AI/인터넷/게임

이승훈

02) 6915-5680
dozed@ibks.com

3일 간격으로 쏟아진
중국의 초대형 오픈웨이트
AI 모델

[AI/인터넷/게임]
중국 AI 충격, AI 캐즘의 기간을 단축시킬 수도

1. 중국 AI의 원투펀치 공격

딥시크(DeepSeek), 알리바바(Qwen), 바이두(ERNIE) 등 중국 기업과 연구소가 고성능 오픈웨이트(Open-Weight) 모델을 연이어 공개하며 글로벌 AI 생태계에 큰 파장을 일으키고 있다. 오픈웨이트란 소스코드뿐 아니라 모델의 핵심 자산인 가중치까지 공개해 누구나 다운로드, 파인튜닝, 상업적 활용이 가능한 모델을 의미한다.

- 문샷 AI Kimi K3 (7/16): 896개 전문가 중 토큰당 16개만 활성화하는 MoE(Mixture of Experts) 기반 2.8조 파라미터 모델이다. 코딩 및 에이전트 벤치마크 상위권을 자체 제시하며, 일부 지표에서 Opus 4.8와 GPT-5.5(high)를 상회하고, Fable 5와 GPT-5.6 Sol에는 뒤진다고 주장한다. 100만 토큰 컨텍스트와 네이티브 비전(이미지와 영상 입력을 텍스트와 함께 처리)을 지원한다.

- 알리바바 Qwen3.8-Max (7/19): Qwen이 최초로 공개한 1조 파라미터를 넘긴 멀티모달 모델(2.4조 파라미터)이다. 선도 프론티어와 대등, Fable 5에만 열위를 표방하나 벤치마크, 모델카드, 활성 파라미터, 라이선스 등은 아직 미공개이다.

표 1. K3 자체 발표 벤치마크

벤치마크	K3	의미
SWE Marathon (장기 코딩)	42	Opus 4.8(40.0), GPT-5.6(39.0), Fable 5(35.0) 상회
Program Bench (구조적 완성도)	77.8	전 모델 중 1위 (GPT-5.6 대비 +0.2)
SWE-bench Verified (실무 코딩)	76.80%	전작 K2(65.8%) 대비 대폭 개선
Frontend Code Arena (UI 코드)	1,679 (1위)	직전 대비 17계단 상승

자료: 기업 발표 자료, IBK투자증권

2. 과거 DeepSeek R1 충격과 다른 이유

파라미터 확대가 부르는 메모리 수요 폭증

K3 공개 직후, 2025년 1월 DeepSeek R1 발표 당시처럼 엔비디아 등 미국 AI 관련 기업의 시가 총액이 급락했다. 그러나 두 사건의 성격은 근본적으로 다르다. DeepSeek의 본질은 '싸게 만들기'였으나, K3와 Qwen3.8은 효율을 개선하되 '더 크게 만들기'로 승부한다. 2조 파라미터급 모델은 전체 파라미터를 GPU 메모리에 상주시켜야 하므로 메모리 인프라에 막대한 부담을 준다. 모델이 커질수록 필요한 GPU와 메모리의 수는 늘어난다. 추가로 100만 토큰의 긴 컨텍스트를 처리할 때 발생하는 KV 캐시까지 더해져 메모리 수요가 급팽창한다. 결과적으로 파라미터의 확산은 알고리즘 효율 개선을 압도하는 수준으로 메모리 수요를 폭증시키며, 중국 모델의 확산세는 오히려 고대역폭, 고용량 메모리(HBM 및 DRAM) 수요를 키우는 요인으로 작용한다.

3. 토큰 가격이 하락할수록 AI 수요는 기하급수적으로 증가

1) 토큰 비용 하락을 위한 노력

전용 칩 설계와 효율화를 통해 '토큰 당 원가'를 낮추는 미국 빅테크

미국 빅테크는 추론에 드는 '토큰당 원가'를 낮추기 위해 두 방향으로 노력 중이다.

- 전용 칩 직접 설계: 구글(TPU), 아마존(Trainium), 메타(MTIA), 마이크로소프트(MAIA)가 각자 AI 전용 칩을 만들어 쓴다. 비싼 외부 칩(엔비디아 GPU) 비중을 줄여 계산 비용을 낮춘다.

- 같은 일을 더 적은 계산으로: 캐시 재사용, 숫자 정밀도를 낮추는 양자화(Ex, 32비트를 8비트로 변환), 작업을 여러 장비에 나누는 분산 추론(다수 GPU, 통신 속도 중요)으로 효율을 끌어올린다.

2026년 7월, 구글의 프로젝트로 보도된 신규 칩 'Frozen v2'는 제미나이 구조 일부를 실리콘에 각인해 최신 TPU 대비 전력당 토큰 처리량을 6~10배 높이는 것을 목표로 하며, 2028년 배치를 앞두고 있다. 구글은 이에 발맞춰 2028년까지 TPU 생산량을 2배 이상 확대할 것으로 예상된다. 다만 신형 TPU 파드 역시 대규모 HBM을 소비하므로 'Frozen v2'의 전력 효율 혁신 역시 전체 메모리 수요 확대를 야기할 것이다.

2) 토큰 비용 하락 → 추론 수요 증가 (제번스의 역설)

싸질수록 더 쓴다, AI 판 제번스의 역설

효율 개선은 총소비를 줄이기보다 오히려 늘린다. 현재 AI 시장에서는 '제번스의 역설(Jevons Paradox)'이 강력하게 작동하고 있다. 에이전틱 워크로드는 계획 수립, 도구 호출, 검증, 재시도 과정에서 토큰을 대량 소모하기 때문에 일반 챗봇 대비 수십 배 이상의 자원을 태운다. 이로 인해 다수의 기업이 이미 올해 AI 예산을 초과하는 것으로 파악된다. 그러나 토큰 비용 단가가 하락할 경우 억눌려 있던 수요를 폭증시키는 방아쇠가 될 것이다. 결론적으로 토큰 단가 하락은 단순한 비용 절감이 아닌 수요 확장의 촉매로 작용하며, 이는 다시 설비투자(CapEx) 상향을 정당화하는 자기강화 루프를 만들 것이다.

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.