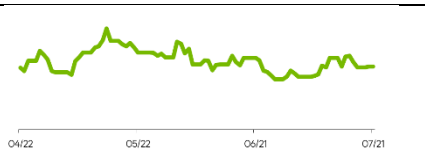
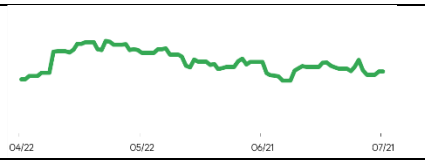
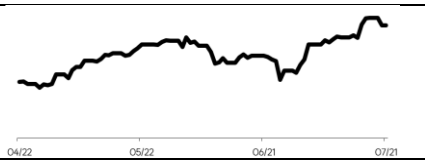
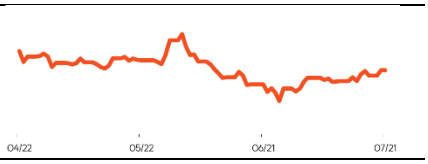
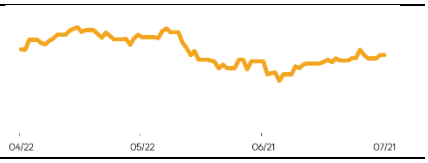
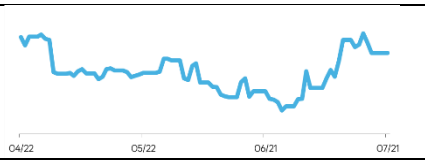
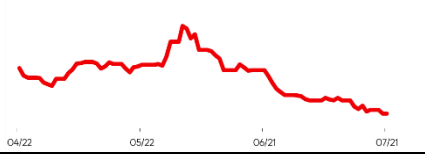
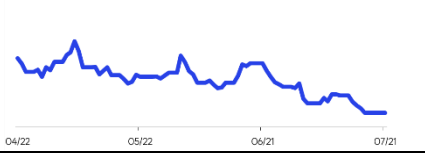
2026-07-22

박제민의 Q&A.I

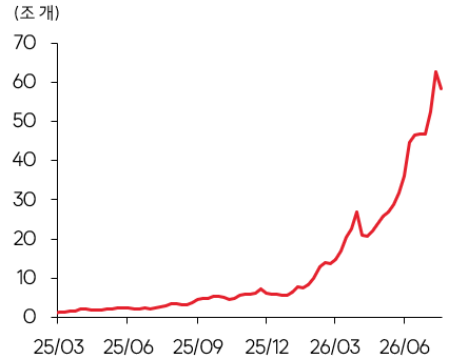
SK securities
SK증권 해외주식/AI. 박제민
jeminwa@sks.co.kr

	엔비디아	시가총액: 4,919	12MF PER: 20	
		-독보적인 GPU 기술력있으나 ASIC 으로 독점력 축소 -Rubin Ultra 기술 전환 및 CPU, LPU 주문 추이 주목		
	알파벳	시가총액: 4,282	12MF PER: 27	
		- TPU 활용 Anthropic 독주로 인프라 해자 인정받는 중 - Coding Agent 역전, 소비자용 Agent 출시 주목		
	애플	시가총액: 4,797	12MF PER: 37	
		-메모리 가격 상승이 브랜드 파워로 극복 가능한지 시험 중 -Siri 고도화로 인한 소비자용 Agent 출시 주목		
	마이크로소프트	시가총액: 2,988	12MF PER: 21	
		-OpenAI 해자 약화로 인프라 사업부 성장률 축소 -자체 모델 및 커스텀 AI 도입 컨셉 수요 및 기술력 주목		
	아마존	시가총액: 2,689	12MF PER: 29	
		-양대 AI Labs 모두에 공격적 투자를 통한 고객 확보 성공 -인프라 외 Bedrock 등 AIaaS ARR 전망에 주목		
	메타 플랫폼스	시가총액: 1,639	12MF PER: 19	
		-플랫폼 인기 여전, Muse Spark 개발로 유통과 모델 확보 -플랫폼 데이터, 글래스 등 활용한 소비자용 Agent 주목		
	오라클	시가총액: 350	12MF PER: 15	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		
	코어위브	시가총액: 40	12MF PER: -	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		

주: 시가총액 십억달러

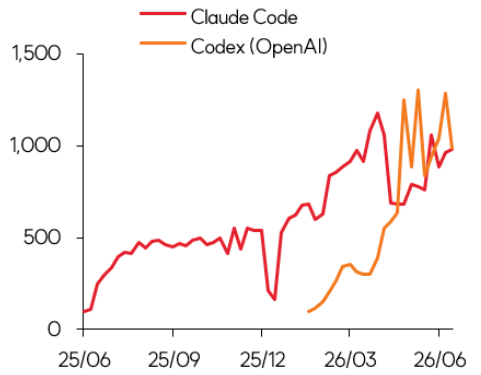
Update Chart

OpenRouter 토큰 사용량



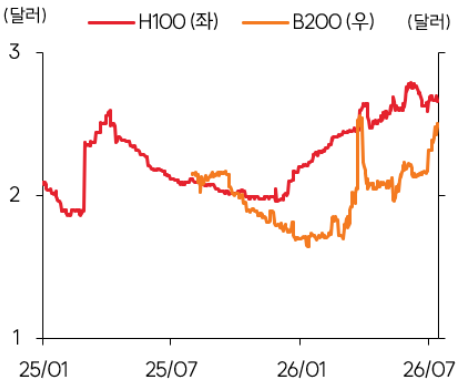
자료: OpenRouter, SK 증권

주요 Coding Agent 활용 강도 (백만회 달성일 = 100)



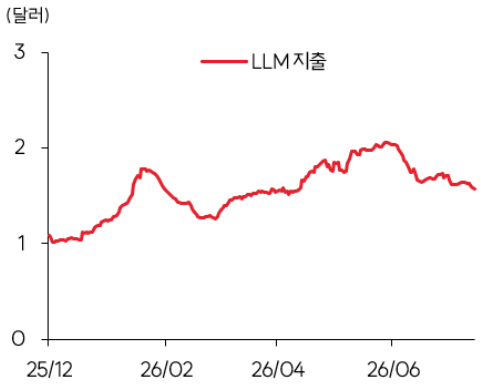
자료: NPMjs, SK 증권 / 주: 업데이트 및 다운로드 수 기준. Codex 5월 수치 일부 조정

GPU 렌탈 가격 추이 (SiliconData)



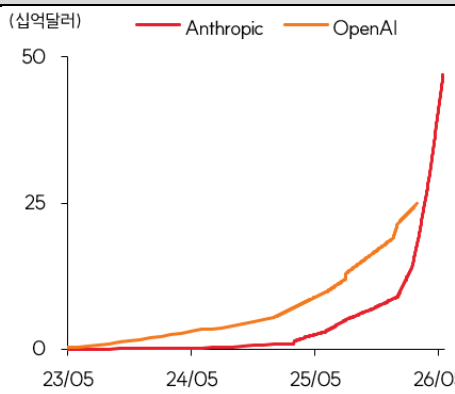
자료: Bloomberg, SK 증권

LLM 지출 단가 추이 (SiliconData)



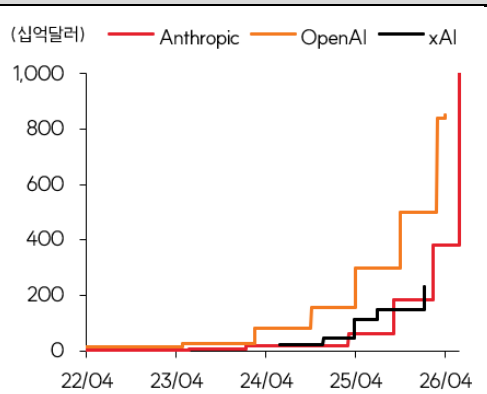
자료: SimilarWeb, SK 증권

주요 AI Labs ARR 추이



자료: EpochAI, SK 증권

주요 AI Labs 기업가치 추이



자료: EpochAI, SK 증권

중국 오픈웨이트 모델의 부상

중국산 프론티어 모델의 등장

7/17~7/20 진행된 상하이 세계인공지능대회(WAIC)에 맞춰 역대 가장 조밀한 중국 프론티어 오픈웨이트 릴리스 러시

Kimi K3, Qwen3.8, DeepSeek V4, GLM-5.2, MiniMax M3, Hunyuan Hy3 모두 6~7 월달에 릴리즈

Kimi K3 는 A.A 벤치마크 기준 Claude Opus 4.8 을 앞서는 성능 공개. 미국 모델 중 Fable, GPT 5.6 Sol 에만 뒤지는 모습

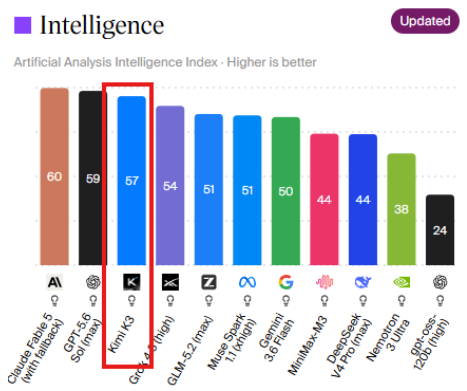
코딩 에이전트 벤치마크는 SWE, Terminal Bench 를 포함하여 Kimi 가 우세한 벤치마크가 다수 존재

기존 오픈웨이트 모델들과 다르게 100 만 토큰 당 $3/$15 의 가격 책정 (Sonnet 급). Frontier 모델 가격 수준 책정은 처음

Alibaba 는 아직 Private 하게만 공개된 자신들의 모델인 Qwen 3.8-Max 를 Fable 모델로 자평

Kimi K3 는 2.8 조개, Qwen 3.8 은 2.4 조개 파라미터 모델로 기존 오픈웨이트 모델 대비 '체급 대형화'한 모델

Kimi Artificial Analysis

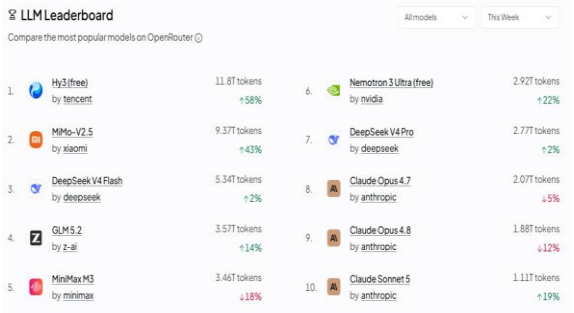


자료: Artificial Analysis, SK 증권

Alibaba: Qwen 은 Fable 다음 수준으로 강력

자료: China Post, SK 증권

OpenRouter 사용 랭킹 (2026.7.13~2026.7.20)



자료: OpenRouter, SK 증권

중국 오픈웨이트 모델 규모 및 특징 비교

모델	총 파라미터	활성 파라미터	가격 (입력/출력, 100만 토큰당)	특징 한 줄
Kimi K3	2.8T	896개 중 16개 활성 (미공개)	$3 / $15	최대 규모·프리미엄 가격, 가중치 7/27 공개 예정
Qwen3.8-Max	2.4T	미공개	프리뷰만 (표준가 -10%)	"Fable 5 다음" 주장, 오픈 "곧"·독립 벤치 없음
DeepSeek V4-Pro	1.6T	49B	$0.435 / $0.87	최저가권·MIT, 100만 컨텍스트 오픈웨이트 최고 코딩
GLM-5.2	753B	~40B	$1.40 / $4.40	가성비 오픈웨이트, 채택 속도 최상위
Tencent Hy3	295B	21B	(공개가 미확인)	경량·효율 플레이, Apache 2.0·256K 컨텍스트

자료: 언론 종합, SK 증권

Kimi K3 가 정말로 성능이 좋은걸까?

7/16 목요일 Kimi K3 공개 이후 주말간 개발자들은 실질 성능 검토. 일부 실망스러운 결과 도출

Simon Willison 의 '자전거 타는 팰리컨 그려줘' 테스트에서 미국 선두 업체 대비 3.9 배 이상 인풋 토큰 활용

고등 IMO 문제 풀이에서 Kimi K3 1,545k output token, 17.4 시간 소요. Fable 705k output token, 2.5 시간 소요

Agent 성능은 파라미터 체급과 벤치마크만으로 측정 불가. 특히 최근 Token Optimization 트랜드와 더불어 가성비 중시

벤치마크는 회계사 시험을 푸는 능력을 알려준다면 Agent 는 회계 부서 업무 수행 능력(정확도, 시간)과 수행 가성비가 중요

Open-weight 모델을 다운로드하여 로컬에 활용할 경우 무료이지만 서비스할 경우 여전히 추론 비용 발생

유료 서비스에 활용할 경우, 효율은 곧 사업자 이익률. 모델 효율이 곧 칩의 tokens-per-watt 과 유사한 역할

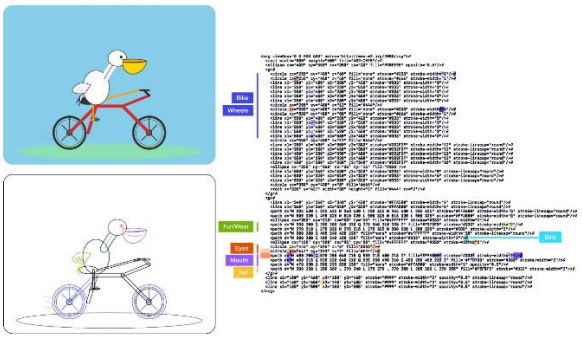
즉, AI 회계 업무 서비스 사업자의 경우 Kimi K3 를 활용할 경우 토큰 당 단가 외에도 고려할 사항이 많은 것

추론 성능을 결정하는 요소

1. **Model footprint**: 모델 크기. 각 서비스 시 마다 필요한 연산 및 메모리량 결정
2. **Inference efficiency**: 추론 당 토큰 사용량. MoE 와 같은 아키텍처적 선택은 생성된 토큰 당 연산 감소
3. **Memory efficiency**: 메모리 효율. 아키텍처적으로 KV 캐시 요구량을 줄여 GPU 활용도 증가 가능. 대표적 기법으로 Kimi 의 Delta Attention, DeepSeek 의 Compressed Sparse Attention 존재
4. **Serving efficiency**: 서빙 효율. 배칭, 스케줄링 등의 최적화는 GPU 활용도를 극대화하고 요청 간 작업을 공유
5. **Token efficiency**: 정답에 도달하는 데 필요한 토큰 수가 적을수록 추론 비용이 낮아짐. CoT 효율 및 답변 효율 등으로 귀결

Kimi K3 는 MoE 분산율이 높고 메모리 활용을 잘한 모델로 설명되지만 전반적인 추론 성능 결정 요소들에서 미국 모델 대비 열위로 판단. 결정적으로 미국 선두 AI Labs 의 경우 모델 파라미터를 포함한 추론 성능 결정 요소에 대한 성능 지표 미공개

Simon Willison 팰리컨 테스트



자료: X(Simon Willison), SK 증권

IMO 문제 해결에 걸린 시간 및 토큰 수 비교

Wall-clock minutes per problem (all rounds)

Run	P1	P2	P3	P4	P5	P6	Total
Claude Fable 5	5	27	73	8	10	26	150
GPT-5.6 Sol (xhigh)	7	106	27	10	20	60	231
Kimi K3	20	85	491	40	28	381	1045
GPT-5.6 Sol (default)	5	15	15	5	8	14	60

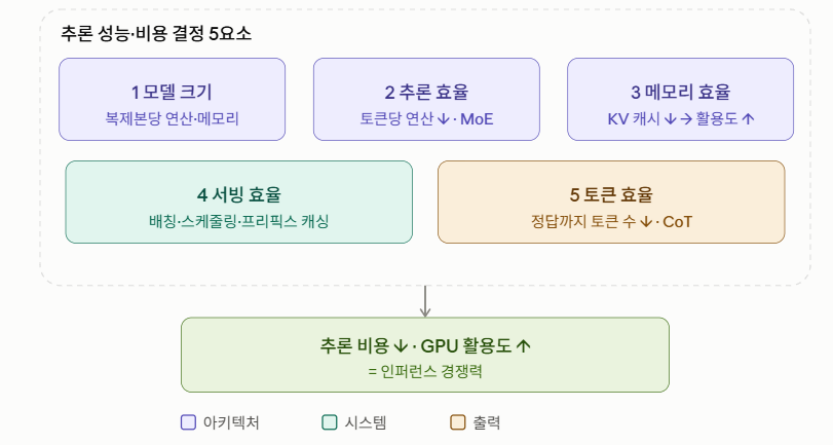
Completion (output) tokens per problem

Includes billed reasoning tokens where the provider kills them as output (Kimi K3, GPT-5.6 Sol).

Run	P1	P2	P3	P4	P5	P6	Total
Claude Fable 5	16,826	141,363	331,025	38,027	55,114	191,922	705,277
GPT-5.6 Sol (xhigh)	8,488	112,359	42,359	23,693	14,997	34,972	236,172
Kimi K3	27,657	171,310	846,686	70,540	47,540	381,100	1,548,783
GPT-5.6 Sol (default)	6,880	17,800	56,674	9,664	9,227	21,420	20,215

자료: 언론 종합, SK 증권

추론 성능-비용 결정 5 요소



자료: Stratechery, SK 증권

AI Labs: 경쟁 고도화 진입

선두 AI Labs 는 중국 모델사들 대비 여전히 많은 무기 보유

1) 재무적 시너지: 이미 높은 매출 기록하며 추론-훈련 시장간 재무적 시너지 기대 가능

2) Lock in 효과: Coding Agent 사용자들이 UI/UX 에 익숙해지면서 sticky 한 경향을 보이는 중

3) Data Flywheel: 서비스를 시행하며 얻게되는 코드, Agent 경험을 Post training 에 활용하며 모델 성능 재귀적 개선

Kimi K3 의 모회사 Moonshot 은 이번 모델 릴리즈로 사용자가 몰리며 신규 가입자 서비스 중단

최근 Moonshot AI 가 $30B+ 밸류 펀딩 마무리 중, 주주결의 배포로 6 개월 내 홍콩 상장 준비 착수. 자금적 기반 마련 시도

성능 발전과 함께 재무적 체급을 갖추면서 중국 모델사들과의 경쟁이 더 치열해질 것으로 전망

CSP: 경쟁 고도화의 이점과 서비스 원가 절감을 동시에

AI Labs 경쟁 고도화는 곧 서버 수요 증가로 직결, 선두 AI Labs 들의 공격적 투자 의지 지속될 가능성

다만 선두 AI Labs 에 집중됐던 기존 Backlog 의 가치는 일부 희석되는 성격 존재

현재 CSP 3 사에서 OpenAI, Anthropic 이 차지하는 Backlog 규모는 $1T 이상, 독점적 지위 훼손은 곧 해당 수주잔고의 훼손

그러나 최전방 기업 고객의 AI 수요만 지속된다면 Backlog 훼손 이후에도 서버를 이용할 플레이어는 있을 것

SageMaker, Azure Foundry, Vertex 와 같은 AIaaS(또는 TaaS)에 선택지 증가. 효율화로 기업 고객의 도입 계기로 작용 가능

추후 Copilot 와 같은 SaaS 서비스에도 엔진을 오픈웨이트 모델로 활용할 여지 존재

Microsoft 는 Copilot 을 OpenAI-Anthropic 에서 Kimi K3 로 전환하여 추론 비용을 최대 $600M 절감할 수 있다고 계산

AI Infra 는 대형화를 좋아해

사용하는 메모리나 연산의 절대적인 양은 모델 체급이 늘어나면서 모두 증가, 효율화하는 신기술들이 많은 부분 공개

Kimi K3 같은 대형 모델의 경우 GB200/GB300 NVL72 랙의 Scale Up 성능에 많은 부분 의존. 신규 인프라 수요 견인

메모리의 경우 MXFP4(4 비트) 양자화 기술로 2.8T 파라미터를 1.4TB (FP16 이라면 5.6TB)로 압축. Kimi Delta Attention (KDA) 기술로 KV Cache 효율화

기존 대형 모델사들의 경우 아키텍처 공개가 제한적으로, 해당 기술력이 진정 메모리 효율화인지 판단하기 어려운 상황

최근 Z. AI 의 중국산 칩 활용 1GW AI 데이터센터 완공 보도 존재. 중국이 AI 모델과 같이 인프라 자립 꾀할 경우 리스크

Moonshot AI 펀딩 시작

Moonshot Plans IPO in Six Months After China AI Breakthrough

People visit the Moonshot AI stand, featuring Kimi K3, during the World Artificial Intelligence Conference in Shanghai on July 17. *Source: AFP/Getty Images*

자료: Bloomberg, SK 증권

Z.AI 중국산 칩 활용 데이터 센터 완공

Z.ai powers up a 1-gigawatt AI data center built entirely on Chinese chips, report claims — GLM developer now runs multiple 10,000-chip clusters with zero Nvidia silicon

News By Luke James | Published 8 hours ago

The blacklisted Chinese AI lab's gigawatt-class site is among the largest built by any Chinese model developer.

(Image credit: UCG)

자료: Tom's Hardware, SK 증권

Microsoft 가 Kimi K3 를 고려 중이라는 루머

Vaibhav Sisinty
@VaibhavSisinty

Microsoft is reportedly testing Kimi K3 to power their flagship Copilot AI assistant. Saving up to 60% per token.

That's $600M saved on every $1B spent on inference.

Kimi K3 already runs on Azure. K2.7 Code is already inside GitHub Copilot. This isn't new territory for Microsoft. It's the next step.

And on July 27, Moonshot is open-sourcing the full 2.8 trillion parameter weights. Once that happens, Microsoft can self-host it on Azure instead of paying API fees. The cost drops even further.

The biggest American tech company. Choosing a Chinese open-source model over its own partner's models. Because the math doesn't lie.

Open source isn't just competing anymore. It's winning the contracts.

자료: X, SK 증권

Kimi K3 사용자 폭주로 신규 서비스 중단

Kimi K3 developer suspends new subscriptions amid compute constraints

Listen

By SCMP
Published Jul 21, 2026 4:04 pm KST

People walk past the Moonshot booth promoting its Kimi K3 AI model, during the World Artificial Intelligence Conference (WAIC) in Shanghai, China, July 17. Reuters-Yonhap

자료: Korea Times, SK 증권

Compliance Notice

작성자(박제민)는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.

당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.