01 — BACKGROUND

프로젝트 배경

Why Now? 사내 AI 지식 서비스가 필요한 이유

현재 문제점

- 2025.08 사내 외용 생성형 AI 서비스 차단 조치
- H Chat Pro 응답 지연 및 처리 중단 발생
- Confluence 위키 수동 검색 — 평균 15분/건
- 지식 사일로 — 부서 간 정보 단절
- 신규 입사자 온보딩 지식 습득에 2-3주 소요

솔루션 전략

- Confluence REST API v2 — 고정밀 데이터 추출
- Azure OpenAI (GPT-4o) — 맞춤형 RAG 파이프라인
- Microsoft Teams Bot — 업무 흐름 통합 UX
- 사내 지식만 활용 — 데이터 주권 확보
- 즉시 답변 + 출처 명시 — 신뢰성 확보

15분	68%	2-3주	수동
현재 평균 검색 시간	정보 미검색 비율	온보딩 지식 습득	지식 업데이트 추적

시스템 아키텍처

Azure 클라우드 기반 계층형 마이크로서비스

Client Layer
Azure Bot Service · Microsoft Teams Bot · React Web UI

API Gateway — FastAPI
Python 3.12 + Pydantic v2 · 비동기 (Async/Await) · Azure App Service

RAG Engine — LangChain
문서 전처리 · 청킹 · 검색 체인 비동기 처리 · 프롬프트 템플릿 관리

Vector Store — Azure AI Search
벡터 + 메타데이터 필터링 · 하이브리드 검색 · 고성능 인덱싱

AI Model — Azure OpenAI
GPT-4o (답변 생성) · text-embedding-3-large (3,072D 임베딩)

Data Source — Confluence API v2
위키 본문 · 레이블 · 계층 구조 · Webhook + 30분 폴링 동기화

Infrastructure Stack

Layer	Technology
Client	Teams Bot + React Web UI
API Gateway	FastAPI (Azure App Service)
RAG Engine	LangChain (Python 3.12)
Vector DB	Azure AI Search
AI Model	Azure OpenAI (GPT-4o)
Embedding	text-embedding-3-large (3,072D)
Data Source	Confluence REST API v2
Auth	Azure AD (Entra ID) SSO
Monitoring	Application Insights + Power BI

FastAPI 백엔드 & Confluence 통합

API 엔드포인트 설계와 데이터 동기화 전략

서비스 API

메서드	엔드포인트	목적
POST	/api/chat	RAG 기반 지능형 질의응답
GET	/api/search	의미론적 검색 및 문서 탐색
POST	/api/index	스페이스별 인덱싱 및 임베딩
POST	/api/webhook/teams	Teams 봇 메시지 핸들링
GET	/api/health	시스템 상태 모니터링

Confluence API v2 엔드포인트

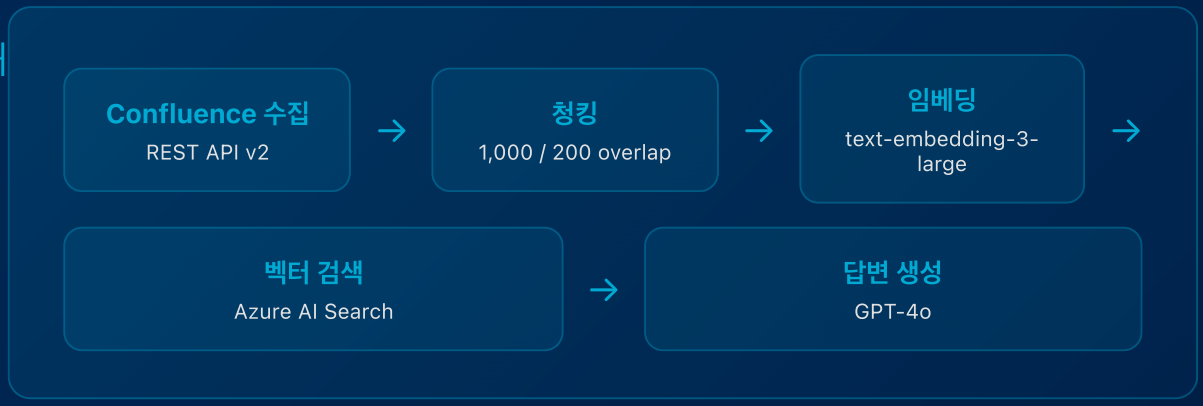
엔드포인트	용도
GET /wiki/api/v2/spaces	지식 지도 파악
GET /wiki/api/v2/pages	페이지 메타데이터
GET /pages/{id}?body-format=view	정제된 HTML 본문
GET /pages/{id}/children	지식 계층 정보
GET /pages/{id}/labels	카테고리 분류

동기화 전략

1. 초기 전체 크롤링 (Full Crawl)
2. Webhook + 30분 폴링 (Fallback)
3. 변경 청크 단위 부분 재임베딩

RAG 파이프라인 설계

Retrieval-Augmented Generation 상세 아키텍처



청킹 & 임베딩 파라미터

1,000
Chunk Size (tokens)

200
Overlap (tokens)

3,072
벡터 차원

0.1
Temperature

GPT-4o 프롬프트 제어

시스템 프롬프트

"귀하는 사내 지식 어시스턴트입니다. 반드시 제공된 Confluence 컨텍스트에만 기반하여 답변하십시오. 답변 시 출처 페이지 제목과 URL을 명시해야 합니다. 컨텍스트에 정보가 없다면 명확히 모른다고 답하십시오."

추론 제어

Temperature 0.1: 사실적 정확도 최우선
Max Tokens 2,000: 상세 설명 + 출처
Top-K 5: 상위 5개 청크 컨텍스트

Teams 통합 & 사용자 경험

업무 흐름을 벗어나지 않는 Seamless 지식 검색

자연어 질의응답

Bot Framework SDK 활용. Adaptive Cards로 정제된 답변과 출처 링크 제공. 마크다운 포맷 지원.

슬래시 명령어

`/filter [Space_Key]` — 특정 프로젝트/부서 범위 내 검색 제한. 스페이스 기반 스코핑.

멀티턴 대화

대화 문맥(Context) 유지. 연속 질문에 매끄러운 답변. 세션 기반 히스토리 관리.

피드백 루프

'도움됨/도움되지 않음' 버튼. 답변 품질 정량 평가 → 프롬프트 최적화 데이터 수집.

인터랙션 플로우

1. **사용자 질의**
   Teams 채팅에서 자연어 질문 입력
2. **의미론적 검색**
   Azure AI Search에서 관련 문서 청크 검색 (Top-K 5)
3. **GPT-4o 답변 생성**
   검색 결과 + 시스템 프롬프트 기반 답변 생성
4. **Adaptive Card 응답**
   답변 본문 + 출처 링크 + 피드백 버튼
5. **피드백 수집**
   사용자 만족도 → 품질 개선 데이터 축적

핵심 활용 시나리오

AI 지식 어시스턴트가 해결하는 업무 유형

사내 정책/규정 검색

Before: 15분 (수동 위키 검색) → **After:** 30초

97% 절감

HR 정책, 보안 규정, 업무 가이드를 자연어로 즉시 검색. 출처 페이지 URL 자동 제공.

기술 문서 질의응답

Before: 20분 (문서 탐색) → **After:** 1분

95% 절감

API 문서, 아키텍처 설명, 트러블슈팅 가이드를 대화형으로 검색.

프로젝트 지식 공유

Before: 동료 문의 대기 → **After:** 즉시 답변

자동화

프로젝트 히스토리, 의사결정 배경, 기술 선택 이유를 즉시 파악.

신규 입사자 온보딩

Before: 2-3주 → **After:** 3-5일 75% 단축

조직 구조, 업무 프로세스, 시스템 사용법을 AI가 멘토처럼 안내.

부서 간 지식 연결

Before: 사일로 → **After:** 통합 검색

사일로 해소

멀티 스페이스 검색으로 부서 간 정보 단절 해소. 크로스 팀 협업 촉진.

변경 사항 추적

Before: 수동 추적/누락 → **After:** 자동 알림

누락 Zero

중요 문서 변경 시 자동 감지 → 관련자에게 Teams 알림 → 영향 분석 제공.

운영 최적화 프레임워크

VEQG 4대 평가 차원과 Action Plan

V — Volume

BU/팀별 세션 수 · 쿼리 빈도 분석 · 도입 성숙도 측정

E — Efficiency

응답 성공률 vs 토큰 · API Latency · Dwell Time 분석

Q — Quality

사용자 만족도 점수 · 이상치 탐지 · 품질 저하 구간 식별

G — Growth

WAU/MAU 추이 · 서비스 확산 전략 · 사용 패턴 변화

단기 Action Plan (0-3개월)

조직 특화 프롬프트 템플릿 배포 · 지연 시간 개선 캐싱 최적화 · 주간 사용량 모니터링 체계 · 답변 품질 베이스라인 설정

중기 Action Plan (3-6개월)

부서별 토큰 예산 할당 · 사용량 기반 맞춤형 교육 · 경영진 보고용 KPI 대시보드 (Power BI) · 프롬프트 A/B 테스트

보안 및 거버넌스

엔터프라이즈 3중 보안 체계

Managed Identity

코드 내 자격 증명 노출 원천 차단. API Key 의존 탈피. Azure 네이티브 인증으로 Production Standard 보안.

인프라 보안

Azure AD (Entra ID) SSO. Role-based Access Control (RBAC). 네트워크 격리 (VNet). TLS 1.3 암호화.

콘텐츠 감사

Azure OpenAI 콘텐츠 필터링. 전체 질의 Audit Log 저장. 데이터 거버넌스 준수. DLP 정책 적용.

보안 레이어 강도

인증 (Entra ID)	완전
네트워크 격리	VNet
데이터 암호화	TLS
콘텐츠 필터링	자동
감사 로깅	완전

보안 원칙

Zero Trust — 모든 요청을 검증
최소 권한 — 필요한 접근만 허용
감사 추적성 — 모든 질의/답변 로그
데이터 주권 — 사내 데이터만 활용

ROI 분석

지식 검색 자동화의 비용 절감 효과

항목	Before	After	절감
정책/규정 검색	2.1억원	0.3억원	1.8억원
기술 문서 검색	1.8억원	0.2억원	1.6억원
온보딩 비용	1.2억원	0.3억원	0.9억원
부서 간 문의	1.5억원	0.3억원	1.2억원
지식 업데이트 관리	0.8억원	0.1억원	0.7억원
합계	**7.4억원**	**1.2억원**	**6.2억원**

70%
검색 시간 절감

3.8x
투자 대비 ROI

7개월
투자 회수 기간

6.2억
연간 순 절감

정성적 효과

직원 만족도 향상 (정보 접근성) · 온보딩 가속화 (75% 단축) · 부서 간 사일로 해소 · 지식 자산 가시화

비용 구조

연간 운영 비용과 Azure 서비스 비용 상세**연간 운영 비용**

항목	비용	비율
Azure OpenAI (GPT-4o)	0.7억원	44%
Azure AI Search	0.3억원	19%
Azure App Service	0.2억원	12%
운영 인력 (2명)	0.2억원	12%
모니터링/보안	0.1억원	7%
교육/온보딩	0.1억원	6%
합계	**1.6억원**	**100%**

투자 **1.6억** → 절감 **6.2억** = **ROI 3.8x**

Azure 서비스 월간 비용 (예상)

서비스	Tier	월 비용
Azure OpenAI	GPT-4o Standard	$4,500
Azure AI Search	Standard S1	$750
App Service	P2v3	$300
Azure Bot Service	Standard	$50
Application Insights	Standard	$100
합계		**$5,700**

비용 최적화 전략

프롬프트 캐싱: 반복 질의 패턴 캐시
토큰 예산 관리: 부서별 월간 상한
임베딩 재사용: 변경분만 재임베딩
GPT-4o-mini: 단순 질의 자동 라우팅

리스크 관리

예상 리스크와 완화 전략

리스크	확률	영향	완화 전략
AI 환각 (Hallucination)	높음	높음	RAG 기반 사실 검증 + Temperature 0.1 + 출처 명시 필수 + "모른다" 응답 허용
기밀 정보 노출	중간	높음	스페이스별 접근 제어 + RBAC + 기밀 등급 분류 + DLP 정책 + 콘텐츠 필터링
데이터 동기화 지연	중간	중간	Webhook 실시간 + 30분 폴링 Fallback + 동기화 상태 모니터링 대시보드
API 비용 초과	중간	중간	부서별 토큰 예산 + 프롬프트 캐싱 + GPT-4o-mini 자동 라우팅 + 월간 상한
사용자 저항	낮음	중간	Teams 기반 Seamless UX + 파일럿 성공 사례 공유 + 챔피언 사용자 육성

핵심 원칙: AI 답변에는 반드시 **출처(Confluence 페이지 URL)**를 명시. 컨텍스트에 없는 정보는 **"해당 정보를 찾을 수 없습니다"**로 응답. 모든 질의/답변은 **감사 로그**에 저장.

4단계 롤아웃 전략

MVP에서 전사 플랫폼까지 단계적 확산

P1 **MVP 구축 (1-3개월)** Core

1개 부서 (20명) · 핵심 RAG + Teams Bot
기술 문서 1개 스페이스 대상
목표: 일 50건 질의 처리, 만족도 >3.5

P2 **운영 최적화 (4-6개월)** Optimize

3개 부서 (60명) · KPI 대시보드 + 피드백 루프
프롬프트 A/B 테스트 · 토큰 예산 관리
목표: 일 200건 처리, 만족도 >4.0

P3 **전사 확산 (7-12개월)** Scale

전 부서 (200명+) · 멀티 스페이스 지원
부서별 커스텀 프롬프트 · 변경 알림 자동화
목표: 일 500건 처리, 검색 시간 70% 절감

P4 **AI Agent 확장 (12개월+)** Advanced

그룹사 확대 · AI Agent 기능 추가
문서 자동 생성 · 지식 그래프 · 다국어 지원
목표: 전사 AI 지식 플랫폼

20명	60명	200+명	그룹사
P1: MVP	P2: 최적화	P3: 전사	P4: 확장

90일 실행 로드맵

인프라 구축부터 MVP 운영까지

주차	목표	산출물	성공 기준
1-2	Azure 인프라 구축	OpenAI, AI Search, App Service 배포, Entra ID 연동	보안팀 검증 통과
3-4	RAG 파이프라인 v0.1	Confluence 크롤링 + 임베딩 + 벡터 검색 + GPT-4o 연동	답변 정확도 >80%
5-6	Teams Bot 개발	자연어 질의 + Adaptive Card 응답 + 피드백 버튼	E2E 플로우 동작
7-8	파일럿 시작	1개 부서 온보딩, 기술 문서 스페이스 대상	일 30건 처리
9-10	품질 최적화	피드백 반영, 프롬프트 튜닝, 청킹 전략 조정	만족도 >3.5
11-12	MVP 안정화	운영 매뉴얼, 모니터링 대시보드, 확장 계획	일 50건, 만족도 >4.0

4주
RAG 파이프라인 완성

8주
파일럿 시작

12주
MVP 안정화

HYUNDAI AUTOEVER

CONFLUENCE AI KNOWLEDGE SERVICE

지식 기반의 일하는 방식을 혁신합니다

Confluence의 지식을 실시간으로 활용하고,
Teams에서 벗어나지 않는 Seamless 경험을 제공합니다.

4주	8주	12주
RAG 파이프라인	파일럿 시작	MVP 안정화

보안과 거버넌스가 검증된 **엔터프라이즈 AI 지식 서비스**