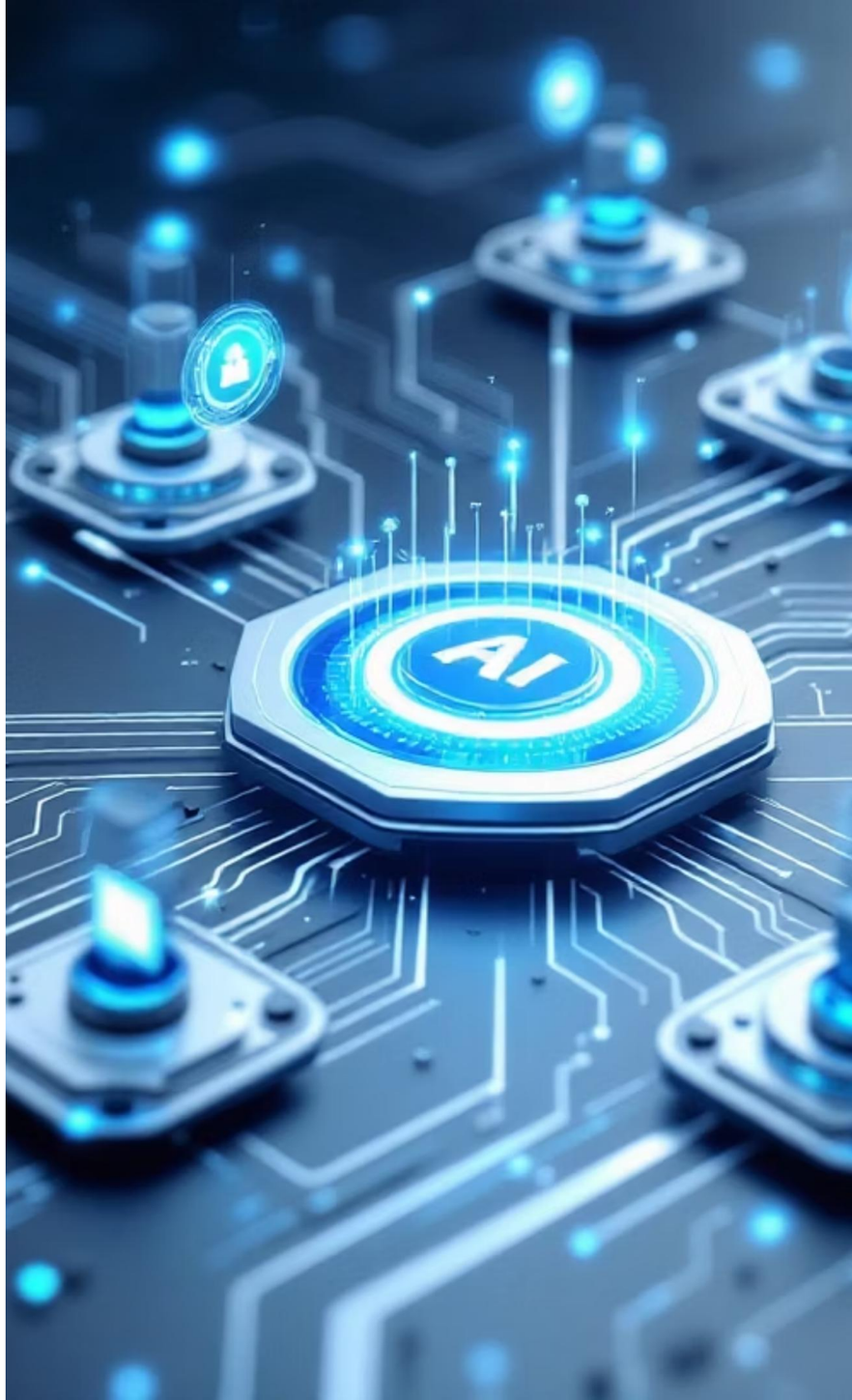
HAMONIZE

https://hamonize.com

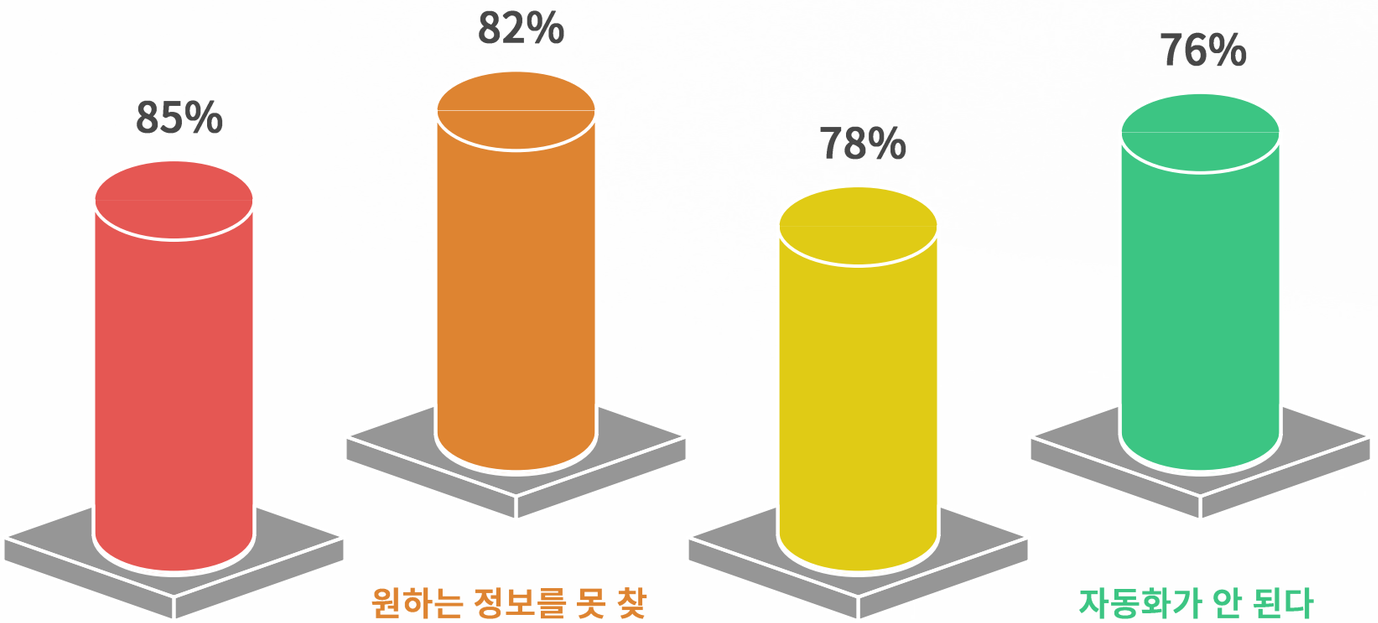
하모나이즈, 당신만의 소버린 AI 플랫폼

PII Safeguard, Knowledge Graph, ML-Cache RAG, Kernel WORM 등
독창적 특허기술로 AI 도입의 모든 애로사항을 완벽히 해결합니다.



문의 : 02)2039-3977 / sales@invesume.com / M. 010.3838.8378 김형채 대표

기업이 실제 겪는 AI 4대 문제



결과적으로 AI 도입 실패율 70% 이상

애로사항
데이터 보안 및 프라이버시 우려
할루시네이션 (허구 정보 생성)
네트워크 의존성
운영 비용 증가
맞춤형 솔루션 부족
데이터 소유권 문제
실시간 처리 성능

무엇이 필요할까요?


좋은 LLM
(계획 수립, 추론, 대화 처리)
지식 기반 검색(RAG)
(문서/데이터 실시간 반영)
Tools
(검색, 파일 처리 등 실행)
오케스트레이터
(에이전트 및 툴 조율)
인증 및 보안
(사용자 인증, 접근 제어)
메모리 / 컨텍스트 관리
(기억 유지, 컨텍스트 유지,)
Agent Runtime
(안전한 실행 환경)
Process 엔진
(워크플로우, 상태 관리)

엔터프라이즈 AI 플랫폼 기능을 만족하는가?

엔터프라이즈 AI 요구사항	현장의 고객 니즈	하모나이즈 핵심 기능
Conversational Agent Workspace	AI와 지속적 문맥을 가진 업무형 대화 수행	AI 채팅, 세션, 히스토리 요약, 단기/장기 기억
Document Intelligence & Multimodal Understanding	문서·이미지·레이아웃을 읽고 구조화	OCR, 문서 업로드, 멀티모달 RAG, 문서 레이아웃 이해
Explainable Knowledge Retrieval	근거와 연결관계를 보여주는 신뢰형 검색	RAG, 출처 인용, GraphRAG, CRAG
Agent Orchestration & Tool Use	여러 모델/도구를 자동 조합해 업무 수행	슈퍼바이저 에이전트, 툴 호출, 웹검색, 선택형 모델 워크플로우
Reliability & Quality Control	할루시네이션을 줄이고 결과를 자동 보정	Self-Reflection, CRAG, 검증 루프
Governance, Security & Audit	권한·로그·정책 통제가 가능한 기업형 운영	JWT, RBAC, API Keys, activity_logs, PII 처리
Operations, Performance & Scalability	서비스 상태를 관리하고 확장 가능	systemd, PM2, Redis 큐, Worker, 백업/모니터링
Model Lifecycle & Optimization	자체 모델을 학습·개선·선택 운영	LLM Fine-tuning, LoRA, finetune-cli, 선택형 모델 라우팅

특장점 - 챗봇이 아니라, 지식과 업무를 이해하고 실행하는 AI

문서 이해 고도화

OCR, 문서 레이아웃 이해, 멀티모달 RAG 지원. PDF, 이미지, 표, 도식 등 복합 문서 구조를 업무 맥락으로 해석 가능

신뢰성 높은 RAG 체계

GraphRAG, CRAG, 출처 인용, 설명가능한 검색 제공. 단순 유사도 검색이 아니라 문서 관계와 답변 근거까지 반영

독창적 검색 품질 고도화

예상질문 생성 기반으로 실제 사용자 질문 패턴 반영. **4-Score 기반 리랭킹**(키워드 정확도, 의미 관련성, 내용 품질, 최신성) 종합 평가. 고객 조직의 문서/질문 특성에 맞춘 **독자 임베딩모델 파인튜닝** 지원

업무 결과물 자동 생성

보고서, 문서, 폼, 신청서 등 산출물 자동 생성. 단순 답변형 AI가 아니라 실제 업무 결과물을 만들어 주는 구조

에이전트/워크플로우 자동화

Assistant(MCP 특화), AI 자동화 모듈, Supervisor Agent, Tool Use 제공. 질의응답을 넘어 실제 업무 단계 수행과 도구 호출 가능

엔터프라이즈 보안

PII Safeguard, RBAC, 감사로그, 보안 가드레일, 정책 기반 통제 구조. 폐쇄망·온프레미스 환경에서 데이터 주권과 통제성 확보

경쟁제품 비교

제품	문서 레이아웃·OCR·멀티모달	GraphRAG	CRAG급 신뢰화	보안/가드레일/감사	보고서·문서 자동 생성	에이전트/툴/워크플로우	예상질문 생성	4-Score 리랭킹	독자 임베딩모델 파인튜닝
하모나이즈	**지원**	**지원**	**지원**	**지원**	**지원**	**지원**	**지원**	**지원**	**지원**
올거나이즈 Alli	부분	공개확인 어려움	부분	지원	부분	지원	미지원	공개확인 어려움	미지원
업스테이지	지원	미지원	공개확인 어려움	공개확인 어려움	부분	부분	미지원	미지원	미지원
와이즈넛 WISE iR AG V2	지원	공개확인 어려움	부분	미지원	공개확인 어려움	부분	미지원	미지원	공개확인 어려움
솔트룩스 루시아 온 / A.RAG	부분	부분	부분	지원	부분	지원	미지원	미지원	공개확인 어려움
나무기술 NAA	공개확인 어려움	공개확인 어려움	공개확인 어려움	부분	부분	지원	미지원	미지원	공개확인 어려움

표기 기준: **지원 / 부분 / 공개확인 어려움 / 미지원**. 경쟁사 기능은 **공식 공개 자료 기준**으로만 판단.

AI 모델이 자신의 환경을 이해하도록 최적화

> "최신 AI 모델을 도입했지만, 우리 회사만의 전문 용어나 특수한 업무 맥락을 전혀 이해하지 못해 답변의 품질이 실망스럽습니다. 결국 실무에 활용하기 어렵습니다."

데이터셋 생성

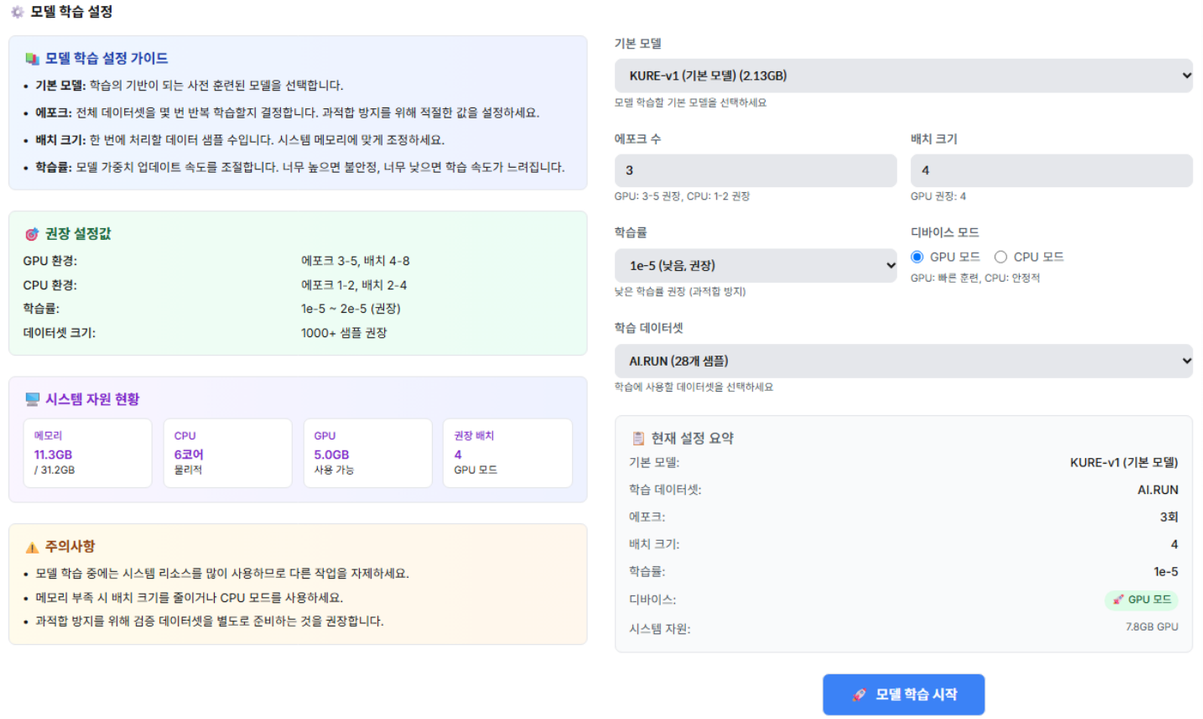
자동으로 학습을 위한 질문-답변 쌍을 생성

학습 및 관리

학습 옵션을 최적화하고 진행 상황을 모니터링

성능 평가

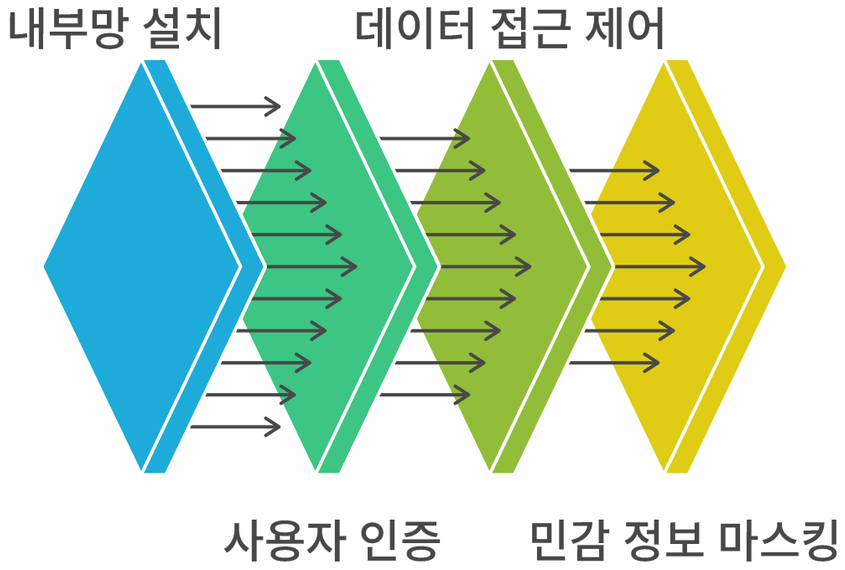
모델 성능을 객관적인 지표로 분석



사내 정보 유출, AI 도입의 가장 큰 걸림돌

> "업무에 AI를 활용하고 싶지만, 회사의 중요한 데이터가 외부로 유출될까 봐 걱정입니다.
> 클라우드 기반 AI 서비스는 보안 정책상 사용하기 어렵습니다."

하모나이즈는 엔터프라이즈 보안 및 규제 정책을 완벽하게 준수합니다.



1단계 보안

초기 보호를 위해 TLS1.3, AES-256, RBAC 및 감사 로그를 포함합니다.

1

2단계 보안

향상된 보안을 위해 온프레미스 솔루션, 에어갭 및 HSM에 중점을 둡니다.

2

3단계 보안

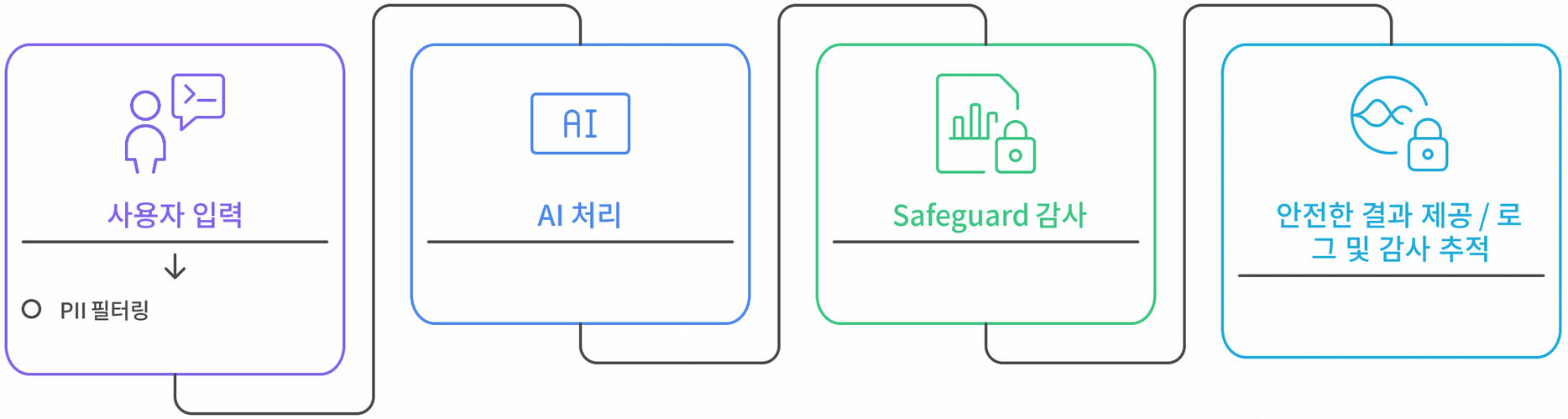
고급 방어를 위해 제로 트러스트, 256 암호화 및 예측 위협 차단을 구현합니다.

3

안전한 AI 활용을 위한 이중 보호 체계

> 1) PII 필터링은 사용자의 입력과 문서 내 개인정보·민감정보를 사전에 식별하여 불필요한 노출을 차단.
> 2) Safeguard 기반 감사 기능은 AI의 응답과 처리 과정을 점검·기록하여, 정책 위반 여부와 위험 응답 가능성을 사후 검증.

하모나이즈는 기업이 안심하고 사용할 수 있는 통제 가능한 AI 환경을 제공합니다.



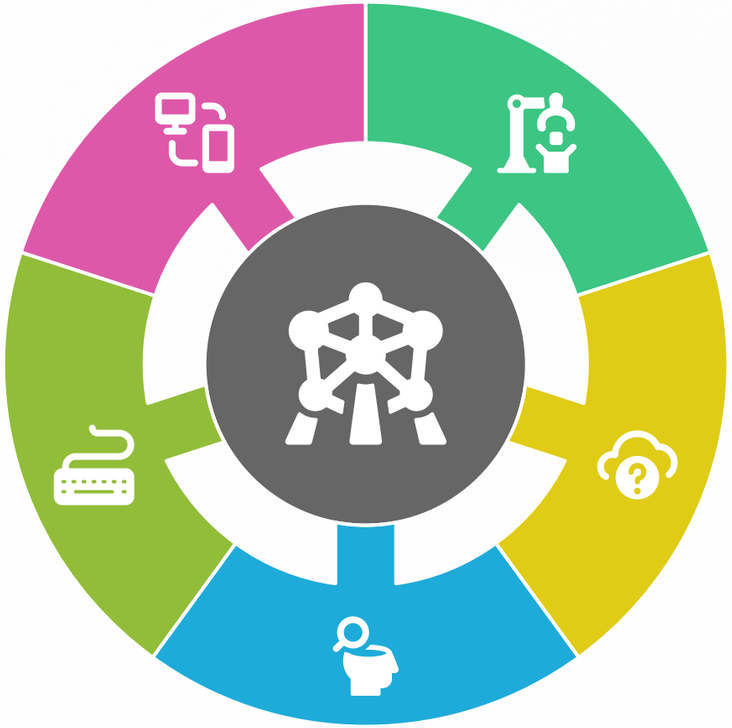
"복잡성, 부정확성, 불신, 자동화 부재를 한 번에 해결“

Smart Tune

복잡한 RAG 최적화 프로세스를 자동화하여 운영을 간소화합니다.

하이브리드 도구호출

Tool calling 지원/미지원 모델 모두 도구 사용 가능한 기술을 제공합니다.



모델 자동 선택

최적의 AI 모델을 자동으로 선택하여 효율성을 향상시킵니다.

질문 기반 임베딩

질문을 기반으로 데이터를 임베딩하여 관련성을 높입니다.

듀얼 모델 AI 추론

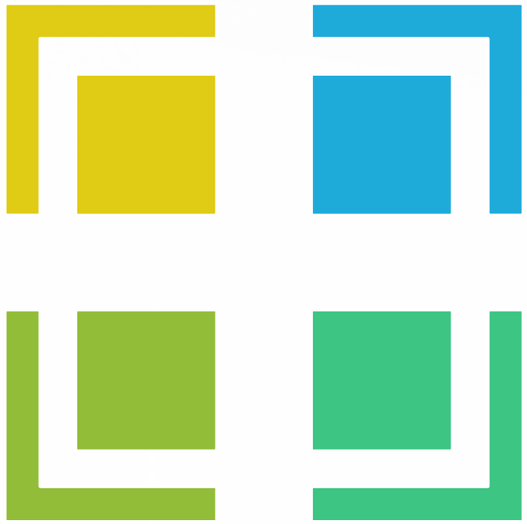
CoT, CAG 등 AI 추론 프로세스를 통해 고품질의 응답을 제공합니다.

차세대 지식 관리 RAG

> "회사에 분명 관련 자료가 있을 텐데, 어디에 누가 가지고 있는지 몰라 매번 새로 만들거나 헤맵니다.
> 과거 자료를 잘 활용해서 업무 효율을 높이고 싶습니다."

출처 명시 및 신뢰도

AI의 모든 답변은 어떤 문서의 몇 페이지를 참고했는지 명확한 출처를 함께 제공합니다.



AI 기반 의미 검색

한국어 특화 임베딩 기술을 활용하여 질문의 의도를 파악하고 정확한 내용을 찾아냅니다.

개인화 및 최적화

Embedding FineTuner를 통해 회사 용어와 데이터에 맞춰 RAG 성능을 지속적으로 최적화합니다.

자동 문서 처리

지정된 폴더에 문서를 넣기만 하면 AI가 자동으로 내용을 분석하고 학습합니다.

원하는 정보 95% 정확도

모델의 성능(정밀도, 재현율, F1 Score)을 객관적인 지표로 비교, 분석하고 가장 성능이 좋은 모델을 선택하여 현업에 바로 적용할 수 있습니다. 우리 회사 데이터와 업무에 100% 최적화된 '우리 회사만의 AI'를 갖게 되어, 차원이 다른 검색 품질과 답변 정확도를 경험할 수 있습니다.

특화 임베딩 및 파인튜닝

질문 기반 임베딩과 도메인 특화 파인튜닝

DB 및 문서 통합 로더

내부 데이터베이스와 모든 형식 문서가 완벽하게 통합.

3레벨 하이브리드 검색기

유사도, 질문, 키워드 결합 검색으로 정확도가 75%에서 95%로 향상

자율 판단 및 실행 에이전트

> "단순히 묻고 답하는 것을 넘어, AI가 스스로 계획을 세우고 여러 단계를 거쳐 업무를 처리해 주었으면 합니다.
> 예를 들어 '경쟁사 분석 보고서 초안 작성해줘' 라고 말하면 자료 검색부터 문서 작성까지 알아서 해주는 거죠."

코드 생성 및 실행

데이터 분석 및 웹 크롤링을 위한 즉석 코드 생성

강력한 도구

파일 변환 및 이미지 처리와 같은 작업을 자동화하기 위한 100개 이상의 도구

멀티 에이전트

다단계 작업 수행

복잡한 작업을 해결하기 위해 여러 도구를 순차적으로 사용

지능형 모델 선택

작업 유형에 따라 최적의 AI 모델을 자동으로 선택

전문가 수준의 지능형 문서 자동 생성

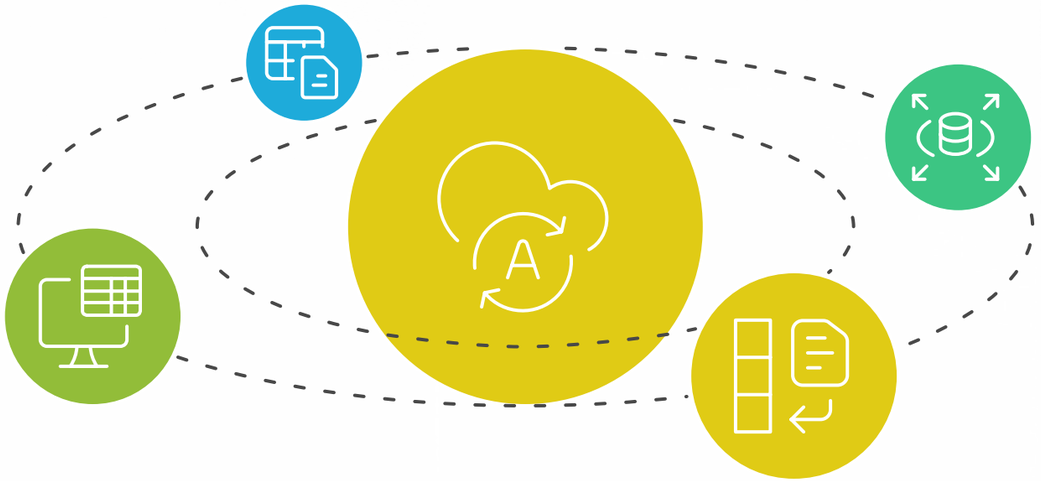
> "사업 계획서, 주간 보고서, 분석 보고서 등 반복되는 문서 작업에 너무 많은 시간을 쏟고 있습니다.
> 자료 찾고, 정리하고, 형식에 맞춰 작성하는 일이 비효율적이라고 느낍니다."

템플릿 기반 자동화

다양한 보고서 템플릿을 기반으로 동작합니다.

RAG 및 웹 검색 연동

사내 데이터베이스와 최신 웹 트렌드를 통합합니다.



시각 자료 자동 생성

차트와 다이어그램을 자동으로 생성하여 보고서를 향상시킵니다.

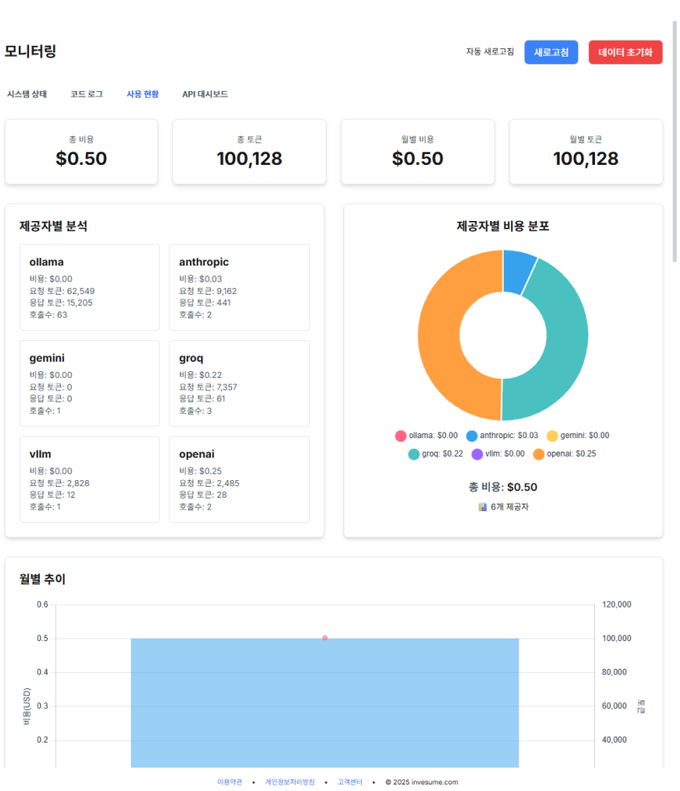
다양한 포맷 지원

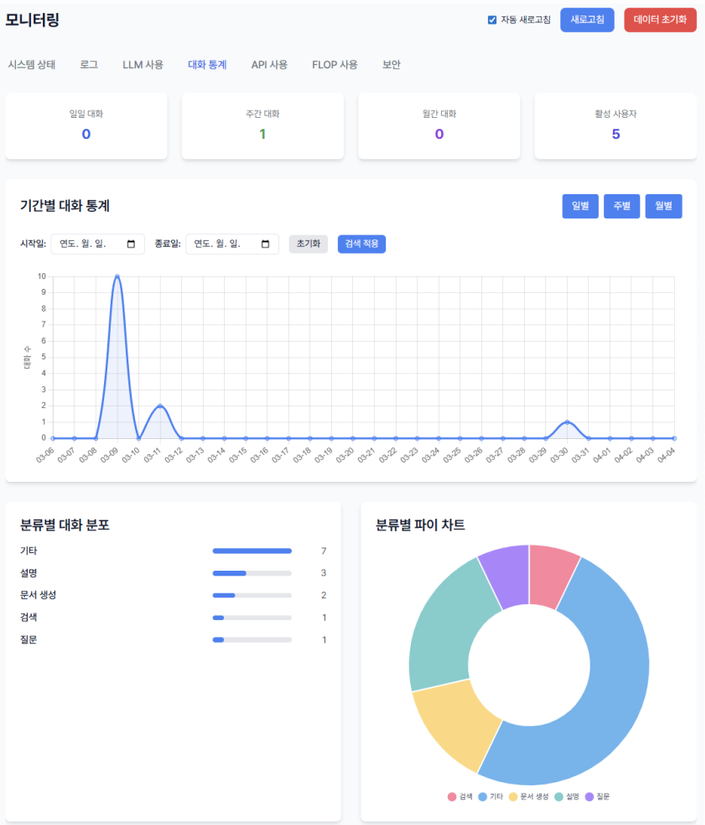
PDF, HWP, DOCX, PPTX 등 다양한 포맷으로 다운로드를 제공합니다.

직관적인 대시보드의 투명한 가시성

> "AI를 도입해도 제대로 쓰고 있는지, 성능은 괜찮은지, 문제는 없는지 파악하기가 어렵습니다.
> AI가 '블랙박스'처럼 느껴져 관리에 어려움을 겪습니다."

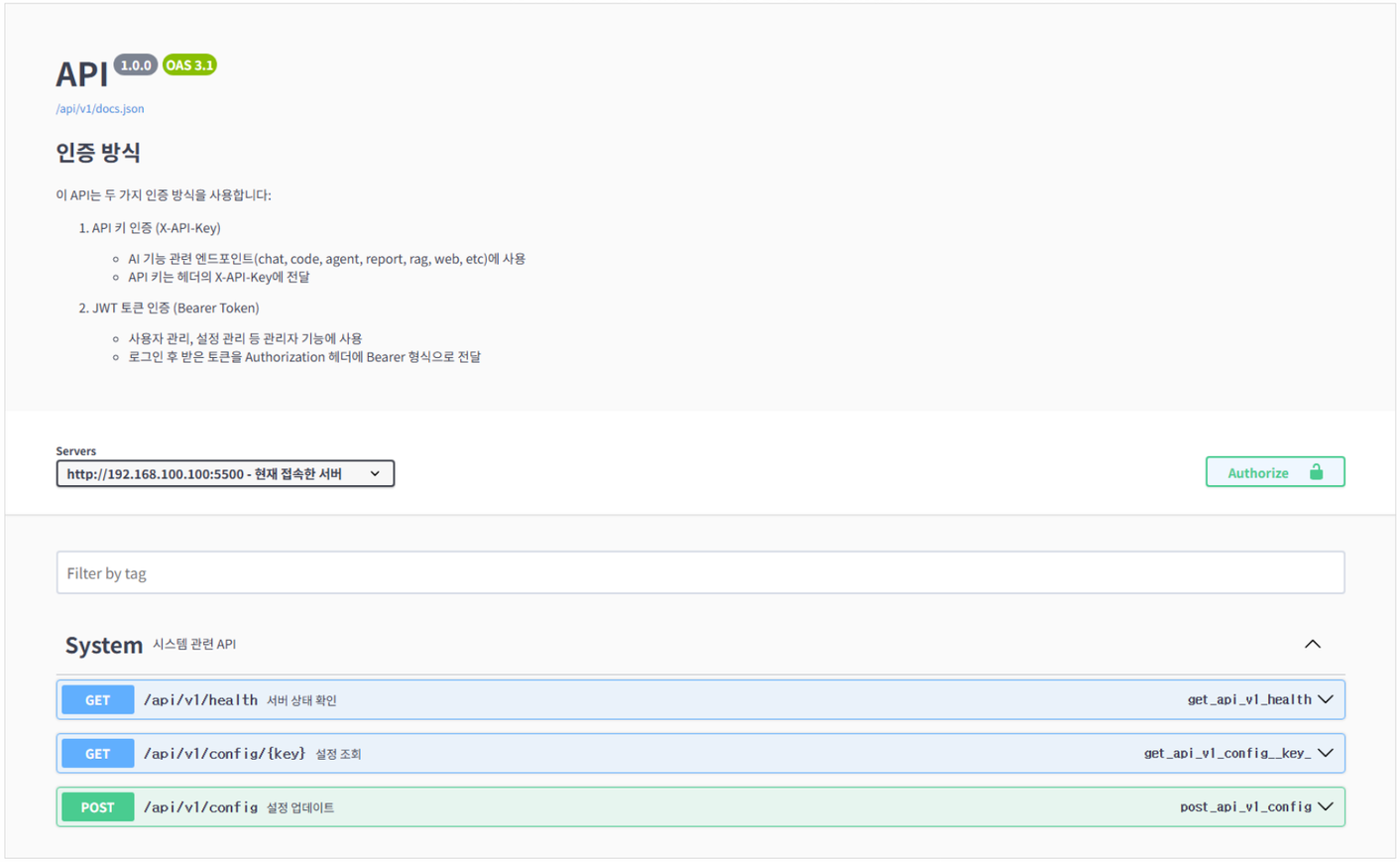
1. **투명한 운영**
   직관적인 대시보드를 통해 AI 활동을 명확하게 식별
2. **RAG 및 모델 관리**
   지식 기반을 관리하고 모델을 평가합니다
3. **사용자 관리**
   사용자 권한을 모니터링하고 관리합니다
4. **종합 설정**
   정책에 맞게 AI 설정을 사용자 정의합니다





어떤 시스템이라도 쉬운 연계 가능

> 모든 기능은 고객 시스템 연계를 위한 Open API 3.01 표준을 준수하는 API 제공으로 쉽게 사용자의 시스템에 연계할 수 있습니다.



신뢰할 수 있는 검증된 품질과 기술 수준

1등급

GOOD Software

소프트웨어품질인증서

Certificate of Software Quality

상호또는성명 Name of Company / Name of Applicant	:	㈜인베슘(198-86-00029) Invesuem, Inc.
소프트웨어의명칭 Name of Software	:	하모나이즈 v1.0 HAMONIZE v1.0
인증등급 Certification Level	:	1등급 (1등급이 2등급 보다 더 높은 등급입니다.)
인증번호 Certification No.	:	21-0642
제조자및제조국가 Manufacturer and Country of Manufacture	:	㈜인베슘/대한민국 Invesuem, Inc./Republic of Korea
인증연월일 Date of Certification	:	2021년(Year) 12월(Month) 23일(Day)
기타 Additional Information	:	

1등급

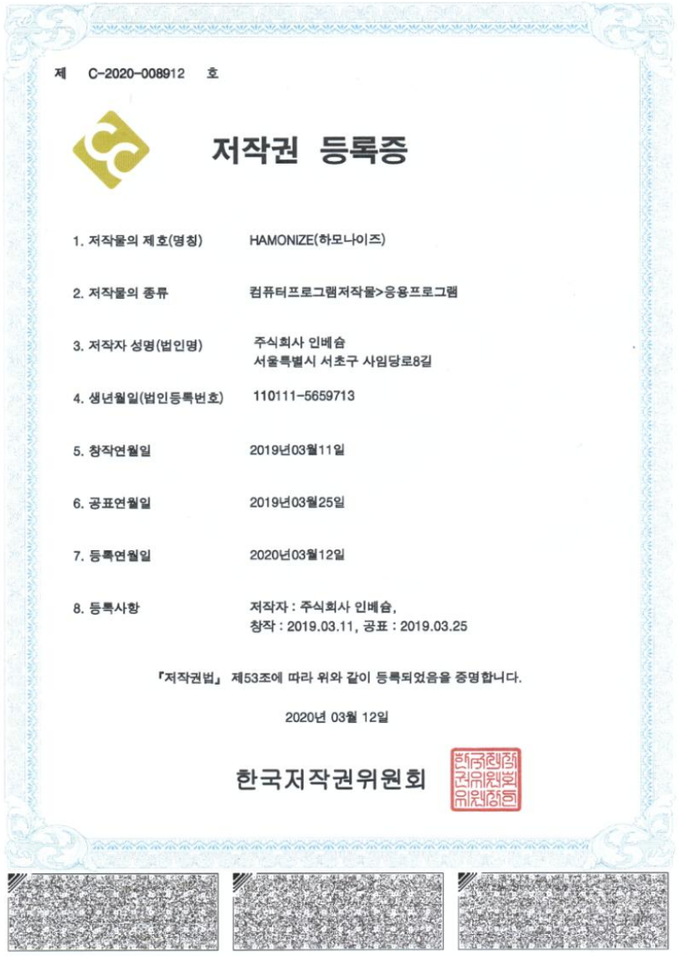
GOOD Software

위 소프트웨어는 「소프트웨어 진흥법」 제20조제3항,
같은 법 시행규칙 제6조1항에 따라 인증되었음을 증명합니다.

I hereby confirm that the quality of the foregoing software has been certified under Article 20.3 of the Software Promotion Act and Article 6.1 of its Enforcement Decree.

2021 년(Year) 12 월(Month) 23 일(Day)

TTA 한국정보통신기술협회장
CEO & President
Telecommunications Technology Association



제 C-2020-008912 호

저작권 등록증

1. 저작물의 제호(명칭)	HAMONIZE(하모나이즈)
2. 저작물의 종류	컴퓨터프로그램저작물>응용프로그램
3. 저작자 성명(법인명)	주식회사 인베슘 서울특별시 서초구 사임당로8길
4. 생년월일(법인등록번호)	110111-5659713
5. 창작연월일	2019년03월11일
6. 공표연월일	2019년03월25일
7. 등록연월일	2020년03월12일
8. 등록사항	저작자 : 주식회사 인베슘, 창작 : 2019.03.11, 공표 : 2019.03.25

「저작권법」 제53조에 따라 위와 같이 등록되었음을 증명합니다.

2020년 03월 12일

한국저작권위원회

특허. Orchestration Engine

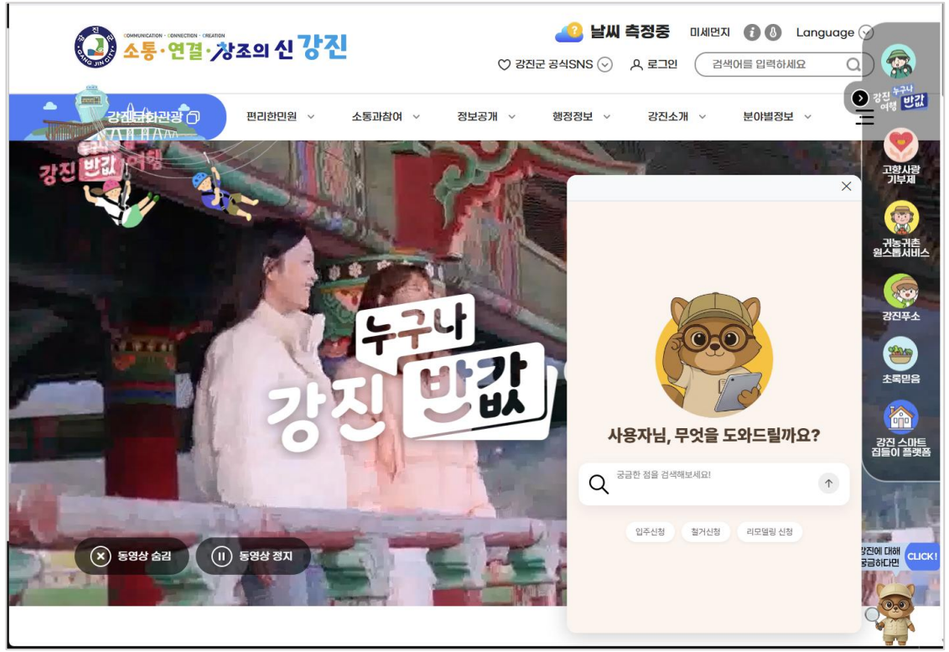
- 다중 AI 모델의 동적 오케스트레이션 기술 작업 특성에 따른 최적 모델 선택 알고리즘
- AI 모델에 따른 프롬프트 템플릿 최적화와 토큰 사용량 제어

특허. Code Executor

- 독자적 하이브리드 런타임 아키텍처
- 프로세스 격리를 통한 안정적인 코드 실행 환경 관리 (Sandbox)
- 코드에 대한 Abstract Syntax Tree (AST) 분석을 기반으로 패키지 의존성 관리

강진군청

챗봇을 넘어, 신청서 자동작성까지 수행하는 행정형 AI



주요 애로사항

- 빈집관리와 리모델링 신청절차는 행정 절차에서 민원인 문의가 많았음
- 단순 FAQ 챗봇으로는 실제 신청 과정까지 연결되지 못해 업무 효율 개선에 한계
- 담당 공무원은 반복적인 안내와 신청서 작성 지원에 많은 시간을 투입해야 했음

적용 사례

- 빈집관리와 리모델링 신청절차 전용 도구를 별도로 개발
- AI 챗봇 → Tools 호출 → RPA 연동 → 신청서 자동작성 흐름으로 업무를 연결
- 사용자의 질의응답 과정에서 필요한 조건을 수집하고, 행정 절차에 맞는 신청 양식을 생성

개선 효과

- 민원인은 절차 이해와 서류 준비 부담을 줄이고 담당자는 지원 업무를 줄일 수 있음
- 행정서비스가 “정보 제공” 에서 “업무 처리 지원” 단계로 고도화됨

> "질문에 답하는 챗봇이 아니라, 신청 업무를 끝까지 수행하는 행정형 AI 에이전트 사례"

한국수력원자력

‘읽지 못하는 AI’에서, 실시간 대답하는 교육형 AI로 전환

주요 애로사항

- 기존 이러닝 시스템에는 PDF, PPT, 이미지, 표 등 다양한 강의자료가 축적되어 있었음
- 기존 AI는 레이아웃, 표 구조, 이미지 맥락까지 충분히 이해하지 못해 답변 품질이 낮았음
- 교육생이 실시간으로 질문해도, 강의자료의 핵심 내용을 정확히 찾아 설명하지 못하는 문제

적용 사례

- 문서 레이아웃 분석모델을 적용해 강의자료의 제목, 본문, 표, 이미지, 도식 구조를 분리·이해
- 멀티모달 지원모델을 활용해 텍스트뿐 아니라 시각자료 맥락까지 함께 해석
- Graph RAG를 통해 강의 주제, 개념, 관련 문서 간 연결관계를 반영

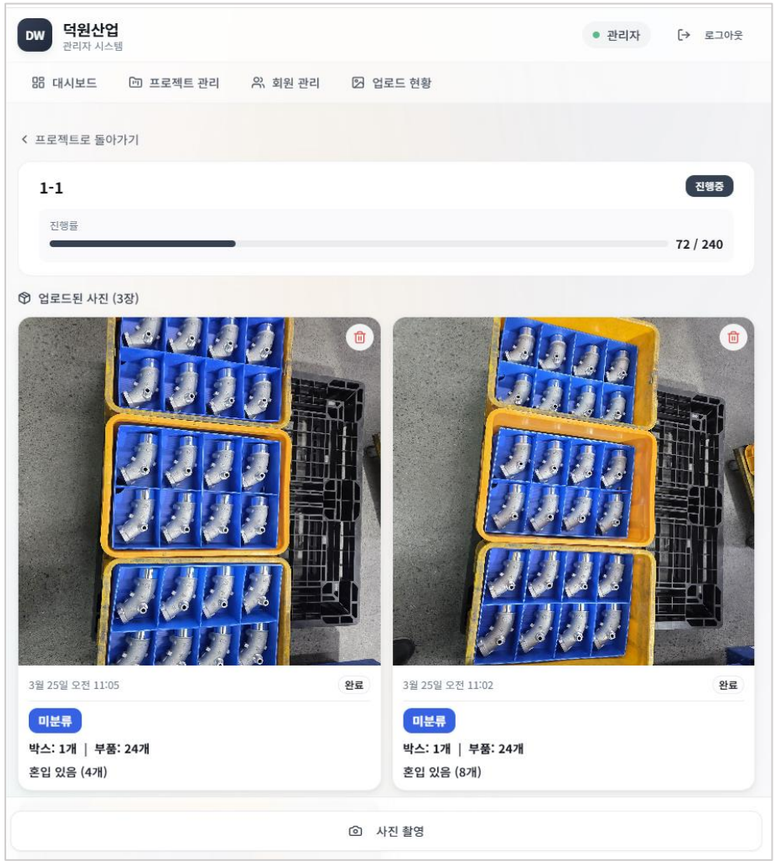
개선 효과

- 강의자료 이해 기반의 답변 정확도 향상
- 단순 키워드 검색이 아닌 맥락 기반 학습지원 가능

"문서를 읽는 AI가 아니라, 강의자료의 구조와 맥락을 이해하는 AI로 교육 응답 품질을 높인 사례"

덕원산업

현장 검수·품질관리, OCR과 Vision AI로 자동 분석 체계 구축



주요 애로사항

- 검수데이터와 현장 이미지·문서가 혼재되어 품질 판단이 어렵고
- 현장 데이터 분석이 수작업에 의존하며
- 시각 정보와 문서 정보를 함께 해석해야 하는 업무 특성 존재

적용 사례

- OCR과 Vision LLM을 활용하여 현장 데이터 자동 분석
- 검수데이터와 시각 정보를 함께 해석하여 지능형 품질관리 체계 구축
- 품질평가 결과를 일관된 기준으로 정리하고 재사용 가능하게 만듦

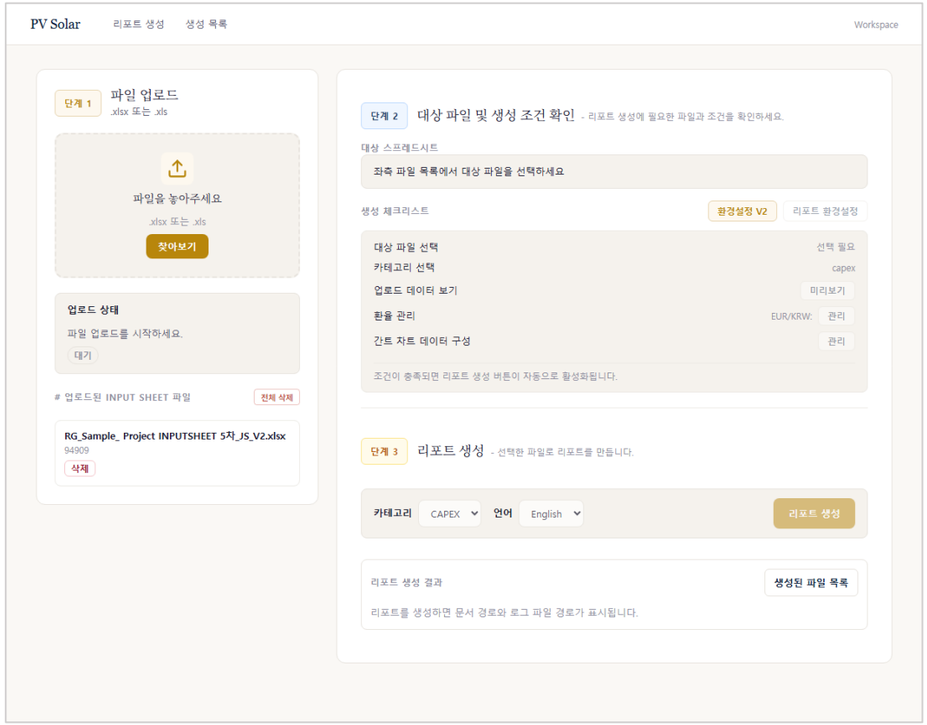
개선 효과

- 현장 검수 업무의 속도와 일관성 향상, 이미지·문서 혼합 데이터 분석 자동화
- 품질관리 업무의 표준화 및 정확도 개선

"현장 이미지와 검수 데이터를 AI가 함께 해석해 품질관리를 자동화한 사례"

오리지네이티드

복잡한 태양광 업무 데이터를 연결한 업무형 문서생성 AI



주요 애로사항

- 태양광 관련 업무는 다양한 입력데이터와 복잡한 계산 로직을 거쳐야 실제 업무 문서 작성이 가능
- 담당자가 수작업으로 데이터를 입력·계산·정리하고 결과 문서를 만드는 과정이 비효율적
- AI 챗봇은 복잡한 연산 규칙과 회사 고유의 문서 형식을 반영하기 어려웠음

적용 사례

- 별도의 업무 연산 처리 레이어를 통해 실제 업무에 필요한 계산 로직 수행
- 계산 결과를 바탕으로 기업 고유의 PowerPoint 템플릿에 맞춘 문서를 자동 생성
- AI가 단순 문답이 아니라 업무 계산 + 결과 문서화까지 연결하는 서비스로 구현

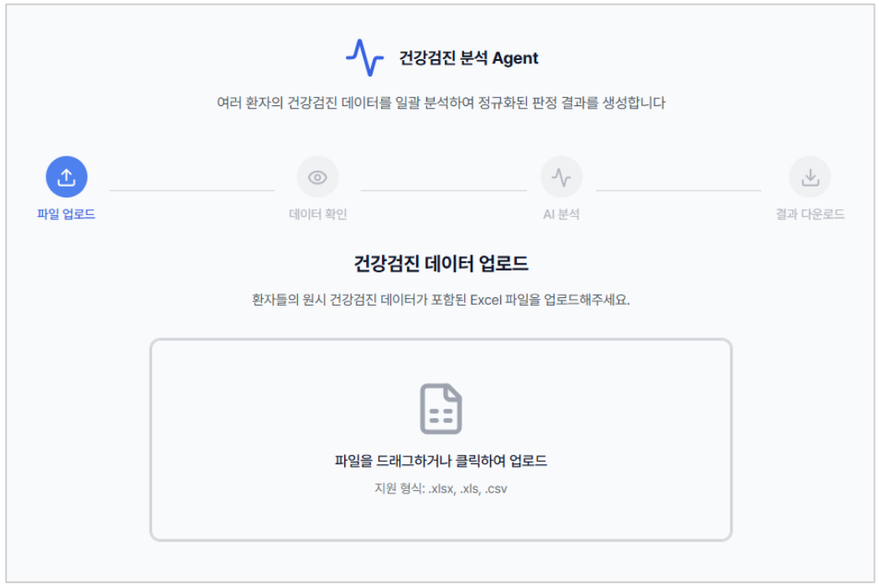
개선 효과

- 결과물의 형식과 품질을 기업 표준에 맞게 일관되게 유지
- 담당자는 계산과 작성 대신 검토와 의사결정에 집중 가능

> "복잡한 입력과 연산을 거쳐, 회사 표준 문서까지 자동 완성하는 실무형 AI 사례"

KYOCERA 건강검진센터

대량 건강검진 기록, 분석·리포트형 AI 자산으로 전환



주요 애로사항

- 연간 4만 건 이상의 건강검진 기록이 누적되지만 데이터 관리·활용 효율이 낮음
- 대량 기록을 사람이 직접 확인·정리하는 데 시간이 많이 소요됨
- 데이터 기반 리포트 생성 과정이 비효율적임

적용 사례

- 건강검진 데이터를 AI 분석 가능 구조로 정리
- 기록에서 필요한 정보와 패턴을 빠르게 추출
- 반복적 데이터 정리 업무를 줄이고 결과 해석 중심 업무로 전환

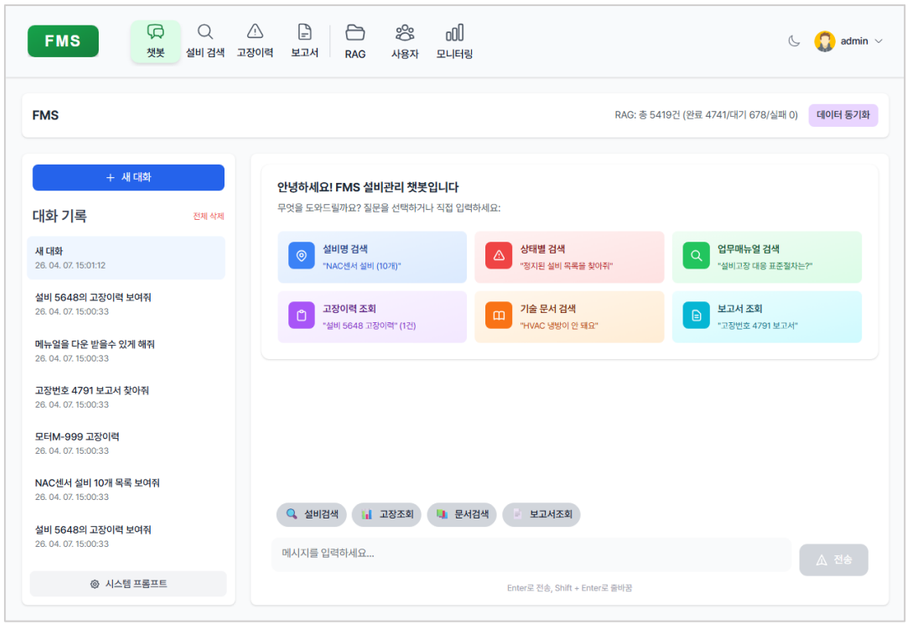
개선 효과

- 데이터 관리 효율 향상, 대량 기록 기반 분석
- 단순 저장 데이터가 활용 가능한 의료·헬스케어 지식 자산으로 전환

> "대량 검진 기록을 AI가 읽고 정리해, 실질적인 헬스케어 서비스로 연결한 사례"

비전아이티

설비관리 지식, 담당자 경험 의존에서 AI 자동점검 지원



주요 애로사항

- 설비 관련 정보와 QA 데이터가 분산되어 장애 원인 확인이 어려움
- 고장 데이터베이스는 있지만 실무자가 필요한 정보를 빠르게 찾기 어려움
- 점검 보고서 작성이 경험 많은 담당자에게 의존적

적용 사례

- 도메인 특화 설비 DB와 QA 데이터를 바탕으로 설비 챗봇 구축
- 고장 데이터베이스와 과거 사례를 검색해 자동 점검 보고서 생성
- 실무자는 AI가 정리한 결과를 바탕으로 빠르게 검토·조치

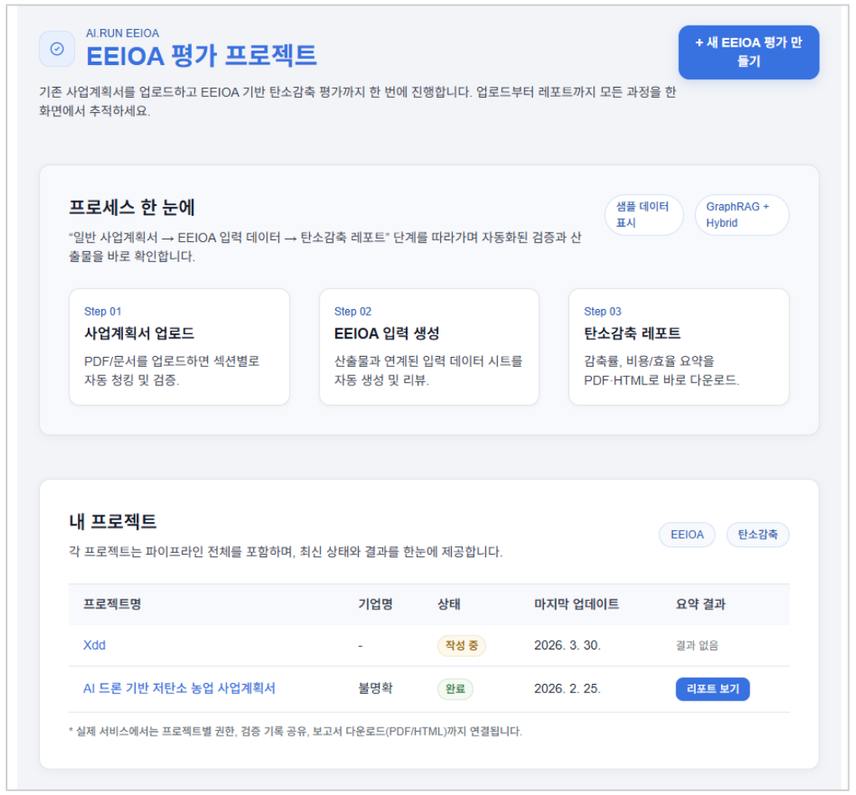
개선 효과

- 설비 관련 정보 접근 속도 향상, 장애 원인 확인과 점검 보고 업무 효율화
- 숙련자 의존도를 낮추고 지식의 재사용성 강화

> "설비 담당자의 경험을, 누구나 활용할 수 있는 AI 점검 지원 체계로 만든 사례"

교보증권

ESG·규제 대응 리포트, AI 평가 에이전트로 혁신



주요 애로사항

- ESG 의무보고 대응을 위해 다양한 데이터에서 탄소배출량 수집·정리 필요
- 국제표준 기준에 맞춘 보고서 작성에 많은 시간과 인력 소요
- 데이터는 존재하지만 평가와 리포트 문서화 과정이 비효율적

적용 사례

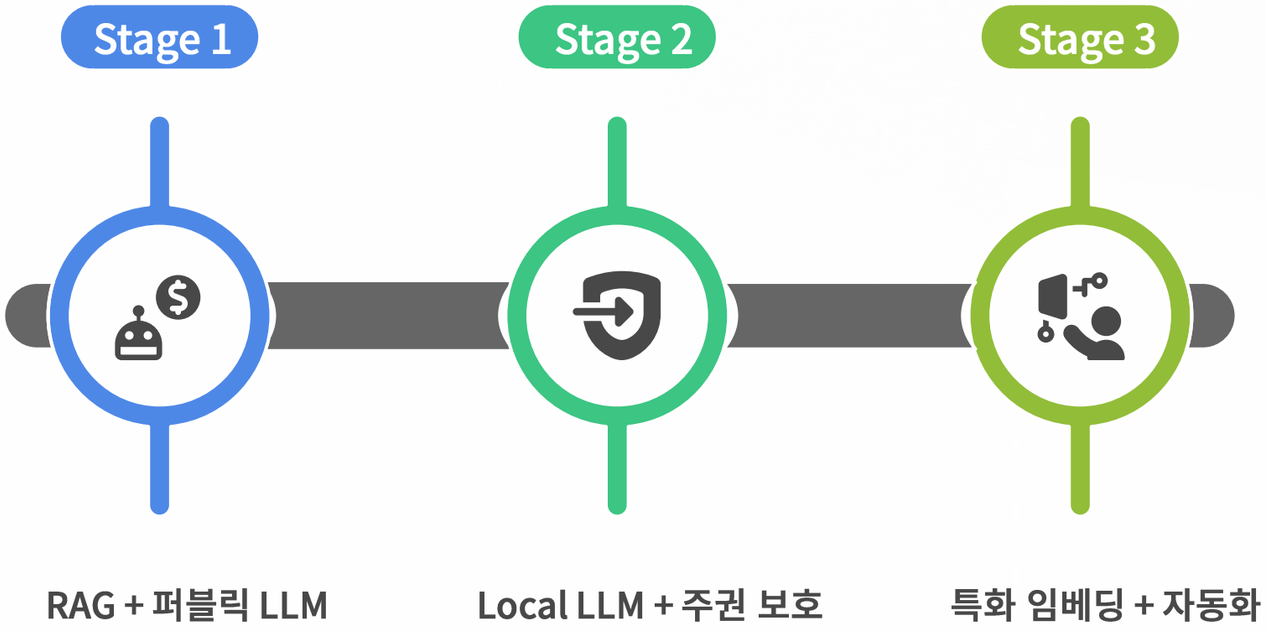
- EEIOA 평가 기준을 반영한 AI에이전트 적용
- 관련 데이터와 기준 문서를 AI가 함께 참조하여 탄소배출량 자동 집계
- 국제표준 기준에 맞춘 리포트 초안 자동 생성

개선 효과

- ESG 보고 준비 시간 단축, 수작업 집계 오류 가능성 감소
- 규제 대응형 보고체계 자동화로 업무 신뢰성과 일관성 향상

> "복잡한 ESG 평가와 보고를, AI가 기준에 맞춰 정리해주는 금융권 사례"

소버린 AI 구축 로드맵



단계	기능	하드웨어
Stage 1	RAG + 퍼블릭 LLM	No GPU
Stage 2	Local LLM + 주권 보호	Inference AI 서버 (L40S 4GPU 이상)
Stage 3	도메인 특화 전용 모델	학습용 AI 서버

🌏 소버린 AI(Sovereign AI) 는 데이터와 AI 모델을 자국(또는 조직) 주권 하에 두어 독립적으로 운영·통제하는 AI를 의미.

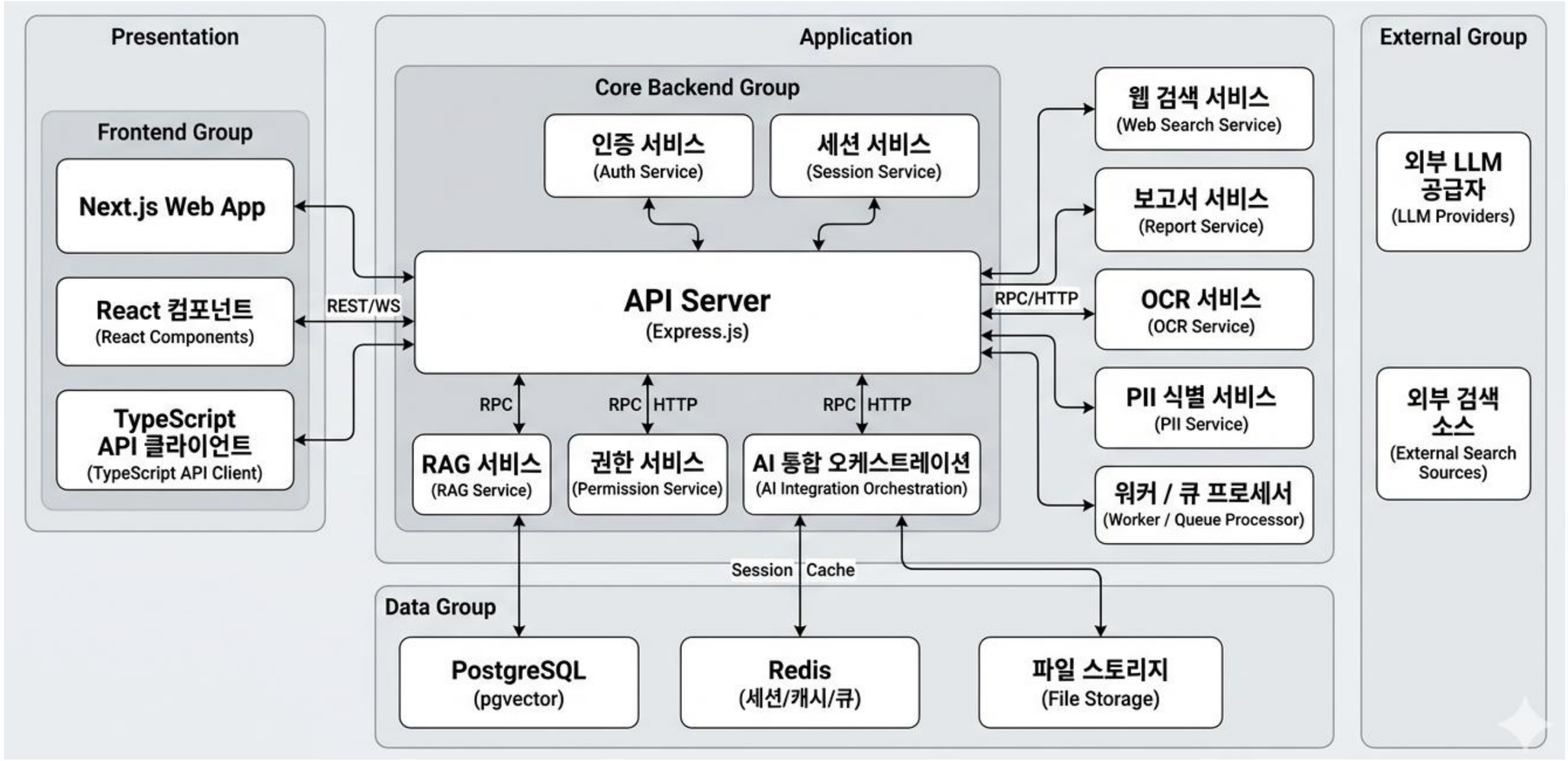
온프레미스 최소/권장 사양

하모나이즈는 **검색/웹 서버와 추론서버를 분리하지 않고 1대 통합 서버**로도 운영 가능하며, 최소사양은 **L40S 4장**, 권장사양은 **L40S 8장 또는 H100 3장** 이 적절.
(통합 서버 1대 기준, 동시접속 증가 시 GPU 확장 필요.)

구분	최소사양	권장사양
구성 방식	검색/웹 + 추론 + RAG + 문서파싱 통합 1대 서버	검색/웹 + 추론 + RAG + 문서파싱 통합 1대 서버
OS	**Ubuntu 24.04 이상**	**Ubuntu 24.04 이상**
CPU	32 Core 이상	64 Core 이상
GPU	**L40S × 4ea 이상**	**L40S × 8ea 이상** 또는 **H100 × 3ea 이상**
메모리	256GB 이상	512GB 이상
디스크	2TB NVMe 이상	4TB NVMe 이상
Network	1Gbps Ethernet 이상	10Gbps Ethernet 이상 권장
적용 대상	온프레미스 기본 운영, 일반 업무형 챗봇 + RAG + 문서 파싱	기관/기업 실서비스, 문서 업로드·비전·리포트·RAG 동시 운영

- **실제 병목은 AI 추론과 큐 대기**: 전체 응답 시간의 대부분이 AI 호출에 집중. 웹 최적화보다 GPU 인프라와 큐 제어가 핵심

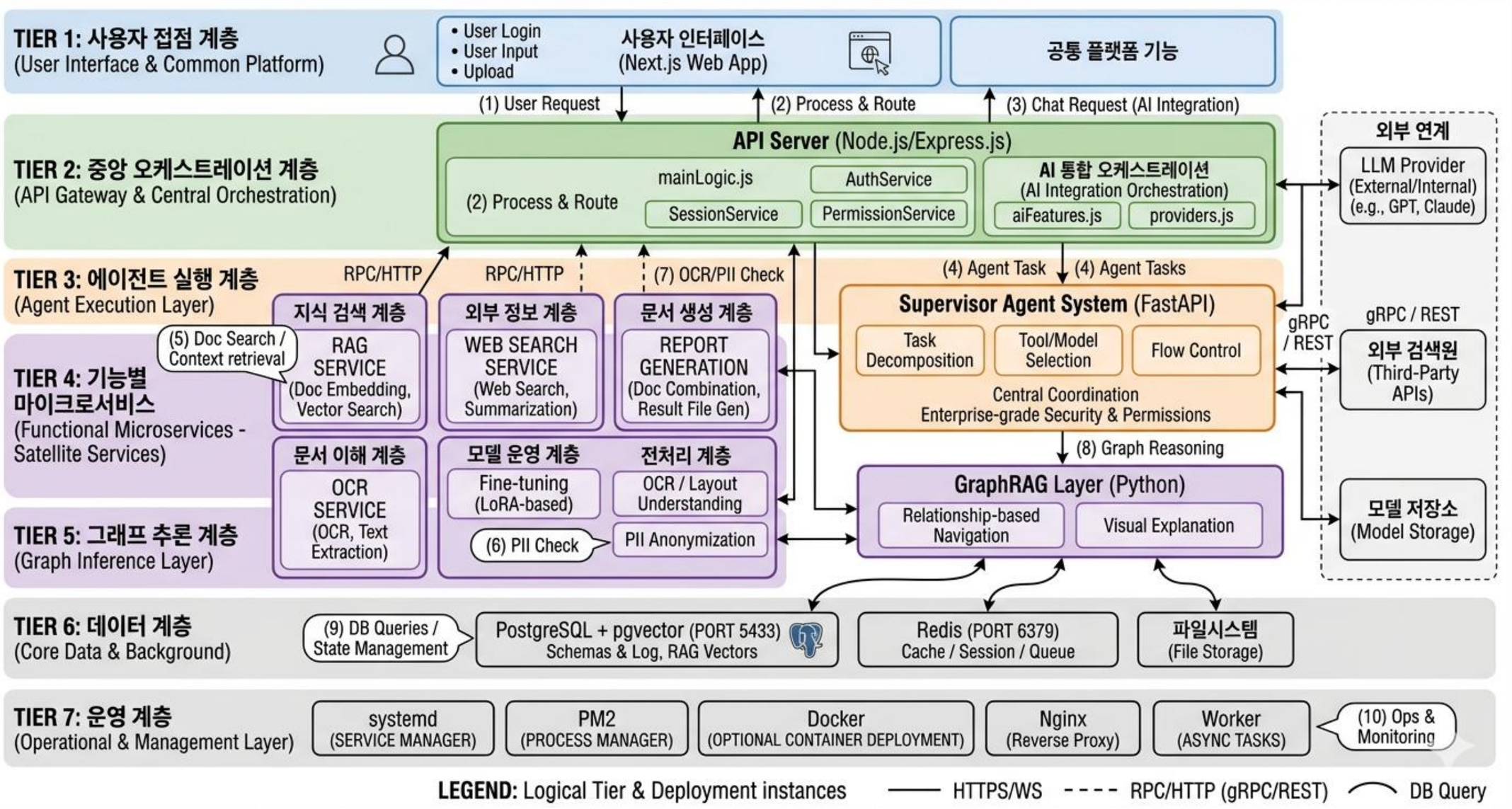
별첨. 플랫폼 아키텍처


Presentation
Frontend Group
Next.js Web App
React 컴포넌트
(React Components)
TypeScript
API 클라이언트
(TypeScript API Client)
REST/WS
Application
Core Backend Group
인증 서비스
(Auth Service)
세션 서비스
(Session Service)
API Server
(Express.js)
RPC
RPC
HTTP
RPC
HTTP
RAG 서비스
(RAG Service)
권한 서비스
(Permission Service)
AI 통합 오케스트레이션
(AI Integration Orchestration)
RPC/HTTP
웹 검색 서비스
(Web Search Service)
보고서 서비스
(Report Service)
OCR 서비스
(OCR Service)
PII 식별 서비스
(PII Service)
워커 / 큐 프로세서
(Worker / Queue Processor)
Session
Cache
Data Group
PostgreSQL
(pgvector)
Redis
(세션/캐시/큐)
파일 스토리지
(File Storage)
External Group
외부 LLM
공급자
(LLM Providers)
외부 검색
소스
(External Search
Sources)

별첨. 주요 기능 목록

주요 기능	엔터프라이즈 AI 실제 요구사항	기능 설명	책임 컴포넌트/서비스
대화형 업무 지원	사용자가 AI와 실시간 대화	Web + API + WebSocket 기반 채팅, 세션 저장, 문맥 유지	Next.js, API Server, PostgreSQL, Redis
대화 기억 및 컨텍스트 유지	이전 대화와 업무 이력을 이어서 활용	세션/인증 세션 저장, 히스토리 관리, 장기 기억 연계	PostgreSQL, Redis, 세션 계층
문서특화 멀티모달 RAG	PDF/스캔/이미지 문서를 읽고 검색 가능한 지식으로 변환	업로드 → OCR → 텍스트/레이아웃 추출 → 벡터 인덱싱 → 검색	OCR Server, RAG Server, PostgreSQL + pgvector, 파일시스템
설명 가능한 지식 검색	답변과 함께 출처, 근거, 관계를 제시	RAG + 출처 인용 + CRAG + Graph RAG	RAG, Graph RAG, PostgreSQL + pgvector
그래프 기반 지식 탐색	단순 유사도 검색을 넘어 관계 기반 탐색 수행	Graph 기반 관계 시각화·탐색	Graph RAG, RAG/DB 연계
AI 문서 생성 자동화	최신 정보를 결합해 보고서/문서 생성	RAG 검색 + 웹검색 + Report 생성 파이프라인	Report Server, Web Search, PostgreSQL, 파일시스템
웹 정보 통합	내부 지식 외 최신 외부 정보를 결합	독립 Web Search 서비스 호출 및 결과 통합	Web Search Server, Redis, API Server
슈퍼바이저 에이전트 실행	작업 단계를 조율하고 적절한 도구/모델을 선택	오케스트레이션, 툴/모델 라우팅	API Orchestrator, Agent System
선택형 멀티모델 워크플로우	요청 유형에 따라 다른 모델을 선택적으로 사용	providers 및 모델 선택 로직, MOE 지향 운영	OpenAI/Anthropic/Gemini/Groq/Ollama 연계
결과 품질 보정	응답 품질을 스스로 평가하고 재시도	Self-Reflection, CRAG, 쿼리 재작성	오케스트레이션 계층
보안/감사/정책 통제	기업 규정에 따라 사용자·행위·키를 통제	Auth, Permission, API Keys, Logs, Activity Logs	PostgreSQL, API Server
모델 개선 및 내재화	모델을 튜닝해 도메인 적합도 향상	llm-finetune, finetune-cli, lora	Fine-tuning 서비스, 모델 저장소, 파일시스템

별첨. 시스템 아키텍처


TIER 1: 사용자 접점 계층
(User Interface & Common Platform)
• User Login
• User Input
• Upload
사용자 인터페이스
(Next.js Web App)
공통 플랫폼 기능
(1) User Request
(2) Process & Route
(3) Chat Request (AI Integration)
TIER 2: 중앙 오케스트레이션 계층
(API Gateway & Central Orchestration)
API Server (Node.js/Express.js)
(2) Process & Route
mainLogic.js
AuthService
SessionService
PermissionService
AI 통합 오케스트레이션
(AI Integration Orchestration)
aiFeatures.js
providers.js
외부 연계
LLM Provider
(External/Internal)
(e.g., GPT, Claude)
TIER 3: 에이전트 실행 계층
(Agent Execution Layer)
RPC/HTTP
RPC/HTTP
(7) OCR/PII Check
(4) Agent Task
(4) Agent Tasks
TIER 4: 기능별 마이크로서비스
(Functional Microservices - Satellite Services)
지식 검색 계층
(5) Doc Search / Context retrieval
RAG SERVICE
(Doc Embedding, Vector Search)
외부 정보 계층
WEB SEARCH SERVICE
(Web Search, Summarization)
문서 생성 계층
REPORT GENERATION
(Doc Combination, Result File Gen)
Supervisor Agent System (FastAPI)
Task Decomposition
Tool/Model Selection
Flow Control
Central Coordination
Enterprise-grade Security & Permissions
gRPC / REST
gRPC / REST
외부 검색원
(Third-Party APIs)
(8) Graph Reasoning
문서 이해 계층
OCR SERVICE
(OCR, Text Extraction)
모델 운영 계층
Fine-tuning
(LoRA-based)
전처리 계층
OCR / Layout Understanding
(6) PII Check
PII Anonymization
GraphRAG Layer (Python)
Relationship-based Navigation
Visual Explanation
모델 저장소
(Model Storage)
TIER 5: 그래프 추론 계층
(Graph Inference Layer)
TIER 6: 데이터 계층
(Core Data & Background)
(9) DB Queries / State Management
PostgreSQL + pgvector (PORT 5433)
Schemas & Log, RAG Vectors
Redis (PORT 6379)
Cache / Session / Queue
파일시스템
(File Storage)
TIER 7: 운영 계층
(Operational & Management Layer)
systemd
(SERVICE MANAGER)
PM2
(PROCESS MANAGER)
Docker
(OPTIONAL CONTAINER DEPLOYMENT)
Nginx
(Reverse Proxy)
Worker
(ASYNC TASKS)
(10) Ops & Monitoring
LEGEND: Logical Tier & Deployment instances
HTTPS/WS
RPC/HTTP (gRPC/REST)
DB Query