Global Tech

Kimi 쇼크의 실체

글로벌 테크 박재환 6158 / jaehwan124@

❝ 중국 MoonShot AI는 7월 16일 총 2.8조개의 파라미터와 100만 토큰의 컨텍스트 윈도우를 갖춘 Kimi K3를 공개. 일부 벤치마크에서는 Claude Fable 5와 GPT-5.6 Sol 등 미국 최상위 모델과 대등하거나 더 높은 성능을 기록. 반면 출력 토큰 100만개 당 비용은 15달러로 GPT-5.6 Sol의 절반 수준. 이에 시장에서는 K3를 낮은 비용으로 미국 모델을 추격한 '딥시크 모먼트 2.0'으로 해석하며 **AI 인프라 과잉 투자 우려가 재차 확대.**

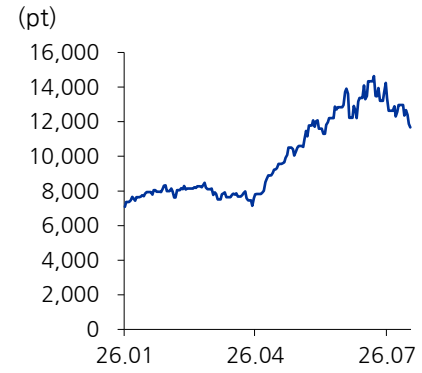
❝ 그러나 K3는 당시 제한된 컴퓨팅 자원으로 비용을 낮춘 경량 모델인 딥시크와는 차이가 존재. K3는 모델 축소나 인프라 절감을 선택하지 않고, 오히려 파라미터를 이전 모델 대비 2.7배 확대하고, 100만 토큰의 대규모 컨텍스트를 지원하도록 설계. 이는 **중국 AI 모델이 단순 가격 경쟁을 넘어 모델의 스케일과 성능을 경쟁하는 단계로 진입했음을 시사.**

❝ **낮은 서비스 가격이 곧 낮은 추론 인프라 부담을 의미하지 않음.** K3는 MXFP4 기준 모델 파라미터 저장에 약 1.5TB의 메모리가 필요하며, KV Cache와 연산 버퍼 등을 고려하면 추가적인 메모리 수요가 발생. 이에 Moonshot은 원활한 모델 구동을 위해 64개 이상의 가속기로 구성된 Super Node 사용을 권고. 이는 **K3가 대규모 AI 인프라를 요구하며 기존의 중국 모델들과 궤를 달리 하고 있음을 시사.**

❝ K3의 KDA(Kimi Delta Attention) 역시 우려로 작용. KDA는 불필요한 KV Cache를 누적을 줄여 롱 컨텍스트의 메모리 부담을 완화. 다만 과거 문맥에 대한 정확도를 보완하기 위해 K3는 기존 Attention 계열인 Gated MLA와 KDA를 결합하여 사용. 또한 KDA는 KV Cache에만 관여하며, 모델 파라미터 자체를 로딩하는 가속기와 메모리 수요에는 영향이 제한적. MoonShot이 K3 모델 공개 직후 신규 구독을 제한한 점 역시 고성능 모델 구동에 필요한 AI 인프라가 여전히 병목으로 작용함을 시사. 따라서 **중국발 고성능 모델의 출현과 확산은 AI 반도체 수요에 긍정적인 측면이 더 크다고 판단.**

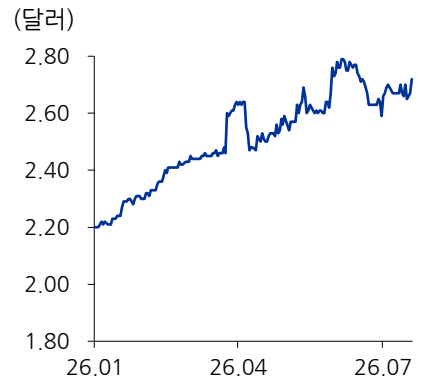
❝ K3의 세부 파라미터와 실제 운영 비용 등이 확인되지 않았다는 불확실성은 존재. 그러나 중국 AI 모델이 스케일링의 방향성으로 전환하고 있으며, 미국 프론티어 모델과의 '성능 경쟁' 단계에 진입했다는 점은 명확한 지점. **K3의 파라미터 확대와 성능 개선은 모델의 스케일을 키울수록 성능이 향상된다는 스케일링의 법칙이 여전히 유효함을 시사. Kimi K3는 AI 캐팩스에 대한 위협보다 글로벌 모델 경쟁 심화와 AI 인프라 확보 경쟁 확대를 증명하는 사례로 판단.**

SOX 지수 추이



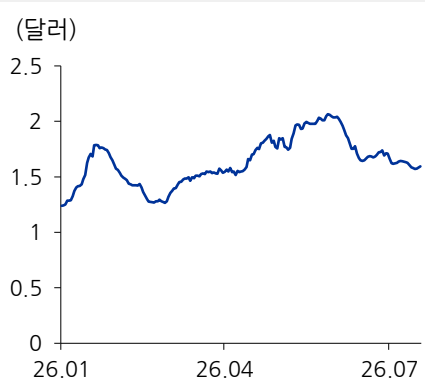
자료: Refinitiv, 유진투자증권

H100 렌탈가 추이



자료: Bloomberg, 유진투자증권

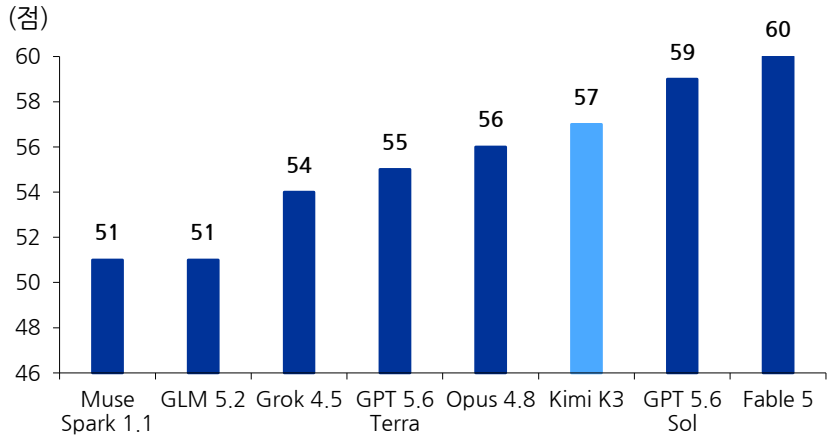
토큰 지출 인덱스 추이



자료: Bloomberg, 유진투자증권

Kimi K3 지능 인덱스는 57 점으로 미국 프론티어 모델과 유사한 수준을 기록

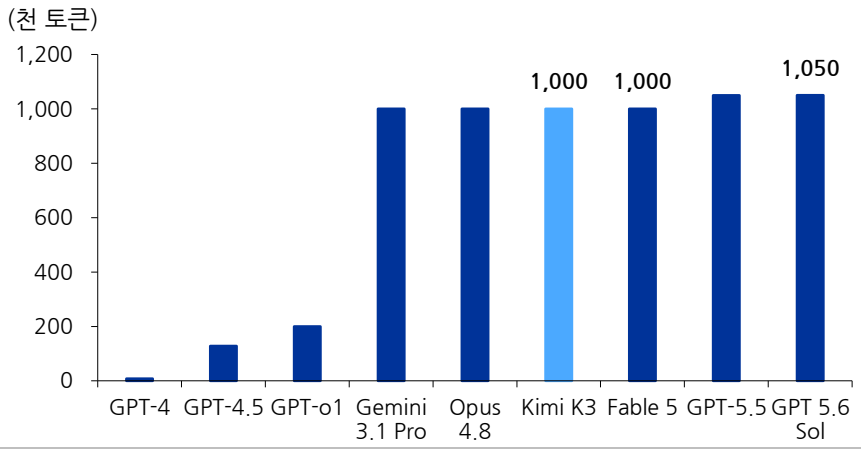
주요 지능 인덱스 비교



자료: Artificial Analysis, 유진투자증권

Kimi K3 는 최대 100 만 토큰의 컨텍스트 윈도우를 지원

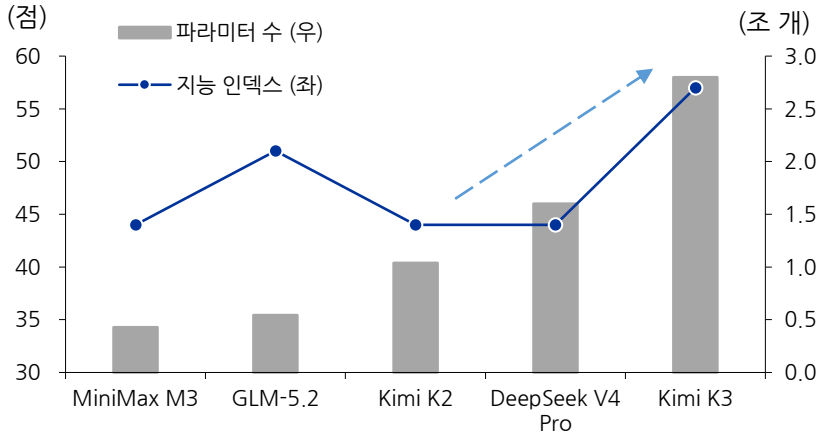
주요 AI 모델 컨텍스트 윈도우



자료: 각 사, 유진투자증권

파라미터 확대와 함께 성능도 상승하며 스케일링의 법칙의 유효성을 재확인

중국 AI 모델 파라미터, 지능 인덱스



자료: Artificail Analysis, 유진투자증권

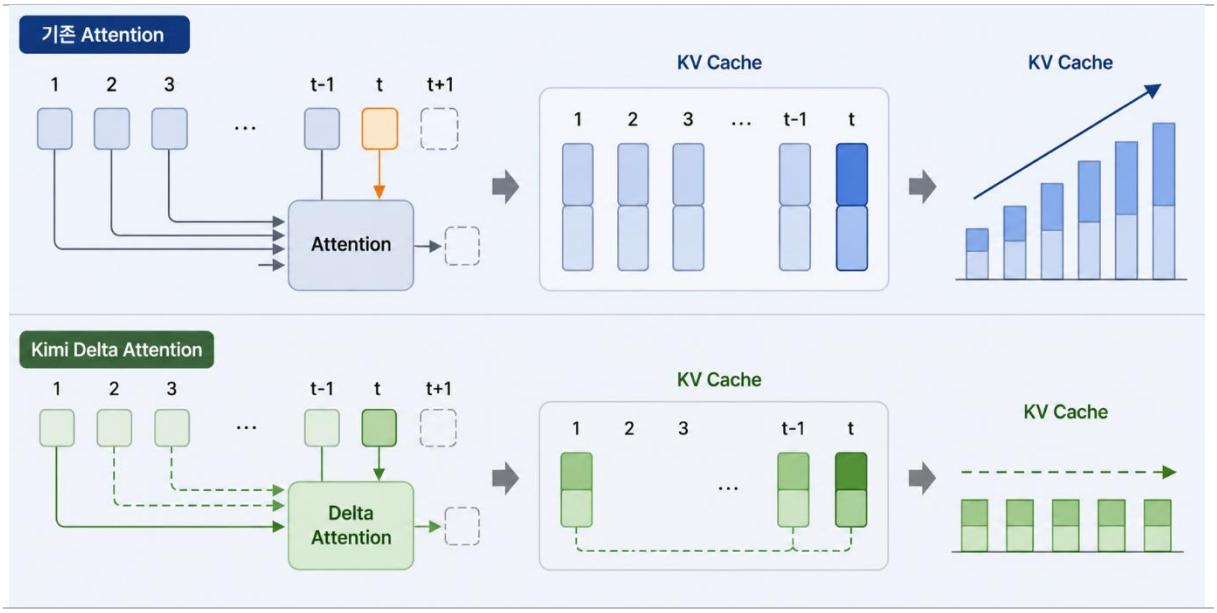
KDA(Kimi Delta Attention)이란?

KDA 는 이전 토큰의 Key 와 Value 를 모두 KV Cache 에 저장하는 대신, 과거 문맥을 일정한 크기의 상태로 압축해 유지하는 Attention 구조. 일반적인 Attention 은 새로운 토큰을 생성할 때마다 기존 토큰의 Key 와 Value 를 KV Cache 에 추가하고, 이후 추론 연산에서 이를 반복적으로 참조. 이에 따라 컨텍스트가 길어질수록 KV Cache 용량과 메모리 부담도 토큰에 비례해서 증가.

반면 **KDA 는 새로운 정보가 입력되면 기존 상태와의 차이를 계산하고, 변화된 부분을 중심으로 갱신하는 방식.** 정보의 중요도에 따라 기억을 유지하거나 제거하는 게이트 구조도 적용. 현재 작업 목표나 핵심 지시사항은 장기간 유지하는 반면, 이미 종료된 Tool Call 이나 중요도가 낮아진 정보는 빠르게 축소. 이에 따라 **KDA 가 적용된 영역은 컨텍스트가 길어져도 KV Cache 가 지속적으로 누적되지 않아 롱 컨텍스트의 메모리 부담을 줄일 수 있음.**

다만 과거 입력을 원문 그대로 보존하지 않고 압축된 상태로 유지하고, 디테일한 정보는 회수가 불가능하기에, **오래전 특정 문장이나 수치를 정확히 다시 찾는 능력은 일반적인 Attention 대비 약화**될 수 있음. Kimi K3 는 이를 보완하기 위해 KDA 만 단독으로 사용하지 않고, 과거 토큰을 직접 참조할 수 있는 Gated MLA 를 함께 적용. **KDA 가 대부분의 긴 문맥을 낮은 메모리 부담으로 처리하고, Gated MLA 가 특정 정보에 대한 정확한 검색을 보완하는 하이브리드 구조.**

기존 Attention vs. KDA



자료: 유진투자증권

글로벌 밸류체인

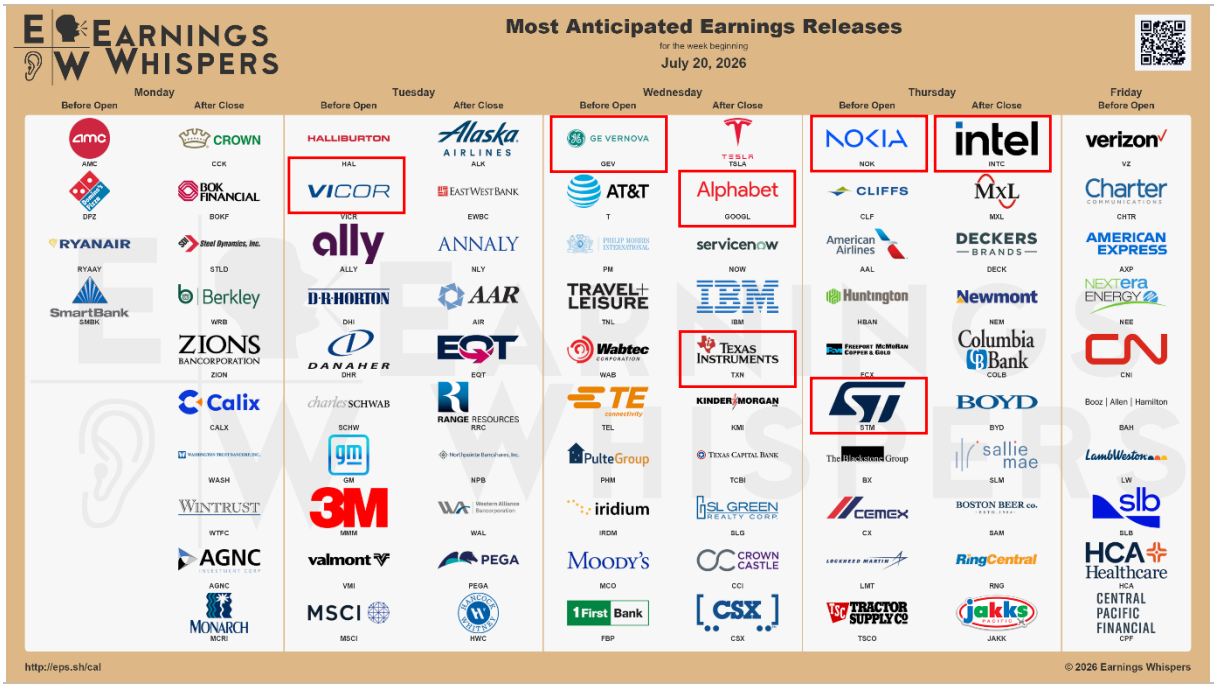
	세분화	주가	시가총액	주가 변동(%)			밸류에이션		
		각국 통화	백만 달러	1W	1M	6M	PER (TTM)	PER (FWD)	PSR (FWD)
아마존	AI 캐팩스 Big5	247	2,659,480	0.8	1.2	7.0	35.2	25.2	3.1
구글		347	4,217,934	-2.9	-5.8	7.7	35.5	25.8	9.4
마이크로소프트		394	2,925,466	2.3	3.8	-13.4	24.9	21.2	7.9
메타		646	1,639,847	-3.5	11.9	6.9	19.2	17.3	6.2
오라클		126	364,120	-10.1	-31.4	-29.7	20.6	15.7	4.1
네비우스	네오 클라우드	178	45,120	-19.1	-38.0	79.0	-	-	9.0
코어위브		73	40,088	-17.6	-37.9	-23.1	-	-	2.6
아이렌		34	12,015	-18.3	-43.9	-38.0	-	-	5.8
디지털오션		119	13,897	-8.9	-31.4	137.7	479.8	102.3	11.1
어플라이드 디지털	AI DC 코로케이션	26	7,370	-17.2	-44.6	-27.3	-	-	12.7
Hut8		91	11,252	-10.5	-26.5	56.9	-	-	33.5
테라울프		18	8,999	-17.3	-37.3	36.2	-	-	18.3
사이퍼디지털		18	7,183	-20.6	-39.8	-2.0	-	-	19.4
코어사이언티픽		21	6,657	-10.9	-28.2	14.1	-	-	8.5
엔비디아	AI XPU	203	4,908,002	-3.9	-3.7	13.9	35.5	20.4	11.2
브로드컴		371	1,764,230	-7.3	-9.9	11.5	57.1	23.6	12.5
AMD		496	808,387	-11.1	-7.7	113.8	171.6	55.8	14.4
인텔		95	477,671	-13.5	-29.1	95.7	378.6	85.1	7.9
마벨		189	165,057	-20.0	-39.2	136.4	125.5	41.3	13.0
ARM		267	285,380	-17.4	-39.2	149.3	380.2	122.1	47.5
삼성전자	메모리	248,000	978,457	-13.4	-29.2	74.0	19.7	4.5	1.8
SK하이닉스		1,817,000	873,924	-17.4	-34.0	148.9	16.9	4.0	3.0
마이크론		849	958,798	-13.3	-25.1	132.6	19.1	5.8	4.1
샌디스크		1,355	200,635	-29.3	-38.0	199.0	46.3	7.3	4.4
키옥시아		52,110	175,822	-32.3	-52.0	242.7	51.1	5.1	2.9
TI	아날로그 반도체	284	258,485	-8.8	-12.0	49.8	48.7	34.5	11.9
ADI		375	182,833	-5.1	-13.6	27.0	38.2	26.7	11.2
인피니온		63	94,546	-10.0	-22.7	56.5	76.8	27.7	4.6
모놀리식 파워		1,312	64,459	-3.0	-16.1	26.8	94.6	50.7	16.3
바이코		238	10,831	-12.7	-28.3	51.0	79.4	61.4	15.7
나비타스		11	2,753	-14.9	-52.3	16.2	-	-	57.8

자료: Bloomberg, 유진투자증권

글로벌 밸류체인

	세분화	주가	시가총액	주가 변동(%)			밸류에이션		
		각국 통화	백만 달러	1W	1M	6M	PER (TTM)	PER (FWD)	PSR (FWD)
TSMC	파운드리	2,320	1,864,077	-4.9	-3.7	30.7	26.9	19.1	9.5
UMC		130	50,659	-15.3	-10.7	91.7	32.5	27.3	5.6
SMIC		70	107,684	-11.0	-8.8	-6.4	115.6	54.2	8.9
글로벌 파운드리		57	31,539	-16.7	-33.0	31.8	42.5	29.6	4.2
타워 세미콘덕터		234	26,461	5.1	-18.4	81.5	108.9	54.8	12.2
ASE	OSAT	597	82,513	-10.9	-2.6	101.7	55.0	32.0	3.2
엠코		63	15,601	-10.7	-30.4	28.3	36.0	28.8	2.0
ASML	반도체 장비	1,546	685,710	0.3	-6.8	35.6	56.0	33.8	11.9
AMAT		530	420,529	-12.1	-14.2	66.4	53.7	35.1	10.8
램리서치		313	391,804	-10.6	-19.5	40.9	58.8	41.3	13.5
KLA		213	277,910	-8.1	-18.0	43.2	58.4	45.6	16.9
테라다인		322	50,462	-10.4	-26.4	43.9	58.4	46.4	11.4
어드반테스트		27,505	123,960	-7.8	-13.3	28.0	53.4	38.6	13.4
암페놀	인터커넥트	151	186,011	-4.9	-7.8	-0.7	39.2	30.4	5.4
크레도 테크놀로지		203	37,795	-21.4	-25.4	32.3	80.0	33.3	15.3
아스테라랩스		304	52,043	-26.5	-27.2	65.2	212.2	89.1	29.6
루멘텀홀딩스		733	57,013	-8.6	-13.8	105.4	137.4	46.9	11.4
코히어런트		278	54,309	-14.5	-28.7	43.5	105.3	36.7	6.1
어플라이드 옵토		102	8,218	-14.6	-36.7	160.9	-	56.3	6.0
코닝		155	132,812	-19.0	-20.7	67.0	73.5	45.6	6.7
키사이트		316	53,986	-1.9	-13.1	49.0	45.1	29.2	7.4
비아비 솔루션스		38	9,350	-11.4	-19.1	104.7	450.5	31.6	5.3
델	서버	396	256,092	-8.9	-3.2	256.8	31.3	21.2	1.5
HPE		46	60,675	-5.6	-3.4	124.7	31.5	11.9	1.3
SMCI		24	15,640	-14.6	-21.1	-23.0	12.7	8.4	0.3
셀레스티카		301	34,561	-16.3	-19.1	-3.0	39.9	26.3	1.6
아리스타 네트웍스		169	212,314	-9.8	-0.6	32.2	58.0	44.5	17.4
시스코 시스템스		112	441,204	-7.7	-6.4	52.6	36.7	23.8	6.5
패브리넷		478	17,141	1.5	-16.6	-0.9	40.9	29.2	3.2
시에나		374	52,999	-18.7	-12.6	55.2	105.9	47.9	7.5
시놉시스	EDA	384	73,581	-13.7	-15.6	-24.8	96.7	24.3	7.3
케이던스 디자인		330	91,050	-14.1	-14.8	7.5	66.1	40.5	14.3

주간 실적발표 일정



자료: Earnings Whisper, 유진투자증권

Compliance Notice

당사는 자료 작성일 기준으로 지난 3 개월 간 해당종목에 대해서 유가증권 발행에 참여한 적이 없습니다
당사는 본 자료 발간일을 기준으로 해당종목의 주식을 1% 이상 보유하고 있지 않습니다
당사는 동 자료를 기관투자가 또는 제 3 자에게 사전 제공한 사실이 없습니다
조사분석담당자는 자료작성일 현재 동 종목과 관련하여 재산적 이해관계가 없습니다
동 자료에 게재된 내용들은 조사분석담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다
동 자료는 당사의 제작물로서 모든 저작권은 당사에게 있습니다
동 자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다
동 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다

투자기간 및 투자등급/투자의견 비율

종목추천 및 업종추천 투자기간: 12 개월 (추천기준일 종가대비 추천종목의 예상 목표수익률을 의미함)		당사 투자의견 비율(%)
· STRONG BUY(매수)	추천기준일 종가대비 +50%이상	3%
· BUY(매수)	추천기준일 종가대비 +15%이상 ~ +50%미만	94%
· HOLD(중립)	추천기준일 종가대비 -10%이상 ~ +15%미만	3%
· REDUCE(매도)	추천기준일 종가대비 -10%미만	0%

(2026.06.30 기준)