

## <SW분야 산학협력프로젝트(제안서)>

과 제 명	RAG 기법을 활용한 개인용 어시스턴트 개발		
참여기업	PyTorch 리드메인테이너	담당자(직위)	박정환 멘토
팀원요건	<div><div><div>○ ChatGPT, Gemini, Claude과 같은 대규모 언어 모델(LLM) 활용에 관심이 있는 학생</div><div>○ Backend: Python을 활용한 RESTful API 활용 경험이 있는 학생</div><div>○ Frontend: Web Service 또는 Mobile App 개발 경험이 있는 학생</div></div></div>		
추진배경	<div><div><div>○ ChatGPT 등장 이후 많은 사람들이 관심을 가지고 LLM을 개발 및 활용하고 있으나, 동시에 정확하지 않은 정보를 제공하는 환각(Hallucination) 현상이 문제가 되고 있음</div><div>○ 환각 문제를 해결하기 위한 방법 중 하나로 외부 소스로부터 가져온 정보를 활용하는 검색-증강 생성 기법(RAG, Retrieval-Augmented Generation)가 제안되고 있음</div><div>○ 텍스트 파일이나 PDF, 또는 개인이 관심있는 다양한 데이터 소스를 활용하는 개인용 어시스턴트를 개발하며 RAG 기법에 대해 알아보하고자 함</div></div></div>		
프로젝트 목표 및 내용	<div><div><div>(아래 내용은 초기 프로젝트 목표로, 참여 학생들의 배경 지식 및 경험에 따라 변경 가능)</div><div>○ 진행 상황에 따라 크게 3단계로 나누어 프로젝트를 진행 예정</div><div><div>- 1단계(필수): 상용 대규모 언어모델(LLM)을 활용한 개인용 어시스턴트 개발</div><div>- 2단계(필수): 기본 RAG 기법을 적용 → 다양한 데이터 소스를 활용하는 어시스턴트 개발</div><div>- 3단계(선택): 고급 RAG 기법을 적용 → 정교하고 정확한 데이터 소스를 활용하도록 개선</div></div></div></div>		

Input

Query

How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

Indexing

Documents

Chunks Vectors

embeddings

Retrieval

Relevant Documents

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Generation

Question : How do you evaluate the fact that the OpenAI's CEO, ..... dynamics?

Please answer the above questions based on the following information :

Chunk 1 :

Chunk 2 :

Chunk 3 :

Combine Context and Prompts

without RAG

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

with RAG

.....This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Answer

Output

User

그림 1. RAG 기술 개요

- 1단계: 초기 목표는 개인용 어시스턴트를 개발하는 것으로, 아래와 같은 과제들을 포함
  - LLM 적용: OpenAI나 Google 등의 API를 활용
  - 사용자 인터페이스: Streamlit 및 Dash 등을 활용한 프론트엔드 구성
  - 프롬프트 정리: 정리된 답변을 내놓을 수 있도록 프롬프트 엔지니어링 진행
- 2단계: RAG 기법을 적용하여 외부 데이터 소스를 활용하도록 추가 개발 (그림 1. 참조)
  - 데이터 정리: PDF, Web Page 등을 추출하여 조각을 나누고 임베딩을 추출하여 정리
  - 데이터 조회: 사용자 질의로부터 임베딩을 추출하여 Vector DB에 조회
- 3단계: 고급 RAG 기법을 적용한 개선 (또는 1~2단계에서 도출된 과제 개선 등)
  - 고급 RAG 기법 적용 (그림 2. 참조)
  - Multi-Model이나 Multi-Modality 적용, 사용성 개선(UI), 서비스화 등을 고려하여 개선

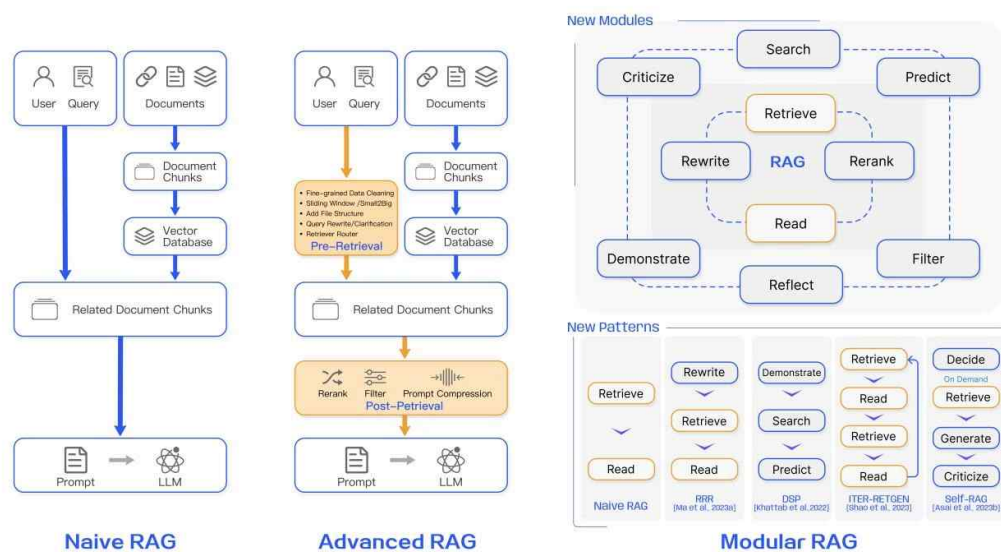


그림 2. RAG 의 3 가지 주요 패러다임

## 기대효과

- 임베딩(Embedding) 및 RAG에 대한 이해도 향상
- LLM의 효율적 활용 방법에 대한 이해도 향상
- 향후 LLM 및 Multimodal LLM 활용 프로젝트를 위한 기초 함양