

초점

October 2026 No.2

토큰 이코노미의 부상과 시사점

이경선 연구위원

정보통신정책연구원 디지털정책연구실

토큰 이코노미의 부상과 시사점*

이경선 연구위원

정보통신정책연구원 디지털정책연구실, leeks@kisdi.re.kr

요약

- 최근 AI 산업에서는 토큰이 AI 서비스의 사용량·연산 규모·비용을 측정하는 핵심 단위로 활용되는 ‘토큰 이코노미(token economy)’가 부상
- 이러한 변화의 배경에는 AI 산업의 경쟁축이 모델 개발에서 서비스·수익화로 확장되고 AI의 활용 확대와 추론 모델·에이전트의 확산으로 추론 연산 수요가 증가하면서 비용구조까지 변화하는 구조적 전환이 존재
 - 토큰당 가격과 추론 비용은 빠르게 하락하고 있으나, AI 활용 확대에 따라 전체 토큰·연산 수요와 AI 지출은 오히려 증가
 - 특히, AI는 이용량과 과업 복잡성이 증가할수록 추가적인 연산과 비용이 발생하면서 과업 당 토큰·연산 투입량이 서비스 비용을 좌우하는 주요 변수로 부상
- 다만 토큰 자체가 AI의 지능이나 경제적 가치를 직접 의미하는 것은 아님
 - 동일한 토큰이라도 모델·과업·처리방식에 따라 비용과 가치가 상이
 - 따라서 공급자에게는 토큰이 원가와 자원 사용을 관리하는 핵심 단위로 기능하는 반면 고객에게는 토큰 자체보다 서비스·과업을 통해 얻는 가치가 비용 지불의 기준이 될 가능성
- 결국 토큰 이코노미의 핵심은 더 많은 토큰을 생산·소비하는 것 자체가 아니라, 증가하는 토큰 소비를 실제 경제적 가치와 수익으로 전환하는 데 있음
 - AI 산업화가 진전될수록 경쟁의 초점은 ‘토큰당 비용’에서 ‘과업당 비용’, ‘과업당 창출가치’로 확장될 가능성
 - 향후 AI 경쟁력은 AI를 실제 과업에 얼마나 효율적으로 적용하고, 이를 생산성 향상·서비스 가치·기업 수익으로 연결할 수 있는가에 의해 좌우될 것

* 본고는 필자가 KISO 저널에 기고한 ‘AI 추론 수요의 확대와 토큰 이코노미의 부상’(2026.8.31)을 바탕으로 작성

이 개 요

- 최근 AI 산업에서는 ‘토큰 이코노미(token economy)’가 새로운 화두로 부상
 - 토큰 이코노미는 AI의 활용이 확대되면서 토큰이 AI 서비스의 사용량·연산 규모·비용을 측정하는 핵심 단위로 자리잡는 흐름
- 이는 토큰당 가격이 빠르게 하락하고 있음에도 AI 모델 관련 총지출이 크게 증가하고 있는 최근의 시장 변화와 맞물려 있음
 - OECD 자료에 따르면 2025년 1월 AI 모델의 이용가격*은 2023년 1월 대비 약 80% 하락(André et al. 2025.6)
 - * 모델 성능을 고려하여 조정한 이용가격
 - 가트너는 2030년이 되면 1조 파라미터 규모 LLM(large language model)의 추론 비용이 2025년 대비 90% 이상 하락할 것으로 전망(Gartner, 2026.3.25)
 - 반면 전 세계 AI 모델 지출은 2025년 약 155억 달러에서 2026년 326억 달러, 2027년에는 592억 달러로 가파르게 성장할 것으로 전망(Gartner, 2026.5.19)
- 즉, AI 산업에서는 단위비용의 하락이 활용 확대를 촉진하고 총수요 증가로 이어지는 ‘제번스의 역설(Jevons paradox)’과 유사한 현상이 나타나고 있음
 - 제번스의 역설은 기술 발전으로 자원 이용 효율이 높아져 단위당 비용은 낮아지더라도 비용 절감에 따른 이용 확대가 전체 자원 소비를 오히려 증가시킬 수 있다는 개념으로,
 - 기술의 효율 개선이 실제 이용 확산으로 연결되는 과정에서 나타날 수 있음
- AI의 비용 하락과 활용 범위의 확대가 동시에 진행되는 최근의 변화는 AI 산업에 새로운 전환점을 의미할 수 있다는 점에서 본고에서는 토큰 이코노미의 부상 배경과 그 의미를 보다 구체적으로 살펴보고자 함

02 배경: 학습에서 추론·활용으로

- 토큰 이코노미의 부상 배경에는 AI 산업의 경쟁축이 모델 개발에서 서비스·수익화로 확장되고 AI의 활용 확대와 추론 모델·에이전트의 확산으로 추론 연산 수요가 증가하면서 비용구조까지 변화하는 구조적 전환이 존재
 - AI 모델의 성능 향상과 추론 비용의 하락은 다양한 서비스와 업무에서의 AI 활용을 확대하고, 이는 다시 토큰 사용량과 추론 연산 수요의 증가로 연결

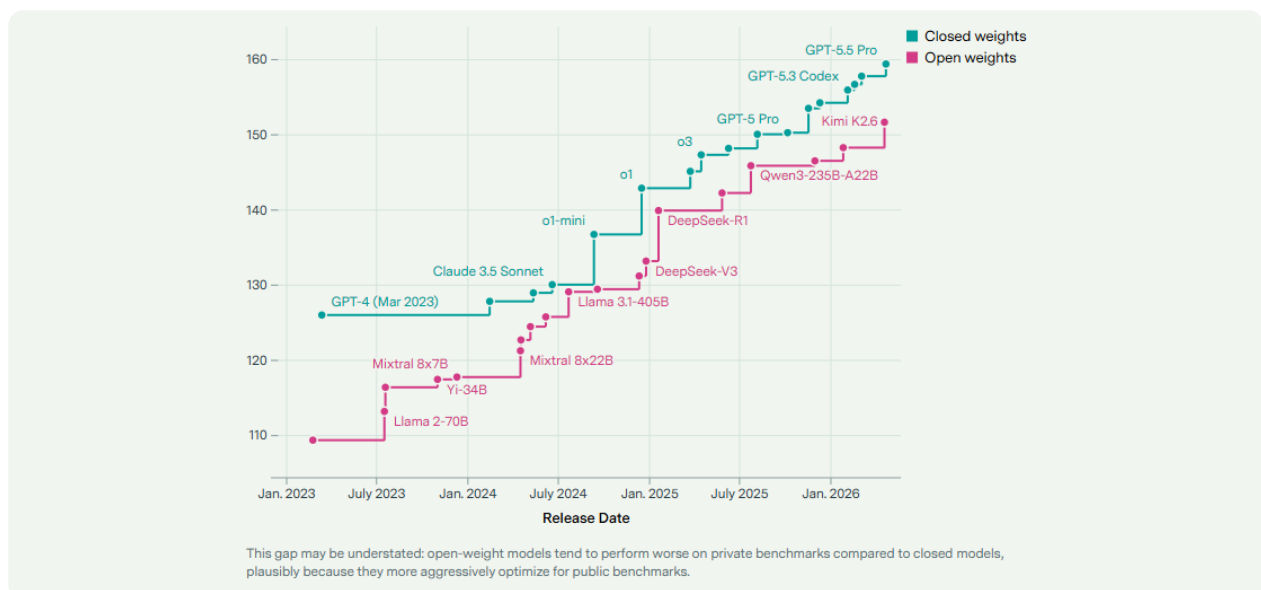
I 경쟁축의 확장: 모델 경쟁에서 서비스·수익화 경쟁으로

- 오픈소스 진영의 빠른 추격으로 AI 모델 간 성능 격차가 빠르게 좁혀지면서 AI 경쟁의 중심이 ‘모델 자체의 성능’에서 ‘실제 서비스 적용과 수익화’로 확장
 - 챗지피티(ChatGPT)의 등장 이후 한동안 AI 산업의 경쟁은 ‘누가 더 우수한 파운데이션 모델을 개발할 수 있는가’를 중심으로 전개
 - 그러나 프론티어 모델*이 여전히 성능 우위를 유지하고 있음에도 오픈소스 진영의 추격 속도가 빨라지면서 모델 성능만으로는 지속가능한 경쟁우위 확보가 점차 어려워지고 있음

* 현재 이용가능한 AI 모델 중 가장 높은 수준의 성능과 역량을 갖춘 최첨단 모델

※ Epoch AI(2026.5.29)에 따르면 압도적 성능을 가진 프론티어 모델도 이제 4개월 후면 오픈소스 진영에서 더 낮은 비용으로 구현 가능

I 그림 1 I 폐쇄형-오픈소스 모델 간 Epoch Capabilities 인덱스 점수 비교



출처: EPOCH AI(2026.5.29.)

I 연산의 무게중심 이동: 학습에서 추론으로

- AI 활용이 확대됨에 따라 연산 수요의 중심도 모델을 만드는 ‘학습(training)’에서 실제로 모델을 사용하는 ‘추론(inference)’으로 이동
 - AI 모델을 만드는 과정이 ‘학습’이라면, 만들어진 모델을 실제 서비스에 적용하여 사용자의 입력에 대한 결과를 생성하는 과정은 ‘추론’
 - ‘학습’은 대규모 데이터를 반복 처리하면서 모델의 파라미터를 지속적으로 조정하는 과정으로, 순전파(forward propagation)*와 역전파(back propagation)**를 모두 수행하기 때문에 대규모 연산이 집중적으로 발생
 - * 입력값을 모델에 넣고 여러 층의 계산을 거쳐 예측 결과를 만드는 과정
 - ** 예측 결과와 정답 간 오차를 바탕으로 모델의 각 파라미터가 결과에 얼마나 영향을 미쳤는지 계산해 이를 조정하는 과정
 - 반면 ‘추론’은 학습이 완료된 모델*을 활용해 새로운 입력에 대한 결과를 생성하는 순전파 중심의 연산으로, 서비스 이용 시마다 반복적으로 발생
 - * 학습이 완료된 모델은 사전학습과 미세조정 등을 거쳐 모델 가중치(파라미터)가 확정된 상태로, 추론 과정에서는 일반적으로 파라미터를 추가 조정하지 않고 입력에 대한 결과를 생성
- 생성형 AI의 추론 과정에서는 새로운 토큰을 생성할 때마다 모델의 가중치와 이전 문맥 정보를 참조해 연산을 수행하기 때문에 답변이 길어질수록 전체 추론 연산량도 증가
 - 개별 추론의 연산량은 대규모 모델 학습보다 작지만, 이용자 수·호출 횟수·답변 길이가 늘어나면 전체적으로 막대한 연산 수요로 집적
 - 즉, AI 활용 확대가 토큰 사용량 증가, 추론 연산 수요 증가로 직접 연결되는 구조
- 여기에 추론 모델(reasoning model)의 확산은 하나의 과업을 처리하는데 필요한 토큰과 연산량을 더욱 증가시키는 요인으로 작용
 - 일반적인 생성형 AI 모델도 응답을 생성하는 과정에서 추론(inference)* 연산을 수행하지만, 추론 모델(reasoning model)은 최종 답변을 도출하기 위해 문제를 여러 단계로 분석하거나 중간 결과를 검토·수정하는 등 추가적인 추론(reasoning)** 과정을 수행
 - * inference: 학습된 모델이 입력을 받아 출력을 생성하는 실행 과정
 - ** reasoning: 답을 내기 위해 문제를 단계적으로 분석·검토하는 사고 과정
 - 이에 따라 하나의 요청을 처리하는 데 더 많은 토큰과 연산이 사용될 수 있으며, 사용자가 보는 최종 응답은 짧더라도 이를 생성하는 과정에서는 상당한 연산이 발생 가능

- AI 에이전트의 확산 역시 과업당 추론 수요 증가를 촉발
 - 기존 생성형 AI가 주로 ‘요청 → 추론 → 응답’을 통해 사용자의 질문에 답하는 방식이었다면,
 - 사용자를 대신해 행동까지 수행하는 AI 에이전트는 하나의 과업을 처리하기 위해 ‘요청 → 추론 → 검색·외부 도구 호출 → 결과 확인 → 재추론 → 추가 행동’의 과정을 반복적으로 수행
 - 이에 따라 하나의 사용자 요청이 여러 차례의 모델 호출과 도구 사용으로 이어지면서 과업당 토큰 사용량과 연산량이 크게 증가할 수 있음
- 특히 추론 모델과 AI 에이전트가 결합될 경우, 복잡한 추론과 반복적인 모델·도구 호출이 중첩되면서 하나의 과업을 수행하는 데 필요한 연산량은 더욱 증가
- 이에 따라 AI 활용이 본격화될수록 대규모 추론 수요를 빠르고 경제적으로 처리할 수 있는 인프라 역량*이 중요한 경쟁 요소로 부상
 - * 대규모 추론에서는 GPU의 연산 능력뿐 아니라 메모리 대역폭과 네트워크 등 시스템 전반의 처리 효율이 중요
 - 특히, 범용적인 과업에서 모델 간 성능 격차가 줄어들수록 단순한 사용자 규모뿐 아니라 토큰당 비용, 추론 효율, 하드웨어 가동률 등 증가하는 AI 수요를 얼마나 효율적이고 경제적으로 처리할 수 있는가가 경쟁력을 좌우

I 비용구조의 변화: 사용량 연동 비용의 비중 확대

- AI의 활용 확대와 추론 연산 수요의 증가는 서비스 사용량에 따라 반복적으로 발생하는 비용을 확대하면서 AI 산업의 비용구조에 변화를 초래
 - 모델 학습 비용은 특정 모델을 개발·학습하는 과정에서 대규모로 선투입되고, 학습이 완료된 후에는 서비스 이용량에 따라 반복적으로 발생하지 않는다는 점에서 선투자·고정비 성격이 강함
 - 반면 추론·활용 단계에서는 서비스 이용 시마다 연산 비용이 발생하며, 이용자 수, 호출 횟수, 처리되는 토큰 수가 늘어날수록 전체 비용도 함께 증가
 - 또한, 같은 이용량이라도 과업이 복잡해질수록 더 많은 추론 연산과 모델·도구 호출이 필요해 과업당 서비스 원가가 높아질 수 있음
 - 이는 한번 개발된 소프트웨어를 추가 사용자에게 제공하는데 드는 한계 비용이 매우 낮았던 기존 소프트웨어와는 구별되는 특징
- 즉, AI 서비스 이용 확대가 곧 추가적인 추론 비용 발생으로 이어지기 때문에, 이용 증가를 수익성 있는 사업모델로 연결하는 것이 중요한 과제로 부상

03 토큰 이코노미의 주요 개념과 산업적 의미

- 토큰 이코노미는 AI 서비스가 반복적으로 생산·소비되는 과정에서 토큰이 사용량과 추론 처리 규모를 계량하고 이를 비용 및 가격과 연결하는 경제적 단위로 기능하는 구조
 - 시스코(Cisco)는 토큰 이코노미를 “토큰을 작업의 기본 단위이자 현대 IT 스택의 주요 통화로 취급함으로써 인공지능의 비용, 가치 및 규모를 관리하는 데 사용되는 경제적 프레임워크”로 정의(Cisco, 2026)
- 즉, 토큰 이코노미의 의미는 토큰이 단순한 기술적 처리 단위를 넘어 경제적 계량 단위로 기능하며 이를 중심으로 AI 서비스의 경쟁·수익구조와 생태계의 성장 메커니즘이 변화하고 있다는 점에서 찾을 수 있음

I 경제적 계량 단위로서의 토큰

- 토큰은 본래 AI가 텍스트 등 정보를 처리하는 기술적 단위였으나, AI 활용이 확대되면서 AI 사용량과 자원 소비를 계량하는 경제적 단위로서 의미가 확대
- 특히, 사용자의 요청에 따라 실시간으로 연산자원이 소비되고 서비스 비용이 토큰 소비량과 밀접하게 연결되면서 토큰은 AI 사용량과 이에 수반되는 자원 소비를 측정하는 공통 단위로 기능
 - AI는 서비스가 개발된 이후에도 이용될 때마다 새로운 연산이 투입되며 동일한 한 번의 요청이라도 과업의 복잡성이나 처리 과정에 따라 필요한 연산량과 토큰 사용량이 크게 달라질 수 있어 사용자 수나 요청 횟수만으로 실제 자원 소비를 설명하기 어려움
 - 이에 따라 하나의 과업을 처리하는 데 얼마나 많은 토큰과 연산이 투입되는지가 추론 서비스의 처리 규모와 비용을 설명하는 중요한 변수로 부상
- 단, 토큰을 AI 지능이나 경제적 가치를 직접 측정하는 단위로 볼 수는 없음
 - 동일한 수의 토큰이라도 어떤 모델이 생성했는지, 어떤 과업에 사용되었는지에 따라 제공하는 성능과 경제적 가치가 다름
 - 또한, 모델의 구조, 추론 방식, 하드웨어 및 처리 방식 등에 따라 동일한 양의 토큰이라도 필요한 비용이 상이
- 즉, 토큰은 균질한 상품이라기보다 AI 사용량과 연산 비용을 연결하여 추론 서비스의 생산·소비 규모를 비교·측정할 수 있게 해주는 공통의 경제적 계량 단위로 이해하는 것이 적절

I 시의 생산요소화와 수익모델 다변화

- AI 모델의 성능향상과 추론·에이전트 기능의 고도로 AI가 단순한 정보 제공 도구를 넘어 실제 업무와 과업 수행에 직접 활용되면서 다양한 산업과 업무의 생산과정에 투입되는 범용적 생산요소로 확산되는 추세
- AI가 실제 업무 수행에 반복적으로 투입될수록 서비스 원가도 사용자 수보다는 실제 사용량과 과업 복잡성에 더욱 밀접하게 연동
- 이에 따라 AI 서비스의 과금방식과 수익모델도 실제 사용량과 창출가치를 반영하는 방향으로 다변화
 - 기존 소프트웨어는 정액제 또는 사용자 수 기반 과금이 일반적이었으나, AI 서비스에서는 토큰·호출량 등 실제 사용량을 직접 가격에 반영하는 방식이 확대
 - 동시에 고객의 비용 예측 가능성을 높이기 위해 정액 구독에 일정 사용량을 포함하고, 이를 초과하면 추가 과금하는 혼합형 모델도 확산
 - AI 에이전트와 같이 과업 수행과정에서 여러 차례의 모델 호출과 도구 사용이 발생하는 서비스에서는 내부적인 토큰 소비량을 고객에게 그대로 전가하기보다 과업의 완료나 성과를 기준으로 가격을 책정하는 방식도 등장

표 1 | AI 서비스의 주요 수익·과금 모델과 사례

과금 방식	구조	실제 사례	토큰 이코노미 관점의 의미
사용량 기반	입력·출력 토큰 등 실제 사용량에 비례해 과금	OpenAI API는 모델별로 입력·출력 토큰 100만 개당 가격을 제시 ¹	토큰이 원가와 매출을 직접 연결하는 경제적 단위로 기능
정액 + 사용량 혼합형	월 구독료에 일정 사용량을 포함하고 추가 사용분은 별도 과금	Cursor는 월구독 요금제에 포함사용량을 제공하면서 초과·일부 기능에 사용량 기반 요금을 적용 ²	사용량 변동성이 큰 AI 서비스에서 고객의 예측 가능성과 공급자의 원가 회수를 동시에 추구
사용자 수 + 성과 기반 혼합형	기본 SW는 사용자 수로 과금하고 AI 기능은 성과 발생 건수로 별도 과금	Intercom은 기본 플랫폼은 사용자 수 기반으로 과금하면서 FinAI Agent는 outcome당 과금 ³	기존 SaaS 가격체계 위에 AI가 실제로 만들어낸 결과 단위가 추가
성과 기반	토큰이나 호출량이 아니라 완료된 업무·결과에 대해 과금	Sierra는 에이전트 서비스에 대해 토큰이 아니라 결과에 대해 지불하는 outcome-based 모델을 제시 ⁴	내부적으로 토큰 비용을 관리하되 고객에게는 과업 가치 자체를 판매

1 <https://developers.openai.com/api/docs/pricing?tab=suite>

2 <https://prod.cursor.com/help/account-and-billing/overages>

3 <https://www.intercom.com/pricing>

4 <https://sierra.ai/product/horizon>

- 즉, 공급자에게는 토큰이 사용량과 원가를 측정·관리하는 핵심 단위로 기능하나, 고객에게는 토큰 자체보다 서비스·과업을 통해 얻는 가치가 비용 지불의 기준이 될 가능성
 - 공급자 내부에서는 토큰을 중심으로 원가와 자원 사용을 관리하더라도, 고객에게는 사용량·과업·성과 등 다양한 단위로 가격을 제시하는 구조가 확대될 가능성
- 다만 AI 서비스의 원가가 사용량에 민감해지고 있다고 해서 낮은 가격에 토큰을 공급하는 것 자체가 곧 경쟁력을 의미하는 것은 아님
 - 범용 모델의 성능 대비 가격이 빠르게 하락하고 있지만 최고 수준의 프런티어 모델에는 여전히 프리미엄이 존재한다는 점에서 궁극적으로는 고객이 지불할 만한 성능과 과업 가치를 제공할 수 있는지가 중요
- 결국 AI 서비스의 수익성은 단순히 토큰당 가격을 낮추는 데 있기보다, 과업 수행에 필요한 토큰과 연산을 효율적으로 관리하면서 고객이 지불할 만한 가치를 창출할 수 있는가에 의해 좌우

I AI 생태계로의 파급효과

- AI의 활용 확대는 AI 산업이 대규모 인프라에 대한 선행 투자에 의존하여 성장하던 단계에서 실제 활용 수요가 AI 생태계 전반의 투자와 공급 확대를 견인하는 구조로 발전하는 기반이 될 수 있음
 - AI 서비스와 에이전트의 확산으로 증가한 추론 수요가 다시 컴퓨팅과 인프라 투자를 촉진하면서 실제 수요와 투자가 연결되는 성장의 선순환 구조가 형성될 수 있기 때문
- 즉, 'AI 서비스·에이전트 활용 확대 → 추론 수요 증가 → 컴퓨팅 수요 증가 → 인프라 투자 확대 → AI 공급역량 확충 및 비용 절감 → AI 서비스 개선 및 활용 확대'로 이어지는 지속가능한 성장구조로의 전환이 가능
- 특히 AI 에이전트가 사용자를 대신해 다른 AI 모델과 디지털 서비스를 자동으로 호출·이용하게 되면 AI 서비스의 수요 주체가 인간 사용자에서 AI 에이전트로까지 확대되며 수요가 증가할 가능성
 - 이는 AI 간 서비스 호출과 거래를 증가시키면서 인간 사용자의 직접적인 요청을 넘어서는 새로운 형태의 AI 수요를 창출할 수 있음
- 물론, 추론 수요의 증가 자체가 경제적 가치 증가를 의미하지는 않기 때문에 이러한 선순환이 지속되기 위해서는 늘어난 AI 사용과 연산 투입이 실제 생산성 향상, 새로운 서비스 가치 창출, 기업의 수익으로 연결되는 '가치 증명'이 필요

04 소결 및 시사점

- 토큰 이코노미의 부상은 AI가 연구개발 중심의 기술에서 지속적으로 생산·소비되는 경제적 서비스로 전환되고 있음을 시사
- 이에 따라 AI 산업의 경쟁도 ‘누가 가장 좋은 모델을 보유하고 있는가’에서 ‘필요한 AI 역량을 얼마나 효율적으로 공급하고 이를 실제 활용으로 확산시키는가’의 문제로 확대
 - 토큰은 AI의 최종적인 경제적 산출물이 아니며 토큰당 비용이 낮아지고 소비량이 증가하더라도 그것이 생산성 향상이나 기업의 수익 증가로 이어지지 않는다면 AI 산업의 지속적인 성장을 담보하기는 어려움
- 결국 AI 산업화가 진행될수록 경쟁의 초점은 토큰당 비용에서 과업당 비용으로, 나아가 과업당 창출 가치로 확장될 가능성
- 따라서 토큰 이코노미가 던지는 궁극적인 질문은 ‘늘어난 토큰 소비가 실제 경제적 가치와 수익으로 연결되는가’이며,
- 향후 AI 경쟁력은 더 많은 토큰을 생산·소비하는 능력 자체보다, AI를 통해 실제 과업을 효율적으로 수행하고 이를 얼마나 높은 경제적 가치로 전환할 수 있는가에 의해 좌우될 것

 참고문헌

이경선(2026.8.31), “AI 추론 수요의 확대와 토큰 이코노미의 부상”, KISO 저널, 이슈 인사이드, <https://journal.kiso.or.kr/contents/detail/1690>.

André, C., Bélin, M., Gal, P., & Peltier, P.(2025.6), “Developments in Artificial Intelligence Markets: New Indicators Based on Model Characteristics, Prices and Providers”, OECD Artificial Intelligence Papers, No. 37, OECD Publishing, Paris.

Cisco(2026), “What is AI Tokenomics?”, <https://www.cisco.com/site/us/en/learn/topics/artificial-intelligence/ai-tokenomics.html>.

Cursor 사이트, <https://prod.cursor.com/help/account-and-billing/overages>.

EPOCH AI(2026.5.29.), “Open models lag state-of-the-art closed models by 4 months”.

Gartner(2026.3.25), “Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025”.

Gartner(2026.5.19), “Gartner Forecasts Worldwide AI Spending to Grow 47% in 2026”.

Intercom 사이트, <https://www.intercom.com/pricing>.

OpenAI 사이트, <https://developers.openai.com/api/docs/pricing?tab=suite>.

Sierra 사이트, <https://sierra.ai/product/horizon>.

KISDI Perspectives 발간 내역

KISDI Perspectives는 국내외 정보통신미디어 관련 주요 정책 및 동향을 분석한 리포트입니다.

문의: 노희윤 부연구위원(정보통신정책연구원 미디어정책연구실, hyooooon@kisdi.re.kr, 043-531-4042)