Deloitte Insights

AI-generated Image

November 2026



AI 토크노믹스

토큰 기반의 TCO 분석과 AI 비용 최적화 전략

Deloitte Insights

Deloitte.

Download on the App Store
GET IT ON Google Play

'딜로이트 인사이트' 앱에서
경영·산업 트렌드를 만나보세요!

목차

01. AI 토크노믹스의 필요성 04

1.1 예측이 어려운 AI 지출 구조 04
1.2 투자 성과 검증의 불투명성 05
1.3 AI 비용 관리 체계의 필요성 05

02. 기존 AI 비용관리 체계의 한계 06

2.1 AI 비용 산정 방식의 한계 06
2.2 AI 조달 방식에 따른 비용 가시성과 운영상의 장단점 07

03. 토큰 비용을 결정하는 세 가지 요인: 기술 스택, 호스팅 방식, AI 모델 09

3.1 AI 기술 스택: 어떤 인프라를 이용하는가 11
3.2 호스팅 방식: 어디서 AI를 구동하는가 13
3.3 AI 모델: 개방형인가, 폐쇄형인가 16

04. 딜로이트의 AI 총소유비용(TCO) 분석: 호스팅 방식별 비용 비교 17

4.1 설계: 기준 설정과 연차별 사용 시나리오 17
4.2 결과: 사용량 증가에 따라 역전되는 비용 우위 20
4.3 시사점: 사용 규모에 맞는 인프라 선택과 전환 유연성 확보 23

05. AI 비용 최적화: 모델·설계·운영 단계별 과잉지출 절감 방안 24

5.1 업무에 맞는 모델과 인프라 선택 24
5.2 설계·모델 최적화·추론 과정의 불필요한 자원 사용 절감 25
5.3 모니터링·비용 배분·사용 한도 기반의 운영 거버넌스 확립 26

06. 경영진 실행 과제 27

리더 메시지



박경순 상무

AI Tokenomics 서비스 전문팀
한국 딜로이트 그룹

02 6676 1726
kyungspark@deloitte.com

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 달라지는 것으로 나타났습니다. 따라서 인프라는 현재 단가보다 향후 사용량을 고려해 선택하고, AI 사용 방식의 최적화와 예산·사용 한도·승인 체계를 통해 비용을 지속적으로 통제해야 합니다.

본 리포트는 AI 비용 관리와 투자 성과 검증을 위한 판단 기준과 경영진의 실행 과제, 이를 지원하는 딜로이트의 AI 비용 관리 서비스를 제시합니다. AI 투자의 경제성을 높이고 실질적인 성과를 창출하려는 경영진께 도움이 되기를 바랍니다.

AI 토크노믹스 서비스 소개 →

AI 토크노믹스 리포트 요약본 →



01
AI 토크노믹스의 필요성

1.1 예측이 어려운 AI 지출 구조

기업 IT 예산 중 AI 관련 지출 비중이 빠르게 커지고 있다. 일부 기업에서는 이 비중이 절반에 달한다.[1] AI 수요가 확대되면서 2026년 전 세계 클라우드 인프라 서비스 지출도 전년 대비 27% 증가가 예상된다.[2] 지정학적 불확실성이 커지며 데이터 주권과 인프라 자립도 기업의 핵심 과제로 부상했다.[3] AI 투자 규모와 변동성 확대로 최고정보책임자(CIO)뿐만 아니라 최고재무책임자(CFO)와 이사회도 투자 판단과 책임의 주체가 됐다.

기존 IT 예산은 소프트웨어 사용권이나 서버 단위로 산정되어 예측하기 쉬웠던 반면, AI 지출은 비선형적으로 증가해 예측이 어렵다. AI 지출을 예측하려면 텍스트, 이미지, 오디오 같은 데이터를 처리하는 정보의 최소 단위인 '토큰' 사용량을 파악해야 한다. AI와 상호작용할 때마다 토큰이 소비되고, 이를 처리하는 데 비용이 드는 구조이기 때문이다. 하지만 토큰 사용량은 AI 추론 방식에 따라 달라진다. 예를 들어, 여러 추론 단계를 거치면서도 맥락을 기억할 수 있는 고급 추론 모델은 단순 응답 모델보다 연산량이 훨씬 많다.

실제로 AI 컴퓨팅 수요와 토큰 처리량은 빠르게 늘고 있다. 엔비디아는 앞으로 AI 컴퓨팅 수요가 10억 배 규모로 늘어날 것으로 전망했다. 구글은 2025년 10월 월간 토큰 처리량이 1,300조 개로 1년 전의 130배에 달한다고 발표했다.[4] 토큰 소비가 빠르게 늘면서 기업의 자본투자 및 에너지 사용 규모와 부담도 커지고 있다.

1.2 투자 성과 검증의 불투명성

AI 투자가 가시적인 성과가 나타나기까지는 시간이 필요하고, 성과 측정도 어렵다. 딜로이트 US Tech Value 서베이에 따르면, 비즈니스·기술 의사결정권자 548명 중 45%가 기초 수준의 AI 자동화도 투자수익률(ROI)을 달성하려면 최대 3년이 걸릴 것으로 예상했다. 고도화된 AI 자동화는 3년보다 더 오래 걸릴 것으로 예상됐다.[5] 글로벌 재무 리더 1,326명 대상으로 실시한 Finance Trends 서베이에서도 AI 투자가 명확하고 측정 가능한 가치를 창출하고 있다고 답한 응답자는 28%에 그쳤다.[6]

1.3 AI 비용 관리 체계의 필요성

기업이 AI 투자 비용을 측정하고 성과를 달성하고자 한다면, 토큰 단위로 AI 지출을 추적·예측·최적화하는 비용 관리 체계 마련이 우선되어야 한다. **본 리포트는 이러한 토큰 기반 AI 비용 관리 체계를 'AI 토크노믹스'라고 지칭한다.** 토큰으로 비용을 분석하면 GPU 사용량, 호스팅 방식, 데이터 저장 방식 등의 AI 투자 옵션이 구체적인 재무 수치로 환산되고, 매출과 마진, 생산성을 만드는 데 실제로 얼마가 드는지 확인할 수 있다. 결국 토큰 단위로 비용을 파악하고, 이를 성과와 연결하는 역량이 핵심이다. AI 경쟁력은 도입 속도뿐 아니라 비용 구조 관리 체계에도 달려 있다. 경영진의 과제는 통제 불가능한 비용의 누적을 막고, 토큰 사용이 측정 가능한 성과로 이어지게 하는 것이다.





02
기존 AI 비용 관리 체계의 한계

2.1 AI 비용 산정 방식의 한계

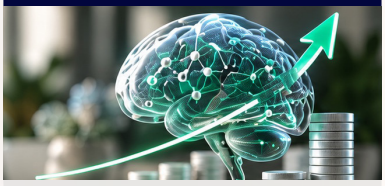
기업들은 컴퓨팅 비용을 산정할 때 주로 라이선스, 사용자 수, 서버 사용 시간 등을 기준으로 사용한다. 하지만 기업들은 이 기준으로 AI의 실제 사용량과 비용 발생 구조를 충분히 설명하지 못하고 있다. 결국 전력 사용량을 kWh로 측정하듯 토큰이라는 공통의 단위로 AI 사용량, 비용, 성과를 측정해야 한다 (그림 1 참조).

그림 1. 토큰 기반 AI 비용 산정 구조



이처럼 비용은 토큰 사용량과 토큰당 단가에 따라 정해지는데, 토큰 수요와 비용은 다음 세 가지 특성 때문에 예측하고 통제하기 어렵다.

비선형 수요



복잡한 추론 모델은 성능이 높지만, 단순 추론보다 훨씬 많은 토큰을 소비할 수 있다.

사용량 변동



시험 사용(experimentation)의 빈도와 규모, 워크로드 설계, 프롬프트 방식에 따라 토큰 사용량이 크게 달라진다.

가격 변동



모델과 인프라 효율성에 따라 토큰 처리 비용과 이용 단가가 달라진다.[7]

토큰 비용은 사용 방식뿐 아니라 인프라 구성에 따라서도 달라진다. GPU, 스토리지, 네트워크를 어떻게 구성하느냐에 따라 토큰 하나를 처리하는 속도와 단가가 달라지기 때문이다. 그러므로 기업 리더는 인프라 선택이 비용에 미치는 영향을 이해하고, 토큰 처리에 투입한 비용이 매출·마진·생산성 개선으로 이어지는지 평가해야 한다.

2.2 AI 조달 방식에 따른 비용 가시성과 운영상의 장단점

AI의 주요 조달 방식은 SaaS, API, AI 팩토리다. SaaS는 완성된 소프트웨어를 주로 구독료나 사용자당 요금으로 이용하는 방식이다. API는 기업이 AI 모델에 직접 요청을 보내고, 처리된 토큰 수에 따라 요금을 지불하는 방식이다. AI 팩토리는 전용 인프라와 소프트웨어·서비스를 통합해 AI를 운영하는 방식으로, 기업이 인프라를 직접 소유하거나 외부 인프라를 활용할 수 있다.

기업은 비용, 지연 시간, 보안 요건 등을 고려해 조달 방식을 선택하며, 이 선택에 따라 비용이 청구·집계되는 방식뿐 아니라 비용 발생 원인을 파악하고 사용을 조정할 수 있는 범위도 달라진다(표 1).[8]

표 1. AI 조달 방식별 비용 가시성과 운영상의 장단점

조달 방식	토큰 사용량 · 비용 가시성	장점	단점
SaaS (패키지 소프트웨어)	토큰 사용량과 처리 비용이 구독료에 포함돼 직접 드러나지 않는 경우가 많음	운영이 간편함	모델·인프라와 토큰 소비에 대한 기업의 통제력이 제한적
API	토큰 사용량과 적용 단가를 확인할 수 있음	사용량 기반 지출을 추적하기 용이함	워크로드 설계와 프롬프트 길이에 따라 지출이 달라지며, 제공자의 인프라를 직접 통제하기 어려움
AI 팩토리 (자체 인프라 운영)	측정 체계를 통해 토큰 처리량과 인프라 비용을 직접 파악할 수 있음	데이터 주권에 대한 통제력이 높고, 충분한 규모와 활용률을 확보하면 단가를 낮출 수 있음	초기 투자 부담이 크고, 기술·운영 역량이 필요함

AI 팩토리의 정의와 선택 기준

딜로이트는 AI 팩토리를 전용 컴퓨트·네트워크·스토리지와 최적화된 소프트웨어·서비스의 결합으로 정의한다. 이는 학습부터 추론까지 AI 라이프사이클 전반을 고성능·대규모로 지원한다. 핵심 산출물은 AI가 생성하는 지능이며, 토큰 처리량으로 측정한다. 이러한 산출물은 기업의 의사결정과 자동화, 새로운 AI 솔루션 개발을 뒷받침한다.

기업은 외부 서비스를 계속 이용할지, 자체적으로 AI 팩토리를 구축해 운영할지 결정해야 한다. 어떤 방식이 적합한지는 워크로드 규모, 데이터 민감도, 향후 사용량을 얼마나 예측할 수 있는지에 따라 크게 달라진다.



오프프레미스(SaaS·API): 초기 파일럿이나 수요 변동성이 큰 워크로드, 데이터 민감도가 낮은 업무에 적합하다. 초기 인프라 투자 없이 바로 시작할 수 있고, 사용 규모를 유연하게 조정할 수 있다는 것이 강점이지만, 사용량이 늘면 토큰당 비용이 자체 운영보다 높을 수 있다.



AI 팩토리: 워크로드 규모가 크고 수요가 예측 가능하며, 지연 시간에 민감한 업무에 적합할 수 있다. 특히 자체 운영 비용이 외부 서비스 이용 비용보다 낮아지는 시점부터 비용 측면의 이점이 커진다. 초기 투자나 인프라 확보에 부담이 따르지만, 인프라를 충분히 활용하면 토큰당 비용을 낮추고 데이터 주권을 보다 직접적으로 관리할 수 있다. 구축 경로도 다양하다. 기업이 자체 시설에서 인프라를 운영하는 온프레미스나, 외부 데이터센터에 자체 장비를 두는 콜로케이션뿐 아니라, AI 전용 인프라에 특화된 신생 클라우드 사업자인 네오클라우드를 활용할 수도 있다.[9]

대부분의 글로벌 기업은 두 방식을 결합한 하이브리드 방식을 취하고 있다. 작고 예측하기 어려운 시험적 워크로드는 API로 처리하고, 규모가 크고 가치가 높은 워크로드는 사용량과 비용 구조가 안정되는 시점에 맞춰 AI 팩토리로 옮기는 식이다. 두 방식을 결합해도 비용 관리 문제는 남는다. 외부 서비스는 비용 구조를 파악하고 사용을 통제하는 데 제약이 있고, 자체 운영은 인프라 비용을 토큰 처리량에 배분하고 가동률을 관리해야 한다. 기업은 각 방식의 한계를 보완하도록 사용량·비용·성과를 연결하는 관리 체계를 마련해야 한다.

03

토큰 비용을 결정하는 세 가지 요인: 기술 스택, 호스팅 방식, AI 모델

AI 조달 방식이 비용의 청구 형태를 결정한다면, 실제 토큰 처리 비용은 인프라 구성과 운영 방식, 모델 선택에 따라 달라진다. 같은 AI 모델을 사용하더라도 사용자 수, 토큰 사용량, GPU 사용 시간 등 과금 기준이 다르므로, 요금만으로는 비용 구조를 충분히 파악하기 어렵다. 비용을 비교하려면 그 바탕이 되는 기술 스택, 호스팅 방식, AI 모델을 함께 살펴봐야 한다. 이 장에서는 세 가지 요인이 토큰 비용에 미치는 영향과 각 선택에서 고려해야 할 사항을 분석한다.



그림 2. 토큰 비용 결정 요인



3.1 AI 기술 스택: 어떤 인프라를 이용하는가

AI 기술 스택이란 AI를 사용하는 데 필요한 GPU, 스토리지, 네트워킹 등 인프라 구성요소를 뜻한다. 이러한 인프라의 선택과 구성은 토큰 처리 비용에 반영되지만, 기업이 세부 비용 구조를 얼마나 파악할 수 있는지는 조달 방식에 따라 다르다.

SaaS는 여러 비용이 하나의 계약 금액에 묶여 있어 세부 비용 구조가 드러나지 않는 경우가 많아, 기술 스택의 각 구성 요소를 구분해 TCO를 분석하기 어렵다.

API는 기술 스택 관련 비용이 토큰당 요금이나 처리량 기준 요금에 간접적으로 반영된다.



기본 요금

입·출력 토큰의 요율은 모델에 따라 다르며, 일반적으로 토큰 100만 개당 요금으로 표시된다.



할인·요금 옵션

예약 용량(미리 약정한 만큼 할인 받는 방식), 프롬프트 캐싱(반복되는 입력을 재사용해 비용을 아끼는 방식), 배치 처리(여러 요청을 묶어 처리해 단가를 낮추는 방식) 같은 할인 옵션이 제공된다. 일부 기업 고객에게는 사용자 기반 요금이 함께 적용되기도 한다.



추가 비용

데이터 저장 비용이나 클라우드 밖으로 데이터를 전송할 때 발생하는 반출 비용이 더해지면 API의 총비용이 늘어날 수 있다.

자체 호스팅은 외부에 토큰별 요금을 지불하지 않고 인프라 구축과 운영에 비용을 직접 투입한다. GPU 구매·임대, 스토리지 구성 등에 따른 자본지출과 운영지출을 토큰 처리량과 함께 분석해 토큰당 비용을 산정한다.

표 2. AI 기술 스택이 토큰 비용에 미치는 영향과 AI 팩토리 도입 시 비용·투자 부담

구성요소	토큰 비용에 미치는 영향	AI 팩토리 도입 시 비용·투자 부담
컴퓨트 (GPU 및 가속기)	• 최신 GPU와 고대역폭 메모리는 토큰당 처리 시간을 단축하지만, 구매·임대 비용이 높음	인프라 구매·임대 비용이 가장 큰 비중을 차지하며, 기업이 이를 직접 부담하고 빠른 신제품 출시 주기를 고려해야 함
스토리지 (고속 데이터 접근)	• AI 워크로드는 성능을 유지하고 비용을 관리하기 위해 비휘발성 메모리와 병렬 파일 시스템을 활용해 테라바이트 규모의 데이터를 연속적으로 처리함 • 기존 스토리지는 데이터 접근 지연으로 GPU 대기 시간을 늘려 토큰당 비용을 높임	비휘발성 메모리·병렬 파일 시스템·벡터 데이터베이스에 대규모 투자 필요
네트워킹 (GPU 인터커넥트*)	• 수천 개 GPU로 학습하려면 초저지연 인터커넥트가 필수적임 • 통신 지연이 길면 GPU 유휴 시간과 토큰당 비용이 함께 높아지며, 기존 방식은 토큰 비용을 오히려 높임	네트워크 인프라 비용을 기업이 직접 부담
전력·냉각 (AI 랙의 에너지 수요)**	• 차세대 GPU 랙은 1대당 250~300kW 소비 (일반 서버는 10~15kW) • 전력 비용은 온프레미스에서는 직접 청구되고 클라우드에서는 요금에 포함되어, 어느 쪽이든 토큰 처리 비용에 반영됨	높은 전력·냉각 운영비와 액체 냉각 설비가 필요
시설 (물리적 인프라 요건)	• AI 랙의 무게는 최대 약 1,360kg으로, 기존 랙보다 약 40% 무거울 수 있음.[10] • 바닥 보강과 고도화된 냉각 설비가 필요할 수 있으며, 관련 비용이 토큰 처리 비용에 반영됨	바닥 보강·랙 설치 비용을 기업이 직접 부담함
인력·운영 비용	• IT 운영·관리, 소프트웨어·라이선스, 애플리케이션 개발·통합, 데이터 관리·거버넌스, 추론·서빙, 보안·컴플라이언스, 사용자 교육·변화관리 비용이 포함됨	MLOps*** 전반의 비용, 전문역량센터(COE)**** 운영 및 역량 강화 비용, 오케스트레이션 프레임워크와 MLOps 도구 (데이터·오케스트레이션·보안), 컴플라이언스 비용을 기업이 직접 부담

출처: 딜로이트 분석(프로젝트 경험 기반)

*__GPU 인터커넥트(InfiniBand, NVLink 등):__ GPU 간 데이터 전송을 지원하는 연결 기술. 통신이 느리면 GPU의 대기 시간이 늘어 처리 효율이 떨어진다.

**__랙(rack):__ 서버와 네트워크 장비 등을 장착하는 표준 구조물. 데이터센터에서 주로 캐비닛 형태로 설치된다.

***__MLOps(Machine Learning Operations):__ AI 모델의 배포·모니터링·업데이트 등을 체계적으로 관리하는 운영 방식. 소프트웨어 개발과 운영을 통합하는 DevOps 개념을 머신러닝에 적용한 것이다.

****__전문역량센터(Center of Excellence, COE):__ 특정 분야의 전문 인력을 모아 표준을 수립하고 다른 부서를 지원하는 전담 조직.

3.2 호스팅 방식: 어디서 AI를 구동하는가

토큰 비용은 AI 모델을 어디서, 어떻게 구동하는지에 따라서도 달라진다. 같은 대규모 언어모델(LLM)이라도 온프레미스, 네오클라우드, 하이퍼스케일러, API 중 어떤 호스팅 방식을 택하느냐에 따라 비용 구조가 달라진다.

온프레미스는 기업이 자체 시설에서 인프라를 운영하는 방식이다. 콜로케이션은 외부 데이터센터의 공간·전력·냉각 설비 등을 이용하면서 기업의 장비를 운영하는 방식이다. 네오클라우드는 GPU 등 AI 전용 인프라 제공에 특화된 클라우드 사업자를 말한다. 하이퍼스케일러는 AWS·Microsoft Azure·Google Cloud와 같은 대형 범용 클라우드 사업자다. API 이용은 모델을 구동하는 인프라를 직접 운영하지 않고, 제공자의 AI 모델을 호출해 사용한다.



표 3. AI 호스팅·이용 방식별 비용 구조와 운영 특성

구분	온프레미스	네오클라우드	하이퍼스케일러	API 이용
지출 구조	초기 자본지출은 크지만, 운영지출은 상대적으로 낮음	운영지출만 발생	운영지출만 발생	운영지출만 발생
이용 단가와 과금 기준*	**최저** GPU 시간당 약 1~2달러	**중간** GPU 시간당 평균 약 1~4달러이나 변동성이 큼	**높음** GPU 시간당 약 3~7달러이나 지역·모델 등에 따라 상이	**매우 높음** 출력 토큰 100만 개당 0.40~100달러 이상
확장성	**중간** 장비 조달·전력 확보·설치에 시간이 소요	**높음** 수요에 따라 자원을 유동적으로 할당함	**중간~높음** 수요에 따라 확장 가능하며, 확장 가능한 규모가 매우 큼	**매우 높음** 제공자가 전적으로 관리
지연 시간	**가장 낮음** 하드웨어와 네트워크를 직접 구성해 조정할 수 있음	**낮음** AI 전용 설계를 활용할 수 있으나 물리적 배치는 통제할 수 없음	**중간** 물리적 인프라와 워크로드 배치 통제력이 거의 없음	**중간~높음** 제공자의 인프라·네트워크를 통제할 수 없고 원거리 통신이 수반됨
통제·맞춤화	**매우 높음** **완전한 통제**	**중간** 물리적 인프라와 유지보수는 통제할 수 없으나, 배포하는 모델·소프트웨어에 대한 통제력은 높음	**중간** 네오클라우드와 동일하게 취급	**매우 낮음** 인프라를 통제할 수 없으며, 모델 튜닝과 응답 형식에 대한 통제도 제한적
보안·데이터 주권	**최고 데이터** 암호화·전송·저장을 직접 통제	**높음** 콜로케이션과 같은 수준이며 암호화 수준은 더 높음	**중간** 데이터 유출 위험이 있으며 정확한 호스팅 위치 통제가 어려움	**낮음** 제공자의 아키텍처와 거버넌스 체계를 통제할 수 없음
배포 기간	**오래 걸림** 장비 조달·설치 등으로 수개월이 걸릴 수 있음	**즉시**	**즉시**	**즉시**
유지보수 책임	**주로 기업이 담당** 시설·전력 등 관리형 서비스는 공동 책임	**공동** 물리적 인프라는 제공사, 나머지는 고객이 담당	**공동** 물리적 인프라는 제공사, 나머지는 고객이 담당	**AI 모델 제공자**
적합한 워크로드	안정적이고 처리량이 큰 워크로드	탄력적 컴퓨트**, PoC***, 비용에 민감한 워크로드. 민감한 데이터를 처리하는 워크로드를 위한 추가 기능을 제공할 수 있음	탄력적 컴퓨트, PoC	신속한 시험적 활용, 에이전트, RAG****

출처: 딜로이트 분석. 공개·비공개 추정치, 공개 GPU 가격 자료, 생성형 AI API 요금 비교 자료, 하이퍼스케일러 비용 계산기 등을 기반으로 작성

*이용 단가와 과금 기준: 온프레미스·네오클라우드·하이퍼스케일러는 GPU 시간당 단가(GPU 1대를 1시간 사용하는 데 드는 비용), API는 출력 토큰 100만 개당 요금

**탄력적 컴퓨트: 수요에 따라 연산 자원을 조정해야 하는 업무

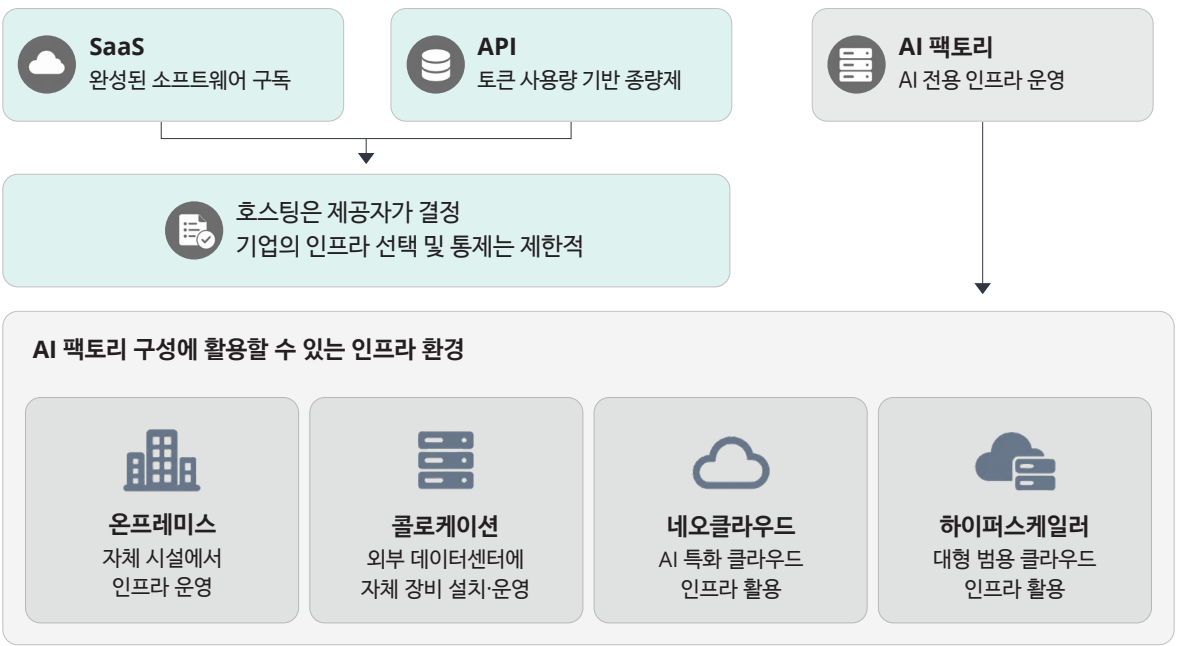
***PoC(Proof of Concept): 아이디어나 기술이 실제로 작동하는지 소규모로 시험해보는 검증 단계

****RAG(검색증강생성): AI가 답변을 생성하기 전 데이터 검색으로 응답 품질을 높이는 기술

호스팅 방식에 따라 인프라 구성도 달라지며, 이는 토큰 처리 효율에 영향을 미친다. 일정한 공간에 얼마나 많은 연산 자원을 배치하는지인 컴퓨트 밀도, GPU 간 통신 경로와 연결 구조에 따른 네트워크 근접성, 저장장치가 단위 시간에 처리할 수 있는 데이터양인 스토리지 처리량 등이 주요 요인이다.

기업이 호스팅 방식을 선택하고 통제할 수 있는 범위도 조달 방식에 따라 달라진다. SaaS와 API 이용자는 기반 인프라와 비용 구조를 직접 통제하기 어려워 같은 작업이라도 제공자에 따라 비용이 달라질 수 있지만, 자체 운영 기업은 다양한 호스팅 방식을 직접 선택할 수 있으며 이 선택이 토큰당 비용을 결정하는 주요 요인이 된다(그림 3).

그림 3. 조달 방식별 호스팅 선택 범위



3.3 AI 모델: 개방형인가, 폐쇄형인가

AI 모델이 개방형인지 폐쇄형인지도 토큰 비용에 영향을 미친다. 개방형 모델은 공개되어 있어 기업이 직접 배포하고 조정할 수 있는 모델을 뜻한다. 폐쇄형 모델은 모델 자체를 공개하지 않고 API나 서비스 형태로 제공한다.

조달 방식에 따라 선택할 수 있는 모델과 비용 구조도 달라진다. SaaS 이용자는 벤더가 구축한 모델을 그대로 사용하고, API 이용자는 제공자를 선택할 수 있지만 모델의 비용 구조는 정할 수 없다. 자체 운영 기업은 배포 가능한 모델과 인프라를 함께 선택해 비용·유연성·데이터 주권 사이의 우선순위를 직접 조정할 수 있다.[11]



개방형 모델

개방형 모델은 모델 자체의 이용료가 없거나 낮은 경우가 많지만, 직접 운영하려면 인프라와 운영 비용이 발생한다. 기업은 모델을 자체 환경에 배포함으로써 운영에 대한 통제력을 높이고, 업무에 맞게 조정하며, 데이터 주권을 보다 직접적으로 관리할 수 있다.

개방형 모델은 기업 고유의 데이터나 민감한 데이터로 모델을 추가 학습하는 파인튜닝(fine-tuning, 사전 학습된 모델을 특정 업무에 맞게 조정하는 방식)에도 적합하다. 또한 특정 제공자에 대한 의존을 줄여 벤더 종속(vendor lock-in, 다른 제공자로 전환할 때 기술적·계약적 부담이 커지는 현상)을 완화하며, 장기적으로 토큰 비용을 낮추는 데에도 유리하다.

대표적인 모델군으로는 Meta의 Llama와 Mistral 등이 있다. NVIDIA NIM Microservices처럼 벤더가 개방형 모델을 표준화되고 안전한 배포 단위로 패키징하는 프레임워크도 등장해, 기업이 맞춤형으로 구축하던 통합 작업이 체계화되고 있다.



폐쇄형 모델

폐쇄형 모델은 API에서 토큰 사용량에 따른 종량제 요금이 적용되는 경우가 많다. 기업이 직접 모델을 배포하거나 인프라를 구축하지 않고도 이용할 수 있고, 기본 제공 기능만으로 높은 성능을 활용할 수 있다는 장점이 있다. 제공자의 운영 지원을 받을 수 있다는 점도 특징이다. 대표적인 모델군으로는 Anthropic의 Claude, Google의 Gemini, OpenAI의 GPT, xAI의 Grok 등이 있다.

이러한 이점에도 폐쇄형 모델은 토큰당 비용이 상대적으로 높은 편이고, 토큰 사용량이 변동하면 지출을 예측하기 어렵다. 맞춤화가 어렵다는 한계와 함께 데이터 저장에 대한 우려, 벤더 종속 위험도 검토해야 한다.



04

딜로이트의 AI 총소유비용(TCO) 분석: 호스팅 방식별 비용 비교

4.1 설계: 기준 설정과 연차별 사용 시나리오

AI 비용 구조에는 제본스의 역설(Jevons' paradox)이 나타난다. 제본스의 역설은 효율이 개선될수록 오히려 총 소비량이 늘어나는 현상이다.[12] 같은 토큰을 더 적은 비용으로 처리할 수 있게 되면서 토큰당 비용은 빠르게 하락하고 있다. 딜로이트는 평균 추론 비용이 2025년 백만 토큰당 0.04달러에서 2030년경 약 0.01달러까지 내려갈 것으로 전망한다.[13]

실제로 토큰당 비용이 내려가는 동안에도 기업의 AI 지출은 오히려 늘고 있다.[14] 에이전틱 시스템과 멀티에이전트 워크플로가 확산되면서 토큰 수요가 기하급수적으로 늘어, 인프라 효율 개선 속도를 앞지르는 경우가 많기 때문이다. 비용 관리 체계를 제대로 갖추지 않으면 총비용은 계속 커질 수 있다.

토큰 소비 증가에 따른 비용 부담은 기업, 벤더, 최종 사용자에게 각각 다르게 작용한다. 토큰 소비가 빨라질수록 이 지출을 누가 최종적으로 떠안느냐가 관건이며, 워크로드가 커지고 복잡해지면 그 구조도 달라진다. 딜로이트는 이 비용 부담이 어느 시점에 어디로 옮겨 가는지 확인하기 위해 기술 스택 전반의 토큰 TCO를 분석했다. 분석은 토큰 가격을 결정하는 세 가지 핵심 요인에 따라 TCO가 어떻게 달라지는지 검증했으며, 세 요인은 다음과 같다.



기술 스택

GPU, AI 모델,
아키텍처에 따른 비용 차이



호스팅 방식

사용량과 복잡도가 증가할 때
호스팅 방식별 비용 차이



사용량 확대

사용자 수나 활용 사례별
추론 복잡도, 깊이가 증가할 때
나타나는 토큰 사용량과 지출 변화

분석 목적은 TCO를 좌우하는 핵심 요인, 사용량 확대에 따른 비용 변화, 호스팅 방식 간 비용 우위가 바뀌는 시점을 확인하는 데 있다. TCO 모델은 이상적인 가정이 아니라 실제 기업 운영 조건을 반영하도록 설계했다.[15] 이 모델은 다양한 구성을 지원하지만, 본문에서는 일반적인 엔터프라이즈 워크로드를 대표하는 시나리오로 요약했다.

분석조건

- ☑ **컴퓨트 스택:** GPU 서버 - NVIDIA HGX B200(NVLink·NVSwitch 지원), CPU - AMD EPYC 9654
- ☑ **LLM:** Llama 3.3 70B FP8 TP2, GPT-4o
- ☑ **호스팅 방식:** 온프레미스(AI 팩토리)*, API 이용, 전문 네오클라우드 제공자(NCP, Neocloud Provider)**

* 본 분석에서 AI 팩토리는 온프레미스로 구축한 경우를 가리킨다.
** NCP는 시간당 요금과 기간별 예약 계약을 모두 제공하나, 본 모델에서는 시간당 요금만 적용

분석은 사용량이 늘어남에 따라 GPU를 8개 단위로 증설하는 것으로 가정했다. 다음 표는 이 모델에 적용한 연차별 시나리오를 보여준다.

표 4. 연차별 AI 활용 시나리오: 파일럿에서 대규모 추론까지

단계	연차별 AI 활용 시나리오: 파일럿에서 대규모 추론까지
1년차: 파일럿	• 챗봇, FAQ 어시스턴트 같은 단순하고 반복적인 문의 응대에 AI 도입 • 고객서비스·HR 문의·기본 IT 헬프데스크를 지원하는 경량 대화형 AI • 짧고 정형화된 질의응답 위주이며 대화 맥락 유지는 최소화
2년차: PoC/소규모 도입	• 문서 요약, 사내 문서 검색처럼 조직의 지식을 활용하는 업무로 확장 • 정책 문서·제안서·계약서를 검색·요약하는 사내 어시스턴트 활용 • 문서 의미를 바탕으로 관련 정보를 찾는 시맨틱 검색(semantic search)과 이전 대화 맥락을 유지하는 멀티턴(multiturn) 대화 포함
3년차: 대규모 추론	• 분석 코파일럿(분석가의 인사이트 도출을 돕는 AI) 등 의사결정 지원 업무로 고도화 • 컨설턴트·애널리스트·감사인의 인사이트 도출, 보고서 초안 작성, 여러 데이터 소스에 걸친 분석 지원 • 추론·구조화된 출력·엔터프라이즈 시스템 연동 포함

기술 변화에 대응하는 인프라 투자 전략



AI 하드웨어가 빠르게 발전하면서 기술적 노후화 주기가 기존 감가상각 주기를 크게 앞지르고 있고, GPU 세대교체도 잦아지고 있다. 최신 AI 모델의 기능을 구현하는 데 필요한 성능이 기존 선도 GPU의 역량을 금세 넘어서고, 구형 하드웨어에 대한 지원은 줄고 있다. GPU 제조사가 연 단위 출시 주기로 전환하면서 교체 시점은 더 앞당겨졌다. 기업은 빠른 업그레이드의 이점과 기술 경쟁에서 뒤처질 위험 사이에서 균형을 계속 맞춰야 한다.

GPU 성능 향상으로 고도화된 추론, 방대한 텍스트 요약, 1시간 분량의 영상 분석처럼 많은 맥락 정보를 한 번에 처리하는 AI 활용이 가능해졌다. 에이전틱 추론을 포함한 이러한 작업은 대규모 GPU 메모리와 최신 하드웨어를 필요로 할 수 있으나, 기업의 멀티모달·에이전틱 추론 도입은 아직 초기 단계다. 추론 작업은 특히 중형 모델의 경우 구형 GPU에서도 원활히 처리되는 경우가 많다.

토큰 가격이 하락하면서 자체 구축과 외부 서비스 이용의 비용 구도도 빠르게 바뀌고 있다. 기업은 고정된 가정에 의존하기보다 향후 수요와 기술 변화를 반영한 인프라 전략을 마련해야 한다. 업그레이드 시점을 신중하게 계획하고 비용 구조를 점검해, 시장이 안정되는 과정에서도 투자의 타당성이 유지되는지 확인해야 한다.

4.2 결과: 사용량 증가에 따라 역전되는 비용 우위

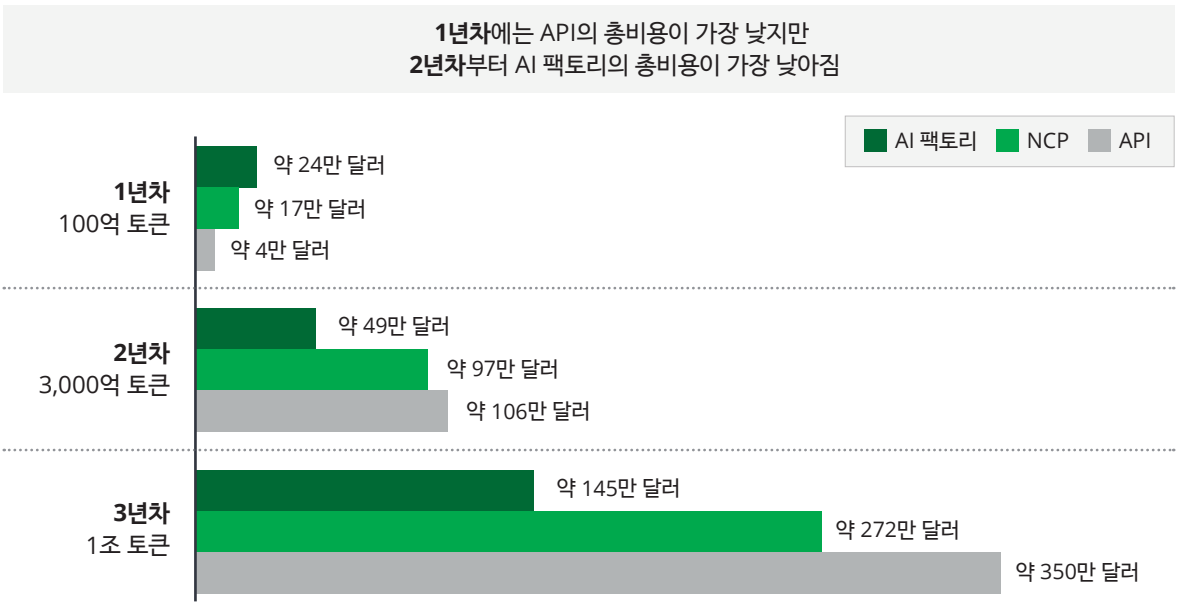
딜로이트는 호스팅 방식별 TCO를 비교하기 위해 하드웨어 가동률, 에너지 비용, 시설 비용 등 가치사슬 전반의 변수에 현재 시장 상황을 반영해 시뮬레이션을 수행했다. 분석 결과, 비용이 가장 낮은 방식은 토큰 사용량에 따라 달랐다. 초기에는 API 이용이 유리했지만, 사용량이 늘면서 자체 운영 AI 팩토리의 비용 우위가 나타났다.

1. 연차별 총비용 분석

연간 토큰 사용량이 100억 개 수준인 1년차에는 API가 비용 측면에서 유리했다. 사용한 만큼 요금을 지불하므로, 자체 인프라를 확보했을 때 발생할 수 있는 유휴 설비 비용을 부담하지 않기 때문이다.

2년차부터는 비용 우위가 역전됐다. 사용량과 추론 부하가 늘면서 AI 팩토리는 고정 인프라 비용을 더 많은 토큰에 분산하고 가동률을 높여 API보다 낮은 비용을 보였다. 3년간의 연간 TCO를 합산하면 API의 누적 TCO는 AI 팩토리의 약 2.1배다(그림 4). 시뮬레이션 결과, 종량제 API와 NCP는 단순하고 사용량이 적은 업무에, 자체 운영 AI 팩토리는 복잡하고 사용량이 많은 업무를 장기간 운영할 때 비용 효율적이다.

그림 4. 연차별 사용량 증가에 따른 호스팅 방식별 연간 TCO



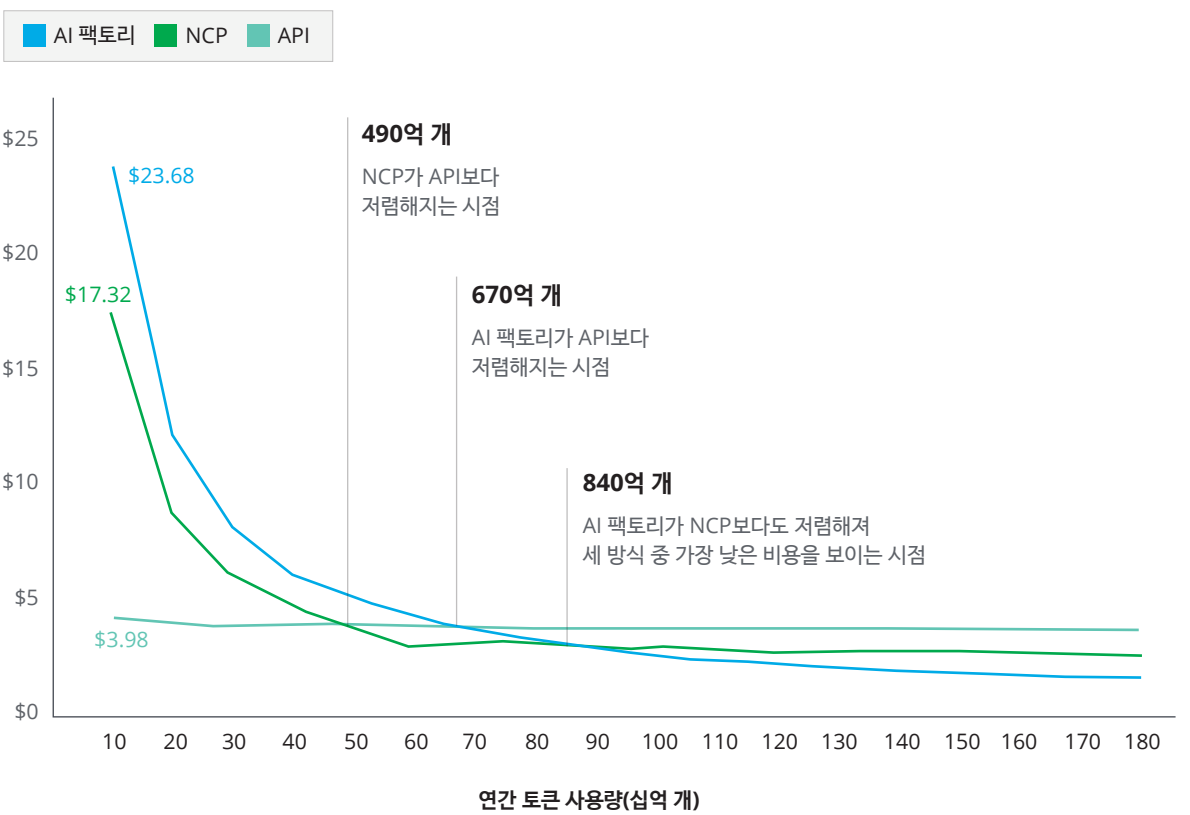
출처: 딜로이트 시뮬레이션

2. 사용량별 토큰당 비용 분석

연간 토큰 사용량이 증가할 때 토큰 100만 개당 TCO가 어떻게 달라지는지도 분석했다. 각 방식의 연간 토큰 처리에 필요한 설비 구축비(자본지출)와 연간 운영비를 합산하고, 이를 토큰 처리량으로 나눠 비교했다.

사용량이 적은 구간에서는 API가 가장 저렴했다. API는 비용이 사용량에 비례해 늘어나 토큰당 비용이 거의 줄지 않은 반면, AI 팩토리와 NCP는 사용량이 커질수록 토큰당 비용이 빠르게 하락했다. NCP는 연간 490억 토큰부터, AI 팩토리는 670억 토큰부터 API보다 저렴해졌고, 840억 토큰을 넘어서면 AI 팩토리가 NCP보다도 저렴해져 세 방식 중 가장 낮은 비용을 보였다(그림 5).

그림 5. 연간 토큰 사용량별 호스팅 방식의 토큰당 비용과 비용 우위 변곡점

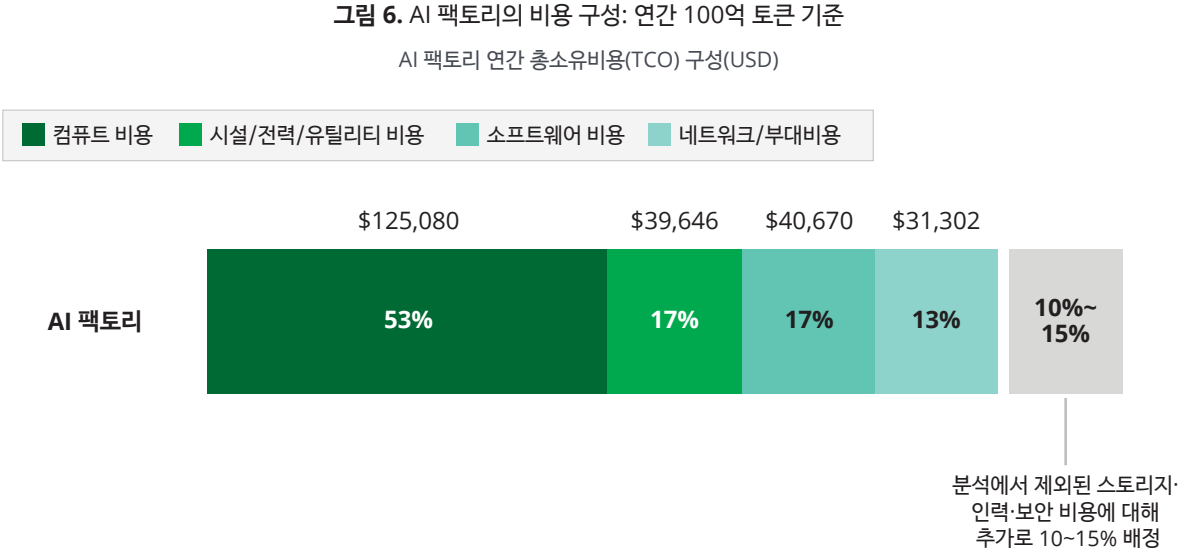


출처: 딜로이트 시뮬레이션

3. AI 팩토리 비용 구성 분석

100억 토큰 규모에서 AI 팩토리의 연간 TCO는 컴퓨트 비용이 53%를 차지하고, 시설·전력·유틸리티, 소프트웨어, 네트워킹·부대비용은 각각 13~17%로 비슷한 비중을 보인다. 컴퓨트를 제외한 항목이 비용의 약 절반(47%)인 셈이다(그림 6).

그림 6. AI 팩토리의 비용 구성: 연간 100억 토큰 기준



시설·전력·유틸리티

- AI 인프라 운영의 핵심 요소. 특히 에너지 소비가 많은 GPU는 전력 수요를 크게 늘림
- 액체 냉각, 워크로드 스케줄링, 전력사용효율(PUE) 최적화 등을 통해 에너지 효율을 개선하면 운영상 이점과 비용 절감 가능

소프트웨어

- AI 소프트웨어 제품군의 프레임워크·라이브러리·지원 서비스 등을 이용하기 위한 기업용 라이선스 비용
- GPU와 워크로드의 오케스트레이션·관리 비용은 미포함

네트워크·부대비용

- 대규모 AI 워크로드 운영에 필요한 고대역폭 스위치·라우터·인터커넥트 등의 비용
- 일반적으로 전체 TCO의 10~20% 차지

출처: 딜로이트 시뮬레이션

GPU가 에너지 집약적이라 전력 수요가 급증하므로 시설·전력·유틸리티는 AI 인프라 운영에서 중요한 역할을 한다. 액체 냉각, 워크로드 스케줄링, 전력사용효율(PUE) 최적화 등으로 에너지 효율을 높이면 운영상 이점과 비용 절감 효과를 크게 얻을 수 있다.

온프레미스 AI 팩토리는 초기 투자 부담이 크지만, 아키텍처를 신중하게 설계하고 기술 스택 전반을 최적화하면 장기적으로 상당한 비용 절감과 이점을 얻을 수 있다. 이 모델은 기존 온프레미스 데이터센터를 AI용으로 개조하는 조직이 흔히 택하는 공랭식 시스템을 가정했다. 액체 냉각을 도입한다면 전력·유틸리티 비용에 일회성 설치비(케이지 구축, 냉각 개조, 배관, 냉각·분배장치 도입)도 필요에 따라 반영해야 한다.

4.3 시사점: 사용 규모에 맞는 인프라 선택과 전환 유연성 확보

분석 결과, 소규모 워크로드에서는 API 이용 방식이 비용 측면에서 유리했으나 사용량과 추론 복잡도가 증가하면 비용 구조가 달라졌다. 자체 운영의 비용 이점은 워크로드 특성과 기업의 우선순위에 따라 커질 수 있다. 기업은 현재의 이용 단가뿐 아니라 향후 수요와 운영 조건까지 함께 검토해야 하며, 다음 세 가지에 유의해야 한다.



워크로드 파악

업무별 사용량과 특성을 측정하고 중요도에 따라 우선순위를 정해야 한다.
이 결과가 인프라에 직접 투자할지, 외부 서비스를 이용할지를 정하는 기준이 된다.



현재·미래 수요 분석

현재 사용량과 향후 수요는 호스팅 전략과 모델 선택에 직접적인 영향을 미친다.
사용 규모가 커지면 토큰당 비용과 TCO가 크게 달라질 수 있으므로 이를 함께 검토해야 한다.



벤더 종속을 줄이는 설계

기술과 수요 변화에 대응할 수 있도록 유연하게 설계해야 한다. 모듈형 하이브리드 아키텍처를 검토하고 AI 전략을 정기적으로 재점검해, 기술적 제약보다 업무 요구와 재무적 판단에 따라 호스팅 방식을 선택할 수 있어야 한다.

05

AI 비용 최적화: 모델·설계·운영 단계별 과잉지출 절감 방안



토큰 가격이 하락하더라도 AI 활용이 확대되고 워크로드가 복잡해지면 총지출은 증가할 가능성이 크다. 기업은 모델·인프라 선택, 워크로드 설계, 운영 관리의 세 단계에서 불필요한 지출을 줄일 수 있다. 비용 최적화는 단순히 단가를 낮추는 것을 넘어 예측하기 어려운 지출을 관리하고, 벤더 종속으로 인한 제약을 줄이는 일도 포함한다.

5.1 업무에 맞는 모델과 인프라 선택

과잉지출의 흔한 원인은 업무에 필요한 수준보다 큰 모델이나 인프라를 사용하는 데 있다.[16] 프론티어급 모델은 범용성이 높지만 토큰 소비 부담도 클 수 있다. 특정 분야의 업무에서는 소형 모델이나 파인튜닝한 모델로도 더 적은 토큰을 사용하면서 비슷한 정확도를 얻을 수 있다.

인프라도 업무 특성에 맞춰 선택해야 한다. 고성능 GPU는 대규모 학습에 필요하지만, 이상 탐지나 분류 등 일부 작업에는 과도할 수 있다. 이러한 작업에는 중급 GPU나 CPU가 비용 측면에서 더 적합할 수 있다.

여러 모델을 병행하는 전략도 유용하다. 높은 성능이 꼭 필요한 업무에만 대형 모델을 사용하고, 나머지는 소형 모델이나 개방형 모델로 처리하면 성능 요구를 충족하면서 비용을 관리할 수 있다.

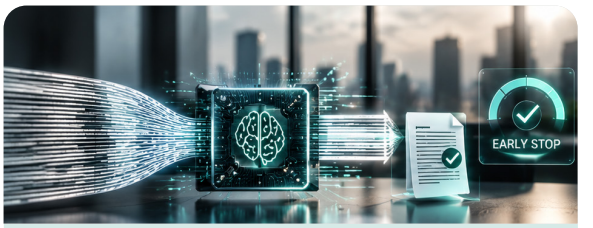
5.2 설계·모델 최적화·추론 과정의 불필요한 자원 사용 절감

적합한 모델을 선택하더라도 설계가 비효율적이면 토큰이 낭비될 수 있다. 단순한 로직으로 충분한 작업에 긴 추론 과정을 돌려 토큰을 과소비하는 AI 에이전트가 대표적이다. 다음 네 가지 접근을 통해 불필요한 추론과 중복 처리를 줄이고, 업무 수행에 필요한 토큰과 연산 자원을 효율적으로 사용할 수 있다.



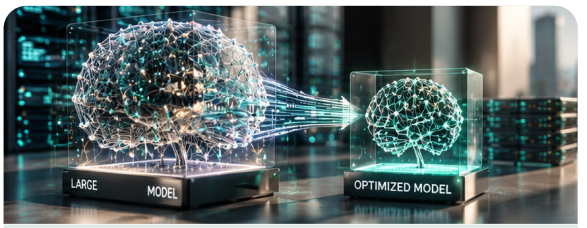
간결한 설계

조건별로 처리 경로를 결정하는 의사결정 트리(Decision Tree), 사전에 정한 규칙에 따라 작업을 처리하는 규칙 기반 엔진, 한 번에 처리하는 입력 분량을 제한하는 컨텍스트 윈도우 상한을 활용한다.



알고리즘 기법

조기 종료는 정해진 정확도 기준에 도달하면 처리를 중단한다. 프롬프트 절단은 입력 길이를 줄이고, 압축 트랜스포머는 성능을 유지하면서 불필요한 토큰 사용을 줄인다.



모델 최적화

양자화는 모델 가중치의 정밀도를 낮춰 연산 부담을 줄이고, 프루닝은 불필요한 파라미터를 제거한다. 지식 증류는 소형 모델이 대형 모델의 출력이나 동작을 학습하도록 한다. 전이학습은 사전 학습한 모델을 활용해, 처음부터 전체 모델을 학습하지 않고도 새로운 업무에 맞게 조정하는 방식이다.



추론 설계

검색증강생성(RAG)은 필요한 정보를 검색해 활용함으로써 과도한 맥락 입력의 필요성을 줄인다. 프롬프트 엔지니어링은 입력을 간결하게 구성하고, 응답 캐싱은 동일한 질의에 대한 반복 처리 비용을 줄인다. 배치 처리는 여러 요청을 묶어 처리량을 높인다. 전통적인 결정론적 모델과 확률적 추론 모델을 함께 활용해 업무 복잡도에 맞춰 토큰 사용량을 조정한다.

5.3 모니터링·비용 배분·사용 한도 기반의 운영 거버넌스 확립

앞선 기술적 최적화만으로는 충분하지 않다. AI 지출을 지속적으로 모니터링하고 통제하는 운영 체계가 뒷받침돼야 한다. 관리가 소홀하면 여러 사업부에서 발생한 비용이 눈에 띄지 않게 누적될 수 있다. 기업은 클라우드 비용 관리에 적용했던 재무적 관리 원칙을 AI에도 적용해야 한다.

오케스트레이션은 워크로드의 배치와 실행을 조정하는 관리 수단의 하나다. GPU 활용률 약 85%를 목표로 삼으면 유휴 자원 없이 인프라를 효율적으로 쓰고 있는지 확인할 수 있다.

GPU 사용 시간, 스토리지, 데이터 반출, 토큰 사용량을 통합 모니터링하면 지출 현황을 실시간으로 파악할 수 있다. 조직·업무·프로젝트 정보를 태그로 연결하면 사업부별 비용 배분과 내부 청구(charge-back)가 가능해지고, 공식 승인이나 관리 체계 밖에서 쓰이는 그림자 AI(shadow AI)의 비용이 통제 없이 쌓이는 것도 막을 수 있다.

예산 알림, 컨텍스트 윈도우 제한, API 사용량 상한 등은 과도한 사용을 방지하는 통제 수단이다. 재무·기술·사업 조직이 함께 사용량과 비용을 관리하는 FinOps를 적용하면 수요 예측과 투자 승인도 연결할 수 있다. 예상 토큰 수요와 투자수익률(ROI)을 검토하고, 정해진 비용·성과 기준에 따라 프로젝트를 승인하는 방식이다.

기업은 시장 가격이나 토큰 수요 증가 자체를 통제할 수는 없지만, 다음 네 가지는 관리할 수 있다.

01



현재 사용하는 모델과 인프라의 효율 개선

02



업무에 비해 과도한 모델·인프라에 대한 지출 방지

03



예측하기 어려운 토큰 소비와 비용 증가 통제

04



전환 유연성을 제한하는 계약 조건과 기술적 선택 검토

AI 비용을 에너지나 자본 배분과 같은 수준으로 관리하는 기업은 지출 통제력을 유지하면서 AI 활용을 확대할 수 있다. 관리가 부족한 기업은 과다 지출할 가능성이 높을 뿐 아니라, 전략적 유연성을 제한하는 모델과 벤더에 종속될 위험도 안게 된다.



06
경영진 실행 과제

사용자당 라이선스나 고정형 가상머신 요금은 예측 가능한 워크로드와 안정적인 사용 패턴을 전제로 하지만, AI 워크로드는 사용 방식과 추론 복잡도에 따라 비선형적으로 확대될 수 있다. 기업은 토큰 사용량과 처리 비용을 함께 살펴봐야 한다.

토큰은 AI의 사용량·비용·성과를 연결해 분석하는 핵심 기준이다. 어떤 업무에 비용이 투입되는지, 자원이 얼마나 효율적으로 사용되는지, 지출이 실제 가치 창출로 이어지는지 파악하는 출발점이 된다. 비용과 성과를 연결해 평가하려면 기술적 선택과 재무적 판단을 함께 검토해야 하므로, CIO는 비용과 투자 성과를, CFO는 모델과 인프라의 운영 특성을 이해해야 한다.

관리 체계가 없으면 AI 지출은 기존 SaaS 갱신 계약에 포함되어 파악되지 않은 채 증가하거나, 예측하기 어려운 API 청구서로 급증하거나, 되돌리기 어려운 인프라 약정에 묶일 수 있다. 이러한 비용의 누적은 단순한 예산 초과를 넘어 기업의 경쟁력과 전략적 유연성에 영향을 미칠 수 있다. 경영진은 AI 사용을 가치 창출과 연결하고, 활용 확대에 맞춰 비용 관리 원칙을 적용해야 한다.

경영진이 고려해야 할 네 가지 시사점

- 비용의 최종 부담 주체와 가격 반영 구조를 파악해야 한다. 비용 변동은 가치사슬의 어느 단계에서든 나타날 수 있으며, 벤더의 가격 정책이나 라이선스 구조를 통해 기업에 반영될 수 있다.

- 사용 규모에 맞춰 조달 방식을 선택해야 한다. 소규모 워크로드에는 SaaS나 API가 적합할 수 있다. 규모가 크고 수요가 예측 가능한 워크로드에서는 자체 운영이 유리할 수 있다.

- 인프라·설계·사용량을 함께 관리해 비용 증가를 억제해야 한다. 인프라 최적화, 워크로드 설계, 사용량 통제를 통해 불필요한 지출을 줄일 수 있다.

- 인건비 절감액과 AI 처리 비용을 함께 평가해야 한다. 토큰 비용이 인건비 절감액을 초과할 수 있으므로, 자동화 효과만으로 투자 가치를 판단해서는 안 된다. 에이전트가 복잡해질수록 사람이 개입하는 방식(human-in-the-loop)이 업무 효율을 좌우할 수 있다.

하이브리드 인프라와 FinOps를 통한 실행 기반 마련

하이브리드 아키텍처는 비용과 전략적 요구에 맞는 환경에 워크로드를 배치할 수 있는 유연성을 제공한다. 민감한 데이터를 처리하는 업무는 온프레미스에, 수요 변동이 큰 시험적 활용은 클라우드에, 지연 시간에 민감한 추론은 엣지에 배치하는 방식이다. 모델 역시 업무에 필요한 성능에 맞춰 선택하고, 에이전트 설계와 프롬프트에서 발생하는 불필요한 처리를 줄여야 한다. 토큰 낭비는 기업 가치의 손실로 이어진다.

클라우드 비용 관리에 활용해 온 FinOps, 즉 재무·기술·사업 조직이 함께 사용량과 비용을 관리하는 운영 방식을 AI에도 적용해야 한다. FinOps는 토큰 사용량과 비용을 지속적으로 측정하고 통제하는 기반을 제공한다. 하이브리드 인프라와 FinOps를 결합하면, 업무에 맞는 운영 환경을 확보하면서 AI 활용을 예측 가능하게 확대할 수 있다.

실시간 토큰 사용량 모니터링, 예산 알림, 사업부별 비용 배분·청구, 프로젝트별 투자수익률(ROI) 기준은 단순한 지원 업무가 아니라 AI 활용이 지속 가능하도록 돕는 재무적 통제 장치다. 이 장치를 갖춘 기업은 토큰 소비를 측정 가능한 기업 가치로 연결하면서 AI 활용을 확대할 수 있다. 갖추지 못한 기업은 AI 프로젝트가 여러 사업부로 분산되고 워크로드가 당초 범위를 넘어 커지면서 비용이 눈에 띄지 않게 누적되고, 비용 증가와 계약상 제약으로 전략적 유연성을 잃을 수 있다.

토큰 경제는 AI 전략의 부수적인 요소가 아니라 운영 모델의 핵심이다. 하이브리드 인프라와 FinOps는 이를 지속 가능하게 만드는 기반이다. 경영진은 자본·에너지·인재와 같은 수준의 엄격한 관리 원칙을 AI에도 적용해야 한다.

출발점: AI 토큰 사용 현황 및 비용의 파악과 통제

AI 활용을 확대하는 국내 기업은 자체 인프라 구축을 검토하기에 앞서, 현재 어떤 서비스를 누가, 어떤 업무에 사용하는지 파악할 필요가 있다. 사용 현황과 사용량, 비용을 확인하고 관리 책임을 정하는 것이 출발점이다.

사용 현황과 비용을 파악하려면 청구 금액을 합산하는 것만으로는 부족하다. 공급업체와 제품마다 청구 단위와 계약 조건이 다르고, 사용자 수·크레딧·토큰·쿼리 등 사용량을 나타내는 단위도 서로 다르다. 청구 내역과 사용 기록을 조직·업무별로 연결해, 비용 발생 지점과 지출 목적까지 파악해야 한다.

이러한 기반이 마련되면 사용하지 않는 서비스, 업무에 비해 과도한 모델, 예산을 초과하는 사용을 점검할 수 있다. 이후 사용 규모와 수요의 예측 가능성이 높아지면 모델 변경, 호스팅 전환, AI 팩토리 구축 여부도 근거를 갖고 검토할 수 있다.

현재의 사용 현황과 비용을 파악하는 일은 단기적인 지출 통제에 그치지 않는다. 워크로드와 사용 규모를 이해하고, 장기적인 AI 운영 전략을 수립하기 위한 기반이 된다.

그림 7. AI 토크노믹스 관리 체계

사용·청구 데이터를 통합하여 비용과 정책을 관리하고, 승인된 조치의 실행 결과를 다음 예산 권한·관리 기준에 반영



주석

1. Tim Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?," Deloitte, October 16, 2025.
2. Omdia, Global cloud infrastructure spending rose 29% in Q4 2025 as hyperscalers scaled AI infrastructure investment
3. Bram De Schouwer, Thomas Kessler, and Aurélien Descamps, "Cloud sovereignty: Succeeding in the evolving landscape," Deloitte, March 12, 2025.
4. Sundar Pichai, "Q3 earnings call: Remarks from our CEO," Google's The Keyword, October 29, 2025.
5. Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?"
6. Steve Gallucci et al., "Finance Trends 2026: Navigating the expanded scope of finance," Deloitte Insights, October 6, 2025.
7. Steven Rosenbush, "AI economics are brutal. Demand is the variable to watch," Wall Street Journal, October 14, 2025.
8. Chris Thomas et al., "Is your organization's infrastructure ready for the new hybrid cloud?," Deloitte, June 30, 2025.
9. Chris Thomas, Ganesh Seetharaman, and Diana Kearns-Manolatos, "AI workloads are surging. What does that mean for computing?," Deloitte, August 21, 2025.
10. Eka Linwood, "Design parameters for data center facilities," STRUCTURE, January 1, 2023.
11. Frank Nagle, "Revealing the hidden economics of open models in the AI era," The Linux Foundation, November 19, 2025.
12. Greg Rosalsky, "Why the AI world is suddenly obsessed with a 160-year-old economics paradox," NPR's Planet Money, February 4, 2025.
13. LLM Token Cost, NVIDIA Token Calculator, accessed November 18, 2025.
14. Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?"
15. Lambda, "AI cloud pricing," accessed December 12, 2025; CoreWeave, "GPU cloud pricing," accessed December 12, 2025.
16. Thomas et al., "Is your organization's infrastructure ready for the new hybrid cloud?"

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
AI Tokenomics 서비스 전문팀
한국 딜로이트 그룹

02 6099 4888
changsulee@deloitte.com

박경순 상무
AI Tokenomics 서비스 전문팀
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

앱

Download on the App Store

GET IT ON Google Play

카카오톡 채널

'딜로이트 인사이트' 앱과 카카오톡 채널에서
경영·산업 트렌드를 만나보세요!

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