<SW분야 산학협력프로젝트(제안서)>

과 제 명	Advanced RAG(Retrieval Augmented Generation) 기반 AI 서비스(ex. Digital Forensic, Research Assistant) 개발		
참여기업	(주)데이터스트림즈	담당자(직위)	이동욱 수석연구원

팀원요건

○ 필요 인원 : 4~5명 (하기 역할 통합되어 운영 가능)
- 프로트엔드 개발 팀원 1~2명
- 백엔드 개발 팀원 1~2명
- RAG+LLM 기능 개발 2~3명

○ 프로젝트 배경기술 : LLM Inference Serving (ollama, vLLM, etc), LangChain/LlamaIndex, Streamlit, Chainlit, AutoRAG, Kubernetes, Python, etc.

○ 프로젝트 진행을 위해 요구되는 기술 :
- 각 담당영역에서 요구되는 기술에 대한 경험/학습 필요
- 웹 프론트엔드 (React, Svelte, Streamlit, Chainlit 등)
- 웹 백엔드 (SpringBoot, Django, FastAPI 등)
- OpenAI API 사용 또는 Open Source LLM 사용 기술
- LangChain, LlamaIndex 등 Framework 사용 기술

추진배경

○ ChatGPT의 임팩트 이후, 다양한 언어모델(Large Language Model, LLM)들이 등장하여, 학계 및 산업계 전반으로 커다란 영향력을 보였으며, 현재는 'LLM, LMM, LAM' 등의 지칭되는 Large Model 자체에 대한 연구와, 모델 밖의 세상에 존재하는 데이터를 참조/활용하기 위한 RAG(Retrieval Augmented Generation)이 AI 혁신/활용의 중심에 있음

Large Model + RAG

○ Large Model의 경우, 효율적으로 Fine-Tuning 방법과 Architecture의 개선 등의 연구가, RAG의 경우 Retrieval을 하기 위한 Query Transformation, Embedding, Retrieval Method 및 ReRanker, Filter, Compressor 등에 대한 연구가 진행되고 있음

○ 본 프로젝트는 이러한 AI 연구 분야 중, RAG 사용을 위한 기반 기술을 습득하고, 서비스(ex. Digital Forensic, Research Assistant, etc)를 개발하는 것을 목표로 함

○ 참고문헌

- Retrieval-Augmented Generation for Large Language Models: A Survey (https://arxiv.org/abs/2312.10997)
- SoK: Exploring the Potential of Large Language Models for Improving Digital Forensic Investigation Efficiency (https://arxiv.org/html/2402.19366v1)
- ChatGPT for digital forensic investigation: The good, the bad, and the unknown (https://www.sciencedirect.com/science/article/pii/S266628172300121X)

프로젝트 목표 및 내용

○ RAG는 참고하고자 하는 데이터들로부터, 질의(Query)에 대한 관련성 높은 데이터를 추출하고, 이를 Large Model에 전달하여, 원하는 답을 찾는 시스템

○ Digital Forensic의 경우, 디지털 장치로부터 관련 데이터를 수집하고, 이 데이터들을 수사관이 분석툴을 이용해 키워드 검색 등의 수동적인 작업 또는 스크립트 작성을 통해 진행되는 분석과정에 RAG를 적용하는 것임

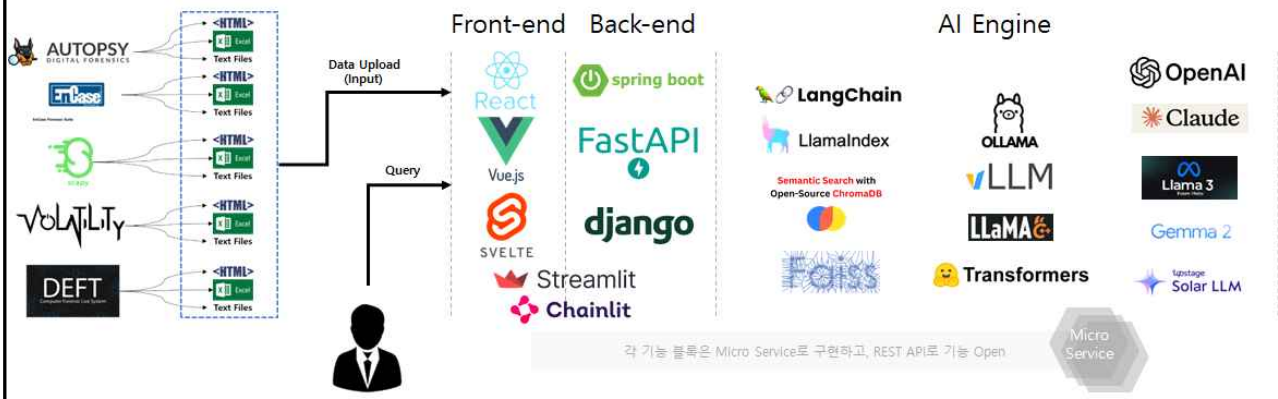
- 디지털 장치로부터 수집한 데이터를 RAG 시스템에 입력으로 주고, 수사관이 질문을 던지면, 관련된 데이터를 검색하고, 그 검색데이터를 Large Model에 질문과 함께 전달하여 자연어를 통해 분석을 진행하는 시스템

○ Research Assistant의 경우, 연구를 진행하고자 하는 분야의 데이터(논문, 책, 블로그 등)를 RAG 시스템에 입력으로 주고, 연구자 또는 학습자가 질문을 하면, 관련 데이터를 찾아 Large Model에 질문과 함께 전달하여 응답을 받는 시스템

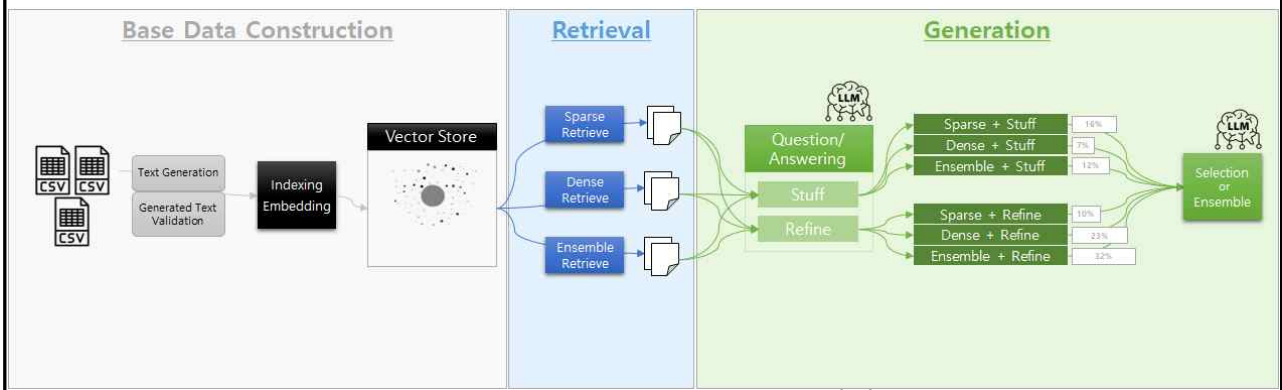
- 특정 언어(Rust, Go 등)를 학습하고자 하는 경우, 해당 언어의 공식 문서 및 Textbook을 RAG 시스템에 저장소에 넣고, 모르는 내용을 질문하거나, Quiz(문제출제+답) 출제를 요청하여 사용자가 자체 테스트를 진행할 수 있도록 하는 시스템 구축이 가능할 것임

○ Digital Forensic, Research Assistant 외에도 RAG를 적용한 AI 서비스 구현 가능

○ 본 프로젝트는 이러한 AI 서비스를 구현하는 과정에 RAG를 적용하는 기술을 학습하는 것을 목표로 하며, AutoRAG(Markr.AI)에서 제시하고 있는 바와 같이, RAG를 구성하는 다양한 기술요소를 이해하는 것을 목표로 함



<시스템 구성 (Digital Forensic 예시)>



<AI Engine Part, RAG system 예시>

기대효과

○ RAG(Retrieval Augmented Generation)에 대한 이해 (Naive/Advanced/Modular)

○ AI 기술(LM+RAG)을 활용한 서비스 개발 능력 배양