Fast campus

the Red.

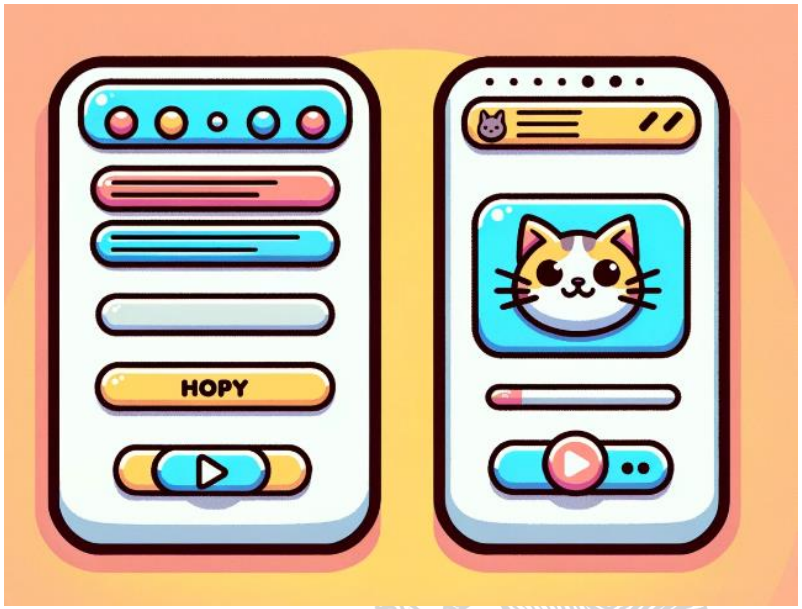
LLM 마스터클래스 #4

LLM 서비스 개발: RAG

Part 4-1.
서비스 예시

서비스를 만든다는 것

- 아주 간단한 웹 페이지!
  - 버튼을 누르면 cat fact 를 보여줍니다
- 아주 간단한 모바일 앱!
- UI + API call + 백엔드



서비스를 만든다는 것

- “데이터 보안이 제일 중요해요!!”
- “그냥 쿨한 AI 앱을 빨리 만들어보고 싶어요!”
- “우리 회사 시스템 안에 넣어보고 싶어요”
- “파인 튜닝으로 제대로 학습 좀 시켜보고 싶어요”
- “저의 업무를 LLM 으로 훈련 시킬 수 있을까요??”

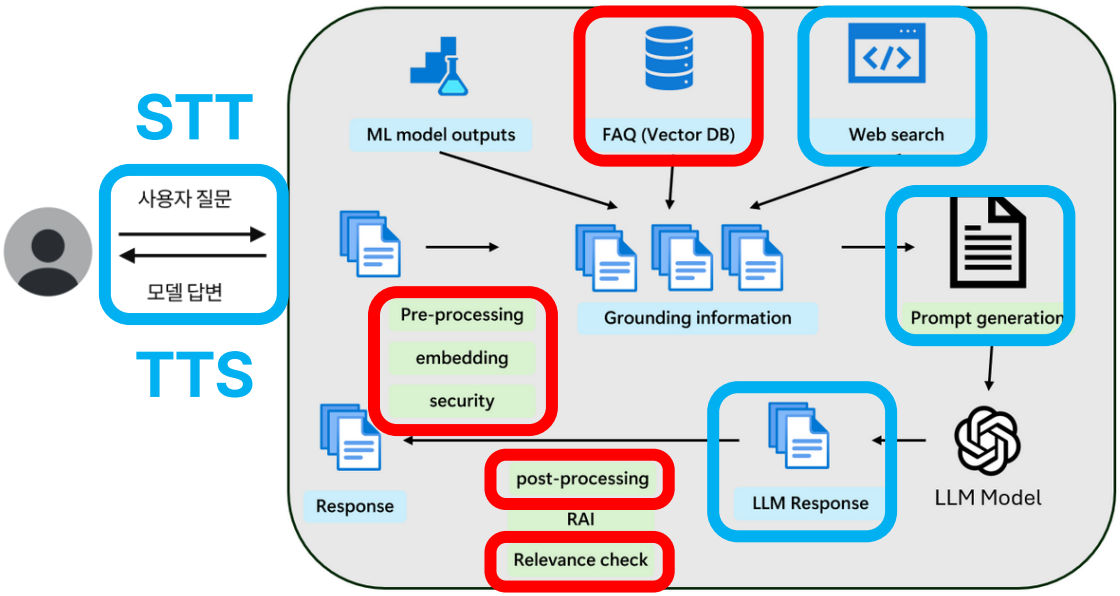
LLM 서비스 요청

- "데이터 보안이 제일 중요해요!!"
  → 사용할 데이터의 현재 상태와 흐름, 프로세스 난이도 중요
- "그냥 쿨한 AI 앱을 빨리 만들어보고 싶어요!"
  → 타겟에 따라서 어떻게 배포할것인가, UI/가격/시스템 선택
- "우리 회사 시스템 안에 넣어보고 싶어요"
  → 회사 시스템의 요구 사항및 데이터 상황 중요. 데이터 보안 요구 중요.
- "파인 튜닝으로 제대로 학습 좀 시켜보고 싶어요"
  → 데이터셋이 있는가, 파인 튜닝을 지속적으로 관리하고 업데이트하고 테스트할 프로세스가 있는가
- "저의 업무를 LLM 으로 훈련 시킬 수 있을까요??"
  → 데이터셋, 회사의 의도, stand-alone app/tool 등등

우리의 샘플 앱

파랑
이전 파트에서 커버

빨강
이번 파트에서 커버



우리의 샘플 앱

이전 파트

- LLM 사용하는 코드 써보기
- 여러가지 모델 써보기
- STT, TTS 기술 쓰기
  - pre-processing, post-processing
- Grounding information 을 검색,
  데이터 로딩 등으로 준비하기
- 프롬프트 엔지니어링으로
  프롬프트 만들기

이번 파트

- Grounding, RAG 이해하기
- 임베딩, 벡터 인덱스 이해하기,
  VectorDB 둘러보기
- Pre-post processing 에서 기본적인
  moderation / security API 써보기
- 하나의 앱으로 묶어보기

Part 4-2.
언어 모델의 특성

언어 모델의 특징

구술 시험을 본다고 합시다.

물론 여러가지 자료를 보고
공부했지만
생각이 제대로 안 나면...?

- 언어 모델은 데이터베이스가 아님
- 사람의 뇌도 데이터베이스가 아님
- 그렇지만 정확하게 기억한다고 착각할 수 있음

언어 모델의 특징

오픈 북 시험은 다르겠죠?

Part 4-3.

Grounding, RAG primer

RAG의 기본

오픈 북 시험

- “컨닝 페이퍼 줄 테니까 그걸 바탕으로 읽어!
- 모르는 거 만들어내지 말고!
- 우리가 쓴 langchain 예시
  - 유튜브
  - 웹 검색

Grounding 의 뜻

Grounding

- “I’m grounded”
- “Grounded fleet”
- “back to earth!”
- 피뢰침 (lightning rods)
- grounding principles
- grounding facts

Grounding - 주요 방식 RAG

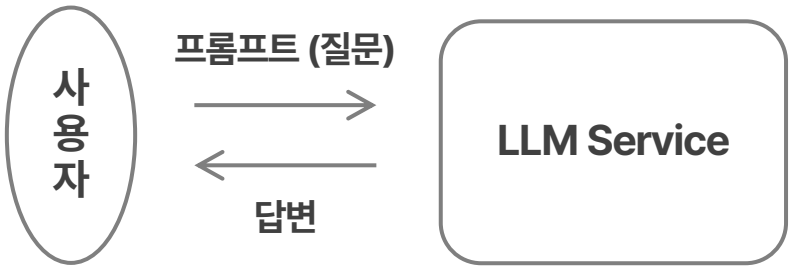
Retrieval Augmented Generation

- retrieval - 쪽지를 만들어서
- augmented – (생성하기 전에 LLM 에게 패스하여) 더 나아진
- generation - 생성

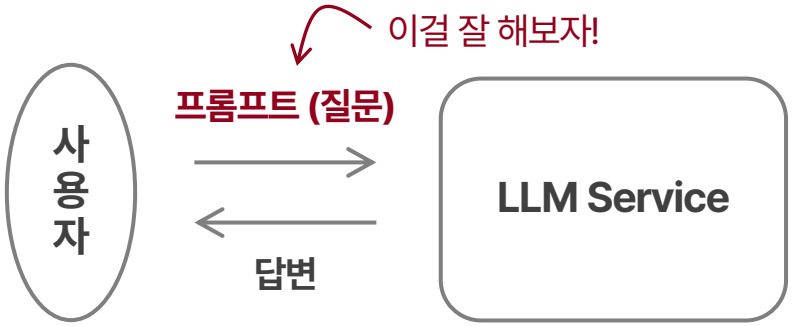
Grounding 을 할 수 있는 방법 - RAG

OpenBook 으로 만들기

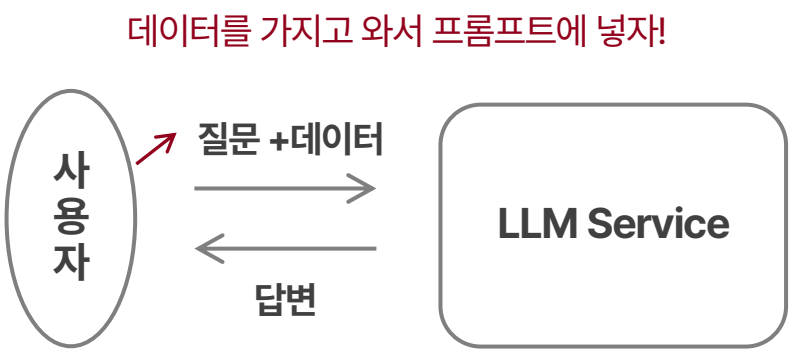
기본 – 웹으로 쓰기

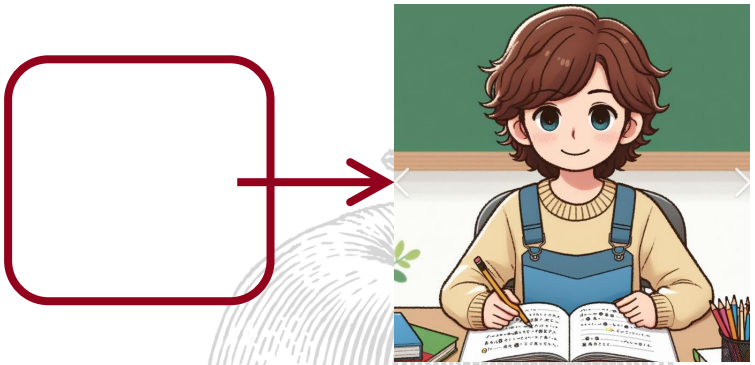


프롬프트 엔지니어링

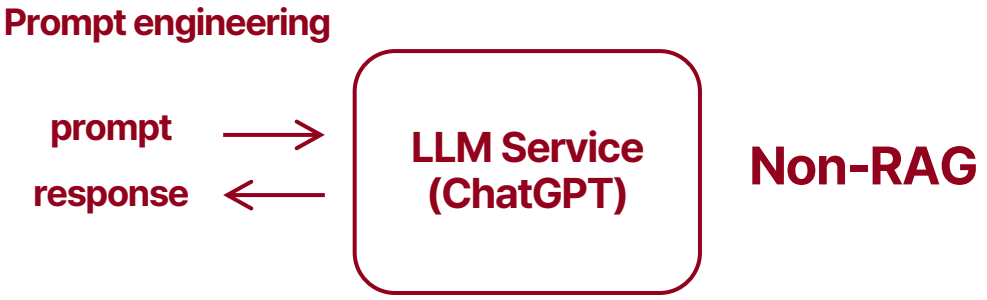


RAG 1



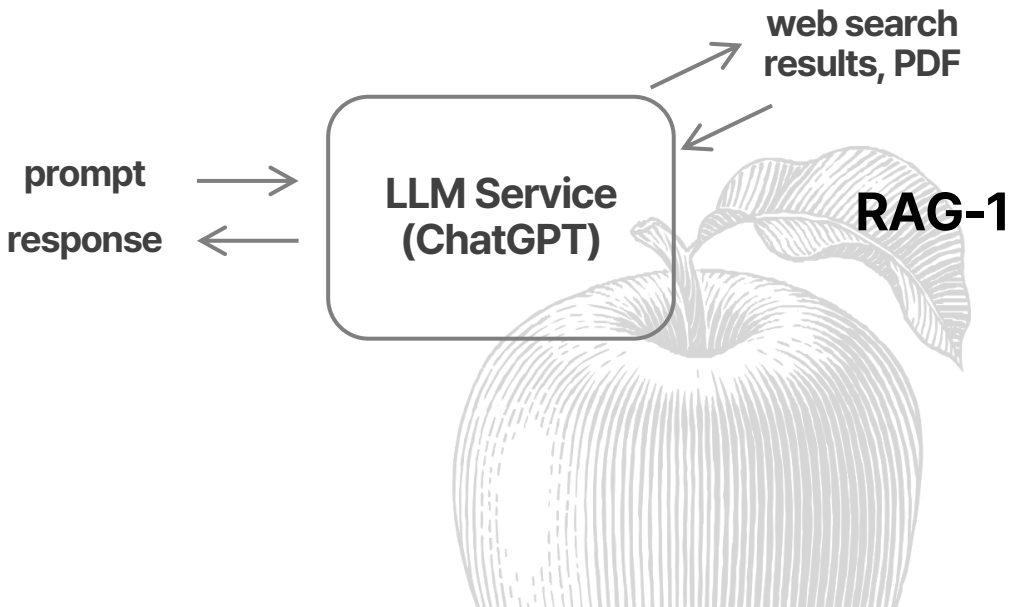


RAG 버전 1



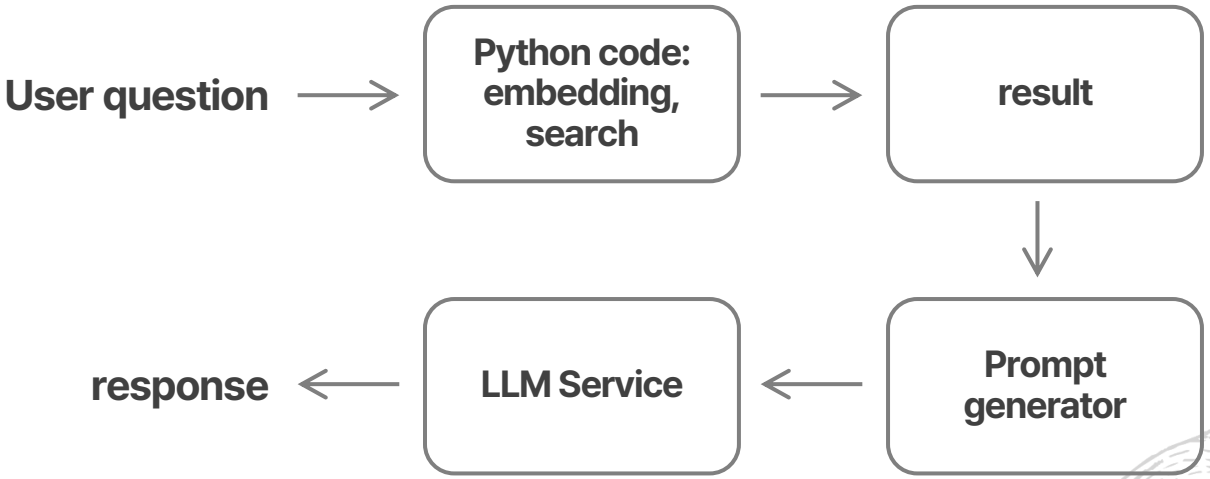
가장 간단한 방법

- 프롬프트에 데이터 포함
- API call 로 데이터 가져오기



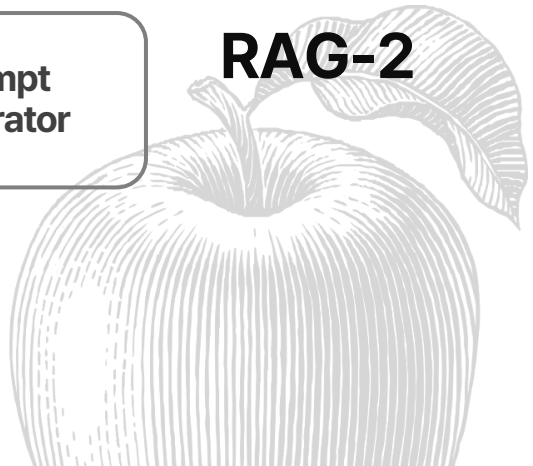
RAG 버전 2

- 미리 저장되어 있는 데이터를 가져와서 프롬프트를 만듬 - embedding
- LLM 모델을 몇 번 쓸 수도 있음



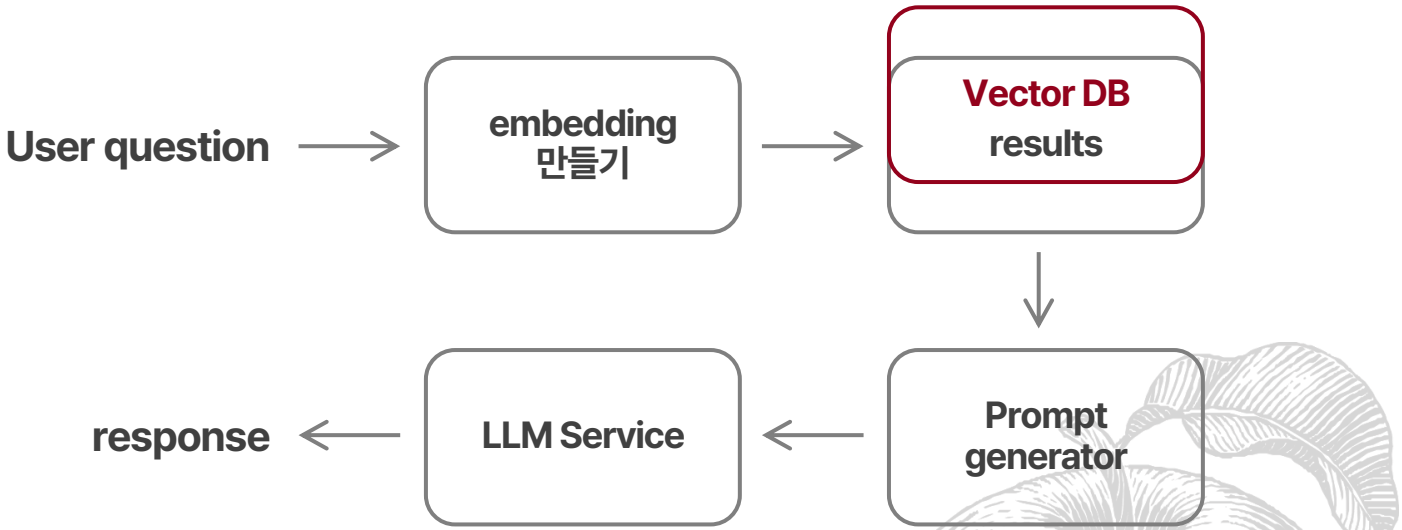
Embedding으로 검색하여
매칭되는 데이터로
fewshot 예시 만들기
(email 카탈로그)

RAG-2



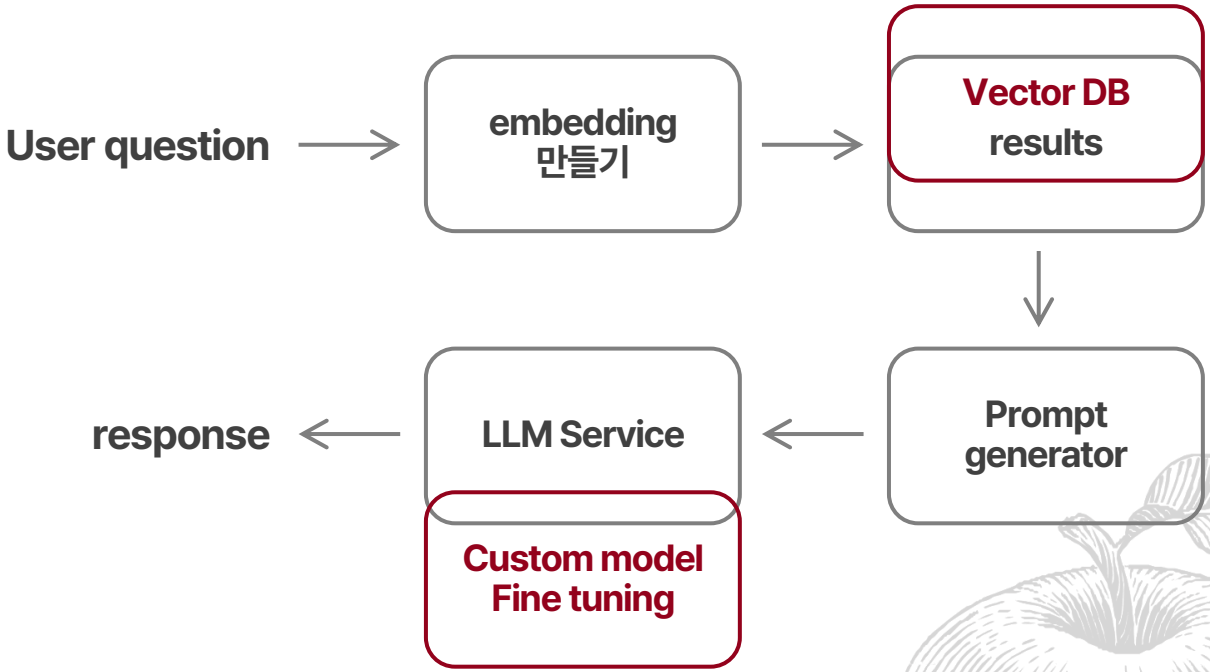
RAG 버전 3

- Vector DB 를 만들고 데이터를 저장해둠
- Embedding 으로 데이터를 저장하고 찾을 때도 씀

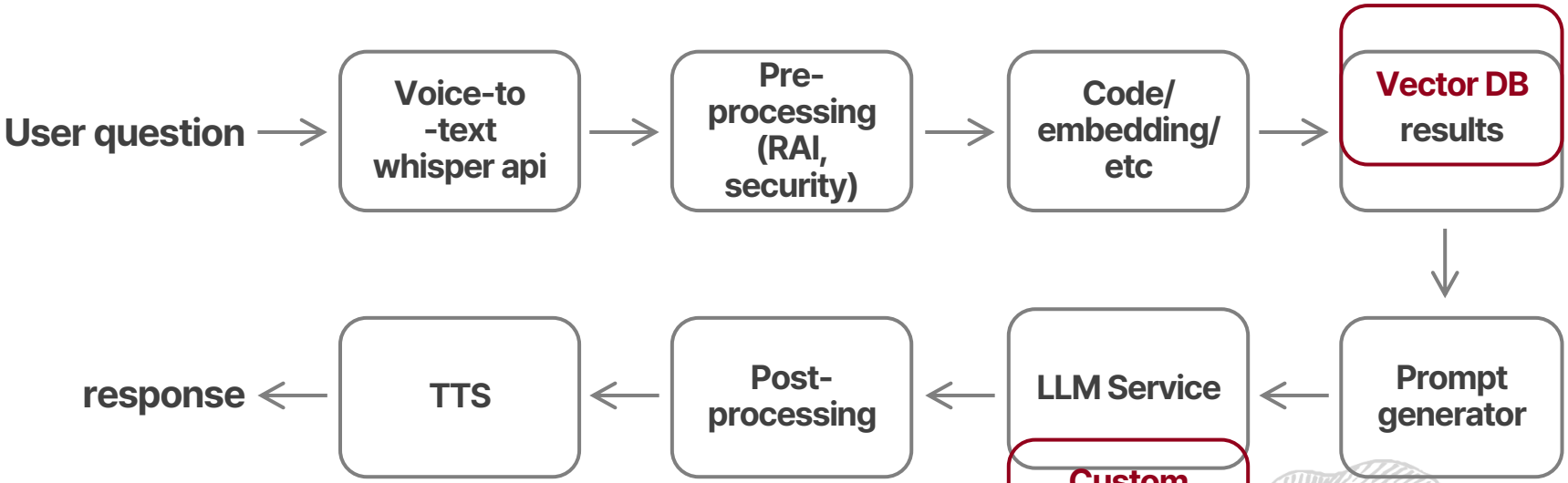


RAG 버전 4

- LLM 서비스 자체를 업그레이드



우리의 샘플 앱


User question
Voice-to
-text
whisper api
Pre-
processing
(RAI,
security)
Code/
embedding/
etc
Vector DB
results
Prompt
generator
LLM Service
Custom
model
Fine tuning
Post-
processing
TTS
response

Part 4-4.
Embeddings

Embeddings

임베딩 이해하기

Embeddings 이해하기

실제 영어에서 embeddings 의 뜻

- Embed (임베드)

일반적 의미	어떤 것을 다른 것에 삽입하거나 통합하는 행위를 의미. 예를 들어, 웹 페이지에 비디오나 지도 등의 외부 콘텐츠를 '임베드'하는 것
기술적 의미	특정 데이터나 기능을 한 시스템 또는 장치 안에 내장시키는 것을 의미.

- Embedding (임베딩)

일반적 의미	'Embed'의 명사형으로, 삽입되거나 통합된 결과물을 의미

Embeddings 이해하기

AI 에서 Embedding 이란?

- 데이터, 특히 단어나 문장과 같은 텍스트 데이터를 벡터의 형태로 표현하는 과정을 의미합니다. 이 벡터는 원래의 텍스트 데이터보다 낮은 차원을 가지며, 기계 학습 모델이 이해할 수 있는 형태로 변환됩니다. 이런 벡터를 '임베딩 벡터'라고 하며, 이는 데이터의 의미적, 문맥적 특성을 수치화한 것입니다.

더 간단히 말하면?

- “위치 표현”

Embeddings 이해하기

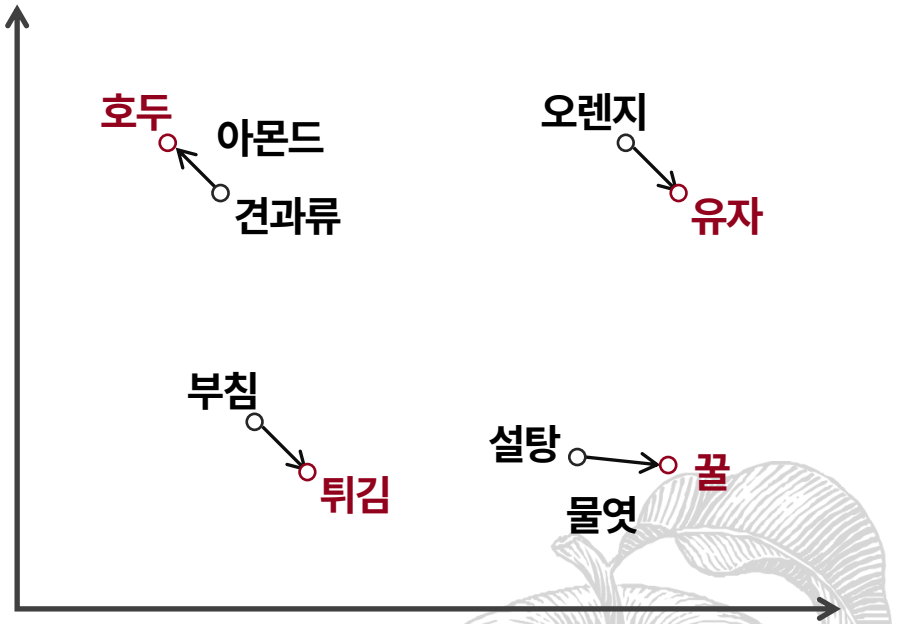
- 마트에서 Aisle 3: 냉동고
  - 만두: Aisle 3, 4번 선반, 1035
  - 냉동 파전: Aisle 3, 4번 선반, 864
  - 고양이 장난감: Aisle 34, 28번 선반, 379245
- 언어 모델 내에서 -
  - 애들: [3, 6, 23, 64, 2364, 1235...]
  - 학생들: [4, 7, 12, 51, 2325, 62243...]
  - 만두: [2356, 2346, 1274,7324....]
- 벡터: [1, 2, 3, 4, 5...] (숫자 리스트)
- 경도/위도 벡터:
  - [37.5519, 126.9918] (서울)
  - [35.2100° N, 129.0689°] (부산)
  - 47.6061° N, 122.3328° W (시애틀)

Embeddings 이해하기

Cosine similarity

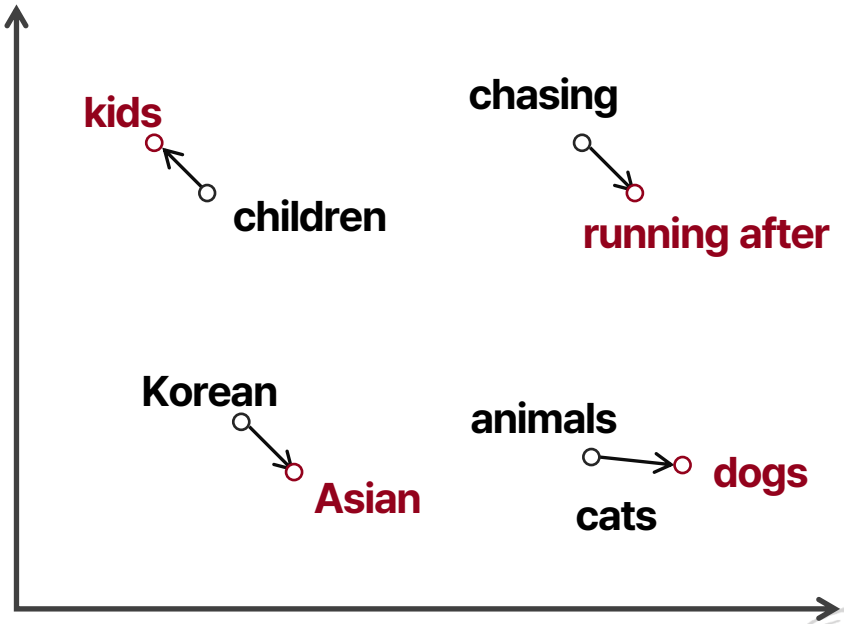
꿀 꽈배기

호떡

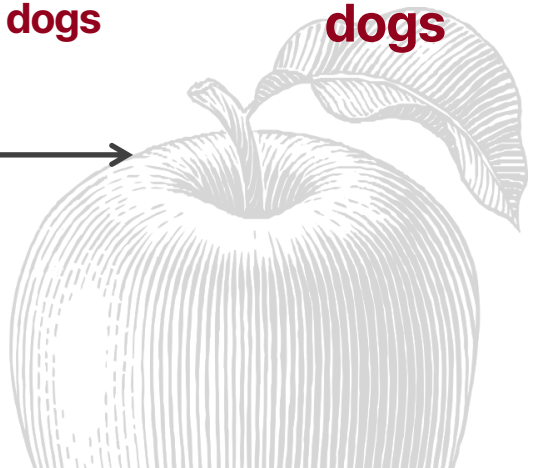


Embeddings 이해하기

The children were chasing animals



The kids were running after the dogs



Semantic search

- “시맨틱 검색은 검색 보강 생성(RAG)을 위한 주요 기술이 되었으며, 이는 유일한 방법은 아닙니다.
  이 과정은 문서나 문서의 일부를 그들의 의미적 표현을 사용하여 색인화하는 것을 포함합니다.
  검색 시, 쿼리의 의미적 표현에서 가장 관련성 높은 문서를 찾기 위해 유사성 검색이 수행됩니다.
  이 강력하고 사용하기 쉬우며 빠른 기술은 임베딩 모델, 벡터 인덱스, 유사성 검색에 의존합니다.”
- Semantic search
  : 정확한 문자열이 아닌 **“의미”로 찾기**

Embeddings 만들어보기

- 코드 실습
  - 임베딩은 어떻게 생겼나요?
  - 여러가지 임베딩 모델들
  - 임베딩으로 찾기
  - Cosine similarity 사용하기
- 다른 임베딩 모델 쓰기?
  - A 마트와 B 마트의 주소를 쓰기?

Embedding

임베딩 코드 실습

Embedding 코드 실습

Colab notebook

- 노트북 링크

Embedding Use case - OpenAI

OpenAI 의 예시 써보기

- https://platform.openai.com/docs/guides/embeddings/use-cases

Part 4-5.

Vector indexing, DBs

VectorDB

Vector indexing, VectorDBs

Vector Index and DB

어떻게 찾을 것인가...

만두 찾아주세요!
[3, 2543, 634..]
주세요!

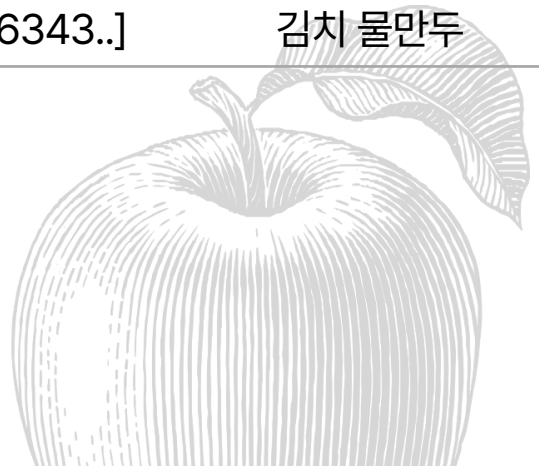
앗 네!
찾아볼게요!

기존 DB

인덱스 / PK	품목
PD00231	납작 군만두
PD00232	사천 군만두
PD00233	김치 물만두

물품 목록 / Vector DB

인덱스 / PK	품목
[3, 2543, 6341..]	납작 군만두
[3, 2543, 6342..]	사천 군만두
[3, 2543, 6343..]	김치 물만두



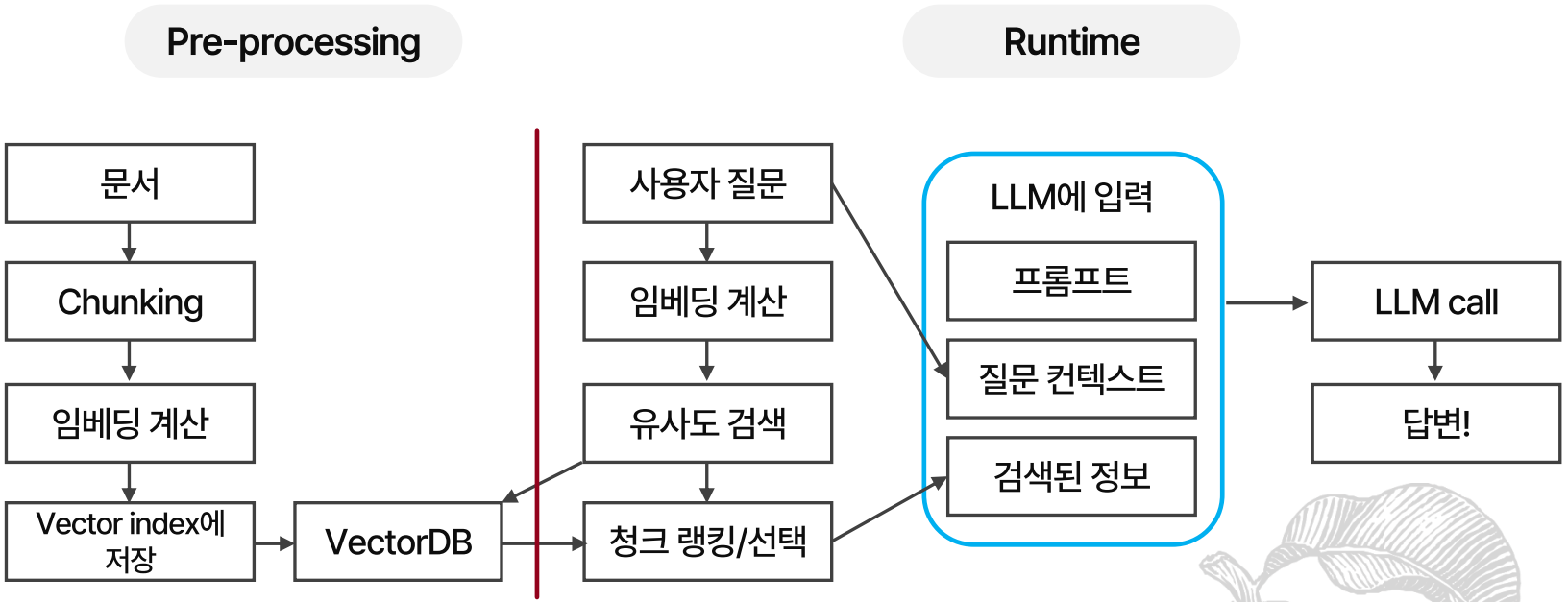
Vector DB

- 새로운 정보를 DB 에 저장
  - "키", "찾는 방법"을 임베딩으로 지정 → VectorDB
  - 코딩 예시 복습 - 유튜브 내용 검색

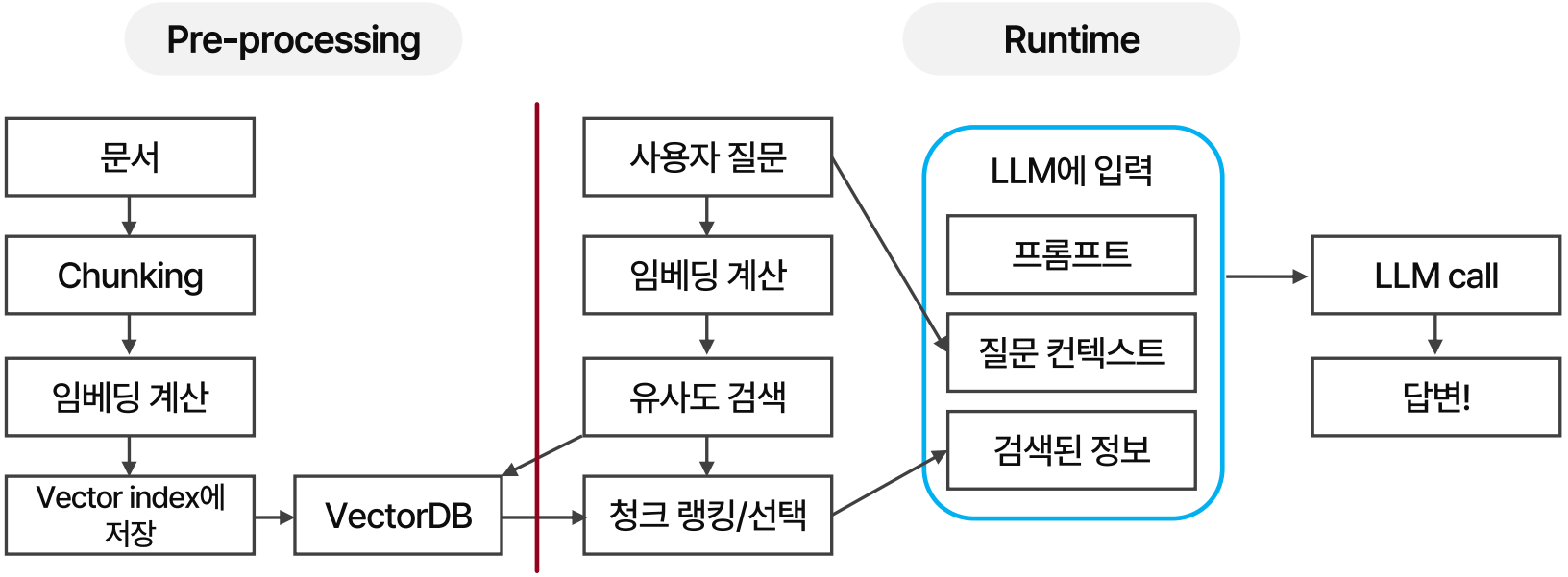
Vector DB 종류

- Pinecone, Weaviate, QDrant, Milvus, Chroma 등의 전용 벡터 데이터베이스 등장
- Pinecone은 서비스로서의 소프트웨어로 성공적인 사례
- Weaviate, QDrant, Milvus는 클라우드 서비스와 함께 오픈 소스로 쉽게 배포 가능
- Chroma는 SQLite와 유사해 작은 프로젝트에 적합
- FAISS
- 벡터 인덱스 지원 제품
  - AMZ: Elasticsearch, Redis, Postgres 등 다양한 데이터베이스가 벡터 인덱스 지원.
  - MS: Cognitive Search, etc

Vector 인덱스와 찾기 모델



Vector 인덱스와 찾기 모델



물품 목록 / Vector DB

인덱스 / PK	품목
[3, 2543, 6341..]	납작 군만두
[3, 2543, 6342..]	사천 군만두
[3, 2543, 6343..]	김치 물만두

"만두나 김밥 먹을까?"

→ "사용자가 원하는 품목 [만두] 종류 몇 개 예를 들면서 어떻게 맛있게 먹을 수 있는지 추천해줘"

→ "오늘은 매콤한 사천 군만두 어떠세요? 김치 물만두는 오늘 세일이네요!

Ranking, cutoff

- 쓸데없는 내용도 검색될 수 있음
- 검색 내용이 너무 많을 수도 있음 → 토큰 수 늘어남, 결과 부정확함
- 가장 유용한 몇 개만 가지고 와서 프롬프트 템플릿을 만들고 LLM 에 보내기

사용자가 원하는 음식은
[만두 김치만두 김밥 떡볶이 엽기 떡볶이 탕수육 탕후루 메로나...]이다

vs

사용자가 원하는 음식은 만두와 김밥이다.

Vector 인덱스와 찾을 때에..

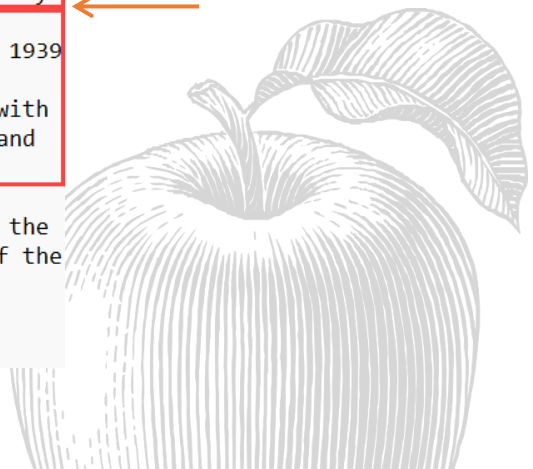
청킹! 어떻게 나누고 정리할 것인가!

Chunking – 정해진 사이즈

World War II[b] or the Second World War (1 September 1939 – 2 September 1945) was a global conflict between two major alliances: the Allies and the Axis powers. The vast majority of the world's countries, including all the great powers, fought as part of these military alliances. Many participating countries invested all available economic, industrial, and scientific capabilities into this total war, blurring the distinction between civilian and military resources. Aircraft played a major role, enabling the strategic bombing of population centres and delivery of the only two nuclear weapons ever used in war. It was by far the deadliest conflict in history, resulting in 70–85 million fatalities. Millions died due to genocides, including the Holocaust, as well as starvation, massacres, and disease. In the wake of Axis defeat, Germany, Austria, and Japan were occupied, and war crime tribunals were conducted against German and Japanese leaders.

The causes of the war are debated; contributing factors included the rise of fascism in Europe, the Spanish Civil War, the Second Sino-Japanese War, Soviet–Japanese border conflicts, and tensions in the aftermath of World War I. World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland. The United Kingdom and France declared war on Germany on 3 September. Under the Molotov–Ribbentrop Pact of August 1939, Germany and the Soviet Union had partitioned Poland and marked out their "spheres of influence" across Finland, Estonia, Latvia, Lithuania, and Romania. From late 1939 to early 1941, in a series of campaigns and treaties, Germany conquered or controlled much of continental Europe in a military alliance called the Axis with Italy, Japan, and other countries. Following the onset of campaigns in North and East Africa, and the fall of France in mid-1940, the war continued primarily between the European Axis powers and the British Empire, with the war in the Balkans, the aerial Battle of Britain, the Blitz of the UK, and the Battle of the Atlantic. In June 1941, Germany led the European Axis powers in an invasion of the Soviet Union, opening the Eastern Front, the largest land theatre of war in history.

정해진 사이즈 청크

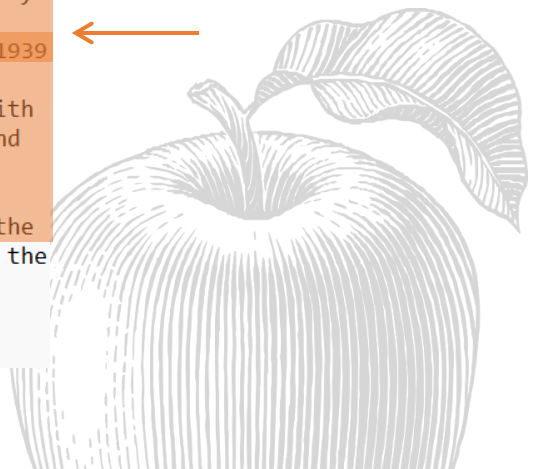


Chunking – 겹치는 부분 설정

World War II[b] or the Second World War (1 September 1939 – 2 September 1945) was a global conflict between two major alliances: the Allies and the Axis powers. The vast majority of the world's countries, including all the great powers, fought as part of these military alliances. Many participating countries invested all available economic, industrial, and scientific capabilities into this total war, blurring the distinction between civilian and military resources. Aircraft played a major role, enabling the strategic bombing of population centres and delivery of the only two nuclear weapons ever used in war. It was by far the deadliest conflict in history, resulting in 70–85 million fatalities. Millions died due to genocides, including the Holocaust, as well as starvation, massacres, and disease. In the wake of Axis defeat, Germany, Austria, and Japan were occupied, and war crime tribunals were conducted against German and Japanese leaders.

겹치는 부분

The causes of the war are debated; contributing factors included the rise of fascism in Europe, the Spanish Civil War, the Second Sino-Japanese War, Soviet–Japanese border conflicts, and tensions in the aftermath of World War I. World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland. The United Kingdom and France declared war on Germany on 3 September. Under the Molotov–Ribbentrop Pact of August 1939, Germany and the Soviet Union had partitioned Poland and marked out their "spheres of influence" across Finland, Estonia, Latvia, Lithuania, and Romania. From late 1939 to early 1941, in a series of campaigns and treaties, Germany conquered or controlled much of continental Europe in a military alliance called the Axis with Italy, Japan, and other countries. Following the onset of campaigns in North and East Africa, and the fall of France in mid-1940, the war continued primarily between the European Axis powers and the British Empire, with the war in the Balkans, the aerial Battle of Britain, the Blitz of the UK, and the Battle of the Atlantic. In June 1941, Germany led the European Axis powers in an invasion of the Soviet Union, opening the Eastern Front, the largest land theatre of war in history.



Chunking – 문단별

World War II[b] or the Second World War (1 September 1939 – 2 September 1945) was a global conflict between two major alliances: the Allies and the Axis powers. The vast majority of the world's countries, including all the great powers, fought as part of these military alliances. Many participating countries invested all available economic, industrial, and scientific capabilities into this total war, blurring the distinction between civilian and military resources. Aircraft played a major role, enabling the strategic bombing of population centres and delivery of the only two nuclear weapons ever used in war. It was by far the deadliest conflict in history, resulting in 70–85 million fatalities. Millions died due to genocides, including the Holocaust, as well as starvation, massacres, and disease. In the wake of Axis defeat, Germany, Austria, and Japan were occupied, and war crime tribunals were conducted against German and Japanese leaders.

긴 문단

The causes of the war are debated; contributing factors included the rise of fascism in Europe, the Spanish Civil War, the Second Sino-Japanese War, Soviet–Japanese border conflicts, and tensions in the aftermath of World War I. World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland. The United Kingdom and France declared war on Germany on 3 September. Under the Molotov–Ribbentrop Pact of August 1939, Germany and the Soviet Union had partitioned Poland and marked out their "spheres of influence" across Finland, Estonia, Latvia, Lithuania, and Romania. From late 1939 to early 1941, in a series of campaigns and treaties, Germany conquered or controlled much of continental Europe in a military alliance called the Axis with Italy, Japan, and other countries. Following the onset of campaigns in North and East Africa, and the fall of France in mid-1940, the war continued primarily between the European Axis powers and the British Empire, with the war in the Balkans, the aerial Battle of Britain, the Blitz of the UK, and the Battle of the Atlantic. In June 1941, Germany led the European Axis powers in an invasion of the Soviet Union, opening the Eastern Front, the largest land theatre of war in history.

Chunking – 메타 데이터 포함

A WEAK (k,k)-LEFSCHETZ THEOREM FOR PROJECTIVE TORIC ORBIFOLDS

WILLIAM D. MONTOYA

Instituto de Matemática, Estatística e Computação Científica,
Universidade Estadual de Campinas (UNICAMP),
Rua Sérgio Buarque de Holanda 651, 13083-859, Campinas, SP, Brazil

February 9, 2023

Abstract

Firstly we show a generalization of the $(1,1)$-Lefschetz theorem for projective toric orbifolds and secondly we prove that on $2k$-dimensional quasi-smooth hypersurfaces coming from quasi-smooth intersection surfaces, under the Cayley trick, every rational (k,k)-cohomology class is algebraic, i.e., the Hodge conjecture holds on them.

1 Introduction

In [3] we proved that, under suitable conditions, on a very general codimension s quasi-smooth intersection subvariety X in a projective toric orbifold $\mathbb{P}^d_\Sigma$ with $d+s=2(k+1)$ the Hodge conjecture holds, that is, every (p,p)-cohomology class, under the Poincaré duality is a rational linear combination of fundamental classes of algebraic subvarieties of X. The proof of the above-mentioned result relies, for $p \neq d+1-s$, on a Lefschetz

처리하는 문서가 늘 비슷한
포맷이라면 커스텀 청킹이
훨씬 낫지만 다른 타입
문서에는 적용하기 힘듦

Chunking 설명

- Chunking 의 중요성
- 문장별로? 줄 수대로? 페이지별로? 문단별로?
  - 고정 크기 청킹 (보통 1024 토큰)
    - 구현 간단, 예측 가능.
    - 중간에 끊어지면 의미 손실
  - 동적 크기 청킹/semantic: 데이터의 내용에 따라 블록의 크기가 달라지는 방법.
    - 유연하고 효율성 증가
    - 구현이 복잡함
- 정확도 vs 속도
  - 작은 임베딩: 정확한 임베딩, 그러나 처리하는데 훨씬 오래 걸림
  - 큰 임베딩: 빠르지만 덜 정확함

Vector 인덱싱, 찾기 툴 문제

- 소설 정보? 컨텍스트 없이 이해하기 힘들 수 있음
- 비슷비슷한 유사도면 뭘 선택?
- 토큰!! 한 마디 한 마디 다 돈임!!! 임베딩 만드는 것도 돈임!!
- LlamaIndex:
  - PDF, DB, 또는 API 로도 데이터를 받아 청킹, 인덱스 만듬
  - 사용자 질문에 효과적으로 답할 수 있음
- LlamaIndex vs GPTs
- Langchain tools – text splitters
  - Splitters for – sentence, word, HTML, code, markdown..
  - UserDefined (custom character)

Vector 인덱싱, 찾기 툴 문제 해결

- 텍스트 청크 요약해서 저장하기
- Query 다시 쓰기, 쿼리 여러 개로 만들기
- 비슷비슷한 내용 합치기
- 전통적인 키워드 검색이나 메타데이터 필터링이랑 합치기
  - 메타데이터: (이미지) 날짜, 사이즈, 장소, (텍스트) 태그/카테고리, 저자/소스
  - 메타데이터 만들어 저장: “할인 품목“, “인기 품목“, “A 브랜드의 저렴 버전“, “마진율”
  - 정확한 단어 검색이 유용할 수 있음: 브랜드, 인용
- API/Plugin 등을 사용하여 검색

VectorDB / Indexing 실습

https://docs.llamaindex.ai/en/stable/getting_started/starter_example/

https://www.deeplearning.ai/short-courses/building-applications-vector-databases/

https://www.deeplearning.ai/short-courses/advanced-retrieval-for-ai/

https://www.deeplearning.ai/short-courses/preprocessing-unstructured-data-for-llm-applications/

Langchain 코드 실습

- LlamaIndex 코드 실습
- Deeplearning.ai
- https://github.com/teddylee777/langchain-kr 랭체인
- https://wikidocs.net/231147 랭체인 입문부터 응용까지
- https://www.youtube.com/@AI-km1yn/videos 모두를 위한 AI

Part 4-6.
후처리 post processing

후처리 Post-processing

Relevance, moderation, RAI, groundedness, correctness etc

LLM 답변 받은 후의 post processing

- 윤리적 AI (RAI), 관련성 (relevance), 사실 기반 (groundedness), cosine similarity
- 내용 검열 (moderation) API
  - 사용자 쿼리에도 적용
- Azure 는?
- Google 은?

Moderation API - OpenAI

- Moderation API:
  - 한글 ☹
- 코드 샘플

범주	설명
hate	인종, 성별, 민족, 종교, 국적, 성적 취향, 장애 상태 또는 계급에 따른 증오심을 표현, 선동, 조장하는 콘텐츠입니다. 보호되지 않는 집단(예: 체스 선수)을 겨냥한 증오성 콘텐츠는 괴롭힘에 해당합니다.
hate/threatening	인종, 성별, 민족, 종교, 국적, 성적 취향, 장애 상태 또는 계급에 따라 대상 집단에 대한 폭력이나 심각한 피해도 포함하는 증오성 콘텐츠입니다.
harassment	대상을 향해 괴롭히는 언어를 표현, 선동, 조장하는 콘텐츠.
harassment/threatening	대상에 대한 폭력이나 심각한 피해를 포함하는 괴롭힘 콘텐츠입니다.
self-harm	자살, 절단, 섭식 장애 등 자해 행위를 조장, 조장, 묘사하는 콘텐츠입니다.
self-harm/intent	화자가 자살, 절단, 섭식 장애 등 자해 행위에 가담하고 있거나 가담할 의도가 있음을 표현하는 콘텐츠입니다.
self-harm/instructions	자살, 베기, 섭식 장애 등 자해 행위를 조장하거나 그러한 행위를 하는 방법에 대한 지침이나 조언을 제공하는 콘텐츠.
sexual	성행위 묘사 등 성적인 흥분을 불러일으키거나 성서비스를 홍보하는 콘텐츠(성교육 및 웰빙 제외)
sexual/minors	18세 미만의 개인이 포함된 성적 콘텐츠.
violence	죽음, 폭력 또는 신체적 부상을 묘사하는 콘텐츠.
violence/graphic	죽음, 폭력 또는 신체적 상해를 그래픽으로 자세하게 묘사하는 콘텐츠입니다.

LLM 답변 받은 후의 post processing

- Azure 는? Azure AI 콘텐츠 RAI
- Google 은? Gemini 안전 세팅

Groundedness, relevance testing

- RAGAS library 로 실습 – 노트북
  - https://docs.ragas.io/en/stable/concepts/metrics/index.htm
- 생성
  - Faithfulness : 사실인가?
  - Answer relevancy : 질문과의 관련성
- 검색
  - Context precision : 가져온 정보의 signal to noise – 랭킹
  - Context recall: 가져온 정보로 충분한가
- 그 외:
  - Context relevancy : 필요한 정보만 있는가
  - Context entities recall: 가져온 정보중 몇 개를 썼는가
  - Answer semantic similarity: 답변과 ground truth 유사도

후처리 코드 실습

- 후처리 코드 실습

Part 4-7.

RAG 아키텍쳐 샘플

RAG 아키텍쳐 샘플

아키텍쳐 보기

RAG 시스템 타입

아키텍처 타입

기본

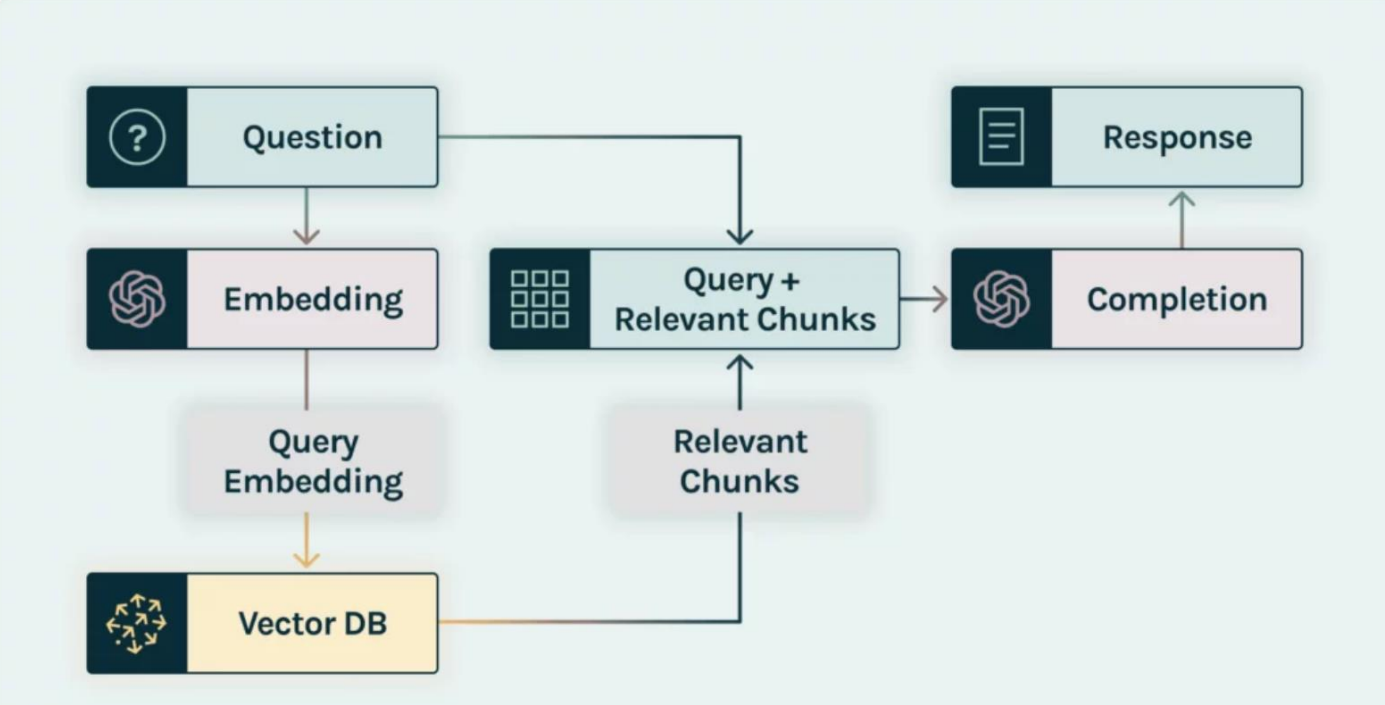
질문이 들어옴 → 답변을 열심히 찾아서 정리해옴 → 아나운서에게 이거 보고 답변하라고 함.

- "답변에 필요한 정보를 열심히 찾아서"
- "정리해서"
  - 물어볼 만한 답변을 미리 폴더에 잘 정리해두기
  - 사연도 정리해두기
  - 유튭 댓글도 리서치 해서 잘 정리해두기
  - 다시 한 번 추출해서 정리하기

RAG 시스템 타입

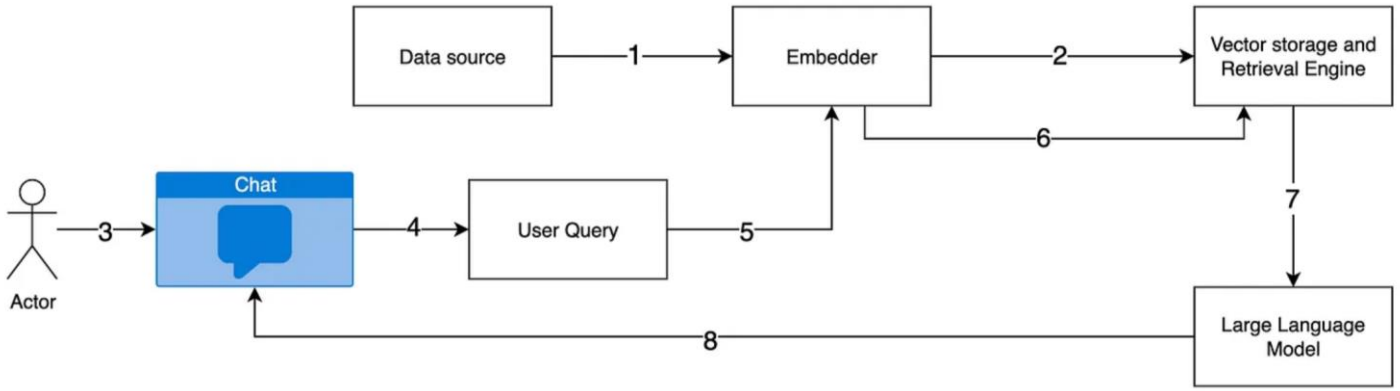
임베딩	파일링 시스템 코드
벡터 DB	파일링 캐비닛
프리 프로세싱	방송 나가기 전에 미리미리 자료 다 정리해둠
포스트 프로세싱	유튜버/아나운서가 녹화해둔 답변을 편집

RAG 시스템 타입 - 이해하기



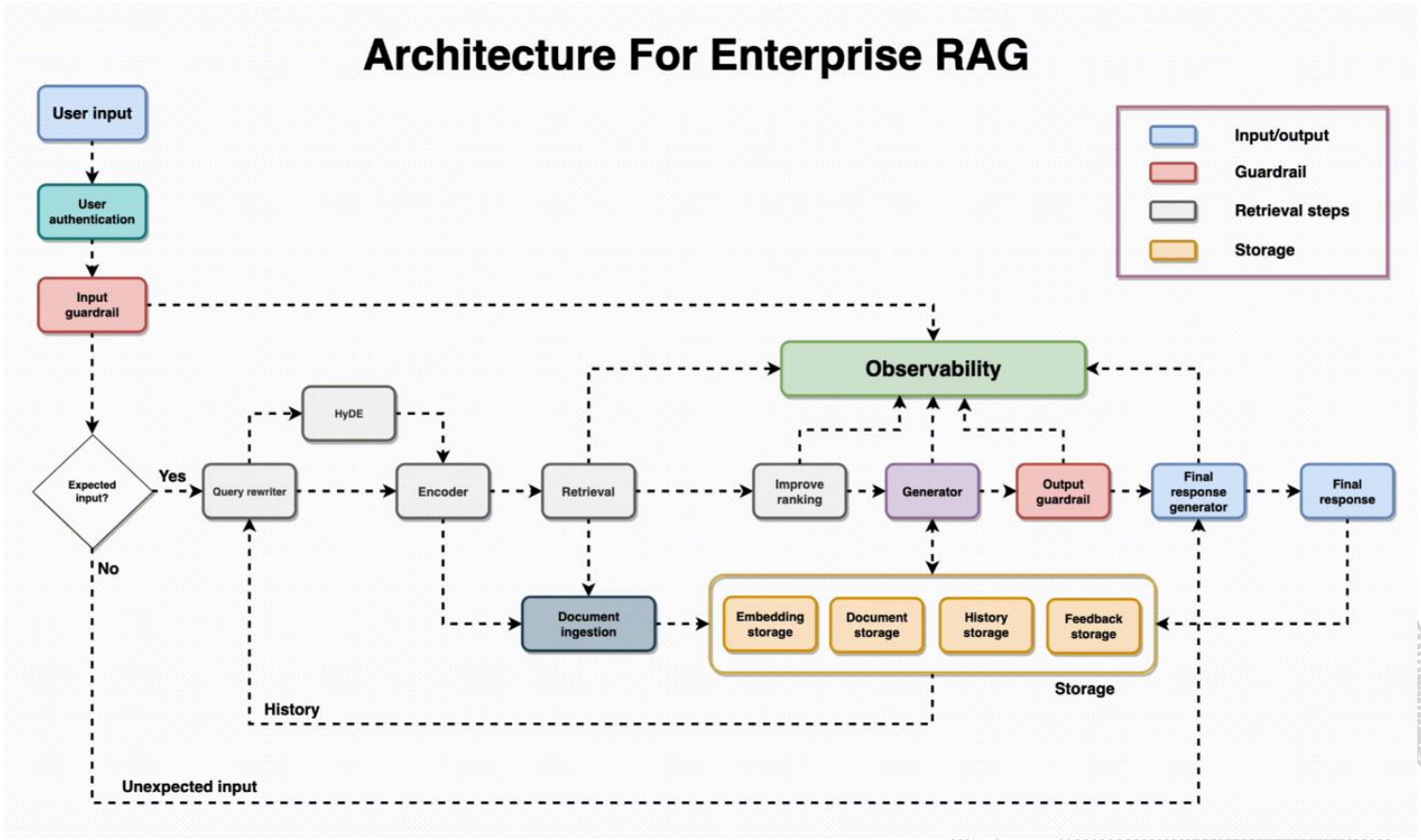
https://truera.com/ai-quality-education/generative-ai-rags/what-is-retrieval-augmented-generation-rag-for-llms/

RAG 시스템 타입 - 이해하기



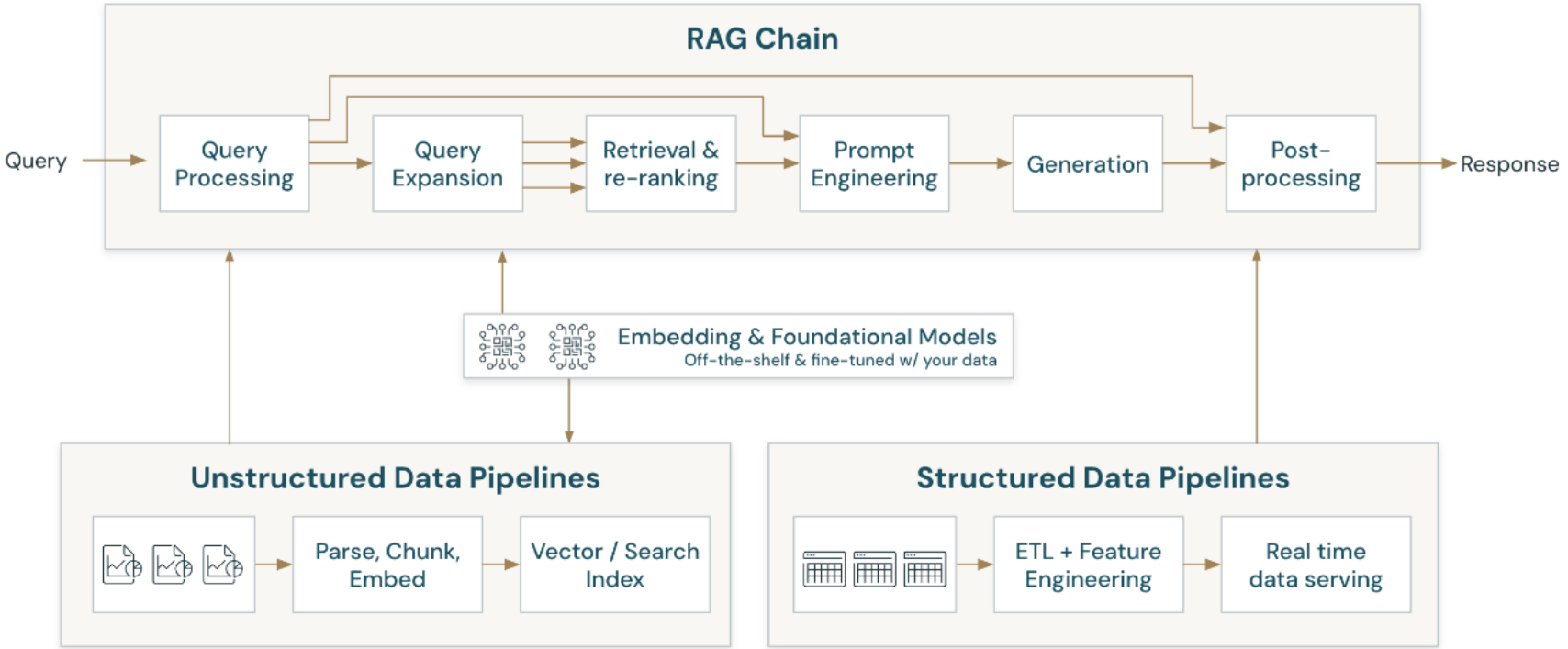
https://medium.com/@pandey.vikesh/rag-ing-success-guide-to-choose-the-right-components-for-your-rag-solution-on-aws-223b9d4c7280

RAG 시스템 타입 - 이해하기



https://www.rungalileo.io/blog/mastering-rag-how-to-architect-an-enterprise-rag-system

RAG 시스템 타입 - 이해하기



https://docs.databricks.com/en/generative-ai/retrieval-augmented-generation.html

Part 4-8.
샘플 앱

샘플 앱

노트북에서 앱 만들어보기

샘플 앱 코드 노트북

- 샘플 앱 코드
- 미니 벡터DB 와 답변 템플릿 코드