

랩 위드의 AI 기반 솔루션 >

쉬운 랩 위드(주) AI 인프라 설명 _ 3단 라우터 & 자체 sLLM



기술사업화 랩 위드
2026.07.30. 17:34 조회 6

댓글 0 URL 복사

핵심 요약 — 한 줄로 먼저

- 3단 라우터 : EXAONE(1단) → Grok(2단) → Claude(3단·최종) — 난이도별 자동 분배
- 자체 sLLM : rnd-midm (KT Mi:dm 2.0 기반) — R&D 사업화 초안 전용, v6 운영 , 현재 v7 학습 중
- RAG 지식DB : 벡터 지식DB 3,400건 · 한국어 특화 임베딩 하이브리드 검색 _ 두 시스템 공통 앞 단 연동
- 핵심 가치 : 완전 온프레미스 · 기술사업화 특화 · 고객 기밀 외부 전송 없음

왜 이렇게 만들었나 — 핵심 가치 3가지

가치	내용
정보 보안	고객의 R&D·재무·특허 자료는 민감 정보. 전 구성요소가 사내 서버에서만 구동되어 외부(외부 API·클라우드)로 데이터가 나가지 않음.
비용 최적화	모든 질문에 고성능 AI를 쓰면 비용이 폭증. 쉬운 일은 로컬 AI(EXAONE), 어려운 일은 외부 AI(Grok·Claude)로 자동 분배해 질(質)은 유지하고 비용은 최소화.
도메인 특화	범용 AI는 기술사업화 계획서의 표준 목차·정량목표·사업화 전략을 잘 모름. 회사가 실제 보유한 문서(14GB)로 직접 학습시킨 전용 모델(rnd-midm)을 별도 운영.

지식DB / RAG — 공통 앞 단

벡터 지식DB는 3단 라우터와 초안공장(rnd-midm) 양쪽 앞 단에 공통으로 위치하며, AI가 답하기 전에 관련 근거를 먼저 찾아 프롬프트에 주입합니다.

구성 요소	내용
벡터 지식 검색 엔진	한국어 특화 임베딩 모델 기반 · 하이브리드 검색 (키워드+의미)
지식DB 규모	3,400건+ · 연구계획서 · 특허 · 시장조사 · 사업화 보고서
적용 방식	관련 근거 추출 → 프롬프트 주입 → 환각 억제
하드웨어	임베딩 모델 GPU 상주 (RTX 32GB · 상시 가동)

3단 라우터 — 자동 모델 선택 시스템

개요

들어온 질문의 난이도를 자동으로 판별해서, 3개의 서로 다른 AI 중 가장 적합한 AI에게 넘기는 자동 분배 구조입니다.

 비유: 병원의 '1차 진료(EXAONE) → 전문의(Grok) → 대학병원(Claude)' — 사안의 무게에 맞는 곳으로 자동 배정

AI 질의 분류기(EXAONE) → IF/ELSE 분기 → 3개 LLM 노드 구조로 구현. 분류기 자체도 EXAONE이 담당하므로 분배 처리 부담이 없습니다.

▶ 3개 계층 상세

단계	모델	역할	처리 대상	특징
1단	EXAONE 3.5	분류기·단순 처리	FAQ·요약·내부 초안·정보 탐색	GPU 상주 · ~69 tok/s · 로컬
2단	Grok (xAI)	중간 난이도 주력	시장조사·BM·브리핑·IR 초안	빠른 속도 · 실시간 웹 · 외부 AI
3단	Claude	최종 고품질 산출물	사업계획서·IR 최종본·전략·고난도 추론	최고 품질 · 긴 맥락 유지 · 외부 AI

▶ 로컬 AI로 Grok을 대체하지 않는 이유

실측 결과 : qwen3-deepseek-r1 14B _ 콜드 로딩 110~120초, 워밍 속도 EXAONE의 절반 이하, 16GB GPU OOM 위험, 품질 Grok·Claude 미만. → 현행 3단 유지가 최적. 로컬 14B는 24GB GPU 업그레이드 시에나 유효.

▶ 3역할 분담 — 라우터 배정 원칙

역할	담당 모델	설명
탐색·어드바이저	EXAONE 3.5	질의 분류·지식DB 검색·상담 응답 (3단 라우터 1단)
보고서 초안	rnd-midm	R&D 사전기획·시장조사·사업화전략 초안 자동 생성 (내부 컨설턴트 전용)
최종 검수·산출물	Claude	고객·투자자 제출용 최종 산출물 품질 검수·완성 (최종 산출물만 사용)

 자체 sLLM — rnd-midm (R&D 사업화 특화 모델)

▶ 개요

rnd-midm은 랩워드가 직접 보유한 실제 기술사업화 문서로 파인튜닝한 전용 소형 AI입니다.
범용 AI가 모르는 연구개발사업계획서의 표준 구조와 문체를 학습해, 컨설턴트가 보고서 초안을 빠르게 뽑는 데 씁니다.

▶ 어떻게 만들었나

항목	내용
베이스 모델	KT Mi:dm 2.0 (11.5B) _ 국산 · MIT 라이선스(상업 활용 자유). 국산 LLM 8종 동일 조건 실측 비교 결과 R&D 문서 생성 품질 1위 선정 (2026-07-18 대표 결정)
학습 방식	QLoRA 4bit + LoRA _ 16GB GPU 한 대로 학습 가능한 경량 기법
학습 데이터	회사 보유 실제 문서 (PDF 2,000여 건, 14GB) _ 연구계획서·시장조사·결과보고서. PII(전화·이메일·주민번호·성명) 마스킹 후 학습용 가공

운영 방식	rnd-midm:latest 태그로 로컬 AI 실행 엔진에 등록 · 초안공장 앱 호출 시 로드
-------	--

▶ 버전 현황 (2026-07-30)

버전	내용	v7 목표	학습 규모	상태
v5	B+batch2 통합	-	train 1,972 / eval 218	완료
v6	전체 코퍼스 확장 (R&D 계획서 유형 우선)	-	train 5,769 / eval 639	✅ 현재 운영
v7	근거충실 + 사업계획서 항목별 정합	RAG 형식 동일 학습 · 환각 억제 · 배경/목표/요소기술/방법론/사업화 항목별 균형	생성 중	⌚ 학습 중 ~7/31 완료

- ⚠️ v7 학습데이터 생성이 GPU를 점유 중. 이 기간 초안공장 응답 지연·타임아웃 가능.
품질 최종 평가는 v7 등록 후 진행.
- ⚠️ rnd-midm은 function calling 미지원 → 도구 호출이 필요한 에이전트/어드바이저 앱에는 사용 불가.
해당 앱은 EXAONE·Grok·Claude가 담당.

🔗 3단 라우터 vs 자체 sLLM — 핵심 차이점

비교 항목	3단 라우터	자체 sLLM (rnd-midm)
본질	모델 선택·오케스트레이션 시스템	R&D 사업화 특화 경량 언어모델
구성	EXAONE(1단) → Grok(2단) → Claude(3단)	rnd-midm 단독 (KT Mi:dm 2.0 기반)
역할	어떤 모델로 처리할지 자동 결정	보고서 초안 텍스트 직접 생성
사용처	어드바이저·챗봇·에이전트·브리핑 전반	보고서 초안공장 전용 (컨설턴트 내부)
제약	없음 (범용)	function calling 미지원, 에이전트 불가
보안	3단 Claude 호출 시 외부 API 사용	100% 로컬 — 외부 전송 없음
현재 상태	라이브 운영 중	v6 운영 · v7 학습 중 (~7/31)
관계	rnd-midm과 별개 파이프라인	3단 라우터와 병렬, RAG 지식DB 공유

- 💡 핵심 관계 : 두 시스템은 별개 파이프라인. 3단 라우터는 난이도별 질의 처리, rnd-midm은 R&D 초안 생성에 특화.
벡터 지식DB(RAG)가 앞 단에서 두 시스템에 근거를 공급.

장점

- 🔒 우리 자료는 우리 서버 안에서만 처리됩니다
완전 온프레미스 — 고객 기밀 유출 위험 없음. 정부과제·기업 R&D 자료를 안심하고 맡길 수 있음.

⚙️ 질의 성격에 따라 가장 적합한 AI가 자동으로 처리합니다

3단 라우터로 최적화 — 로컬 AI(EXAONE)가 대부분을 처리하고, 고부가 산출물은 외부 AI(Grok·Claude)가 담당.

🔗 기술사업화에 특화된 전용 AI입니다

범용 챗봇이 아니라, 회사가 실제 수행한 사업계획서·시장조사·결과보고서로 직접 학습시킨 rnd-midm — 표준 목차·정량목표·사업화 전략을 그대로 생산.

📊 실증된 성능

국산 LLM 8종 실측 비교로 KT Mi:dm 2.0 선정, v5→v6→v7로 지속 고도화. 근거기반(RAG) 설계로 환각 억제.

▶ 용어 풀이

용어	쉬운 설명
온프레미스	외부 클라우드가 아니라 회사 안에 둔 서버에서 직접 돌리는 방식
sLLM	small Language Model. 특정 업무에 맞게 작고 빠르게 특화한 AI
파인튜닝(미세조정)	이미 똑똑한 AI에게 우리 회사 자료를 추가로 가르쳐 우리 업무를 잘하게 만드는 것
RAG(검색증강생성)	AI가 답하기 전에 회사 지식 DB에서 관련 근거를 먼저 찾아와 그 근거를 바탕으로 답하게 하는 방식 → 지어내기(환각)를 줄임
환각(hallucination)	AI가 근거 없이 그럴듯하게 지어내는 오류
3단 라우터	질문 난이도에 따라 가장 적합한 AI에게 자동으로 배정하는 스마트 분배 시스템

랩위드(주) AI 인프라

기술사업화 컨설팅 품질 고도화 전용 · AI 솔루션 판매 아님

💡 AI도 효과적인데 — 기술사업화 전문 데이터셋까지 더해졌습니다

학습 데이터: 기술개발 사업계획서 · 기술사업화 자료 2,000건(14GB) · 정제된 R&D 실문서

컨설턴트 질의 입력

🔒 보안 게이트웨이 (HTTPS)

AI 오케스트레이션 엔진

📖 지식DB / RAG — 공통 앞 단 · 3단 라우터·초안공장 모두에 전문 근거 공급

벡터 지식 검색 엔진

3,400건 · 한국어 특화 임베딩 · 하이브리드 검색

연구계획서 · 특허 · 시장조사 · 보고서

근거 추출 → 프롬프트 주입 → 환각 억제

🔄 3단 라우터 — 질의 자동 분배

질의 난이도 자동 판별 → 최적 AI 배정

1단 · EXAONE 3.5 — 로컬

단순 분류 · 요약 · 내부 검색 · 초안 보조

사내 서버 GPU 상주 · 외부 전송 없음

2단 · Grok (xAI) — 외부 AI

시장조사 · BM 분석 · 브리핑 · IR 초안

중간 난이도 · 범용 · 대량 처리

3단 · Claude

최종 사업계획서 · IR · 전략 보고서

고난도 추론 · 최고 품질 보장

최종 응답 · 보고서 출력

🎯 3역할 분담 (라우터 배정 원칙)

탐색 = EXAONE (1단 · 로컬)

초안 = rnd-midm (사내 전용)

검수 = Claude (최종 산출물만)

💡 보고서 초안공장 — 컨설턴트 내부 전용

rnd-midm (자체 sLLM)

베이스: KT Midm 2.0 (11.5B)

MIT 라이선스 · 국산 LLM 8종 비교 선정

학습: 4bit 양자화 + LoRA 경량 파인튜닝

데이터: 기술사업화 실문서 2,000건(14GB)

정제된 R&D 계획서 · PII 마스킹 처리

버전 현황

v5 완료

v6 운영중

v7 학습중 (~7/31)

⚠️ function calling 미지원 → 에이전트 앱 사용 불가

R&D 보고서 초안 완성

항목	3단 라우터 (EXAONE → Grok → Claude)	자체 sLLM (rnd-midm)
목적	질의 난이도별 최적 AI 자동 배정	R&D 사업화 특화 보고서 초안 생성
보안	3단 Claude 호출 시 외부 API 사용	100% 로컬 · 외부 전송 없음
사용처	어드바이저 · 챗봇 · 시장분석 · 브리핑	보고서 초안공장 전용

기술사업화 컨설팅의 깊이 있는 지원 → www.labwith.kr

기술사업화 랩 위드

댓글을 남겨보세요

등록

<https://cafe.naver.com/f-e/cafes/31194010/articles/797?boardtype=L&menuid=8&referrerAllArticles=false>

5/5