

IBKS Spot Comment

AI/인터넷/게임

이승훈

02) 6915-5680
dozed@ibks.com중국 AI 모델의 폭발적인
사용 증가

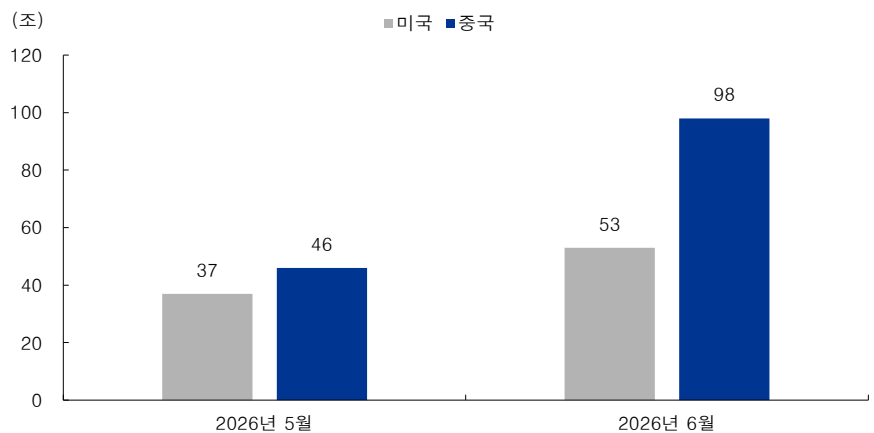
[AI/인터넷/게임]

중국 AI 모델의 글로벌 역습

1. 중국 모델, 토큰 사용량에서 미국 추월

상위 20개 AI 모델을 기준으로 미국과 중국 간의 토큰 사용량을 비교한 결과, 2026년 5월에는 중국이 46조 토큰, 미국이 37조 토큰을 기록하며 중국이 근소한 우위를 보였다. 그러나 한 달 뒤인 2026년 6월에는 중국이 98조 토큰으로 급증하며 미국의 53조 토큰과 격차를 두 배 가까이 벌렸다. 결국 폭발적인 중국 AI 모델의 도입 및 사용 증가는 가격 경쟁(토큰 단가 하락)을 부추길 수 있으며, 이는 막대한 자금을 투자한 미국 빅테크 기업들의 투자 회수를 더욱 어렵게 만드는 잠재적 위협 요인으로 여겨진다. 그러나 실제 돈이 지급되는 AI 워크로드의 토큰 점유율은 미국 3대 LLM모델이 90% 이상인 것으로 나타난다.

그림 1. 미국과 중국의 토큰 사용량 비교



자료: Apollo Chief Economist, Yahoo Finance, IBK투자증권

중국 모델의 핵심은 성능
향상과 토큰당 비용 우위

2. 역전의 동력은 성능 상향 평준화와 압도적 가격 우위

핵심 변수는 토큰당 성능비다. AI 에이전트는 단순 챗봇 대비 토큰 소비량이 수십 배 많아 토큰 과금 기반의 미국 프런티어 모델 사용 시 비용 부담이 기하급수적으로 커진다. 중국 모델의 토큰당 비용은 미국 프런티어의 수십분의 1 수준이다. DeepSeek는 2026년 5월 V4 Pro 가격을 75% 영구 인하했고, 샤오미는 플래그십 API 가격을 최대 99% 인하했다. 중국 오픈소스 모델이 미국 대비 60~90% 저렴한 것으로 알려져 있다.

전문 언론사는 모델 혼용으로 비용을 최대 95% 절감한 사례가 확인된다고 보도했다. 실제 전환 사례도 축적되고 있다. Snowflake는 GLM-5.2가 Claude와 대등한 성능을 더 낮은 가격에 제공한다고 평가하기도 했다.

3. 중국 주요 LLM 3사

중국 주요 LLM 기업

- 1) DeepSeek
- 2) Qwen
- 3) GLM

실사용(토큰), 기술력, 자본시장 대표성 등을 종합할 때 DeepSeek(딥시크), Alibaba Qwen(큐원), Zhipu AI GLM(즈푸)가 중국의 주요 LLM 기업으로 판단된다.

표 1. 중국 주요 LLM 기업

구분	상장 여부	핵심 내용
DeepSeek	비상장	항저우 소재. 광원평의 헤지펀드 High-Flyer 전액 출자로 설립. 2026년 6월 창사 첫 외부투자 유치(약 74억달러, 약 11.5조원). 포스트머니 밸류에이션 약 590억달러. 텐센트, CATL, 국가AI산업투자기금 참여. 국가 기금만 의결권 보유하며 기타 투자자는 5년 락업 및 무의결권 LP 지분
Qwen (Alibaba)	상장	Alibaba Group Holding 산하. Alibaba Cloud가 개발 주체. 클라우드-칩-모델 수직계열화 유일 완비
GLM (Zhipu AI)	상장	HKEX 2513(Knowledge Atlas Technology). 2026년 1월 IPO(공모가 HK\$116.20)로 세계 첫 LLM 전업 상장사. 상장 후 최대 2,400% 급등, 시총 HK\$1조(약 1,280억달러) 돌파. 2026년 7월 HK\$1,588에 약 40억달러 유상증자. 상하이 STAR마켓 2차 상장 추진

자료: 언론 기사, 기업 발표자료 종합, IBK투자증권

1) 성능 비교 (vs Claude-Gemini-GPT)

성능이 비슷할수록 가격이
결정 변수로 작용

코딩 실전 능력의 표준 벤치마크인 SWE-bench Verified 기준, 중국 최상위 모델은 미국 최상위 폐쇄형 대비 약 8~10%p 열위이나, Gemini 3.1 Pro와는 동률 수준까지 근접했다. 특히 80.2~80.6% 구간에 중국 4개 모델(DeepSeek V4 Pro Max, MiniMax M3, Qwen3.7 Max, Kimi K2.6)과 Gemini 3.1 Pro가 0.4%p 이내로 밀집해 있으며, 해당 구간의 출력 가격은 \$2.4~12/M로 5배 차이가 난다. 성능이 수렴할수록 가격이 결정 변수가 되는 구조다.

표 2. LLM 별 코딩 성능 비교

모델	SWE-bench Verified	비고
Claude Opus 4.8 (미국)	88.60%	상용 코딩 최상위 (Mythos 5는 95.5%, 승인기관 한정)
GPT-5.5 (미국)	88.70%	Claude와 동급 최상위
Gemini 3.1 Pro (미국)	80.60%	프런티어 최저가 티어
DeepSeek V4 Pro Max	80.60%	오픈웨이트 최고점, Gemini와 동률
Qwen3.7 Max	80.4%	AA Intelligence Index 46, 코딩 글로벌 top 10
GLM-5.2 (Zhipu)	Code Arena 프론트엔드 세계 2위	AA Intelligence Index 51로 오픈웨이트 1위 Gemini 3.1 Pro(46) 상회

자료: llm-stats, Artificial Analysis Intelligence, IBK투자증권

2) 토큰 비용 비교

에이전트나 복잡한
워크로드일수록
격차 극대화

1백만 토큰당 공식 API 가격(입력/출력, USD) 기준이다. 출력 토큰이 대량 발생하는 에이전트나 리포트 생성 워크로드일수록 격차가 극대화된다. 단순 계산으로 월 1억 출력 토큰 워크로드의 경우 DeepSeek V4 Pro는 87달러, Claude Opus 4.8은 2,500달러, 원화 기준 약 13만원 대 388만원(1,550원/달러)이다.

표 3. 중국 LLM 별 토큰 비용 비교

모델	입력 / 출력 (\$/M)	Claude Opus 4.8 대비 (출력 기준)
DeepSeek V4 Flash	0.14 / 0.28	약 89배 저렴
DeepSeek V4 Pro	0.435 / 0.87	약 29배 저렴 (2026.5 75% 영구 인하 반영)
MiniMax M3	0.60 / 2.40	약 10배 저렴, SWE 80%+ 최저가
GLM-5.2 (Zhipu)	1.40 / 4.40	약 6배 저렴, 프런티어급 최저가
Qwen3.7 Max	1.25~2.50 / 3.75~7.50	약 3~7배 저렴 (프로바이더별 상이)

자료: 기업 자료 종합, IBK투자증권

3) 주요 기업별 학습/추론 인프라 현황

중국 AI 기업의 방향성은
탈NVIDIA

공통 방향성은 탈NVIDIA다. 미국 수출통제로 Blackwell급 접근이 차단된 상태에서 3사 모두 화웨이 Ascend 또는 자체 칩으로 이동 중이며, 중국 기업들은 향후 1년 내 AI 가속기 예산의 46%를 자국산으로 전환할 계획이다. 화웨이는 중국 데이터센터 칩 시장의 약 절반을 점유하고 있다. 이는 중국 AI 밸류체인 투자 관점에서 모델사보다 화웨이 생태계의 낙수효과가 더 직접적임을 시사한다.

표 4. 주요 기업별 학습/추론 인프라 비교

기업	학습/추론 인프라
DeepSeek	초기 R1은 NVIDIA H800(대중 수출통제 이전 물량)으로 학습. 2026년 4월 출시 V4부터 화웨이 Ascend(950p)로 학습 워크로드를 전환. 중국산 칩에서 1.6조 파라미터 학습을 입증. 외부 퍼블릭 클라우드에 의존하지 않는 자체 인프라 노선으로, 내몽골 우란차부에 첫 자체 데이터센터 구축 중(경자산→중자산 전환)
Qwen	Alibaba Cloud 자체 인프라에서 학습 및 추론 일괄 수행. 자체 설계 칩(T-Head 계열 Zhenwu M890)과 자체 슈퍼노드 서버(Panjiu AL128)를 투입한 풀스택 수직계열화. NVIDIA 칩 병행 사용. 2026년 데이터센터 투자는 알리바바 단독 약 530억달러 규모(아시아·중남미·중동 중심)
GLM (Zhipu)	2025년 1월 미 상무부 Entity List 등재로 NVIDIA 첨단칩 접근 차단. 중국산 실리콘 기반으로 학습. 추론은 자사 Zai 클라우드 API 및 중국 국유 클라우드사 라이선싱으로 제공. 유상증자 자금의 상당분을 컴퓨트 인프라 확충에 배정

자료: Reuters, Wccftch, 기업 발표 자료, IBK투자증권

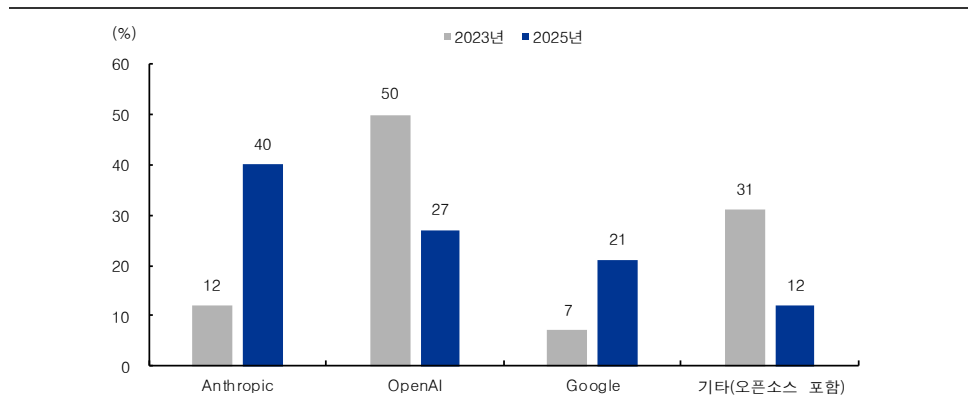
하지만, 실제 돈을 버는
기업은 미국 3사

토론 사용량 역전에도 불구하고, 실제 돈이 지불되는 미국 엔터프라이즈 시장은 미국 3사가 약 90%를 장악한 별개의 전장이다. 2026년 1분기 기준 미국 기업 LLM API 지출 점유율은 Anthropic 40%(2023년 12%에서 급등), OpenAI 27%(2023년 50%에서 하락), Google 21%(7%에서 상승) 순이다. 오픈소스 모델의 엔터프라이즈 비중은 오히려 19%에서 11~13%로 하락했다. 성능 격차가 유의미한 고부가 워크로드에서는 폐쇄형 프런티어 모델의 지배력이 강화되고 있다는 의미다.

AI 모델 시장의 이원화
시작

시장은 이원화됐다. (1) 미국 엔터프라이즈의 고부가·규제 민감 워크로드는 Anthropic 중심의 폐쇄형 프런티어가 장악하고, (2) 개발자·스타트업·글로벌 사우스의 가격 민감·대량 워크로드는 중국 오픈웨이트가 잠식한다.

그림 2. 미국 LLM API 지출 점유율 비중



자료: 각 기업 자료, IBK투자증권

Compliance Notice

동 자료에 게재된 내용들은 외부의 압력이나 부당한 간섭없이 본인의 의견을 정확하게 반영하여 작성되었음을 확인합니다.

동 자료는 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.

동 자료는 조사분석자료 작성에 참여한 외부인(계열회사 및 그 임직원등)이 없습니다.

조사분석 담당자 및 배우자는 해당종목과 재산적 이해관계가 없습니다.

동 자료에 언급된 종목의 지분율 1%이상 보유하고 있지 않습니다.

당사는 상기 명시한 사항 외 고지해야 하는 특별한 이해관계가 없습니다.