



LLM (Large Language Model) 학습/추론 최적화개론



한국컴퓨팅산업협회는 고성능컴퓨팅 인력양성 교육을 통해
중소기업의 고성능 컴퓨팅(HPC) 이용에 대한 진입장벽을 낮추고, 활용을 확산하고자 합니다.
많은 관심과 참여 바랍니다.

교육 개요

교육기간 2026.10.06.(화)~2026.10.08(목), 10:00~17:00(21시간)

접수기간 2026. 09.01(화) ~ 10월 05일(월)

교육대상

- ChatGPT, Gemini 등 LLM 서비스 API를 활용한 개발 경험이 있는 개발자 또는 개발 관리자
- LLM 서비스 API와 오픈소스 LLM의 차이를 이해하고, 오픈소스 LLM 기반의 모델 학습·추론 최적화를 적용하고자 하는 개발자
- “LLM(Large Language Model) 개발 개론~중급” 과정 이수자 또는 이에 준하는 LLM 개발 기초역량 보유자

교육비 무료

교육장소 판교 HPC 이노베이션 허브 교육실 [약도보기](#)

권장이수과목 ▶ 온라인 무료 교육 신청하기

분야	과목명
AI/빅데이터	자연어 처리 기초
	생성형 AI 활용 프롬프트 엔지니어링

교육 과정 소개

· 교육수준 : NCS 5(중급)

일시	단원명	교육내용	강사
10.06.(화)	HPC 이노베이션 허브	• HPC 클러스터 시스템 이해 • 허브 시스템 자원 사용법	KCIA
	LLM 최적화 이해	• LLM 서비스 개발에서 발생하는 주요 문제 상황 - 모델 품질, 응답속도, API 및 GPU 비용, 서비스 확장성 • LLM 최적화의 필요성과 적용 시점 • LLM 최적화의 주요 유형 - Prompt Optimization - Data Optimization - Model Optimization - Inference Optimization • LLM 서비스 API 비용 구조 이해 및 비용 산정 • API 기반 LLM과 오픈소스 LLM 기반 서비스 구조 비교 • 실습 대상 Task 및 3일간 실습 시나리오 소개	
	Task 정의와 LLM 최적화를 위한 평가	• LLM Task 정의의 필요성 • 복합 문제의 Task Divide & Conquer • LLM 최적화를 위한 평가항목 정의 - 품질, Latency, Throughput, 비용 • LLM 평가 데이터 구성 방법 • 정량·정성 평가 방법 • LLM as a Judge의 기본 개념 • 상품 요약 Task를 활용한 Baseline 평가 실습 • 최적화 전 성능·속도·비용 기준값 측정	
10.07.(수)	LLM Task 성능 최적화	• Task 특성에 따른 오픈소스 LLM 선정 • 학습 및 평가 데이터 준비 • Auto Prompt Engineering 개념 및 적용 • 데이터 품질 개선 및 Data Scaling • LLM as a Judge를 활용한 성능 평가 • Single Task Supervised Fine-Tuning(SFT) • SFT 기반 Task 성능 개선 실습 • Reasoning SFT의 개념과 적용 사례 • GRPO의 개념 및 활용 방식 • SFT·Reasoning SFT·GRPO의 적용 조건과 Trade-Off 비교 • Baseline 모델과 최적화 모델의 성능 비교	 김진성
10.08.(목)	LLM Task 추론 성능 최적화	• LLM Inference 서비스 시나리오 이해 • LLM 추론 성능 측정 기준 - TTFT(Time to First Token) - Latency - Throughput - Tokens/sec - GPU Memory • Model Distillation 개념 및 Trade-Off • Teacher Model과 Student Model 구성 • Reasoning Instruction Dataset을 활용한 Distillation 방법 • vLLM 기반 LLM Serving 구조 이해 • vLLM을 활용한 추론 환경 구축 및 실행 • 동시 요청 및 Batch 조건에 따른 추론 성능 측정 • 최적화 전·후 Latency 및 Throughput 비교 • 모델 성능·속도·GPU 자원·비용 간 Trade-Off 분석 • 서비스 요구사항에 따른 최적 모델 및 Serving 방식 선정	

※ 교육생 실습 환경 GPU 제공 예정

※ 주최 측 사정에 의해 상기 내용은 변경될 수 있음

교육 신청



지원 내용

-  **교육비**
-  **수료증 발급** (교육 80%이상 이수 및 설문 완료시)
-  **교재 및 자료**

안내 사항

- (필수요건) 재학/재직 증명서 또는 본인을 증명할 수 있는 기타 신분증 제출 후, 교육 등록 확정 안내 메일 확인 필수
- 중식 및 주차비 미지원
- 최소 신청 인원(10명) 미달 시 폐강될 수 있음
- 중소기업 재직자 및 미취업자 우선 선발

구분	내용	경고
교육 취소	교육시작일 기준 7일 이전 ~ 전날 취소시	경고 1회 부여
	교육시작 당일 취소시	경고 2회 부여
수료 여부	미수료시	경고 2회 부여

※ 경고 2회 누적시 향후 2개월간 교육 참여 제한

문의처

· 한국컴퓨팅산업협회 전락사업팀

 edu@k-cia.or.kr

 031-5182-9045

신청하기 ▶