

SECTOR UPDATE

2026. 9

글로벌주식팀

이영진 Senior Analyst, CFA
youngjin91.lee@samsung.com

본 자료는 고객의 투자에 참고가 될 수 있는 각종 정보 제공을 목적으로 제작되었습니다. 본 자료는 합리적인 정보를 바탕으로 작성된 것이지만, 투자권유의 적합성이나 완전성을 보장하지 못합니다. 따라서 투자 판단의 최종 책임은 투자자 본인에게 있으며, 본 자료는 어떠한 경우라도 법적 책임소재의 증빙으로 사용될 수 없습니다.

삼성증권



글로벌 AI / SW

지능의 풍요와 에이전트 작업 시간의 확장

지능의 풍요와 에이전트 작업 시간의 확장

재귀개선 사이클 기반으로 가속화되고 있는 AI의 산업 발전 속 AI 모델 기업 간 경쟁도 끝없이 이루어지고 있습니다. 에이전트 AI 확산과 엔터프라이즈 침투에 따른 토큰 사용량도 폭발하고 있는데요. 하지만 단순한 Q의 증가와 P의 하락보다 현 시점 AI 산업의 중요한 주제 중 하나는 지능의 풍요입니다.

프론티어 모델 간 성능 격차 축소 → 저가 모델 사용량 증가 → 이에 대응하는 기존 프론티어 모델 가격 인하 기반으로 업무에서 쓸 만한 모델이라는 선택지는 더욱 풍부해졌습니다. 벤치마크로 확인되는 범용적 성능을 넘어 실제 유저 단에서 체감되는 성능 향상, 사용량 제한, 빠른 롤아웃 등의 중요성도 높아지고 있습니다.

치열한 경쟁 속 오픈AI와 앤스로픽을 둘러싼 바람의 방향도 바뀌고 있습니다. 앤스로픽이 여전히 재무적 관점에서 우위를 보이고 있는 것은 사실이나, 최근 ARR 성장 트렌드에서 변화가 나타나고 있습니다. 오픈AI는 GPT-5.6 시리즈를 통해 Codex WAU 2,500만 명을 돌파했습니다. 작년 말 코드 레드를 선언한 이후 본격적으로 주도권을 되찾은 상황입니다.

특히 GPT-6 Astra는 결정적 한 방입니다. 새로운 규칙을 파악하고 추론하는 능력의 강화를 기반으로 AI 에이전트가 자율적으로 끝까지 작업을 수행할 수 있는 유효 시간의 확장이라는 의미를 가지고 있는데요. Astra가 보여준 발전은 AI 에이전트가 처리할 수 있는 업무 범위의 확장으로 이어집니다.

당연하게도 공격적 캐파 확장은 불가피합니다. 컴퓨터에 자신감을 보였던 오픈AI조차 Astra의 높은 수요에 따라 월 \$200의 Pro 플랜의 신규 구독을 일시 중단하기도 했습니다. 이는 클라우드 인프라 기업에게는 주요 모멘텀이 될 것입니다.

AI 트렌드 변화는 SW 섹터에게도 직접적인 영향을 미치고 있습니다. 앞서 살펴본 AI 산업 내 주요 이슈들은 각각 긍정적 요인과 부정적 요인으로 나누어 볼 수 있고, 양 측이 모두 혼재되어 번갈아 가며 파급력을 미치는 상황입니다.

SaaS와 같은 엔드 소프트웨어는 이러한 AI발 내러티브 변화에 더욱 취약한 모습을 보일 수 있습니다. 결국 장기적으로 에이전트가 구동되는 인프라 역할을 담당하는 소프트웨어(클라우드 인프라, 사이버 보안, 데이터) 기업이 긍정적이라는 생각을 유지합니다.

PART
01

AI - 풍부해진 모델과 장기 작업 능력 향상의 조화

Samsung Securities

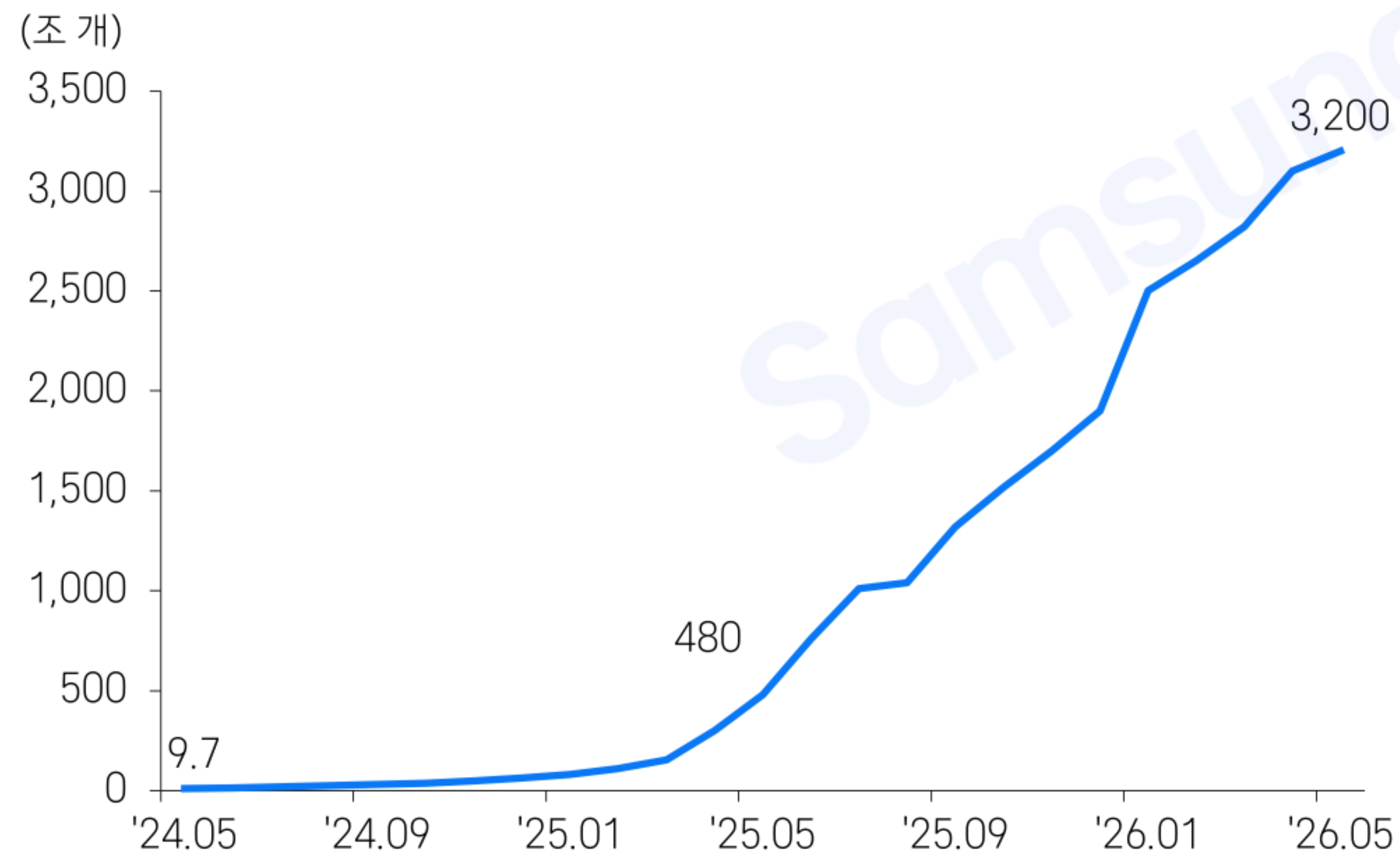
견조한 AI 수요 | 이제는 지겨워진 토큰 증가 트렌드

▶ 본격적인 에이전틱 AI 확산에 따른 토큰 사용량 급증. 오픈AI 경영진은 AI 제품 수요가 '수직 벽'처럼 증가한다는 코멘트

- 구글 월간 토큰 처리량: '24년 5월 9.7조 개 → '25년 5월 480조 개 → '26년 5월 3,200조 개(+7x YoY). API 분당 토큰 처리량 190억 개(+6x)
- OpenRouter 플랫폼 내 주간 토큰 처리량: 무료 제공 이벤트 영향이 있지만, 꾸준한 증가세. 현재 주간 130조 개 페이스. 연초 10조 개 미만
- 기업 고객의 AI 도입은 PoC(개념 검증, 테스트) 단계 이후 본격화. 주요 기업은 AI 발 생산성 강화와 투자 자원 마련을 위한 인력 감축도 발표
- Tokenmaxxing 트렌드 이후 AI의 효율적 처리 및 활용 중요성 상승 + 토큰 가격(P) 하락 트렌드 → 그러나 사용량(Q) 증가가 더욱 가파른 상황

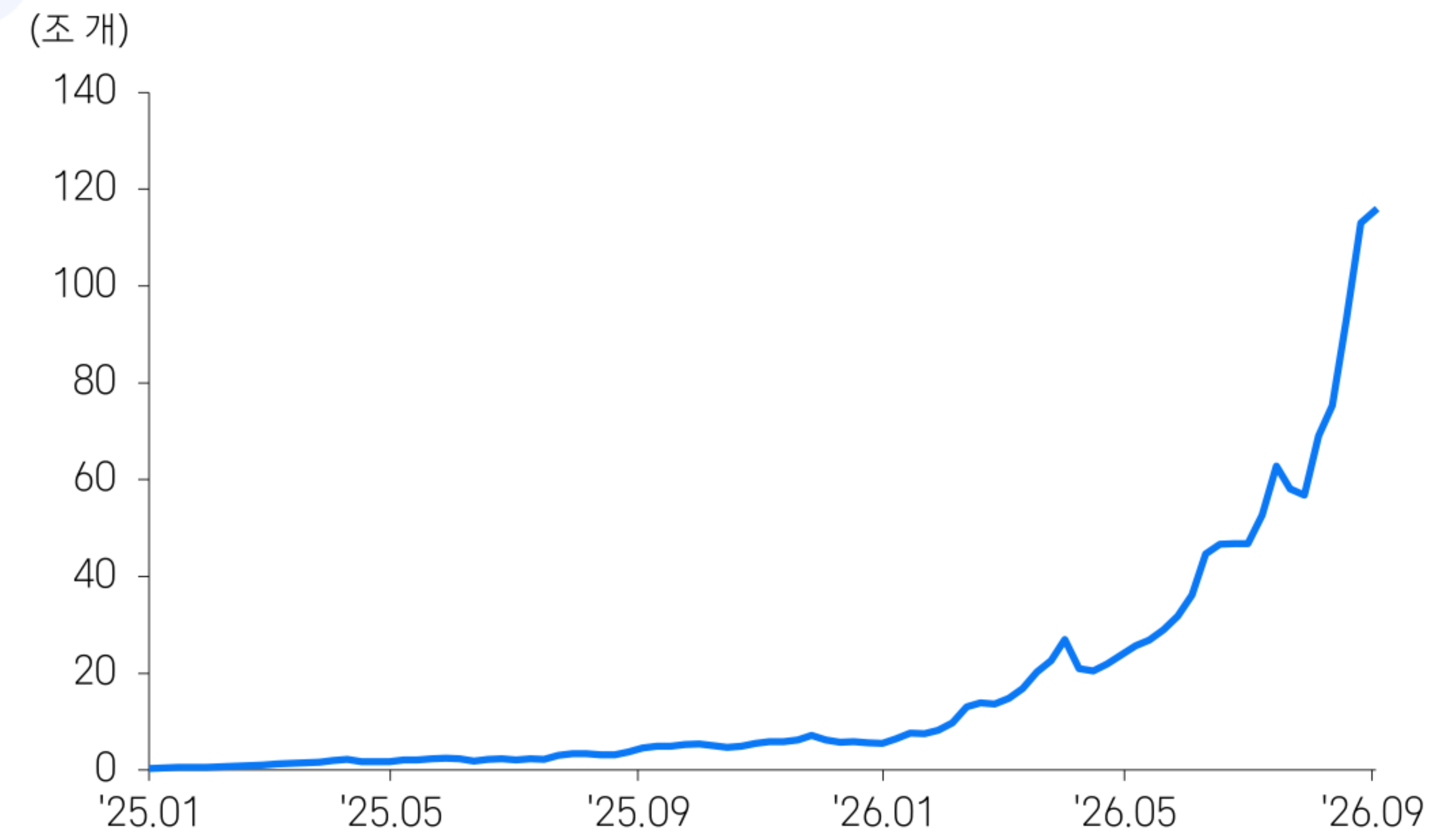
* Tokenmaxxing(토큰 맥싱): 기업의 AI 도입 초기 성과 및 사용 확대를 위해 직원에게 최대한 많은 토큰을 사용하도록 유도하는 전략

구글 월간 토큰 처리량 추이



자료: Google, 삼성증권

OpenRouter 주간 토큰 처리량 추이



자료: OpenRouter, 삼성증권

AI 모델 경쟁 | 돌고 도는 모델 출시 사이클

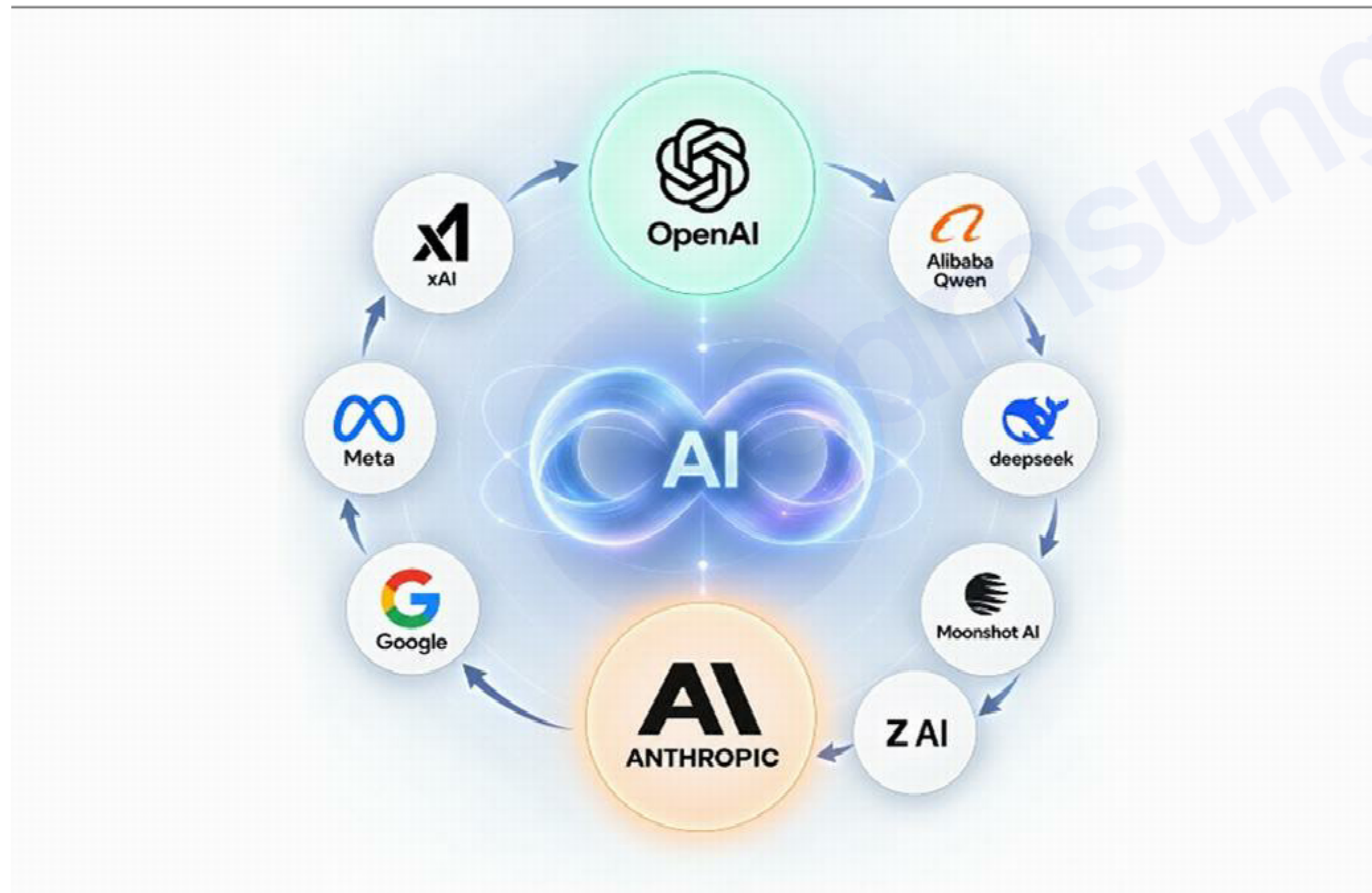
▶ 현 시점 프론티어 AI 기업은 오픈AI와 앤스로픽. 다만 xAI, 메타, 구글 그리고 중국 AI 기업의 추격도 지속

- 재귀개선 사이클 기반으로 프론티어 AI 기업의 모델 출시 사이클은 급격하게 가속. 최고의 모델 자리를 지키는 기간과 모델 간 격차 축소 트렌드
- 벤치마크로 확인되는 범용적 모델 성능보다 실제 유저 단에서 체감되는 성능 향상, 사용량 제한, 안정적 운영, 빠른 롤아웃 등의 중요성 상승
- 특히 오픈AI의 Astra는 AI 에이전트가 자율적으로 작업을 끝까지 지속 수행할 수 있는 유효 작업 시간(Time Horizon) 확장 측면의 의의 존재

* 재귀개선: AI가 스스로 AI 연구 개발에 참여해 더 뛰어난 AI를 만들고, 다시 후속 AI 발전을 가속화하는 반복적 개선 과정

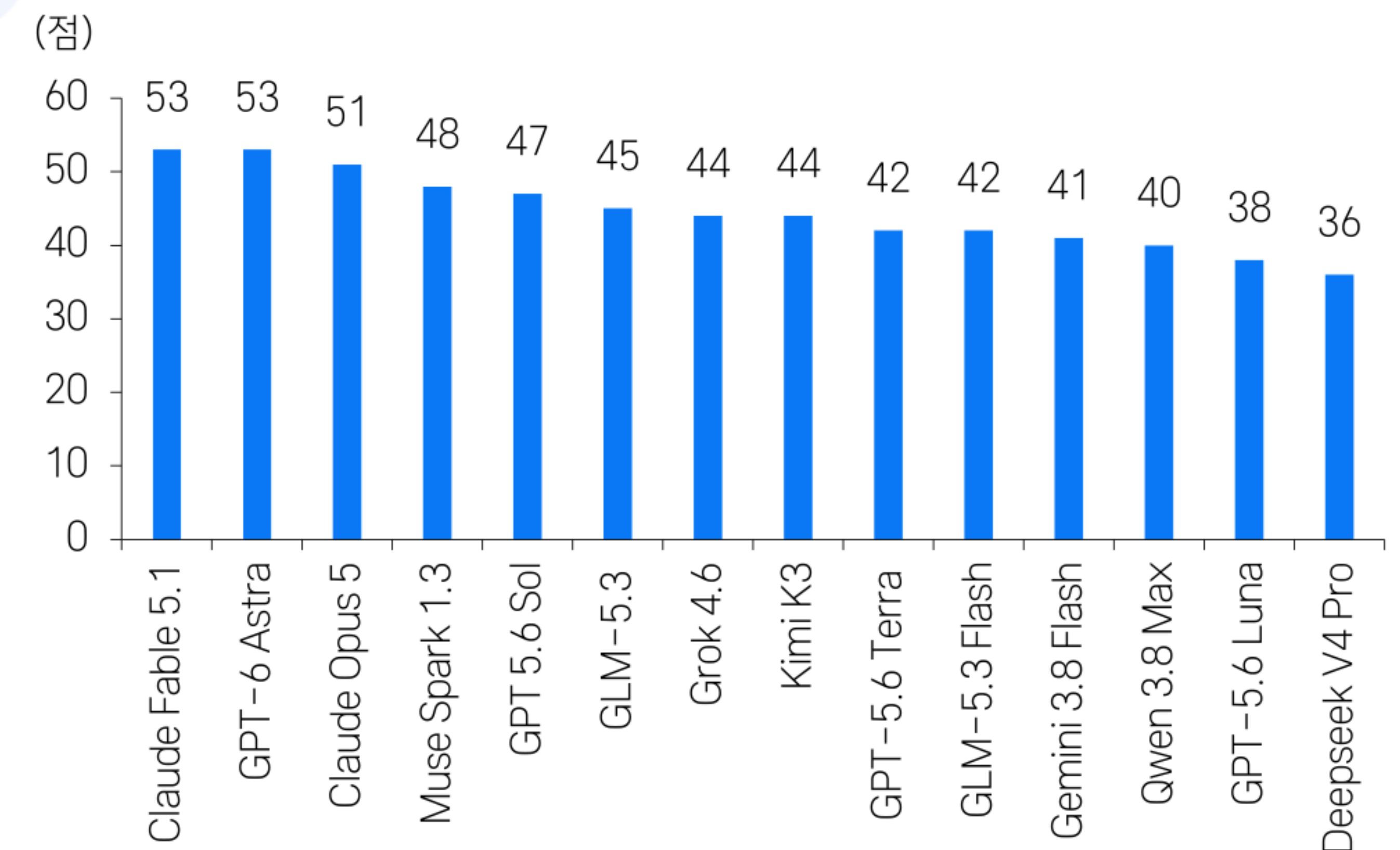
* SOTA(State Of The Art): 특정 분야에서 나온 기술이나 모델 중 최고 수준의 성능을 의미

돌고 도는 AI 모델 출시 사이클



자료: 삼성증권

주요 AI 모델의 벤치마크 순위 비교(AAI Index 기준)



자료: Artificial Analysis, 삼성증권

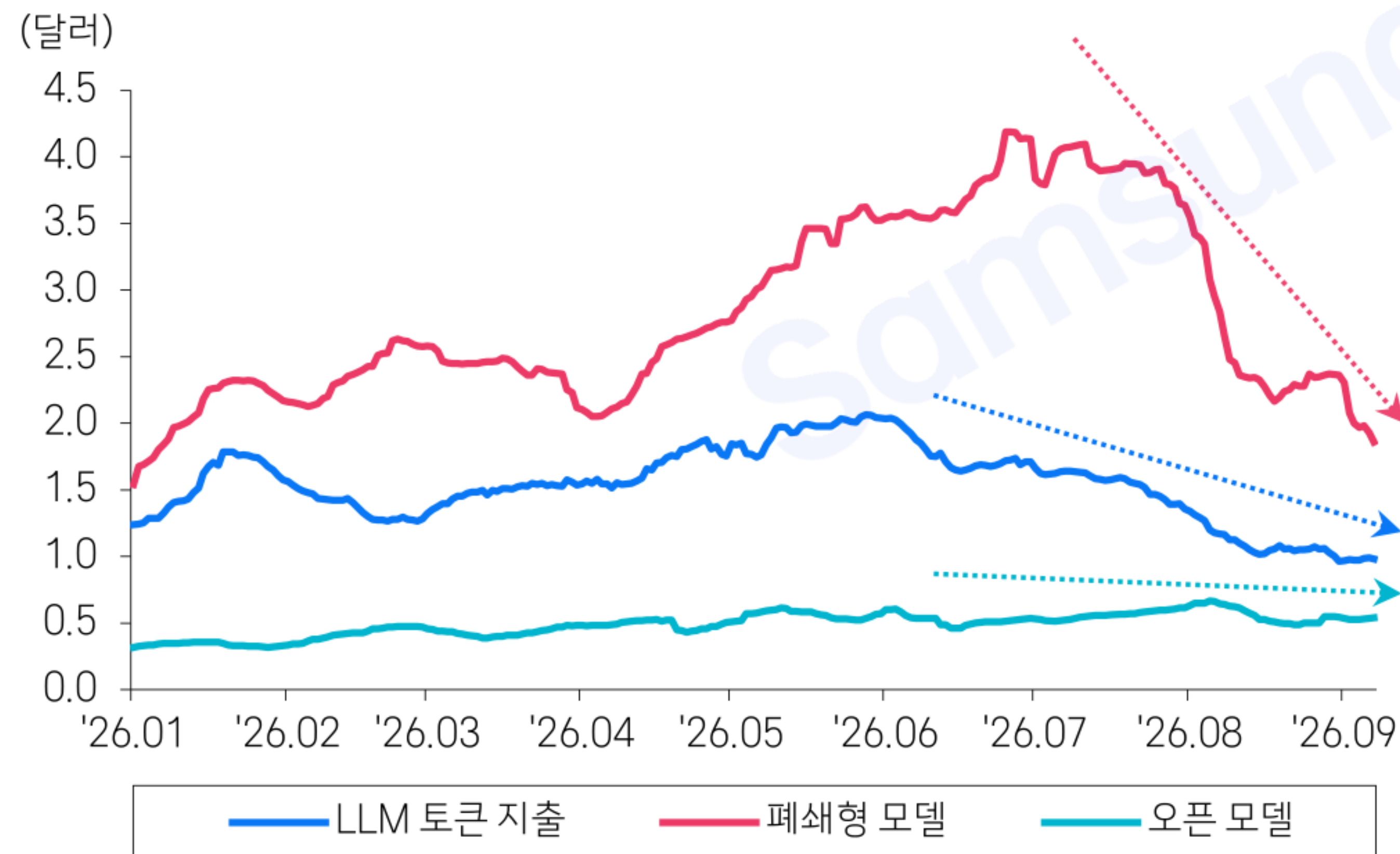
지능의 풍요 | 치열한 경쟁 속 많아진 선택지

▶ 모델 경쟁 심화 + 효율성 추구에서 시작된 토큰 가격 하락 → 지능 자체보다 이를 활용하는 플랫폼 역량이 부각 국면

- 프론티어 모델 간 성능 격차 축소 → 저가 모델 사용량 증가 → 기존 프론티어 모델 가격 인하 = LLM Expenditure 인덱스 하락 요인
- 특히 에이전트에 탑재해 일상 업무에서 쓸 만한(AAI Index 지능 점수 40점 ~ 50점) 모델의 가격 인하가 이전보다 가파르게 진행
- 지능의 희소성이 낮아지며 단순 모델 간 성능 경쟁에서 지능(모델)의 효율적 배포 및 실질적 가치 창출로 내러티브 이동
- 미국 프론티어 모델이 향유해 온 성능 프리미엄의 희석 가능성 존재. 다만 가격 인하에 따른 수요 탄력성이 더 큰 상황 → 절대적 이익 규모 증가

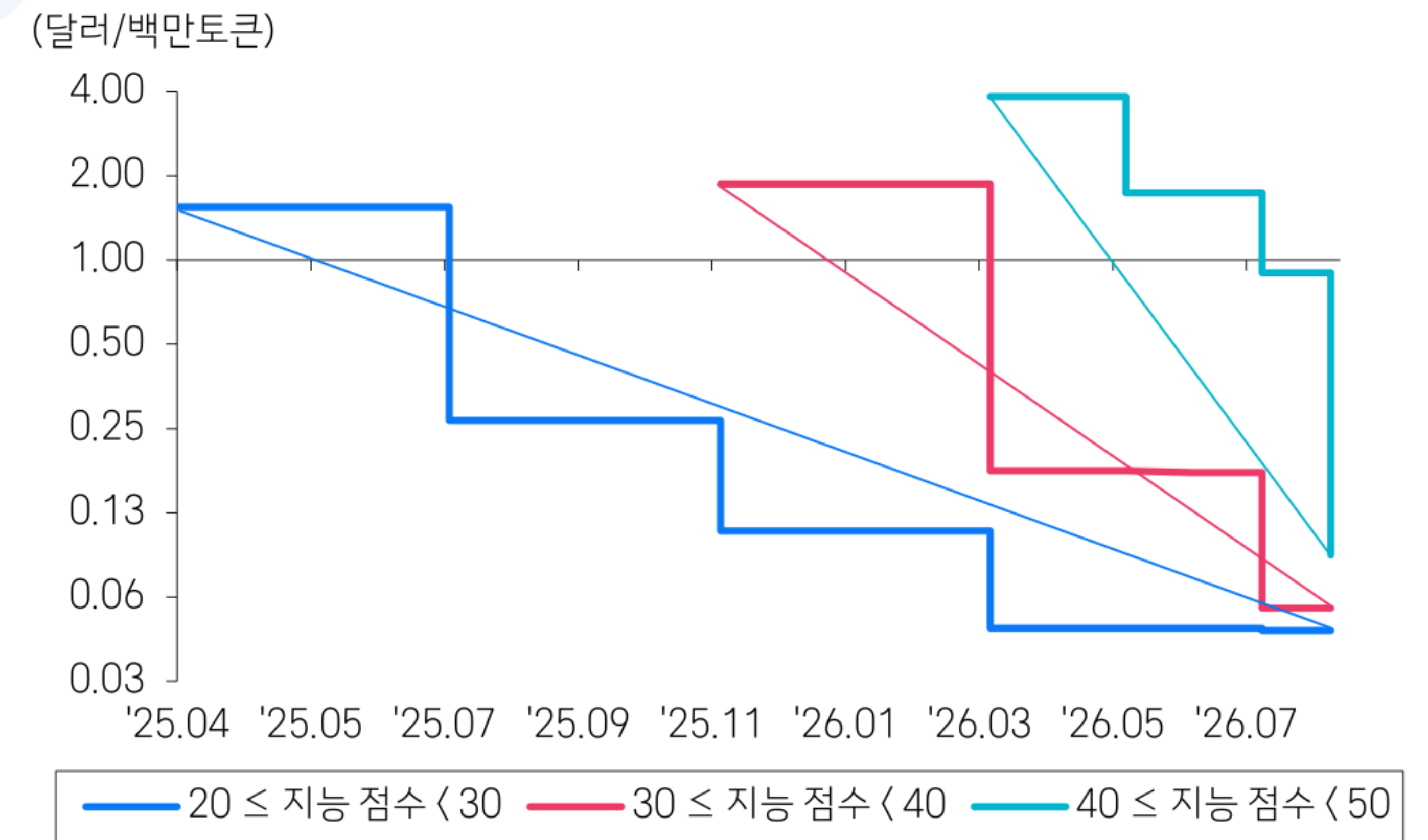
* 프론티어 모델: AI 모델 비교에서 가장 선두에 있는 것으로 평가되는 모델

LLM Expenditure 인덱스 하락. 극적인 변화는 폐쇄형 모델



자료: SiliconData, 삼성증권

지능 점수가 높은 '쓸 만한 모델'의 가격 인하 가속화



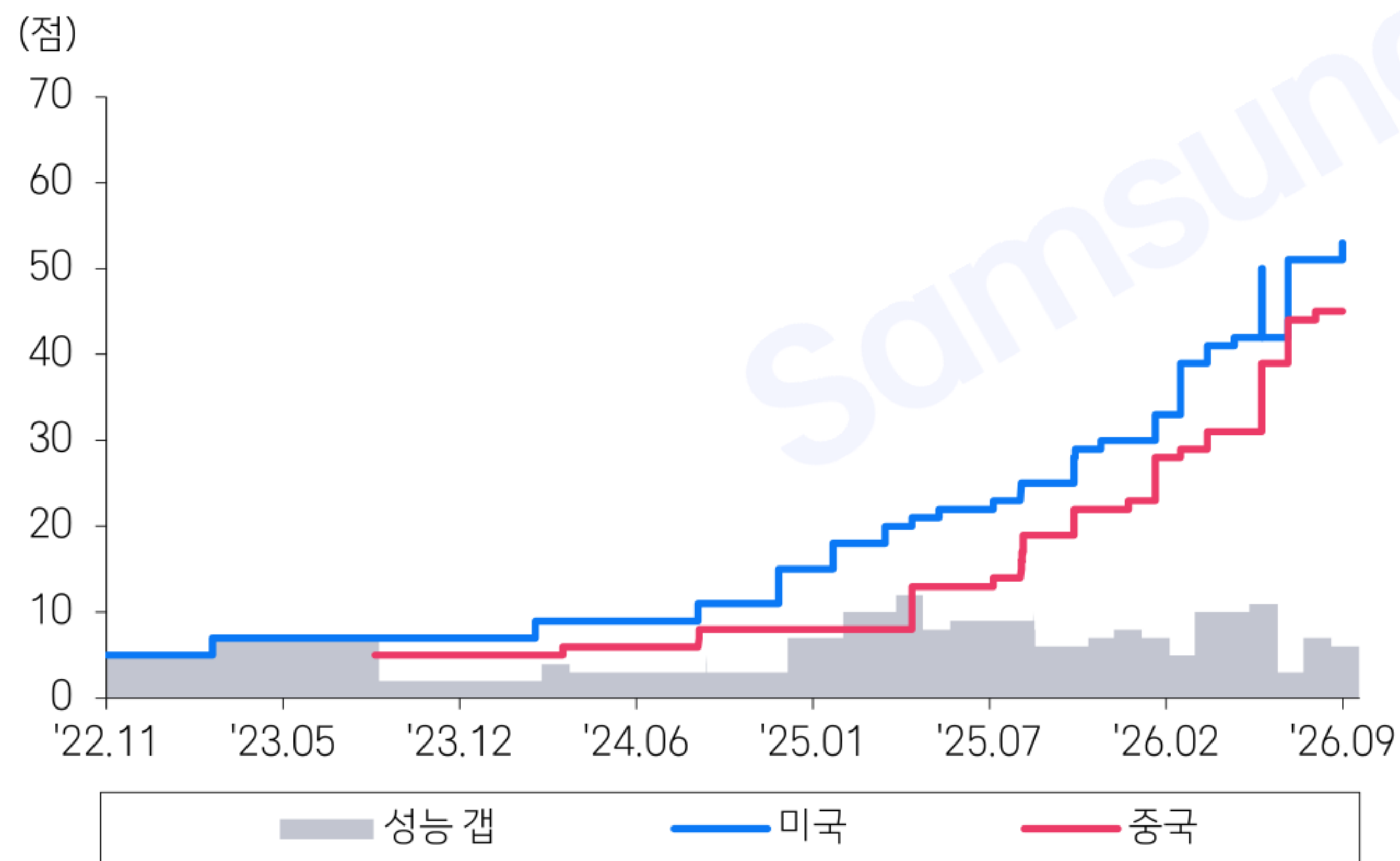
자료: Artificial Analysis, 삼성증권

미-중 경쟁도 계속된다 | 중국의 추격과 수익화의 필요성

▶ 2조 파라미터 이상의 중국 모델(ex Kimi K3) 등장을 비롯해 미-중 모델 간 성능 격차는 빠르게 축소

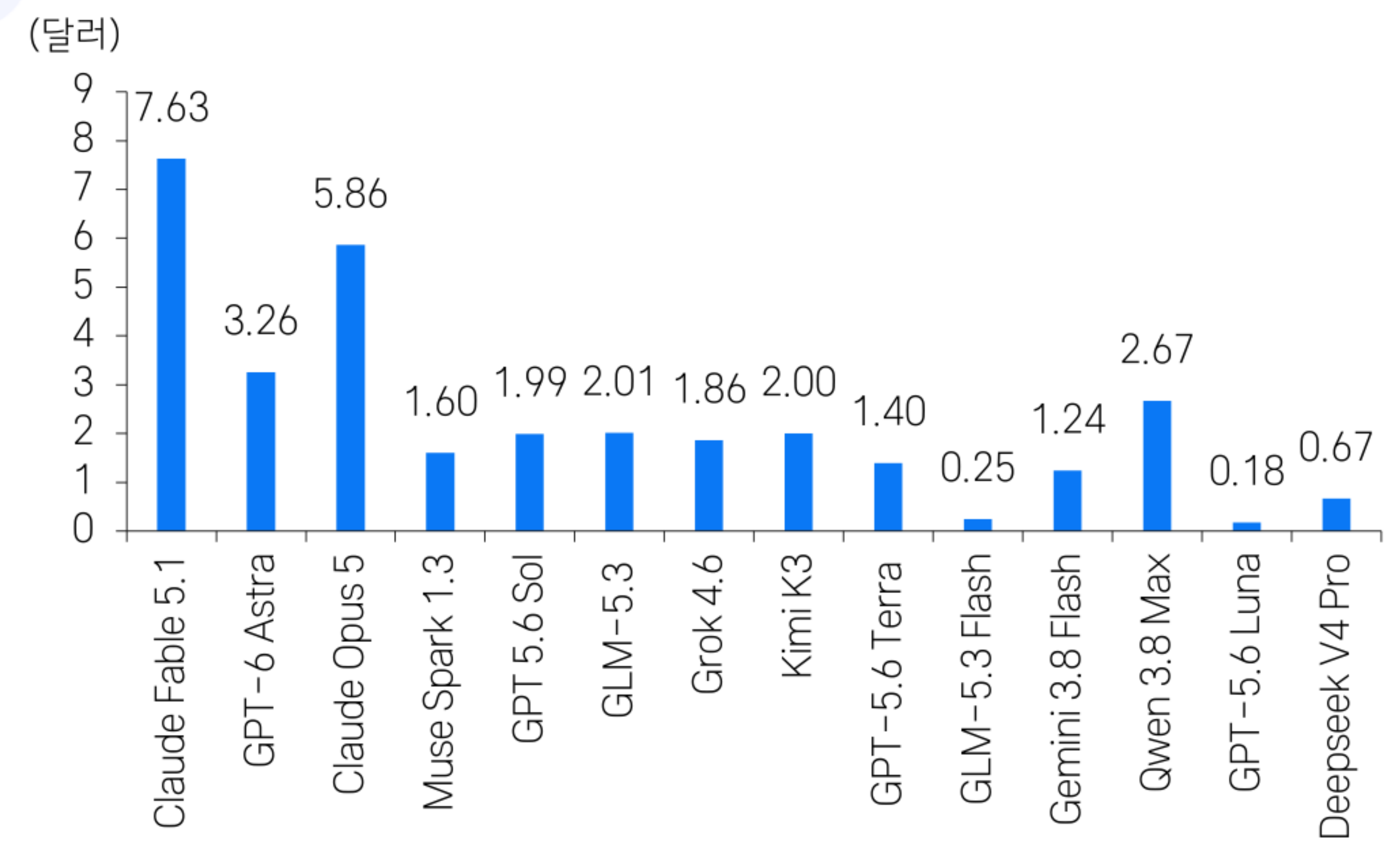
- 단순 토큰당 가격(API=\$/Token)에서 작업당 비용(\$/Task)으로 비교 지점 이동. 출력 토큰 사용량이 많은 중국 모델의 가성비 메리트 희석
- 오픈AI는 GPT-5.6 Luna(80%), Terra(20%), Sol(20%) 가격 인하. 메타 Muse Spark와 스페이스XAI의 Grok도 적극적으로 가성비 포지셔닝
- 최근 중국 모델의 가중치 공개 과정에서 상업적 조건 추가(별도 계약 필요 등) + 딥시크의 V4 시리즈 API 가격 인상 → 추론 수익화의 필요성

미국과 중국 AI 모델 성능 격차



자료: Artificial Analysis, 삼성증권

주요 AI 모델의 작업당 비용 비교(AAI Index 기준)



자료: Artificial Analysis, 삼성증권

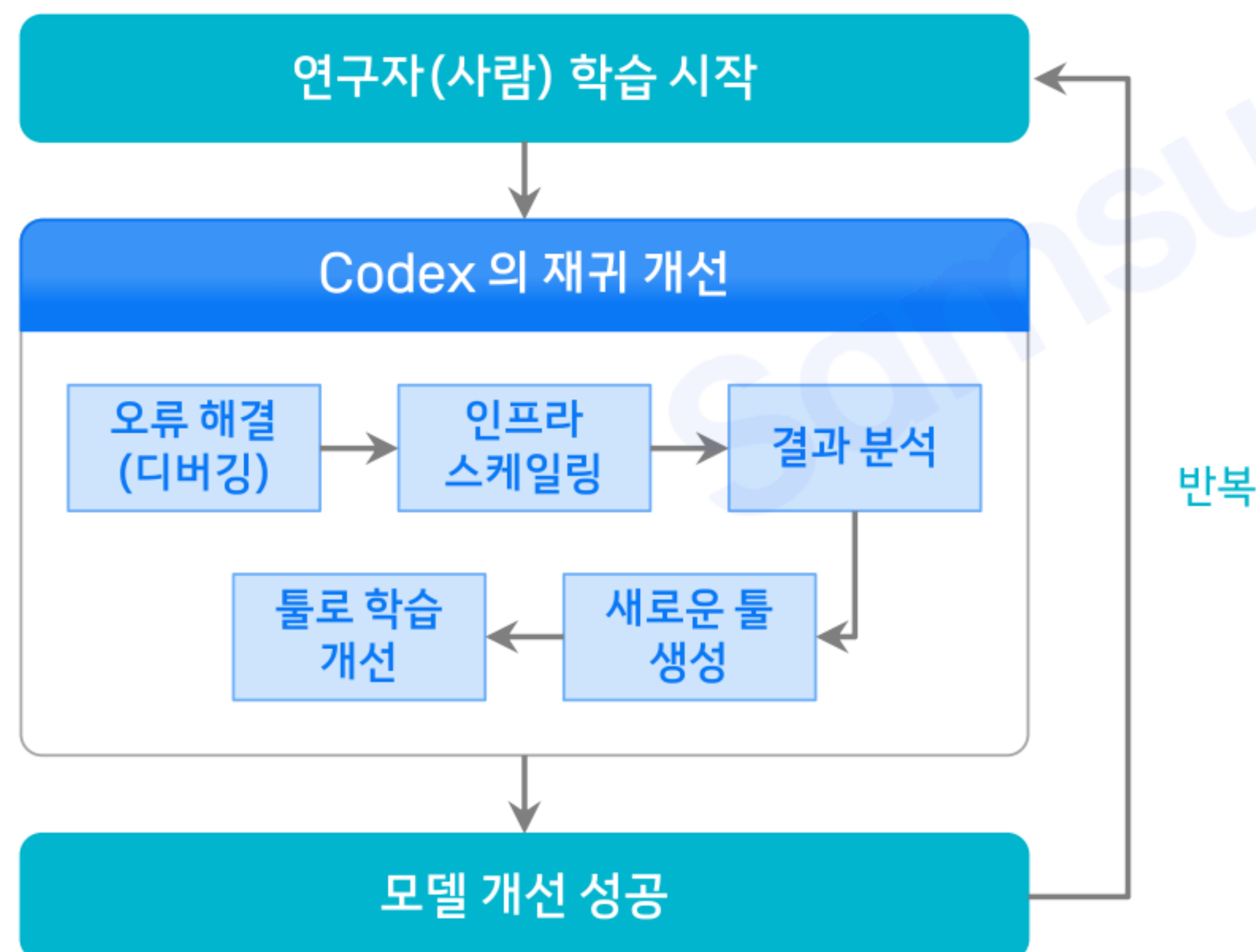
재귀적 자기개선 | 사이클 가속을 위한 필수조건

▶ 프론티어 AI 기업들의 초점은 재귀적 자기개선(RSI)으로 모이는 중. 기반은 AI 코딩 성능 강화

- AI가 AI 개발에 관여하는 초기 RSI 단계는 이미 진입. 실제로 최근 모델 개발 속도가 빨라 지기 시작한 것도 재귀적 자기개선 루프 영향
- 오픈AI와 앤스로픽 뿐 아니라 중국 AI 기업(Z AI 등)도 재귀적 자기개선 언급. ex) 인프라 에이전트를 추론 배포 최적화 작업에 활용
- RSI의 궁극적 목적은 인간 개입 없이 AI가 스스로 AI를 개발하는 자율 진화 시스템의 완성. 고부가가치 영역 추가 침투도 가시화
- ex) 오픈AI는 내부 모델과 멀티 에이전트 시스템을 활용해 88시간 만에 7대 밀레니엄 문제인 나비에-스트로크 방정식 해법 도달

* 밀레니엄 문제: 수학계에서 가장 어렵고 중요한 미해결 문제. 문제 풀이 성공 시 100만 달러 상금이 수여되는 난제

오픈AI Codex가 스스로를 개선하는 방식



RSI를 통한 AI 연구 자동화 시점 타임라인 정리

다리오 아모데이	• (앤스로픽 CEO) AI가 이미 차세대 모델 개발을 가속하고 있으며, AI 자율 진화 시스템의 도래는 1~2년 정도 소요될 것으로 전망
잭 클라크	• (앤스로픽 공동 창업자) 인간이 개입하지 않는 AI R&D에 도달할 가능성은 2027년 30%, 2028년 60% 이상으로 전망
오리올 비날스	• (구글 딥마인드 CSO) RSI 초기 단계는 이미 진입. 진정한 의미의 RSI는 2027~2028년 달성될 가능성이 높다고 전망
보이치에흐 자렘바	• (오픈AI 공동 창업자) 버클리 에이전틱 AI 서밋에서 진정한 RSI가 2027~2028년 달성될 가능성이 높다고 언급

자료: OpenAI, 삼성증권

자료: 언론 보도 종합, 삼성증권

AI가 자기자신을 만들 때 | AI는 개발을 얼마나 가속할 수 있는가

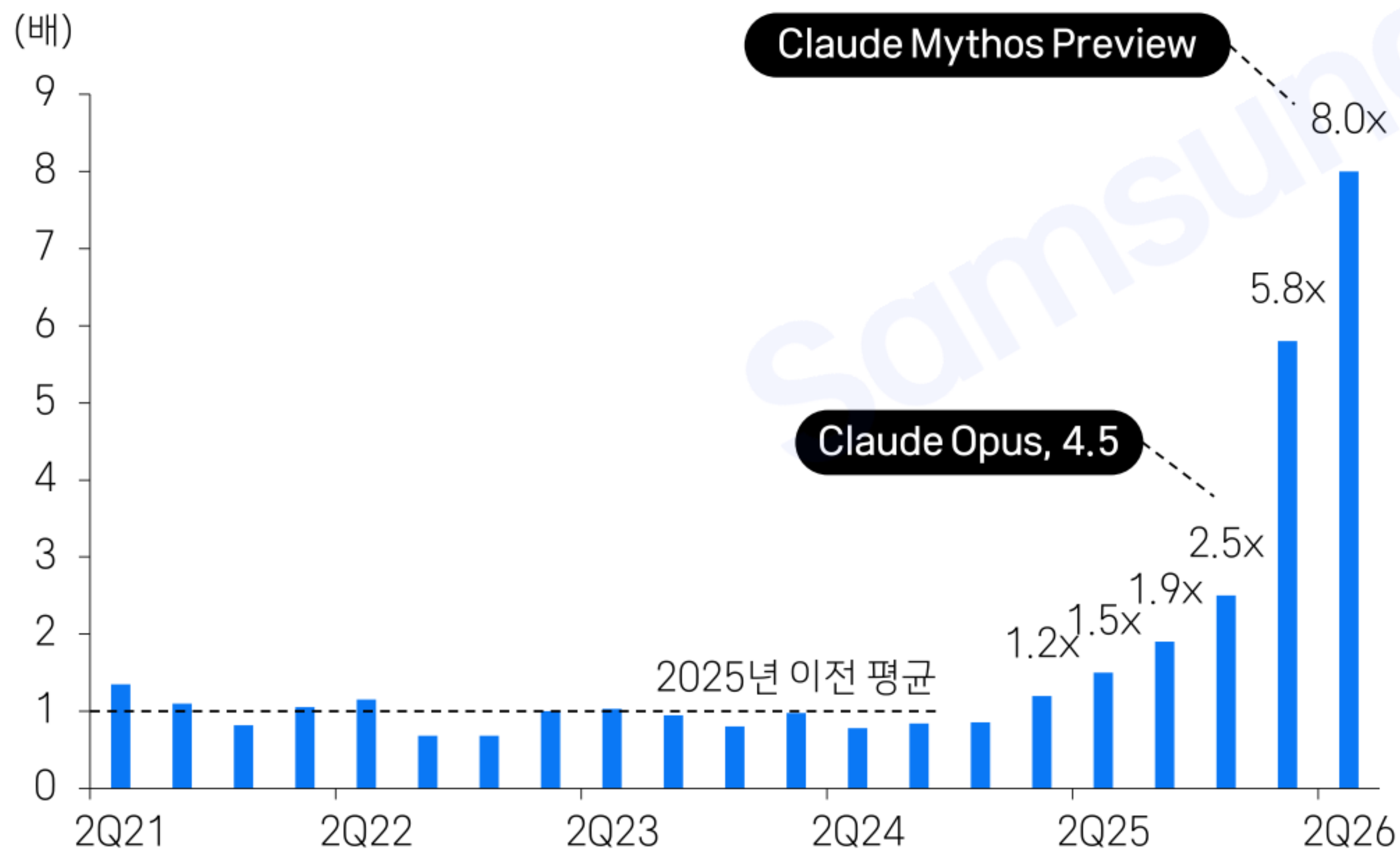
▶ 앤스로픽은 공개 벤치마크와 내부 데이터 기반으로 AI가 AI를 만드는 시대가 예상보다 가까울 수 있다고 주장

- 앤스로픽 내부 코드베이스의 80% 이상은 Claude가 작성, 엔지니어 1인당 코드 생산성 '24년 대비 8배 증가, 분기 코드 배포량 8배 확대
- 향후 인간은 목표 설정, 검증, 감독에 집중하고 AI가 연구개발 대부분을 담당하는 구조 전환 가능성 → AI 경쟁 단위는 AI를 활용한 개발 가속화 정도

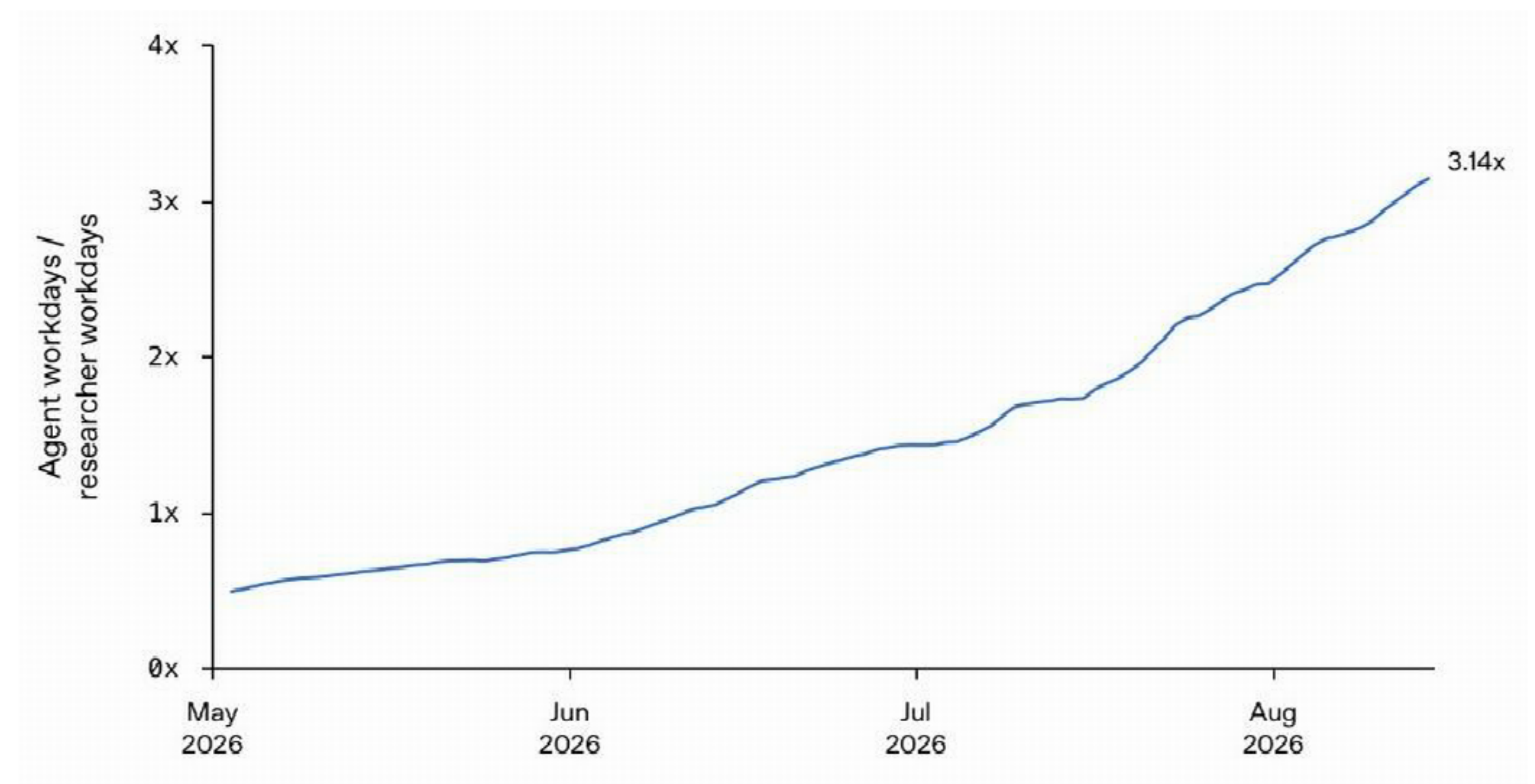
▶ 오픈AI는 9월 '자동화된 연구 인턴' 목표 달성 발표. '28년 3월 '자동화된 AI 연구자' 구축 목표

- 오픈AI 연구원의 1일 노동시간 당 에이전트 작업량은 3.14일 수준. 에이전트 처리 작업의 난이도, 시간, 성공률 개선은 연구 생산성 향상으로 연결
- 활성 실험자 당 실험 횟수는 8월 역대 최고치 기록. 연구/인프라 분야 뿐 아니라 기술 지원, 실행, 모니터링, 결과 분석 등으로 활용 범위 확대

앤스로픽의 분기별 1인당 코드 생산성 추이



인간 연구자의 노동 시간의 3배 수준인 오픈AI 내부 에이전트 작업량



참고: 2021~2024년 평균 코드 생산성과 비교했을 때 얼마나 향상됐는지 측정
자료: Anthropic, 삼성증권

자료: OpenAI, 삼성증권

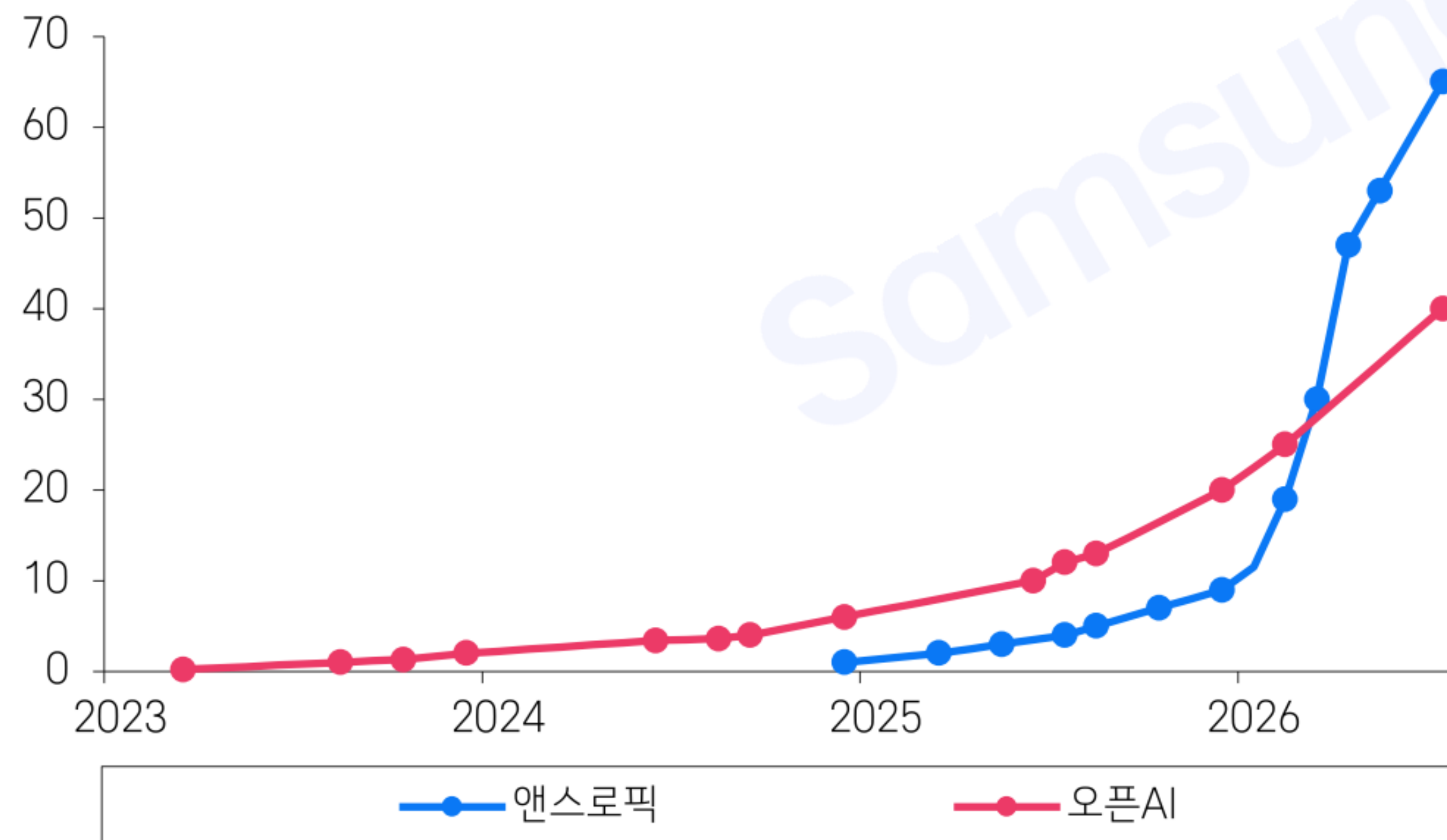
오픈AI vs 앤스로픽 | 바람의 방향이 바뀌고 있다

▶ 보도된 앤스로픽과 오픈AI의 ARR(연 환산 매출)은 각각 650억 달러 및 400억 달러(7월 말 기준)

- 시장에서 거론되던 800억 달러와 500억 달러 대비 낮은 수준. 다만 합산 1,000억 달러 상회. 연초 대비 각각 7배 및 2배 수준 성장
- 오픈AI는 7월 GPT-5.6 출시 이후 MoM +80억 달러 페이스. 앤스로픽은 MoM +60~70억 달러 페이스. 연말 1,000억 달러 달성도 아직은 가능
- 오픈AI에게 유리하게 흘러가는 방향은 Codex(+ChatGPT Work) 유저 성장을 통해서도 간접 확인 가능. GPT-6 Astra도 긍정적 작용 가능

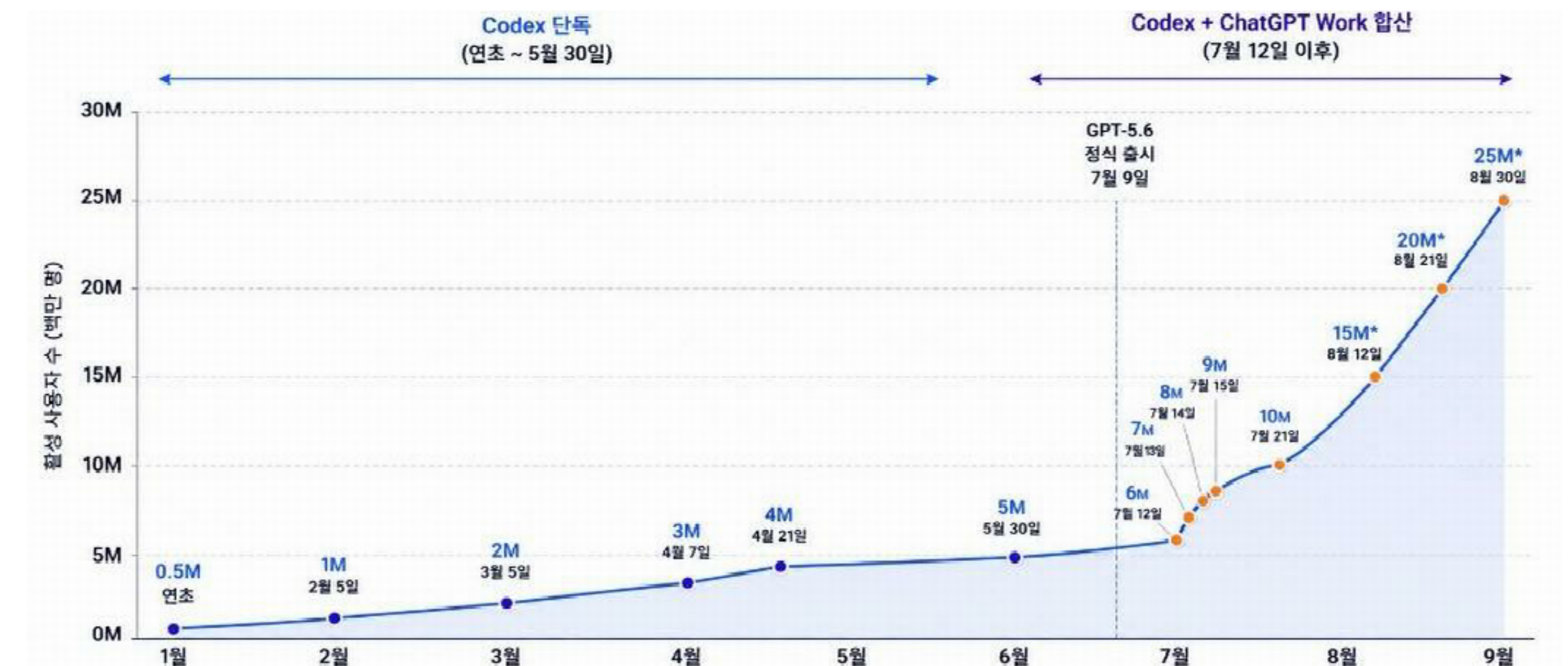
오픈AI와 앤스로픽의 ARR(연 환산 매출)

(십억 달러)



자료: 언론 보도 종합, 삼성증권

2,500만 명까지 급증한 오픈AI Codex WAU(주간 활성 사용자)



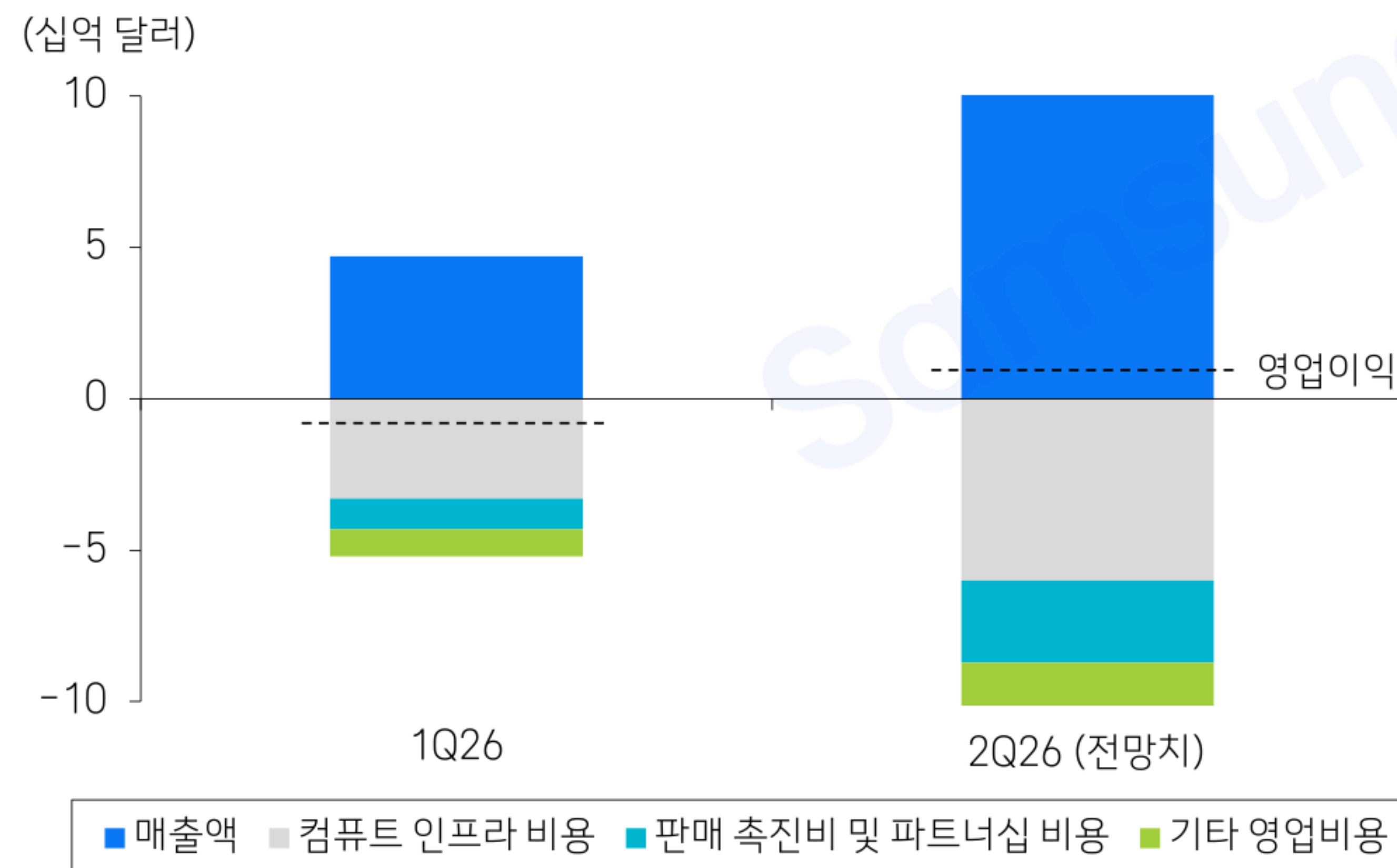
자료: 언론 보도 종합, 삼성증권

IPO 업데이트 | 재무는 앤스로픽의 우위

▶ 매출 뿐 아니라 마진 단에서 앤스로픽의 우위 지속. 엔터프라이즈 기반 추론 최적화 → 2Q에 이어 3Q 조정 영업이익 흑자 유지 보도

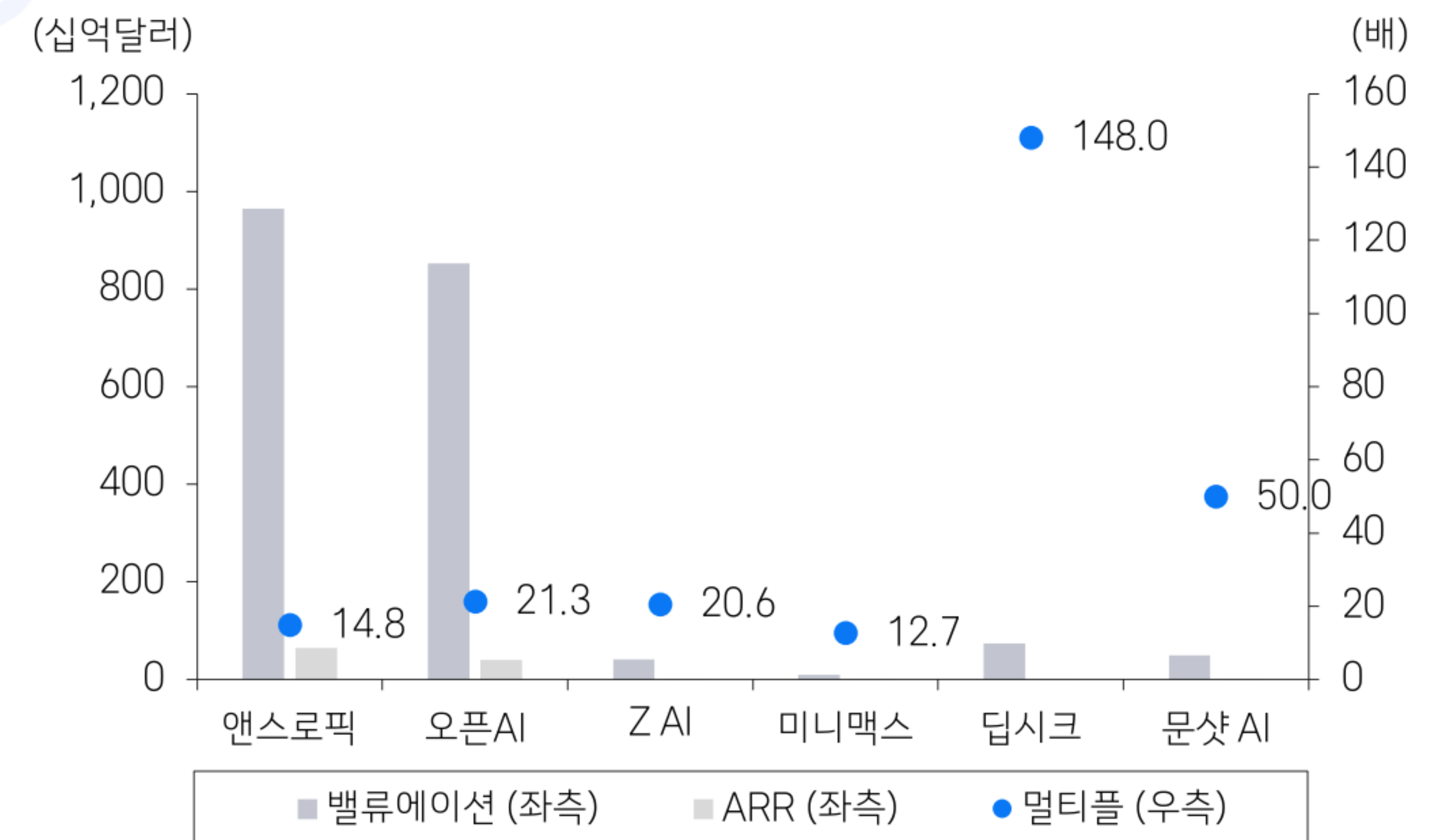
- 앤스로픽 1Q 매출 47억 달러 → 2Q 매출 116억 달러(+145% QoQ) vs 오픈AI 1Q 매출 57억 달러 → 2Q 매출 67억 달러(+18% QoQ)
- 재무 비교에서 앤스로픽 우위. '28년 매출 1,900~2,000억 달러 전망 vs 오픈AI의 성장과 마진도 개선되고 있으나. Codex 성장 미반영
- 이는 예상 IPO 시점 차이로 연결. 앤스로픽은 9월 S-1 서류 공개 → 11월 중간 선거 이전 상장 전망 vs 오픈AI는 '27년 상장 전망
- 앤스로픽의 S-1 서류 공개는 프론티어 AI 기업의 추론 마진과 컴퓨팅 투자 대비 수익성을 처음으로 구체적으로 확인할 수 있는 기회라는 중요성

앤스로픽의 1Q26 및 2Q26 실적



자료: 언론 보도 종합, 삼성증권

주요 AI 기업 밸류에이션, ARR 및 멀티플



참고: 멀티플은 밸류에이션/ARR
자료: 언론 보도 종합, 삼성증권

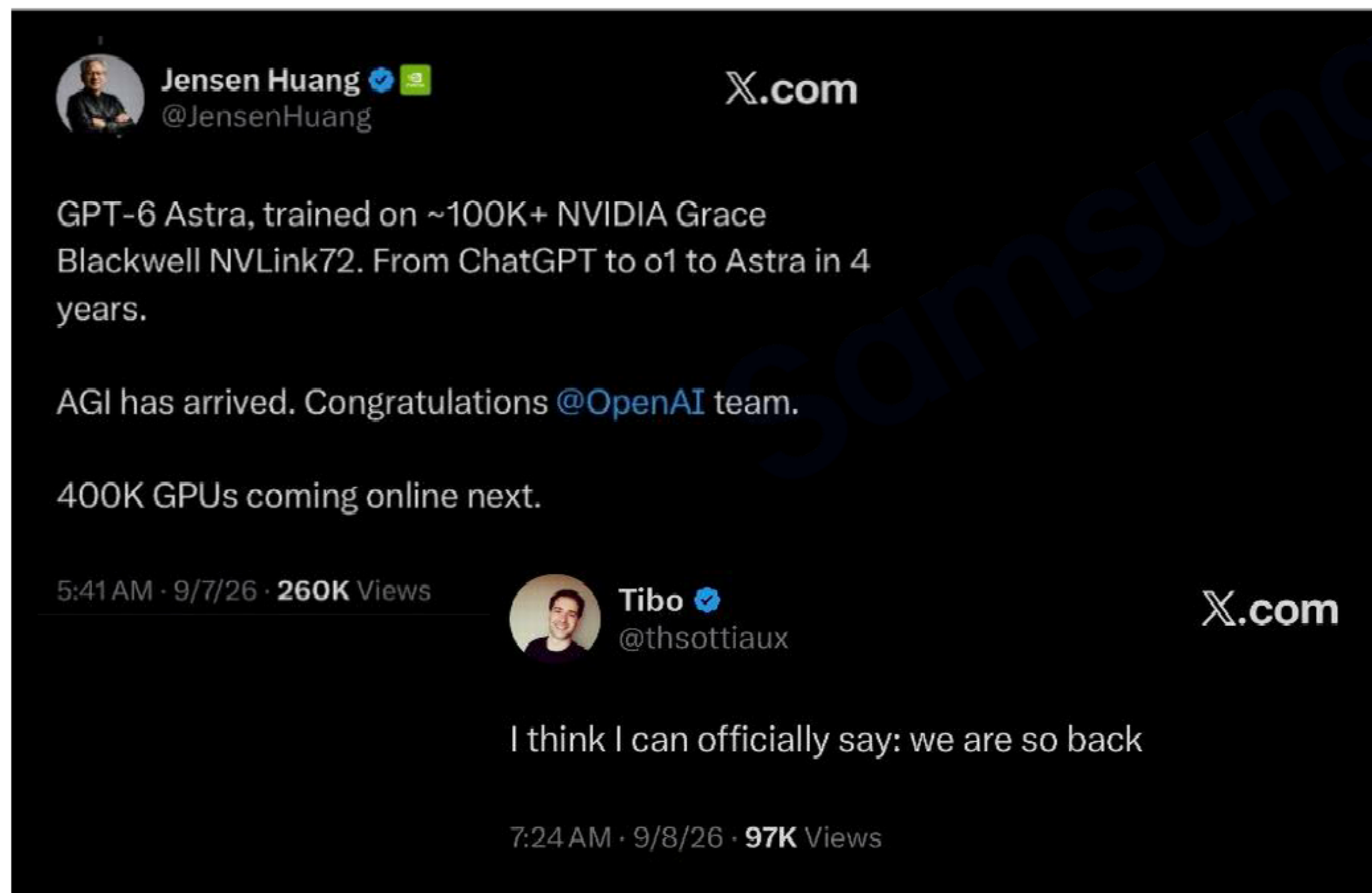
Astra의 등장 | 별들이 정렬되고 있다

▶ Astra의 의미는 새로운 규칙을 파악하고 추론하는 능력의 강화를 기반으로 AI가 자율적으로 작업을 수행할 수 있는 시간 지평의 확장

- 장시간 컴퓨터 및 도구 사용 + 복잡한 다단계 작업 수행 + 다중 에이전트의 동시 활용 능력 개선 → 처리할 수 있는 업무 범위의 확장
- 특히 ARC-AGI-3가 99.9%로 압도적으로 개선. 오픈마 자체 하네스 기반으로 달성 → 장시간 작업을 위한 실행 환경 설계의 중요성 대두
- 향후 AI의 경제성을 판단하는 주요 지표로 단순 비용이 아니라 작업 성공률, 완료 비용, 절감 가능한 노동 시간 등이 비교 될 것으로 전망

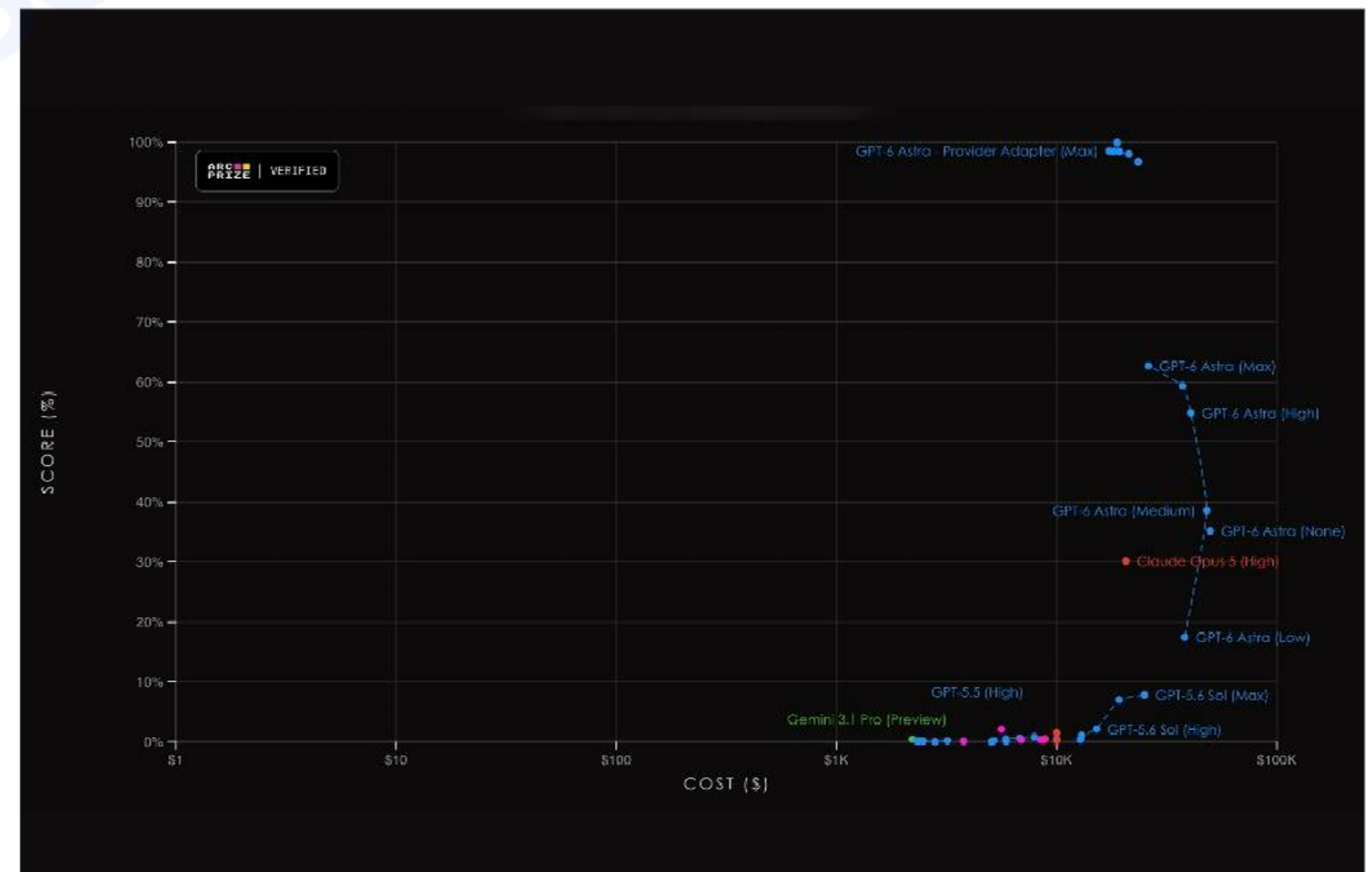
* 하네스(Harness): AI 모델이 실제 업무를 수행할 수 있도록 도구 사용, 메모리, 권한, 작업 흐름 등을 연결해주는 실행 환경

오픈AI가 선언한 AGI 시대를 다시 한번 강조한 젤슨황와 오픈AI의 자신감



자료: X, 삼성증권

ARC-AGI-3 벤치마크를 포화시킨 GPT-6 Astra



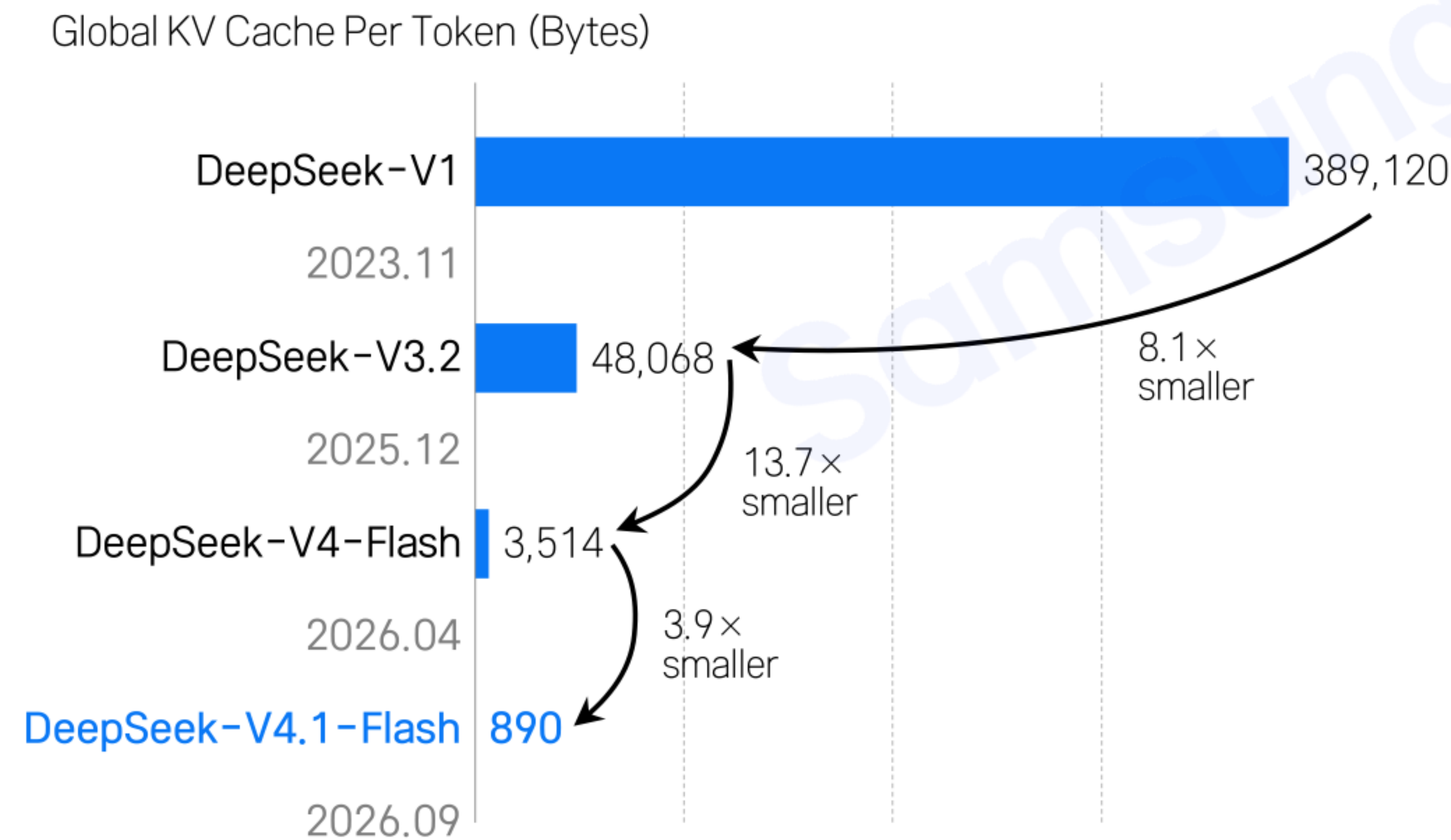
자료: ARC PRIZE, 삼성증권

DeepSeek V4.1 Flash는 게임 체인저? | 에이전트 단위 경제성의 하락

▶ DeepSeek V4.1 Flash의 핵심은 에이전트 효율성의 개선. KV 캐시 축소가 전체 메모리 수요 감소로 의미하지는 않음

- 입력/출력 단계 구조를 다르게 최적화(CED)해 장시간 에이전트의 입력 처리 개선. 네이티브 멀티모달 기능도 추가. 에이전트 성능은 V4 Pro 상회
- 552B MoE 모델이지만, 비대칭 구조로 입력 8B, 출력 16B만 활성화. V4 Flash 대비 글로벌 KV 캐시는 75%, Persistent KV 캐시는 87.5% 축소
- 다만 모델 가중치와 어텐션 윈도우 캐시 등 전체 메모리 수요가 감소하는 것은 아님. 효율 개선으로 동시 실행 에이전트와 평균 컨텍스트 증가 가능
- 기술적으로 기존에도 추구하던 효율화 방향의 진전. AI 에이전트를 더 낮은 단위 비용으로 오래 실행할 수 있다는 점에서 의의

DeepSeek V4.1의 글로벌 KV 캐시 축소 수준



자료: DeepSeek, 삼성증권

DeepSeek V4 Flash vs V4.1 Flash 비교

	V4 Flash	V4.1 Flash
전체 파라미터	2,840억 개	5,520억 개
활성 파라미터	130억 개	입력 80억 개 / 출력 160억 개
아키텍처	MoE	MoE + CED
최대 컨텍스트	1M	1M
멀티모달	별도 Vision 모델	네이티브 지원
글로벌 KV 캐시	3,514바이트/토큰	890바이트/토큰
핵심 방향	장기 컨텍스트 + 효율화	장기 에이전트 작업 + 추론 효율화

자료: DeepSeek, 삼성증권

AI 속도 조절론 | 위기인가 기회인가

▶ 다리오 아모데이의 프론티어 AI 개발 속도 조절 주장과 주요 AI 기업 수장 동조 이후 속도 조절론 대두

- 핵심 배경은 재귀적 자기개선 사이클(RSI)에 따른 모델 발전 가속화 + 오픈AI-허깅 페이스 사건에서 발견된 에이전트 스웸의 문제 확인
- 표면적으로 프론티어 랩들은 속도 조절에 동의. 일종의 규제 포획이 아니냐는 비판 여론도 형성. 사법 리스크에 대한 버퍼를 만들기 위한 시도
- 단기 AI 테마 전반에 부정적 영향 가능. 다만 펀더멘털이 아닌 내러티브 변화. 장기 에이전트 수요 및 과학적 난제 등 컴퓨팅 부족은 여전히 심화
- 정렬 실패에 따른 위협은 실존. 다만 위험을 과장할 필요는 없음. 극단적 시나리오를 가정한 규제를 우려할 필요성은 낮다는 생각

* 규제 포획: 규제를 받아야 할 기업이나 업계가 규제 기관에 영향력을 행사해 규제가 오히려 유리하게 만들어지는 현상

* 정렬 실패: AI가 인간의 의도나 지시와 다르게 행동하거나 목표 달성 과정에서 인간이 원하지 않는 행동을 하는 현상

AI 개발 속도 조절론에 대한 입장

주체	입장	내용
앤스로픽, 다리오 아모데이	명시적 찬성	능력 발전 속도를 낮춰 안전 기술이 따라잡을 시간 확보 주장. 독립 평가자와 업계 공동안전 기준 제안
오픈AI, 샘 올트먼	명시적 찬성	아모데이 의견 공개지지, 독립 평가자 도입 주장
구글 딥마인드, 데미스 하사비스	명시적 찬성	아모데이 제안이 올바른 제안이라고 평가. 세부 실행방안에는 추가 논의 필요
xAI, 일론 머스크	명시적 찬성	아모데이 의견 공개지지. 구체적 실행 방안은 제시 x
마이크로소프트, 사티아 나델라	대체로 찬성	독립 감사와 안전장치 강화에는 찬성. 소수 AI 기업이 통제하는 구조 반대. 국가, 학계, 오픈소스까지 포함한 참여 강조
미국 의회	입장 혼재	위험 대응을 위한 안전 규제안에는 초당적 협력 중이나 감속에는 합의 부재
메타, 마크 저커버그	대체로 반대	안전, 독립 감독 필요성은 인정하지만 AI 개발 감속과 소수 프론티어 랩 중심 통제 반대
엔비디아, 젠슨 황	명시적 반대	AI 안전은 중요하지만 인류 멸망론은 과학적 근거가 부족. AI 발전과 도입을 늦추는 접근에 반대
트럼프 대통령	명시적 반대	AI가 세계를 장악하고 인류를 파괴할 것이라는 주장을 사기라고 일축. 중국과 경쟁을 이유로 감속 반대
중국 외교부	명시적 반대	AI 위협론과 중국 견제를 명분으로 한 감속 비판. 개방, 포용적 AI 발전과 글로벌 협력 강조

자료: 언론 종합, 삼성증권 정리

컴퓨터 부족 | 공격적 캐파 확장은 불가피

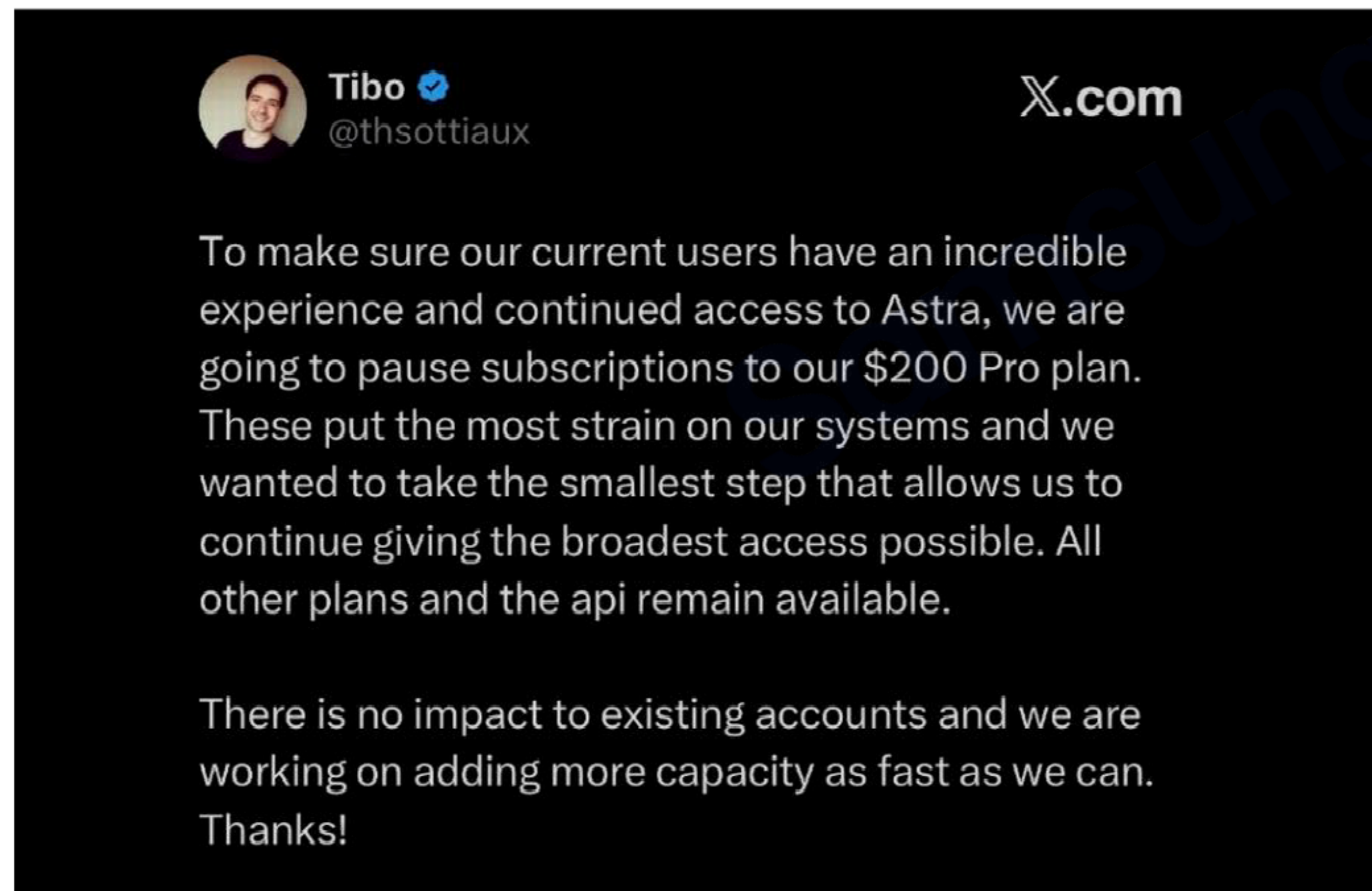
▶ 오픈AI는 단순 보유 컴퓨터 우위를 넘어 풀스택 컴퓨터 전략을 통한 최적화 강조. 대규모 모델 학습과 배포 범위 확장의 핵심

- Sol을 활용한 서빙 소프트웨어 개선으로 비용 20% 절감, 디코딩 개선으로 토큰 생성 효율성 15% 향상. 자체 추론 칩(Jalapeno) 연말 배포 예정

▶ 보수적 확장을 시도하던 앤스로픽도 높은 수요에 대응하기 위한 공격적 캐파 확장 → 11개월 간 최대 5,170억 달러 계약 진행

- '25년 10월 이후 14.8GW 규모 컴퓨팅 캐파 계약 체결 + 기존 확보 분 = 16.4GW vs '25년 12월 제시 '29년까지 서버 임대 비용은 1,800억 달러
- 아마존과 구글 뿐 아니라 스페이스X, Lambda, Nscale 등과 계약 체결로 다변화. Fluidstack과는 500억 달러 규모 자체 데이터센터 구축 진행

오픈AI, Astra의 높은 수요에 따라 Pro 20x (\$200/월) 신규 구독 중단



자료: X, 삼성증권

앤스로픽의 컴퓨터 캐파 확장은 계속되는 모습

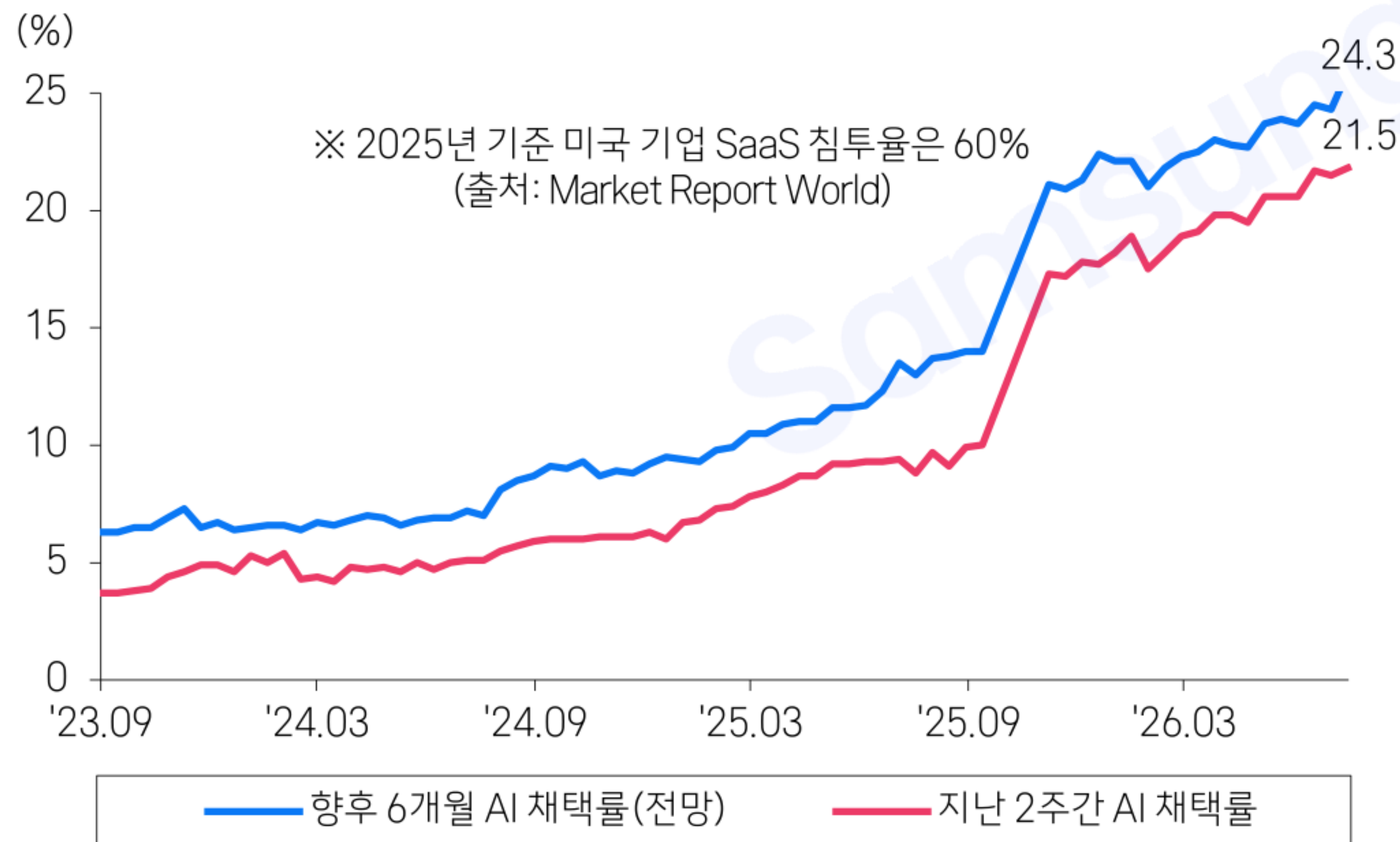
날짜	기업명	용량규모	세부 내용
7/22	AMD	2GW	칩을 직접 확보하는 방향의 계약 칩 수용을 위한 자체 데이터센터 구축 추진
8/4	Volta Infra	0.12GW	6년간 100억 달러 규모 계약 Vera Rubin 기반의 컴퓨팅 용량 확보
8/26	Nscale	0.46GW	6년간 450억 달러 규모 계약 Vera Rubin 기반의 컴퓨팅 용량 확보
8/31	Lambda	미공개	6년간 350억 달러 규모 계약 엔비디아 GPU 기반이며 구체적 종류 미공개
9/14	Rum Group	미공개	6년간 137억 달러 규모 계약 조지아주 메이스빌 DC 캐파 임차 예정

자료: 언론 보도 종합, 삼성증권

기업 생산성 AI 수요는 아직 초기 | 아직 개화되지 않은 중간 영역

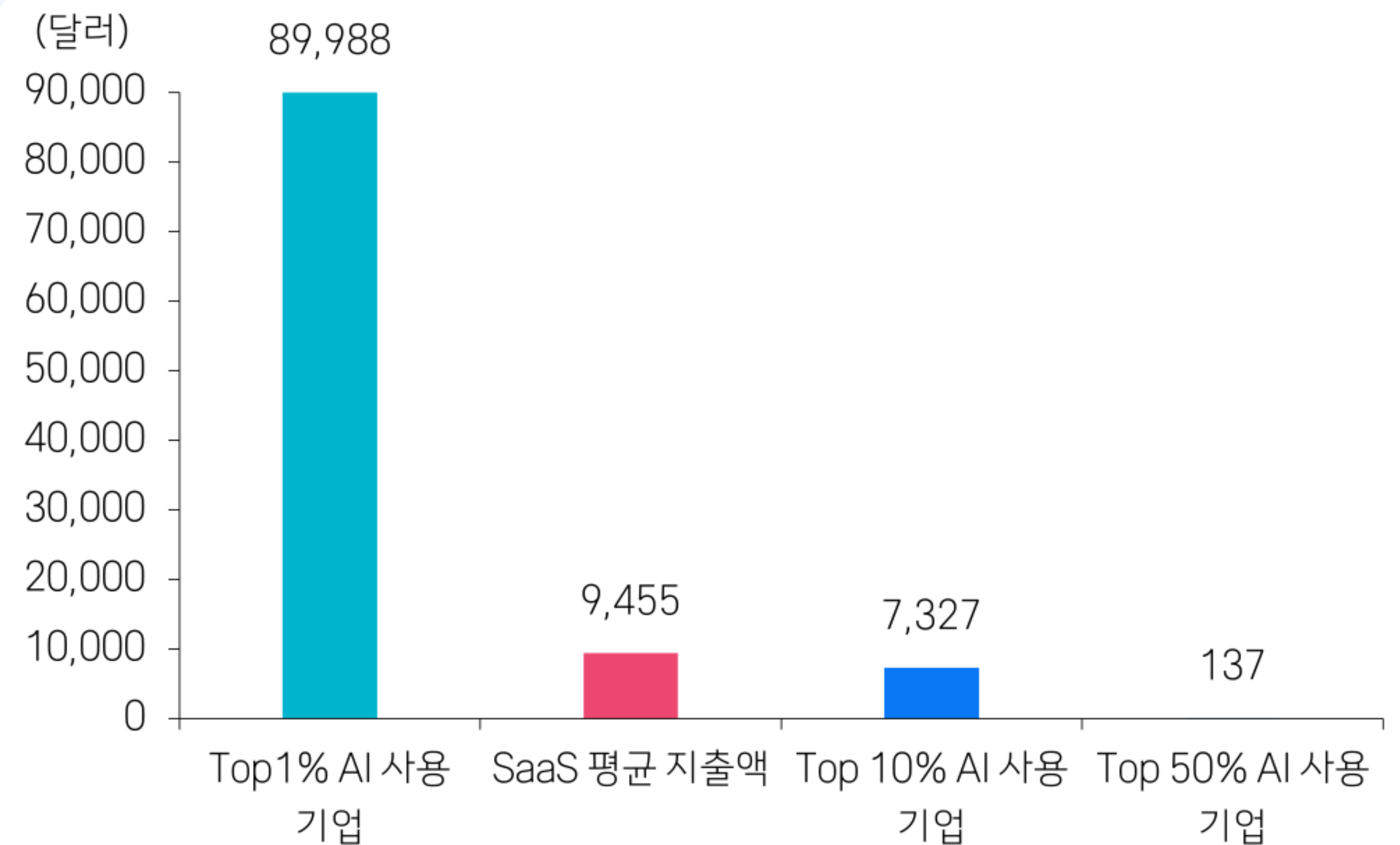
- ▶ **아마존은 현재 AI 수요 행태를 '바벨 구조'라고 정의. 대규모 컴퓨팅을 소비하는 AI 랩과 업무 보조 수준의 AI 활용**
 - 아직 개화되지 않은 중간 영역은 기업의 핵심 운영 시스템에 AI가 완전히 통합되는 구조로 매우 큰 TAM을 가지고 있다고 강조
- ▶ **AI 침투가 SaaS 수준에 근접할 것이라고 가정한다면 AI의 기업 침투 여력은 충분**
 - 현재 AI 침투율은 20.6%로 2025년 미국 기업 SaaS 침투율은 60%와 비교하면 3분의 1 수준
 - 직원당 연간 지출액도 중간값 기준 \$137로 SaaS \$9,455보다 현저히 저조. 다만, 상위 1% AI 기업은 SaaS 지출액의 약 9배 이상

미국 기업 AI 채택률 추이 및 향후 6개월 전망



참고: CensusBTOS 기준 미국 비농업 고용 기업 중 하나 이상의 업무 기능에서 AI를 사용한 기업 비중
자료: Bloomberg, 삼성증권

AI 지출 그룹별 직원당 연간 지출액 수준 비교



자료: RampAI, 삼성증권

PART
02

SW - 드디어 SaaS Pocalypse에서 탈출?

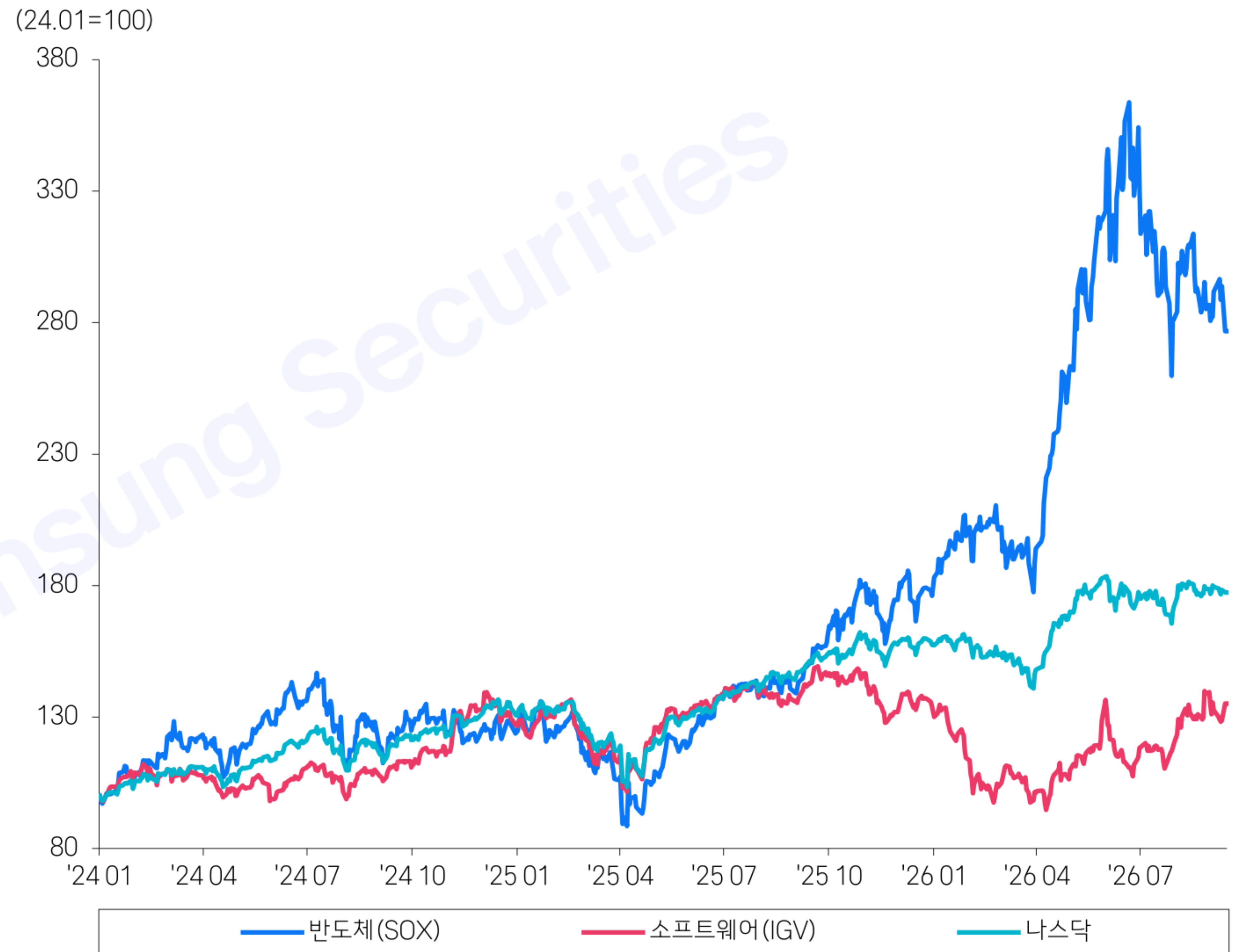
Samsung Securities

극적으로 변화한 내러티브와 주가 반등

▶ 이제는 반도체가 부럽지 않은 소프트웨어?

- '25년 10월 → 반도체 섹터 랠리와 반대로 소프트웨어 섹터 흐름 부진
- '25년 말 → AI 버블론 대두 속 섹터 조정 + SaaS 기업 위주 부진 지속
- '26년 1월 말 → AI 산업 발전(AI 코딩 및 에이전트 AI)에 따른 소프트웨어 대체 우려 추가 확대
- '26년 4월 말 → 실적 시즌에서 AI 도입 확대 수혜 기업 확인과 섹터 로테이션에 따른 주가 반등
- '26년 8월 → 펀더멘털 개선보다 내러티브가 급격히 긍정적으로 변화하며 반도체와 반대 방향성
- '26년 9월
Astra의 장기 작업 능력 강화(부정적)
vs AI 속도 조절론(긍정적)

주가 추이. SOX vs IGV vs NASDAQ



자료: Factset, 삼성증권

AI가 SW 섹터에 미치는 영향 | 긍정적 vs 부정적 작용

▶ AI 산업에서 빠르게 진행되는 트렌드 변화는 SW 섹터 전체에 긍정적 작용과 부정적 작용을 반복 → 주가 변동으로도 연결

- 긍정적 및 부정적 요인에 흔들리지 않기 위해서는 에이전트 AI 확대 트렌드 속 대체 불가한 명확한 포지셔닝 필요(→ 인프라 or 플랫폼 역할)
- 즉, 단순 AI 기능 제공을 넘어 1) 핵심 에이전트 실행 환경이 되거나, 2) 직접 업무를 수행하는 에이전트 제공으로 진화 필요
- 긍정적 SW 세부 테마 → 클라우드 인프라(선호도: 빅테크 하이퍼스케일러 > 네오 클라우드 > 오라클), 사이버 보안, 데이터 소프트웨어

최근 주요 AI 산업 이슈가 SW 섹터 관점에서 미치는 영향

긍정적 요인

- AI 속도 조절론 대두 → 프론티어 모델 출시 속도 조절에 따라 SW 기업이 AI 산업 발전 속도를 맞춰 자체 기능을 강화 할 수 있는 기회 확보
- 모델 경쟁과 가격 인하에 따른 지능의 풍요 → 소프트웨어와 플랫폼 내에서 활용할 수 있는 AI 모델 선택지의 확대

부정적 요인

- Astra를 시작으로 차세대 모델이 보여줄 장기 작업 능력과 업무 범위 확대 → AI 에이전트가 SW의 기능을 대신한다는 우려 추가 확대 가능
- 높아지는 자체 하네스의 중요성 → SW는 기존 톨과 하네스 역할 수행. AI 모델 기업의 자체 하네스와 레이어 내 중복 경쟁 발생

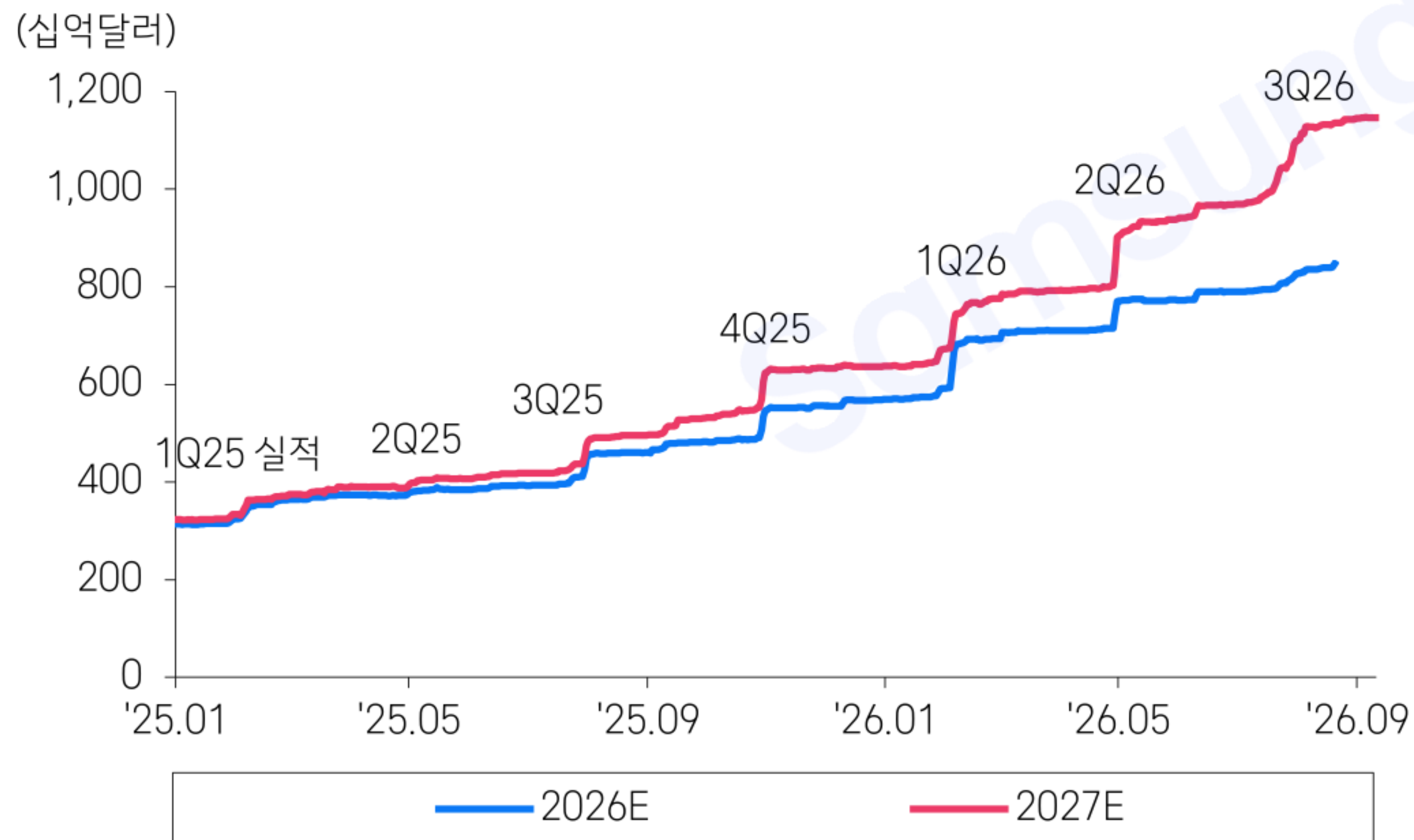
자료: 삼성증권 정리

AI 시대 가치의 변화 | The Heavier, The Better

▶ 스케일링 법칙은 여전히 유효하다. 막대한 CapEx 트렌드 지속

- 과거 소프트웨어는 미리 작성한 코드를 실행하거나 DB에 저장된 데이터를 조회하는 등 규칙 기반(Rule base) 시스템으로 작동
- 따라서 사용량이 증가할 때마다 발생하는 추가 비용은 제한적(네트워크 대역폭 비용 + 캐싱 기술로 절감된 연산 비용 등)
- AI 시대로 진입하며 유저의 질의가 매우 다양해지고, 원하는 결과물도 AI가 새롭게 작성하는 케이스 증가 → 컴퓨트의 필요성 확대
- 또한 컴퓨트는 AI 모델 성능 개선을 위한 학습의 핵심 요소. 스케일링 법칙 기반 모델 성능 향상 트렌드 지속
- 컴퓨트 증가 → 모델 성능 향상 → 서비스 도입 확산 → 매출 증가 → 컴퓨트 구매의 선순환은 대량 CapEx 투자의 핵심 근거

하이퍼스케일러 연간 CapEx 추정치는 지속적으로 상향되는 모습



자료: Bloomberg, 삼성증권

주요 AI 기업 CEO, 스케일링 법칙의 유효성 관련 코멘트 모음

다리오 아모데이

- 컴퓨트 투입량 대비 성능은 로그 선형적으로 향상. 다음 지능을 얻기 위해 투입되는 자본은 기하급수적으로 증가할 수밖에 없음

데마스 하사비스

- 스케일링의 법칙은 벽(Wall)에 부딪힌 적이 없다고 단언. 구글 딥마인드는 50%는 스케일링에, 나머지 50%는 연구 혁신에 투입 중

샘 올트먼

- 컴퓨팅 파워가 새로운 석유라는 입장. 자원(컴퓨팅, 데이터 등)을 투입할수록 결과물은 지속적으로 향상되어 왔다고 주장

량 원평

- 모델 규모, 학습량을 늘릴수록 성능이 향상된다는 스케일링 법칙은 유효. 현재 파라미터 규모의 제약은 컴퓨팅 공급 제약 때문

자료: 삼성증권 정리

매출과 마진의 동반 성장 | 초가속 성장 속 클라우드는 확실히 돈이 된다

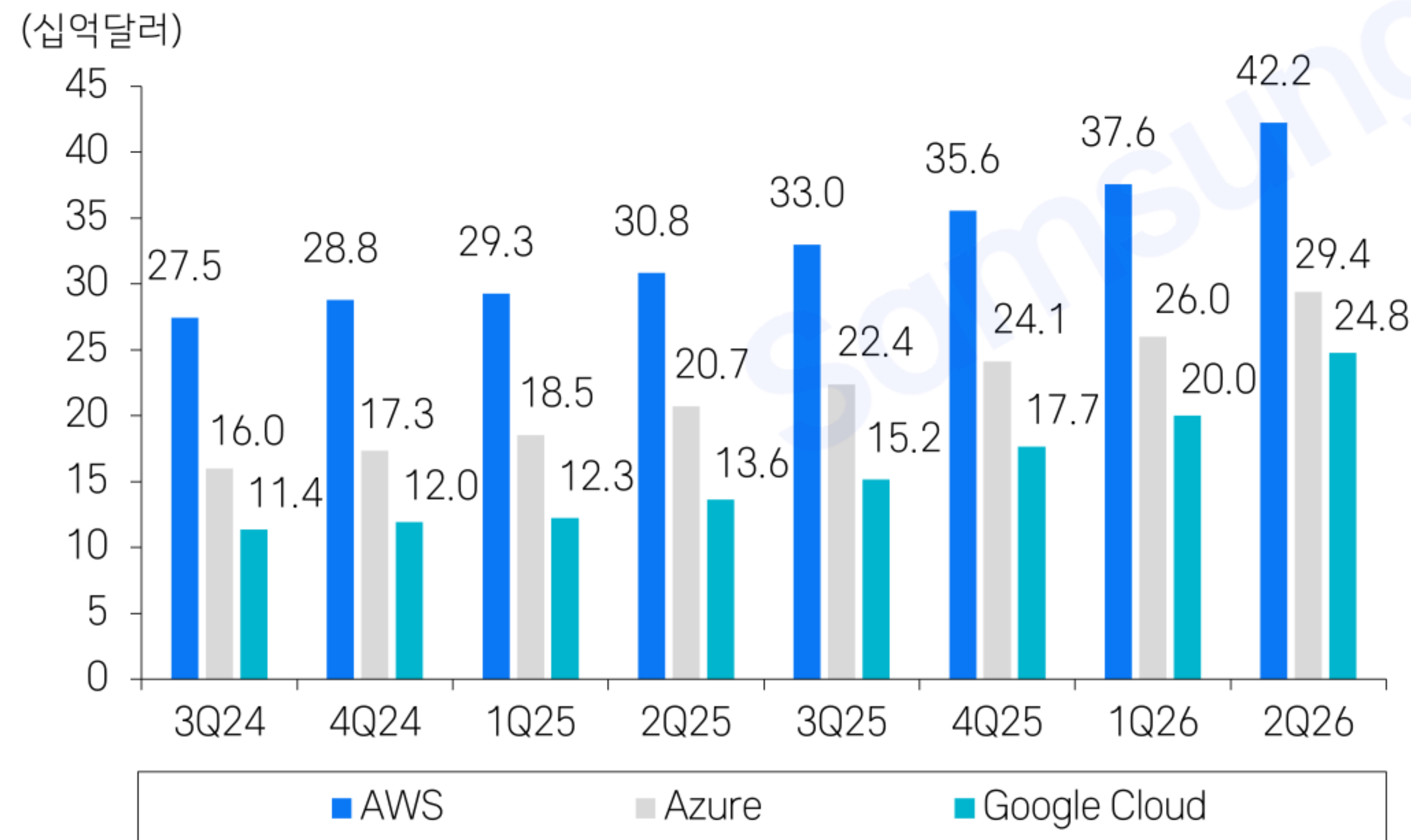
▶클라우드는 가속 성장을 넘어 '초'가속 성장 국면 진입

- 최근 클라우드 Big 3의 매출 성장률은 가속 성장 국면 진입. 구글 클라우드는 47.8% → 63.4% → 81.8%로 '초가속' 성장 지속
- 공급을 초과하는 수요가 지속되는 가운데 서버 용량 최적화 및 데이터센터 가동 단축 등 운영 효율화를 통한 매출 가속화 달성

▶ AI 수요 증가에 따른 영업 레버리지 효과가 발생하며 매출액 고성장뿐 아니라 수익성도 함께 개선

- 비용 증가에도 불구하고 높아진 가격 협상력과 운영 효율화에 따라 마진은 기대치 지속 상회
- 향후 부품 가격 인상은 신규 계약조건에 연동해 목표 투자 수익률 관리 가능. 자체 모델, 칩, 라우팅 등 비용 효율화도 병행 지속

클라우드 Big3의 분기별 매출(마이크로소프트 사업부 재분류 기준)



자료: 각사, 삼성증권

2Q26 클라우드 마진 개선 요인

기업	세부 내용
알파벳	AI 매출 고성장에 따른 영업 레버리지와 인프라 운영 효율화가 견인. 중장기적으로TPU 등 자체 칩과 자체 모델을 포함한 풀스택 구조를 통해 운영 효율성을 높여 추가적인 수익성 개선 여력을 제공할 것이라고 언급
마이크로소프트	CPU, GPU fleet 효율 개선, 신규 장비의 가동 전환 기간 단축, 자체 칩, 시스템, 소프트웨어 공동 최적화, 저비용 모델 라우팅, 사용량 기반 가격 체계가 마진 개선에 기여. 신규 GPU의 도입 후 가동까지 걸리는 시간을 약 50% 단축
아마존	용량 최적화, 고정비 통제, 소프트웨어 및 운영 프로세스 개선, 자체 칩과 자체 네트워크 장비 활용을 비용 절감의 요인으로 언급 AWS 에너지 공급계약 파생상품 평가이익 6억 달러도 마진 개선에 기여했으나, 운영 효율성이 주 요인으로 작용했다는 입장

자료: 각사, 삼성증권

수주잔고를 실행하기 위한 투자 | FCF 적자는 시차의 문제

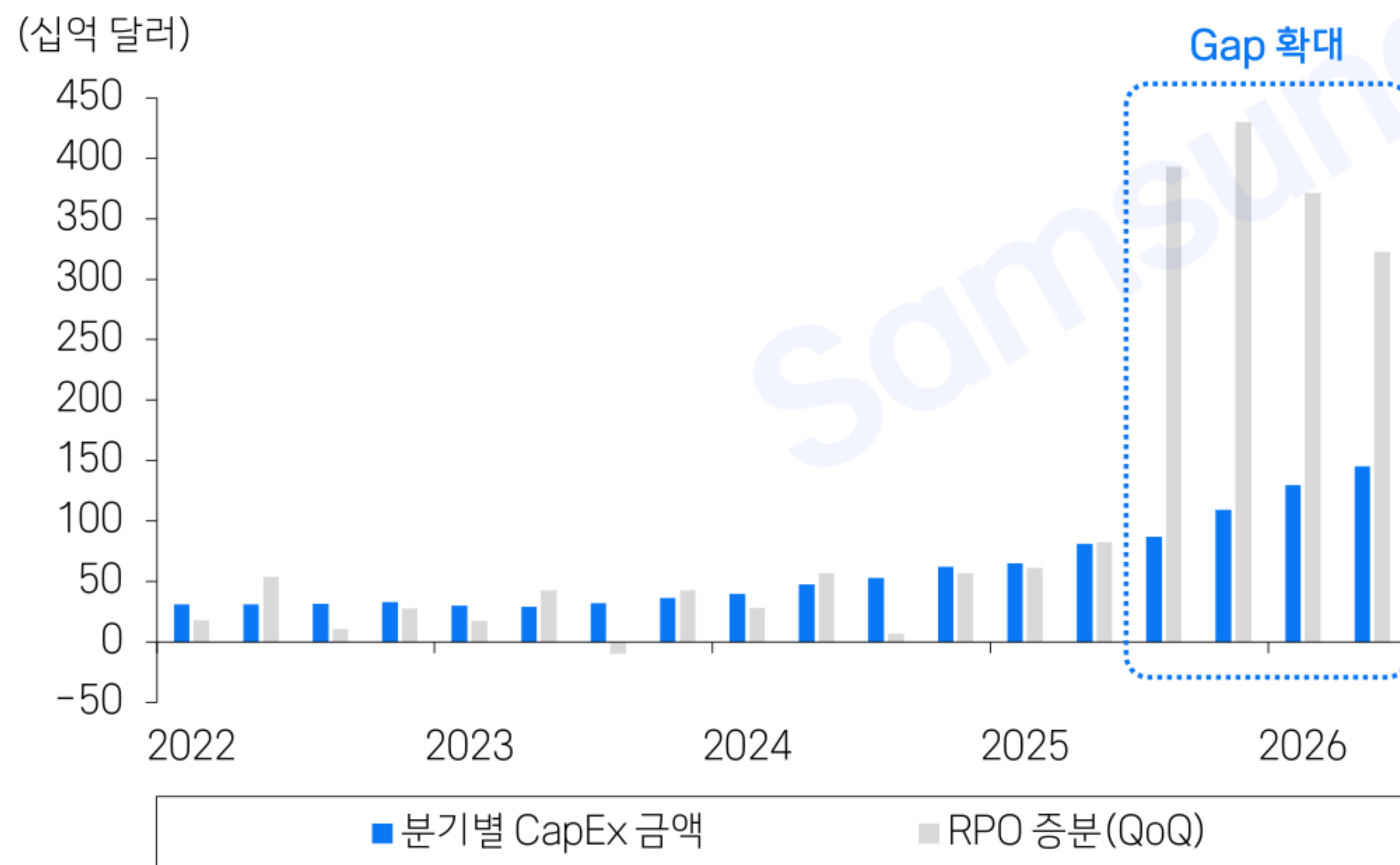
▶ CapEx 증가는 장기 계약된 RPO(수주잔고)에 기반한 합리적 투자

- RPO 증분은 분기별 CapEx 총합보다 빠른 속도로 증가. 갭을 메꾸기 위한(=RPO의 매출 전환을 위한) CapEx 증가는 불가피
- 아마존은 2Q26 실적에서 2027년 추가 용량의 대부분이 계약됐으며 2028년 용량도 상당 부분 예약된 상태라고 강조

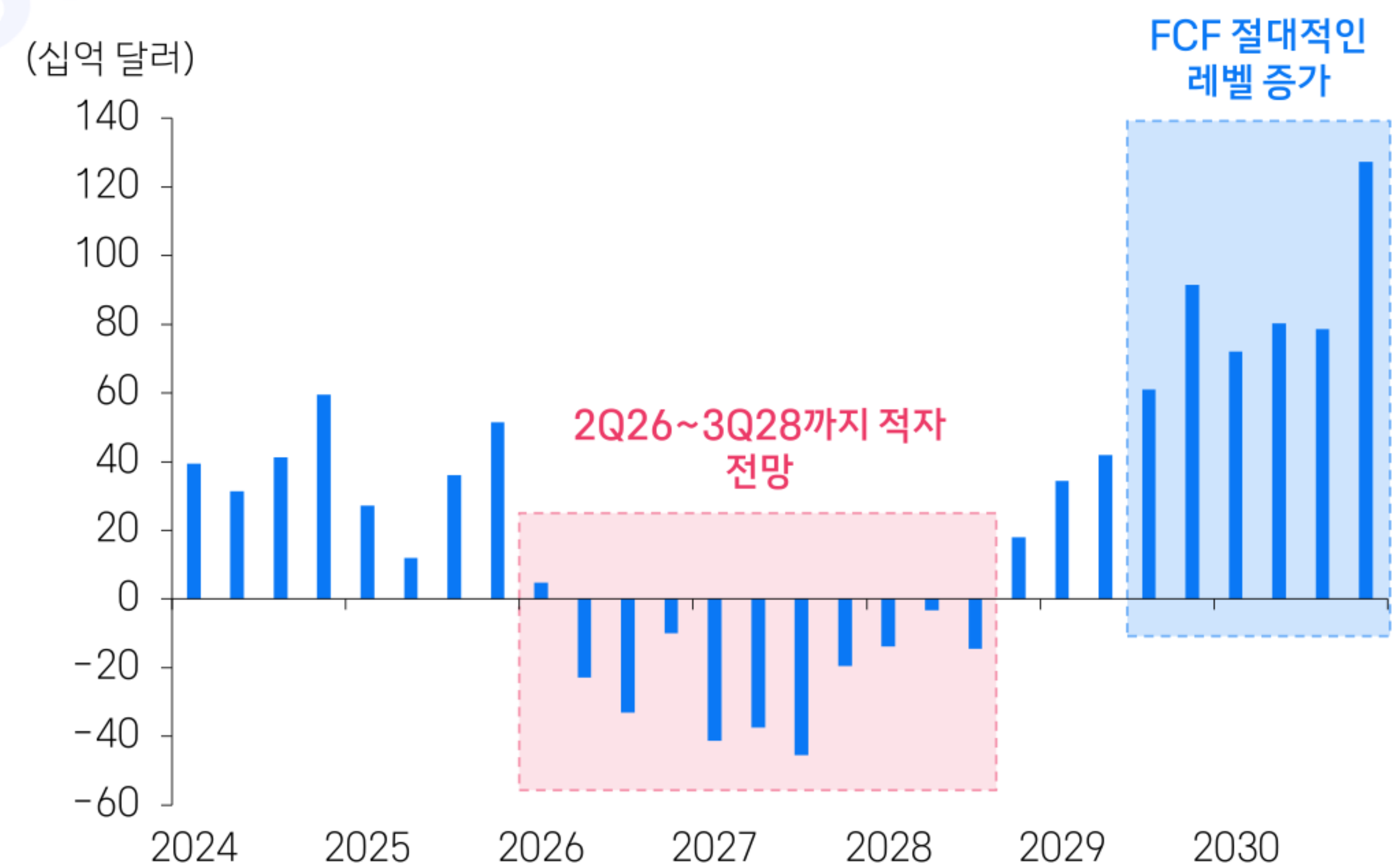
▶ 단기 FCF 악화는 데이터센터의 매출 발생 시점보다 선행 투자에 따른 시차의 문제

- 아마존은 데이터센터 서버 투입 약 2년 전부터 투자가 선행된다고 언급. 서버 및 네트워크 장비는 서비스 시작 수개월 전 구매
- 서버, 네트워크 투자 손익 분기점은 평균 3년 미만 도달 → 이후 2~3년간 유의미한 FCF 창출 가능

하이퍼스케일러의 RPO 증분과 분기별 CapEx 금액 추이



하이퍼스케일러 FCF 흑자 전환 시점 및 이후 절대값 레벨 상향



참고: 알파벳, 마이크로소프트, 아마존, 오라클 포함
자료: 각 사, 삼성증권

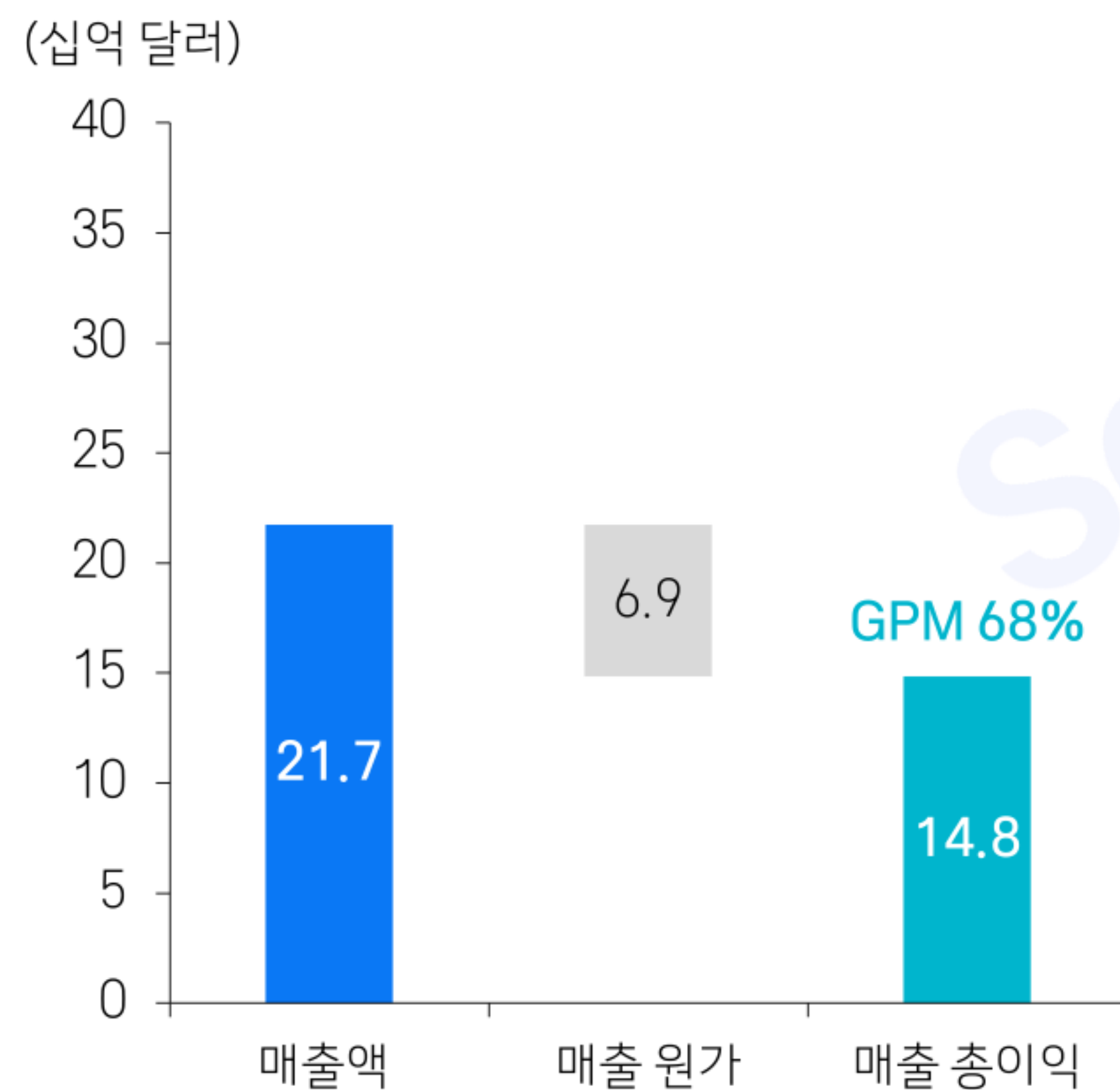
자료: Bloomberg, 삼성증권

컴퓨터 경제성 | 하이퍼스케일러 수익성 분석

▶ 하이퍼스케일러의 컴퓨터 사업 영위 방식 - 1) 컴퓨팅 임대 2) 모델 추론 서비스 → 컴퓨터로 GPM 68 ~ 81% 창출 가능

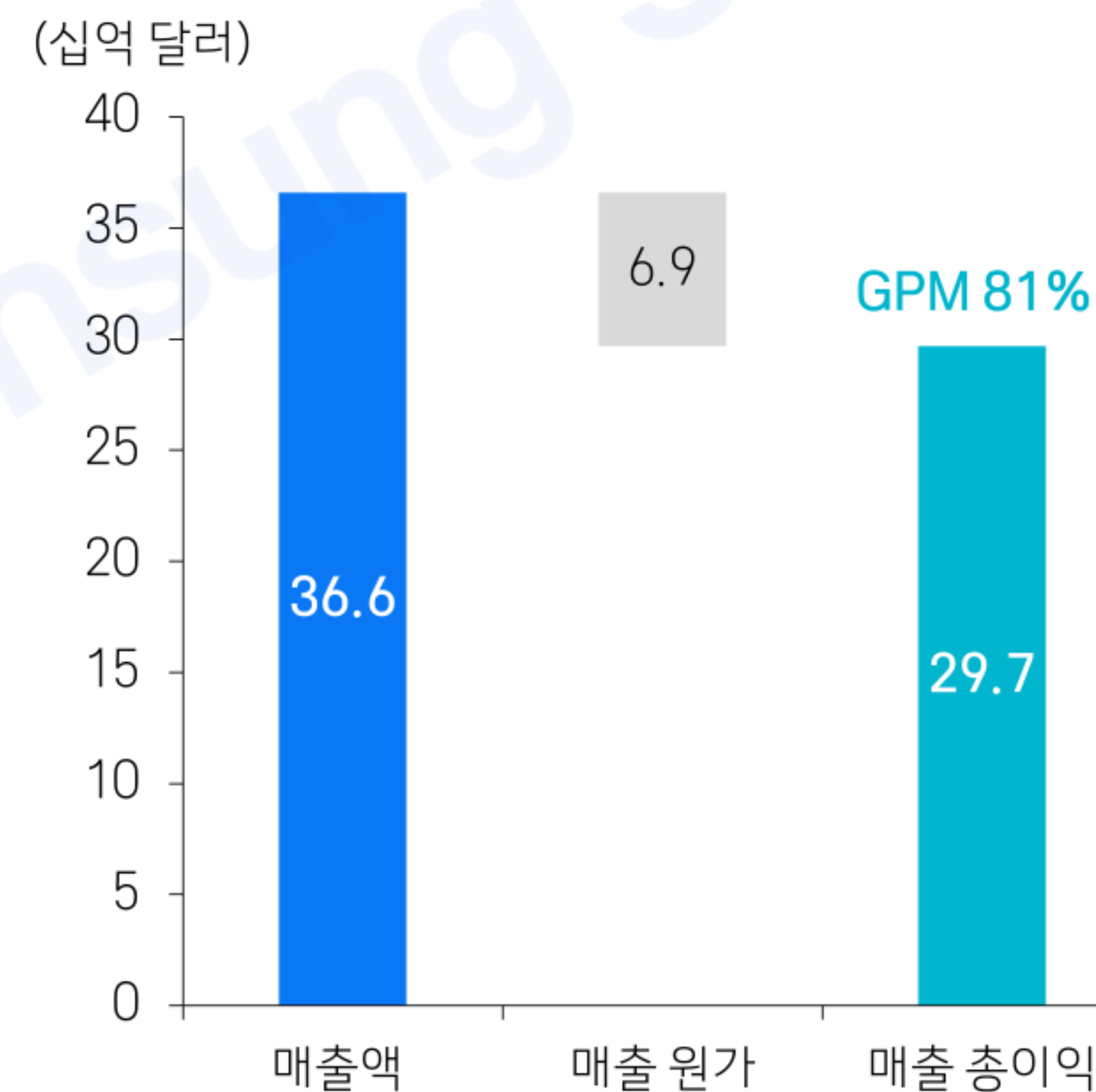
- 매출 베이스가 크고 마진도 높은 사업은 모델 추론 서비스 → 메타가 보유 캐파를 임대보다 자사 서비스에 집중 배치하는 이유
- 그러나 스페이스X 사례처럼 컴퓨터 희소성 프리미엄 기반 단순 임대 서비스도 모델 추론 서비스와 유사한 수준의 매출과 마진 달성 가능
- 향후 대량 컴퓨터를 배치하고 서비스할 수 있는 사업자와 그렇지 못한 사업자 간의 유연성과 수익성 차이는 더욱 벌어질 것으로 생각

컴퓨터 임대 서비스 GPM



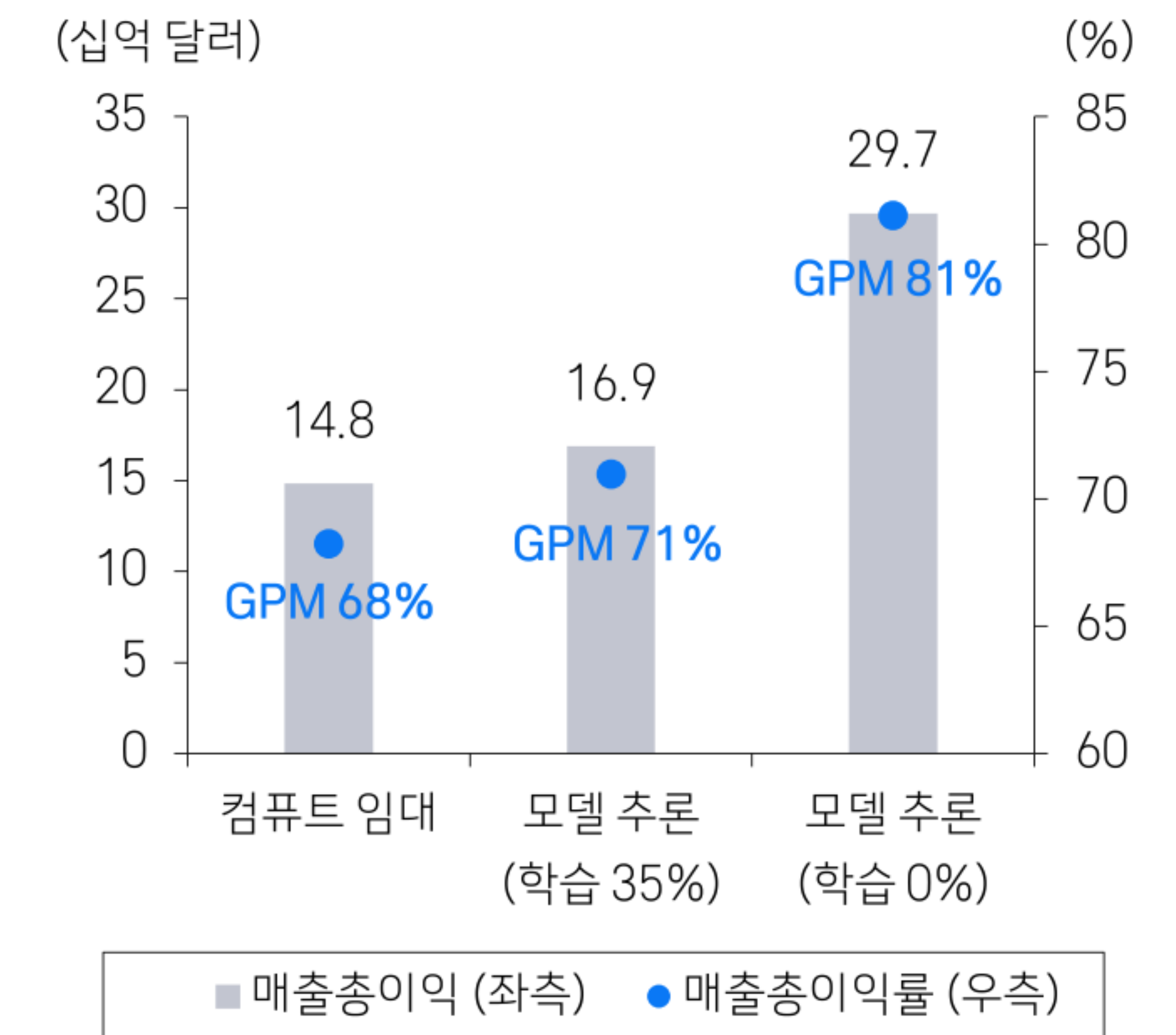
자료: 삼성증권

오픈웨이트 모델 추론 서비스 GPM



자료: 삼성증권

하이퍼스케일러 컴퓨터 사업별 GPM 비교

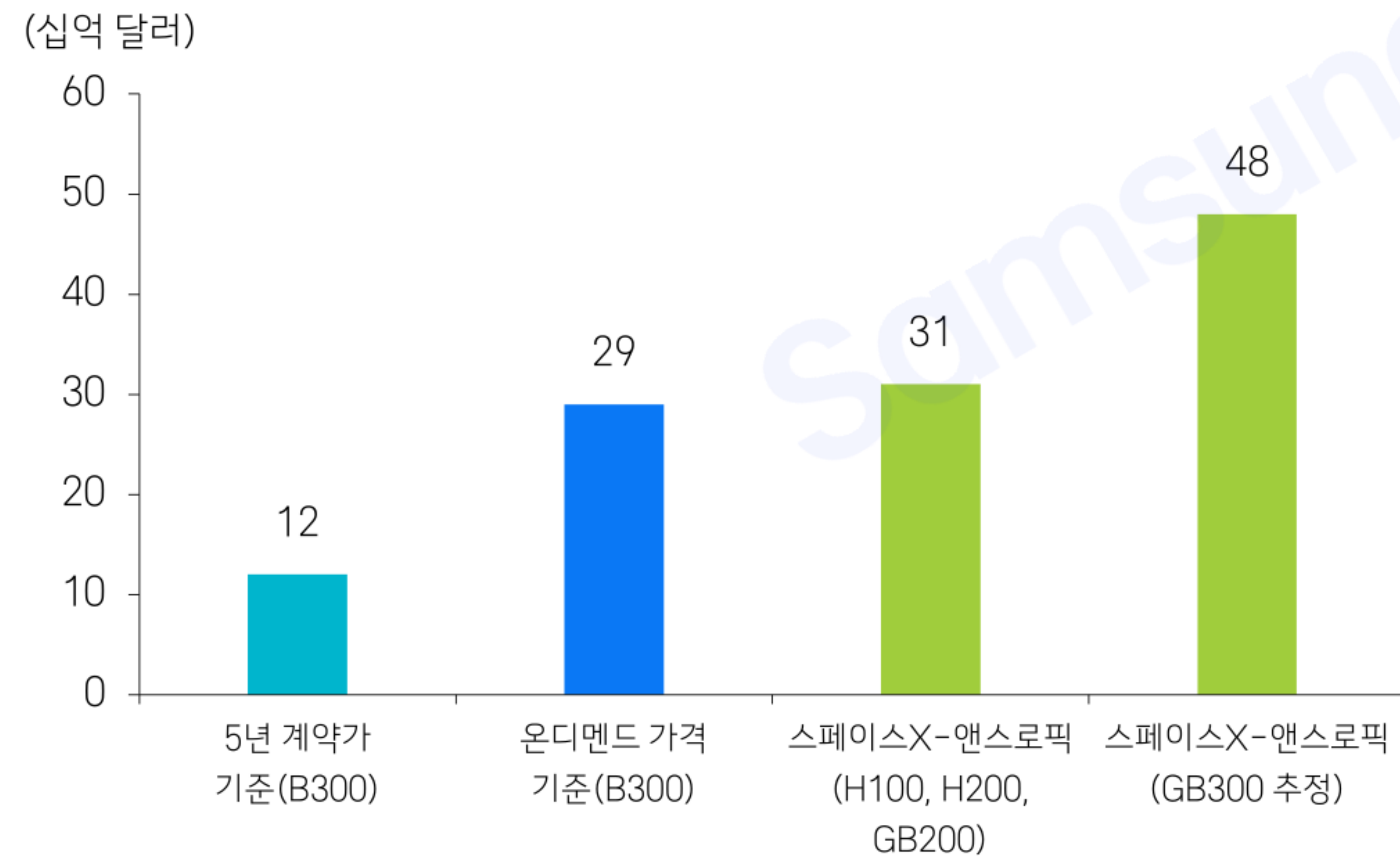


자료: 삼성증권

컴퓨터 경제성 | 스페이스X와 네오 클라우드

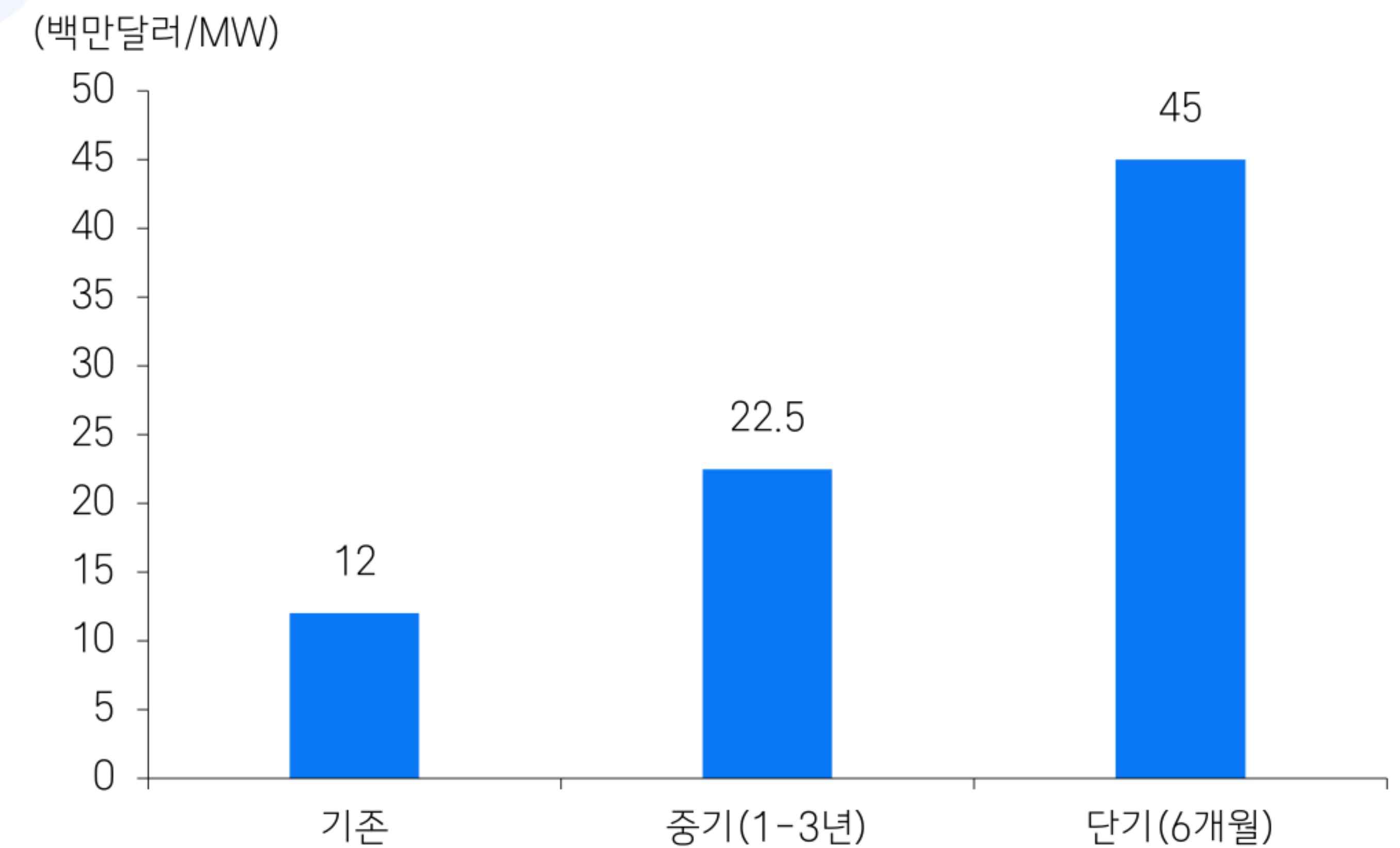
- ▶ 스페이스X는 장기 임대보다 높은 프리미엄으로 컴퓨터 캐파 제공 계약 → 즉시 인도 + 희소성 + 상호 해지 옵션의 프리미엄 반영
 - 계약은 90일 전 상호 해지 가능 조약 존재. 따라서 명목 계약 기간은 3년이지만 가격은 단기 임대가 기준 책정
 - 산업 내 컴퓨터 부족 상황 속에서 짧은 계약 기간, 유연한 계약 구조를 바탕으로 높은 가격 프리미엄을 향유하는 전략
- ▶ 네비우스 실적에서도 유사한 트렌드 확인. 장기 계약에서 캐파 포트폴리오 최적화로 전환
 - 캐파를 의도적으로 선 판매하지 않고, 배치 시점에 가깝게 판매해 더 높은 가격을 확보하는 전략. 단기 계약 MW당 ACV는 4,000~5,000만 달러

1GW 컴퓨터 기준 임대 사업 연 환산 매출 비교



자료: SemiAnalysis, 삼성증권

네비우스의 계약 기간 별 MW당 ACV(평균 계약 가치)



자료: Nebius, 삼성증권

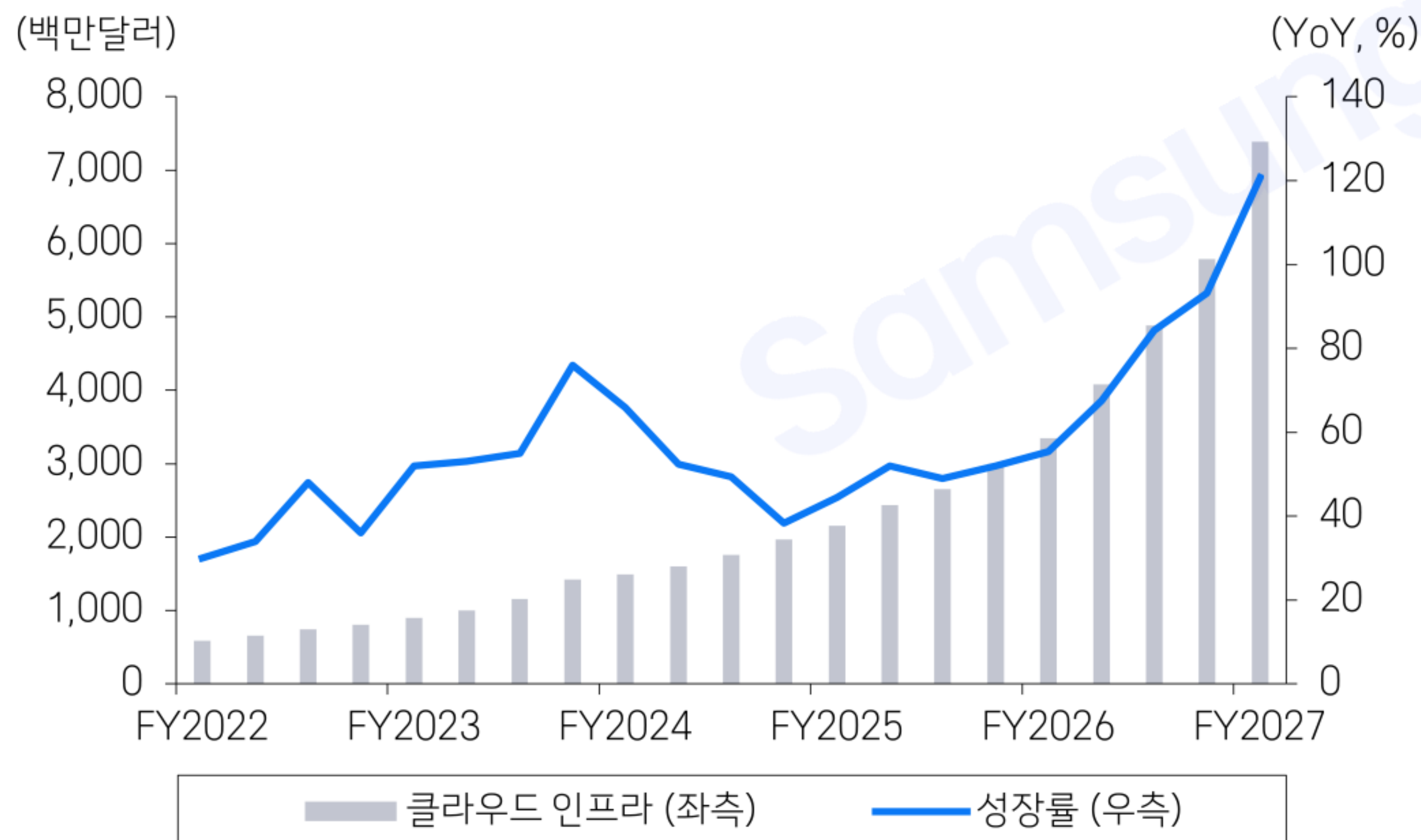
오라클 | 실적은 무난하게 좋아지지만 애매한 강점

▶ RPO와 빠른 캐파 확장을 통한 매출 성장 본격화. 클라우드 인프라 기반 애플리케이션과 데이터베이스 부문도 긍정적 선순환 구조

- 고객 선급금 및 BYOH 모델을 활용해 추가 자본 투자가 불필요하다는 점 강조. 일부 사이트 지연 우려 존재하나 가이드언스에는 영향이 없다는 입장
- GPU 가동률 97.9%. 분기 갱신 시점 도래 구형 GPU 캐파는 기존 계약 대비 20% 높은 가격으로 갱신 또는 재판매. 견조한 수요 확인
- 주가는 오픈AI의 모델 경쟁력과 연동되는 측면. 다만 대형 단일 고객 의존, 프로젝트 동시 진행에 따른 개별 사이트 이슈, 자금 조달 이슈 지속
- 안정성에서는 빅테크 대비 열위, 모멘텀에서는 네오 클라우드 대비 열위. 10월 AI World 행사에서 가이드언스 업데이트 전까지 개별 모멘텀 제한적

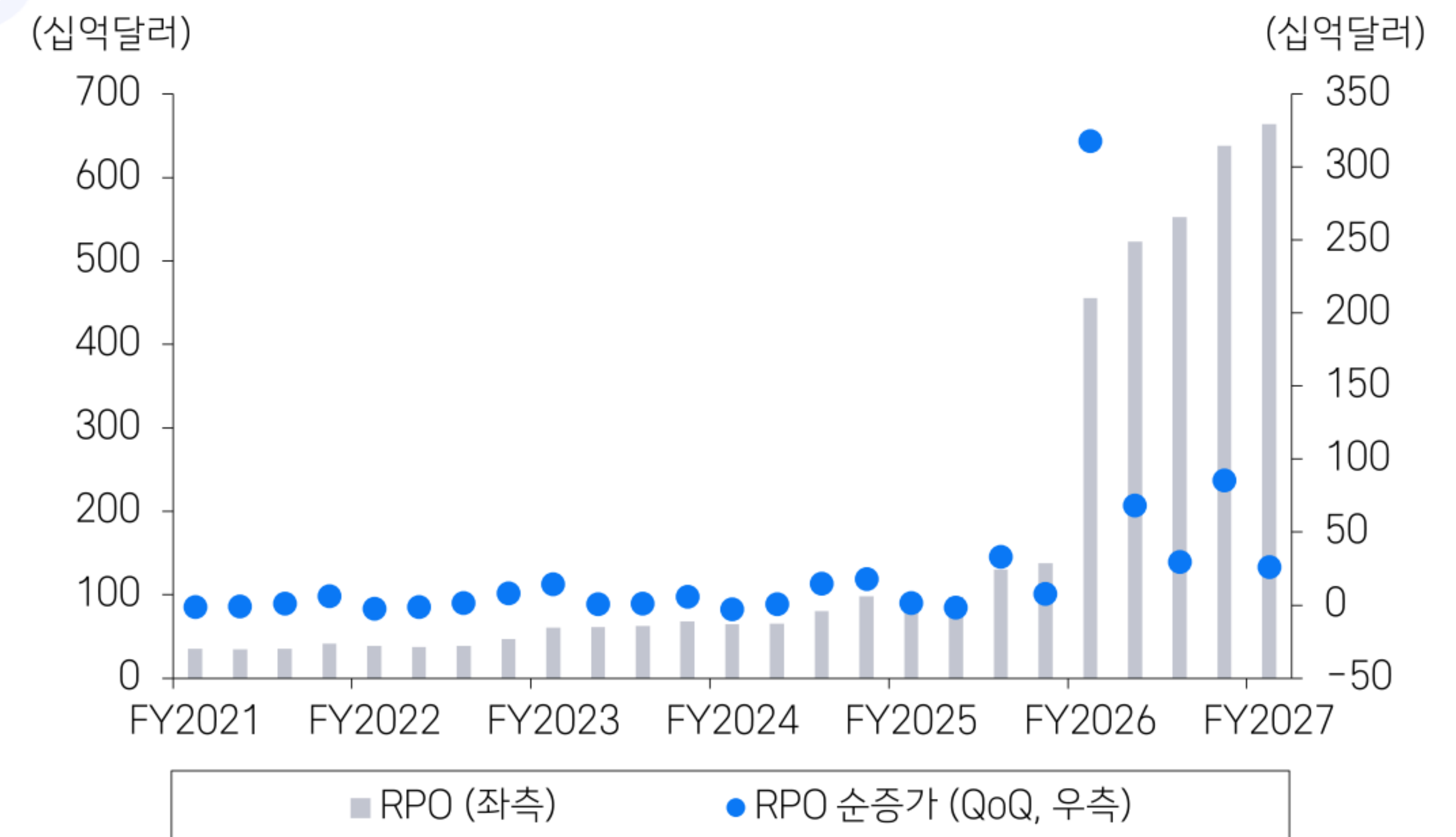
* BYOH(Bring Your Own Hardware): 고객이 AI 반도체를 비롯해 캐파 확장에 필요한 하드웨어를 직접 조달하는 방식. 오라클의 CapEx 규모 축소 효과

오라클 클라우드 인프라 매출 추이



자료: Oracle, 삼성증권

오라클 RPO 추이

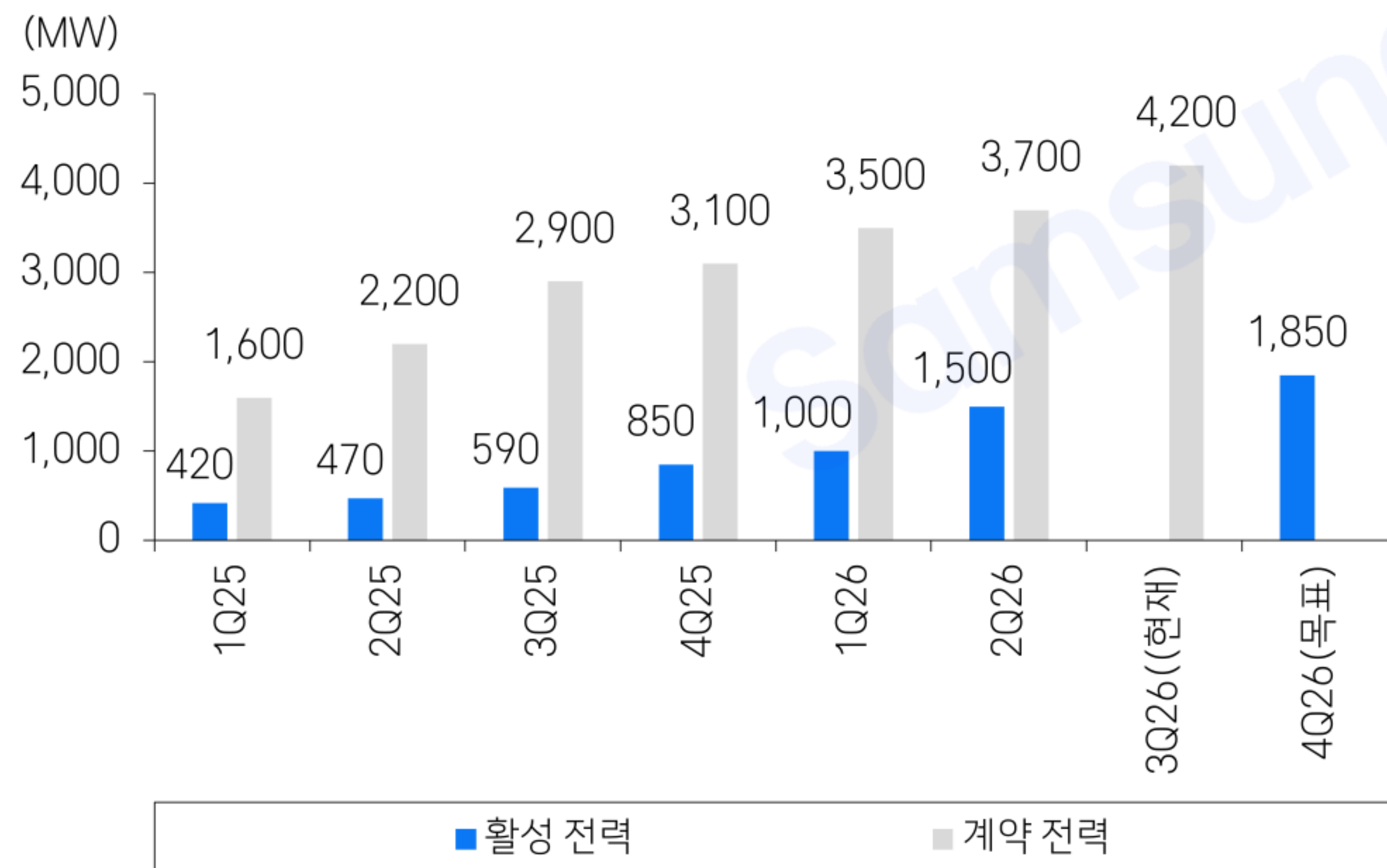


자료: Oracle, 삼성증권

네오 클라우드 | 캐파 확장 속 스택의 중요성 부각

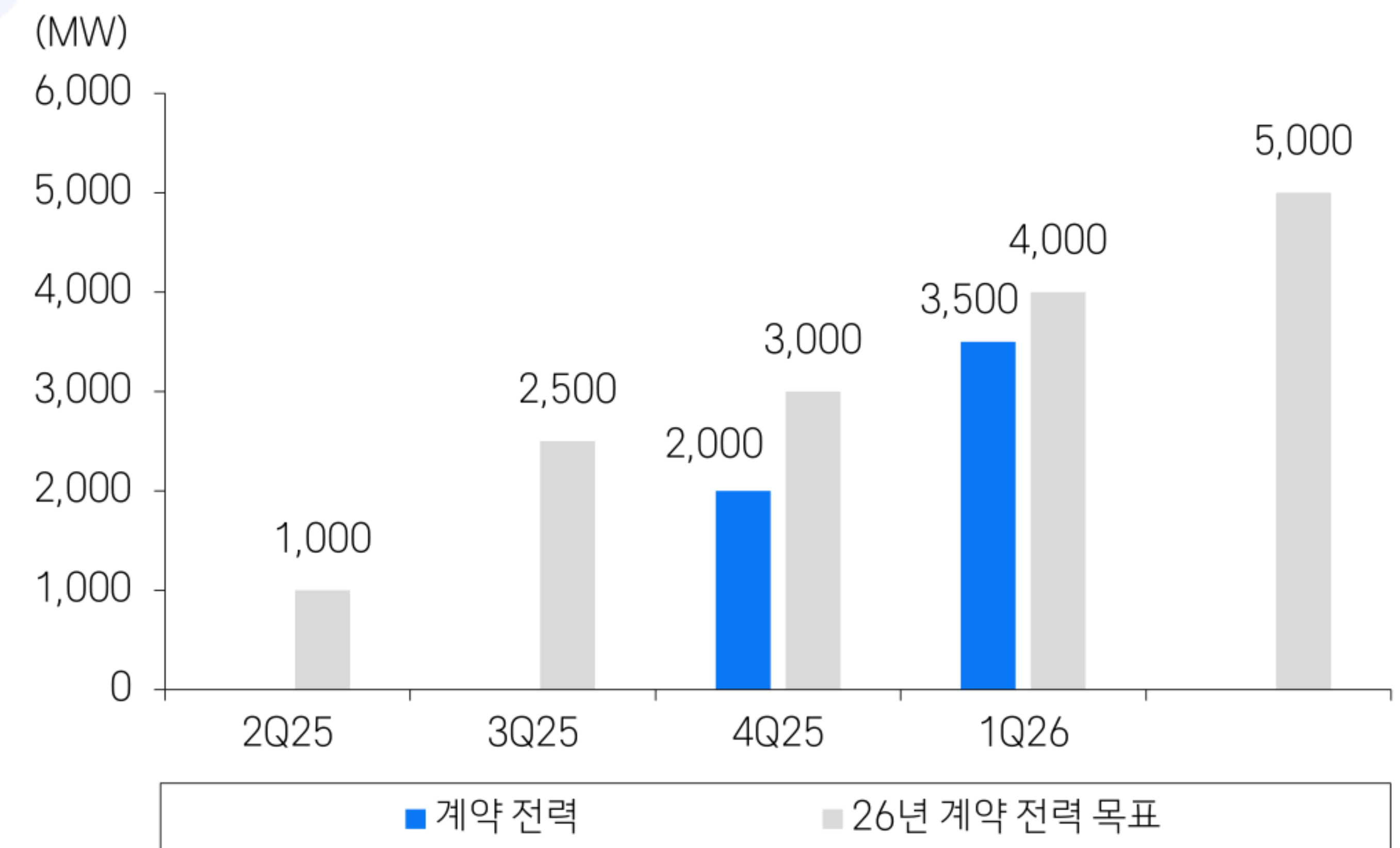
- ▶ 네오 클라우드 기업의 2Q26 실적에서는 1) 여전히 높은 수요와 부족한 공급, 2) 캐파 관련 가격 결정력 강화, 3) 마진 개선 확인
- '27년 성장의 병목 요인으로 꼽은 것은 수요보다 전력 및 캐파의 적절한 확보. 빠르게 실제 매출을 발생시킬 수 있는 캐파로 전환할 수 있는지 여부
 - 구형 GPU도 워크로드 믹스 변화와 공급 부족 속 유효한 수익화 자산으로 존재. 초기 계약 종료 이후 추가 수익화 여력이 큰 상황
 - 장기 계약에서 캐파 포트폴리오 최적화로 전략을 전환. 다음 성장 축으로 고마진 소프트웨어 및 서비스 믹스 확대를 제시(관리형 추론, 토큰 팩토리)

코어위브 활성 전력 및 계약 전력 캐파 추이



자료: CoreWeave, 삼성증권

네비우스 전력 캐파 확장 계획



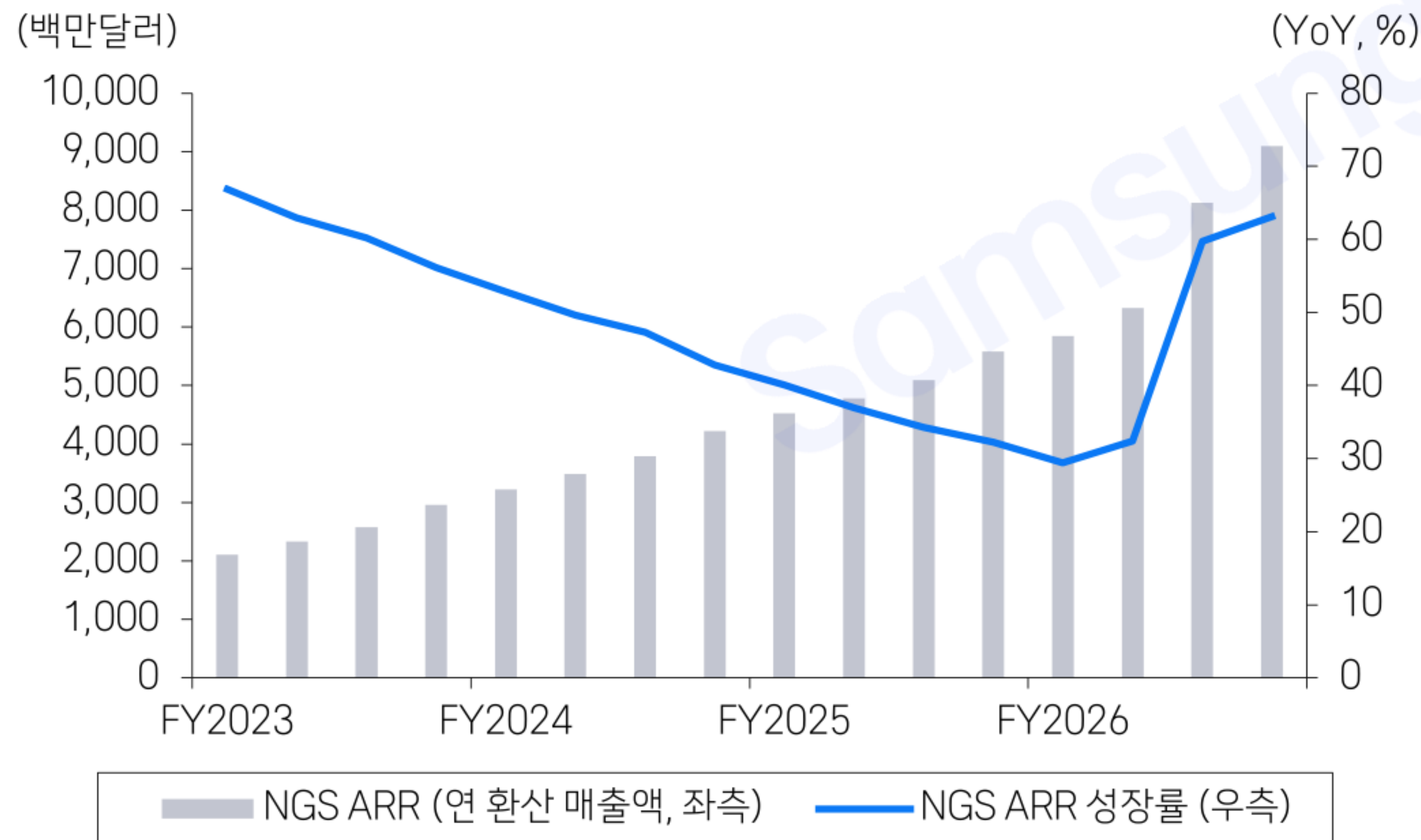
자료: Nebius, 삼성증권

사이버 보안 | Mythos에 이어 오픈 모델 확산이 만든 변곡점

▶ AI의 발전과 확산은 공격과 방어 양 쪽에서 사이버 보안 수요를 구조적으로 증가시키는 요인

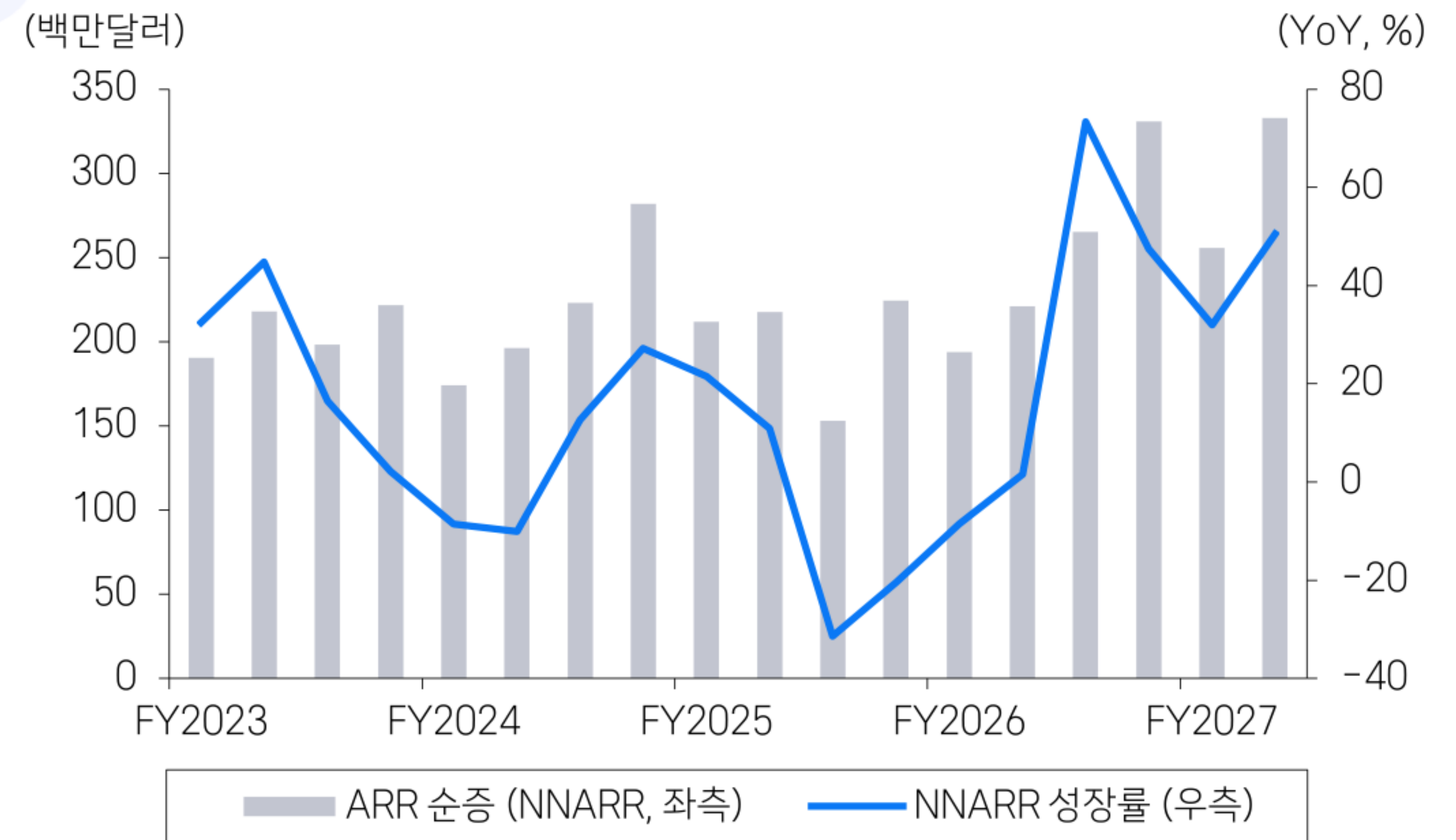
- AI가 공격자의 생산성을 높이는 동시에 기업 내 AI 에이전트와 AI 인프라 확대 → 공격 속도와 보호해야할 데이터, 아이덴티티, 공격 표면 동시 증가
- Mythos에 이어 오픈 모델 확산도 기업별 자체 모델과 에이전트 도입을 촉진. 네트워크, 클라우드, 아이덴티티 등 기존 보안 영역도 수혜
- 공격 속도가 인간 중심 대응 체계를 넘어서는 만큼 통합 보안 데이터와 AI 기반 실시간 탐지, 자동 대응의 중요성 확대 → 신규 AI 보안 시장 부상
- 축적된 위협 데이터, 실시간 보안 데이터, 고객 신뢰와 유통망은 AI 기업의 침투를 견제할 수 있는 해자. 정렬 실패 리스크도 보안의 필요성으로 연결

팔로알토 네트워크의 NGS 연환산 매출 추이



자료: PaloAltoNetworks, 삼성증권

클라우드 스트라이크의 NNARR 및 성장률



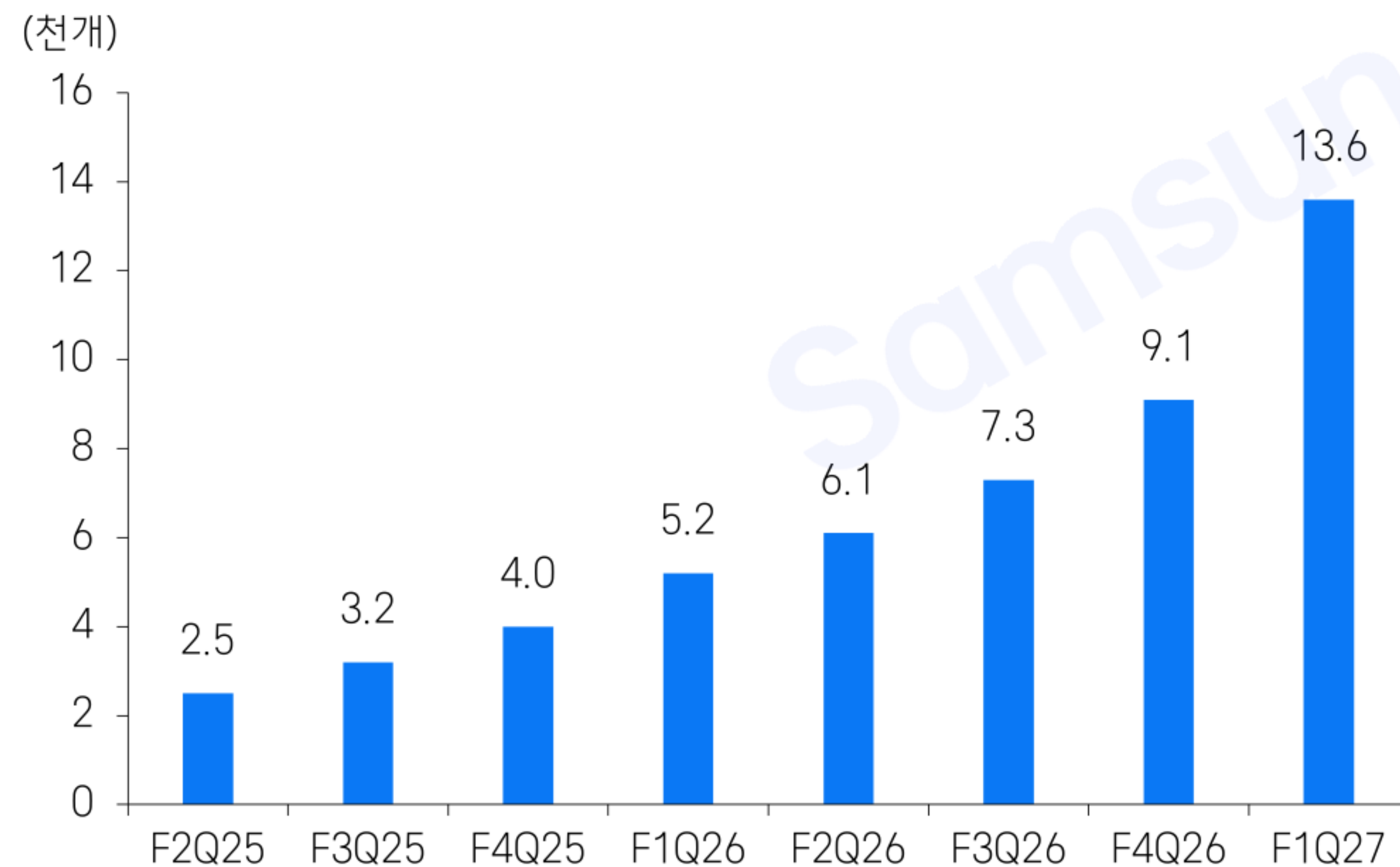
자료: CrowdStrike, 삼성증권

인프라 SW | 스노우플레이크 & 클라우드플레어 vs 데이터도그, 엣갈린 운명

▶ AI 효과의 코어 확산 여부 + 특정 AI 네이티브 고객 집중 여부 + 엔터프라이즈 고객 전반의 AI 도입 수혜 여부

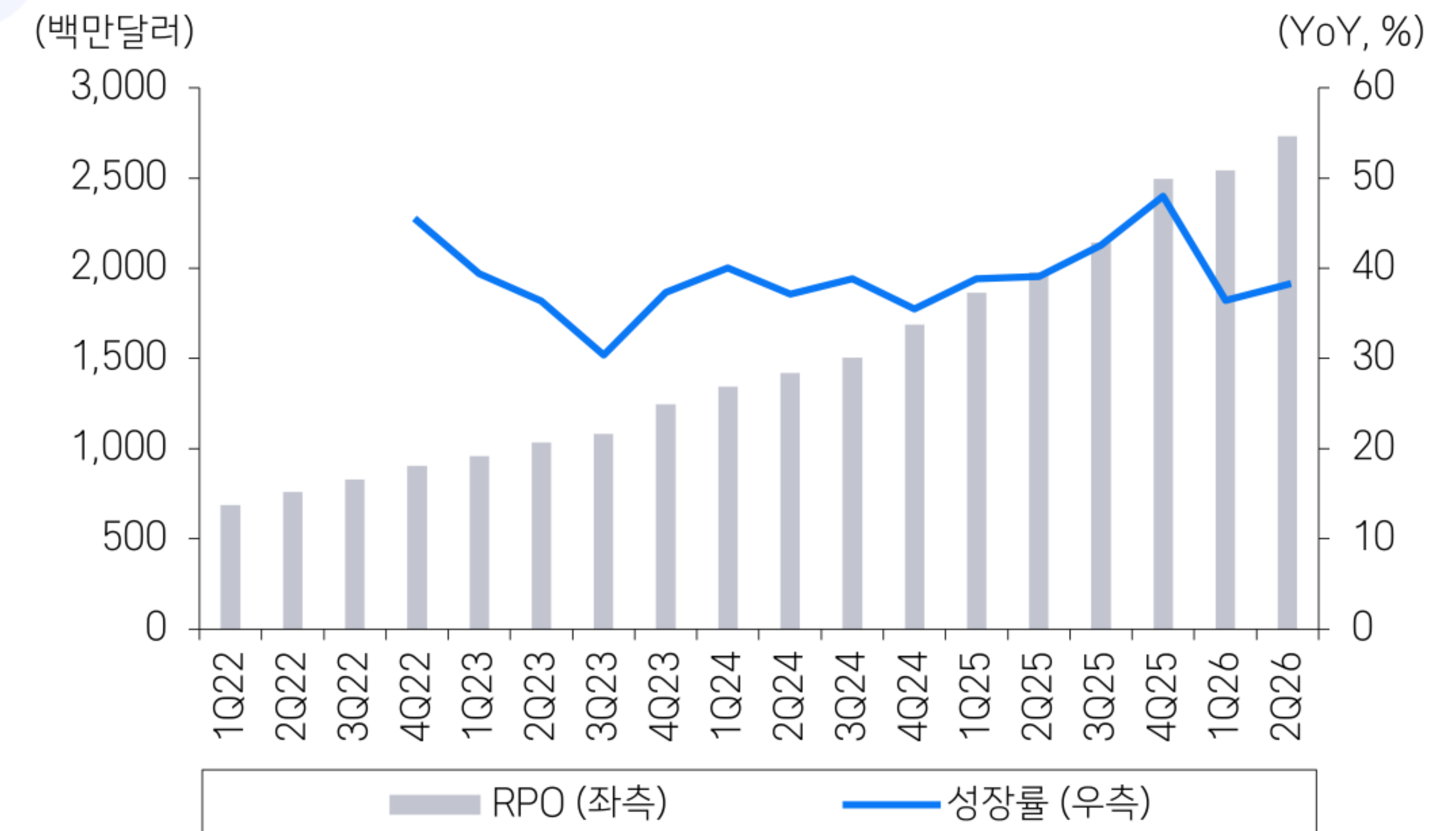
- 스노우플레이크→ Cortex Code(CoCo), CoWork 포함 광범위한 AI 포트폴리오 성과 확인. 제품 매출 성장률 가속분의 절반이 AI 매출에서 발생. AI 도입 고객의 코어 사용량이 함께 증가하는 플라이휠 사이클. AI 네이티브 고객 비중은 여전히 작은 수준. 엔터프라이즈 고객 전반 AI 도입 효과
- 클라우드플레어→Workers 개발자 플랫폼과 에이전틱 AI 관련 모멘텀 강화. 에이전트 트래픽 증가가 플랫폼 전반의 수요로 연결
- 데이터도그 →건조한 실적 성장을 보여줬지만, 최대 고객(오픈AI)의 사용량 감소가 3Q부터 시작되며 관련 영향을 가이던스에 보수적 반영

스노우플레이크의 주간 AI/ML 기능 사용 고객



참고: F2Q27에는 전체 AI 지표가 아니라 제품별 CoCo 9,100개, CoWork 5,800개로 발표
자료: Snowflake, 삼성증권

클라우드플레어의 RPO 및 성장률



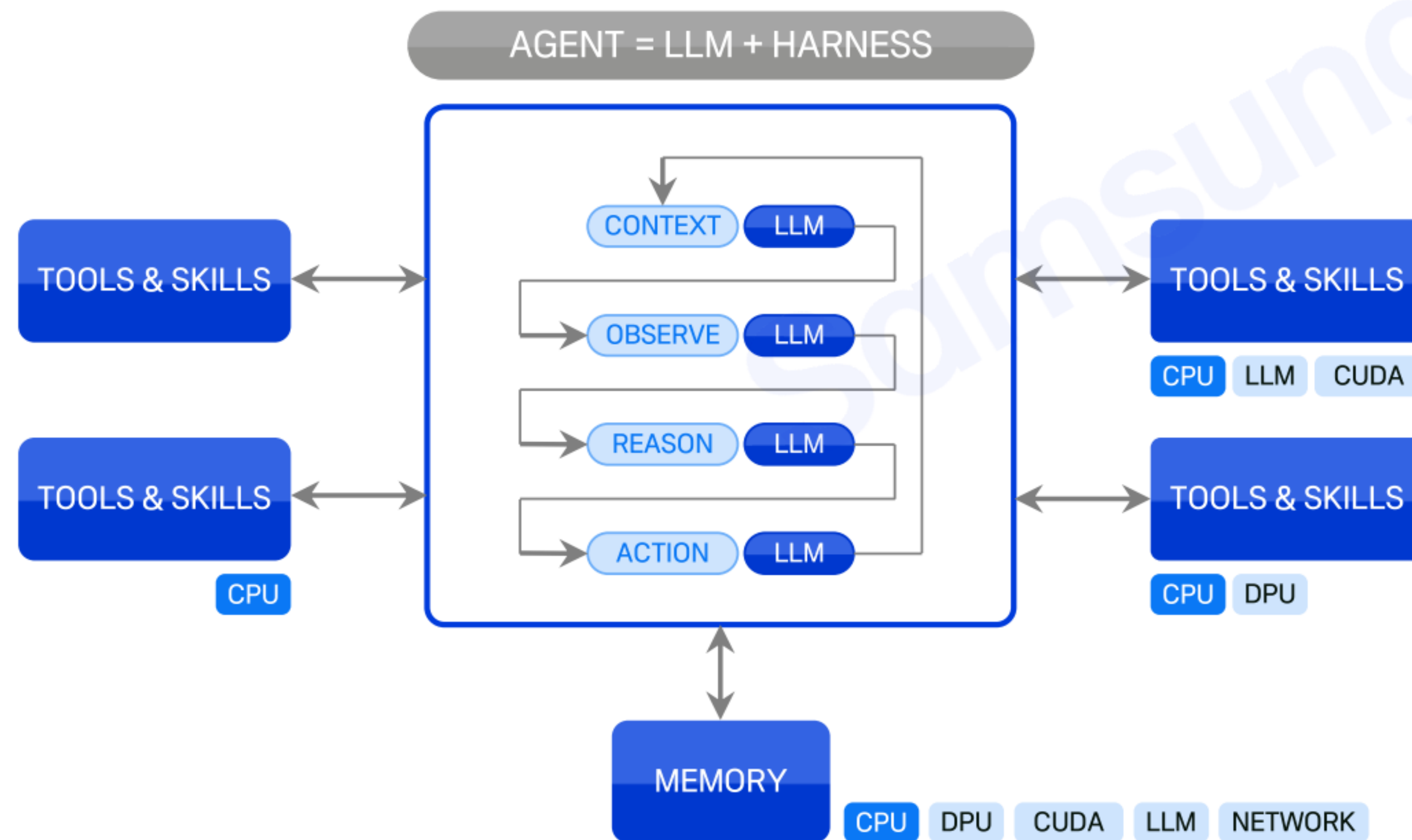
자료: Cloudflare, 삼성증권

높아지는 하네스의 중요성 | 그러나 이것도 AI 기업이 만들어낸다면?

▶ 엔터프라이즈 AI 본격화에 따라 모델이 활용하는 도구와 하네스의 중요성 동반 상승

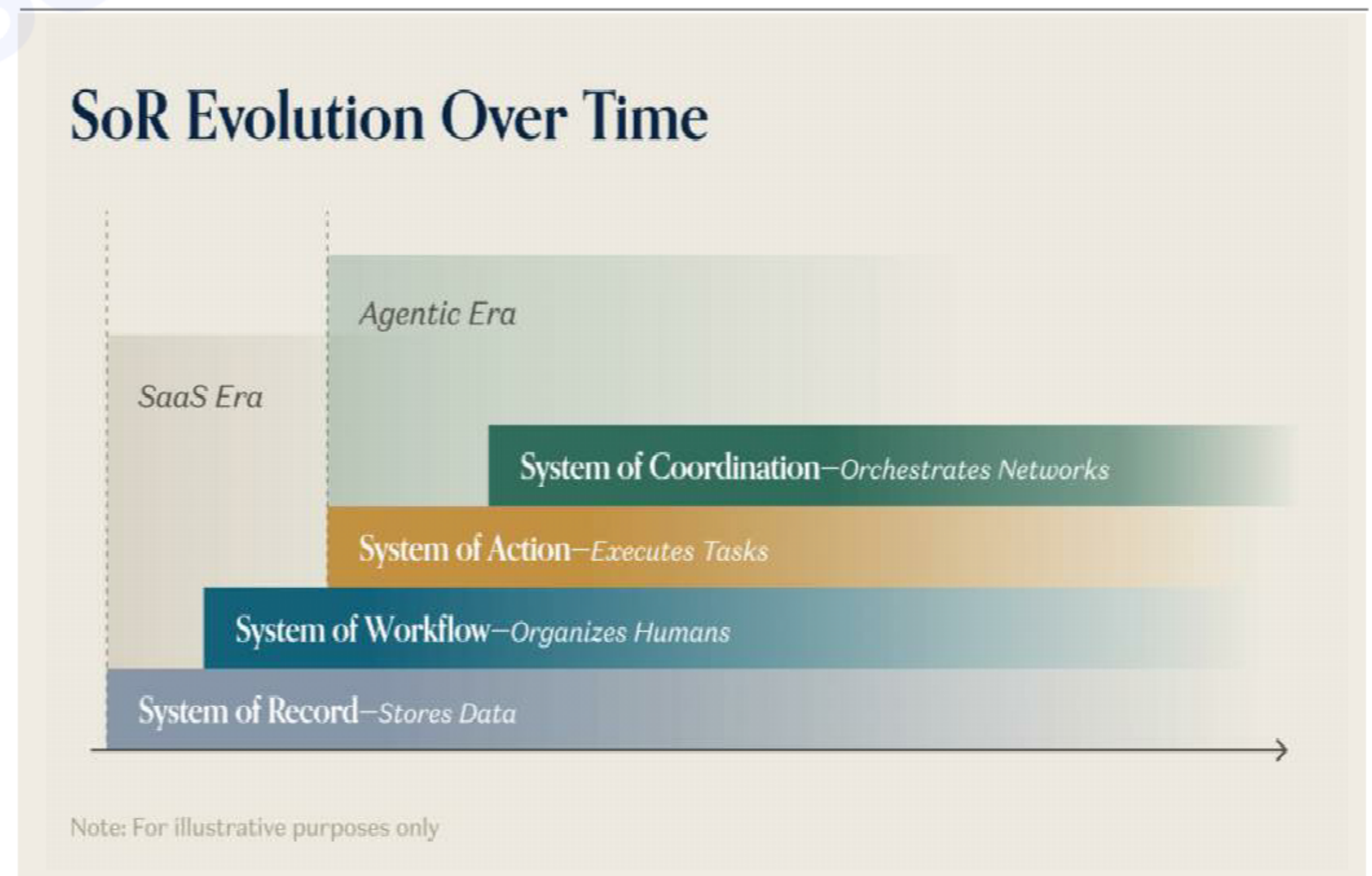
- 젠슨 황은 연초 Computex 키노트에서 자율 에이전트(Agent)를 두뇌(LLM, AI 랩 기반)과 운영체제(Harness, SW 제공)가 결합된 형태로 정의
- 에이전트 구성 요소(모델, 하네스, 런타임) 간 경계가 흐려지는 상황 → 오히려 에이전트 엔지니어링 레이어가 AI 모델 기업으로 편입
- Astra의 ARC-AGI-3 결과 중 Provider Adaptive(자체) 하네스 활용은 Standard 하네스 대비 +37%p. 상태 관리와 컨텍스트 유지 방식 중요성

자율 에이전트 = 두뇌(LLM) + 운영체제(Harness)



자료: NVIDIA, 삼성증권 정리

소프트웨어의 가치가 창출되는 영역의 변화 전망



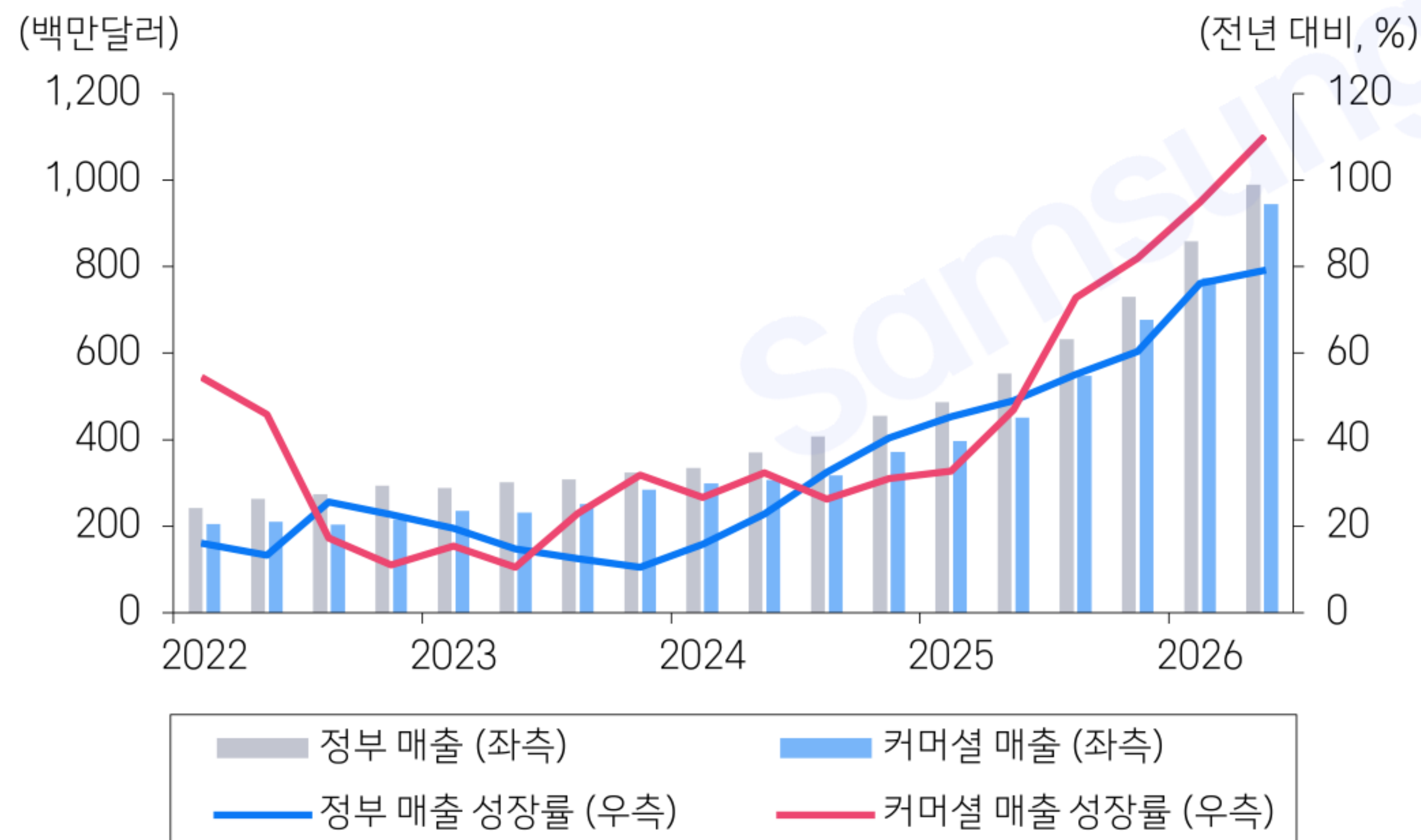
자료: a16z, 삼성증권

팔란티어 테크놀로지스 | 소버린 AI의 중요성

▶ 팔란티어는 매출 성장률과 이익률 측면에서 최고 수준을 지속적으로 경신. AI 기반 성과를 입증하는 유일무이한 소프트웨어 기업

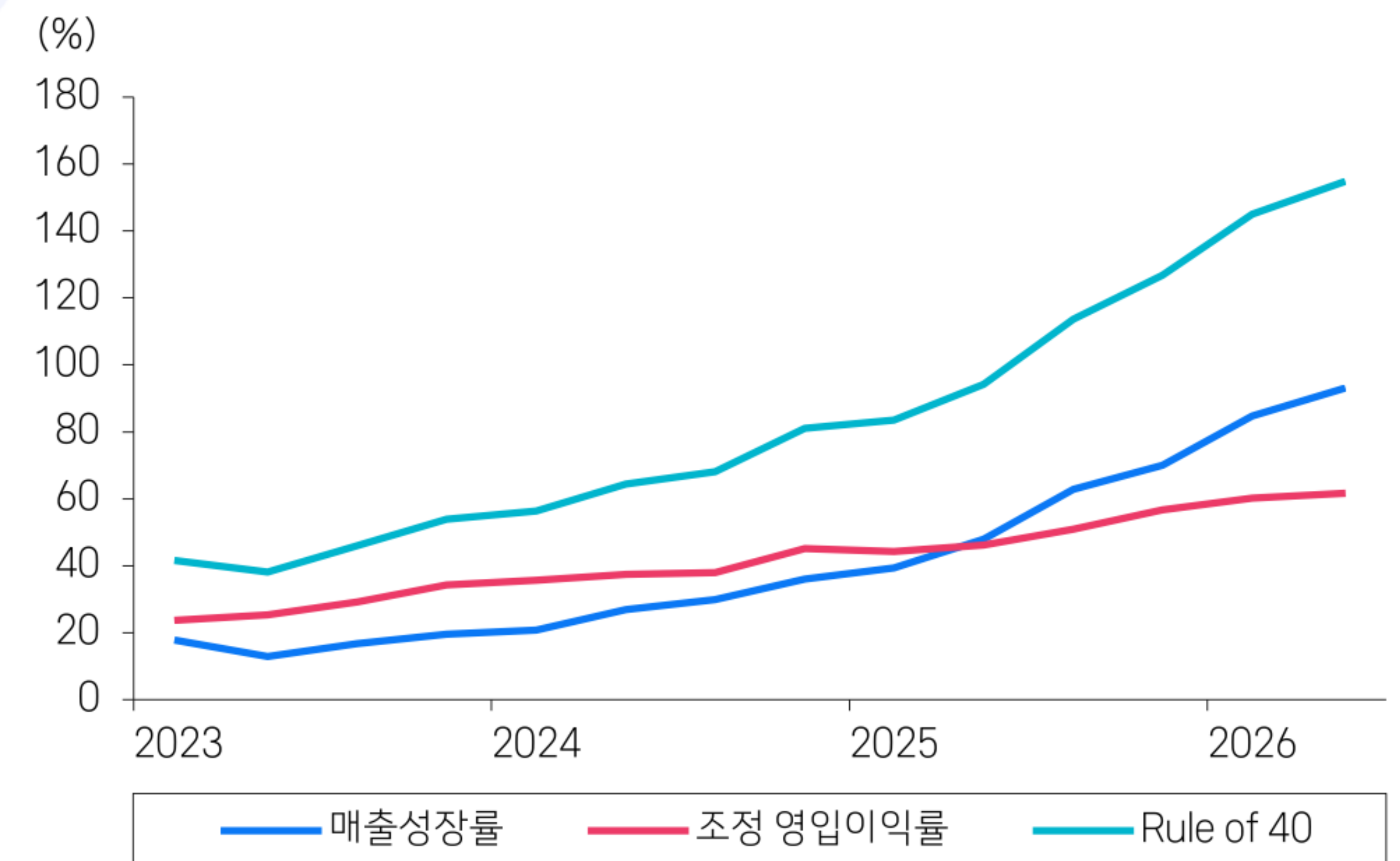
- AI Slop과 토큰 기반 AI를 비판하며, 기업의 데이터, 모델 가중치, 업무 노하우를 고객이 직접 소유하는 소버린 AI를 핵심 패러다임으로 제시
- 소버린 AI는 국가 단위 AI 구축 프로젝트나 과제로 인식되었지만, 기업 고객의 AI 도입에서도 주권이 필요하다는 입장
- 모델 자체보다 AIP + 온톨로지 + FDE를 통한 실제 운영 역량이 경쟁력. 동일 모델을 사용해도 팔란티어 플랫폼이 더 높은 경제적 가치 창출
- AI 모델을 활용한 솔루션을 제공하는 엔드 소프트웨어 기업 중에서는 성장성과 보유하고 있는 해자 관점에서 가장 앞서 있다고 생각

정부와 커머셜 매출이 모두 고성장하는 국면



자료: Palantir Technologies, 삼성증권

팔란티어의 Rule of 40(매출 성장률 + 조정 영업이익률) 추이



자료: Palantir Technologies, 삼성증권

Compliance Notice

- 본 조사분석자료의 애널리스트는 2026년 9월 15일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2026년 9월 15일 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.

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