Retrieval-Augmented Generation (RAG)

Fast campus

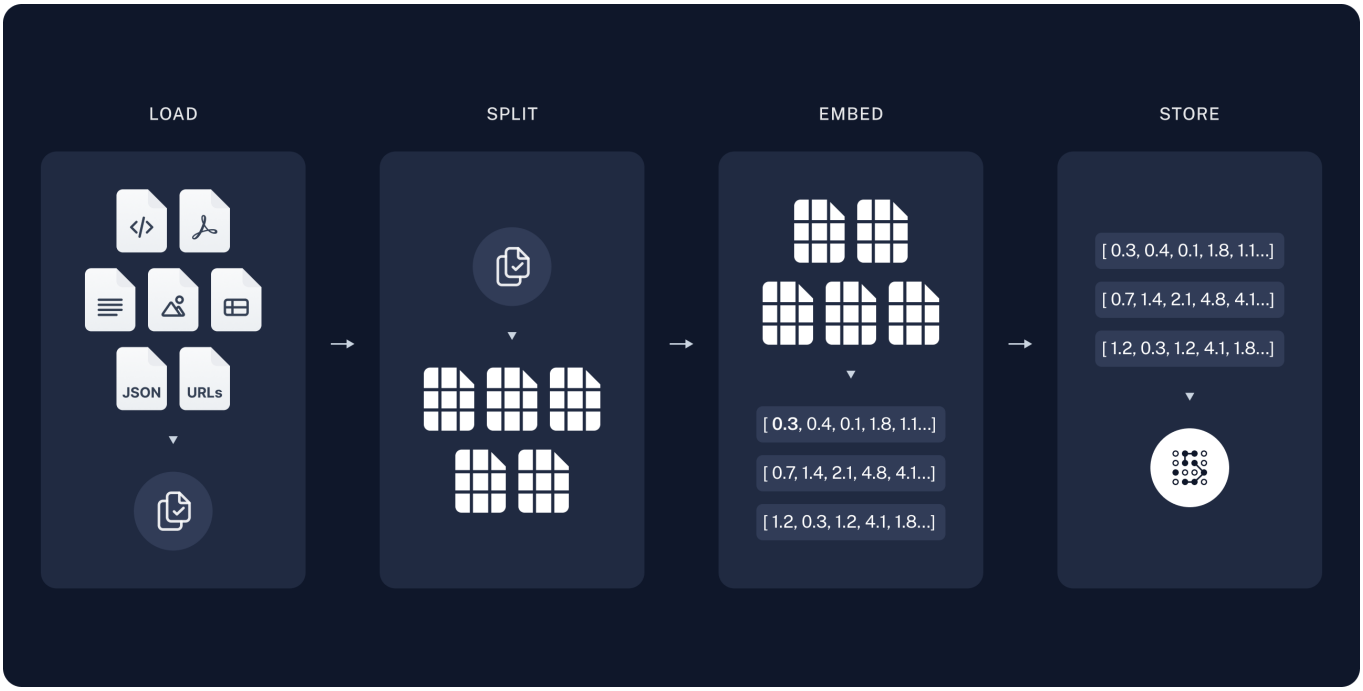
Retrieval-Augmented Generation(RAG)는 자연어 처리(NLP) 분야에서의 혁신적인 기술로, 기존의 언어 모델의 한계를 넘어서 **정보 검색**과 **생성을 통합하는 방법론**입니다.

기본적으로, RAG는 풍부한 정보를 담고 있는 **대규모 문서 데이터베이스에서 관련 정보를 검색**하고, 이를 통해 언어 **모델이 더 정확하고 상세한 답변을 생성**할 수 있게 합니다.

예를 들어, **최신 뉴스 이벤트나 특정 분야의 전문 지식과 같은 주제**에 대해 물어보면, **RAG는 관련 문서를 찾아 그 내용을 바탕으로 답변**을 구성합니다.

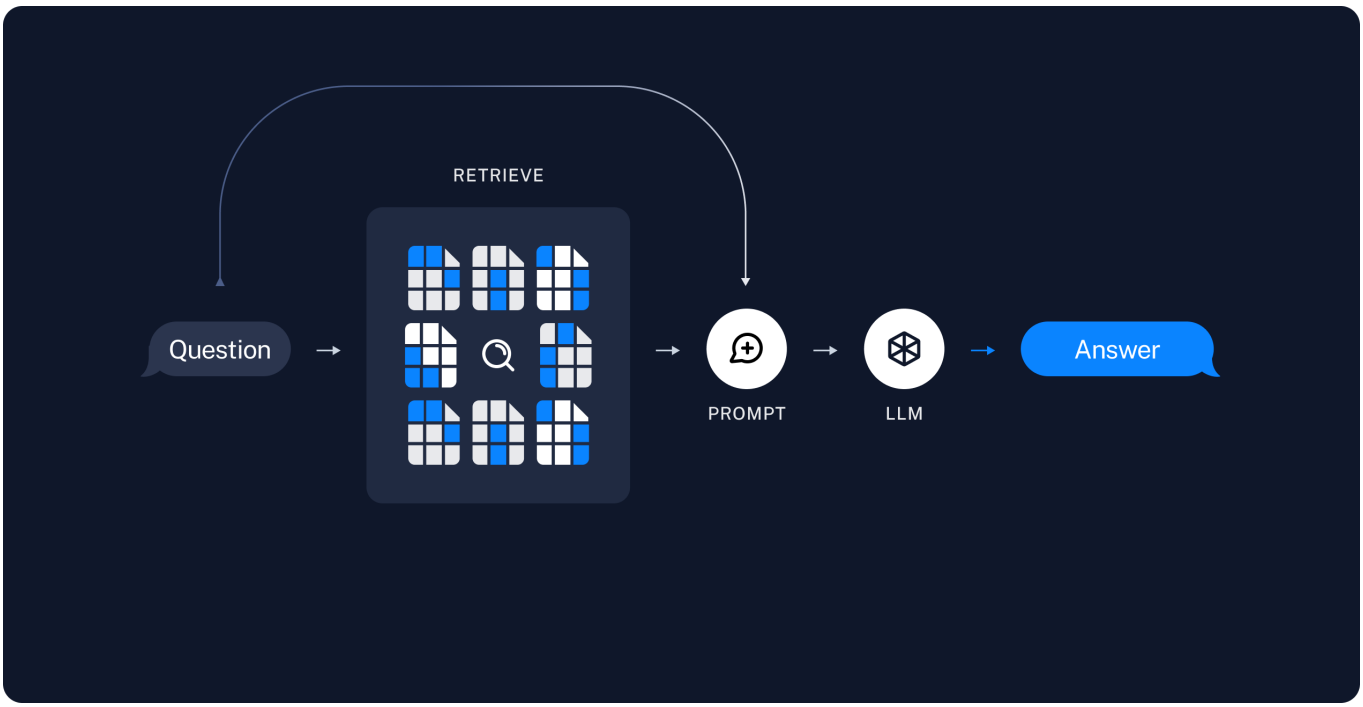
RAG의 8단계 프로세스

사전 준비단계



1. 도큐먼트 로드 (Document Loader): 이 단계에서는 외부 데이터 소스에서 필요한 문서를 로드하고 초기 처리를 합니다. 이것은 마치 책을 여러 권 챙겨 도서관에서 공부하는 것과 비슷합니다. 학생이 공부하기 전에 필요한 책들을 책장에서 골라오는 과정입니다.
2. 텍스트 분할 (Text Splitter): 로드된 문서를 처리 가능한 작은 단위로 분할합니다. 큰 책을 챕터별로 나누는 것과 유사합니다.
3. 임베딩 (Embedding): 각 문서 또는 문서의 일부를 벡터 형태로 변환하여, 문서의 의미를 수치화합니다. 이는 책의 내용을 요약하여 핵심 키워드로 표현하는 것과 비슷합니다.
4. 벡터스토어(Vector Store) 저장: 임베딩된 벡터들을 데이터베이스에 저장합니다. 이는 요약된 키워드를 색인화하여 나중에 빠르게 찾을 수 있도록 하는 과정입니다.

런타임(RunTime 단계)



5. 검색기 (Retriever): 질문이 주어지면, 이와 관련된 벡터를 벡터 데이터베이스에서 검색합니다. 질문에 가장 잘 맞는 책의 챕터를 찾는 것과 유사합니다.
6. 프롬프트 (Prompt): 검색된 정보를 바탕으로 언어 모델을 위한 질문을 구성합니다. 이는 정보를 바탕으로 어떻게 질문할지 결정하는 과정입니다.

7. **LLM (Large Language Model)**: 구성된 프롬프트를 사용하여 언어 모델이 답변을 생성합니다. 즉, 수집된 정보를 바탕으로 과제나 보고서를 작성하는 학생과 같습니다.
8. **체인(Chain) 생성**: 이전의 모든 과정의 하나의 파이프라인으로 묶어주는 체인(Chain) 을 생성합니다.