Gen AI / LLM 활용 및 성능 개선 (AI Advanced)

1. 교육 일시

7/21(월) - 7/23(수) 전일 교육

2. 교육 목표 및 내용

교육 목표	• 서비스에 필요한 Gen AI / LLM 관련 기술을 이해하고 적용할 수 있다. • Gen AI 모델 학습 및 고도화를 통해 원하는 품질을 찾을 수 있다.
기대 효과	• Gen AI / LLM 을 서비스 관점의 적용 방안을 파악하고, 실서비스에 대한 개선이나 새로운 서비스를 개발할 수 있다.

3. 교육 진행

※ AM 10:00 - PM 7:00 까지 기본 교육 시간으로 진행됩니다. • 오프라인 3일 과정

교육 시간표

이론+실습	**프로젝트**

	Day 1	Day 2	Day 3
10:00 - 11:00	**- 개념/모델 이해** ㄴ Deep learning ㄴ Language model ㄴ Transformer ㄴ GPT1, GPT2, GPT3, ChatGPT (RLHF) ㄴ Close Model / Open Model / Reasoning Model ㄴ Model Evaluation	**- RAG** ㄴ Loader ㄴ Chunking ㄴ Embedding Models ㄴ Retrieval ㄴ Vector Database ㄴ Evaluation **- RAG 성능 개선 / 최적화**	**- Function Call 심화** ㄴ Function Calling 기본 개념 및 이해 ㄴ Parallel Multiple Function Calls ㄴ LLM Pipeline

11:00 - 12:00	ㄴ Memory bound ㄴ Optimization **- On device AI** ㄴ Ollama ㄴ sLLM ㄴ Constraints (HW, Cost) ㄴ Optimization: MoE, Kernel, Quantization	**- RAG 성능 개선 / 최적화** ㄴ Query translation ㄴ Decomposition ㄴ Routing ㄴ Query structuring ㄴ Indexing ㄴ Retrieval ㄴ Graph RAG ㄴ Contextual Retrieval	ㄴ AI Agent (LangGraph 활용) ㄴ Integration - LLM 관련 오픈소스 소개
12:00 - 13:00			
13:00 - 14:00	**- Prompt Engineering 주요 기법들** ㄴ few shot, CoT, Reasoning Tokens, ReAct	- 출처 내용 기반 답변 요약 기능	**<Project : Perplexity Clone 서비스 개발>** **(AI Agent, 컨텐츠 요약 구현)** - LangGraph: GraphRAG / Agentic RAG - 질문에 대한 출처를 제공 - 출처 내용 기반 답변 요약 기능 추가 **(검색 추천 기능 구현)** - 출처 내용 기반 답변 요약 기능 추가 **(프로젝트 결과 공유 및 선정 후 발표)**
14:00 - 15:00	**- Prompt Engineering 심화** ㄴ 타 서비스 시스템 프롬프트 분석 ㄴ Prompt Format, Prompt Generator ㄴ 프롬프트 비용, 프롬프트 압축 ㄴ 프롬프트 평가	**<Project : Perplexity Clone 서비스 개발>** **(RAG)** - 질의에 대한 출처 제공 - 출처 내용을 요약하여 답변 제공	
15:00 - 16:00			
16:00 - 17:00	**(환경구축, Prompt Engineering)** - 프로젝트 설명 및 환경 구축 - Model 연동 및 프로젝트 환경 구축 - Prompt Engineering 적용 및 평가 - Automated Promp Generation		
17:00 - 18:00		ㄴ Finetuning 원리/기법 이론 설명 ㄴ Finetuning vs RAG ㄴ finetune 등 관련 기술들 (LoRA, PPO, trl) ㄴ OpenAI Fine-tuning API ㄴ 경량 모델 파인튜닝 (1h) **- 평가** ㄴ정량 평가 (Rouge, BLEU, Benchmarks Datasets) ㄴ 정성 평가 (human eval, metrics)	**- Multimodal 이해** ㄴ 적용된 서비스/사례 소개/학습법 ㄴ OpenAI API 연동 서비스 **- AI 서비스 개발 방향성 (LLMOps stack)** - Q&A - Closing
18:00 - 19:00			

※ Perplexity AI : LLM을 사용하여 질문에 대한 정확한 답변을 제공하는 AI 기반 검색 엔진