전북특별자치도 기반 생성형 AI시스템 고도화

오픈소스 URL : https://github.com/jbnu-KimGilMoIm

2025학년도 2학기 SW 캡스톤디자인 경진대회

팀 명 김길모임

팀 원 김현아, 길민준, 모승종, 임나윤 (컴퓨터인공지능학부 3 · 4학년)

지도교수 김윤경 교수

산업체 전북특별자치도청

개발 동기 및 목적

■ 개발 동기

전북특별자치도청은 행정문서 작성, 민원 대응, 내부 보고 등 대부분의 업무가 텍스트 기반의 반복 작업으로 구성되어 있어 공무원들이 문서 · 작성 · 검색 편집에 소모하고 있다. 그러나 외산 LLM 기반 상용 서비스는 개인정보 보호와 내부 자료 유출 위험, 구독료 부담 등으로 도입이 어렵고, 생성형 AI 서비스를 자체적으로 운영하기에는 GPU 인프라 · 기술 인력이 부족한 문제가 존재하였다.

■ 개발 목적

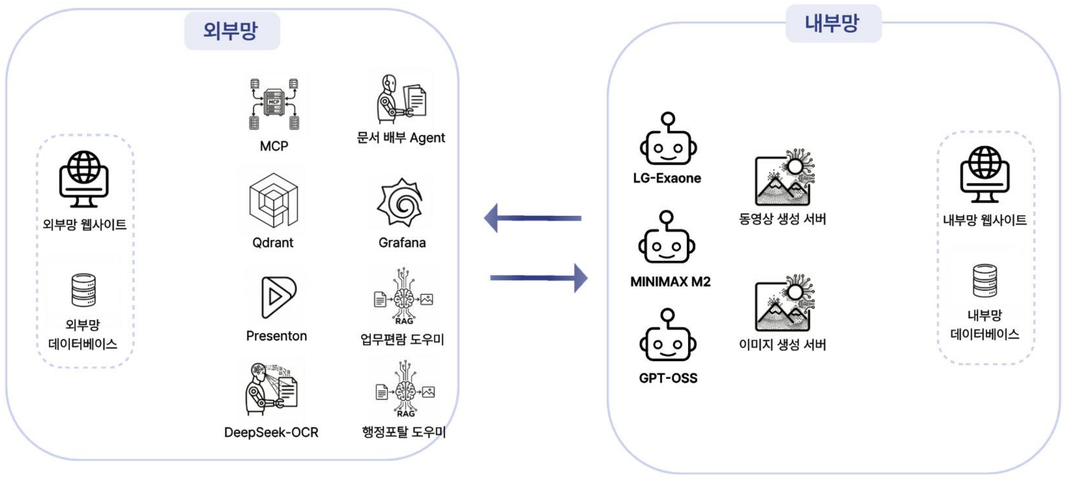
온프레미스 인프라 구축	외부망 의존 없는 내부망 독립형 LLM 시스템
	ROCm(AMD) 환경 기반의 고성능 모델 안정적 서빙
행정 문서 처리 완전 자동화	HWP/HPWX 보고서 및 회의록 자동 생성 (서식 준수) MCP기반 Office(PPT, Excel, Word) 문서 생성 및 자동 배부
	부서/유형별 문서 자동 배부 기능 구현
사용자 중심의 지능형 업무 지원	RAG 고도화를 통한 고정확도 행정 문서 검색
	모바일-STT 연계를 통한 현장 즉시 보고 및 워크플로우 통합

개발 내용

■ 주요 개발 내용

GPU 인프라 & 운용 최적화	• AMD MI300X/ROCm 환경 구축 및 커널 호환성 해결 • Flash/Paged Attention 적용 및 KV Cache 최적화 • Grafana 모니터링 및 PostgreSQL HA 구성
AI 모델 및 RAG 고도화	• 행정 특화 RAG: DeepSeek OCR 및 Semantic Chunking으로 인식률 극대화 • 문서 배부 Agent: Skeleton Matching + Hybrid Similarity (정확도 98.4%)
	• Qdrant + gRPC 기반 대규모 문서(3만건 이상) 고속 인덱싱
서비스 확장 & 파이프라인	• 멀티미디어: Qwen/ComfyUI 기반 이미지,영상 생성 워크플로우
	• 모바일 연동: Conduit 기반 앱 개발, 음성(STT)을 통한 즉시 기록 지원

■ 전체 아키텍쳐



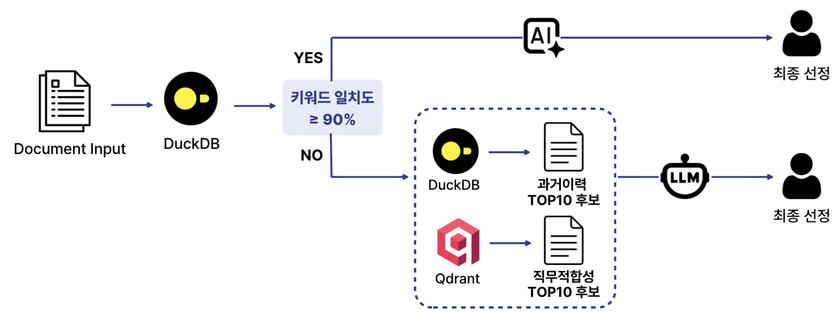
- 모든 시스템 분리하여 모듈화
- 20개 이상의 컨테이너가 API통신만 수행

⇒ API 통신만 허용

주요 기술

구분	핵심 내용
Infra & Serving	• **AMD MI300X/ROCm 최적화:** 커널 호환성 해결 및 드라이버 튜닝 • **vLLM & SGLang:** Paged Attention 기술을 적용한 초고속 LLM 추론 서빙
Data & RAG	• **Qdrant & DuckDB:** 대규모 벡터 검색과 키워드 검색을 결합한 하이브리드 엔진 • **DeepSeek-OCR:** 표/이미지 포함 문서의 텍스트 정밀 추출 및 구조화
Agent & Service	• **MCP (Model Context Protocol):** LLM이 파일 시스템을 제어하는 에어전트 서버 자체 구현 • **HWPX Engine:** XML 직접 제어를 통한 서식(템플릿) • **Conduit:** iOS/Android 하이브리드 앱 개
DevOps & Monitoring	• **Grafana:** GPU/CPU 리소스 및 토큰 사용량 실시간 시각화 • **High Availability:** PostgreSQL 이중화 및 Nginx 기반 부하 분산 처리

■ DuckDB & Qdrant 기반 문서 배부 알고리즘



결과 및 분석

■ 정량적 성과

- 검색 속도 혁신: Qdrant 및 Semantic Chunking 도입으로 RAG 응답 속도 평균 2초 내외 달성
- 업무 효율 극대화: 3단계 자동 배부 알고리즘을 통해 기존 1일 이상 소요되던 문서 분류 · 배부 시간을 수 분 단위로 단축
- 높은 정확도 검증: 하이브리드(키워드+벡터) 매칭 기술 적용 결과, 문서 배부 정확도 98.4% 기록

■ 시스템 완성도

- 행정 업무 자동화: Agent 기반의 HWP보고서 자동 생성 및 모바일-STT 연계를 통해 행정 업무 전과정의 One-Stop 워크플로우 구현
- 안정적인 운영 환경: ROCm 기반 GPU 모델 서빙 안정화 및 Grafana PostgreSQL(HA) 도입으로 공기관 수준의 신뢰성 · 보안성 확보
- 확장성 및 기술 자산화: 독자적인 GPU 기반 온프레미스 AI 구축 노하우를 확보하여, 향후 재난 대응 · 민원처리 등 다양한 행정 서비스로의 확장 기반 마련

도청 내부 테스트에서 RAG · 문서배부 · HWP 생성 기능 안정 동작 ⇒ 행정 실무 적용 가능성 확인

전북대학교 SW중심대학사업단