



(주) 천명솔루션
CM SOLUTION CO., LTD.

보안과 생산성을 동시에!

CM-AI BOX

기업 맞춤형 구독형 Private Local LLM Rental Service





왜 Private Local LLM인가?

데이터 보안 리스크와 비용 부담을 줄이는 Private Local LLM 도입 전략



100%
폐쇄형 운영



구독형



통합 운영 지원

— 도입이 필요한 이유



01 데이터 반출 및 규제 리스크

외부 LLM 사용 시 기업 기밀
유출 우려와 컴플라이언스
대응 한계



02 비용 예측의 불확실성

토큰 기반 과금으로 예산
통제가 어렵고 활용 확대 시
비용 증가



03 구축·운영 부담

초기 인프라 비용과
모델 최적화, 전문 인력
투입 부담

핵심 제안

구독형 프라이빗 렌탈 소프트웨어 솔루션

합리적인 고정 비용으로 Private Local LLM을 도입하세요



데이터 외부 전송 없는 100% 폐쇄형 운영



H/W + LLM 엔진 + 운영 관리 통합 제공



설치, 업데이트, 장애 대응 기본 포함





CM-AI BOX

All-in-One 서비스 범위

도입부터 운영까지 한 번에 제공하는 엔터프라이즈 AI 서비스

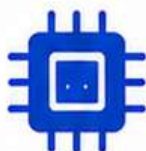


01 서비스 구성



하드웨어 공급

Local LLM 최적화
인프라 노드 제공



모델 서빙·운영

Ollama 엔진 기반
Gemma 최적화



소프트웨어 공급

RAG 지원
챗봇 SW 제공



보안 및 모니터링

접근 제어, 감사로그,
시스템 헬스 체크



기술 지원

업데이트, 백업,
장애 대응 프로세스

02 기본 제공 및 확장 옵션

기본 제공



- ✓ 프라이빗 인프라 1식 렌탈
- ✓ 자체 개발 RAG 챗봇 SW 제공
- ✓ Gemma 기반 자연어 모델 탑재
- ✓ 정기 업데이트 및 장애 대응

확장 옵션

- ✓ 데이터 저장공간 확장
- ✓ 고가용성(HA) 구성
- ✓ 전용 기술 지원(On-site)
- ✓ 고급형 업그레이드(AI Agent 사용)



03 도입 로드맵

STEP 01



사전 진단

인프라 환경 조사



STEP 02



PoC 수행

개념 검증 및
성능 테스트



STEP 03



운영 전환

데이터 연동 및
최적화



STEP 04



정식 운영

KPI 기반
성과 측정

고급형 AI-Agent 모델 특성

답변을 넘어 직접 실행하는 기업형 AI 도우미

01



답변을 넘어, 직접 실행하는 AI

반복형 실행 업무에서
검색·수집·작업 실행까지 수행

03



내 프로젝트를 이해하는 맞춤형 지원

선택한 프로젝트의 구조와
소스 코드를 분석해
정확한 해결책 제공

02



계획부터 검증까지 이어지는 자동화

AI가 작업을 계획하고
필요한 파일을 찾아
순차적으로 실행

04



로컬 실행 기반의 안전한 작업

작업 범위를 선택한
프로젝트 내부로 제한하고
승인 기반으로 제어



핵심 도입 효과



코드 작성 자동화

반복 코드 작성 시간을
줄이고 개발 속도 향상



디버깅 및 버그 수정 지원

오류 원인 분석과 수정
방법을 빠르게 제안



테스트 코드 생성

중요한 동작과 예외 상황을
확인하는 테스트 코드 자동 생성



문서/리팩토링 자동화

문서 요약 및 작성, 코드
리팩토링으로 품질 향상



보안형 사내 AI 지식검색 솔루션

Ollama · Gemma · RAG 기반의 프라이빗 LLM 서비스

사내 문서와 지식을 안전하게 연결해 정확한 답변을 제공하는
로컬 AI 플랫폼입니다. 사용자 인증부터 감사로그까지 전 과정을
관리하며, 외부 유출 없이 신뢰할 수 있는 업무형 AI 환경을 구현합니다.



주요 기능



강화된 보안

데이터 외부유출 차단,
로컬 환경 운영



정확한 답변

사내 지식 기반 검색과
근거 중심 응답



운영 효율

API 연동과 자동화로
빠른 업무 적용



유연한 확장

문서 업데이트와
모델 확장에 유리

상세 요청 처리 프로세스

01



사용자 인증 ·
권한 확인



02



프롬프트
정책 검사



03



RAG
검색



04



컨텍스트
조합

05



Gemma
추론



06



답변
생성



07



로그 및
감사기록 저장

핵심 기술 구성과 RAG 파이프라인

정확한 답변을 위한 강력한 기술 구성과 체계적인 데이터 파이프라인



01 핵심 기술 구성



Ollama
(Inference Engine)

- 로컬 모델 서빙 최적화
- REST API 기반 순쉬운 연동
- 동시 요청 처리 지원



Gemma
(Intelligence Model)

- 코드 · 기술 문서 이해에 강점
- 텍스트 · 이미지 · 문서까지 확장 가능
- 업무 환경에 맞는 모델 선택 가능

02 RAG 및 데이터 파이프라인

STEP 01



문서 수집

사내 문서, 가이드,
기술 문서 취합

STEP 02



데이터 정제

불필요한 태그 및
노이즈 제거

STEP 03



텍스트 청킹

의미 단위로
텍스트 분할

STEP 04



임베딩

고차원 벡터
데이터로 변환

STEP 05



벡터 DB 저장

유사도 검색을 위한
인덱싱 저장

03 핵심 운영 원칙



근거 문서
우선 답변



출처 (Source)
기반 응답 제공



모르는 질문은
'모름'으로 안내



지속적인 문서
업데이트 지원



데이터 외부유출 차단
내부망 기반으로 안전한
AI 환경 운영



실시간 가드레일 적용
정책 기반 필터링으로
위험 응답 최소화



사내 지식 기반 답변
사내 문서 중심의 답변으로
정확성과 신뢰성 확보



안정적인 성능과 엔터프라이즈 보안

SWE-Bench 해결률



69.6%

코드 문제 해결 성능

개발 생산성



30~50%

반복 업무 자동화

AI 도입 후



32~42

Man-Month

기존 60 MM 대비 절감

1 성능 지표(KPI)



평균응답시간 (Simple)

3초 이내



평균응답시간 (RAG)

10초 이내



동시 처리량 (Throughput)

최대 5명 동시 접속 유지



추론 정확도 (Contextual)

환각 현상 10% 미만 유지

2 확장 전략



수평 확장
(Scale-out)

사용자 증가 시
노드 추가



수직 확장
(Scale-up)

GPU/VRAM 증설로
성능 향상



동시 사용자 5명
초과 시 추가 노드 권장

RAG 비중 증가 시
Vector DB 분리 가능

3 엔터프라이즈 보안



접근 권한 관리
(RBAC)

역할 기반 접근 제어로
안전한 계정 운영



데이터 암호화
(TLS 1.3 / AES-256)

전 구간 암호화로
데이터 보호 강화



감사 로그 기록

사용자 질의, 일시,
대상 시스템 접속 기록



데이터 거버넌스 /
보존 · 삭제 정책

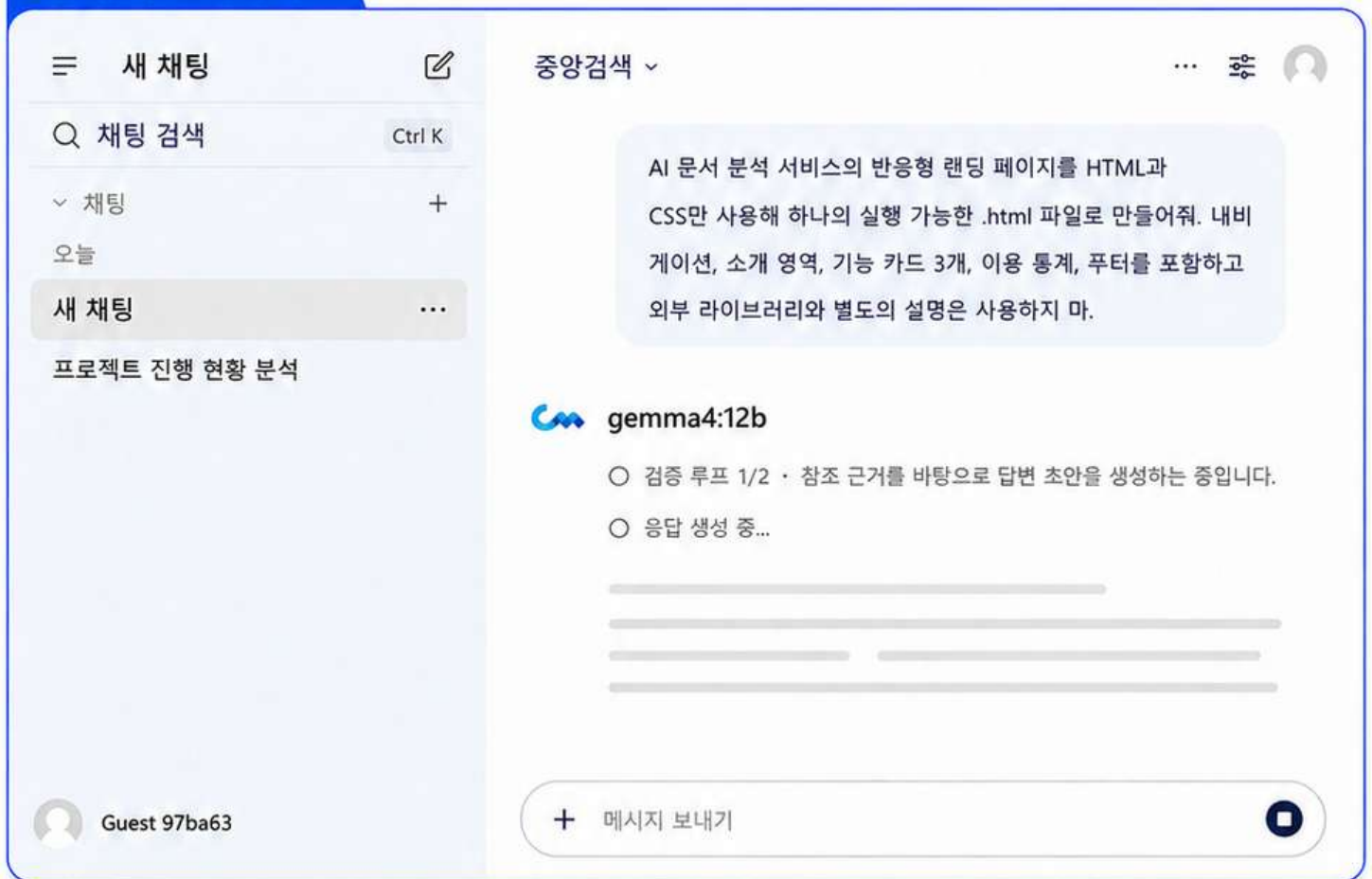
보존 기간 설정 및 만료 시
즉시 파기 정책



기업 보안 가이드라인에 맞춰 커스터마이징 가능



실제 데모 이미지



이런 기업에 적합합니다

 보안이 중요한 조직 민감한 데이터를 외부로 유출 없이 안전하게 관리하고 싶은 조직	 사내 지식검색이 필요한 조직 문서, 매뉴얼, 기술자료 등을 빠르고 정확하게 찾아 활용하고 싶은 조직	 개발 생산성을 높이고 싶은 팀 코드 생성, 테스트, 요약 자동화로 개발 효율을 극대화하고 싶은 팀	 온프레미스 AI 도입을 검토 중인 기업 클라우드 의존 없이 자체 환경에서 AI를 운영하고 싶은 기업
--	--	--	--

문의 및 데모체험

(주) 천명솔루션

영업 / 기술지원 / 제안 문의



02-6959-0666



sales@cmsolution.co.kr



www.cmsolution.kr



궁금한 점이 있으시면 언제든지
편하게 문의해주세요.