(10주차) RAG(Retrieval-Augmented Geration)

명지대학교
MYONGJI UNIVERSITY

RAG(Retrieval-Augmented Generation) 개요

RAG 개요

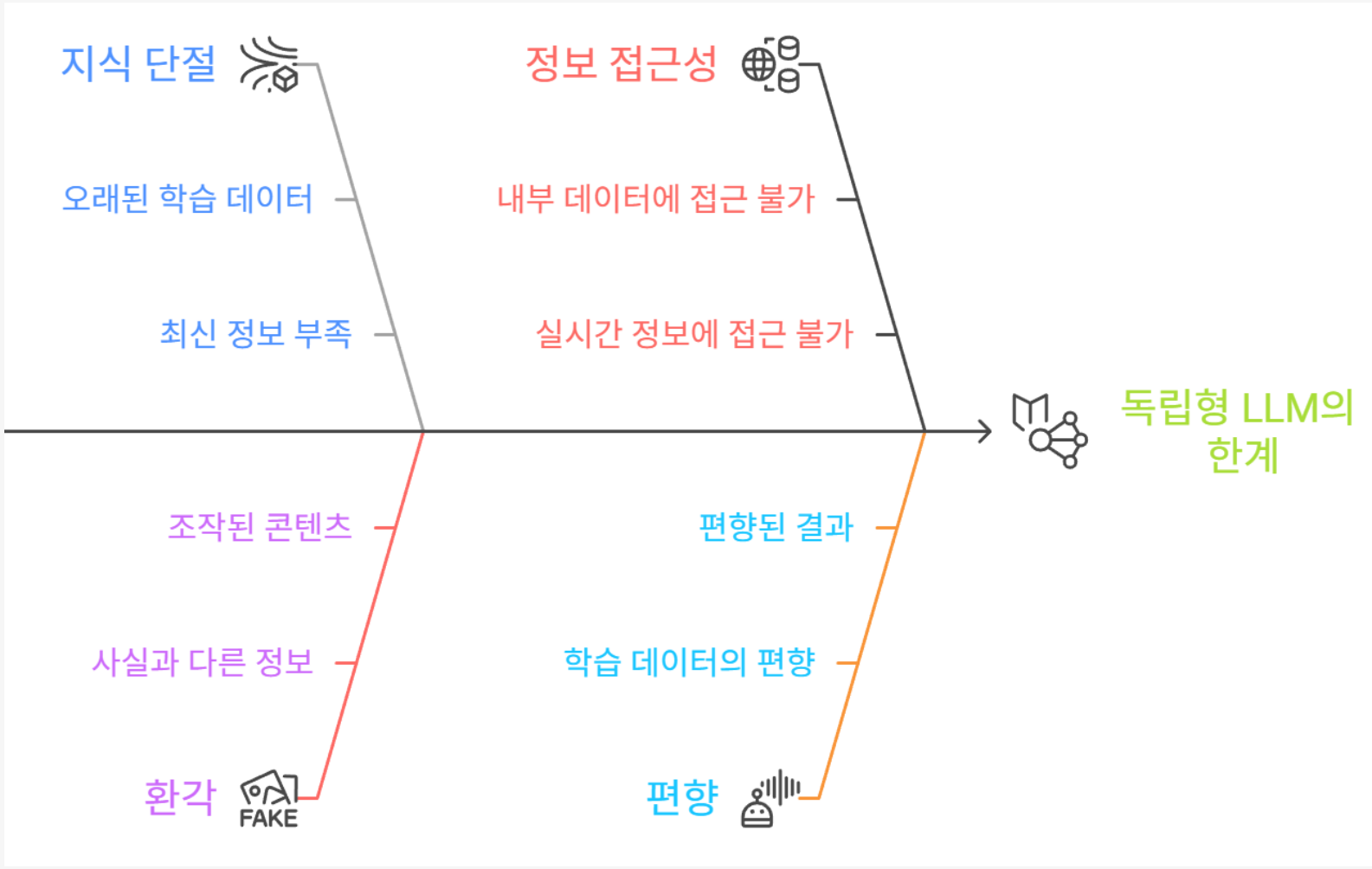
RAG(Retrieval-Augmented Generation)란?

외부 정보 검색 기능을 LLM과 통합한
하이브리드 인공지능 아키텍처

텍스트 생성의 정확성, 최신성, 검증 가능성을 향상

RAG(Retrieval-Augmented Generation) 개요

LLM의 한계



RAG(Retrieval-Augmented Generation) 개요

독립형 LLM의 한계

LLM

- **Fine-tuning:** 기존 딥러닝 모델의 weight를 조정하여 원하는 용도의 모델로 업데이트 합니다.
- **N-shot Learning:** 0개~n개의 출력 예시를 제시하여, 딥러닝이 용도에 알맞은 출력을 하도록 조정합니다.
- **In-context Learning:** 문맥을 제시하고, 이 문맥 기반으로 모델이 출력하도록 조정합니다.

RAG(Retrieval-Augmented Generation) 개요

RAG의 구성요소

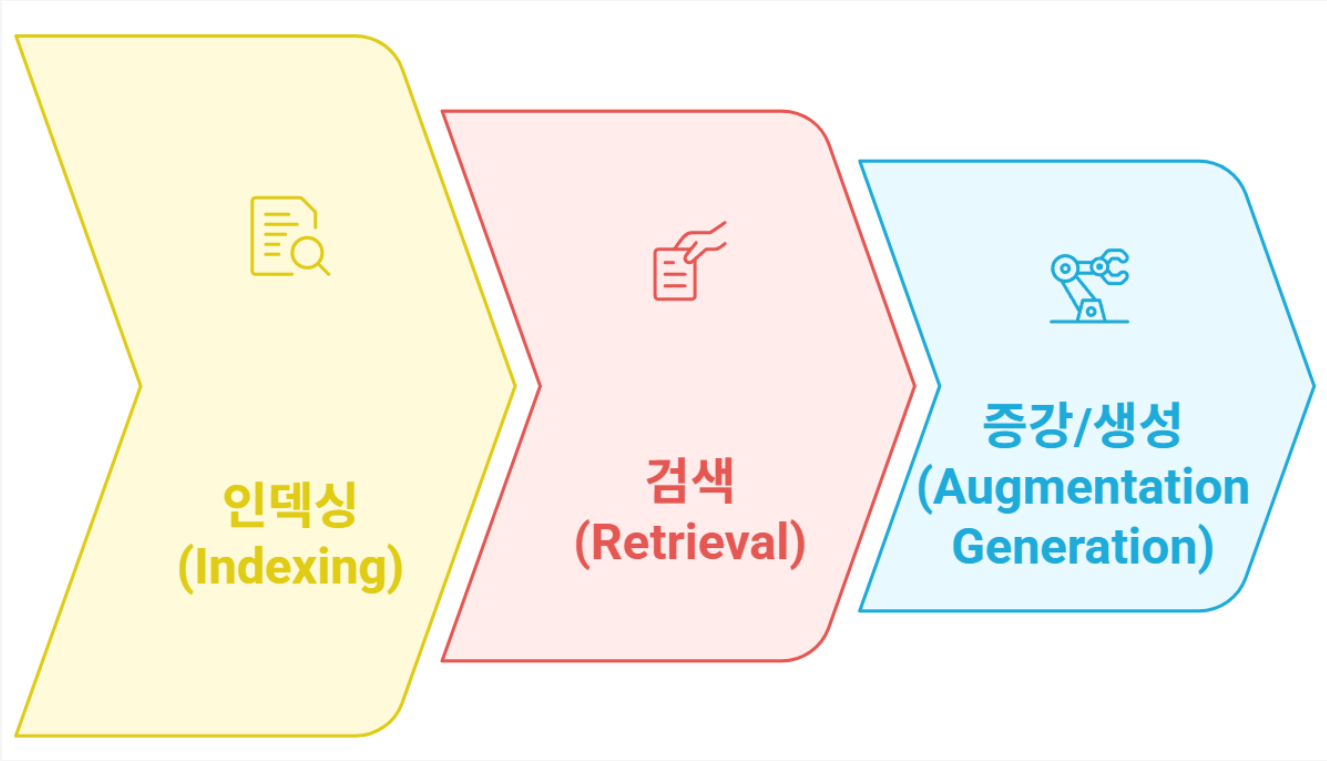
검색기(Retriever), 생성기(Generator), 둘 사이의 상호작용을 포함하는 증강(Augmentation)으로 구성

1	**질의 입력** 사용자의 질의 입력
2	**검색기** 입력된 질의(질문)에 대해 외부 지식 소스에서 관련 문서/정보 검색 데이터베이스, 위키피디아 등 문서 코퍼스(Corpus).
3	**생성기** 원본 질의와 검색된 컨텍스트(정보)를 입력받아, 일관성 있는응답으로 합성 어텐션(Attention) 메커니즘 등을 사용하여 검색된 정보를 효과적으로 통합
4	**증강** 검색된 외부 지식을 활용하여 생성기로 전달되는 입력을 증강 생성되는 결과물의 사실 기반 정도 및 관련성 향상

RAG(Retrieval-Augmented Generation) 개요

RAG의 워크플로우

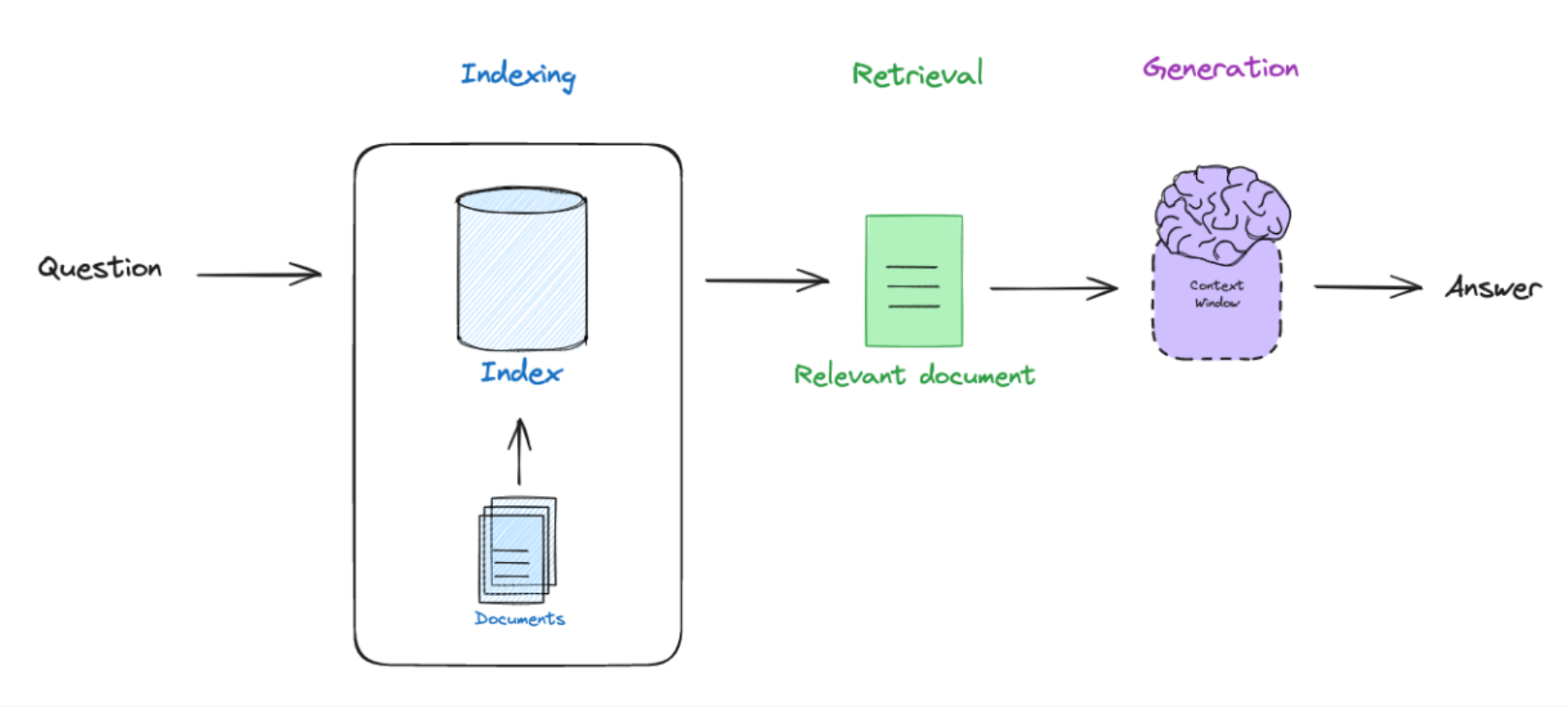
RAG의 워크플로우는 크게 인덱싱(Indexing), 검색(Retrieval) 및 증강/생성(Augmentation and Generation) 절차로 수행



RAG(Retrieval-Augmented Generation) 개요

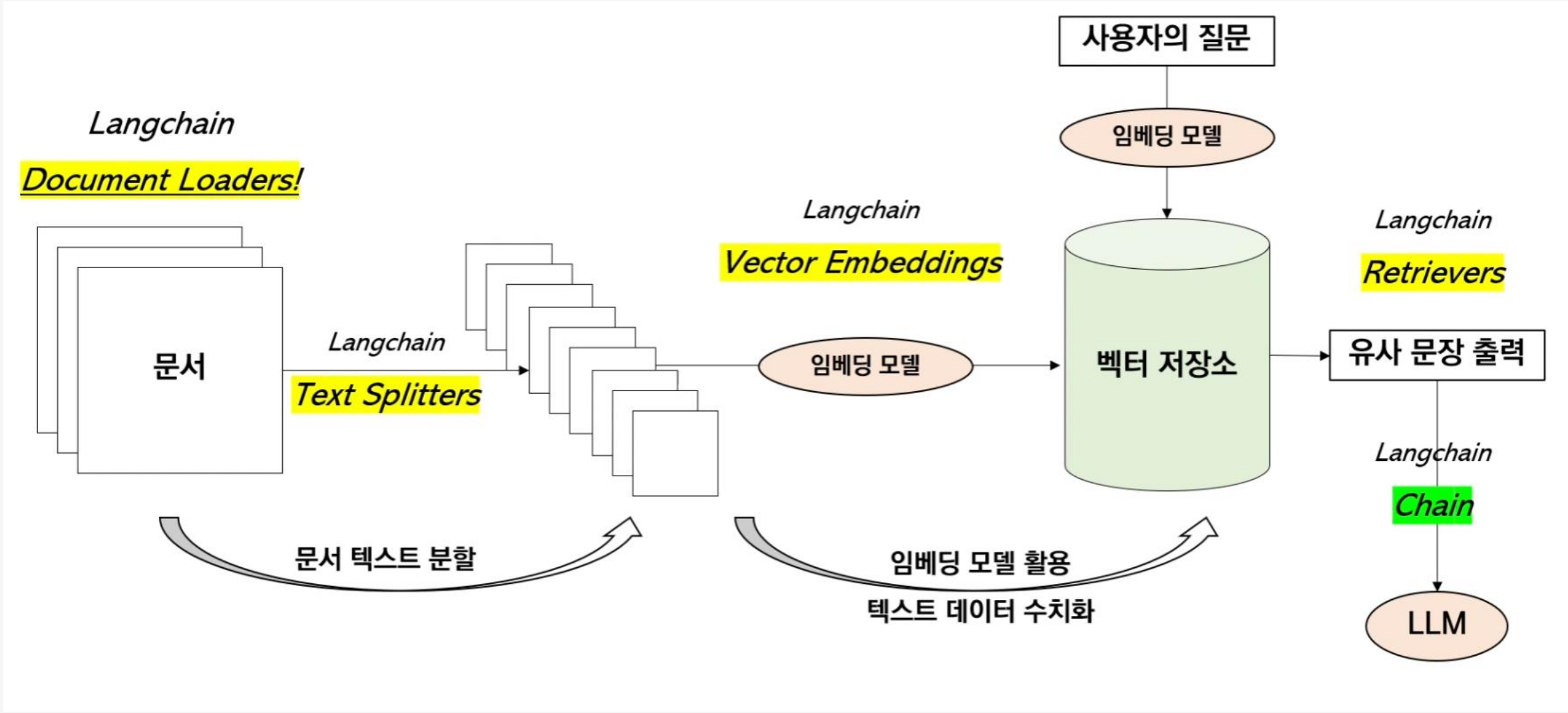
RAG의 워크플로우

RAG의 워크플로우는 크게 인덱싱(Indexing), 검색(Retrieval) 및 증강/생성(Augmentation and Generation) 절차로 수행



RAG(Retrieval-Augmented Generation) 개요

RAG의 워크플로우-LangChain 예시



RAG(Retrieval-Augmented Generation) 개요

RAG의 워크플로우 - 인덱싱(Indexing)

인덱싱 (Indexing)

- 외부 지식 소스를 효율적인 검색을 위해 사전 처리하는 준비 과정
- 데이터 전처리: 원본 데이터를 정제하고 구조, 다양한 형식(PDF 등)의 문서 처리 과정 포함
- 청킹(Chunking): 대량의 문서를 문맥을 유지하면서 더 작고 관리하기 쉬운 Chunk로 분할, 청킹 전략은 RAG 시스템 성능에 중요한 영향
- 임베딩(Embedding): 텍스트 청크를 임베딩 모델을 사용하여 벡터 표현(임베딩)으로 변환
- 벡터 저장소 인덱싱(Vector Store Indexing): 생성된 임베딩을 빠른 유사도 검색을 위해 특화된 벡터 데이터베이스에 저장

RAG(Retrieval-Augmented Generation) 개요

RAG의 워크플로우 - 검색 (Retrieval)

검색 (Retrieval)

- 추론 시점(사용자 질의 입력 시)에 관련 정보 검색
- 질의 처리 (Query Processing): 사용자 질의를 인덱싱 단계와 동일한 임베딩 모델로 임베딩
  필요시 질의 변환 또는 확장 기법 사용 가능
- 유사도 검색 (Similarity Search): 질의 임베딩과 벡터 저장소 내 청크 임베딩 비교
  (주로 코사인 유사도 사용)
  가장 유사도 높은 상위 K개의 청크 검색

RAG(Retrieval-Augmented Generation) 개요

RAG의 워크플로우 - 증강 및 생성 (Augmentation and Generation)

증강 및 생성 (Augmentation Generation)

- 컨텍스트 증강, 프롬프트 구성 및 최종 답변 생성
- 컨텍스트 증강 (Context Augmentation): 검색된 관련 청크들을 원본 사용자 질의와 결합
- 프롬프트 구성 (Prompt Construction): 결합된 정보를 LLM 생성기를 위해 템플릿을 사용하여 프롬프트 형태로 구성
- 응답 합성 (Response Synthesis): 증강된 프롬프트를 생성기 LLM에 입력하여 최종 답변 생성

RAG(Retrieval-Augmented Generation) 개요

RAG 주요 오픈소스 프레임워크

RAG 오픈소스 프레임워크

- LangChain: 체인/에이전트 구성에 중점을 두며, 로더, 스플리터, 임베딩, 벡터 저장소, LLM 등 다양한 구성 요소에 대한 광범위한 통합과 그래프 기반 오케스트레이션(LangGraph)을 제공
- LlamaIndex: 특히 RAG 및 LLM과 데이터 인터페이스에 중점을 두며, 데이터 수집, 인덱싱 검색 및 질의(QueryEngine)를 위한 추상화를 제공
- Haystack : 검색, 생성, 평가를 위한 구성 요소와 사전 정의된 파이프라인 템플릿을 포함하여 NLP 파이프라인(RAG 포함) 구축에 중점

(참고) LangChain

LangChain 개요

LangChain이란?

대규모 언어 모델(LLM) 활용 애플리케이션 개발을 위한
오픈소스 프레임워크

LLM을 다양한 외부 데이터 소스 및 기능과 통합

(참고) LangChain

LangChain의 기능

LangChain의 기능

- 데이터 연결 및 인식 (Data-aware applications):

LLM을 다양한 외부 데이터 소스(문서, 데이터베이스, API 등)와 연결

모델이 특정 컨텍스트나 최신 정보를 기반으로 응답하도록 지원

검색 증강 생성(Retrieval Augmented Generation, RAG)의 핵심 기술

- 에이전트 기능 (Agentic applications):

LLM을 활용하여 주변 환경 관찰, 행동 결정 및 실행하는 '에이전트' 구축 지원

에이전트는 도구(tools)를 사용하여 계산, 검색, 외부 API 호출 등 다양한 작업 수행 가능

(참고) LangChain

LangChain의 기능

LangChain의 기능

- **모듈화 및 표준화**:

  모듈화: LLM, 프롬프트, 체인, 인덱스, 메모리 등 다양한 구성 요소를 모듈 형태로 제공

  개발자가 필요에 따라 조합하여 복잡한 애플리케이션 구축 용이

  표준화: 다양한 LLM 제공자(OpenAI, Hugging Face, Google Vertex AI 등)에 대한 일관된 인터페이스 제공

  모델 간 전환 용이성 증대

(참고) LangChain

LangChain의 구조

🦜🔗 LangChain

LLM : 초거대 언어모델로, 생성 모델의 엔진과 같은 역할을 하는 핵심 구성 요소

예시: GPT-3.5, PALM-2, LLAMA, StableVicuna, WizardLM, MPT,…

Prompts : 초거대 언어모델에게 지시하는 명령문

요소: Prompt Templates, Chat Prompt Template, Example Selectors, Output Parsers

Index : LLM이 문서를 쉽게 탐색할 수 있도록 구조화 하는 모듈

예시: Document Loaders, Text Splitters, Vectorstores, Retrievers,…

Memory : 채팅 이력을 기억하도록 하여, 이를 기반으로 대화가 가능하도록 하는 모듈

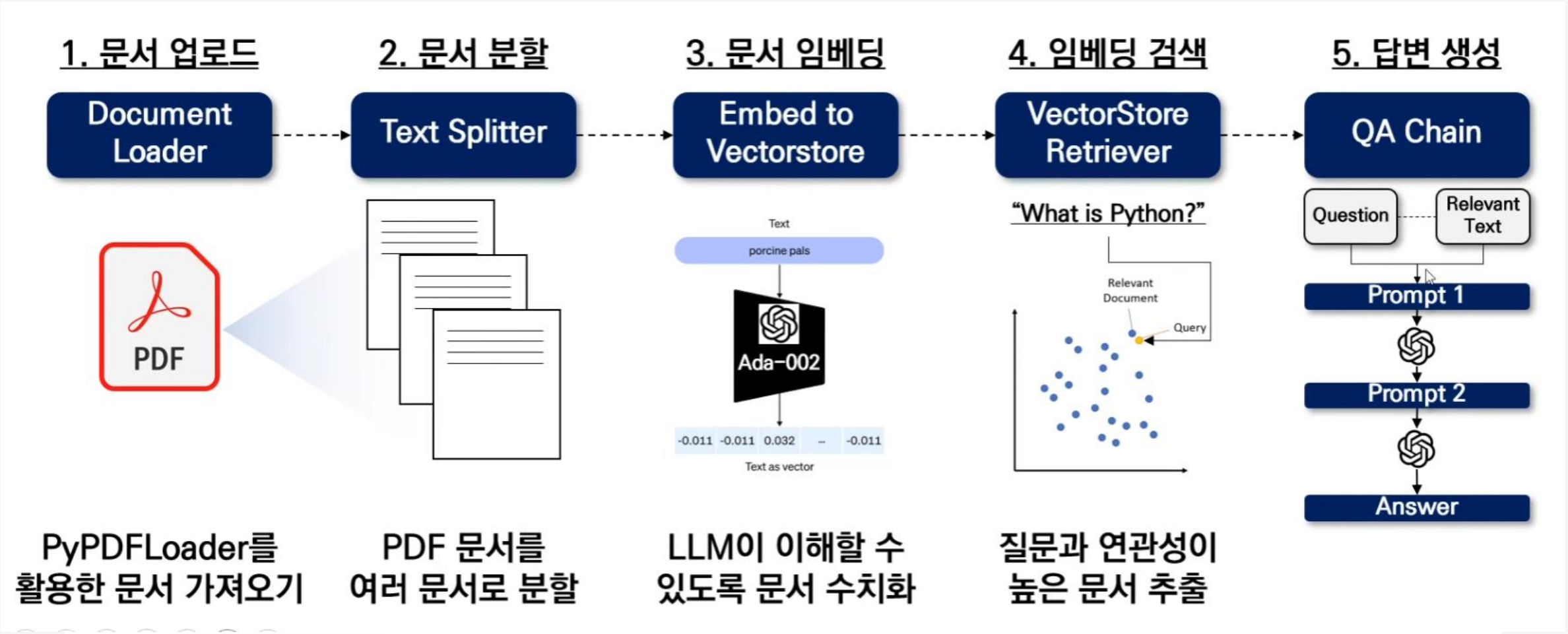
예시: ConversationBufferMemory, Entity Memory, Conversation Knowledge Graph Memory,…

Chain : LLM 사슬을 형성하여, 연속적인 LLM 호출이 가능하도록 하는 핵심 구성 요소

예시: LLM Chain, Question Answering, Summarization, Retrieval Question/Answering,…

(참고) LangChain

LangChain RAG 프로세스-PDF문서 예시



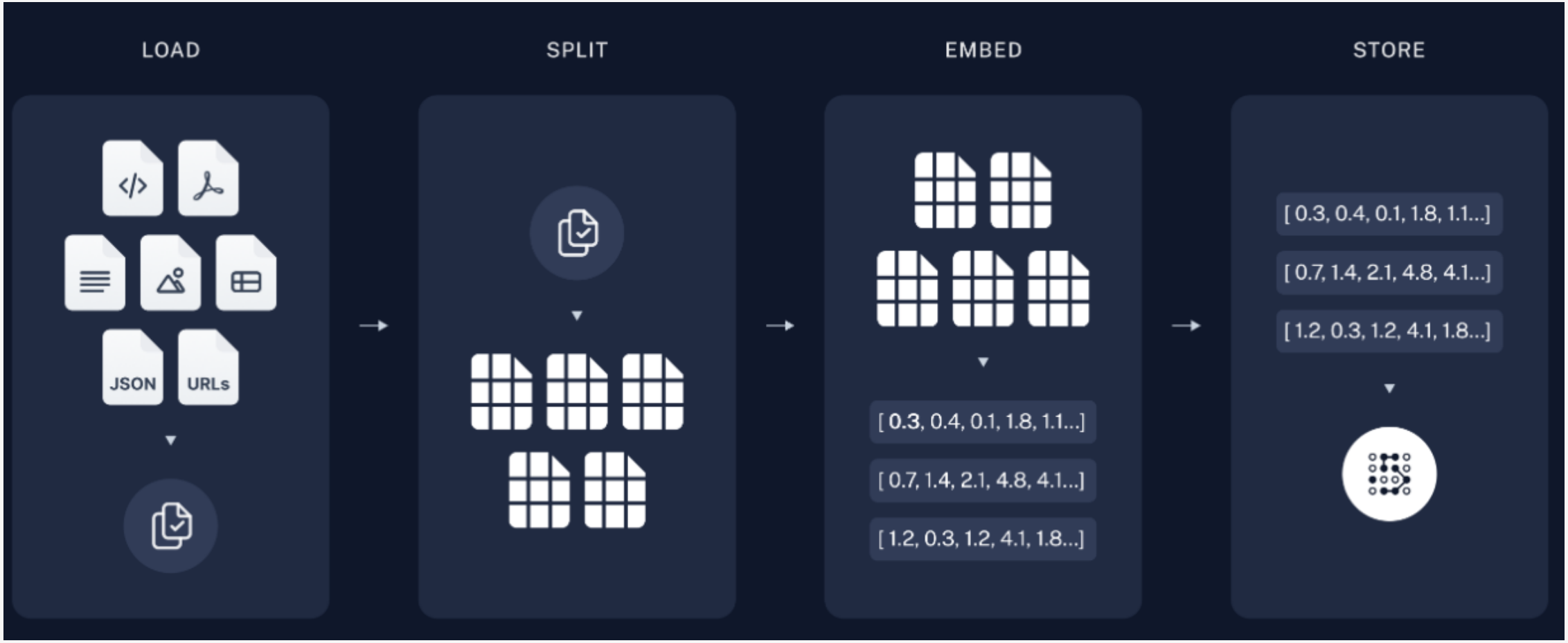
(참고) LangChain

LangChain RAG 프로세스

1	문서 로드(Load)	문서(pdf, word), RAW DATA, 웹페이지, Notion 등의 데이터를 읽기
2	분할(Split)	불러온 문서를 chunk 단위로 분할
3	임베딩(Embedding)	문서를 벡터 표현으로 변환
4	벡터DB(VectorStore)	변환된 벡터를 DB에 저장

(참고) LangChain

LangChain RAG 프로세스



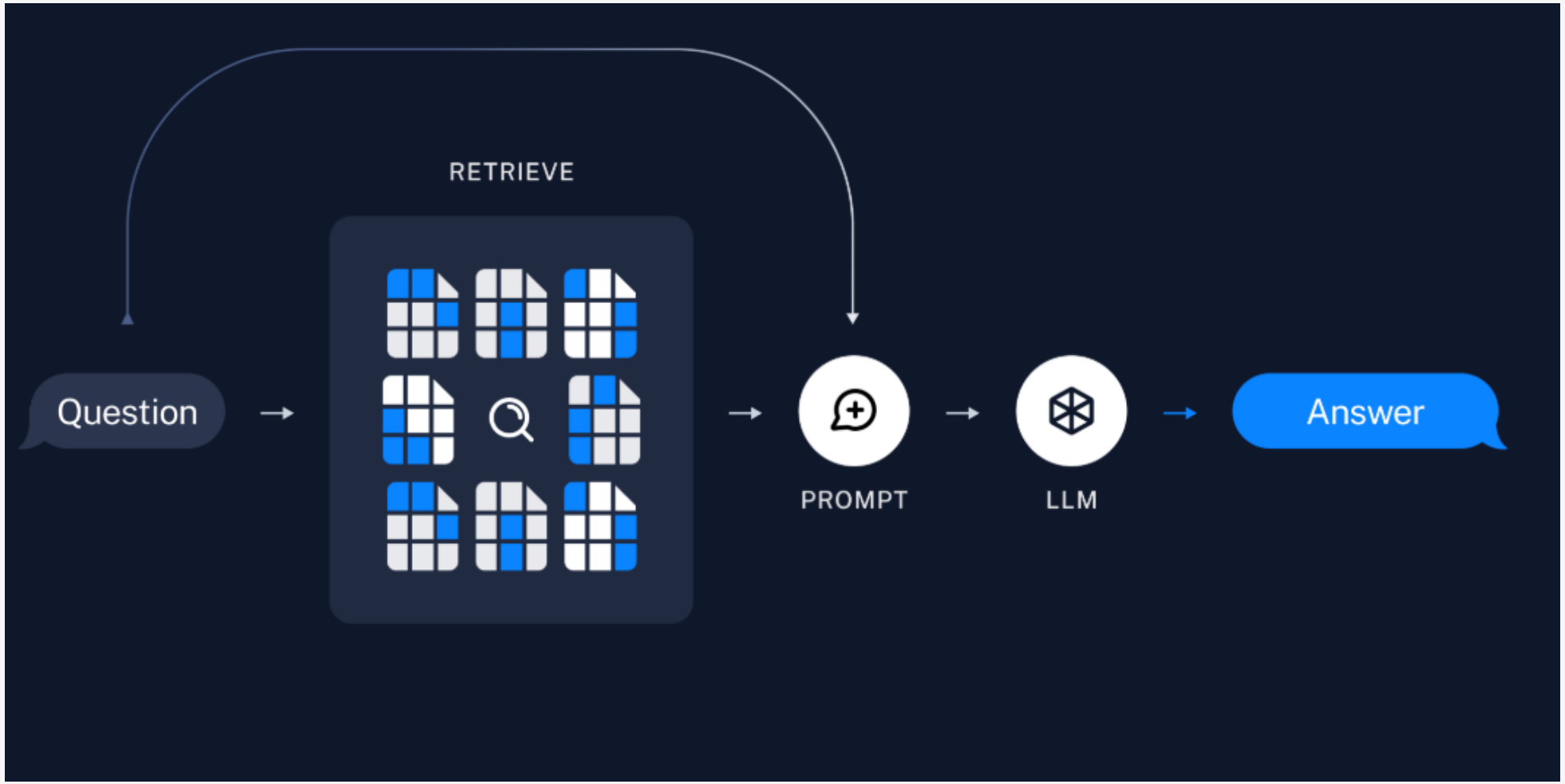
(참고) LangChain

LangChain RAG 프로세스

5	검색(Retrieval)	유사도 검색(similarity, mmr), Multi-Query, Multi-Retriever
6	프롬프트(Prompt)	검색된 결과를 바탕으로 원하는 결과를 도출하기 위한 프롬프트
7	모델(LLM)	모델 선택(GPT-3.5, GPT-4, etc)
8	결과(Output)	텍스트, JSON, 마크다운

(참고) LangChain

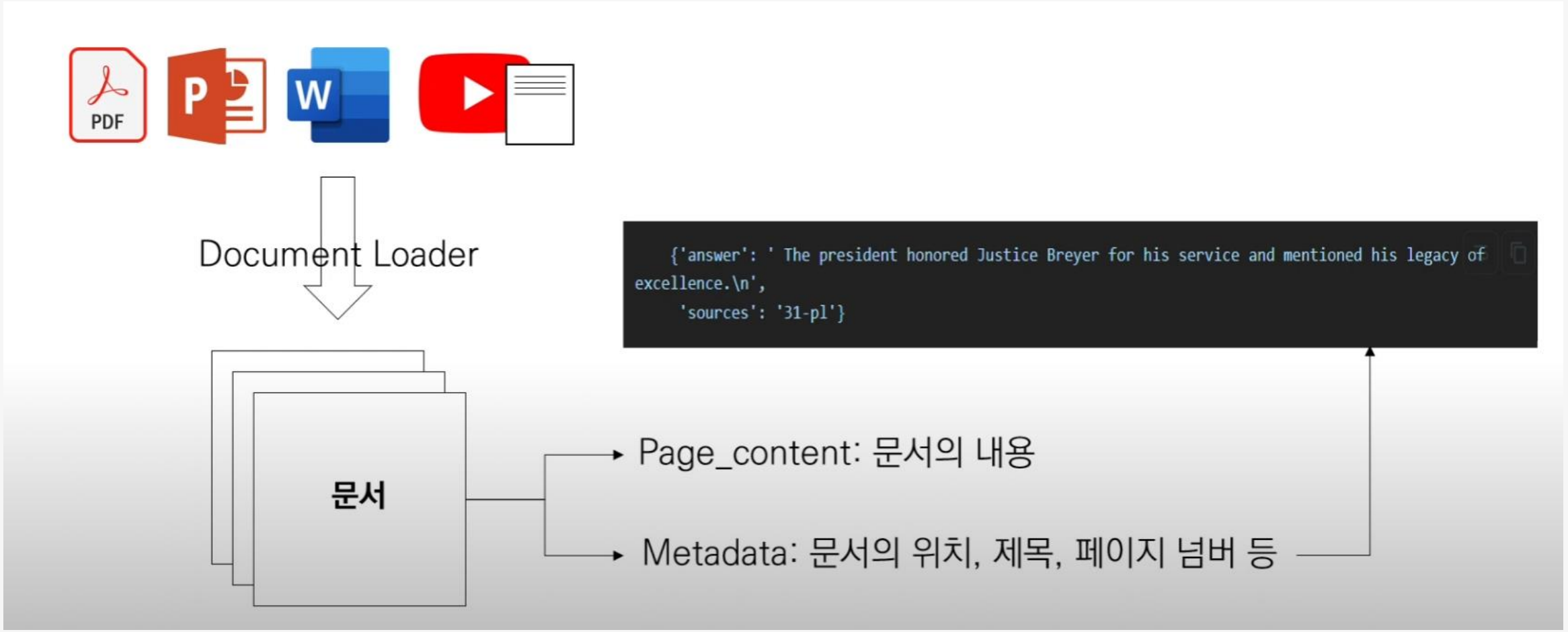
LangChain RAG 프로세스



(참고) LangChain

Document Loaders

Document Loaders는 다양한 종류의 문서를 메모리로 로드



(참고) LangChain

TextSplitter

대부분의 경우에 RecursiveCharacterTextSplitter를 통해 분할

CharacterTextSplitter

2023년 들어 지구 평균온도가 관측 사상 최고치를 기록(WMO, 2023)하는 등 지구 온난화가 가파른 속도로 진행되고 있다. 전 세계적인 기후변화 충격은 세계 경제에 부정적인 영향을 미치고, 그 영향이 국내 경제에도 수출입 경로를 통해 파급될 수 있다. 수입경로 측면에서는 기후변화에 따른 농축수산물 공급충격이 국내 수입가격 상승으로 이어져 국내 경제에 영향을 미칠 수 있다. 수출경로 측면에서는 기후변화 피해에 따른 교역상대국의 소득 감소가 국내 수출품에 대한 수요 감소로 이어질 수 있다. 이에 본 연구는 해외 기후변화의 물리적 피해가 국내 경제에 파급되는 영향을 수출입경로를 중심으로 분석하였다.

분석 결과, 기후변화로 인한 장기간의 점진적 온도상승(만성리스크)은 글로벌 농축수산물 공급 감소와 글로벌 수요 감소를 통해 국내 산업의 생산 위축과 부가가치 감소를 유발하는 것으로 나타났다. 특히 ① 수입 농축수산물에 대한 의존도가 높은 음식료품 제조업(-6.1~-18.2%, 2023~2100년 누적 기준 부가가치 변동폭), 음식 서비스업(-10.2~-17.9%)과 ② 수출 비중이 높은 자동차(-6.6~-13.6%), 정유(-5.8~-11.6%), 화학(-5.0~-10.2%) 산업에서 생산 위축이 발생하여 부가가치가 감소하는 것으로 나타났다. ···

Max_token = 500

2023년 들어 지구 평균온도가 관측 사상 최고치를 기록(WMO, 2023)하는 등 지구 온난화가 가파른 속도로 진행되고 있다. 전 세계적인 기후변화 충격은 세계 경제에 부정적인 영향을 미치고, 그 영향이 국내 경제에도 수출입 경로를 통해 파급될 수 있다. 수입경로 측면에서는 기후변화에 따른 농축수산물 공급충격이 국내 수입가격 상승으로 이어져 국내 경제에 영향을 미칠 수 있다.
수출경로 측면에서는 기후변화 피해에 따른 교역상대국의 소득 감소가 국내 수출품에 대한 수요 감소로 이어질 수 있다. 이에 본 연구는 해외 기후변화의 물리적 피해가 국내 경제에 파급되는 영향을 수출입경로를 중심으로 분석하였다.

구분자 1개 기준으로 분할하므로,
max_token을 지키지 못하는 경우 발생

RecursiveCharacterTextSplitter

2023년 들어 지구 평균온도가 관측 사상 최고치를 기록(WMO, 2023)하는 등 지구 온난화가 가파른 속도로 진행되고 있다. 전 세계적인 기후변화 충격은 세계 경제에 부정적인 영향을 미치고, 그 영향이 국내 경제에도 수출입 경로를 통해 파급될 수 있다. 수입경로 측면에서는 기후변화에 따른 농축수산물 공급충격이 국내 수입가격 상승으로 이어져 국내 경제에 영향을 미칠 수 있다. 수출경로 측면에서는 기후변화 피해에 따른 교역상대국의 소득 감소가 국내 수출품에 대한 수요 감소로 이어질 수 있다. 이에 본 연구는 해외 기후변화의 물리적 피해가 국내 경제에 파급되는 영향을 수출입경로를 중심으로 분석하였다.

분석 결과, 기후변화로 인한 장기간의 점진적 온도상승(만성리스크)은 글로벌 농축수산물 공급 감소와 글로벌 수요 감소를 통해 국내 산업의 생산 위축과 부가가치 감소를 유발하는 것으로 나타났다. 특히 ① 수입 농축수산물에 대한 의존도가 높은 음식료품 제조업(-6.1~-18.2%, 2023~2100년 누적 기준 부가가치 변동폭), 음식 서비스업(-10.2~-17.9%)과 ② 수출 비중이 높은 자동차(-6.6~-13.6%), 정유(-5.8~-11.6%), 화학(-5.0~-10.2%) 산업에서 생산 위축이 발생하여 부가가치가 감소하는 것으로 나타났다. ···

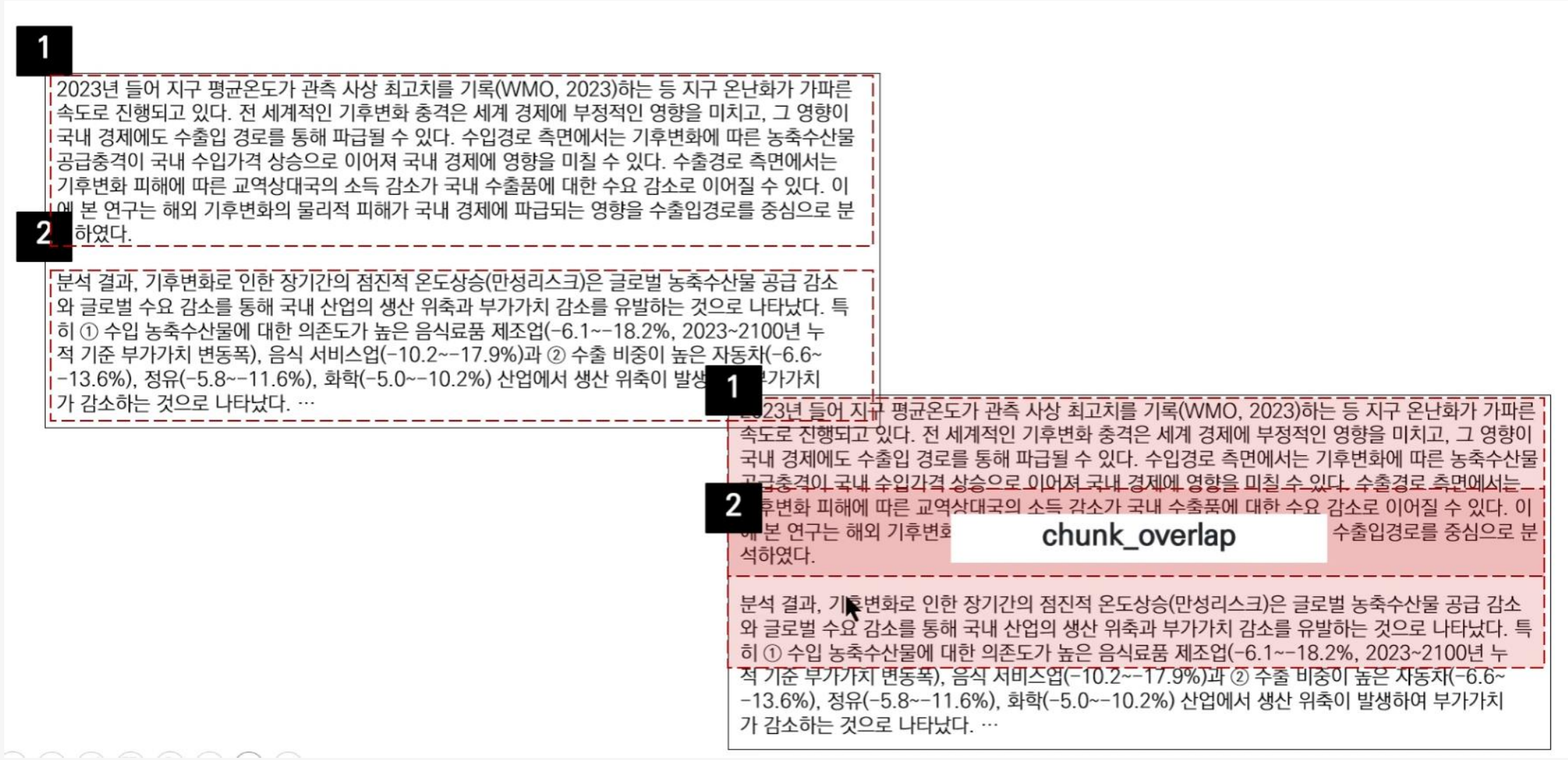
Max_token = 500

2023년 들어 지구 평균온도가 관측 사상 최고치를 기록(WMO, 2023)하는 등 지구 온난화가 가파른 속도로 진행되고 있다. 전 세계적인 기후변화 충격은 세계 경제에 부정적인 영향을 미치고, 그 영향이 국내 경제에도 수출입 경로를 통해 파급될 수 있다. 수입경로 측면에서는 기후변화에 따른 농축수산물 공급충격이 국내 수입가격 상승으로 이어져 국내 경제에 영향을 미칠 수 있다.

줄바꿈, 마침표, 쉼표 순으로 재귀적으로 분할하므로,
max_token 지켜 분할

(참고) LangChain

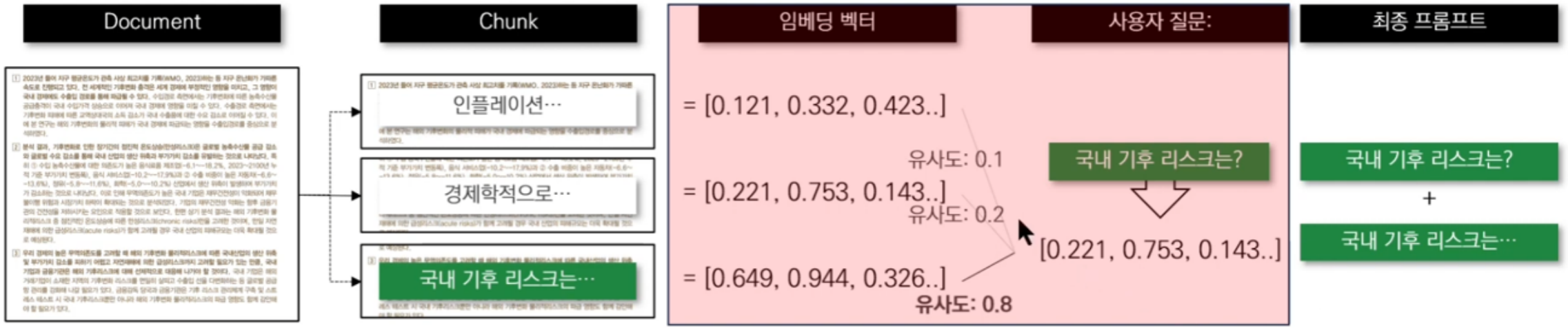
Chunk Overlap



(참고) LangChain

Text Embeddings

Text Embeddings 는 자연어를 숫자로 변환하여 질의와 Chunk간 유사도 비교



(참고) LangChain

Embedding 모델

구분	기업명	모델명	장단점
유료 임베딩 모델	OpenAI Cohere Amazon . : :	text-embedding-ada-002 embed-multilingual-v2.0 titan-embed-text-v1 . : :	- 사용하기 편리하지만 비용 발생 - API 통신 이용하므로 보안 우려 - 한국어 포함 많은 언어 임베딩 지원 - GPU 없이도 빠른 임베딩
로컬 임베딩 모델	HuggingFace	bge-large-en-v1.5 multilingual-e5-large instructor-xl **ko-sbert-nli** **KoSimCSE-roberta-multitask**	- 무료지만 다소 어려운 사용 - 오픈소스 모델 사용하므로 보안 우수 - 모델마다 지원 언어가 다름 - GPU 없을 시, 느린 임베딩