

Implementation and Advancement of RAG Using LangChain

LangChain을 활용한 RAG 구현 및 고도화

RAG (Retrieval-Augmented Generation)

목차 RAG (Retrieval-Augmented Generation)

1. RAG 이해를 위한 기초 이론
2. RAG (Retrieval-Augmented Generation)

1. RAG 이해를 위한 기초 이론



- 텍스트 생성, 이해, 번역 등 다양한 NLP 작업을 수행
- 매우 큰 텍스트 데이터 세트에서 학습
- 높은 정확도와 자연스러운 언어 생성

Gemini



you.com



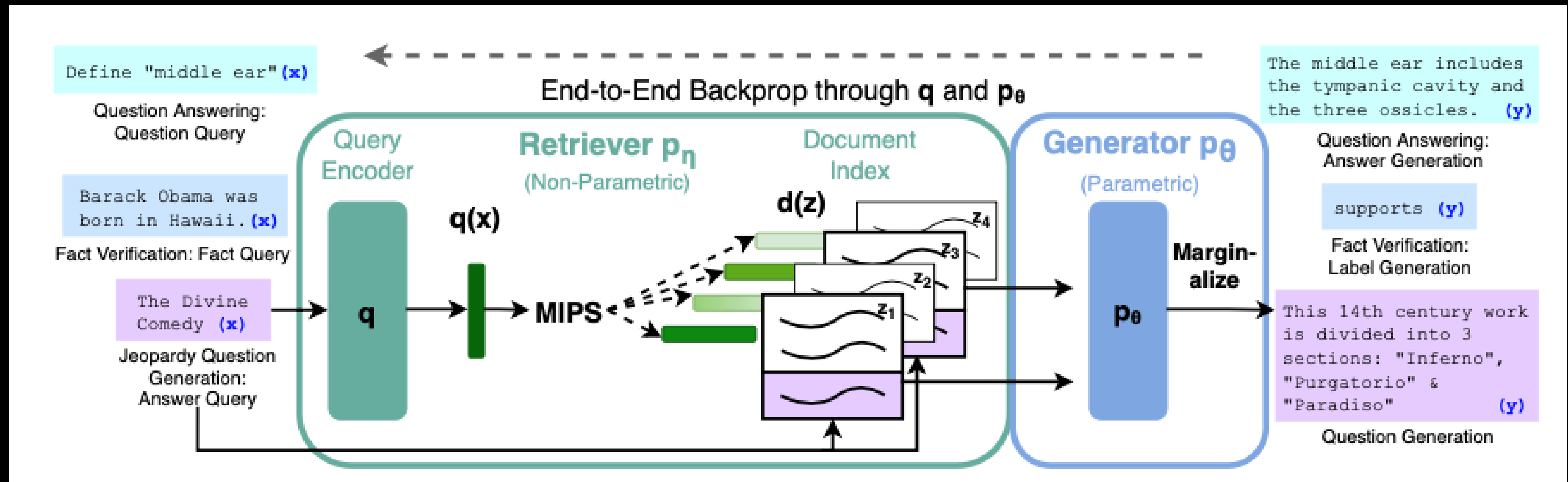
- 대부분의 Task에서 강점을 보임
- Data만 있다면 대부분의 task에 적용 가능
- 학습 방식과 프롬프트에 따라 다양한 Task에 바로 적용 가능(Zero shot learning)



- 수 많은 학습 데이터를 요구함 (수십TB 급)
- 훈련 데이터를 모두 암기하는 것은 아님
- 학습하지 않은 Open domain QA 에 약한 모습을 보임
- 일부 상식적인 부분을 이해하지 못함
- Fine-tuning에 매우 큰 비용 소모



- RAG(Retrieve-Augmented Generation)
 - 정보 검색(retrieval)과 응답 생성(generation)을 결합한 모델
 - 사용자의 질문이 주어지면, Retriever는 관련된 정보나 문서를 데이터베이스에서 검색
 - 검색된 정보로 질문에 대한 답변 생성
 - 보다 풍부하고 정확한 정보를 제공 가능





- 환각 현상(Hallucination) 감소
- 도메인 적응성 개선
- Open domain QA성능 향상
- 참고한 Knowledge base가 적절한지 판단 가능
- 정보 검색에 강함



RAG vs. Fine Tuning

LangChain을 활용한
RAG 구현 및 고도화

RAG

1. RAG 이해를 위한 기초
이론

	RAG	Fine-tuning
장점	외부 지식을 추가하여 정확도, 신뢰도를 높일 수 있다.	특정 작업에 대한 성능을 높일 수 있다. (예: 요약 작업 성능 향상)
	새로운 정보를 추가할 때, 추가 학습이 필요하지 않다.	매우 구체적인 태스크에 유용하며, 일관된 품질을 제공한다.
단점	검색된 정보의 품질에 의존한다.	많은 양의 학습 데이터가 필요하다.
	검색 시스템을 동원하여야 하기 때문에 더 많은 컴퓨팅 자원을 사용할 수 있다.	학습한 데이터 외의 질의에는 좋은 답변을 얻을 수 없다.

2. RAG (Retrieval-Augmented Generation)



RAG (Retrieval-Augmented Generation, 검색 증강 생성) 기법을 활용해서 문서를 근거로 하는 QA 챗봇을 개발할 수 있다.



올해 키오스크 시장의 전망을 알려줘



주제: 국내 키오스크(무인정보단말기) 이용 실태조사 및 개선 방안에 대한 분석

소개: 최근 시장조사국 시장감시팀의 조사 결과, 전반적으로 '외식업'에서는 많은 피해가 발생하고 있다고 나왔습니다. 이러한 상황은 모든 연령대가 외식업 키오스크를 가장 많이 이용하기 때문으로 추정됩니다.

주요 내용:

- '키오스크 기능 표준화' 필요성과 연령대별 불편사례를 보여줍니다.
 - [표7-1-1] '키오스크 기능 표준화' 필요성을 나타내는 테이블을 제시하며, 20대부터 60대까지 모든 연령대에게 도움이 될 것이라고 답변하는 명의 비율이 가장 높습니다.
 - [표7-1-2] 연령대별 키오스크 불편 사례 1~3순위를 나타내는 테이블도 제시하며, '기기 오류', '조작 어려움' 등의 불편함이 20대부터 60대까지 모든 연령대에게 있다고 나타냅니다.
- 디지털 약자층을 포함한 모든 소비자의 키오스크 접근성과 편의성 확보를 위해서는 점진적으로 개선이 필요하다고 설명합니다.

키오스크(무인정보단말기) 이용 실태조사

VII 문제점 및 개선방안

1. 키오스크의 기능 표준화를 통한 이용 편의성 제고

- (문제점) 키오스크 관련 소비자 불만·피해 발생 원인 분석 결과 키오스크 '이용 미숙(18건/18.8%)', '이용법 안내 부족(13건/13.5%)'으로 인한 사례가 많았음.
- 소비자 설문조사 결과, 전 연령대가 키오스크를 통한 주문 순서 등 기능을 표준화하면 키오스크 사용에 도움이 될 것이라고 응답함.

[표7-1-1] '키오스크 기능 표준화' 필요성(N=500)

(단위: 명, %)

키오스크 기능 표준화	20대	30대	40대	50대	60대~
도움이 될 것이다	79	87	77	87	94
보통이다	14	12	17	11	5
도움이 되지 않을 것이다	7	1	6	2	1

- 또한 전 연령대가 키오스크의 조작이 어렵다고 응답하였는데, 이 역시 업종별 또는 브랜드별로 키오스크의 화면 구성 및 조작 방법 등이 모두 다르기 때문으로 판단됨.

[표7-1-2] 연령대별 키오스크 불편 사례 1~3순위 (중복응답, N=233)

(단위: %)

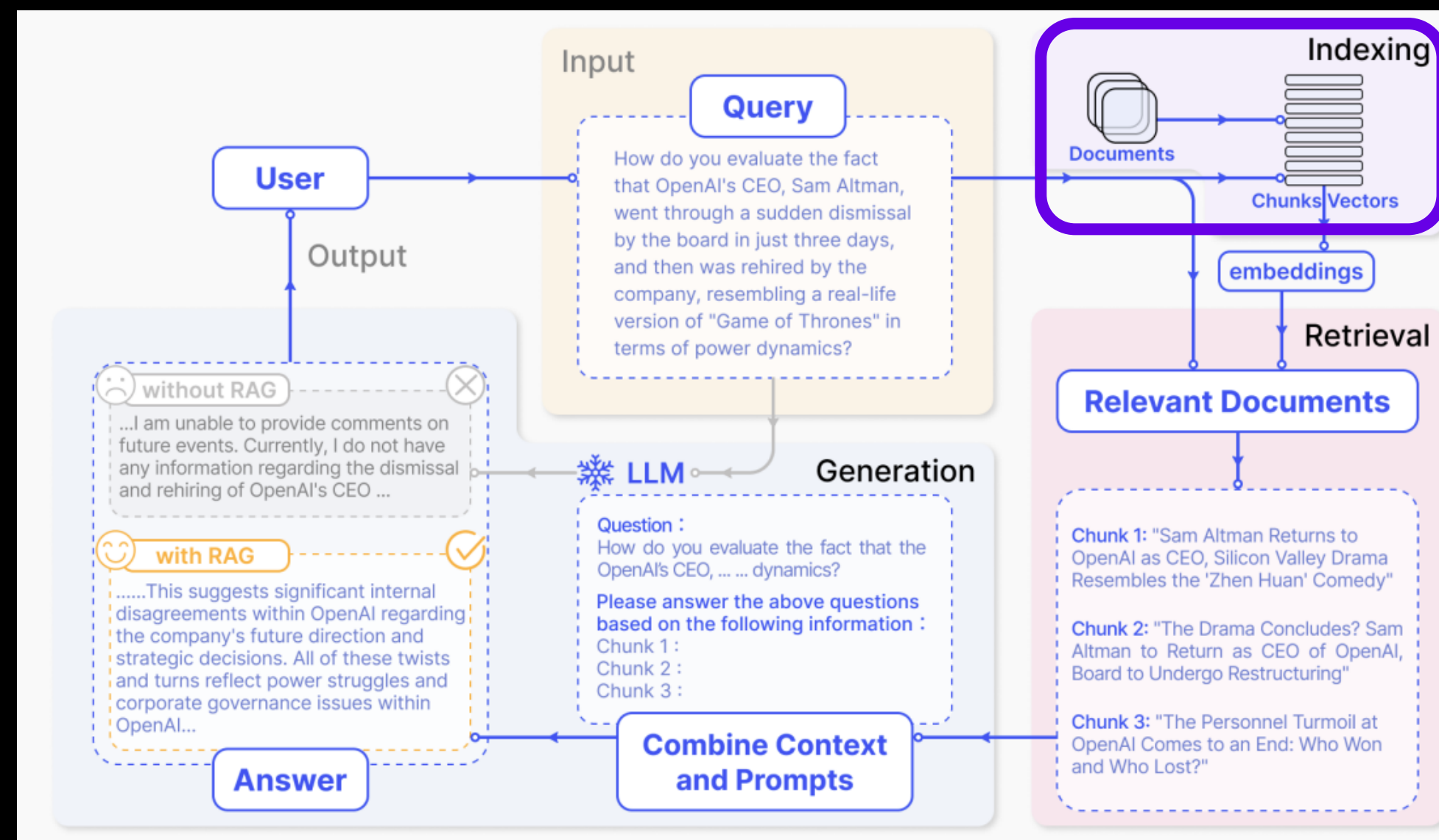
구분	20대	30대	40대	50대	60대~
1순위	기기 오류 (70.8)	뒷사람 눈치 (52.9)	뒷사람 눈치 (66.7)	뒷사람 눈치 (61.5)	조작 어려움 (53.6)
2순위	뒷사람 눈치 (47.9)	조작 어려움 (51.0)	검색 어려움 (46.2)	조작 어려움 (51.3)	검색 어려움 (42.9)
3순위	조작 어려움 (35.4)	기기 오류 (45.1)	조작 어려움 (41.0)	할인·적립 (38.5)	뒷사람 눈치 (41.1)



이름이 내포한 뜻과 같이,
검색을 통해 LLM의 생성 과정을 증강하는 기법

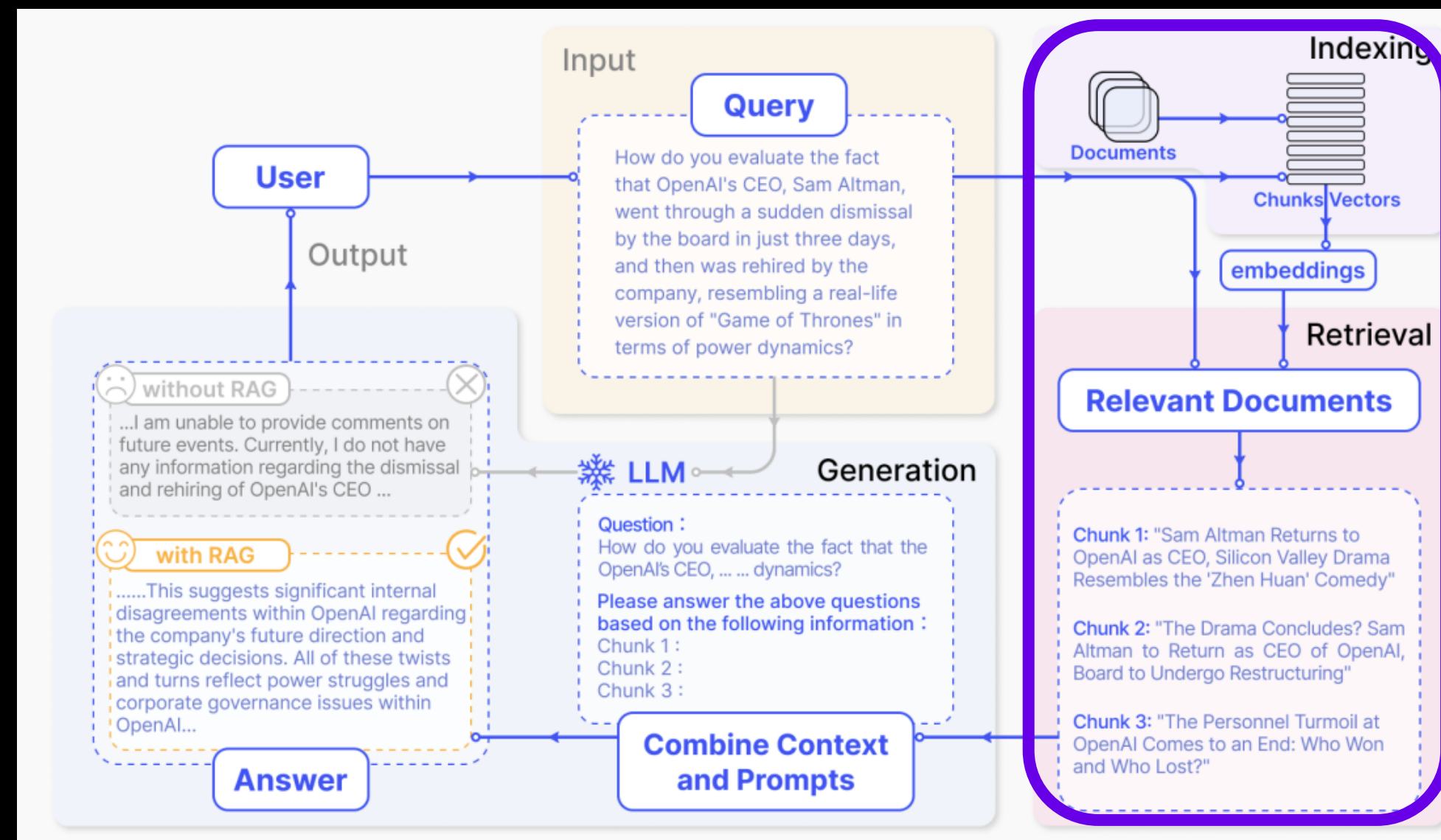


- RAG의 대략적인 진행 과정은 다음과 같다:
 1. 사전에 정보를 담고 있는 문서를 일정 크기(Chunk)로 나눠서 Vector DB로 저장한다.
 2. 사용자의 입력과 유사한 K개의 문서를 검색 (Retrieve) 한다.
 3. 사용자의 입력에 검색된 문서를 더해 증강된 (Augmented) 프롬프트를 LLM에 입력한다.
 4. LLM은 검색된 문서 정보를 바탕으로 상대적으로 더 정확한 답변을 생성한다.



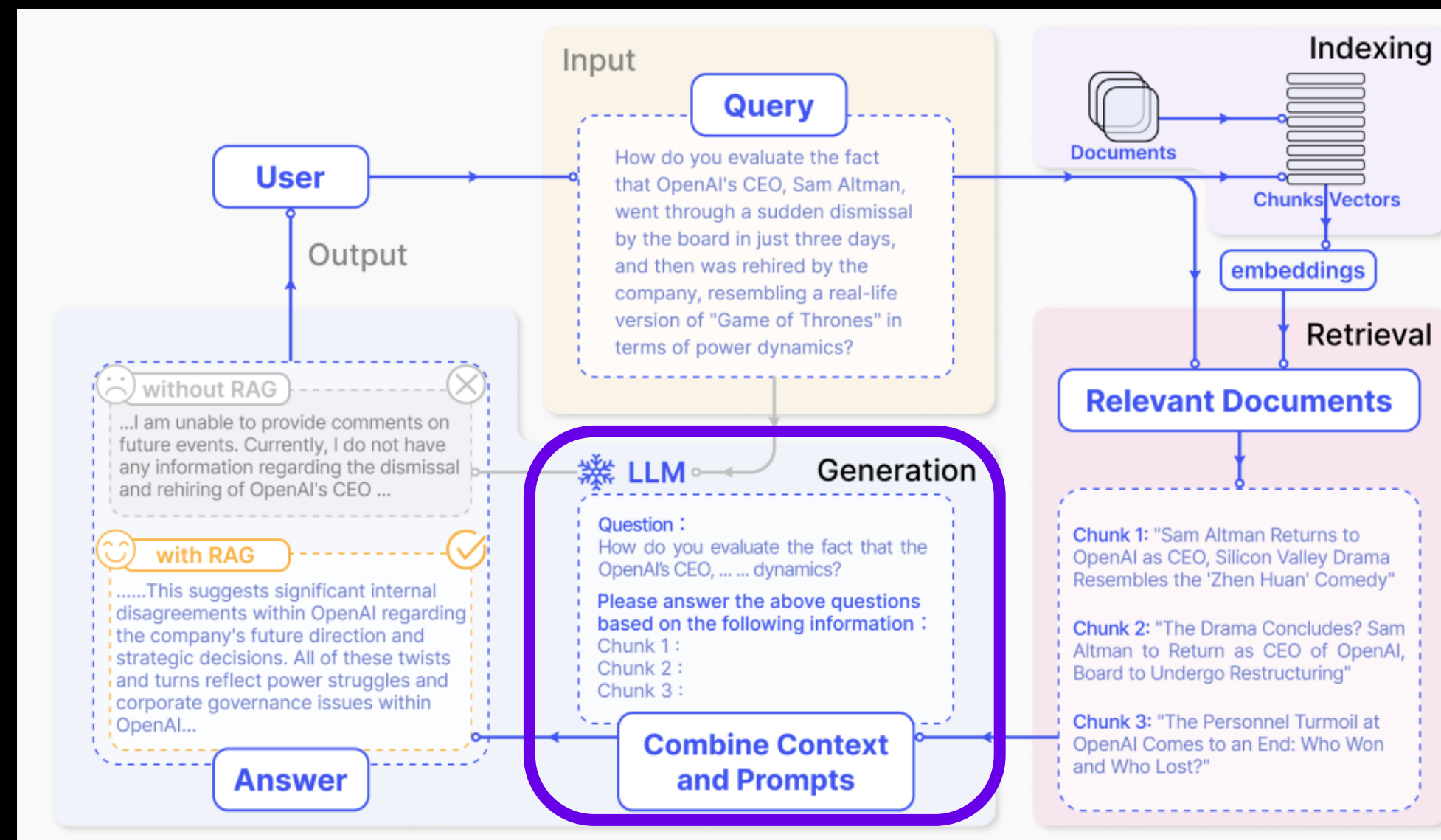


- RAG의 대략적인 진행 과정은 다음과 같다:
 1. 사전에 정보를 담고 있는 문서를 일정 크기(Chunk)로 나눠서 Vector DB로 저장한다.
 2. 사용자의 입력과 유사한 K개의 문서를 검색 (Retrieve) 한다.
 3. 사용자의 입력에 검색된 문서를 더해 증강된 (Augmented) 프롬프트를 LLM에 입력한다.
 4. LLM은 검색된 문서 정보를 바탕으로 상대적으로 더 정확한 답변을 생성한다.



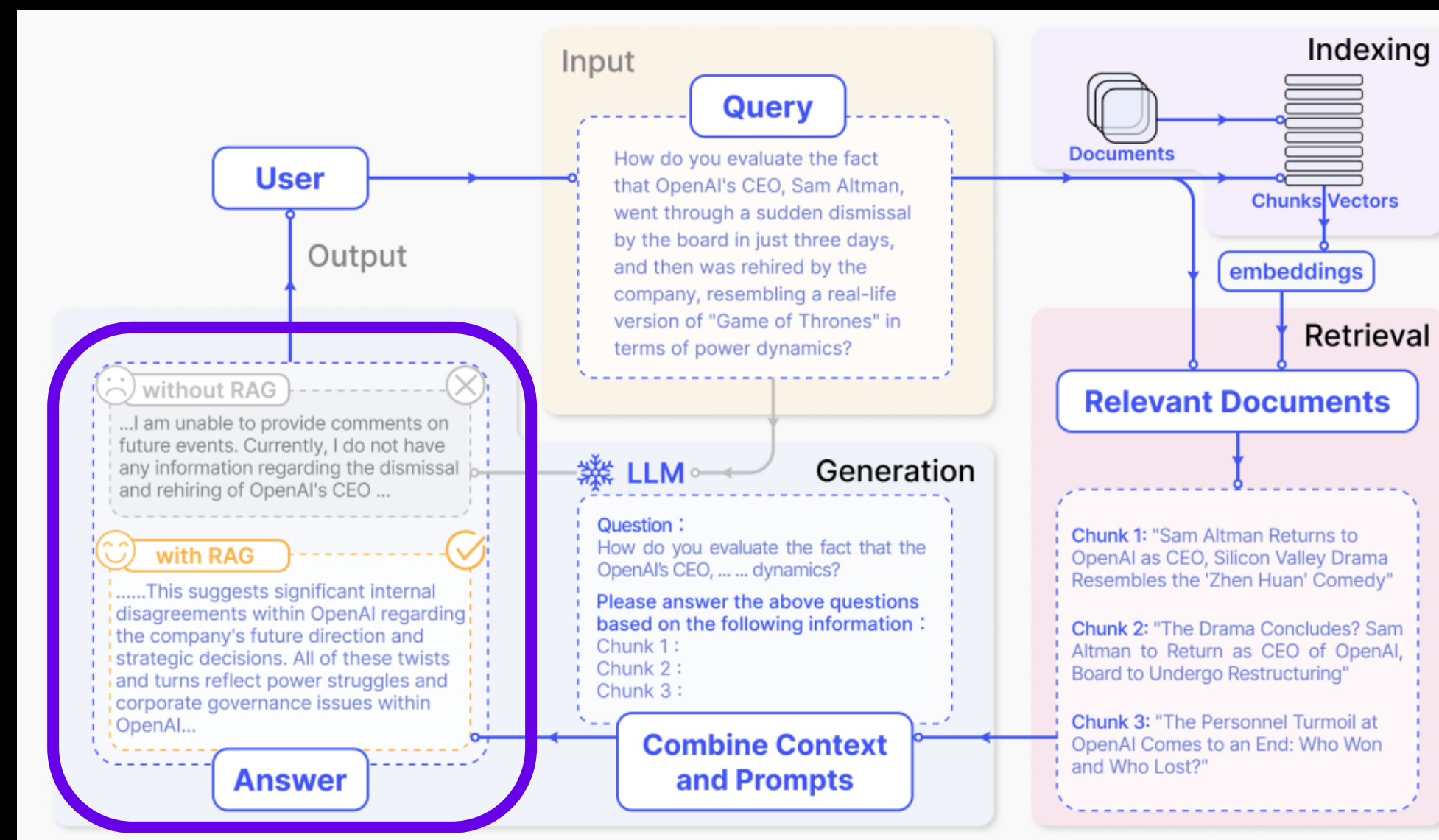


- RAG의 대략적인 진행 과정은 다음과 같다:
 1. 사전에 정보를 담고 있는 문서를 일정 크기(Chunk)로 나눠서 Vector DB로 저장한다.
 2. 사용자의 입력과 유사한 K개의 문서를 검색 (Retrieve) 한다.
 3. 사용자의 입력에 검색된 문서를 더해 증강된 (Augmented) 프롬프트를 LLM에 입력한다.
 4. LLM은 검색된 문서 정보를 바탕으로 상대적으로 더 정확한 답변을 생성한다.





- RAG의 대략적인 진행 과정은 다음과 같다:
 1. 사전에 정보를 담고 있는 문서를 일정 크기(Chunk)로 나눠서 Vector DB로 저장한다.
 2. 사용자의 입력과 유사한 K개의 문서를 검색 (Retrieve) 한다.
 3. 사용자의 입력에 검색된 문서를 더해 증강된 (Augmented) 프롬프트를 LLM에 입력한다.
 4. LLM은 검색된 문서 정보를 바탕으로 상대적으로 더 정확한 답변을 생성한다.

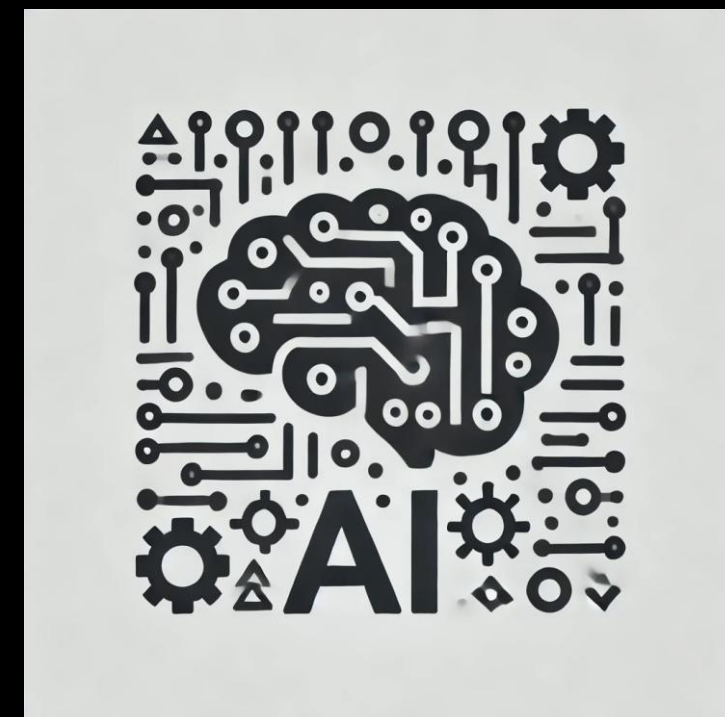
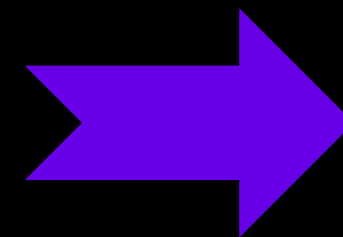




- LLM을 비롯하여 자연어를 처리하는 AI는 자연어를 있는 그대로 입력 받는 것이 아니라, 특정 알고리즘을 따라 정해진 크기의 벡터로 벡터화 된 정보를 입력 받는다.

안녕하세요,
오늘은 기분이 좋은 날입니다.

(입력 불가능)





- LLM을 비롯하여 자연어를 처리하는 AI는 자연어를 있는 그대로 입력 받는 것이 아니라, 특정 알고리즘을 따라 정해진 크기의 벡터로 벡터화 된 정보를 입력 받는다.





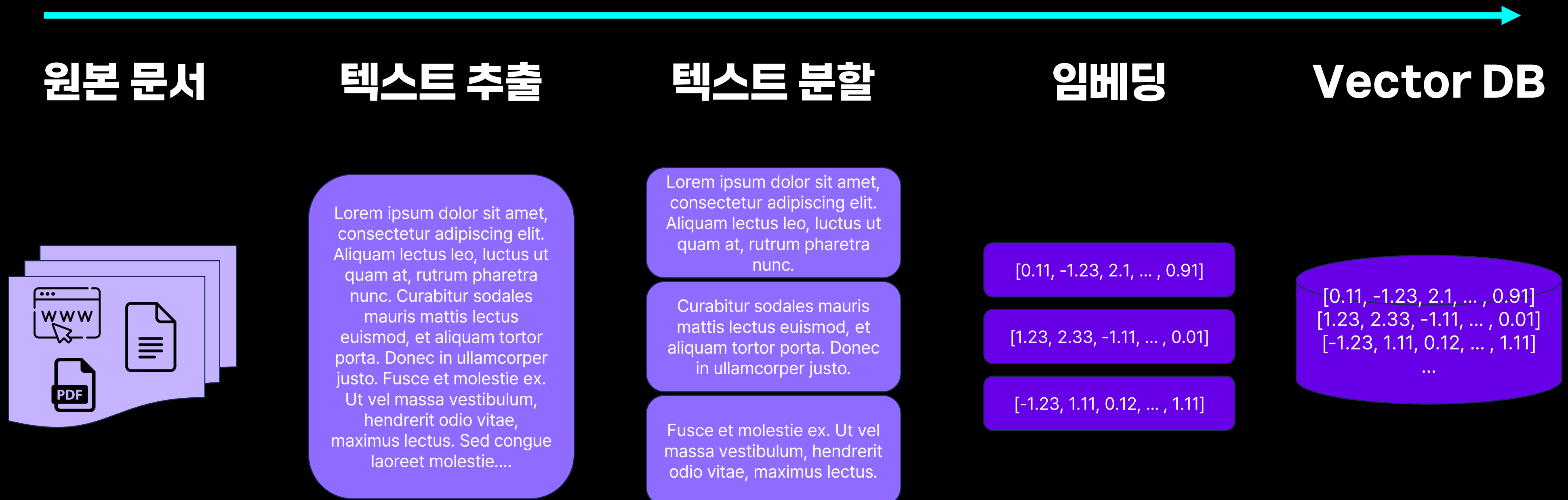
Vector DB

LangChain을 활용한
RAG 구현 및 고도화

RAG

2. RAG (Retrieval-
Augmented
Generation)

- Vector DB는 정보를 담고 있는 문서를 전처리 및 분할한 후 **Embedding** 과정을 거친 **Vector**를 저장한 DB를 의미한다.

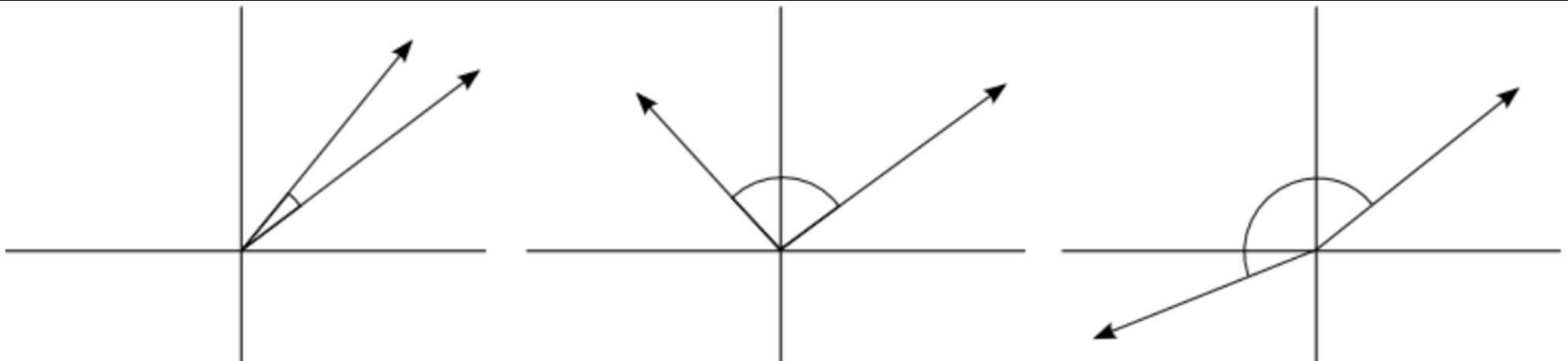




- 필요한 정보를 검색하는 작업
- 데이터베이스, 인터넷, 또는 다른 정보 저장소에서 관련 정보를 찾아내는 과정
- 사용자의 쿼리에 가장 잘 맞는 데이터를 식별하고 추출하는 기술과 알고리즘
- 웹 검색 엔진, 디지털 도서관, 온라인 데이터베이스, 정보 검색 시스템 등 다양한 분야에서 중요한 역할 수행



- 코사인 유사도

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$


- L2 distance

$$\begin{aligned} d(\mathbf{p}, \mathbf{q}) &= d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}. \end{aligned}$$

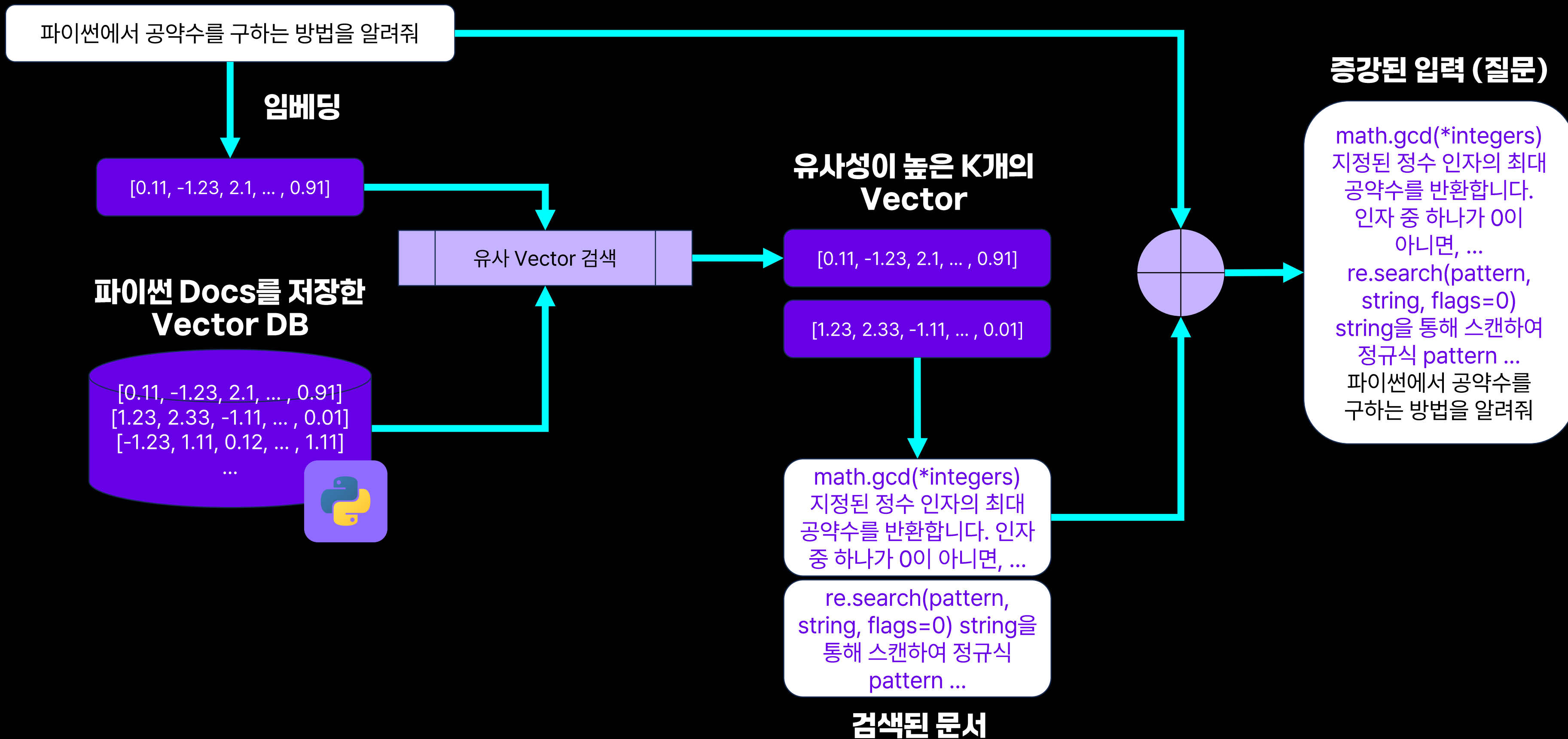


Retrieval (검색) 예시

LangChain을 활용한
RAG 구현 및 고도화

RAG

2. RAG (Retrieval-
Augmented
Generation)





Retrieval (검색) 예시

LangChain을 활용한
RAG 구현 및 고도화

RAG

2. RAG (Retrieval-Augmented Generation)

