



SK증권 해외주식/AI. 박제민
jeminwa@sks.co.kr

Token Optimizing 의 시대: LLM 지출 단가 하락 사이클

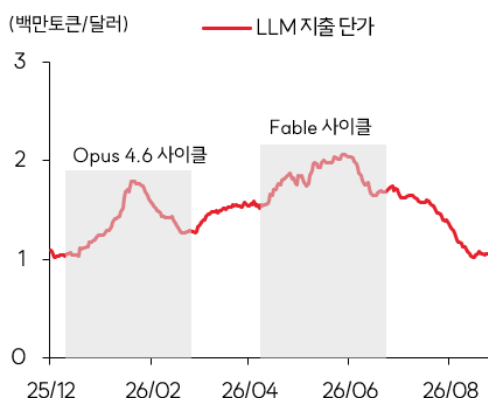
고객은 Frontier 프리미엄을 오래 지불하지 않는다

LLM 지출 단가(총 지출액 ÷ 총 토큰=백만토큰당 달러)가 25/12 \$1.0 > 26/06 \$2.0 > 26/08 약 \$1.05 으로 하락 복귀
단가 상승 구간은 신규 플래그십 출시와 일치: 2월 고점은 Opus 4.6 출시 직후, 6월 고점은 Fable 5 출시 직후
단가 하락 구간은 저가 모델 물량 확대와 일치: 7/30 Luna 80% 인하, 8/13 3.7 Flash 반값, 8/21 Sol 인하가 순차 반영
지출 단가는 가격표가 아니라 믹스 지표, 비싼 모델이 나오면 고객이 물려 단가가 오르고 저가 모델로 이동하면 내리는 구조
두 사이클에서 보이는 프론티어 프리미엄이 지속되는 기간은 2달 내외. 이후에는 사용자 유인 감소한다고 판단

한 달간 프론티어 3사 모두 가격 하락

Anthropic Opus 5 (7/24): In/Out \$5/\$25 로 전작 Opus 4.8 과 동일 가격, Fable 5(\$10/\$50) 대비 절반 가격에 근접 성능
같은 날 Effort 설정(low/medium/high) 도입, 가격표는 동일하나 태스크당 청구액을 고객이 직접 조절 가능한 구조로 전환
OpenAI 1차 (7/30): 출시 3주 만에 GPT-5.6 Terra 20% 인하(\$2/\$12), Luna 80% 인하(\$0.20/\$1.2)
Anthropic (8/10~11): Sonnet 5 도입가 \$2/\$10 을 표준가로 확정, 9/1 예정이던 \$3/\$15 인상 취소
Google (8/13): Gemini 3.7 Flash 를 도입가 \$0.75/\$3.75 로 3.6 Flash 출시가(\$1.5/\$7.5)의 절반 수준으로 출시
OpenAI 2차 (8/21, Reuters): 플래그십 GPT-5.6 Sol 20% 이상 인하, \$5/\$30 > \$4/\$20
주요 언론사들은 가격 인하를 고가 모델 배포를 꺼리는 기업 고객의 비용 민감도와 중국 스타트업 경쟁으로 보도

LLM 지출 단가 추이



자료: Bloomberg, SK 증권

OpenAI의 공격적 가격 프로모션

OpenAI

Lower prices for GPT-5.6 Sol

Same Sol intelligence. Lower API pricing. More room to build.

Build with GPT-5.6 Sol

We've lowered API prices for GPT-5.6 Sol. Following the [recent Terra and Luna price reductions](#), this makes the intelligence you use for your hardest work more affordable. GPT-5.6 Sol's promotional pricing is available at least through [November 21, 2026](#).

Starting today, you get the same Sol intelligence at a lower price. Sol now costs \$4 per million input tokens and \$20 per million output tokens, 20% lower input pricing and 33% lower output pricing.*

자료: OpenAI, SK 증권

DeepSeek 는 반대로 가격 인상?

DeepSeek (8/6): API 전반 "상당한" 인상 예고, 인상 폭·시점 미공개, 개발자에게 "사용량을 미리 계획하라"는 공지
 현행 V4-Flash \$0.14/\$0.28 은 5/22 영구 인하 이후 ByteDance·Tencent 가 따라 내린 중국 시장의 바닥가
 7월 중순 이미 피크·비피크 요금제 도입(베이징 9~12시, 14~18시 2배 요금), 두 달 내 두 번째 가격 조정
 배경은 수요 폭주와 컴퓨트 부족, 하루 8조 토큰 처리량으로 보도 되는 중
 하루 8조 토큰이면 원가 이하 판매가 물리적으로 불가능한 구간, "토큰 가격은 무한히 내려간다"는 서사의 하한선이 처음 숫자로 드러난 사례로 판단
 최근 \$50B 규모 1GW 내뿜는 데이터센터와 \$8B 펀딩 추진 병행 뉴스, IPO 관련 루머도 지속 보도되는 중
 즉, 프론티어 랩은 경쟁 압력으로 마진을 깎아 내리고 저가 랩은 컴퓨트 부족으로 올리는 양방향 수렴이 진행 중인 것으로 판단

고객의 Token Optimizing, 단가는 내려가는데 올라가는 청구서

FinOps Foundation 2026 에서 엔터프라이즈 API 24억 건 분석: 블렌디드 단가 1Q25 \$18.4 > 1Q26 \$6.07 로 67% 하락
 동시에 기업 73%가 AI 비용 계획 초과 응답, 평균 기업 AI 예산 2024년 \$1.2M > 2026년 \$7M
 AI 지출 관리 실무자 비중 2025년 31% > 2026년 98%, 클라우드 비용 관리 조직이 1년 만에 토큰 비용 관리 조직으로 전환
 원인은 Agentic Workflow 의 토큰 승수: 채팅은 1회 왕복, 에이전트는 수십 스텝 루프, 매 스텝마다 누적 컨텍스트 재전송

DeepSeek 가격 인상 보도

DeepSeek Plans 'Significant' Price Increase for AI Services

By Vlad Savov and Saritha Rai
 August 6, 2026 at 11:53 AM GMT+9
 Updated on August 6, 2026 at 1:02 PM GMT+9

Save Translate

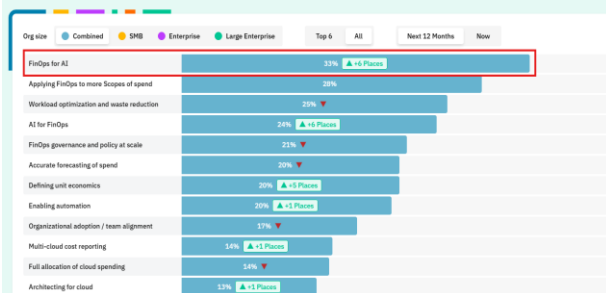
Takeaways by Bloomberg AI

DeepSeek plans to implement a significant price increase across its AI services, an unusual shift from the Chinese player whose low-cost models have put pressure on US and domestic rivals.

자료: Bloomberg, SK 증권

FinOps 내에서 AI 지출 관리 관심이 높아지는 중

Current and future priorities



자료: datafinops, SK 증권

중국 빅테크 실적발표: 비슷한 문법으로 후행

알리바바: AI 성장에도 본업 악화, 투자 가속화로 그룹 이익 악화

FY27 1Q 매출 RMB 269B(\$39.6B), +9% YoY, 클라우드와 쿼터머스가 견인

클라우드 외부 매출 +45% YoY, 22분기 최고치 기록, 9분기 연속 가속

조정 EBITDA RMB 27.3B(\$4.0B), -30% YoY, AI 기술 투자 확대가 클라우드 이익 개선을 상쇄

영업현금흐름 RMB 22.9B(\$3.4B), +11% YoY, FCF -RMB 44.7B(-\$6.6B), 전년 대비 적자 폭 2.4배 확대

CapEx RMB 67.7B(\$10.0B), 3개년 RMB 380B(\$56B) 계획 중 6월 말 누적 RMB 190B(\$28B) 집행, 진도 50%

CapEx 급증 사유 3가지: 1) 하드웨어 사이클 변동(분기별) 2) Agent 수요 대응 CPU 조달 확대 3) 광범위한 부품 가격 상승

바이두: AI 성과 애매, 구글과 다르게 검색 광고 성과 부진

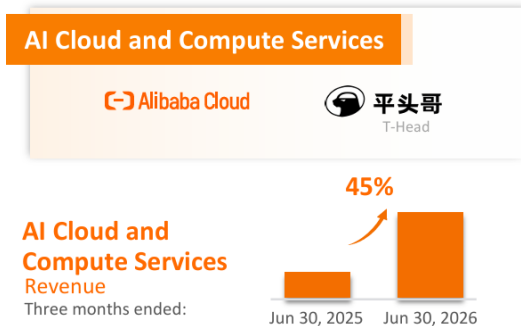
두 회사 모두 "검색 광고 회사에서 AI 풀스택 회사로"라는 같은 전환을 진행 중

구글은 검색·클라우드·자체집·로보택시 네 축에서 모두 결과가 나오기 시작했고, 바이두는 클라우드와 로보택시는 잘 되는데 본업인 검색 광고가 빠지는 상황

구글은 AI Overviews 와 AI Mode 를 검색에 통합하면서 오히려 검색 횟수가 늘고 광고가 2Q26 +17% 성장. 바이두도 AI 검색으로 사용자 활동은 크게 늘었지만 광고 수익화를 일부러 보류 중이라 광고 매출은 하반기까지 압박 지속 전망

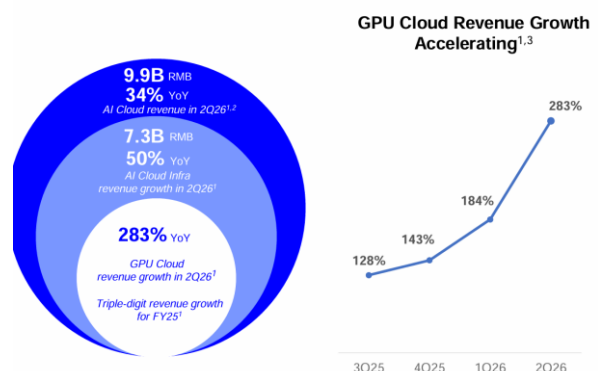
구글을 포함한 미국 빅테크들은 클라우드 백로그 제시. 바이두는 AI Cloud +50%, GPU Cloud +283% 성장률만 제시

알리바바 AI 관련 사업부 성장을 가속화



자료: Alibaba, SK 증권

바이두 AI 관련 사업부 성장을 가속화



자료: Baidu, SK 증권

알리바바 vs 아마존: 같은 프레임, 다른 신뢰도

두 기업은 규모부터 상이. AWS의 이번 분기 QoQ 증분 매출 \$4.6B = 알리바바 외부 클라우드 분기 매출 전체(\$5.2B)

CapEx에서도 아마존 분기 \$53.1B = 알리바바 3개년 총 예산 (\$56B), 아마존은 알리바바의 3년치를 한 분기에 집행

유사점 1. ROIC 프레임: Jassy "3년 미만 손익분기, 내용연수 5~6년, 이후 2~3년 FCF" vs Toby Xu "3년 내 손익분기, 내용연수 5년, 이후 2년 FCF". 문장 구조 동일, 알리바바가 아마존 프레임 차용으로 판단

유사점 2. 부품 인플레이: 아마존 CapEx 가이드 \$200B > \$220B 상향 사유 "메모리 비용 상승" 명시 / 알리바바 "칩 부품 가격 상승". 아마존은 기존 장기계약 가격 고정으로 원가 상승 흡수

차이점 1. 수익성 층위: AWS 39.3%는 감가상각 후 영업이익률, 알리바바 12%는 감가상각 전 EBITDA 마진, 직접 비교 불가. 알리바바 2030 목표 총마진 20%가 AWS 현재 영업이익률보다 낮음

차이점 2. 본업 캐시카우: 아마존 광고 \$19.8B +26%, 북미 리테일 영업이익 \$9.1B, 채권 발행으로 조달 / 알리바바 이커머스 EBITDA 포함, CMR -7%, "선불금 수취·공동 구축" 명시. 자기 대차대조표만으로 페이스 유지가 빠듯하다는 신호로 해석

가속기를 넘어서는 중국 ASIC

알리바바의 자체칩 부서 T-Head: ARM CPU·AI 가속기·인터넥트까지 서버 내 주요 칩 자체 설계 완료

M890(내부명 PPU1.5): HBM 144GB, 인터넥트 800GB/s, FP32~FP4 학습·추론 단일 칩. 산술상 H2O의 3배급 주장
핵심은 칩이 아닌 슈퍼노드 128장 슈퍼노드 구성, 64장까지 풀 대역폭. 이번 실적발표에서 M890 슈퍼노드 인스턴스 상용 출시
2조+ 파라미터 추론 구동, Kimi K3·Qwen 3.8Max가 외부 고객 MaaS에 실사용 중

자체 스위치 칩으로 128장 슈퍼노드를 프로덕션에 올린 것은 시스템 단위에서 엔비디아를 카피하기 시작했다는 의미로 판단

PPU 누적 출하 56만 장(5월), CEO "전세대 50만 장+ 출하". 고객 400개+(5월) > 650개+(8월 초)

병목은 파운드리. Alibaba Cloud 내 자체칩 배포 비중 아직 낮음. 하반기 공급 램프업 예정에도 공급을 걱정거리로 언급

알리바바 자체 칩 및 자체 모델 구축



자료: 언론 자료, SK 증권

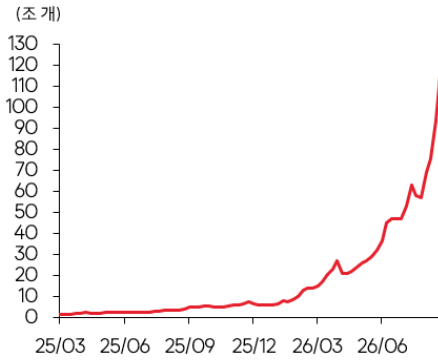
시스템 단위 설계에서 자체칩 비중 증가



자료: 언론 자료, SK 증권

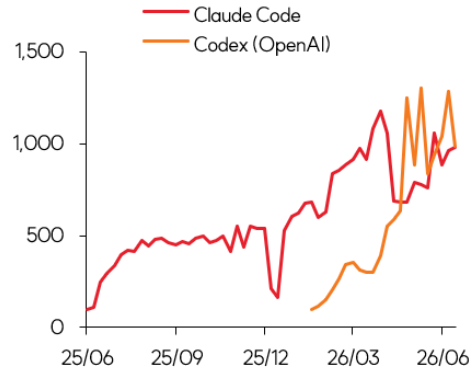
Update Chart

OpenRouter 토큰 사용량



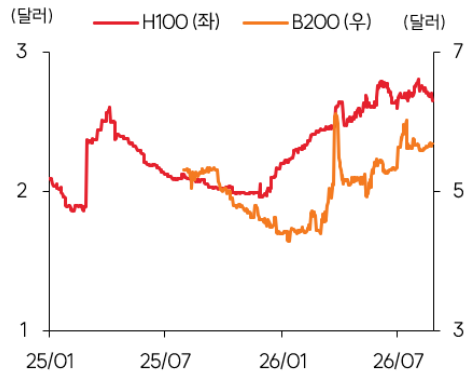
자료: OpenRouter, SK 증권

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)



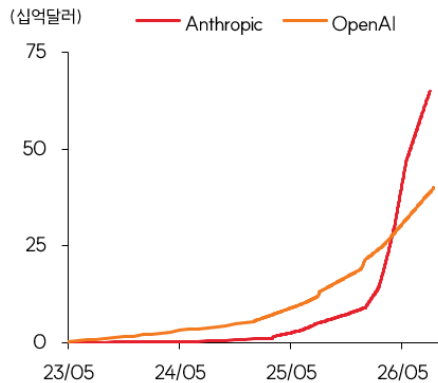
자료: Bloomberg, SK 증권

LLM 지출 단가 추이



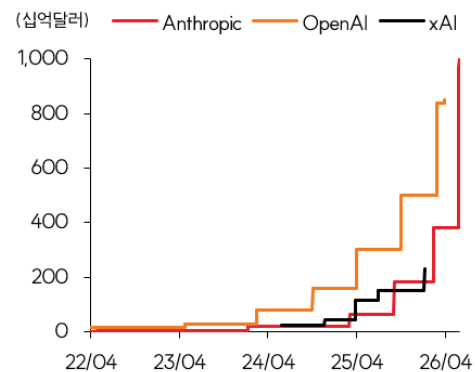
자료: Bloomberg, SK 증권

AI Labs ARR 추이



자료: EpochAI, SK 증권

주요 AI Labs 기업가치 추이



자료: EpochAI, SK 증권

Compliance Notice

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

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

당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.