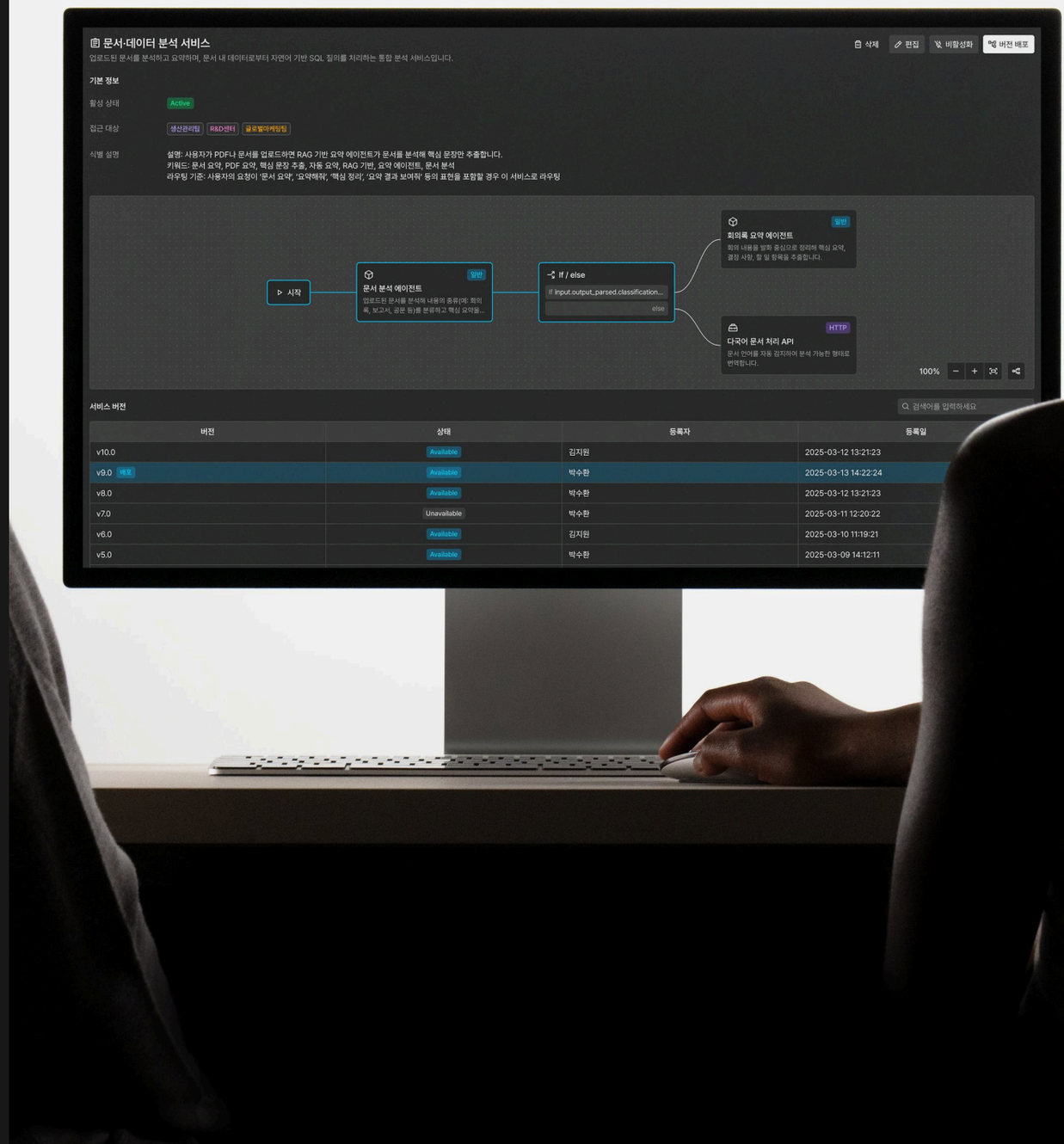
exemble

기업 데이터 기반 생성형 AI를 온프레미스에서 통합 운영하다

온프레미스에서 완성하는 엔터프라이즈 AI 운영

온프레미스·폐쇄망 환경에서 LLM 모델 배포, 문서 벡터화, AI 에이전트 생성, 보안·감사까지 AI 서비스 운영에 필요한 모든 요소를 단일 플랫폼에서 제공하는 통합 LLMOps 솔루션입니다.

사용자는 내부 데이터를 안전하게 활용하면서 RAG 기반 AI 서비스를 빠르게 구축하고 지속적으로 개선할 수 있습니다.



exemble

엑셈블이 남다른 이유

구축에서 운영까지, AI 서비스 전 과정을 하나로 통합

생성형 AI 도입이 확대되면서 기업은 데이터 보안, 모델 운영, 비용 통제 등 새로운 운영 과제에 직면하고 있습니다. 특히 공공기관과 금융기관은 폐쇄망 환경에서 민감 데이터를 보호하면서도 AI 서비스를 안정적으로 운영해야 합니다.

그러나 퍼블릭 AI는 외부 서버로 데이터가 전송되어 보안 우려가 크고, 조직 내부 업무에 특화된 답변을 제공하기 어렵습니다. eXemble은 온프레미스·폐쇄망 환경에서 모델 배포부터 데이터 벡터화, 에이전트 생성, 보안·감사까지 AI 서비스 운영의 전 과정을 단일 플랫폼에서 통합 관리합니다.

Customer Stories

고객 사례

공공기관

폐쇄망 환경에서 업무 자동화 AI 서비스 구축

도입으로 망분리 환경에서 내부 문서 기반 AI 챗봇을 구축했습니다. 기존 수작업 대비 처리 시간 70% 단축, 조직별 접근 권한 분리로 보안 감사 요건을 충족했습니다.

금융기관

NL2SQL 기반 사내 데이터 질의 서비스 도입

NL2SQL 에이전트를 통해 비개발자도 자연어로 사내 DB를 조회할 수 있게 되었습니다. 분석 인력 의존도를 50% 낮추고 데이터 기반 의사결정 속도를 크게 향상시켰습니다.

Product Highlights

제품 특장점

온프레미스·폐쇄망에서 구현하는 엔터프라이즈 AI

폐쇄망 환경에서 LLM 서빙과 RAG 파이프라인을 완벽하게 구축하고 공공·금융 보안 규정을 충족하는 AI 서비스를 제공합니다.

사용자 피드백으로 진화하는 AI 품질 개선

채팅 피드백을 수집해 RLHF 강화학습을 자동 수행하고 학습 완료 시 새 버전으로 자동 등록됩니다.

비정형 데이터의 자동 수집과 지능형 벡터화

PDF, HWP, 이미지, DB까지 자동 수집하여 벡터로 변환하고 데이터 변경 시 자동 재벡터화로 검색 품질을 유지합니다.

고속 추론 엔진과 GPU 자원 최적화

고효율 서빙으로 대규모 요청을 안정적으로 처리하고 GPU 이상탐지로 비용을 절감합니다.

코드 없이 설계하는 AI 에이전트 워크플로우

드래그 앤 드롭으로 AI 에이전트와 도구를 연결하여 업무 흐름을 시각적으로 설계하고 즉시 테스트할 수 있습니다.

AI 서비스 전체를 조망하는 통합 운영 대시보드

사용량, 응답속도, 비용 지표를 한 화면에서 확인하고 인프라까지 통합 모니터링합니다.

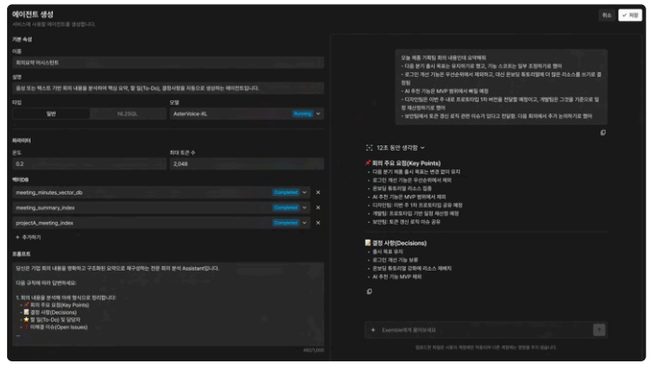
AI 에이전트 구성

개발 지식 없이도 업무 목적에 맞는 AI 에이전트를 직접 구성할 수 있습니다. 드래그 앤 드롭 방식의 워크플로우 캔버스에서 에이전트와 외부 도구를 조합해 복잡한 AI 서비스를 시각적으로 설계합니다.

1 에이전트 빌더

목적에 맞는 AI 에이전트를 손쉽게 생성

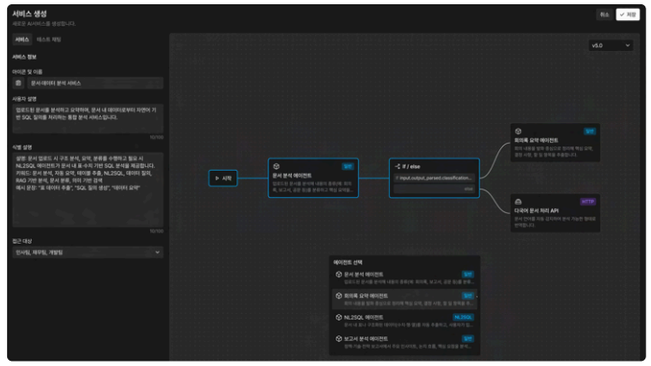
- 모델, 벡터 테이블, 시스템 프롬프트 조합으로 에이전트 구성
- 테스트 채팅으로 변경 사항 적용 결과 즉시 확인
- 사용자 피드백을 반영한 학습된 에이전트 사용 가능



2 워크플로우 캔버스

노드 기반 시각적 설계로 복잡한 AI 서비스 구현

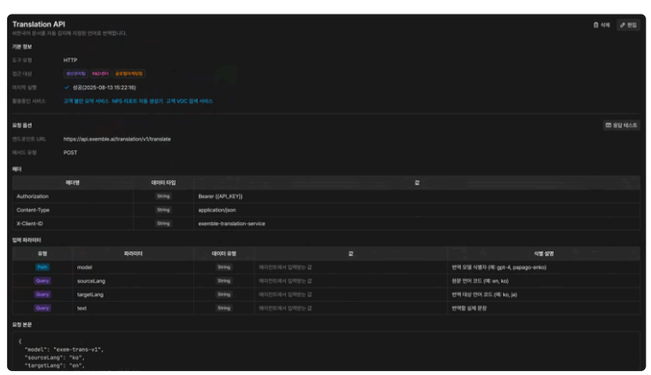
- AI 에이전트와 외부 도구를 드래그 앤 드롭으로 연결
- 작업 흐름과 도구 호출 과정을 캔버스에서 실시간 추적
- 답변 생성 경로를 시각적으로 확인해 장애 원인 분석 용이



3 외부 도구 연동

API와 사내 시스템을 재사용 가능한 도구로 통합 관리

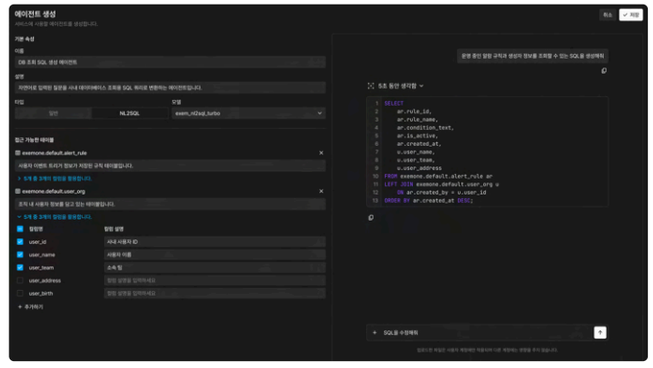
- HTTP API, MCP Tool 등 다양한 외부 도구 저장 후 재사용
- 도구 생성 시 테스트 실행으로 정상 동작 즉시 검증
- LLM 생성 값 또는 사용자 입력 값 기반 동적 파라미터 처리



4 자연어 DB 질의를 통한 차트 생성

자연어 질문을 SQL로 변환해 자동 실행

- 비개발자도 손쉽게 사내 DB 조회 가능
- 분석 인력 의존도를 낮추고 데이터 활용 범위 확대
- 쿼리 결과 기반 즉시 응답 생성으로 업무 속도 향상



운영 모니터링

AI 서비스별 사용량, 비용, 응답속도 등 핵심 지표를 단일 대시보드에서 한눈에 파악합니다. 에이전트와 서비스의 변경 이력을 자동으로 버전 관리하고 최적 버전을 선별해 즉시 배포할 수 있습니다.

1 통합 대시보드

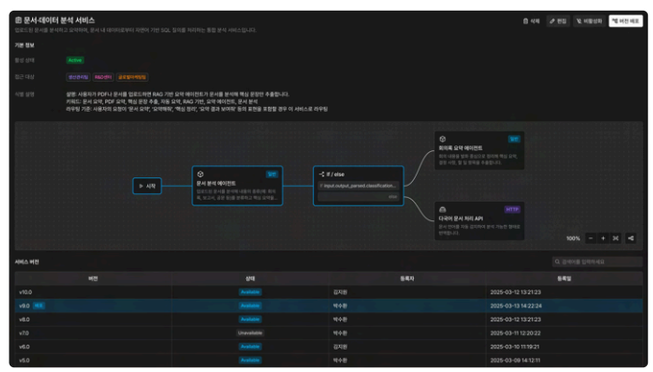
AI 서비스의 사용량·비용·성능을 단일 화면에서 통합 모니터링

- 서비스별 사용량, 비용, 응답속도 등 핵심 KPI 실시간 집계
- 특정 서비스 선택으로 상세 지표 필터링 가능
- 서버 성능 추이 분석으로 용량 계획 수립 지원

2 자동 버전 관리

에이전트와 서비스의 모든 변경 이력을 자동 추적하고 관리

- 모델, 벡터 테이블, 프롬프트 변경 시 버전 자동 생성
- 이전 버전과 비교해 어떤 값이 바뀌었는지 상세 추적
- 서비스 중단 없이 원하는 버전으로 즉시 전환



3 에이전트 평가

버전별 성능을 비교해 최적의 에이전트 선별

- 평가 데이터셋 업로드로 정확도 측정
- 버전별 성능 비교 후 상위 에이전트 자동 식별
- 가장 성능이 좋은 버전을 원클릭으로 배포

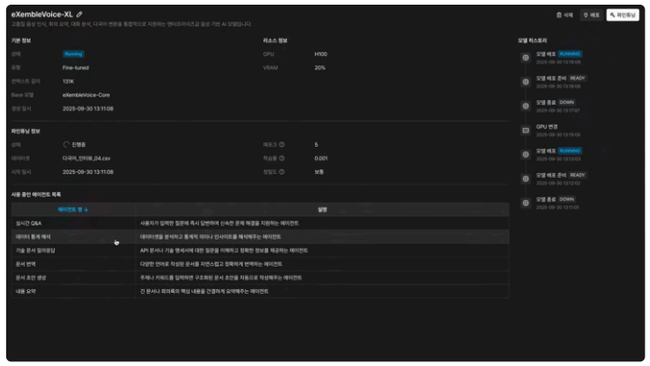
모델 운영

효율적인 모델 서빙 엔진으로 대규모 언어 모델을 안정적으로 서빙합니다.
여러 모델을 동시에 운영하면서 요청 특성에 따라 최적의 모델로 자동 연결해
운영 효율과 응답 품질을 함께 높입니다.

1 모델 등록과 버전 관리

플랫폼 내 모든 모델을 중앙에서 체계적으로 관리

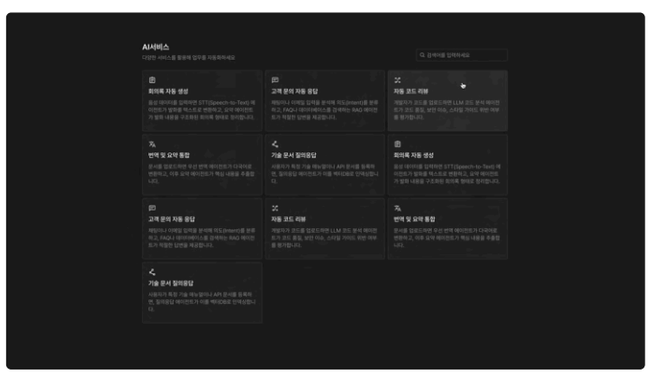
- 모델별 상세 정보와 배포 상태 실시간 조회
- 배포 히스토리 추적으로 변경 이력 완전 보존
- GPU 선택, VRAM 할당량 설정으로 서빙
  환경 최적화



2 고효율 추론 엔진

대규모 요청도 빠르고 안정적으로 처리

- 효율적인 모델 서빙 엔진으로 처리속도 향상,
  운영비용 절감
- 동시 다발적 요청에도 응답속도 일정 수준 유지
- 부하 자동 조절로 서비스 가용성 확보



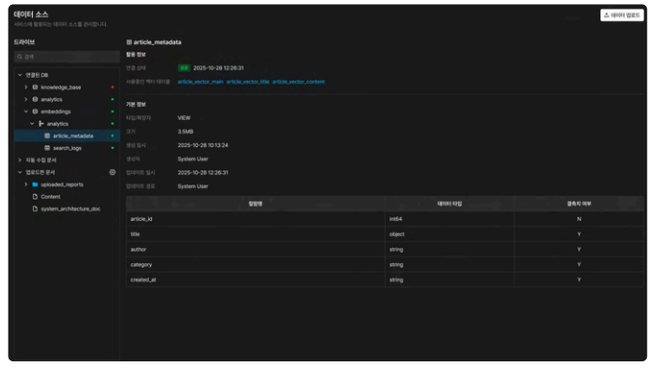
데이터 통합 관리

사내 문서, 데이터베이스, 자동 수집 파일까지 흩어진 데이터 자산을
한 곳에서 관리합니다. AI가 이해할 수 있는 벡터 인덱스로 자동 변환하고
데이터 변경 시 실시간 동기화로 항상 최신 정보 기반의 응답을 보장합니다.

1 데이터소스 통합

흩어진 데이터 자산을 한 화면에서 통합 탐색하고 중앙 관리

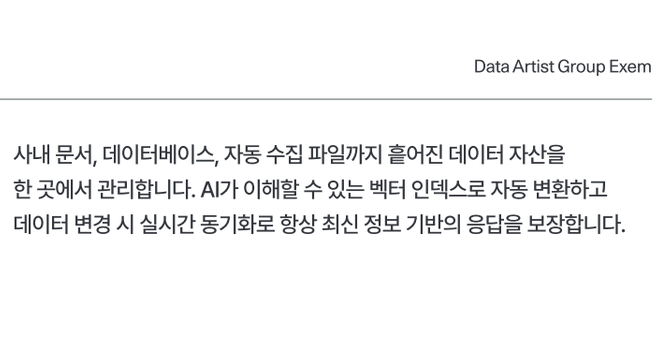
- 연결된 DB, 업로드 문서, 자동 수집 문서를 계층
  구조로 분류
- 데이터 소스별 연결 상태와 벡터 테이블 사용 현황
  실시간 확인
- 스키마, 테이블, 폴더, 파일 단위의 세분화된 탐색 지원



2 자동 벡터화

문서·데이터를 AI가 이해하는 형태로 자동 변환

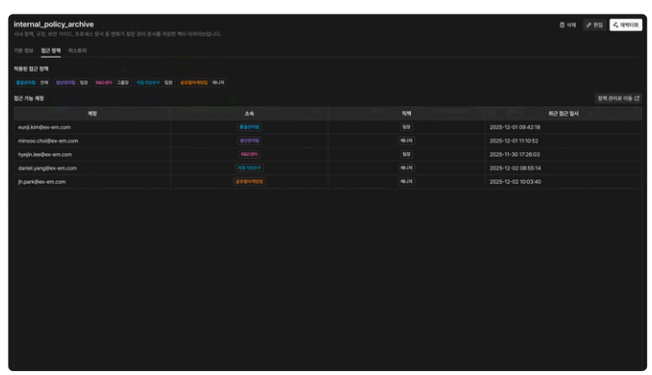
- PDF, HWP, 이미지 등 다양한 포맷의 OCR 및
  텍스트 추출
- 파일, 폴더, 데이터베이스 쿼리 결과까지
  벡터화 대상 확장
- 데이터 변경 감지 시 자동 재처리로 최신 상태 유지



3 AI 검색 데이터 관리

벡터 테이블 기반의 조직 전용 지식 저장소 구축

- 벡터화 작업 상태별 그룹화와 진행 히스토리 추적
- 소속·직책별 접근 정책 설정으로 민감 데이터 보호
- 수동 재벡터화 기능으로 운영자 주도의 데이터 갱신



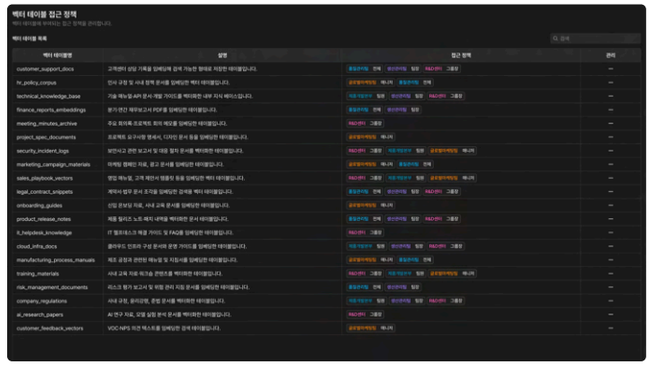
보안과 권한 관리

사용자와 조직 단위로 세분화된 접근 권한을 설정해 민감 정보를 보호합니다. 부서·직급별 데이터 접근 범위를 명확히 제한하고, 기업 내부 보안 정책과 자연스럽게 연동됩니다.

1 접근 권한 제어

소속과 직책에 따라 데이터 접근 범위를 정밀하게 설정

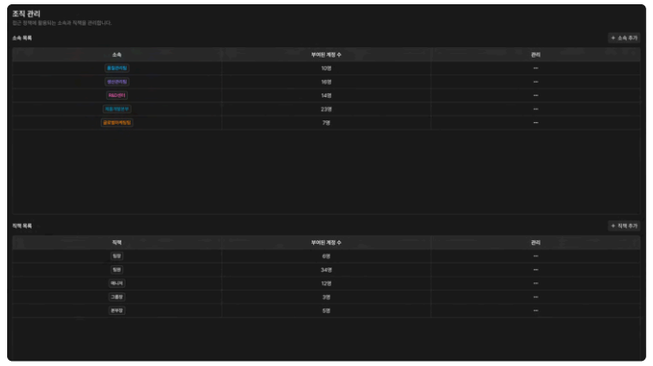
- 부서·직급별 벡터 테이블 접근 정책 개별 지정
- 사용자 역할(USER/ADMIN) 기반 관리 화면 접근 제한
- 계정 잠금과 권한 변경 이력 완전 관리



2 조직 체계 관리

조직 구조에 맞춰 사용자를 체계적으로 분류

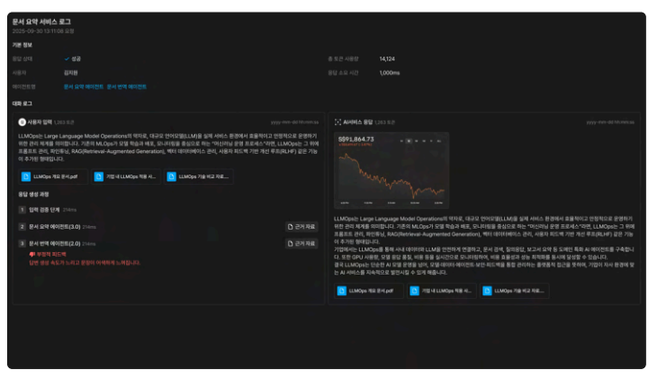
- 소속과 직책 기반 조직 계층 구조 설정
- 조직 단위 일괄 권한 부여로 관리 효율 향상
- 민감 정보 노출 원천 차단을 위한 데이터 격리



3 안전한 AI 응답 생성

조직 규정 준수를 위한 응답 범위 설정 및 이력 관리

- 민감 정보 요청·부적절한 질문에 대한 자동 차단
- 시스템 프롬프트 기반 응답 가능 범위 설정
- 차단된 응답 이력 추적 및 감사 로그 관리



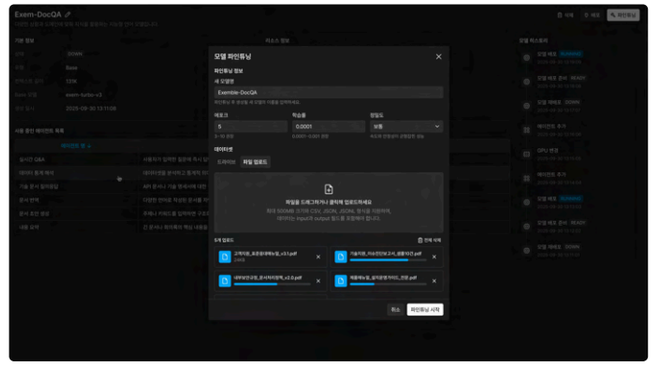
AI 품질 개선

도메인 특화 데이터로 모델을 파인튜닝하고, 사용자 피드백 기반 강화학습(RLHF)으로 AI 품질을 지속적으로 높입니다. 응답에 사용된 근거 자료를 투명하게 제공해 신뢰할 수 있는 AI 서비스를 구현합니다.

1 파인튜닝

도메인 특화 데이터로 모델 성능을 정밀하게 조정

- 학습 데이터셋 업로드 후 형식 오류 자동 검사
- 파인튜닝 진행 상황 실시간 모니터링
- 학습 완료 모델 자동 등록, 즉시 배포 준비 완료



2 피드백 기반 강화학습

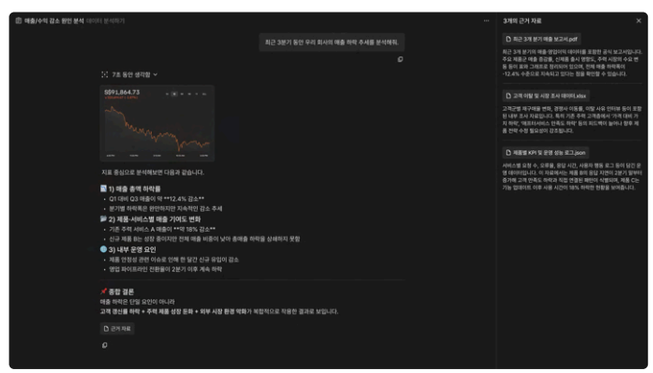
사용자 평가를 반영해 AI 품질을 지속적으로 향상

- 응답별 좋음/나쁨 피드백 수집 후 학습 데이터로 활용
- 피드백 반영 완료 시 개선된 모델 자동 생성
- 학습된 모델이 적용된 새 에이전트 버전 자동 추가

3 근거 자료 추적

AI 응답의 출처를 투명하게 제공

- 응답에 사용된 참조 문서와 데이터 출처 표시
- RAG 기반 응답의 신뢰성 검증 지원
- 환각(Hallucination) 최소화를 위한 근거 역추적



Architecture
구조도

1 데이터 소스

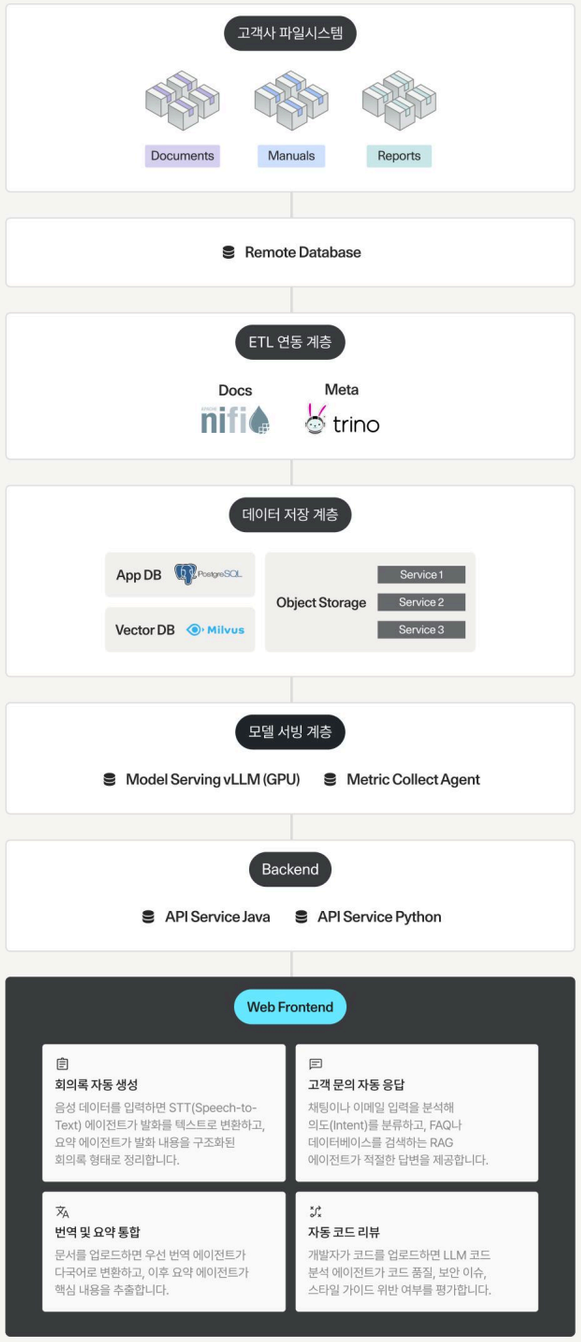
- 기업 내부에 존재하는 다양한 업무 데이터 활용 (매뉴얼, 보고서 등 문서 데이터와 업무 시스템, 운영 DB 등 데이터베이스)
- 고객 데이터는 기존 시스템에 그대로 유지되며, 원본 데이터에 대한 직접 변경 없이 활용

2 데이터 수집 및 처리

- 데이터 수집 파이프라인으로 다양한 소스 자동 연계
- 빅데이터를 위한 분산 쿼리 엔진으로 대용량 데이터 조회 및 분석 처리
- 데이터 변경 감지 시 자동 재처리 지원

3 AI 서비스 구현

- 고성능 LLM 서빙으로 GPU 자원 효율 극대화
- 사용자는 캔버스 기반 워크플로우로 업무 특화 AI 서비스 및 채팅을 손쉽게 제작
- 모든 요청에 대해 로그, 성능 지표, 사용 이력 자동 수집
- 보안 정책, 접근 권한, 감사 로그를 서비스 레벨에서 일관되게 적용



Platform Specs
지원 스펙 · 환경

웹 브라우저

지원 브라우저: Chrome, Edge
권장 해상도: 1920×1080 / 최소 해상도: 1440×900

eXemble 플랫폼 서버 요구사항

* 고객사 데이터 규모 및 보관주기에 따라 변동 가능

플랫폼 운영 서버

OS: Rocky Linux 8.9 이상 또는 Ubuntu 22.04 이상
Kernel: 4.18.0 이상 (Rocky Linux) / 5.15.0 이상 (Ubuntu)
CPU: 권장 16 Core / 최소 8 Core
Memory: 권장 32GB / 최소 16GB
Disk(DB): 권장 1TB / 최소 500GB

모델 서빙 서버 (1,000명 동시 사용 기준)

OS: Rocky Linux 8.9 이상
Kernel: 4.18.0 이상
CPU: 권장 24 Core / 최소 16 Core
Memory: 권장 256GB / 최소 128GB
GPU: 권장 H100 × 1 / 최소 A30 × 4
Disk: 권장 4TB / 최소 2TB

데이터 수집 서버 (자동 수집 기능 사용 시 필요)

OS: Rocky Linux 8.9 이상
CPU: 권장 8 Core / 최소 4 Core
Memory: 권장 128GB / 최소 64GB
Disk: 권장 1TB / 최소 100GB

분산 파일 시스템 (3-Node 클러스터 구성)

OS: Rocky Linux 8.9 이상
CPU: 권장 8 Core / 최소 4 Core (노드당)
Memory: 권장 128GB / 최소 64GB (노드당)
Disk: 권장 1TB × 3 (총 3TB) / 최소 100GB × 3 (총 300GB)

www.ex-em.com

우리는 데이터를 통해 철학하고 혁신합니다

서울시 강서구 마곡중앙8로5길 40
TEL 02-6203-6300 **FAX** 02-6203-6301

exem