Market Insight | 2026.7.9

그림: Midjourney(New Chatpter: Sovereignty)

AI 시장의 컬러가 바뀌었다

Good Enough 모델, 추론 슈퍼사이클, 그리고 AI 주권

한종목

chongmok.han@miraeasset.com

CONTENTS

Executive Summary

AI 시장 컬러 전환과 핵심 수혜주

2026년 하반기, AI 시장의 분위기가 바뀌고 있습니다. 지난 3년간 시장의 질문은 "누가 가장 똑똑한 AI 모델을 만드는가" 하나였습니다. **이제 질문이 바뀌었습니다. "누가 싸진 지능을, 기업이 안심하고 쓸 수 있게 만들어 주는가"입니다.**

인구의 극히 일부만이 Fable(최상위 AI 모델)이나 곧 출시될 GPT-5.6을 사용하고 있는 반면, 나머지 모든 사람들의 AI 경험은 소형 모델 수준에 머물러 있습니다. Google 검색의 AI 요약, Meta의 AI, ChatGPT 무료 버전 정도입니다. 쉽게 말해 소수만 스포츠카를 타고 대다수는 경차를 타는 세상인데, 그 스포츠카마저 지금은 규제와 안전장치 때문에 일반 판매가 막혀 있는 상황입니다.

이 격차는 문제가 아니라 국면입니다. **최고급 지능이 묶여 있는 동안, 시장은 '충분히 쓸만한(good enough)' 모델, 즉 1등은 아니지만 실제 업무의 대부분을 처리할 수 있는 값싼 모델로 회사 업무를 실제로 뜯어고치는 단계**에 들어섰습니다. **이 격차를 메우는 기간을 본 보고서는 버퍼 타임(**buffer time, 완충 기간)이라 부릅니다.

그리고 이 기간에 폭발하는 것이 추론(inference) 수요입니다. 추론이란 이미 만들어진 AI 모델을 실제로 돌려서 답을 뽑아내는 일로, 자동차에 비유하면 '차를 만드는 일(학습)'이 아니라 '차를 몰고 다니는 일'입니다. 차가 싸지고 흔해질수록 도로 위 주행량이 폭발하듯, 지능이 싸질수록 추론량이 폭발합니다.

여기에 이번 보고서를 통해, 새로 추가하는 축으로 세 가지를 제시합니다.

첫째는 돈줄입니다. 엔비디아가 7월 초 공식화한 새 사업 모델은 전 세계 신생 AI 인프라 사업자들에게 사실상 은행 보증을 서 주기 시작했습니다.

둘째는 주권입니다. 팔란티어의 'AI 주권' 선언과 CEO Alex Karp의 발언("토큰이 그렇게 가치 있다면, 파는 쪽은 왜 가치가 아니라 사용량 단위로 과금하는가")은 기업들이 지능을 빌려 쓰는 단계를 지나 생산수단을 소유하려는 다음 국면을 예고하며, 이 선언은 팔란티어와 엔비디아의 파트너십이라는 실물 동맹으로 즉시 뒷받침되었습니다.

셋째는 데이터입니다. 인터넷의 공짜 데이터가 고갈되는 이른바 '데이터 장벽' 때문에 최상위 지능의 확장 속도 자체에 상한이 걸렸고, **연구소들이 갈망하는 최고급 훈련 재료의 최대 매장지가 다름 아닌 기업 내부라는 사실이 AI 주권 테제의 증거**가 되고 있습니다.

결론을 한 문장으로 줄이면 이렇습니다. 다음 승자는 벤치마크 1등 회사가 아니라, 기업의 데이터와 권한이 지나가는 길목을 차지한 회사입니다. 이 기준으로 세 가지 시험(본문에서 설명합니다)을 전부 통과한 상장기업은 셋입니다. **통제를 파는 팔란티어(PLTR), 관문을 파는 아마존(AMZN), 그리고 컴퓨트와 돈줄을 쥔 엔비디아(NVDA)입니다.**

I. 시장의 컬러가 바뀌었다

1. 두 개의 AI 경제

지금 지구에는 두 개의 AI 경제가 있다. 한쪽에는 극단적 파워유저의 경제가 있다. AI에게 작업을 통째로 맡기는 '에이전트' 방식으로 AI를 쓰는 인구는 전 세계의 0.2%에 불과하다. 그런데 이 0.2%의 소비량이 무시무시하다. 비개발자 한 명이 컴퓨터 자원(CPU 코어 500개와 GPU 5개)을 하루 종일 점유하는 사례가 나오고, 대기업에서는 엔지니어 1인당 연간 1,000만 달러어치의 AI 사용료를 태우며, 인터넷 인프라 기업 Cloudflare의 집계로는 사람이 아닌 AI 에이전트가 만드는 인터넷 트래픽이 이미 사람의 트래픽을 추월했다.

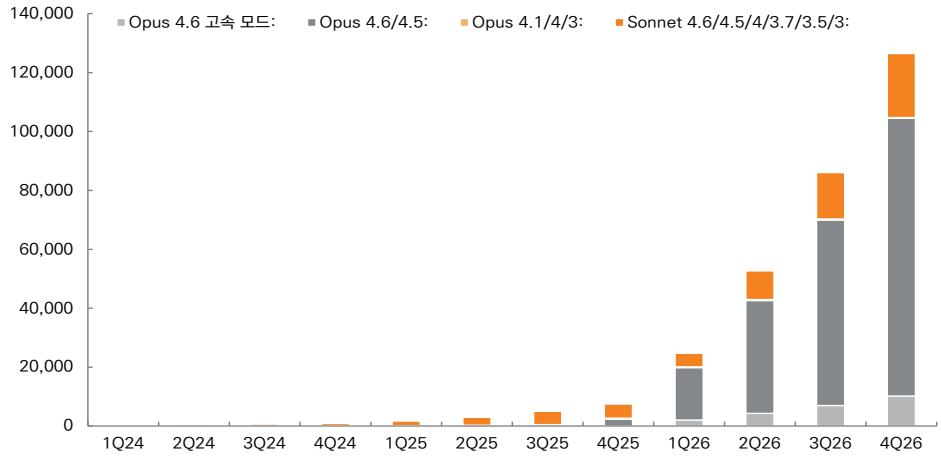
다른 쪽에는 나머지 99.8%의 경제가 있다. 이들의 AI 경험은 무료 챗봇과 검색창의 AI 요약이 전부다. 최상위 모델이 실제로 무엇을 할 수 있는지 체감한 적이 없으니, "AI가 일자리를 뺏는다"는 말도, 수천억 달러가 데이터센터에 쏟아지는 것도 이해가 되지 않는다. AI 거품론의 대중적 연료가 바로 이 체감 격차다.

투자자에게 이 격차는 두 가지를 말해 준다. 첫째, 아직 시작도 안 한 수요의 크기다. 0.2%가 이미 만든 추론 매출이 올해 말 2,000억 달러를 넘길 전망이다. 나머지 인구가 절반만 따라와도 숫자의 자릿수가 바뀐다. 둘째, 수요가 채워지는 방식이다. 99.8%가 AI에 진입하는 통로는 최고급 모델이 아니라 자기 업무에 '충분히 쓸만한(good enough)' 값싼 모델이다. 다음 성장 구간의 주인공은 최고급 지능이 아니라 값싼 지능의 대량 배포다.

2. 돈은 최상위 모델로 흐르는데, 일은 값싼 모델이 하고 있다.

2026년 상반기의 성적표는 언뜻 반대를 말하는 것처럼 보인다. AI 매출의 90% 이상은 최상위(프론티어) 모델이 가져갔다. Anthropic은 2분기에 흑자(주식보상비용 제외 기준)를 냈고, 주력 모델의 이익률은 80%를 넘는다. 그런데 같은 기간 전 세계에서 실제로 처리된 AI 작업량(토큰)의 약 80%는 무료로 공개된 오픈소스 모델이 감당했다. 돈은 최상위 모델로 흐르는데, 일은 값싼 모델이 하고 있는 것이다.

그림 1. Anthropic의 모델별 API 매출 추정 (단위: 백만 달러)



자료: SemiAnalysis, 미래에셋증권 리서치센터

이 어긋남의 원인은 단순하다. 상반기 최상위 모델 매출의 대부분은 '코딩'이라는 단 하나의 용도에서 나왔다. 코드는 틀리면 안 되니 기업들이 최고 모델에 기꺼이 웃돈을 낸다. 반면 문서 분류, 고객 응대, 사내 검색 같은 대량 업무는 이미 값싼 오픈 모델로 넘어갔다. 법률 AI 기업 Harvey의 사례가 상징적이다. 공개된 오픈 모델을 자사의 법률 데이터로 추가 훈련시키고, 질문을 적절한 모델로 자동 배분하는 장치(라우터)를 붙였더니 최상위 모델보다 싸고 더 좋은 결과가 나왔다.

이 어긋난 구조는 영원한 균형이 아니라 과도기의 사진 한 장이다. '매출은 최상위가 다 가져간다'를 영원한 진리로 믿으면 독립 AI 연구소들의 1조 달러급 몸값이 정당화되고, '일은 오픈소스가 다 한다'를 영원한 진리로 믿으면 그 연구소들은 전부 공매도 대상이 된다. 진실은 둘 다 아니다. 부(富)는 두 흐름이 반드시 통과하는 교차로로 이동한다. 추론을 돌리는 인프라, 그리고 여러 모델을 골라 연결해 주는 '관제 계층'이 그 교차로다.

표 1. 관제 계층을 이루는 다섯 가지

용어	쉽게 말하면	핵심 역할
Harness	한 모델의 일하는 방식	프롬프트, 도구 사용, 메모리, 컨텍스트, 중간 점검 방식을 설계
Routing	업무 배차	어떤 요청을 어떤 모델에 보낼지 결정
Orchestration	전체 작업 지휘	여러 모델·도구·에이전트·워크플로를 함께 조정
Control Plane	기업의 관제실	권한·정책·승인·감사·배포·모델 교체를 통제
Ontology	회사의 디지털 운영 지도	데이터와 실제 객체·관계·권한·행동을 연결

자료: 미래에셋증권 리서치센터

3. 세 개의 사건, 하나의 신호

2026년 6~7월, 이 이동을 증명하는 사건이 세 건 연달아 일어났다.

(1) Fable/Mythos 셧다운(6월 12일)

미국 정부가 안보를 이유로 Anthropic의 최상위 모델에 대한 외국인 접근을 제한한 사건은 중요한 사실을 드러냈다. 셧다운은 AI 수요를 없앤 것이 아니다. 수요의 주소지를 공용 API에서 기업 전용 관리형 환경으로 옮겼다. 외부 API 하나에 업무를 직접 연결하면 정부 서한 한 장으로 시스템 전체가 멈출 수 있다. 공용 API가 단일장애점(SPOF)이 될 수 있다는 사실이 실증된 것이다.

더 중요한 디테일은 최고급 모델이 묶인 동안에도 한 단계 아래의 충분히 좋은 모델은 계속 영업했다는 점이다. 최고급 지능만 제약되고 Good Enough 지능은 정상적으로 돈을 벌었다. 이는 지능의 양극화가 단순한 성능 차이가 아니라 제도적 배포 차이로 굳어질 수 있음을 보여준다.

(2) Meta의 클라우드 임대업 진출(7월 1일)

Meta가 AI 컴퓨팅 용량을 외부에 판매하겠다는 계획을 공개하자 시장은 크게 반응했다. 필자의 분석으로는 이 발표에 과장이 섞여 있다고 본다. Meta는 지금 컴퓨팅이 남는 회사가 아니라 오히려 모자라서 외부 업체들에 480억 달러 이상을 내고 빌려 쓰는 회사이며, 팔겠다는 물량도 지금이 아니라 2027~28년에야 완공될 미래의 용량과 구세대 GPU다.

그런데 버퍼 타임의 관점에서는 흥미로운 반전이 하나 숨어 있다. 구세대 GPU는 최첨단 경쟁에는 못 쓰지만, '충분히 쓸만한 모델'을 싸게 돌리는 데는 오히려 딱 맞는 장비다. 최신 경주용 차량은 아니어도 택시 영업용으로는 충분한 중고차인 셈이다. good enough 추론 붐은 Meta의 약점을 어느 정도 강점으로 바꿔 준다.

그리고 이 사건이 보여주는 더 큰 신호는 따로 있다. 광고로 먹고살던 회사조차 'AI 인프라 임대'를 새 수익원으로 선언할 만큼, 추론 인프라의 몸값이 올랐다는 사실 그 자체다.

(3) 엔비디아의 새 사업 모델 공식화(7월 초)

AI 칩의 지배자 엔비디아는 칩을 파는 데서 그치지 않고, 칩을 사는 AI 인프라 사업자의 최소 현금흐름을 보강하고 그 대가로 수익에 참여하는 새로운 구조를 내놓았다.

첫 파트너도 미국 빅테크가 아니라 호주와 인도네시아의 신생 사업자들이다. 이 사건이 보여주는 것은 단순한 GPU 판매 확대가 아니다. 추론 수요의 다음 병목이 컴퓨트 자체에서 컴퓨트를 지을 돈으로 이동하고 있다.

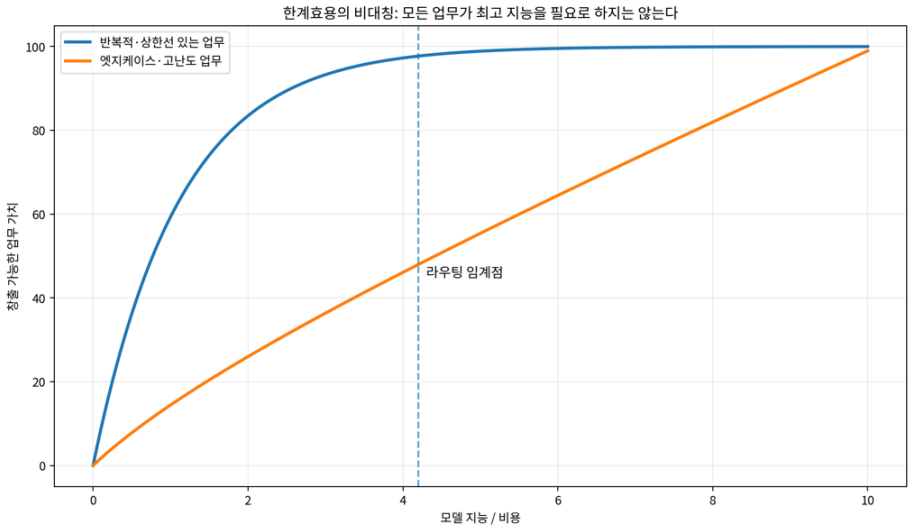
세 사건을 한 문장으로 줄이면 이렇다. 지능은 흔해지기 시작했지만, 지능을 안전하게 배포하고 그 배포에 돈을 대는 능력은 여전히 희소하다. 병목은 지능의 "생산"이 아니라 지능의 "유통"(배포, 통제, 금융)으로 이동했다.

II. "Good Enough 경제학"과 추론 슈퍼사이클

1. 한계효용과 달러당 지능

대부분의 기업은 회사 장부 정리에 수학 박사를 고용하지는 않는다. 장부 정리로 만들 수 있는 가치에는 상한선이 있어서, 그 이상의 지능은 돈 낭비이기 때문이다. 반대로 탈모 치료제 개발처럼 성공하면 가치가 거의 무한하게 열리는 문제에는 최고 지능이 필수다. 핵심은 이것이다. 기업 업무의 80% 이상이 전자, 즉 '상한선이 있는 일'이다. 문서 분류, 영수증 처리, 고객 응대에서는 더 똑똑한 모델을 써도 더 많은 돈을 벌지 못한다.

그림 2. 지능의 한계효용: 모든 업무가 최고 지능을 필요로 하지는 않는다
문서 분류·고객 응대 같은 반복 업무는 모델 지능이 일정 수준을 넘으면 업무 가치가 빠르게 포화.
더 비싼 최고급 모델을 써도 얻는 이익이 거의 없다.
반면 연구· 중대한 의사결정 같은 엣지케이스 업무는 모델이 똑똑해질수록 추가 가치 계속 커짐.
따라서 기업은 모든 업무에 최고급 모델을 쓰는 대신,
대부분은 값싼 Good Enough 모델로 처리하고 어려운 일만 최고급 모델로 보내는 라우팅이 경제적.
결국 AI의 목표 함수는 '가장 똑똑한 모델'이 아니라,
업무 난이도에 맞는 지능을 가장 낮은 비용으로 배분하는 시스템으로 이동한다.



자료: 미래에셋증권 리서치센터

최고 성능이 모든 가치를 가져가는 멱법칙(power law, 상위 1%가 전체를 독식하는 분포)은 극단적 예외 상황이 가치나 책임의 99%를 차지하는 도메인에만 적용된다. 예술, 스포츠, 최첨단 소프트웨어가 그렇다. 그러나 이 세상을 이루는 수만 가지 일은 그런 구조를 갖지 않는다. 가치의 95%가 '충분히 괜찮음'으로 채워지는 곳에서, 추가적인 몇%p의 성능 향상을 위해 수십 배의 웃돈을 낼 사람은 없는 것이다.

그래서 산업의 목표 함수가 바뀌고 있다. '가장 똑똑한 지능'에서 '1달러당 가장 많은 지능(intelligence per dollar)'으로 말이다. 기업들의 AI 시범 도입이 실전 배치로 넘어가면서, 재무 부서가 AI 예산에도 다른 지출과 똑같은 투자수익률(ROI) 잣대를 들이대기 시작했다. 2026년에 본격화된 조용한 혁명이다.

2. 디플레이션의 속도와 리눅스 모멘트

숫자가 혁명의 속도를 보여준다. AI를 한 번 돌리는 비용(추론 비용)은 2년 만에 최대 97% 떨어졌고, 같은 성능 기준으로 1년에 60배씩 싸지고 있다. 하드웨어 세대 교체, 매주 갱신되는 구동 소프트웨어, 모델 구조 자체의 효율화가 겹친 결과다. 지능이 전기와 인터넷이 걸었던 길, 즉 특별한 기술에서 흔한 기반 설비로 가는 길을 압축 재연하고 있는 셈인데, 역사의 교훈은 명확하다. 전기가 흔해졌을 때 부의 가치는 발전소가 아니라 전력망과 가전제품으로 갔다.

무료로 공개된 오픈소스 모델(Llama, DeepSeek, Qwen 등)은 벤치마크 1위를 할 필요가 없다. 기업 업무의 80%를 처리할 수 있는 '충분히 쓸만한' 수준에만 도달하면 판이 바뀐다. 컴퓨터 운영체제 Linux가 그랬다. Windows를 성능으로 이긴 적은 없지만, 공짜이고 충분히 좋았기에 전 세계 서버 시장을 장악했다. AI의 리눅스 모멘트가 지금이라는 평가가 많다. 이에 대한 실물 증거는 Meta의 내부 관행이다. 크고 비싼 모델은 '교사'로만 쓰고, 실제 서비스에는 교사에게 배운 작고 싼 '증류(distilled) 모델', 곧 큰 모델의 능력을 작은 모델에 압축해 옮긴 것을 배치한다. 지능을 만드는 일(학습)과 지능을 쓰는 일(추론)이 분리된 것이고, 이 분리가 돈의 축을 모델 자체에서 모델을 돌리는 인프라로 옮겨 놓았다.

(1) Chamath의 테스트

오픈소스가 시장을 장악하는 '리눅스 모멘트'를 입증하는 정량적인 사례도 나왔다. 유명 투자자 Chamath Palihapitiya가 본인 회사에서 실시한 노후 시스템 현대화 작업에서 진행한 실증 테스트가 대표적이다. 최고급 모델인 Claude Opus 4.8에 외부 제어판(별도의 오케스트레이션)을 결합하자, 단독 사용 대비 비용은 1.4배 저렴해지고 속도는 1.5배 향상됐다. 반면 같은 제어판에 값싼 오픈 모델(Zhipu의 GLM 5.2)을 붙이자, 속도는 3배 느려졌지만 비용이 무려 16.4배나 폭락했다. 단 1회의 예비 실험이라는 한계는 있지만, 그가 던진 질문은 정곡을 찌른다. "오픈 모델을 쓰면 비용이 16분의 1로 줄어드는데, 주주 돈을 집행하는 상장사가 왜 비싼 폐쇄형 API를 고집하는가?"

(2) NVIDIA의 실증 (feat. LangChain)

그리고 불과 며칠 뒤, 이 문제의식을 훨씬 더 공식적인 사례가 뒷받침했다. NVIDIA와 LangChain은 오픈 모델인 Nemotron 3 Ultra에 장기 에이전트용 제어 계층(Deep Agents Harness)을 맞춤 조정한 결과, 자체 평가에서 0.86의 성능을 1회당 4.48달러에 구현했다고 발표했다. 가장 가까운 성능을 낸 폐쇄형 최상위 모델의 비용은 43.48달러로 약 10배 높았다. 핵심은 Nemotron이 절대적으로 가장 똑똑한 모델이라는 데 있지 않다. 모델이 도구를 어떻게 사용하고, 어떤 맥락을 읽고, 중간 결과를 어떻게 점검할지를 주변 소프트웨어가 업무에 맞게 설계하자, 훨씬 싼 오픈 모델도 프런티어 모델에 근접한 실제 에이전트 성능을 냈다는 점이다.

두 사례가 가리키는 방향은 같다. 기업이 사야 하는 것은 언제나 가장 똑똑한 모델이 아니라, 자기 업무의 품질 기준을 가장 낮은 비용으로 통과하는 시스템이다. 모델의 부족한 몇 퍼센트를 더 비싼 지능으로 메울 것인지, 아니면 더 나은 관제와 업무 설계로 메울 것인지가 새로운 경제적 선택지가 된 것이다. 엔터프라이즈 AI의 가치는 점점 모델 단품이 아니라 '맞춤형 관제 계층 + 충분히 쓸만한 모델'이라는 조합에서 나온다.

그림 3. NVIDIA 공식 트윗: 'Good Enough + 맞춤형 관제'의 선언
Nemotron 3 Ultra 자체를 다시 학습시키는 대신, 모델 주변의 '일하는 방식'을 조정한 것
그 결과 싼 오픈 모델도 메모리 사용법, 도구 설명 등을 잘 설계하면 프런티어급 성능에 근접.
더 중요한 점은 이런 환경을 모두 기업이 직접 수정하고 통제할 수 있어,
폐쇄형 API와 달리 기업 고유의 업무 지식이 내부 자산으로 남는다는 것.

자료: 엔비디아, LangChain, 미래에셋증권 리서치센터

그림 4. 수 백개의 에이전트 중에서 엔비디아 "Nemotron 3 Ultra"가 가격대비 정확도 제일 우수.
최고 성능의 Opus 4.8은 정확도가 약 87%로 소폭 높았지만, 비용은 43.48달러로 거의 10배 비쌈.
즉, 기업 입장에서는 불과 1%p 안팎의 추가 성능을 위해 10배의 비용을 지불할지,
훨씬 싼 오픈 모델로 업무 기준을 통과할지 선택.
'가장 똑똑한 모델'보다 '필요한 품질을 가장 낮은 비용으로 수행하는 시스템'으로 이동할 것이다

자료: LangChain, 미래에셋증권 리서치센터

3. 논리의 급소, 그리고 버퍼 타임

단, 필자가 주장하는 "Good enough" 모델 사이클에는 숨은 가정이 하나 있다. "업무의 난이도가 지금 그대로"라는 가정이다. 지능이 싸지면 기업은 더 어려운 일, 즉 단순 분류가 아니라 맥락을 읽고 스스로 결정하는 일에 AI를 쓰려 들 것이고, 그러면 최상위 모델 수요도 함께 되살아날 것이다. 실제로 올 상반기 최상위 모델로 매출이 쏠린 것이 이 메커니즘(코딩이라는 고난도 업무의 등장)이었다.

그러나 주장의 '급소'를 인정해도 투자 결론은 바뀌지 않는다. 업무 상향의 속도 자체에 상한이 걸려 있기 때문이다. 최상위 모델이 더 어려운 업무로 확장하려면 그 업무의 훈련 데이터가 필요한데, 인터넷의 공짜 데이터는 이미 바닥났다. 업계 분석에 따르면 인터넷 전체에서 실제로 쓸모 있는 공개 텍스트는 약 300조 토큰 규모로, 프론티어 연구소들은 이를 거의 다 소진했다. 인터넷은 인류 문명이 AI에게 준 일회성 보조금이었던 셈이다. 따라서, AI 모델이 회계나 법률 분야에서 코딩만큼 인간 전문가보다 잘하지 못하는 이유는 지능의 한계가 아니라 데이터 커버리지의 한계인 것이다. 새 도메인의 전문가 데이터는 사람을 고용해 비싸게 '제조'해야 한다. 요컨대 최상위 지능의 확장 속도는 이제 데이터 수집 속도에 묶여 있다.

결국, 프론티어 AI가 계속 더 똑똑해져도, 그 지능이 기업 현장 전체를 곧바로 장악하지는 못한다는 말로 귀결된다. 최상위 지능의 AI는 규제 때문에 '널리 쓰이지' 못하고, 데이터 부족 때문에 '모든 업무를 잘하게' 되지도 못하기 때문이다. 그래서 그 사이에 값싸고 통제 가능한 Good Enough 모델이 기업 업무를 먼저 차지하는 시간이 생긴다. 그것을 필자는 버퍼 타임이라 부르고, 이 시간적 간극이 짧게 끝나기 힘들 것이라 생각한다. 다시 말해, 이 시공간에 투자 기회가 있을 것이다.

그림 5. 최상위 AI는 규제와 접근 제한 때문에 모든 국가와 기업에 즉시 배포되기 어렵다.
동시에 공개 웹 데이터의 고갈은 프런티어 모델이 새로운 산업과 업무로 확장되는 속도를 늦춘다.
두 제약 덕분에 생기는 시간이 버퍼 타임, 기업은 그 공백을 Good Enough 모델의 배포로 메운다.
이때의 투자 기회는 최상위 지능보다, 값싼 지능을 배포하고 라우팅하며 통제하는 길목에서 찾는다_

자료: 미래에셋증권 리서치센터

III. 추론 붐을 이끄는 길목과 돈줄

1. AI의 중앙은행: 엔비디아의 최저 수익 보증 모델

'네오클라우드'란 (주로) 엔비디아의 GPU를 대량으로 사서 데이터센터에 설치해 놓고, 그 컴퓨팅 파워를 시간 단위로 빌려주는 임대업자다. 건물을 지어 사무실을 임대하는 부동산업과 구조가 같다. 다만 건물 대신 GPU를 임대한다.

문제는 이 임대업이 돈을 빌리기가 어렵다는 점이다. 클러스터(GPU 수만 개를 묶은 단지) 하나를 세우려면 세 가지가 동시에 맞물려야 한다. 첫째는 자본, 곧 수십억 달러의 대출이다. 둘째는 구매 확약(offtake)으로, "완공되면 우리가 몇 년간 빌려 쓰겠다"는 대형 고객의 장기 계약이다. 셋째는 장비를 놓을 데이터센터(건물과 전력)다. 그런데 은행은 장기 고객 계약이 없으면 대출을 내주지 않는다. 그러나 클러스터를 빌리려는 고객은 자금과 데이터센터가 확보되지 않은 사업자와 장기 계약을 맺으려 하지 않고, 데이터센터 사업자는 자금과 고객이 확인되지 않으면 희소한 전력과 공간을 내주지 않는다. 모두가 다른 두 주체가 먼저 움직이기를 기다리는 삼각 교착이다.

지금까지 이 교착을 푸는 열쇠는 하나뿐이었다. Microsoft나 Meta 같은 초우량 대기업이 "우리가 5년간 통째로 빌리겠다"고 보증을 서 주는 것이다. 은행은 임대업자가 아니라 그 뒤의 대기업 신용을 보고 돈을 빌려줬다.

표 2. AI 데이터센터 건설 및 융자시 3각 의존 및 교착관계

구분	은행 / 대주단	대형 고객	데이터센터 / 전력
은행 / 대주단	-	장기 구매확약 필요	전력·공간 확보 확인 필요
대형 고객	대출 실행 확인 필요	-	실제 용량 확보 필요
데이터센터 / 전력	자금조달 확인 필요	장기 수요 확약 필요	-

자료: 미래에셋증권 리서치센터

이 방식에는 한계가 있었다. 기존 AI 인프라 금융은 '5년짜리 대형 고객 계약'이 있어야 성립했다. 그러나 시장에서 가장 빠르게 늘어나는 추론 업체와 AI 스타트업은 몇 달짜리 유연한 GPU 임대를 원한다. 실제 수요는 넘치지만 계약기간이 너무 짧아 은행이 장기 대출의 근거로 인정하지 못했고, 그 결과 수요가 인프라 공급으로 전환되지 못했다. 엔비디아의 최저 수익 보증은 바로 이 '단기 수요와 장기 금융 사이의 시간 불일치'를 메우는 신용 브릿지라 할 수 있다.

표 3. 엔비디아가 신용 브릿지가 된다

주체	엔비디아 보증 전	엔비디아 보증 후
은행	단기 고객은 못 믿음	최소 현금흐름 바닥을 믿고 대출
GPU 클라우드	5년 고객 없으면 건설 어려움	단기 고객 중심으로도 건설 가능
AI 스타트업	장기 계약 요구받음	수개월 단위로 GPU 사용 가능
엔비디아	GPU 일회성 판매	보증 + 수익 배분 구조

자료: 미래에셋증권 리서치센터

엔비디아가 7월 초 공식화한 모델의 구조는 쉽게 말해 'GPU 클라우드 사업자'의 매출에 바닥을 깔아주는 것이다. 고객을 충분히 확보하지 못해 GPU가 놀 경우, 엔비디아가 일정 기간 미리 정한 가격으로 남는 컴퓨팅을 사주겠다고 약속한다. 그러면 은행은 신생 GPU 클라우드 업체의 신용이 아니라 엔비디아의 보증을 보고 장기 대출을 실행할 수 있다. 대신 사업이 잘돼 실제 임대 매출이 보증선을 크게 넘어가면, 엔비디아는 초과 수익의 일부(40~60%로 추정)를 나눠 갖는다. 한마디로 '안되면 엔비디아가 일부 바닥을 받쳐주고, 잘되면 위쪽 수익을 함께 먹는' 구조다. 이 때문에 엔비디아는 GPU를 한 번 팔고 끝나는 공급자에서, GPU가 수년간 만들어내는 임대 수익에 계속 참여하는 금융 파트너로 진화한다.

표 4. 엔비디아의 보증: 네오클라우드 A사의 매출에 일종의 바닥(floor)을 깔아주는 backstop

상황	A사의 실제 고객 매출	엔비디아의 역할
장사가 잘됨	100	아무것도 안 사줘도 됨
장사가 보통	70	보증선 이상이면 개입 없음
장사가 부진	40	부족한 컴퓨팅 일부를 사줌
장사가 매우 부진	20	약정 범위 안에서 매출 바닥을 받쳐줌

자료: 미래에셋증권 리서치센터

표 5. "망할 위험은 내가 일부 막아줄 테니, 대박 나면 약 절반 정도는 나눠줘"

사업 결과	GPU 클라우드 업체	엔비디아
매우 부진	최소 수익 방어	손실 위험 부담
보통	정상 영업	개입 제한적
대박	큰 수익 창출	초과 수익 일부 공유

자료: 미래에셋증권 리서치센터

엔비디아의 첫 공식 파트너는 호주의 Sharon AI(GPU 최대 4만 개, 보증 총액 약 49억 달러)와 인도네시아의 Firmus(가속기 최대 17만 개)다. 업계에서 이 구조를 두고 "엔비디아가 AI의 중앙은행이 되고 있다"는 표현까지 통용되기 시작했다. AI 인프라 시장의 유동성 공급 역할을 엔비디아가 자처하고 나선 것이다.

우선, 이것은 good enough 추론 붐의 자금줄이라 생각한다. 이 프로그램의 명시적 목표가 "소수 대기업 너머로 컴퓨팅 접근을 개방하고, 1년 미만 단기 임대를 공급하는 것"이다. 그런데 단기 임대의 수요자는 바로 추론 서비스 업체와 스타트업들이다. 버퍼 타임에 폭발하는 추론 수요가 돈줄의 벽에 막히지 않도록, 엔비디아가 직접 신용을 공급하는 것이다.

둘째, 이것은 '한 번 팔고 끝'을 '두고두고 받는 돈'으로 바꾸는 사업 모델 전환이다. 칩을 팔면 매출은 그 분기로 끝나지만, 그 칩이 6년간 벌어들이는 임대 수익의 40~60%를 나눠 받으면 반복 수익(recurring revenue)이 된다. 토큰 처리량이 늘수록 엔비디아의 배분 수익도 늘어나는, 추론 붐 그 자체에 대한 과금 장치다.

셋째, 지정학이다. 첫 파트너가 미국이 아니라 호주와 인도네시아라는 사실이 말해 주듯, 엔비디아는 각 나라에서 자국 통제하의 AI 인프라, 이른바 소버린 AI를 세우려는 로컬 챔피언들을 키우고 있다. 미국 대기업들이 자체 칩으로 엔비디아와 맞서기 시작한 세계에서, 구매자층을 전 세계에 만들어 두는 것은 젠슨 황이 수년간 공언해 온 '다극 체제' 전략이다.

추론 슈퍼사이클의 돈줄이 열렸고, 그 수도꼭지를 쥔 것은 은행이 아니라 엔비디아다.

2. 오케스트레이션: 갈아타기 비용의 진짜 위치

AI 모델은 본질적으로 기억이 없는 계산기다. 질문을 넣으면 답을 뱉을 뿐, 이 직원이 이 데이터를 볼 권한이 있는지, 이 결재는 누구를 거쳐야 하는지, 지난 분기에 같은 문제를 어떻게 처리했는지 모른다. 그런 회사의 '맥락 기억'을 쥔 쪽(관제 계층)이 AI 시대의 운영체제가 되고, 모델은 6개월마다 갈아 끼우는 부품이 된다.

관제 계층에는 AI가 언제 틀렸고, 사람이 무엇을 고쳤으며, 어떤 판단이 최종적으로 승인됐는지가 모두 기록된다. 이 기록은 시간이 갈수록 그 회사만의 전용 AI를 만드는 가장 귀한 학습 재료가 된다. 자율주행차가 운전자의 개입과 위험 상황을 계속 모아 다음 버전의 주행 성능을 높이듯, 기업 AI도 사람의 수정과 승인 기록을 먹으며 점점 그 회사의 업무 방식에 맞게 진화한다. 결국 중요한 것은 처음부터 가장 똑똑한 모델을 가진 자가 아니라, AI가 일하고 실패하고 교정되는 전 과정을 기록하는 자다. 관제를 장악한 플랫폼이야 말로 기업의 AI 개선 선순환(flywheel)까지 함께 쥔다고 볼 수 있다.

다만 기업들이 원하는 것은 단순히 더 좋은 관제 도구가 아니라 선택권을 잃지 않는 것이다. 한 AI 모델 회사에 종속되는 것도 두렵지만, 한 클라우드 사업자에게 모든 데이터와 업무를 맡기는 것 역시 위험하다. 그래서 어떤 모델을 쓸지, 어디에서 돌릴지, 얼마를 쓸지, 무엇을 외부에 내보낼지를 기업이 직접 결정할 수 있게 해 주는 독립형 관제 플랫폼이 투자 기회로 성립한다.

그리고 바로 이 지점에서 팔란티어가 'AI 주권(sovereignty)'이라는 이름으로 가장 선명한 선언을 내놓았다.

3. 주권의 선언: “생산수단을 소유하라”

팔란티어는 7월 초 'AI 주권(sovereignty)'에 관한 9개 조항의 공식 입장을 발표했다. 핵심은 어렵지 않다. AI를 쓰되, 중요한 선택권까지 남에게 넘기지 말라는 것이다.

그림 6. 팔란티어가 7월 초 갑자기 트윗으로 써버린 “AI 주권에 대한 생각”

자료: 팔란티어, 미래에셋증권 리서치센터

첫째, 데이터다. 기업의 데이터에는 단순한 문서만 들어 있는 것이 아니다. 어떤 고객을 우선하는지, 어떤 예외를 허용하는지, 누가 무엇을 승인하는지, 실패했을 때 어떻게 고치는지 같은 그 회사만의 운영 방식이 들어 있다.

이 데이터를 외부에 넘기는 것은 기록을 맡기는 데 그치지 않는다. 미래의 AI가 배울 가장 중요한 원재료까지 함께 넘기는 것이다.

둘째, 가중치(weights)다. 가중치는 AI가 학습을 통해 얻은 능력이 압축된 결과물이다. 기업이 자기 데이터로 모델을 추가 훈련했는데도 그 결과물을 다른 회사가 통제한다면, 기업은 자기 비용으로 만든 지능을 스스로 소유하지 못하는 셈이다. 결국 데이터를 소유하고, 그 데이터로 만든 지능도 소유해야 한다는 말이다.

셋째, '관제'다. 어떤 모델을 쓸지, 어떤 데이터에 접근하게 할지, 어떤 행동은 사람의 승인을 받게 할지, 실패하면 무엇으로 교체할지를 기업이 직접 통제해야 한다. 결국 팔란티어가 말하는 AI 주권은 데이터·가중치·관제에 대한 선택권을 기업 내부에 남겨두는 것이다.

이런 문제의식은 CEO Alex Karp가 최근에 발언한 직설적인 비판으로 이어진다. "기업이 원하는 것은 토큰 자체를 많이 쓰는 것이 아니라, 실제로 돈을 벌거나 비용을 줄이는 것이다." 그런데 AI 모델 회사는 대개 토큰을 많이 사용할수록 더 많은 매출을 올린다. 즉, 고객은 더 적은 비용으로 문제를 해결하고 싶고, 공급자는 더 많은 사용량을 원한다.

둘의 이해관계가 일치하지 않고, 핵심은 AI의 경제적 가치와 토큰 소비량은 분명히 다르다는 점이다. 토큰을 많이 쓰는 시스템이 반드시 좋은 시스템은 아니다. 오히려 잘 설계된 소프트웨어와 업무 구조는 불필요한 호출을 줄이고, 싼 모델과 비싼 모델을 적절히 나누어 쓰며, 같은 성과를 더 적은 비용으로 낼 수 있다.

그래서 Karp는 실제로 돈이 쌓이는 곳이 두 군데라고 주장한다. "하나는 기업의 업무와 권한을 이해하는 애플리케이션 계층, 다른 하나는 그 모든 연산을 실제로 수행하는 컴퓨트 계층이다." 전자가 팔란티어의 온톨로지이고, 후자가 엔비디아와 AWS 같은 인프라다.

그리고 이 선언은 말에 그치지 않았다. 팔란티어와 엔비디아의 파트너십은 이 주권 논리를 실제 제품 구조로 옮긴다. 기존 방식에서는 기업이 폐쇄형 모델을 외부 API로 호출한다. 기업 데이터는 모델 회사의 시스템으로 들어가고, 모델은 외부에서 관리되며, 기업은 사용한 토큰만큼 계속 돈을 낸다. 팔란티어와 엔비디아가 제시하는 방향은 다르다.

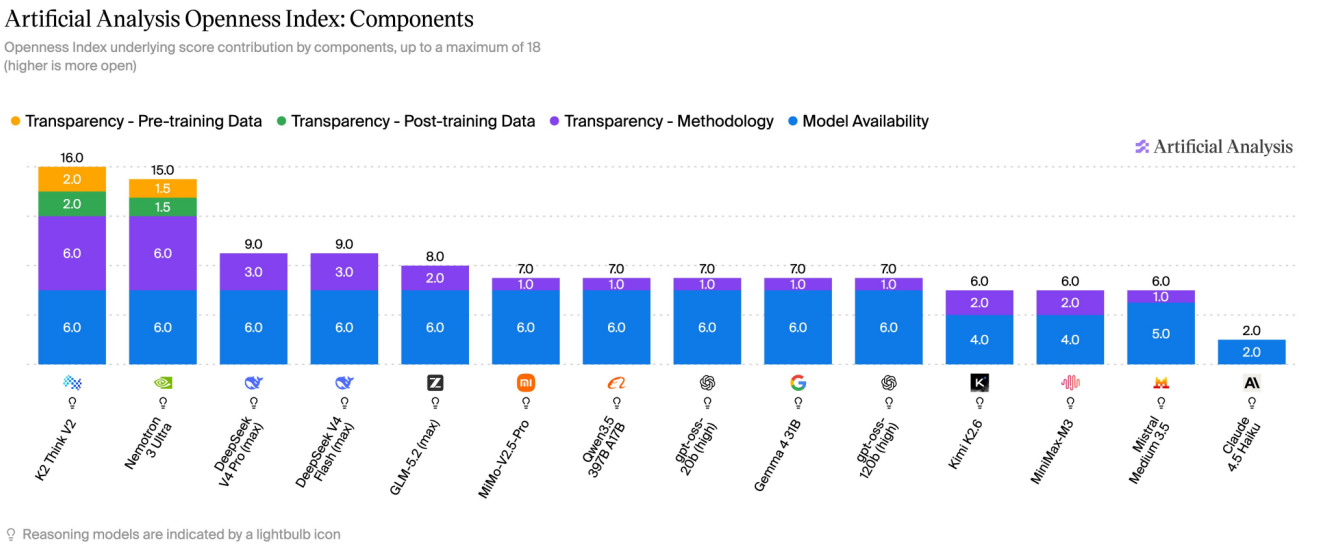
- 기업이 개방 모델을 가져온다.
- 자기 데이터로 추가 훈련한다.
- 그 결과 만들어진 가중치를 자기 통제 아래 둔다.
- 팔란티어의 온톨로지가 그것을 실제 업무와 연결한다.

즉, 모델을 빌려 쓰는 데서 끝나는 것이 아니라 자기 데이터로 만든 지능을 자기 시스템 안에 남기는 구조다. 본 보고서에서 말한 '주권 스택'이 바로 이것이다. "데이터 → 가중치 → 관제 → 실제 업무"로 이뤄지는 이 네 층을 남에게 통째로 맡기지 않는 것이다.

엔비디아가 직접 만든 오픈소스 모델인 Nemotron이 여기서 중요한 이유도 단순하다. 기업은 단지 가중치를 내려받을 수 있는 모델만 원하는 것이 아니다. 이 모델이 무엇을 배웠는지, 어떤 방식으로 만들어졌는지, 내부 검증이 가능한지까지 알고 싶어 한다. Artificial Analysis의 개방성 지수에서 Nemotron 3 Ultra가 높은 점수를 받는 이유도 이 때문이다.

중국계 오픈웨이트 모델 상당수는 가중치는 공개하지만, 어떤 데이터로 어떻게 훈련했는지는 충분히 공개하지 않는다.

그림 7. Artificial Analysis 개방성 지수: AI 모델이 얼마나 상세하게 공개했는지를 18점 만점으로 평가.
엔비디아 Nemotron 3 Ultra는 15점으로, 모델 접근성(6점)과 학습 방법론(6점)을 완전히 공개해 최상위권.
반면 DeepSeek·Qwen 등 모든 중국 '오픈 모델'은 어떤 학습 데이터로 훈련했는지 어떻게 훈련했는지 제조법은 공개하지 않음.
참고로 1위를 기록한 K2 Think V2는 UAE/아부다비에서 만든 '완전 주권형' 오픈소스 모델.



자료: Artificial Analysis, 미래에셋증권 리서치센터

반면 대기업은 모델을 내부 시스템에 넣기 전에 이런 질문을 던진다.

- 어떤 데이터가 학습에 들어갔는가?
- 모델의 행동을 내부에서 검증할 수 있는가?
- 그 위에 우리 회사의 데이터를 다시 학습시킬 수 있는가?

표 6. 대부분의 "오픈소스" 모델들은 완전 오픈이 아니다

항목	오픈 웨이트 (Open Weights)	완전 오픈소스 (Full Open Source)
가중치	공개	공개
코드	일부만 공개되는 경우 많음	대부분 공개
훈련 데이터	거의 비공개	공개 (또는 상세한 구성 공개)
훈련 과정	거의 비공개	상세 레시피 공개
재현 가능성	불가능	가능
주요 용도	실행 + 미세조정	연구, 재현 가능성, 새로운 모델 개발

자료: 미래에셋증권 리서치센터

중국의 AI 모델처럼 가중치만 열려 있어서는 이 질문에 충분히 답하기 어렵다. 그래서 개방성은 단순한 철학이 아니라 기업 배포의 조건이 될 수 있다. 모델을 실제 기업 내부망에 넣고, 검증하고, 자기 데이터로 다시 훈련하려는 수요가 늘수록, 더 깊게 더 넓게 열린 "진정한 의미의 오픈소스 모델"의 가치가 커진다. Nemotron과 팔란티어의 결합이 중요한 이유도 여기에 있다. 하나는 기업이 통제할 수 있는 모델을 제공하고, 다른 하나는 그 모델을 기업의 데이터·권한·업무에 연결한다.

여기서 한 가지 모순처럼 보이는 문제가 남는다. AI가 싸질수록 사용량이 폭발하는 제본스의 역설이 있지만, 정작 팔란티어의 CEO Alex Karp는 토큰을 많이 쓰게 만드는 사업 모델을 비판했다. 그러나 둘은 실제로 모순되지 않는다. 전체 토큰 사용량은 폭발할 수 있지만 토큰 자체를 사고파는 장사의 마진은 떨어질 수 있다고 생각하기 때문이다. 예를 들어 전기 사용량은 계속 늘어나도, 단순히 전기를 중간에서 되파는 사업은 큰 마진을 남기기 어렵다. 대신 발전 원가가 압도적으로 낮거나, 전기를 이용해 고부가가치 서비스를 만드는 쪽이 돈을 번다.

AI도 마찬가지다. 토큰이 싸질수록 사용량은 늘어난다. 그러나 차별성 없이 남의 모델을 호출해 다시 파는 사업은 가격 경쟁에 빠진다. 이제부터 진짜 부가 가치는 크게 보면 두 곳에 남는다. 토큰이 상품(commodity)이 될수록 돈은 양 끝단으로 이동한다. 첫째는 "성과와 통제를 파는 계층"이다. 기업의 데이터와 권한을 이해하고, 어떤 모델을 어디에 써야 하는지 결정하는 곳이다. 팔란티어가 여기에 해당한다. 둘째는 폭발하는 연산량을 가장 낮은 원가로 처리하거나, 그 물동량 자체에 반복적으로 과금하는 계층이다. 아마존과 같은 하이퍼스케일러나 엔비디아가 여기에 해당한다.

물론 팔란티어의 'AI 주권론'을 그대로 받아들일 필요는 없다. 그러나 이 주장이 힘을 얻는 이유는 현실의 사건이 그 방향을 뒷받침하기 시작했기 때문이다. 외부 API 하나에 의존한 시스템(정부에 의한 Mythos/Fable 셧다운)은 모델 접근이 막히는 순간 함께 멈춘다. 반면 데이터와 관제를 기업이 직접 쥐고 있다면, 모델을 다른 것으로 교체하고 업무를 계속할 수 있다. 지능을 빌려 쓰는 시대가 끝나고, 지능의 생산수단을 소유하려는 시대가 열리고 있다. 이 차이가 AI를 둘러싼 투자 시장의 새로운 색깔을 만들고 있다고 본다.

4. 데이터: 다음 병목이자 마지막 보물

버퍼 타임을 길게 늘어트리는 주범인 '데이터 장벽'은, 뒤집어 보면 새로운 산업 하나를 만들고 있다. 공개 웹 데이터만으로는 더 이상 모델 성능을 빠르게 끌어올리기 어려워지자, 프론티어 연구소들은 이제 전문가 데이터를 돈 주고 '제조'하기 시작했다. 의사, 변호사, 회계사, 엔지니어를 고용해 실제 문제를 어떻게 풀고 어떤 판단을 내리는지 기록하게 하는 것이다.

이 때문에 학습용 데이터 자체가 새로운 산업으로 커지고 있다. 과거의 데이터 라벨링이 사진 속 사물을 표시하는 작업이었다면, 이제는 전문가의 판단 과정과 문제 해결 방식을 AI가 배울 수 있는 형태로 바꾸는 일이 핵심이 됐다. Mercor 같은 회사가 주목받는 이유도 여기에 있다. 데이터센터에 막대한 자본을 쏟는 흐름만큼 이제는 고급 지식을 대규모로 만들어내는 것이 절실해진 셈이다.

* *Mercor: 의사·변호사·엔지니어 같은 전문가를 대규모로 모집해, AI 연구소가 필요로 하는 고품질 훈련 데이터를 만들어주는 회사. 전문가들이 AI의 답변을 평가·수정하고, 문제 풀이 과정과 피드백을 제공하면 Mercor가 이를 받아 RLHF·모델 평가·에이전트 훈련 데이터로 가공해줌. 쉽게 말하면, 전문가의 지식과 판단'을 AI 훈련 데이터로 만들어 파는 회사.*

표 7. Mercor의 ARR 추이: 17개월 만에 40배 상승

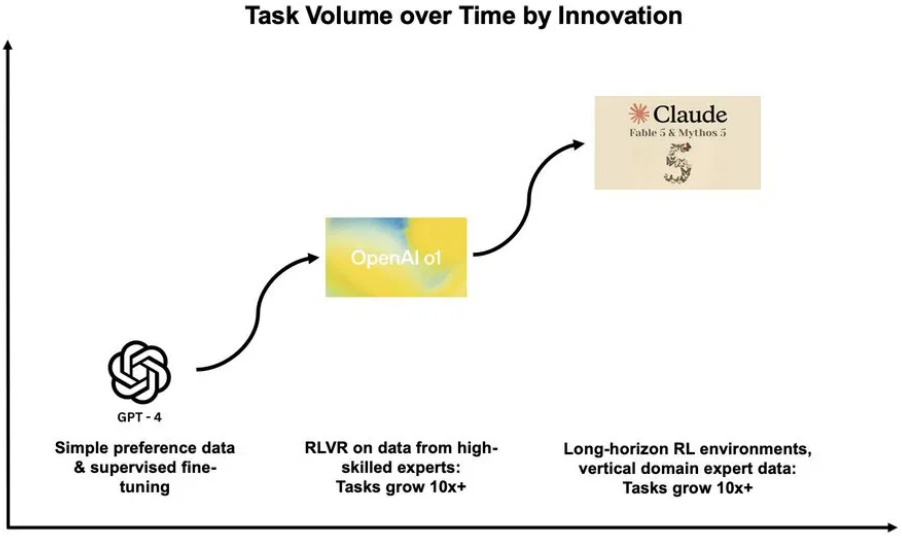
시점	Annualized Revenue Run-rate	변화
2025.01.27	$50M	기준점
2025.02.20	$75M	약 3주 만에 +50%
2025.03	$100M	$100M 돌파
2025.09 초	$450M+	6개월 만에 4.5배+
2025.09	$500M	공식적으로 $500M 도달
2026년 초	$1B+	약 반년 만에 다시 2배
2026.06	$2B+	다시 약 반년 만에 2배

자료: 미래에셋증권 리서치센터

Mercor의 투자사인 Benchmark의 임원 Everett Randle은 "토큰이 추론과 모델 사용의 단위라면, 태스크(task)는 모델 개선의 단위"라면서, 이 Task Economy가 Next 1조 달러 AI 시장이 될 것이라 전망했다. 여기서 태스크란 강화학습의 '연습' 단위다. 모델에게 과제와 환경(예를 들어 계약서 검토라는 과제와 법률 데이터룸이라는 환경)을 주고, 결과물을 채점 기준으로 검증하는 하나의 세트로, 모델에게 '무엇을 할지'가 아니라 '어떻게 할지'를 가르친다.

* *Task Economy: 각 기업이 자기 직원·전문가의 암묵지와 업무 방식을 AI가 반복 연습할 수 있는 '문제 세트'로 바꾸는 산업.*

그림 8. AI 모델의 발전이 'task 수요'를 폭발시키는 과정
GPT-4 때는 사람이 AI 답변을 보고 '좋다·나쁘다'를 평가하는 선호 데이터만으로도 모델을 개선
reasoning 모델 시대에는 AI가 어려운 문제를 많이 연습해야 하므로 태스크가 10배 이상 증가.
장기 에이전트 시대에는 어려운 문제를 만드는 것을 넘어서,
기업들의 '실제 업무 환경'을 다 만들어야 해서 태스크가 또 10배 이상 증가.
즉, 모델이 더 똑똑해질수록 데이터 수요가 줄어드는 것이 아니라
오히려 더 복잡하고 비싼 '연습문제'가 폭발적으로 필요해짐.



자료: Everett Randle, 미래에셋증권 리서치센터

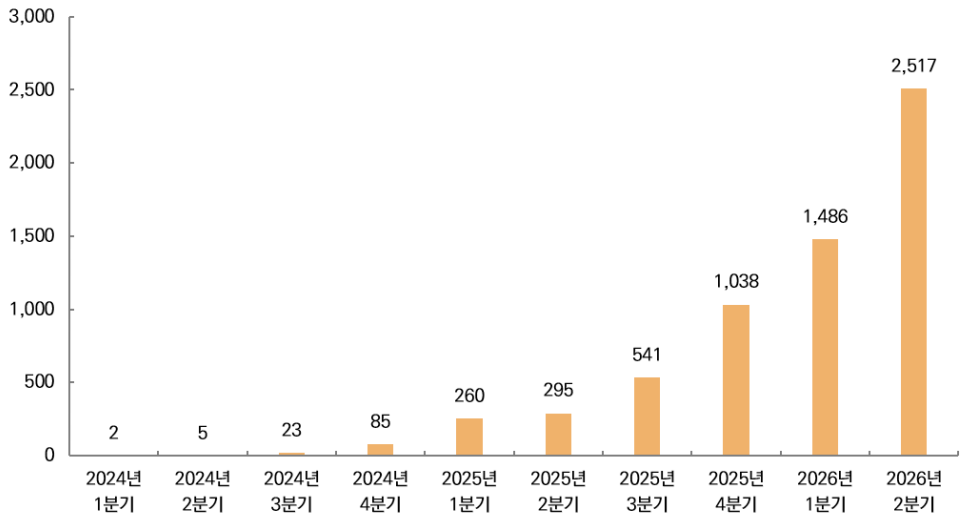
표 8. Task Economy 이해하기

Task 경제에서 어떻게 돈을 쓰나?	실제로 하는 일	기업이 얻는 것
① 전문가 동원	애널리스트, 변호사, 의사 등 현업 전문가를 대규모로 투입해 실제 문제와 모범 답안, 판단 기준을 만듦	직원 머릿속에만 있던 암묵지와 판단 노하우
② 실제 업무 환경 구축	기업 내부 문서, 데이터베이스, 사내 시스템, 업무 도구를 AI가 실제처럼 사용할 수 있는 훈련 환경으로 만듦	AI가 실제 업무를 수행하는 연습장
③ 대규모 태스크 생성	계약 검토, 투자 판단, 리스크 심사 등 실제 업무를 수십만~수백만 개의 과제로 만들어 AI에게 반복 수행시킴	기업 업무를 세분화한 고유한 AI 훈련 데이터
④ 전문가 평가·수정	AI의 결과를 전문가가 채점하고, 틀린 부분을 고치며, 왜 틀렸는지와 더 나은 판단 방식을 기록함	AI가 배울 수 있는 고품질 피드백과 검증 데이터
⑤ 모델 학습·평가	축적된 태스크와 피드백을 활용해 모델을 추가 학습하고, 실제 업무에서 얼마나 잘하는지 계속 평가함	범용 모델 아닌 자사 업무에 특화된 '전용 AI'

자료: 미래에셋증권 리서치센터

Task Economy와 관련한 실제 수치도 가파르게 상승 중이다. 프론티어 연구소들의 데이터 지출은 전년 대비 10배씩 늘고 있고, 태스크 플랫폼의 분기별 전문가 투입 시간은 2024년 1분기 2,000시간에서 2026년 2분기 약 252만 시간으로 10개 분기 만에 1,000배 이상 뛰었다. 다시 말해, AI를 더 똑똑하게 만들기 위해 인간 전문가의 지식과 판단을 '훈련 데이터'로 제조하는 업무량이 2년 반만에 1,260배 늘었다는 의미다.

그림 9. Mercor의 분기별 전문가 투입 시간은 지난 2분기 251.7만 시간으로 증가. (단위: 천 시간)
연구소가 인터넷 데이터를 읽는 단계에서 벗어나,
사람의 머릿속 지식과 판단을 돈을 주고 '제조'하는 단계로 넘어갔다는 증거.
AI의 다음 병목이 컴퓨트만이 아니라
'고급 인간 지식을 얼마나 빠르게 훈련 데이터로 바꿀 수 있는가'로 이동하고 있다는 강력한 증거.



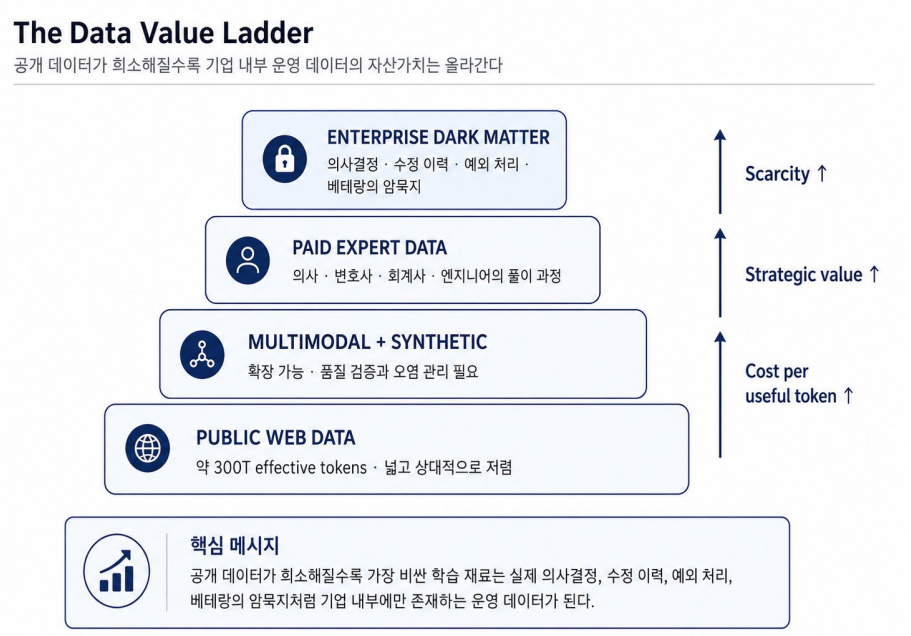
자료: Mercor, Everett Randle, 미래에셋증권 리서치센터

또한 결정적으로, 이 시장의 구매자는 이제 AI 연구소만을 의미하지 않는다. 기성품의 범용 AI 모델과의 차별화를 위해 개별 기업들이 태스크 지출에 1억 달러 이상의 금액을 배정하기 시작했다. Randle의 표현대로 "미래 AI 역량에 필요한 인류 지식의 99%는 사람들의 머릿속에 있고", 그 머릿속 지식을 먼저 데이터로 바꾸는 쪽이 해자를 갖기 때문이다.

물론, 지능의 상품화(commidity)는 모든 영역에서 똑같이 진행되지 않는다. 문서 분류, 요약, 고객 응대 같은 루틴 업무에서는 값싼 모델이 계속 확산될 수 있지만 수학, 사이버보안, 법률, 생명과학처럼 비싼 전문가 데이터가 집중되는 영역에서는 프론티어 모델이 다시 큰 격차를 벌릴 수 있다. 즉, 경제의 많은 업무에서는 good enough 모델이 승리하더라도, 특정 도메인에서는 데이터 해자를 가진 최고급 모델의 가격 결정력이 남을 수 있다.

그럼에도 더 중요한 사실은, AI 연구소들이 가장 탐내는 데이터의 최대 매장지는 인터넷이 아니라 기업 내부에 있다는 점이다. 어떤 고객을 먼저 챙기는지, 어떤 상황에서 예외를 허용하는지, 숙련자가 '이상 신호'를 어떻게 알아채는지 같은 지식은 대부분 공개돼 있지 않은 암묵적 지식이다. 조직의 시스템과 업무 절차, 베테랑의 머릿속에 흩어져 있다.

그림 10. 가장 희소한 데이터는 시장에서 살 수 있는 전문가 지식이 아니다
실제 의사결정·수정 이력·예외 처리·베테랑의 암묵지처럼 기업 내부에만 존재하는 운영 기록이다.
공개 데이터가 고갈될수록 기업의 일상 업무에서 생성되는 '각 기업들이 가진 암흑물질'이 미래 AI를 차별화하는 가장 값비싼 자산으로 올라선다.



자료: 미래에셋증권 리서치센터

그래서 앞서 언급한 관제 계층(control plane)은 바로 이 '암묵지'를 기록으로 바꾼다는 점에서 핵심인 것이다. AI가 틀린 지점, 사람이 고친 내용, 최종 승인된 선택이 쌓일수록 기업은 자기만의 고유한 학습 데이터를 갖게 된다. 따라서 연구소의 데이터 갈증이 커질수록 기업은 한 가지 사실을 깨닫게 된다. "우리의 데이터는 단순한 기록물이 아니라, 우리만의 미래의 지능을 만드는 원재료다."

이게 팔란티어의 AI 주권 논리와 맞닿는 지점이다. 그래서 데이터를 넘기는 것은 파일을 맡기는 정도가 아니라 그 데이터로 더 나은 지능을 만들 권리까지 함께 넘긴다는 것을 의미한다.

IV. 투자 지도와 핵심 수혜주 3선

1. 세 가지 시험

앞서 확인한 구조적 변화가 실제 이익으로 이어질 수 있는지를 세 가지 시험으로 걸러냈다. 세 조건을 모두 통과한 상장기업으로 우선 세 곳을 꼽아봤다.

1. 갈아타기 비용의 비대칭성: 고객이 밑단의 AI 모델은 쉽게 바꿀 수 있어도, 이 회사의 플랫폼이나 인프라는 떠나기 어려운가?

2. 디플레이션 수용성: 토큰 가격과 단위당 컴퓨트 비용이 계속 하락해도, 늘어난 사용량과 처리량 덕분에 오히려 이익이 커지는 구조인가?

3. 주권과 규제의 해자: 데이터 주권과 규제 요구가 복잡해질수록, 고객이 이 회사에 더 의존하게 되는가?

이는 앞서 확인한 세 구조를 그대로 종목 선별 기준으로 바꾼 것이다. 갈아타기 비용은 모델보다 관제에 쌓이고, 가격 하락은 처리량 폭발을 부르며, 모델 규제와 AI 주권 요구는 이를 감당할 수 있는 기업들의 해자를 넓힌다. 이 세 시험을 모두 통과한 기업이 팔란티어, 아마존, 엔비디아다. 셋은 서로 다른 층에 서 있지만 공통점이 있다. 지능 그 자체가 아니라, 지능이 기업과 국가에 배포되는 길목을 장악한다.

2. 팔란티어: “우리 데이터를 지키면서 AI 어떻게 통제할까”에서 필요

(1) 시험 1. 모델은 바꿀 수 있지만, 업무의 기억은 옮기기 어렵다

팔란티어의 온톨로지는 기업의 데이터, 권한, 업무 규칙을 하나의 구조로 엮은 일종의 ‘회사의 디지털 지도’다. 누가 어떤 데이터를 볼 수 있는지, 어떤 판단은 누구의 승인을 거쳐야 하는지, AI가 실패했을 때 사람이 어떻게 고쳤는지가 그 위에 쌓인다. 이 구조가 한번 업무에 깊이 들어가면 밑단의 모델을 바꾸는 것은 쉬워도 플랫폼을 떠나는 것은 어렵다. 모델 교체는 자유롭지만, 회사의 기억과 권한 체계를 통째로 옮기는 것은 어렵다.

6월 Mythos/Fable 셧다운 사태는 이 차이를 실증했다. 외부 API 하나에 업무를 직접 연결한 기업은 모델 접근이 막히는 순간 함께 멈췄다. 반면 데이터와 관제를 직접 쥔 기업은 밑단의 모델을 다른 것으로 교체하고 업무를 이어갈 수 있었다. 이 비대칭성이 팔란티어의 가장 강한 락인이다.

(2) 시험 2. 지능이 싸질수록, 통제의 값은 오히려 올라간다

팔란티어는 토큰을 팔지 않는다. 토큰 위에서 무엇을 어떤 모델에 맡기고, 누가 승인하며, 무엇을 기록할지를 판다. 오픈 모델이 공짜에 가까워질수록 “이 값싼 지능을 우리 회사 시스템에 어떻게 안전하게 붙일 것인가?”라는 질문이 중요해진다.

모델이 많아질수록 선택과 통제가 어려워지고, AI 사용량이 늘어날수록 권한·감사·승인 체계의 중요성도 커진다. 그래서 팔란티어는 지능의 디플레이션을 피하는 것이 아니라, 그 디플레이션이 만들어내는 복잡성에 과금한다.

(3) 시험 3. 주권 요구를 가장 먼저 제품 언어로 바꿨다

팔란티어의 AI 주권 선언은 자사 제품을 파는 내러티브이기도 하다. 그러나 중요한 것은 이 서사가 실제 제품 구조와 결합하고 있다는 점이다. 엔비디아와의 파트너십이 대표적이다. 기업은 개방 모델을 가져와 자기 데이터로 추가 훈련하고, 만들어진 가중치를 자기 통제 아래 둘 수 있다. 팔란티어의 온톨로지는 그 모델을 실제 데이터·권한·업무에 연결한다.

즉, "개방 모델 + 자기 데이터 + 자기 가중치 + 온톨로지"라는 소버린 AI 스택의 조립자가 되는 것이다. 팔란티어의 가장 큰 투자 논거는 바로 여기에 있다. AI 시장이 모델 성능 경쟁에서 누가 데이터와 가중치, 업무 권한을 통제하느냐의 경쟁으로 이동할수록 이 회사의 전략적 위치가 좋아진다. 결국 팔란티어의 해자가 얼마나 오래 유지되는지는 소프트웨어 기능 자체보다 기업 보안 인증, 접근 제어, 감사 체계, 실제 고객의 축적된 업무 데이터가 얼마나 쉽게 대체될 수 있는지에 달려 있다.

(4) 실적 연결 고리: 무엇이 숫자로 찍혀야 하는가

AI 업무량 증가 → Ontology/AIP 안으로 더 많은 핵심 업무 유입 → 대형 계약·고객당 지출 확대 → 매출 성장 + 영업 레버리지

이에 따라 팔란티어의 실적에서 항상 필수적으로 확인해야할 지표는 3가지다. “기존 고객의 계약 규모가 계속 커지는지, 대형 계약과 계약잔고가 늘어나는지, 고성장과 동시에 영업이익률이 확대되는지”다. 참고로 이 세 가지는 이제껏 모두 우상향 추세를 그리는 중이다.

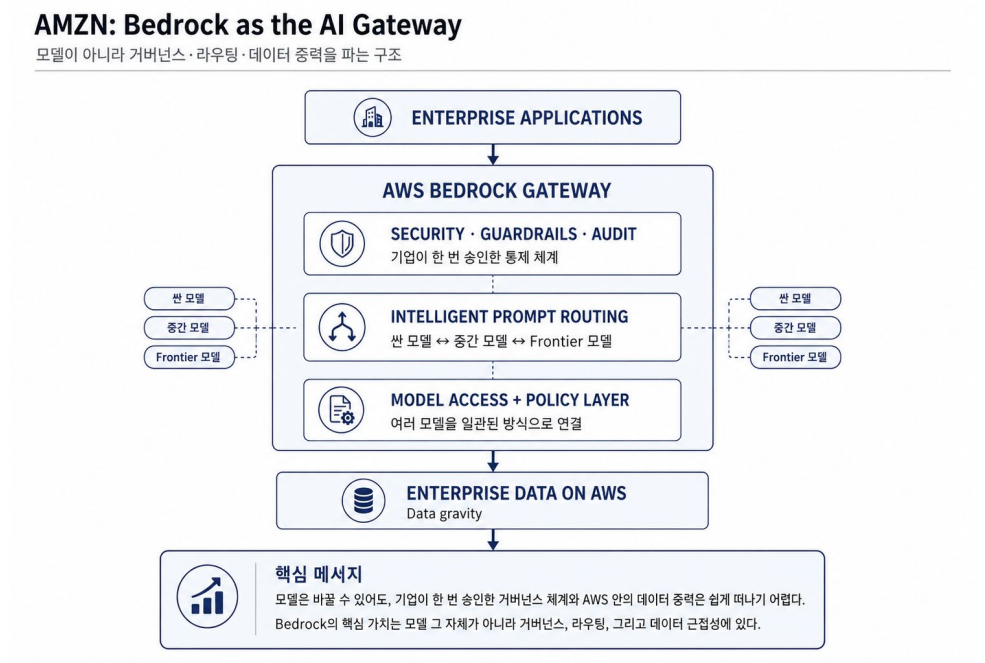
3. 아마존: 여러 값싼 지능을 가장 낮은 원가로 돌리는 곳

(1) 시험 1. Bedrock의 락인은 모델이 아니라 거버넌스다

AWS의 Bedrock은 여러 회사에서 제공하는 AI 모델을 한 플랫폼 안에서 골라 쓰게 해 주는 플랫폼이다. 중요한 것은 모델 선택권 자체가 아니다. 기업 보안팀이 한번 접근 제어, 데이터 정책, 감사 체계를 승인하고 나면, 그 안에서 모델을 바꾸는 것은 쉽지만 플랫폼 밖으로 나가는 것은 훨씬 어렵다.

여기에 AWS의 '데이터 중력'이 붙는다. 기업의 핵심 데이터가 이미 AWS에 들어가 있다면, AI 연산 역시 그 데이터 가까이에서 수행하는 것이 비용과 지연 시간 면에서 합리적이다. 그래서 AWS의 락인은 이렇게 작동한다. 모델은 바꿀 수 있지만, 데이터와 보안 체계가 있는 관문은 떠나기 어려운 것이다. 이 구조는 프론티어 모델 하나의 승패와 무관하게 AWS가 계속 길목에 남을 수 있게 한다.

그림 11. AWS Bedrock의 핵심은 특정 모델을 파는 데 있지 않고,
기업이 승인한 보안·가드레일·감사 체계 안에서 여러 모델을 사용할 수 있게 하는 것.
쉬운 업무는 싼 모델로, 어려운 업무는 중간·프런티어 모델로 보내는 라우팅을 통해
기업은 성능과 비용을 동시에 최적화할 수 있다.
밑단의 모델은 계속 교체할 수 있지만, 기업의 보안 정책과 핵심 데이터가 이미 AWS 안에 쌓일수록
그 관문을 떠나는 비용은 높아진다.



자료: 미래에셋증권 리서치센터

(2) 시험 2. 지능의 가격이 떨어질수록 원가 구조가 중요해진다

아마존은 디플레이션을 두 가지 방식으로 흡수한다. 첫째는 가동률 경제학이다. 전 세계 고객의 서로 다른 시간대와 업무량을 겹쳐 GPU와 가속기를 최대한 놀리지 않는 능력은 글로벌 초대형 클라우드의 핵심 원가 우위다. 토큰 단가가 떨어질수록 작은 사업자는 마진 압박을 받지만, 높은 가동률을 유지하는 사업자는 낮은 가격에서도 살아남을 수 있다.

둘째는 자체 칩이다. 추론용 Inferentia와 학습용 Trainium은 외부 GPU만 사다 임대하는 구조보다 원가를 직접 통제할 수 있게 한다. 값싼 오픈 모델을 자체 칩 위에서 대량으로 돌리는 세계는 아마존의 원가 구조와 잘 맞는다.

즉, good enough 모델의 확산은 AWS에 단순한 트래픽 증가가 아니라, 자체 칩과 높은 가동률의 경제성을 증명하는 시장이다.

(3) 시험 3. 규제가 복잡할수록 관문은 더 중요해진다

나라별 데이터 보관 규정, 정부 인증, 접근 권한, 보안 정책을 모두 맞추는 것은 쉽지 않다. AWS는 수십 년 동안 이 인프라를 쌓아왔다. 셧다운 국면에서도 아마존의 위치는 독특했다. Anthropic의 투자자이자 인프라 파트너이고, Bedrock을 통한 유통자이면서, 자체 모델과 칩을 가진 경쟁자이기도 하다.

어떤 모델이 이기든 기업 수요가 AWS 안에서 움직인다면 아마존은 길목에 남는다.

(4) 가장 흉한 현금흐름이 오히려 파종기의 증거일 수 있다

아마존의 최근 12개월 잉여현금흐름은 AI 설비투자 확대로 전년 대비 95% 급감해 약 12억 달러까지 떨어졌다. 시장은 이를 수익성 훼손으로 본다.

아마존은 지금 AI 인프라를 직접 깔고, 직접 칩을 만들고, 그 위에서 모델을 유통하는 회사가 되려 한다. 현재의 현금흐름 악화가 장기적인 컴퓨트 원가 우위와 관문 지배력으로 전환된다면, 지금은 수확기가 아니라 파종기로 봐야 하는 것이다.

물론 핵심 리스크는 명확하다. 막대한 CapEx가 충분한 사용량과 매출로 이어지지 않으면, 이 논리는 무너진다. 결국 아마존의 투자 판단은 AI 인프라 가동률과 AWS AI 매출이 자본지출을 얼마나 빠르게 흡수하는가에 달려 있다.

(5) 실적 연결고리와 다른 CSP들과의 비교

앞으로 투자자는 "추론 수요 증가 → AWS AI 사용량·가속기 가동률 상승 → 고정비와 감가상각 흡수 → AWS 매출 성장 + 장기 마진 개선"의 그림이 지켜지는지만 확인하면 된다.

그런데 왜 MSFT는 차순위인가? 기업용 소프트웨어와 클라우드의 락인은 강하다. 다만 본 보고서가 찾는 것은 최대한 모델에 중립적인 관문과 주권 구조다. 자체 모델 생태계와 전략적 이해관계가 강할수록 완전한 중립 관문이라는 서사는 약해진다.

구글은 아마존과 같이 자체 모델과 자체 칩의 원가 우위는 강력하다. 그러나 본 보고서의 핵심인 기업별 관제·권한·멀티모델 중립성의 수혜는 PLTR·AWS보다 직접적이지 않다고 판단했다.

또한 오라클은 모델 선택과 AI 애플리케이션 관문이라는 전체 생태계의 폭은 상대적으로 더 검증해야 할 필요가 있어 아마존보다는 약한 주자라 생각한다.

4. 엔비디아: AI 인프라 건설의 돈줄을 쥔다

공급망을 직접 추적하는 일부 리서치 진영은 올 하반기 엔비디아 데이터센터 매출을 컨센서스보다 약 20% 높게 보고 있다. "AI GPU 붐"이 4년차인데도 시장이 여전히 물량을 과소평가하고 있다는 주장인 것이다. 그러나 더 중요한 것은 물량이 아니다. 시장이 아직 완전히 가격에 넣지 않은 것은 사업 모델의 변화로 생각한다.

(1) 시험 1. CUDA 위에 금융 락인이 추가된다

기존 엔비디아의 해자는 CUDA를 중심으로 둔 모델-칩 공동 최적화였다. 그러나 이번 달 초에 나온 최저 수익 보증(backstop)은 그 위에 새로운 락인까지 만든다. GPU 클라우드 사업자는 엔비디아의 신용을 바탕으로 대출을 받고, 그 자금으로 엔비디아 GPU를 산다. 사업이 잘되면 초과 수익 일부를 엔비디아와 나눈다.

이 구조에 들어온 사업자가 다른 칩으로 갈아타는 것은 단순한 하드웨어 교체가 아니다. 대출 구조와 보증 관계, 수익 배분 계약까지 다시 짜야 한다. 즉, CUDA가 기술적 락인이라면, backstop은 금융적 락인이다. 보증을 받은 GPU 클라우드 사업자는 엔비디아 생태계의 장기 파트너가 된다.

(2) 시험 2. 일회성 판매를 처리량 연동 수익으로 바꾼다

칩을 한 번 파는 사업은 원리상 가격 하락과 세대 교체에 취약하다. 그러나 수익 배분 모델은 다르다. GPU 임대 수요가 늘고 가동률이 올라갈수록 엔비디아가 가져갈 수 있는 초과 수익도 커진다. 토큰 단가가 떨어져 전체 사용량이 폭발하면, 엔비디아는 칩 판매뿐 아니라 그 칩이 만들어내는 임대 수익에도 참여할 수 있다.

그래서 이 구조의 핵심은 "GPU 판매 → 신용 보강 → 인프라 건설 → 임대 수익 → 수익 배분"으로 이어지는 긴 수익 사슬이다. 엔비디아는 추론 슈퍼사이클의 물동량 자체에 반복적으로 과금하는 장치를 만들고 있다.

(3) 시험 3. 세계가 쪼개질수록 새로운 고객이 생긴다

소버린 AI는 엔비디아의 지정학적 수혜 논리다. 미국 빅테크 클라우드를 쓰기 어렵거나 쓰고 싶지 않은 국가들은 자국 안에 AI 인프라를 만들려 한다. 그러나 모든 국가가 자체 칩과 소프트웨어를 처음부터 개발할 수는 없다. 그래서 등장하는 절충안이 글로벌 기술은 쓰되, 데이터와 운영 통제는 로컬에 남긴다는 구조다.

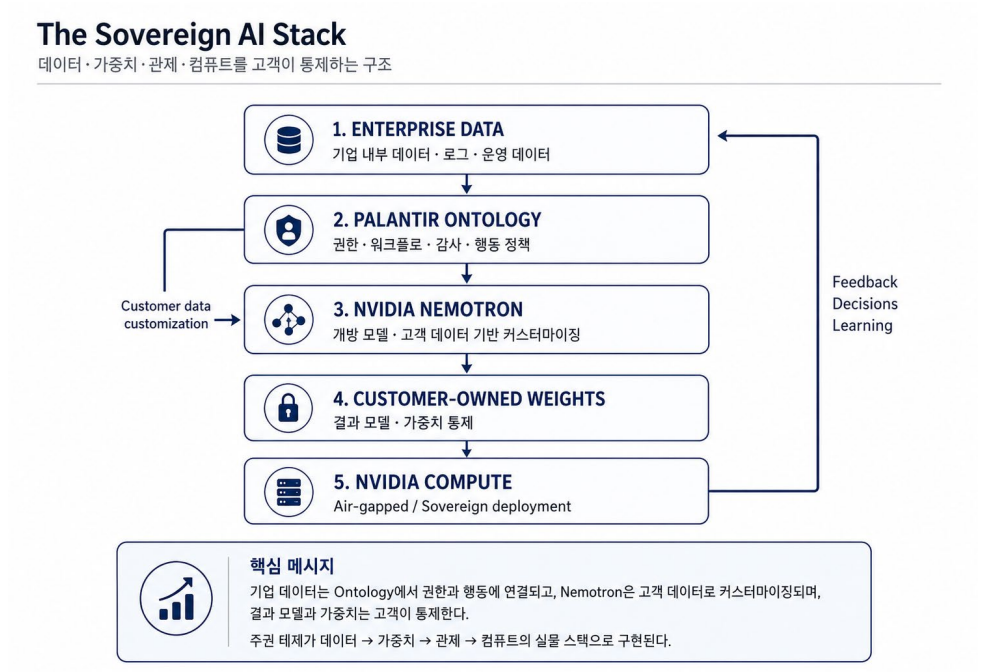
엔비디아는 GPU와 소프트웨어를 공급하고, 경우에 따라 금융 구조까지 지원한다. 세계가 국가별로 쪼개질수록 새로운 로컬 클라우드와 소버린 AI 프로젝트가 생기고, 그 공통분모를 엔비디아가 공급한다.

(4) 엔비디아의 히든 병기: 오픈소스

여기에 하드웨어와 금융에 이어 세 번째 다리가 추가됐다. 오픈소스 모델이다. 그들의 모델 Nemotron은 기업이 검증하고 자기 데이터로 추가 훈련해 사용할 수 있는 모델을 목표로 한다. Artificial Analysis의 개방성 지수에서는 Nemotron 3 Ultra가 18점 만점에 15점으로, 학술기관 모델을 제외한 상업 모델 중 가장 높은 수준으로 평가됐다.

이 전략의 의미는 단순하다. 폐쇄형 연구소들이 토큰을 팔 때 엔비디아는 모델을 공짜에 가깝게 풀고, 그 모델이 돌아갈 칩을 팔고 그 칩을 살 금융을 열어주고 그 위에서 발생하는 임대 수익에 참여한다는 것이다. 엔비디아는 지능의 상품화를 누구보다 원하는 기업이고 지능의 상품화를 자기 하드웨어와 금융 생태계를 확장하는 무기로 쓰는 것이다.

그림 12. 엔비디아 + 팔란티어: Sovereign AI Stack
기업 내부 데이터는 팔란티어의 온톨로지를 통해 권한·업무 흐름·승인·행동 규칙과 연결된다.
이 데이터를 바탕으로 엔비디아 Nemotron 오픈 모델을 기업 업무에 맞게 조정하고,
그 결과 만들어진 모델과 가중치는 고객이 직접 통제한다.
기업 고객에게 맞춤화된 AI 모델은 고객이 원하는 격리된 환경과 컴퓨트 위에서 실행되고,
실제 업무의 결정·수정·피드백은 다시 다음 학습 데이터로 축적된다.
즉, AI 주권은 "'데이터 → 관제/하네스 → 오픈 모델 및 가중치 → 컴퓨트를 직접 소유"하는 스택.



자료: 미래에셋증권 리서치센터

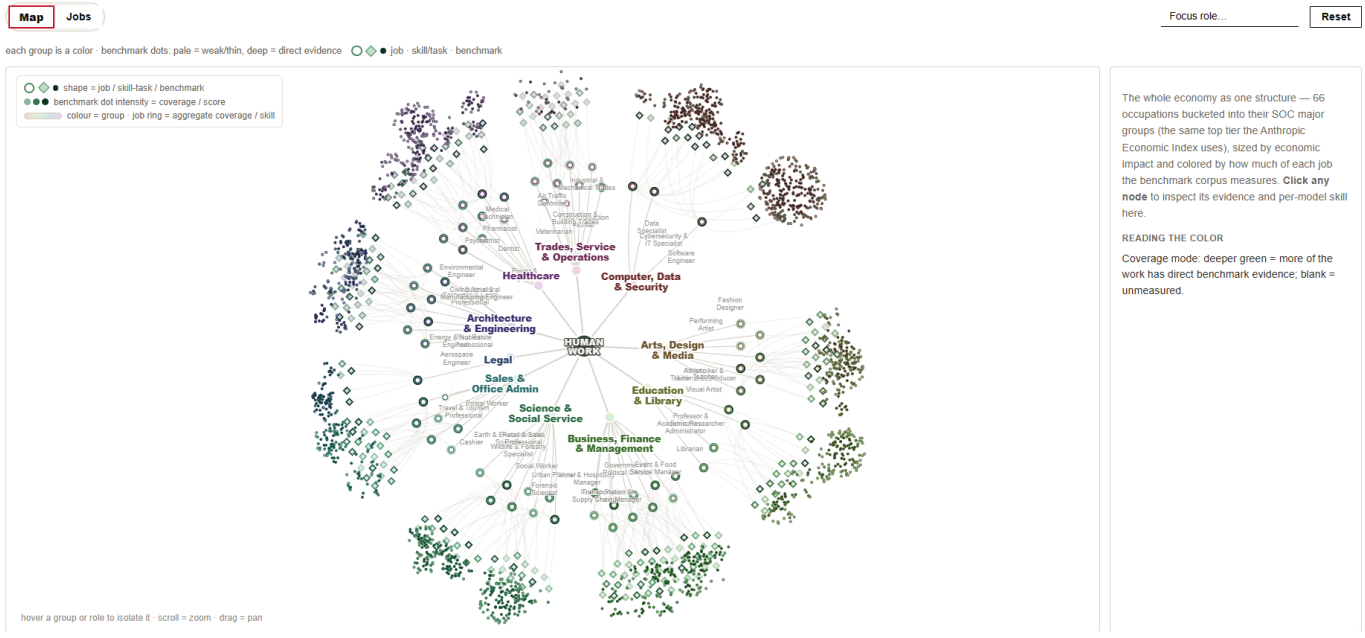
V. 반증과 결론

1. 본 보고서의 주장이 틀리는 경우는 프론티어의 '역습'이 있을 때임

압도적인 범용 지능의 등장으로, 모델 하나가 업무 분류, 도구 선택, 권한과 보안 정책까지 스스로 처리하면 외부 관제 계층은 껍데기가 된다. 즉 '모델은 부품이고 관제 계층이 중요하다'는 본 보고서의 전제가 무너진다. 다만 이 극단 시나리오에도 방어 논리가 있다. 권한과 감사는 기술이 아니라 지배구조의 문제다. 시험 감독관과 수험생을 같은 존재에게 맡길 수 없듯, AI가 자기 행동을 스스로 승인하고 감사하는 구조를 규제기관과 기업이 받아들이기 어렵다.

사실 진짜 위험은 훨씬 현실적인 형태로 올 수 있다. 완전한 AGI가 아니라, 비싼 전문가 데이터로 무장한 프론티어 모델이 도메인 영역에 하나씩 하나씩 도장깨기를 하는 시나리오다. good enough 모델이 활약할 수 있는 유효 영역 자체를 잠식해 들어오는 경우다.

그림 13. 프론티어 모델의 도장깨기의 전황: 인류의 '일' 중에서 AI 벤치마크가 실제로 채점할 수 있는 부분이 얼마나 되는가?
10여 개 직군(헬스케어, 법률, 건축·엔지니어링, 컴퓨터·데이터·보안, 영업·사무, 예술·미디어 등)이 존재.
각 직군 아래 66개 직업(원), 그 직업을 구성하는 스킬·태스크(마름모), 그리고 그 태스크를 측정하는 벤치마크(작은 점)가 존재.
진한 색은 그 일에 대한 직접적 벤치마크 증거가 존재, 옅거나 빈 곳은 측정된 적 없음(unmeasured)을 의미.
컴퓨터·데이터·보안 클러스터는 점이 빽빽하고 진한데 트레이드·서비스·운영, 헬스케어의 상당 부분, 예술 계열은 옅거나 비어 있음.



자료: BenchmarkList, 미래에셋증권 리서치센터

이 시나리오를 이해하려면 먼저 'AI는 왜 하필 코딩에서 가장 먼저 초인 수준에 도달했는가'에 대해 고민해보면 된다. 코딩이 쉬워서가 아니다. 인터넷에 공짜 훈련 데이터가 산더미로 쌓여 있었다(GitHub의 수십억 줄 코드). 또한, 정답 확인이 자동으로 된다(코드는 돌려 보면 맞는지 틀리는지 즉시 판명되므로, 모델이 무한히 자가 훈련할 수 있다). 요컨대 코딩은 '데이터가 공짜이고 채점이 공짜인' 희소한 대형 도메인이었고, 그래서 빠르게 정복됐다.

AI 연구소들은 이 성공에서 공식을 배웠다. "도메인 데이터 + 검증 가능한 피드백 = 능력 도약". 문제는 여러 도메인에서의 데이터는 전혀 공짜가 아니라는 점이다. 인터넷의 쓸모 있는 공개 텍스트는 이미 소진됐고, 모델이 회계나 법률에서 코딩만큼 못하는 이유는 지능의 한계가 아니라 데이터 커버리지의 한계다.

그래서 연구소들은 이제 데이터를 '제조'한다. 의사, 변호사, 회계사, 엔지니어를 고용해 문제 풀이 과정을 통째로 기록하게 하는 것이다. 이 지출은 현재 연 70억 달러 규모이고 2030년까지 1,000억 달러 이상으로 커질 전망이며, 창업 3년 만에 수백만 명의 전문가 라벨러를 조직해 연 매출 20억 달러 안팎에 이른 Mercor 같은 회사가 그 실물이다. 실제로 특정 연구소의 수학, 특정 연구소의 사이버보안처럼, 비싼 데이터가 집중 투입된 영역에서는 프론티어와 Good Enough 간의 격차가 다시 벌어지고 있다.

이 시나리오가 실현되는 모습은 이렇다. 프론티어가 법률 도메인의 전문가 데이터에 수억 달러를 태운다. 6개월 뒤, 법률 업무에서 'good enough'의 기준선 자체가 올라간다. 어제까지 오픈 모델로 충분했던 계약서 검토가, 이제는 프론티어만 통과할 수 있는 품질 기준의 업무가 된다. 그 도메인에서 값싼 모델은 다시 밀려나고, 프론티어 모델은 프리미엄 가격의 결정력을 되찾는다. 같은 일이 회계에서, 의료에서, 보험 심사에서 순차적으로 반복된다. good enough의 영토 80%가 한꺼번에는 아니어도 한 조각씩 수복당하는 것이며, 본 보고서의 버퍼 타임은 도메인별로 '조기 마감'될 수 있다.

다만 이러한 시나리오에도 제동 장치가 있다. 첫째, 경제성의 필터다. 전문가 데이터 제조에는 큰 돈과 시간이 들므로, 정복은 가치 상한이 극히 높은 도메인(법률, 의료, 금융과 같은 규제 산업)부터 순차적으로 진행될 것이다. 문서 분류나 고객 응대 초벌처럼 가치 상한이 낮은 루틴 업무는 비싼 데이터를 태울 경제적 이유가 없어, good enough의 본진은 오래도록 남을 수 있다.

둘째, 기업 데이터의 방벽이다. 연구소가 시장에서 살 수 있는 것은 '일반 전문가'의 지식까지다. 특정 회사의 업무 절차와 베테랑의 노하우는 마치 그 기업의 임직원들이 갖고 있는 '암흑물질'과 같은 것으로써 시중에 팔리는 물건이 아니다. 법률회사인 Harvey가 자사 데이터로 오픈 모델을 훈련해 프론티어 모델을 이긴 사례는 기업이 자기 데이터로 맞불을 놓을 수 있음을 보여준다. 게다가 개별 기업들은 1억 달러 이상의 예산으로 본인들의 암묵지를 task 데이터로 바꾸기 시작했다. 이는 "Harvey식 맞불 작전"이 하나의 사례가 아니라 전반적인 전략이 되고 있다는 뜻이다.

2. 결론: 지능이 아니라 길목이 희소해졌다

현재의 AI 투자는 더 이상 누가 가장 좋은 모델을 만드는지를 맞히는 과학 경연이 아니다. 이익률과 투자수익률과 규제 준수를 따지는, 지루하지만 돈이 되는 기업용 인프라 경제학이 되었다. AI가 기업의 실전 배치 단계로 들어오면서 경쟁의 중심이 지능의 생산에서 지능의 유통과 소유로 이동하고 있다.

이때 필자가 주장한 '버퍼 타임'에서 이러한 전환 생태계가 핵심으로 부상한다. '버퍼 타임'의 논리를 마지막으로 요약한다. 최고급 지능은 정부의 규제 탓에 모든 곳에 바로 배포되지 못하고(셧다운의 제도화), 고품질 데이터 부족 때문에 모든 산업과 업무로 즉시 확장되지도 못한다. 물론 그 사이 기업은 기다리지 않고 그 공백을 '충분히 좋은' 지능의 대량 배포로 메운다(리눅스 모멘트).

"good enough" 모델의 배포는 4가지를 동시에 낳는다. 추론 수요의 폭발(제본스), 관제 계층의 부상(갈아타기 비용을 차지), 건설 자금의 새 공급자(엔비디아의 중앙은행화), 그리고 소유권의 각성(AI 및 데이터 주권)이다. 인구의 0.2%와 99.8% 사이의 체감 격차는 거품의 증거가 아니라, 아직 청구되지 않은 수요의 크기로 봐야 한다.

팔란티어, 아마존, 엔비디아. 셋은 소프트웨어, 클라우드, 반도체+금융이라는 서로 다른 층에 서 있다. 그러나 "지능은 흔해졌고 지능의 안전한 배포와 그 소유권은 희소하다"는 사실에 과금한다는 점에서는 한 팀이다. 사실 이 리포트를 쓰면서, 남기고 싶은 단 하나의 문장을 생각해봤더니 꽤 단순했다.

"When intelligence is everywhere, sovereignty becomes scarce."

표 9. 이 보고서의 유효기간을 결정짓는 주요 지표 일곱가지

무엇을 체크해야 하나?	좋은 신호	위험 신호
값싼 AI가 최고급 AI를 계속 따라잡는가?	오픈·저가 모델이 몇 달 차이로 최고급 (공개) 모델을 계속 추격	격차가 1년 이상 벌어지고 루틴 업무도 따라잡지 못함
AI가 싸질수록 사람들이 정말 더 많이 쓰는가?	토큰 가격은 내려가도 전체 사용량은 더 빠르게 증가	가격이 내려가는데도 총 사용량까지 감소
기업이 모델보다 '관제 플랫폼'에 더 의존하는가?	핵심 업무와 데이터가 Bedrock·팔란티어 플랫폼 안으로 이동	실험(PoC)만 늘고 실제 업무 및 데이터 이전은 정체
GPU 임대시장은 건강한가?	가동률이 안정적이고 엔비디아의 지급 보증이 발동되지 않음	GPU 임대료 급락 또는 엔비디아가 실제 손실 보전
기업들이 'AI 주권'을 강하게 요구하는가?	자체 가중치, 온프레미스, 데이터 비학습 조건이 계약에 출현	이런 요구가 실제 구매 조건으로 확산되지 않음
기업 내부 데이터의 값이 정말 올라가는가?	전문가 데이터 지출과 기업 데이터 활용 계약이 증가	데이터 확보 비용과 기업의 데이터 보호 요구가 약해짐
강자의 지배력이 규제의 표적이 되는가?	별다른 제재 없이 시장 지배력 확대	엔비디아나 CSP에 대한 공식 반독점 조사 시작

자료: 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 조사분석 대상법인의 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.