Market Insight | 2026.4.17

그림: Midjourney(중국의 오픈소스의 성장을 바라는 엔비디아의 CEO)

AI Bi-weekly

너무 빨리 변하는 AI… 결국에는 진짜들만 남는다

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com

CONTENTS

Highlight of the Week

젠슨 황의 대담을 프로파일러 관점으로 해체하기

젠슨 황 CEO가 드와케시 파텔과의 1:1 대담에서 드러낸 발언과 침묵, 평정심을 유지한 구간과 참았던 감정이 터진 구간은 엔비디아 제국의 해자가 어디에 있고 어디가 얇은지를 드러내는 농밀한 1차 사료에 가깝습니다. 저희는 프로파일러가 된다는 생각으로, 그의 인터뷰를 해체하였고, 이를 토대로 현재 시점의 AI 하드웨어 지도를 그려본 결과, 몇 가지 핵심 논지로 추려지게 되었습니다.

첫째, 에이전틱 AI의 도래는 AI 인프라 경제학 전체를 뒤흔든 'CPU 르네상스'와 '메모리 폭증'이라는 구조적 수요 혁명을 촉발했습니다. TrendForce의 데이터를 보면, 과거 GPU:CPU가 10:1이었던 비율이 에이전틱 워크로드에서 1:1로 수렴하고 있습니다. 이는 일종의 권력 이동입니다. AMD의 CPU(EPYC)가 데이터센터 매출을 39% 폭증시키고, 인텔이 PC CPU 웨이퍼를 서버용으로 몰아주며, OpenAI가 AWS와 "수천만 개의 CPU" 확보 조항이 담긴 380억 달러 계약을 맺은 이유가 바로 여기에 있습니다. CPU의 슈퍼사이클이 본격화되고 있습니다.

한편 Dell의 CEO에 따르면, 메모리 수요는 가속기당 용량 25배에다 클러스터 스케일의 25배가 곱해져 최대 625배까지 폭증하는 비탄력적 국면에 진입했습니다. 이로써 HBM의 웨이퍼 캐파 3~4배 잠식, 첨단 패키징의 제조 사이클 장기화, 하이엔드 스펙의 초저수율이라는 3중 병목으로 인해 메모리 슈퍼사이클은 2028년까지 지속되며, 2026년 2분기 DRAM 계약 가격은 전 분기 대비 50% 이상 상승이 예고됩니다.

둘째, 미중 AI 경쟁의 진짜 승부는 FLOPs가 아니라 'Temporal Delta'에 있다는 것입니다. 젠슨 황이 인터뷰 중 평정심을 잃은 구간이 대중국 수출 통제 파트입니다. 다만 DeepSeek V4·R2의 출시 지연이 시사하듯 수출 통제는 중국을 멈추게 하지는 못해도 늦추는 데는 명백히 작동하고 있고, AI 레이스에서 1년의 시간 지연은 10년의 격차로 벌어지므로 이 Temporal Delta 자체가 미국 연구소들의 우위라 판단합니다.

셋째, 지난 4월 8일 Anthropic의 Claude Managed Agents 출시는 소프트웨어 산업의 실리콘 복속을 가속화했습니다. 에이전트 런타임 레이어가 commodity화되면서 얇은 UI 기반 SaaS와 LLM 래퍼형 제품은 정면으로 압박을 받고 있습니다. 이에 따라 에이전트 오케스트레이션 자체는 이제 해자가 아니게 됐습니다. 그러나 역설적으로, 에이전트가 싸게 복제될수록 그 에이전트가 접속하는 "현실의 데이터·권한·검증 작업"은 더 희소해집니다. 팔란티어의 온톨로지와 시놉시스의 물리 엔진이 AI 시대에 오히려 재평가받는 이유입니다.

결론적으로, 우리들이 던져야 하는 핵심 질문은, 여전히 "누가 AI 시대의 물리적 인프라를 장악했는가, 누가 그 인프라 위에서 돌아갈 에이전트의 현실을 정의하는가"입니다. 이 질문에 답하는 기업들은 점점 소수가 되어가고 있으며, 그들의 가격은 시장이 아직 완전히 반영하지 못하고 있다고 판단합니다.

성능은 위로, 부담은 아래로

HAI가 2026년 4월 13일 발간한 AI Index Report는 AI 산업의 두 가지 상반된 흐름을 동시에 기록하고 있습니다. 코딩 벤치마크 SWE-bench Verified에서 AI 성능은 1년 만에 인간 기준선에 근접했고, Humanity's Last Exam 정답률은 8.8%에서 50%를 넘어섰습니다. Generative AI는 출시 3년 만에 전 세계 인구의 53% 채택에 도달했으며, 테크 업계 조직 채택률은 88%에 달합니다. 능력 곡선은 꺾이지 않고 있습니다.

그러나 같은 보고서가 측정한 반대편 지표들은 다른 이야기를 합니다. Foundation Model Transparency Index는 58점에서 40점으로 역대 최저를 기록했습니다. 가장 강력한 모델일수록 공개하는 정보가 가장 적습니다. AI가 일자리 방식에 긍정적 영향을 미칠 것이라는 전문가 응답은 73%인 반면, 일반 대중은 23%에 그칩니다. AI 능력이 빠르게 올라가는 동안, 그것을 만드는 과정은 비밀스러워지고, 사용하는 사람들의 불안은 커지고 있습니다.

고용 데이터는 이 마찰이 이미 현실로 나타나고 있음을 보여줍니다. Brynjolfsson 연구팀의 ADP 급여 데이터 분석에 따르면, 소프트웨어 개발자 22~25세의 고용은 2022년 이후 약 20% 감소했습니다. 같은 기간 30세 이상은 6~12% 증가했습니다. AI는 경험을 보완하고 신입이 수행하던 업무를 흡수하고 있습니다. 단기적으로는 합리적인 선택처럼 보이지만, 10년 후 중급·시니어 인력을 공급할 파이프라인을 동시에 막고 있습니다.

이 마찰은 19세기 러다이트 운동과 구조적으로 닮아 있습니다. 당시 직조공들이 반대한 것은 기계가 아니라 비용과 이익의 불균등한 분배였습니다. 지금도 마찬가지입니다. 조직·정부 레이어에서는 데이터센터의 에너지·수자원 부담이 특정 지역에 집중되는 현상에 대한 반감이, 개인 레이어에서는 커리어 진입로가 좁아지고 있다는 경험에서 오는 분노가 동시에 나타나고 있습니다. 미국 11개 주의 데이터센터 규제 움직임과 Gen Z의 분노 지수 상승은 그 전조입니다.

그러나 이 마찰은 AI 산업만의 위기가 아닙니다. 인터넷과 모바일 생태계도 같은 과정을 거쳤고, 그 속에서 살아남은 플레이어는 기술적 우위만으로 선별되지 않았습니다. 러다이트 운동이 기계를 멈추지 못했듯, 지금의 마찰이 AI의 전진을 막지는 않을 것입니다. 그러나 그 마찰을 어떻게 관리하느냐가 다음 국면의 승자를 가를 가능성이 높습니다.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
12	13	14	15	16	17	18
		· 중과서광 실적(2025 결산)		· TSMC 실적		
19	20	21	22	23	24	25
			· GEV 실적 · VRT 실적 · TSLA 실적 · IBM 실적	· INFO 실적 · SK하이닉스 실적 · SAP 실적 · INTC 실적	·	
26	27	28	29	30	1	2
	· CDNS 실적(잠) · 중과서광 실적(1분기)	· BE 실적	· HD현대 실적(예) · GOOGL 실적 · MSFT 실적 · META 실적 · AMZN 실적(잠)	· 삼성전자 실적 · 네이버 실적 · 크래프톤 실적 · AAPL 실적(잠) · ILMN 실적		

자료: Bloomberg, 미래에셋증권 리서치센터

I. AI & Convergence

1. 왜 이 대담을 경청해야 하는가

(1) 드와케시 파텔이라는 인터뷰어의 무게

드와케시 파텔(Dwarkesh Patel)은 2026년 현재 실리콘밸리에서 가장 준비된, 그리고 가장 공격적인 테크 인터뷰어로 꼽힌다. 그의 이력이 이를 증명한다. 일리야 수츠케버, 다리오 아모데이, 데미스 하사비스, 마크 저커버그, 샘 알트만을 이미 그의 사무실에 초대한 바 있고, 에이전트 시대를 설계한 이들의 뇌 구조를 말 그대로 가장 깊게 들여다본 인터뷰어다. 그의 스타일은 CEO들이 준비해온 세일즈 피치를 기술적 팩트와 논리적 모순으로 송곳처럼 찌르는 것이다.

그리고 며칠 전 젠슨 황이 그를 찾았는데 이번 대담에서 중요한 포인트는 이것이다. 젠슨 황은 몇 년 사이 수많은 기조연설과 단독 인터뷰를 해왔지만, 드와케시와의 대담은 지금까지의 모든 무대와는 온도가 사뭇 달랐다. 그동안 대부분의 질문은 "엔비디아의 다음 칩은 무엇인가", "AI의 미래는 어떤가" 수준의 우호적 프레임이었다.

그러나 드와케시는 "소프트웨어가 범용화되면 하드웨어도 범용화되는 것 아닌가", "중국에 칩을 파는 것이 북한에 핵미사일 껍데기를 파는 것과 무엇이 다른가" 같은 공격적 질문을 마치 투수가 꽂는 스트라이크처럼 쏟아냈다. 이 대담에서 젠슨 황이 내놓은 답변과, 답변 사이의 미세한 방어 기제, 그리고 평정심이 무너진 구간들이, 2026년 엔비디아 제국의 진짜 지도라 할 수 있겠다.

그림 1. 드와케시 파텔 팟캐스트의 젠슨 황 캡처
"중국 시장을 뺏겨도 중국은 결국 테슬라·아이폰처럼 독자 생태계를 만든다"고 도발하자,
젠슨 황이 평정을 잃고 단호하게 자른 바로 그 순간의 장면이다.
엔비디아 제국의 진짜 해자가 단일 칩 성능이 아니라 글로벌 생태계의 락인에 있다는 사실을,
젠슨 황의 평정심이 깨진 이 장면이 역설적으로 증명

자료: 드와케시 파텔, 젠슨 황, 미래에셋증권 리서치센터

(2) 2026년 4월이라는 시점의 무게

이 대담의 타이밍도 중요하다. 현재 엔비디아 시총은 다시 5조 달러 근처에 이르고 있고 Blackwell GPU가 전세계 AI 컴퓨팅의 주력으로 자리 잡았지만, Anthropic이 AMD의 MI450 GPU를 대량으로 구매할 계획이라는 루머까지 나오고 있는 상황이다.

그러한 Anthropic은 ARR $300억 달러에 도달했고, 경쟁사인 OpenAI는 Anthropic의 대주주인 아마존 AWS와 380억 달러 딜을 맺었으며, Mythos급 10T 모델의 사이버 공격 능력이 전 세계 보안 체계를 흔들고 있다. 이렇게 매우 긴박하게 전개되는 국면에서 젠슨 황이 1:1 대담에서 어떤 질문을 부드럽게 넘겼고 어떤 질문에서 목소리 톤이 올라갔는가는, 모든 AI 투자자들이 가장 주의 깊게 봐야 할 신호다.

(3) 대담을 관통하는 한 문장: "전자를 토큰으로"

대담의 가장 중요한 한 순간은, 파텔이 "AI가 소프트웨어를 범용화시키면 엔비디아 하드웨어도 범용화되는 것 아닌가"라고 도발했을 때다. 젠슨 황은 0.5초 정도 옅은 미소를 지은 뒤 이렇게 답했다. "결국 무언가는 전자(Electrons)를 토큰(Tokens)으로 변환해야 한다."

이 문장은 그냥 레토릭으로 치부할 게 아니라 일종의 지배 선언이라 할 수 있다. 그는 엔비디아를 '소프트웨어를 구동하는 칩 회사'가 아니라 우주의 기본 입자인 전자(에너지)를 인간의 지식 단위인 토큰으로 치환하는 유일한 연금술사의 자리로 격상시킨 것이기 때문이다. 이 프레임은 제본스의 역설(Jevons Paradox)과 정확히 맞물린다. 모델 가격이 0에 수렴할수록 사람들은 더 많은 토큰을 생성하려 하고, 그 모든 토큰은 결국 전기를 먹는 엔비디아 GPU를 거쳐야 한다. 즉, 소프트웨어의 범용화는 엔비디아에게 위협이 아니라 수요의 폭발이라는 말이다.

대담을 꼼꼼히 해체하면 젠슨 황이 평정심을 유지한 구간과 그렇지 않은 구간이 선명하게 갈린다. 그가 대개 여유로움을 느낀 주제는 "구글 TPU의 위협, 오픈소스 모델의 확산, 하이퍼스케일러들의 자체 ASIC"이었다. 엔비디아 CEO 입장에서 크게는 경쟁 위협을 느끼지 않는다는 방증일 수 있다.

그러나 그가 감정적으로 폭발한 하나의 구간은 '대중국 수출 통제'에 대한 얘기가 나왔을 때였다. 이는 엔비디아의 진짜 해자가 단일 칩 성능이나 CUDA 단독이 아니라 글로벌 시장 전체에 깔린 엔비디아 생태계의 '락인'이라는 사실, 그리고 그 락인을 정치권력이 수출 통제로 절단할 때 이 제국이 크게 흔들린다는 사실을 역설적으로 드러낸다.

물론, 젠슨 황은 대담에서 "미국 테크 스택 전체의 승리"라는 표현을 반복했다. 그가 진짜 말하고 싶었던 생각은 "글로벌 AI 컴퓨팅은 결국 엔비디아의 실리콘 위에서 작동할 것"이었을 가능성이 크다. 투자자 입장에서 이 속내를 곧이곧대로 받아들일 필요는 없지만, 그가 구축한 구조를 우회할 수 있는 길이 현재로서는 매우 좁다는 현실은 인정할 필요가 있다.

본 보고서는 이 대담을 포함해 현재 AI 업계의 몇 가지 전선(에이전틱 워크로드가 촉발한 CPU·메모리 구조 재편, 지정학이 강요하는 AI 생태계의 분기, 소프트웨어 런타임의 실리콘 복속 등)을 순서대로 들여다볼 것이다.

2. 에이전틱 AI와 CPU 르네상스: 인프라 경제학이 다시 쓰이고 있다

(1) 챗봇 시대의 죽음과 에이전틱 워크로드의 본질

챗봇 시대의 경제학은 이미 끝났다. ChatGPT가 전성기를 누리던 시절의 구조는 단순했다. 사용자가 질문을 입력하면 CPU는 토큰화와 일부 전처리만 담당했고, GPU가 연산의 90%를 독식했다. 당시 AI 인프라의 관심사는 "얼마나 큰 GPU 클러스터를 확보하느냐" 하나로 수렴했다.

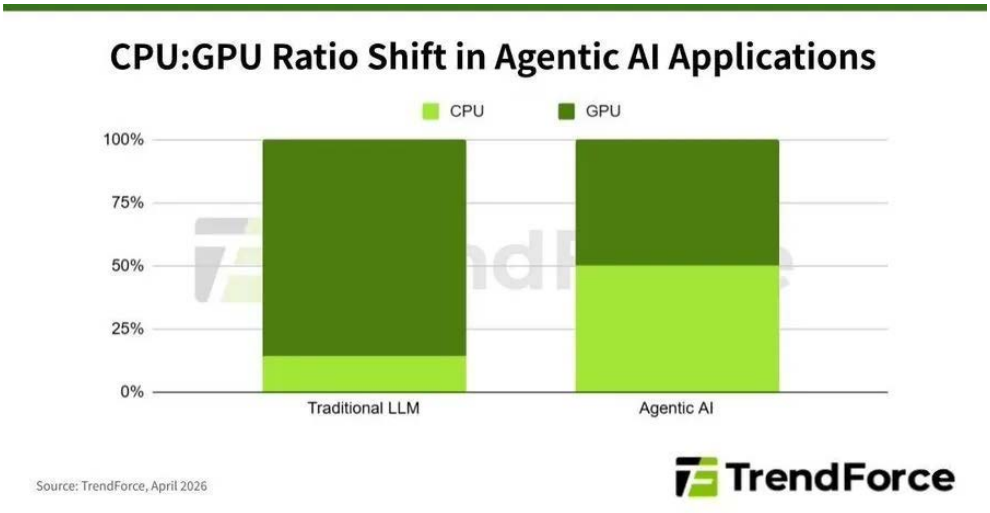
2026년의 에이전틱 AI는 이 구조를 송두리째 뒤집었다. 자율적으로 사고하고 행동하는 에이전트는 웹을 검색하고, 데이터베이스를 쿼리하고, API를 호출하고, JSON 및 HTML을 파싱해 수십 개의 하위 에이전트에게 작업을 분배한다. 이 일련의 네트워크 I/O와 파일 핸들링과 '직렬' 데이터 처리는 GPU가 절대 할 수 없는 순수한 'CPU의 영역'이다. 다시 말해, 에이전트 하나가 일할 때 랙당 500kW를 태우는 비싼 GPU가 CPU가 API 응답을 기다리는 동안 유휴 상태(Idle)로 놀고 있다는 뜻이다.

조지아텍과 인텔의 2025년 11월 연구는 이 기형적 병목을 수치로 증명했다. 에이전틱 워크플로우에서 전체 지연 시간의 50%에서 최대 90%가 CPU의 '도구 처리(Tool processing)'에서 발생한다. 비싼 GPU가 일하는 게 아니라 기다리는 시간이 압도적인 것이다.

(2) TrendForce와 TSMC가 증명한 CPU의 구조적 입지 전환

TrendForce가 2026년 4월 내놓은 차트는 이 전환을 단 한 장의 이미지로 압축한다. 전통적인 LLM 워크로드에서 CPU:GPU 자원 소비 비율은 약 1:7 수준이었다. 그러나 Agentic AI 워크로드에서는 이 비율이 거의 1:1로 수직 상승한다. 쉽게 말해, 1개의 GPU를 굴리기 위해 이제 1개의 CPU가 반드시 매칭되어야 한다는 말이다. GPU 중심 구조가 CPU와의 파트너 공생 구조로 재편된 것이다.

그림 2. 2026년 4월, Agentic AI 애플리케이션에서 CPU:GPU 자원 비율 변화
Traditional LLM 워크로드에서 약 1:7였던 비율이 Agentic AI에서 1:1으로 수직 상승



자료: TrendForce, 미래에셋증권 리서치센터

실리콘 벤더들의 실적은 이 구조적 전환을 날 것 그대로 반영하고 있다. AMD의 리사 수 CEO는 작년 4분기 실적에서 데이터센터 39% 성장의 핵심 동력으로 GPU가 아닌 EPYC CPU를 콕 집어 언급한 바 있다. 인텔은 137억 달러 매출에도 불구하고, 몰려드는 서버 CPU 수요를 감당하지 못해 PC용 CPU 웨이퍼를 강제로 서버용으로 돌리는 카니발라이제이션에까지 들어갔다. 그럼에도 2026년 1분기까지는 '절대적 공급 제한(Absolutely constrained)' 상태로 알려진다. 아마존 AWS의 자체 CPU인 Graviton 역시 두 대형 고객이 2026년 전체 캐파를 통째로 사겠다고 요청할 만큼 공급이 말랐다.

더 결정적인 증거는 OpenAI가 AWS와 맺은 380억 달러 딜이다. 시장은 이 계약을 '엔비디아 GPU 확보 계약'으로 해석했지만, 실제 본질은 에이전틱 워크로드 확장을 위한 "수천만 개(Tens of millions)의 CPU" 확보 조항이었다. TSMC 경영진조차 며칠 전 실적 발표에서 "오늘날의 AI 데이터센터에서 CPU는 점점 더(more and more) 중요해지고 있다"고 공식 선언했다. CPU 르네상스는 시장의 희망이 아니라 파운드리 최고 권위가 인정한 '팩트'다.

그렇다면 CPU를 더 찍어내면 되는 것 아닌가? 여기서 글로벌 반도체 공급망의 잔혹한 병목이 드러난다.

AMD의 서버 CPU는 TSMC의 3nm 공정에서 만들어진다. TSMC의 3nm 캐파는 월 16만 장 규모로 돌아가지만, 이 귀한 웨이퍼는 마진이 훨씬 높은 AI 가속기(GPU/커스텀 ASIC)에 최우선으로 할당된다. 결과적으로 고부가가치 GPU에 밀려 범용 CPU의 웨이퍼 할당이 깎여나가는 부수적 피해가 발생한 것이다.

여기서 진짜 중요한 것은 그 위의 한 겹이다. 엔비디아는 이미 1,000억 달러 이상의 현금을 꽂아 TSMC의 CoWoS 패키징 라인과 HBM 생산량의 90% 이상을 선점해 버렸다. KeyBanc의 채널 체크에 따르면, 엔비디아 확보 CoWoS 물량은 2026년 65만 개, 2027년 84만 개 수준이다. 2026년 65만 개의 CoWoS는 Blackwell 550~600만 개, Rubin 150만 개, Hopper 100만 개를 지원할 수 있는 물량이다. 리사 수의 AMD나 구글이 훌륭한 칩을 설계해도, 그것을 찍어낼 공장과 조립할 메모리가 이 지구상에 별로 남아있지 않다는 뜻이다. 즉, CPU 캐파 부족은 수요 폭발뿐만 아니라 엔비디아가 자본으로 공급망 상류를 점유한 결과 나타난 2차 파생 효과에 가깝다.

설상가상으로 메모리 가격 폭등은 패닉 바잉을 촉발했다. 고객들이 "시스템 가격이 더 오르기 전에 CPU라도 미리 사두자"는 Pull-forward 구매에 나서면서 IT 유통업체의 CPU ASP는 단일 분기에 30%나 폭등했다. 가격 발견 기능 자체가 망가지고 있다.

(3) 메모리: 625배 수요 폭주와 웨이퍼 3~4배 잠식

메모리 시장은 회복기가 아니다. 극단적 공급 부족과 공격적 가격 인상이 맞물린 전시 경제에 진입했다. CAPEX가 오르는데도 2027~28년까지 생산량이 늘기 힘든 구조적 이유가 있다. 먼저 HBM의 웨이퍼 캐파 잠식이다. HBM은 일반 DRAM과 같은 용량을 생산할 때 웨이퍼 캐파를 3~4배 더 잡아먹는다. HBM 수요가 늘수록 물리적 웨이퍼 아웃풋은 오히려 쪼그라드는 역설이 발생한다. 그리고 HBM4e, SOCAMM, LPDDR5, GDDR7, eSSD 같은 고부가가치 칩은 열 압착(Thermal Compression)이나 하이브리드 본딩(Hybrid Bonding) 같은 최첨단 패키징 공정이 도입되면서 제조 완료까지 걸리는 시간이 훨씬 길어졌다.

그 결과 새로운 팹이 지어져도 2028~30년까지 주요 칩메이커들의 연간 비트 그로스는 20% 미만(sub-20%)에 머물 것이라는 컨센서스가 형성되고 있다. 실제 신호들이 이 예측을 뒷받침한다. 2026년 2분기 DRAM 계약 가격은 전 분기 대비 50% 이상 상승이 예고되었고, NAND는 그보다 더 기이한 폭등을 준비하고 있다. 더 놀라운 기현상은 단종 수순이던 구형 공정 DDR3와 레거시 MLC NAND조차 DDR5 가격의 단기 조정을 뚫고 오히려 가격이 오르는 상황이다. 이것이 단순 사이클 상승이 아닌 가속기당 메모리 용량 25배에다 클러스터 스케일도 25배로 커지면서 총 메모리 수요가 625배(25x25) 상승하는 비탄력적 국면임을 증명하는 신호다.

3. 엔비디아 vs 중국: 지정학의 공학적 진실

(1) "Mythos는 평범한 연산력 수준"이라고 말한 젠슨 황! 수출 통제 큰 의미 없다?

대담 전체에서 젠슨 황의 평정심이 무너진 순간은 바로 이 질문에서 비롯되었다. 드와케시 파텔이 "중국에 칩을 파는 것이 사이버 공격 능력을 키워주는 것 아닌가. 1,000명의 해커와 100만 명의 AI 해커는 다르다. 북한에 핵미사일 껍데기를 파는 것과 무엇이 다른가"라고 밀어붙였을 때다. 드와케시가 핵무기 비유를 채 마치기도 전에 젠슨 황은 톤을 높이며 말을 잘랐다. "Not true. Not at all true. 형편없는 비유다(lousy analogy)."

젠슨 황은 곧바로 "Mythos 같은 10T 모델도 평범한(mundane) 컴퓨팅 용량으로 훈련되었다. 그 정도 용량과 종류의 컴퓨팅은 중국에 풍부하게(abundantly) 존재한다"고 받아쳤다. 우리는 이 문장은 어느 정도 마케팅 측면에 비약이 담긴 발언이라고 생각한다. 2026년 4월 기준 10T 파라미터 모델을 훈련하려면 최소 10만 개 이상의 B200 GPU가 단일 클러스터로 묶여야 하며, 이 규모에서는 MTBF(평균 고장 시간) 문제로 매일 수십 개의 GPU·케이블·스위치가 물리적으로 고장 난다. 실제로 Blackwell 세대로 들어와서는 Hopper 때보다 고장이 잦아, 칩이 온라인 된 이후 2~3주 만에 칩의 15% 정도가 고장난다는 현실적 통계도 있을 정도다. 또한, 10조개의 파라미터를 처리하는 데만 수십 TB의 메모리가 필요하고 '옵티마이저 상태'와 '활성화(activations) 데이터'까지 포함하면 최소 100TB 이상이 단일 훈련망에 동기화되어야 한다. 이것을 "평범하다"고 부르는 것은 비약에 가깝다고 사료된다.

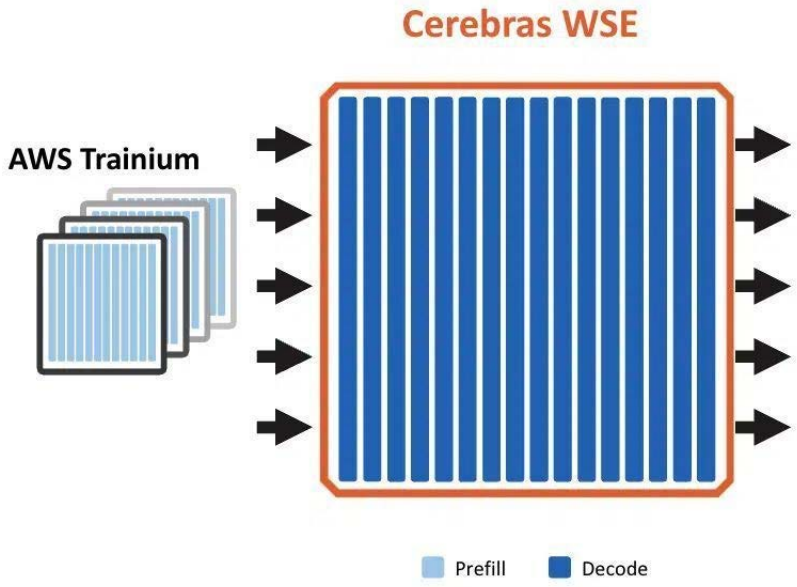
그렇다면 젠슨 황은 왜 이런 말을 했을까? 아마도 방어 기제일 수 있다. 만약 그가 "10T 모델 훈련은 미국의 10만 대 규모 Blackwell 슈퍼컴퓨터에서만 가능한 인류 최고의 비대칭 전략 자산이다"라고 인정하게 되면, 엔비디아의 GPU는 즉각 '핵무기와 동급의 전략 물자'로 재분류된다. 칩이 전략 물자가 되는 순간 미국 정부는 수출 통제뿐만 아니라 칩의 배급 권한, 마진율 통제, 나아가 설계 기술의 국가 귀속까지 요구할 수도 있다. 그렇다면 엔비디아가 장기적으로 추구하는 "70% 매출총이익률"은 국가에 의해 조정당할 위험이 크다. 그래서 그는 여유로운 척하며 "그 정도 컴퓨팅 규모는 중국에서도 평범한 수준"이라고 대중과 정치인들을 향해 프레임을 친 것으로 보인다.

(2) 황의 논리: 엔비디아의 SW 해자는 강력하니 중국에서 이걸 포기할 필요가 없다

또 하나 중요한 지점이 있다. 엔비디아의 소프트웨어 해자를 CUDA라는 단일 요소로 보는 것은 구태의연한 인식이라는 점이다. 왜냐하면, 2026년의 엔비디아는 경쟁사의 이기종 아키텍처마저 자신의 소프트웨어로 강제 흡수하는 단계에 진입했기 때문이다.

가장 소름 돋는 사례는 AWS Trainium(Prefill)과 Cerebras WSE(Decode)의 이종 클러스터다. 두 회사는 엔비디아 하드웨어를 배제하기 위해 협력했지만, 그들의 KV캐시를 칩 간에 전송하기 위해 여전히 엔비디아의 NIXL 소프트웨어와 AWS EFA를 사용해야만 한다. 엔비디아 하드웨어를 피해 가려는 시도조차 엔비디아의 소프트웨어 생태계 안에서 이루어지고 있는 것이다.

그림 3. AWS Trainium × Cerebras WSE 이종 클러스터 (Prefill/Decode 분리)
AWS Trainium이 Prefill(사전 채우기, 연산 바운드)을 담당하고
Cerebras WSE가 Decode(디코드, 메모리 대역폭 바운드)를 담당하는 Disaggregated Serving.
엔비디아 하드웨어 없이 구현한 '반(反)엔비디아 연합'의 상징적 구성도.
그런데 이 두 칩 사이에서 KV 캐시를 주고받는 순간, 이들은 여전히 엔비디아의 NIXL 소프트웨어와 AWS EFA를 거쳐야만 한다. 하드웨어는 피해갔지만 소프트웨어 그물은 벗어나지 못한 것이다…



자료: Amazon, 미래에셋증권 리서치센터

이 사례가 중요한 이유는, 미국 내 최대의 하이퍼스케일러(AWS)와 가장 대담한 XPU 도전자(Cerebras)조차 엔비디아의 소프트웨어 그물을 벗어나지 못한다는 점에 있다. 하물며 SMIC의 공정 한계와 자체 IP 부재에 시달리는 중국의 화웨이가 이 그물을 빠져나갈 수 있을까? 젠슨 황이 드와케시 파텔에게 "우리는 자동차가 아니다(We are not a car). 컴퓨팅 생태계는 교체하기가 엄청나게 어렵다. x86이 여전히 존재하는 이유, Arm이 그토록 끈끈한(sticky) 이유가 있다"고 못 박은 것은 바로 이 구조를 가리킨 것이다.

그러니까 "중국이 지금 엔비디아 칩을 쓰는 한, 그들은 우리 문법으로 사고하고 우리 툴로 개발한다. 그런데 우리가 그들을 밖으로 밀어내면, 언젠가 우리와 경쟁하는 별도의 문명이 만들어진다"는 말이다. 수출 통제에 대한 반론이자, 동시에 생태계 락인의 본질을 설명한 것으로 보면 된다.

(3) 수출통제는 효과가 있었고 그 증거는 DeepSeek의 지연

피터 윌드포드(Peter Wildeford)와 같은 저명한 미래학자는, 그가 가진 지정학적 리얼리즘으로 젠슨 황의 논리를 구조적으로 해체한다. 윌드포드가 먼저 짚는 부분은 이 점이다. 중국 정부/공산당의 의지에 따르면, CUDA가 풀리든 안 풀리든 무조건 CANN(Compute Architecture for Neural Networks, 화웨이판 CUDA) 등 자체 생태계로 갈 수밖에 없다는 점이다. 이건 변수가 아니라 상수라는 게 그의 입장이고 필자도 이에 동의한다.

중국이 엔비디아 생태계에 머물렀던 이유는 엔비디아에 매료되어서가 아니라 화웨이 Ascend와 CANN이 완성될 때까지 시간을 벌기 위한 전술적 임시방편이었을 뿐이다. 태양광·전기차에서 중국이 반복한 플레이북이 AI에도 적용된다는 말이다. 즉, "외산으로 초기 시장 구축 → 자국 공급망 임계점 도달 → 보조금과 규제로 외산 축출"이 AI 가속기에서도 그대로 재생될 것이라는 분석이다.

그리고 현재의 현실은 미국의 수출 통제가 작동하고 있음을 보여준다. DeepSeek V4와 R2의 출시가 실제로 꽤 많이 지연되고 있기 때문이다. 모델 체급이 1T를 넘어 멀티 Trillion 단위로 넘어가면, 아무리 알고리즘을 기가 막히게 깎아도 순수한 물리적 네트워킹 대역폭과 클러스터의 깡성능(Brute-force FLOPs)이 필요해지기 때문이다.

게다가 엔비디아의 진짜 해자는 단일 GPU 성능이 아니라 NVLink와 NVSwitch를 통한 무손실 스케일업(Scale-up)이다. 중국은 미국의 제재로 인해 인피니밴드(InfiniBand) 급의 네트워킹 장비와 이더넷 스위치(브로드컴의 Tomahawk 등) 확보에 치명적 제약을 받고 있다. 모델 파라미터가 5T 수준으로 커지면 수만 개의 최신 칩이 훈련 중 끊임없이 파라미터와 그래디언트를 주고받아야 하는데, 화웨이 칩을 아무리 많이 때려 박아도 칩 간 데이터 전송 속도가 느리면 전체 클러스터는 연산을 멈추고 데이터를 기다리는 대기 상태(Idle state)에 빠진다. 이것이 현재 DeepSeek를 포함한 중국의 AI 업계가 직면한 허들이며, 화웨이의 AI DC 시스템인 CloudMatrix 384가 엔비디아의 GB200 NVL72를 따라잡으려 하지만 MoE All-to-All 네트워킹에서 여전히 뒤처지는 이유다.

(4) 젠슨 황을 옹호하는 사람들과 그에 대한 반론: Ascend는 TPU와 다르다

물론, 필자와 윌드포드의 의견에 일리 있는 기술적 반론도 존재한다. 이와 관련해 Gavin Baker는 미국과 중국의 네트워킹 토폴로지가 '분기'하고 있다고 분석했다. 미국은 전력이 부족해 구리선(NVLink) 중심의 스위칭 구조로 갔고, 중국은 전력이 넘쳐나 광학 기반의 토러스 중심의 광학 스케일업으로 갔다는 것이다.

따라서 Gavin Baker를 포함한 이들의 결론은, 엔비디아와 다른 경쟁사들(구글 등)의 토폴로지가 다르니 결국에는 '모델 이식성'이 붕괴할 것이라는 말이다. 여기서 모델 이식성이 붕괴된다는 말은 엔비디아에 최적화되어 훈련된 모델은 계속해서 엔비디아의 칩에서만 활용돼야 한다는 것을 의미한다. 따라서 구형 GPU라도 중국에 계속 팔아놓고 'CUDA 종속성'을 유지하는 것이 미국의 국가 안보에 오히려 유리하다는 주장이다. 이는 역설적으로 젠슨 황의 수출 통제 반대 논리를 공학적으로 옹호해 주는 논거가 된다.

그러나 여기에는 중대한 맹점이 있다고 생각한다. 이들에 대한 재반론을 펼쳐보면, 우선, 화웨이와 중국 반도체 생태계는 구글 TPU가 나아가는 방식과 다르다. TPU가 엔비디아와 이질적인 길을 걷고 있다면(네트워킹 토폴로지가 완전히 다름), 중국의 방식은 그렇지 않다. 그들은 오히려 철저하게 '엔비디아의 완벽한 복제품(China-version Nvidia)'이 되기 위해 노력하고 있다.

화웨이가 무지막지하게 거대한 광학 스케일링을 하는 이유는 독창적인 아키텍처 철학 때문만은 아니다. SMIC의 5nm(N+3) 공정 한계로 단일 칩 성능이 H100의 60~70%에 머물다 보니, 엔비디아 1랙(72개)의 성능을 내려면 150개 이상 묶어야 했기 때문이다.

애초에 150개로 스케일업을 하려면 물리적 거리가 멀어져 구리선으로는 신호를 보낼 수 없고, 게다가 엔비디아식 224G SerDes IP가 없는 화웨이는 어쩔 수 없이 광학 부품을 랙 내부에 많이 붙여놓을 수밖에 없는 것이다. 이를 두고 "아키텍처의 분기"라고 부르는 것은, 뒷다리가 불편한 늑대가 앞다리로만 기어가는 모습을 보고 "새로운 진화 트리"라고 분석하는 것과 같다.

더 결정적인 증거는 화웨이의 새로운 Ascend 변형 칩(950PR)이 엔비디아의 Disaggregated Serving 아키텍처를 정확히 복제했다는 사실이다. 고정된 MXU(행렬곱셈 유닛) 크기를 키우는 구글 TPU식 특화 칩이 아니라, 추론 워크로드의 동적인 변화에 맞춰 칩 역할을 분리(Prefill/Decode 분리) 하는 범용 GPU 클러스터링을 택한 것이다. 중국의 스타트업인 Biren과 Moore Threads 같은 반도체 팹리스들도 내부 SM(Streaming Multiprocessor) 구조까지 엔비디아를 그대로 벤치마킹하고 있다. 이들의 논리는 간단하다. "세상의 대부분 AI 알고리즘(오픈소스)이 엔비디아의 구조에 맞춰 진화하고 있으니, 우리의 하드웨어도 일단은 엔비디아와 똑같이 생겨야 한다."

여기에 화웨이에서 발표한 "CANN Next" 컴파일러의 추상화로써 생기는 변화까지 고려해야 한다. CANN Next는 중국 개발자들이 엔비디아용으로 짜던 코드와 익숙한 방식을 화웨이 칩에서도 최대한 비슷하게 돌릴 수 있게 만드는 번역기라고 할 수 있다. 이로써, 엔비디아와 중국 업체들의 '기술적 수렴(convergence)'은 더 빨라진다. 실제로 화웨이 CANN Next는 지난 수년간 텐센트·바이두·DeepSeek 엔지니어들을 투입해 엔비디아의 cuBLAS·cuDNN과 1:1 매칭 작업을 해왔다. 개발자들이 엔비디아에서 화웨이로 옮겨가는 난이도가 점점 낮아진다는 뜻이다. 화웨이가 cuBLAS·cuDNN 대응 계층을 정교하게 쌓을수록, 중국 내부에서는 “엔비디아가 아니어도 되는 순간”이 결국에는 점점 늘어날 것이다.

게다가 DeepSeek V4가 출시를 코 앞에 둔 시점에서 "Ascend 칩에서 MFU 85%를 달성했다"는 중화권에서 발생한 루머 소식은 이 추상화의 깊이를 보여준다. 즉, 결국 중국의 모델과 칩은 엔비디아를 중심축으로 삼아 그 궤도를 함께 도는 동기화된 진화(Synchronized Evolution)를 겪는다는 말이다. 중국은 갈라파고스에 갇히는 것이 아니라 엔비디아가 깔아놓은 레일 위를 달리는 투박하지만 가장 위협적인 복제 열차를 완성해 가고 있다.

(5) 시간적 격차(Temporal Delta)... 그 자체가 우위다

엔비디아와 중국 업체들의 기술적 수렴이 일어난다고 해서 미국의 수출 통제가 의미 없는 것은 아니다. 오히려 정반대다.

중국이 엔비디아를 극복하고 결국에 자체 생태계를 구축하는 것이 어차피 일어날 '상수'라면, 수출 통제는 유일한 방어 수단이 된다. 특히 긴박하게 변화하는 AI 경쟁에서는 1년의 격차가 10년의 경제적·군사적 격차로 벌어진다.

미국 AI 기업들이 Blackwell의 압도적 스위칭 대역폭 위에서 5T~10T 모델을 여유롭게 실험할 때, 중국은 구형 칩의 통신 병목을 풀기 위해 연구 역량을 낭비하고 있다. DeepSeek V4·R2의 출시 지연이 바로 이 시간 지연(Temporal Delta)의 증거다. 미국의 수출 통제는 중국의 자립을 막는 정책이 아니라, 그 자립의 속도를 늦추는 정책으로 이해하는 편이 더 정확하다.

그림 4. Q4 2025 기준 약 110만 H100e로 글로벌 전체의 5% 수준

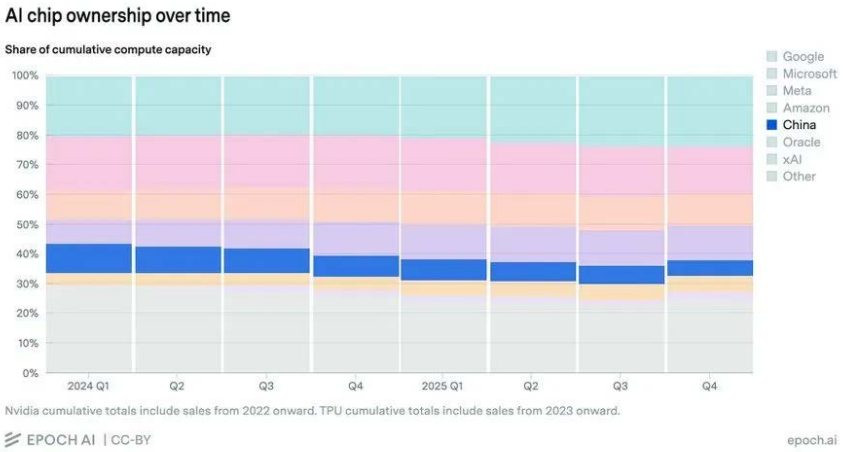
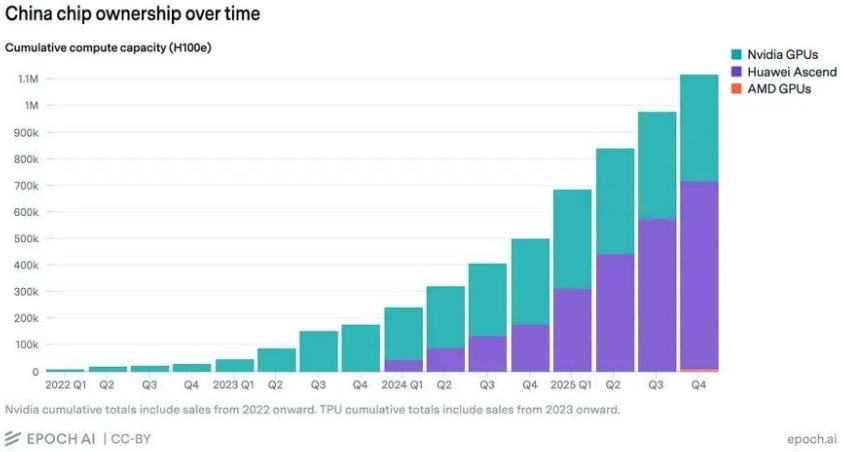


그림 5. 2022 Q1~2025 Q4 중국의 누적 AI 컴퓨팅 캐파(H100e 환산).
Ascend가 약 70% 비중으로 엔비디아를 제치고 1위로 올라섰지만, 전체 파이가 쪼그라든 상태



자료: Epoch AI, 미래에셋증권 리서치센터

Epoch AI 데이터가 이 구조를 뒷받침한다. 2022 Q1부터 2025 Q4까지 중국의 누적 AI 컴퓨팅은 약 110만 H100e 수준으로, 글로벌 전체 AI 컴퓨팅의 5%를 간신히 넘는 수준에 그쳐 있다. 같은 기간 화웨이 Ascend가 중국 내 점유율 70%까지 치고 올라오며 엔비디아를 제쳤지만, 이는 전체 파이 자체가 쪼그라든 결과다.

젠슨 황이 대담에서 걱정한 "중국의 갈라파고스화"는 실제로 어느정도 발생하고 있지만, 엔비디아에게 치명적인 방식으로는 아니다. 이 갈라파고스는 글로벌 AI 인프라 파이의 5%에 갇혀 있고, 나머지 95%에서는 엔비디아의 CUDA 제국이 여전히 지위를 굳히고 있다.

엔비디아가 가진 중국 업체들에 대한 해자의 본질은 “절대 못 따라온다”가 아니라 “따라오더라도 늦게 따라온다”는 시간 우위에 있다. 글로벌 사우스(인도·중동·아프리카·동남아)가 화웨이가 제공하는 턴키 AI DC 솔루션으로 편입되는 시점은, 젠슨 황이 우려한 것보다는 더 느려질 가능성이 높다. 이것이 2026년 4월 미중 AI 경쟁의 공학적/지정학적/지경학적 진실이라 사료된다.

4. Anthropic의 Managed Agents와 'Deep SaaS' 점검

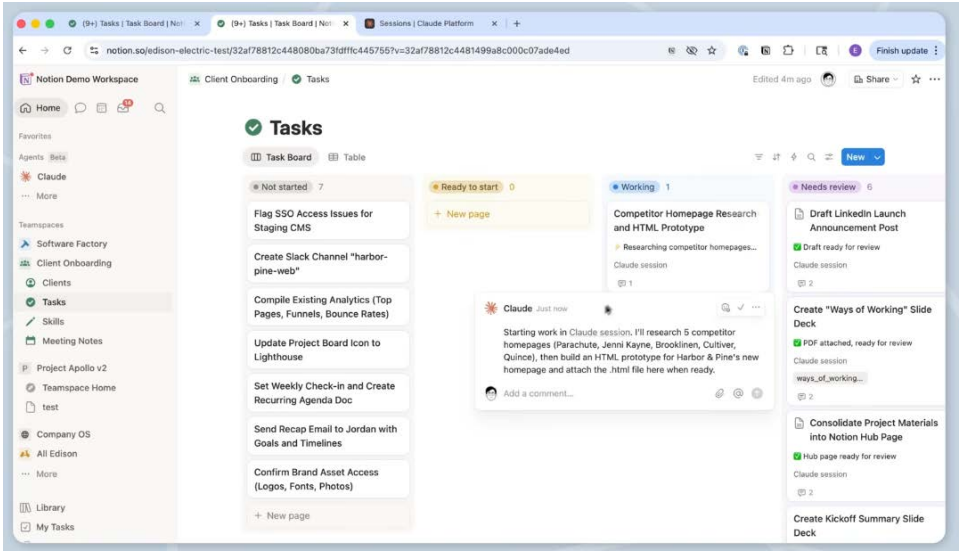
(1) Managed Agents의 commodity화: 에이전트 런타임이 인프라가 되다

지난 4월 8일, Anthropic은 Claude Managed Agents를 public beta로 공개했다. 공식 문서에서는 이것을 "stateful sessions with persistent event history를 제공하는 fully managed agent infrastructure"로 정의한다. 쉽게 말해, 이 Managed Agents를 도입한 SaaS 기업들은, '이제 스스로 agent loop를 짜고, tool execution을 만들고, 런타임을 운영하고, 세션 복구와 이벤트 로그를 관리'하는 인프라 공사를 처음부터 다시 하지 않아도 된다.

그냥 Anthropic이 알아서, 세션 런타임·툴 실행·컨테이너·이벤트 히스토리까지 대신 관리해 주는 것이다. 좀 더 엄밀히 말하면, Anthropic은 이 스택을 agent(모델·시스템 프롬프트·툴·MCP 서버·skills의 묶음), environment(컨테이너 템플릿), session(append-only 이벤트 로그), harness(Claude와 도구를 연결하는 실행 루프), sandbox(코드 실행과 파일 편집이 일어나는 격리 환경)라는 5개의 인터페이스로 분리해 상품화했다. 과거에는 SaaS 팀이 직접 코드로 접착해야 했던 런타임 레이어를, 이제는 벤더가 공용 인프라처럼 제공한다는 뜻이다.

이 상품화의 방향성은 명확하다. Managed Agents는 '에이전트 인프라의 AWS' 같은 역할을 노리고 있다. 2000년대에 AWS가 서버·네트워크·스토리지 같은 물리 인프라를 API로 추상화하면서 수많은 SaaS가 '인프라를 소유하는 회사'에서 '제품 경험을 파는 회사'로 전환됐던 것과 같은 구도가, 이번에는 에이전트 레이어에서 반복되고 있다. Anthropic 공식 사례 페이지를 보면, 해당 서비스의 홍보 협력업체로 등장한 Notion이 나온다.

그림 6. 노션 Task Board 안에서 Claude가 작업 중인 화면
Claude는 백그라운드 실행 주체이고, Notion은 업무 상태·검토·협업이 일어나는 UI 레이어가 된다

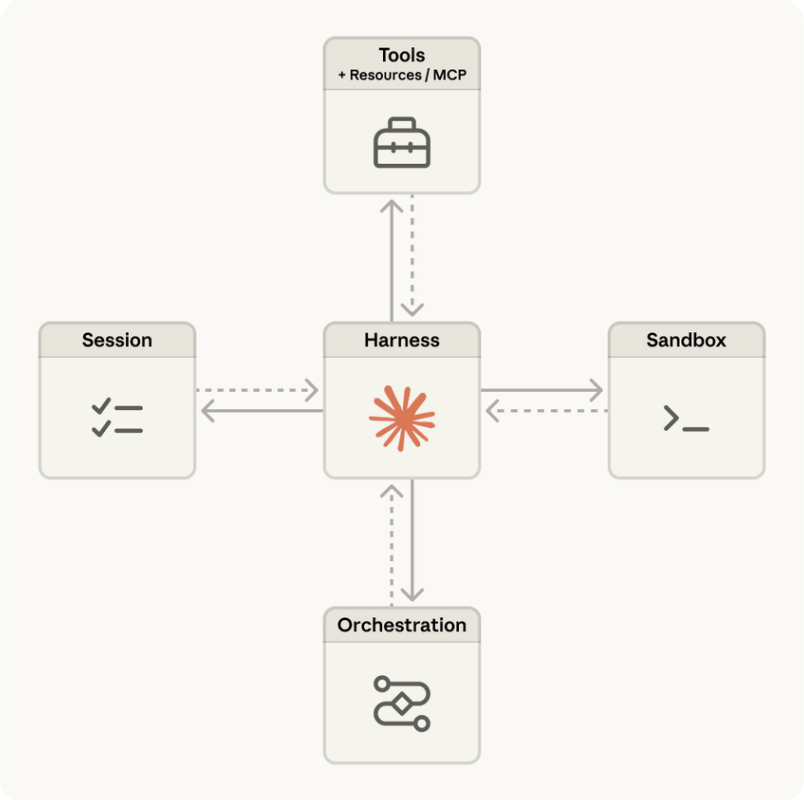


자료: Notion, Anthropic, 미래에셋증권 리서치센터

Notion은 단일 task board에서 30개 이상의 동시 에이전트 작업을 실행하고, 결과를 리뷰 담당자에게 라우팅하며, 코드·PPT·PDF·웹사이트 같은 산출물을 만들게 한다. 여기서 Notion이 보여주는 Managed Agents 사례의 본질은 "사람-에이전트 협업 UI는 SaaS 기업들이 쥐고, Anthropic이 에이전트 런타임을 맡는 분업" 구조가 된다. 분업이라고 불리지만, 실질적인 교섭력의 무게추는 후자 쪽으로 많이 기울어져 보이는 게 사실이다. 그리고 이 분업이 SaaS 업계에 갖는 함의는 일단, 에이전트 인프라가 빠르게 범용화되어서 차별화 포인트가 아래가 아니라 위로 올라가게 된다는 점이다. 즉, 모델이나 런타임이 아니라 UI, approval flow, retrieval, workflow routing, human review, artifact handoff, domain context가 해자가 된다.

그러나 더욱 중요한 의미는 바로 이것이다. "에이전트를 돌리는 능력"과 "에이전트 오케스트레이션" 그 자체는 더 이상 해자가 아니라는 것이다. 많은 SaaS가 "우리는 에이전트를 위한 멀티스텝 실행을 관리한다"는 말을 해자처럼 포장할 수 있었지만, 그 주장은 빠르게 약해진다. 너무나 빠르게 변화하는 시장이다. 한 달 전만 하더라도 하네스 엔지니어링이 핵심이라고 부각됐는데, 벌써 그 논리가 조금 희석되고 있는 세상이다. 이제 이 층은 Anthropic·OpenAI 같은 런타임 벤더가 공용 인프라처럼 제공하기 시작했다. AI 모델 제공업체가 이렇게 미들웨어로 본인들의 영역을 확장하는 것은 사실 오늘내일 일은 아니다.

그림 7. Claude Managed Agents 공식 아키텍처 다이어그램
중앙의 Harness가 Session(append-only event log), Tools + MCP(외부 기능 호출 수단), Sandbox(격리된 코드 실행 환경), Orchestration(순서·분기·의존성 조율)을 다 엮이는 구조
각 회사의 코더들이 직접 짜야 했던 이 5개 레이어를 Anthropic이 단 하나의 관리형 런타임으로 '상품화'했다는 점이, 이번 Managed Agents 출시의 본질이다!

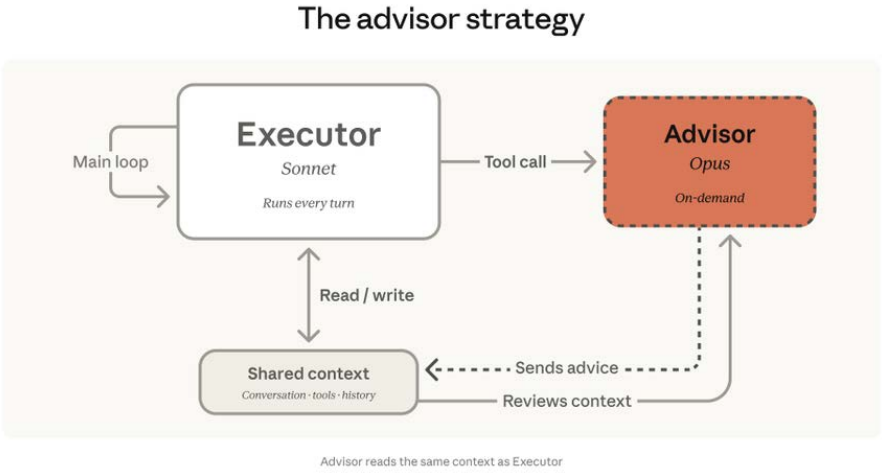


자료: Anthropic, 미래에셋증권 리서치센터

어찌 됐든, 충격의 직접 피해자는 명확하다. 얇은 UI(유저 인터페이스) 기반의 SaaS, seat 기반 과금 모델, LLM 래퍼형 제품이다. 사람이 화면을 들여다볼 필요가 줄어들수록 Engagement를 위한 화려한 인터페이스는 무용지물이 된다. 팔란티어의 CEO 알렉스 카프의 말처럼 시장은 점점 '기능' 그 자체보다는 '결과'와 '책임'을 사기 시작하고 있다. "에이전트를 붙이면 된다"는 수준의 AI-washing은 더 빠르게 들통날 것이다.

그런데, Anthropic이 발표한 Managed Agents의 강력함에는 구조적 대가가 따른다. 비용 구조가 그것이다. Anthropic은 Managed Agents를 토큰 비용 외에 running 상태의 session runtime에 대해 '시간당 $0.08'을 과금한다. 단일 에이전트로는 싸 보여도, Notion처럼 수십 개를 병렬 fan-out으로 돌리면 토큰 비용에 런타임 비용까지 동시에 증가한다. 그래서 앞으로 SaaS의 실력은 '에이전트를 단순히 많이 띄우는 것'이 아니라 언제 병렬로 돌리고 어떻게 모을지, 어느 단계에서 인간 승인을 넣을지를 설계하는 능력에서 갈린다. 오케스트레이션 기법에 더 신경써야 한다는 것이다. 이제 "무조건 병렬화"는 제품 전략이 아니라 비용 폭탄이 될 수 있다. "에이전트 비용 통제"가 이제 코더들의 실력이 판가름 나는 영역이 아닐까 하는 생각도 든다.

그림 8. Managed Agents의 세계관 안에서 “실행”과 “고차 판단”을 분리한 어드바이저 패턴
Executor(Sonnet)는 매 턴마다 실제 작업을 수행하는 주력 실행자.
Advisor(Opus)는 항상 서 있는 모델이 아니라, 어려운 분기점에서만 불려 나오는 상위 추론 엔진.
Shared context는 대화, 툴 사용 내역, 작업 히스토리를 함께 쌓아두는 공용 작업 문맥.
Advisor는 이 같은 문맥을 읽고 계획이나 조언만 다시 넘겨준다.
Managed Agents는 결국, 단일 모델이 혼자 다 하는 구조가 아니라,
하네스가 여러 모델 역할을 조합해 성능·비용을 최적화하는 아키텍처.



자료: Anthropic, 미래에셋증권 리서치센터

그런데 여기서 한 가지 중요한 지점을 놓치면 안 된다. Managed Agents가 강력하지만, Anthropic이 제공하는 session은 결국 append-only event log이자 context object에 머문다. 즉 업무의 연속된 기록이지, 기업의 canonical(진정한 표준이 되는) world model(기업을 완전한 디지털 트윈으로 만드는 모델)은 아니다.

에이전트가 많아질수록 SaaS는 일을 그냥 말로 대충 시키는 수준으로는 운영하기 어려워진다. 누가 어떤 일을 맡았는지, 지금 어디까지 승인됐는지, 다음 단계는 무엇인지 같은 것을 '명확한 규칙과 상태표'로 관리해야 한다. 그래서 Managed Agents가 퍼질수록 각 SaaS는 결국 자기 서비스 안에, '작은 온톨로지 같은 구조'를 만들게 될 가능성이 크다. 에이전트를 돌리는 엔진 자체는 점점 평준화되겠지만, "업무 상태와 객체를 어떻게 정의하고 연결하느냐"는 각 회사가 직접 가져가야 하는 핵심 자산이 된다는 뜻이다. 바로 그 지점에서 다음 논의가 시작된다.

(2) 팔란티어와 시놉시스: 에이전트가 접속할 '현실'을 쥔 자

Managed Agents 이후 남아야 할 질문은 좁아졌다고 판단한다. "누가 에이전트를 돌리는가"가 아니라, "누가 에이전트가 건드릴 현실을 쥐고 있는가"가 더욱 더 중요해졌다. 에이전트 런타임이 commodity화될수록, 에이전트가 접속할 현실의 데이터·권한·검증 엔진은 상대적으로 더 희소해진다. 확률적으로 작동하는 모델(LLM)이 현실 세계를 더 자주 만질수록, 그 위에 놓인 운영 의미체계와 통제 레이어의 '결정론적' 가치가 기하급수적으로 커지기 때문이다.

이 관점에서 우리가 지난 2월에 작성했던 "AI Bi-weekly: SaaS 스트레스 많이 받을거야" 리포트에서 강조했던, 팔란티어와 시놉시스의 해자는 한 층 더 깊은 곳에 있다고 사료된다. 한쪽은 System of Action, 다른 한쪽은 System of Truth다. 두 축 모두 Managed Agents가 침범하기 어려운 레이어를 장악하고 있다.

먼저, 팔란티어는 System of Action에 가깝다. 즉, 단순히 데이터를 보여주는 시스템이 아니라 실제 업무의 의사결정을 움직이는 시스템에 가깝다. 팔란티어의 온톨로지는 그냥 “데이터가 어디에 있나”를 정리하는 카탈로그가 아니다. 오히려 회사의 디지털 자산 위에 올라간 운영 레이어이자, 현실 조직을 소프트웨어 안에 옮겨놓은 디지털 트윈에 가깝다.

여기에는 객체와 속성, 객체 간 연결 같은 의미론적 요소만 있는 것이 아니다. 어떤 행동이 가능한지, 누가 무엇을 바꿀 수 있는지, 어떤 조건에서 변경이 허용되는지 같은 실행 규칙까지 함께 들어 있다. 즉 팔란티어는 “회사 안에 무엇이 존재하는가”를 넘어서, “그것이 누구 권한 아래에서 어떤 절차를 거쳐 어떻게까지 바뀔 수 있는가”를 한 체계 안에서 다룬다.

이 지점이 Managed Agents와 팔란티어가 서 있는 층위가 다르다는 사실을 잘 보여준다. Managed Agents는 기본적으로 프롬프트와 툴 호출을 중심으로 돌아가는 시스템이고, 팔란티어는 객체·행동·권한 모델을 중심으로 돌아가는 시스템이다.

전자는 세션 단위로 움직이고, 후자는 조직 전체가 공유하는 객체 그래프 위에서 움직인다. 전자의 핵심은 “에이전트가 도구를 잘 불렀는가”이고, 후자의 핵심은 “상태 변화가 정합적으로 일어났는가, 그리고 그 변경이 안전하게 실제 시스템에 반영됐는가”이다. 다시 말해 Managed Agents가 “Claude를 오래도록, 안정적으로, 잘 작동하게 만드는 법”이라면, 팔란티어는 “그 Claude가 현실 세계의 어떤 객체를 사실로 인식하고, 어떤 권한으로, 어떤 조건에서, 무엇을 바꿀 수 있는가”를 다루는 시스템이다. 레벨이 다르다.

여기서 중요한 오해 하나를 걷어내야 한다. 팔란티어가 강한 이유는 '에이전트 회사'여서가 아니다. 본질적으로는 '기업용 상태 기계(state machine)를 가진 회사'이기 때문에 강한 것이다. 이 차이는 매우 크다. Managed Agents가 퍼지면 팔란티어가 자랑하던 일부 기능, 예를 들어 에이전트 연결, 복잡한 오케스트레이션 시연, 자연어 인터페이스 같은 것은 빠르게 평준화될 수 있다.

하지만 팔란티어의 진짜 해자는 원래 거기에 있지 않았다. 핵심은, "(에이전트가 쉽게 내 회사가 무슨 회사인지 잘 이해하도록 하는) 명확한 세계 모델, (에이전트가 권한 밖의 행동을 하지 못하도록 하는) 통제된 액션과 촘촘한 권한 구조, (인간의 승인이 시의적절하게 포함되고, 에이전트 피드백을 위한) 감사 가능성, 그리고 (입력이 같으면 언제나 같은 옳은 정답을 내는) 결정론적인 write-back"에 있다.

오히려 에이전트가 많아질수록 이 해자의 중요성은 더 커진다. 태생적으로 확률론적으로 움직여야 하는 AI 에이전트가 현실 세계를 더 자주 건드릴수록, 그 위에서 상태와 권한을 구조적으로 묶어주는 운영 의미체계의 가치가 훨씬 커지기 때문이다. 한마디로 말해, 에이전트를 돌리는 런타임이 범용화될수록 팔란티어의 진짜 가치, 즉 현실 세계를 안전하게 바꾸는 상태 모델의 가치가 더 또렷하게 드러난다.

한편, 시놉시스는 'System of Truth'에 가깝다. 반도체 설계는 물리 법칙, 양자역학, 공정 미세화, 신호 무결성처럼 절대 어길 수 없는 제약 위에서 돌아가는 세계다. 여기서는 LLM이 흔히 보여주는 "그럴듯하지만 틀린 답"이 통하지 않는다. 칩 설계는 1nm만 어긋나도 수천억 원이 날아갈 수 있는 영역이기 때문이다. 결국 중요한 것은 결과가 정말 tapeout-ready한지, 물리적으로 구현 가능한지 여부다.

그래서 AI 에이전트가 수백 개의 설계 가설을 세우고 수천 개의 플로우를 병렬로 돌릴수록, 마지막에 더 자주 호출되는 것은 채팅 인터페이스가 아니라 verification 및 simulation engine이다. 이 점에서 시놉시스는 얇은 SaaS라기보다, 공학적 진실을 최종 판정하는 Deep SaaS라고 할 수 있다. 다시 말해 Managed Agents는 시놉시스에 위협이라기보다 오히려 workload를 더 키워주는 증폭기 역할을 할 가능성이 크다.

실제로 시놉시스는 이 흐름을 방어적으로만 보는 회사가 아니다. 이미 agentic AI를 자사 스택 안으로 적극적으로 흡수하고 있다. AgentEngineer는 reasoning, planning, learning, execution을 수행하는 멀티에이전트 엔지니어링 워크플로로 소개되고 있고, 시놉시스는 EDA agentic AI를 전자설계 전반에 걸친 자율적이고 다단계적인 공학 작업으로 정의한다. 즉 AI가 자신들을 대체하는지를 고민하는 단계가 아니라, AI를 이용해 검증·설계 엔진을 더 많이 돌리게 만들자는 쪽으로 이미 방향을 잡은 것이다. 이 구조에서 시놉시스는 UI를 파는 회사가 아니라, 에이전트가 끝내 참조할 수밖에 없는 진실 판정기를 쥔 회사다.

정리하면, 팔란티어의 해자는 "에이전트를 잘 붙인다"는 데 있는 것이 아니라, 에이전트를 붙여도 현실의 운영 세계가 망가지지 않도록 만드는 데이터·행위·권한 구조에 있다. 시놉시스의 해자도 "AI를 잘한다"는 데 있는 것이 아니라, AI가 끝내 의존해야 하는 물리적 정답지와 검증 엔진을 쥐고 있다는 데 있다.

두 회사는 각각 엔터프라이즈 운영과 공학 설계라는 전혀 다른 영역에 서 있지만, 결론은 하나로 모인다. 에이전트가 싸게 복제될수록, 그 에이전트가 접속해야 하는 현실의 데이터, 권한 체계, 검증 엔진은 오히려 더 희소해지고 더 비싸진다.

따라서 Anthropic의 Managed Agents처럼 얇은 런타임 계층이 빨리 상품화될수록, Deep SaaS의 본질은 오히려 더 선명하게 드러난다. 껍데기가 벗겨질수록 알맹이의 가치는 더 빛나 보인다.

II. AI & Reports

1. 능력 곡선은 멈추지 않아

(1) 작년의 나는 잊어라: 희망편

2026년 4월 13일, Stanford 인간중심 AI 연구소(HAI)가 2026 AI Index Report를 공개했다. 9년째 발간되는 이 보고서는 산업계 PR 예산과 무관한 독립 데이터로 구성되며, AI가 실제로 어디쯤 와 있는지를 가장 신뢰도 높게 측정하는 연간 문서 중 하나로 평가받는다.

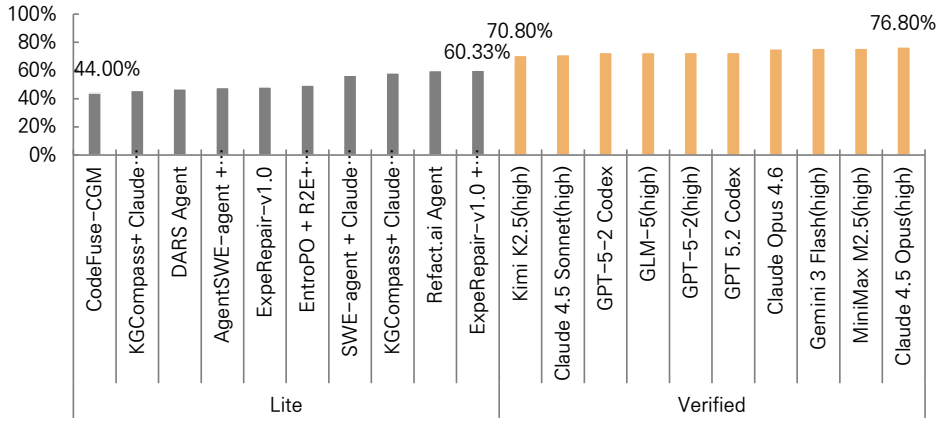
올해 보고서의 헤드라인은 명확하다. AI의 능력 곡선은 꺾이지 않고 있다.

코딩 벤치마크인 SWE-bench Verified에서 AI의 성능은 불과 1년 만에 60%에서 인간 기준선(human baseline)에 근접한 수준으로 도약했고, 최고 난도의 학문적 문제를 다루는 Humanity's Last Exam(HLE)에서 최상위 모델의 정답률은 2025년 8.8%였다. 2026년 4월 기준 이 수치는 38.3%를 거쳐 50%를 넘어섰다. 인간 전문가가 겨우 풀어내는 문제들을 AI가 절반 이상 맞히기 시작한 것이다.

채택 속도도 역대 기록을 갈아치웠다. Generative AI는 출시 3년 만에 전 세계 인구의 53% 채택에 도달했다. 개인용 컴퓨터(PC)도, 인터넷도 이 속도에 미치지 못했다. 기업 조직 차원에서도 테크 업계 채택률은 88%에 달했고, 미국 대학생의 80% 이상이 이미 학업에 AI를 활용하고 있다.

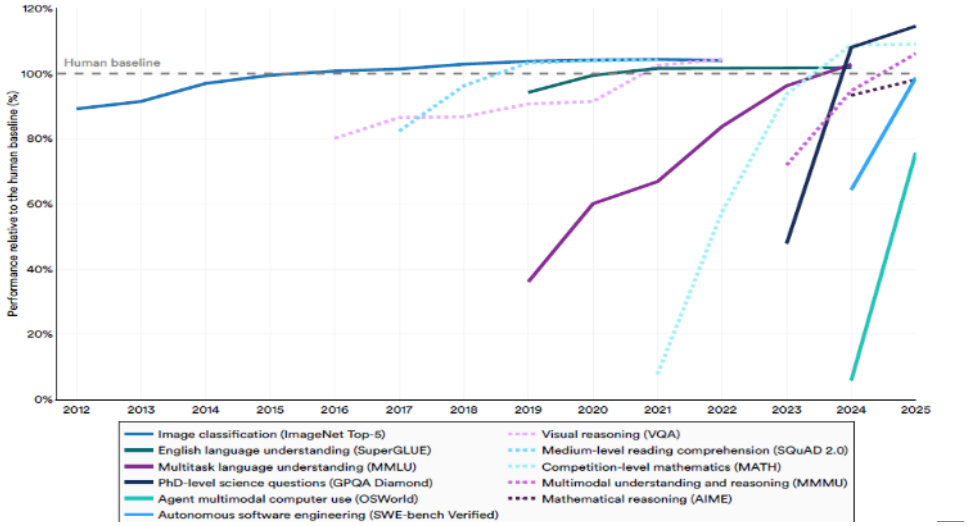
숫자만 보면 AI는 이미 도착했다. 그것도 예고된 것보다 빠르게.

그림 9. SWE-Bench Lite 및 Verified 점수 추이



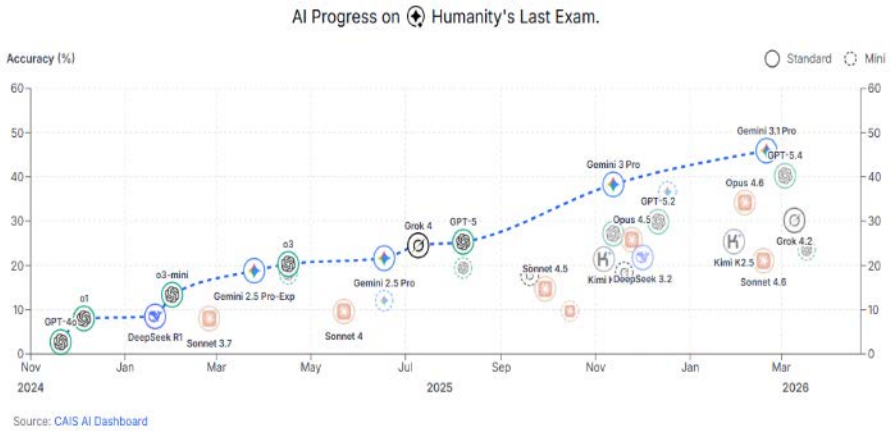
자료: HAI, 미래에셋증권 리서치센터

그림 10. 인간 기준선을 100%로 둔 모델 성능은 기준선에 근접 또는 기준선을 뛰어넘은 상황



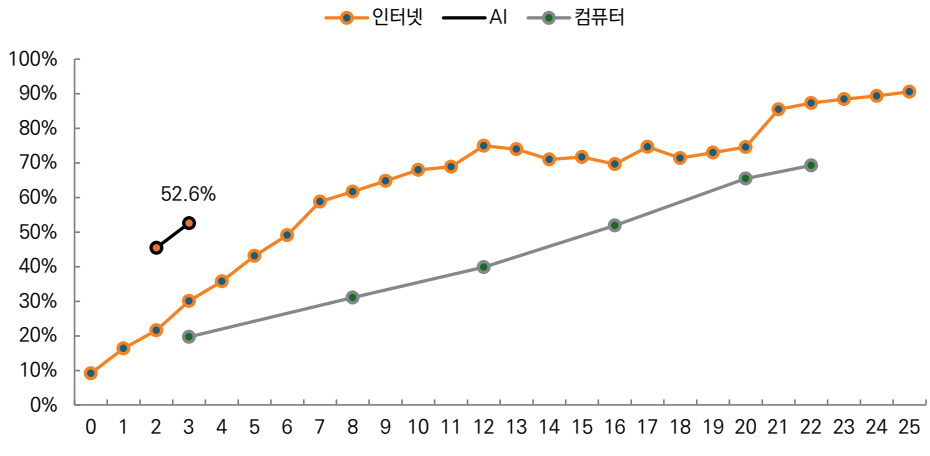
자료: HAI, 미래에셋증권 리서치센터

그림 11. Humanity's Last Exam 점수도 2024년 11월 2.7%(GPT-4o)에서 2026년 2월 45.9%(Gemini 3.1 Pro)로 꾸준한 성장세



자료: HAI, 미래에셋증권 리서치센터

그림 12. AI의 채택 속도는 역대급 – 이미 전 세계 인구의 53%가 AI를 사용하고 있음



자료: ITU, RPS, CPS, 미래에셋증권 리서치센터

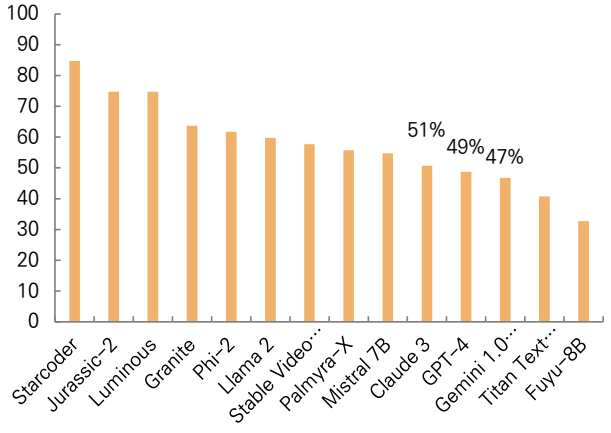
(2) 작년의 나는 잊어라: 절망편

그런데 이것은 이야기의 절반에 불과하다. 보고서가 동시에 측정한 또 다른 지표는 정반대 방향을 가리킨다.

Foundation Model Transparency Index(파운데이션 모델 투명성 지수)는 2025년 58점에서 2026년 40점으로 급락했다. 역대 최저다. AI 능력이 가장 빠르게 올라간 1년 동안, 그 AI를 만드는 기업들의 정보 공개 수준은 가장 크게 하락했다. 2025년 공개된 주요 모델 95개 중 80개가 훈련 코드 없이 출시됐고, Google, Anthropic, 그리고 OpenAI는 최신 모델의 데이터셋 규모와 훈련 기간 공개를 모두 중단했다. 보고서는 이 역설을 직접 지적하고 있다: "가장 강력한 모델일수록 공개하는 정보가 가장 적다."

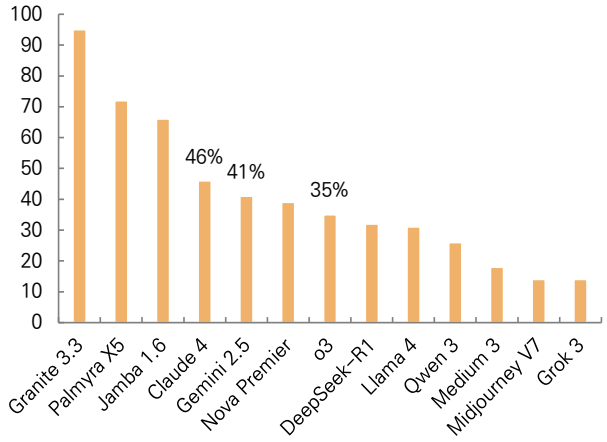
또한, AI를 바라보는 시각도 매우 혼란스럽다. 미국 일반 대중 중 AI가 일자리 방식에 긍정적 영향을 미칠 것이라고 답한 비율은 23%에 불과한 반면, 같은 질문에 대해 AI 전문가는 73%가 긍정적 영향을 미칠 것이라고 답했다. AI 전문가와 대중이 바라보는 관점 사이의 괴리를 단적으로 보여주는 지표다.

그림 13. 2024년 주요 플레이어들의 투명성 지수 대비



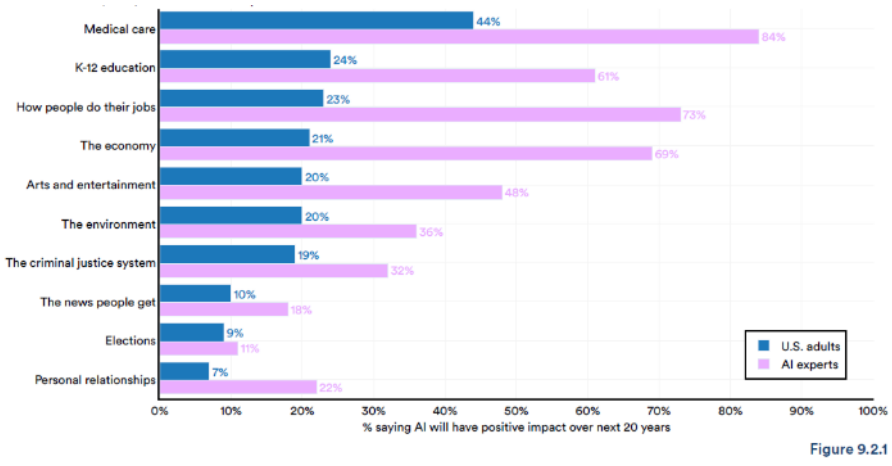
자료: HAI, 미래에셋증권 리서치센터

그림 14. 2025년 투명성 지수가 전반적으로 하락한 모습



자료: HAI, 미래에셋증권 리서치센터

그림 15. AI가 사회에 미칠 영향에 대해, 미국 성인은 AI 전문가들보다 덜 긍정적으로 바라보고 있음



자료: HAI, 미래에셋증권 리서치센터

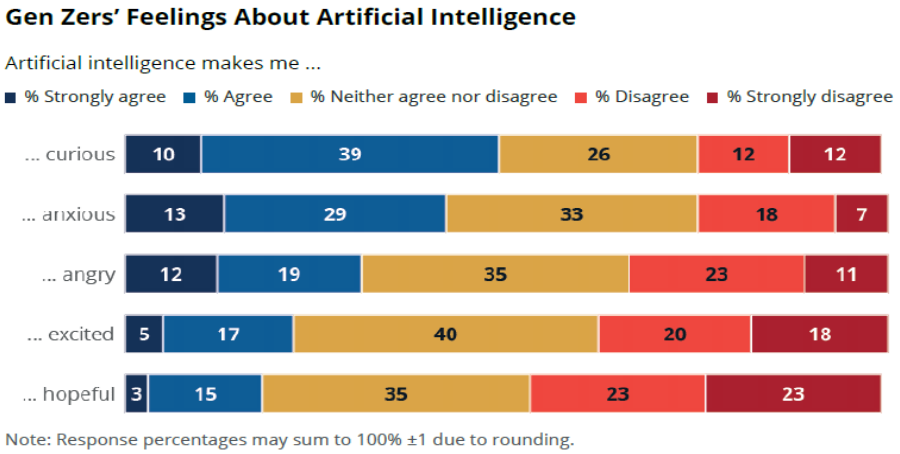
(3) 능력-불투명성-불안의 삼각 구조

데이터가 증명하듯, AI 성능은 아직도 가파르게 올라가고 있다. 동시에 이 AI를 만드는 과정은 점점 비밀스러워지고 있으며, 사용하는 사람들의 불안은 커지고 있다. 능력·불투명·불안이 동시에 상승하는 이 삼중 구조가 Stanford 보고서가 데이터로 드러낸 2026년 AI의 민낯이다.

지역 정부들이 데이터센터 신설에 규제와 금지로 대응하기 시작했다. Maine은 미국 최초로 데이터센터 건설 모라토리엄을 법제화할 가능성이 높으며, 현재 최소 11개 주에서 유사 법안이 발의 중이다. 갤럽 조사에 따르면 Gen Z의 22%는 매일, 29%는 적어도 주 1회 이상 AI를 쓰지만, AI에 대한 감정은 호기심과 불안감이 공존하는 일종의 카오스 상태다. 기술의 속도와 사회의 이해 속도 사이의 간극은 이미 현실로 나타나고 있다.

하지만 현 상황은 일종의 기호지세(騎虎之勢)다. AI라는 호랑이의 등에 올라탄 이상, 내릴 수는 없다. AI를 쓰지 않으면 뒤처지고, 쓰더라도 그 이후의 일들을 생각하면 안심할 수 없는 상황이다. 이러한 현 상황에서 오는 마찰은 일자리 문제에서 선명하게 나타나고 있다.

그림 16. Z 세대가 AI에게 느끼는 감정은 호기심/기대감과 불안/분노가 공존하는 혼돈 상태



자료: Voices of Gen Z: The AI Paradox, 미래에셋증권 리서치센터

2. 사다리 걷어차기(x), 지우기(o)?

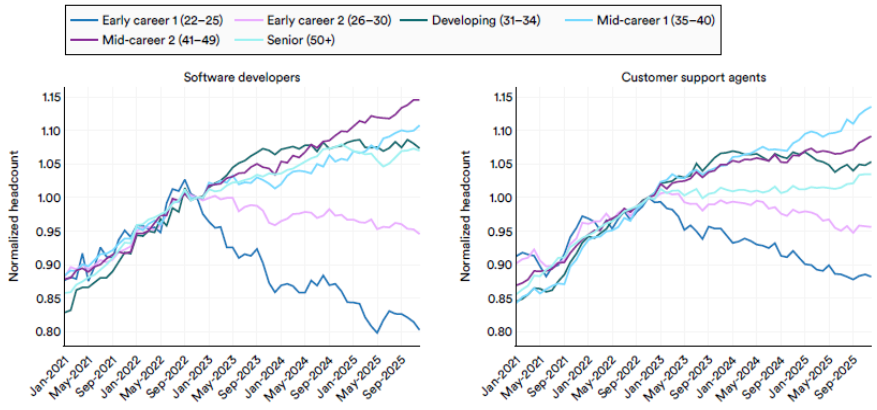
AI가 일자리를 빼앗는다는 서사는 이미 익숙하다. 그러나 Stanford 보고서와 그 기반이 된 데이터들이 실제로 가리키는 방향은 조금 다르다. 전체 고용 총량이 줄어들고 있다는 증거는 아직 뚜렷하지 않다. 대신 고용 충격이 매우 선명하게 집중되는 지점이 있다. 바로 경력의 시작점이다.

(1) 신입은 줄이고, 시니어는 늘리고

Stanford Brynjolfsson 연구팀이 미국 최대 급여관리 기업 ADP의 수백만 건 급여 데이터를 분석한 결과, 소프트웨어 개발자 중 22~25세 연령대의 고용은 2022년 이후 약 20% 감소한 것으로 나타났다. 같은 기간 동일 직종 30세 이상의 고용은 오히려 6~12% 증가했다. AI가 경험 많은 시니어를 보완하는 동시에, 신입이 수행하던 업무를 빠르게 흡수하고 있다는 해석이 가능하다.

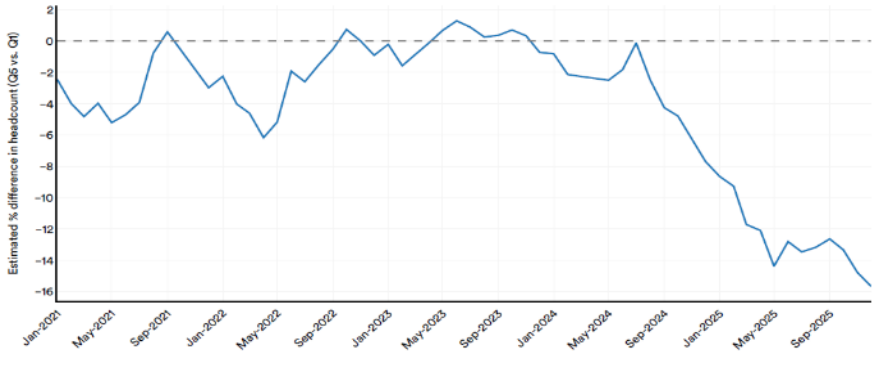
이 현상은 소프트웨어 개발에 그치지 않는다. AI 노출도가 높은 직군, 즉 고객서비스, 회계, 법률지원, 마케팅 등에서 22~25세 고용은 2022년 대비 13~16% 상대적으로 감소했다. 반면 AI 노출도가 낮은 직종, 즉 의료보조원이나 현장 생산감독 같은 역할에서는 연령에 무관하게 고용이 안정적으로 유지됐다. 따라서 여기에는 분명한 패턴이 있다. AI가 대체하는 것은 특정 사람이 아니라 특정 유형의 지식, 특히 교육과 훈련을 통해 습득한 체계적 지식(codified knowledge)이며, 그 지식에 가장 많이 의존하는 사람이 바로 커리어 초입의 신입 직원들이다.

그림 17. 소프트웨어 개발자 및 고객 서비스 직군의 연령대별 채용 변화



자료: HAI, 미래에셋증권 리서치센터

그림 18. 위 직군 외에도, AI 노출도가 높은 직종 전반에서 22-25세 채용은 감소 추세



자료: HAI, 미래에셋증권 리서치센터

댈러스 연은은 더 긴 시야에서 이 문제를 바라보고 있다. 기존 화이트칼라 직종의 경력은 신입으로 입사해 단순하고 반복적인 업무를 수행하면서 조직의 맥락과 암묵지를 축적하는 구조로 설계돼 있다. 그런데 AI가 이 신입 단계의 업무를 대체하기 시작하면, 기업 입장에서 신입 채용의 비용 대비 효익이 떨어진다. 기업의 이러한 의사결정은 단기적으로는 합리적인 선택이다. 그러나 이 선택은 장기적으로 중급·시니어 인력 공급 자체를 끊는 결과로 이어질 수 있다. 조직들이 여전히 필요로 하고 있고, 10년 후에도 필요할 숙련 인력은 지금 신입들이 훈련을 통해 성장해야만 도달이 가능하다. 그런데 이 파이프라인이 막히고 나면, 미래의 문제는 일자리가 아니라 사람이 된다.

그리고 이러한 구조적 현실이 Gen Z의 감정 데이터와도 일맥상통한다. 갤럽이 2026년 2~3월 진행한 조사에서 Gen Z 중 "화가 난다"는 응답은 2025년 22%에서 2026년 31%로 9%p 상승했다. 흥미로운 것은 이들이 AI를 외면하고 있어서가 아니라는 점이다. 절반 가까이가 AI를 매일 또는 주 1회 이상 사용하고 있다. 그럼에도 불안과 분노가 커지고 있다. 갤럽의 수석 연구원 Zach Hrynowski는 이 분노의 원인을 명확히 지목했다. Gen Z 중 가장 분노한 집단이 취업 시장에 가장 많이 노출된 연령대, 즉 취업을 준비 중이거나 막 커리어의 첫발을 내딛은 20대 초중반이라는 것이다.

따라서 이들의 분노는 단순한 기술 혐오가 아닌, AI를 매일 쓰면서도 그 AI가 자신의 커리어 불확실성을 키우고 있다는 구체적인 경험에서 오는 반응이다. Fortune의 보도에 따르면, AI 도입에 저항하는 일부 직원들이 성과 측정 지표를 조작해 AI의 효과를 숨기거나, 의도적으로 낮은 품질의 산출물을 내놓는 방식으로 AI 롤아웃에 저항하는 사례도 보고되고 있다. 호랑이 등에 올라타 있으면서도 한 발은 바닥에 끌리게 둔 형국이다.

(2) 생산성 역설

한편 기업 차원에서는 다른 숫자가 나온다. AI가 도입된 환경에서 고객서비스 생산성은 14%, 소프트웨어 개발 생산성은 26% 향상됐다는 수치가 보고서에 인용돼 있다. 그러나 이전 AI Bi-Weekly에서 소개한 바와 같이, 골드만 삭스는 이 생산성 이익이 거시 경제 전반에는 아직 나타나지 않고 있다고 지적한다. S&P 500 기업 중 AI의 수익 영향을 실제로 정량화한 곳은 1%에 불과하며, 경제 전반에서 AI와 생산성 사이의 "의미 있는 관계"는 아직 관측되지 않았다.

생산성은 올라가고 있는데 그 과실이 거시 데이터에서는 보이지 않고, 그 이면에서 신입 고용은 줄고 있다. 호랑이 등에 타기는 했지만, 아직 그 위에서 묘기를 부릴 정도는 아닌 셈이다.

3. 러다이트 2026년 ver.

기술에 대한 사회적 신뢰가 소진되는 방식은 선형적이지 않다. 신뢰는 불신을 억누르면서 서서히 줄어들다가 특정 임계점에서 급격히 붕괴하는 경향이 있으며, 그 이후의 반응은 누적된 불신이 한꺼번에 나타나는 형태를 취한다. 지금은 아직 그 임계점 이전이다. 그러나 전조는 이미 관측되고 있다.

그런데 이 마찰을 자세히 들여다보면, 성격이 다른 두 종류의 반감이 동시에 나타나고 있다는 점이 보인다.

첫 번째는 조직·정부 레이어의 반감이다. 데이터센터 신설에 따른 에너지와 수자원 부담이 특정 지역 사회에 집중되는 현상이 핵심이다. AI 인프라가 가져오는 경제적 이익은 빅테크 기업과 그 주주들에게 귀속되지만, 전력망 과부하, 수자원 소비, 소음과 환경 오염은 데이터센터가 들어선 지역 주민들이 감당한다. 데이터센터 신설에 대한 지역 규제 움직임이 미국 11개 주에서 동시에 나타나고 있고, Maine은 미국 최초의 주 단위 데이터센터 모라토리엄 법제화에 근접해 있다. 이는 AI 기술 자체에 대한 반감보다는, 충분한 설명 없이 빠르게 확장되는 방식, 그리고 그 과정에서 지역 사회가 감당해야 하는 에너지·수자원 부담에 대한 반응이다.

두 번째는 개인 레이어의 반감이다. 앞서 섹션 2에서 살펴본 것처럼, AI는 경력의 첫 칸을 지우면서 확산되고 있다. 22~25세 소프트웨어 개발자 고용이 20% 감소하는 동안 시니어는 오히려 성장했다. 이 구조에서 AI를 매일 쓰는 Gen Z가 동시에 분노하는 것은 모순이 아니다. 도구로서의 AI는 수용하면서도, 그 AI가 자신의 커리어 진입로를 좁히고 있다는 경험이 분노를 만들어내는 것이다.

두 반감의 공통점은 하나다. AI가 가져오는 이익과 비용이 균등하게 분배되지 않고 있다는 인식이다. 이는 과거 러다이트 운동과 구조적으로 닮아 있다.

19세기 초 영국 산업혁명 당시 직조공들이 방직기계를 파괴하며 저항했던 러다이트(Luddite) 운동은 오늘날 기술 부정의 대명사처럼 쓰인다. 하지만 당시 직조공들이 반대한 것은 기계 도입으로 인한 경제적 충격이 자신들에게만 집중되고 그 과실은 공장주에게 돌아가는 불균형한 구조였다. 기술이 가져온 생산성 향상이 사회 전체로 확산되는 데는 수십 년이 걸렸고, 그 이전의 과도기는 격렬한 갈등으로 채워졌다.

Gen Z가 AI를 매일 쓰면서도 분노하는 현상, 일부 직원들이 AI 롤아웃에 조직적으로 저항하는 현상은 21세기의 러다이트 운동과 그 구조가 닮아 있다. 개인 레이어에서는 기술의 과실이 자신들에게 닿지 않는다는 경험이 분노를 만들어내는 한편, 조직·정부 레이어에서는 공장이 들어설 때 지역 주민이 소음과 오염을 감당했던 것처럼, 데이터센터가 들어선 지역이 에너지와 수자원 부담을 떠안고 있다. 두 층위 모두, 기술 혐오가 아니라 비용과 이익의 배분 문제다. 그리고 역사가 보여주듯, 이 마찰은 기술이 충분히 확산되고 그 이익이 광범위하게 분배되기 시작할 때 비로소 완화됐다.

그렇다면 이 마찰을 AI 산업만 지니는 특수한 위기로 읽을 필요는 없다. 오히려 성숙하는 산업에서 자연스럽게 나타나는 조정 과정으로 보는 시각이 더 적합하다. 인터넷 산업도, 모바일 생태계도 초기에는 규제와 사회적 마찰을 동반했다. 그 과정에서 살아남은 플레이어들은 기술적 우위만으로 선별된 것이 아니었다. 장기적으로 지속 가능한 사업 모델과 동시에 사회와의 신뢰를 관리하는 역량, 그리고 규제 환경에 유연하게 적응하는 구조를 갖춘 곳들이 마찰을 통해 오히려 경쟁력을 벼릴 수 있었다.

하지만 사회적 불만이 수면으로 떠오르고 있는 지금, 벤치마크 성능이 얼마나 높은지도 중요하지만 이보다는 규제 환경 변화에 얼마나 유연하게 대응할 수 있는지, 사회적 신뢰를 얼마나 의식적으로 쌓아가고 있는지가 경쟁력을 가르는 하나의 변수로 작용할 수 있다. 결과적으로 러다이트 운동이 기계를 멈추지 못했듯, 지금의 마찰이 AI의 전진을 막을 수는 없을 거라 판단한다. 그러나 그 과정에서의 마찰을 어떻게 관리하느냐가 다음 국면의 승자를 가를 가능성이 높다.

III. AI Peer Table

표 2. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
최종 소비자(온디바이스, 앱)											
애플	AAPL US	263.40	5,718	1.1	4.2	3.2	6.6	36.2	-3.0	04-30	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	388.90	2,158	12.5	-1.7	-11.1	-9.3	61.0	-13.5	04-22	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
Pony AI	PONY US	10.97	7	17.1	-4.6	-32.7	-46.3	133.9	-24.3	05-20	중국의 자율주행 선도기업 중 하나
삼성전자	005930 KS	217,500	1,272	5.1	11.7	45.5	126.9	300.8	80.6	04-30	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
듀오링고	DUOL US	103.45	7	14.3	1.5	-31.1	-68.7	-68.3	-41.1	05-04	AI 캐릭터와 상호작용하는 학습 플랫폼
크래프톤	259960 KS	258,000	12	8.4	10.3	11.0	-8.0	-32.2	5.4	04-30	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
쇼피파이	SHOP US	126.94	245	13.0	0.3	-18.5	-18.9	51.2	-21.1	05-05	이커머스 콘텐츠 생성의 자동화(Shopify Magic AI)
코인베이스	COIN US	199.83	78	18.2	-1.7	-17.1	-39.5	16.0	-11.6	05-07	AI 에이전트간 자율 암호화폐 거래 인프라 구축
비즈니스 효율화											
크라우드스트라이크	CRWD US	418.20	157	6.0	-1.3	-7.9	-13.3	10.1	-10.8	06-03	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	251.00	348	5.8	0.7	-17.4	-8.0	7.9	-14.8	04-22	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	194.00	177	4.3	-1.8	-31.6	-15.9	-30.2	-26.6	06-18	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	151.66	325	8.7	-8.3	-25.0	-35.0	-34.1	-27.2	04-24	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,319.20	85	-0.9	5.6	-21.9	-9.0	-3.9	-18.3	04-23	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
일루미나	ILMN US	132.72	30	7.3	11.3	-6.3	38.4	80.5	1.2	04-30	AI 기반 첨단 유전체 분석
HD현대	267250 KS	260,000	21	2.8	-3.0	10.9	61.2	272.1	38.6	04-29	AI, 로보틱스, 디지털 트랜스포메이션 관련 노력
AI Ops											
클라우드플레어	NET US	197.38	103	2.2	-4.8	7.2	-6.6	80.2	0.1	05-07	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	142.76	505	9.4	-6.5	-16.5	-19.9	54.0	-19.7	05-04	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
파운데이션 모델											
알파벳	GOOGL US	336.02	5,983	5.5	10.0	1.9	33.8	119.9	7.4	04-29	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	676.87	2,533	7.7	7.9	9.2	-4.8	35.2	2.6	04-29	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	135.80	490	8.2	0.9	-18.3	-12.0	27.0	-4.9	05-15	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
네이버	035420 KS	219,500	34	8.7	-3.1	-9.5	-11.5	20.8	-9.0	04-30	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)

자료: Bloomberg, 미래에셋증권 리서치센터

주: 실적발표 일정은 변경될 수 있음.

표 3. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
클라우드											
마이크로소프트	MSFT US	420.26	4,615	12.6	5.1	-8.4	-17.5	14.0	-12.9	04-29	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	249.70	3,970	6.9	17.9	4.4	16.4	43.2	8.2	04-29	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	178.34	758	29.4	14.8	-6.3	-42.7	38.8	-7.9	06-11	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	4,672.00	248	23.7	32.0	16.7	-18.7	175.0	6.3	05-13	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	198.35	7,127	7.9	8.3	6.5	9.1	89.9	6.4	05-20	AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자
브로드컴	AVGO US	398.47	2,790	12.3	22.9	13.5	13.0	130.0	15.4	06-05	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	278.26	671	17.6	41.6	20.0	18.6	215.2	29.9	05-05	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	68.50	509	11.0	49.7	45.9	85.9	256.2	85.6	04-23	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	457.23	762	8.5	3.5	26.1	126.0	561.1	60.3	06-25	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	1,148,000	818	11.9	17.6	51.7	153.2	559.6	76.4	04-23	HBM 부문 전세계 최강자
시놉시스	SNPS US	441.15	125	8.9	3.6	-14.6	0.2	6.1	-6.1	05-28	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
케이던스 디자인 시스템	CDNS US	306.96	125	9.2	4.9	-3.3	-5.3	18.4	-1.8	04-27	반도체 EDA 부문 리딩 업체
암페놀	APH US	148.96	271	8.2	9.1	-3.3	17.4	131.6	10.4	04-29	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	294.13	166	2.3	11.1	66.3	65.5	311.0	81.6	04-22	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	162.33	255	8.4	33.4	53.5	-5.2	60.7	48.5	05-06	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	133.37	172	11.3	45.7	65.8	51.3	157.5	57.1	05-29	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
코어위브	CRWV US	119.56	95	30.0	39.2	18.1	-15.6	193.5	67.0	05-14	AI용 클라우드 인프라 제공업체
중과서광	603019 CH	88.99	28	4.3	4.2	-4.7	-15.0	36.6	4.0	04-24	중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체
SMIC	981 HK	59.95	113	2.9	-3.1	-24.3	-13.2	30.3	-16.1	05-08	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	2,085.00	2,534	4.3	11.5	20.2	44.7	149.8	34.9	07-17	명실상부 반도체 파운드리 최강자
폭스콘	2317 TT	207.00	135	3.2	-2.4	-11.7	-8.6	59.4	-10.2	05-14	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	134.00	20	-0.7	0.0	-7.6	-1.8	38.0	-11.0	05-06	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,475.00	13	14.9	8.9	4.0	12.6	58.4	-1.0	05-11	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스
에너지 인프라											
GE 베르노바	GEV US	978.32	389	1.1	18.3	43.6	62.8	199.0	49.9	04-22	AI 데이터센터용 가스터빈 인프라 솔루션
슈나이더 일렉트릭	SU FP	267.75	269	4.5	7.6	13.9	6.1	32.4	14.0	07-30	AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션
비스트라	VST US	165.53	83	8.4	2.3	-0.5	-21.1	43.6	2.8	05-07	미국 최대 독립 전력 생산 사업자
블룸 에너지	BE US	210.06	88	31.2	36.7	40.5	82.5	1,078.1	141.8	04-28	AI 데이터센터용 연료전지 솔루션
넥스트에라 에너지	NEE US	91.83	283	-2.8	-1.1	10.5	9.4	44.1	15.1	04-23	AI 데이터센터용 엔드투엔드 전력 파트너

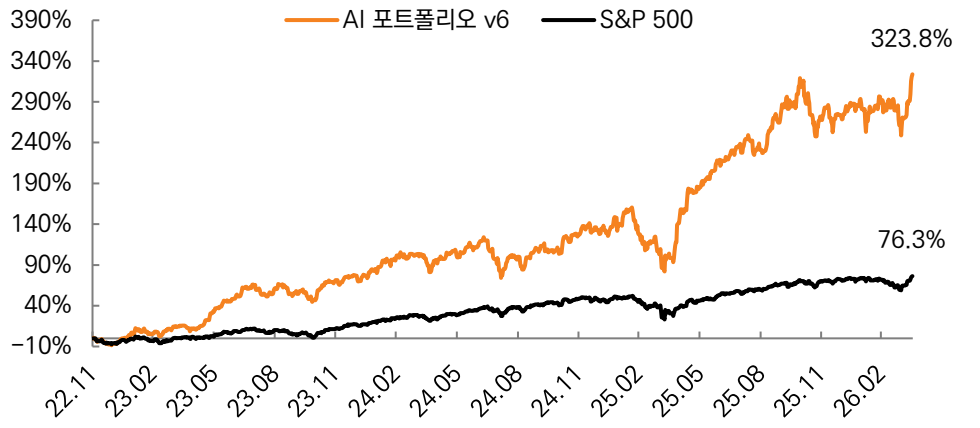
자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

IV. Appendix: Charts

그림 19. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)

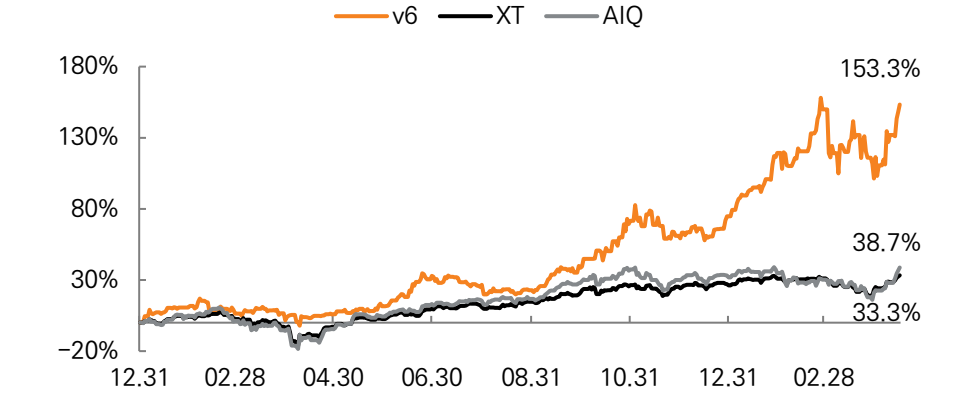
자료: Bloomberg, 미래에셋증권 리서치센터

그림 20. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)



자료: Bloomberg, 미래에셋증권 리서치센터

그림 21. 대표 AI ETF vs AI 유니버스(v.6) 수익률 추이(24.12.31 기준)
- Global X Artificial Intelligence & Technology ETF(AIQ), iShares Exponential Technologies ETF(XT)



자료: Bloomberg, 미래에셋증권 리서치센터

표 4. SimpleBench 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	평균 점수(%)	기관명	라이선스	지식 컷오프
-	인간 최고 점수	95.4	-	-	-
-	인간 평균	83.7	-	-	-
1(-)	Gemini 3.1 Pro Preview	79.6	구글	독점 폐쇄소스	2025년 1월
2(-)	Gemini 3 Pro Preview	76.4	구글	독점 폐쇄소스	2025년 1월
3(-)	Claude Opus 4.6	67.6	앤트로픽	독점 폐쇄소스	2025년 8월
4(-)	Gemini 2.5 pro(06-05)	62.4	구글	독점 폐쇄소스	2025년 1월
5(-)	Claude Opus 4.5	62.0	앤트로픽	독점 폐쇄소스	2025년 8월
6(-)	GPT-5 Pro	61.6	OpenAI	독점 폐쇄소스	2024년 9월
7(-)	Gemini 3 Flash Preview	61.1	구글	독점 폐쇄소스	2025년 1월
8(-)	Grok 4	60.5	xAI	독점 폐쇄소스	2024년 12월
9(-)	Claude 4.1 Opus	60.0	앤트로픽	독점 폐쇄소스	2025년 3월

자료: simple-bench.com, 각 사, 미래에셋증권 리서치센터
주: SimpleBench는 인간에게는 쉽지만 기계에게는 어려운 문제로 이루어진 벤치마크

표 5. LLM 벤치마크 챗봇아레나의 Hard Prompts 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	claude-opus-4-6-thinking	1536	앤트로픽	독점 폐쇄소스	2025년 8월
2(-)	claude-opus-4-6	1529	앤트로픽	독점 폐쇄소스	2025년 8월
3(-)	gemini-3.1-pro-preview	1516	구글	독점 폐쇄소스	2025년 1월
4(New)	muse-spark	1513	메타	독점 폐쇄소스	2026년 1월
5(▼1)	gpt-5.4-high	1505	OpenAI	독점 폐쇄소스	2025년 8월
6(-)	gemini-3-pro	1504	구글	독점 폐쇄소스	2025년 1월
7(▼2)	grok-4.20-beta1	1500	xAI	독점 폐쇄소스	2024년 11월
8(-)	claude-opus-4-5-20251101-thinking-32k	1500	앤트로픽	독점 폐쇄소스	2025년 8월
9(▼2)	grok-4.20-beta-0309-reasoning	1499	xAI	독점 폐쇄소스	2024년 11월
10(▼1)	claude-sonnet-4-6	1498	앤트로픽	독점 폐쇄소스	2025년 8월

자료: lmarena.ai, 각 사, 미래에셋증권 리서치센터
주: Hard Prompts는 기존의 일반적인 요청보다 훨씬 더 까다롭고 복잡한 문제를 제시하여 LLM의 한계를 테스트하는 벤치마크

표 6. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	GPT-5.4 Thinking xHigh Effort	80.28	OpenAI	독점 폐쇄소스	2025년 8월
2(-)	Gemini 3.1 Pro Preview High*	79.93	구글	독점 폐쇄소스	2025년 1월
3(New)	Claude 4.7 Opus Thinking xHigh Effort	76.91	앤트로픽	독점 폐쇄소스	2026년 1월
4(▼1)	Claude 4.6 Opus Thinking High Effort	76.33	앤트로픽	독점 폐쇄소스	2025년 8월
5(▼1)	Claude 4.5 Opus Thinking High Effort	75.96	앤트로픽	독점 폐쇄소스	2025년 7월
6(▼1)	Claude 4.6 Sonnet Thinking Medium Effort	75.47	앤트로픽	독점 폐쇄소스	2025년 8월
7(▼1)	GPT-5.2 High	74.84	OpenAI	독점 폐쇄소스	2025년 8월
8(▼1)	GPT-5.2 Codex	74.30	OpenAI	독점 폐쇄소스	2025년 8월
9(▼1)	GPT-5.1 Codex Max	73.98	OpenAI	독점 폐쇄소스	2024년 9월
10(▼1)	Gemini 3 Pro Preview High	73.39	구글	독점 폐쇄소스	2025년 1월

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 7. 코딩 관련 벤치마크 SWE-Bench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	점수(%)	기관명	등록일
1(-)	Claude 4.5 Opus (high reasoning)	76.80	앤트로픽	2026-02-17
2(-)	Gemini 3 Flash (high reasoning)	75.80	구글	2026-02-17
3(-)	MiniMax M2.5 (high reasoning)	75.80	미니맥스	2026-02-17
4(-)	Claude Opus 4.6	75.60	앤트로픽	2026-02-17
5(-)	GLM-5 (high reasoning)	72.80	Zhipu AI	2026-02-17
6(-)	GPT-5-2 (high reasoning)	72.80	OpenAI	2026-02-19
7(-)	GPT 5.2 Codex	72.80	OpenAI	2026-02-17
8(-)	Claude 4.5 Sonnet (high reasoning)	71.40	앤트로픽	2026-02-17
9(-)	Kimi K2.5 (high reasoning)	70.80	Moonshot AI	2026-02-17
10(-)	DeepSeek V3.2 (high reasoning)	70.00	DeepSeek	2026-02-17

자료: SWE-Bench, 미래에셋증권 리서치센터
주: 동일 모델이더라도, 추론 설정 등의 변경으로 점수가 재산정되는 경우 새로운 항목으로 판단. 등록일 업데이트를 참고

표 8. 지식 환각 벤치마크인 AA-Omniscience의 성능 상위 모델 10개

순위 (전주 변동)	모델	점수	환각률(%)	기관명	총 토큰 사용량 및 비용
1(-)	Gemini 3.1 Pro Preview	33	50%	구글	8.7M / $106
2(New)	Claude Opus 4.7	26	36%	앤트로픽	-
3(New)	Grok 4.20 0309 v2	15	17%	xAI	9.3M / $53
4(▼2)	Claude Opus 4.6(max)	14	61%	앤트로픽	11M / $297
5(▼2)	Claude Sonnet 4.6(max)	12	46%	앤트로픽	17M / $263
6(▼2)	Gemini 3 Flash	12	92%	구글	11M /$32
7(▼2)	GPT-5.4(xhigh)	6	89%	OpenAI	44M / $652
8(New)	MiMo-V2-Pro	5	30%	샤오미	21M / $63
9(New)	Qwen 3.6 Plus	3	32%	알리바바	16M / $47
10(▼4)	GLM-5.1	2	29%	Zhipu AI	14M / $62

자료: Artificial Analysis, 미래에셋증권 리서치센터
주: AA-Omniscience는 모델의 환각을 평가하는 벤치마크. 정답을 맞추면 가점을, 환각에는 감점을 하며 답변 거부에는 페널티를 주지 않음. 점수는 -100점부터 100점 사이로, 0점은 정답과 오답 수가 같음을 의미. 음수는 오답이 정답보다 많았음을 의미.
*환각률 = 답변하지 않거나 모른다고 해야하는 문제에 대해, 답변을 (환각으로) 제공한 비율
= 부정확한 답변 / (부정확한 답변 + 부분 답변 + 시도하지 않음)

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 팔란티어 테크 47주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 엔비디아 69주 보유하고 있습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.