Asset Allocation

Compliance Notice

본 조사자료는 고객의 투자에 정보를 제공할 목적으로 작성되었으며, 어떠한 경우에도 무단 복제 및 배포 될 수 없습니다. 또한 본 자료에 수록된 내용은 당사가 신뢰할 만한 자료 및 정보로 얻어진 것이나, 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

자산배분/해외크레딧
Analyst 이영주 leeyoungju@hanafn.com

하나증권 리서치센터

2026년 7월 2일 | Global Asset Research

자산배분의 창(窓)

토큰 경제 (Tokenomics)로 이어지는 AI 생태계

AI 경쟁은 성능에서 운영의 안정성과 수익성으로

그동안 AI 경쟁은 더 뛰어난 모델을 개발하는 기업을 중심으로 전개됐다. 더 높은 성능과 긴 컨텍스트, 더 많은 파라미터가 경쟁력을 결정하는 핵심 기준이었다. 그러나 AI가 기업 업무 전반으로 확산되면서 경쟁의 초점도 변화하고 있다.

AI 생태계는 크게 인프라 (Infrastructure), AI 모델 (Model), 애플리케이션 (Application)의 세 계층으로 나눌 수 있다. 인프라 기업은 GPU·데이터센터·네트워크 등 AI 연산 기반을 제공하고, AI 모델 기업은 이를 활용해 추론 서비스를 제공한다. 애플리케이션 기업은 AI 모델을 활용해 기업 고객에게 서비스를 제공하며 최종 수익을 창출한다.

이 중, 애플리케이션 기업에게 중요한 것은 더 뛰어난 AI 모델보다도 이를 안정적으로 서비스하고 지속 가능한 수익으로 연결하는 역량이 될 것이다. 이 과정에서 토큰 (Token)은 AI 생태계의 핵심 경제 단위로 자리 잡으며, 이를 중심으로 새로운 경제 구조 (Tokenomics)가 형성되고 있다.

Goldman Sachs는 월간 토큰 사용량이 2030년까지 현재 대비 약 24배 증가한 120 Quadrillion Tokens에 이를 것으로 전망했다. 반면 토큰당 추론 비용은 연간 60~70% 하락할 것으로 예상한다. 단위 비용은 낮아지지만, 사용량 증가 속도가 이를 크게 웃돌면서 AI 활용 규모는 계속 확대될 것으로 예상되고 있다.

애플리케이션 기업의 비용 구조도 달라지고 있다

애플리케이션 계층의 소프트웨어 기업들은 전통적으로 사용자 수(Seat)를 기준으로 구독료를 받는 SaaS 모델을 가져왔다. 기존 SaaS에서는 고객이 증가해도 추가 원가는 제한적이었고, 일반적으로 70~90% 수준의 Gross Margin을 유지할 수 있었다.

반면 생성형 AI에서는 AI 기능을 제공할수록 AI 모델 호출이 증가하면서 토큰 사용량과 추론비도 함께 늘어난다. 즉, 기존 SaaS 대비 AI 서비스는 사용량 증가가 곧 원가 증가로 이어지는 구조를 갖게 된다. 그래서 AI Native SaaS의 Gross Margin은 50~60% 수준으로 낮아지고 있으며, 추론비는 매출의 약 23%를 차지하는 것으로 추정되고 있다.

Reuters 조사에서도 71%의 기업이 AI 관련 비용을 이미 예산 초과한 것으로 나타났다. 결국 AI 기능이 확대될수록 토큰 사용량도 함께 늘어나면서 'Token Tax'가 애플리케이션 기업의 새로운 비용 부담으로 부상하고 있다.

이에 따라 과금 방식도 변화할 수밖에 없다. 증가하는 Token Tax (토큰 사용에 따른 추가 비용)를 기존 구독 모델만으로는 회수하기 어려워졌기 때문이다. 다양한 기업들이 AI 시대에 맞춰 Consumption 기반, Credit 기반의 과금 방식을 확대하며 AI 비용을 사용량에 맞춰 회수하는 구조를 강화하고 있다.

토큰 경제는 AI 생태계의 경쟁 방식을 바꾸고 있다

토큰은 인프라, AI 모델, 애플리케이션을 연결하는 공통의 경제 단위다. 토큰 사용량이 증가할수록 추론 수요도 함께 확대되며 GPU, 데이터센터, 네트워크 등 인프라 투자도 늘어나고 있다. 현재 AI 투자 사이클에서 인프라 계층이 가장 직접적인 수혜를 받고 있는 이유도 여기에 있다.

다만 산업이 성숙할수록 관심은 단순한 토큰 사용량보다 토큰을 얼마나 효율적으로 처리하고 수익으로 연결하는지로 이동할 가능성이 높다. 인프라 기업은 연산 처리 효율, AI 모델 기업은 추론 효율, 애플리케이션 기업은 AI 서비스를 안정적인 매출과 이익으로 연결하는 역량이 경쟁력을 결정하는 기준이 될 것으로 보인다.

AI 생태계의 지속성을 결정하는 것은 '수익화'

인프라 투자만으로 AI 산업이 지속적으로 성장하기는 어렵다. 애플리케이션 기업이 AI 서비스를 통해 안정적인 수익을 창출해야 기업의 AI 도입이 확대되고, 이는 AI 모델 활용과 인프라 투자 증가로 이어진다. 다시 향상된 성능과 낮아진 비용은 AI 활용을 더욱 확대시키며 생태계의 선순환을 만든다.

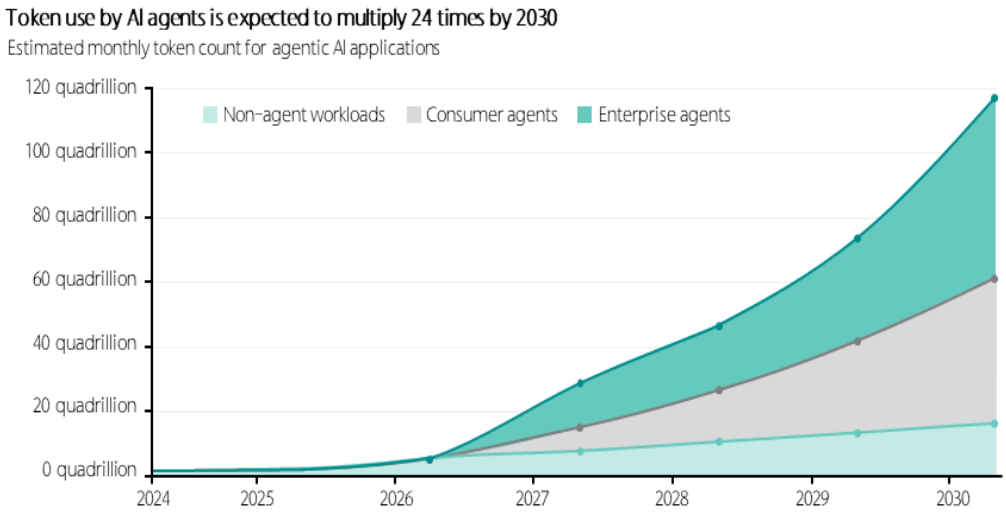
결국 시장의 관심도 토큰의 양(Volume)에서 토큰의 수익화(Monetization)로 이동할 것으로 보인다. 투자 포인트 역시 AI를 하나의 산업으로 접근하기보다 애플리케이션의 수익화가 AI 모델과 인프라 투자로 이어지는 선순환이 지속될 수 있는지에 맞춰질 가능성이 높다.

** 2030 년 월간 토큰 사용량 : 현재 대비 약 24 배 증가 (120 Quadrillion Tokens)

• 토큰당 추론 비용 : 연간 60~70% 하락 전망

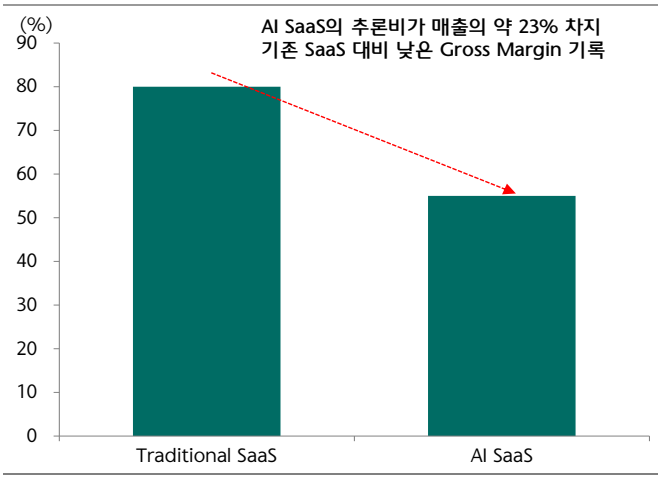
• 비용 감소보다 사용량 증가가 더 빨라 AI 연산 수요는 지속 확대

도표 1. 토큰 사용량 증가 전망



자료: GoldmanSachs, 하나증권

도표 2. 기존 SaaS Vs AI Native SaaS (Gross Margin 및 매출 구조도 변화)



구분	Traditional SaaS	AI SaaS
과금 기준	Seat 기반	Usage 기반
과금 방식	Subscription	Consumption / Credit
비용 구조	고정비 중심	추론비 증가
운영 레버리지	사용자 증가 → 마진 확대	사용량 증가 → 비용 증가

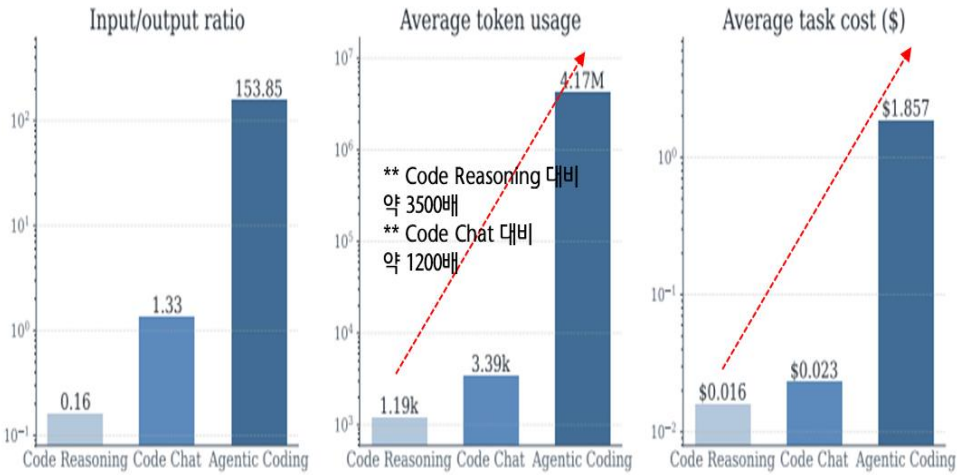
자료: GoldmanSachs, ICONIQ Research Paper, 하나증권

** AI 에이전트는 단순 질문을 넘어 계획 수립, 자료검색, 코드 수정, 결과 검증을 반복 수행

* 이 과정에서 이전 작업 내용을 계속 불러와야 하므로 토큰 사용량이 기존 AI 활용방식보다 폭발적으로 증가

*비용증가 대부분은 답변을 생성하는 출력 (Output) 보다 누적된 작업 반복 처리하는 입력 (Input) 토큰에서 발생

도표 3. 연산 과정별 토큰 사용량 및 비용



주) Code Reasoning: AI가 한 번의 추론으로 문제를 해결
Code Chat: Ai가 대화를 주고받으며 문제를 해결
Agentic Coding: AI가 스스로 계획을 세우고 여러작업을 수행
자료: Bai, L. et al. (2026) How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks, 하나증권

도표 4. 기업별 과금 구조 변화

기업	기존 구조	AI 이후 변화	현재 방식 (사실 기반)
Salesforce	Seat 기반 Subscription	AI 에이전트 사용량에 따라 과금	Consumption + Flex Credits (Agentforce Actions 기준)
Adobe	Subscription	생성형 AI 사용량에 따라 차감	Generative Credits
Microsoft	Microsoft 365 Seat	Copilot 라이선스 + 일부 API 사용량 과금	Subscription + Usage (API)
OpenAI	API 중심	토큰 사용량 기반	Usage-based (Input / Output Token)
Anthropic	API 중심	토큰 사용량 기반	Usage-based (Input / Output Token)
Google (Gemini API)	API 중심	토큰 사용량 기반	Usage-based (Token)
Amazon Bedrock	Cloud 사용량	모델 호출량 기준	Pay-as-you-go

자료: 하나증권