



2026. 7. 19 (일)

Kimi K3의 본질: H/W(메모리) 롱-플랫폼 슷

- 본질은 공개형 모델의 폐쇄형 프론티어 모델 Catch up. 단기에 AI 프론티어 모델(GPT, Claude 등) 익스포저 큰 빅테크 CapEx에 대한 우려는 인정
- 다만 이를 AI 하드웨어 미국 밸류체인 전체 투자 우려로 확대 해석하는 것은 기우. AI 인프라 투자 주체가 빅테크 이외에도 커지고 있음을 이해할 필요
- 고성능 공개형 모델로 ISP, 온프레미스에서 AI H/W 추가 투자 필요성 강화
- 거대 공개형 모델 구현 및 확산에 대응하는 AI 인프라 투자 확대는 필연적, 메모리 포함 AI 인프라 투자 가속 요인

투자전략

Analyst 황수욱

soowook.hwang@meritz.co.kr

미국 AI 프론티어 모델 vs 중국 AI 모델 격차 축소의 확인

7월 19일 Moonshot AI의 최신 AI 모델 Kimi K3 공개. 2.8조개의 파라미터, 100만 개 토큰 컨텍스트, 네이티브 멀티 모달 기능을 갖춘 초대형 모델. 공개된 API 성능은 Artificial Analysis Intelligence Index 기준으로 최상위권 기록. 가장 최신 모델인 GPT 5.6 Sol, Claude Fable 5와 유사한 수준이며, 기존에 해당 모델과 프론티어 경쟁을 하던 다른 미국 프론티어 모델을 아웃퍼폼

토큰당 훈련 및 연산 단가는 다른 프론티어 모델들과 유사해 딥시크 때와 같은 토큰당 비용 이슈와는 구분. 딥시크 충격은 600만 달러라는 제한적인 훈련비용 주장이 핵심이었음. 반면 K3는 현재 훈련비용, 훈련 FLOPs, 사용 GPU 종류 및 수량, 활성 파라미터 수를 공개하지는 않음. 오히려 Moonshot은 K3 배치에 64개 이상 가속기로 구성된 슈퍼노드 활용을 권장하는 등 하드웨어 효율화와는 별개

이번 발표의 핵심은 그동안 미국이 주도해오던 AI 성능 경쟁에 중국 모델이 아주 근접하게 따라잡은 사건. 미국 폐쇄형 프론티어 모델(GPT, Claude 등)들의 성능 차별화에서 발생하는 희소성 프리미엄이 약해질 수 있다는 우려 확대

하반기 전망에서도 언급했던, 중국 AI 약진 리스크가 현실화된 이벤트로 판단. 이에 따라 확대될 수 있는 우려는 OpenAI나 Anthropic향 익스포저가 큰 빅테크들의 성장성에 대한 우려가 당분간 부각될 수는 있음. 미국 AI 프론티어 기업들의 가격 결정력이 약해질 수 있다는 우려, 점유율을 내줄 수 있다는 우려 등이 이에 해당할 것. 다만 Kimi K3의 단가가 다른 모델 대비 크게 낮지 않다는 점도 상기할 필요

다만 AI 인프라 밸류체인(GPU, 메모리, 네트워크 등)에 대한 우려는 착각일 것

다만 AI 프론티어 모델이 이끄는 AI 인프라 투자 전반의 위축으로 해석하는 것은 오해일 것으로 생각. 즉, 메모리, GPU, 네트워크 등 AI 하드웨어 투자에 대한 눈 높이는 낮출 필요가 없고 오히려 호재로 해석해야 한다는 의견

중국 AI 모델의 미국 AI 프론티어 모델 추격이 미국 AI 인프라 기업에게 악재일 수 있다는 생각의 전제는 1) AI CapEx가 빅테크/하이퍼스케일러 중심이며, 2) 중국 AI 모델은 중국 AI 인프라 하드웨어를 활용한다는 관점일 것

두가지 모두 착각. 먼저 AI CapEx의 중심은 하이퍼스케일러에서 Non-하이퍼스케일러로 이동하는 추세로 판단. 엔비디아의 매출액 성장세와 매출 구성으로부터 힌트를 얻을 수 있음. <그림 3>에서 처럼 엔비디아의 매출액은 3Q25 이후 성장세가 가속화. 동기간 <그림 4>와 같이 하이퍼스케일러향 데이터센터 매출 비중은 감소하고, AI 클라우드, 일반 기업, 엔터프라이즈향 매출 비중이 늘어나고 있음

2023년 이후 2024년까지는 AI 데이터센터 투자가 빅테크가 경쟁에서 지지 않겠다는 네러티브 투자 중심이었음. 수익화 근거가 부족했기 때문에 빅테크의 현금을 중심으로 투자. 그런데 2025년 3분기 이후는 부채 기반 레버리지 투자가 확대 중. 데이터센터의 수익성이 확인된 이후 현금 없이도 빅테크 뿐만 아니라 여러 주체가 향후 수익성을 담보로 계약하는 부채(레버리지) 기반의 데이터센터 투자를 할 수 있게 됨

두번째로 중국 AI 모델의 확장에는 미국 AI 하드웨어가 수혜를 볼 수 있음. 일단 Kimi K3는 고성능 하드웨어의 필요성을 시사. 64개 이상 가속기로 구성된 슈퍼 노드를 권장하는데, DGX B200 8대 클러스터에서도 K3 구동이 가능하지만, 사실상 GB200, GB300 NVL72를 염두에 둔 언급으로 보임. 기술문서가 공개되어야 구체적으로 확인 가능하겠으나 성능과 모델 크기상 Rubin NVL72가 이상적일 가능성

미국 AI 하드웨어 수혜 가능성에 대해 이해하려면 아직 한국 주식시장에 보편적으로 확산되지 못한 AI 추론 생태계에 대한 이해가 필요함. 추론 서비스 사업자, Inference Service Provider(ISP)에 대한 이해인데 필자는 2025년 GTC에서 네오클라우드의 비즈니스화를 보았고, 2026년 GTC에서는 ISP의 비즈니스화를 보았다고 강조해왔음

AI 인프라 서비스 사업자(ISP)는 Kimi 같은 공개형 AI 모델을 직접 개발하는 대신, 이를 자사 GPU 서버에 설치한 뒤 기업들이 인터넷을 통해 쉽게 사용할 수 있도록 빌려주고 사용량만큼 요금을 받음. 비유하면 공개형 모델은 누구나 무료로 가져올 수 있는 ‘레시피’이고, ISP는 대형 주방과 조리도구를 갖춘 뒤 고객 대신 음식을 만들어 판매하는 식

고객은 비싼 GPU를 직접 구매하거나 복잡한 시스템을 관리할 필요 없이 API 사용료나 월 구독료만 내면 되고, ISP는 GPU 임대료, 토큰 사용료, 기업 맞춤형 모델 구축비, 보안·운영 관리비 등을 통해 수익을 얻음. 결국 공개형 모델이 많아질수록 모델 자체의 가격은 낮아질 수 있지만, 이를 안정적으로 빠르게 실행해 주는 서버·전력·소프트웨어 운영 서비스의 수요는 오히려 커질 수 있음

핵심은 글로벌에서 공개형 모델로 ISP 비즈니스를 하는 기업들이 미국 밸류체인이라는 것. Fireworks AI, Together AI 등이 대표적인 고속 성장 ISP 기업들인데, 이들은 공개형 모델을 서비스하더라도 연산에는 대부분 미국 기업이 설계한 AI 가속기를 사용하고 있음

이 외에도 고성능 공개형 모델의 확산은 오히려 AI CapEx 수요를 자극할 수 있음. 토큰 가격 문제가 첨예해지는 환경에서 ISP뿐만 아니라 온프레미스에서 공개형 모델 구현을 통한 자체적인 API 연산을 하는 수요도 커지는 중. 그런데 K3와 같은 고성능 공개형 모델은 기존 공개형 모델보다 온프레미스에서 해당 모델을 구현하는데 더 높은 사양, 더 큰 IT 설비 용량을 요구할 것으로 추정(그림 6)

전략: AI H/W(GPU, 메모리) 롱, 빅테크(AI 프론티어 및 하이퍼스케일러) 숏

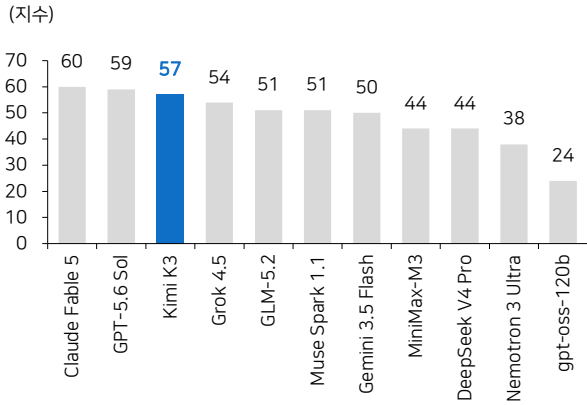
현재 상장 주식을 투자하는 입장에서는 어느 AI 모델이 최종 승자가 되는지를 맞출 필요가 없음. 다만 미국 AI 모델이 시장 헤게모니가 약화된다고 본다면 해당 기업들에 대한 임대 계약 비중이 높은 하이퍼스케일러 플랫폼은 수익성, 성장성 눈높이가 낮아질 수 있음. 이 또한 중기적으로 하이퍼스케일러들이 공개형 모델의 연산 비즈니스를 확대한다면 우려 요인이 아니게 될 일들

그러나 폐쇄형, 공개형 모델을 활용하는 모든 AI 인프라 사업자들이 사용해야 하는 AI 하드웨어 인프라 사업자(메모리, GPU, 네트워크 등)에게는 이번 K3 공개가 어찌됐든 중기적으로 수요의 상방이 열리는 이벤트로 해석하는 것이 타당

K3의 등장을 경쟁자가 늘어나는 산업의 외연 성장을 촉진하는 이벤트로 본다면, 이 경쟁자들에게 청바지와 곡괭이를 제공하는 기업을 사는 발상이 타당

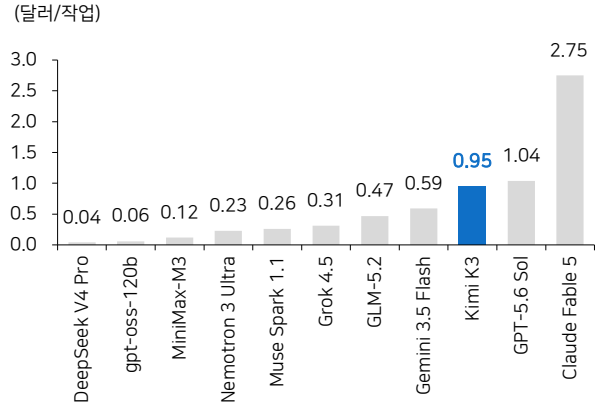
이상의 아이디어는 시장에서 1) AI CapEx가 빅테크/하이퍼스케일러뿐만 아니라 확산 중이며, 2) 중국 AI 모델은 중국 AI 인프라 하드웨어를 활용한다는 생각까지 공감하게 된다면 더욱 더 공고해질 것. 산업의 구조상 공개형 모델이 진보할수록 해자는 AI 플랫폼(모델) 기업이 아닌 AI 인프라 확보 기업에게로 이전됨. AI 인프라 제조 기업(메모리, GPU, 네트워크, 전력 인프라) 집중 전략이 중기적으로 여전히 우위라는 의견 유지

그림1 Artificial Analysis Intelligence Index 최상위에 K3



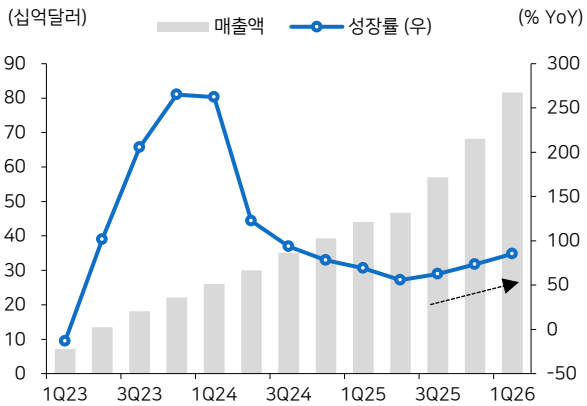
자료: Artificial Analysis, 메리츠증권 리서치센터

그림2 토큰 비용이 답시크 때처럼 저렴한 것은 아님



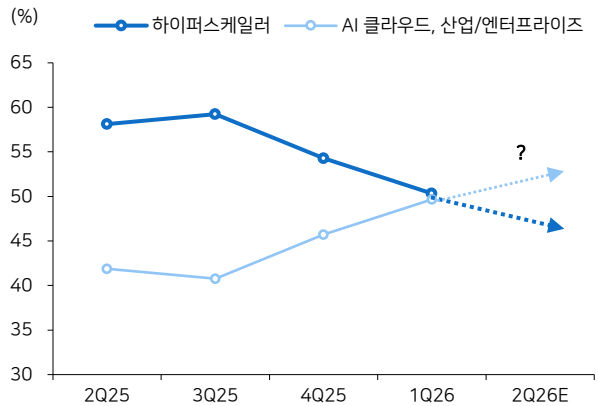
자료: Artificial Analysis, 메리츠증권 리서치센터

그림3 엔비디아 매출액 성장률



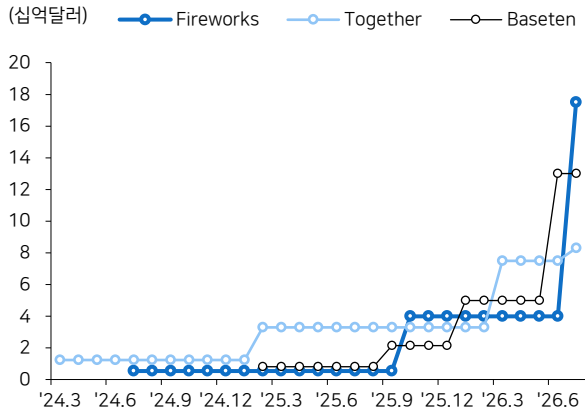
자료: NVIDIA, 메리츠증권 리서치센터

그림4 엔비디아 데이터센터 매출 고객 비중: Non-빅테크 증가



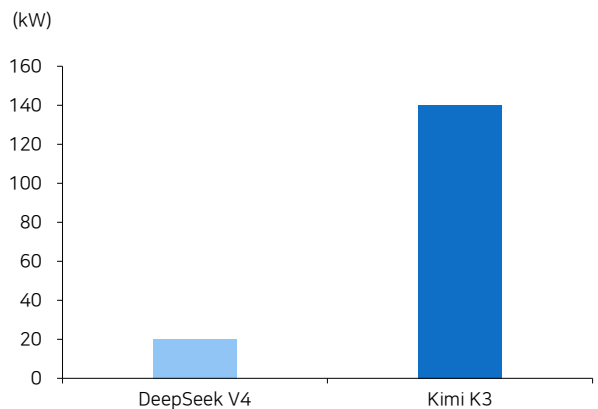
자료: NVIDIA, 메리츠증권 리서치센터

그림5 ISP(Inference Service Provider) 기업가치평가액



자료: 언론 보도 종합, 메리츠증권 리서치센터

그림6 최신 공개형 모델 하나를 구동하는데 필요한 용량 추정



주: V4 추정치는 이론적 최소 단일 인스턴스 추정치. 실제 안정적 온프레미스 배포는 30~100kW이상도 필요할 가능성
자료: 메리츠증권 리서치센터 추정

Compliance Notice

- 본 조사분석자료는 제3자에게 사전 제공된 사실이 없습니다.
- 당사는 자료작성일 현재 본 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 추천 종목과 재산적 이해관계가 없습니다.
- 본 자료에 게재된 내용은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 신의 성실하게 작성되었음을 확인합니다.

본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 리서치센터의 추정치로서 오차가 발생할 수 있으며 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 본 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 투자 결과와 관련한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 당사의 허락 없이 복사, 대여, 배포 될 수 없습니다.
