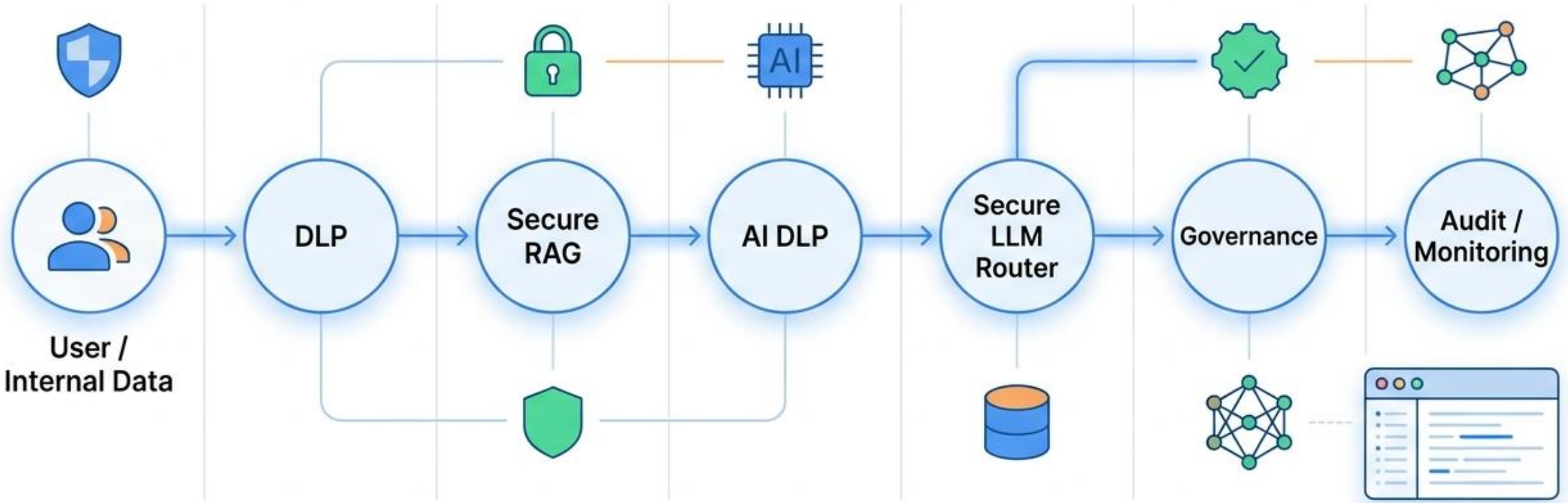
국정원 AI 보안 가이드 2.0 대응
Secure RAG 및 LLM 라우팅 통합 보안 플랫폼 제안

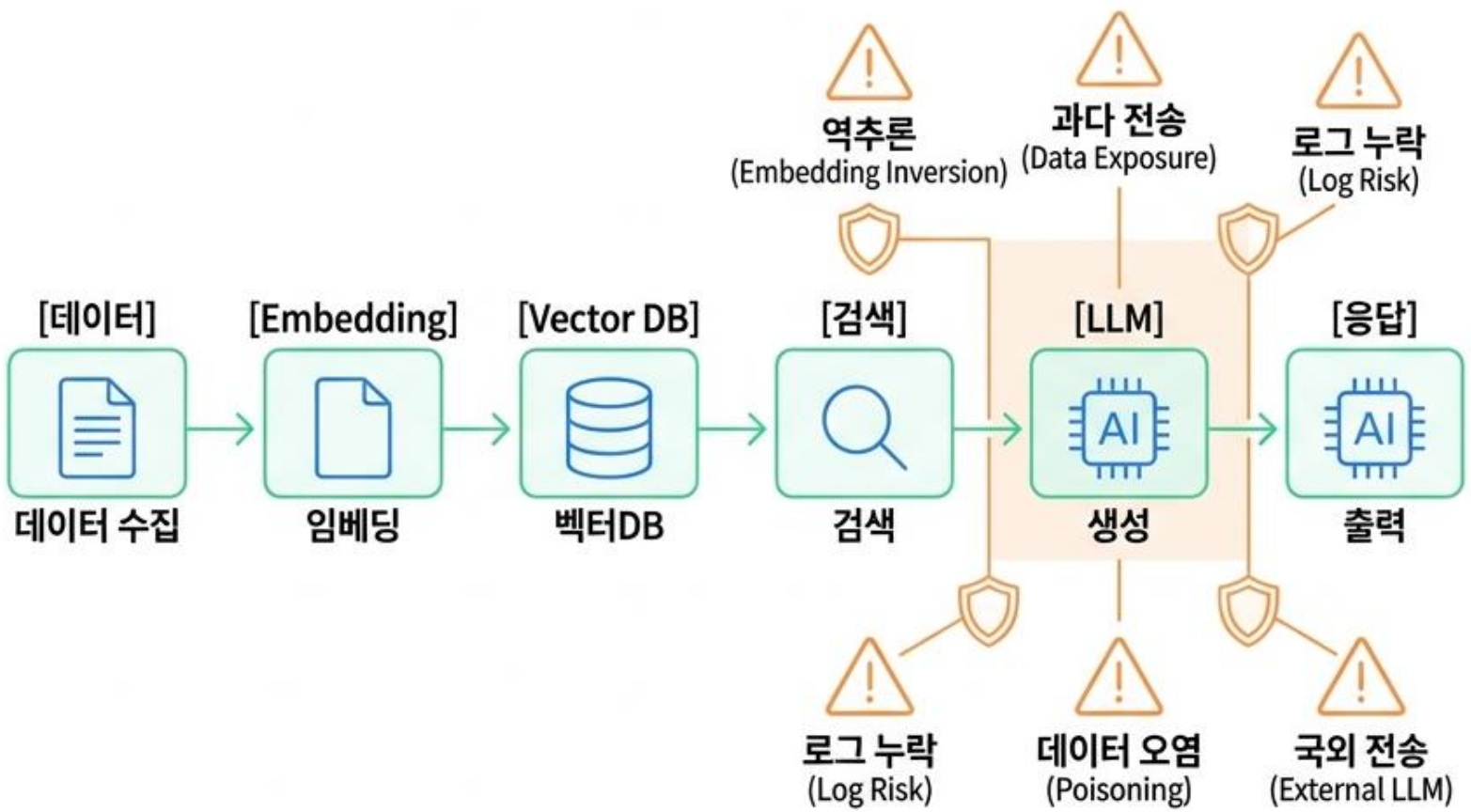
(주)스캐터엑스

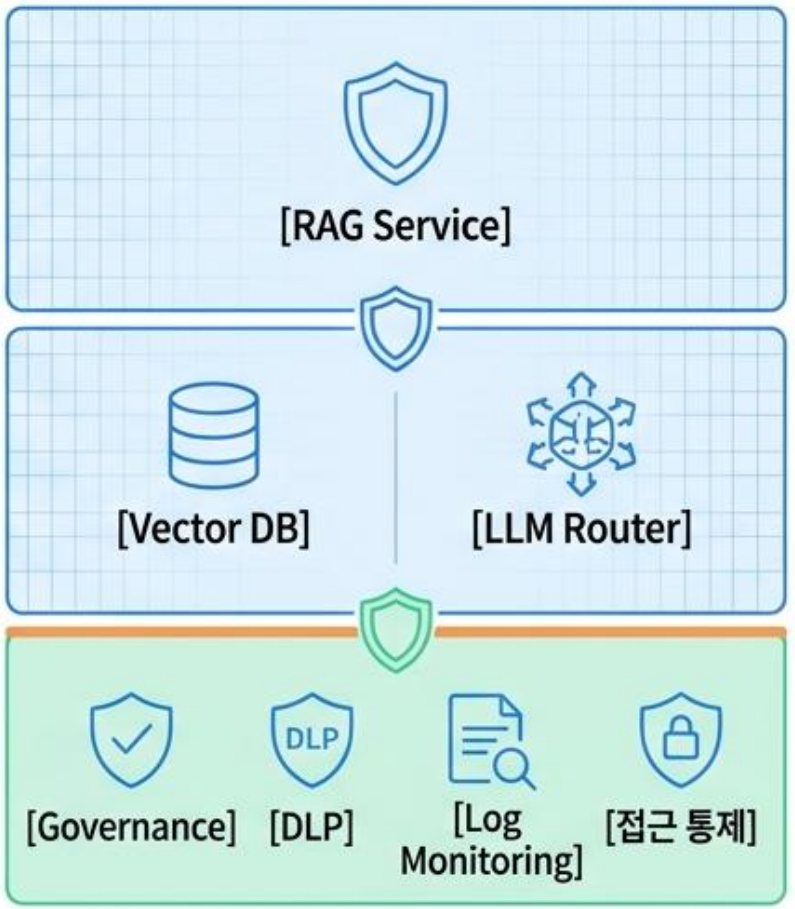


공공 생성형 AI 확산에 따른 보안 위협 고도화에 체계적 대응이 필요합니다.



RAG 기반 공공 AI 도입은 수명주기 전반의 구조적 보안 통제가 필수입니다.





게이트키퍼형 DLP는 RAG 기반 AI의 수명주기 보안을 통제하기에 구조적으로 한계가 있습니다.

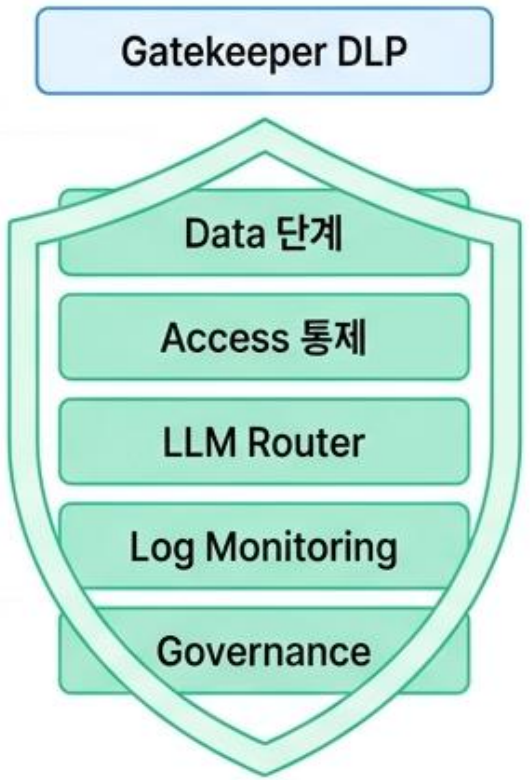
Gatekeeper DLP



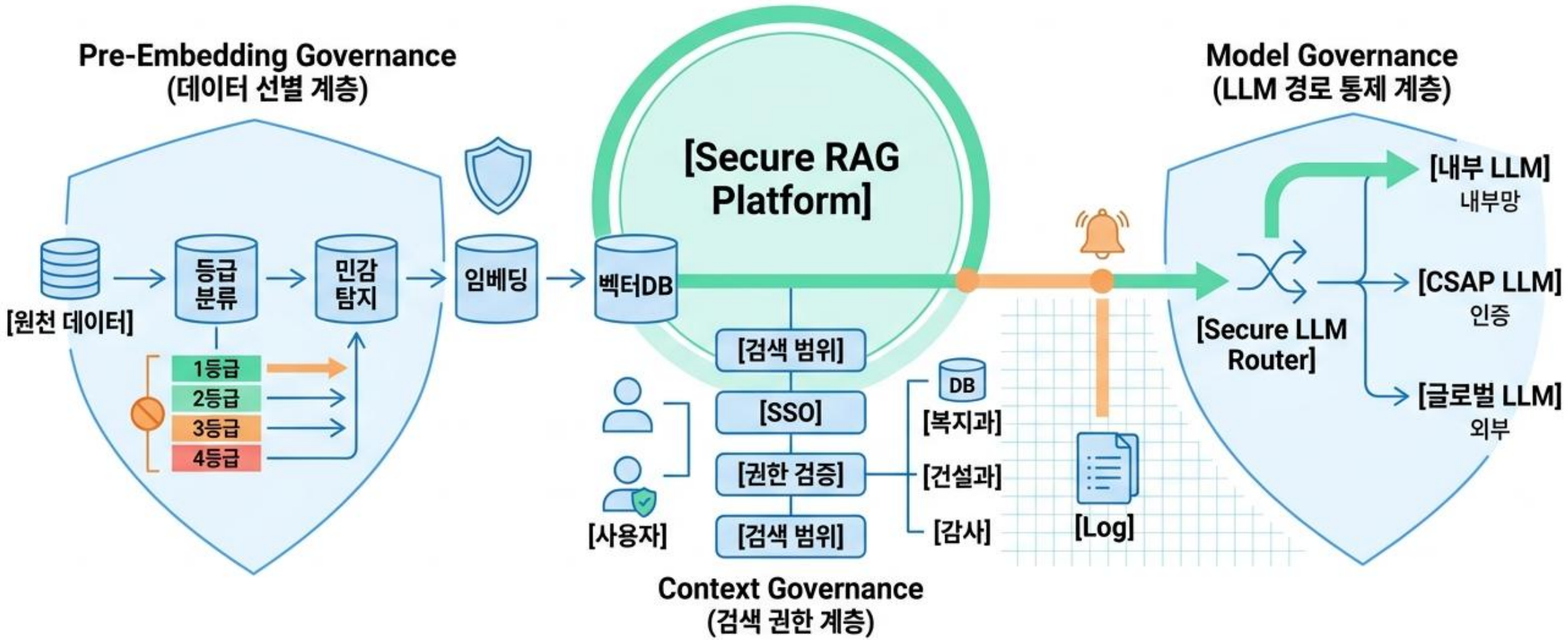
RAG Pipeline Structure



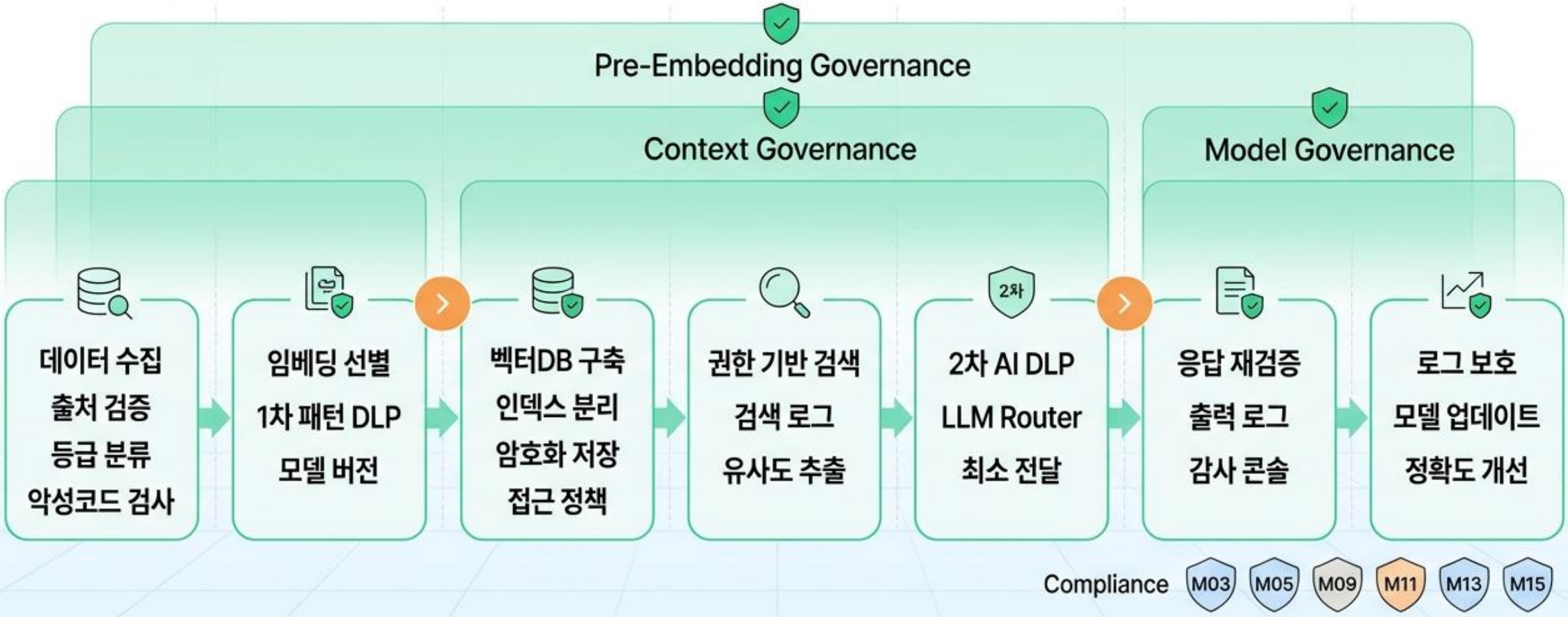
Lifecycle Control Comparison



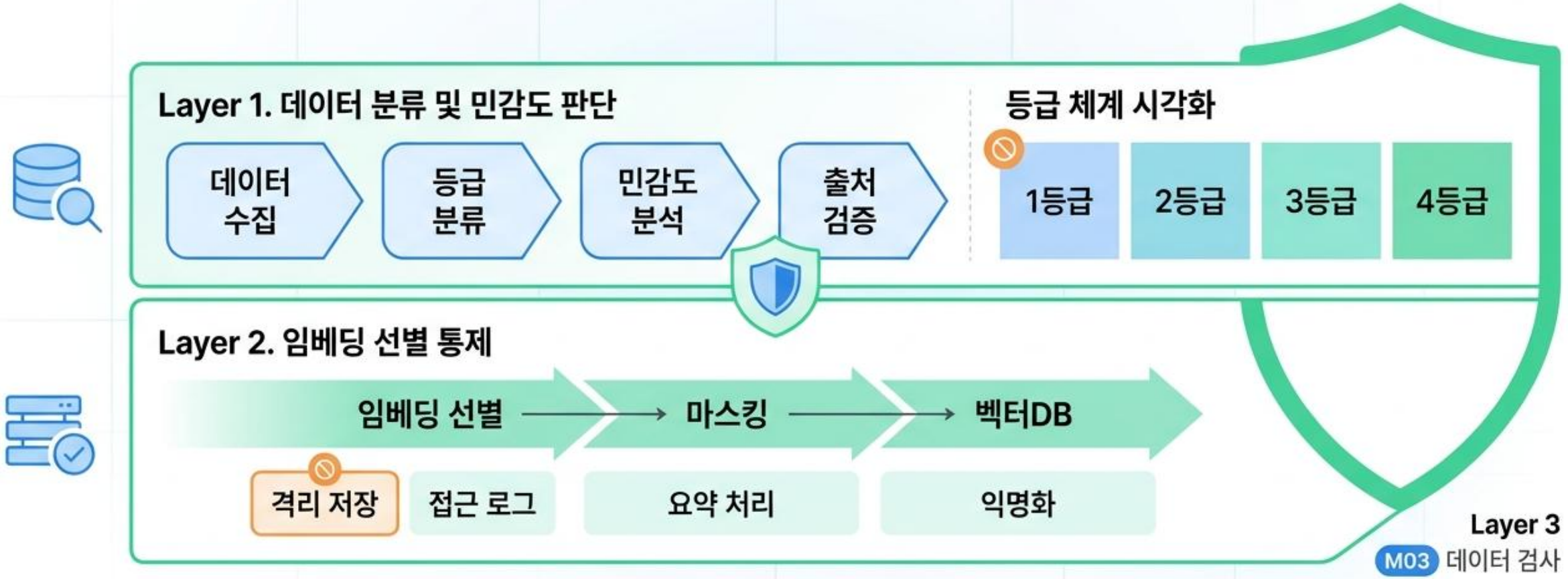
Secure RAG는 데이터 임베딩부터 LLM 응답까지 전 과정을 통제하는 공공 전용 AI 보안 아키텍처입니다.



Secure RAG는 데이터 수집부터 감사 단계까지 전 생애주기를 통제하는 공공 전용 AI 보안 아키텍처입니다.



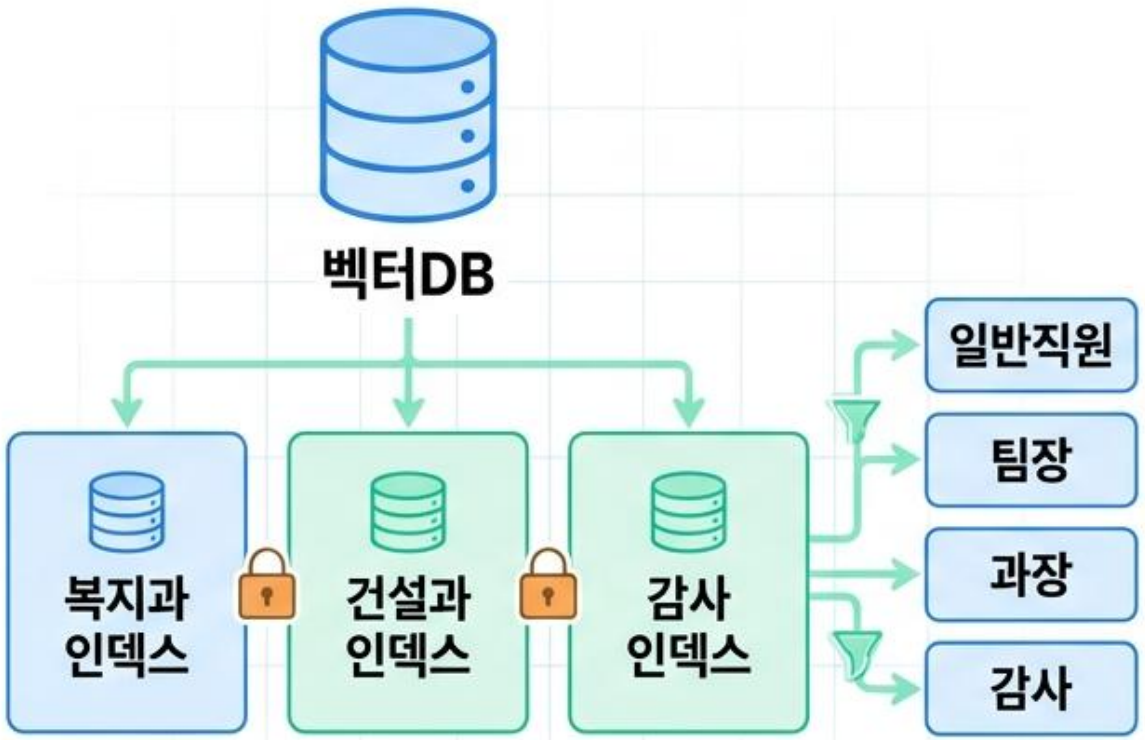
Secure RAG는 데이터 수집부터 임베딩 선별까지 통제하는 다계층 보안 아키텍처입니다.



벡터DB는 부서별 접근통제와 다계층 암호화를 적용한 프로덕션 보안 구조로 설계되어야 합니다.

1 부서·권한별 인덱스 분리

M05 접근통제



2 다계층 암호화 구조

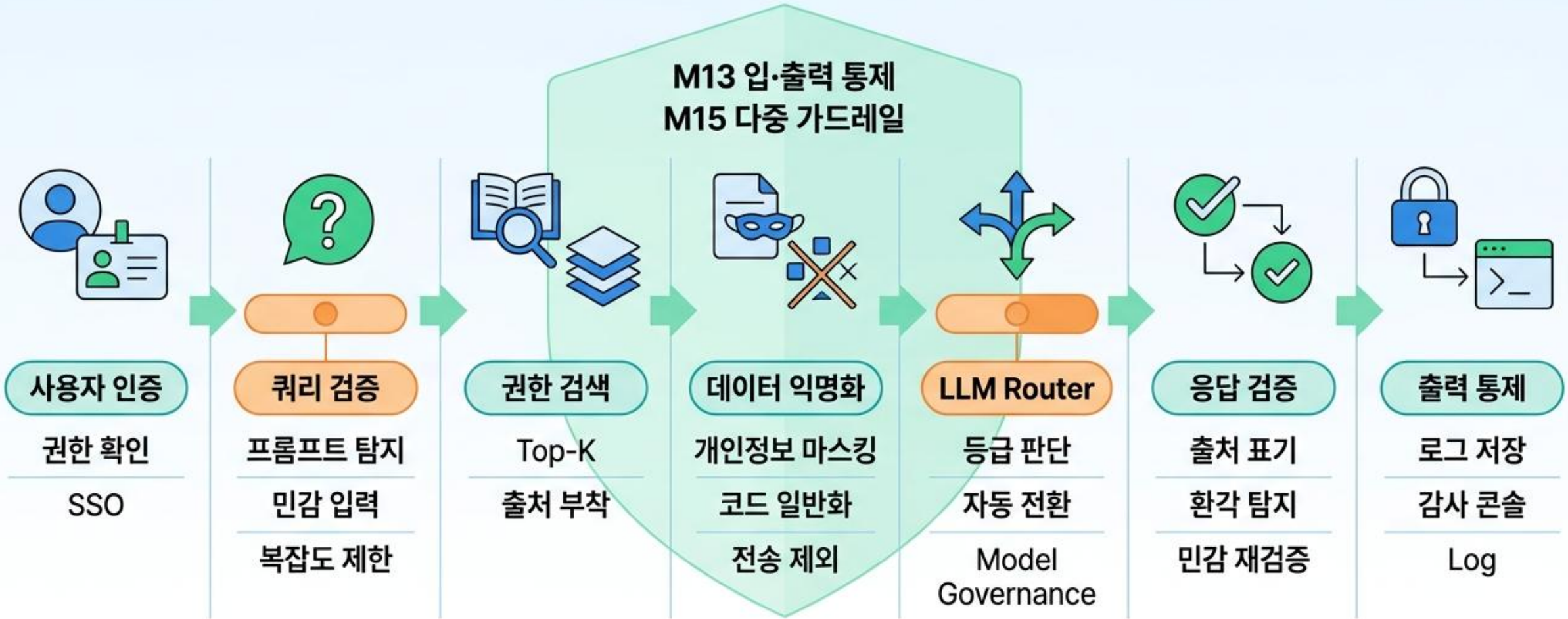


로그 및 감사 체계

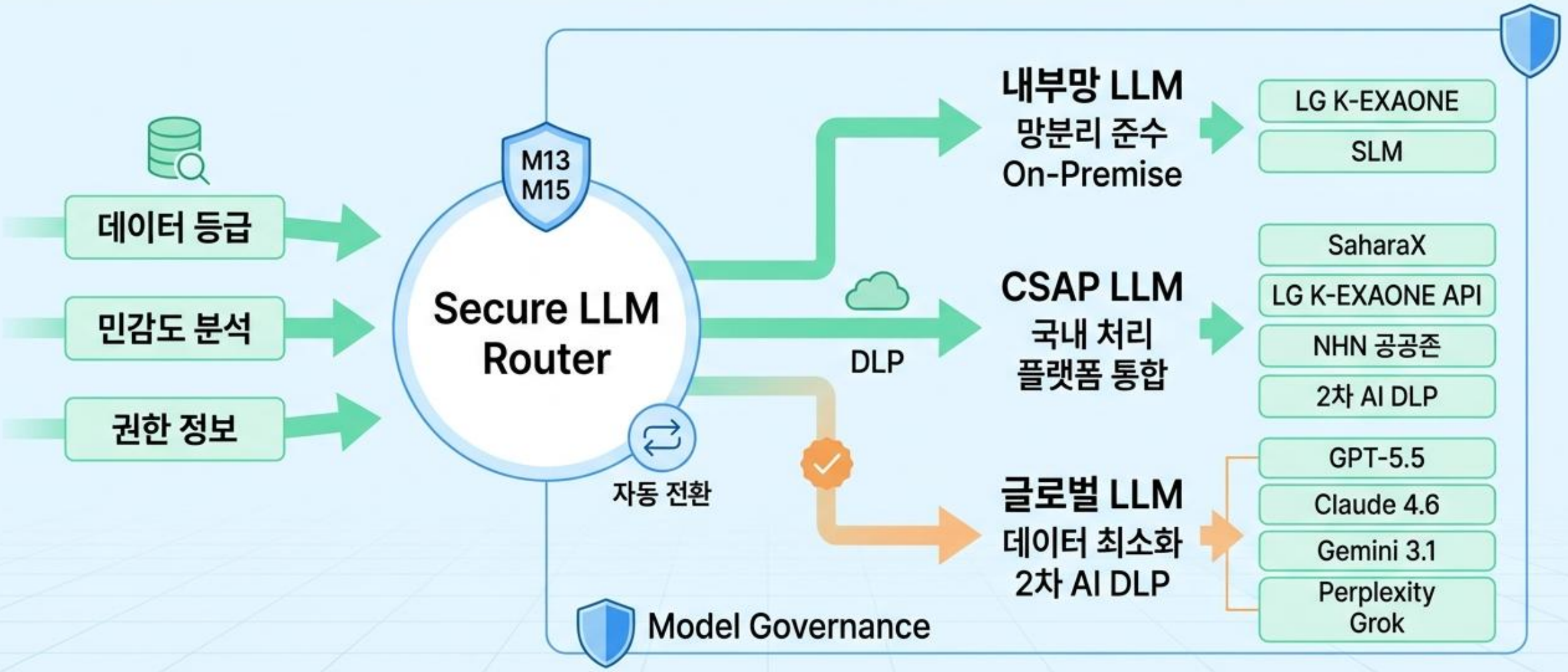
M09 로깅·모니터링



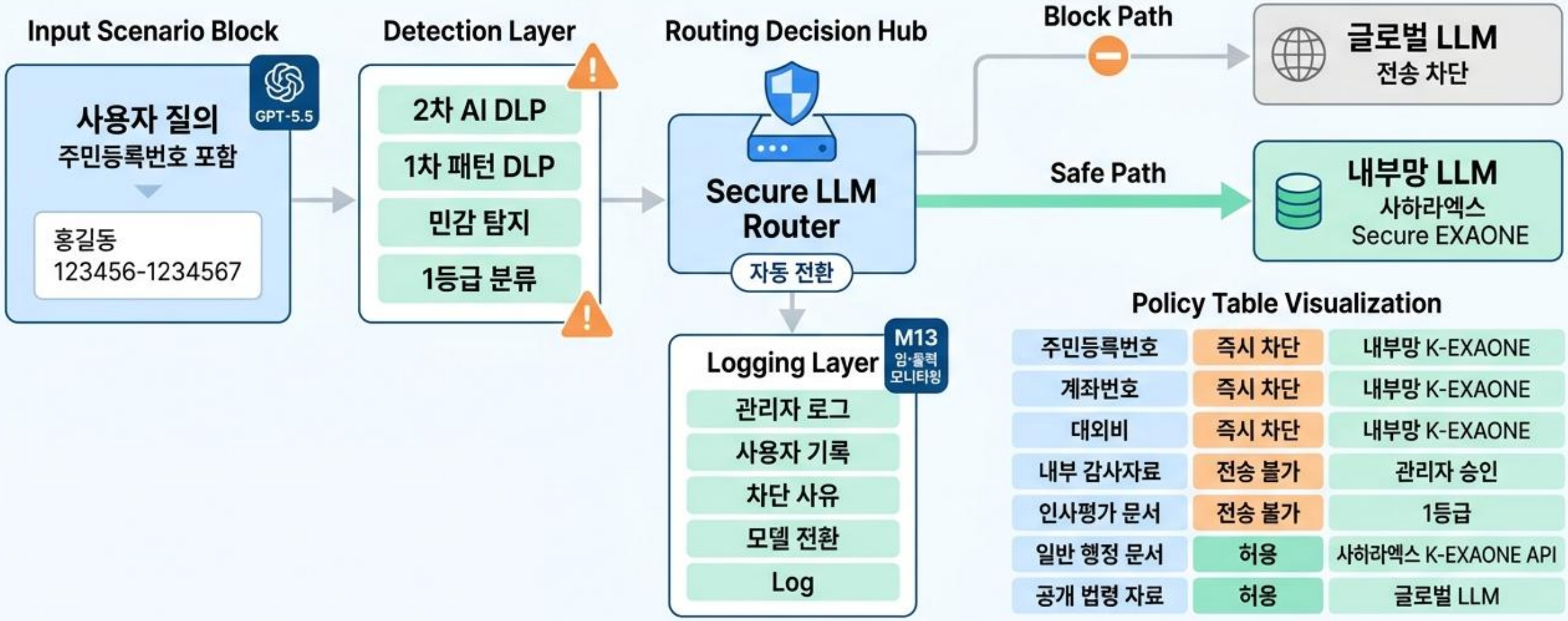
Secure RAG는 질의 입력부터 응답 출력까지 다중 가드레일 기반 7단계 보안 통제를 수행합니다.



Secure LLM Router는 데이터 민감도 기반으로 안전한 LLM 경로를 자동 선택하는 정책 강제 통제 구조입니다.



Secure LLM Router는 민감정보가 포함된 질의를 글로벌 LLM으로 전송하지 않고 안전한 경로로 자동 전환합니다.



Secure LLM Router는 정보 등급에 따라 모델 선택을 강제하는 정책 기반 통제 구조입니다.

권고	4등급	글로벌 LLM	경고 표시
강제 전환	2등급 3등급	내부망 LLM 사하라엑스 Secure EXAONE API	자동 전환
완전 차단	1등급	전송 불가	관리자 승인

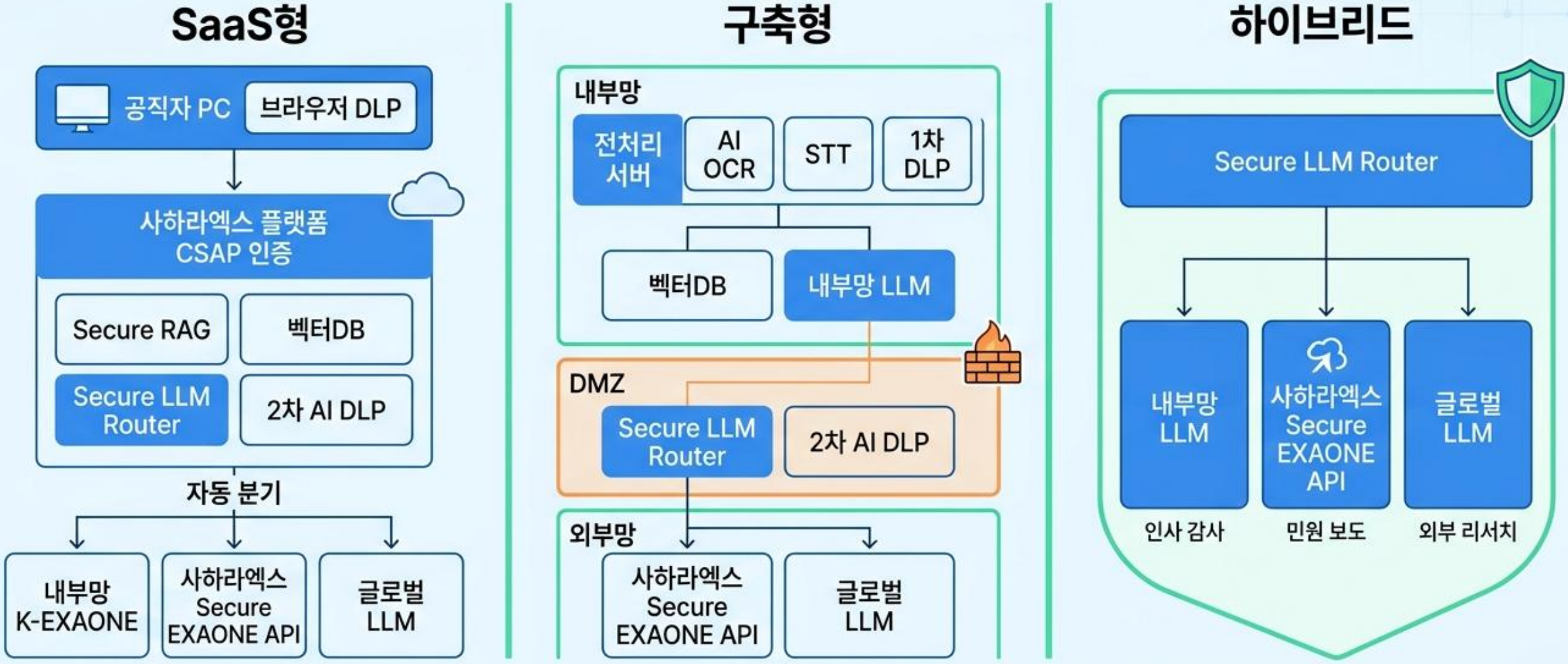
Governance

- **중앙 관리자** — 기관 정책
- **부서 관리자** — 예외 설정
- **일반 사용자** — 범위 내 선택

M15 다중 가드레일

M20 민감 명령 승인 절차

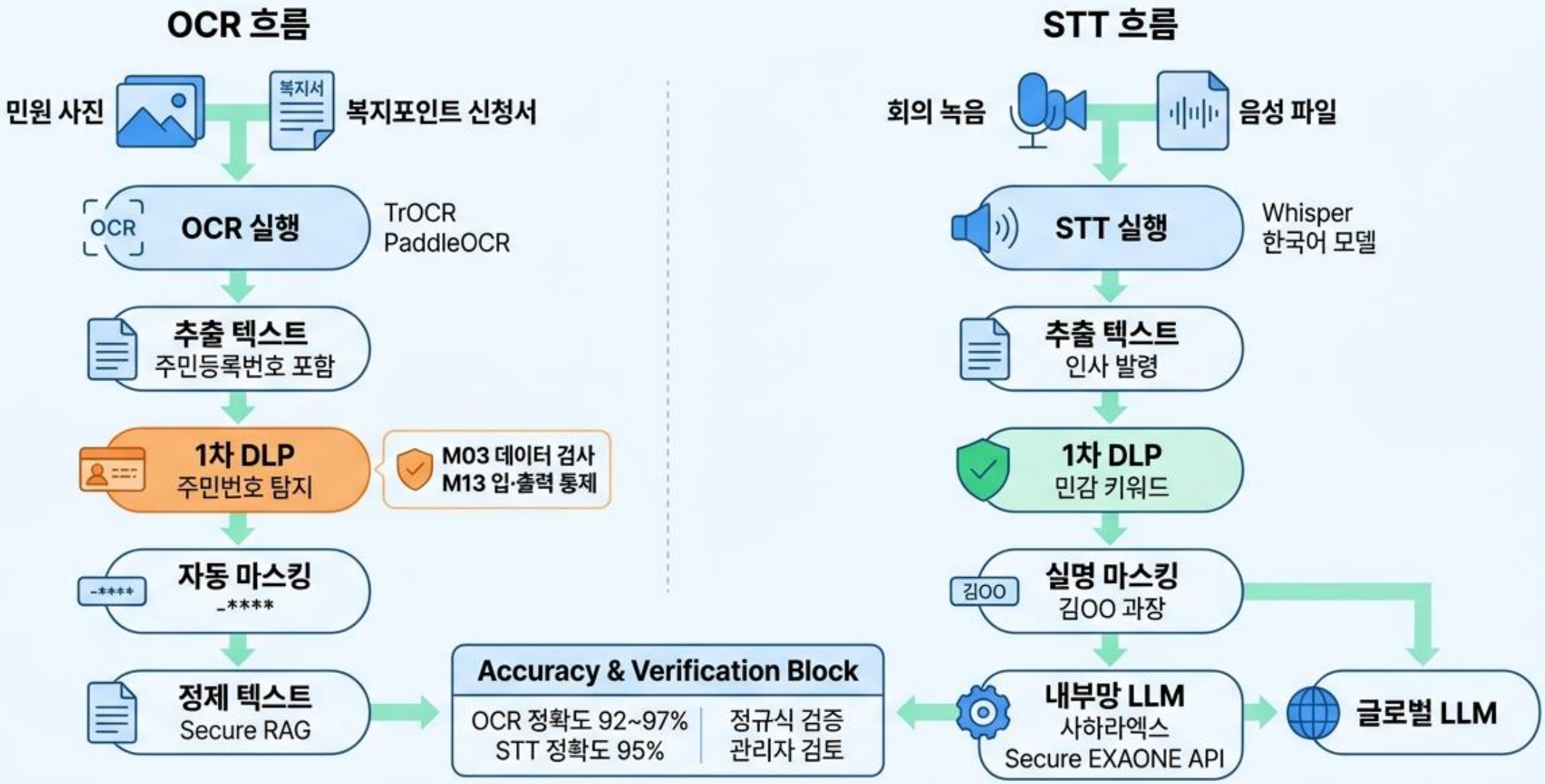
Secure LLM Router는 Saas형과 구축형을 모두 지원하여 기관 보안 수준에 맞는 배치 모델을 제공합니다.



비정형 데이터는 RTX 5090 기반 내부망 전처리를 통해 외부 반출 없이 안전하게 처리됩니다.



OCR 및 STT 처리 이후 DLP를 적용하여 민감정보를 차단한 텍스트만 AI로 전달합니다.



RTX 기반 내부망 전처리는 공격 표면을 최소화하여 비정형 데이터 위협을 구조적으로 차단합니다.

Attack Surface Reduction ASR

비정형 경로 차단

원본 차단 | VLM 차단 | 프롬프트 차단



내부망 격리

내부망 전처리 — RTX 서버 — 인터넷 차단 — API 없음

재검증 통제

OCR | STT | 1차 DLP | 재탐지 | 승인 통제

운영 보안 강화 블록

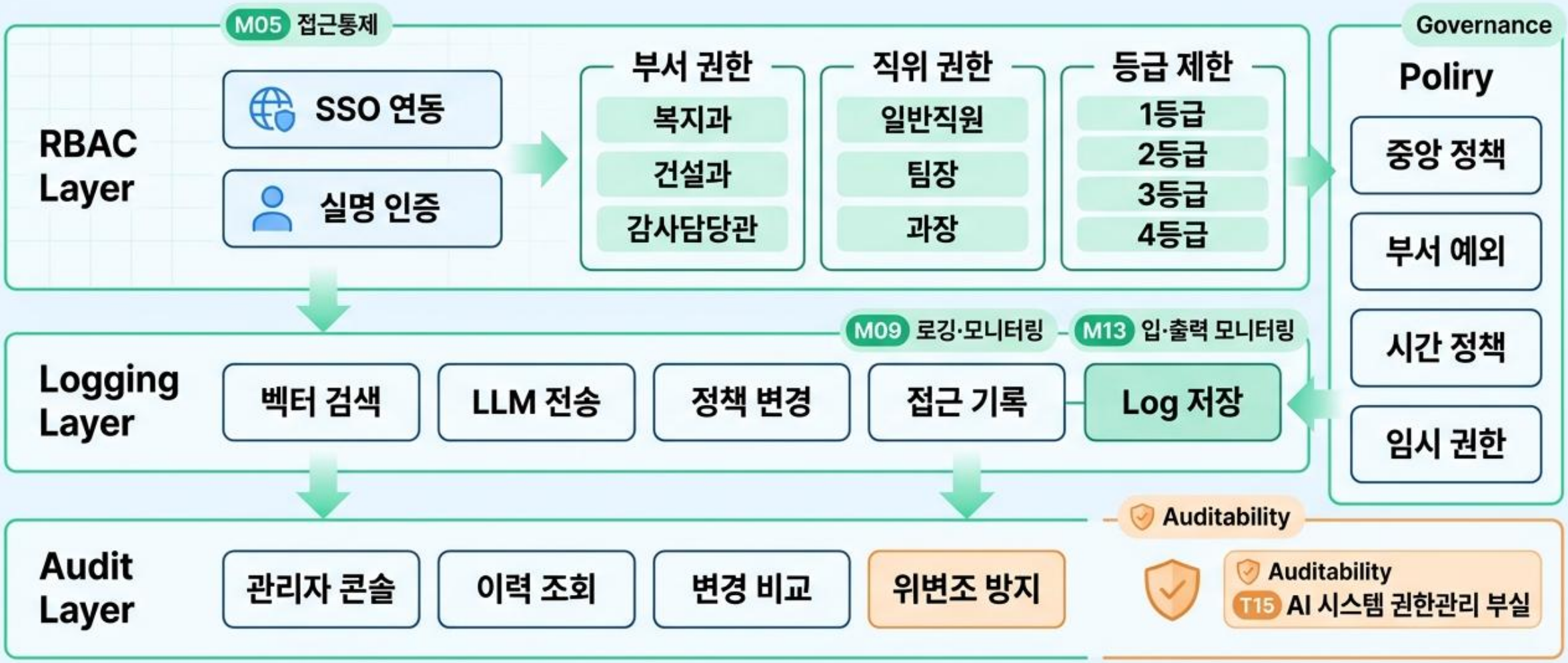
접근 통제 | MFA | 접근 로그 | 체크섬 검증 | 취약점 점검 | 패치 자동화 | 암호화 저장 | 자동 삭제 | 백업 통제

Threat 대응 매핑

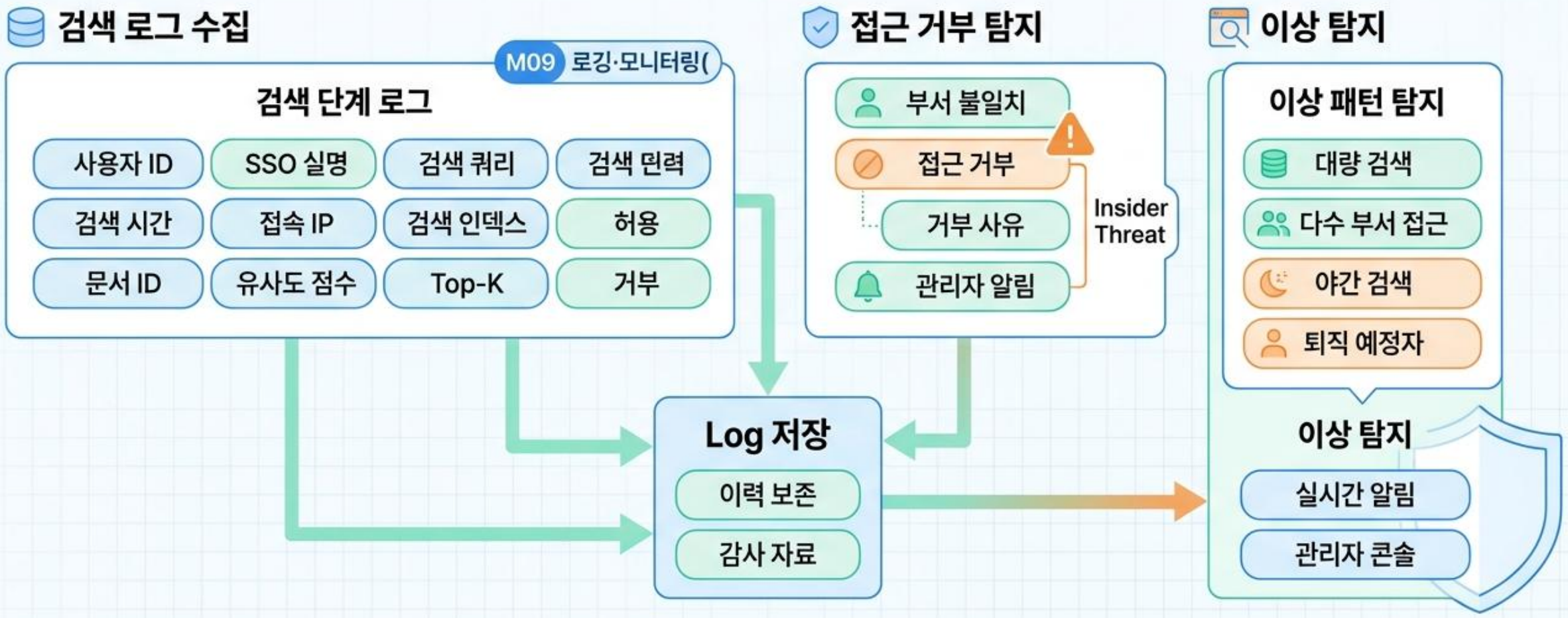
이미지 숨김	패턴 탐지
QR 유도	URL 검증
음성 민감정보	자동 마스킹
EXIF 유출	메타 삭제
스테가노그래피	원본 미전송

- M15 다중 가드레일
- M23, M24 적대적 입력 대응
- M25 취약점 점검

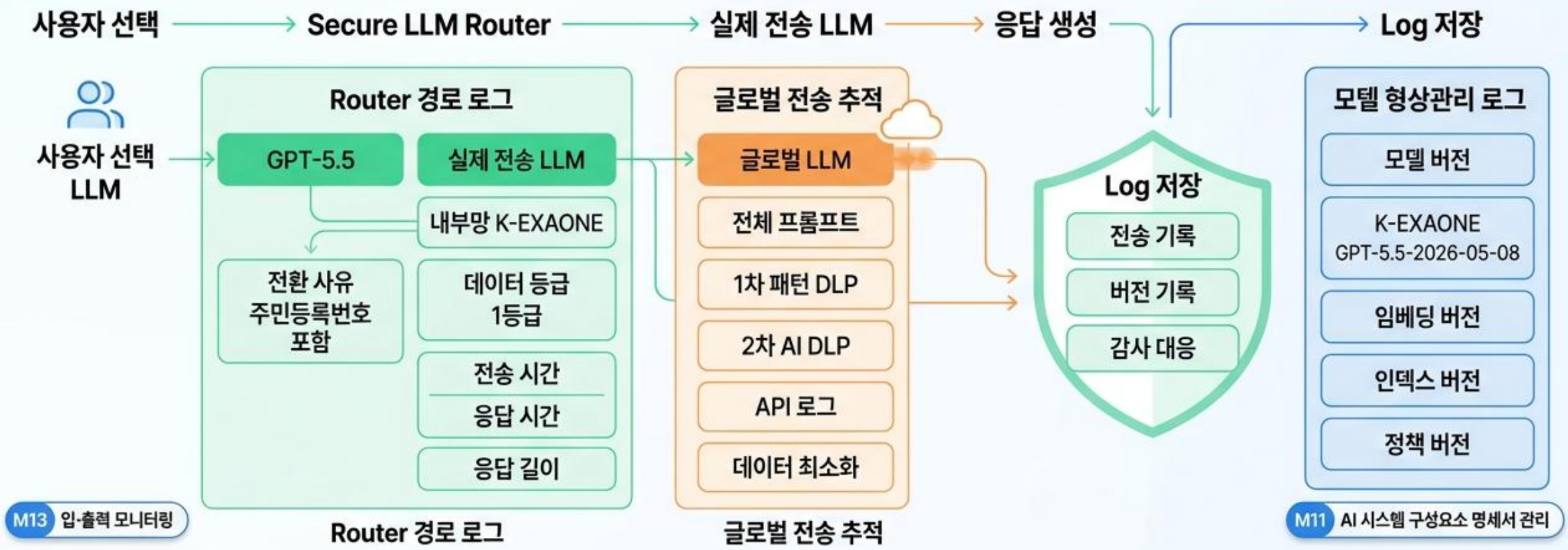
Secure RAG는 부서·직위 기반 접근통제와 전 구간 로그 기록을 통해 감사 가능성을 구조적으로 보장합니다.



벡터 검색 전 과정은 실명 기반 로그로 기록되어 감사 대응과 내부자 위협 탐지를 가능하게 합니다.

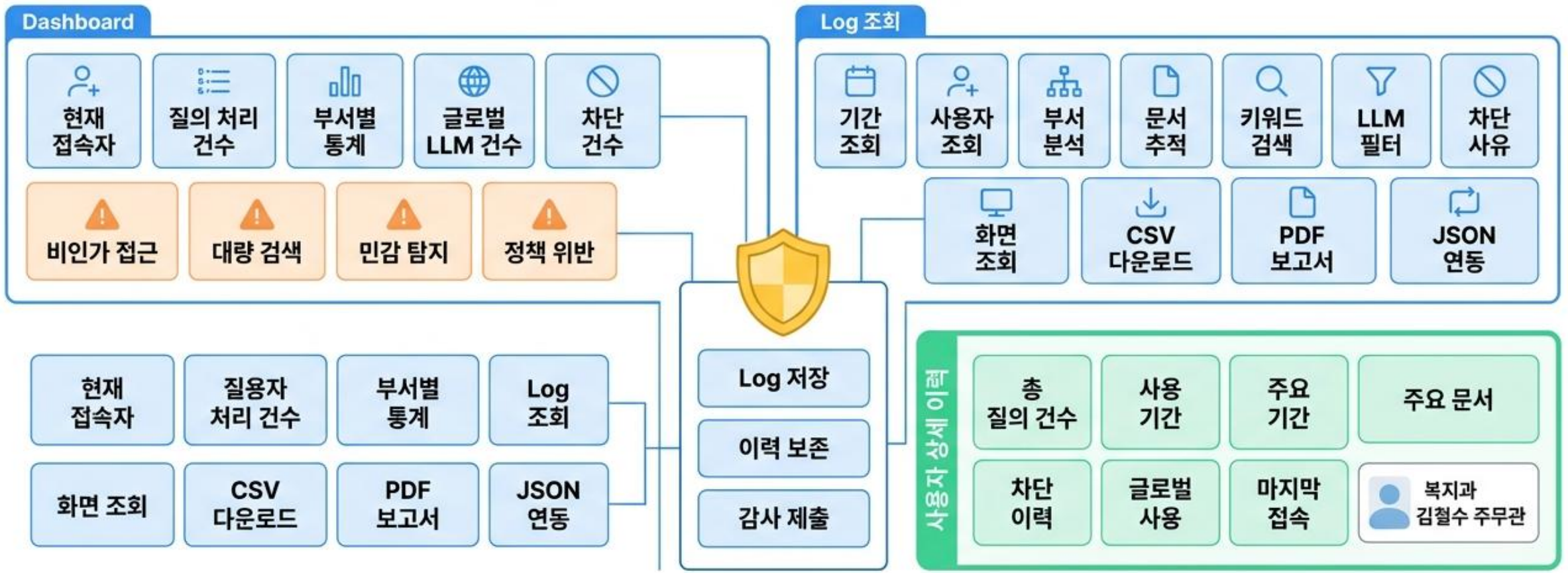


LLM 전송 전 과정은 모델 선택부터 응답 생성까지 모두 기록되어 사후 추적이 가능합니다.

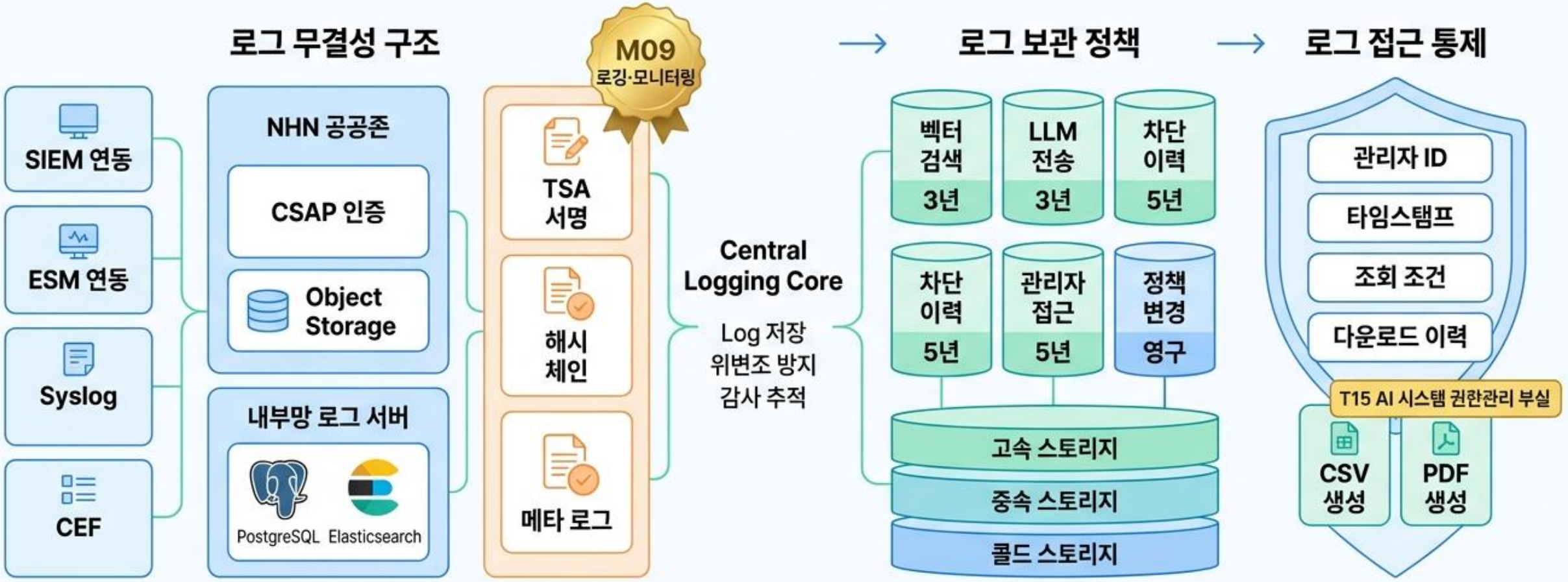


'관리자 감사 콘솔은 실시간 모니터링과 상세 이력 추적을 통해 감사 대응을 지원합니다.'

통합 대시보드 → 감사 조회 → 사용자 상세 이력



모든 로그는 위변조 방지 구조와 보관 정책에 따라 안전하게 저장되고 감사 추적이 가능합니다.



Secure RAG는 국정원 AI 보안 가이드 2.0은 핵심 통제 항목을 구조적으로 충합니다.

M03 데이터 검사

- 데이터 검사
- 민감 분류
- 키워드 탐지
- 악성코드 검사
- 출처 검증
- 격리 저장

완전 충족

M05 접근통제

- 접근통제
- SSO 연동
- 인덱스 분리
- 직위 제한
- 등급 제한
- 자동 차단

완전 충족

M09 로깅·모니터링

- 로깅 모니터링
- 검색 이력
- 전송 로그
- 차단 기록
- 감사 콘솔
- SIEM 연동
- 해시 체인

완전 충족

M13 입·출력 통제

- 입출력 통제
- 1차 패턴 DLP
- 2차 AI DLP
- 자동 마스킹
- 프롬프트 탐지
- 데이터 최소화

완전 충족

M15 다중 가드레일

- 다중 가드레일
- 브라우저 DLP
- 임베딩 선별
- 접근 통제
- Secure LLM Router
- 출력 검증
- 내부망 전처리
- 로그 기록

완전 충족

추가 대응

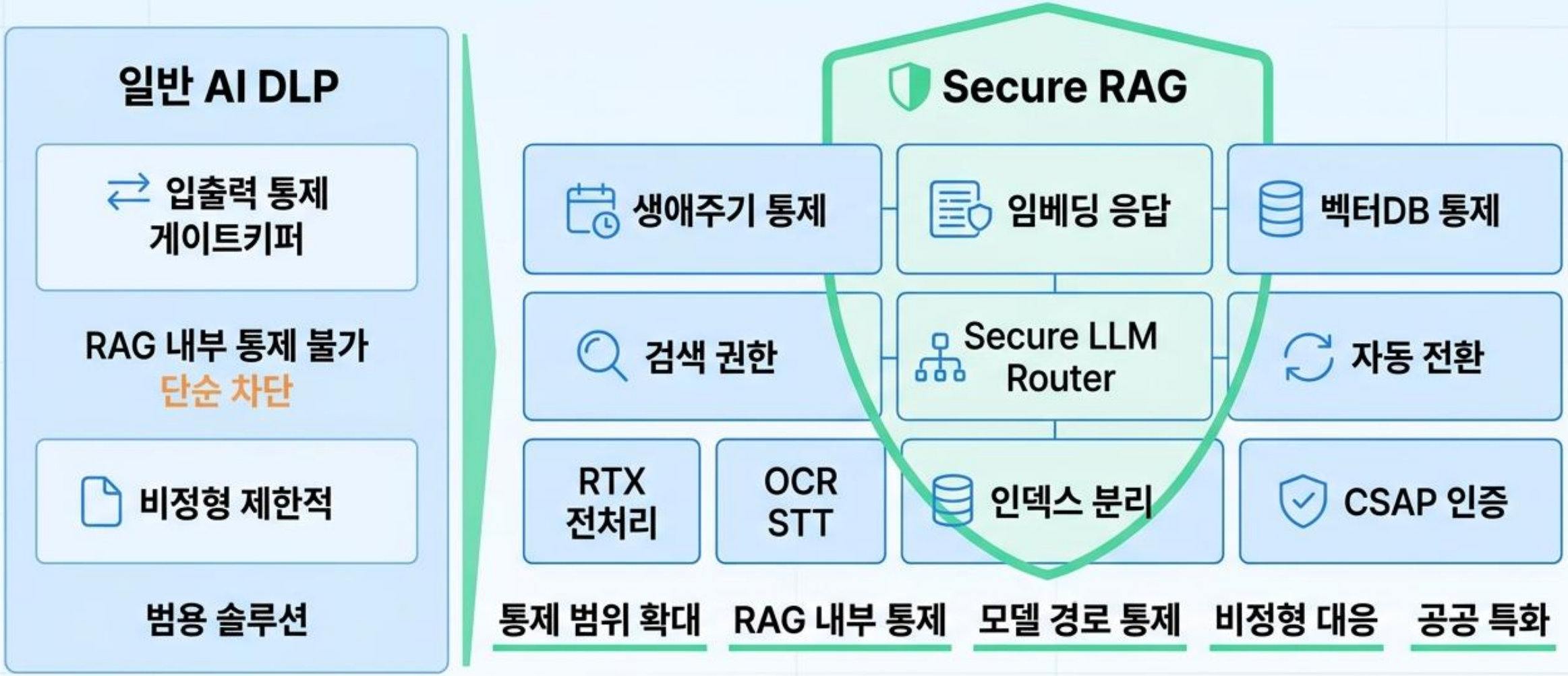
- M01 출처 검증
- M11 형상관리
- M20 관리자 승인
- M23 입력 대응
- M25 취약점 점검

2026 사이버보안 실태평가 가산점

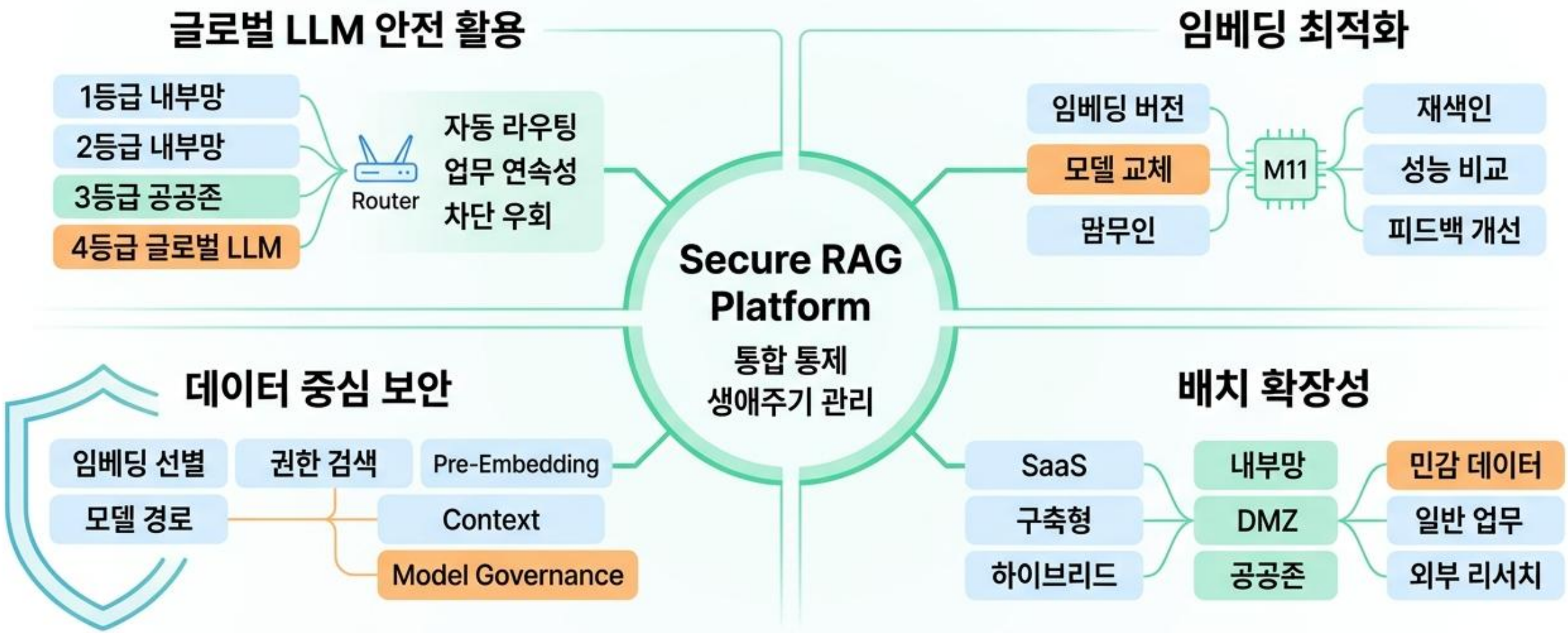
- AI 보안업무 도입
- 예산 확보
- N2SF 연계

가산점 확보

Secure RAG는 일반 AI DLP를 넘어 데이터 생애주기 전체를 통제하는 공공 특화 보안 구조입니다.



Secure RAG는 공공기관이 AI를 안전하게 활용할 수 있도록 확장 가능한 통합 통제 플랫폼입니다.



Secure RAG는 공공기관이 AI 시대에 안전하게 혁신할 수 있도록 지원하는 통합 통제 플랫폼입니다.



Strategy Block 1

범정부 AI 연계

공통 기반
통합 운영

Strategy Block 2

N2SF 대응

차등 보안
정책 정렬

Strategy Block 3

실태평가 가산점

AI 보안업무
예산 근거

Secure RAG

공공 전용 플랫폼

Strategy Block 5

Governance

공통 기반
통합 운영

Strategy Block 2

N2SF 대응

차등 보안
정책 정렬

Strategy Block 4

지자체 공유

광역 기초 통합
중복 방지

Strategy Block 5

민간 확산

금융·의료·교육
공공 검증

게이트키퍼 DLP | 입구 출구 통제 vs **Secure RAG** | 데이터 흐름 통제

본 제안은 국정원 및 행정안전부 가이드라인과 관련 보안 연구 문헌을 기반으로 설계되었습니다.

국가정보원 AI 보안 가이드북 2.0

M03 M05 M09 M11 M13 M15 M20 M23~24 M25

수명주기 위협 | LLM 선택 정책 | 내부망 구축 원칙

행정안전부 초거대 AI 가이드라인 2.0

데이터 등급 분류 | AI 도입 기준

데이터 최소화 | CSAP 인증

보안 가이드 및 법령

한국인터넷진흥원 생성형 AI 보안 가이드 | 개인정보보호법 시행령 제30조

외부 LLM 통제 | 접속기록 보관

보안 가이드 및 법령

한국인터넷진흥원 생성형 AI 보안 가이드 | 개인정보보호법 시행령 제30조

외부 LLM 통제 | 접속기록 보관

산업 및 언론 자료

CIO Korea | 데일리안

아이뉴스24 | 컴트루테크놀로지

CSAP 인증 | AI 보안 가산점

연구 문헌

RAG 보안 연구

말뭉치 오염 방지 | 임베딩 역전 공격

프롬프트 인젝션 | 출처 표기