Deloitte Insights

November 2026



AI 토크노믹스

토큰 기반의 TCO 분석과 AI 비용 최적화 전략

Deloitte Insights

AI-generated Image

Deloitte.

Leader's Message



박경순 상무

Forensic & Resilience 본부
경영자문부문 | 한국 딜로이트 그룹

AI 경쟁력은 이제 '얼마나 쓰는가'보다 '어떻게 관리하는가'에 달려 있습니다.

AI 확산과 함께 기업의 AI 지출도 빠르게 늘고 있습니다. 상당수 기업이 예산 초과를 겪고 있지만, AI 투자가 측정 가능한 가치를 창출한다고 답한 재무 리더는 절반에 못 미칩니다. AI 비용은 이제 경영진이 예측·통제하고 투자 성과를 검증해야 할 핵심 경영 과제입니다.

본 리포트는 AI 비용 관리의 출발점으로 '토큰'에 주목합니다.
토큰은 AI 사용량을 측정하고 비용을 분석하는 공통 기준입니다.
딜로이트 분석 결과, 세 가지 인프라 방식(자체 인프라, AI 전용 클라우드, 외부 서비스)의 비용 우위는 AI 사용량에 따라 달라집니다.
따라서 **향후 사용량을 고려한 인프라 선택과 AI 사용 방식의 최적화, 예산·사용 한도·승인 체계를 통한 지속적인 비용 관리가 필요**합니다.

AI 비용을 체계적으로 관리하고 투자 성과를 높이기 위한
경영진의 판단과 실행에 도움이 되기를 바랍니다.

AI 토크노믹스 서비스 소개 →

AI 토크노믹스 리포트 →

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

왜 AI 비용을 토큰으로 관리해야 하는가?

⚠ AI 지출 확대에 따른 비용 관리의 어려움



AI 지출 급증

2026년 글로벌 클라우드 인프라 지출 전년 대비 **27% 증가** 예상



비용 예측의 불확실성 증가

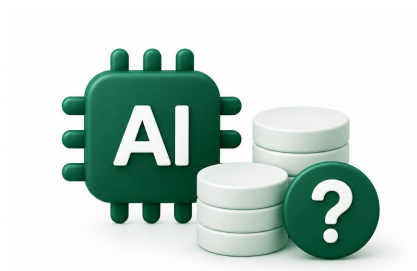
응답 기업의 **93%**가 AI 예산 대비 지출 초과 경험



투자 성과 검증 불투명

AI 투자가 측정 가능한 가치를 창출한다고 답한 재무 리더는 **28%에 불과**

⚠ 현재 AI 비용 관리의 한계



비용 예측 한계

AI 사용량이 늘어날수록 향후 비용 규모 산정에 어려움



성과 검증의 한계

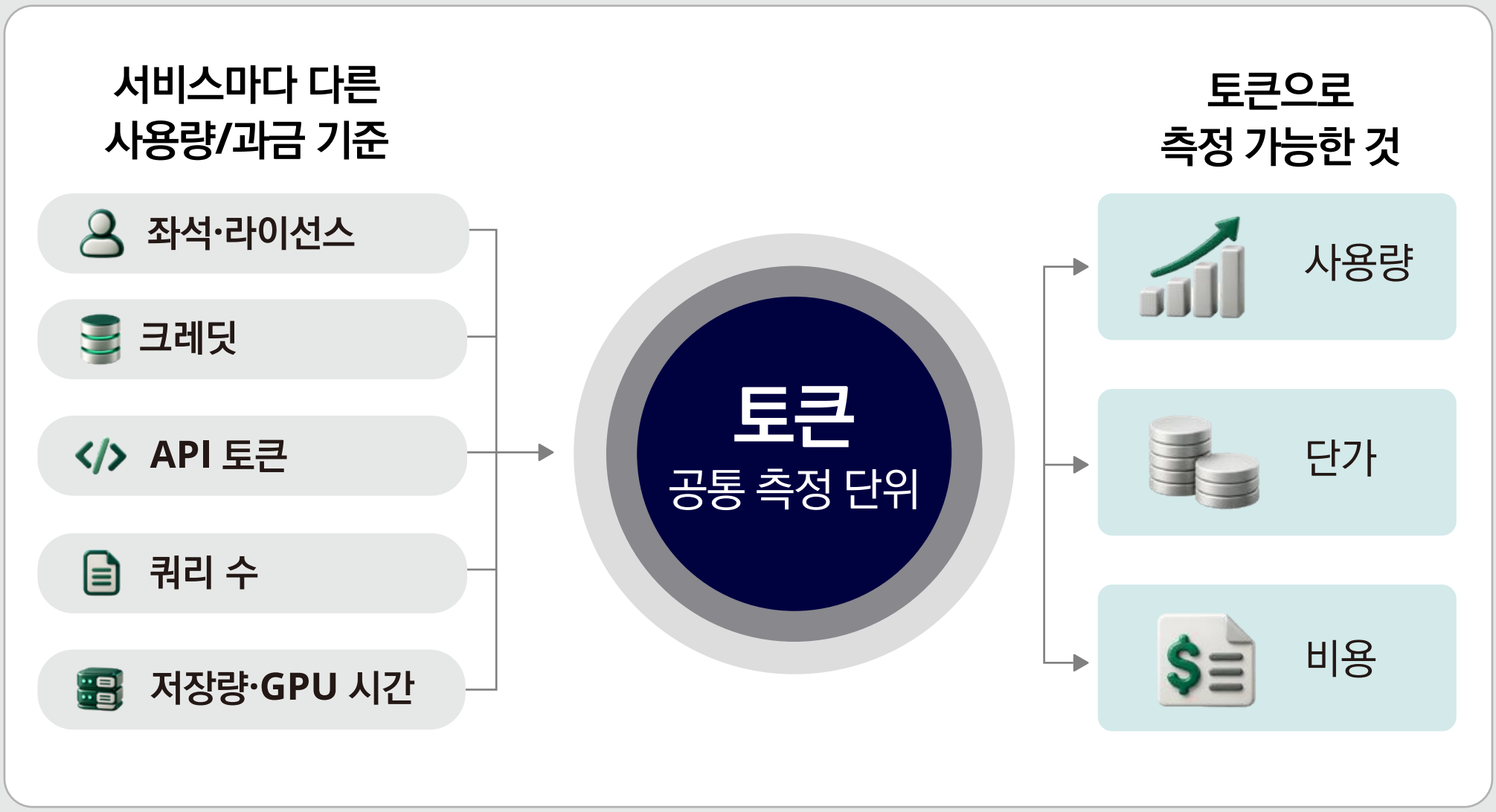
모델·서비스별 효율과 투입 비용 대비 실제 업무 성과를 비교·평가하기 어려움

서로 다른 AI 사용량과 비용을 일관된 기준으로 비교 관리할 공통 기준이 필요

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

AI 사용량과 비용을 연결하는 공통 측정 단위, 토큰



토큰 기반 측정의 효과

가시성 확보	조직·업무별 AI 비용을 파악하고 지출 원인 분석
통제 강화· 비용 누수 차단	예산 초과를 사전 관리하고 불필요한 AI 지출 최소화
관리 책임 명확화	AI 비용 관리의 책임 주체와 승인·통제 절차 정립

토큰 기반 측정을 통해 AI 비용의 가시성과 통제력을 높이고, 관리 책임을 명확화

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

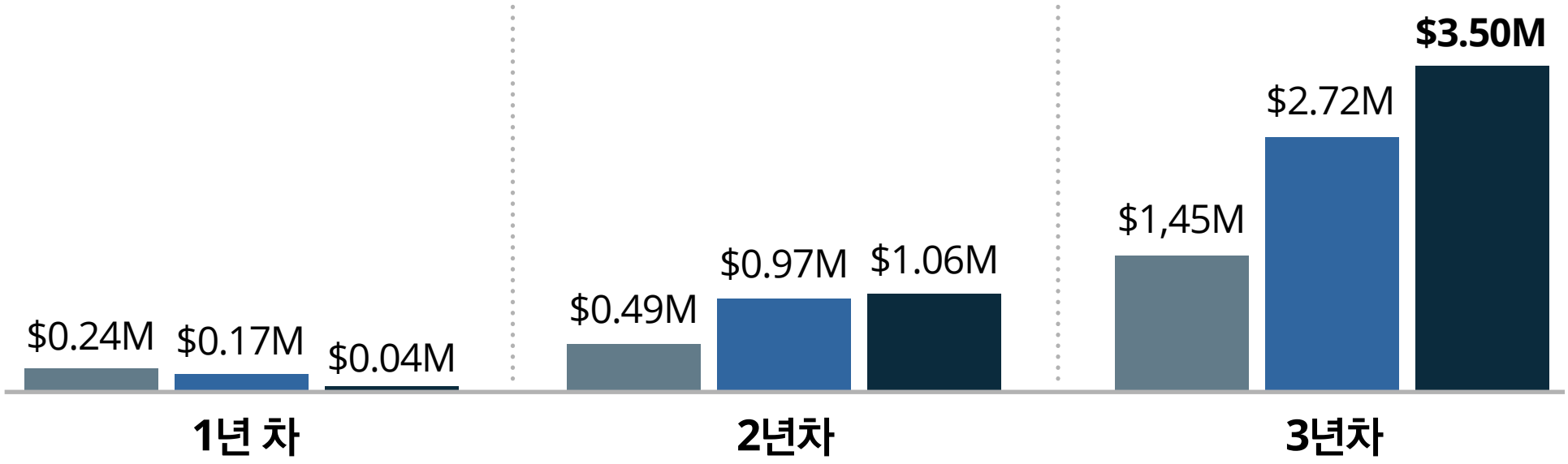
AI 사용량에 따른 인프라 비용 우위 변화

AI 인프라별 비용 구조와 경제성 비교

1년 차 파일럿
100억 토큰

2년 차 PoC
3,000억 토큰

3년 차 대규모 추론
1조 토큰



3년 누적 TCO

API	NCP	AI 팩토리
$4.60M	$3.86M	$2.18M

API 비용은 AI 팩토리의 **2.1배**

AI 사용량이 증가할수록 비용 우위가 API에서 AI 팩토리로 이동

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

AI 비용 최적화 방안 ①

AI 비용 절감을 위한 사용 방식 최적화



업무 난이도와 처리 방식에 맞는 모델·컴퓨팅 자원 선택

소형·파인튜닝 모델 활용, 분류·이상탐지는 중급 GPU·CPU,
고성능 모델은 필요한 업무에만(멀티모델)



불필요한 생성형 AI 호출과 입력 정보 최소화

의사결정 트리·규칙 기반 엔진, 컨텍스트 윈도우 상한



불필요한 연산과 토큰 처리의 조기 차단

조기 종료, 프롬프트 절단, 압축 트랜스포머



반복 처리와 장시간 추론을 줄이는 구조 설계

RAG, 프롬프트 엔지니어링, 응답 캐싱, 배치 처리,
결정론적·확률적 모델 병행



모델 크기와 연산량을 줄여 처리 효율 개선

양자화, 프루닝, 지식 증류, 전이학습

**"토큰 단가가 내려가도 사용량이 늘면 총지출은 커지므로
업무에 필요한 만큼의 모델·인프라·추론만 쓰도록 사용 방식을 설계"**

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

AI 비용 최적화 방안 ②

AI 비용의 지속적 통제를 위한 관리체계 구축

측정

통합 모니터링

GPU 사용시간·스토리지·데이터 반출·
토큰 사용량을 한 곳에서 실시간 파악



배분

책임 귀속·내부 청구

조직·업무·프로젝트 태그를 활용해
사업부별 비용을 배분하고
그림자 AI 비용 누적 차단

예시 관리대상 비용 **1,000만원**

직접 귀속	700만원
공통비 배부	200만원
미귀속 별도관리	100만원

통제

예산·사용 한도

예산 알림, 컨텍스트 윈도우 제한,
API 사용량 상한으로 과도한 사용 방지

예시

- **80% 소진** 알림
- **초과 전망** 조정안 승인
- **한도 도달** 제한 또는 승인된 예외

개선

재무·기술·사업 공동 관리

예산·토큰 소요와 ROI 기준으로 프로젝트를 승인하고,
미사용 계정·중복 서비스·약정 조건을 정기적으로 조정



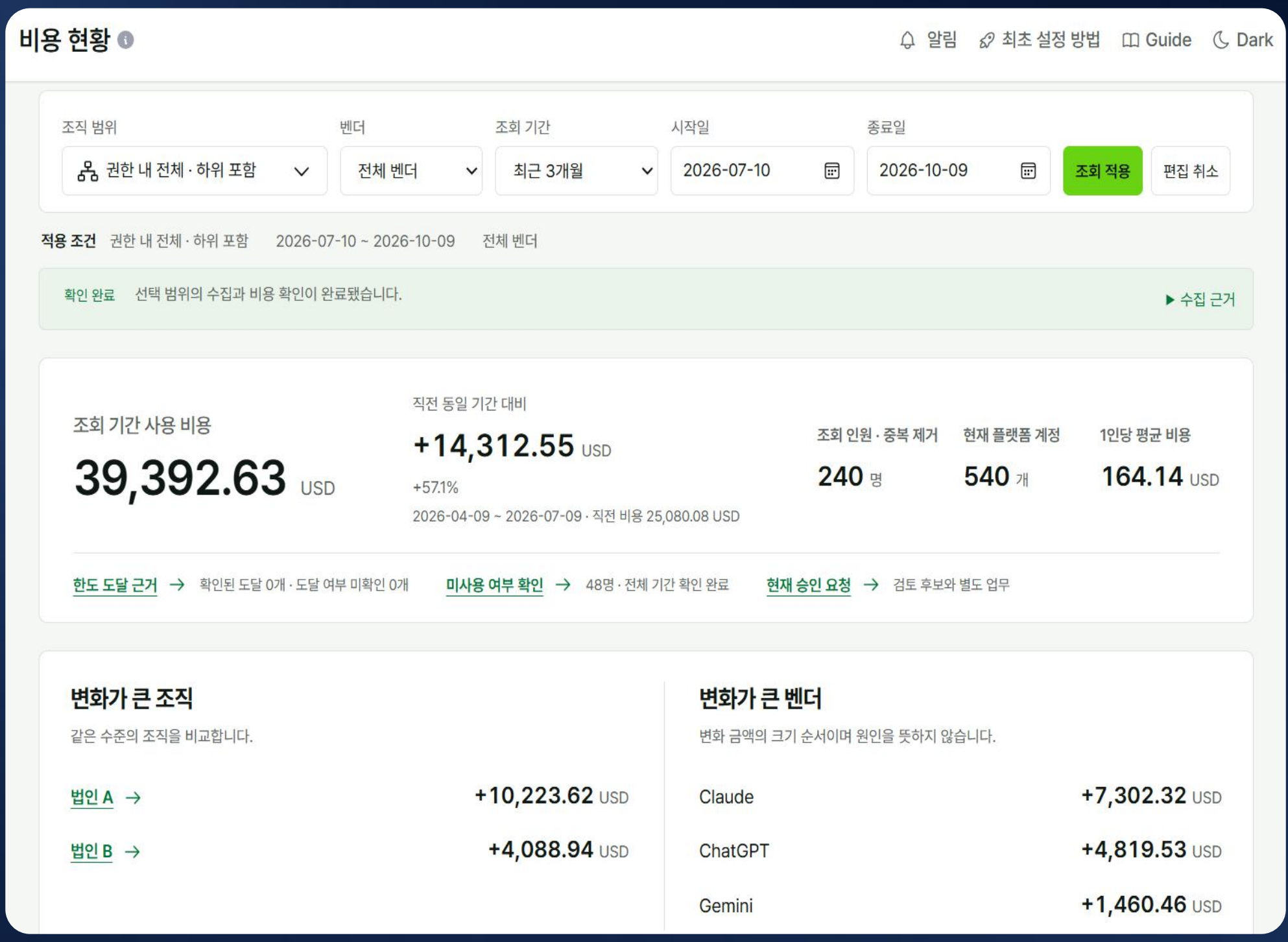
**AI 비용 데이터를 조직별 책임과
예산 정책에 연결해 지속적으로 관리**

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

AI 비용 관리체계를 실제 운영으로 연결하는

딜로이트 AI Tokenomics



* 이해를 돕기 위한 예시이며, 실제 고객 데이터가 아닙니다.



데이터 통합

사용·청구 내역,
조직·업무 정보



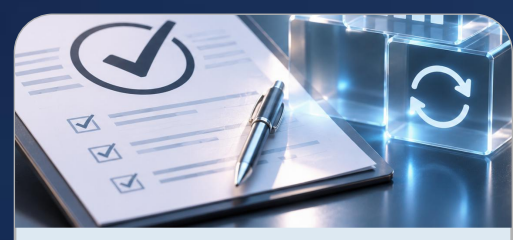
비용·정책 관리

비용 집계·배부,
예산·권한 정책



의사결정·실행

예산·배부 승인,
한도·권한 설정



결과 반영

실제 지출·
예외 처리 확인 후
기준 보완

운영 중 발견한 차이와 예외를
다음 관리 기준에 반영

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

AI 비용관리 전반을 아우르는

딜로이트 AI Tokenomics

기업의 우선과제에 맞춰 필요한 영역부터 적용



비용 가시화·책임 귀속

조직 · 업무별 책임 비용을 하나의 기준으로 확인



예산 배분·관리

예산 사용 전망과 승인 기준에 따른 예산 관리



사용 권한 통제

업무별 허용 범위와 처리 기록 관리



비용 효율화

사용 실적을 바탕으로 갱신 수량과 약정 조정



운영체계 정립·내재화

역할·승인 권한·운영 주기 정립 및 자립 운영 체계 마련

기업별 비용관리 수준과 우선 과제에 맞춰 적용 범위를 구성

상세한 내용은 웹페이지에서 확인하세요

© 2026. For information, contact Deloitte Anjin LLC

Deloitte.

한국 딜로이트 그룹 전문가

딜로이트 AI Tokenomics 서비스 전문팀

AI 사용이 전사로 확산되면서, AI 비용은 더 이상 IT 부서의 운영비로만 볼 수 없게 되었습니다.
AI Tokenomics 서비스는 흩어진 사용·청구 데이터를 조직과 업무 기준으로 연결해
비용의 책임을 분명히 하고, 예산과 사용 권한을 정책으로 관리합니다.
실행 결과는 다시 다음 관리 기준에 반영해, AI 비용이 재무·거버넌스 체계 안에서
지속적으로 관리되도록 지원합니다.

이창성 상무

Forensic & Resilience 본부
경영자문부문 | 한국 딜로이트 그룹

02 6099 4888
changsulee@deloitte.com

박경순 상무

Forensic & Resilience 본부
경영자문부문 | 한국 딜로이트 그룹

02 6676 1726
kyungspark@deloitte.com

Deloitte Insights

세일즈&마케팅
권지원 대표
jekwon@deloitte.com

성장전략부문
서정욱 부대표
juseo@deloitte.com

딜로이트 인사이트 편집장
박경은 Director
kyungepark@deloitte.com

연구원
이소윤 Consultant
soyunlee@deloitte.com

디자이너
박근령 Manager
keunrpark@deloitte.com

Contact us
krinsightsend@deloitte.com

앱스토어, 구글플레이/카카오톡에서 '딜로이트 인사이트'를 검색해보세요.
더욱 다양한 소식을 만나보실 수 있습니다.

Deloitte.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

Deloitte Asia Pacific Limited is a company limited by guarantee and a member firm of DTTL. Members of Deloitte Asia Pacific Limited and their related entities, each of which are separate and independent legal entities, provide services from more than 100 cities across the region, including Auckland, Bangkok, Beijing, Hanoi, Hong Kong, Jakarta, Kuala Lumpur, Manila, Melbourne, Osaka, Seoul, Shanghai, Singapore, Sydney, Taipei and Tokyo.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms or their related entities (collectively, the "Deloitte organization") is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.

본 보고서는 저작권법에 따라 보호받는 저작물로서 저작권은 딜로이트 안진회계법인("저작권자")에 있습니다. 본 보고서의 내용은 비영리 목적으로만 이용이 가능하고, 내용의 전부 또는 일부에 대한 상업적 활용 기타 영리목적 이용시 저작권자의 사전 허락이 필요합니다. 또한 본 보고서의 이용시, 출처를 저작권자로 명시해야 하고 저작권자의 사전 허락없이 그 내용을 변경할 수 없습니다.

© 2026. For information, contact Deloitte Anjin LLC