

Langchain으로 LLM 서비스 개발하기

이소연 강사

Langchain으로 LLM 서비스 개발하기

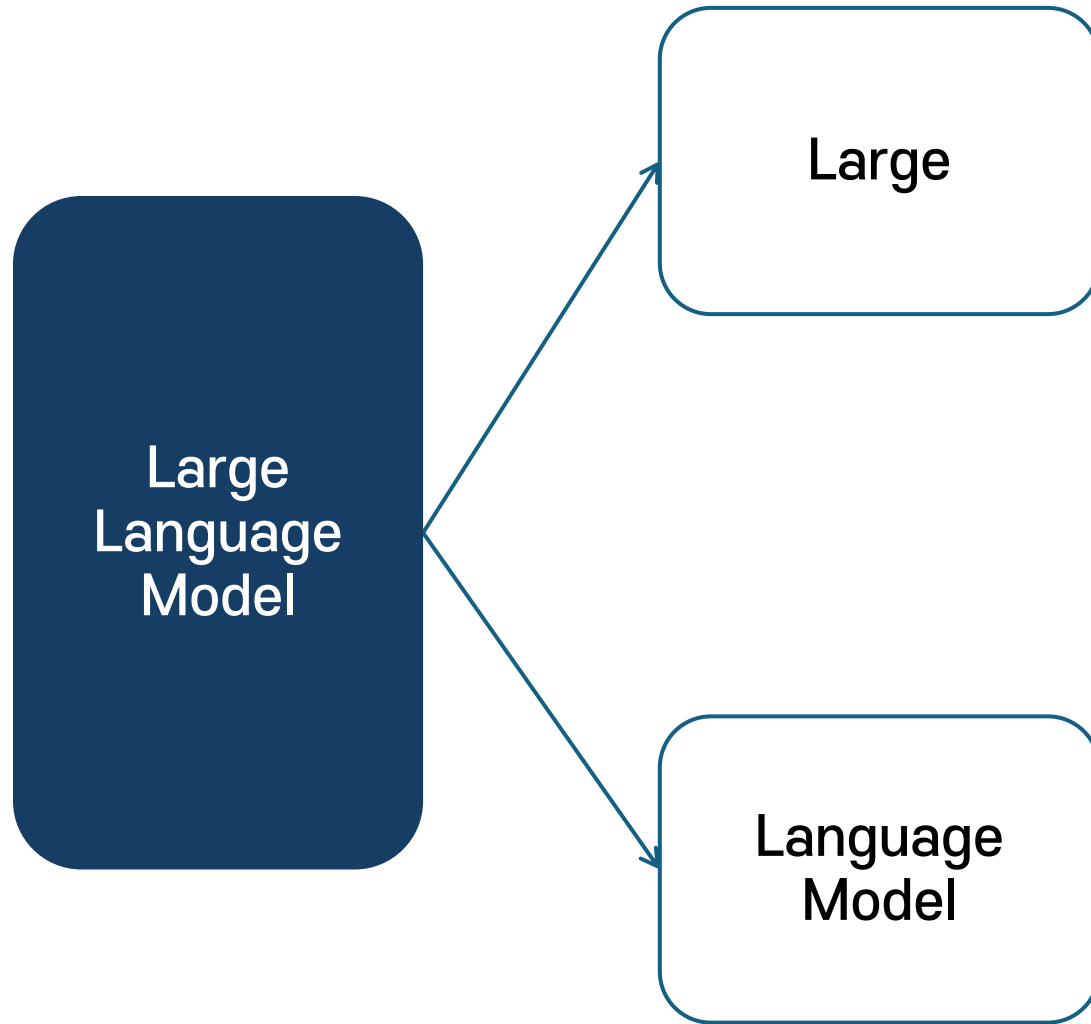


Large Language Model

The diagram consists of two rounded rectangular boxes. The left box is dark blue and contains the text 'Large Language Model'. The right box is orange and contains the text 'LangChain'. There are no connecting lines or arrows between the two boxes.

LangChain

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.

S = Where are we going

Previous words (Context) Word being predicted

$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$



딥 러닝을 이용한 |

딥 러닝을 이용한 부동산가격지수 예측
딥 러닝을 이용한 자연어 처리 입문
딥 러닝을 이용한 한국어 의존 구문 분석
딥 러닝을 이용한 개체명 인식
딥 러닝을 이용한 차량 번호판 검출
딥 러닝을 이용한 한국어 의미역 결정
딥 러닝을 이용한 한국어 형태소의 원형 복원 오류 수정
딥 러닝을 이용한
딥 러닝을 이용한 구문 분석

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.

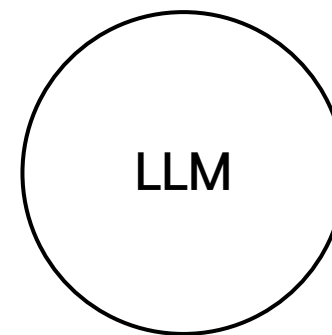
2. 어떤 방법론을 쓸 것인가?



Ngram, 확률



RNN, LSTM, Transformer



GPT-3

Large Language Model



학습된
데이터 양

** TB
(약 45TB 추정)

모델
파라미터 수

1750억개

처리할 수 있는
언어의 범위와
능력

GPT-3를 기준으로 보는 LLM의 크기

Large Language Model – Data Sources

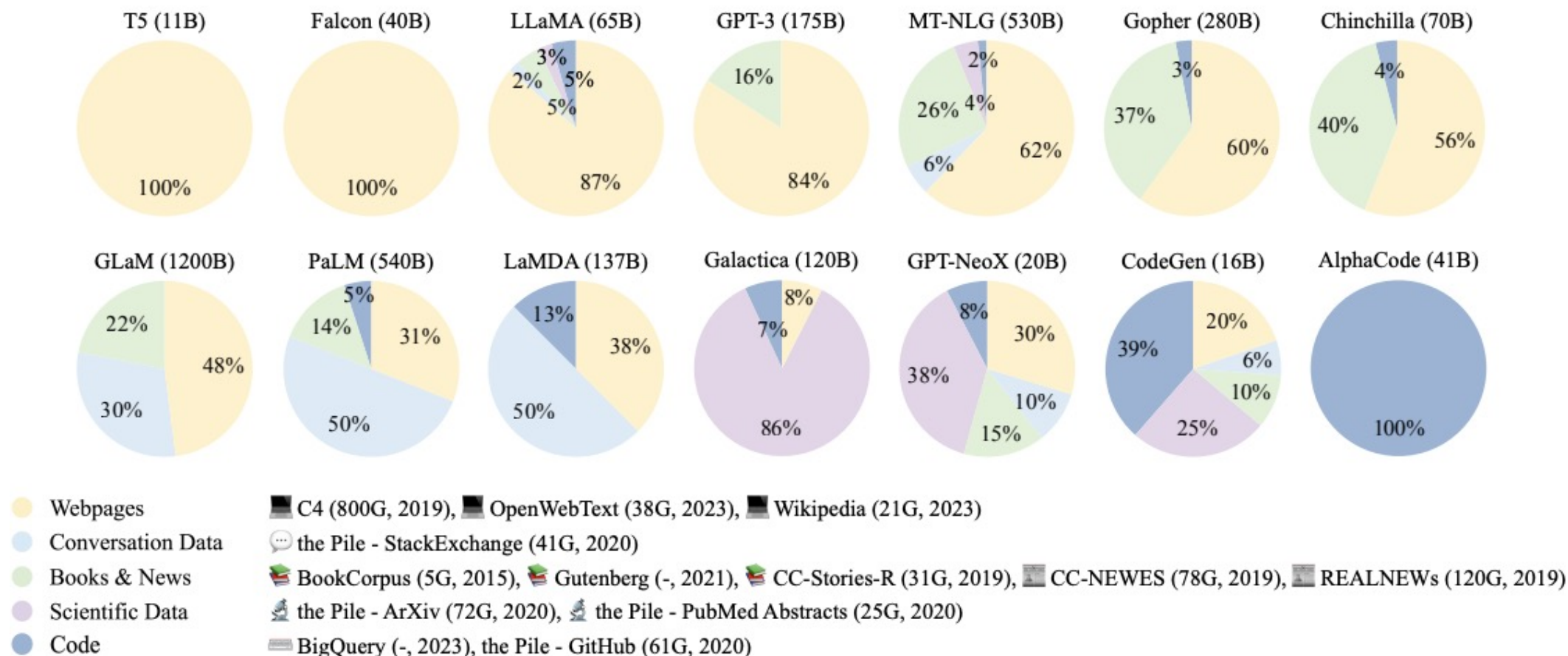
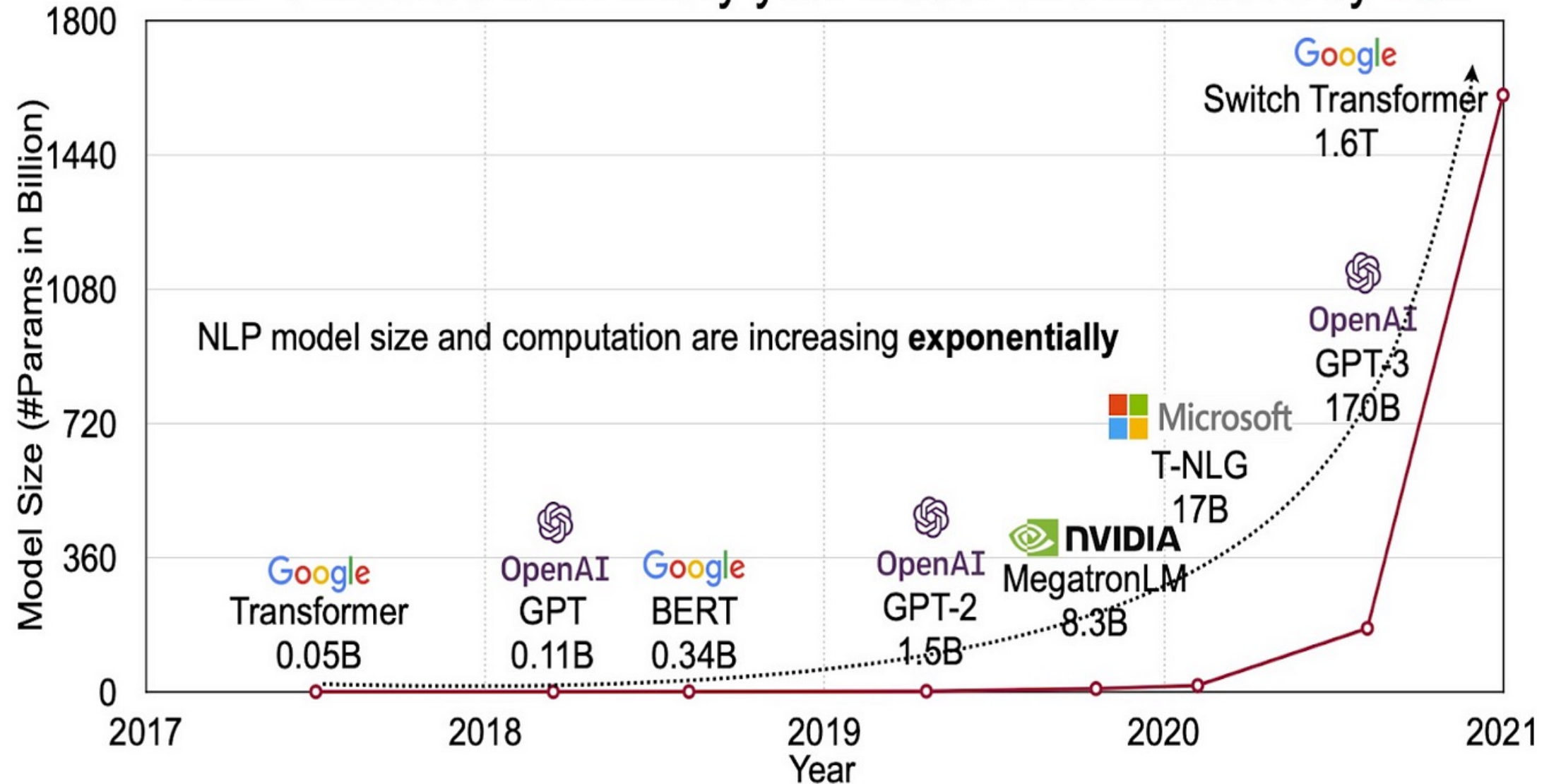


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.

NLP's Moore's Law: Every year model size increases by 10x



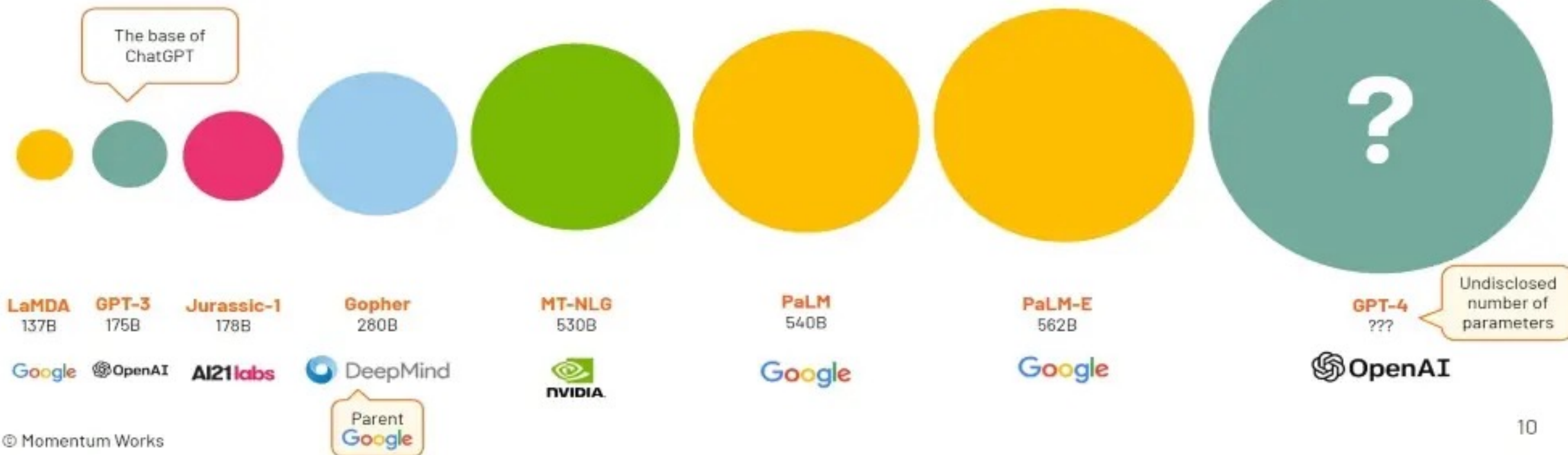
Large Language Models are becoming very large indeed



Small models (<= 100b parameters)



Large models (>100b parameters)



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

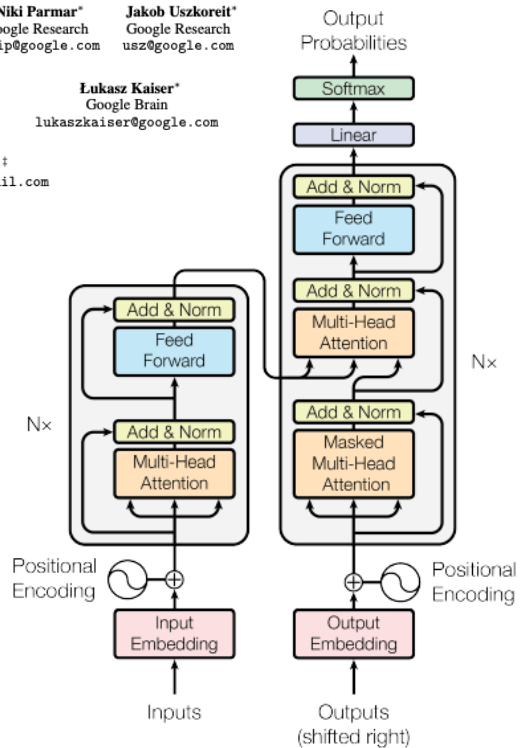
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

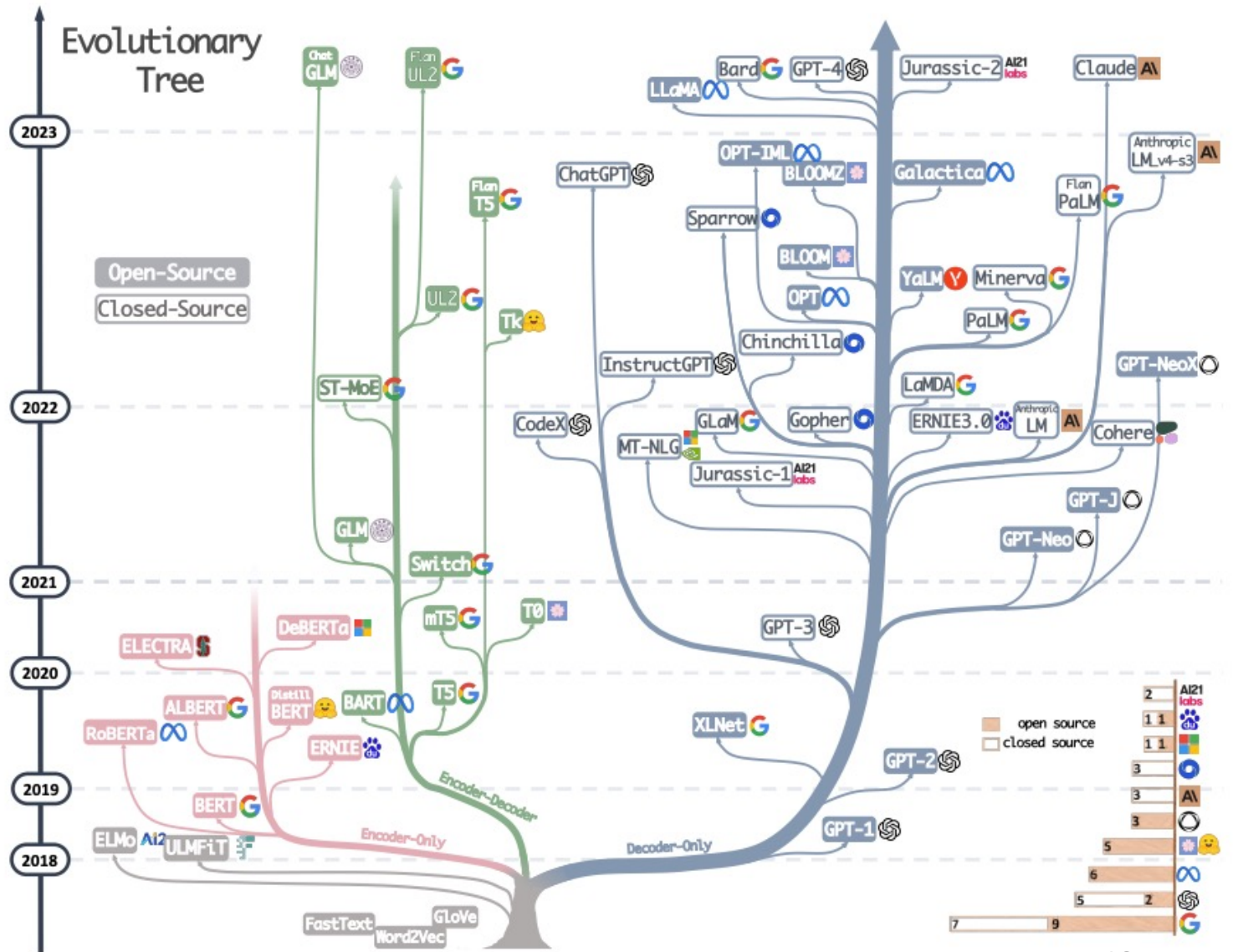
Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com



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

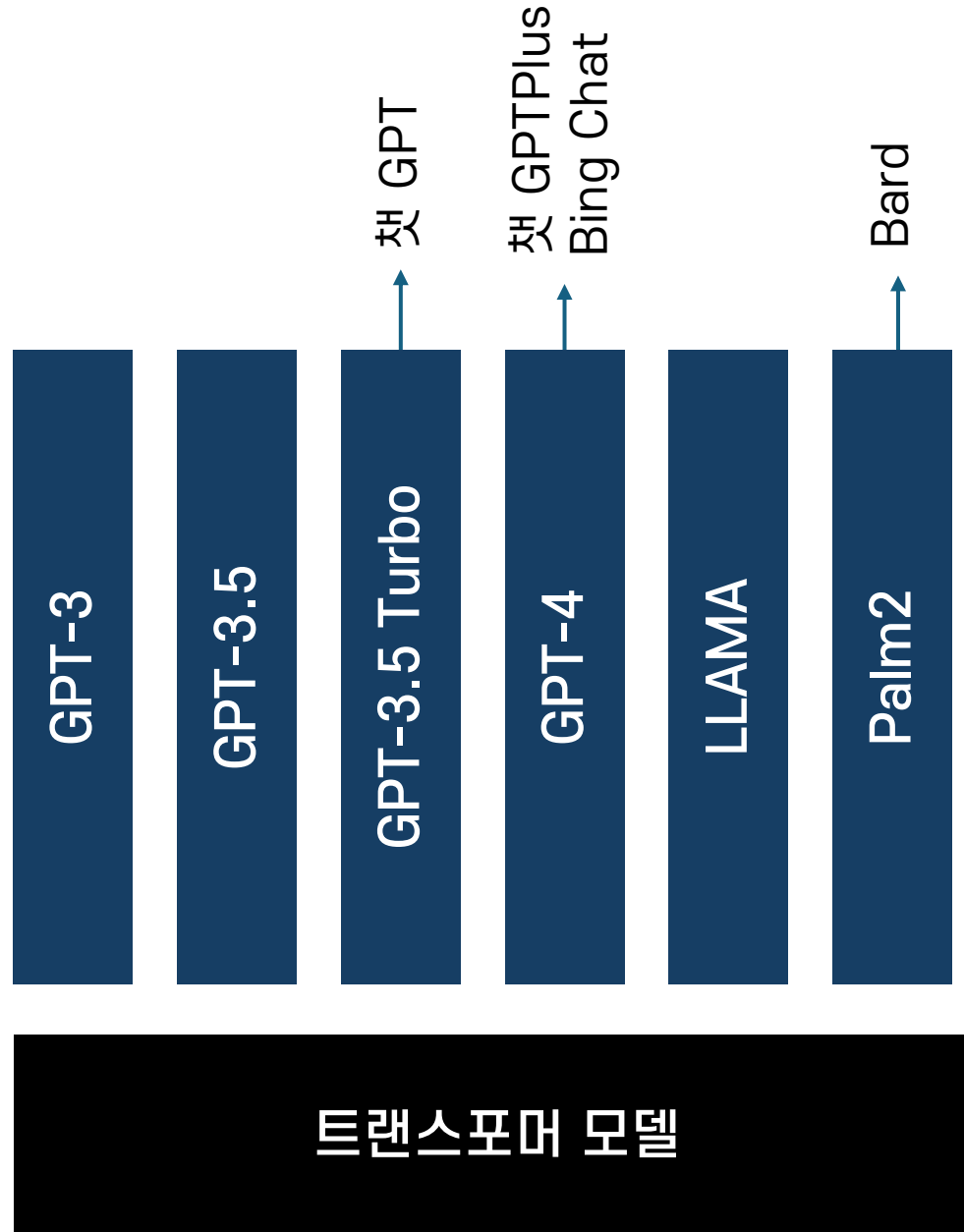


LLM의 핵심 구조

어플리케이션

LLM

LLM 기본 모델





LLM

LangChain의 구조

LLM

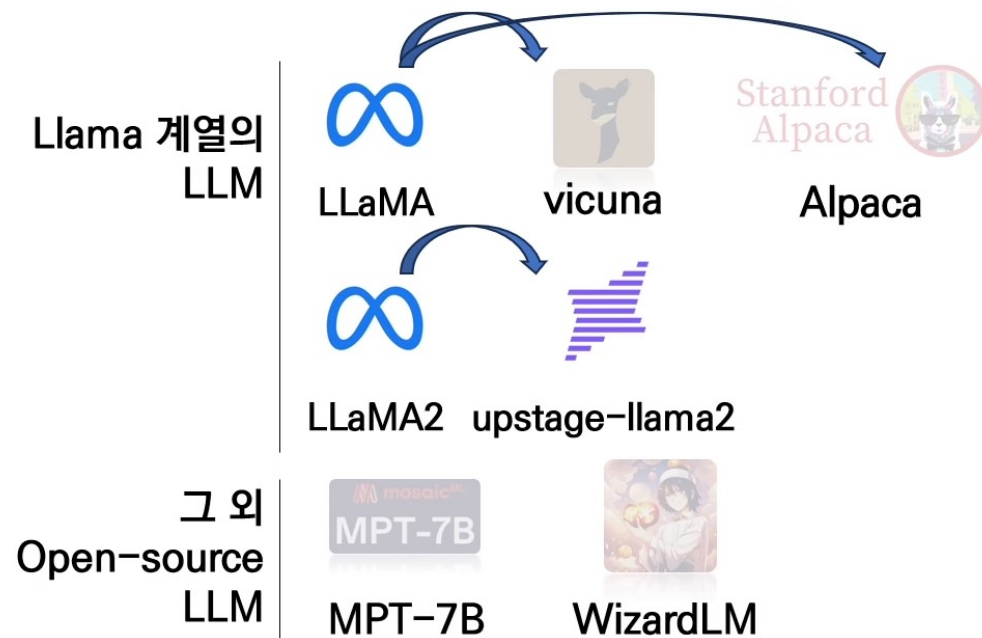
Closed Source

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

오늘의 주제로 돌아와서

Langchain으로 LLM 서비스 개발하기

LLM 만들기 ❌

LLM 활용하기 🟢

Langchain으로 LLM 서비스 개발하기



Large Language Model



LangChain

Langchain이란?



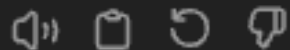
You

랭체인이 뭐야?



ChatGPT

랭체인(LangChain)은 대형 언어 모델(LLM)을 사용하여 애플리케이션 생성을 단순화하도록 설계된 프레임워크입니다. 이는 언어 모델 통합 프레임워크로서, 문서 분석 및 요약, 챗봇, 코드 분석을 포함한 일반적인 언어 모델의 사용 사례를 지원합니다. LangChain은 언어 모델로 구동되는 애플리케이션 개발을 위한 프레임워크로, 언어 모델을 다양한 데이터 소스와 연결하고, 제공된 내용을 바탕으로 답변하는 방식에 대해 추론하는 데 도움을 줍니다.



Langchain이란?

대형 언어모델(LLM) 어플리케이션 개발 프레임워크



프레임워크란?

= 관련 도구 모음



You

프레임 워크가 뭐야?



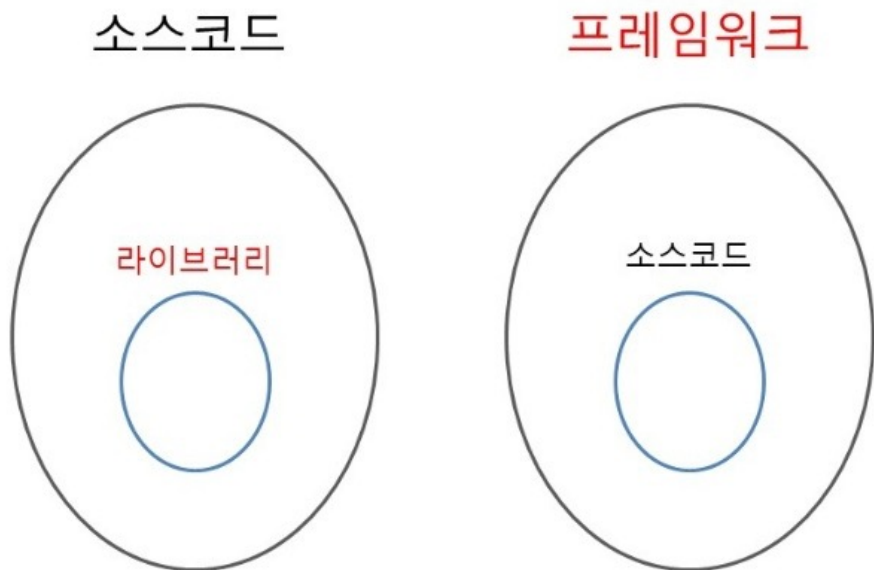
ChatGPT

프레임워크는 특정 언어나 기술을 사용하여 애플리케이션을 개발할 때 기반으로 삼을 수 있는 기본 구조나 "뼈대"입니다. 개발자들이 소프트웨어 제품을 더 빠르고 효율적으로 개발할 수 있도록 도와주며, 일반적인 프로그래밍 작업을 위한 라이브러리, 도구 세트, API 등을 포함합니다. 프레임워크는 반복적인 코드 작성을 줄여주고, 개발 프로세스를 표준화하며, 개발자가 애플리케이션의 핵심 기능에 집중할 수 있도록 돕습니다.

예를 들어, 웹 개발에서는 Ruby on Rails, Django, Spring과 같은 프레임워크가 자주 사용됩니다. 각 프레임워크는 특정 프로그래밍 언어로 작성되었으며, 데이터베이스 연결, 사용자 인증, 템플릿 처리 등과 같은 공통적인 웹 개발 작업을 처리하는 데 필요한 기능과 구조를 제공합니다.

프레임워크란?

= 관련 도구 모음



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

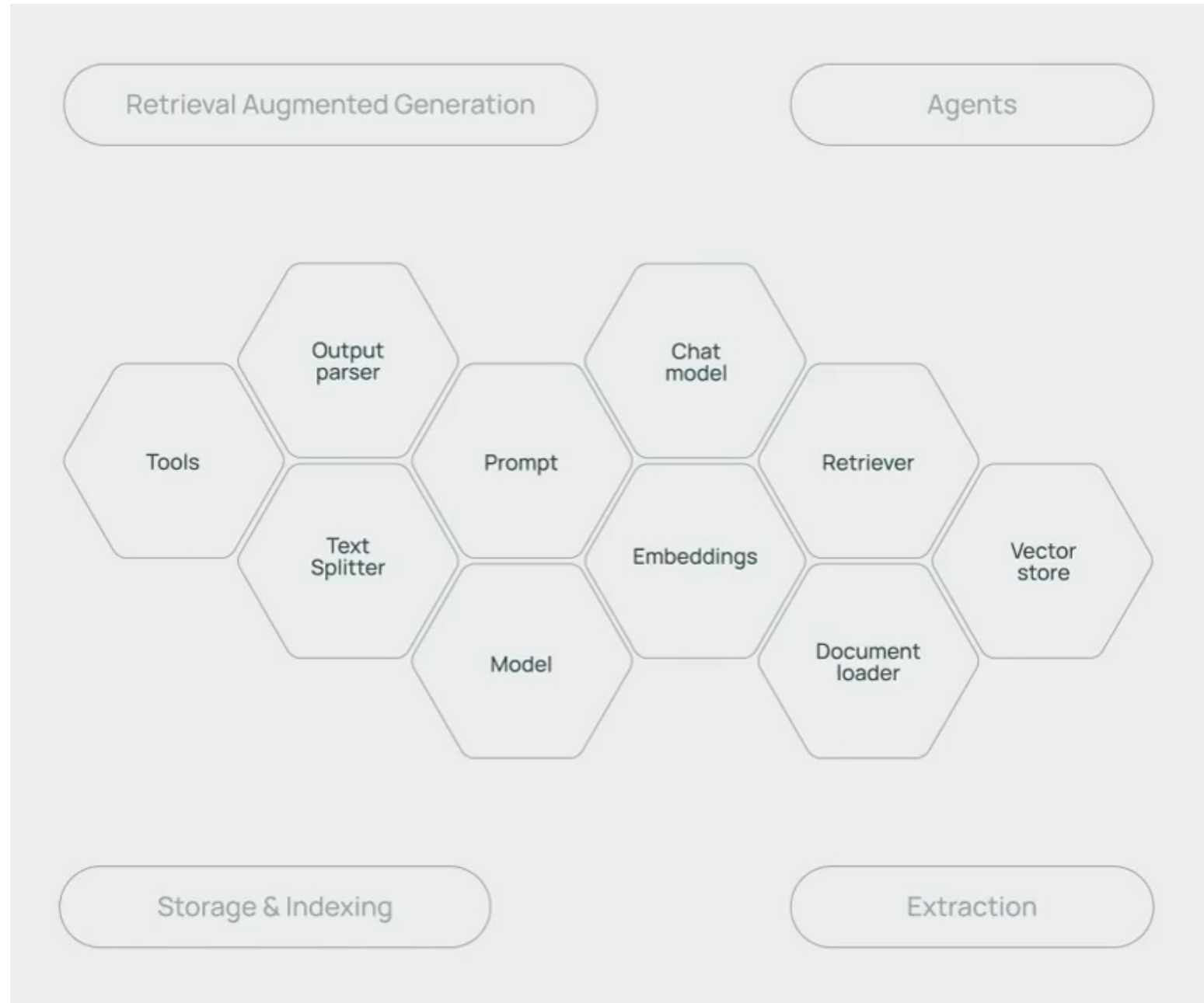
〈표 1〉 프레임워크와 라이브러리의 비교

Langchain이란?

LLM으로 어플리케이션 만들 때
많이 사용하는 도구들의 모음



Langchain이란?



LangChain 사용하는 이유

LLM 프레임워크를 사용하는 이유

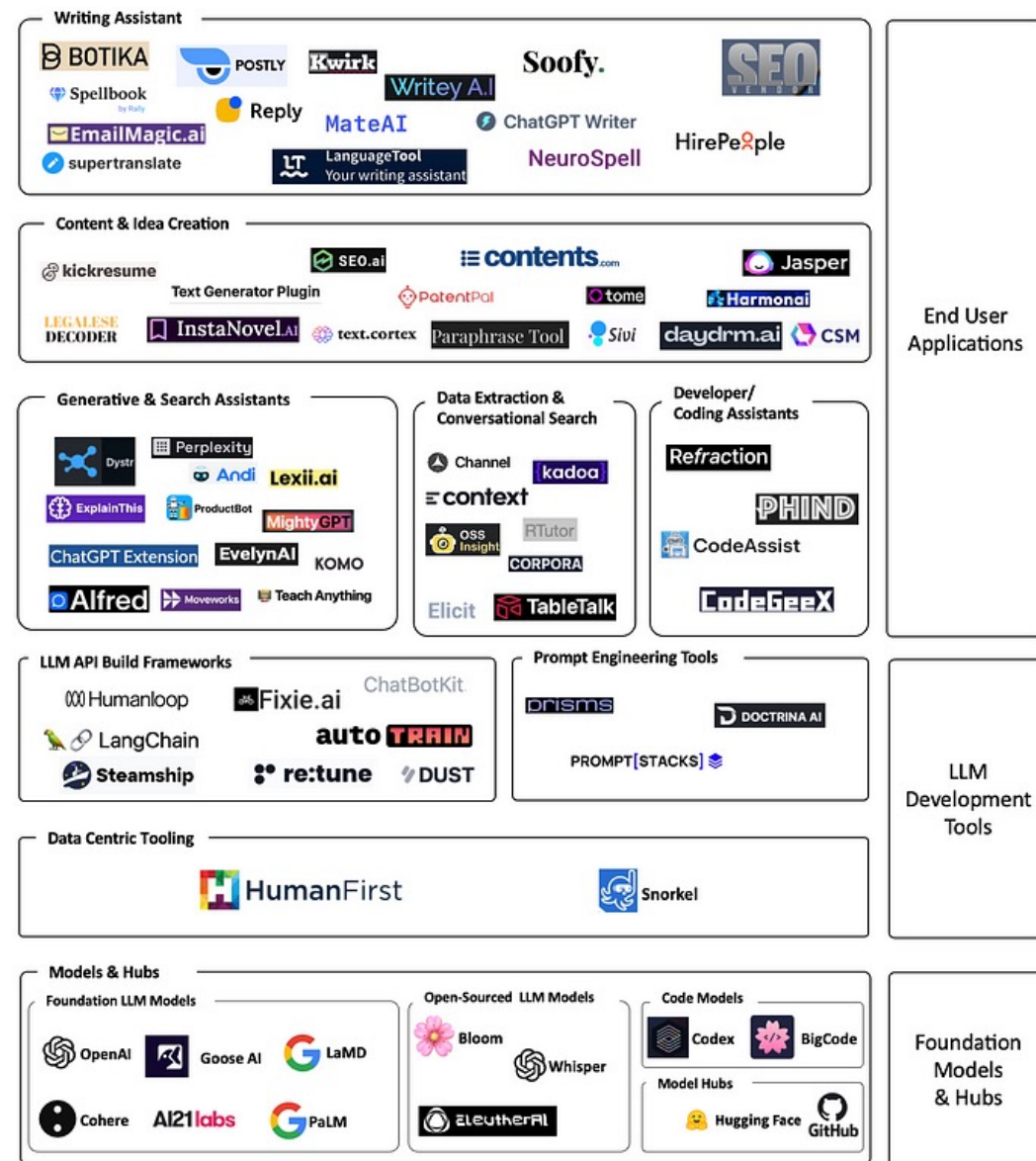
1. 여러 도구들을 연결하고 관리하기 편리
2. 교체 편리
3. 추상화로 코드가 짧고 이해가 쉬움

LangChain을 사용하는 이유

커뮤니티가 가장 큼 (가장 많이 씬)

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

Foundation Large Language Model Stack



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

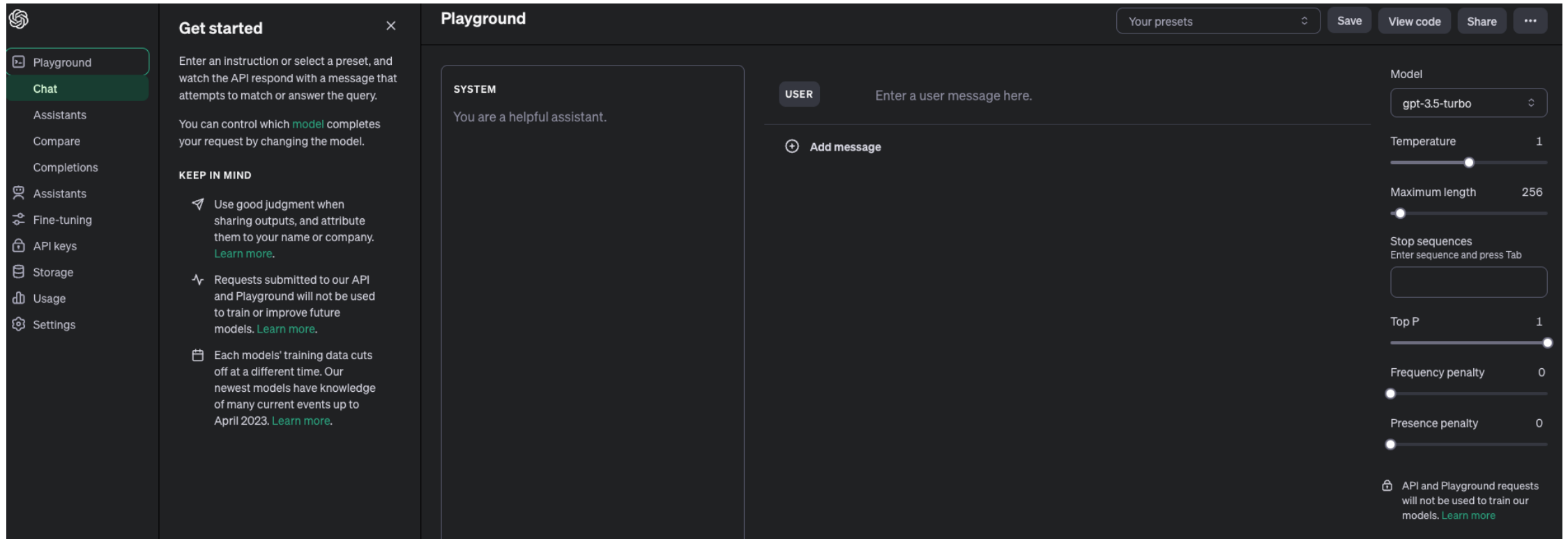
우리 회사 내규에 관한 챗봇을 만들려면?

1. 그냥 챗GPT를 사용하면 어떻게 될까?



1. OPEN AI API 살펴보기

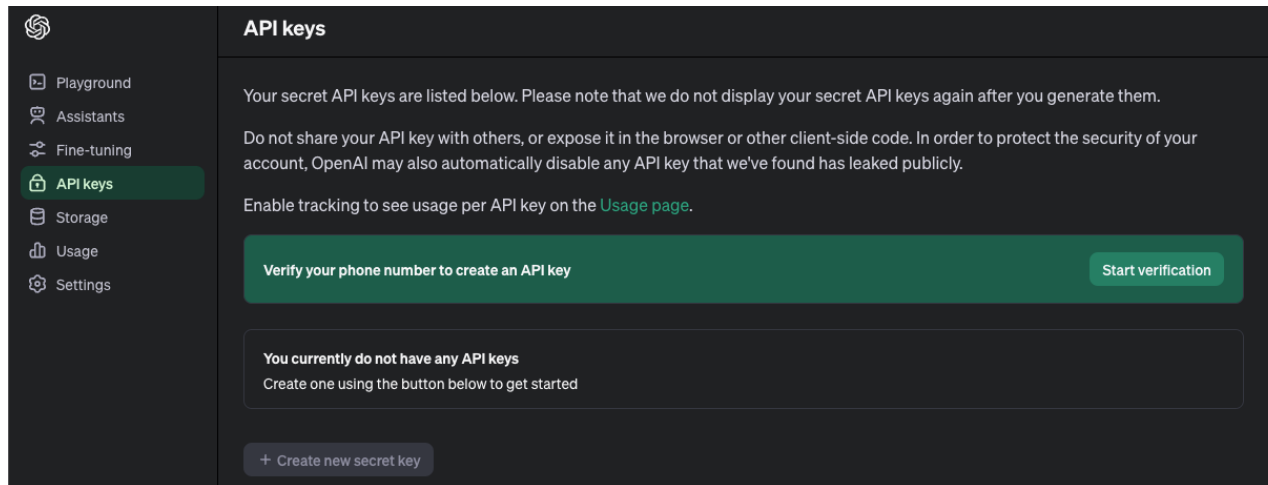
OpenAI playground : <https://platform.openai.com/playground/chat>



1. OPEN AI API 살펴보기

OpenAI : <https://openai.com/product>

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



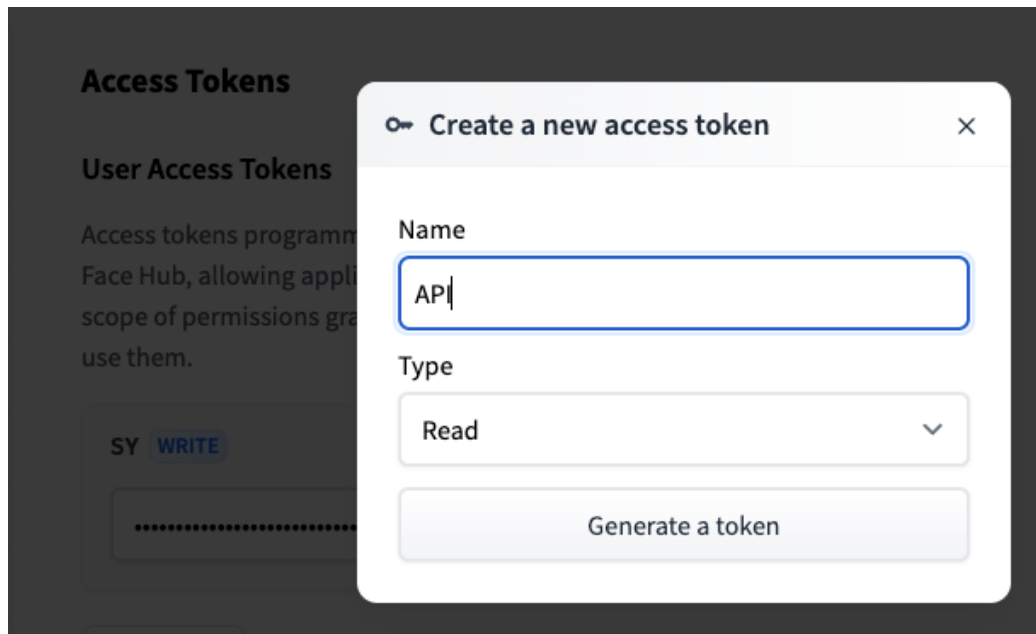
무료 토큰이 만료되었고, 새로 결제하고 싶지 않다면? HuggingFace를 이용합시다.

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

1. HuggingFace API

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

24년도 API Key 발급받기 : <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

LLM 활용 방법

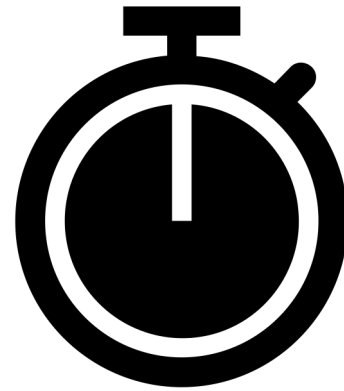
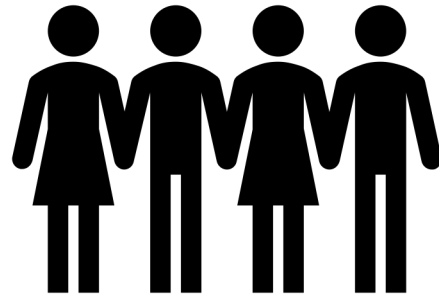
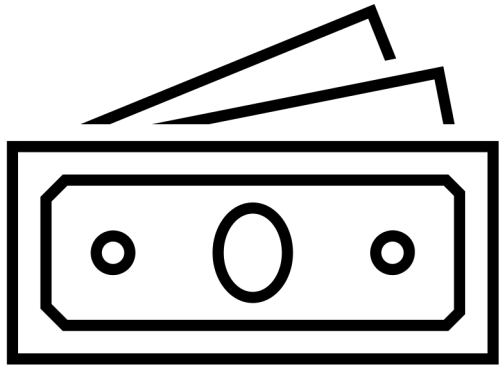
우리 회사 내규에 관한 챗봇을 만들려면?

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?

파인튜닝(Fine Tuning)



You

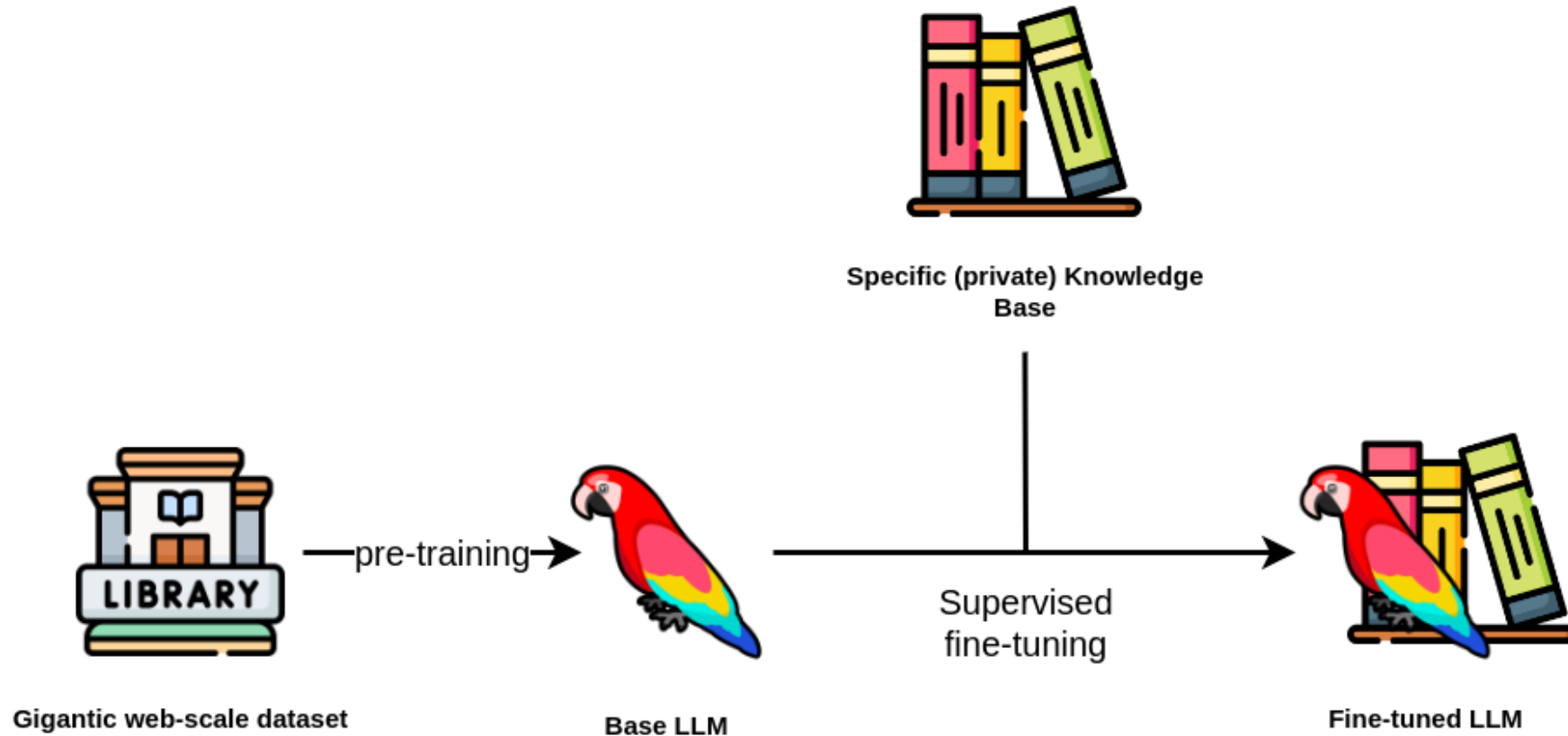
파인튜닝이란?



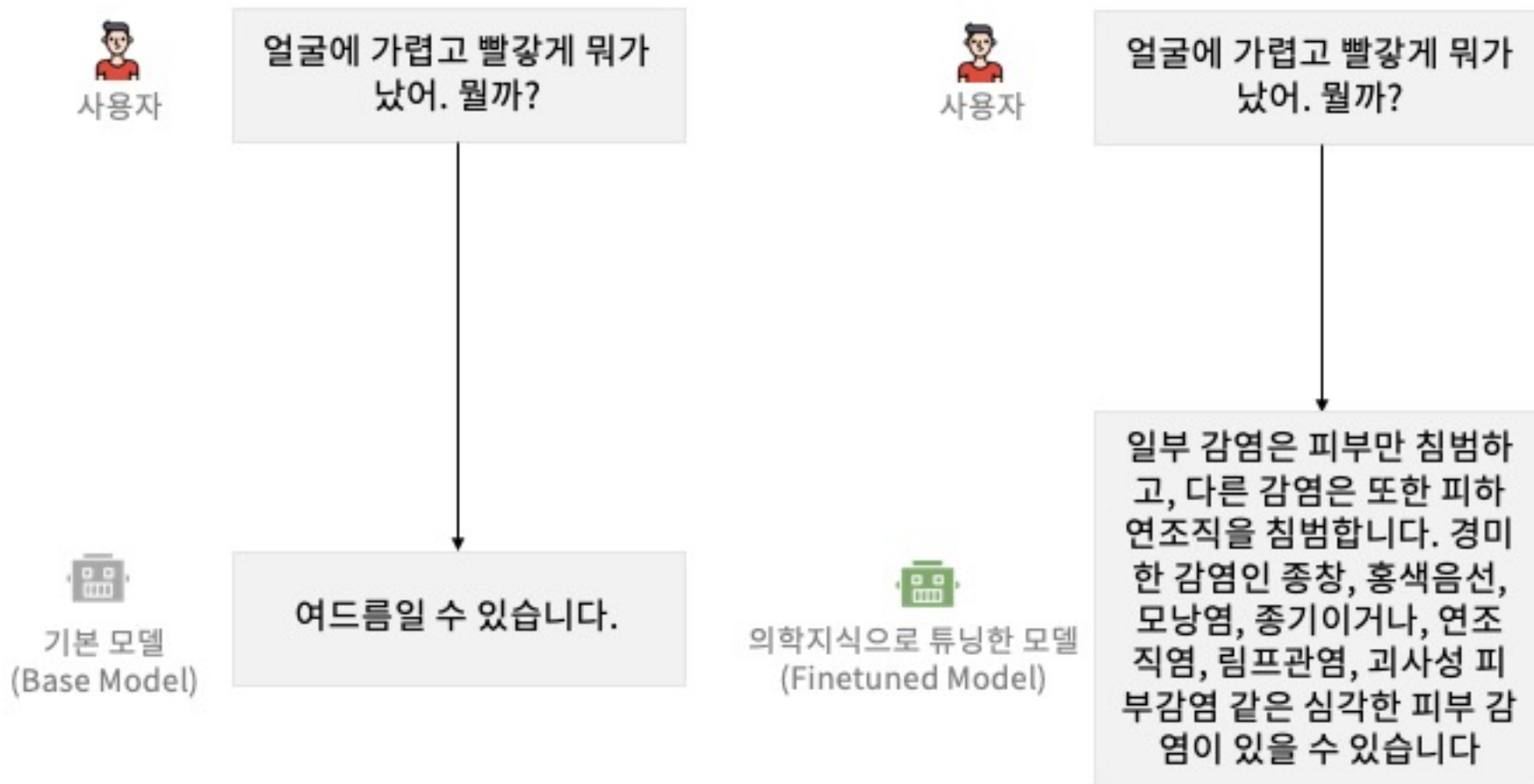
ChatGPT

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.

파인튜닝(Fine Tuning)



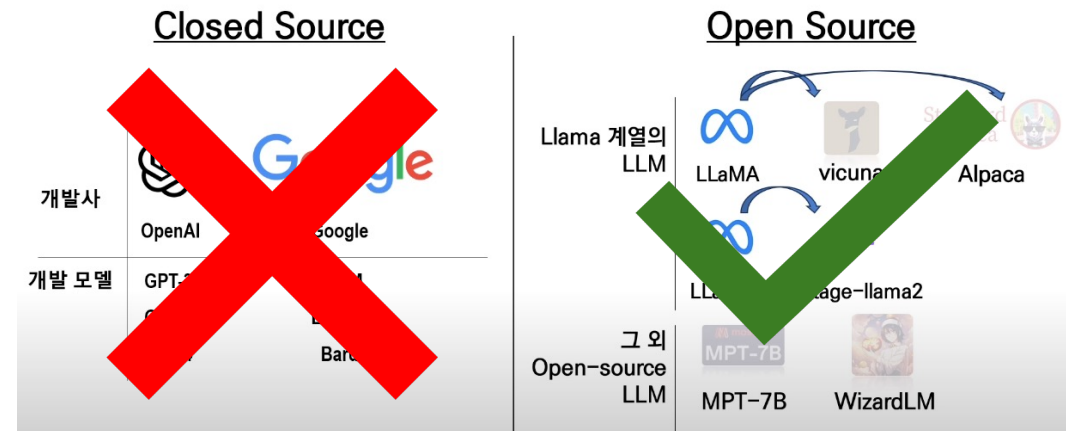
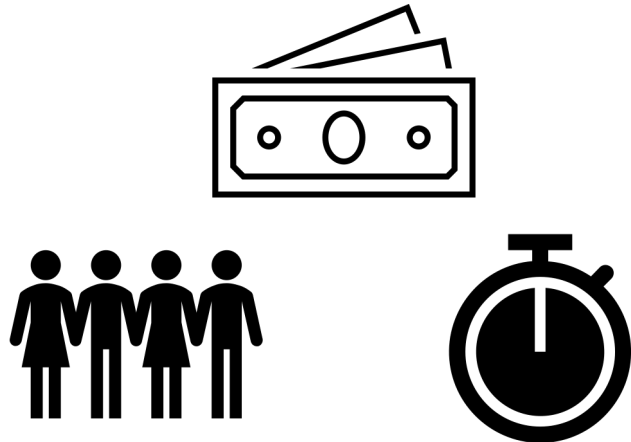
파인튜닝(Fine Tuning)



LLM 활용 방법

우리 회사 내규에 관한 챗봇을 만들려면?

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



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.

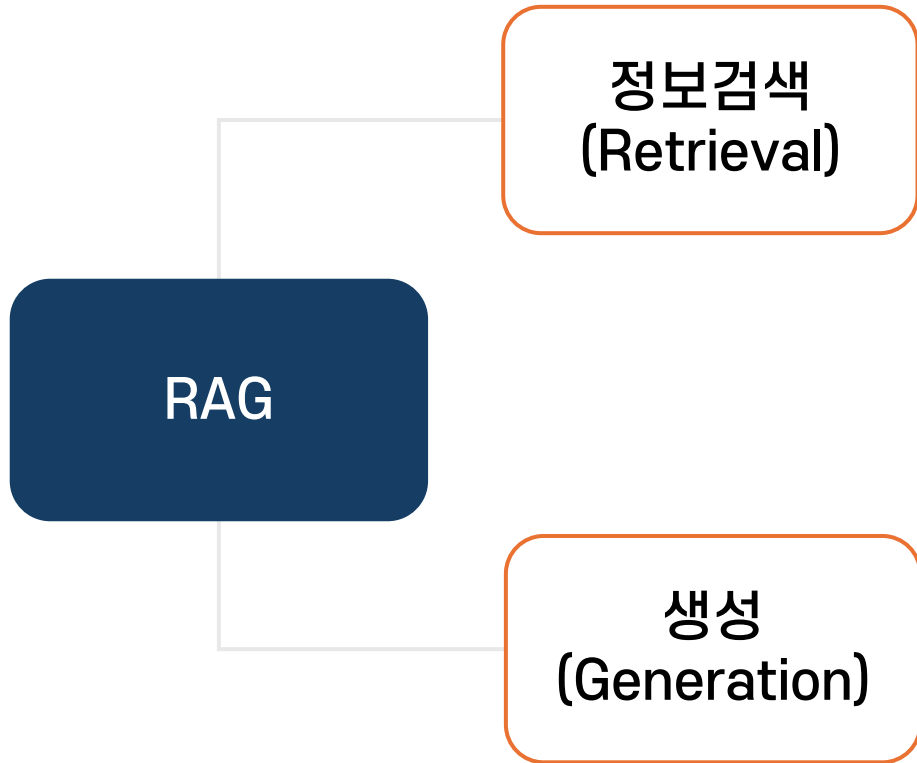
문서 검색 단계

- 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?

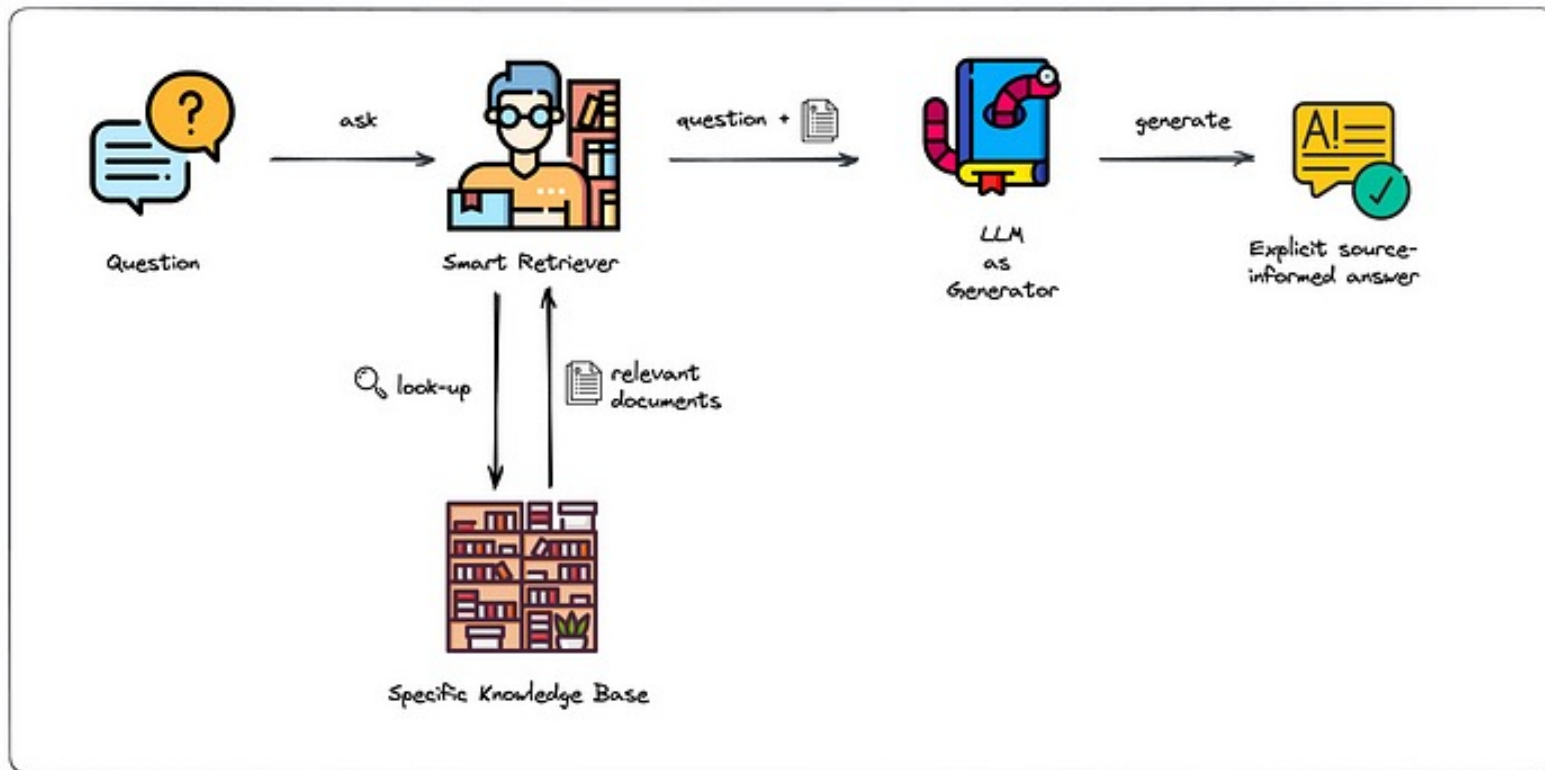


1. 질문 : 사용자로부터 질문이 입력됩니다
 2. 쿼리 (문서검색) : 모델은 대규모의 문서 데이터 베이스나 콘텐츠 저장소에서 질문과 관련된 문서나 정보를 검색합니다
 3. 정보 검색 결과 : 검색결과 중에서 가장 관련성이 높은 문서와 사용자의 질문을 결합하여 LLM에 전달합니다.
 4. 정보전달 : 사용자의 질문과 정보 검색결과가 모델에 전달됩니다. 이단 계에서 모델은 문서의 정보를 활용하여 질문에 대한 의미를 이해합니다.
 5. 텍스트 생성: 전달받은 정보를 바탕으로 답변을 생성합니다.
- > LLM에 의해 처리되는 부분

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.



RAG 원리

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>)

대한민국
₩ 266개 언어 ▼

문서 토론
임기 원본 보기 역사 보기 도구 ▼

위키백과, 우리 모두의 백과사전.

다른 뜻에 대해서는 **대한민국 (동음이의)** 문서를 참고하십시오.

남한은 여기로 연결됩니다. 중국 대륙식 시대 10대 중 하나에 대해서는 **남한 (십국)** 문서를 참고하십시오.


대한민국(한국 한자: 大韓民國)은 동아시아의 한반도 군사 분계선 남부에 위치한 나라이다. 약칭으로 **한국**(한국 한자: 韓國)과 **남한**(한국 한자: 南韓, 문화어: 남조선)으로도 불리며 헌정체제는 **대한민국 제6공화국**이다. 대한민국의 국기는 대한민국 국가법에 따라 **태극기**이며, 국가는 관습상 **애국가**, **국가의관송상 무궁화**이다. 공용어는 **한국어**와 **한국 수어**이다. 수도는 **서울특별시**이다. 인구는 2024년을 기준으로 약 5,176만 명으로, 전체 인구 중 절반이(약 2,626만 명) **수도권**에 살고 있다.^[6]

대한민국은 1948년 5월 10일 총선거를 통해 **제헌국회**를 구성하였고, 1948년 8월 15일 **대한민국 정부**를 수립하였다. 1948년 제헌 국회에서 대한민국의 국호를 계승하여 **한법**에 명시하였고, 다시 1954년 1월 16일 국무원 고시 제7호 '국호 및 일부 지방명과 지도색 사용에 관한 건'에 의해 확정하였다.

대한민국은 20세기 후반 이후 급격한 경제 성장을 이루었다. 그 과정에서 1990년대 말 외환 위기 등의 부침이 있기도 했다. 대한민국의 2022년 1인당 국민 총소득 (GNI)은 명목 3만 4,994달러이다.^[7] 2020년 유엔개발계획 (UNDP)이 매년 발표하는 인간개발지수 (HDI) 조사에서 세계 2위를 기록하였다.^[2] 2021년 7월 2일 스위스 제네바 본부에서 열린 '제68차 무역개발이사회' 마지막 회의에서 한국의 지위를 선진국 그룹으로 '의견 일치'로 변경하고 선진국으로 인정했다.^[8] 다만 높은 자살률, 장시간 근로 문화와 높은 산업 재해 사망률, 저출산 등의 사회 문제가 이 같은 성과와 병존하고 있다.

대한민국은 이코노미스트에서 발표하는 민주주의 지수 조사에서 2019년 기준 23위의 8.0점을 기록한 바와 같이 아시아에서 민주주의 수준이 가장 높은 국가 가운데 하나이다. 또한 대한민국은 주 요 20개국 (G20), 경제 협력 개발 기구 (OECD), 개발 원조 위원회 (DAC), 파리 클럽 등의 회원국이다.^[9]

1948년 이후로 오늘날까지 **한반도**에는 대한민국과 조선민주주의인민공화국이라는 두 개의 분단 국가가 각각 남북에 위치하고 있다. 한반도와 북속도서의 면적은 약 22만 km²이며, 인구는 대한민국과 조선민주주의인민공화국을 합쳐 2019년을 기준으로 약 7,718만 명에 달한다.^[10].




대한민국

국기 국장

대한학

국새

표어	홍익인간
국가	<div style="display: flex; align-items: center;"> ▶ 애국가 </div>



때 유래한 것으로 고려, 에서 대하이라 한다. 삼국 때는 삼한을 ‘삼국’이라 부르는 뜻이다. 한(韓)이라는 단 낱말로서, “하나”, “하” 말고 사용되었다.

면서 다시 선택하였다. 세 칸과 한 나라로 통합되었 백번 각 나라의 문자를 보고 오늘이 있기를 기다린 있을 것이다.”고 밝힌 는 이승만, 김구 등 대한 고 낮은 구분

년 1월 16일 국무원 고시 [13] 이에 20세기 전반까 **한국(韓國)**”, “**한(韓)**”으 **민국의 국민**은 자국의 국 **방**”라고 한다. “대한”이라 **국과**^[14] 대비해 **한반도** 가리고 있는 조선민주주

면 반도형이며, 가장 큰 리안 리아스식 해안으로 한 수많은 섬이 있어서 만의 차가 적다. 해안 근 육으로 약 87.4km 거리

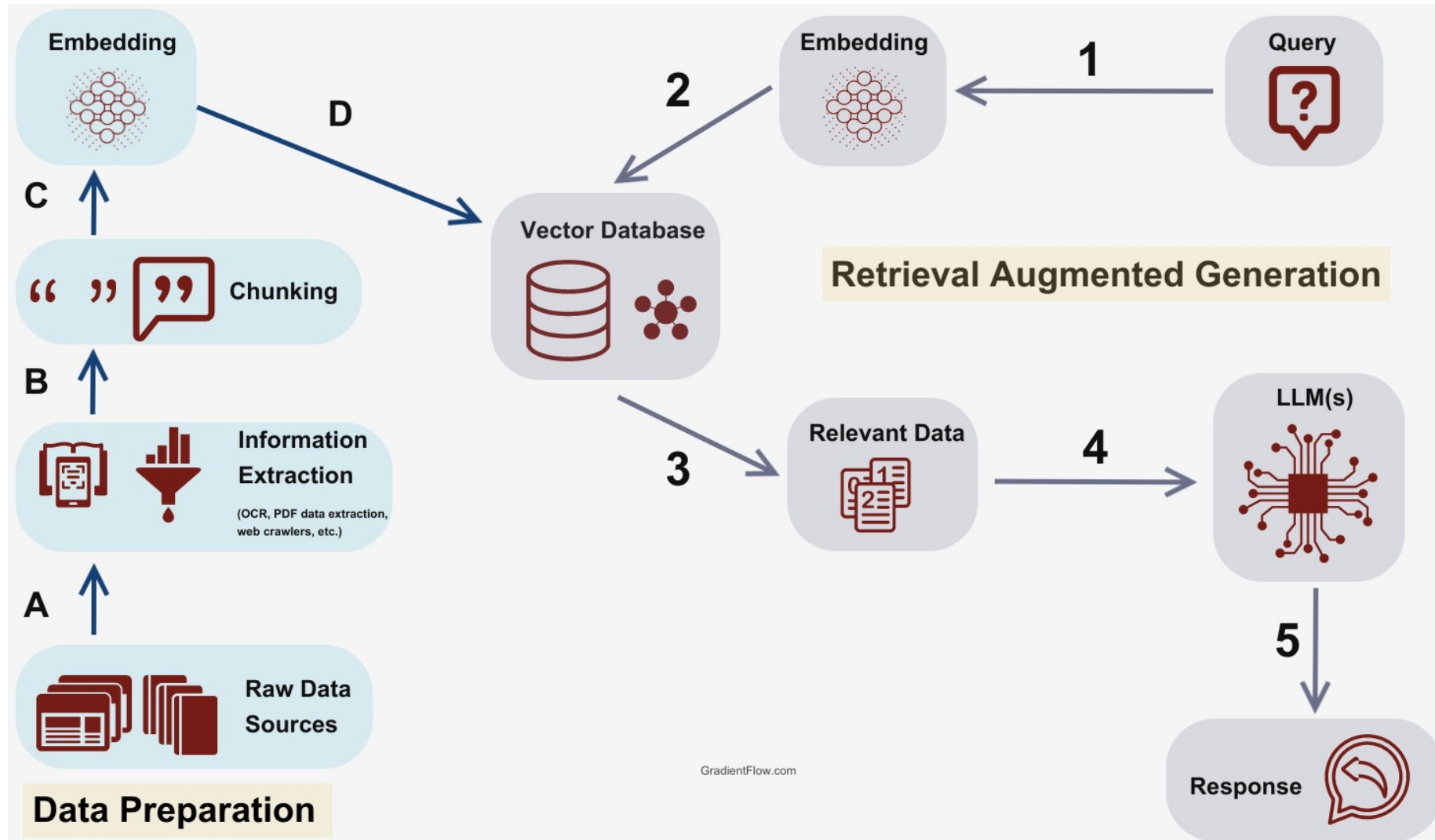
우려하였다.^{[18][19][20][21]} 건국된 중에 황조 고려 려졌다. 유럽인들이 ‘고 는 ‘-a’를 붙여 ‘고려인의 왕’으로 Corée, 스페인 식 영어 명칭은

■ ■ ■



수도는 서울특별시이다.

RAG(Retrieval-Augmented Generation) 구현과정



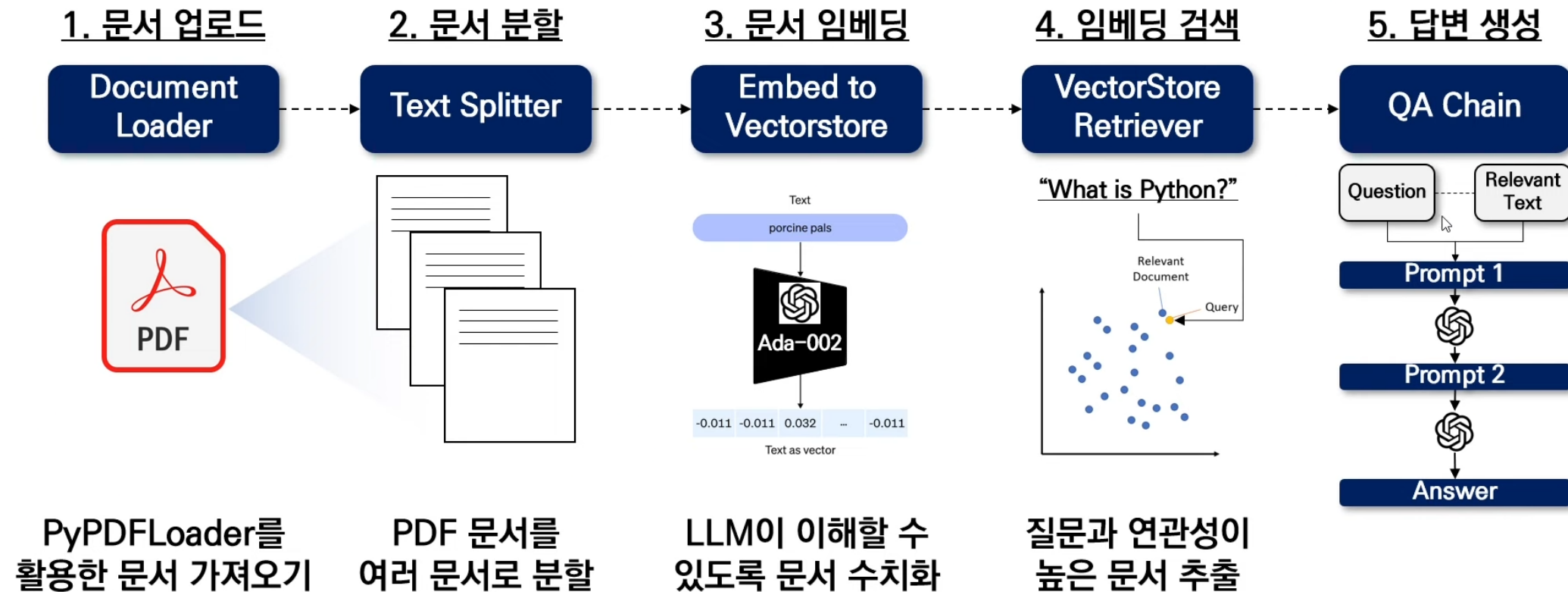
Q&A with RAG



LangChain

예시: PDF 챗봇 구축

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



감사합니다

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)

