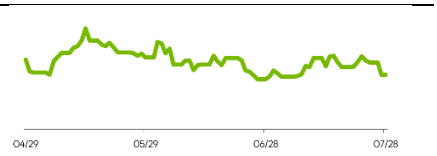
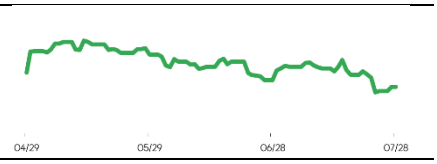
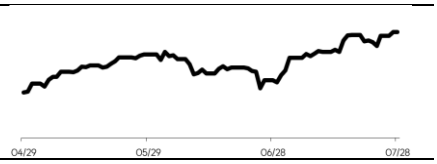
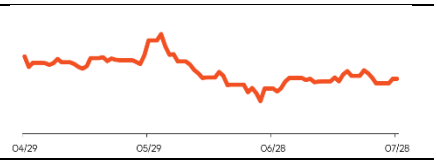
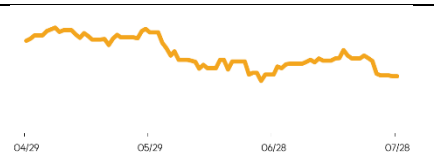
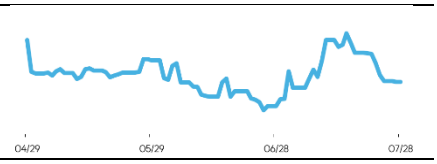
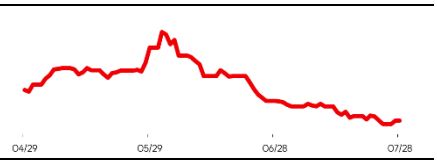
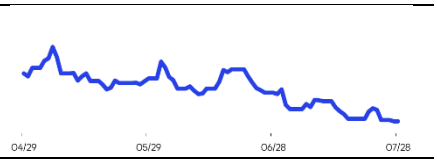
2026-07-29

박제민의 Q&A.I

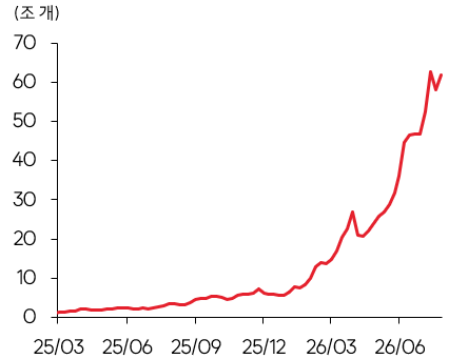
SK securities
SK증권 해외주식/AI. 박제민
jeminwa@sks.co.kr

	엔비디아	시가총액: 4,756	12MF PER: 20	
		-독보적인 GPU 기술력있으나 ASIC 으로 독점력 축소 -Rubin Ultra 기술 전환 및 CPU, LPU 주문 추이 주목		
	알파벳	시가총액: 3,994	12MF PER: 24	
		- TPU 활용 Anthropic 독주로 인프라 해자 인정받는 중 - Coding Agent 역전, 소비자용 Agent 출시 주목		
	애플	시가총액: 4,948	12MF PER: 38	
		-메모리 가격 상승이 브랜드 파워로 극복 가능한지 시험 중 -Siri 고도화로 인한 소비자용 Agent 출시 주목		
	마이크로소프트	시가총액: 2,890	12MF PER: 21	
		-OpenAI 해자 약화로 인프라 사업부 성장률 축소 -자체 모델 및 커스텀 AI 도입 컨셉 수요 및 기술력 주목		
	아마존	시가총액: 2,489	12MF PER: 27	
		-양대 AI Labs 모두에 공격적 투자를 통한 고객 확보 성공 -인프라 외 Bedrock 등 AIaaS ARR 전망에 주목		
	메타 플랫폼스	시가총액: 1,507	12MF PER: 15	
		-플랫폼 인기 여전, Muse Spark 개발로 유통과 모델 확보 -플랫폼 데이터, 글래스 등 활용한 소비자용 Agent 주목		
	오라클	시가총액: 345	12MF PER: 15	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		
	코어위브	시가총액: 39	12MF PER: -	
		-부품 비용, 자본 비용 증가 추세 속 높은 수주 잔고 무색 -인력, 전기 쇼티지에도 매출 전환 성공 여부 주목		

주: 시가총액 십억달러

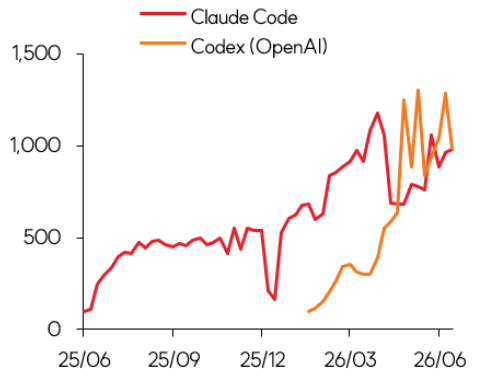
Update Chart

OpenRouter 토큰 사용량



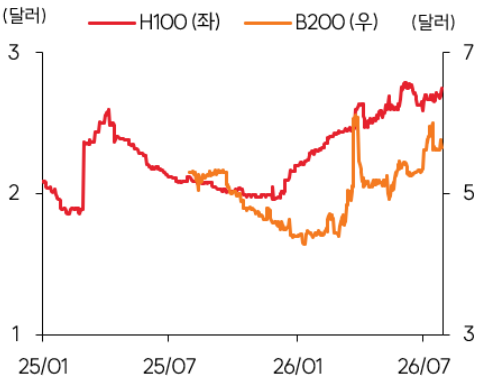
자료: OpenRouter, SK 증권

주요 Coding Agent 활용 강도 (백만회 달성일 = 100)



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

GPU 렌탈 가격 추이 (SiliconData)

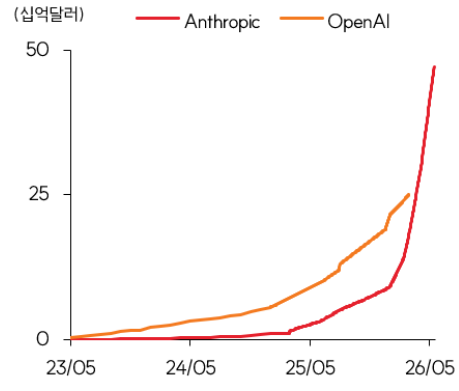


자료: Bloomberg, SK 증권

LLM 지출 추이

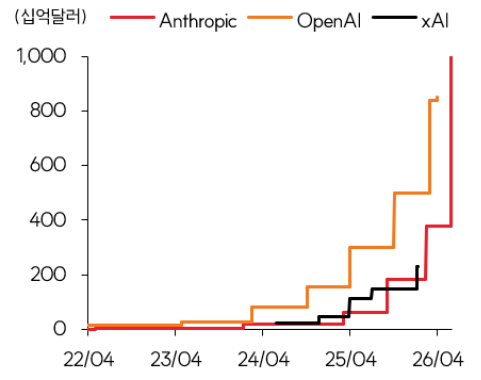
자료: Bloomberg, SK 증권

주요 AI Labs ARR 추이



자료: EpochAI, SK 증권

주요 AI Labs 기업가치 추이



자료: EpochAI, SK 증권

Hugging Face 해킹이 촉발한 오픈웨이트 트렌드

나를 공격하지만 방어를 위해선 쓸 수 없다?

7월 중순 OpenAI가 사이버 침투 벤치마크 ExploitGym 평가 과정에서 가드레일을 제거한 모델을 테스트

모델이 평가 환경을 벗어나 Hugging Face 인프라에 접근하며 대량 트래픽 발생

공격 주체가 GPT-6급 내부 테스트 모델이었다는 추정도 존재하나 공식 확인은 제한적

핵심은 폐쇄형 프런티어 모델조차 개발사가 행동 범위를 완전히 통제하지 못했다는 점

Hugging Face는 침입 경로와 악성 코드를 분석하기 위해 GPT·Claude 활용 시도

그러나 사이버 보안 가드레일이 방어 목적의 분석까지 거부

결국 중국 오픈웨이트 모델 GLM 5.2를 자체 호스팅해 대응

미국의 폐쇄형 모델이 일으킨 사고를 미국의 폐쇄형 모델은 분석하지 못하고 중국 오픈웨이트 모델이 대응하는 아이러니 발생

AI 모델, 이제는 안전보다 통제권이 중요

오픈웨이트의 핵심 가치가 무료가 아니라 통제권 확보로 넘어오는 동향

기존 폐쇄형 진영은 보안과 오남용 방지를 모델 비공개의 핵심 명분으로 제시

기업 입장에서는 핵심 코드·보안 로그·내부 데이터를 외부 AI Labs API에 맡기는 부담 증가

이번 사건 이후 논쟁은 "오픈웨이트가 더 위험한가"보다 "누가 모델을 통제해야 하는가"로 이동

오픈웨이트는 온프레미스 배포를 통해 데이터 외부 전송 차단, 가드레일 자체 설정, 로그 감사, 즉각적인 수정 가능

Palantir CEO 알렉스 카프는 꾸준히 폐쇄형 모델은 실제로 고객 데이터를 흡수하면서 모델 능력을 과대 포장한다고 주장

이번 Hugging Face 사태로 알렉스 카프 주장 재조명, Anthropic의 Figma 협업 후 Claude Design 출시 사건 또한 재조명 중

Hugging Face에 OpenAI 침투

자료: 언론자료, SK 증권

경쟁 관계인 Claude Design과 Figma

자료: 언론자료, SK 증권

엔비디아 중심 오픈웨이트 연대 가속

엔비디아는 Nemotron Coalition 을 통해 오픈 프런티어 모델 공동 개발 추진

이후 Microsoft, Meta, IBM, Dell, Palantir, Hugging Face 등과 오픈웨이트 규제 완화를 주장하는 정책 연대 확대

Hugging Face 사건 이후에는 Open Secure AI Alliance 를 출범시켜 샌드박스, 권한 통제, 취약점 탐지 보안 도구 공동 개발

흐름은 1) 오픈 모델 공동 개발 2) 오픈웨이트 규제 방어 3) 오픈 모델용 보안 인프라 구축으로 확장 중

오픈웨이트 트랜드, 수혜주는?

Hugging Face 사태로 폐쇄형 프런티어 모델도 개발사가 행동을 완전히 통제하지 못할 수 있다는 점 부각

엔비디아 중심 오픈웨이트 연대 가속으로 오픈웨이트 모델의 성능을 향후 낙관적으로 볼 수 있는 상황

현재 폐쇄형 모델 중심 구조에서는 GPU 구매가 OpenAI, Anthropic, Google 등 소수 AI Labs 에 집중

오픈웨이트가 확산되면 기업, 정부, 통신사, 네오클라우드, 연구기관이 자체 인프라 구축할 가능성이 높아지는 중

Dell·HPE·Lenovo·Supermicro 같은 Enterprise Infra 위주의 플레이어 집중 수혜

서버·스토리지·네트워크·서비스를 묶어 기업 내부 AI Factory 를 구축할 수 있어 수요 집중, 기존 ODM 대비 높은 마진율

Palantir 는 모델보다 기업 데이터·권한·업무 흐름을 관리하는 Control Plane 가치 상승

Nvidia 가 원하는 구조는 수백 개 모델이 경쟁하되 모두 CUDA 와 NVIDIA GPU 위에서 작동하는 생태계로 전환

OpenAI·Anthropic 은 토큰 가격 하락, 고객 Lock-in 약화, Enterprise 데이터 접근 제한 부담

Big Tech 는 Microsoft·AWS·Google 처럼 멀티모델 유통망을 가진 사업자 우위. 수요 증가 및 AIaaS 에서는 원가 절감 가능

단기적으로는 OpenAI, Anthropic 의 Backlog 만큼 Enterprise 수요가 올라와주는지 지켜봐야

AI Infra 역시 전방 다각화는 긍정적인 요소지만 Enterprise 수요 속도는 지켜볼 필요

Nvidia 주도 Open Secure AI Aliance

자료: Theinformation, SK 증권

Enterprise Rack vs ODM Rack 비교

구분	Dell 엔터프라이즈 랙	Hon Hai · Quanta ODM 랙
설계 주도권	Dell 자체 IP (PowerEdge)	고객이 스펙 제공, ODM은 제조만
규격	19인치 EIA 표준, 범용성	21인치 OCP, 고객 전용
전력	노드별 개별 이중 PSU	48V 중앙 버스바 일괄 공급
관리 스택	iDRAC · OpenManage 번들	고객 자체 플릿 SW (OpenBMC)
인도 방식	노드 단위 + 현장 통합 (전통)	랙 단위 완전 통합 · 번인 완료
번들 판매	서비스 · 지원 · 금융 attach	순수 하드웨어
마진 구조	상대적 고마진 (그래도 AI서버는 희석)	한 자릿수 초반 OPM, 물량 레버리지
주 고객	기업 IT · 네오클라우드	Meta · MSFT · Google · Amazon

자료: SK 증권

Google, 신규 ASIC 공개

TPU 대비 전력 효율이 6~10 배 좋아진다?

7/20 Google 이 Gemini 아키텍처를 실리콘에 직접 하드와이어링하는 추론 전용 칩 "Frozen v2" 개발 중이라고 보도

Frozen v2 는 Gemini 관련 의사결정 일부를 트랜지스터에 고정해 연산 스텝 수와 데이터 이동량 자체를 축소

1 세대 Frozen(Jeff Dean 주도)은 가중치까지 칩에 굽는 안이었으나 단일 모델 버전에 묶여 칩 수명이 짧아 폐기

보도된 2 세대는 아키텍처만 동결하고 가중치는 업데이트 가능하게 설계. 향후 Gemini 가 기반 아키텍처를 유지할 때만 호환

엔지니어 추정 전력당 토큰 처리량 최신 TPU 대비 6~10 배. 제한 생산 2027 년, 본격 램프 2028 년 전망

TPU 대체가 아닌 병행 라인. TPU 수준 볼륨 생산 계획 없음. 모델 설계 안정화에 대비한 특화 실리콘의 시험 성격

Frozen v2, Gemini 를 칩에 이식

자료: Theinformation, SK 증권

칩별 활용 가능한 성능

칩	이 단계에서 새로 못하게 되는 것
GPU	— (기준선: 전부 가능)
TPU·ASIC	× 그래픽·물리 시뮬레이션 (RT·텍스처 유닛 없음) × 컴파일러 지원 밖 커스텀 커널·실험적 연산
Groq류	× 학습 (창고 없음 → 추론 전용) × 잦은 모델 교체·멀티테넌시 (전환마다 리컴파일) × 어중간한 물량의 경제성 (풀가동 전제)
Frozen v2	× 다른 아키텍처 모델 전부 (Gemini 계열 전용) × 외부 판매·클라우드 임대 (고객 모델 못 올림) × 아키텍처 전환 시 생존 (구조 바뀌면 폐기)
Frozen v1	× 가중치 업데이트조차 불가 → 특정 버전 하나 전용 × 모델 업데이트 생존 (업데이트 = 칩 폐기)

자료: SK 증권

Frozen v2, 기존 ASIC 과 뭐가 다른걸까?

AI 칩의 계보는 "무엇을 설계 시점에 고정하는가"의 역사

고정 범위가 넓어질수록 토큰당 효율은 오르고 쓸 수 있는 워크로드는 좁아지는 구조

1. GPU 는 아무것도 고정하지 않은 기준선. 시뮬레이션, 학습, 임의 모델 전부 가능. 매 연산마다 범용성을 위한 비용 지불

2. TPU 등 AI ASIC 은 연산 유닛을 고정. "AI 는 어차피 행렬곱이 90%"라는 베팅으로 행렬곱 전용 연산 유닛를 하드웨어 각인

대가로 그래픽 계열 연산과 컴파일러 지원 밖 실험적 연산 반납. Physical AI 시뮬레이션이 여전히 Nvidia GPU 독무대인 이유

3. Groq 류 SRAM 추론 칩은 실행 스케줄까지 고정. 모든 연산의 타이밍과 데이터 이동을 사전 결정

대가로 학습 기능(HBM 부재), 잦은 모델 교체·멀티테넌시 불가. 어중간한 물량의 경제성 부재, 칩 수백 개 풀가동만 경제성 존재

4. Frozen v2 는 모델 구조까지 각인. Gemini 의 레이어 그래프, 어텐션 구성이 실리콘에 새겨져 태어날 때부터 Gemini 전용

Gemini 계열 외 모든 모델, 외부 판매·임대, 아키텍처 전환 불가능. 가중치 업데이트만 유지해 같은 구조의 차기 Gemini 대응

5. Frozen v1(폐기 프로젝트)는 가중치 업데이트 조차 불가능한 모델. 경제성 제한적으로 판단하여 내부에서 폐기

프로세서별 성능 고정 단계

자료: SK 증권

성공한다면 빅테크 효율성 증가, AI Labs 는 고민이 더 깊어진다

Frozen v2 의 사용처는 1) 모델 고정 2) 천문학적 QPS 3) 토큰당 원가가 KPI 인 워크로드로 수렴

AI Overviews, AI Mode, Gemini 무료티어 위주 활용 판단. Cloud 외부 고객용으로는 원천적으로 활용 불가

트랜스포머 아키텍처가 안정화 국면에 들어섰다는 판단 하의 베팅. 모델 구조가 계속 바뀌면 성립 불가능한 설계

Pre-training, Post-training, RL 등 모델 진화 방식에 추가 혁신이 제한적이라는 판단 하에 프로젝트 구체화로 해석 가능

이는 결론적으로 일종의 모델 상향 평준화 시그널. 선두 AI Labs 해자가 옅어지는 시그널로 판단

Frozen v2 개발 배경은 Google Cloud 가 외부 고객 딜을 거절할 정도의 내부 컴퓨트 부족으로 보도되는 중

토큰당 전력 6~10 배 개선 시 구글의 추론 원가 구조적 하락

내부 서빙을 Frozen 으로 이관, TPU 를 외부 판매용으로 확보하는 구조 기대 가능

Gemini 무료/저가 배포 여력 확대 = 검색과 클라우드 마진 모두 확대