201935607 전기공학과 김대현

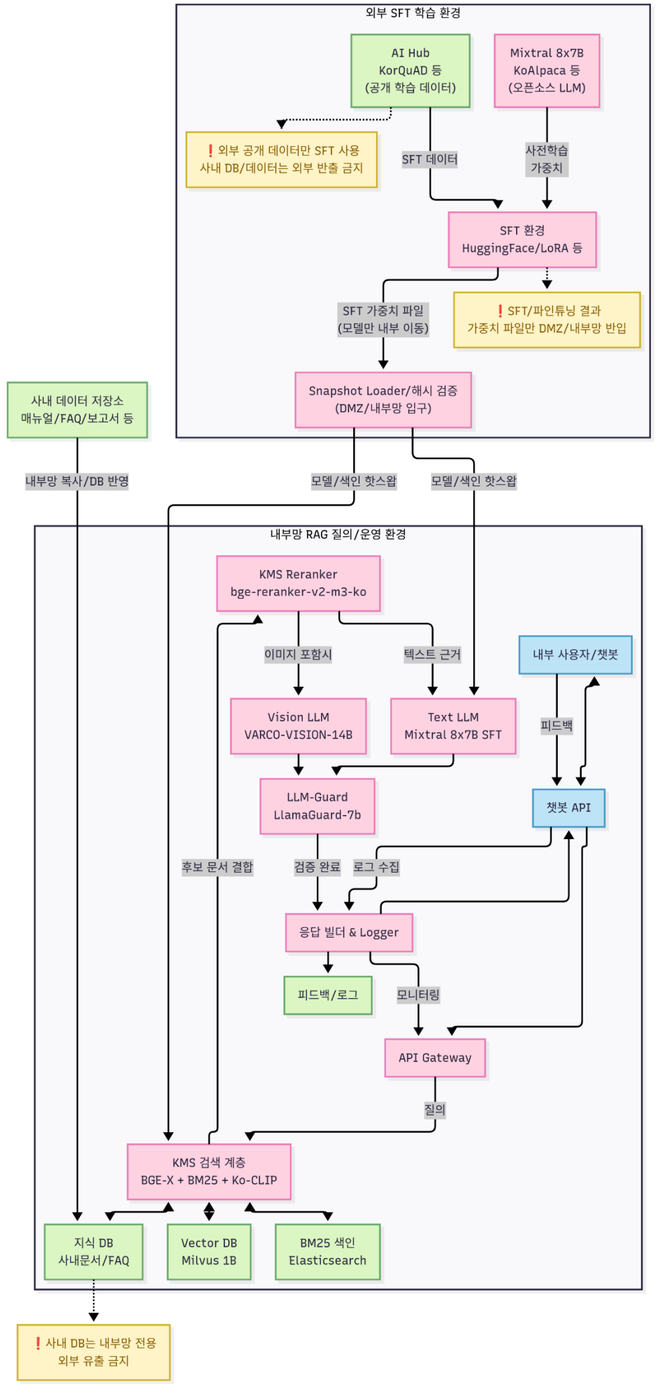
KMS with Modular RAG

무제한 자원 기반 Modular RAG 시스템 설계

LLM 기반 KMS 구성도

- 하늘색: 사용자/챗봇 UI
- 분홍색: AI/모델/파인튜닝/서빙 관련
- 연두색: 데이터/DB/저장소
- 노란색: 참고/경고 설명

- 외부 SFT 학습 환경: 공개 LLM, 공개 데이터, SFT only (사내데이터 반출 X)
- Snapshot Loader/DMZ: 모델 가중치만 내부 반입
- 내부망: 사내 데이터, 지식DB, VectorDB, 검색계층, LLM, RAG 전 처리 파이프라인
- 사내 데이터 저장소: 내부망에만 존재



KMS 구성도 설명

1. 아키텍처 원칙

- 외부 Supervised Fine-Tuning(SFT) 환경과 내부 RAG 환경을 물리/논리적 구분
- 외부 환경: 오픈소스 LLM, 공개 데이터, SFT를 활용
  - 결과로 생성된 모델 가중치 파일만 내부망으로 반입
  - 사내 DB/데이터는 절대 외부로 반출되지 않음
- 내부 환경: 사내 데이터, 벡터 DB, 색인, LLM RAG 파이프라인 운영
  - 실제 서비스/검색/생성/피드백/로깅/모니터링 모두 내부에서만 동작

2. 외부 SFT 학습 환경

- 오픈소스 LLM: Mixtral 8x7B, KoAlpaca 등 LLM 모델
- 공개 학습 데이터: AI Hub, KorQuAD 등 공개된 한국어 학습 데이터셋
- SFT 환경: HuggingFace, LoRA 등 파인튜닝(SFT)을 수행하는 환경
- Snapshot Loader/해시 검증: DMZ(Demilitarized Zone) 또는 내부망 입구에 위치하며, 외부 SFT 환경에서 생성된 모델 가중치 파일만 검증(예: 해시 검증) 후 내부망으로 안전하게 반입하는 역할
- 사내 데이터 저장소: 매뉴얼, FAQ, 보고서 등 사내의 원천 데이터를 저장하는 공간. 내부망에서만 복사/DB 반영되어 지식 DB에 저장

KMS 구성도 설명

3. 내부망 RAG 질의/운영 환경

- 내부 사용자/챗봇: 사내 사용자 및 챗봇이 질의(Query)와 피드백을 주고받는 인터페이스
- API Gateway: 모든 질의가 Gateway를 통해 RAG 파이프라인으로 전달되며 전처리 기능을 수행
- KMS 검색 계층: BM25(키워드 기반 색인), BGE-X(임베딩 기반), KoCLIP(이미지 검색) 등 다양한 검색 기술을 활용하고 여러 DB를 동시에 참조하여 관련성 높은 후보 문서를 선별
- KMS Reranker: dragonkue/bge-reranker-v2-m3-ko 모델과 같은 Cross-Encoder를 사용하여 KMS 검색 계층에서 선별된 Top-K 후보 근거 문서들을 재랭킹
- Text LLM/ Vision LLM:
  - Mixtral 8x7B SFT (Text LLM): 외부 SFT 환경에서 파인튜닝된 가중치를 반입하여 텍스트 기반 답변 초안을 생성
  - VARCO-VISION-14B (Vision LLM): 멀티모달 질의(이미지 포함) 시 이미지 근거를 바탕으로 답변 초안을 생성
- LLM-Guard: LlamaGuard-7b 모델을 활용하여, 생성된 답변 초안에 대한 팩트(Factuality) 검증, PII(개인 식별 정보) 필터링, 정책 위반 여부 등을 검증
- 응답 빌더 & Logger: 최종 응답을 가공하고, 사용자 피드백 및 시스템 로그를 수집하며, 모니터링 등의 기능을 담당. 가공된 최종 응답을 챗봇 API로 반환
- 피드백: 실사용 로그, 사용자 피드백 등 시스템의 지속적인 개선을 위해 데이터를 저장하는 저장소
- 지식 DB/Vector DB/BM25 색인: 각각 임베딩 벡터 데이터베이스(Milvus 1B), 키워드 기반 Elasticsearch 색인, 사내문서 및 FAQ 원본을 저장하는 지식 데이터베이스. 사내 데이터 저장소에서 내부망 DB로만 복사/반영되며, 외부로의 유출은 엄격히 금지

KMS 구성도 설명

4. 흐름 요약

- 외부에서 공개 LLM과 공개 데이터를 사용하여 SFT(파인튜닝)를 진행
  - 파인튜닝의 결과로 생성된 모델 가중치 파일만 Snapshot Loader를 통해 내부망으로 안전하게 전달
- 사내 데이터는 내부망에서 지식 DB로만 복사/적재 (외부 유출 불가)
- 내부 사용자/챗봇의 질의는 API Gateway를 통해 내부망 환경의 RAG 파이프라인으로 전달
  - KMS 검색 계층에서 지식 DB/Vector DB/BM25 색인을 참조하여 후보 문서를 검색
  - 선별된 후보 문서는 KMS Reranker를 거쳐 재랭킹
  - 재랭킹된 텍스트 근거는 Text LLM으로, 이미지 포함 질의는 Vision LLM으로 전달되어 답변 초안 생성
  - 생성된 답변 초안은 LLM-Guard에서 보안 및 정확성 검증
  - 검증 완료된 답변은 응답 빌더&로거에서 최종 가공된 후 챗봇 UI/사용자에게 제공되며 피드백과 로그 수집 (내부 저장)

5. 구조적 장점

- 운영망/학습망 완전 분리:
  - 시스템의 안정성과 보안성을 극대화하며, 각 환경의 독립적인 관리 가능
- 데이터 경로/책임/보안 이슈 명확화:
  - 데이터의 유입/유출 경로와 각 컴포넌트의 책임을 한눈에 파악할 수 있어 보안 리스크 관리에 용이
- 확장성 고려:
  - 추후 RLHF(Reinforcement Learning from Human Feedback) 루프, Shadow Deploy, 사내 관리 툴 연동 등 추가적인 기능 확장에도 유연하게 대응할 수 있도록 설계

LLM 선택 및 모델 컴포넌트

❖ 주요 LLM 및 관련 컴포넌트들은 “무제한 자원”환경을 가정하고 정확도 (휴먼 F1) 95 %+ 및 동의어·표·이미지 포함 질의 Recall 99 % 달성을 목표로 선택되었음.

1. Generator (Text LLM)

- 선택 모델: mistralai/Mixtral-8x7B-Instruct-v0.1를 SFT한 버전
- https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
- 선택 이유:
  - 뛰어난 성능과 효율성: Mixtral 8x7B는 MOE(Mixture of Experts) 아키텍처를 사용하여 8개의 전문가 모델 중 질의 당 2개의 전문가만 활성화하므로, 46.7B 매개변수를 가지면서도 추론 시에는 12.9B 매개변수의 효율성을 보임. 이는 대규모 모델의 강력한 성능을 유지하면서도 "무제한 자원" 가정 하에 최적의 운영 효율성 제공
  - 고품질 지시 추종 능력: SFT를 통해 사용자 지시를 더욱 잘 이해하고 따르도록 훈련하여, KMS의 복잡한 질의에 대한 정확하고 자연스러운 답변 생성에 매우 적합.
  - 추가 LoRA (SFT 코퍼스 30 % 한국어)

2. Generator (Vision LLM)

- 선택 모델: VARCO-VISION-14B https://huggingface.co/NCSOFT/VARCO-VISION-14B
- 선택 이유:
  - 멀티모달 질의 완벽 대응: 이미지/도표가 포함된 사내 문서(예: 설계도, 통계 그래프 포함 보고서)에 대한 질의에 답변하기 위해 필수적인 Vision LLM. KMS의 활용 범위를 텍스트뿐만 아니라 복잡한 시각 정보로 확장
  - 대규모 비전-언어 통합 이해: 총 15 B 규모(언어 14 B Qwen2.5-14B + ≈0.6 B Vision 타워)를 통해 복잡한 시각 정보를 언어 정보와 통합적으로 이해하고 추론할 수 있는 강력한 능력을 제공
- 주의: CC BY-NC 4.0은 사내 상용 서비스 전면 도입 시 별도 허가가 필요

LLM 선택 및 모델 컴포넌트

3. Retriever (텍스트/표 임베딩))

- 선택 모델: BGE-M3 아키텍처 확장·한국어 30 % 가중 pre-train
- https://huggingface.co/BAAI/bge-m3
- 선택 이유:
  - 초고성능 임베딩 및 다국어 최적화: BGE-M3가 이미 100 + 언어·멀티모드 지원. 무제한 자원으로 파라미터·데이터 모두 4 × 증대 → 장문・희귀어 recall 극대
  - 멀티모달 데이터 처리: 텍스트뿐만 아니라 표와 같은 구조화된 데이터 임베딩도 지원하도록 하여, 사내 KMS의 다양한 문서 형식에 대응
  - 현재 BGE-M3는 구조화 테이블을 특별 처리하지 않음. 테이블 세포 분할-시퀀스화 후 임베딩하거나 TAPEX/TabERT류와 혼합하는 추가 모듈이 필요
  - 단일 Retriever 모델 대신 여러 전문가 Retriever를 조합하는 MoE 구조를 활용하여, 다양한 유형의 질의(텍스트, 표, 도메인 특화 등)에 대한 검색 정확도와 Recall을 극대화

4. Hybrid Sparse Retriever

- 선택 방식: BM25, 형태소 분석, BGE-X 내부의 Self-Learned Sparse 헤드 결합
- 선택 이유:
  - 키워드 및 의미 기반 검색 시너지: 임베딩 기반 검색(BGE-X)의 의미론적 이해와 함께, BM25의 키워드 매칭 강점을 활용하여 검색 정확도를 높임
  - 한국어 특화 검색 강화: 형태소 분석을 통해 한국어 특유의 교착어 특성 및 동의어, 표기 변형에 대한 대응력을 높여 법률, 의학 등 전문 용어가 많은 사내 문서 검색에 효과적. 특히, Self-Learned Sparse 헤드는 학습을 통해 희소(sparse) 특징의 중요도를 자체적으로 학습하여 검색 정확도를 더욱 높임

LLM 선택 및 모델 컴포넌트

5. Cross-Encoder Re-ranker

- 선택 모델: dragonkue/bge-reranker-v2-m3-ko
- https://huggingface.co/dragonkue/bge-reranker-v2-m3-ko
- 선택 이유:
  - 한국어 최적화 Cross-Encoder: 다국어 기반 BGE-reranker-v2-m3를 한국어 데이터(AI-Hub 기계독해 + 하드 네거티브)로 추가 SFT해 국내 도메인에 바로 투입 가능
  - 검증된 성능・벤치마크 1위: AutoRAG Korean Embedding Benchmark에서 F1 0.9123 / Recall 0.9649(Top-k 1)로 동급·상위 모델을 모두 앞섬

6. LLM-Guard

- 선택 모델: meta-llama/LlamaGuard-7b
- https://huggingface.co/meta-llama/LlamaGuard-7b
- 선택 이유:
  - 정책·언어 맞춤형 확장 용이: 모델이 Instruction-tuned classifier라 위험 taxonomy 프롬프트만 교체하면 새로운 사내 정책(예: PII, 내부 영업비밀)으로 바로 확장할 수 있고, Zero/Few-shot 예시만 더해도 커스텀 클래스 지정이 가능
  - LoRA·QLoRA 방식으로 한국어 안전 데이터 몇 천 문장만 추가해도 민감 표현 탐지 Recall을 손쉽게 높일 수 있음
- 주의: Llama 2 Community 라이선스 → MAU 700 M 이상 기업은 사전 승인 필요 / 대안으로 OpenAI Guardrails, ZTA-PPO 가능

한글 KMS용 LLM학습을 위한 학습용데이터 세트

추천도	데이터셋	한국어 용량/품질	URL / 라이선스	장점	잠재 한계
★★★	AI Hub 통합 텍스트·도메인 코퍼스	820 GB 이상, 40+ 도메인 (법률·금융·상담·매뉴얼 등)	NIA Open (상업 재배포 허용, 출처 표시) (aihub.or.kr)	정제·라벨 품질 높음, 메타데이터 풍부, 멀티모달 세트 동시 제공	회원제 다운로드, 세트별 포맷 상이, 일부 저작권 제외 요건
★★☆	KorQuAD 1.0/2.0	120 K QA 쌍, 위키 근거 포함	CC BY 4.0 (korquad.github.io)	구조화 QA → SFT · RAG 동시 학습	위키 편중, 대화 스타일 빈약
★★☆	KLUE 11 Task 모음	13 만 샘플(문장·문단)	CC BY-SA 4.0 (github.com)	NLI·STS·NER 등 다양한 언어 능력 태스크	라이선스 SA 조항(파생물 같은 라이선스)
★★	OSCAR ko + Common Crawl ko	수 TB 웹 스크랩	CC BY 4.0 / CC BY-SA (oscar-corpus.com, commoncrawl.org)	규모 최대, 도메인 다양	강한 노이즈, 대규모 필터링 필수
★★	XOR-TyDi QA ko	6 K 다국어 QA	Apache 2.0 (ai.google.com)	다국어 코드스위칭, 번역 태스크 겸용	규묘 작음, 한국어 비율 < 5 %
★★	LAION-KOR	100 M+ 이미지-텍스트	CC BY 4.0 (laion.ai)	멀티모달 SFT/RAG 용이	텍스트 노이즈, NSFW 필터 필수

ML Framework - PyTorch

1. PyTorch 선택 이유
   - 유연한 실험·디버깅
     - 동적 계산 그래프 덕분에 복잡한 LLM 구조를 직관적으로 구성하고 수정 가능
     - SFT나 미세조정 과정에서 빠른 프로토타이핑 및 오류 추적에 용이
   - 압도적 생태계 및 커뮤니티 지원
     - Hugging Face Transformers, DeepSpeed, PyTorch Lightning 등 핵심 툴들이 대부분 PyTorch 기반
     - LLM 학습 관련 자료, 코드, 지원이 가장 풍부함
   - 최신 연구 및 모델 구현과의 호환성
     - Mixtral, RAFT, Mask-DPO 등 최신 LLM 논문 및 커뮤니티 구현이 PyTorch 중심
     - 실험 재현성과 연구 확장성 확보에 유리
   - 강력한 분산 학습 지원
     - torch.distributed, FSDP (Fully Sharded Data Parallel), DeepSpeed (특히 ZeRO-3) 등으로 대규모 학습 효율화
     - H100 × 8 환경에서 Mixtral-8×7B 전체 학습 가능
   - GPU 성능 최적화
     - NVIDIA CUDA와의 궁합 최상
     - bf16, Tensor Core 활용으로 학습 속도 및 자원 효율 극대화

학습 과정 (데이터 -> SFT -> 배포)

1. 데이터 파이프라인

- AI Hub·KorQuAD·KLUE + 웹 크롤링 → 정제·QA 구조화
  → Train / Val / Test 분할
- Hard-negative 마이닝, RAFT용 (Q, Docs, Answer) 트리플 생성

2. 모델·토크나이저 로드

- mistralai/Mixtral-8x7B-Instruct-v0.1 가져오기
- 한국어 특화 토큰 추가 후 **토크나이저 재학습**

3. SFT 전략

- 풀 파인튜닝(8×H100 80 GB) — 최대 성능
- RAFT : RAG용 문서·질문-답변으로 추가 미세조정
- Mask-DPO : 휴먼 피드백 기반 응답 선호도·사실성 정렬
  (PII 필터는 별도 LlamaGuard에서 수행)

4. 학습 인프라

- GPU : 8×H100 80 GB(LLM SFT) + 4×H100(Retriever 임베딩 모델)
  + A100/H100 (Vision 14 B)

5. 분산·최적화

- DeepSpeed ZeRO-3 + bf16 / FSDP, gradient checkpointing,
  optimizer offload 등 최신 분산 학습 및 메모리 최적화 라이브러리
  활용하여 학습 속도 및 효율성 극대화

6. 평가 & 저장

- 정량 : RAGAS, MTEB-KO
  (Retriever, Re-ranker, Generator 각 컴포넌트 성능 측정)
- 정성 : 휴먼 F1 ≥ 95 % (어노테이션 문장 수: 5000개)
- 학습된 모델 가중치는 Snapshot-Loader(해시 검증)통해 내부망 반입

학습 과정 (데이터 -> SFT -> 배포)

7. 서빙·모니터링

- LLM / Triton Inference Server 활용 + KV cache·batch stream, 이미지 입력 없으면 Vision LLM 스킵 등 Latency 최적화
- A/B Shadow Deployment, 모델·색인 Hot-swap으로 무중단 서비스 및 성능 검증
- LlamaGuard-7B LoRA 커스텀 시 PII Recall≥ 98 % 목표 Open-PII-2024 데이터셋

8. 성능 향상 옵션

- Expert-of-Experts MoE Retriever:
  - BGE-X (3B) + 64-expert MoE retriever 조합으로 장문 및 희귀어 Recall 99%+ 달성
- GraphRAG (지식 그래프 기반 RAG):
  - 사내 지식 그래프 기반 지식 탐색으로 복잡한 사실 관계 질의 정확도 향상
- Long-Context Streaming:
  - Mixtral 128k token RoPE 확장(실험적) 등 긴 컨텍스트 효율적 처리
- 리트리버-제너레이터 동시 RL (Self-RAG):
  - End-to-End 강화 학습을 통해 RAG 시스템 전체 성능 자율적 향상

실제 적용 분야 & 기대 효과

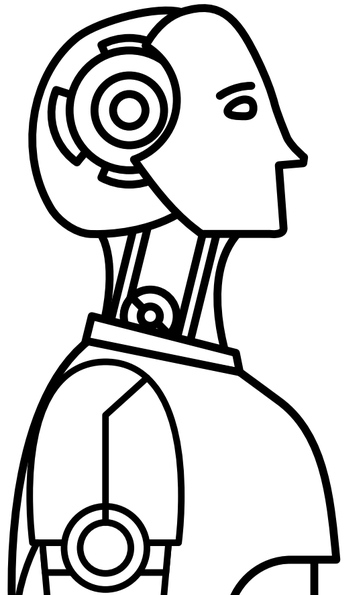
도메인	대표 시나리오	주요 효과
IT 지원	장애 KB 검색, 코드·API Q&A	MTTR▼, 개발 속도▲
경영·사무	휴가·비용 FAQ, 규정 Q&A	반복 문의 자동화, 규정 준수율▲
법무	계약 조항·판례 검색	검토 시간▼, 리스크▼
영업·마케팅	제품/경쟁사·시장 정보	맞춤 제안서 작성 속도▲
CS/콜센터	상담원 어시스트, FAQ 봇	응답 일관성▲, 인력 부하▼
교육·온보딩	신입 가이드, 직무 러닝	온보딩 기간▼, 역량▲

종합 효과 - 정보 탐색 시간 대폭 단축, 지식 일관성 제고, 환각 감소, 운영 비용 절감, 직원 생산성 상승.

지표	현행	6개월 후(예상)	변화
업무 생산성	100	125	▲ 25 %
평균 MTTR	10 h	6 h	▼ 40 %
FAQ 자동응답 커버리지	35 %	60 %	▲ 70 %

업무 시나리오	Before	After (RAG 적용)
장애 KB 검색 🐞	12 min	**40 sec**
휴가 규정 FAQ 📄	2 min	**10 sec**
계약 조항 추출 ⚖️	30 min	**3 min**

감사합니다



시스템 구축 비용

예산 항목 (1 년)	수량 / 가정	단가 (USD)	소계 (USD)	백만 (KRW)	비고
GPU 서버 (자체 보유)					
· 8 × H100 80 GB (LLM SFT)	1 노드	$38 500 × 8	$308 000	420.7	카드당 350 W TDP
· 4 × H100 80 GB (Retriever)	1 노드	$38 500 × 4	$154 000	210.4	—
· 8 × A100 80 GB (Vision LLM)	1 노드	$30 000 × 8	$240 000	327.8	A100 가격 하단치
스토리지·네트워크 (NVMe 100 TB + 200 GbE)	1 세트	—	$30 000	41.0	벤더 견적 평균
데이터센터·전력(전년)	42U 랙 10 kW	—	$14 000	19.1	콜로 랙 $599/mo + 10 kW×$160
클라우드 백업 (AWS p5.48 336h)	—	$31.464/h	$10 600	14.5	AWS 가격
소프트웨어 서포트 (HF Pro Team)	1 년	—	$5 000	6.8	OSS 스택은 0 $
인건비 (서울·총보수)					
· ML 엔지니어	3 명	$65 000	$195 000	266.4	연 ₩76–93 M 평균
· 데이터 엔지니어	1 명	$60 000	$60 000	82.0	—
· DevOps / SRE	1 명	$70 000	$70 000	95.6	—
· 라벨링 파트타임	5 명×6 개월	$1 800/mo	$54 000	73.8	QA·PII 검수
CapEx 소계	—	—	$732 000	999.9	
OpEx 소계 (1 년)	—	—	$403 600	551.3	
총 1차 구축 예산	—	—	$1 135 600	1 551.2	≈ 155 억 원 ±15 %