

[기술 백서]

공직자 생성형 AI 비서 초거대 플랫폼 구현 방법론

광역이 주도하고 기초와 함께 쓰는 AI 플랫폼의 이상적인 구현 방안



흩어져 있는 Scatter
SCATTER

+ **X**
숨겨진 비밀 Value & Insight

[기술 백서] 공직자 생성형 AI 비서 초거대 플랫폼 구현 방법론

부제: 광역이 주도하고 기초와 함께 쓰는 AI 플랫폼의 이상적인 구현 방안

서문 (Preface)

공직자를 위한 진정한 AI 비서를 향한 여정

존경하는 독자 여러분께,

본 기술 백서를 펼쳐주신 데 깊은 감사의 말씀을 드립니다.

인공지능 기술이 급속도로 발전하는 시대, 많은 지자체가 생성형 AI 도입을 추진하고 있지만, 막상 현장에서는 "데이터를 제대로 이해하지 못하는 AI", "보안이 불안한 AI", "비싼 비용에 비해 실효성이 떨어지는 AI"라는 난관에 부딪히고 있습니다. 저희 스캐터엑스는 지난 수년간 이러한 문제들과 정면으로 마주하며, 공공 행정 현장에서 진정으로 필요한 AI가 무엇인지 고민해왔습니다.

그 고민의 결과, 저희는 **"데이터의 정확한 이해 없이는 AI의 지능도 무의미하다"**는 결론에 도달했습니다. 아무리 뛰어난 AI 모델이라도 행정 문서 특유의 복잡한 표, 수치, 맥락을 제대로 인식하지 못한다면 공직자들에게 외면받을 수밖에 없습니다. 또한 아무리 성능이 좋아도 보안이 취약하다면 공공기관에서는 사용할 수 없으며, 아무리 기능이 많아도 비용이 천문학적이라면 지속 가능하지 않습니다.

이러한 문제의식을 바탕으로, 저희는 다음 세 가지 핵심 기술을 개발하고 실제 공공기관 현장에 적용해왔습니다.

첫째, 다차원 RAG 임베딩 기술입니다.

공공기관의 대표적 문서 포맷인 HWP 파일을 PDF 로 변환하지 않고 직접 파싱하여, 문서 내 표의 행과 열이 담고 있는 의미, 숫자 간의 관계, 조건부 로직까지 이해하는 기술입니다. 단순히 텍스트를 벡터화하는 기존 RAG 를 넘어, 텍스트·표·이미지가 혼재된 복합 문서의 구조를 입체적으로 파악합니다. 이 기술은 파주시청, 연천군청을 포함한 6 개 공공기관에서 이미 검증되었으며, "예산 명세서의 특정 항목 전년 대비 증감률"과 같은 복잡한 질의에도 정확히 답변할 수 있습니다.

둘째, AI DLP(Data Loss Prevention) 기술입니다.

외부의 강력한 글로벌 LLM 을 활용하고 싶지만 보안이 걱정되는 지자체의 고민을 해결하기 위해 개발했습니다. 단순한 키워드 차단이 아닌, 프롬프트의 문맥을 이해하여 기밀 정보 유출을 실시간으로 차단하며, 클라우드에 의존하지 않고 로컬 환경에서 경량화되어 구동됩니다. 현재 국정원 보안 가이드라인 충족을 위한 CSAP 인증을 2 월 중 획득 예정이며, 공공·금융 시장의 엄격한 보안 요구사항을 만족시키기 위해 최선을 다하고 있습니다.

셋째, 멀티 LLM 하이브리드 아키텍처입니다.

"하나의 모델에 종속되면 그 모델이 노후화될 때 모든 것을 다시 구축해야 한다"는 문제를 해결하기 위해 설계했습니다. 내부망에서는 보안이 검증된 모델을, 외부망에서는 최신 글로벌 LLM 을 보안 등급에 따라 유연하게 사용할 수 있으며, 특정 모델에 장애가 발생해도 자동으로 다른 모델로 전환(Fail-over)되어 서비스 연속성을 보장합니다. 파주시청, 연천군청, 금천구청, 관악구청, 전력거래소 등에서 1 만여명의 공직자들께서 사용하는 멀티 LLM 환경을 성공적으로 운영하고 있습니다.

그러나 저희는 이것이 완성이 아니라 시작이라고 생각합니다.

본 백서에서 제안하는 '**광역-기초 연합형 생성형 AI 플랫폼**'은 저희가 그동안 쌓아온 기술과 경험을, 더 많은 지자체와 공직자들이 혜택을 받을 수 있도록 확장하고자 하는 청사진입니다. 광역자치단체가 데이터 플랫폼과 보안 인프라를 구축하고, 기초자치단체는 중복 투자 없이 이를 함께 활용하는 모델. 데이터는 내부에서 완벽히

장악하고, 추론은 외부의 거대 지능을 안전하게 빌려오는 하이브리드 전략. 이것이 저희가 제안하는 가장 현실적이고 지속 가능한 해법입니다.

이 백서는 저희 혼자만의 생각이 아닙니다. 여러 지자체 담당자분들과의 수많은 회의, 현장에서 겪은 시행착오, 그리고 무엇보다 "진짜 쓸 수 있는 AI"를 원하는 공직자분들의 목소리가 모여 만들어진 결과물입니다. 부족한 점이 많겠지만, 이 백서가 생성형 AI 도입을 고민하시는 지자체 관계자분들께 작은 나침반이 되기를 진심으로 바랍니다.

저희 스캐터엑스는 앞으로도 겸손한 자세로 현장의 목소리에 귀 기울이며, 공공 행정의 디지털 전환에 미력하나마 기여하고자 합니다. 본 백서에 대한 여러분의 소중한 의견과 조언을 기다리겠습니다.

감사합니다.

2026 년 1 월 22 일

(주)스캐터엑스 대표이사 강명수

0. Executive Summary (요약)

본 백서는 광역자치단체와 기초자치단체가 공동으로 활용할 수 있는 '**광역-기초 연합형 생성형 AI 플랫폼**'의 기술적 아키텍처와 운영 거버넌스를 정의한다.

현재 공공부문의 AI 도입은 개별 기초지자체의 중복 투자, 고비용 GPU 인프라의 비효율성, 그리고 데이터 보안 및 정확성 문제라는 삼중고에 직면해 있다. 특히, 6조 8,500억 개 이상의 파라미터를 가진 초거대 추론 모델을 내부망에 구축하는 것은 재정적으로 불가능하며, 정제되지 않은 데이터 학습은 필연적으로 할루시네이션(Hallucination)을 유발한다.

이에 본 백서는 "**데이터는 내부에서 완벽하게 장악하고(Understanding), 추론은 외부의 거대 지능을 빌려온다(Reasoning)**"는 하이브리드 전략을 제안한다. 이를 실현하기 위해 ①국정원 보안 가이드라인(2025.12)을 준수하는 AI DLP 체계, ②데이터 정확성을 위한 4단계 정제 파이프라인, ③단일 다차원 RAG 및 5가지 서비스 유형, ④비용 효율적인 멀티 LLM 전략, ⑤정밀 빌링 및 성과 분석 시스템을 핵심 솔루션으로 제시한다.

1. 서론: 공공 AI 도입의 패러다임 전환

1.1. 현황 및 문제점: '구축'의 함정

현재 많은 지자체가 경쟁적으로 자체 LLM 구축을 시도하고 있다. 그러나 이는 두 가지 치명적인 문제에 봉착한다.

첫째, **비용의 비대칭성**이다. 공직자들이 기대하는 GPT-4 또는 GPT-5 수준의 고도화된 추론 및 기획 능력을 갖춘 모델을 구동하기 위해서는 H100 등 고가의 GPU 수백 장이 필요하다. DeepSeek의 사례를 보면 GPT-5급 성능을 구현하기 위해서는 약 6조 8,500억 개의 파라미터가 필요하며, 이를 서비스 가능한 수준으로 운영하려면 막대한 GPU 클러스터가 요구된다. 이는 기초지자체 단위 예산으로는 감당할 수 없는 규모다. 실제로 일부 지자체와 중앙부처에서 GPU 8장으로 추론 모델을 운영하려다 서비스 중단 사례가 발생했다.

둘째, **기술의 진부화**다. AI 모델은 3~6개월 단위로 비약적인 성능 향상이 이루어진다. 수십억 원을 들여 구축한 모델은 1년도 지나지 않아 구형이 되며, 지속적인 재학습 비용은 초기 구축비를 상회한다. 일부 광역자치단체에서는 초기에 도입한 5개 LLM 중 3개를 폐기하고 새로운 모델로 교체하는 과정에서 막대한 비용과 시간이 소요된 사례도 있다.

1.2. 데이터 사일로(Silo)와 품질 이슈

더 큰 문제는 데이터다. 각 부서의 행정 데이터는 HWP, PDF 등 비정형 형태로 산재해 있으며, 이를 그대로 벡터 DB에 넣을 경우 AI는 행정 문서 특유의 복잡한 표, 수치, 맥락을 이해하지 못한다. 실무 검증 결과 "AI가 답변을 잘 못하는 근본 원인은 지자체 내부 데이터가 올바르게 정리되지 않았고, 효율적으로 정리되어 있지 않으며, AI가 이해할 수 있도록 데이터 가공이 안 됨"이라는 점이 명확히 확인되었다.

특히 업무 매뉴얼 내 표와 숫자를 기존 RAG가 제대로 인식하지 못하는 문제가 발생했다. 예를 들어 세입예산명세서와 같이 복잡한 표 구조를 가진 문서에서 특정 사업명과 금액, 전년 대비 증감액을 정확히 매핑하지 못해 실무 활용도가 떨어졌다.

이는 엉뚱한 답변을 내놓는 할루시네이션의 주원인이 되며, 결국 공직자들의 외면을 받게 된다.

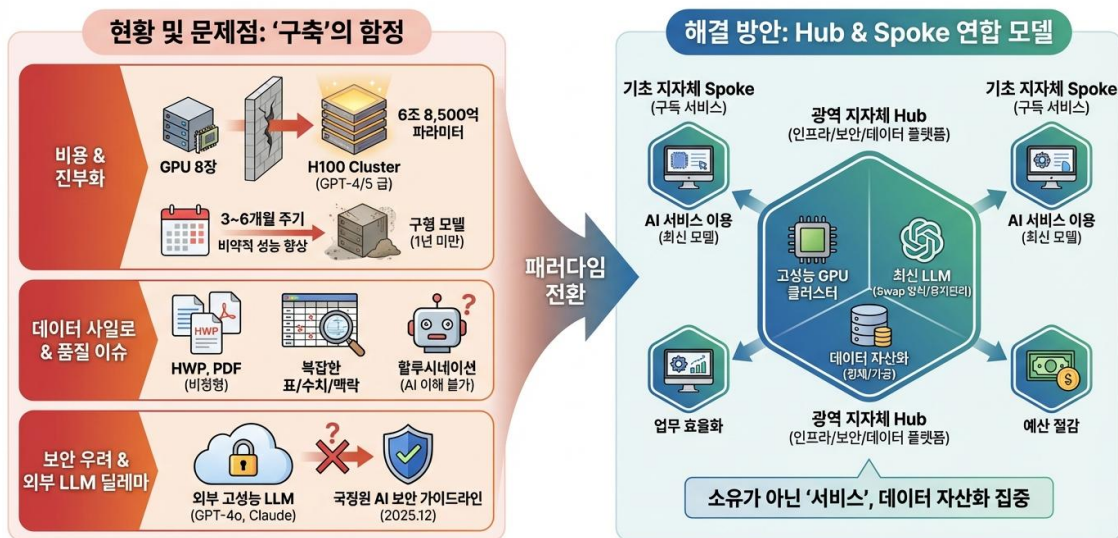
1.3. 보안 우려와 외부 LLM 활용의 딜레마

고성능의 글로벌 LLM(GPT-4o, Claude 등)을 활용하면 성능 문제는 해결되지만, 행정 데이터를 외부로 전송하는 것에 대한 보안 우려가 크다. 국정원은 2025년 12월 16일 AI 보안 가이드라인을 발표하여 외부 LLM 활용 시 필수 보안 요구사항을 제시했지만, 많은 지자체가 이를 어떻게 구현해야 할지 막막한 상황이다.

1.4. 해결 방안: Hub & Spoke 연합 모델

따라서 패러다임의 전환이 필요하다. 광역자치단체가 고성능 인프라와 보안 체계, 데이터 플랫폼(Hub)을 구축하고, 기초자치단체는 이를 구독(Spoke)하는 방식이다. 모델은 소유하는 것이 아니라 '서비스'로 이용하며, 지자체는 '데이터 자산화'에 집중해야 한다. 선진 사례에서는 LLM을 영구 구매하는 것이 아니라 매년 유지관리 계약을 통해 최신 버전으로 교체(Swap)하는 방식을 채택하고 있다.

공공 AI 도입의 패러다임 전환: '구축'에서 '구독'으로



2. [Security] 국정원 가이드라인 준수 보안 아키텍처

외부의 강력한 추론 모델을 활용하기 위해서는 그에 상응하는 강력한 보안 체계가 선행되어야 한다. 본 플랫폼은 2025 년 12 월 16 일 발표된 국정원 「AI 보안 가이드라인」을 완벽하게 준수하는 보안 아키텍처를 기반으로 설계되었다.

2.1. 보안 컴플라이언스 개요

국정원의 AI 보안 가이드라인은 15 개 주요 위협과 30 개 대응 방안을 제시하고 있다. 본 플랫폼은 특히 다음의 핵심 요구사항을 기술적으로 구현한다.

- **M05: 접근 통제** - 사용자/데이터/경로별 권한 제어
- **M09: 입력 데이터 검증** - 프롬프트 인젝션 방지
- **M13: 데이터 유출 방지** - AI DLP 를 통한 민감정보 차단
- **M23: 사용 이력 기록 및 관리** - 로깅 및 감사 시스템

2.2. AI DLP (Data Loss Prevention) 시스템

기존의 정규식(Regex) 기반 차단은 '주민번호 형식' 등 단순 패턴만 걸러낼 뿐, 문맥 속에 숨겨진 민감 정보는 탐지하지 못한다. 본 플랫폼은 LLM 기반의 AI DLP 를 도입한다.

AI DLP 의 작동 원리:

1. **입력 필터링:** 사용자가 프롬프트를 입력하면 AI DLP 가 먼저 분석한다.
2. **문맥 인식 탐지:** "우리 시의 내년도 비공개 예산안 요약해줘"와 같은 프롬프트에서 '비공개 예산'이라는 맥락을 인지하여 차단하거나 경고를 발생시킨다.
3. **자동 마스킹(Masking):** 입력 데이터 내의 PII(개인식별정보 - 주민번호, 전화번호, 주소 등)를 탐지하여 [NAME], [ID_CARD], [PHONE] 등으로 자동 치환 후 외부 모델로 전송한다.
4. **출력 필터링:** AI 답변을 수신한 후에도 다시 한번 검증하여 민감 정보가 포함되어 있지 않은지 확인한다.

5. **복원:** 사용자에게 답변을 보여줄 때는 마스킹된 정보를 원래대로 복원하여 자연스러운 문장으로 제공한다.

준수 항목: 가이드라인의 '데이터 유출 방지(M13)' 및 '입력 데이터 검증(M09)' 항목을 기술적으로 충족한다.

2.3. 로깅 및 감사(Auditing) 체계

AI 활용의 투명성과 사후 추적성을 보장하기 위해 모든 트랜잭션을 기록한다.

로깅 시스템의 구성:

- **실시간 로깅:** 사용자 ID, 접속 IP, 요청 프롬프트(비식별화 처리), AI 응답 결과, 사용된 모델 및 토큰 수, 처리 시간 등을 실시간으로 수집한다.
- **무결성 보장:** 로그 데이터는 블록체인 또는 위변조 방지 스토리지에 저장하여 사후 조작이 불가능하도록 한다.
- **이상 징후 탐지:** 특정 사용자가 단시간에 과도한 API 를 호출하거나, 프롬프트 인젝션(Prompt Injection - AI 의 보안 설정을 무력화하려는 시도) 공격 패턴이 감지될 경우 즉시 세션을 차단하고 관리자에게 알림을 발송한다.
- **감사 리포트:** 월별/분기별로 AI 사용 현황, 보안 이벤트, 차단 내역 등을 자동으로 생성하여 정보보호 담당자에게 제공한다.

준수 항목: 가이드라인의 '사용 이력 기록 및 관리(M23)' 항목을 충족한다.

2.4. 망 분리 및 연계 아키텍처

물리적 망 분리 환경에서 클라우드 기반 LLM 을 안전하게 사용하기 위한 구조다.

3-Zone 아키텍처:

- **내부망(Private Zone):** 민감 데이터 저장소, 내부 구축형 LLM(EXAONE, OSS 등), 관리자 콘솔, 부서별/개인 RAG 가 위치한다. 인터넷과 완전히 차단되어 있다.

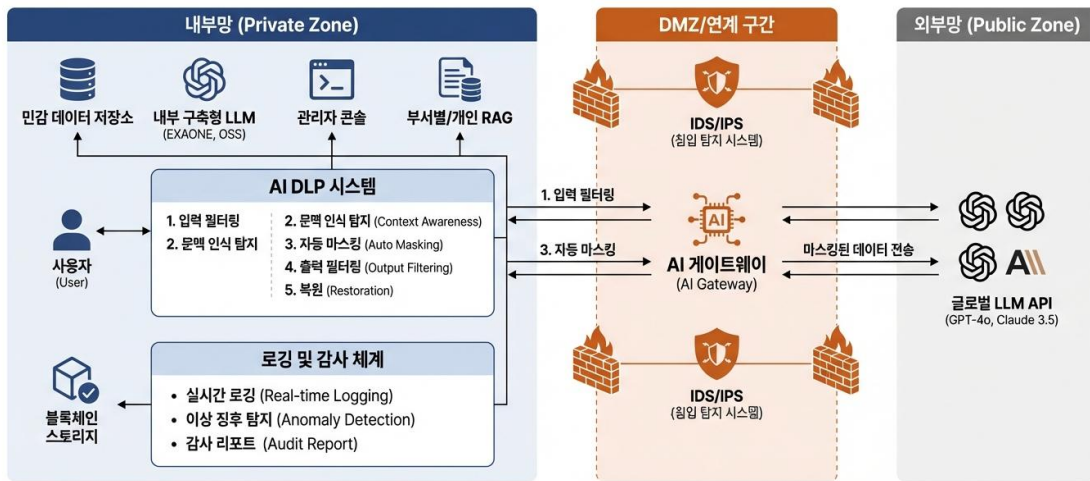
- **DMZ/연계 구간:** AI 게이트웨이(Gateway)가 위치한다. 내부망에서 AI DLP 검증을 통과한 데이터만이 이 구간을 통해 외부 API 로 전송된다. 방화벽과 침입탐지시스템(IDS/IPS)이 이중으로 구성된다.
- **외부망(Public Zone):** 글로벌 LLM API(GPT-4o, Claude 3.5 등)와 통신한다. 이 구간에서는 절대로 원본 민감 데이터가 노출되지 않도록 설계된다.

데이터 흐름 예시:

1. 공직자가 내부망에서 "비공개 인사 자료 요약"을 요청
2. AI 라우터가 '비공개' 태그를 인식하고 내부망 LLM 으로 라우팅
3. 내부 LLM 이 처리하여 답변 제공 (외부로 데이터 전송 없음)

국정원 가이드라인 준수 보안 아키텍처

2025년 12월 16일 발표 「AI 보안 가이드라인」 완벽 준수



3. [Data] 데이터 정확성 확보를 위한 4 단계 정제 방법론

"AI 는 데이터의 거울이다." 데이터가 정확하게 구조화되지 않으면 아무리 좋은 모델도 무용지물이다. 본 플랫폼은 행정 데이터의 특수성을 반영한 **4 단계 정제 파이프라인**을 적용한다.

3.1. 데이터 중심 AI(Data-Centric AI) 철학

실무 검증 결과, AI 의 답변 품질은 모델의 성능보다 '**데이터의 정제 수준**'이 결정한다는 것이 확인되었다. 특히 "업무 매뉴얼 내 표와 숫자를 기존 RAG 가 제대로 인식하지 못하는 문제"가 핵심 과제로 도출되었다.

Step 1. 보안 분류 (Security Layer)

목적: 데이터 수집 단계에서부터 보안 사고 가능성을 원천 차단한다.

방법:

- 문서 수집 시 AI 가 내용을 분석하여 보안 등급(공개/대외비/비공개)을 자동 태깅한다.
- 국가보안법, 개인정보보호법 등 관련 법령에 저촉되는 키워드나 맥락을 탐지한다.
- 담당자가 수동으로 보안 등급을 조정할 수 있는 인터페이스를 제공한다.

활용:

- '비공개' 태그가 붙은 데이터는 물리적으로 외부망 전송 경로가 차단되며, 오직 내부망에 구축된 sLLM 에서만 처리되도록 라우팅된다.
- 이는 AI DLP 와 연동되어 이중 보안 체계를 형성한다.

Step 2. 부서 맥락화 (Department Layer)

목적: 행정 용어의 동음이의어 문제와 부서별 관점 차이를 해결한다.

방법:

- 문서에 소유 부서, 상위 국/실, 관련 부서 정보를 메타데이터로 부여한다.
- 조직도 정보와 연동하여 부서 간 계층 관계를 반영한다.
- 동일한 용어라도 부서에 따라 다른 의미로 해석될 수 있음을 AI 에 학습시킨다.

활용 사례:

- "예산 현황 알려줘"라는 동일한 질문에 대해:
- 복지과 직원에게는 '복지 급여 예산'을 우선 검색
- 도로과 직원에게는 '도로 보수 예산'을 우선 검색
- 기획예산과 직원에게는 '전체 예산 편성 현황'을 제공

효과:

- 부서별 업무 맥락을 정확히 이해하여 답변의 정확성이 크게 향상된다.
- 출산 장려 지원금과 같이 광역·기초 자치단체가 각각 예산을 편성하여 3 단계로 지급하는 복잡한 구조도 명확히 구분하여 안내할 수 있다.

Step 3. 업무 세분화 (Task Layer)

목적: 사용자의 질문 의도에 맞는 최적의 문서 유형을 매칭한다.

방법:

- 문서의 성격을 다음과 같이 세분화하여 분류한다:
- 법령/조례/규칙
- 지침/가이드라인
- 사업 계획서
- 결과 보고서
- 통계표/수치 데이터
- 회의록/협약 내용
- 민원 Q&A 사례집

활용 사례:

- "이 민원을 어떻게 처리해야 하나요?" → 법령/지침 + 민원 사례집 우선 검색
- "내년도 사업 아이디어를 제안해주세요" → 타 지자체 계획서 + 우수 사례 우선 검색
- "작년 실적 대비 올해 목표는?" → 결과보고서 + 통계표 우선 검색

효과:

- 질문의 의도를 정확히 파악하여 가장 적합한 참조 문서를 선별함으로써 답변의 적합성이 향상된다.

Step 4. 보고서 통합 (Report Layer)

목적: 단편적인 정보 검색을 넘어, 논리적이고 풍부한 근거를 갖춘 보고서 작성을 지원한다.

방법:

- 내부 행정 데이터뿐만 아니라 다음의 외부 데이터를 연결한다:
- 중앙정부 정책 자료
- 타 지자체 우수 사례
- 뉴스 기사 및 여론 동향
- 통계청 공공 데이터
- 학술 논문 및 연구 보고서
- 이들 간의 인용 관계, 연관성을 지식 그래프(Knowledge Graph)로 형성한다.

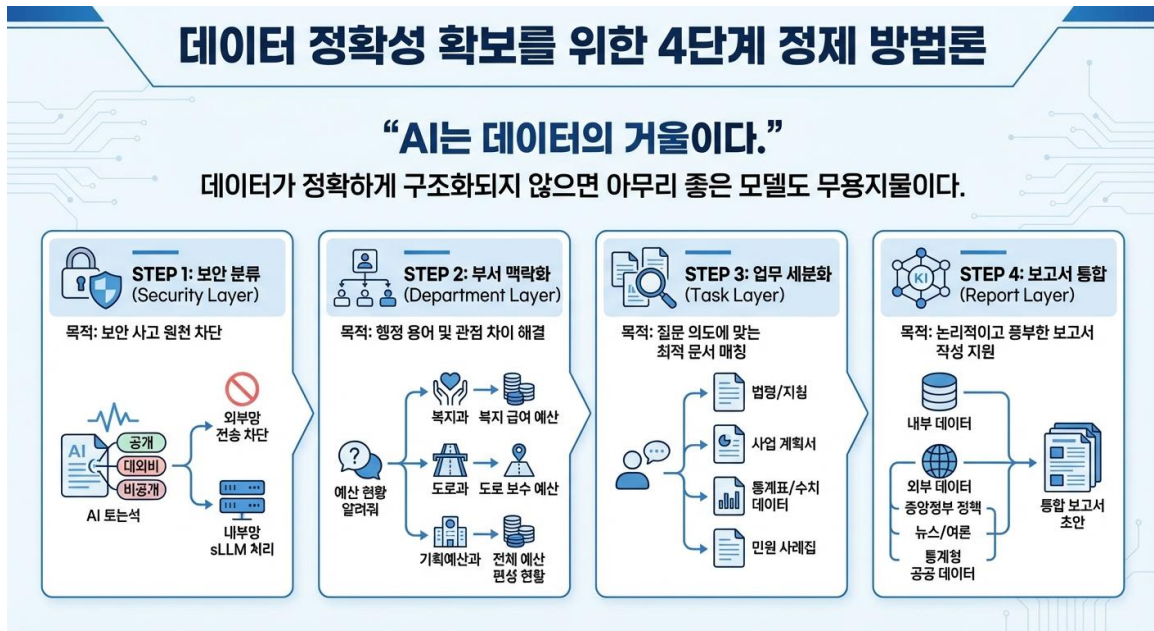
활용 사례:

- "청년 일자리 사업의 기대효과를 작성해줘"라는 요청 시:
- 내부 사업 계획서의 목표
- 타 지자체의 유사 사업 성과 통계
- 관련 뉴스 기사 (긍정적 여론)
- 고용노동부의 청년 고용 정책 방향

- 을 모두 종합하여 논리적인 보고서 초안을 생성

효과:

- 단순 사실 나열이 아닌, 설득력 있는 근거 기반 행정이 가능해진다.



4. [RAG] 단일 다차원 RAG 및 5 가지 서비스 유형

4.1. 단일 다차원 RAG (Single Multi-dimensional RAG)

기존 RAG(검색 증강 생성)는 텍스트를 평면적으로 검색하는 데 그쳐, 표나 복잡한 구조를 이해하지 못했다. 본 플랫폼은 **다차원 임베딩** 기술을 적용한다.

다차원 임베딩의 기술적 구현:

- **텍스트 차원:** 일반적인 문장과 단락을 의미 벡터로 변환
- **표 구조 차원:** 표의 행/열 관계, 헤더 정보를 별도로 인식하여 구조적 쿼리 지원
- **수치 데이터 차원:** 금액, 비율, 증감률 등 숫자를 별도 차원으로 처리하여 수치 비교 연산 가능
- **이미지/도표 차원:** OCR 및 이미지 인식을 통해 도표 내 정보 추출
- **메타데이터 차원:** 앞서 정의한 4 단계(보안-부서-업무-보고서) 정보를 차원으로 포함

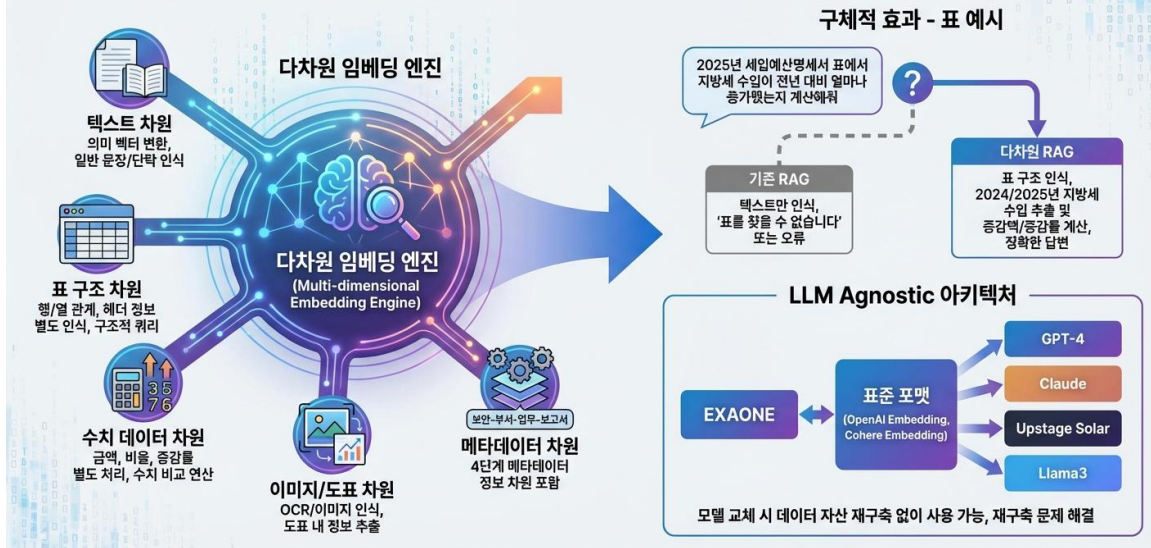
구체적 효과:

- "2025 년 세입예산명세서 표에서 지방세 수입이 전년 대비 얼마나 증가했는지 계산해줘"
- 기존 RAG: 텍스트로만 인식하여 "표를 찾을 수 없습니다" 또는 엉뚱한 수치 제시
- 다차원 RAG: 표 구조를 인식하고 2024 년 지방세 수입과 2025 년 수입을 정확히 추출하여 증감액과 증감률을 계산하여 답변

LLM Agnostic 아키텍처:

- 이 RAG 시스템은 특정 LLM(예: Llama3)에 최적화되지 않은 표준 포맷(OpenAI Embedding, Cohere Embedding 등)으로 구축된다.
- 따라서 추후 LLM 모델을 EXAONE 에서 Upstage Solar 로 교체하거나, GPT-4 에서 Claude 로 변경하더라도 데이터 자산은 재구축 없이 그대로 사용할 수 있다.
- 일부 광역자치단체에서 3 개 LLM 을 폐기하면서 겪은 RAG 재구축 문제를 원천적으로 해결한다.

단일 다차원 RAG (Single Multi-dimensional RAG) & 5 Service Types



4.2. 5 가지 맞춤형 RAG 서비스

단일 플랫폼 내에서 사용자의 권한과 목적에 따라 5 가지 가상 RAG 를 동적으로 구성하여 제공한다.

1. 공통 RAG (Common Knowledge Base)

- **대상 데이터:** 국가 법령, 광역 조례, 온나라/새울 시스템 매뉴얼, 공통 행정 서식
- **관리 주체:** 광역자치단체
- **활용:** 전 직원이 공유해야 하는 표준 지식. 법령 검색, 시스템 사용법 안내 등

2. 부서 RAG (Department Knowledge Base)

- **대상 데이터:** 해당 부서의 업무 매뉴얼, 인수인계 자료, 과거 사업 내역, 부서 특화 지침
- **관리 주체:** 각 기초지자체의 부서
- **활용:** 부서원끼리만 공유되는 격리된 데이터 공간. 신규 직원 교육, 업무 인수인계 지원
- **보안:** 타 부서는 절대 접근 불가 (Multi-tenancy 격리)

3. 업무 RAG (Project Knowledge Base)

- **대상 데이터:** TF 팀, 축제 준비위원회, 재난 대응 상황실 등 특정 프로젝트 관련 자료
- **관리 주체:** 해당 프로젝트 책임자
- **활용:** 한시적으로 구성되는 협업 지식 베이스. 프로젝트 종료 후 아카이빙 또는 삭제
- **특징:** 여러 부서의 직원이 함께 참여하는 크로스펑셔널(Cross-functional) 협업 지원

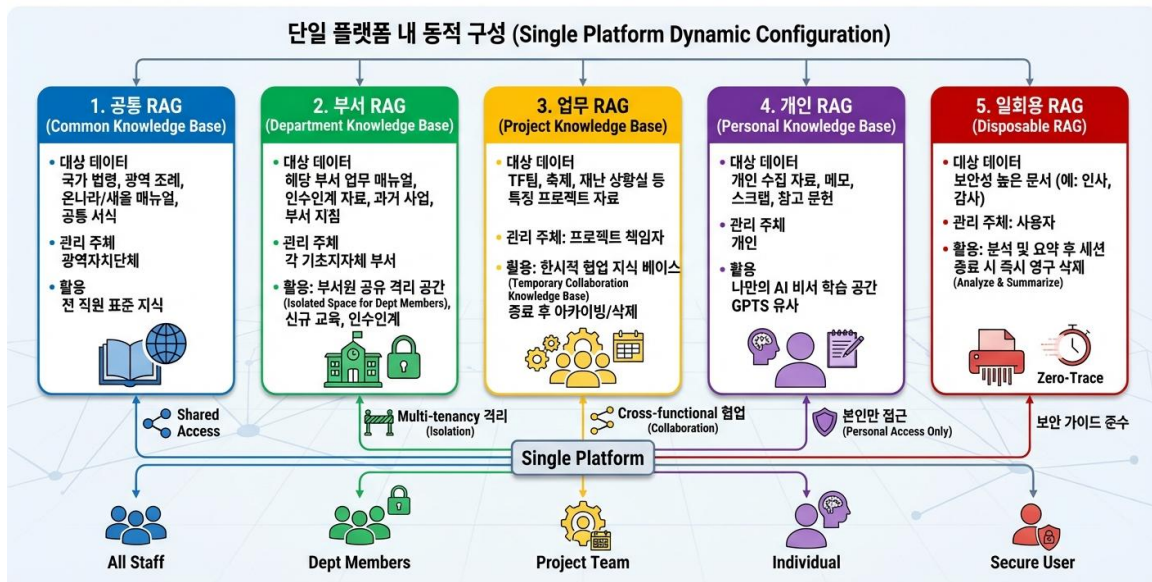
4. 개인 RAG (Personal Knowledge Base)

- **대상 데이터:** 공무원 개인이 수집한 자료, 메모, 스크랩, 참고 문헌
- **관리 주체:** 개인
- **활용:** 나만의 AI 비서를 학습시키는 공간. ChatGPT 의 GPTS 와 유사한 개념
- **보안:** 본인만 접근 가능

5. 일회용(Disposable) RAG (Volatile Secure Mode)

- **대상 데이터:** 보안성이 극도로 높은 문서 (예: 인사 발령안, 감사 자료)
- **관리 주체:** 사용자
- **활용:** 파일을 업로드하여 분석 및 요약 수행한 후, 세션이 종료되면 해당 데이터와 벡터값은 즉시 영구 삭제(Zero-Trace)된다.
- **보안:** 국정원 보안 가이드의 '데이터 최소 보유 원칙' 준수
- **사용 사례:** "이 감사 자료에서 핵심 이슈만 뽑아줘" → 분석 완료 후 로그아웃 시 자동 삭제

4.2. 5가지 맞춤형 RAG 서비스



5. [LLM] 비용 효율적인 하이브리드 멀티 LLM 전략

모든 업무에 최고사양의 LLM 을 쓰는 것은 낭비이며, 모든 업무를 저사양 모델로 처리하는 것은 불가능하다. 본 플랫폼은 **역할(Role)에 따른 철저한 이원화 전략**을 취한다.

5.1. 전략의 핵심: 역할 분담

실무 검증을 통해 도출된 핵심 전략은 다음과 같다:

- **Understanding(이해)은 내부에서:** 데이터 독해, 팩트 추출, 구조 파악
- **Reasoning(추론)은 외부에서:** 논리적 사고, 창의적 기획, 복잡한 문제 해결

이 전략은 비용과 성능의 최적 균형점을 찾는 것이다.

5.2. Internal LLM (The Understanding Engine)

모델 선정:

- **LG EXAONE 100B, OpenAI OSS 100B** 등 한국어 성능이 우수한 고성능
경량/오픈소스 모델

운영 전략: GPU 최소화

- 학습(Training)이 아닌 추론(Inference) 전용으로 최적화(Quantization)하여, H100 소량 또는 L40S 클러스터 등 합리적인 인프라에서 구동한다.
- 양자화(Quantization) 기술을 적용하여 FP16 또는 INT8 수준으로 모델 크기를 압축함으로써 GPU 메모리 요구량을 50% 이상 절감한다.
- 배치 처리(Batch Processing) 최적화를 통해 동시 다중 사용자 요청을 효율적으로 처리한다.

핵심 역할:

- **한국어 행정 문서 독해:** HWP, PDF 등 공공기관 특유의 문서 형식을 정확히 파싱하고 이해한다.

- **팩트 추출 및 구조화:** 복잡한 표, 수치, 조직 구성 현황 등을 정확히 인식하여 구조화된 데이터로 변환한다.
- **내부 비공개 데이터 처리:** 보안등급 '비공개'로 분류된 데이터의 요약, 검색, 1 차 초안 작성을 전담한다.
- **질문 의도 파악:** 사용자의 프롬프트를 분석하여 내부 처리 가능 여부를 판단하고, 필요시 외부 LLM 으로 라우팅한다.

특징:

- 데이터의 '이해(Understanding)'와 '정리'에 특화되어 있다.
- 외부 인터넷 연결이 전혀 필요 없어 망분리 환경에서 완벽하게 작동한다.
- 공직자들이 일상적으로 수행하는 업무(규정 검색, 민원 답변 초안, 간단한 요약)의 약 70%를 처리할 수 있다.

5.3. External LLM (The Reasoning Engine)

모델 선정:

- **GPT-4o/5, Claude 3.5 Sonnet, o1** 등 현존하는 최고 성능의 글로벌 상용 모델 (API 연동)

운영 전략: Pay-per-use (종량제)

- AI DLP 를 통과한 '공개' 등급 데이터에 한해서만 호출하며, 사용한 토큰만큼만 비용을 지불한다.
- 자체 GPU 인프라 구축 대비 약 80% 이상의 초기 투자비(CAPEX)를 절감할 수 있다.
- 최신 모델이 출시될 때마다 API 엔드포인트만 변경하면 즉시 업그레이드되므로 기술 노후화 문제가 없다.

핵심 역할:

- **고도 추론:** 내부 LLM 이 정리한 팩트를 바탕으로 논리적 인과관계 분석, 정책 효과 예측 등 복합적 사고 수행.

- **창의적 기획:** 새로운 사업 아이디어 도출, 보도자료 작성, 슬로건 제안 등 창의성이 요구되는 업무.
- **복잡한 문제 해결:** "A 정책과 B 정책이 상충할 때 최적의 절충안은?" 같은 다차원 의사결정 지원.
- **다국어 번역 및 글로벌 벤치마킹:** 해외 우수 사례 조사, 영문 보고서 작성 등.

특징:

- 데이터의 '추론(Reasoning)'과 '창의성'에 특화되어 있다.
- 6 조 8,500 억 개 파라미터급 모델을 자체 구축할 경우 필요한 천문학적 GPU 비용을 API 사용료(월 수백만~수천만 원)로 대체한다.
- 공직자들이 "ChatGPT 보다 못하다"고 느끼는 성능 격차를 완전히 해소한다.

5.4. 하이브리드 오케스트레이션 (AI Router)

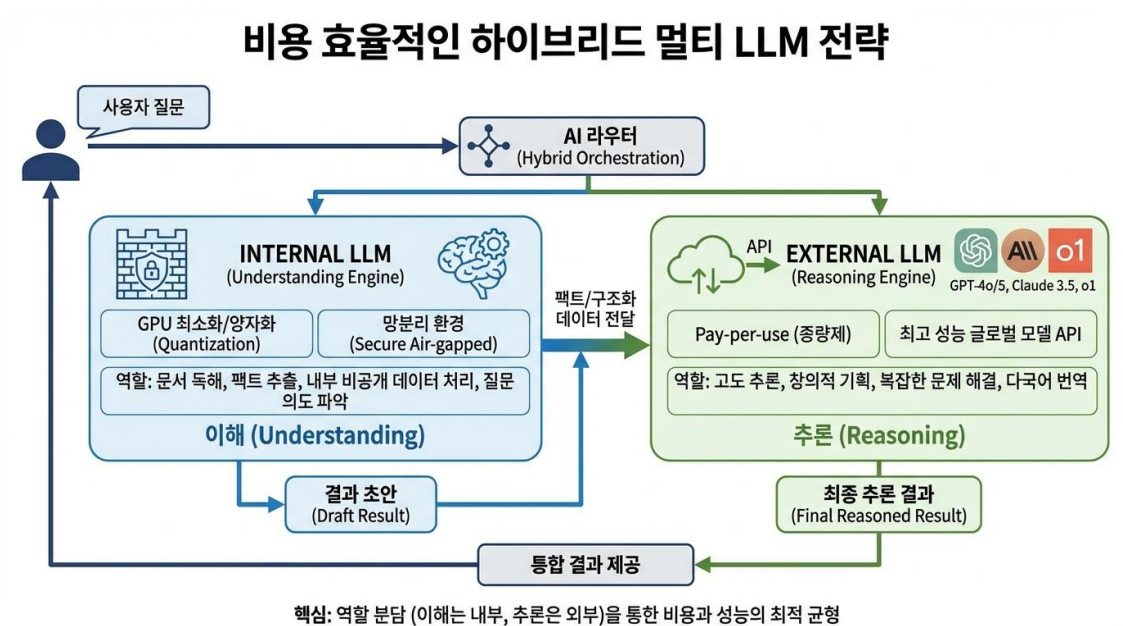
사용자가 질문을 입력하면, 플랫폼의 **AI 라우터(Router)**가 질문의 난이도와 보안 등급을 분석하여 최적의 모델로 연결한다.

라우팅 로직:

1. **보안 등급 확인:** 질문에 참조될 데이터의 보안 태그를 확인한다.
 - '비공개' 데이터 포함 → 무조건 내부망 LLM
 - '공개' 데이터만 → 다음 단계 진행
2. **업무 복잡도 분석:** 질문의 유형을 AI 가 자동 분류한다.
 - 단순 검색/요약형 → 내부망 LLM (비용 절감)
 - 추론/기획형 → 외부망 LLM (성능 극대화)
3. **하이브리드 처리:** 복잡한 질문의 경우 단계별로 분업한다.
 - 1 단계: 내부 LLM 이 관련 데이터를 검색하고 정리
 - 2 단계: 정리된 팩트를 외부 LLM 에 전달하여 추론 및 기획안 도출
 - 3 단계: 결과를 통합하여 사용자에게 제공

라우팅 예시:

- "온나라 시스템 사용법 알려줘" → 내부 LLM (공통 RAG 검색)
- "우리 시 청년 일자리 사업의 기대효과를 논리적으로 설명해줘" → 하이브리드 (내부 LLM 이 사업 데이터 정리 + 외부 LLM 이 논리 구조화)
- "비공개 인사 자료 요약해줘" → 내부 LLM (외부 전송 차단)



6. [Governance] 통합 빌링 및 성과 분석(ROI)

시스템

광역자치단체(제공자)와 기초자치단체(사용자) 간의 지속 가능한 협력을 위해서는 공정하고 투명한 비용 정산과 성과 입증이 필수적이다.

6.1. 계층적 사용량 집계 (Hierarchical Metering)

집계 구조:

- [기초지자체 → 국(실) → 과 → 팀 → 개인]의 5 단계로 데이터를 상세하게 드릴다운(Drill-down)하여 추적한다.

집계 대상 지표:

분류	측정 항목	분석 목적
추론(Reasoning)	외부 LLM 토큰량	고난이도 기획 업무 비중 분석
이해(Understanding)	내부 LLM 토큰량	일상 행정/보안 업무 비중 분석
지식(Knowledge)	RAG 검색 횟수	데이터 활용도 및 근거 기반 행정 빈도
자산(Asset)	데이터 저장량(GB)	부서별 지식 자산화 기여도
활동(Activity)	세션/트랜잭션 수	활성 사용자 및 헤비 유저 식별

대시보드 기능:

- 각 지자체 담당자는 웹 대시보드를 통해 실시간으로 다음 정보를 확인할 수 있다:
- 이번 달 예상 청구 금액 vs 잔여 예산
- 부서별/개인별 AI 사용량 순위
- 가장 많이 사용된 모델 및 RAG 유형
- 시간대별 트래픽 히트맵

6.2. 하이브리드 과금 체계

비용의 성격에 따라 분담 방식을 이원화한다.

구축비 분담금 (CAPEX - 고정비):

- **대상:** 내부망 GPU 서버, 스토리지, 공통 RAG 구축 및 유지보수 비용
- **산정 기준:** 기초지자체의 공무원 정원(User License) 또는 재정 규모에 비례
- **청구 방식:** 연 1 회 선납 또는 분기별 분납
- **특징:** 예측 가능한 고정 비용으로 예산 수립이 용이함

종량제 사용료 (OPEX - 변동비):

- **대상:** 외부망 API 토큰 사용량 및 내부망 연산 점유율
- **산정 기준:** 실제 사용한 토큰 수 × 모델별 단가 (협의 결정)
- **청구 방식:** 월 단위 후불 정산 (Pay-as-you-go)
- **특징:** "안 쓰면 안 내는" 합리적 구조로 과금 공정성 확보

예산 통제 기능:

- 월별 예산 한도를 설정하면, 80% 도달 시 자동 알림 발송
- 100% 도달 시 외부 LLM 사용을 일시 중단하고 내부 LLM 만 사용하도록 자동 전환 (서비스 중단 방지)

6.3. 성과 및 ROI(투자 대비 효과) 분석

단순히 "얼마를 썼다"가 아니라 "얼마나 행정 효율을 높였나"를 증명한다.

정량적 지표:

1) 시간 절감 효과

- **측정 방법:** 업무 유형별(번역, 요약, 초안 작성, 자료 조사) 표준 수행 시간 대비 AI 활용 시 단축된 시간을 산출
- **환산:** 절감 시간 × 공무원 평균 시간당 인건비 = 비용 절감액

- **예시:** "이번 달 기획예산과는 AI 를 통해 총 120 시간의 문서 작성 시간을 단축하여 약 600 만 원의 인건비를 절감함"

2) 비용 대체 효과

- **측정 방법:** 번역 외주, 자료 조사 용역, 디자인 외주 등을 AI 로 대체하여 절감한 직접 예산을 산출
- **예시:** "보도자료 번역을 외주 대신 AI 로 처리하여 연간 2,000 만 원 절감"

정성적 지표:

1) 데이터 자산화 지수

- **측정 방법:** 각 부서가 플랫폼에 업로드하고 공유한 양질의 데이터(매뉴얼, 사례집 등) 건수 및 조회수를 평가
- **활용:** 우수 부서에 인센티브 제공하여 지식 공유 문화 확산

2) 답변 채택률 (Acceptance Rate)

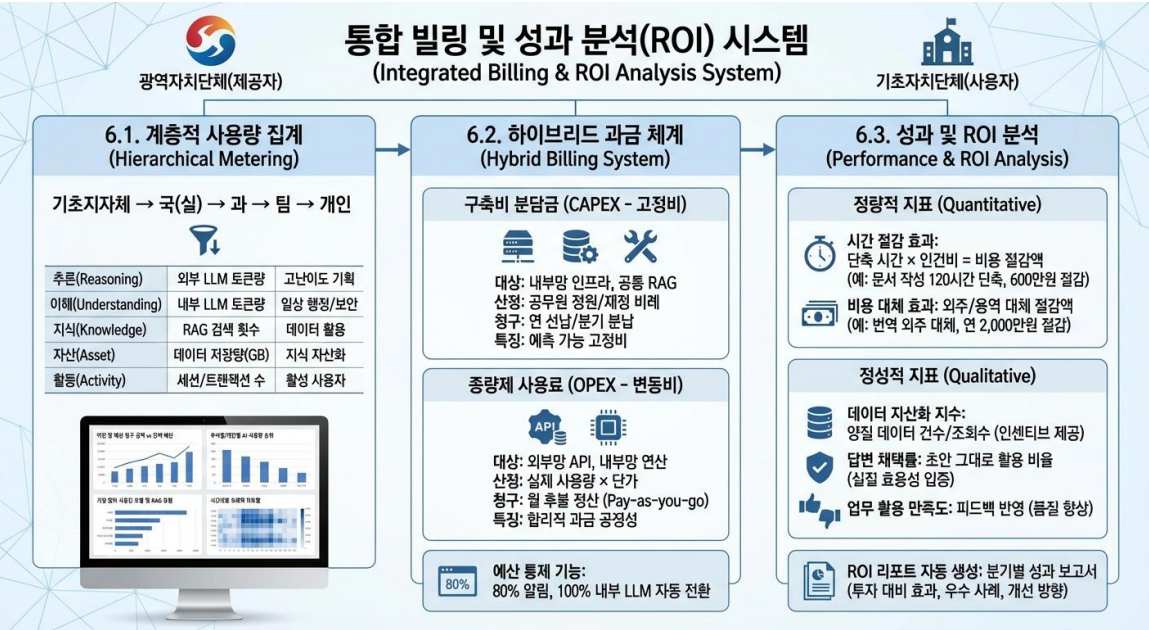
- **측정 방법:** AI 가 생성한 초안을 사용자가 수정 없이 그대로 활용한 비율을 분석
- **활용:** 모델의 실질적 효용성을 입증하고, 채택률이 낮은 업무 영역을 파악하여 개선

3) 업무 활용 만족도

- **측정 방법:** AI 답변 하단에 '좋아요/싫어요' 버튼 배치 및 분기별 설문조사 실시
- **활용:** 사용자 피드백을 반영하여 프롬프트 템플릿 개선 및 RAG 품질 향상

ROI 리포트 자동 생성:

- 분기별로 다음 내용이 포함된 성과 보고서를 자동 생성하여 단체장 및 의회에 제출:
- 총 투자 비용 (구축비 + 사용료)
- 시간 절감 효과 (시간 → 금액 환산)
- 비용 대체 효과
- 부서별 활용 우수 사례
- 향후 개선 방향



7. 추진 로드맵 및 결론

7.1. Agile 도입 전략: PoC First

수십억 원 규모의 대규모 ISP(정보화전략계획)를 수립하고 일괄 구축하는 방식은 실패 확률이 높다. 본 플랫폼은 **점진적 검증과 확산(Agile Approach)**을 추진 전략으로 채택한다.

1 단계: PoC (Proof of Concept) - 6 개월

- **예산:** 약 2,000 만 원 내외 (부가세 별도)
- **범위:** 광역 본청 핵심 부서(기획예산과, 정보화담당관) 및 선도 기초지자체 1~2 곳
- **검증 목표:**
 - 4 단계 데이터 정제 파이프라인의 실효성 검증
 - 다차원 RAG 의 표/수치 인식 정확도 측정
 - AI DLP 의 민감정보 차단율 테스트
 - 하이브리드 LLM 라우팅 로직의 비용 효율성 분석
- **성공 기준:** 기존 대비 답변 정확도 30% 향상, 보안 사고 0 건

2 단계: Pilot (시범 사업) - 6 개월

- **예산:** 1 차 본 사업 예산 (수억 원 규모)
- **범위:** PoC 결과를 바탕으로 데이터 플랫폼 전면 구축, 희망 기초지자체 5~10 곳으로 확대
- **구축 내용:**
 - 광역 데이터 센터에 내부망 LLM 서버 구축
 - AI DLP 및 게이트웨이 시스템 구축
 - 통합 빌링 시스템 가동
- **성공 기준:** 참여 지자체 만족도 80% 이상, ROI 150% 이상

3 단계: Full Launch (전면 확산) - 1 년

- **범위:** 광역 내 전 기초지자체 대상 서비스 확산

- **고도화:**
- 멀티 에이전트(Multi-Agent) 서비스 오픈 (AI 끼리 협업하여 복잡한 업무 처리)
- 행정 시스템(온나라, 새울) API 연동으로 자동 기안서 작성
- 음성 인터페이스 추가 (전화 민원 자동 응대)

7.2. 구독형(Subscription) 생태계 조성

AI 모델은 '소유'하는 것이 아니라 '**지속적으로 교체(Swap)**'하는 것이다.

LLM 운영 방식의 전환:

- **기존 방식:** 특정 모델을 구매하여 파인튜닝(Fine-tuning) 후 영구 사용
- 문제점: 3 개월 뒤 구형 모델이 되어 성능 저하, 재학습 비용 천문학적
- **제안 방식:** 연 단위 유지관리 계약을 통해 최신 버전으로 지속 교체
- 장점: 항상 State-of-the-Art 성능 유지, 재학습 불필요, 예산 예측 가능

광역자치단체의 역할:

- 내부망 LLM: 연 1 회 최신 오픈소스 모델로 업그레이드
- 외부망 API: 글로벌 LLM 제공사와 포괄 계약 체결하여 최신 모델 즉시 적용
- 기초지자체는 별도 계약 없이 구독료만으로 최신 기술 향유

7.3. 결론

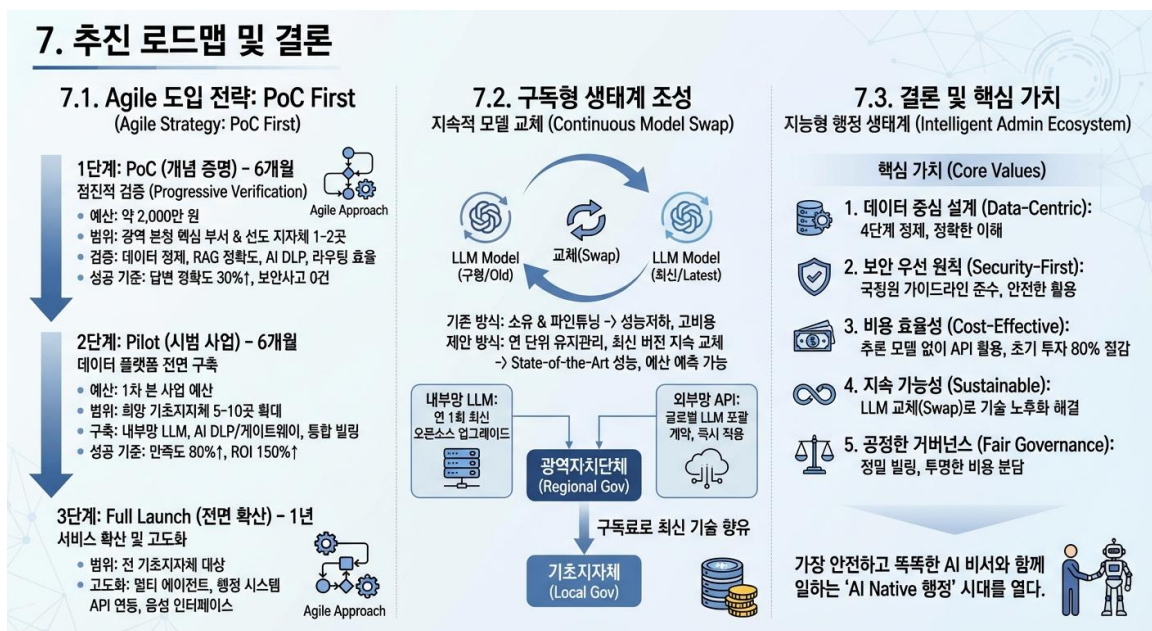
본 백서에서 제안하는 **[광역-기초 연합형 생성형 AI 플랫폼]**은 단순한 시스템 구축 사업이 아니다. 이는 광역자치단체의 '**인프라 및 보안 리더십**'과 기초자치단체의 '**현장 행정 데이터**'가 결합하여 만들어내는 거대한 지능형 행정 생태계다.

핵심 가치:

1. **데이터 중심 설계:** 4 단계 정제 파이프라인을 통해 AI 가 행정 문서를 정확히 이해하도록 함
2. **보안 우선 원칙:** 국정원 가이드라인을 완벽 준수하여 외부 LLM 을 안전하게 활용

3. **비용 효율성:** 추론 모델을 구축하지 않고 API 로 활용하여 초기 투자비 80% 절감
4. **지속 가능성:** LLM 교체(Swap) 방식으로 기술 노후화 문제 해결
5. **공정한 거버넌스:** 정밀 бил링 시스템으로 투명한 비용 분담 및 성과 입증

이 플랫폼을 통해 우리는 예산 낭비를 막고, 보안 우려를 해소하며, 모든 공직자가 가장 안전하고 똑똑한 AI 비서와 함께 일하는 '**AI Native 행정**'의 시대를 열 수 있을 것이다.



최종 메시지:

"지금 수백억을 들여 자체 AI 를 만들어도 6 개월 뒤면 고철이 됩니다. 우리는 AI 를 '만드는' 것이 아니라, 가장 잘 '부리는' 플랫폼을 만들어야 합니다. 그것이 바로 데이터 중심의 하이브리드 연합 플랫폼입니다."