정정 안내

「공공부문 생성형 AI 플랫폼 구축 전략 가이드」 정정 사항

2026년 3월 10일 배포한 본 자료에 대해, 전북특별자치도청의 요청을 반영하여 2026년 3월 11일 아래와 같이 정정했습니다. 본문과 다른 내용이 있으면 이 정정 안내가 우선합니다. 혼선을 드린 점 사과드립니다.

1	**인력 명칭**	"전북대학교 대학원 석사과정" → **"대학교 3~4학년 학생"**
2	**DeepSeek 모델 사용**	DeepSeek R1·V3.1을 사용한 바 있으나, 국정원 보안성 검토 과정에서 사용 금지 안내를 받아 **현재는 DeepSeek 계열 모델을 사용하지 않음**
3	**파라미터 수치**	"1.7조 개"는 근거 확인 불가, "6.8조 개"는 **6,850억 개(685B)**의 오기. 관련 수치 표현 전면 정정
4	**AMD GPU 운영·성능**	"AMD GPU 8장으로 실운영 불가"는 사실과 다름. vLLM 0.17(AMD 공식 지원)·AITER·FlashAttention v2 최적화로 **GPU 4장에서 397B VLM 모델을 200명 동시 서빙**(1인당 80 TPS), 비디오 모델 제외 시 400명까지 가능. 과거 속도 저하는 대형 모델이 보조 작업까지 처리했기 때문이며, 소형 작업 모델 분리로 해소
5	**"소형 모델로 타협" 표현**	소형 모델로 타협한 것이 아니라 **397B급 대형 모델을 안정적으로 운영 중**. 7~14B 모델은 행정 업무에 충분하지 않다고 판단
6	**인프라 증설 계획**	2027년 엔비디아 루빈 칩 8장 구매 예정
7	**데이터 처리 관련 입장**	전북특별자치도청은 **모델 자체 성능을 가장 중요하게 고려**함
8	**RAG 구성**	"일반적인 RAG"가 아니라 BGE-M3(1024차원, 한국어 특화) + 별도 리랭커 + 의미 기반 GraphRAG(2026년 1월 적용) 구조
9	**예산 명세서 인식 오류 원인**	"검색 출처 제한으로 예산 명세서 소실"이 아니라 **Top-K=3 제한**이 원인. 현재 Top-K=100, 상위 50위까지 모델 입력으로 수정 완료
10	**OCR·데이터 고도화**	GLM-OCR 적용 중. 향후 문서 분할 후 고성능 OCR 별도 적용 예정
11	**사용량 관리**	개인별 사용량 제한 기능 구현 완료. 과금은 시·군 협의 후 적용 예정
12	**민감정보 처리**	하드웨어 기반 개인정보 탐지 솔루션 보유, 도민정보시스템실 서버 전체 적용
13	**동시 사용자 처리**	397B 모델로 400명까지 처리 가능. 7~14B 모델은 행정 환경에서 성능 부족
14	**과금 체계 표현**	"쓴 만큼만 정확히 내는 과금 체계"가 아니라 **개인별·모델별 사용 제한 기능 구현 완료** 상태
15	**최신 오픈소스 모델 예시**	예시로 든 Llama 4.x는 비교적 이전 모델임. 현재 Qwen3.5를 AMD AITER 기반 최적화 경로(Flash Attention·FP8 MoE·Expert Parallel·Prefix Caching 등)로 운영 중

앞으로 기관 검증 절차 강화, 기술 수치 사전 교차 확인, 기관 입장과 기업 견해의 명확한 구분을 지키겠습니다.
(주)스캐터엑스

광역·기초·공기업·공공기관 맞춤형

생성형 AI 플랫폼 고도화 전략 가이드

전라북도청 실증 사례를 통해 본
다차원 RAG 및 사용량 기반 정밀 빌링 도입 방안

전라북도청 생성형 AI 플랫폼 구축의 우수성

광역 최고 수준 인프라와 산학 협력 기반 예산 절감 모델

① 광역 최고 수준 AI 인프라 구축
Top-tier AI Infrastructure

- 내부 행정망 기반 멀티 LLM 플랫폼
- 최신 오픈소스 모델 신속 도입
  Qwen 3.5 397B
  DeepSeek V3
- AMD MI300x GPU 8장 확보
- 고성능 워크스테이션 운영
- 광역자치단체 최고 수준 인프라



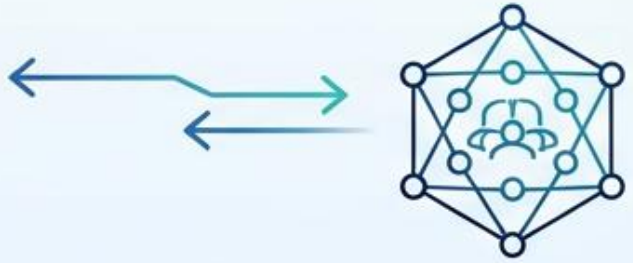
② 실무 중심 사용자 기능 구현
User-Centric AI Services

- 외부 검색 API 연동(퍼플렉시티 등)
- 이미지 생성 기능
- 문서 요약 기능
- 업무지원 특화 인터페이스
- 사용자 친화적 설계



③ 산학 협력 기반 예산 절감 모델
Academia-Industry Collaboration Model

- 전북대학교 대학원 연구진 협력
- 외주 개발 비용 최소화
- 지속적 연구 및 기술 고도화
- 자생적 개발 생태계 구축
- 비용 효율적 공공 AI 모델



예산 제약을 극복한 자생적 공공 AI 혁신 모델

Sustainable & Cost-Effective Public AI Innovation

전라북도청 생성형 AI 플랫폼 구축의 안타까운 한계

인프라 제약과 데이터 구조화 부재의 구조적 문제

인프라의 한계 (Infrastructure Limitation)

- 동시 접속 40명 수준에서 서비스 중단
- GPU 과부하 발생



- AMD GPU 8장으로 대형 모델 운영 불가
- 1.7조·6.8조 파라미터
  대형 추론 모델 실운영 불가
- 고성능 모델 제거 → 소형 모델 위주 타협 운영



- GPU 가격 급등 2.7억 원 → 6.7억 원
  - 예산 기반 인프라 증설 어려움



물리적 인프라 한계로 인한 성능 타협 불가피

데이터의 벽 (Data & RAG Limitation)



- 기존 RAG 구조 키워드 기반 검색 시맨틱(코사인 유사도) 중심
- 검색 출처 최대 18개 제한 (18개)
- 맥락 단절(Context Loss) 발생



- 방대한 매뉴얼·예산 명세서 전체 비교 불가
- 법령·조례 입체적 연계 분석 한계
- 단순 질의응답 수준에 머무름



- 행정 온톨로지(구조화) 부재
- 국-과-팀 관계도 메타데이터 미구축
- 복잡한 표 구조 이해 한계

모델 문제가 아닌 데이터 구조화 한계

"성공적인 AI 행정을 위해서는 인프라 확장과 데이터 구조화 고도화가 필수"

Infrastructure Scalability & Data Ontology are Key to Public AI Advancement

‘통합 AI 플랫폼 시대를 대비한 핵심 고도화 전략’

인프라 보호를 위한 정밀 빌링과 데이터 구조화를 위한 다차원 RAG

문제 인식 영역

- 전북도청 + 14개 시·군 통합 AI 플랫폼
- 동시 접속 증가 → 트래픽 폭증 위험
- 현 구조 유지 시 인프라 마비 가능성
- 무임승차 및 과도한 사용 통제 필요



중앙 해결 구조 – 2대 핵심 전략

① 사용량 기반 정밀 빌링 시스템

- 기관별 사용량 실시간 측정
- 사용한 만큼 비용 분담
- 특정 기관 과도 사용 차단
- 인프라 안정성 확보
- 공정한 비용 배분 구조



인프라 보호와 공정한 비용 구조의 핵심 장치

② 3072차원 메타데이터 RAG 고도화

- 단순 유사도 검색 한계 극복
- 행정 온톨로지 자동 생성
- 국-과-팀 조직도 반영
- 업무 분장 기반 관계도 자동 구성
- 복합 추론 가능
- 깊이 있는 행정 분석 지원



정확하고 깊이 있는 복합 행정 추론의 기반 기술

“통합 AI 시대, 인프라 보호와 데이터 구조화 없이는 지속 가능하지 않다”

Scalable Infrastructure & Advanced RAG are Essential for Sustainable Public AI

‘공공 행정 데이터 구조화 기반 다차원 시큐어 RAG 고도화 전략’

표 구조 파싱 · 3072차원 메타데이터 · 멀티모달 통합 검색

정형 문서 고도화

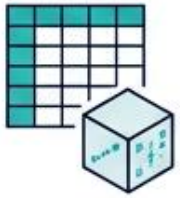
문제 인식 (기존 방식 한계)



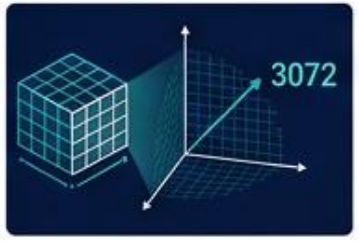
- 공공 문서 70% 이상 HWP 형식
- 복잡한 표 구조 + 문단 계층(제1장·제1절·제1조)
- 기존 RAG의 무작위 청킹(Chunking)
- 헤더(항목)와 데이터(값) 관계 단절
- 문맥 손실(Context Loss)
- 예산 분석 오류 발생



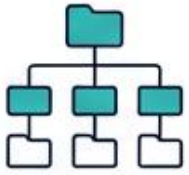
해결 구조 – 다차원 시큐어 RAG



- HWP 표 구조 원형 파싱(Parsing)
- 병합 셀·계층 구조 인식
- 행·열 관계 정확한 전달
- 문서당 최대 3072차원 메타데이터 자동 추출
  (제목, 작성일, 작성 부서, 보안 등급, 관련 법령)
- 자동 인덱싱(Indexing)



조직도 기반 온톨로지 자동화



- 국-과-팀 조직도 연동
- 업무 분장표 자동 매핑
- 부서 간 관계도 자동 생성(Ontology Automation)
- 복합 질문 대응
- 3개년 예산 비교 분석 가능
- 단절 없는 입체적 탐색



비정형 데이터 통합 검색

핵심 기술 구성



- OCR(광학 문자 인식) 통합
- 객체 탐지(Object Detection)
- 이미지·도면 텍스트 추출
- 구조·치수·객체 지몽 인식

“텍스트 + 표 + 조직도 + 이미지 통합 지능형 행정 AI”

Multimodal Secure RAG for Advanced Public Administration



법령 교차 검증 구조



AI 분석

- 필로티 구조 불법 증축 탐지
- 불법 연결 통로 인식
- 건축법·주차장법·소방법 교차 분석
- 위반 소지 자동 가이드 제시
- 텍스트 + 이미지 통합 추론

지속 가능한 공공 AI를 위한 보안 강화 및 RAG 진화 전략

N2SF 대응 AI DLP 분기별 업데이트 · 연 최대 4회 RAG 교체 체계

국정원 N2SF 대응 AI DLP 전략
(N2SF-Compliant AI DLP Strategy)

정책 배경
- 국가 망 보안체계(N2SF) 가이드라인
- 데이터 중요도 기반 통제
  ('기밀', '민감', '공개', '공개')
- 제로트러스트 보안 아키텍처

AI DLP 핵심 기능
- 프롬프트 입력 즉시 분석
- 0.1초 이내 탐지
- CPU 기반 경량 LLM 적용
- 주민등록번호·계좌번호 탐지
- 대외비 키워드 실시간 차단
- 문맥 기반 민감정보 식별



신종 위협 대응
- 프롬프트 인젝션 공격
- 스테가노그래피 공격
- 지속적 공격 기법 고도화
- DLP 룰셋 분기별 무상 업데이트
- 국정원 가이드라인 변화 자동 반영

별도 추가 예산 없이
국가 최고 수준 보안 유지

AI

**"보안은 분기별로 강화하고,
기술은 지속적으로 진화한다"**

Secure by Design, Evolving by Quarter

RAG 기술 진화 대응 전략
(RAG Evolution & Upgrade Strategy)

기술 환경 변화
- 3~6개월 단위 기술 진화
- Contextual RAG
- Agentic RAG
- GraphRAG
- 검색 정확도 지속 향상

기존 방식의 한계
- 영구 라이선스 구매 구조
- 1년 내 구형화 위험
- 매몰 비용 발생
- 예산 비효율

제안 모델 – 연간 구독 기반 고도화
- 연간 라이선스 기반
- 분기별 성능 점검
- 연 최대 4회 무상 기술 교체(Swap)
- 최신 RAG 지속 적용
- 기관은 본연의 업무에 집중



시스템의 진화는 공급사가 책임

하이브리드 LLM 아키텍처 및 지속 진화형 교체 전략

내부 소형 LLM 보안 처리 · 외부 대형 LLM 복합 추론 · 객관적 성능 관리 체계

① 하이브리드 LLM 역할 분담 구조 (Hybrid LLM Architecture)

내부 소형 LLM (sLLM)

- 내부 행정망 전용 운영
- 7B~14B 오픈소스 모델 활용
- 대외비 문서 요약
- 내부 결재 기안 작성
- 개인정보 포함 민원 분석
- 수백 명 동시 접속 처리
- 제한된 GPU 자원 효율 활용
- 내부 데이터 외부 유출 차단

보안 중심 업무는 내부에서 안전하게 처리



외부 대형 LLM API

- Gemini 3.1
- GPT-5.3
- Claude Sonnet 4.6
- 복합 추론 전담
- 방대한 예산 비교 분석
- 타 지자체 법령 교차 검증
- 정책 시뮬레이션
- 무한 확장 컴퓨팅 파워

중심: 복합 추론은 외부의 계산 능력을 활용

② Artificial Analysis 인덱스 기반 성능 관리 (Objective Performance Tracking)



- Artificial Analysis 인덱스 도입
- 글로벌 LLM 순위 실시간 공개
- 추론·수학·코딩·지시 이행률 종합 평가
- 객관적 점수 기반 관리
- 감사 대응 수치 입증 가능
- 주관적 평가 배제



중성: 객관적 지표로 시스템 경쟁력 검증

③ 연 4회 최신 LLM 교체(Swap) 전략 (Quarterly LLM Upgrade Strategy)



- 분기별 정기 점검
- 글로벌 상위 3개 모델 평균 대비
- 10점 이내 격차 유지
- 성능 하락 시 즉각 교체
- Llama 4.X, Qwen 3.X 등 신규 모델 반영
- 연 최대 4회 무상 교체



중성: AI는 시간이 갈수록 더 강력해지는 구조

"보안은 내부에서, 통찰은 외부에서, 성능은 객관적으로 관리한다"

Secure Inside, Powerful Outside, Measured by Global Standard

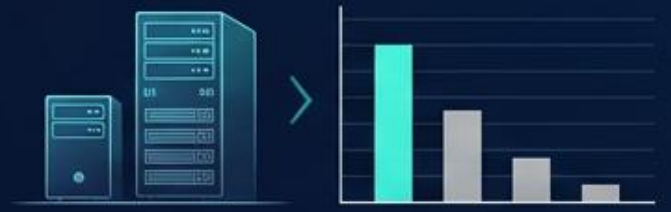
인프라 공유를 위한 연합 AI 플랫폼 구조

광역-기초 통합 모델 · 본사-지사 트래픽 분산 아키텍처

문제 인식 영역
(배경 설명)

기관 규모별 디지털 격차 확대

- 대형 기관 – 수백억 원 예산
- 소규모 지자체 – IT 예산 평균 5억 원 미만
- 인구 10만 명 미만 기초지자체 89개



디지털 격차 해소를 위한
공동 인프라 필요

광역-기초 데이터 통합 아키텍처

- 광역자치단체 = 고성능 GPU 인프라 구축
- 다차원 RAG 플랫폼 중앙 운영
- 산하 시·군 전용 보안망 접속

다타 슈너키

- 초기 구축 비용 부담 없음
- 사용량 기반 합리적 과금
- 동일한 최고 성능 AI 서비스 공유



- 광역 + 기초 행정 데이터 통합
- 조례·민원 사례 유기적 연결
- 정책 시뮬레이션 가능
- 우수 정책 타 지자체 적용 분석



하나의 플랫폼, 모두의 행정 역량 상향 평준화

공기업·공공기관 본사-지사 연동 구조

본사:

- 대형 복합 추론 LLM
- 전사 통합 데이터베이스
- 심층 분석 전담



지사:

- 경량화 소형 LLM(sLLM) 노드
- 민원 처리 특화
- 일상 문서 생성 전담
- 지역 단위 업무 처리

- 트래픽덩이 자동한 모로 자등직으로 분산한해 처리
- 발상한 트래픽도 알로없 이뭉 처리



대규모 트래픽도 지능적으로 분산 처리

“공동 인프라 구축으로 비용은 절감하고, 행정 역량은 공동으로 강화한다”

Shared Infrastructure, Equalized Digital Capability

공정한 비용 분담을 위한 사용량 기반 정밀 빌링 시스템

100% 종량제 · 실시간 3대 사용량 측정 · 무임승차 방지 구조

기존 정액제 방식의 한계
(Flat-Rate Limitation)

- 인구 비례 과금
- 직원 수 비례 과금
- 행정 처리 편의성 중심 구조
- 실제 사용량 미반영
- 적게 사용하는 기관의 역차별
- 무임승차(Free-riding) 발생
- 과도 사용 기관의 GPU 독점
- 전체 시스템 속도 저하 위험



정액제는 공정하지 않다

100% 종량제 기반 정밀 빌링
(Pay-as-you-go Precision Billing)

- 오직 사용한 만큼만 과금
- 100% 종량제(Pay-as-you-go)
- 기관별 독립 컨테이너 운영
- 사용량 완전 분리 추적
- 정부 가이드라인 2.0
- 장부가드라인 2.0 권장 체계 부합
- 통신망 수준 정밀 과금 엔진 적용

ScatterX × 엔텔스(NTELS)

통신 빌링 엔진 기술 적용
초 단위 정밀 계산



쓴 만큼 정확히 내는 구조

3대 실시간 측정 지표 (Three Real-time Billing Metrics)

① API 호출 횟수
- LLM 질의 횟수
- RAG 외부 검색 횟수

② 데이터 저장 용량
- 벡터DB 사용량
- 문서 저장 용량 (GB 단위)

③ GPU 점유율
- LLM 추론 시간
- 이미지 분석 시간
- 초 단위 컴퓨팅 시간 측정

00h 43s	00h 48s
00h 23s	01h 54s



고부하 작업 = 상응 비용 부담

"완벽한 공정성 없이는 연합 플랫폼도 지속될 수 없다"

Fair Usage, Sustainable Platform

‘투명한 실시간 정산과 AI 예산 예측 기반 인프라 보호 체계’

실시간 대시보드 · 분기별 자동 정산 · 예산 시뮬레이터 · 트래픽 사전 통제

실시간 대시보드 및 자동 정산 시스템

- 기관별 실시간 사용량 공개
- 일별·월별 사용량 추이 그래프
- 누적 과금 금액 즉시 확인
- 직관적 UI 제공

투명성은 신뢰의 시작



이번 달 사용 금액: ○○원

- 분기별 자동 정산
- 자동 청구서 발송



분기별 자동 정산 흐름도

- 초 단위 상세 사용 내역(Log) 제공
- 이의 제기 시 투명한 데이터 검증



로그 데이터 타임라인

AI 예산 예측 시뮬레이터

- 과거 1년 사용 패턴 분석
- 부서별 이용 추세 반영
- 계절별 민원 증가 요인 반영
- 내년도 예상 청구액 자동 산출
- 예산 편성 지원
- 안정적 재정 계획 수립

과거 데이터 → AI 분석 → 미래 예측 그레프

“내년도 예상 청구액: 000만 원”



예산은 감이 아니라 데이터로 예측

트래픽 사전 통제 및 인프라 보호

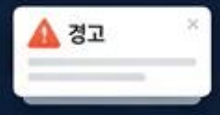
- 예산 한도 80% 도달 시 자동 경고
- GPU 허용치 초과 사전 감지
- 관리자 자동 알림 발송
- 트래픽 급증 사전 차단
- 인프라 마비 예방
- 플랫폼 운영 기관 보호



경고 게이지



서버를 보호하는 방패 이미지



경고 알림 팝업

마비 전에 경고, 초과 전에 통제

투명한 정산과 사전 통제로 365일 안정적 운영

Transparent Billing, Predictive Budgeting, Proactive Protection

소규모 기초자치단체 맞춤형 AI 도입 전략

SaaS 단독 도입 또는 광역 연합 플랫폼 합류



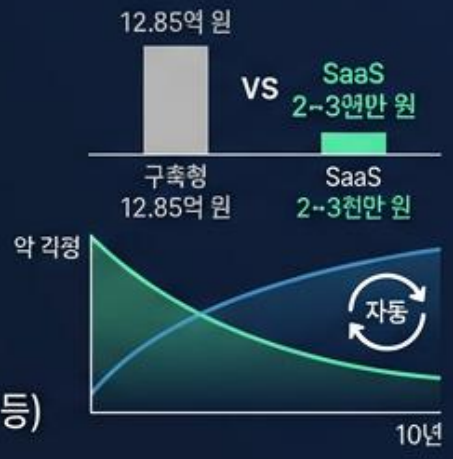
- 인구 10만 명 미만 89개 지자체
- 연간 IT 예산 제한적
- 자체 구축 수십억 원 부담
- 대민 밀착형 업무 다수
- 대민 밀착형 업무 다수
- AI 도입 효과 체감도 높음

적은 예산, 큰 효과



SaaS 단독 도입 방식
(SaaS Subscription Model)

- 초기 구축비 12.85억 원 부담 해소
- 연간 2,000만 ~ 3,000만 원 구독료
- 10년 TCO 약 2~3억 원 수준
- 즉시 도입 가능
- 연 4회 LLM 점검 및 교체
- 분기별 AI DLP 보안 업데이트
- 전담 인력 불필요
- 국내 인증 클라우드 기반(NHN Cloud 등)
- 국정원 보안 가이드라인 충족



가장 빠르고 부담 없는 시작



광역-기초 연합 플랫폼 합류
(Regional Alliance Platform Model)

- 광역 고성능 플랫폼 공동 활용
- 인구 비례 정액제 폐지
- 100% 사용량 기반 과금
- 보안 컨테이너 완전 격리
- 데이터 용량·추론 시간 기준 1원 단위 과금
- 예산 낭비 없음
- SaaS 백업 병행 가능
- 경제적·안전한 하이브리드 환경







최고 성능을 공정한 비용으로 공유

SaaS 단독 도입 → 신속·간편·저비용 **광역 연합 합류 → 고성능·공동 인프라·완전 종량제**

Start Small or Join the Alliance – Both Are Sustainable

중대형 기초자치단체 3개년 자체 구축형 하이브리드 모델

단계적 투자 · 자체 역량 강화 · 유휴 자원 공유 선순환 구조

중조시자청

확보예산

- 인구 10만 명 이상
- 1차 사업 예산 5억 원 이상 확보 가능
- 3개년 단계적 구축 권장
- 층 투자 규모 20억 ~ 25억 원

⇒ 단계적 투자로 지속 가능한 자체 플랫폼 완성

3개년 로드맵 타임라인 (Three-Year AI Roadmap)

1차년도 – 기반 구축 단계
(Foundation Phase)

- 예산 약 5억 원
- 데이터 플랫폼 구축
- 시큐어 RAG 도입
- 멀티 LLM 연간 라이선스
- 초기 하드웨어 장비 구축
- 안정적 운영 기반 확보

인프라 기반 확립

2차년도 – AI 에이전트 확산 단계
(Expansion Phase)

AI

- 예산 약 8~15억 원
- 부서별 전용 AI 에이전트 10개 이상 개발
- 민원·복지·예산·건축 등 부서 특동화
- 실무 업무 효율 극대화

업무 자동화 확산

3차년도 – 생성형 BI 고도화 단계
(Advanced Analytics Phase)

AI

AI

- 예산 약 8~15억 원
- 자연어 기반 데이터 분석
- 정책 시뮬레이션 지원
- 생성형 BI 고도화
- 데이터 기반 의사결정 체계 완성

정책 의사결정 지능화

유휴 자원 공유 선순환 모델

- 인구 30만 이상 대규모 지자체 적용 가능
- 유휴 GPU 자원 공유
- 인근 소규모 지자체 연계

GPU

GPU

닝낭칙규모 기초지자체

인근소규모 기초지자체

사용량

- 사용량 기반 정밀 빌링 적용
- 사용한 만큼 비용 회수
- 유지관리비 일부 충당
- 선순환 재정 구조

자체 구축 → 공동 활용 → 비용 회수 → 지속 가능

“3년 후, 단순 도입을 넘어 지역 AI 허브로 도약”

Build, Expand, Lead – From Adoption to Regional AI Hub

공기업·공공기관 본사-지사 연동형 AI 플랫폼 및 정밀 과금 체계

망 분리 독립형 구축 · 부서별 GPU 점유 기반 내부 과금



- 수많은 지사와 대규모 인력
- 동시 접속 트래픽 폭증 가능성
- 전사적 통합 관리 필요
- 3개년 총 투자 250억 원 규모
- 3가원 운영 효율 극대화

대규모 조직일수록 정밀한 자원 관리가 필수

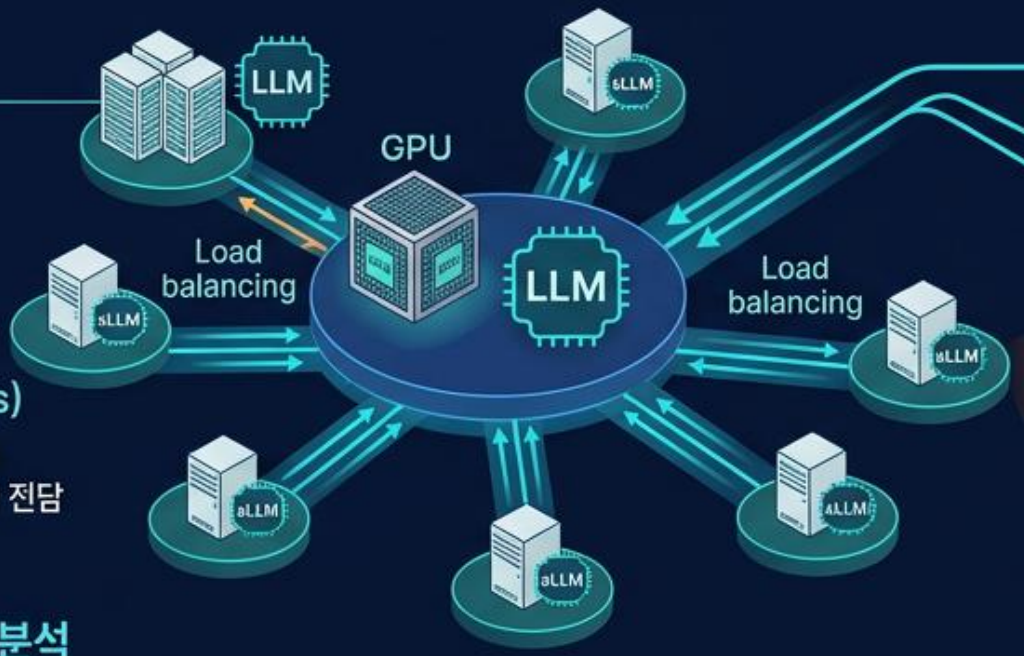
본사(Main AI Core)

- 대형 복합 추론 LLM 구축
- 전사 통합 데이터베이스
- 고난도 분석 전담
- 중앙 통제 및 정책 관리

각 지사(Branch Nodes)

- 경량화 소형 LLM 분산 배치
- 지역 민원 및 일상 문서 생성 전담
- 트래픽 1차 분산 처리
- 본사 과부하 최소화

분산 처리, 중앙 집중 분석



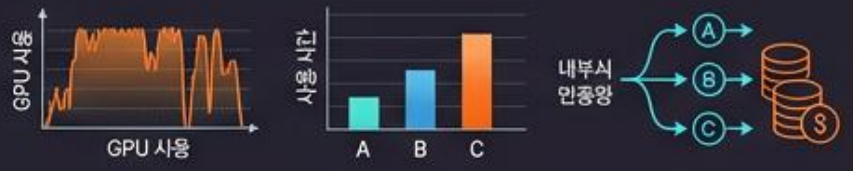
Headquarters-Branch Distributed Architecture

3개년 토조드맵 — 3년년 투자

분산 처리, 중앙 집중 분석

부서별 정밀 내부 과금 체계 (Internal Charge-back System)

- 엔텔스(NTELS) 빌링 엔진 적용
- GPU 점유 시간 초 단위 측정
- 지사별·부서별 사용량 분리
- 정액제 폐지
- 실제 점유 시간 기반 내부 과금(Charge-back)
- AI 남용 방지
- IT 자원 운영 효율 극대화



쓴 만큼 내부 정산, 자원은 공정하게 배분

- 대규모 트래픽 안정적 분산
- 전사 자원 통합 관리 체계 확립
- 불필요한 AI 사용 억제
- 망분리 독립형 보안 강화
- 망 분리 독립형 보안 강화
- 지속 가능한 엔터프라이즈 AI 운영

Enterprise AI with Distributed Control & Precision Charge-back

10년 TCO 기반 예산 편성 및 지속 가능 운영 전략

영구 라이선스 vs 연간 라이선스 · 장기 예산 안정성 · 국비 확보 전략

영구 라이선스 vs 연간 라이선스 비교
(Perpetual vs Annual License Model)

영구 라이선스
(Perpetual License)



- 초기 대규모 구축비 부담
- 1년 내 기술 노후화 위험
- 3~6개월 단위 기술 혁신 미반영
- 매몰 비용 발생
- 업그레이드 추가 예산 필요



연간 라이선스
(Annual Subscription Model)



- 초기 구축비 부담 완화
- 10년 TCO 획기적 절감
- 연 최대 4회 최신 LLM 무상 교체(Swap)
- 지속 업그레이드 구조
- 비용은 안정, 성능은 지속 향상



비용은 낮추고, 기술은 계속 진화

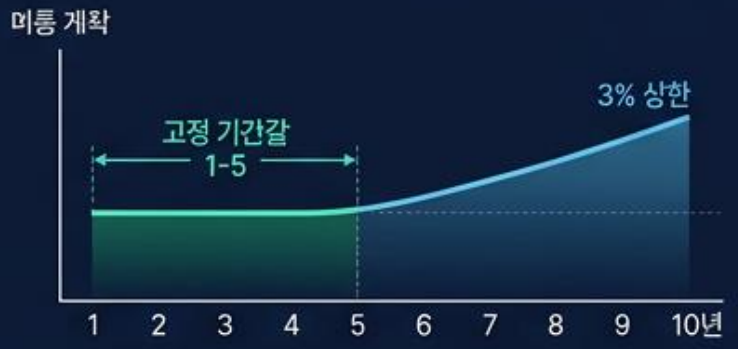
장기 예산 예측 및 물가 연동 안정화 정책
(Long-term Budget Predictabiiity)

AI 예산 예측 시뮬레이터

- 과거 사용 패턴 분석
- 부서별 트래픽 증가율 반영
- 차년도 예상 청구액 자동 산출
- 예산 편성 근거 데이터 제공

물가 연동 정책

- 1~5년 차 라이선스 단가 동결
- 6년 차 이후 CPI 연동
- 연 최대 3% 상한
- 6개월 전 사전 통보 의무화
- 중기 재정계획 안정성 확보



예측 가능한 예산, 안정적인 운영

중앙정부 재원 확보 전략
(Government Funding Strategy)

- 행정안진부 공공 AI 지원 사업
- 과기부 AI 바우처 사업
- 특별교부세 활용 진략
- 초기 구촉비 50% 국비 확보 가능
- 다년도 계속 사업 등록
- 3개년 이상 안정적 예산 확보
- 행정 잘차 간소화



국비 활용으로 초기 부담 최소화

"10년을 내다보는 예산 전략이 AI 혁신의 지속성을 완성한다"

Predictable Budget, Sustainable AI Transformation

‘기술 파트너 선정 기준 및 필수 계약 가이드’

과업지시서 명문화 · 성능 보증 · 보안 준수 · 예산 안정성 · 데이터 주권



- 모든 구속력은 과업지시서에서 시작
- 특수조건 및 SLA 조항 명문화 필수
- 기술 약속 → 법적 강제 조항 필요
- 도입 후 방치 방지

명시되지 않으면 보장되지 않는다

① 연합 솔루션 도입 요건



보안 · 성능 · 공정성 3대 측

스케터엑스(ScatterX)
- 시큐어 RAG
- 멀티 LLM
- 국정원 N2SF 완전 준수
- Contextual RAG· Agentic RAG 즉시 적응
- 연간 라이선스 기반 업그레이드
- 노코드(No-code) 에이전트 빌더

엔텔스(NTELS)
- 100% 종량제 정밀 빌링
- 통신망 수준 과금 엔진
- API 호출 횟수
- 벡터DB 저장 용량
- GPU 점유율(초 단위)
- 인구 비례 정액제 배제

보안 + 성능 + 공정성 통합 요구

② 객관적 성능 보증 및 연 4회 교체 의무



- 아티피셜 어넬리시스(Artificial Analysis) 기준 준용
- 글로벌 상위 3개 모델 평균 대비
- 10점(또는 20%) 이내 격차 유지 보증
- 법적 강제 조항 명시
- 성능 미달 시 연 최대 4회 무상 교체(Swap)


잠수 비교 그래프


분기별 교체 순환 구조

객관적 지표 기반 법적 성능 보증

③ N2SF 준수 및 AI DLP 분기별 업데이트



- 국정원 N2SF 가이드라인 준수
- DLP 룰셋 분기별 무상 업데이트
- 프롬프트 인젝션 차단
- 스테가노그레피 공격 대응
- 최신 위협 방어 로직 적용

보안은 정기적 의무 업데이트로 유지

④ 물가 연동 조정 및 사전 통보 조항



- 1~5년 가격 동결
- 6년 차 이후 CPI 연동
- 연 최대 3% 상한
- 6개월 전 사전 통보 의무화
- 업그레이드 무상 제공

예산 예측 가능성의 법적 보장

⑤ 데이터 주권 및 컨소시엄 검증 기준



- CSAP 인증 국내 클라우드만 사용
- 해외 서버 전송 금지
- 계약 종료 후 30일 내 완전 삭제
- 부채비율 200% 이하
- 영업이익률 10% 이상
- 대규모 빌링 구축 경험
- 다차원 RAG·멀티 LLM 운영 역량 입증

데이터 주권과 재무 건전성 동시 확보

“기술은 계약으로 완성된다”

Governed by Contract, Sustained by Performance

결론 및 향후 전망

선도적 도전에서 대한민국 공공 AI 표준 생태계로

문제 인식



- 2026년 디지털 전환의 분기점
- 생성형 AI = 행정 패러다임 전환
- 예산 제약
- 엄격한 보안 규제
- 기술적 불확실성

위기는 곧 전환의 기회

전라북도청의 선도적 도전



- 척박한 환경 속 선도적 도입
- 최신 오픈소스 모델 적용
- 산학 협력 기반 자생적 혁신
- 인프라 부하 경험
- 인프라 부하 경험
- 데이터 구조화 한계 체험
- 대한민국 공공 AI의 성장통 선행 경험



실패가 아닌 선구적 성장의 여정

비전

대한민국 공공 AI 연합 생태계

대한민국 공공 AI 연합 생태계



- 광역-기초 연합 플랫폼
- 데이터 행정 허브
- 소규모 지자체 포함
- 대형 공기업까지 확장
- 단 한 곳의 소외 없는 AI 생태계
- 안진하고 공장한 표준 모델

대한민국 4대 공공부문을 아우르는 AI 표준 생태계

핵심 솔루션 키

① 다차원 시큐어 RAG
(Multidimensional Secure RAG)



- 행정 문서 맥락 완전 이해
- 온톨로지 기반 구조화
- 복합 추론 지원
- 단철 없는 분석

② 100% 종량제 기반 정밀 빌링 시스템
(Precision Billing System)



- 무임승차 원천 차단
- 사용량 기반 공정 분담
- 인프라 마비 예방
- 지속 가능한 연합 운영

기술과 공정성이 결합될 때 완성되는 표준



대한민국 4대 공공 부문을 아우르는 AI 표준 생태계

“국민을 위한 AI, 공정하게 공유되는 혁신”

A Fair and Secure AI Ecosystem for Public Innovation