



SK증권 해외주식/AI. 박제민
jeminwa@sks.co.kr

Google Update: 인력 이탈로 불거지는 Gemini 이슈

두 달 만에 Gemini 테크리드 3인 중 2인 공석

(8/5) Gemini 공동 책임자 Jeff Dean 퇴사, neolab 'Discovery Loop' 설립 — 구글 최고 수준 엔지니어 3인 동반 이탈

DeepMind 공동 창업자 겸 전 CEO Demis Hassabis 역할 축소

잔류한 유일한 Gemini 공동 책임자 Koray Kavukcuoglu(전 DeepMind CTO)가 DeepMind/Gemini 총괄 승계

6월 Shazeer 의 OpenAI 이동까지 감안 시 테크리드 3인 중 2인이 2개월 만에 공석

DeepMind CEO 직책 자체 폐지, 후임 미임명

Google 은 해당 뉴스가 불거진 8/5 장중 -5.4% 저점 기록

Flash 만 덩그러니.. Gemini 3.5 Pro 는 언제?

반년 전인 2025년 말 Gemini 3는 전 경쟁사를 상회하며 OpenAI의 'code red' 선언을 유발한 모델

그러나 Gemini는 현재 지속적인 모델 성능 부재, 지연에 시달리는 중

(7/21) Gemini 3.6 Flash 공개, 성능은 중국 오픈웨이트 모델 하회 + 전작 3.5 Flash 대비 개선폭 미미

Pro 라인 현재 공백 상태. 이번 Flash 역시 효율 지표만 개선: 출력 토큰 17% 절감, 소요 시간 50% 절감, 가격 17% 인하

Artificial Analysis 창업자, 최근 순위 재편에서 구글이 처음으로 상위 5개 랭에서 탈락했다고 언급

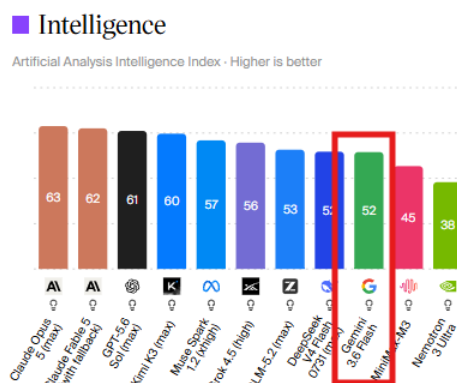
Gemini 3.5 Pro가 요원해진 상황에서 Gemini 4.0 관련 코멘트가 나오는 중

구글 핵심 인력 4인이 창업한 Discovery Loop



자료: 언론 자료, SK 증권 / 좌측부터 Jeff Dean, Sanjay, Quoc, Oriol

Gemini AA Intelligence Benchmark 순위



자료: Artificial Analysis, SK 증권

Gemini 3.5Pro 는 언제부터 밀린걸까?

5/19 (I/O): Pichai 무대 발언 "다음 달" > 6 월 타겟 제시, 실제로는 3.5 Flash 만 출시

7/16: Bloomberg, 현·전직 직원 10 명 인용, 코딩 성능 내부 기준 미달, 6 월 말 학습데이터 업데이트가 오히려 성능 악화

7/21: Flash 3 종(3.6 Flash / 3.5 Flash-Lite / 3.5 Flash Cyber)만 출시, 발표문 말미에 Gemini 4 사전학습 착수 언급

2개월간 3 회 이상 지연 끝에 Pro 라인 자체가 스킵된 것

지속된 모델 지연은 컴퓨트 배분권이 문제?

(8/5) SemiAnalysis, 리더십 개편을 계기로 DeepMind 의 프론티어 랩 지위에 사실상 사망 선고

프론티어 복귀 확률 0 으로 규정, Gemini 3 Pro 가 정점이었다고 판정하며 강한 비판

진단의 핵심은 인사가 아닌 배분권. 컴퓨트 배분 주도권이 리서치에서 클라우드 세일즈로 완전 이전된 구조적 전환

향후에도 '근본적인 신념 부족'으로 SOTA 도달 불가능 주장, Thomas Kurian 은 컴퓨트 배분 통제권을 GCP 위주로 활용할 것

NeoLabs 인수가 유일한 길이나 Dylan Patel 은 해당 가능성도 사실상 0 으로 진단하며 구글 관료 문화 비판

Consumer Agent 는 완전히 다른 환경일 가능성이 높다

Coding Agent 가 Gen.AI 의 유일한 전방이라면 Dylan Patel 의 비판은 유효

그러나 구글이 궁극적으로 집중할 것으로 보이는 Consumer Agent 영역은 현재와 다른 벤치마크 필요할 것으로 판단

AI Input 확보 디바이스(AR 안경), 다양한 개인적 맥락의 포착 및 해석, 낮은 Latency, 장기 메모리 등이 필요

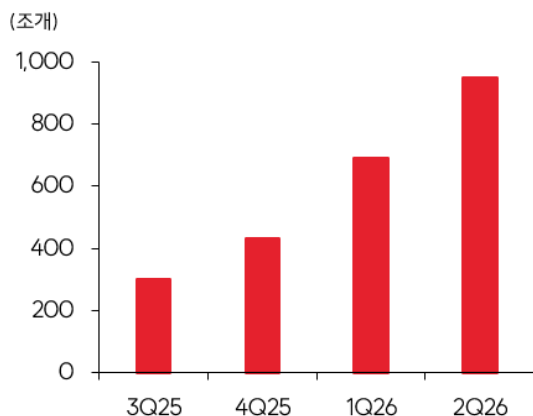
업무용 Agent 가 하네스 완성도로 성능이 갈리는 전장이라면, Consumer Agent 는 생활 맥락을 읽어내는 별개의 전장

Consumer 영역의 방향성이 미확정인 만큼, 일단 인프라 선투자로 대응하는 판단을 했을 가능성

반면 동일 Consumer Tech 기업인 Meta 는 SOTA 확보를 위한 공격적 투자 지속 중

소비자의 선택을 누가 가져가는지 주목 필요

구글 Gemini 1P API 토큰 사용량 (월간)



자료: Google, SK 증권

Semianalysis 의 강한 Deepmind 비판



자료: Semianalysis, SK 증권

SpaceX, 이제 로켓이 아니라 컴퓨트를 판다

실적 발표에서 보여준 공격적 용량 확보 움직임

2Q26 nameplate 컴퓨팅 용량 1.4GW 로 마감, 연말 2GW 이상 전망

2027 년 말 누적 온라인 컴퓨트는 하이싱글 GW 수준 목표

전력·냉각 기준 20GW 온라인화 목표, 발전소 단에서 최소 15GW 이상 확보 예상

GPU 소요분 대비 과잉 수준의 전력·냉각·전기 설비 선확보 전략

로켓·위성 기술력을 지상 데이터센터에 이식 중, 독보적 데이터센터 증설 속도 목표에도 강한 자신감

\$100B ARR 과 '1년 미만 회수'

2Q26 AI 부문 매출 \$2.6B, 이 중 \$1.6B 가 Google 과의 컴퓨트 계약

2Q26 AI 컴퓨팅 Capex \$15.8B, 향후 2 개 분기도 유사 수준 가이드

신규 컴퓨팅 자본 투입에 대해 1년 미만 회수 기간 주장

12 월 예상 매출 기준 연말 \$100B ARR 도달 궤도 제시

분기 \$15.8B 가 4 개 분기 지속 가정 시 2026 년 AI Capex \$63B, 연말 ARR \$100B 대비 산술적으로 1년 내 회수 성립

최근 Colossus·Colossus II 사이트에 대한 클라우드 계약을 맺음. 대상자는 Google(Anthropic 향)으로 추정

투자자들이 추정하는 단가는 시간당 \$12 수준의 가격으로 기존 시장 가격 대비 1.5~2 배 수준의 ASP 로 계약

실질 유효기간은 이미 짧은 상황

명목 계약기간 2026 년 10 월~2029 년 6 월, 그러나 구글 컨콜에서는 Pichai CEO 는 a six-month deal 로 지칭

10/1 개시 + 12/31 직후 통지 + 90 일 = 정확히 4/1. 첫 해지 가능 시점 종료가 구글의 기본 시나리오로 판단

SpaceX 입장에서 Colossus 1·2 대비 Grok 수요처 미확보 상태, 추후 회수 옵션 확보 + 높은 단가로 거절 유인 부재

높은 단가가 유지될 수 있을까?

출발점은 Anthropic 의 1H26 급격한 성장, Coding 이 Agent 로 작동하기 시작하자 개발자 수요 급증

Anthropic 이 단기간 대량 컴퓨트 확보에 나선 배경: 1) 초기 고객 확보가 이후 Binding 으로 연결 2) 토큰당 과금 구조상 유입되

는 대로 높은 추론 마진 3) 확보 고객이 사전학습 데이터·하네스 워크플로우 기반 제공

당시 CSP 3사와 네오클라우드 전부 공급 제약 상태

Google 은 단기 손실을 감수하고도 Anthropic 확보 의지, TPU 대량 활용 고객이자 관계 심화 기대

결과적으로 'SpaceX 대용량 서버를 높은 단가에 단기 임대'하는 계약 체결

계약에 상호 해지 조항 존재. 자체 증설 완료 시 고가 컴퓨트 반납 후 Anthropic 에 직접 공급 전환 가능

즉, 지금의 임대분이 향후 구글 직접 매출로 넘어오는 설계인 것

이를 위해 구글이 컴퓨트 일부를 Anthropic 쪽으로 선회했을 가능성, Gemini 3.5 Pro 지연과도 연결 관계 존재할 것으로 판단
Anthropic은 궁극적으로 Enterprise Tech 기업으로 구글 Consumer 축과의 직접 충돌 낮음

AWS 3년 vs SpaceX 1년

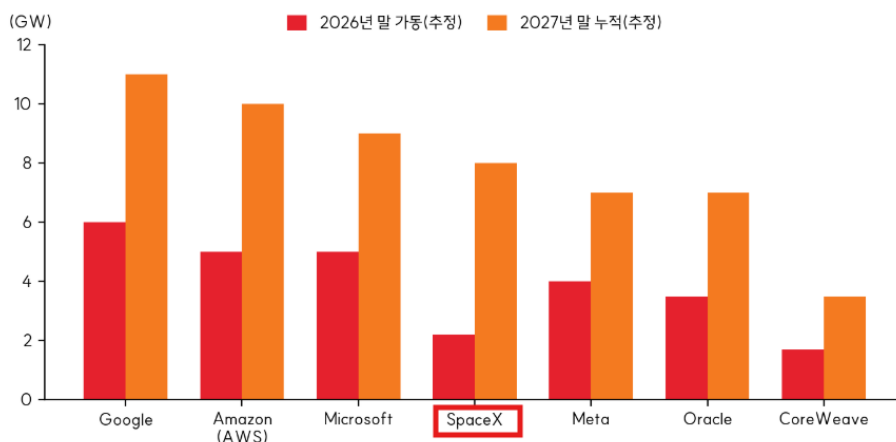
SpaceX는 해당 계약을 근거로 'AI Capex 회수 기간 1년 미만' 주장. 특수 거래 조건을 일반화한 세일즈 화법으로 판단
유인도 존재: 지분 보유 인력의 보호예수 미종료 구간
체결 계약 기준 산식이므로 허위는 아니며, 특수성만 설명되지 않는 구조
대조군은 AWS — 최근 계약 기준 3년 회수 기간 제시, 대상은 OpenAI·Anthropic이 유력
AWS 계약이 IaaS보다 PaaS 성격으로 더 많은 것을 제공함에도 단가는 더 낮음
즉, SpaceX-Google 단가는 시장 가격이 아니라 공급 제약 국면의 일시적 프리미엄인 것으로 판단

왜 엔비디아만인가?

서버를 전량 NVIDIA로 구축 결정

- 엔지니어 모델 지향: Physical AI는 CUDA 위에서 학습·추론돼야 하는 영역, 아직 ASIC 해자 낮음
 - 엔비디아의 Backstop(수요·잔여가치 보증성 지원) 기대
- 'Grok이 단연 최고의 엔지니어가 될 것'으로 코멘트 데이터 우위 기반 버티컬 포지셔닝
Grok 5에는 SpaceX가 25년간 생산한 전체 데이터 코퍼스 반영 예정

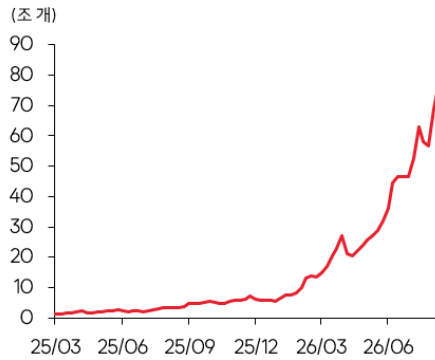
SpaceX 포함한 빅테크 DC Capa 비교 (추정치)



자료: 각 사, SK 증권 추정

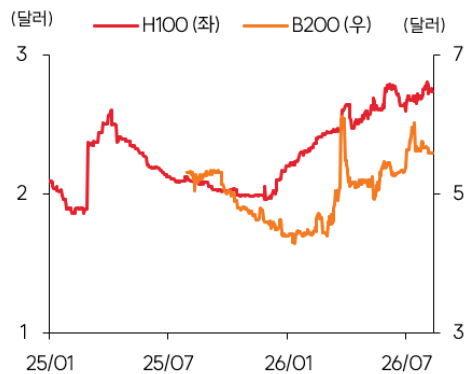
Update Chart

OpenRouter 토큰 사용량



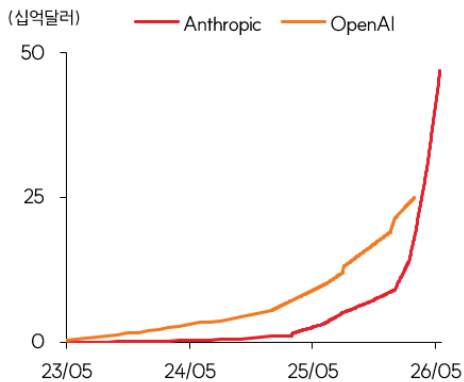
자료: OpenRouter, SK 증권

GPU 렌탈 가격 추이 (SiliconData)



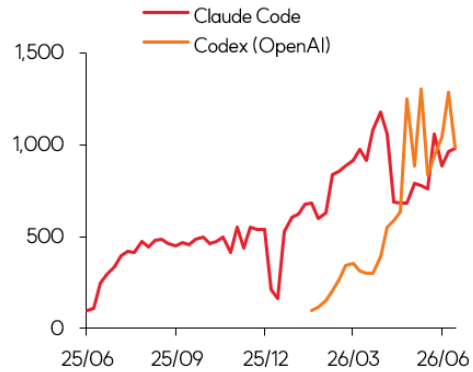
자료: Bloomberg, SK 증권

AI Labs ARR 추이



자료: EpochAI, SK 증권

주요 Coding Agent 활용 강도 (백만회 달성일 = 100)



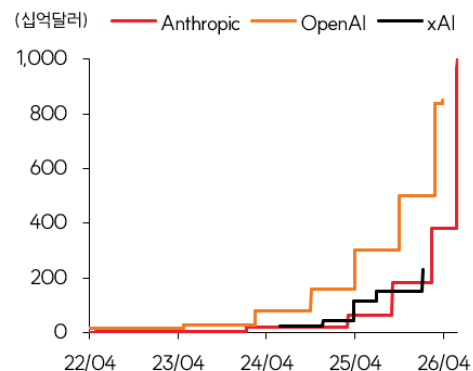
자료: NPMjs, SK 증권 / 주: 업데이트 및 다운로드 수 기준. Codex 5월 수치 일부 조정

LLM 지출 단가



자료: SimilarWeb, SK 증권

주요 AI Labs 기업가치 추이



자료: EpochAI, SK 증권

Compliance Notice

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

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

당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.