softly.ai

We help people focus on what matters the most

- **과제명**: Retrieval-augmented Generation Pipeline의 성능 개선과 배포 (3개월, 3인)

- **과제 배경**
  - LLM 이후 NLP 분야의 기술적 격변
  - General Assistant 로써의 LLM의 한계 극복 및 AI Assistant Product 고도화의 필요성

- **과제를 통해 얻을 수 있는 것**
  - LLM 학습 전반 과정의 이해 (Pre-training / RLHF (SFT-RM-PPO))
  - LLM 을 활용한 AI Product 개발 및 배포 경험 및 AI의 결과물을 사용자를 위해 개선하는 과정 전반

- **필요 지식**
  - AI 모델의 사전학습(pre-training) 및 파인튜닝(fine-tuning) 관련 기초적인 이해도
  - Information Retrieval 및 LLM 관련 기초적인 이해도 (e.g.) elastic search, LLaMa 등)
  - AI Product 개발에 필요한 Back-end, DevOps (MLops) 관련 기초적인 이해도

- **과제 내용**
  - 주어진 query에 대한 관련 지식 retriever 모듈 개발
  - retriever의 반환 결과를 활용한 LLM의 답변 생성 및 성능 향상 (prompt-engineering / fine-tuning)
  - RAG pipeline의 end-to-end 성능 평가를 위한 벤치마크 데이터 제작 및 성능 평가
  - end-to-end pipeline의 성능 개선을 위한 다양한 아이디어 적용

- **참고 사항**
  - 프로젝트 참여 인원의 배경지식과 경험을 고려하여 과제 내용 및 목표가 하향/상향 조정될 수 있습니다.
  - 프로젝트 수행에 활용되는 모든 데이터 및 코드는 대중에게 공개된 내용을 기반으로 하며, 발표에 사용되는 내용은 사전에 프로젝트 담당자와 협의가 필요할 수 있습니다.

- **담당자**
  - 박성준 (SoftlyAI, CEO)
  - 연락처: 010-9605-2655 / sungjoon.park@softly.ai