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

공공부문 생성형 AI 플랫폼 구축 전략 가이드 요약본

1. 배경 및 정책 환경

2026 년도 국가 AI 예산이 역대 최대 규모인 10 조 1 천억 원으로 편성되고, 행정안전부와 한국지능정보사회진흥원(NIA)이 180 억 원 규모의 중앙·지방정부 AI 혁신 지원 사업을 추진하는 등 공공부문의 생성형 AI 도입이 본격화되고 있습니다. 디지털플랫폼정부위원회는 「공공부문 초거대 AI 도입·활용 가이드라인 2.0」을 배포하여 범정부 AI 공통 기반 구현, 멀티모달 AI, 에이전틱 AI 등 최신 기술 동향의 적극 반영을 권장하고 있습니다. 그러나 엄격한 보안 요구사항(국정원 N2SF 가이드라인 등)과 한정된 예산, 기술적 불확실성이라는 삼중고 속에서 실제 도입은 난항을 겪고 있습니다.

2. 전라북도청 실증 사례 분석

전라북도청은 AMD MI300x GPU 8 장과 Qwen 3.5 397B, DeepSeek V3 등 최신 오픈소스 모델을 활용하여 광역자치단체 최고 수준의 생성형 AI 플랫폼을 선도적으로 구축했습니다. 전북대학교 대학원 연구원들과의 산학 협력을 통해 외주 개발 비용 없이 시스템을 지속 고도화하는 비용 효율적 모델을 완성했습니다. 그러나 시범 운영 과정에서 동시 접속자가 불과 40 명 수준에 도달하자 GPU 과부하로 서비스가 중단되었고, 1.7 조~6.8 조 파라미터 대형 추론 모델의 실운영이 물리적으로 불가능하다는 한계에 직면했습니다. 또한 기존의 단순 키워드 중심 RAG 방식은 HWP 행정 문서의 복잡한 표 구조와 부서 간 관계(온톨로지)를 AI 가 제대로 인식하지 못해 데이터 맥락이 소실되는 치명적 문제를 드러냈습니다.

3. 핵심 기술 솔루션

다차원 시큐어 RAG: 공공기관 문서의 70% 이상을 차지하는 HWP 파일의 표 구조를 원형 그대로 파싱하여 제목, 작성 부서, 관련 법령 등 3072 차원의 메타데이터를 자동 추출합니다. 공공기관의 명확한 조직도(국-과-팀)를 활용해 별도 컨설팅 없이 데이터 관계도를 자동 구축하며, 고성능 OCR 및 객체 탐지 기술을 통합하여 이미지·도면 등 비정형 데이터까지 처리합니다. 국정원 N2SF 가이드라인을 준수하는 AI DLP 를 기본 탑재하여 0.1 초 이내에 민감 정보를 차단하고, 프롬프트 인젝션 등 신종 위협에 대응하는 탐지 엔진을 분기별로 무상 업데이트합니다.

하이브리드 멀티 LLM 체계: 대외비 문서 요약 등 보안 필수 업무는 내부 행정망의 경량화된 소형 LLM(7B~14B)이 전담하고, 복잡한 정책 시뮬레이션 등 고도 추론 업무는 외부 대형 LLM API 를 안전한 보안 터널로 연동하여 내부 GPU 과부하를 원천 방지합니다. 'Artificial Analysis 인덱스'를 객관적 평가 기준으로 삼아 시스템 LLM 성능이 글로벌 상위 3 개 모델 평균 대비 항상 10 점 이내 격차를 유지하도록 관리하며, 기준 미달 시 연 최대 4 회까지 최신 모델로 무상 교체합니다.

4. 연합 플랫폼 및 정밀 빌링 시스템

전국 226 개 기초자치단체 중 인구 10 만 명 미만의 소규모 지자체 89 곳을 위해 광역자치단체가 선도하는 데이터 통합 연합 플랫폼을 제안합니다. 대규모 공기업을 위한 본사-지사 연동 구조도 설계하여 트래픽을 효율적으로 분산시킵니다. 기존의 불공정한 인구 비례·직원 수 비례 정액제를 전면 폐지하고, 코스닥 상장기업이자 SKT 빌링 시스템 공급사인 엔텔스(NTELS)와 스캐터엑스(ScatterX)가 공동 제공하는 정밀 빌링 시스템을 도입합니다. API 호출 횟수, 벡터 DB 저장 용량, 초 단위 GPU 점유율 등 3 대 핵심 지표를 실시간 계측하는 100% 종량제 과금 체계를 확립하여 무임승차를 완벽히 방지하고 공정한 비용 분담을 실현합니다. 실시간 대시보드로 사용량을 투명하게 공개하고, AI 예산 예측 시뮬레이터로 안정적인 예산 편성을 지원합니다.

5. 기관 유형별 맞춤형 전략

소규모 기초지자체(인구 10 만 미만): 연간 2~3 천만 원의 SaaS 단독 도입 또는 광역 연합 플랫폼 합류 권장. **중대형 기초지자체**(인구 10 만 이상): 3 개년 20~25 억 원 규모의

자체 구축형 하이브리드 모델 제안. **광역자치단체**: 산하 시·군 연합 플랫폼 주도를 위한 3 개년 100~120 억 원 규모 투자, 2 차년도에 엔텔스 정밀 빌링 엔진 도입 필수. **공기업·공공기관**: 본사-지사 연동 및 부서별 과금 체계 구축을 위한 3 개년 250 억 원 규모 전략 제시. 막대한 매몰 비용이 발생하는 영구 라이선스 방식 대신 연간 라이선스를 채택하고, 도입 후 5 년간 가격 동결 및 6 년 차부터 소비자물가지수(CPI) 연동(연 최대 3% 상한) 조정을 통해 10 년 총소유비용(TCO)을 획기적으로 절감합니다.

6. 실행 로드맵 및 계약 가이드

발주 시 과업지시서에 스캐터엑스와 엔텔스의 연합 솔루션 수준을 충족하는 도입 요건을 명시하고, Artificial Analysis 기준 성능 보증 및 연 4 회 무상 교체 의무화 조항을 강제해야 합니다. 국정원 N2SF 지침에 따른 AI DLP 분기별 업데이트, 장기 예산 예측성을 위한 물가 연동 조정 및 6 개월 전 사전 통보 조항, 국내 클라우드 보안 인증(CSAP) 서버 사용에 따른 데이터 주권 보장 및 완전 삭제권 등을 명시하여 투명하고 공정한 데이터 행정 허브를 완성해야 합니다. 전라북도청을 비롯한 선도 기관들의 훌륭한 시도와 고귀한 희생이 단일 기관의 실험으로 끝나지 않고, 대한민국 4 대 공공부문 전체가 소외 없이 최상의 AI 를 누리는 안전하고 공정한 연합 생태계의 표준으로 자리매김하기를 기대합니다.