The RAI Pipeline

엔터프라이즈 AI의 가치는 모델이 아니라 통제 계층에서 나온다

Han Kim · IOV Labs (아이오브연구소)

오픈 연구 · MIT · 2026-07-17

이해관계 명시. 저자는 한국형 기업 AI 게이트웨이 제품(RAI, `rai.kr`)을 운영하는 IOV 소속이다. 본 연구는 제품 광고가 아니라, 그 제품이 구현한 아키텍처 패턴을 형식화하고 그 비용 효과를 공개 가격으로 재현 가능하게 계산한 것이다. 비용 수치는 특정 고객의 실측이 아니라 2026 공개 list 가격 + 투명한 워크로드 가정에 기반한 **모델**이다.

초록. 엔터프라이즈 AI 도입 실패는 대개 모델 품질이 아니라 통제·통합의 부재에서 온다. 본 연구는 실무에서 검증된 5단계 처리 파이프라인(인증·권한 → DLP 마스킹 → 사내문서 RAG → 모델 라우팅·캐싱 → 감사·집계)을 하나의 합성 함수로 형식화하고, 그중 달러로 측정 가능한 계층(라우팅·캐싱)의 비용 효과를 2026 실가격으로 정량화한다. 대표적 엔터프라이즈 RAG 요청(입력 5,000토큰·출력 500토큰) 기준, 파이프라인은 요청당 비용을 $0.0225에서 $0.0077로 **65.9% 낮춘다**: 프롬프트 캐싱이 36.0%, 그 위에 모델 라우팅(요청의 30%만 프런티어)이 나머지를 더한다. 라우팅 단독 효과는 46.7%로, 실무에서 관측된 "단가 47%↓"와 일치한다. 절감은 두 레버(캐시 가능 입력 비중·프런티어 라우팅 비중)로 완전히 결정되며, 민감도 격자로 그 범위를 공개한다. 우리는 이 파이프라인을 "가치 = 안전 × 정확 × 경제성 × 통제가능성"의 곱으로 읽고, 어느 계층도 0이면 전체가 0임을 논한다. 마지막으로 이 공식이 입력측만 방어함을 지적하고, 출력 검증을 6번째 계층으로 제안한다.

문제: 모델이 아니라 통제가 도입을 가른다

엔터프라이즈 AI 경쟁은 모델 품질을 중심으로 벌어진다. 그러나 실제 도입의 병목은 다른 데 있다. IOV Labs의 선행 연구는 AI 도입 시 체감 생산성은 +20%지만 실측은 오히려 −19% 느려질 수 있음을 보고했다(METR 기반). 그 격차는 모델이 아니라 그 주변의 통제·검증 시스템의 부재에서 온다. 개인 계정으로 결제되는 "Shadow AI"는 기밀을 모델사로 흘리고, 비용은 분산되며, 답의 근거는 회사 진실이 아니라 모델의 일반 지식이다.

본 연구의 주장은 단순하다. **엔터프라이즈 AI의 가치는 모델이 아니라 그 모델을 감싼 처리 파이프라인에서 나온다.** "도구를 깐다"가 아니라 "업무에 접목한다"는 말의 실체가 이 파이프라인이다.

파이프라인의 형식화

RAI 파이프라인은 모든 요청을 다섯 계층으로 순차 처리한다(그림 1). 순서는 임의가 아니라 하중을 견디는 구조다.

$$\text{답변} = \text{감사}(\text{모델}(\text{라우팅·캐싱}(\text{RAG}(\text{DLP}(\text{인증·권한}(\text{요청}))))))$$

- ① **인증·권한.** 신원을 세션에서 도출한다(모델 출력이 아니라). 그래야 프롬프트 주입이 테넌트를 넘지 못한다. 반드시 첫 계층이다.
- ② **DLP 마스킹.** 주민번호·계좌 등 PII를 모델에 보내기 전에 가린다. 데이터 주권(PIPA 국외이전)의 요건이며, 모델 앞이어야 한다.
- ③ **사내문서 RAG.** ERP·거래명세·정책을 벡터 검색해 답의 근거를 회사 진실에 고정한다. 환각을 줄인다.
- ④ **모델 라우팅·캐싱.** 민감·고난도는 프런티어(Claude), 대량·일상은 저가 모델로 라우팅하고, 반복되는 시스템 프리픽스는 캐싱한다. 비용의 계층이다.
- ⑤ **감사·집계.** 모든 질의·응답·비용을 기록하고 부서별로 집계한다. "도입되는 AI"와 "Shadow AI"를 가르는 계층이다.

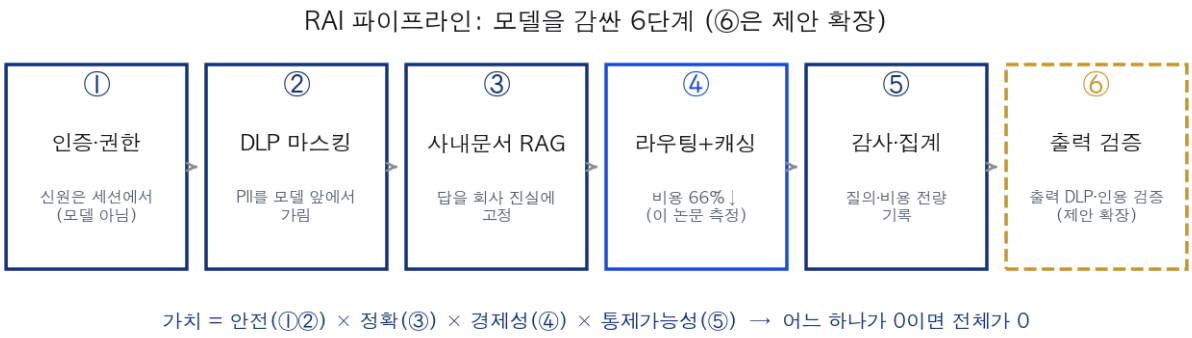

그림 1: RAI 파이프라인: 모델을 감싼 5+1 계층. ⑥ 출력 검증은 본 연구가 제안하는 확장이다.

가치는 곱으로 읽힌다: **가치 = 안전(①②) × 정확(③) × 경제성(④) × 통제가능성(⑤)**. 어느 계층이 0이면(예: 감사 없음 → 통제가능성 0) 전체 가치가 0이다. 이것이 다섯 계층이 모두 필요한 이유이며, 단일 기능 카피가 이 파이프라인을 복제하지 못하는 이유다.

측정: 비용 계층(④)의 정량화

다섯 계층 중 달러로 곧장 측정되는 것은 ④다. 여기에 초점을 맞춰 수치를 낸다.

입력값

모든 가격은 2026 공개 list다(그림 2). Claude Sonnet 4.6(프런티어) $3/$15 per Mtok, 캐시 읽기는 입력의 10%($0.30). Claude Haiku 4.5(대량) $1/$5. 프롬프트 캐싱은 반복되는 시스템 프리픽스의 입력 토큰을 90% 할인한다. 대표 요청은 캐시 가능 시스템 프리픽스 3,000토큰 + 요청별 가변(질의·RAG 청크) 2,000토큰 + 출력 500토큰, 프런티어 라우팅 30%로 잡는다.

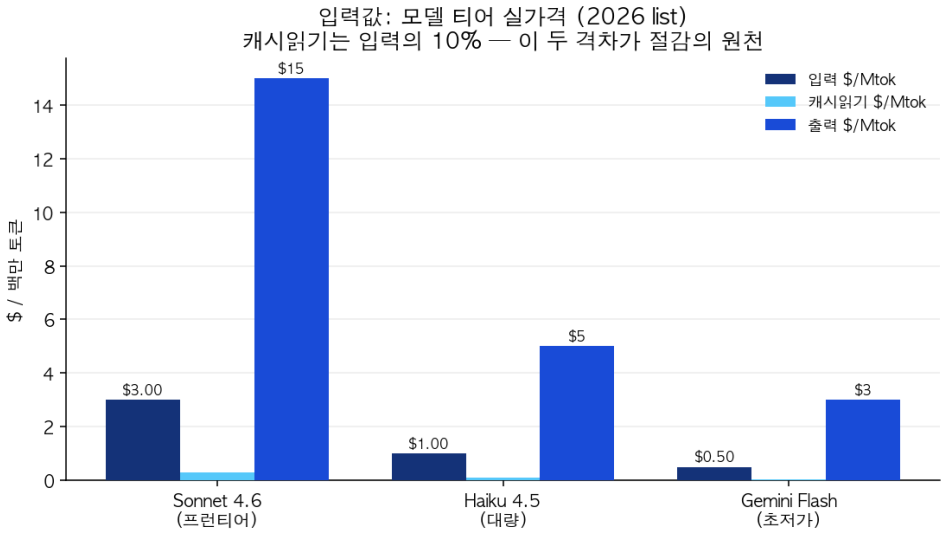

그림 2: 모델 티어 실가격(2026). 프런티어-저가 격차와 캐시읽기 10%가 절감의 두 원천이다.

결과: 요청당 −65.9%

기준(전량 프런티어·무캐시)은 요청당 $0.0225다. 프롬프트 캐싱만으로 $0.0144(−36.0%), 그 위에 모델 라우팅을 더하면 $0.0077(**−65.9%**)이 된다(그림 3). 라우팅 단독 효과는 −46.7%로, 실무에서 관측된 "단가 47%↓"와 사실상 일치한다. 즉 실무의 "최대 60%↓" 주장은 과장이 아니라, 오히려 대표 가정에서는 보수적이다.

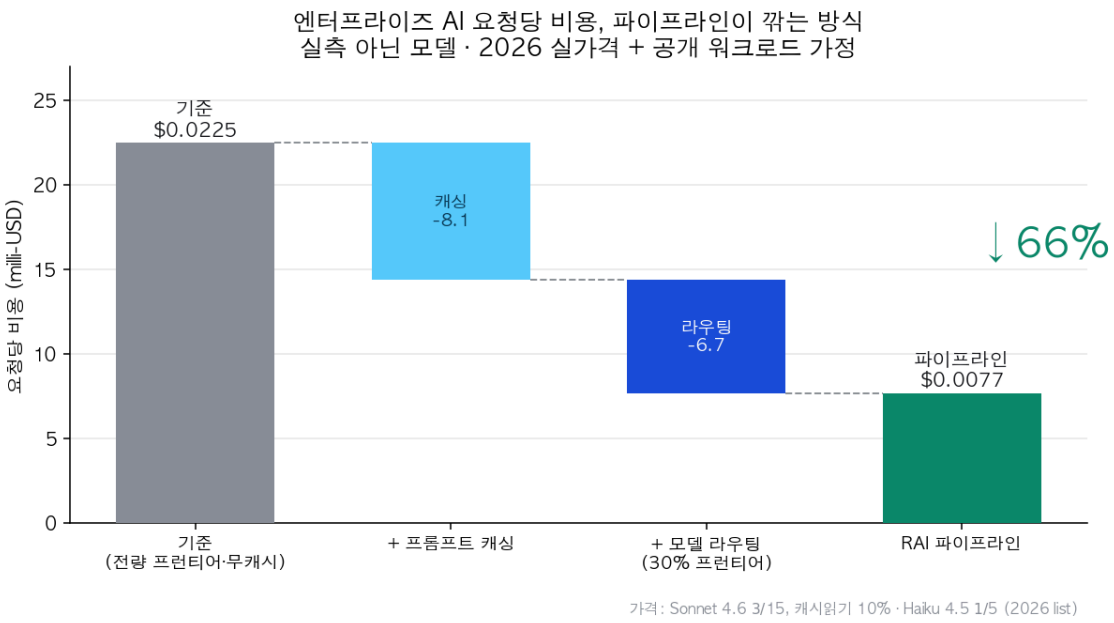

그림 3: 요청당 비용 워터폴. 캐싱과 라우팅이 순차로 비용을 깎는다.

절감은 두 레버로 결정된다

절감률은 오직 두 변수, 캐시 가능 입력 비중과 프런티어 라우팅 비중에 의존한다(그림 4, 그림 5). 시스템 프롬프트가 반복될수록(캐시 비중↑) 절감이 커지고, 프런티어에 덜 보낼수록(라우팅 믹스↓) 절감이 커진다. 이 둘만 알면 절감률이 완전히 정해진다. 그래서 우리는 헤드라인 하나가 아니라 격자 전체를 공개한다.

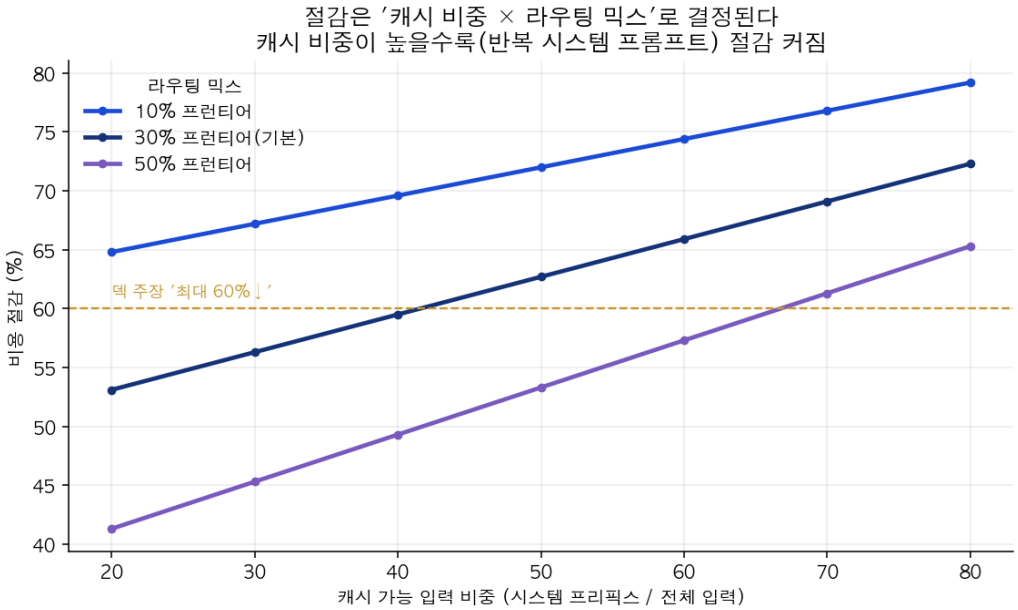

그림 4: 절감 vs 캐시 비중, 라우팅 믹스별. 점선은 실무 주장 “60%↓”.

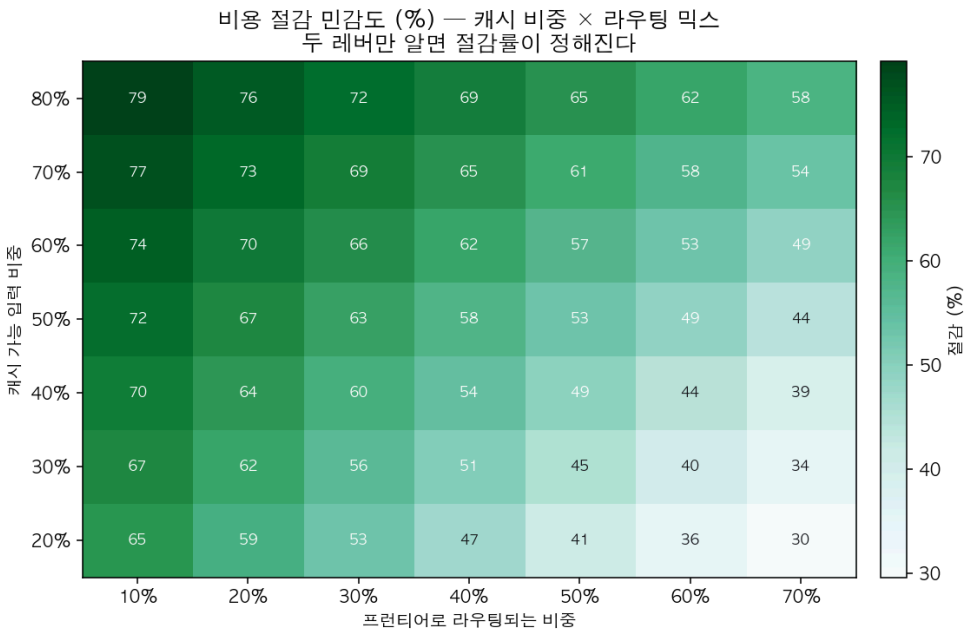

그림 5: 민감도 격자(%). 캐시 비중 × 라우팅 믹스.

조직 규모 효과

요청량이 클수록 절감액은 선형으로 커진다(그림 6). 월 10만 요청 조직은 기준 대비 월 약 $1,482, 연 약 $17,800를 절감한다. 이는 API 비용만이며, 계층 ②⑤가 주는 규제·거버넌스 가치(조달 통과·기밀 보호)는 별도다.

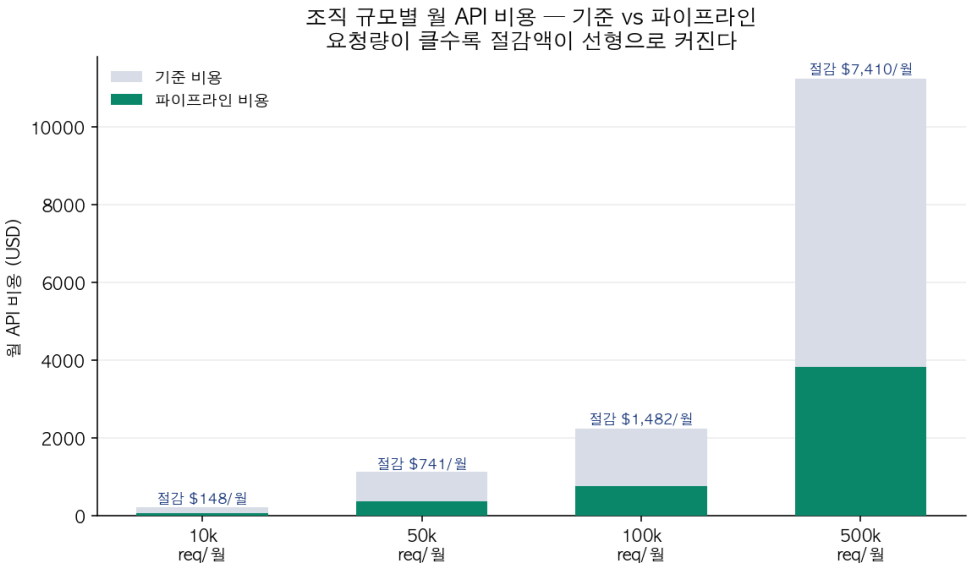

그림 6: 조직 규모별 월 API 비용. 기준 vs 파이프라인.

한계와 확장: 출력측이 비어 있다

이 파이프라인은 강력하지만 **입력측 방어**에 치우쳐 있다. 세 가지를 분명히 한다.

첫째, 출력 검증이 없다. ②의 DLP는 입력만 가린다. 모델이 컨텍스트의 PII를 출력으로 흘리거나 ③의 근거를 날조하면 지금 구조는 막지 못한다. 우리는 **⑥ 출력 검증**(출력 DLP + 인용 충실도 검증)을 6번째 계층으로 제안한다. 이는 IOV Labs의 자기선호·완료착시 연구가 공통으로 가리키는 결론과 같다: 모델의 자기보고를 신뢰하지 말고 시스템이 검증하라.

둘째, ②↔③ 순서 긴장. 질의를 RAG 앞에서 마스킹하면 검색 품질이 떨어지고, 검색 뒤에 마스킹하면 사내문서의 PII까지 잡힌다. 최적은 "질의 마스킹 → RAG → 검색결과 재마스킹"의 이중 구조일 가능성이 크며, 이는 실측으로 정할 문제다.

셋째, 에이전트 작업엔 완료 검증이 더 필요하다. 파이프라인은 질의응답엔 완결이지만, AI가 작업을 **실행**하면(에이전트) 완료 착시가 발생한다. 그땐 완료를 게이팅하는 검증 계층이 별도로 붙어야 한다.

다음 단계: 모델을 실측으로

비용 모델의 두 가정(캐시 비중·라우팅 믹스)은 RAI의 감사 로그에서 곧장 실측할 수 있다. 실제 요청의 캐시 적중률과 티어 분포를 한 달 집계하면, 본 모델의 65.9%가 그 회사의 실제 절감률로 대체된다. 우리는 이 후속 측정을 공개할 계획이다.

재현. 모델과 그림은 `model.py`, `viz.py`, `gifs.py`로 완전히 재현된다. 입력값(가격·워크로드)을 바꾸면 결론이 바뀐다. 코드·데이터: `github.com/hankimis/rai-pipeline`.

출처. 가격: TLDL LLM Pricing · Anthropic API Pricing · IntuitionLabs(2026). 도입 격차: IOV Labs Enterprise AI Playbook(METR RCT 인용). 자기보고 검증: IOV Labs "The Completion Illusion", "The Judge in the Mirror".