

Tech팀

이종욱 팀장
jwstar.lee@samsung.com

글로벌주식팀

김중한 Senior Analyst
joonghan1.kim@samsung.com

이영진 Senior Analyst
youngjin91.lee@samsung.com



리서치센터 리포트
바로가기

Tech (OVERWEIGHT)

Kimi Moment: 격화되는 프론티어 레이스

- Kimi K3가 촉발한 Kimi 쇼크는 딥시크 쇼크와 다른 새로운 형태의 공포.
- 인프라 투자의 둔화 요인은 아니지만, 미국 프론티어 랩 주도 비즈니스 구조의 변동성 확대로 연결
- AI 모델의 기술 축진은 반도체 수요에 긍정적. Kimi K3의 기술 방향은 고사양 메모리 수요를 촉발. 메모리 주가는 바닥 국면으로 쏠림 현상이 지나가는 중

WHAT'S THE STORY?

Kimi K3의 경쟁력은?: K3는 1) 스케일링 법칙 유효성을 증명한 오픈 웨이트 최대 규모 모델, 2) 중국의 가성비 전략보다 프론티어 경쟁에 본격 합류, 3) API 가격은 높지 않으나, 대규모 토큰 사용에 따라 작업당 비용 강점은 희석, 4) 과추론, 느린 속도, 부진한 실체감, 유저 경험 격차 등 한계 존재. 정보 전달 효율을 높인 아키텍처와 896개 전문가 중 16개만 활성화하는 고희소성 MoE 구조로 높은 스케일링 효율 달성

Kimi 쇼크 vs 딥시크 쇼크: 딥시크 쇼크가 중국 모델의 가성비 쇼크였다면, Kimi 쇼크는 중국이 프론티어 레이스의 선두를 점할 수 있다는 새로운 형태의 공포. 중국 기업이 레이스를 주도할 경우 오픈AI와 앤스로픽의 밸류에 대한 의구심과 중국발 쇼크의 일상화로 연결. 다만 제본스 역설을 기반으로 가격 하락에 따른 총 수요(토큰 사용량)의 증가 추세 지속. CapEx 둔화도 단기에 현실화될 가능성은 낮은 것으로 판단

이제 공은 다시 프론티어 랩에게: 중요한 것은 사용성. 단순 벤치마크로 드러난 성능이나 효율성과는 다른 측면. 원하는 결과물을 얼마나 적은 토큰, 시간, 전력을 사용해 달성했느냐가 중요. 현재 중국 기업 발 혼란은 본격적인 AI 서비스 시대로 넘어가기 직전 과도기의 성장통일 가능성이 높음. 향후에는 어떤 모델을 사용하는지보다 얼마의 비용에 어떤 서비스를 제공했는지가 직접적인 가치로 연결. 프론티어 레이스는 모델과 인프라 경쟁력 두 가지가 모두 갖춰져야 지속 가능

AI모델 기술 축진은 반도체 수요의 티핑포인트: Kimi K3는 프론티어 랩 경쟁을 촉발하고, 기술 개선에 아이디어를 제공. 그 결과 AI 기술 개선 속도가 빨라지는 것이 반도체 수요의 원동력. 반도체 입장에서 AI 고객의 목표가 성능 개선이나 원가 절감이냐가 중요하며 Kimi K3는 성능 개선의 새로운 국면을 제시. AI 모델이 중국 주도라 해도 오픈 모델을 바탕으로 서비스 배포를 촉진시킬 수 있다면 반도체 수요에 긍정적.

메모리 주가 바닥 국면: K3의 특징은 스케일업을 통해 HBM 사용량을 높이고 초희소 MoE를 돌리기 위해 여러 GPU 사이의 대역폭이 중요해짐. AI 기술에서 메모리의 중요도 증가 추세 지속. 삼성전자와 SK하이닉스의 주가는 fwd P/B 각각 1.8배, 2.5배로 단기 바닥 국면 돌입. Tier1 메모리 평균 고점대비 38% 하락하고 M7(ex Tesla) P/E가 필라델피아 반도체지수 P/E를 앞서며 쏠림 현상이 지나가는 중

(다음 페이지에 계속)

Kimi K3의 등장. 파장을 일으키다

Kimi K3의 경쟁력은?

우선 문샷 AI의 신규 모델 Kimi K3 경쟁력을 점검할 필요가 있다. K3의 특징은 1) 스케일링 법칙의 유효함을 증명한 오픈웨이트 최대 규모 모델, 2) 중국 AI 모델 특유의 가성비 전략보다 프론티어 모델 경쟁에 본격 합류, 3) API 가격은 높지 않으나, 대규모 토큰 사용에 따라 작업 당 비용의 강점은 희석, 4) 과추론, 느린 속도, 벤치마크 대비 부진한 실체감, 유저 경험 측면 격차 등 한계도 존재로 정리할 수 있다.

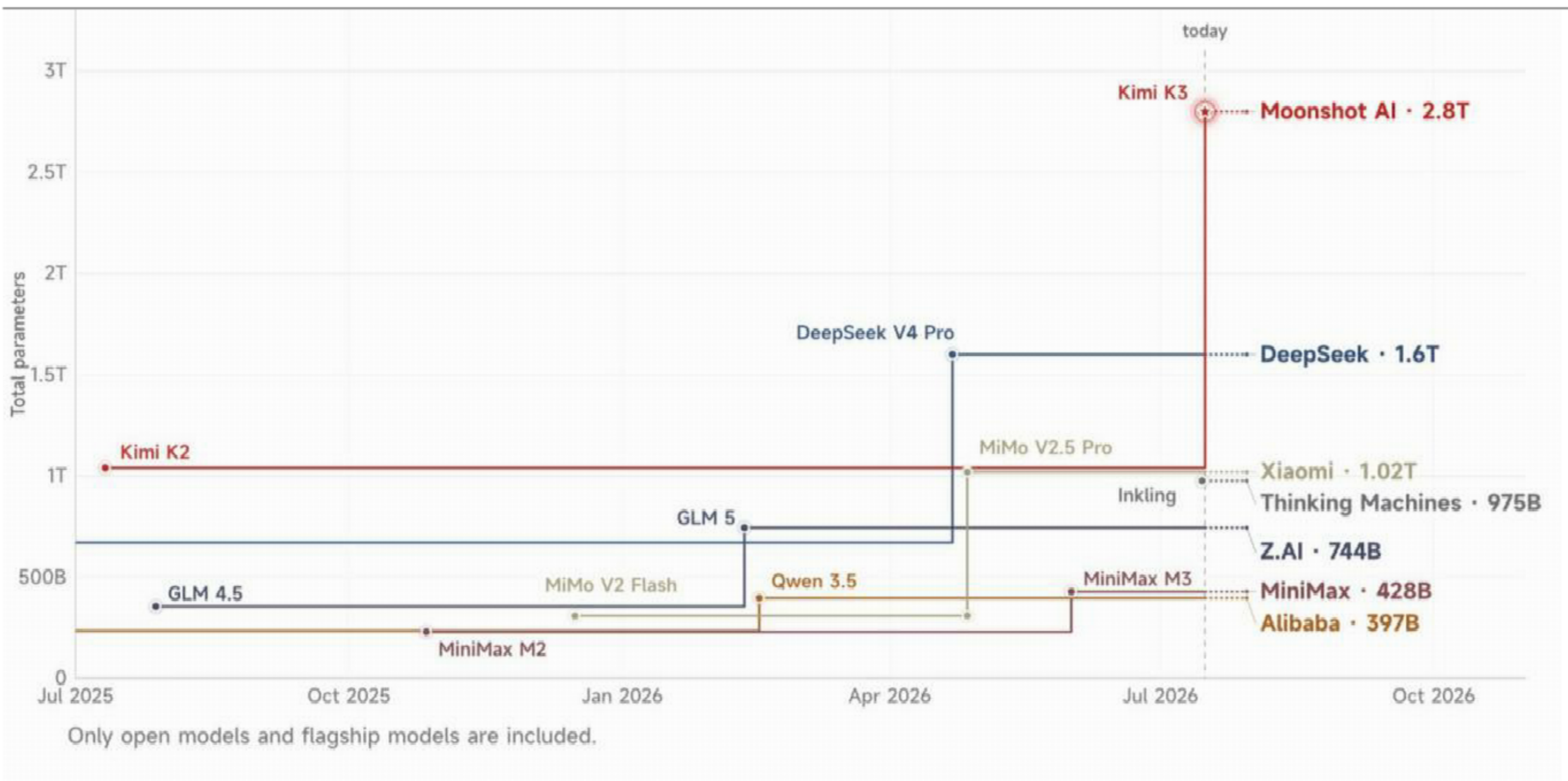
1) 전체 모델 가중치가 7/27일 공개된다는 가정하에 파라미터 규모는 2.8조로 오픈웨이트 최대 규모 모델이다. 기존 가장 큰 규모였던 DeepSeek V4 Pro(1.6T) 대비 2배 가까운 확장이다. 모델 규모 확장에 따라 성능이 개선되는 스케일링 법칙이 유효함을 증명하고 있다.

2) Artificial Analysis Intelligence Index 점수 57점으로 앤스로픽 Fable 5(60점), 오픈AI GPT-5.6-Sol(59점)에 이어 3위를 차지했다. Opus 4.8이나 GPT-5.5를 상회하는 수준이다. 프론트엔드 코딩 등 일부 벤치마크에서는 1위를 차지하기도 했다. SOTA는 아니지만, 프론티어 모델 경쟁에 참여했다. 기존 중국 모델의 두 가지 포지션인 극강의 가성비(딥시크 DeepSeek, 샤오미 MiMo) 또는 프론티어보다 한 티어 낮은 성능에 적당한 가격(Z.AI GLM)과는 또 다른 방향성이다. 특히 5월 말 출시된 Opus 4.8을 넘어선 것은 미-중간 격차가 2달 내외로 크게 축소되었다는 해석의 근거다.

3) API 가격은 인풋 \$3 및 아웃풋 \$15로 Claude Sonnet 5(Index 53점) 정상가(현재는 8월까지 할인)와 동일하다. K2.6(\$0.95 및 \$4) 대비 대폭 상승했다. K2.6보다 Index 평가에서 출력 토큰을 21% 적게 사용했지만, 절대 수준으로 1.3억 토큰을 사용했다. 따라서 작업당 비용은 \$0.95로 GPT-5.6-Sol(\$1.04)와 유사한 수준이다. 물론 Fable 5(\$2.75)나 Opus 4.8(\$1.80)대비는 낮다.

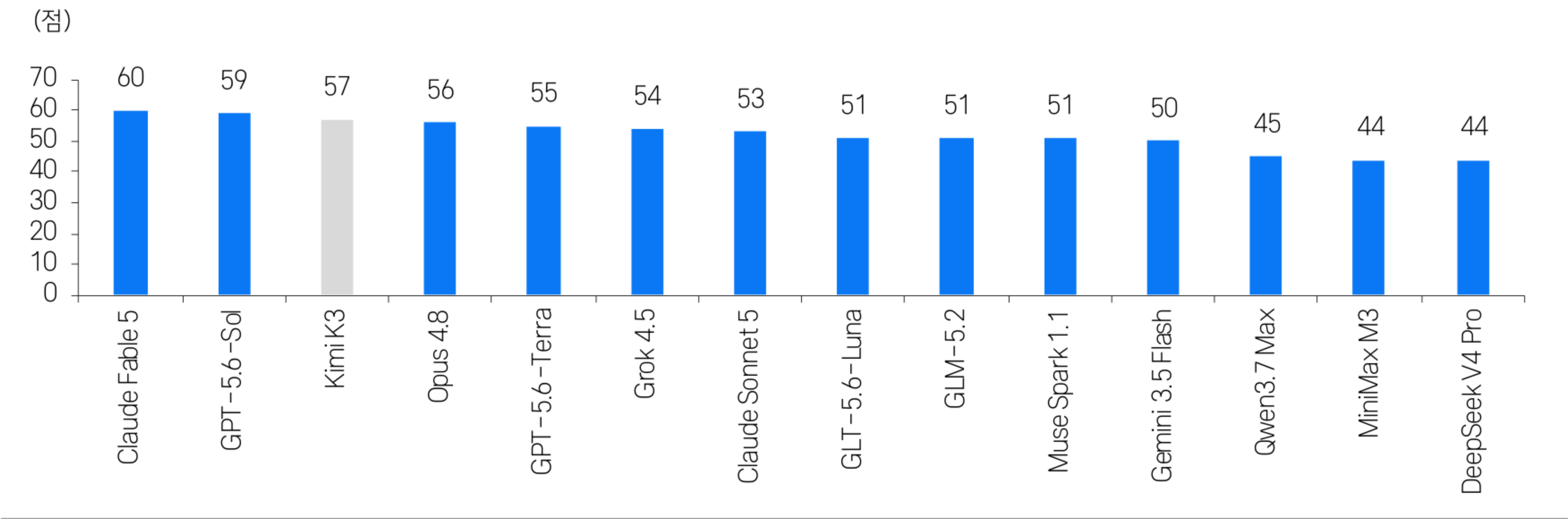
4) Kimi 모델은 전작부터 상당한 토큰 사용과 함께 과추론 이슈가 있었다. 이는 느린 답변 속도로도 연결된다. 중국 모델 특유의 특징이지만, 높은 벤치마크 성능 대비 실제 사용감은 부진하다는 평가도 존재한다. 문샷 AI 자체적으로도 사용자 경험 측면에서 Claude Fable 5와 GPT-5.6-Sol에 비해 여전히 눈에 띄는 격차가 존재한다는 점을 인정하고 있다.

주요 오픈웨이트 모델의 파라미터 규모



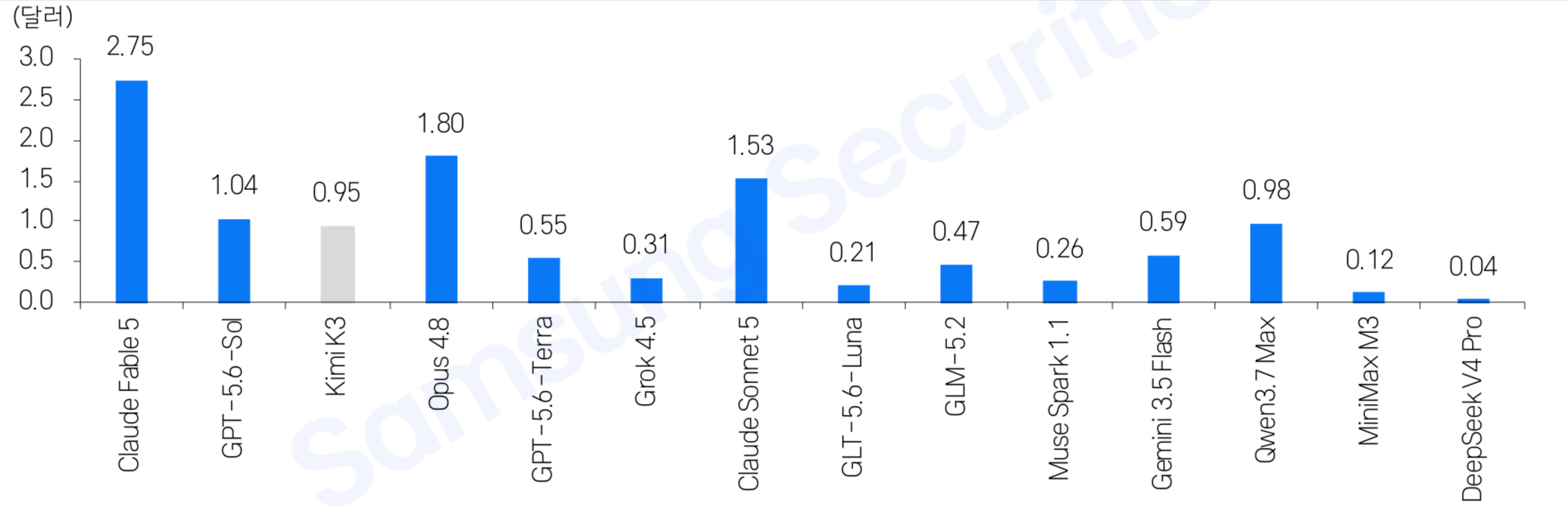
자료: 문샷AI, 삼성증권

주요 모델 Artificial Analysis Intelligence Index 비교



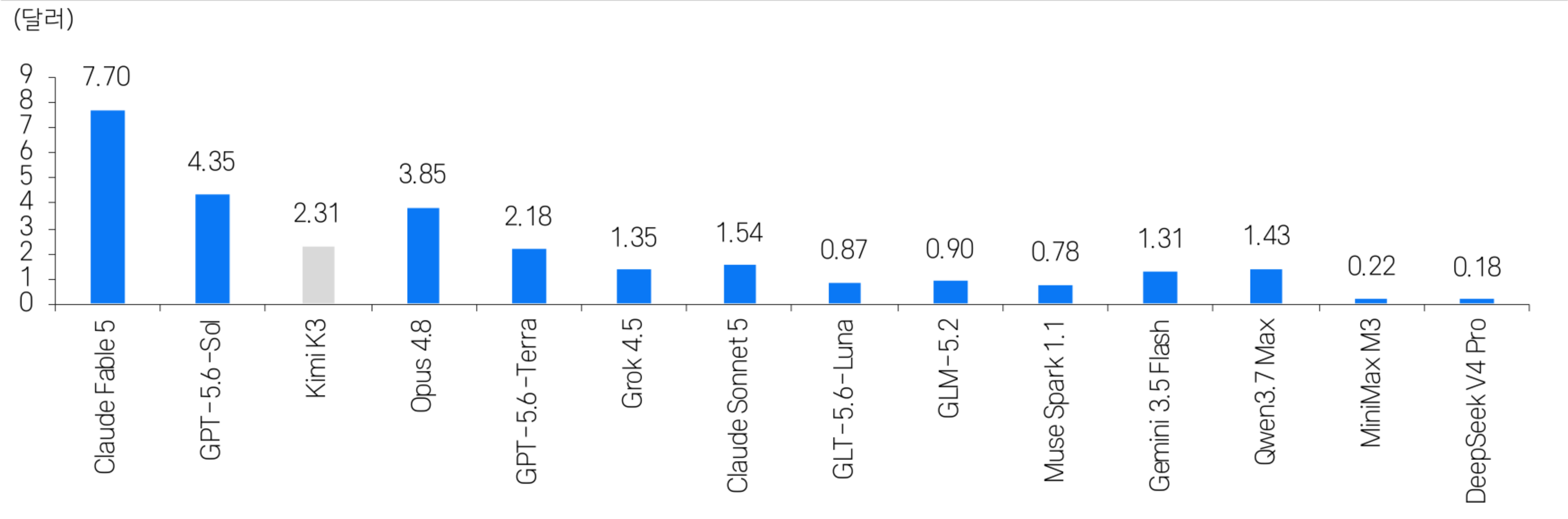
자료: Artificial Analysis, 삼성증권

주요 모델 작업 당 비용 비교



자료: Artificial Analysis, 삼성증권

주요 API 가격 비교(Blended 가격 기준)



참고: Blended 가격은 캐시:인풋:아웃풋 7:2:1 기준

자료: Artificial Analysis, 삼성증권

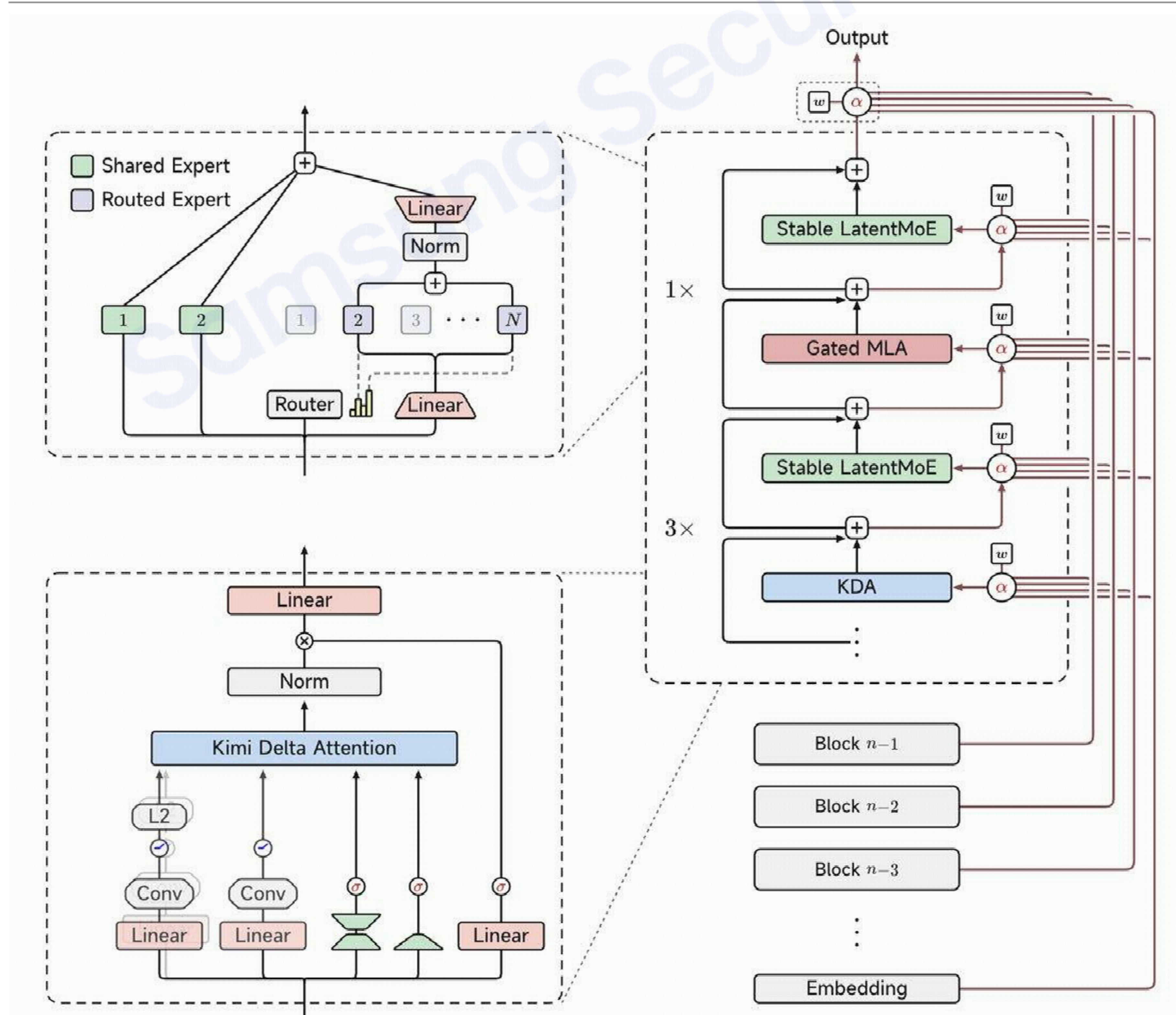
아키텍처 측면 주요 특징

Kimi K3는 KDA와 AttnRes 기반으로 구축되었고, Stable LatentMoE 구조 적용해 Kimi K2 대비 약 2.5배 높은 스케일링 효율을 달성했다는 점을 강조했다. 기술보고서와 웨이트 공개를 통한 검증이 필요하다.

- Stable LatentMoE: 896개의 전문가 중 16개만 활성화하는 고희소성 MoE 구조로, 계산량을 줄이면서 모델 규모를 효율적으로 확장
- Kimi Delta Attention(KDA): 기존 어텐션보다 긴 컨텍스트에서 정보 전달 효율을 높이도록 설계된 새로운 어텐션 구조
- Attention Residuals(AttnRes): 모든 레이어 정보를 누적하는 대신 필요한 중간 표현만 선택적으로 재활용해 정보 전달 효율과 모델 확장성을 높이는 구조

사전학습부터 양자화를 진행한 것도 주요 특징이다. 전작 K2는 사후학습 단계에서 INT4에 적응되었으나, K3는 SFT(지도 미세조정) 단계부터 양자화 인식 훈련을 적용해 MXFP4 가중치와 MXFP8 활성화를 사용하며 하드웨어 호환성을 강화했다. 다만 추론 효율은 더 넓은 고대역폭 네트워킹 영역에서 향상되기 때문에 64개 이상의 가속기를 사용하는 슈퍼노드 구성에서 배포할 것을 권장하고 있다.

Kimi K3 아키텍처



Kimi K3 architecture: the Stable LatentMoE and KDA modules (left), the AttnRes operation α (top right), and the Block Attention Residuals backbone (right).

자료: 문샷AI, 삼성증권

왜 불행은 한번에 몰려오는가?

Kimi 쇼크 vs 딥시크 쇼크, 차이점은?

그렇다면 Kimi 쇼크는 딥시크 쇼크와는 무엇이 다를까? 딥시크 쇼크가 중국 모델의 가성비 쇼크였다면, Kimi 쇼크는 중국이 프론티어 레이스의 선두를 점할 수 있다는 새로운 형태의 공포다. 비슷해 보이지만 분명 다르다.

일부 벤치마크이긴 하나 프론트엔드 코딩에서 미국 프론티어 랩들을 상회하는 성능을 구현했다는 것은 파장을 일으키기에 충분하다. 가뜰이나 구글의 Gemini 3.5 Pro 출시가 코딩 성능 부진으로 인해 미뤄지고 있는 상황에서 발생한 일이기에 더욱 뼈아프다.

제본스의 역설은 작동하지만...

고지능 모델을 서빙하기 위한 인프라, 에너지는 일종의 물리적 병목이다. 중국 기업이라고 피해갈 수 없다. 실제로 Kimi K3의 API 가격은 Claude Sonnet 5와 유사하다. 딥시크의 압도적인 가성비와는 거리가 멀다. 문샷 AI가 Kimi K3 공개 이후 즉각적 컴퓨팅 파워 부족을 겪고 있다는 점도 이를 뒷받침한다.

딥시크 쇼크는 제본스의 역설로 극복되었다. 베이스 시나리오는 이번에도 동일하다. 가격(P)의 하락에 따른 총 수요(토큰 사용량)의 증가 추세는 이어질 것이다. 추론을 중심으로 하는 컴퓨팅 파워 쏠티지는 중국 모델의 선전과 무관하게 지속될 것이고, CapEx 둔화 추세 또한 단기에 현실화될 가능성은 낮을 것으로 판단된다. AI 수요 자극 측면에서 보면, 오픈 소스 중심의 생태계 확장 전략을 지속중인 엔비디아의 일을 중국 기업들이 대신해준 것이나 마찬가지다.

나아가 중국 기업이 프론티어 레이스를 주도하게 될 경우 과거와는 다른 리스크가 새롭게 발생한다.

1) 프론티어 AI 기업의 적정 밸류에 대한 의구심

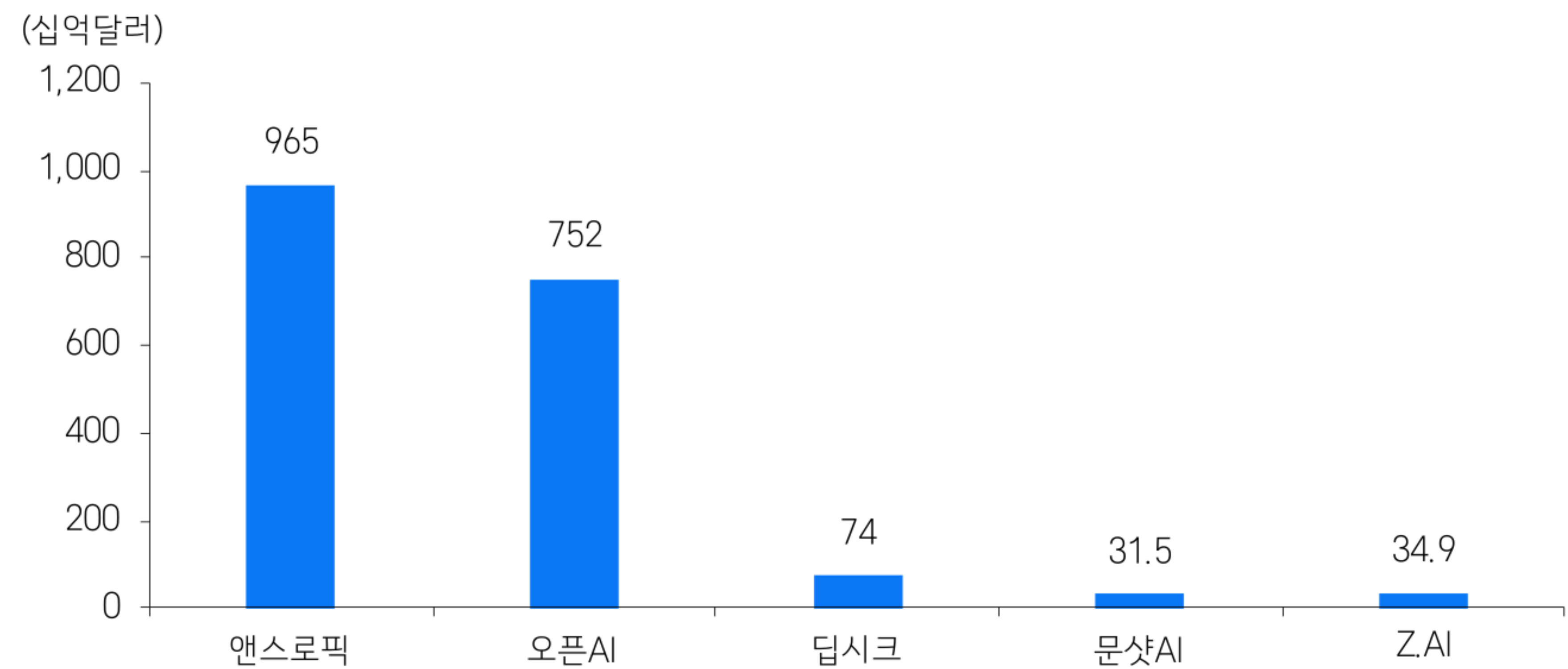
프론티어 AI 랩(오픈AI, 앤스로픽) 밸류의 정당성은 상당 부분 모델 성능에서 기인한다. 현재 시장의 불문율은 최고 성능의 모델을 만들어 압도적 고성장을 지속할 경우 대략 1조 달러 밸류를 부여하겠다는 것이다. 그러나 200~300억달러 밸류로 평가받는 중국 기업이 근소한 격차의 모델을 수주 차이로 만들어내 오픈 웨이트로 풀어낼 수 있다면 기업 가치의 정당화가 어려워진다. 특히 하반기 상장을 앞둔 앤스로픽 입장에서는 상당한 악재다.

현재 빠르게 개화 중인 엔터프라이즈 시장은 크게 코딩과 에이전틱 워크플로우로 구분된다. 현재 프론티어 AI 랩의 핵심 수익원인 코딩 분야가 모델의 성능 개선으로 커머리티화 될 경우 수익성 개선에 대한 전망도 크게 달라질 수 있다. 심지어 에이전틱 워크플로우는 애초에 가성비 모델이 강세인 영역이었다.

타이밍도 나쁘다. 최근 시장의 변동성 확대는 단순히 수요에 대한 불안이 아니다. 막대한 수요를 충족시키기 위한 대량의 CapEx 투자 자체를 리스크로 판단한다. 프론티어 랩의 매출 성장경로 뿐만 아니라 자본시장(채권 발생, 유증) 및 매크로(금리, 침체) 등 모든 요소가 함께 엮여 있다. 실체 없는 메타 쇼크, 일상적인 매크로 지표 하나하나에 민감하게 반응하는 이유다.

예정된 자본 비용 지출이 지속적으로 증가하는 상황에서(반도체 가격 급등, 전력 쏠티지, 데이터센터 반대 시위, 규제, 지연 이슈), 이를 정당화하기 위한 프론티어 랩들의 수익성에 대한 가정에 불확실성이 발생하는 것은 그 자체로 큰 리스크다. 당장 펀더멘털 상의 문제가 발생하지 않아도 끊임없이 기대치를 상회하는 결과물로 증명해야 하는 상황에 놓이게 된다. 펀더멘털이 멀쩡해도, 단기적으로 내러티브 악화에 따른 멀티플 하락 압력이 커지게 된다(멀티플 하락 > 실적 서프라이즈 = 주가 하락).

미-중 프론티어 AI 랩 밸류에이션(시가총액) 비교



참고: 딥시크와 문샷 AI는 신규 펀딩 라운드에서 논의되는 밸류, Z.AI는 시가총액
자료: 언론 종합, Factset, 삼성증권

2) 중국 발 쇼크의 일상화

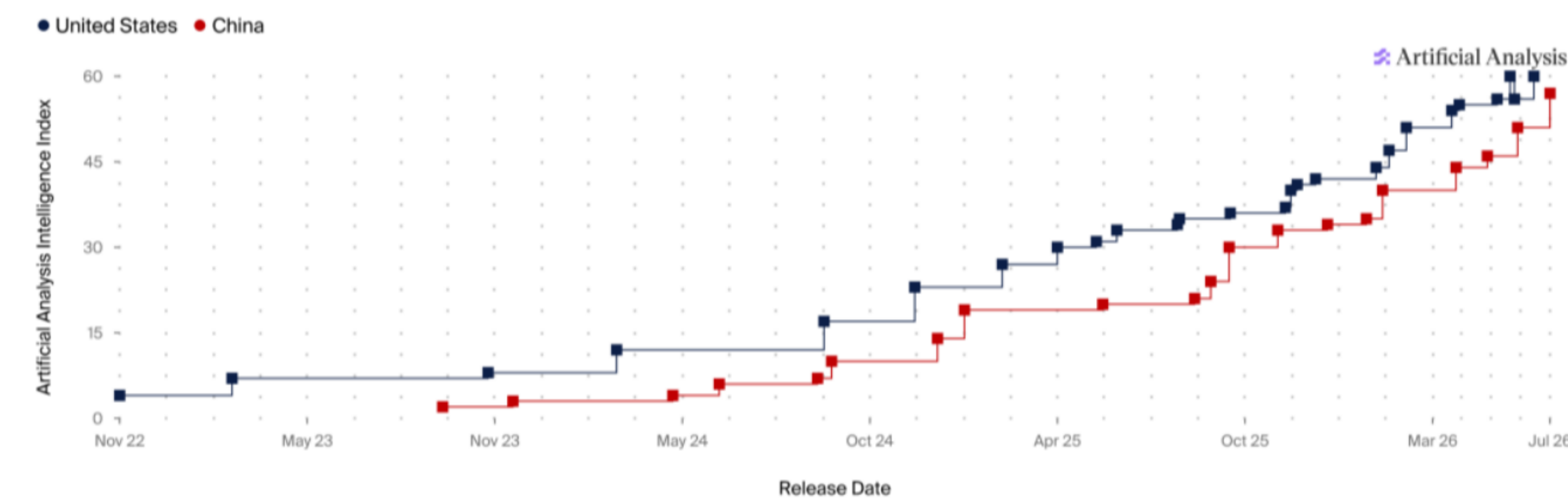
투자자들의 우려를 증폭시키는 것은 중국 연구소들의 상향평준화에 따른 쇼크의 일상화다. DeepSeek, GLM, Kimi까지 앞다투어 놀라운 성능의 모델이 지속적으로 등장하고 있기 때문이다. Z.AI 수석과학자 탕 제는 6월 중순 중국 모델이 Fable 급 수준에 도달하는 시기가 그렇게 오래 걸리지 않을 것이라고 예측한 바 있다. 딥시크 V4 정식 버전이 곧 출시 예정이며, Kimi도 3.1 버전 업데이트를 시사하고 있다.

불과 몇 달 전만 해도 재귀개선 사이클에 올라탄 미국 프론티어 랩들이 중국 기업과의 격차를 벌릴 수 있다는 기대가 있었지만, 현실은 중국 기업들이 프론티어 랩과의 격차를 좁히며 구글, 메타 등 증류를 할 수 없는 기업들을 오히려 압도하고 있다.

Kimi K3로 인해 조금 더 명확해진 사실은 중국 프론티어 랩의 경쟁력이 단순히 증류에만 기인하지는 않는다는 것이다. 1) 최고 수준의 AI 인력, 2) 인프라 부족 상황에서 꽃피운 강화학습 및 최적화 능력, 3) 정부 및 학계와의 긴밀한 협력 등이 뒷받침되고 있다.

일부 벤치마크에서 프론티어 랩의 주력 모델을 상회하기 시작했다는 것은 향후 다른 벤치마크에서 또 다른 역전 사례가 나올 가능성이 높아졌다는 것을 의미한다. 총합은 여전히 미국 프론티어 랩 들이 앞서 있지만, 역전에 대한 공포는 오히려 과거 대비 커진 상황이다.

미-중 프론티어 모델간 격차 추이



자료: Artificial Analysis. 삼성증권

이제 공은 다시 프론티어 랩에게

Kimi K3가 공개된 당일, 오픈AI 샘 올트먼은 “앞으로 12개월이 오픈AI 역사상 최고의 1년이 될 것”이라는 의미심장한 글을 남겼다. 최근 GPT-5.6 및 앤스로픽 대비 풍부한 인프라를 앞세워 고지능 모델의 가성비 전략을 성공시키고 있는 만큼 미래를 상당히 낙관하고 있는 것으로 보인다. GPT-5.6 출시(7/9일) 이후 Codex + ChatGPT Work 사용자는 급증 추세를 지속하며, 기존 500만 명에서 7/12일 600만명 → 7/15일 900만명으로 하루에 백만명 씩 증가하기도 했다.

결국 중요한 것은 사용성이다. 이는 단순 벤치마크로 드러난 성능이나 효율성과는 다르다. 원하는 결과물을 얼마나 적은 토큰, 시간, 전력을 사용해 달성했느냐의 중요성은 지속 높아지고 있다. 지능도 중요하지만 일머리가 더욱 중요한 것과 같다.

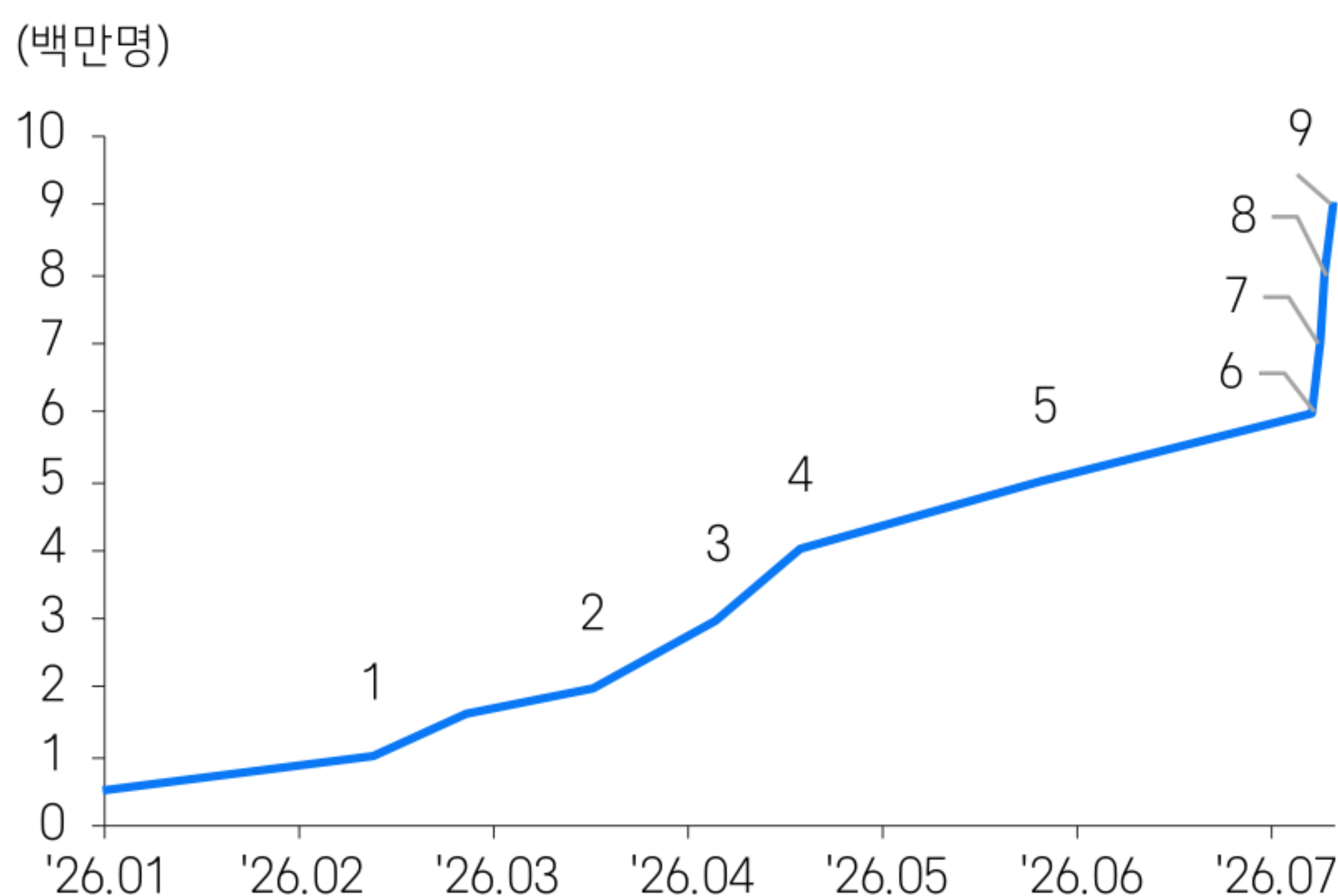
현재 Kimi K3의 경우 상당히 훌륭한 성능에도 불구하고 과추론 및 인덱스 과최적화 이슈가 제기되고 있다. 또한 API 단가는 미국 프론티어 모델보다 낮지만, 실질적 작업당 비용 우위는 크게 축소된다. 가격+속도+성능을 고려하면 탭티어 모델에서 전환할 필요가 낮다는 의견도 존재한다.

또한 중국 기업들은(증류의 도움을 받았을 것으로 추정되는) 학습 부문에서는 지속적으로 높은 효율성을 보여주지만, 의외로 추론 최적화 부문에서는 눈에 띄는 성과를 거두고 있지 못하다. 모델 규모 확장에 비례하겠지만, K3는 전작 대비 API 가격이 크게 상승했고, 아웃풋 토큰은 여전히 많다. 이는 학습과 추론 최적화가 동시에 진행되는 미국 프론티어 랩들의 연구 경향과는 상반되는 결과다. 향후 중국 기업들의 추론 분야 최적화 성공 유무는 중국 기업의 경쟁력 및 인프라 사이클을 가늠하는데 있어 주요 지표가 될 것으로 판단된다.

긍정적으로 생각하면 현재 중국 AI 기업 발 혼란은 본격적인 AI 서비스 시대로 넘어가기 직전 과도기의 성장통일 가능성이 높다. 향후에는 어떤 모델을 사용하는지 보다, 얼마의 비용에 어떤 서비스를 제공했는지가 더욱 직접적인 가치를 만들어낼 것이다.

결국 프론티어 레이스는 모델과 인프라 경쟁력, 두 가지가 모두 갖춰져야 지속 가능하다. 여전히 미국 프론티어 랩들은 경쟁에서 앞서 있으며 중국 기업들의 놀라운 경쟁력에도 불구하고, 아직까지 승패를 단정하기는 이른 국면이다. 다만, 구글, 메타 등 추격을 허용한 기업들에 대한 불확실성은 커질 것이다. 끝없는 증명을 요구하는 무한 경쟁 체제는 그 자체로 AI 테마의 변동성을 키우는 요인이 될 것이다.

오픈AI Codex + 챗GPT Work WAU 추이



자료: 오픈AI, 삼성증권

샘 올트먼 “앞으로 12개월이 오픈AI 역사상 최고의 1년”



자료: X, 삼성증권

Kimi K3, 반도체 수요의 티핑 포인트

기술적 오해와 진실: Kimi K3는 반도체 수요를 촉진

이 모델을 메모리를 덜 쓰는 기술로 해석하는 것은 오해다. 본질은 그 반대라고 생각한다. 디램 사용량이 너무 많아서 아예 쓰지 못했던 모델 성능을 쓸 수 있는 수준으로 효율화한 것이다.

Kimi K3 모델의 핵심은 100만 개의 장문 토큰을, 2.8조 개 파라미터의 초고성능 AI 위에서 실제로 작동할 수 있는 방법을 찾아낸 것에 있다. 표준 어텐션은 문맥이 길어질수록 KV 캐시가 길이의 제곱으로 증가한다. 이 방식으로는 초거대 모델로 100만 토큰을 돌리는 것 자체가 물리적으로 불가능하다. Kimi K3는 그 벽을, 현실에서 배포 가능한 단가 수준까지 끌어내린 것이다. Kimi K3가 반도체 수요 차원에서 던진 메시지는 본 관점에서 해석해야 한다고 본다.

첫째, 컴퓨팅 파워를 줄였다고 보기엔, 기본적으로 지극히 무거운 모델이다.

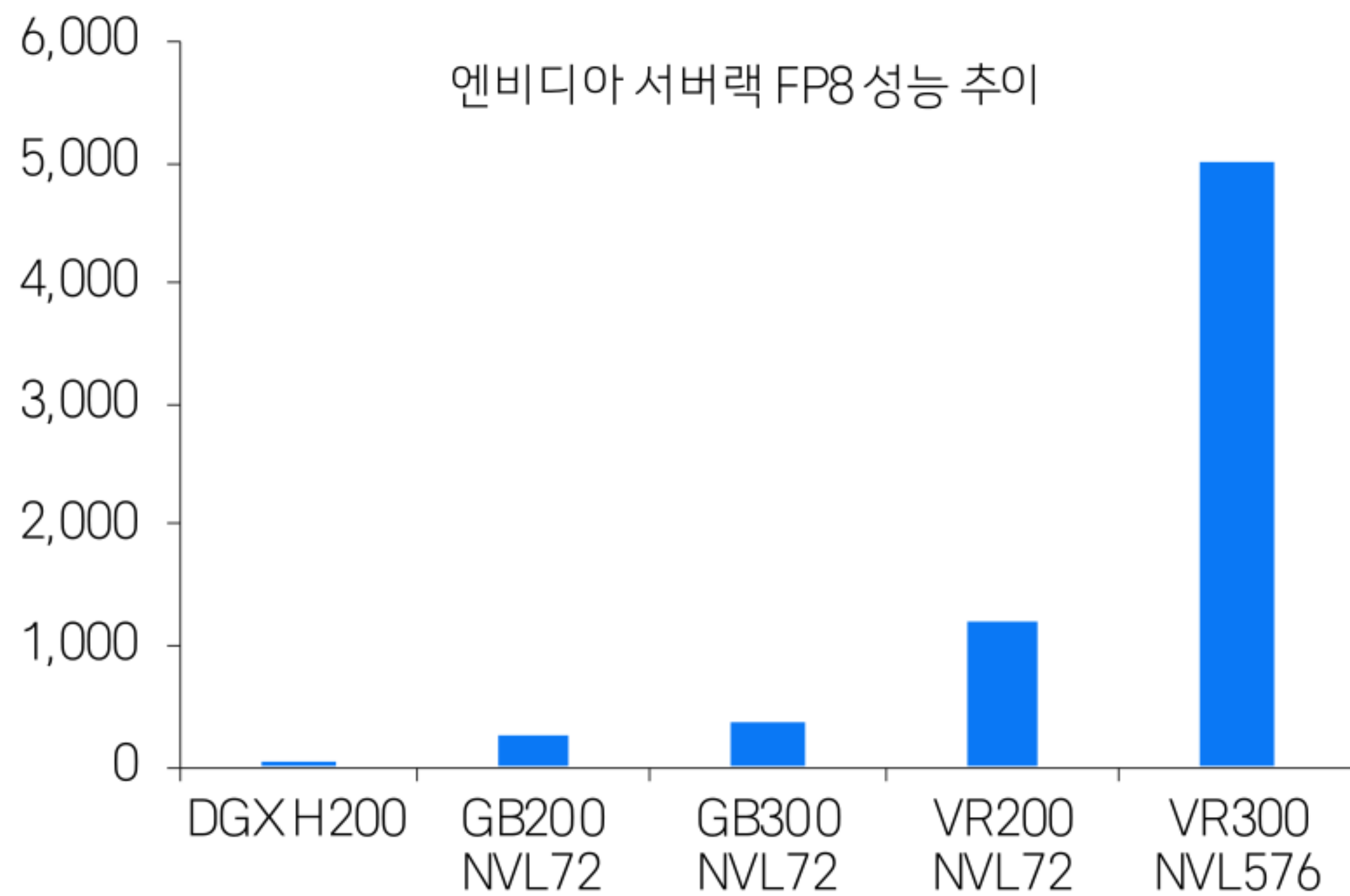
- 문샷 공식 권장 서버는 64개 이상의 가속기로 구성된 슈퍼노드다. 효율화 모델이 랙 스케일 대규모 스케일업 도메인을 권장할 리 없다.
- 모델 가중치를 캐시에 올리는 것만으로 HBM 약 1.5TB를 점유한다. 2.8조 파라미터를 MXFP4 양자화하여 (2.8T×4bit) 나오는 수치이다. 이는 초거대 파라미터의 필연적 결과다.
- KV 캐시를 서버디램, NVMe로 오프로드해야 할 만큼, HBM에 남는 공간이 없다.

총 파라미터는 2.8조로 거대하지만 896개 전문가 중 토큰당 16개만 활성화하는 초희소 MoE(Stable LatentMoE)를 사용한다. 896개나 되는 전문가를 한 GPU에 다 올릴 수 없어 여러 GPU에 흩어 담아야 하고(WideEP), 이는 반도체 관점에서 스케일업(랙 내 고대역폭 연결)을 최우선으로 요구한다는 뜻이다. 간단하게 말해 GPU 동시 연산의 로드는 크지 않고, MoE 사이의 네트워크 수요, 고대역 메모리 요구량이 증가한다. 문샷이 64개 가속기를 연결한 슈퍼노드를 추천하는 이유이다.

둘째, 선형 어텐션(KDA)을 쓰면 KV캐시 사용량을 덜 쓰는 것은 사실이다. 이것이 시장의 오해를 사는 부분이라고 생각한다. 문샷의 Kimi Linear 연구를 보면 KDA는 동일 규모에서 KV 캐시를 최대 75%까지 압축한다고 한다. 그러나 이 효율 추구는 100만 토큰을 감당하기 위해 어쩔 수 없이 선택한 성능 훼손(손실 압축)이다. 절대적인 컴퓨트 자원을 아끼는 것이 주된 목적이 아니다. KDA 효율을 추구하지 않으면 Kimi K3가 추구하는 성능 구현이 힘들다. KV캐시 부담 감소는 모델을 상용화시키려는 노력이었을 뿐 메모리 수요의 하향으로 해석하는 것은 과장된 해석이라고 생각한다.

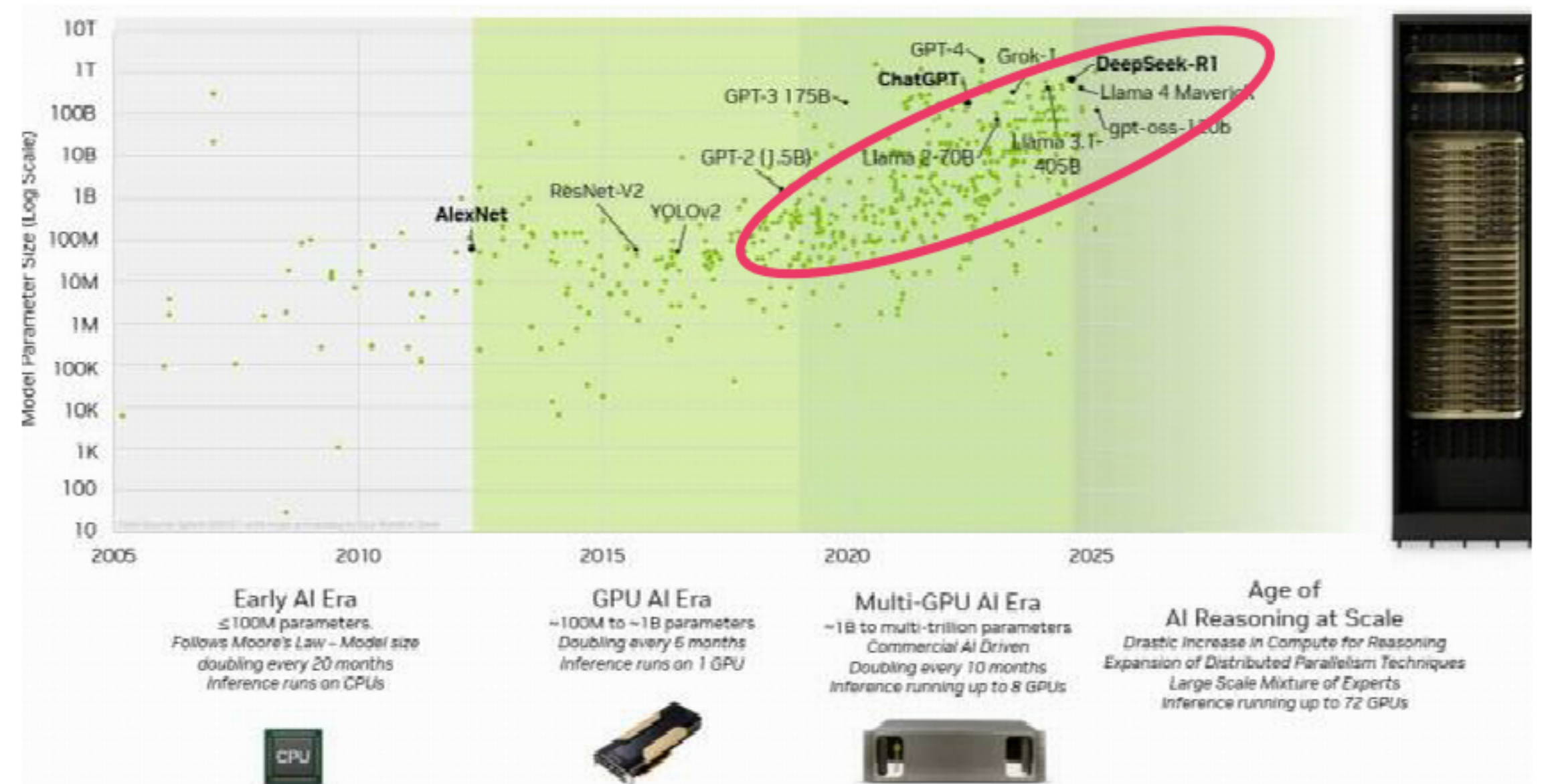
Kimi K3는 엔비디아의 GB200 NVL72 이상을 요구

(PFLOPS)



자료: SemiAnalysis, 삼성증권

KDA와 Wide EP는 스케일아웃 대비 스케일업 수요를 촉발



자료: NVIDIA

경쟁이 앞당긴 수요의 티핑 포인트

AI랩끼리 치열하게 경쟁하는 국면은 오히려 반도체 수요 측면에서는 호재다. 중국 기업들의 AI 기술이 격차를 좁혀오는 것을 본 오픈AI와 엔트로픽은 더 강한 투자를 전개할 가능성이 높다. 지금 상황에서 효율성 경쟁은 중요하지 않다. 결과가 만족스럽지 못하다면 자원을 더욱 투입해야 한다. 미국 하이퍼스케일러들은 미국 AI랩 고객들과 오픈소스의 중국 모델을 쓰려는 고객들을 모두 받아야 하므로 수요는 더욱 강해질 것이라 생각된다.

하지만 중국 AI 모델의 반격을 단순히 점유율 경쟁으로 보는 것은 단편적인 해석이다. 하드웨어 입장에서 경쟁을 통해 AI의 기술 개선 속도가 빨라지는 것이 더 큰 효과를 부른다. 이것은 기술의 초기 국면이기 때문에 나타나는 현상이다. AI기술이 개선되면 최종 소비자들의 사용률이 급격하게 늘어날 것이고, 수요의 S커브에서 티핑 포인트를 지나는 순간이 바로 우리가 기다리던 결정적 장면이다. 이런 측면에서 AI랩들의 성능 경쟁이 반도체 수요의 가장 큰 원동력이다. 오픈AI의 Codex의 사용량 급증에서 가장 전형적인 장면을 목격하고 있다.

앞서 언급한대로 Codex+ChatGPT Work의 사용자는 2월 100만 명에서 7월 14~15일 800만 명으로, 5개월 만에 약 7배 늘었다. 특히 GPT5.6 정식 출시(7월 9일) 직후 3일 만에 300만명이 증가하였다. 600만명에서 800만명으로 급증했다. 급증의 원인은 비용보다 성능에 가깝다. GPT-5.6 Sol은 코딩 에이전트 지수에서 새로운 최고점(80)을 찍으며 출력 토큰은 절반 이하, 소요 시간도 절반 이하, 비용은 약 1/3 낮췄다. 여기에 Codex의 ChatGPT 데스크톱 통합, 리모트 기능, 깃허브 연동 등 서비스 편의성이 더해졌다.

Codex 사용자 급증에서 우리가 얻을 수 있는 교훈은 두 가지이다. 첫째, 지금도 AI의 가격과 성능만 맞으면 소비자는 언제든지 쓸 용의가 있다는 것. 둘째, 급증 현상이 나타난 것으로 볼때 아직 시장은 초기에 불과하다는 것이다. 실제로 이 급증은 오픈AI의 인프라를 압박해 일시적으로 컨텍스트 윈도우를 줄여야 했을 정도인데 이것도 수요가 공급을 앞지르는 초기 시장의 전형 중 하나라고 본다.

최종 소비자가 성능을 원가보다 우위에 두는 한, AI칩과 하이퍼스케일러가 반도체에 돈 쓰는 것을 아까워할 이유가 없다. 초창기 시장이라는 말의 의미는, 성장 S커브가 아직 터지기 전이며, 미국이든 중국이든 누군가 기술만 터뜨려 주면 얼마든지 쓰겠다는 대기 수요가 쌓여 있다는 뜻이다. 조금이라도 성능이 받쳐주면 너도나도 쓸 텐데, Kimi K3가 장문 토큰의 성능에서 진일보한 결과를 보여준 셈이다.

시장을 열어주는 쪽이 나오면, 그것이 미국이든 중국이든 반도체 업체에는 호재다. 중국이 S커브 개화 시점을 앞당겨 주는 것은, 반도체 수요 관점에서 호재로 생각한다.

단, 소프트웨어 파트가 지적인 반대편도 함께 봐야 한다. 수요가 견조해도 자본시장은 이를 정당화하기 위한 막대한 Capex 자체를 리스크로 본다. 그래서 펀더멘털과 별개로 단기 내러티브 악화에 따른 멀티플 하락 압력이 커질 수 있다. 이것은 주가의 방향 문제가 아닌 변동성의 문제이다.

미국과 중국 AI, 공생의 선순환 중

딥시크 R1이 세상에 처음 나왔을 때는 미처 인식하지 못했지만, 이제는 모두가 알고 있다. 기술의 공생 관계가 그것이다.

딥시크 R1은 초거대 파라미터 일변도의 시장에서, 적당한 크기(671B, 활성 37B)에 MoE를 강화하면 추론을 상업적 배포 가능성까지 효율화할 수 있음을 보여주었다. 딥시크의 등장은 미국 AI 산업을 일깨웠고, 결과적으로 Agent AI가 상용화되는 데 기여했다. 그리고 Agent AI는 반도체 수요의 방향을 뺏속부터 바꾸었다 - AI가속기와 HBM수요 뿐 아니라 대량 토큰 소비와 KV캐시를 바탕으로 한 CPU와 서버디램의 수요를 촉발했다.

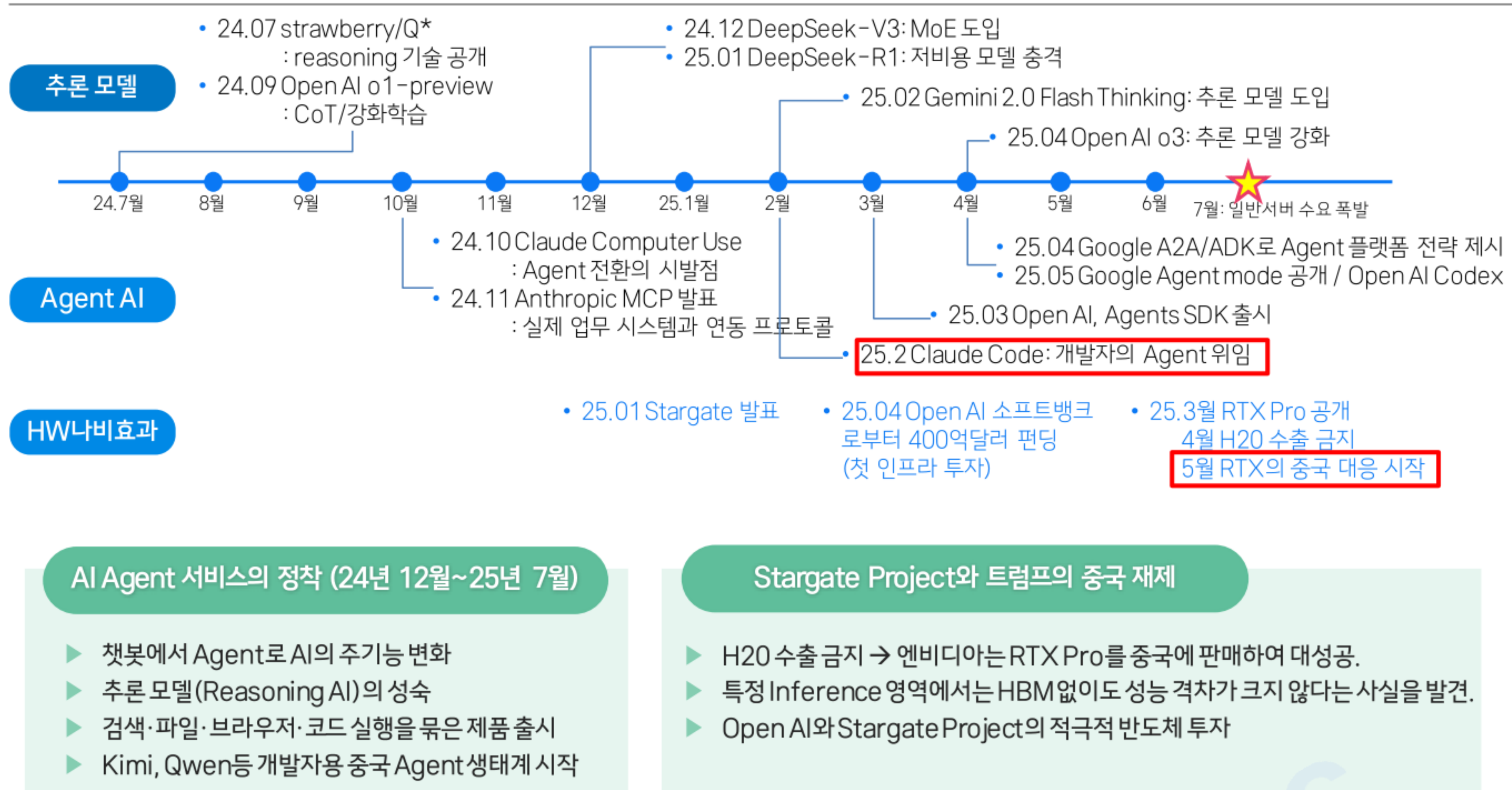
지나고 나서 보니 딥시크 모먼트라는 것은 하나의 반복되는 패턴이 될 것이라 생각한다. 미국이 압도적 자본력과 학계 선도 연구로 방향성을 먼저 제시하면, 중국이 하드웨어 효율을 쥐어짜는 엔지니어링 아키텍처를 완성하고, 미국 AI 기업이 이를 흡수해 서비스로 디플로이하는 기술의 공생과 경쟁 관계다.

- **Reasoning AI:** 1) 방향 설정(미국): GPT o1의 추론 시간 연산 패러다임 제시 → 2) 현실화(중국): 딥시크가 671B MoE 모델로 배포 가능한 비용모델 제시 → 3) 재흡수(미국): 프론티어랩의 적극적인 MoE 채택
- **Long-Horizon Agent:** 1) 방향 설정(미국): 제미니의 장토큰 패러다임 + 선형 어텐션 이론 → 2) 현실화(중국): Kimi K3가 KDA+MoE로 장토큰 자율 에이전트 비용모델 제시 → 3) 재흡수(미국): 선형 어텐션과 극단적 MoE 세분화 예상

문샷의 창업자 양즈린(Yang Zhilin)은 카네기멜론대(CMU) 박사 출신으로, 장문 시퀀스 모델링의 고전인 Transformer-XL과 XLNet의 제1저자다. 즉 그의 연구 DNA 자체가 '초장문 맥락(Long-context)'이다. 문샷이 창업 초기부터 자신들의 전장을 초장문 맥락으로 정의한 것은 우연이 아니다. 2.8조 규모의 초거대 지능에 100만 단어를 다 기억하는 모델을 만들면서 실현 가능한 단가를 만든 모델의 공개는 중국이 미국의 방향성에 동조하는 일종의 티핑 포인트였다.

이제 미국 빅테크들도 차세대 에이전트 모델(차기 GPT·클로드·제미니 라인업)의 인프라 단가를 낮추고 장문 에이전트 성능을 강화하기 위해, Kimi K3가 증명한 이 구조적 레시피를 벤치마킹할 것이다. 그 결과, 지금까지 Agent AI의 한계였던 성능 강화가 곧 나타날 것이다. 이것은 AI의 기술적 한계 때문에 열리지 못했던 새로운 Agent 시장 하나가 막 열리기 직전이라는 뜻이다. 이것이 딥시크가 남긴 교훈이다. 미국 수요든 중국 수요든, 한 쪽이 시장을 열면 나머지도 열린다. 그리고 시장이 열리면 인프라는 더 필요하다.

AI agent의 개화를 촉발한 DeepSeek R1



자료: 삼성증권 정리

반도체 피크아웃은 두 가지 중 하나로 온다

역설적으로, 강세를 주장하려면 언제 이 사이클이 꺾이는지를 먼저 정의해야 한다. 반도체 피크아웃은 다음 두 경로 중 하나로 올 것이라고 생각한다.

첫째, 모든 기술 사이클은 기술이 멈추면서 끝난다. AI의 성능이 더 이상 앞으로 달려나가지 못할 때 AI 사이클은 일단락된다. AI로 바꿀 산업이 소진되거나, 모델이 더 보여줄 것이 없거나, AI 가속기의 성능 개선이 멈추거나, 메모리 대역폭이 더는 늘지 않거나, 반도체 고집적 기술이 바닥나면 AI 사이클은 앞으로 나아가기 어렵다. 그러나 앞서 언급한대로 새 Agent 시장이 막 열리는 정반대의 국면이다. 성능은 여전히 달려나가고 있다.

둘째, 하이퍼스케일러들이 자금을 더 이상 조달할 수 없을 때 사이클이 끝난다. 이것이 이번 사이클에서 유독 강조되는 특징이다. 수요의 원천이 개인 디바이스가 아니라 인프라이고, 금융 자본을 투입해 그 인프라를 짓기 시작했기 때문이다. 그래서 이번 사이클은 프론티어랩 매출뿐 아니라 자본시장(금리, 유동성)이 엮여 있다 앞서 실체 없는 메타 쇼크에 시장이 민감한 이유다. 하지만 역설적으로, 금융 자본 사이클은 이제 막 시작 단계이므로 단기에 끝날 가능성은 낮다.

메모리 섹터에 국한하면 하나를 더 걱정해야 한다. AI 사이클은 계속되지만 메모리가 사이클의 중심에서 이탈하는 경우다(예를 들면 연산이나 네트워크가 병목이 되어 메모리 증설이 후순위로 밀리는 국면). 다만 이 경우는 상대적인 주가 수익률 관점에서의 리스크일 뿐, 급격한 주가 하락을 설명하지는 못한다. 오히려 Kimi K3처럼 무게가 HBM을 포화시키고 KV 캐시를 DDR5와 NVMe로 오프로드하는 구조는, 메모리를 사이클 중심에서 이탈시키는 게 아니라 오히려 중심으로 끌어당기는 사건이다.

이번 이벤트는 주가의 강한 컨빅션을 부른다

우리는 삼성전자와 SK하이닉스의 P/B 1.8배와 2.5배가 단기 저점이라고 전망하며, 이미 현 주가는 단기 저점에 도달했다. 이번에는 주가의センチメント가 바닥 국면에 도달했다고 본다. 시장이 호재조차 악재로 오인하는 네거티브 바이어스로 돌아서 있다는 뜻이다.

1월의 딥시크가 결국 7월 이후 Agent AI 사이클의 원인 중 하나가 되었듯, 이번 Kimi K3 역시 장토큰 자율 에이전트의 성능 강화를 위한 자양분이 될 것이다. 우리는 이번 이벤트를 통해, AI 사이클이 끝나지 않았고 반도체 수요가 오히려 더 강해질 수 있다는 증거를 확보했다.

- Kimi K3는 반도체를 덜 쓰는 모델이 아니라, 못 쓰던 능력을 쓸 수 있게 만든 모델이다
- AI 모델 경쟁은 기술 개선을 촉진하고 반도체 수요의 티핑포인트를 앞당긴다.
- 피크아웃의 트리거 - 기술 정체와 자금 조달 한계는 아직 멀었고 메모리 중요성은 유지되고 있다

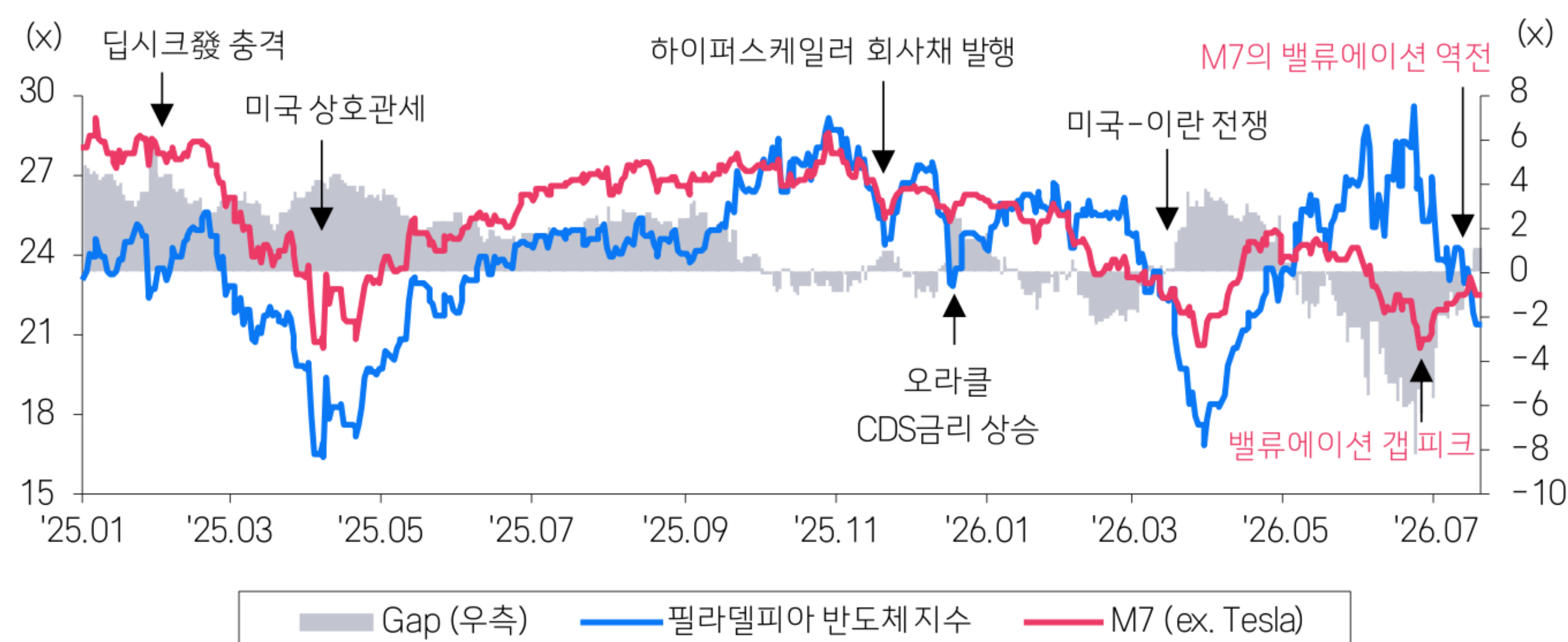
이제 주가 조정을 끝내기 위해 남은 것은 두 가지라고 생각한다.

첫째, 주식 쏠림의 정상화이다. 이번 조정의 가장 큰 이유는 수급 쏠림이라고 본다. 이 쏠림이 해소될 때까지는 기다림이 필요하다. 필라델피아 반도체 지수 P/E가 M7(ex Tesla)의 P/E보다 하락한 것은 쏠림 현상이 해소되고 있다는 증거다.

둘째, 반등의 트리거를 찾는 것이다. 짧게는 클라우드 기업의 실적 발표에서, 길게는 Agent AI 성능 개선을 통한 수요의 실체적 변화에서 올 것이라 믿는다.

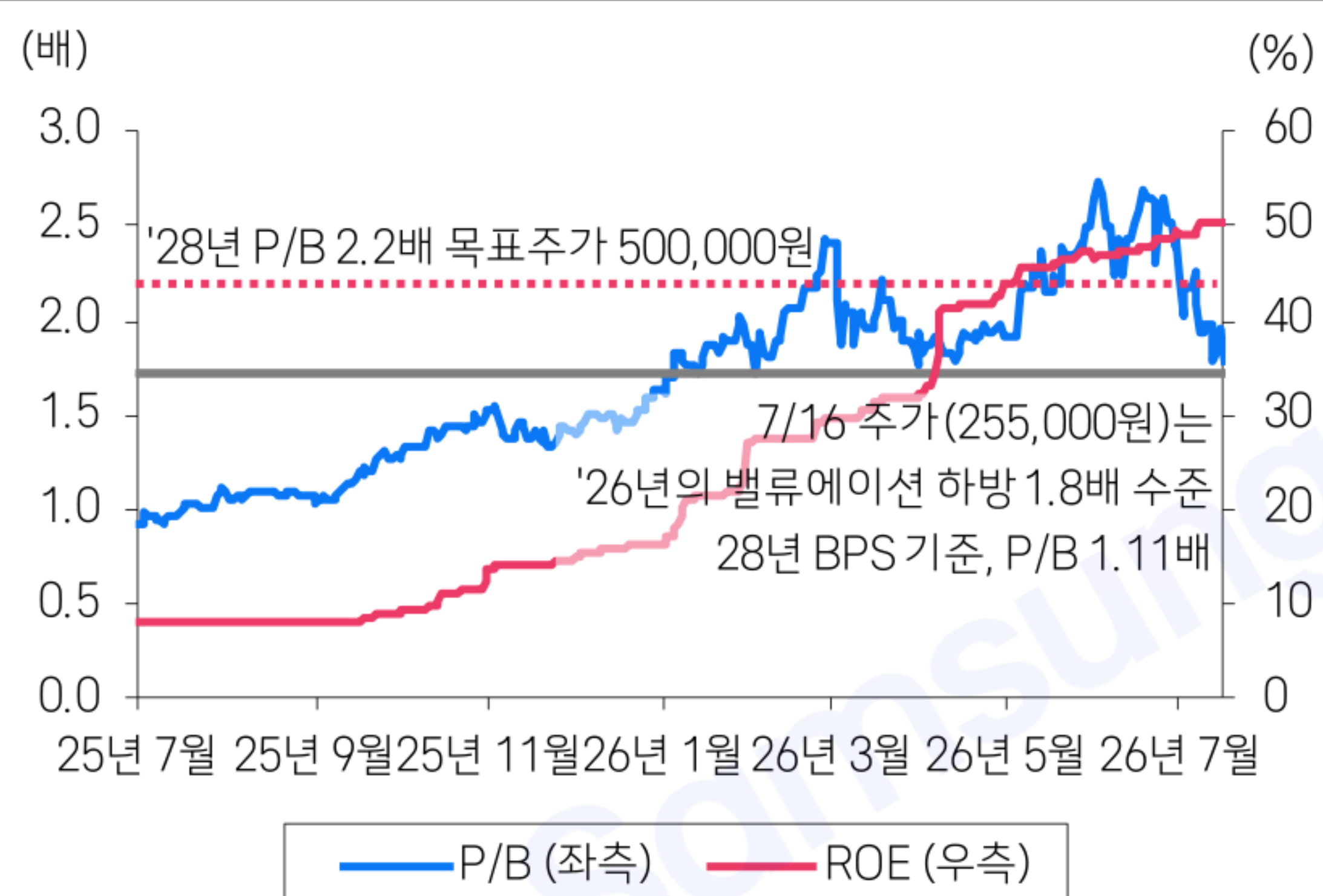
전쟁은 확산되었고, 역설적으로 총알은 더 필요할 것이다. 이번 사이클의 반도체의 수요는 꺾이지 않았다.

필라델피아 반도체 지수 vs M7(테슬라 제외): 12m fwd P/E



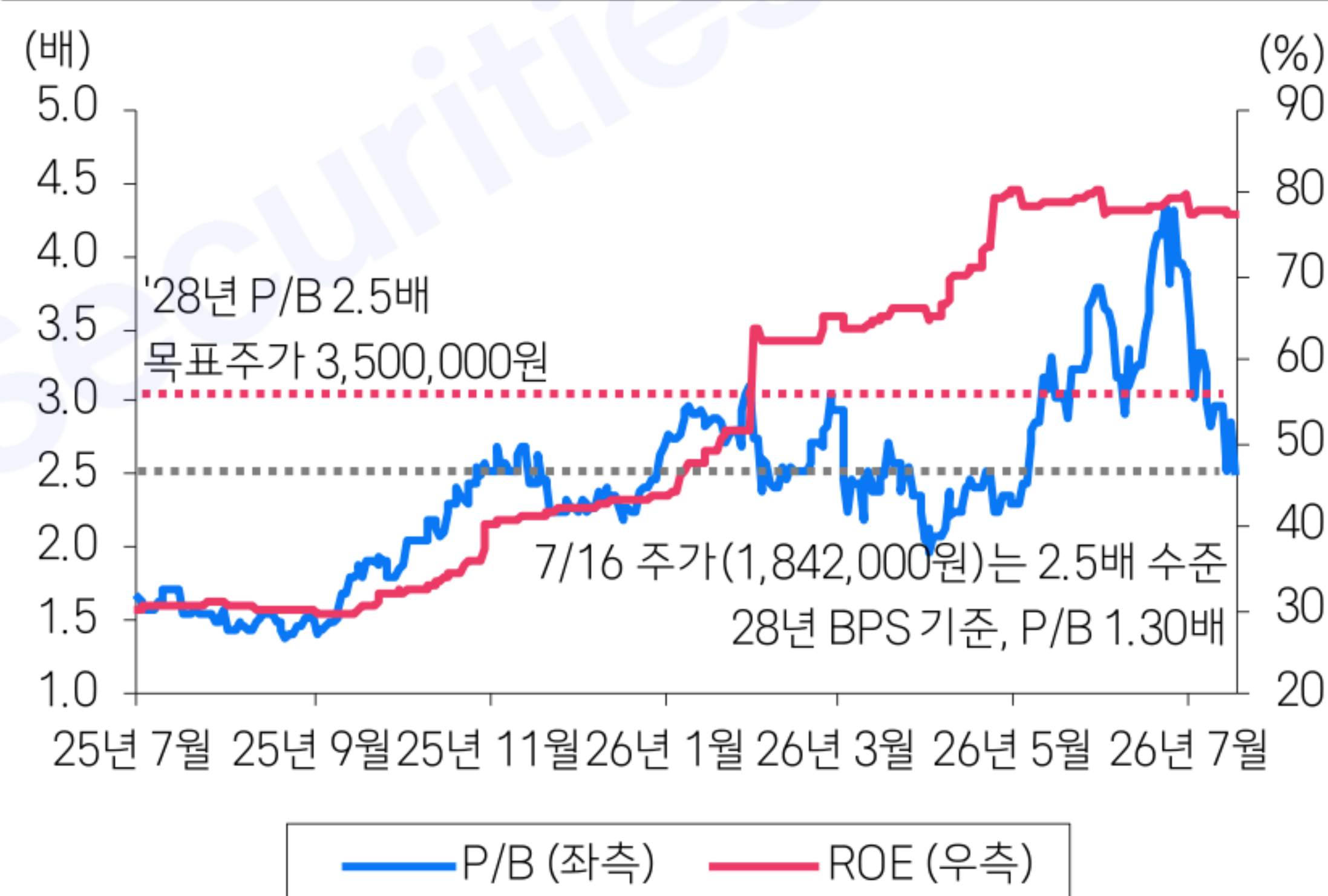
자료: Quantiverse, 삼성증권

삼성전자: 12m fwd P/B vs ROE



자료: Quantiverse, 삼성증권

SK하이닉스: 12m fwd P/B vs ROE

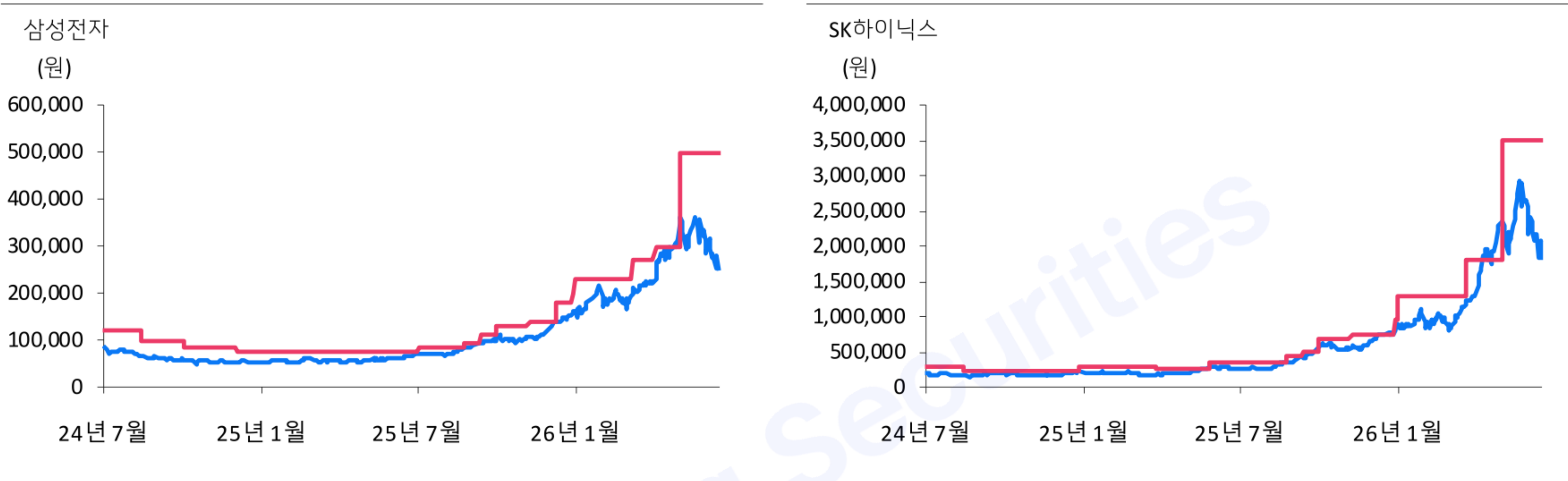


자료: Quantiverse, 삼성증권

Compliance notice

- 당사는 2026년 7월 16일 현재 삼성전자와(과) 계열사 관계에 있습니다.
- 본 조사분석자료의 애널리스트는 2026년 7월 16일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2026년 7월 16일 현재 위 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 조사분석자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.
- 본 조사분석자료는 당사의 저작물로서 모든 저작권은 당사에게 있습니다.
- 본 조사분석자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 조사분석자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.
- 본 조사분석자료는 기관투자가 등 제3자에게 사전 제공된 사실이 없습니다.

2년간 목표주가 변경 추이



최근 2년간 투자의견 및 목표주가 변경 (수정주가 기준)

삼성전자												
일 자	2024/5/28	9/12	11/1	2025/1/2	8/1	9/23	10/13	10/31	12/8	2026/1/9	1/27	1/30
투자의견	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY
TP (원)	120000	100000	83000	74000	85000	93000	110000	130000	140000	180000	200000	230000
괴리율 (평균)	-34.45	-39.49	-33.92	-21.93	-15.05	-6.95	-10.64	-21.96	-16.54	-18.69	-19.23	-19.75
괴리율 (최대or최소)	-26.83	-35.30	-29.28	-1.89	-1.76	1.51	-5.36	-14.54	0.71	-15.39	-18.80	-5.22
일 자	4/7	5/4	6/2									
투자의견	BUY	BUY	BUY									
TP (원)	270000	300000	500000									
괴리율 (평균)	-20.25	-3.39										
괴리율 (최대or최소)	-16.30	16.33										
SK하이닉스												
일 자	2024/7/12	9/12	2025/1/24	4/24	6/25	9/23	10/13	10/30	12/8	2026/1/27	1/30	4/21
투자의견	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY	BUY
TP (원)	280000	240000	280000	250000	340000	430000	500000	700000	750000	950000	1300000	1800000
괴리율 (평균)	-34.15	-23.78	-30.26	-14.53	-17.52	-14.81	-3.33	-19.05	-11.48	-10.42	-26.68	-3.42
괴리율 (최대or최소)	-16.79	-6.04	-21.96	11.40	3.82	-0.47	11.60	-11.43	2.27	-9.37	-10.31	31.28
일 자	6/2											
투자의견	BUY											
TP (원)	3500000											
괴리율 (평균)												
괴리율 (최대or최소)												

투자기간 및 투자등급: 삼성증권은 기업 및 산업에 대한 투자등급을 아래와 같이 구분합니다.

* 2023년 7월 27일부로 기업 투자 등급 기준 변경

기업		산업	
BUY (매수)	향후 12개월간 예상 절대수익률 15% 이상 그리고 업종 내 상대매력도가 평균 대비 높은 수준	OVERWEIGHT(비중확대)	향후 12개월간 업종지수상승률이 시장수익률 대비 5% 이상 상승 예상
HOLD (중립)	향후 12개월간 예상 절대수익률 -15%~ 15% 내외	NEUTRAL(중립)	향후 12개월간 업종지수상승률이 시장수익률과 유사한 수준 (±5%) 예상
SELL (매도)	향후 12개월간 예상 절대수익률 -15% 이하	UNDERWEIGHT(비중축소)	향후 12개월간 업종지수상승률이 시장수익률 대비 5% 이상 하락 예상

최근 1년간 조사분석자료의 투자등급 비율 2026.06.30 기준
매수(85.5%)·중립(14.5%)·매도(0%)

Samsung Securities

삼성증권

삼성증권주식회사

서울특별시 서초구 서초대로74길 11(삼성전자빌딩)
Tel: 02 2020 8000 / www.samsungpop.com

삼성증권 Family Center: 1588 2323

고객 불편사항 접수: 080 911 0900



Member of
**Dow Jones
Sustainability Indices**
Powered by the S&P Global CSA