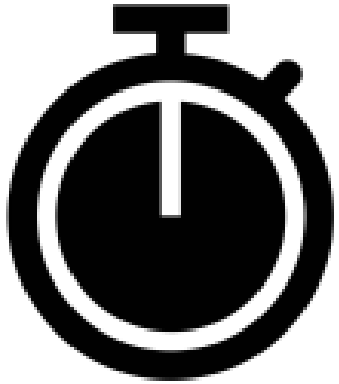
LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

2. 자체 LLM을 만들어볼까?







LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

3. 만들어진 LLM으로 파인튜닝(Fine-Tuning)을 시켜볼까?

파인 튜닝(Fine Tuning)

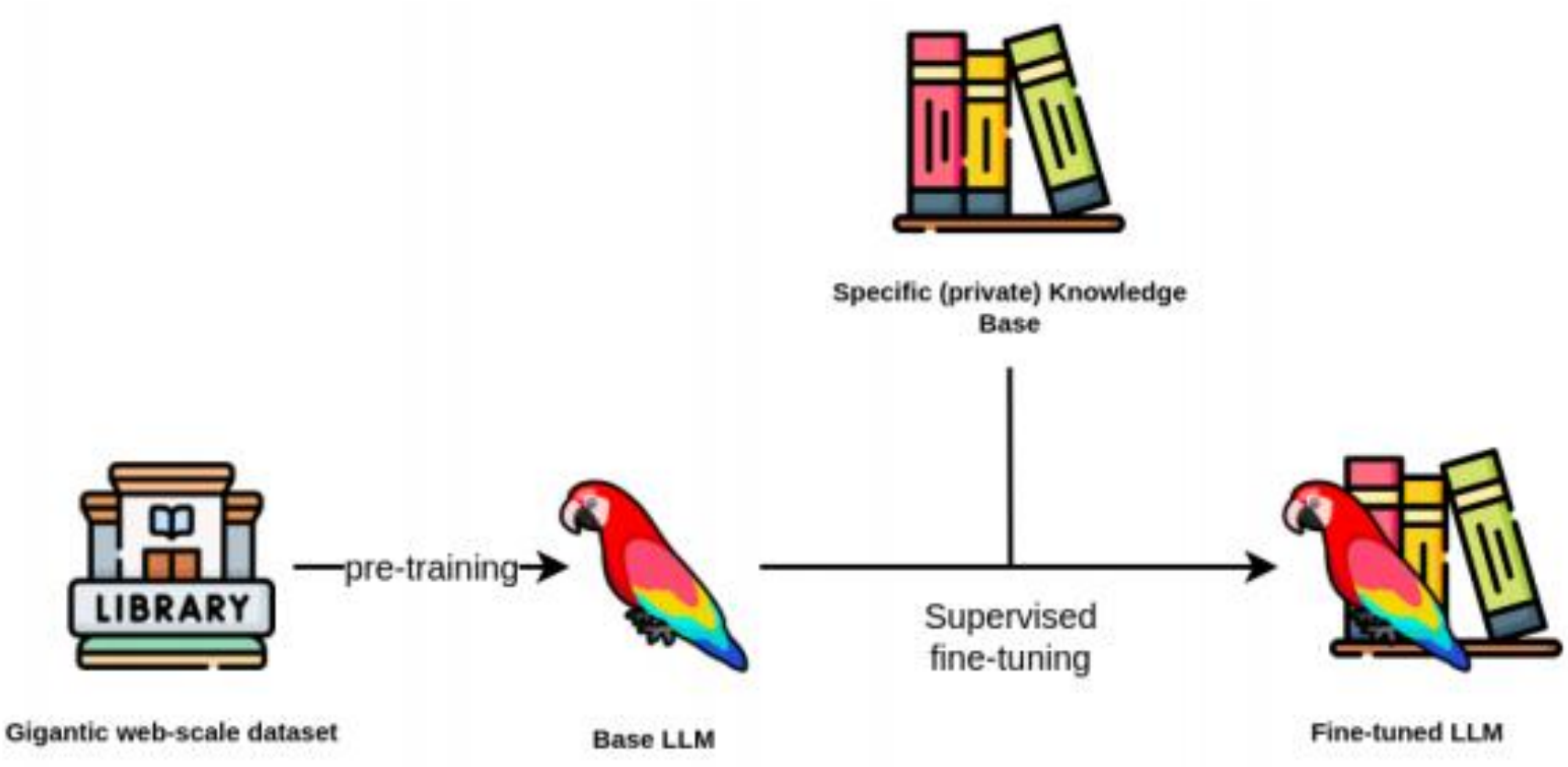
You

파인튜닝이란?

ChatGPT

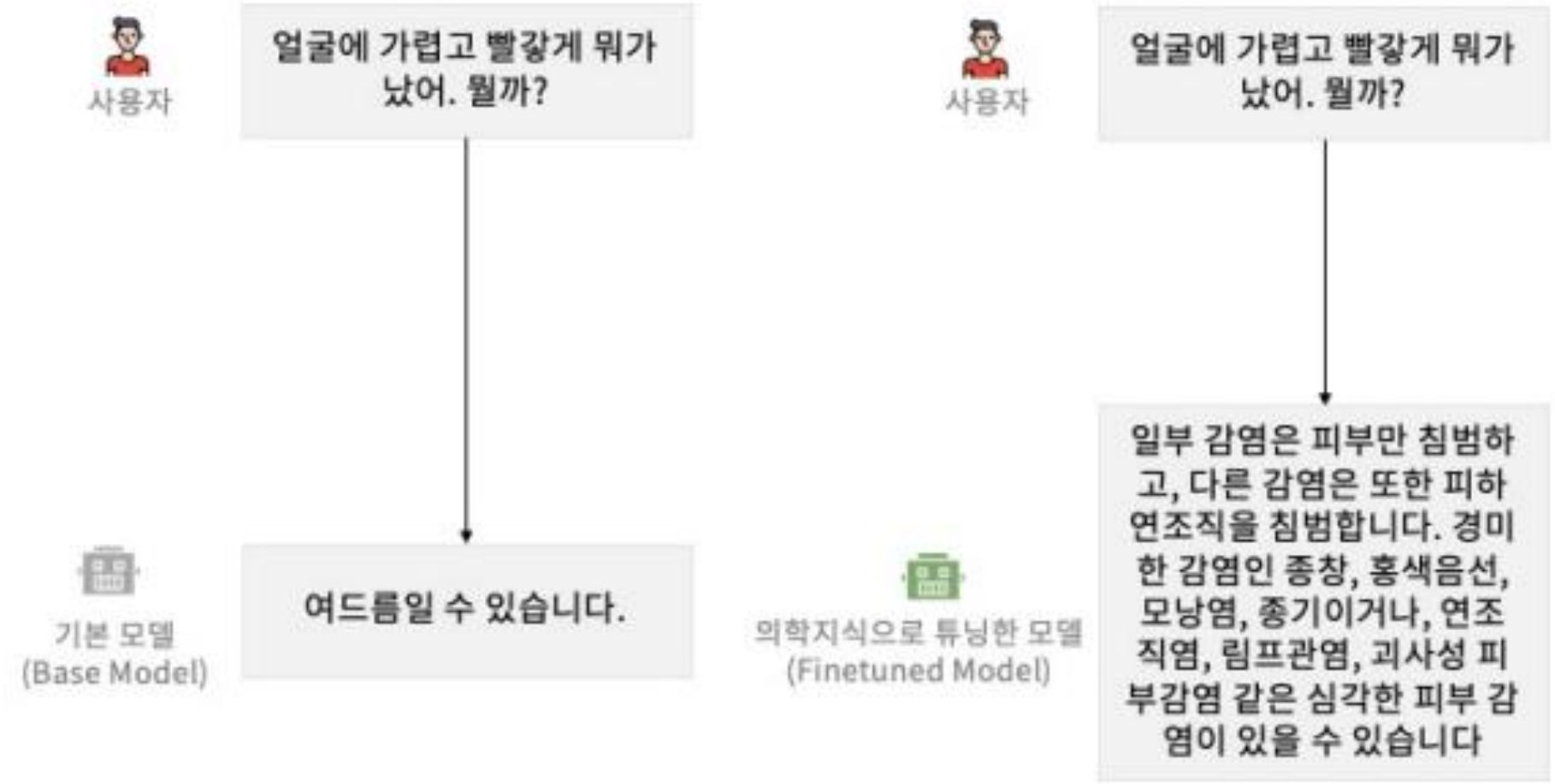
파인튜닝(Fine-tuning)은 이미 대량의 데이터로 사전 학습된 모델을 특정 작업이나 더 적은 양의 데이터로 최적화하여 성능을 향상시키는 과정을 말합니다. 머신러닝과 딥러닝 분야에서 널리 사용되며, 특히 자연어 처리(NLP)나 이미지 인식 같은 분야에서 효과적입니다. 파인튜닝을 통해 사전 학습된 모델은 새로운 데이터셋의 특성을 더 잘 반영할 수 있게 되며, 이는 작업의 정확도나 효율성을 크게 향상시킬 수 있습니다.

파인 튜닝(Fine Tuning)



https://medium.com/mantisnlp/supervised-fine-tuning-customizing-llms-a2c1edbf22c3

파인 튜닝(Fine Tuning)

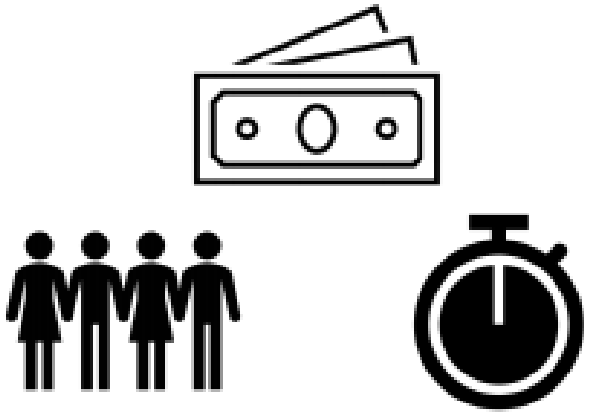


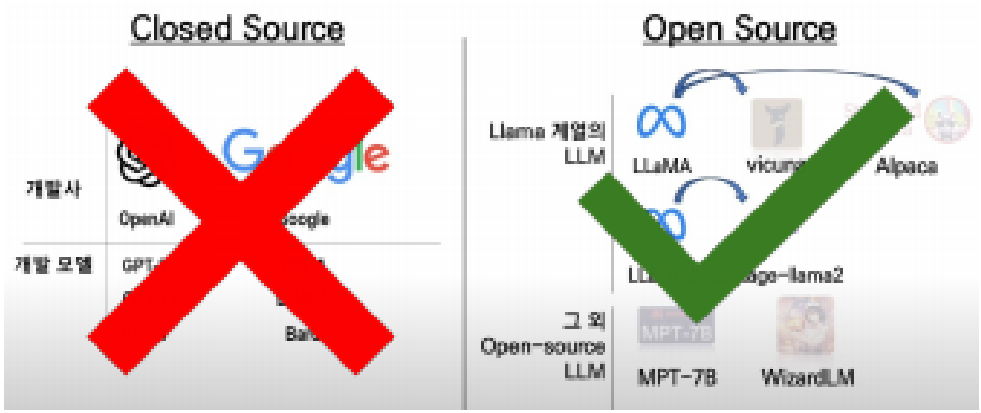
https://brunch.co.kr/@gentlepie/63

LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

3. 만들어진 LLM으로 파인튜닝(Fine-Tuning)을 시켜볼까?





https://brunch.co.kr/@gentlepie/63

LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

4. 추가 학습없이 챗봇을 만들 수는 없을까?

Few Shot Learning

RAG(Retrieval-Augmented Generation)

LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

4. 추가 학습없이 챗봇을 만들 수는 없을까?

Few Shot Learning

RAG(Retrieval-Augmented Generation)

RAG(Retrieval-Augmented Generation) 이란?

You

RAG(Retrieval-Augmented Generation)이란?

ChatGPT

RAG(Retrieval-Augmented Generation)은 자연어 처리(NLP) 분야에서 정보 검색 기법을 활용해 문서 생성을 개선하는 모델 구조를 말합니다. RAG는 크게 두 부분으로 구성되어 있는데, 첫 번째는 문서 검색 단계(retrieval component)이고, 두 번째는 생성 단계(generation component)입니다. 이 구조는 특정 질문이나 프롬프트에 대한 답변을 생성할 때, 관련 정보를 데이터베이스나 문서 집합에서 검색하여 그 정보를 바탕으로 더 정확하거나 정보에 기반한 답변을 생성할 수 있도록 설계되었습니다.

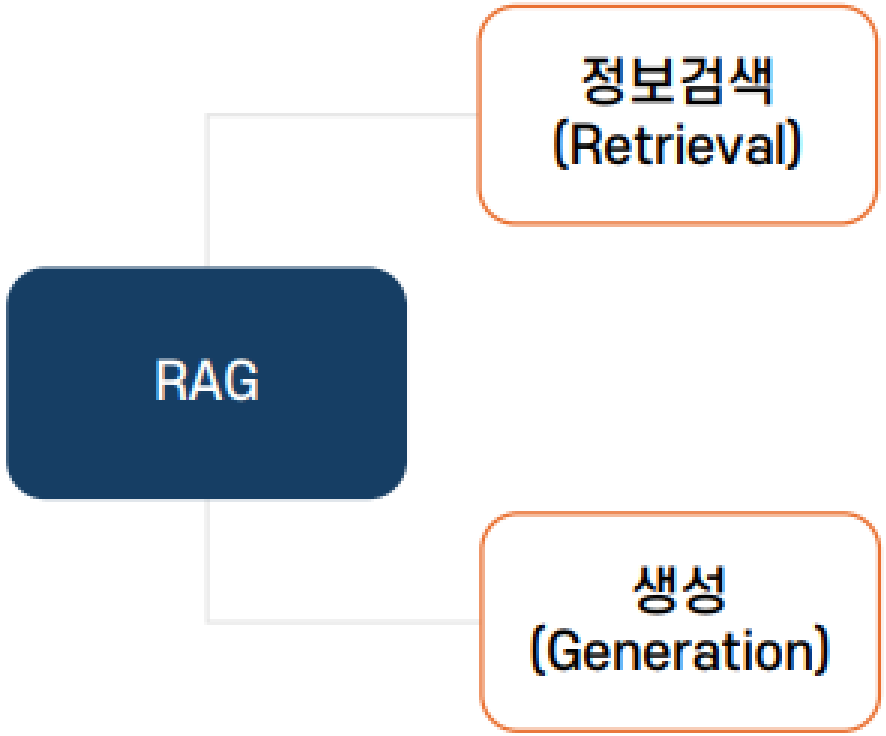
문서 검색 단계

- 이 단계에서는 주어진 질문이나 프롬프트에 가장 관련이 높은 정보를 포함하고 있는 문서를 데이터베이스에서 검색합니다. 검색된 문서들은 답변을 생성하는 데 필요한 배경 정보나 사실 등을 제공합니다.

생성 단계

- 검색된 문서들을 바탕으로, 생성 모델이 이 정보를 활용하여 보다 정확하고 상세한 답변을 생성합니다. 이때 사용되는 생성 모델은 흔히 Transformer 기반의 언어 모델이며, 검색된 문서의 내용을 참고하여 질문에 대한 답변을 생성합니다.

RAG(Retrieval-Augmented Generation) 이란?



1. 질문 : 사용자로부터 질문이 입력됩니다
2. 쿼리 (문서검색) : 모델은 대규모의 문서 데이터 베이스나 콘텐츠 저장소에서 질문과 관련된 문서나 정보를 검색합니다
3. 정보 검색 결과 : 검색결과 중에서 가장 관련성이 높은 문서와 사용자의 질문을 결합하여 LLM에 전달합니다.

4. 정보전달 : 사용자의 질문과 정보 검색결과가 모델에 전달됩니다.
이단 계에서 모델은 문서의 정보를 활용하여 질문에 대한 의미를 이해합니다.
5. 텍스트 생성: 전달받은 정보를 바탕으로 답변을 생성합니다.
-> LLM에 의해 처리되는 부분

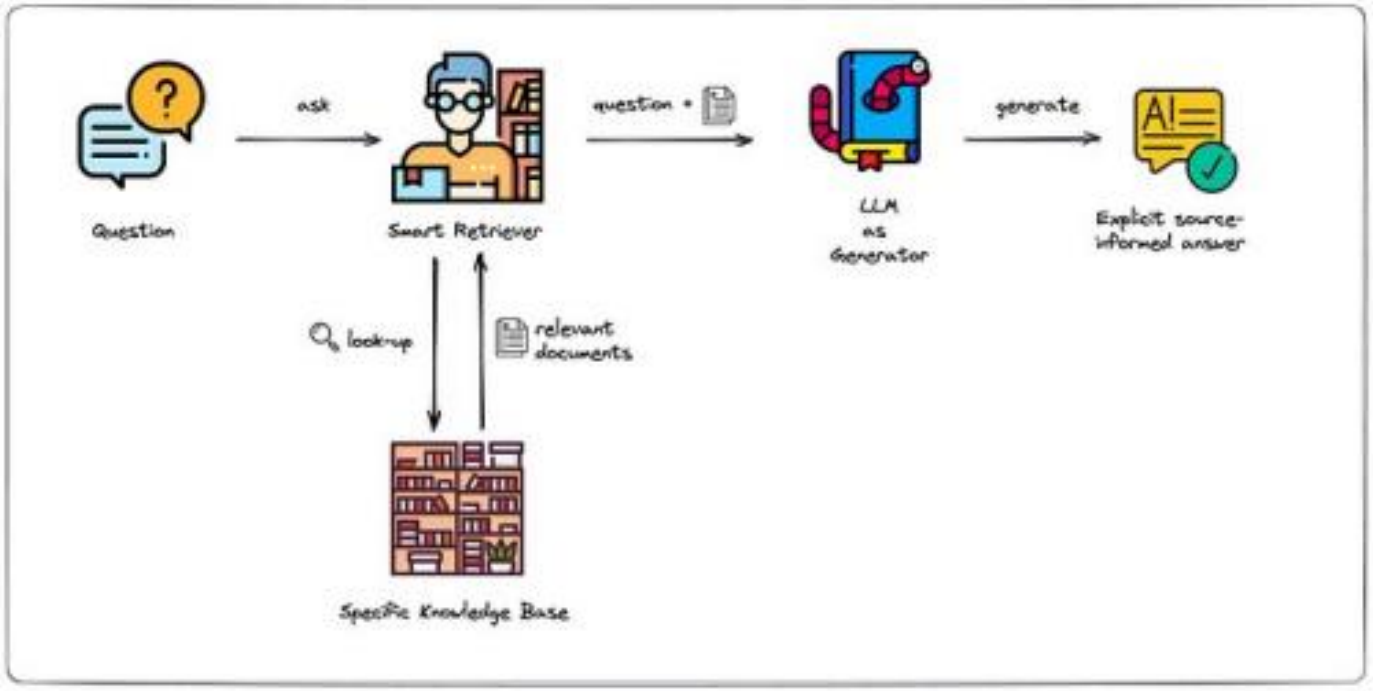
랭체인으로 LLM 기반의 AI 서비스 개발하기(서지영 저| 길벗)

RAG(Retrieval-Augmented Generation) 이란?

RAG(Retrieval-Augmented Generation)이란?

간단하게 말해, RAG는 큰 데이터 베이스나 인터넷과 같은 데이터 소스에서 필요한 데이터를 찾아내고, 그것을 기반으로 텍스트를 생성하는 기술입니다.

=> 이 방식은 LLM이 더 정확하고 신뢰할 수 있는 내용을 생성하도록 도와줍니다.



https://www.ml6.eu/blogpost/leveraging-llms-on-your-domain-specific-knowledge-base

RAG(Retrieval-Augmented Generation) 이란?

정보검색
(Retrieval)

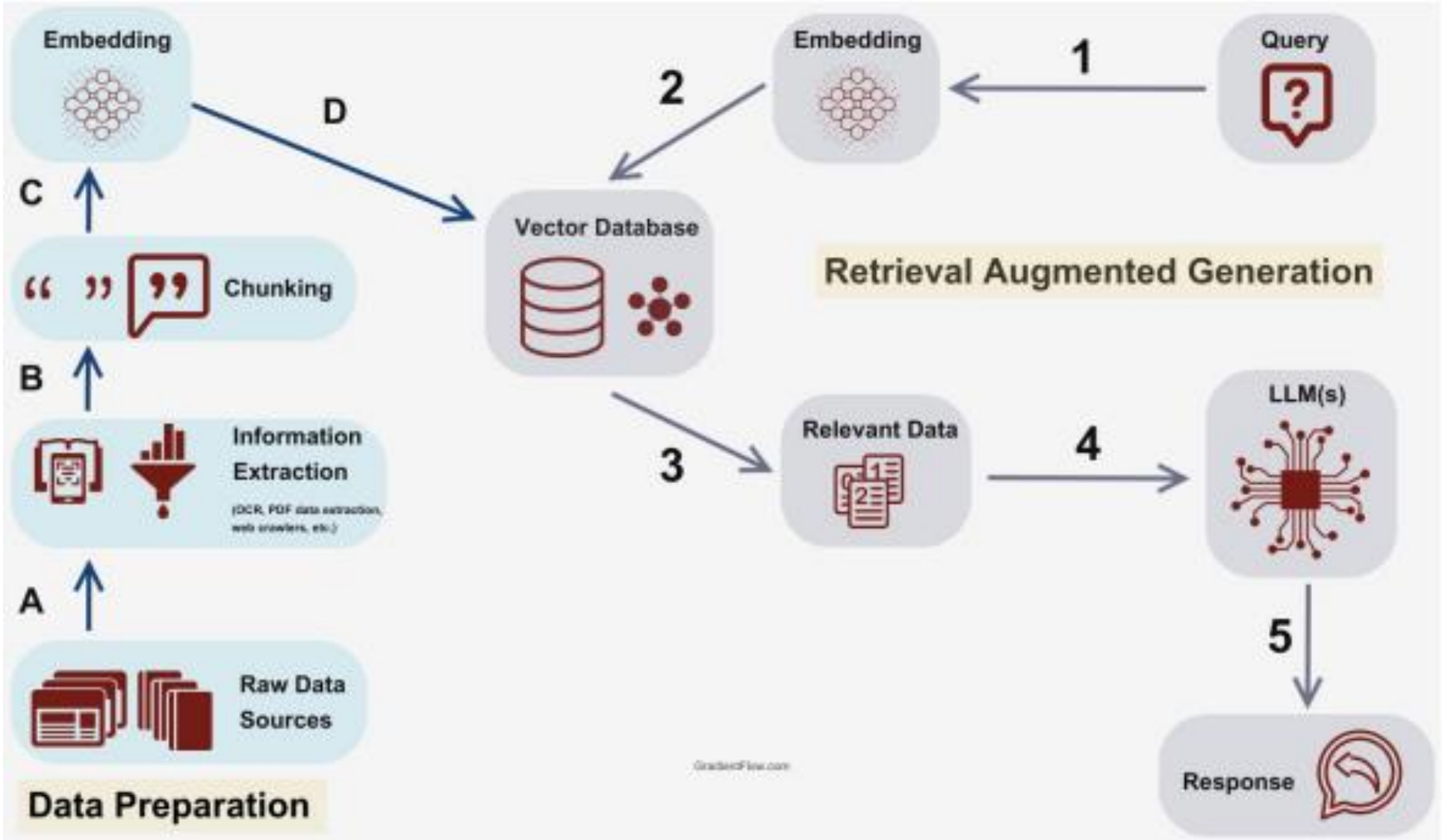
사용자 검색 : 대한민국의 수도는 어디인가요?

주어진 데이터 : 대한민국 위키백과
(https://ko.wikipedia.org/wiki/%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD)

...

수도는 서울특별시이다.

RAG(Retrieval-Augmented Generation) 구현 과정



https://gradientflow.substack.com/p/best-practices-in-retrieval-augmented

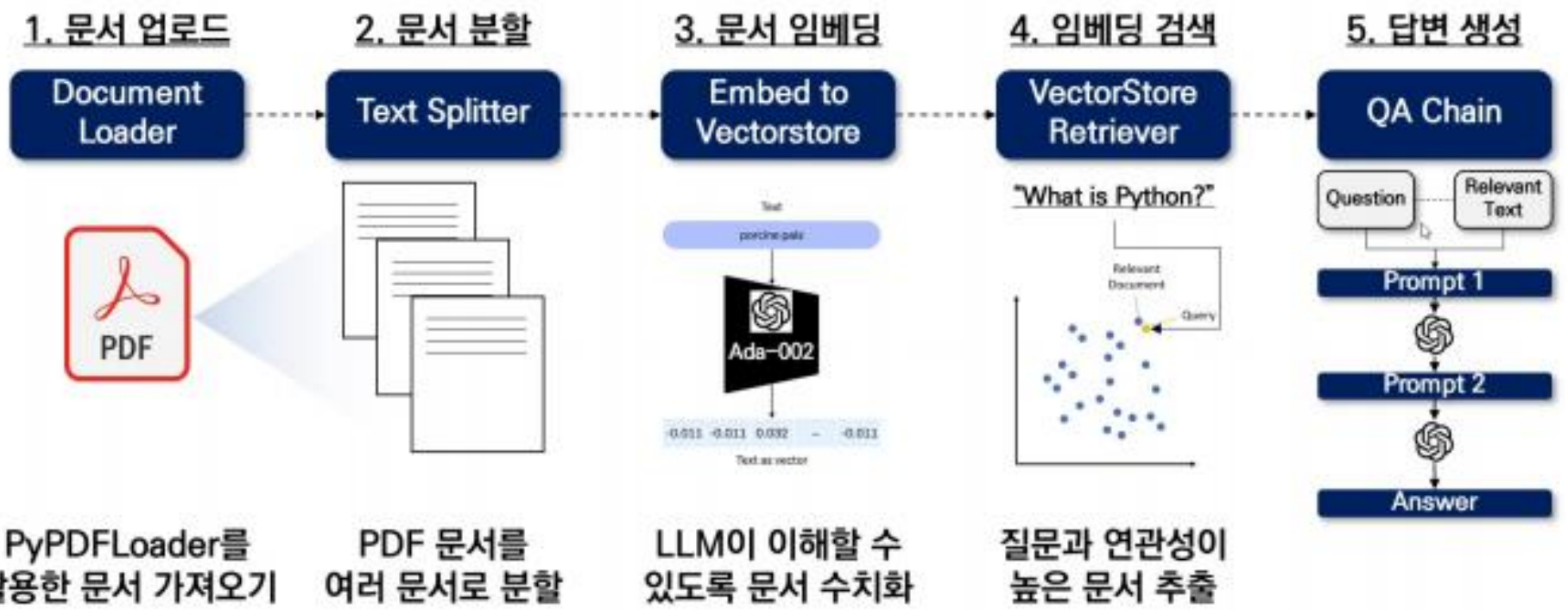
Q&A with RAG

Q&A with RAG

LangChain

예시: PDF 챗봇 구축

문서를 기반으로 챗봇을 구축할 경우, 아래와 같은 과정을 통해 대화가 가능하도록 합니다.



https://www.youtube.com/watch?v=WWRCLzXxUgs [모두의 Ai] LANGCHAIN 편ㄴ

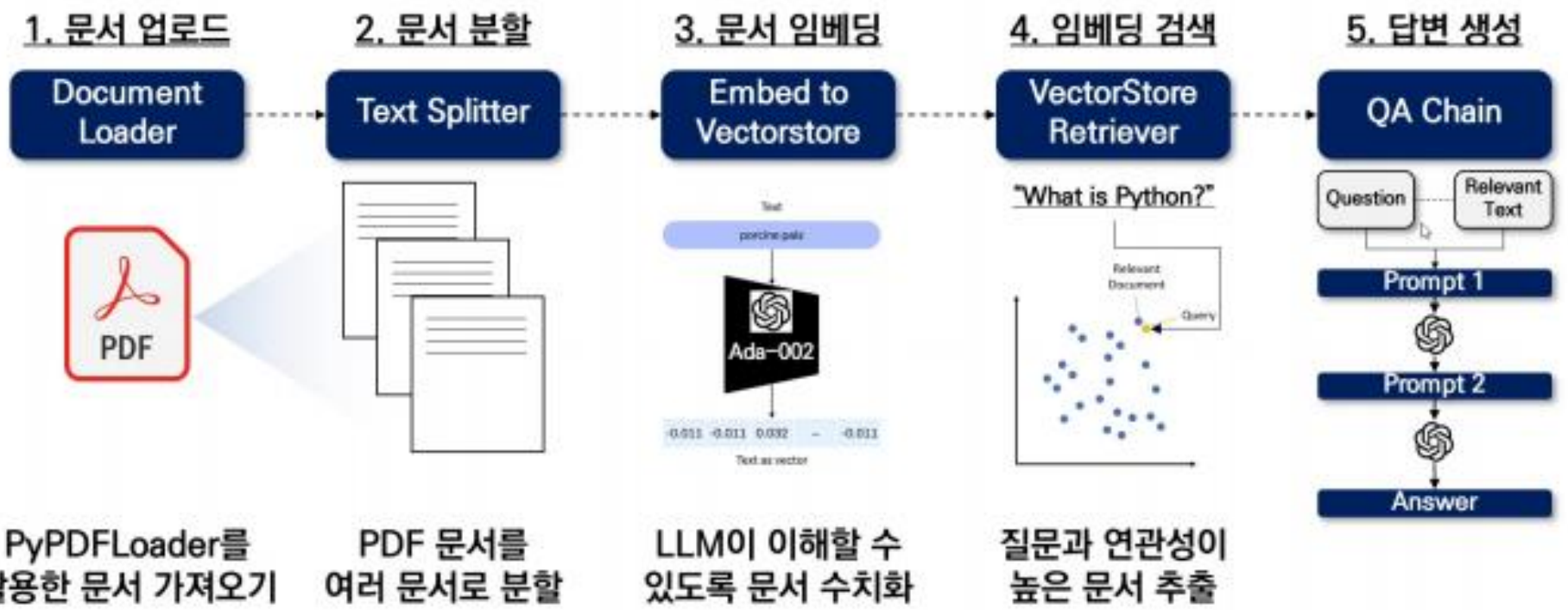
Q&A with RAG

Q&A with RAG

LangChain

예시: PDF 챗봇 구축

문서를 기반으로 챗봇을 구축할 경우, 아래와 같은 과정을 통해 대화가 가능하도록 합니다.



https://www.youtube.com/watch?v=WWRCLzXxUgs [모두의 Ai] LANGCHAIN 편ㄴ

참고자료

- https://platform.openai.com/docs/introduction
- https://python.langchain.com/docs/use_cases/question_answering/
- https://wikidocs.net/233342
- https://github.com/teddylee777/langchain-kr
- https://github.com/langchain-ai/langchain
- https://www.youtube.com/@AI-km1yn
- https://www.youtube.com/watch?v=EWKbZFqiCsE&t=2542s
- https://github.com/gilbutITbook/080413
- 랭체인으로 LLM 기반의 AI 서비스 개발하기(서지영 저 | 길벗)