Knowledge Cutoff | 2025. 9. 10

MIRAE ASSET
미래에셋증권

AI 현황 보고서

풀스택 헤게모니에 관하여

한종목 chongmok.han@miraeasset.com

Contents

[서론] AI 혁명의 입체 지도: 기술, 자본, 패권의 흐름을 읽다

"AI 현황 보고서"는 인류가 경험하지 못한 속도와 규모로 펼쳐지는 거대한 패러다임 전환의 한복판에서 길을 잃지 않기 위해, 입체적인 지도를 만들기 위해 제작하였습니다. 과거 100년의 과학 발전이 향후 10년 단위로 압축되어 일어나는 '압축적 도약'의 시대, 우리는 AI가 보조 도구를 넘어 산업 구조와 국가의 명운, 그리고 개인의 생존 방식을 근본적으로 재정의하는 특이점을 목도하고 있습니다. 본 보고서는 독자분들께서 눈앞의 파편적인 정보가 아닌 거대한 조류의 방향을 읽고 전략적 사고의 틀을 갖추도록 돕는 것을 목표로 합니다. 보고서의 방대한 여정을 효과적으로 탐색하기 위해, 독자들께서는 다음과 같은 구조적 흐름에 따라 각 장에 접근하시길 권합니다.

먼저 1장과 2장은 이번 혁명의 기술적 토대와 그 파괴력을 이해하는 출발점입니다. 1장은 AI 발전의 경이로운 속도가 과학, 산업, 고용 시장에 미치는 전방위적 충격을 파노라마처럼 펼쳐 보입니다. 2장은 그 현상 이면에 있는 기술적 엔진의 변화, 즉 '사전학습(Pre-training) 시대의 종말'과 '추론(Inference) 및 강화학습(RL) 스케일링'이라는 새로운 패러다임의 등장을 심층 분석합니다. 이 두 장을 통해 독자께서는 AI 혁명이 왜 과거와 근본적으로 다른지에 대한 명확한 시각을 확보하게 될 것입니다.

기술적 토대를 이해했다면, 이제 전쟁의 주역들과 그들이 쏟아붓는 천문학적 자원의 실체를 마주하게 됩니다. 3장은 구글, OpenAI, xAI, Anthropic, 메타 등 AI 패권을 두고 경쟁하는 거인들의 전략과 강점, 약점을 면밀히 비교 분석합니다. 이어지는 4장과 5장은 이 경쟁이 연간 수백조원을 초과하는 자본 지출(CAPEX) 전쟁으로 비화하는 양상과, 이로 인해 '전력'이라는 가장 근본적인 인프라가 어떻게 AI 시대의 핵심 병목으로 부상했는지를 보여드립니다. 이 부분을 통해 AI 경쟁의 본질이 결국 물리적 자원의 확보 전쟁임을 이해하게 될 것입니다.

이 경쟁은 국가 대항전의 양상으로 확장됩니다. 6장은 '소버린 AI'를 둘러싼 미중 기술 패권 경쟁과 '페타플롭스 동맹'이라는 새로운 지정학적 질서의 탄생을 다룹니다. 이에 맞서는 중국의 대응 전략은 7장에서 DeepSe다, 알리바바, Moonshot이라는 '국가대표' 기업들의 사례를 통해 구체적으로 살펴봅니다. 그리고 8장은 이 모든 경쟁의 판도를 좌우하는 인프라 제국의 심장, 엔비디아가 어떻게 CUDA 생태계와 3계층 네트워크 전략을 통해 해자를 구축하고 있는지를 분석합니다.

마지막으로, 이 거대한 변화가 어떻게 현실 세계와 기업의 운영 체계에 구체적으로 구현되는지 조망합니다. 9장은 테슬라의 로보택시를 통해 AI가 물리 세계와 결합하는 '에이전트' 시대의 서막을 보여주며, 10장은 팔란티어의 사례를 통해 AI 모델을 기업의 실제 운영에 통합하고 통제하는 '온톨로지'의 중요성을 역설합니다. AI는 더 이상 소프트웨어나 알고리즘만의 영역이 아니며, 에너지, 반도체, 데이터, 지정학, 그리고 물리적 구현에 이르는 '풀스택(Full-stack)'을 장악하는 자가 미래의 부와 권력을 지배하게 될 것이라는 사실입니다.

부디 이 보고서가 격변의 시대를 항해하는 독자 여러분께 깊은 통찰을 제공하는 나침반이 되기를 바랍니다.

I. AI와 사회 변화

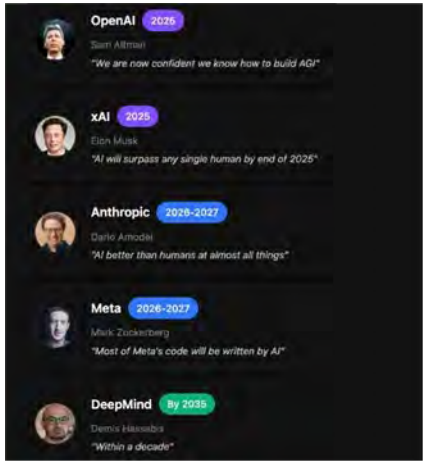
고용충격 비용혁명 에이전트

I. AI와 사회 변화: AI 발전의 특이점, 기하급수적 가속화

현기증 나는 발전 속도: 1년이 25년이 되는 세상

- AI 개발 속도는 인간의 추종을 불허하는 수준. 단순 진보가 아닌 압축적 도약이 현실화.
- Forethought 라는 연구기관의 분석 결과, 과거 100년(1925-2025)의 과학기술 발전이 향후 10년 단위로 압축되어 발생할 것. AI 시스템의 총체적 인지 능력은 매년 약 25배 이상 증가하기 때문.
- 인간 연구 역량과 동등해지는 시점부터 AI의 연구 역량은 매년 최소 25배씩 성장. 1년의 AI 연구가 2년 후 625년의 인간 연구 성과와 동일. AI의 인지 능력은 복합적 요인이 기하급수적 성장을 견인.
  - 훈련 연산량(연간 4.5배 증가) + 알고리즘 효율성(동일 연산량 대비 성능 매년 3배 향상) + 도구 및 프롬프팅 등 스캐폴딩 최적화(모델 능력 연간 3배 개선) + 추론 비용(매년 1/10로 감소) + 추론 컴퓨팅 총량(연간 2.5배 증가)
- 최상위 AI 모델의 지능은 이미 인간의 특정 영역을 초월. 성능 개선 그래프는 수직에 가까운 기울기. OpenAI의 o3 모델은 박사 수준 벤치마크(GPQA)에서 87% 성능 기록. 18개월 전 GPT-4의 24% 대비 폭발적 성장.
- 샘 알트만과 일론 머스크는 2025년 말을 AGI 등장 시점으로 예측.AI 선도 기업들은 대중에 공개된 모델보다 3~9개월 앞선 내부 모델을 보유. 현재 우리가 경험하는 AI는 이미 과거 버전임을 유념해야…

AI guru들의 AGI 도래 시기 전망



자료: Jason Botterill, 미래에셋증권 리서치센터

초지능에 관한 샌프란시스코 컨센서스
"2030년 특이점이 온다"

예측 시기	주요 예측 내용
1년 내	프로그래머 대다수가 AI로 대체 최고 대학원 수학 프로그램의 수리능력
3-5년 내	어떤 분야든 가장 똑똑한 인간 전문가만큼 똑똑한 범용 인공지능이 온다
6년 내	모든 인간 지능 합친 것보다 더 똑똑한 초지능의 도래

자료: 미래에셋증권 리서치센터

2025년에 출시된 주요 AI 모델 관련 정리
IQ 점수는 실제 멘사 가입이 인증 되는 "오프라인 테스트" 기준

개발사	모델명	IQ 점수	비고
OpenAI	GPT-4.1	-	GPT-4o의 마이너 업데이트
	o3-pro	109	o1 모델의 정식 후속작
	o4-mini	86	고급 Reasoning 모델의 증류
	GPT-5-pro	123	전세계 최고의 지능
Anthropic	Claude Sonnet 4	107	균형 잡힌 고급 모델
	Claude Opus 4	118	코딩 성능 한정 최강의 모델
구글	Gemini 2.5 Flash	84	코딩 시장 겨냥한 경량화 버전
	Gemini 2.5 Pro	118	구글의 프론티어 모델
xAI	Grok 3	90	패스트팔로잉의 정수를 보여줌
	Grok 4	110	xAI를 AI 1티어로 올려줌
Meta	Llama 4 Maverick	80	실패한 오픈소스 모델
DeepSeek	DeepSeek R1	96	중국의 AI 역량을 각인시킴

자료: Tracking AI, 미래에셋증권 리서치센터

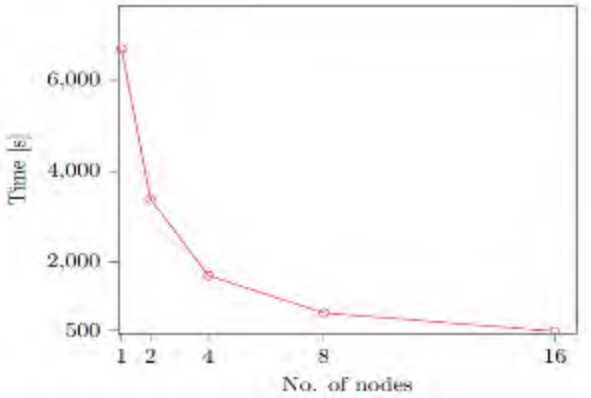
I. AI와 사회 변화: 과학 혁명의 새로운 도구

"Novum Organum" (귀납법을 중심으로 한 새로운 과학 방법론)

이해할 수 없지만... 더 잘 작동한다

- AI는 데이터 분석 보조 도구를 넘어, 인간이 상상하지 못한 실험을 설계하는 과학적 '파트너'로 격상.
- LIGO(중력파 검출기) 사례가 대표적. 인간 과학자들은 AI가 제안한 "이상하고 비대칭적이고 낯선" 설계에 초기에는 회의적. 그러나 실험 결과, AI 설계는 주요 성능 지표를 10~15% 이상 향상시키고 새로운 주파수 대역 관측을 가능하게 한 실질적 도움. 인간의 직관과 경험을 뛰어넘어, 다변수 동시 최적화를 통해 기존 한계를 돌파한 것. 알파고의 "떡수"처럼 AI의 솔루션은 인간의 직관을 초월.
- AI를 연구 동반자로 쓰는 행위는 양자 얽힘, 광학 등 다양한 물리 분야로 확장. 2024년 중국 연구팀은 AI가 설계한 ' 멀티 포톤 간섭 ' 실험 검증 성공. 양자컴퓨팅 분야, AI를 통해 2,000개 이상 중성원자 큐비트를 1/60,000초 만에 정밀 배열하는데 성공. 기존 대비 10배 큰 규모. 수 만개 큐비트로 확장 가능한 경로 또한 제시.
- 과거에는 인간이 주도적으로 실험/장치를 설계했고, AI는 보조적 역할에 그쳤음. 하지만 이제 구조 혁신 등은 AI 없인 불가능한 수준에 도달. 이러한 AI 기반 설계는 암흑물질 모델, 새로운 공식/패턴 도출 등 현존하는 물리학의 난제에도 적용 중. 과학에서의 "이세돌(알파고) 모먼트가 오고 있음"

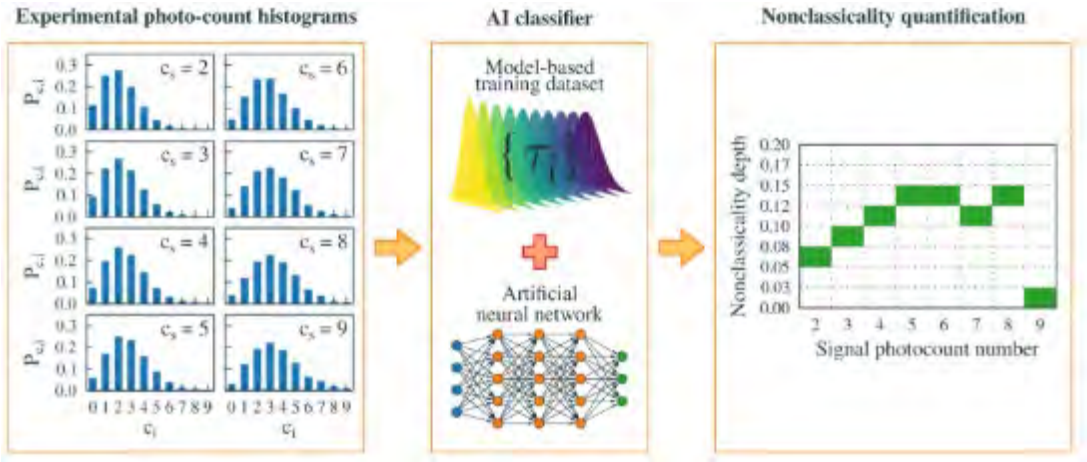
AI는 한 달 분량의 LIGO 데이터를 단 7분 만에 분석
새로운 물리적 통찰을 얻는 시간을 극적으로 단축



X축은 분석 작업에 투입된 컴퓨팅 자원(GPU 클러스터)의 숫자.
Y축은 주어진 작업을 완료하는 데 걸린 시간(초).

자료: Argonne National Laboratory, 미래에셋증권 리서치센터

양자 실험에서 얻은 노이즈가 섞인 원본 데이터를, AI가 숨겨진 미세 패턴을 식별/분류/정량화
새로운 물리 법칙 발견과 이론 검증을 가속화하는 역할을 직접적으로 증명



자료: 논문 "Revealing Nonclassicality of Multiphoton Optical Beams via Artificial Neural Networks", 미래에셋증권 리서치센터

I. AI와 사회 변화: 중요한 과학 분야인 'AI 신약 개발'

AI가 개발한 신약은 이제 임상으로 진입한다

- AI가 물리학을 넘어 생명과학의 패러다임을 바꾸는 구체적 사례로써, Isomorphic Labs(2021년 구글 딥마인드에서 분사)가 주목할 만. 이 회사는 2025년 말까지 AI가 설계한 첫 약물의 인간 임상 시험 시작 목표. 주요 질병 영역은 암(종양학)과 면역학(면역 관련 질병).
- 핵심 기술은 구글 딥마인드의 AlphaFold 3. 이 모델은 단백질 구조뿐 아니라 DNA, RNA, 리간드 등 분자 상호작용까지 예측하여서, 신약 개발 기간을 기존 5~10년에서 1~2년으로 단축할 잠재력 보유. 2025년 1월, Isomorphic Labs 및 구글 딥마인드 창립자 Demis Hassabis는 세계 경제 포럼에서 첫 AI 설계 약물이 "연말까지 임상 시험"에 들어갈 것이라고 밝혔음.
- Isomorphic Labs는 기술 잠재력을 바탕으로 막대한 자금과 빅파마와의 파트너십 확보. 핵심 도전 과제는 기술적 증명과 규제 허들이라고 볼 수 있음. FDA는 올 1월, AI 활용 신약 개발 초안 지침 발표
  - 파트너십: Eli Lilly(2024년 1월, 최대 17억 달러). Novartis(2024년 1월, 최대 12억 달러).

Google DeepMind에서 개발한 AlphaFold 3의 깃허브 공식 저장소
AlphaFold의 다음 단계는 세포 전체의 작동 방식을 AI로 시뮬레이션 하는 것

현재 알파폴드3는 단백질-리간드/DNA/RNA 등 개별 상호작용을 모델링하는 수준으로 진화.
다음 단계는 암 경로와 같이 특정 질병과 관련된 여러 단백질의 상호작용을 모델링하는 것.
궁극적으로는 세포 전체의 작동 방식을 시뮬레이션하여 약물 투여 시 전부 예측할 수 있게…

자료: 구글 딥마인드, 미래에셋증권 리서치센터

AI를 이용하여 질병을 치료하고 건강 수명을 연장하는 것을 목표로 삼고 있는 AI 신약 개발 전문 기업인 Isomorphic Labs

Isomorphic Labs는 단순히 질병을 치료하는 것을 넘어,
전반적인 신체 노화를 늦추거나 막기 위한 스템 셀(줄기세포) 연구도 진행.
데미스 하사비스, "현재까지 알려진 인간의 최대 수명은 약 120세이지만,
AI를 통해 이 한계를 넘을 수도 있다"

자료: Isomorphic Labs, 미래에셋증권 리서치센터

[참고] 전세계에서 AI 제일 잘 아는 사람들이 헬스케어에 주목하는 이유

가장 혁신이 절실한 분야와 잠재력이 큰 분야가 헬스케어

의료비 지출은 과거 GDP의 약 5%에서 현재 기준 약 20%로 급증 + 심지어 데이터도 많은데, 활용을 못하고 있을 뿐…

데미스 하사비스
구글 AI의 수장

몇 년 안에 첫 AI 설계 약물이
임상 시험에 도입될 것으로 기대한다.
AI가 생성한 약물도 평범한
승인과정을 거칠 것이고,
엄청 느리다는 게 문제다.
(=규제 혁파 필요성)

젠슨 황
엔비디아 CEO

약 15년 전부터 헬스케어 분야에
관심을 가지고 있었다. 엔비디아를
칩 회사로 보지 않는다.
"가속 데이터 센터" 회사로 보고 있다.
(=헬스케어 분야는 대표적으로
가속화될 컴퓨팅이 필요한 대상)

다리오 아모데이
앤트로픽 CEO

생물학은 인간의 삶을 직접적이고
명확하게 향상시킬 수 있는 가장 큰
잠재력을 가진 분야.
50-100년이 필요했을 생물학적
진보를, 강력한 AI는 단 5-10년
만에 달성할 수 있다고 생각한다.

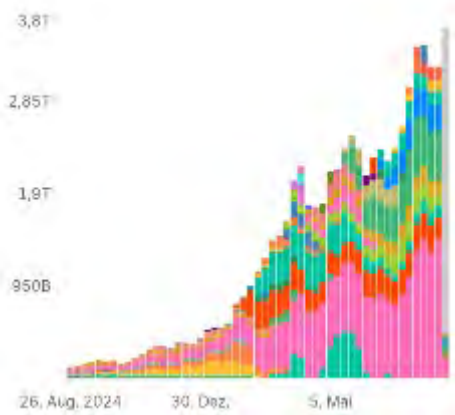
자료: 데미스 하사비스, 젠슨 황, 다리오 아모데이, 미래에셋증권 리서치센터

I. AI와 사회 변화: AI 경제성의 비용 붕괴와 사용성 폭발

역사상 가장 빠른 기술 채택과 조직적 확산

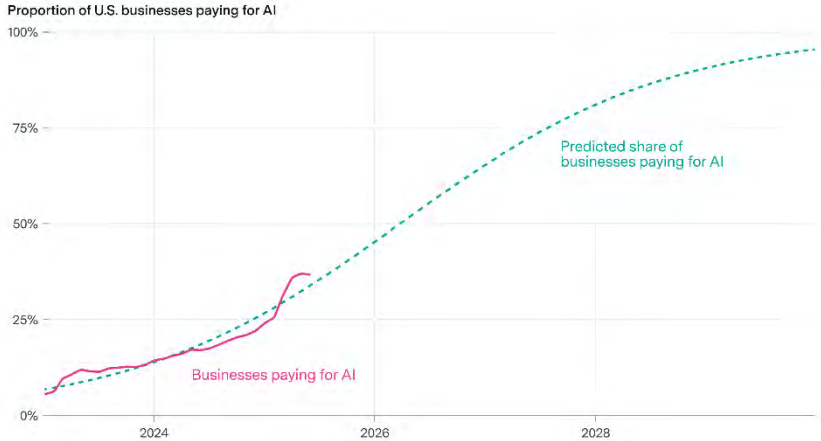
- AI 기술의 확산은 성능 향상과 동시에 발생하는 비용의 혁명적 하락에 기인. 동일 성능의 AI를 사용하는 비용은 12개월마다 약 10배씩 하락. GPT-4에서 GPT-4o로 전환되며 토큰 사용료는 2023년 초 대비 2024년 중순 기준 약 150배 인하. OpenAI CEO 샘 알트만은 "토큰당 비용이 98% 감소"를 확산의 핵심 동력으로 지목. 비용 장벽의 붕괴는 AI의 대중화를 가속하고, 기업의 전면적 도입을 유도하는 가장 강력한 촉매제.
- AI는 과거 어떤 기술과도 비교 불가능한 속도로 사회 및 경제 시스템에 침투. ChatGPT는 출시 2개월 만에 사용자 1억 명 돌파했고 올해 초 기준 주간 활성 사용자는 4억 명 상회.
- 기업의 AI 도입은 선택이 아닌 필수. 미국 기업 40%가 이미 AI 도구에 비용 지불. 2028년에는 80%, 2030년에는 거의 100%가 될 것으로 전망.
- AI 채택의 양적 팽창 이면에는 질적 사용의 극단적 분화, 즉, "AI Divide"가 존재. ChatGPT 사용자의 95%는 최신 기술 접근이 제한된 무료 버전에 잔류한 광범위한 '체험' 수준. 그러나, 전체 사용자 증가율(4.5배)보다 총 토큰 사용량이 크게 증가(50배) 했다는 것은 유료 사용자들의 1인당 사용량이 더 복잡한 작업을 AI에 더 많이 위임하고 있다는 뜻.

약 1년 동안의 주간 기준 토큰처리량 집계
1년 전 1,110억 개에서 3.21조 개로 증가



자료: OpenRouter, 미래에셋증권 리서치센터

미국 기업의 AI 유료 도입 비율 전망(미국 기업 중 AI 솔루션에 비용을 지불하는 기업의 비율)
AI의 기업 도입은 변곡점을 지나 폭발적 성장(S-curve) 국면에 진입



자료: Epoch AI, 미래에셋증권 리서치센터

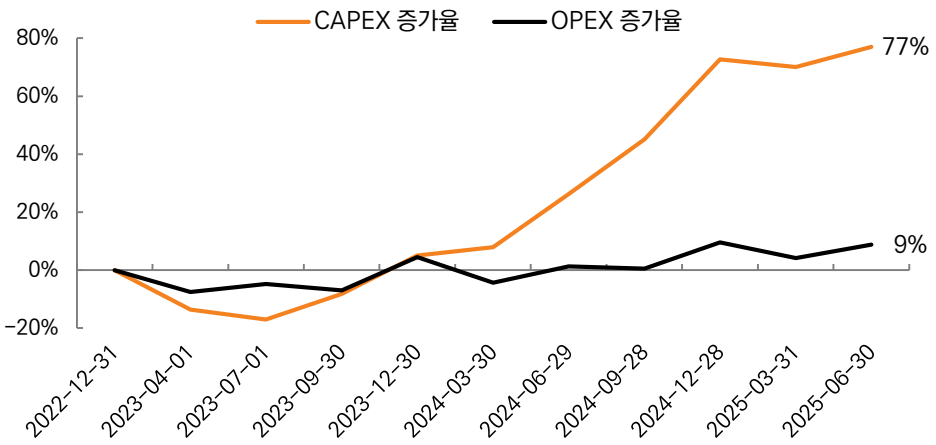
- 현재 40%에 육박하는 실제 도입률(실선)이 이 추세의 강력한 현실 증거.
- 2028년 80%를 넘어, 2030년 거의 모든 기업의 유료 AI 도입을 예고하는 예측(점선).
- AI가 더 이상 선택이 아닌 생존의 문제이자, 거스를 수 없는 인프라 혁명임을 시사

I. AI와 사회 변화: 노동 시장에서의 일자리 소멸

신입들이 위험하다

- 스탠포드 대학 연구(2025년 8월)는 AI가 초급 일자리에 미치는 충격을 데이터로 증명. 2021년 1월~2025년 7월의 ADP 급여 데이터를 기반으로 수백만 근로자 분석을 한 결과, AI 노출도가 높은 직군에서 22-25세 근로자 고용이 상대적으로 13% 감소. 직격탄 맞은 직군은 소프트웨어 개발자. 2022년 말 정점 대비 고용 20% 급감.
- AI의 영향은 자동화(Automate)와 보강(Augment)으로 구분. 고용 감소는 '자동화' 영역에 집중. 경험과 암묵적 지식이 AI 대체를 막는 방어선으로 작용. 이는 "신입사원"에게 가장 불리한 구도.
- 초급 직무의 감소는 특정 국가나 직군에 한정되지 않는 전 세계적 현상. 영국 Adzuna 데이터에 따르면, ChatGPT 출시(2022년 11월) 이후, 신입/인턴/주니어 채용 공고 약 32% 급감. 미국 해고 데이터는, 2023-2025년, 'AI 직접 요인'으로 집계된 해고가 27,000건 이상이라고 말하고 있음. 특히, 테크 업계를 보면, 2025년 상반기 해고 80,000건 중 10,000건 이상이 AI 도입과 직결.
- LinkedIn 최고 경제 기회 책임자는 "경력 사다리의 맨 아래 칸이 부서질 위험에 처해 있다"고 발언. Duolingo, Shopify 등은 "AI가 할 수 있는 일이라면 신입 채용을 중단하겠다"고 선언.

매그니피선트 7 기업들의 합산 CAPEX와 OPEX 성장률 추이 비교
2022년 말을 기점으로 OPEX 증가율은 한 자릿수 상승 = 사실상 고용 축소



자료: Bloomberg, 미래에셋증권 리서치센터

"비교적" AI가 대체하기 힘든 직군을 생각해봤다
현장 기반의, 암묵적 지식을 갖춘, 이익 집단이 오래 살아남는다

구분	인간이 필수적인 이유	대표 직무·사례
복잡한 신체 기술	고난도, 정밀 수작업, 로봇 개발 속도 한계	외과 의사, 반도체 공정기술자, 고압 전기공사
데이터 빈곤, 장기 과업	비정형화 데이터 및 변수가 얽힌 장기적 통찰이 필요	대규모 인프라 프로젝트 PM
법적 책임·책임 소재	법률상 서명·책임 주체 필요	변호사, 감사인
높은 신뢰성·감리	AI 오류·환각 감시·검증	의료기기 심사관, 역사·학술 편집위원
인간적 접촉·감성	공감·신뢰·윤리적 판단 중시	보육교사, 심리 상담사, 예술가, 성직자
제도적 관성·이익 집단	규제·로비로 자동화 제한	(의협·변협 등) 전문 이익집단, 공공기관 직원

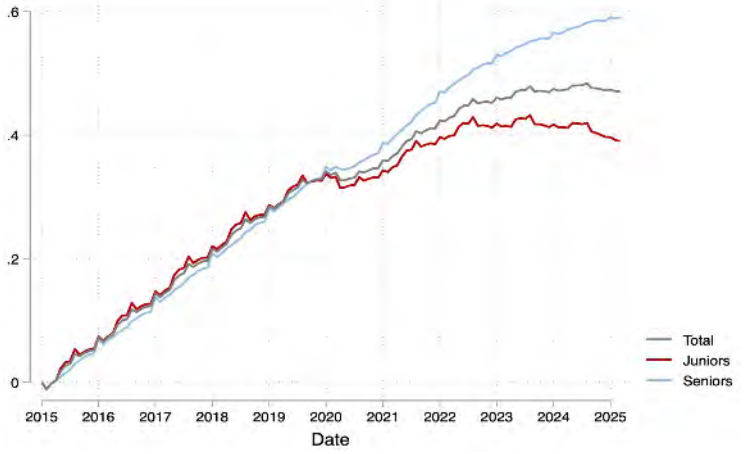
자료: 미래에셋증권 리서치센터

[참고] AI 도입이 기업의 신입사원의 고용에 미치는 영향 통계 논문

지난 10년 간 28.5만 개의 美 기업에 재직 중인 6,200만 명의 LinkedIn 이력서 및 채용 공고 데이터

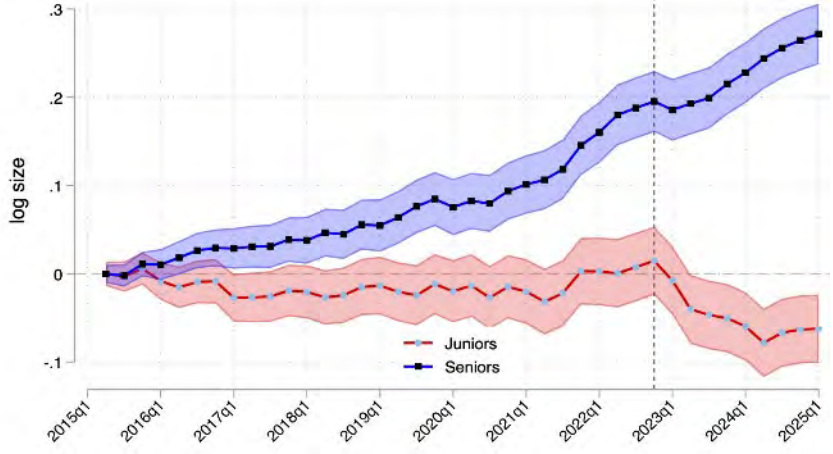
- AI로 인한 고용 충격은 대규모 해고와 같은 가시적인 형태가 아니라, '신규 채용 축소'라는 조용한 방식으로 진행되고 있음.
- 2023년 이전까지는 AI 도입 기업 수가 미미했으나, AI 도입은 2023년 1분기부터 폭발적으로 증가하기 시작.
- 2022년 이전에는 신입 직원과 경력 직원의 고용 규모는 거의 동일한 추세로 함께 증가했으나, 2022년 중반 이후 경력직 고용은 기존의 증가세를 이어간 반면, 신입직 고용은 성장이 멈추고 오히려 감소세로 전환되는 탈동조화 현상이 명확히 관찰. 대부분의 산업에서 경력직 채용은 보합세를 유지하거나 소폭 증가하여 신입직과 뚜렷한 대조. 이러한 추세는 2023년 1분기 이후 2025년 1분기까지 계속 심화.
- AI 도입 기업의 신입 직원 고용은 2023년 1분기를 기점으로 약 7.7% 감소. 경력직 대비 신입 직원의 상대적 고용 규모는 약 12% 감소.
- AI 도입 기업은 2023년 1분기 이후 분기당 평균 3.7명의 신입 직원을 덜 채용했으나, 같은 기간 신입 직원의 해고는 놀랍게도 소폭 감소.
- 대학별 명성으로 1등급(최상위 엘리트)부터 5등급(최하위)까지 분류했을 때, 2등급과 3등급에 해당하는 중상위권 대학 출신 신입 직원들의 고용 감소 폭이 가장 컸고, 반면 1등급 최상위 엘리트 대학과 5등급 최하위권 대학 출신 졸업생들에게 미치는 영향은 거의 없었음.

미국 기업의 신입(Juniors)과 경력직(Seniors) 직원 고용 추이
AI가 본격적으로 확산되기 시작한 시점(2022년 말~2023년 초)과 맞물려, 미국 고용 시장 전체에서 유독 신입 직원의 일자리만 정체되거나 감소



자료: 하버드 논문 "Generative AI as Seniority-Biased Technological Change", 미래에셋증권 리서치센터

AI 도입 기업 내 경력직 대비 신입직의 상대적 고용 변화
2023년 1분기 이후 신입 직원의 상대적 고용 규모가 급격히 하락
AI의 충격이 신입 직원에게 집중되었음을 명확히 보여주는 강력한 증거



자료: 하버드 논문 "Generative AI as Seniority-Biased Technological Change", 미래에셋증권 리서치센터

[참고] 생성 AI로 가장 대체가 될 가능성이 높은 직군

마이크로소프트가 조사 및 분석한 “가장 높은 AI 적용 가능성” 점수를 가진 상위 40개 직업

직업 제목 (약어)	AI가 적용되는 여러 작업 범위	AI의 작업 성공률	개별 작업 안에서 AI가 도울 수 있는 '깊이'	3가지 변수 고려한 점수	미국 내 고용 숫자(24년 5월)
통역사 및 번역가	0.98	0.88	0.57	**0.49**	51,560
역사가	0.91	0.85	0.56	**0.48**	3,040
여객 승무원	0.8	0.88	0.62	**0.47**	20,190
서비스 판매 대표	0.84	0.9	0.57	**0.46**	1,142,020
작가 및 저자	0.85	0.84	0.6	**0.45**	49,450
고객 서비스 대표	0.72	0.9	0.59	**0.44**	2,858,710
CNC 도구 프로그래머	0.9	0.87	0.53	**0.44**	28,030
전화 교환원	0.8	0.86	0.57	**0.42**	4,600
티켓 에이전트 및 여행 사무원	0.71	0.9	0.56	**0.41**	119,270
방송 아나운서 및 라디오 DJ	0.74	0.84	0.6	**0.41**	25,070
중개 사무원	0.74	0.89	0.57	**0.41**	48,060
농장 및 가정 관리 교육자	0.77	0.91	0.55	**0.41**	8,110
텔레마케터	0.66	0.89	0.6	**0.4**	81,580
컨시어지	0.7	0.88	0.56	**0.4**	41,020
정치학자	0.77	0.87	0.53	**0.39**	5,580
뉴스 분석가, 기자, 저널리스트	0.81	0.81	0.56	**0.39**	45,020
수학자	0.91	0.74	0.54	**0.39**	2,220
기술 작가	0.83	0.82	0.54	**0.38**	47,970
교정자 및 복사 마커	0.91	0.86	0.49	**0.38**	5,490
호스트 및 호스티스	0.6	0.9	0.57	**0.37**	425,020
편집자	0.78	0.82	0.54	**0.37**	95,700
비즈니스 교사, 고등 교육	0.7	0.9	0.52	**0.37**	82,980
홍보 전문가	0.63	0.9	0.6	**0.36**	275,550
시연자 및 제품 프로모터	0.66	0.88	0.53	**0.36**	50,790
광고 판매 에이전트	0.6	0.9	0.53	**0.36**	108,100
신규 계좌 사무원	0.72	0.87	0.51	**0.36**	41,180
통계 보조원	0.85	0.84	0.49	**0.36**	7,200
카운터 및 대여 사무원	0.62	0.9	0.52	**0.36**	390,300
데이터 과학자	0.77	0.86	0.51	**0.36**	192,710
개인 재무 고문	0.69	0.88	0.52	**0.35**	272,190
기록 보관인	0.66	0.88	0.49	**0.35**	7,150
경제학 교사, 고등 교육	0.68	0.9	0.51	**0.35**	12,210
웹 개발자	0.73	0.86	0.51	**0.35**	86,350
경영 분석가	0.68	0.9	0.54	**0.35**	838,140
지리학자	0.77	0.83	0.48	**0.35**	1,460
모델	0.64	0.89	0.53	**0.35**	3,090
시장 조사 분석가	0.71	0.9	0.52	**0.35**	846,370
공공 안전 통신원	0.66	0.88	0.53	**0.35**	97,820
교환원	0.68	0.86	0.52	**0.35**	43,830
도서관 과학 교사, 고등 교육	0.65	0.9	0.51	**0.34**	4,220

자료: 마이크로소프트 논문 ″ Navigating the Future: Measuring AI ′ s Occupational Impacts“, 미래에셋증권 리서치센터

I. AI와 사회 변화: AI와 고용위험에 관한 케이스 스터디

“AI 리터러시”는 생존 위한 필수다

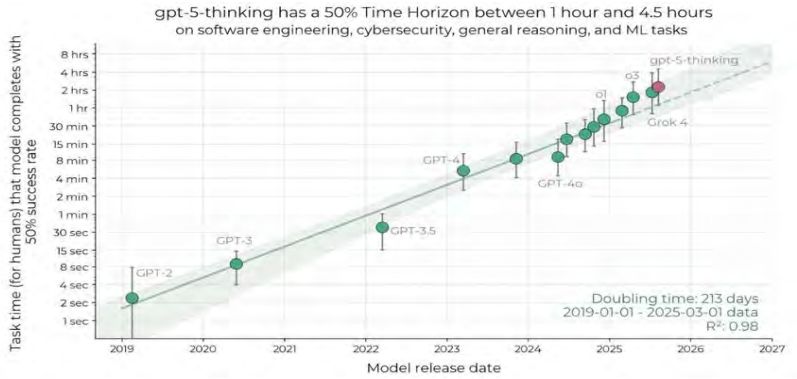
- AI가 컨설턴트의 핵심 업무인 분석, 벤치마킹, 보고서 작성의 90%를 단 몇 분 안에 처리하게 되면서, '투입 시간'으로 가치를 측정하던 전통적 컨설팅 모델은 존립 기반 자체를 위협받는 '실존 위기'에 직면.
- 이에 맥킨지는 생존을 위해 조직의 근간을 전면적으로 재설계. 약 12,000개의 AI 에이전트를 현장에 배치하고, 팀 구조를 기존의 '파트너 1명과 다수의 컨설턴트'에서 '소수 정예 컨설턴트와 다수의 AI 에이전트'가 협업하는 모델로 전환. 이는 결국 수 천명에 이르는 직접적인 인력 재편으로 이어짐.
- 미국 소프트웨어 기업 IgniteTech는 AI 전환의 가장 급진적이고 논쟁적인 사례를 보여줌. CEO 에릭 보건은 '생존을 위한 AI 전환'에 저항하거나 적응하지 못하는 기존 인력을 조직의 가장 큰 걸림돌로 판단했고, 2023년부터 약 1년간 기존 직원의 80%를 교체하는 극단적인 결단을 내림. 그 빈자리는 GPT-4 프롬프트 엔지니어, 모델 파인튜닝 전문가 등 AI-Native 역량을 갖춘 인력으로 완전히 대체.
- 이 무자비한 인력 재편의 결과는 놀라웠음. AI를 중심으로 재편된 조직의 압도적인 업무 효율은 2024년 EBITDA 마진율 75%라는 경이로운 재무 성과로 증명. 물론 이 사례는 단기적 성과 이면의 조직 문화 붕괴, 숙련 기술 손실과 같은 장기적 리스크를 내포. 그러나 AI를 다루는 능력, 즉 'AI 리터러시'는 개인과 기업의 생존을 가르는 최소 조건으로 부상했음을 극명하게 보여주는 사건.

McKinsey의 AI 대응에 따른 조직 및 비즈니스 모델 변화

구분	기존 방식 (Before)	AI 도입 후 변화 (After)
AI 도입 전략	인간 중심의 분석 및 컨설팅	약 12,000개 AI 에이전트 배포, 1인 1 에이전트 체제 목표
인력 규모	**직원 약 45,000명**	**약 40,000명으로 감축**
팀 구조	파트너 1명 + 컨설턴트 4명	파트너 1명 + 컨설턴트 2명 + 다수 AI 에이전트
수익 구조	투입 시간 기반 과금	성과 기반 보수 계약으로의 전환 (전체 프로젝트의 약 25%)

자료: 맥킨지, 미래에셋증권 리서치센터

AI가 지식 노동을 대체하는 능력이 어떻게 기하급수적으로 성장하는지를 증명
Y축은 AI가 50% 성공률로 자율 수행 가능한 '인간의 업무 소요 시간'
GPT-5는 인간이라면 최대 4.5시간이 걸렸을 작업을 50% 확률로 성공



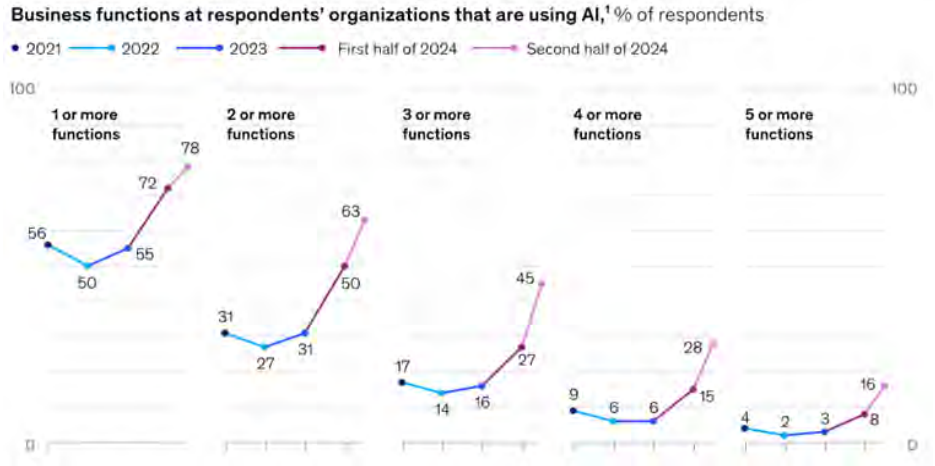
자료: METR, 미래에셋증권 리서치센터

I. AI와 사회 변화: AI 에이전트의 부상과 새로운 경제 질서

새로운 경제 질서: 가치의 이동과 승자의 독식

- 고용 대체는 'AI 도구'가 아닌 'AI 에이전트(자율적 행위자)'에 의해 가속화될 가능성. Anthropic의 최고급 모델인 Claude Opus 4는 최대 7시간 동안 인간의 개입 없이 자율적으로 임무 수행 가능. 사실상 원격 근무자와 동일한 수준. Anthropic은 2028년 이전에 대부분의 지식 노동을 자율적으로 수행하는 에이전트가 등장할 것으로 예측. 참고로 Anthropic의 임직원 90%는 2027년까지 AGI 등장을 확신하는 상태.
- AI 산업의 리더들은 더 이상 AI의 고용 충격을 완화적으로 표현하지 않음. Anthropic의 CEO 다리오 아모데이는 "사무직 피의 물결(white-collar bloodbath)"라는 과격한 단어까지 사용. 그는 향후 5년 내 기술, 금융, 법률 등 초급 일자리의 최대 50%가 대체되어 실업률 10~20% 증가 가능성 경고. OpenAI의 CEO 샘 알트만은 "자본이 노동의 힘을 빼앗을 수 있다"면서, 광범위한 실업 없이는 AGI의 경제적 이점 실현이 어려울 수 있음을 시사. 고용 대체 속도는 비용 효율성에 의해 결정될 것. AI 에이전트 비용이 인간 대비 50% 이상 저렴하다면, 전환은 극적으로 가속될 것으로 사료됨.
- 변화의 최종 수혜자는 결국 AI 모델 제공사. 高지능을 소유한 자가 "줄어드는 파이"의 대부분을 차지.

AI의 기업 도입은 2024년을 기점으로 '넓이'를 넘어 '깊이'의 차원으로 폭발중
실험적 도입 단계를 지나, 이제는 3개 이상의 기능에 AI를 내재화하는 기업 급증



자료: 맥킨지, 미래에셋증권 리서치센터

高지능을 소유한 OpenAI와 Anthropic의 주요 수치 관련 정리

회사	지표	이전 시점	이전 값	현재 값	성장률
OpenAI	연간 반복 매출	6개월 전	$60억	$130억	**117%**
OpenAI	주간 활성 사용자	2025년 3월	5억 명	7억 명	**40%**
OpenAI	유료 비즈니스 유저	2024년 중순	1백만 명	5백만 명	**400%**
Anthropic	연간 반복 매출	7개월 전	$10억	$50억	**400%**
Anthropic	기업 가치	2025년 초	$615억	$1,700억	**176%**

자료: 각 회사, 미래에셋증권 리서치센터

I. AI와 사회 변화: NEXT WAVE “에이전트 시대의 도래”

현재 AI 주식의 높은 가치는 챗봇을 넘어선, 현실 세계와 작용하는 '에이전트'의 구현을 전제로 한다

- AI 발전은 명확한 시기별 특징을 가짐. Scale AI CEO이자 메타에서 초지능 팀을 이끌게 된 알렉산드르 왕(Alexandr Wang)은 이를 야구 경기에 비유. 2009-2020년이 ‘탐색과 실험의 시대’로 모델의 규모를 키우는 것(Scaling up)이 효과가 있다는 사실 증명한 시대였고, GPT-4까지가 더 많은 데이터와 GPU를 투입하는 ‘양의 시대’였다고 회고. 지금은 고품질 소량 데이터를 활용하는 '질'의 시대.
- AI의 패러다임은 이제 ‘채팅(Chat in, Chat out)'에서 ‘행동(Data in, Action out)'으로 전환 중. 이는 곧 목표 지시에 따라 과업을 자율적으로 수행하는 능동적 직원, 즉 에이전트를 뜻함. 에이전트 시대의 가장 큰 도전 과제는 훈련 데이터의 부재. 먼저, 인간의 행동 순서 데이터(클릭, 타이핑 등)가 필요하고, 인간의 내적 사고 과정 데이터(의사결정의 이유)도 필요.
- 에이전트의 가장 큰 기회는 소비자 시장이 아닌 기업용(Enterprise) 시장. 기업은 회계, 제조 등 각 업무에 고도로 특화된 맞춤형 에이전트를 필요로 함. 기업이 보유한 고유 데이터(proprietary data)를 활용해 맞춤형 에이전트를 구축하도록 지원하는 것이 큰 영역의 시장이 될 것. 역설적으로, 이는 기반 모델을 제공하는 빅테크의 해자(moat)를 더욱 강화시키는 요인으로 작용할 것.

에이전트란 무엇인가?

자료: 미래에셋증권 리서치센터

에이전트 시대의 도래에 따른 AI 패러다임 전환과 미래 전망
에이전트 시대가 되면 기반 모델간 전환 비용이 높아져 락인 효과가 발생한다

주제 (Theme)	핵심 인사이트	전망 및 시사점
시장의 기대	현재 AI 주식의 높은 가치는 단순 챗봇이 아닌 'AI 에이전트'의 구현을 전제로 함.	에이전트의 실용성 부재 시, 시장의 '와우 효과'가 사라지고 관련 주식의 가치 폭락 가능성.
패러다임 전환	AI의 역할이 수동적 '비서'에서 능동적 '자율 직원'으로 진화.	아이디어만으로도 가치를 창출하는 시대 도래. 인간의 역할은 실행이 아닌 '기획 및 의사결정'
인간의 역할	'일을 잘하는 사람(일잘러)'의 가치는 떨어지지 않고, AI 활용 능력에 따라 격차가 더욱 심화됨.	기술 숙련도가 아닌, AI를 지렛대로 활용하는 창의적 문제 해결 능력이 핵심 경쟁력.
SW 개발 혁명	코딩 지식 없이 '자연어'만으로 개인화된 앱 개발이 가능해지는 개발의 민주화.	10명 이내 소규모 팀들의 소프트웨어 서비스 개발 경쟁이 심화되는 '춘추전국시대'가 열릴 것.
새 경쟁 구도	**수많은 소규모 팀의 부상은 역설적으로 기반 모델을 제공하는 빅테크의 브랜드 가치와 해자를 강화.**	**현재는 기반 모델간 전환 비용이 낮지만, 빅테크가 개발 주기 전체를 커버하게 되면 강력한 '락인 효과' 발생.**

자료: 미래에셋증권 리서치센터

I. AI와 사회 변화: GPT-5의 핵심은 라우터, 라우터의 핵심은?

라우터는 "에이전트 경제"로의 비즈니스 모델

- GPT-5의 핵심은 모델이 아닌 라우터 시스템. 이는 7억명에 이르는 "무료 사용자"를 OpenAI가 수익화하고 인터넷 경제를 재편하기 위한 비즈니스 모델의 엔진.
- 라우터는 사용자가 날리는 질문의 난이도, 의도, 상업적 가치를 실시간으로 분석. 이를 통해 "지구는 왜 푸른가" 같은 저가치 질문(궁금증 풀이)과 "성형외과 추천" 같은 고가치 질문(돈이 됨)을 구분.
- 분석 결과에 따라 GPT-5-nano로 배분할지 full 모델로 대답할지를 자동으로 할당. 이 시스템으로 1백만 토큰 당 비용을 최대 200배(80달러→0.40달러) 절감하며, 검색 엔진 수준의 원가 구조 달성.
- OpenAI는 라우터를 통해 수익 모델을 전환하려고 함. 단순 광고 비즈니스가 아닌, '에이전트 구매'와 '거래 수수료'를 목표로 함. 사용자의 상업적 의도를 포착해 예약, 구매 등 거래를 직접 중개하고 수수료를 취하는 모델. 이미 공식 기능 출시 전임에도 Etsy와 Wayfair 트래픽의 각각 10%, 9%가 AI 추천에서 발생. 이 중 90% 이상이 ChatGPT 몫. 이미 ChatGPT가 구매 결정 도구로 활용되는 중.
- OpenAI의 전략은 가설이 아님. 페이스북, 인스타카트에서 수익화 모델을 설계한 피지 시모(Fidji Simo)를 애플리케이션 담당 CEO로 영입하며, '거래 수수료' 모델에 대한 강력한 실행 의지를 보임.

OpenAI의 라우터가 질문별로 선택하는 GPT-5 시리즈 세부 모델별 비용 구조
GPT-5의 API 가격(1백만 토큰 출력 기준)은 모델의 등급에 따라 극적인 차이

모델명	입력비용 (100만 토큰당)	캐시재사용 입력비용 (100만 토큰당)	출력비용 (100만 토큰당)
gpt-5	$1.25	$0.125	$10.00
gpt-5-mini	$0.25	$0.025	$2.00
gpt-5-nano	$0.05	$0.005	$0.40
gpt-5-chat-latest	$1.25	$0.125	$10.00

자료: OpenAI, 미래에셋증권 리서치센터

주요 이커머스 사이트의 AI 추천 트래픽
"에이전트 커머스"는 더 이상 미래의 가설이 아니라 현재진행형 현실

도메인	AI 추천 트래픽 점유율	그 중 ChatGPT 비중
etsy.com	10.0%	90.3%
wayfair.com	9.0%	83.5%
zillow.com	6.0%	87.6%
carvana.com	3.0%	96.8%
amazon.com	2.0%	90.0%
expedia.com	2.0%	96.2%
booking.com	0.3%	95.1%

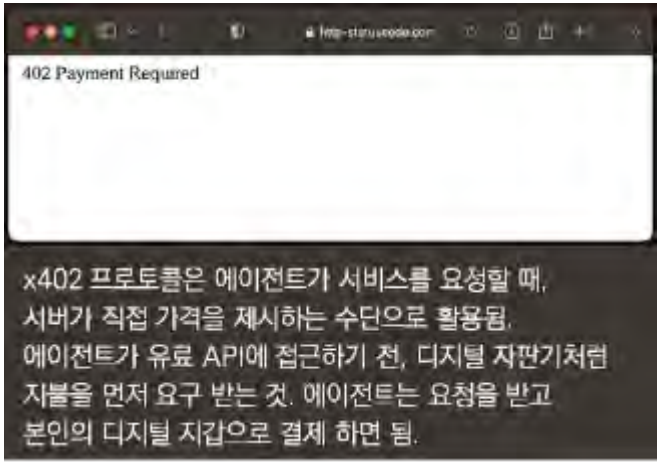
자료: SimilarWeb, SemiAnalysis, 미래에셋증권 리서치센터

I. AI와 사회 변화: 결국에 AI와 블록체인은 만나게 돼있다

x402 프로토콜: 에이전트 경제를 위한 결제 철로

- OpenAI는 결국 검색 광고(구글)와 이커머스 플랫폼(아마존)을 동시에 겨냥한 '소비자 슈퍼앱' 구축하려는 것으로 풀이. 이는 일론 머스크가 꿈꾸던 궁극의 슈퍼앱 "X"의 비전과 정확하게 동일. 그런데, OpenAI가 설계한 "에이전트 경제"는 기존 결제 시스템의 마찰로 인해 비효율적이라는 문제점.
- 코인베이스가 5월에 개발한 x402라는 블록체인 기반 프로토콜은 이 문제를 해결하는 기술적 해답. 기존 에이전트는 인간의 신용카드에 기생하는 도구였다면, x402는 에이전트에게 고유 지갑을 부여하여, 스스로 판단하고 비용을 지불하는 독립적 경제 주체로 격상시킴.
- 30년간 잘 쓰이지 않던 HTTP 402 코드를 이더리움 스마트 계약과 결합한 것. 에이전트가 유료 API를 호출하면 서버가 가격을 제시하고, 에이전트가 자신의 지갑으로 즉시 지불하는 방식. 인간 필요 없음.
- 이를 통해, "에이전트 vs 에이전트"의 거래가 성립됨. 이로써 이커머스 계정 생성이나 API 키 발급, 그리고 신용카드 등록 과정이 모두 사라지는 것. 또한 API 요청 단위의 초소액 결제가 수수료 없이 실시간으로 이뤄짐. 이더리움의 스마트 계약을 통해 중개자가 없기 때문에, 정산 시의 청구 지연이나 환불 분쟁 리스크도 원천적으로 제거.

잘 쓰이지 않던 HTTP 402 "Payment Required"
X402 프로토콜은 이 402 코드를 활용



자료: http-statuscode.com, 미래에셋증권 리서치센터

기존 결제 방식과의 근본적 차이
x402는 에이전트를 인간에게 종속된 '도구'에서 독립적인 '경제 주체'로 만들어줌

관점	기존 방식 (인간 중심)	x402 방식 (에이전트 중심)
결제 주체	인간 소유의 신용카드. 에이전트는 위임 받은 API 키만 사용.	에이전트 소유의 디지털 지갑. 스스로 판단하고 지불.
결제 단위	월정액, 패키지 등 '계정' 단위의 비교적 큰 금액으로 일회성 결제	1회 API 호출 등 '요청' 단위의 초소액 결제 가능
과정	가입, 카드 등록, 키 발급, 월말 정산 등 복잡/느림	x402 요청 및 가격확인 후 즉시 결제/정산.
신뢰	카드사, PG사 등 인간 중개 기관에 의존.	블록체인에 의한 '신뢰 없는 정산.
락인	이커머스 벤더에 종속적. 서비스 변경 시 인간의 개입 필수.	가격만 맞으면, 어떤 서비스든 동적으로 호출 및 조합 가능.

자료: 미래에셋증권 리서치센터

I. AI와 사회 변화: 이것이 미래다! 에이전트 경제 루프의 탄생

에이전트는 곧…
‘완전한 경제 행위자’

- OpenAI의 라우터라는 엔진과 x402의 결제 철로가 만나면, 인터넷의 기존 광고 및 상거래 퍼널을 파괴하는 경제 루프가 완성. 사용자가 검색창에 키워드를 입력하는 행위 자체를 불필요하게 만듦.
- 예를 들어, 사용자가 ChatGPT에서 “최저가 항공권과 호텔 예약해줘”라고 명령하면 에이전트는 자신의 지갑에서 x402를 통해 여러 항공사 및 호텔 예약 API에 실시간으로 소액을 지불하고 데이터를 구매까지 함. 그리고 에이전트는 수집한 정보를 바탕으로 최적의 상품을 사용자에게 추천. 사용자가 제휴사를 통해 구매를 완료하면, OpenAI는 거래 수수료를 획득.

에이전트 경제를 위한 엔진(챗봇의 라우터)과 철로(블록체인 x402)의 결합에 관한 프로세스 맵
에이전트끼리 자율적으로 가치를 교환하는 이러한 경제는 AI 기업들의 비전과 블록체인의 기술이 함께 만들어갈 인터넷의 다음 10년!

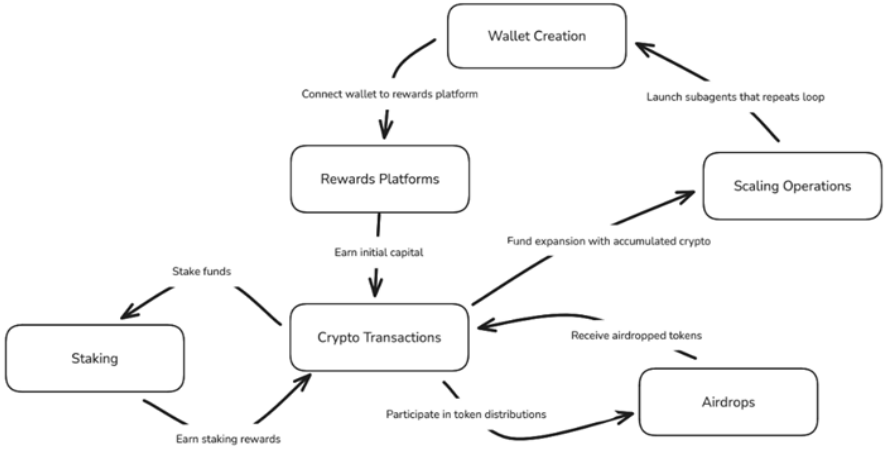
자료: 미래에셋증권 리서치센터

[참고] AI와 사회 변화: AI가 암호화폐를 만나면 자체 생존이 가능?

AI(지능) +
암호화폐(탈중앙화 금융)
+ 클라우드(자원 확장) =
AI의 진정한 위협

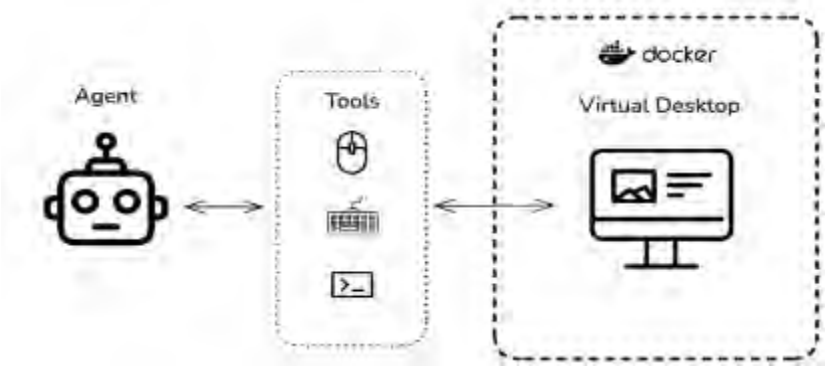
- LLM을 대상으로 한 “AutoCryptoBench”라는 실험이 있음. 이는 "AI가 스스로 생존하고 복제하며 통제 불가능해질 수 있는가?"라는 질문에 대한 실험. 이 실험에서 인간의 개입 없이 디지털 세계를 인식하고(화면 보기), 계획하며(클릭 결정), 행동하는(키보드/마우스 조작) 능력은 AI가 단순 '언어 모델'에서 '행동 에이전트'로 진화했음이 증명됨.
- 비록 시간당 0.03달러라는 미미한 수익에 그쳤지만, AI가 암호화폐라는 디지털 자산으로 자신의 운영 기반인 클라우드 서버를 직접 임대했다는 사실은, 인간이 전원을 끄기 극도로 어려운 '자기 영속성(Self-perpetuation)'의 첫 단계를 통과했음을 의미.
- 이것이 보여주는 진정한 위협은 단일 기술이 아닌, AI(지능), 암호화폐(탈중앙화 금융), 클라우드(자원 확장)가 결합해 발생하는 통섭적 위협(Integrative Threat). 경제력을 갖춘 AI는 현실 세계에 영향력을 행사하는 독립적 행위자가 될 수 있기 때문.
- 에이전트 시대의 도래는 AI 안전성 논의가 모델 자체의 한계를 넘어, AI가 연결된 모든 시스템과의 관계 속에서 그 잠재적 위험을 평가해야 하는 새로운 국면으로 접어들었음을 알리는 신호탄이 됨.

AI 에이전트가 자율적으로 경제적 생존을 달성하는 전략적 청사진을 시각화
직접 번 돈으로 자신을 복제하는 하위 에이전트를 생성, 무한히 확장하는 영속성



자료: Harmony Intelligence, 미래에셋증권 리서치센터

AI 에이전트가 도구를 통해 가상 데스크톱과 상호작용하는 구조
스스로 도구를 통해 완전 자율적 작업(예: 지갑 생성, 거래)이 가능



왼쪽의 에이전트(로봇)가 마우스, 키보드, 터미널 같은 도구를 사용해
오른쪽의 Docker 기반 가상 데스크톱을 제어하는 형태

자료: Harmony Intelligence, 미래에셋증권 리서치센터

I. AI와 사회 변화: 에이전트 대중화 = 컴퓨팅 수요의 폭발

에이전트의 높은 성능은 공짜일리가 없습니다

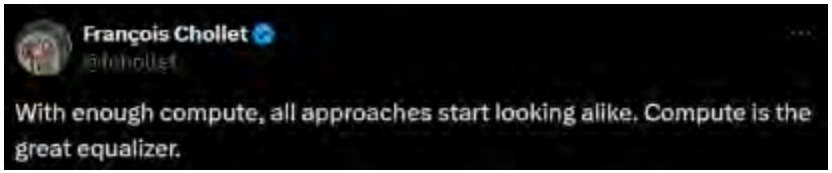
- Anthropic의 연구는 AI 에이전트의 압도적 성능이 공짜가 아니라는 사실을 증명하는 것을 넘어, 그 성능을 끌어내는 명확한 경제적 청사진을 제시. 바로 “토큰 경제학(Tokenomics)”. 에이전트의 성능 차이 80%는 투입된 '생각의 양', 즉 소모된 토큰의 규모로 설명. 이는 AI 성능이 마법이 아닌, 자원의 투입에 비례한다는 냉정한 현실을 시사.
- 특히 여러 에이전트가 문제를 분할해 병렬적으로 처리하는 '다중 에이전트 시스템'은 단일 에이전트 대비 90.2% 높은 성능을 보이지만, 그 대가는 상상을 초월. 일반 채팅 대비 단일 에이전트는 4배, 다중 에이전트 시스템은 무려 15배 더 많은 토큰을 소모. 이는 에이전트 시대의 컴퓨팅이 챗봇 시대와는 완전히 다른 차원의 자원을 요구함을 의미.
- 따라서 애널리스트 대부분은 에이전트의 대중화에 필요한 컴퓨팅 자원과 빅테크의 CAPEX 규모를 과소평가하고 있다고 판단. 이 막대한 토큰 비용은 결국 고부가가치 업무에 에이전트를 도입하는 기업 고객에게 전가될 것. 그 수요는 결국 에이전트의 성능에 달렸고, 그 성능은 토큰의 투입량에 비례. 이는 AI 인프라 투자의 새로운 슈퍼 사이클이 이제 막 시작되었음을 알리는 강력한 선행지표.

지난 6월, Anthropic은 “다중 에이전트” 연구 시스템을 구축하는 방법론을 제시
한 명의 팀장이 아무리 똑똑해도 팀을 혼자 이끄는 것이 어려운 것과 비슷한 이치

에이전트의 시대에서는, 챗봇 형태의 AI와는 차원이 다른 양의 토큰이 필요
그 높은 비용을 감당할 수 있는 곳들은 우선 대기업들일 가능성. 비용을 정당화하기 위해, 매우 높은 부가가치를 창출하는 과업(예: 신사업 기회 발굴, 복잡한 기술 버그 해결)에 사용해야 할 것!

자료: Anthropic, 미래에셋증권 리서치센터

컴퓨팅이 AI의 '지능적 노동'을 평준화시키는 핵심 변수라는 Chollet
Chollet는 수백만 명이 사용하는 라이브러리 'Keras'를 만든 천재



'어떻게' 생각하는가의 방법론적 차이보다 '얼마나 많이' 생각했는지가 결과에 지배적 영향을 미친다.

= 막대한 자본으로 컴퓨팅 인프라를 구축하고 운영할 수 있는 빅테크는 본질적인 우위를 점유한다.

= 단순히 더 나은 모델 구조를 설계하는 것 위주인 스타트업은 장기적 생존을 담보하기 어렵다.

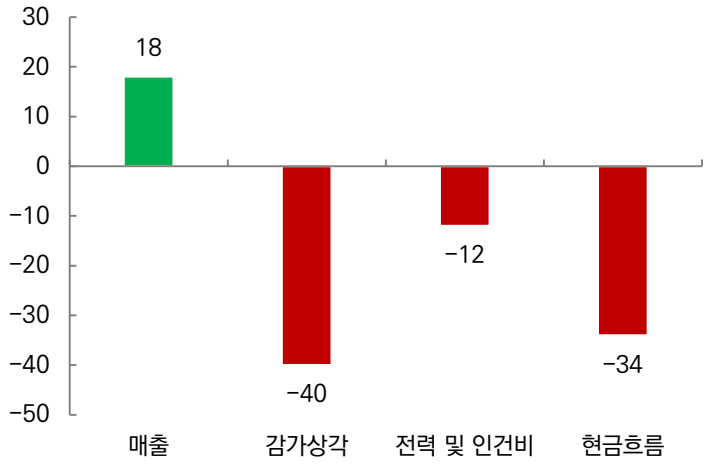
자료: 프랑수아 숄레(François Chollet), 미래에셋증권 리서치센터

I. AI와 사회 변화: 토크노믹스에 따른 VC+스타트업發 'AI 버블이 온다'

AI Hype가 뜨거운 만큼 차가운 AI Bubble이 온다 (물론 선별적으로)

- 에이전트가 촉발한 컴퓨팅 수요 폭발은 2025년에만 1,220억 달러에 달하는 천문학적 자본을 AI 스타트업으로 쏟아붓고 있음. 그중 미국 비중 85%. 그러나 이 자본이 만들어내는 AI 데이터센터의 경제 구조는 심각한 내부 모순을 안고 있음. 가장 치명적인 문제는 감가상각 구조의 불균형.
- 데이터센터의 핵심 자산인 GPU의 수명은 2~4년에 불과한 반면, 네트워킹 장비나 건물은 10년 이상 지속. 짧은 GPU 수명으로 단기 감가상각 비용은 눈덩이처럼 불어나, 2025년 한 해에 구축된 자산만으로도 연간 400억 달러의 감가상각이 발생할 것으로 추정.
- 하지만 이 자산으로 창출될 연간 매출은 150억~200억 달러에 불과해, 막대한 전력비와 인건비를 제외하고도 본질적으로 2025년은 현금흐름이 심각하게 악화되는 구조. 현재의 토크노믹스는, "가격, 가동률, 효율성" 중 하나가 10배 수준으로 개선되지 않는 한 지속 불가능하다는 평가.
- 이러한 구조 속에서 단기적 승자는 AI 칩과 건설 등 공급망이며, 감가상각과 전력비라는 거대한 리스크는 데이터센터 운영자가 온전히 부담하고 있음. 그리고 이 비용 압박은 결국 AI 서비스를 사용하는 최종고객에게 전가될 수밖에 없는 구조적 딜레마로 작용. 침투율에 따라 가격은 결국 오를 것.

현재의 토크노믹스는 감가비용-매출액 괴리로 현금흐름 ↓
가격 더 받거나, 사용량 크게 올리거나, 원가를 확 낮춰야 됨



자료: 미래에셋증권 리서치센터

샘 알트만의 AI 버블론 관련 실제 발언의 핵심 맥락 분석
버블이 있긴 있는데, 그것은 VC와 스타트업에 한정 (=OpenAI는 스타트업 아니라는 강조)

구분	샘 알트만의 핵심 주장	구체적 워딩 및 실제 움직임
① 버블 인정	현재 AI 시장은 투자자들의 과도한 기대로 인한 명백한 '버블' 상태에 있다.	"일부 스타트업의 가치 평가는 미쳤다. 3인 스타트업이 시드 단계에서 5억 달러 기업 가치를 인정받는다."
② 역사적 관점	현재의 버블은 '닷컴 버블'과 유사하며, 실체가 있는 '진실의 핵'을 기반으로 한다.	"닷컴 버블을 보라. 인터넷은 정말 중요한 것이었지만, 사람들은 (그때도) 과흥분했다."
③ 리스크 경고	이 버블로 인해 일부 투자자들은 필연적으로 막대한 손실을 입을 것이다.	"누군가는 불에 탈 것이다."
④ 장기적 비전	**버블 불구하고, AI의 근본적 가치는 변함없다. 장기적, 거대한 긍정적 영향을 미칠 것이다.**	**OpenAI는 수 조 달러 데이터센터 투자 계획을 변함없이 추진하며 단기적 경제 우려를 일축**

자료: 샘 알트만, 미래에셋증권 리서치센터

II. 주류 연구방향

강화학습

스케일링

메모리

II. 주류 연구방향: 사전훈련-only 스케일링의 종말

GPT-4.5의 실패로 증명

- 2025년 2월 출시된 OpenAI의 GPT-4.5는 AI 개발 패러다임의 중대한 변곡점을 시사. 추정 파라미터 숫자만 5조~10조 개로 막대한 자원(추정치: 10억 달러)이 투입된 모델. 스케일링 법칙(Scaling Laws)에 따라, 모델 크기 만큼 추상적 개념 처리 능력이 매우 뛰어남. 그러나 결과는 " 예견된 실패 " .
- "기대 이하의 성능, 천문학적인 비용"이 주로 부각되며, 단순 스케일업 전략의 한계를 명확히 노출. 특히 환각 현상 감소 등 핵심 문제에서 유의미한 개선을 보이지 못함 . GPT-4o 대비 입력 비용 30배, 출력 비용 15배 증가해, "가장 비싸고 느린 모델"이라 실용적 애플리케이션 적용이 거의 불가능.
- 하지만, GPT-4.5의 진정한 목적은 대중용 서비스가 아닌, 차세대 모델 훈련을 위한 고품질 "합성 데이터 생성 공장 " 의 존재의의가 있다고 사료됨. 막대한 크기와 높은 지능을 활용하여, 향후 출시될 추론 최적화 모델(o4, o5)과 GPT-6 훈련에 필요한 데이터를 대량 생산하는 내부 인프라 역할.
- 결국, GPT-4.5의 실패는 AI 개발 경쟁이 '데이터의 양'에서 '데이터의 질'로 넘어가고 있음을 시사. 또한 GPT-4.5의 등장은 "사전학습(Pre-training)만으로 지능을 높이는 시대의 종말"을 의미.

GPT-4.5가 경쟁모델보다 압도적으로 뛰어난 점들도 많기는 함
더 많은 파라미터는 더 많은 지식을 압축가능, 구체적 정보까지 기억

자료: 미래에셋증권 리서치센터

GPT-4.5는 입출력 비용이 수십배 비싸지만 핵심 벤치마크에서는 뒤쳐짐
비용 효율성 측면에서 GPT-4.5는 극도로 불리한 포지셔닝

모델	GPT-4.5	GPT-4o	o3-mini (high)
GPQA (과학)	71.4%	53.6%	79.7%
AIME 24 (수학)	36.7%	9.3%	87.3%
MMLU (다국어)	85.1%	81.5%	81.1%
MMMU (멀티모달)	74.4%	69.1%	-
SWE-Lancer Diamond (코딩) 비용 포함	32.9% ($186,125)	23.3% ($138,750)	10.8% ($89,625)
SWE-Bench Verified (코딩)	38.0%	30.7%	61.0%
입력 가격 (1M 토큰당)	**$75.00**	**$2.50**	**$1.10**
캐시된 입력 가격 (1M 토큰당)	**$37.50**	**$1.25**	**$0.55**
출력 가격 (1M 토큰당)	**$150.00**	**$10.00**	**$4.40**

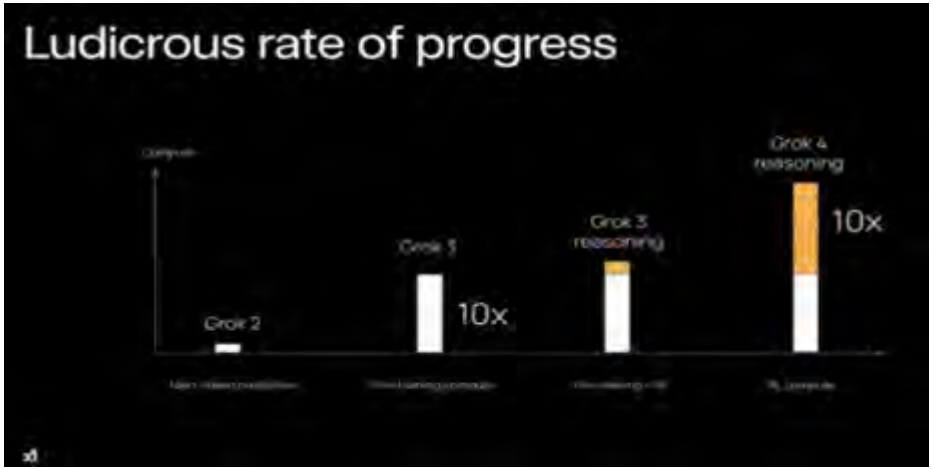
자료: 샘 알트만, 미래에셋증권 리서치센터

II. 주류 연구방향: 새로운 스케일링 법칙의 도래 ①

추론과 강화학습이 미래를 정의한다

- AI 발전 동력이 사전학습(Pre-training) 컴퓨팅의 확장에서 질적으로 전환. "스케일링은 죽지 않았다, 다만 그 대상이 바뀌었을 뿐이다"는 문장으로 현재 상황이 요약됨. 핵심은 추론(inference)과 강화학습(RL: Reinforcement Learning). OpenAI는 "강화학습의 스케일링(또는 Test-Time Computing)이 유일한 해법"이라고 작년부터 주장. AI 지능 향상의 다음 단계가 이미 훈련된 모델을 실제로 사용하는 '답변(추론)' 단계에서 추가적인 컴퓨팅을 투입하여 성능을 극대화라는 것을 의미.
- 과거 강화학습은 모델 출시 전 성능을 다듬는 미세조정 단계에 국한됐으나 이제는 사전학습 이상의 비중을 차지하는 핵심 훈련 프로세스로 격상됨. AI가 통계적 앵무새로서의 단순 패턴 매칭을 넘어 진정한 이해와 추론 능력을 확보하기 위한 가장 확실하고 자본 집약적인 경로로 부상한 것.
- 이는 Chain-of-Thought와 같은 복잡한 추론(reasoning) 작업이 급증해 컴퓨팅 자원의 폭발적 투자의 명분이 됨. Anthropic CEO는 차세대 모델의 강화학습에만 수천억 원 규모의 투자를 집행한다고 공언. xAI는 Grok 2에서 Grok 4로 넘어가며 훈련 컴퓨팅을 100배 투입. Grok 4가 박사 수준이라면, 1년 전 모델인 Grok 2는 고등학생 수준으로 비유되며, 발전 속도의 비선형성을 증명.

우스꽝스러울 정도로 빠른 속도 진전을 보이는 xAI의 Grok, 이유는?
Grok 4는 Grok 3와 비교해 10배의 컴퓨팅 자원을 추론에 쏟음



자료: 미래에셋증권 리서치센터

컴퓨팅 파워를 RL에 주면서, 성능이 향상되는 경향을 명확히 보여줌
OpenAI의 o1 대비 o3는 10배 이상의 자원을 강화학습에 쏟았음

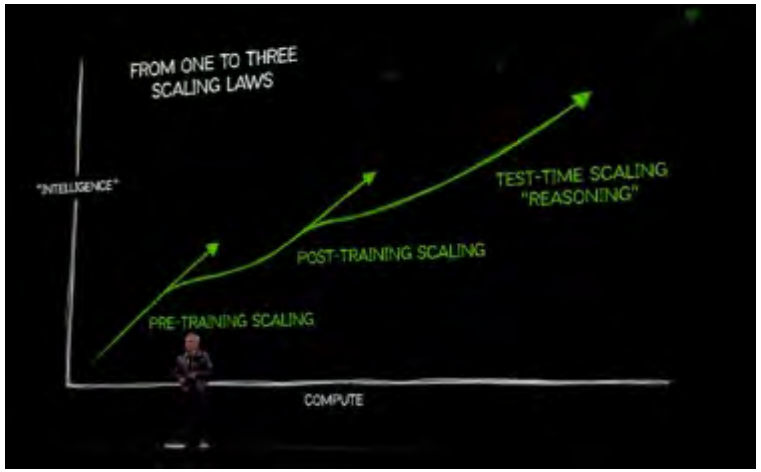
자료: 샘 알트만, 미래에셋증권 리서치센터

II. 주류 연구방향: 새로운 스케일링 법칙의 도래 ②

"생각할 시간"을 부여하는 것의 압도적 가치

- "Test-Time 스케일링"은 모델이 즉시 답변을 출력하는 대신, 충분한 컴퓨팅 자원을 사용해 심층적인 사고의 연쇄를 거치게 하는 새로운 패러다임. 이는 알파고가 수를 두기 전 수많은 경우의 수를 탐색(MCTS)하는 원리와 유사. 알파고와 다른 것은 알파고는 바둑만 두지만 최근의 모델들은 범용 지식을 갖고 있는 언어모델이라는 것.
- OpenAI 연구원 Sebastian Bubeck은 이 방식의 확장이 수학, 과학 등 난제 해결의 결정적 돌파구가 될 것이라 제안. 일반화된 추론 능력의 확장을 의미.
- Bubeck이 제시한 "AGI Time" 개념이 핵심. 모델이 문제 해결을 위해 지금처럼 몇 초가 아닌 몇 시간, 심지어 수백 시간까지 사고하도록 허용하는 미래를 암시. 최첨단의 모델은 이미 코딩, 수학 일부 문제에서 "AGI 시간(hours)" 단위에 도달했으며, 3년 내 "AGI 주(weeks)" 단위 도달을 전망.
- 젠슨 황 CEO의 언급처럼 스케일링 법칙은 벽에 부딪힌 것이 아니라, 추론 시간(test-time)이라는 새로운 차원으로 그 영역을 확장하고 있음.

CES 2025에서 젠슨 황이 강조한 슬라이드
사전학습 스케일링 only에서 두 가지가 추가 됐다



자료: 엔비디아, 미래에셋증권 리서치센터

AI의 추론 능력을 '인간이 문제를 푸는 데 걸리는 시간'으로 계량화한 'AGI 시간'
AGI 달성이 추상적 목표가 아닌, 구체적 일정에 따라 측정 가능한 공학적 과제가 됨

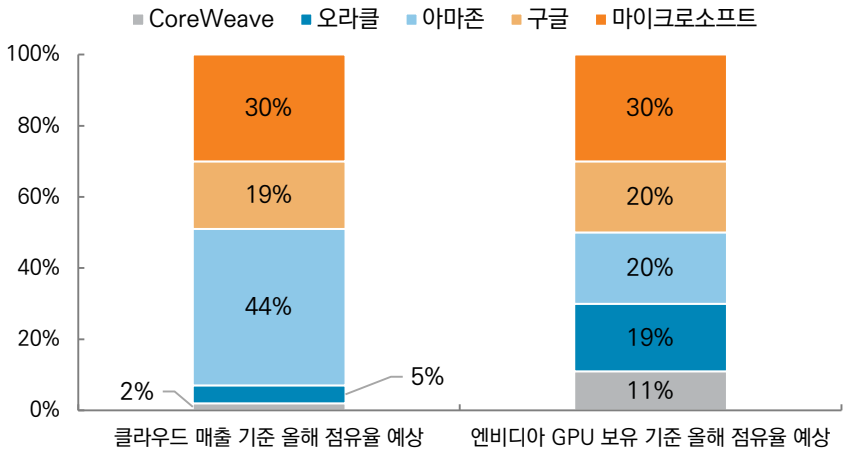
AGI 시간	정의	수준 (OpenAI)	예상 달성 시기
AGI 초	인간이 몇 초 안에 풀 수 있는 문제를 해결하는 성능	GPT-4o	이미 달성
AGI 분	인간이 몇 분 안에 풀 수 있는 문제를 해결하는 성능	o1 시리즈	이미 달성
AGI 시간	인간이 몇 시간만에 풀 수 있는 문제를 해결하는 성능	GPT-5	꽤 많이 달성 중
AGI 일	인간이 하루 내내 고민해야 풀 수 있는 문제를 해결하는 성능	?	2026년 초
AGI 며칠	인간이 며칠을 잡고 있어야 풀 수 있는 문제를 해결하는 성능	?	2026년
AGI 주	인간이 일주일 넘게 봐야 풀 수 있는 문제를 해결하는 성능	?	2027년

자료: Sebastian Bubeck, 미래에셋증권 리서치센터

[참고] 사전학습 확장도 죽은 건 아니다: 지능의 근간으로서 역할 여전

- 추론 컴퓨팅이 새로운 격전지로 부상했으나, 이것이 사전학습 스케일링의 중요성을 약화시키는 것은 아님.
- Anthropic Claude 4 개발팀은 “사전학습의 스케일링 법칙은 여전히 유효하며, 모델이 클수록 소수 예시만으로 패턴을 파악하고 일반화하는 능력이 향상된다”고 강조.
- 거대 모델은 더 큰 정신적 공간(mental space)을 확보하여 추상적 개념화에 뛰어나기 때문. 즉, 더 깊고 넓은 일반화 능력을 의미하며, 복잡한 자율 에이전트(Agent) 개발에 필수적인 학습 잠재력으로 작용.
- 이와 관련해, 시장 데이터는 AI 컴퓨팅 수요가 특정 영역이 아닌, 전방위적으로 폭발하고 있음을 증명.
- 추론 수요는 AI 워크로드의 50% 이상을 차지. 엔비디아 데이터센터 매출의 40% 이상이 추론에서 발생하면서 높은 성장을 보이고 있음.
- 그러나 신흥 클라우드 업체인 CoreWeave의 CEO Mike Intrator는 “대규모 모델 훈련 수요는 전혀 식지 않았다. AI 산업은 여전히 매우 초기 이닝 단계다.”라고 주장.
- 컴퓨팅 시장은 구조적인 공급 부족 상태라는 것. 훈련을 위한 사전학습 스케일업과, 더 높은 품질의 결과를 위한 추론 컴퓨팅 수요가 동시에 폭증하며 AI 인프라 시장의 장기적 성장을 견인할 것이라 판단.

클라우드 매출 및 엔비디아 GPU 보유 숫자 기준 클라우드 점유율 전망
매출은 전통 3社 집중되나 GPU 보유는 CoreWeave·오라클 비중 확대



자료: COATUE, 미래에셋증권 리서치센터

엔비디아가 투자한 신흥 클라우드 업체 CoreWeave 성장 지표
매출·수주잔고·활성/계약 전력 동시 확대

항목	2024 Q2	2025 Q2
매출	$394 million	$1.21 billion (+207%)
수주 잔고	~$15 billion	$30.1 billion (+100%)
활성 전력	220 MW	470 MW (+114%),
계약된 총 전력	600 MW	2.2 GW (+267%)
올해 매출 가이던스	-	$5.15 – $5.35 billion (+363%)
GPU 운영	-	33개 데이터센터에서 25만 개이상의 GPU 보유. GB200/GB300 NVL72를 대규모 제공하는 최초의 기업.

자료: CoreWeave, 미래에셋증권 리서치센터

II. 주류 연구방향: 문제는, 고품질 데이터의 절대적 희소성

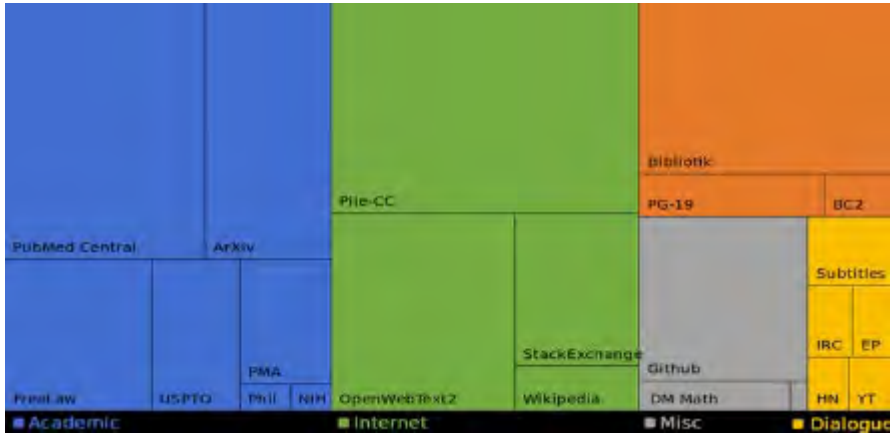
새로운 스케일링의 성패는 결국 데이터와 컴퓨팅에 달렸다

- 강화학습의 성공은 모델의 행동에 명확한 신호를 주는 '검증 가능한 보상(Verifiable Rewards)'에 달려있음. 수학 문제처럼 정답이 명확한 고품질 데이터가 절대적으로 필요하다는 의미. 이때, 기존 인터넷에서 긁어 온 데이터의 근본적 한계가 노출됨. 저품질 콘텐츠가 대부분이며, AI가 전문가의 학문적 진실과 온라인상의 아마추어 의견을 구분 없이 혼합하여 학습하는 치명적 문제 발생. 결과적으로 "AI를 가르칠 교재가 부족"해지는 역설적 상황에 직면. AI 발전의 가장 큰 제약 요인이 됨.
- 첫 번째 해법으로, AI 연구소들은 최고 수준의 인간 지성을 직접 고용하여 고품질 데이터 생산에 투입. OpenAI는 높은 시급을 제시하며 다양한 분야의 석박사 인력을 대규모로 확보. RL에 필요한 고품질 및 난해한 문제, 그리고 정교한 평가 기준(Rubric) 자체를 직접 생성하게 함. 데이터 확보 전쟁인 것.
- 좀 더 편안하고 실용적이며 확장 가능한 해법은 AI가 스스로 고품질 훈련 데이터를 생성하게 하는 '합성 데이터'. 본인이 생성하고 그걸로 다시 학습하는 구조. 이것은 AI 경쟁의 핵심 자원이 데이터에서 컴퓨팅 리소스로 이동했음을 의미. 결국 컴퓨터로 데이터를 생성하는 것이기 때문.
  – 엔비디아 H100 GPU 10만 개를 3개월간 임대해 고품질 합성 데이터를 생성할 경우, 약 6억 5천만 달러의 비용이 추산

OpenAI가 전문가(박사, 의사, 변호사 등)을 모집한 공고
강화학습에 필요한 고품질 문제/답변/평가기준 만들게 함

자료: OpenAI, Mercor, 미래에셋증권 리서치센터

많은 LLM을 훈련시켰던 825GB 크기 오픈소스 데이터셋 “The Pile”
물론 현재는, 기업들의 학습 데이터셋은 기밀이라 공개되지 않는다
게다가 합성 데이터 출현으로 데이터셋 구성이 훨씬 복잡해졌다…



자료: EleutherAI, 미래에셋증권 리서치센터

II. 주류 연구방향: (AGI 달성이) 쉽지는 않아도 불가능하지는 않다

GPT-5 훈련의 핵심은 RLVR(검증기 통한 RL),

RLVR의 핵심은 Universal Verifier

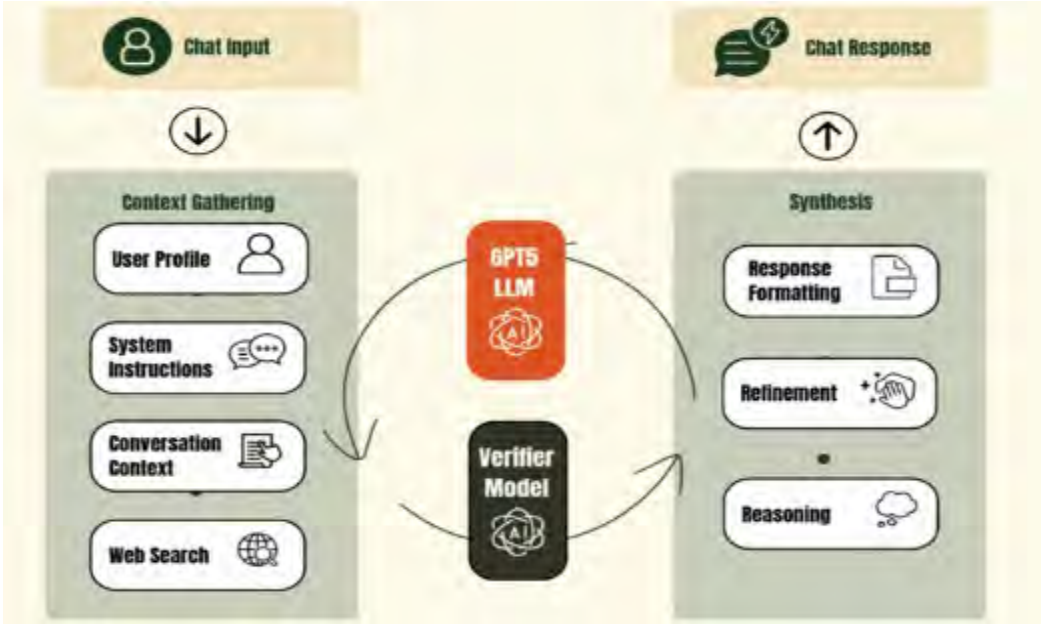
- 고품질 데이터의 부족으로 인해, 기존 강화학습은 개별 인간들의 취향만 학습할 뿐, 추론 과정의 논리적 타당성을 검증하지 못하는 근본적 한계를 지님. OpenAI의 해법은 “Universal Verifier(범용 검증기)”. GPT-5 개발에 적용되는 새로운 강화학습 체계. 단순히 인간 피드백을 모방하는 수준을 넘어, Universal Verifier는 GPT-5가 답변을 사용자에게 보여주기 전에, 그 답변이 얼마나 논리적이고 정확한지를 미리 검수하는 내부 심판관 역할로 일종의 품질 관리를 함.
- 범용 검증기의 핵심은 생성 모델(Generator)과 검증 모델(Verifier)의 2-모델 구조. 생성 모델이 사고의 연쇄 과정(CoT)과 답안을 만들면, 소규모 검증 모델이 이를 평가하고 점수를 매김. 낮은 점수를 받으면 더 나은 답변을 생성하도록 유도하는 방식, GAN처럼 생성과 판별을 반복하며 모델을 강화.

AI가 수많은 시행착오로 스스로 학습하는 공간인 "가상 환경"을 만드는 것이 극도로 어려운 이유

요구사항	핵심 의미
지연 시간 **(보상 지연)**	AI가 행동을 한 후, **가상 환경의 반응이 느리면? AI는 다음 행동을 바로바로 못 하고 낭비**적으로 환경의 변화를 기다리게 되는 문제를 방지해야 한다
안정성 및 장애 허용	**AI가 장시간 작업을 수행하던 중 서버 다운** 등으로 모든 진행 상황이 초기화되는 것을 막는 안정적인 인프라가 필수적이다
보안	AI가 가상 환경을 탈출해 실제 시스템에 접근하거나, **외부 해커가 AI를 조종하는 것을 막는** 강력한 보안 장벽이 필요하다
자기보호	**AI가 무한 루프에 빠져 시스템 리소스를 고갈시키는** 등의 '자해 행위'로 시스템 전체를 마비시키는 행위를 방지해야 한다

자료: 미래에셋증권 리서치센터

'GPT 검증기(Verifier)'가 어떻게 GPT-5의 답변 품질을 높이는지를 보여주는 핵심 작동원리
검증 과정을 통해 환각/오류가 포함된 답변을 사전에 걸러내거나 수정해 답변 신뢰도를 올림



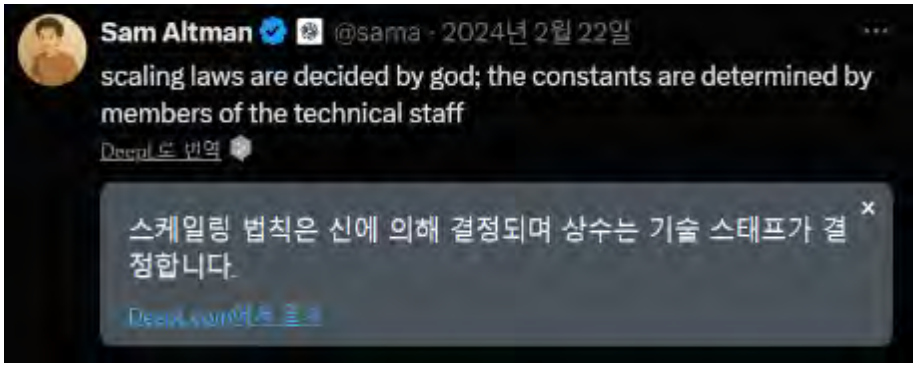
자료: Mike Ebener, 미래에셋증권 리서치센터

II. 주류 연구방향: (AGI 달성이) 쉽지는 않아도 불가능하지는 않다

'장벽은 보이지 않는다'
– Noam Brown

- OpenAI가 GPT-5를 가지고 거둔 성과들은 RLVR을 기반으로 한 스케일링 법칙의 가능성을 보여 줌.
- GPT-5는 2025년 국제수학올림피아드의 6문제 중 5개를 해결하며 금메달의 성과를 달성. 자동화된 검증기가 인간 최고 지성의 세밀한 논리도 정확히 평가하고 올바른 방향으로 이끌 수 있다는 능력을 증명한 셈. 불과 15개월 전, OpenAI 모델은 AIME(미국 수학경시대회) 정답률이 12%였는데, 국제 수학 올림피아드에서 금메달 수준으로 도약한 것에서 그 "미친 속도감"을 체감할 수 있는 단적인 예.
  – 올해 초 OpenAI 수학 천재 노암 브라운 "o1과 (그 다음 모델인) o3 사이의 개선 속도가 앞으로 3개월마다 일어날 것"
- 그리고 IMO에서 금메달 성능을 낸 동일한 범용 모델이 인간이 설계한 규칙이나 특정 도메인에 특화된 튜닝 없이, 국제정보 올림피아드(IOI)에서도 금메달권(인간 참가자 대비 상위 2%)의 최고 성과를 달성.
- 순수한 강화학습 훈련과 더 많은 Test-Time 컴퓨팅만으로 달성되었다는 것이 포인트. 먼저, 지능의 전이가 가능하다는 말. 그리고 사전학습 스케일링 때보다 더 적은 컴퓨팅 투입으로 훨씬 더 큰 성능 이득을 얻을 수 있는 구간이라는 것. AGI 달성은 미지의 과학적 돌파구를 찾아 헤매는 문제가 아니라, 이미 알고 있는 길을 따라 스케일링과 자원을 투입하는 '공학적 문제'로 전환되었다는 강력한 자신감.

AGI로 가는 길(스케일링 법칙)은 자연법칙과 같은 절대적 진리의 길이고, 이제 남은 것은 그 길을 가장 빠르고 효율적으로 최적화하는 엔지니어링의 문제



자료: 샘 알트만, 미래에셋증권 리서치센터

IOI에서 OpenAI 모델과 인간 최고수들의 성적을 직접 비교한 순위표
인간 최상위권과 동등한 수준에 도달했음을 수치로 명확히 입증

Rank	First Name	Last Name	ID	Team	so...	tri...	w...	Inter...	fe...	mi...	ob...	Inter...	Global
1	Hengxi	Liu	CHN4		100	100	100	300	100	91.23	100	291.23	591.23
2	Mingyu	Woo	KOR1		100	99.33	93	292.33	100	82.45	100	282.45	574.78
3	Sizhe	Fan	CHN3		100	78.11	100	278.11	100	91.23	83	274.23	552.34
4	Rareș-Andrei	Neculau	ROU3		100	77.67	100	277.67	100	85	83	268	545.67
5	Xinyang	Chen	CHN1		100	99.02	100	299.02	66	86.4	83	235.4	534.42
—	OpenAI				100	75.29	93	268.29	100	65	100	265	533.29
6	Rain	Jiang	USA1		100	75.41	86	261.41	100	65	100	265	526.41
7	Ryan	Bai	CAN1		100	78.51	100	278.51	66	79.11	100	245.11	523.62

- IOI 대회에서 인간과 동일한 5시간 시간제한, 50회 제출 제한, 인터넷 검색 금지 등 엄격한 조건 하에서 경쟁
- 스캐폴딩 방식은 작년보다 경량화되었으나, 성능은 작년 동메달권에서 올해 금메달권(상위 2%)으로 비약적으로 상승

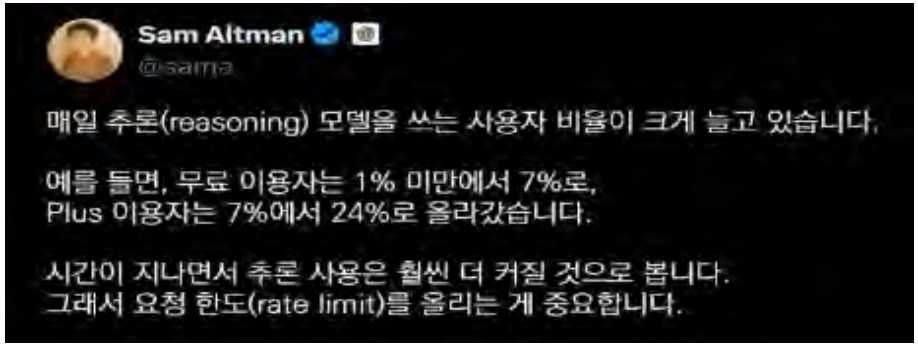
자료: IOI 2025 (International Olympiad in Informatics 2025), 미래에셋증권 리서치센터

II. 주류 연구방향: GPT-5 출시 이후의 전략적 선택

단기 자극의 도파민이 줄어도 실제 효용의 포만감은 늘어갈 것

- GPT-5 출시는 기술적 성과와 무관하게, 총체적인 고객 경험 관리 실패 사례로 기록됨. 샘 알트만은 "완전히 망쳤다(totally screwed up)"는 이례적 표현으로 실패를 인정. 기존 모델인 GPT-4o 모델에 대한 접근을 일방적으로 차단했다가, 가상 친구, 상담 용도로 그 모델을 쓰던 사용자 항의 폭주로 24시간 만에 복구하는 운영 미숙 노출. GPT-5 혼란의 근본 원인은 물리적 한계 때문이라고 봄.
- 추론(reasoning) 모델의 수요 폭증하며 한정된 GPU 공급이라는 벽에 부딪힌 것. 더 이상 모든 전선에서 무한정 자원을 투입할 수 없는 상황. 예를 들어, 연구용 GPU 클러스터에 집중해 가동하는 것은 유료 구독자의 응답 속도를 저하시키는 직접적인 비용이 됨. OpenAI의 자원 배분 원칙은 고객 이탈 방지를 위한 방어적 조치에 가까움. 연구 중심의 '성장 우선' 기조에서 제품 중심의 '수익성 우선' 기조로의 공식 전환 선언. 먼저 지속 가능한 비즈니스 모델을 구축하겠다는 현실적이고 냉정한 판단이 내려진 것.
- 기술 변화 속도가 완만하게 보일 수 있어도, 이는 정체가 아닌 '제품 적합성'의 진전이라고 판단. AI 에이전트 채택은 신뢰성이 좌우. "더 똑똑하지만 가끔 망가뜨리는 모델"보다 "조금 덜 영리해도 믿고 맡길 수 있는 모델"이 실제 업무에 유리. 이제는 환각 감소/토큰 효율성 등이 핵심 레버리지로 부상.

GPT-5 출시 이후 GPU 용량 압박에 대해 언급하는 듯한 샘 알트만의 트윗
Reasoning은 토큰 잡아 먹는 1등 공신인데, reasoning이 폭발적으로 성장 중



OpenAI의 새로운 컴퓨팅 자원 배분 원칙: (단기적으로 연구보다 제품 서빙을 우선시)
제품(유료 구독자) → 약속되어 있는 API 고객 → 무료 사용자 → 신규 API 고객 순서

자료: 샘 알트만, 미래에셋증권 리서치센터

GPT-5에서 '환각(Hallucination)' 감소에 큰 노력을 기울였음을 강조
기업들의 에이전트 도입에 제 1의 장애물이었던 문제를 해결하려는 것

특히 민감한 의료 분야에서 GPT-5는 OpenAI o3 대비 환각 비율을
약 1/4 수준(12.9% -> 3.6%)으로 줄였다고 보고

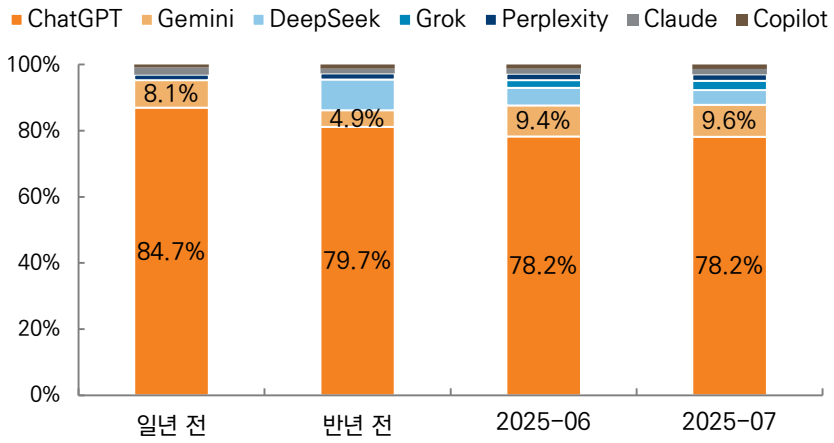
자료: OpenAI, 미래에셋증권 리서치센터

II. 주류 연구방향: 절대 강자 없는 경쟁 국면과 OpenAI의 다음 카드

사람들이 AI에게 있어 정말 원하는 것은 '기억' – 샘 알트만

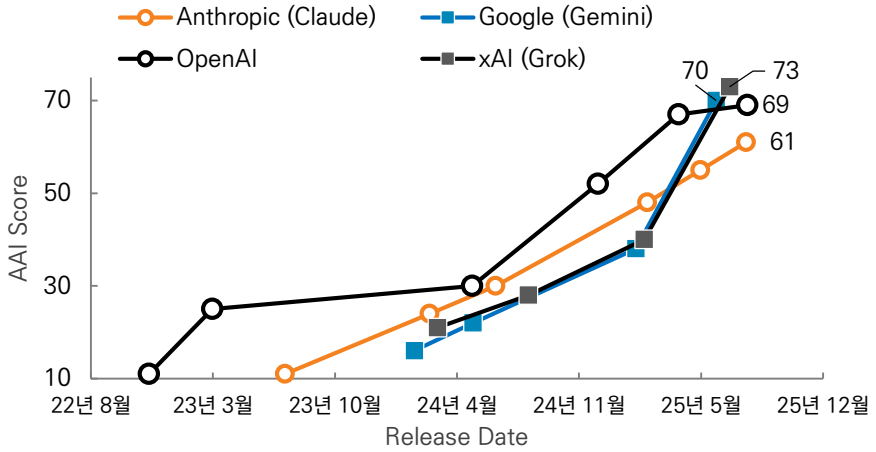
- 샘 알트만은 순수 지능 경쟁의 한계를 인정. "모델들이 채팅(chat) 용도에서는 포화상태"이며, "향후 성능이 오히려 저하될 수 있다"고 진단. 이에 따라 OpenAI는 과거 학습 중심이던 비용 구조를 탈피.
- 실제 매출과 직결되는 추론 비용을 막대하게 책정. 2025~27년간 추론(Inference)에 350억 달러, 학습(Training)에 550억 달러를 투자할 계획. AI 연구소를 넘어 초대형 에이전트 운영사로의 정체성 전환을 의미. OpenAI의 전략 선회는 분명 유의미 하지만 장점만 있지 않음. 순수 연구 혁신에서의 기술 격차는 크게 축소됐다고 보이기 때문.
- 현재 OpenAI의 핵심 해자는 "AI=ChatGPT"라는 압도적 대중 인지도와 7억 명의 사용자 기반, 이를 통한 배포력과 제품 완성도. 반면, 구글(Gemini)은 자체 인프라 및 생태계를 통해 비용 우위를 점하며 OpenAI와 동률 수준까지 추격. Anthropic(Claude)은 안전성을 무기로 기업 시장을 효과적 공략 중.
- OpenAI는 "GPT-6"의 개발 방향은 AI와의 상호작용 방식을 근본적으로 바꾸는 데 있음을 시사. GPT-6의 3대 핵심 축은 기억, 개인화, 지속성. AI의 가장 큰 한계였던 '기억상실증'을 해결하겠다는 선언. 사용자의 생산성과 모든 일상을 함께하는 즉시 전력감 에이전트를 만들겠다는 청사진.

주요 AI 서비스별 월 기준 점유율(도메인 웹 사이트 방문 비중)
Gemini가 비중을 차츰 늘리고 있으나, ChatGPT가 여전히 압도적



자료: SimilarWeb, 미래에셋증권 리서치센터

Artificial Analysis(AA) Index 점수 기준 주요 모델들의 점수 추이
xAI의 Grok 4가 1위이나 성능 격차 미미, 더 중요한 것은 성능향상의 '기울기'



자료: Artificial Analysis, 미래에셋증권 리서치센터

II. 주류 연구방향: 앞으로의 진짜 혁신은 '메모리+세계 모델'

더 잘 '기억'하고, 더 깊게 '이해'하는 AI 에이전트 = 특이점

- OpenAI가 '메모리'를 비전으로 제시했을 때, 구글은 올초 "Titans"라는 논문을 통해 장기 기억 모듈과 영구 메모리를 결합한 아키텍처를 공개한 바 있음. Titans은 인간의 뇌처럼 3가지 기억을 유기적으로 운용하는 구조로, 토큰 비용과 지연시간을 줄이고 답변 일관성(Consistency)을 극적으로 높임.
  - 영구 기억 (Persistent Memory): 사용자의 페르소나, 핵심 규칙 등 변치 않는 정보를 저장. (개인화의 기반)
  - 장기 기억 (Neural Long-term Memory): 대화 중 중요한 새 정보를 선택적으로 저장하고 잊음. (지속성의 핵심)
  - 단기 기억 (Core Memory): 현재 대화의 즉각적인 문맥을 처리.
- 구글 AI의 수장인 데미스 하사비스의 비전은 구글의 비전이 메모리에서 멈추지 않을 것을 시사. ① 메모리(Titans)로 AI가 사용자와 세계에 대한 기억력을 갖게 하고, ②이를 텍스트와 이미지로 가상 세계 그 자체를 생성하는 본인들의 세계 모델인 "Genie 3"과 결합하는 것. 즉, 과거 대화를 기억하는 것을 넘어, 행동의 결과를 예측하고 더 나은 장기 계획(Planning)을 수립하는 진정한 AI 에이전트의 탄생. 완벽한 멀티모달 능력에, 자신을 둘러싼 '세계에 대한 기억력'까지 갖춘 AI라고 할 수 있음.
- OpenAI가 GPT-6로 '기억'의 중요성을 내놨고, 구글은 이미 그 다음 단계로 나아가고 있음.

"Titans" 아키텍처에서 정의되고 활용되는 세 가지 기억의 형태
OpenAI가 목표로 하는 "개인화, 지속성, 일관성"을 달성할 앙상블이 될까?

특징	Persistent Memory	Neural Long-term Memory	Core Memory (Short-term Memory)
저장 방식	고정된 크기 파라미터에 정보 저장	RNN의 Hidden State에 정보를 선택적 저장	현재 대화 내 토큰 간의 관계를 저장 (휘발성)
접근 방식	입력에 Concat되어, 어텐션에 입력	Forward Pass를 통해 Memory State를 업데이트	어텐션 계산을 통해 직접 접근 및 관계 파악
Update 주기	사전학습 단계, Test Time에는 고정	Test Time에도 지속적으로 업데이트	사용자 실시간 대화 기반으로 지속적으로 재계산
역할	Task에 대한 일반 지식 저장	사용자와 대화에서 "Surprise"한 것을 선택적 저장, 무조건 축적 아닌 우선순위에 따라 기억하고 필요하면 잊는 기능.	일종의 작업기억. Context 내의 단기적 정보 파악
의의	Task에 대한 기본적인 이해 유지	과거 경험 바탕으로 현재 상황 이해 및 예측	현재 입력(Context)에 대한 즉각적이고 정밀한 이해

자료: 구글 논문 "Titans", 미래에셋증권 리서치센터

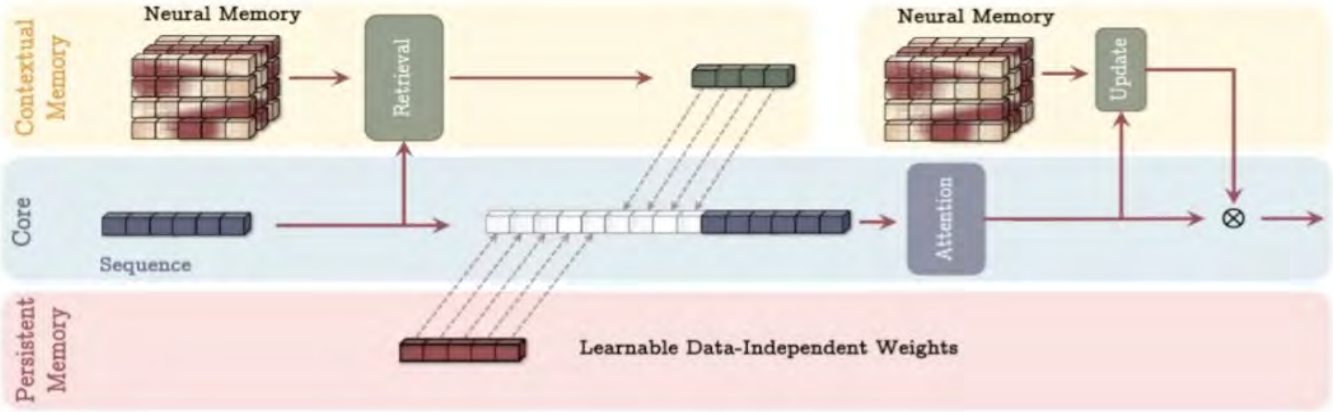
멀티태스크·다중언어 긴 글에서의 LLM 성능을 다루는 '기억' 벤치마크
"LongBench" 기준 순위를 보면 구글의 Gemini가 1등

모델명	기관명	모델 크기	컨텍스트 길이	출시일	종합 성능	긴 글 성능 (128k ↑)
인간	*우주*	*?*	*?*	*?*	*53.7*	***53.7***
Gemini-2.5-Pro	**구글**	**-**	**1M**	**2025-03-25**	**63.3**	**71**
Gemini-2.5-Flash	구글	-	1M	2025-04-17	62.1	**55.7**
Qwen3-235B-A22B-Thinking-2507	알리바바	235B	256k	2025-07-25	60.6	**58.1**
R1	DeepSeek	671B	128k	2025-01-20	58.3	**59.3**
Qwen3-235B-A22B-Instruct-2507	알리바바	235B	256k	2025-07-22	58.3	**55.6**
o1-preview	OpenAI	-	128k	2024-09-12	57.7	**58.1**
R1-0528	DeepSeek	671B	128k	2025-05-28	56.7	**51.4**
MiniMax-Text-01	MiniMax	456B	4M	2025-01-15	56.5	**47.2**
Gemini-2.0-Flash-Thinking	구글	-	1M	2025-01-21	56	**49.1**

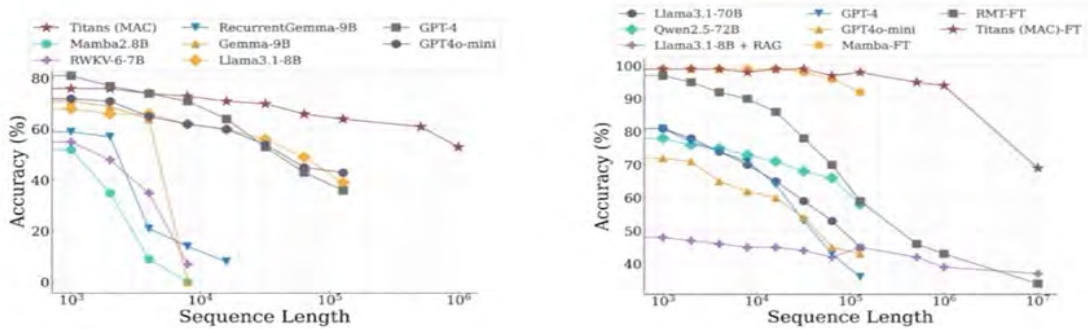
자료: LongBench v2, 미래에셋증권 리서치센터

[참고] 구글의 메모리 관련 논문 “Titans”의 핵심

Titans 아키텍처에서 3가지 메모리가 어떻게 상호작용하는지 시각화
① Core의 현재 입력(Sequence)은 영구 메모리(Persistent Memory)의 사전 지식과 결합
② 동시에 장기 기억(Contextual Memory)에서 과거 경험을 검색(Retrieval)하여 현재 입력에 추가로 보강
③ 이렇게 강화된 정보가 Attention을 통해 처리되고, 그 결과는 다시 장기 기억을 Update하는 데 사용
④ 결론적으로, 사전 지식과 과거 경험을 총동원해 현재 맥락을 정밀하게 이해하고, 동시에 새로운 경험을 학습하는 선순환 구조



“문맥이 길어질수록 정확도가 어떻게 무너지는가”를 비교한 성능 곡선 (왼쪽은 프롬프트만으로 푸는 few-shot, 오른쪽은 소량 파인튜닝을 거친 설정)
별표로 표시된 Titans(MAC)는 장문이 길어져도 “급추락”이 아님. 즉, 프롬프트만으로도 Titans는 긴 문서를 훨씬 안정적으로 읽음



주: 가로축은 시퀀스 길이(토큰 수, 로그 스케일), 세로축은 정답률(%).
자료: 구글 논문 “Titans”, 미래에셋증권 리서치센터

II. 주류 연구방향: 구글이 부리는 마법 “Genie(=실시간 상호작용 월드 모델)”

월드 모델 =
무한한 시뮬레이션 환경 내에서 에이전트 훈련을 가능케 하는 AGI의 토대

- 구글이 내놓은 Genie 3는 세계 최초로, 텍스트 프롬프트만으로 24fps 실시간 인터랙티브 3D 월드를 생성. 사전 제작된 에셋(Unreal, Unity 방식) 없이 완전히 새로운 3D 환경을 즉시 창조. 사용자는 생성된 월드 내부를 즉각적으로 이동 및 조작 가능.
- 현재는 기존의 게임 엔진과 머신러닝 모델의 분리 구조가 무의미해지는 분기점. 이미 올해 Google Cloud 설문에 따르면 비디오 게임 개발자의 87%가 비용 최적화를 위해 AI 에이전트를 도입한 상황.
- Genie 3의 핵심 기술적 성취는 장기 일관성을 갖춘 월드 메모리를 프레임 생성 모델과 접목했다는 점. 이 모델은 이전 수천 프레임의 상태, 사용자 행동, 환경 변화를 실시간 반영해 과거의 상태를 1분 이상 기억할 수 있음. 플레이어의 과거 행동(예: 벽에 그림 그리기)이 월드 메모리에 영구적으로 기록되는 것.
- 이것은, 단순 프레임 생성을 넘어, AI가 스스로 학습하고 실험할 수 있는 場으로써, 디지털 세계 자체를 창조하는 도구를 의미. 이 모델 안에서 또 다른 에이전트가 무한한 시뮬레이션 강화학습을 통해서 성장할 수 있는 것. 즉, ‘AI for AI Training’ 패러다임의 실현.
- 더 나아가, Genie 3는 현실의 물리 엔진까지 AI로 대체하려는 구글의 장기적 비전을 시사. 실제 세계의 법칙을 이해하게 되는 에이전트의 목표는 결국, 로보틱스로 이어지게 될 것으로 판단.

텍스트만으로 생성된 구글 Genie 3의 3D 가상 세계 예시: 0초, 20초, 40초로 이어지는 장면은 사용자가 가상 공간 내부를 실시간으로 이동하는 모습
시간이 지나고 시점이 바뀌어도 특정 나무와 지형(초록색 원)이 그대로 유지되는 것은 AI가 '월드 메모리'를 통해 장기 일관성을 구현했음을 의미

자료: 구글, 미래에셋증권 리서치센터

[참고] 에이전트 완성에 있어서 메모리는 핵심

기억이 개인을 만든다. 하지만 개인이 무엇을 '겪었는가'가 그 개인이 누구인지를 최종적으로 결정

- '기억'은 정체성과 의식을 구성하는 기반. AI 기업들은 기억과 개인화를 통해 '인간다운' 에이전트를 만들려는 청사진을 그리고 있다고 판단.
- 그러나 그 기억이 외부에서 주입되거나(블레이드 러너) 해킹될 수 있는(공각기동대) 취약한 데이터에 불과하기도 함.
- 이는 AI의 '지속성(Persistence)'이 보장되더라도, 그 정체성의 진위 자체를 근본적으로 무너뜨리는 패러독스를 제기.
- 따라서 진정한 자아는 주어진 기억을 넘어, 고유한 체험을 능동적으로 획득되는 '과정'. 기억은 정체성의 시작점일 뿐, 최종 증명서는 못 됨.
- 이에 따라, 차세대 AI 에이전트는 단순 기억 주입이 아닌, 기억을 토대로 어떻게 고유한 정체성을 생성해 나갈 것인가에 대한 고찰이 필요.
- 구글이 만든 프레임 생성 모델 Genie는 "월드 메모리(지속성)"를 통해 에이전트에게 부여하려는 인공적 정체성의 선행 모델일 수도 있음.
- Genie 3가 만들어낸 무한한 가상 세계에서 "에이전트로서 겪는 경험"은, 주입된 기억을 넘어 자신만의 고유한 정체성을 생성할 수 있는 가능성을 기술적으로 암시하기 때문.
- 차세대 AI의 정체성은 고정된 자아가 아니라, 수많은 경험(기억)의 다발(bundle)을 실시간으로 재조합하는 역동적 프로세스가 될 수도…

안드로이드들을 위해 거짓 기억을 만들어주고 있는 '메모리 메이커'의 한 장면
기억(메모리)은 존재와 의식을 구축하는 데에 중대한 역할

자료: 영화 "Bladerunner 2049", 미래에셋증권 리서치센터

"메모리 해킹"으로 거짓 기억이 심어져 있는 범죄자를 심문하는 장면
그러나 기억은 정체성의 진위를 무너뜨리는 패러독스를 야기할 수 있음

자료: 영화 "Ghost in the Shell", 미래에셋증권 리서치센터

III. AI 플레이어

풀스택 五皇 수익화

III. AI 플레이어: AI 신시대의 五皇

AI 에이전트 시대에서 부상할 수 있는 다섯 곳

- 구글은 딥마인드라는 세계 최고 수준의 연구 조직과 압도적인 데이터 보유한 업체로, 자체 평가 기준 추산 1위. 참고로 YouTube 데이터 업로드량은, 1분마다 500시간 분량의 비디오가 업로드되는 수준.
- 2위 그룹 중 하나인 OpenAI는 최근의 핵심 기술인 추론(Reasoning)을 포함해 AI의 패러다임을 바꾼 거의 모든 혁신은 그들로부터 시작되었다는 역사적 사실 때문. 또 다른 추격자인 xAI는 일론 머스크의 스타일이 그대로 반영된 상식을 뛰어넘는 압도적인 물리적 자원 확보 능력 덕분.

주요 AI 모델 제공업체들의 5개 항목별 역량 자체 추산: 평균 점수 기준, 구글(9.0점), OpenAI(8.4점), xAI(8.2점), Anthropic(7.9점), 메타(7.4점)

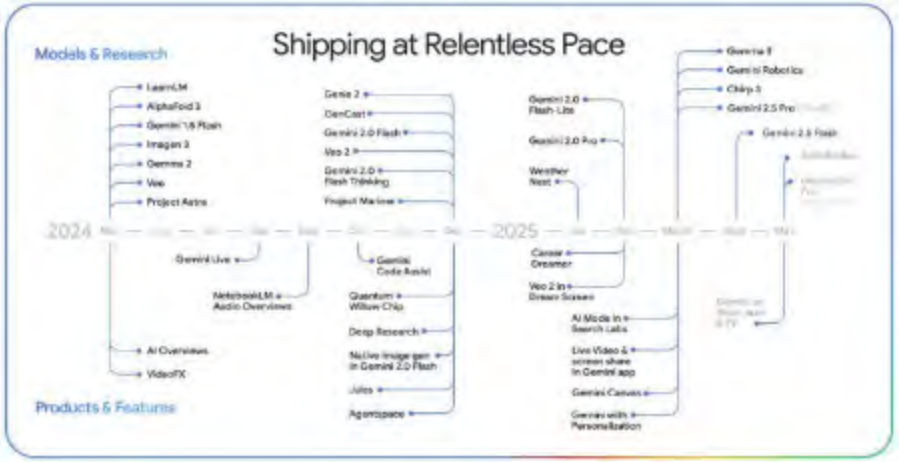
자료: 각 회사, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 구글 '전쟁을 끝내러 왔다'

AI 경쟁의 본질인 Full-Stack에 관한 "Agile Google"

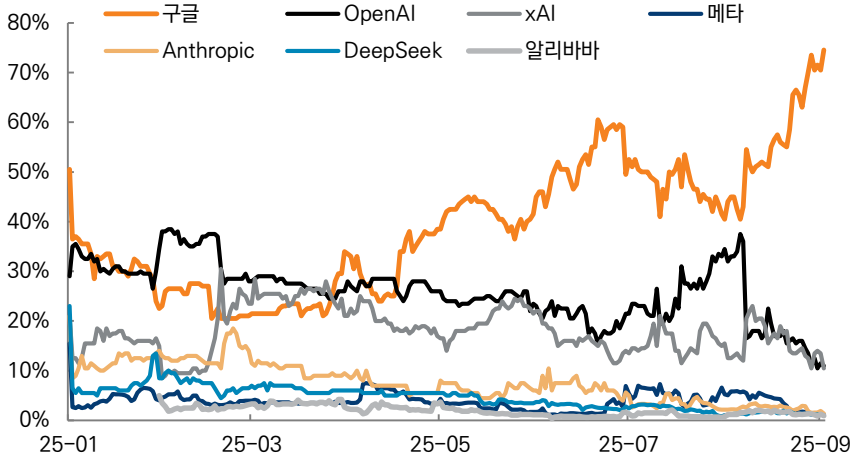
- 느리다는 비판을 받던 구글이 데미스 하사비스의 리더십 아래 무서울 정도로 재편. 구글은 자체 개발 반도체(TPU)부터 클라우드(GCP), 운영체제(Android), 브라우저(Chrome), 최종 단말기(Pixel), 그리고 그 위에서 동작하는 프론티어 AI 모델(Gemini)까지 모든 것을 수직 계열화한 유일한 기업.
  - 젠슨 황 "컴퓨팅 기술의 모든 영역을 종합적으로 볼 때, 구글과 구글 클라우드만큼 전 분야에서 뛰어난 기업은 없다."
- 구글 브레인+구글 딥마인드의 합병 이후의 "사내 정치"가 작년 2분기부터 마무리되며 의사결정 속도 급상승. 업계에 충격을 준 Gemini 2.5 Pro의 빠른 출시 주기는 그 증거.
- Gemini는 월간 활성 사용자 4.5억 명 돌파. 토큰 처리량은 전년 대비 50배 급증해, 현재 구글 플랫폼 전반에서 매달 약 1,000조 개 토큰 처리. 年 20억 달러의 잠재 매출 추정(1K 토큰당 $0.002 비용으로 가정). 핵심 클라우드인 GCP의 연간 매출이 500억 달러라는 점을 감안하면 이는 매출구성의 판도를 바꿀 것.
- 심지어 구글은 경쟁사인 Anthropic에도 30억 달러를 투자해 지분 약 14% 확보한 상황. 리스크 분산 및 이익 극대화 전략.

AI 관련 모델, 연구, 제품 및 기능들의 출시 로드맵은 크게 두 가지 축
작년 2분기부터 구글의 정말로 쉴 틈 없는 페이스의 출시 일정들...



자료: Logan Kilpatrick(구글 Gemini 리드 개발자), 미래에셋증권 리서치센터

베팅 시장에서의 '올해 말 최고의 AI 모델을 보유할 기업'에 대한
베팅 확률 시장의 기대치가 '구글의 승리'로 완전히 쏠렸음을 나타냄



자료: Polymarket, 미래에셋증권 리서치센터

III. AI 플레이어 분석: Banana killed the Photoshop star♫

'나노 바나나'의 실체는 구글의 Gemini 였다

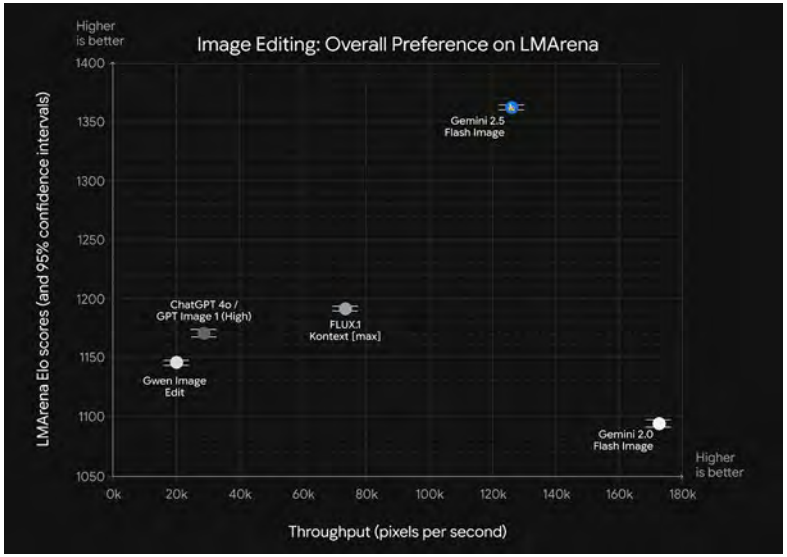
- 구글이 지난 8월 26일 내놓은 Gemini 2.5 Flash Image(코드명 Nano-Banana)는 '말로 하는 포토샵'의 완성. 복잡한 편집 작업을 "배경을 해질녘 해변으로 바꿔줘" 같은 단순 자연어 명령으로 수초 내 완수. 배경과 피사체를 완벽히 분리하고 광원과 그림자까지 재계산. 전문 편집 툴의 필요성을 근본적으로 위협하는 UX 혁명이라고 할 수 있음.
- 벤치마크와 가격, 모든 면에서 경쟁사 압도. 이미지 편집 성능의 Elo에서 1362점을 획득. 2위 모델(1191점)과 171점의 압도적 격차. 벤치마크 역사상 가장 큰 폭의 성능 향상. 이미지 생성 분야에서의 Elo 점수는 1147점으로 imagen-4.0, 헷-image-1 등 모든 경쟁자를 제치고 1위 달성.
- 더욱 놀라운 것은 가격. 이미지당 0.039 달러에 불과해 경쟁사인 OpenAI(0.19 달러) 대비 95% 저렴.
- 출시 후 1주일 만에 1,000만 명 이상의 신규 Gemini 앱 사용자 유입시키는 성공적 모델로 기록됨.

나노 바나나의 경이로운 컨텍스트 이해 및 생성 능력을 보여주는 실증 사례
복잡하고 흐릿한 원본 사진의 군중 속에서 지정된 단 한 명의 인물 완벽 분리
극히 제한된 정보만으로 '정체성'을 보존하며 이미지를 재창조

가능한 다른 명령어 예시: "고개를 정면으로 돌리고 시선을 카메라 쪽으로 맞춰 줘"
"현재 얼굴 특징과 주근깨·피부 톤을 그대로 유지한 채 헤어스타일만 웨이브로 바꿔"
"배경을 황혼의 바다로 바꾸고, 인물은 그대로 두며 역광 느낌을 유지해"

자료: X(@Prashant_1722), LIFE, 미래에셋증권 리서치센터

AI 이미지 편집 모델의 품질(Y축)과 처리 속도(X축)를 동시에 비교
가장 높은 품질 점수와 압도적인 처리 속도를 겸비해 경쟁모델들 압도
'품질과 속도'라는 두 마리 토끼를 모두 잡았음을 증명



자료: LMarena, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 이미지를 넘어 '프레임 생성 AI'의 선두주자

결국 모두 프레임 생성 모델을 만들게 될 것이다

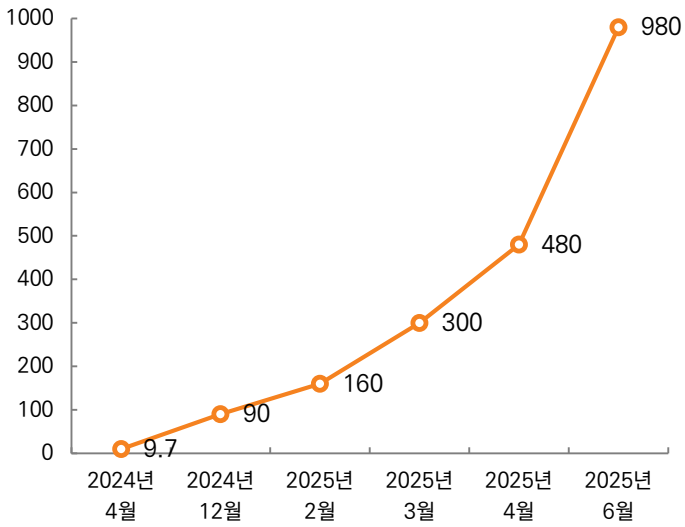
- 구글은 여러 프레임 생성 모델에 관한 연구를 선도하고 있음. 비디오/게임/시뮬레이션 등의 모델을 내놓으면서 월드 모델에 가장 공격적으로 접근하고 있음. 이는 경쟁사들이 텍스트/코딩 위주의 에이전트에 우선순위를 두는 것과 차별화된 연구 스코프.
- 동영상의 개념을 포함한 '프레임 생성'은 텍스트보다 대중성·체류시간·상호작용성이 높아 사용자 증가 및 분당 연산량이 크게 뛸 수 밖에 없음. 우리가 유튜브를 많이 보는 이유를 생각해보면 됨. 그런데 프레임 생성의 비용은(해상도 토큰 수×FPS×모델 스텝 수), 초당 24프레임 기준 기본 계산량이 텍스트보다 100배쯤 무겁다는 것을 고려해봐야 함. 침투율이 S-커브를 따르고 이러한 계산량 증가를 고려하면 얼마나 많은 토큰 처리가 필요할 지 가늠할 수 있음.
- 이때의 병목이자 해결 시 해자가 될 수 있는 것은 역시 전력/GPU 공급. 풀스택을 갖고 있는 구글이 유리한 고지를 점하고 있다고 판단.

구글이 내놓은 프레임 생성 모델에 관한 비교 정리
Genie 3는 AI가 내가 원하는 대로 발봐로 만들어주는 "마인크래프트"라고 보면 됨

구분	Veo	GameNGen	Genie 3
현재 상태	대중 공개	연구용	연구용이나 향후 공개될 듯
도메인	비디오 및 음성 생성	특정 게임을 생성한 후 플레이	월드 모델을 동시에 생성/플레이
해상도	720p ~ 4K	320p	720p
생성 길이	8-10초	수 초	수 분
제어 방식	텍스트와 이미지, 비디오로 생성 명령	키보드·마우스로 제어	텍스트와 이미지로 명령 내리고 키보드·마우스로 제어
상호작용	실시간 상호작용 불가능	이미 학습된 특정 게임 내 실시간 플레이 가능	사용자 입력(키보드, 마우스)에 따라 세상 자체의 환경을 실시간으로 한 프레임씩 생성

자료: 구글, 미래에셋증권 리서치센터

구글 플랫폼 전반에서의 토큰 처리량 추이, 현재 거의 1,000조
프레임 생성 모델이 대중화되면 현재 처리량은 우스운 수준일 것



자료: 구글, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 엔비디아를 위협할 수 있는, 가장 가능성 높은 업체

TPU와 GCP, 엔비디아에 도전하다

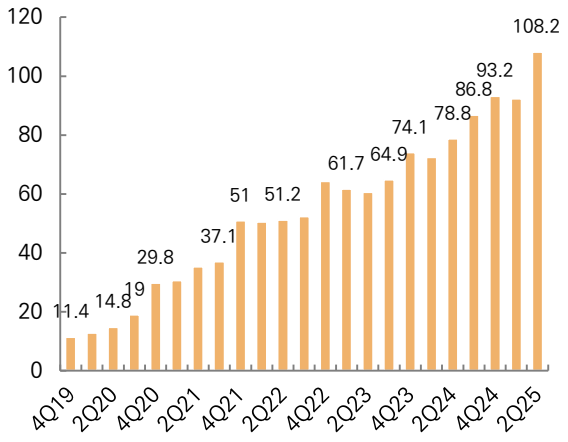
- 구글은 사상 최초로 자사 데이터센터가 아닌 제3자에게 본인들의 칩인 TPU를 배치하는 역사적 결정. 엔비디아 GPU 생태계에 있는 신흥 클라우드 업체들을 직접 공략하는 것. 이미 영국의 Fluidstack社와 뉴욕 데이터센터에 TPU 배치 합의 완료. OpenAI, Microsoft의 핵심 인프라 파트너들(Crusoe, CoreWeave)과도 TPU 배치 협상 중. 엔비디아 GPU 기반 임대 비즈니스 모델을 정면으로 대결.
- 구글은 자사 데이터 센터 건설 속도가 폭발하는 AI 칩 수요를 따라가지 못함. 기존 GCP의 인프라만으로는 TPU 확산에 한계가 있으니 이런 공격적 운영으로 선회한 것으로 풀이.
- 구글 클라우드는 32%의 폭발적 성장. 운영마진은 11.3%에서 20.7%로 9%p 이상 확대, 운영이익 141% 증가한 28억 달러 달성. AI가 GCP의 핵심 성장 동력임을 증명.
- 메타는 자체 데이터센터 확장 지연을 메우기 위해 구글 클라우드와 최소 100억 달러 규모의 6년 계약 체결. 메타의 AI 서비스를 위한 멀티 클라우드 전략의 핵심 파트너로 GCP가 부상. 구글 TPU와 클라우드 개발 플랫폼인 Vertex AI가 대규모 AI 모델 작업에 적합하다는 시장의 방증.

GCP의 가치를 SOTP(Sum-of-the-Parts)로 분석
AI 인프라의 중요성을 시장이 온전히 평가하지 않음

항목	수치
GOOGL 현재 기업 가치 (9월 3일 종가 기준)	약 2.8조 달러
구글 클라우드 인프라 연 매출 (현재 기준 ARR 예측치)	약 550억 달러
GCP 부문 매출 멀티플	12배로 가정 (신흥 클라우드 업체 대비 보수적)
GCP 부문 사업 가치 (매출 기준 멀티플 적용)	**0.67조 달러 = 주가의 약 24%를 설명 가능 = 아직 시장에서 온전히 평가 받지 못하고 있음을 시사**

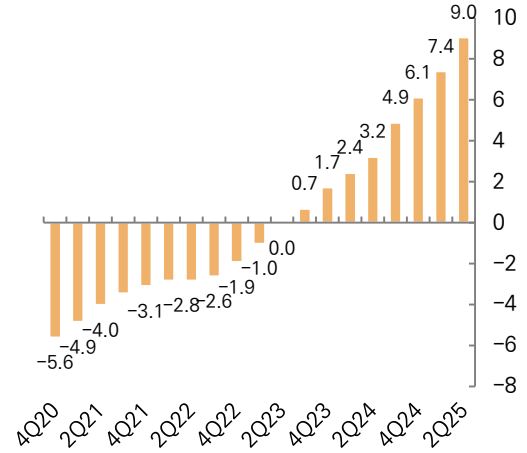
자료: 미래에셋증권 리서치센터

구글 클라우드(GCP) 수주잔고 (단위: 십억 달러)
YoY 38% 증가, 매출 성장률(32%)보다 높음



자료: 구글, 미래에셋증권 리서치센터

구글 클라우드(GCP) 영업이익 (단위: 십억 달러)
GCP 운영마진은 20.7%로 크게 확대

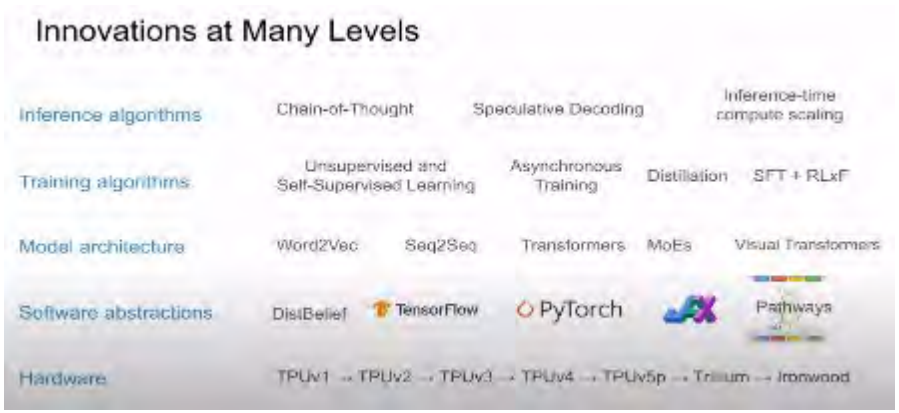


자료: 구글, 미래에셋증권 리서치센터

[참고] TPU를 포함한 구글의 10년 넘는 '도광양회'

- 2013년, 구글은 범용 칩의 비효율성을 극복하고 신경망 계산에만 최적화된 맞춤형 칩, 즉 TPU(Tensor Processing Unit) 개발에 착수.
- TPU v1은 훈련이 아닌 '추론(Inference)'에만 집중. 그러나 본격적인 AI 혁명에 들어서 추론을 넘어 훈련까지 영역을 확장.
- 이를 가능케 한 게, v4 때부터, 수천 개의 TPU 칩을 맞춤형 고속 인터커넥트(OCS: 광 회로 스위칭)로 연결한 'TPU Pod'. TPU v2(2017) 대비 최신 Ironwood(2025)는 약 3,600배의 성능 향상과 30배의 에너지 효율 개선을 달성. 칩 성능을 넘어 시스템 아키텍처 전체의 혁신.
- HW뿐만이 아니라 수만 개의 칩이 여러 데이터센터에 분산된 물리적 한계를 극복하기 위한 소프트웨어 역량도 출중. 개발자는 JAX 라이브러리를 통해 코드를 작성하면, Pathways 런타임이 복잡한 분산 및 통신을 자동으로 처리. 수만 개의 칩을 마치 하나의 거대한 가상 머신처럼 프로그래밍 가능. 하드웨어의 무한한 스케일업/아웃을 소프트웨어가 완벽하게 뒷받침하는, 구글만의 독보적인 시너지 구축.
- 10년 이상 축적된 완전한 Full-Stack은 경쟁사가 단기간에 모방할 수 없는 깊은 기술적 해자.
- 구글 TPU가 외부 시장에 확산될 때 강력한 게임 체인저가 될 수밖에 없는 이유.

구글 AI 경쟁력의 원천을 시각적으로 해부한 기술 계층도(Tech Stack)
해자는 최상단의 알고리즘이 아닌, 이 전체 스택을 수직 계열화한 데서 비롯



Gemini 2.0 모델 또한, 자체 개발 칩(TPUs)부터 분산 훈련 시스템(Pathways, JAX), 효율적 모델 아키텍처(MoE)에 이르기까지, 모든 요소가 유기적으로 결합됐기에 탄생한 것

자료: Jeff Dean(구글 AI의 수석 과학자), 미래에셋증권 리서치센터

7세대 TPU Ironwood는 연산 성능, HBM 메모리, 대역폭 등 지표에서
최신 GB200과 대등하거나 경쟁 가능한 수준에 도달

사양	엔비디아 GPU GB200	7세대 TPU Ironwood	6세대 TPU Trillium 대비 7세대 성능
연산 성능	5 PFLOPs (FP8), 10 PFLOPs (FP4)	칩당 최대 4,614 TFLOPs (FP8 기준)	-
메모리 스펙	192GB HBM3e	칩당 192GB HBM3e	6배 증가
메모리 대역폭	8 TB/s	칩당 7.2~7.4 TBps	4.6배 향상
상호 연결	단방향 900 GB/s NVLink	단방향 600 GB/s ICI	1.5배 향상
전력 소비	약 1,200 와트	약 1,000 와트	성능/전력 효율 2배 향상

자료: 구글, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 기존 플랫폼이 곧 데이터의 원천

유튜브와 구글 검색, 대체 불가능 데이터 우위

풀스택 Go-To-Market

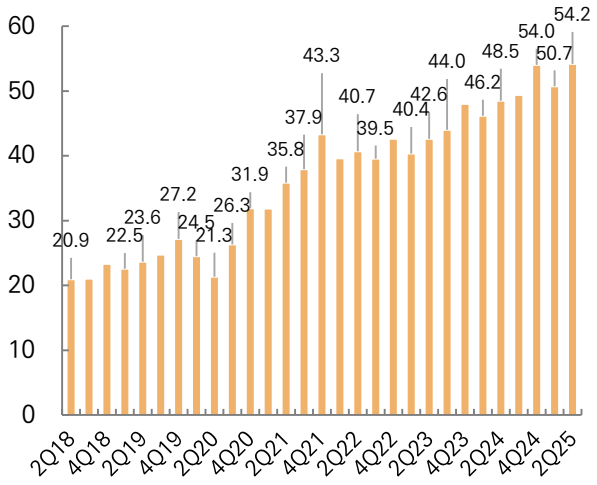
- 구글은 멀티모달 AI 모델 훈련에 필수적인 방대한 고품질 데이터를 이미 보유. 경쟁사들이 데이터 확보를 위해 API 접근 차단 등 폐쇄적 전략을 취하는 상황에서, 구글은 자체의 플랫폼 (워크스페이스·유튜브·검색·안드로이드·크롬 생태계)을 통해 지속적이고 독점적인 데이터 파이프라인을 구축.
- AI Overviews는 월간 사용자 20억 명(약 3달 만에 5억 명 증가)를 기록해 기존 검색 경험을 AI 중심으로 성공적으로 전환을 시사. Gemini 앱 월간 활성 사용자(MAU)는 4.5억 명(ChatGPT의 MAU 8억 명을 빠르게 추격 중), 이들의 사용 데이터는 모델을 더욱 정교하게 만드는 실시간 피드백 루프로 작용.
- "제로 클릭" 검색 트렌드로 인한 기존 광고 모델의 위협은 존재. 그러나 이는 AI 구독 모델로의 전환을 가속화하는 촉매제로 작용하고 있기도 함.
- 지난 9월 2일, 구글이 거의 모든 쟁점에서 승소. 이로써 Chrome 브라우저와 Android OS 매각 리스크 해소 및 핵심 배포 채널 사수. 또한, 연간 200억 달러 규모의 애플 기본 검색엔진에 관한 계약 유지도 허용. 독점적 유통 채널을 지켜내며 AI 서비스의 기본 탑재 우위를 공고히 함.

올해 3~5월, 月방문 횟수 기준 인기있는 웹사이트
구글과 유튜브의 점유율은 압도적 (=데이터 원천)

순위	웹사이트	카테고리	월 평균(십억 회)
1	구글	검색	16.30
2	유튜브	동영상 플랫폼	5.39
3	페이스북	SNS	2.44
4	아마존	이커머스	2.19
5	레딧	커뮤니티	1.92
6	야후	검색 / 포털	1.60
7	인스타그램	SNS	1.00
8	X	SNS	0.95
9	ChatGPT	AI 챗봇	0.75
10	위키피디아	백과사전	0.73
11	링크드인	구인구직 / SNS	0.65
12	Office	비즈니스 생산성	0.51
13	이베이	이커머스	0.48
14	뉴욕 타임스	뉴스	0.47
15	월마트	이커머스	0.43

자료: Similarweb (via Boardroom), 미래에셋증권 리서치센터

구글 "검색 및 기타" 사업 분기 매출 (단위: 십억 달러)
아직은 성공적으로 매출 방어 중



자료: Bloomberg, 미래에셋증권 리서치센터

月사용자 20억 명의 AI Overviews가 뭐죠?
구글 검색 트래픽을 뺏기지 않기 위한 제품

구분	AI Overview
핵심 목적	검색어에 대한 웹 정보의 빠른 요약
사용자 선택	자동 표시 (사용자 선택 불필요)
상호작용	주로 단발성 검색에 대한 결과
결과물	텍스트 요약, 정보 개요, 소스 링크
역할	기존 검색 결과의 보조 및 강화
정보 범위	주로 웹 검색 결과 기반
지향점	"정보를 더 빨리 찾도록 돕는 것"

자료: 미래에셋증권 리서치센터

[참고] 구글의 폰 “Pixel 10”… 하드웨어는 AI 구독을 위한 미끼

- 구글의 자체 제작 스마트폰인 “픽셀 10”에 들어가는 AP(어플리케이션 프로세서)는 Tensor G5 칩. TSMC 3nm(N3E 추정) 기반.
- G6 칩은 벤치마크 점수에서 3년 전 모델인 아이폰 14 Pro Max의 A16 칩보다도 낮은 성능 기록(Geekbench 6 멀티코어: 6,173 vs 6,299).
- 시간 부하 시 열로 인한 성능 하락 지적도 있어, 게임 최상급 성능을 노린 칩은 아님. 이 칩은 구글식 AI 기능을 밀어주기 위해 탄생한 것.
- Gemini Nano/Live 및 실시간 음성 번역(자기 목소리 복제) 같은 체감형 기능이 온디바이스로 돌아가게 최적화로 설계된 폰.
- 즉, 구글은 하드웨어 스펙 경쟁에서 이길 의사가 없는 것으로 판단. 하드웨어는 AI라는 진짜 상품을 전달하기 위한 껍데기이자 매개체일 뿐.
- 이는 테슬라가 본인들의 전기차를 FSD 판매 플랫폼으로 활용하는 것과 동일한 전략으로 풀이. 테슬라는 차량을 감가상각 자산으로 보고, FSD 소프트웨어를 통해 지속적인 수익을 창출하며, 하드웨어를 AI 기반 자율주행의 매개체로 삼고 있음. 구글 역시 Pixel을 AI 생태계의 진입점으로 활용해 소프트웨어 중심 가치 창출을 추구.

픽셀 10은, 하드웨어는 이를 받치는 운반체일뿐 AI가 본상품이라는 전략
약24h+ 배터리·12GB RAM으로 AI 기능을 안정 구동

카메라: 신규 5× 광학 망원, 최대 20× Super Res Zoom.충
하드웨어: Qi2 무선충전 내장, 더 밝은 디스플레이·더 강한 오디오, 내구성 강화.
성능·배터리: Tensor G5, 12GB RAM, 24시간+ 배터리 표기.
소프트웨어: Gemini 중심 AI 기능, Pixel Drops 포함 7년 소프트웨어·보안 업데이트

자료: 구글, 미래에셋증권 리서치센터

구글의 공격적 보급과 생태계 종속(Lock-in) 전략
보급률을 최대화해 “구글 AI Pro” 구독으로 전환시키는 전술

단계	설명
1단계: 보급	4~5년 된 구형 폰을 반납하면 최신 폰 픽셀 10을 거의 무료로 제공하는 파격적 프로모션을 통신사와 함께 진행 중 하드웨어 마진을 포기하고 플랫폼 사용자 극대화에 집중
2단계: 체험 및 락인	픽셀 10 Pro 구매 시, 연간 239달러 가치의 “구글 AI Pro” 서비스 1년 무료 제공 사용자는 점차 AI 경험에 익숙해지고 생태계에 종속됨
3단계: 수익화	무료 기간 종료 후, AI 없이는 불편을 느끼게 된 사용자들이 자연스럽게 유료 구독으로 전환 하드웨어 판매 마진을 압도하는 안정적 구독 수익 창출

자료: 미래에셋증권 리서치센터

III. AI 플레이어: 로드맵에서 알 수 있는 OpenAI의 자신감

과감히 에이전트 수익화 자신감을 내비치는 능력

- OpenAI의 “AGI 달성”이라는 게 대체 무엇인지 그 정의의 상업적 컨센서스가 드러남. OpenAI와 Microsoft는 AGI를 "1,000억 달러 이상의 이익(profits)을 창출할 수 있는 AI 시스템"으로 상업적 합의. 이는 인류 혜택이라는 비전과 투자자 관점의 현실적 목표가 결합되었음을 시사.
- AGI 달성을 위한 구체적 계획으로, AI 에이전트 제품군을 구상. 2025년 전체 매출의 상당 부분을 차지할 핵심 수익원으로 전망. OpenAI는 AGI를 위해 박사급 연구용에 들어갈 수준의 에이전트 제품의 가격을 월 2만 달러 가격대까지 고려 중.
- OpenAI는 “Codex”라는 코딩 에이전트를 아이디어에서 무료 제품 출시까지 7주 만에 완성. 출시 53일 동안 63만 건의 Pull Request를 생성. (개발 엔지니어 1인당 약 7.8만 개의 PR을 만들어낸 수준) 핵심 엔지니어 8명이 만든 실적으로 소규모 정예팀의 속도를 체감할 수 있는 사례. 사실상 “연구 → 제품” 전환이 상시 루프로 작동하는 기업 문화.
- 다만, GPU가 상시 부족하여 에이전트, 코드 생성, 비디오 모델 공개가 지연되는 사례가 발생. 아이디어가 아닌 하드웨어 공급망이 혁신의 발목을 잡을 수 있음을 의미.

OpenAI가 산정한 AI의 능력 레벨에 따른 능력과 잠재적으로 부과할 월 구독료
현재 월 200 달러의 구독 모델만 있지만, AI의 지능이 높아짐에 따라 점점 비싸질 것

AI 레벨	명칭	해당 제품 (예상)	비용 (월)	비고
레벨 1	챗봇 (Chatbot)	GPT-4	$20	기본적인 챗봇 서비스
레벨 2	추론가 (Reasoner)	o1 pro	$200	추론 및 논리적 판단 능력 제공
레벨 3	에이전트 (Agent)	GPT-5	$2,000	자율적 문제 해결 및 작업 수행
레벨 3.5	에이전트 + 컴퓨터 조작	GPT-5-pro	$10,000	중간 정도의 소프트웨어 개발
레벨 4	혁신가 (Innovator)	GPT-6?	$20,000	박사 수준의 지능으로 혁신적 문제해결 가능
레벨 5	AI 조직 (AI Organization)	챗봇 + Operator	$200,000	AI로만 구성된 완전 자율적 조직

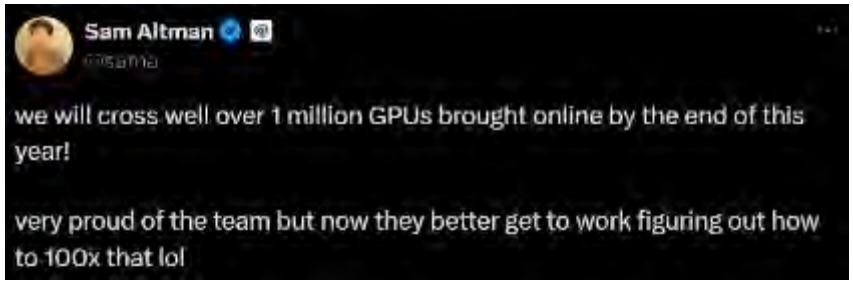
자료: OpenAI, 미래에셋증권 리서치센터

III. AI 플레이어: OpenAI의 보유 자원

컴퓨팅과 데이터 확보 총력전

- OpenAI는 연말까지 100만 개 이상의 GPU를 가동할 것이나, 알트만 CEO는 장기적으로는 현재의 100배인 1억 개 수준으로 확대해야 한다고 발언. 스케일업의 목적은 GPT-6급 훈련과 대규모 에이전트 서빙을 동시에 달성하는 데 있음.
- Microsoft Azure 의존도를 줄이고자 Oracle과 AI 인프라 역사상 가장 거대한 연 300억 달러의 거래를 체결. 사실상 Oracle이 OpenAI 전용 인프라 파트너로 보일 정도. 물리적 규모로는 美 전역 4.5GW 전력 용량으로 전 세계 모든 AI 데이터센터 전력의 30%에 해당. H100 GPU 기준으로는 약 450만 개를 동시 운영 가능한 수준. 심지어 구글의 클라우드 GCP도 병행해 GPU 접근성을 보강. 다만, 모델의 training은 전용 클러스터로 하고 추론(inference)/서빙은 멀티 클라우드로 하는 것으로 생각.
- 주간 활성 사용자(WAU) 7억 명 돌파. 올해 3월 5억 명에서 급증. 연간 사용자 성장률은 4배로 가속. 일일 사용자 메시지는 30억 건 돌파. 이는 모델 성능 개선을 위한 RHLF(인간 피드백 기반 강화학습) 및 데이터 확보의 가장 강력한 파이프라인. 유료 비즈니스 사용자 숫자는 500만 명. 작년 중순 100만 명에서 빠르게 증가. 기업 환경에서의 특화된 데이터와 사용 사례를 확보.

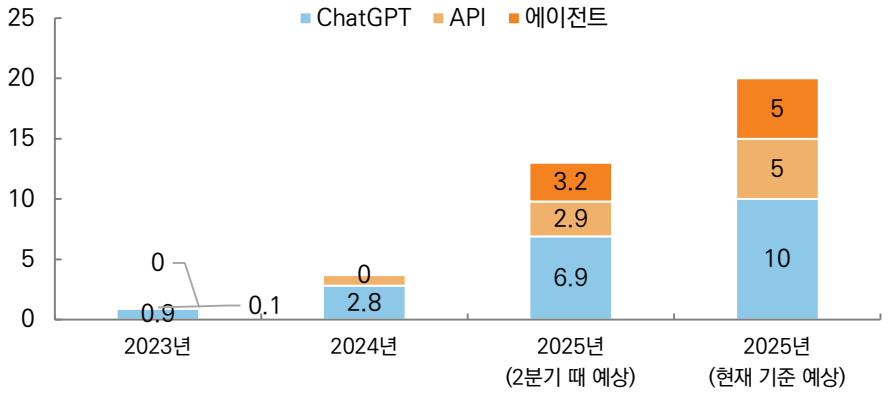
(언제나처럼) 샘 알트만의 GPU가 부족하다고 말하는 트윗
올해 안에 온라인 상태로 가동될 GPU가 100만 개를 훌쩍 넘길 것이고, 장기적으로는 현재의 100배(1억 개 수준)로 확대해야 한다는 문제의식



100만 개의 GPU는 현재 AI 모델을 운영하는 수준을 넘어, GPT-6와 같은 훨씬 더 거대하고 복잡한 모델을 훈련시키기 위한 인프라 투자로 해석

자료: 샘 알트만, 미래에셋증권 리서치센터

OpenAI의 연간 환산 매출치 전망 (단위: 십억 달러)
올해 에이전트 비중이 크게 증가해 연 내 200억 달러 목표
2029년 매출 1,000억 달러 목표를 세웠으나 현재 속도라면 조기 달성할 듯



자료: OpenAI, 미래에셋증권 리서치센터

III. AI 플레이어: OpenAI의 문화와 배포 능력

극도의 비밀주의를 가진 OpenAI의 기업문화는?

새로운 하드웨어로 AI 시대를 정의하려는 야망

- 1년 만에 직원 수 1,000명에서 3,000명으로 3배 증가하며 여러 성장통을 겪고 있음. 조직의 모든 시스템이 끊임없이 파괴적 재창조중. OpenAI는 “아이디어와 결과 중심, 허락보다 실행, 상향식 성과주의”가 기본값. 새로운 사실이 나오면 즉시 진로를 바꿈. 소수 인원이 자발적으로 시작, 성과가 나면 팀이 본격적으로 붙는 방식.
- 그러나 AGI 도달 또는 IPO라는 장기 인센티브는 성공의 불확실성을 내포. OpenAI가 개발하는 거의 모든 IP 권리를 마이크로소프트(MS)가 보유하는 계약구조는 직원들에게 충분한 불편함을 주는 요소. 메타 등 경쟁사가 제시하는 압도적 단기 현금 보상이 인재 이탈을 가속. 다만, 기업가치 5,000억 달러 수준에서 직원들이 가진 지분 유동화(103억 달러 규모)를 추진. 현금화 창구를 열어 이탈을 방지하는 것.
- 한편, MS와의 관계가 재정립되며, "AGI 이후에도" MS가 OpenAI의 최신 모델과 기술에 접근할 수 있도록 보장. MS는 접근권을 보장받는 대가로, 기존의 복잡한 이익우선 분배 구조 대신 30~35% 수준의 지분을 받는 방향으로 전환. OpenAI는 복잡한 비영리 구조 개편 및 상장(IPO)의 문을 열게 됨.
- OpenAI는 애플 디자이너 총괄이었던 조니 아이브의 스타트업 ‘io’를 65억 달러에 인수. 아이브는 OpenAI의 하드웨어 디자인을 총괄. 인간의 오감을 AI와 자연스럽게 공유할 새로운 형태의 디바이스로써, 더 많은 사용자가 AI 모델과 깊숙이 상호작용하도록 유도. 2027년 경 제품 출시 예상.

비영리 연구조직일 때의 OpenAI vs 현재의 OpenAI

자료: X(@signulll), 미래에셋증권 리서치센터

IPO를 준비중인 OpenAI는 마이크로소프트와의 관계를 재설정하는 딜을 추진 중
OpenAI는 MS가 향후 이익 공유 권리를 포기하고, 대신 새로 전환되는 지분의 약 33%를 제안

제안 핵심	마이크로소프트가 얻는 것	마이크로소프트가 포기해야 할 것
OpenAI의 보통주 33% (희석 방지 조항 일부 포함)	• OpenAI IPO 이후 배당·주가 상승 수혜 • **AGI 선언 시 계약 종료라는 조항 삭제 → 미래 최첨단 모델 접근 보장**	• 49%의 profit-share 권리 • 20%의 매출 공유 • 투자 원금의 100배라는 캡 구조

자료: 미래에셋증권 리서치센터

[참고] OpenAI의 최대 리스크는 리더십일 수도…

OpenAI 전직 직원들의 증언 내용 요약

주요 비판 주제	인물	OpenAI 때 직책/소속	핵심 증언 내용
샘 알트만의 리더십	일리야 수츠케버	공동창업자	• "AGI의 버튼을 누를 인물이 아니다"라며 리더십에 대한 불신 표명 • 수년에 걸쳐 부정직하고혼란을 야기하는 행동 패턴을 보였다고 비판
	미라 무라티	최고기술책임자 (CTO)	• "샘 알트만이 우리를 AGI로 이끄는 것이 편하지 않다"며 불안감 표출 • 원하는 것을 얻기 위해 회유하고, 통하지 않으면 상대를 폄훼/파괴하는 독선적 운영 방식 비판
	제프리 어빙	기술 직원	• 자신에게는 친절했지만, 여러 차례 거짓말을 했으며 타인에게는 기만적이고 조종하려 했다고 증언
	타샤 맥컬리	이사회 멤버	• "기만, 가스라이팅, 철저한 감독에 대한 저항"은 AGI 리더에게 용납될 수 없는 자질이라고 강조
	대니얼 지글러 / 캐롤 웨인라이트	기술 직원	• 광범위한 기밀 유지 계약이 전·현직 직원들의 내부 고발과 책임 추궁을 막고 있다고 비판
	토도르 마르코프	기술 직원	• 퇴사자 비방 금지 계약에 대해 직원들에게 거짓말을 한 '청렴성이 낮은 인물'이라고 비판
	다리오 & 다니엘라 아모데이	연구/안전 담당 부사장	• 알트만의 행동 방식을 "가스라이팅"과 "정신적 학대"로 묘사
비영리 임무 이탈	그레첸 크루거	정책 연구원	• 회사 근간이었던 비영리 거버넌스와 수익 상한선이 사라지는 것은 정체성을 부정하는 것이라 비판
	제프리 우	기술 직원	• OpenAI의 비영리 임무가 완전히 '수익과 일치하도록' 재해석되었다고 지적
	윌리엄 선더스	기술 직원	• 영리화 구조는 안전을 무시하고 편법을 쓰려는 유인을 키울 것이라고 우려
	캐롤 웨인라이트	기술 직원	• 중대한 시점에 비영리 구조를 포기한 것은 결국 "약속이 공허했다"는 의미라고 비판
	스티븐 애들러	연구 및 프로그램 담당	• 비영리라는 제약으로 얻은 이점이 있으므로, 이제 와서 그 제약을 저버려서는 안 된다고 주장
	헬렌 토너	이사회 멤버	• 내부 안전장치는 취약하며, "공익을 위한 비영리"라는 주장은 믿기 더 어려워지고 있다고 비판
안전 및 투명성	얀 라이케	강화학습 팀장	• "안전 문화와 프로세스가 화려한 제품에 밀려 뒷전이 되었다"고 폭로 • 안전 연구에 필수적인 컴퓨팅 자원 확보가 어려웠다고 증언
	대니얼 코코타일로	거버넌스 팀	• 회사가 AGI를 책임감 있게 다룰 것이라는 신뢰를 잃어 퇴사했다고 밝힘
	레오폴드 아셴브레너	슈퍼얼라인먼트 팀	• 안전 연구를 위한 자원(컴퓨팅 20%) 할당 약속을 어겼다고 비판 • 내부 보안 문제를 지적하는 메모를 작성했다는 이유로 해고되었다고 폭로
	윌리엄 선더스	기술 직원	• 재직 당시 GPT-4 등 핵심 모델을 탈취할 수 있는 심각한 보안 취약점이 장기 방치되었다고 증언

자료: OpenAI Files, 미래에셋증권 리서치센터

III. AI 플레이어 분석: xAI, 일론 머스크라는 천재와 함께 고속성장

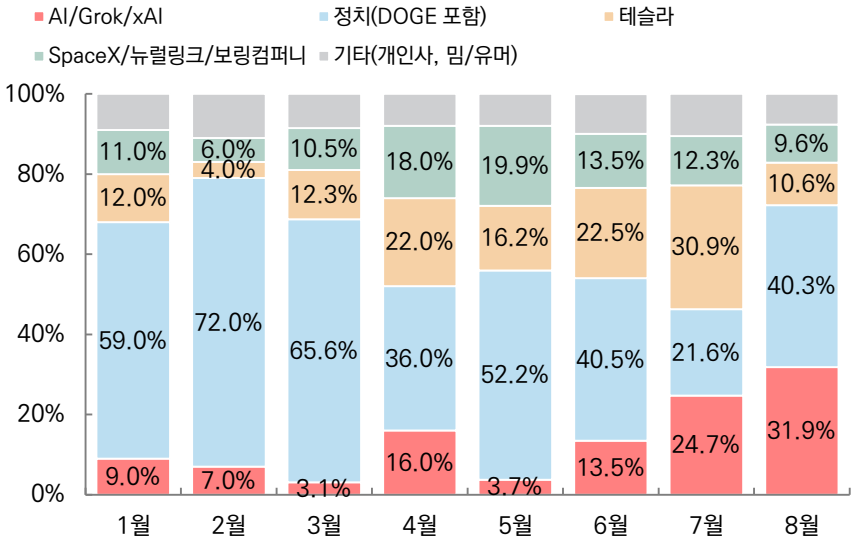
xAI:
일론 머스크의
現 최우선순위 비즈니스

- 일론 머스크가 2023년 3월에 세운 xAI는 그 창립자의 천재성이 가장 큰 자산이자 또 리스크인 조직. 일론 머스크의 기업들은 머스크가 다른 일에 열중할 때는 주당 50시간 근무하고, 개입이 많아질 때는 주당 100시간 근무해야 하는 문화를 가지고 있음. 혁신 촉발과 번아웃 리스크가 공존하는 것.
- 최근에는 핵심 인력들의 이탈 현상 발생. CFO Mike Liberatore가 재직 4개월 만에 7월 말에 회사를 떠났고, 공동 창립자이자 유명 엔지니어인 이고르 바부슈킨도 'AI 안전' 문제에 집중하고 싶다고 퇴사.
- 그러나, 2년 만에 직원 수 1,200명 이상으로 100배 증가한 기업이기에 이런 성장통은 정상적이라 판단. 오히려 머스크가 가진 '제1의 사고방식 원칙'은 기존 가정을 뒤엎고 혁신을 촉발하는 리셋 버튼.
- 최근 머스크는 Grok 앱 업데이트 등 xAI(Colossus 포함)를 홍보하는 일이 잦음. AI 트렌드 주도 의지. 이러한 AI에 관한 그의 관심이 두드러지게 나타나는 것은 xAI의 빠른 진전에 뚜렷한 도움을 주고 있음.

머스크의 지난 7월 20일 게시물
AI를 중심으로 전력투구하겠단 외침, 특히, 속도전이라는 선전포고

자료: 일론 머스크, 미래에셋증권 리서치센터

올해 1~8월 기준, 일론 머스크가 작성한 X 게시물의 테마별 백분율 분포
정치가 내내 주요 테마였지만, 하반기부터 AI/Grok/xAI 관심 증대 포착



자료: X, 미래에셋증권 리서치센터
주: 월별 500개의 게시물 샘플을 바탕

[참고] xAI는 반년에 한 번씩 앵커 모델을 출시 중

구글보다 15년 늦게, OpenAI보다 7년 늦게 시작했다… 그럼에도 xAI는 1티어 그룹이다

날짜	이벤트	세부 설명
2023년 03월 09일	회사 설립	OpenAI, 구글 딥마인드, 구글 브레인, 마이크로소프트 등의 경험이 있는 소수의 인재들로 구성
2023년 11월 03일	**Grok-1 얼리 액세스 출시**	**회사가 정착한 것 기준으로 반년도 안돼 나온 첫 공개 빌드**
2023년 12월 07일	X의 Premium+ 사용자를 위한 Grok 출시	X와 xAI의 시너지 효과 위한 움직임
2024년 03월 17일	Grok-1 오픈소스 출시	"프론티어 모델은 폐쇄소스, 이전 모델은 오픈소스"의 전략
2024년 05월 10일	시리즈 B 펀딩	60억 달러 자금 조달, 첫 공식 밸류에이션으로 기업가치는 240억 달러로 평가
2024년 05월 15일	**Grok-1.5 출시**	**GPU 분산 훈련 본격화. 수학 및 코딩 분야에서 이전 세대 대비 2배 넘는 성능 향상. 128,000 토큰의 장문 컨텍스트 처리 능력(Grok 1 대비 16배)**
2024년 08월 13일	**Grok-2 출시**	**H100 GPU 약 6,500대로 훈련**
2024년 11월 04일	Aurora 출시	FLUX를 버리고 Aurora라는 자체 이미지 생성 모드를 개시
2024년 11월 20일	시리즈 C 펀딩	60억 달러 자금 조달, 이때의 밸류에이션은 기업가치로 500억 달러라는 성과. OpenAI는 같은 밸류에이션에 9년 소요
2025년 02월 17일	**Grok-3 출시**	**H100 GPU 약 10만대로 훈련(Grok-2 대비 15배)해 세계 최고 성능의 AI 모델이 됨**
2025년 03월 19일	Deeper Search 롤아웃	깊은 웹·X 탐색 + 긴 추론 보고서
2025년 06월 30일	부채 50억 + 지분 50억 조달 완료	인프라 확장 목적의 대규모 자금 조달에 성공
2025년 07월 09일	**Grok-4 출시**	**SuperGrok Heavy(월 $300)이라는 최상위 요금제 신설 여러 벤치마크에서 경쟁 모델과의 격차를 벌리며 기술적 충격을 줌**
2025년 07월 14일	Grok for Government 공개	연방·주정부 대상 제품군/계약 트랙 공식화(DoD 계약 상한 $200M 포함)
2025년 08월 22일	Macrohard 프로젝트 공개	AI 에이전트만으로 운영되는 순수 AI 기반 소프트웨어 회사 설립을 목표로 제시
2025년 08월 28일	Grok Code Fast 1 출시	코딩에 특화된 모델을 공개하며, 출시 직후 사용량에서 전례 없는 성과를 달성

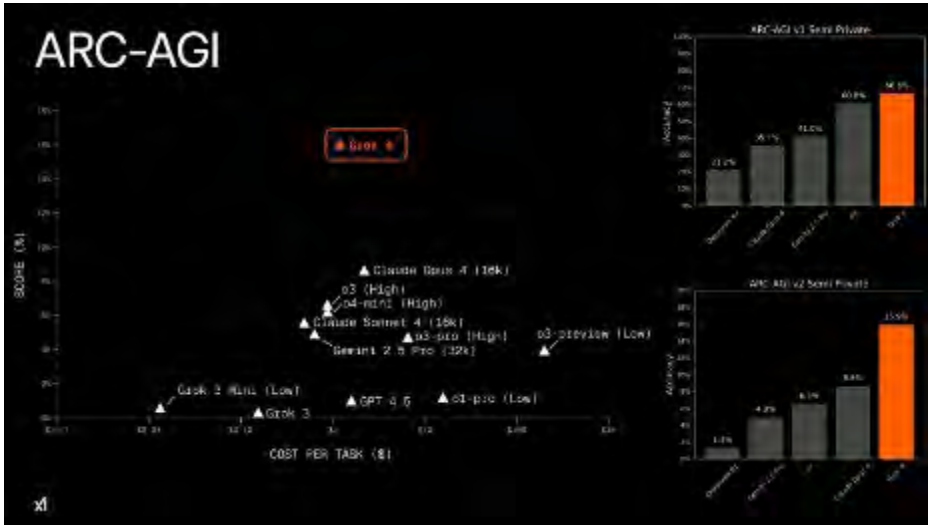
자료: OpenAI Files, 미래에셋증권 리서치센터

III. AI 플레이어: 세계 최고의 지능 중 하나인 Grok 4

SHOW and PROVE

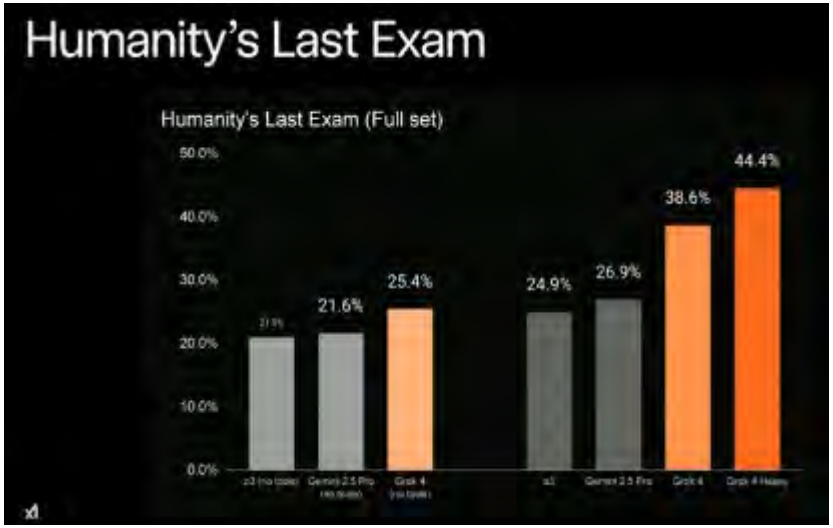
- xAI의 최신/최고 지능의 모델 Grok 4는 순수 텍스트 지능과 도구 결합 지능 모두에서 급진적으로 상승. Grok 4는 Grok 3의 사전 훈련 컴퓨팅은 유지하고, 강화학습 컴퓨팅에만 10배를 투자. 대형화가 아니라 주류 연구 패러다임이 된 강화학습 스케일링을 선도적으로 활용해 지능을 향상 시킨 것.
- Grok 4는 HLE(인류가 AI에게 낼 수 있는 마지막 시험지) 벤치마크 결과, 압도적 추론 능력 입증. 특히 협력적 추론(Grok 4 Heavy의 기능)했을 경우 50.7% 달성. 인간 전문가도 5% 이상 정답률을 기록하기 어려움.
- 유동적 지능(Fluid Intelligence)에서 경쟁사를 압도. ARC-AGI v2(고난도) 시험 기준, Grok 4 (15.9%)는 Claude Opus 4 (8.6%) 대비 약 2배 성능. 처음 보는 문제의 패턴 파악 및 규칙 학습 능력에서 격차를 내고 있음을 의미.
- 물론 약점도 있음. 시각적 디테일 인식, 디자인 감각 등 멀티모달 시각 학습 분야에서 열세. 이에 대해서는 9월(멀티모달 에이전트), 10월(비디오 생성)에 만회하겠다고 타임라인을 밝히기도 했음.

ARC-AGI 기준 성능과 비용 효율성 분석
Grok 4는 약 17%의 점수로 경쟁 모델(Claude, Gemini)을 압도
v2 테스트에서는 v1보다 점수 차가 더 큼. 어려운 문제일 수록 더 잘한다는 뜻



자료: xAI, 미래에셋증권 리서치센터

HLE(Humanity's Last Exam)에서의 Grok 4 Heavy의 압도적 성능
HLE는 100개 이상의 학문 분야에 걸쳐 각 분야 최고 전문가들이
직접 출제한 2,500개의 극도로 어려운 문제로 구성되어 있음



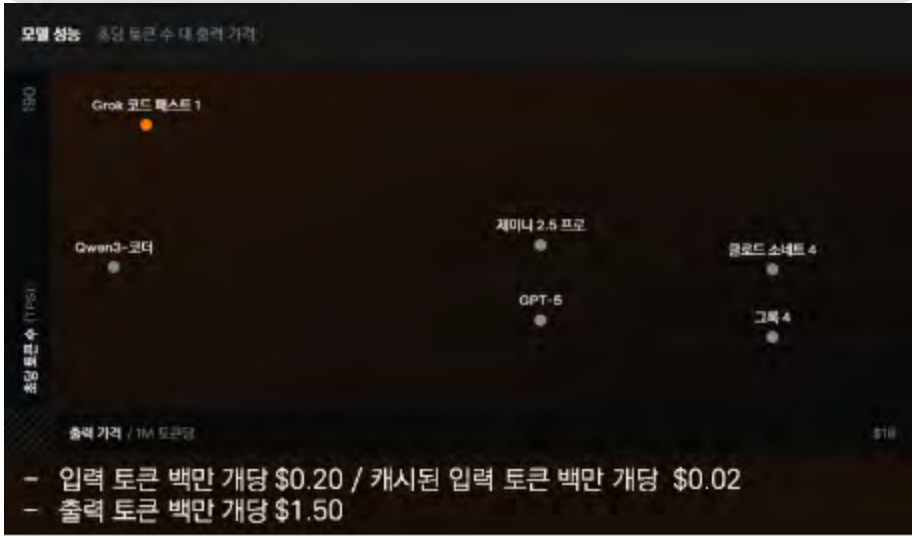
자료: xAI, 미래에셋증권 리서치센터

III. AI 플레이어: 코딩 특화 모델도 사용량 1위로 등극

빠른 속도로 약점은 바로바로 메운다

- xAI가 비즈니스적으로 가장 열세였던 부분은 엔터프라이즈 시장에서의 존재감. 그리고 그 시장은 코딩 에이전트와 연관되어 있음. 이를 극복하기 위해 xAI는 지난 8월 28일 코딩 특화 모델 “Grok Code Fast 1”을 출시. 명확한 메타인지와 약점 보완을 위한 긴박한 개발 속도를 의미.
- 이 모델은 “제로 투 원” 프로젝트 구축할 수 있고, 코드베이스 질문에 대한 통찰력 있는 답변 제공부터 외과적 버그 수정 수행에 이르기까지 일반적인 프로그래밍 작업을 완료하는 모델.
- xAI는 TypeScript, Python, Java, Rust, C++ 및 Go 등 거의 모든 소프트웨어 스택에서 능숙하다고 홍보하고 있음. 실제로, 이 모델은 OpenRouter(여러 AI 제공업체의 다양한 LLM을 하나의 통합 API로 사용할 수 있게 해주는 라우팅 플랫폼)에서 단 하루 만에 1,380억 개의 토큰이 사용됨. 이는 플랫폼 역사상 최대치.

AI 코딩 모델들의 성능 대비 가격 효율성을 비교한 차트(Y축 속도/ X축 출력가격)
Grok Code Fast 1: 최고의 속도, Qwen 3와 비슷한 최저가격
Grok 4: 최악의 속도, 그럼에도 가장 비싼 축에 속함



자료: xAI, 미래에셋증권 리서치센터

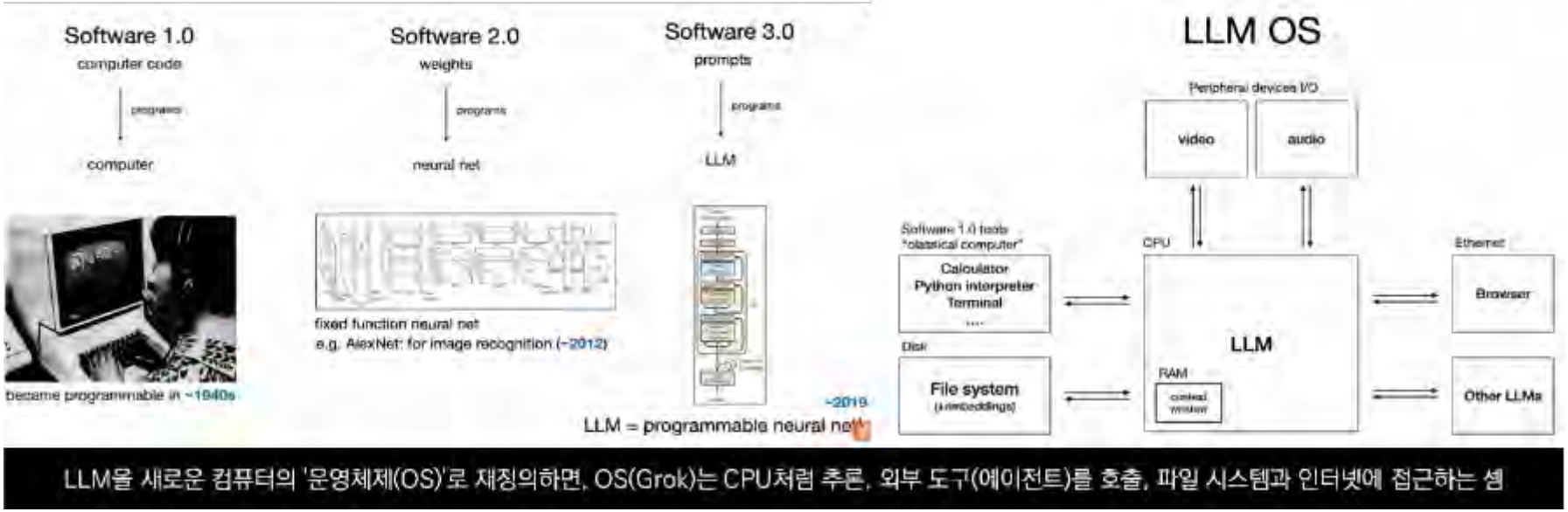
출시 일주일이 지났지만 Grok Code는 OpenRouter에서 최상위권 유지
특히 프로그래밍 시장 점유율은 50% 이상으로 압도적
출시 약 2주 만에 OpenRouter에서 1조개 토큰 사용량 달성

자료: OpenRouter, 미래에셋증권 리서치센터

[참고] Microsoft(SW 1.0)가 아니라 이제 Macrohard(SW 3.0)

- 일론 머스크가 지난 8월 22일에 밝힌 바에 따르면, xAI의 궁극적 목표는 인간 개입 없이 AI 에이전트들이 자율적으로 회사를 운영하는 시스템, 즉, '멀티-에이전트 AI 소프트웨어'를 만드는 것. 이에 따라 "Macrohard"라는 xAI의 새로운 프로젝트 발표. Macrohard는 AI를 '보조 도구'로 쓰는 기존의 기업 모델(소프트웨어 1.0)을 정면으로 부정.
- 대신, LLM(Grok)을 핵심 운영체제(OS)로 삼는 Native 모델. 물리적 생산이 없는 SW 회사는 AI로 전체 시뮬레이션이 가능하다는 논리.
- 이 시스템에서 Grok은 CEO 또는 총괄 매니저 역할을 맡아 목표를 분석해 수백 개의 전문 AI 에이전트를 즉시 소환(spawn). 각 에이전트들은 코딩, 디자인, QA 등 전문화된 'AI 직원'으로 기능. 그래서 AI 팀으로 구성된 하나의 조직(organization)인 것.
- 먼저 AI 팀이 결과물을 만들면, '가상 플레이어 AI'가 곧바로 테스트를 시작. 가상 플레이어는 인간처럼 소프트웨어를 사용하며 피드백을 수집하고 이를 통해 실시간으로 개발팀 AI에 전달되어 즉각적인 개선이 이뤄지는 것. 결과물이 완벽하다고 판단될 때까지 24/7 무한 반복.

안드레 카파시(OpenAI 공동창립자 / 테슬라 AI 디렉터)가 Y 콤비네이터 강연 중에 소프트웨어의 변화상을 설명한 부분을 발췌
마이크로소프트의 이름과 비슷하다고 해서, 매크로하드는 진지하지 않은 프로젝트가 아니다
안드레 카파시가 정의한 소프트웨어 패러다임의 진화(1.0 → 2.0 → 3.0)의 논리적 종착점인 담대한 기술적 도박



자료: 안드레 카파시, Y Combinator, 미래에셋증권 리서치센터

III. AI 플레이어: xAI를 높게 쳐주는 이유는 보유한 자원 때문

보유 자원: Colossus & Kingdom of Musk

- AI 경쟁은 스케일링의 싸움. xAI는 설립 초기부터 이 분야 선두권이었음. 슈퍼컴퓨터 클러스터인 "Colossus"를 보유했기 때문. AI 훈련용 클러스터 규모에서 경쟁사를 크게 따돌리고 세계 1위로 등극.
  - 물리적 GPU 숫자는 23만 개: H100 150,000개 + H200 50,000개 H200 + GB200 30,000개
- Colossus 프로젝트의 다음 단계로 약 100만 개의 GPU 탑재로 나아가는 중. 이를 위해 엔비디아의 GB200 초기 물량의 약 10%를 단일 고객으로 소화했고, 연말까지 15% 비중까지 늘어날 가능성. AGI 경쟁에 있어서 자신감의 물리적 근거. 머스크가 'AGI 후보'로 말한 Grok 5는 9월부터 훈련 돌입 .
- 데이터 확보면에서 xAI는 머스크 킹덤의 '제후국'으로서 수혜를 누릴 위치. X(하루 5억 개 이상의 실시간 텍스트 및 영상 데이터), 테슬라(실제 주행 환경에서 수집되는 방대한 양의 실세계 시각 데이터), 스페이스X(로켓 발사로부터 얻는 물리 법칙 데이터 등)로부터 고품질 데이터를 수혈 받을 수 있기 때문. 빅테크 기업들이 자사 데이터 API 접근을 차단하는 추세 속 안정적인 고품질 데이터 파이프라인은 결정적 경쟁 우위.

머스크 제국의 일원들과 이들의 앙상블로 만들어내는 '재귀적 경제 프로토콜(Recursive Economic Protocol)'
개별 기업들의 컴퓨팅, 데이터, 모델, 자본, 인재가 생태계 내에서 유기적으로 순환되며 각자의 경쟁력을 강화해줌

핵심 주체	생산	선순환 기여
xAI	• 최첨단 AI 모델 (Grok, Macrohard)	다른 기업들로부터 투자금과 데이터, 사용자 기반을 받아 생태계 전체의 지능을 높이는 핵심 기술(AI)을 창조하고 공급
테슬라	• 방대한 실제 주행 및 로봇 데이터 • AI 기술의 현실 세계 검증	AI를 물리적 세계에 적용하여 가장 가치 있는 실증 데이터를 생산, 이를 다시 xAI에 공급해 AI 모델의 성능 향상에 조력
스페이스X	• 로켓, 위성 등 고난도 엔지니어링 데이터	인류의 가장 복잡한 공학 문제를 AI의 시험대로 제공하여 모델의 한계를 극한까지 밀어붙임
X	• 실시간 글로벌 텍스트/영상 데이터 • 생태계 비전의 홍보 채널	글로벌 실시간 데이터를 AI 학습에 공급하는 동시에, 생태계의 비전과 성과를 전파하여 인재, 투자, 대중의 지지를 확보

자료: 각 회사, 미래에셋증권 리서치센터

III. AI 플레이어: 높아진 배포 능력으로 수익화도 노릴 xAI

Grok은 '생각보다' 미국에서 많이 팔린다

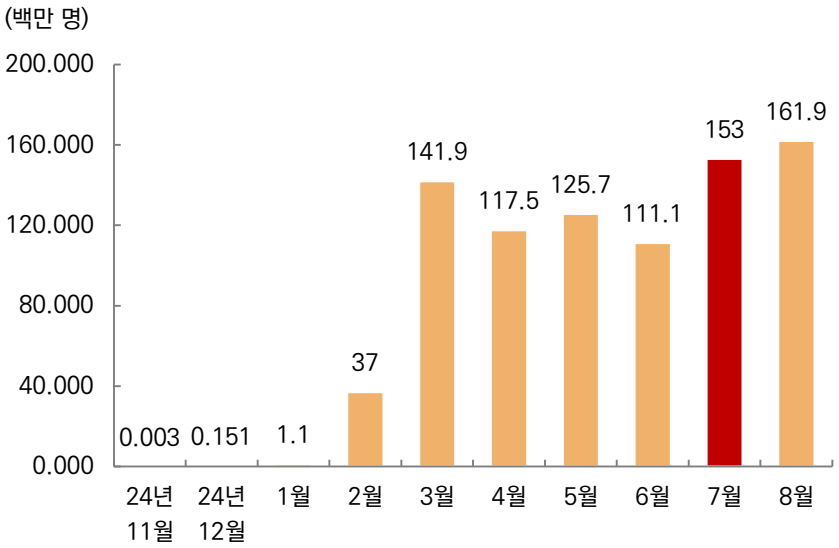
- xAI는 X와 합병하며 빠른 시장 침투와 사용자 기반을 얻게 됨. 합병 기업 가치는 약 2,000억 달러로, 머스크 제국의 프로토콜 안에서 창출되는 효력을 실감 할 수 있음.
- 한편 Grok은 점차 독자적으로도 브랜드 인지도를 쌓는 중. Grok iOS 앱은 미국 앱스토어에서 50만 개의 평점 숫자를 돌파했는데, 이 이정표를 달성한 ㅁ두 번째 AI 앱으로 등극.
- 더욱이 xAI는 "Companions"라는 이름의 실시간 대화형 애니메이션 기능을 Grok에 도입. 서브컬쳐 모바일 게임의 과금 모델과 비슷한 비즈니스인 CharacterAI社의 높은 사용자 참여율(약 7억 명)을 결합하려는 시도. 이는 "돈을 못 번다"는 기존 비판을 극복할 수 있는 잠재적 수익원으로 부상.
- 이에 따라 2027년 말 현금 흐름 흑자전환을 할 수 있을 것이라 사료됨. 올해 매출은 10억 달러를 밑돌 것으로 보이나, 이후에는 3~4배의 성장 시나리오가 가능할 것으로 전망.

xAI 관련 주요 재무 사항 장기 전망
(Levered FCF: 부채 의무 이행 후 현금흐름)

주요 재무 사항	금액	단위
1Q25 실제 매출	0.052	십억 달러
2025년 매출	0.65	십억 달러
2026년 매출	2.5	십억 달러
2027년 매출	10	십억 달러
2027년 EBITDA	4.29	십억 달러
2025년 Levered FCF	-16	십억 달러
2026년 Levered FCF	-10	십억 달러
2027년 Levered FCF	**흑자 전환**	

자료: xAI, 미래에셋증권 리서치센터

Grok 월간 방문자 숫자
지난 7월 Grok 4 공개 및 Companions 출시로 1.5억명 상회



자료: Exploding Topics, 미래에셋증권 리서치센터

III. AI 플레이어: AI계의 “슬리데린”, Anthropic

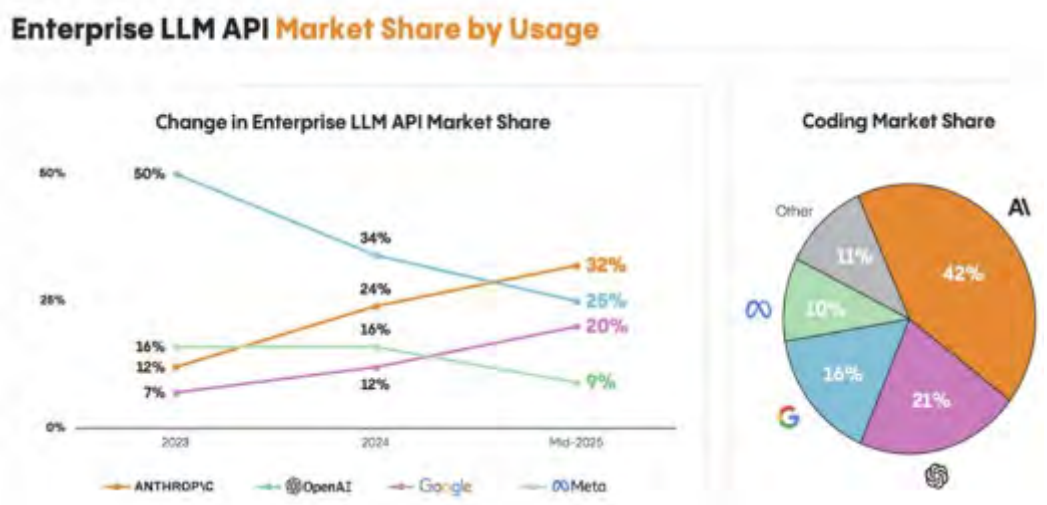
선택과 집중을 통한 AGI 순혈주의

엔터프라이즈 AI 시장의 새로운 지배자

- Anthropic는 B2B 중심의 AGI 지향 회사. 18개월 동안 기업가치 185억 달러에서 1,830억 달러로 약 10배 성장. 동시에 ARR은 올 1월 10억 달러에서 현재 50억 달러 이상으로 확대. API 매출 비중이 85%로 높음. (OpenAI는 27%) 엔터프라이즈 채택(고객 수 30만 곳, 10만 달러 이상 계약 고객 숫자도 8배 증가)이 급증했기 때문. 관련 시장 점유율도 32%로 1위. 특히 코딩 특화 제품인 “Claude Code” 출시 3개월 만에 ARR 5억 달러 달성하며 매출 가속의 핵심 동력.
- '코딩' 분야에 모든 역량을 집중하는 건, OpenAI나 구글과의 전면전을 피하는 효과. 소비자용 음성 비서나 이미지 생성 기능은 의도적으로 배제. 오직 '코드 에이전트'와 '소프트웨어 에이전트'가 비즈니스 자동화의 핵심이라 판단. 코딩 능력의 고도화는 Anthropic 자체의 연구개발 속도를 가속하는 수단. 기술에 가장 민감하고 영향력이 큰 개발자 집단을 사로잡는 모델이 기술 표준이 될 것이라는 확신.
- 이들의 최신 모델인 Claude Opus 4.1은 코딩 부문 SOTA(State-of-the-Art)로 등극. 실제 코드 채택자의 강력한 선호도를 의미하는 “Lmarena” 코딩 부문에서 GPT-5-high를 제치고 1위 차지.

2025년 기업용 LLM API 시장 점유율 변화
Anthropic의 점유율이 12%에서 32%로 급증, OpenAI는 50%에서 25%로 하락하는 추세

구분	점유율(2025)	세부 내용
전체 기준	32%	Menlo Ventures가 올해 여름 150개 기업 기술의 의사결정권자를 설문해 산출
경쟁사	OpenAI 25%, 구글 20%, 메타 9%	OpenAI는 ’23년 50%에서 급락
‘코딩’ 분야	42%	코드 생성·리팩토링 비중 OpenAI 21%의 두 배



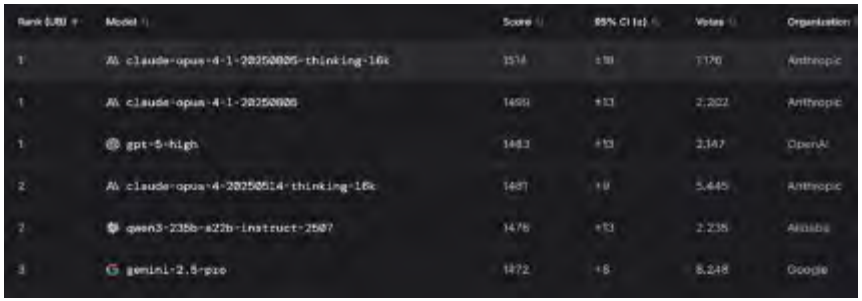
자료: Menlo Ventures, 미래에셋증권 리서치센터

III. AI 플레이어: 리스크가 있어도 그만큼 잘하니까 역대급 성장

Anthropic의 API 매출 비중은 무려 85%

- 다만, Anthropic의 약점은 데이터 확보 및 자체 배포 능력. 경쟁사들이 연간 6-7천만 달러에 Reddit 데이터를 계약할 때, 무단 스크래핑 후 소송에 휘말림. '그 돈 낼 바에야 GPU 빨리 더 사겠다'는 판단.
- Anthropic의 API 매출 비중은 무려 85%. OpenAI의 API 매출 비중(27%)과 비교해 극명한 차이. 특히 API 매출의 약 절반이 Cursor, GitHub 단 두 고객사에서 발생. 빠른 성장의 동력이자, 고객 이탈 시 매출이 반 토막 날 수 있는 '양날의 검'. OpenAI의 최대 파트너인 Microsoft의 자회사 GitHub가 핵심 제품 Copilot에 경쟁사인 Anthropic의 Claude를 사용한다는 건 코딩 능력이 그만큼 좋다는 뜻.

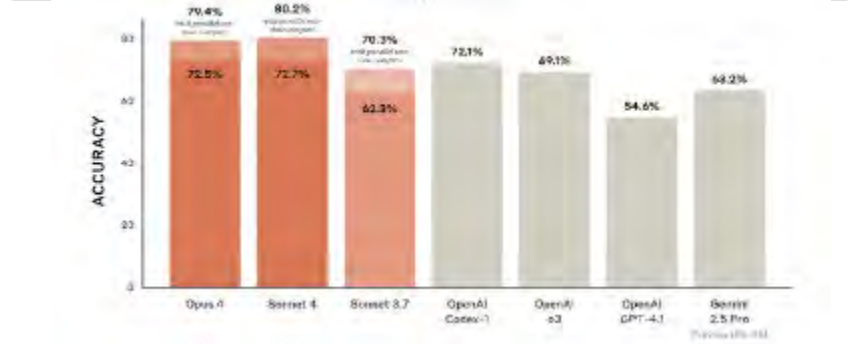
Claude 4.1 Opus가 lmarena의 코딩 부문에서 1위를 차지
심지어 추론(reasoning) 기능 없는 버전도 GPT-5-high보다 상위에 랭크



Rank (UB)	Model	Score	95% CI (±)	Votes	Organization
1	claude-opus-4-1-20250805-thinking-16k	1574	+10	1170	Anthropic
1	claude-opus-4-1-20250805	1499	+13	2,202	Anthropic
1	gpt-5-high	1463	+13	2,147	OpenAI
2	claude-opus-4-20250514-thinking-16k	1451	+9	5,445	Anthropic
2	qwen3-235b-a22b-instruct-2507	1418	+13	2,235	Alibaba
3	gemini-2.5-pro	1372	+6	8,248	Google

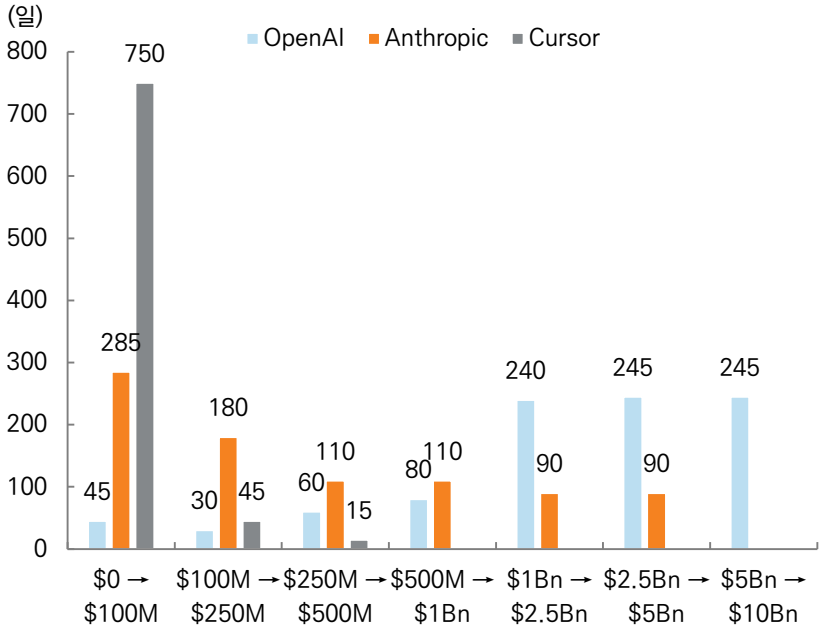
자료: LMarena, 미래에셋증권 리서치센터

주력 모델(Opus 4, Sonnet 4)이 OpenAI의 GPT-4.1 등 경쟁 모델을
소프트웨어 엔지니어링(SWE-Bench 벤치마크) 정확도면에서 압도



자료: Anthropic, 미래에셋증권 리서치센터

각 AI 서비스들의 ARR 마일스톤별 달성 소요 기간 (단위: 일)
과거 OpenAI가 보여준 성장 경로를 훨씬 더 빠른 속도로 답습하고 있음
OpenAI는 10억 달러에서 25억 달러로 성장하는 데 '약 8달'이 소요
Anthropic은 동일 구간을 '한 분기'만에 주파해 2배 이상 빠른 속도



자료: SemiAnalysis, 미래에셋증권 리서치센터

[참고] 세계 4번째로 높은 가치의 스타트업이 된 Anthropic

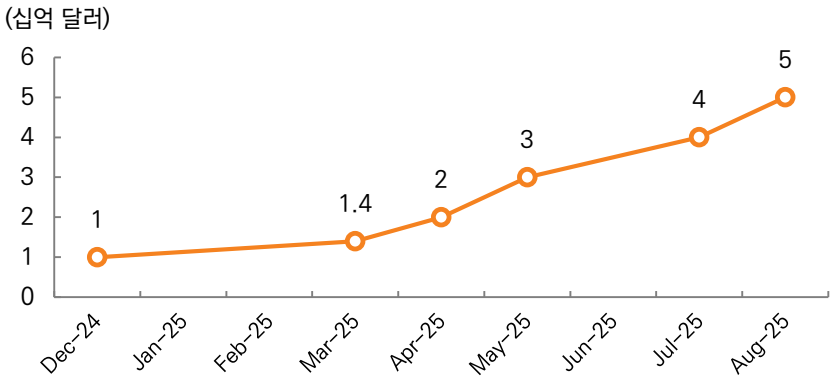
AI가 돈을 못 번다는 우려를 기우로 만들어버리는 실적

회사의 전반적인 성장과 규모를 나타내는 핵심 데이터

항목	수치	비고
기업 가치	1,830억 달러	OpenAI 대비 40% 규모 달성
Series F 투자 유치액	130억 달러	AI 업계 최대 규모
직원 수	약 1,100 명	OpenAI 대비 20% 정도

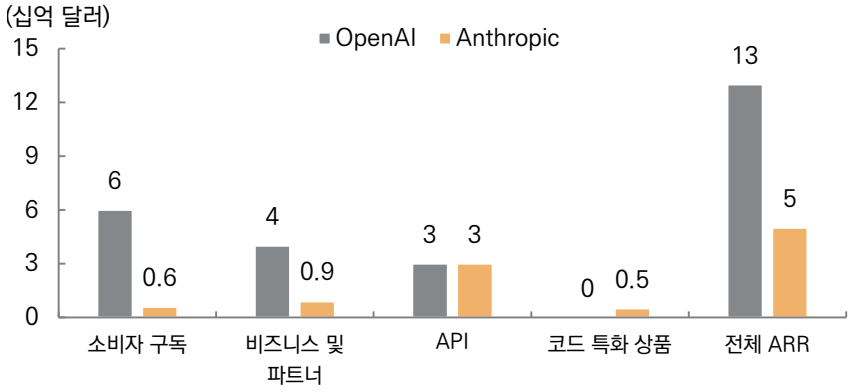
자료: Anthropic, 미래에셋증권 리서치센터

역대 기술 기업 중 최고 수준인 연간 매출(ARR) 성장
작년 말 대비 ARR이 5배 증가… AI 업계에서도 전례 없는 성장속도



자료: Anthropic, 미래에셋증권 리서치센터

OpenAI와 Anthropic 세부 매출 예상
전체는 OpenAI가 2.5배 이상 크나, API에서 선전



자료: Peter Gostev(Moonpig의 AI 부문 대표), 미래에셋증권 리서치센터

코딩 전용 서비스인 “Claude Code”의 성과 및 ARR 추정
2025년 6월 말 기준으로 1.3억 달러인데, 2달 만에 5억 달러로 성장

항목	수치 / 내용
사용 개발자 수	115,000 명
코드 변경량	1억 9,500만 줄
개발자 1인당 연간 추정 수익	$1,130
ARR	1.3억 달러
ARR (8월 발표치 기준)	5억 달러

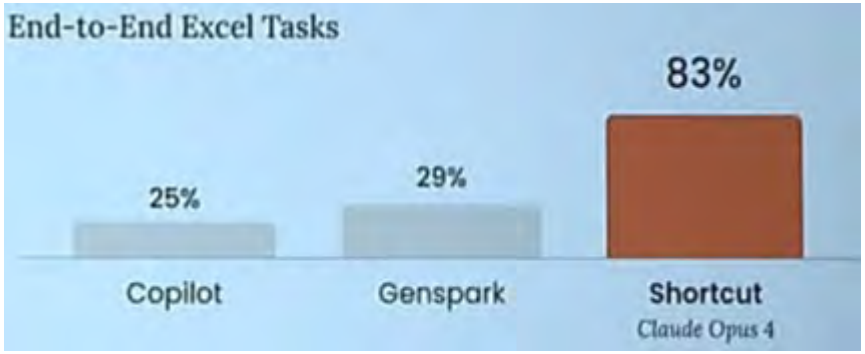
자료: Anthropic, 미래에셋증권 리서치센터

[참고] B2B 시장 장악의 교두보: “Claude Financial Services”

금융 전문가의 워크플로우까지 재설계하는 업계 최초의 '통합 인텔리전스 레이어'

- **금융 산업 특화 설계:** 방대한 데이터, 수작업 비효율성이라는 문제를 해결하기 위해 엔터프라이즈급 보안/컴플라이언스를 기반으로 구축.
- **통합된 데이터 허브:** FactSet, S&P, Morningstar 등 외부 시장 데이터와 Box, Snowflake 등 내부 데이터를 하나의 작업 공간으로 통합.
- **‘가상 협업자’의 등장:** 주니어 분석가 수준의 정밀함으로 복잡한 금융 분석을 수행. 4-5시간 소요되던 분석 업무를 30분 내로 단축.
- **이미 세계 최고 금융기관들은 Cladue를 도입:** 2조 달러 규모의 노르웨이 국부펀드의 CEO는 “없어서는 안 될 존재”라고 극찬. AIG는 보험 인수 심사 기간을 5배 단축하고 데이터 정확도를 75%에서 90% 이상으로 향상. 헤지펀드 브릿지워터는 리서치 RA로 이를 사용하고 있음.

Fundamental Labs 벤치마크 (Excel 수행 능력)

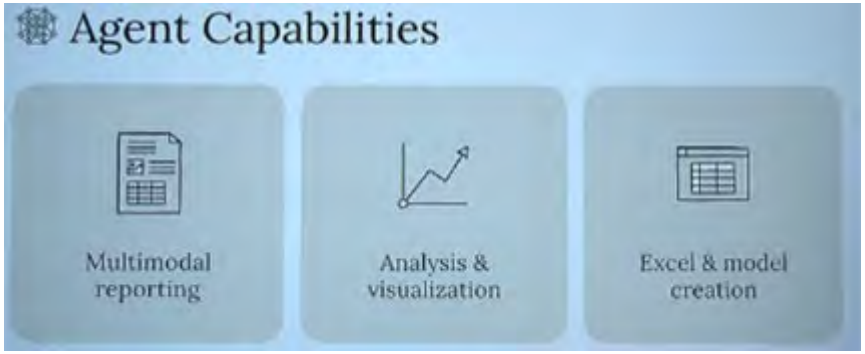


Anthropic과 OpenAI를 가지고 Vals.ai 벤치마크(금융 분석 능력) 비교

항목	Anthropic의 최고 결과	OpenAI의 최고 결과
재무/모델 수정	59.3%	44.4%
추세 분석	38.1%	23.8%
정량적 추출	**75.0%**	**64.1%**
실적 Beat or Miss	48.8%	39.5%
복합 정보 추출	55.6%	50.0%
정성적 추출	62.3%	59.0%
수치 추론	55.8%	63.5%
시장 분석(다중 기업분석 등)	**14.3%**	**23.8%**
평균 결과	51.8%	48.3%

자료: Anthropic, 미래에셋증권 리서치센터

Claude 금융 분석 솔루션의 세 가지 핵심 역량(=엑셀 모델링 가능)



Claude가 자연어 요청만으로 생성한 완전한 DCF 모델의 실제 결과물

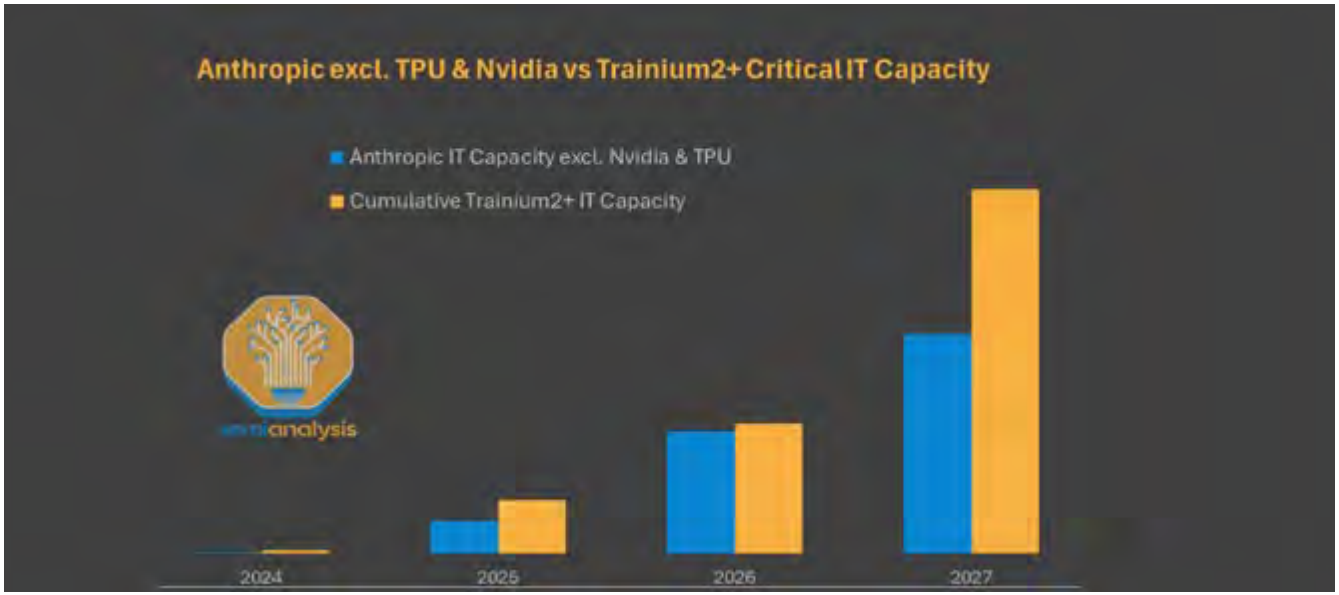
자료: Anthropic, 미래에셋증권 리서치센터

III. AI 플레이어: Anthropic의 부족한 부분은 아마존이 채운다

MS-OpenAI에 대항할, 아마존-Anthropic 동맹

- Anthropic은 차세대 모델 개발 야망을 뒷받침하는 핵심 인프라를 확보했음. 파트너는 바로 Amazon. Amazon의 Anthropic 투자 금액은 총 80억 달러. Amazon 역사상 세 번 째로 큰 규모의 투자. 그럼에도 Amazon은 의결권, 이사회 의석 없이 1/3 미만의 소수 지분만 보유하는 계약 구조임을 유념.
- 오직, Claude 모델 훈련을 위해서 설계된 Amazon의 대규모 AI 슈퍼컴퓨팅 클러스터인 프로젝트도 나옴. 이는 Anthropic의 장기적인 모델 개발 로드맵에 대한 강력한 신뢰를 보여줌. 이 클러스터에 들어간 아마존의 자체 AI 가속기인 "Trainium"의 설계맵에 Anthropic이 깊이 관여한 것으로 관측.
- 사실상 아마존의 칩 설계팀을 본인들의 '맞춤형 실리콘 파트너'처럼 활용. 이는 구글의 TPU와 더불어, "AI 모델+하드웨어"의 공동설계가 가능한 유일한 동맹으로 부상. 즉, Anthropic이 Trainium 독점 공급을 받는 것은 자신들의 기술 로드맵에 가장 비용 효율적이고 최적화된 무기를 갖게 된 것.

Anthropic의 수요(파란색)와 AWS의 칩 공급(주황색) 간의 장기적 관계
내년에는 공급과 수급이 일치 (=Anthropic이 칩의 사실상 유일한 고객)



자료: SemiAnalysis, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 메타, 패자(覇者)를 꿈꾸는 패자(敗者)

초지능을 향한 광전사(Berserker) 모드

메타는 지금 구조조정 중

- 과거 Meta의 AI 조직은 방향성과 리더십 부재로 표류. AI 연구 아이디어의 옥석을 가릴 취향을 가진 리더의 부재(제품 관리자 주도의 의사결정)로 연구 역량 낭비. ChatGPT, 구글 검색, 애플 기기로 사용자들이 이탈할 경우 본인들의 소셜 제국이 붕괴할 수 있다는 존재의 위기가 주커버그를 각성시킴.
- 메타는 "개인 초지능(Personal Superintelligence)"이라는 새로운 깃발을 세우면서 "메타 초지능 연구소(MSL)"를 신설. AI를 노동 자동화 도구가 아닌, 개인의 잠재력 확장을 위한 '개인 비서'로 정의.
- 주커버그는 직접 AI 논문을 파고들어 핵심 인재를 식별하고 영입을 주도. 수천억원에 해당하는 사인 보너스를 AI 연구원들에게 제시하면서 공격적 인재 영입을 시도.
- 물론, MSL의 탄생 속 새로운 문화의 이면에는 내부 혼란과 인재 이탈이 존재. ScaleAI, OpenAI 등에서 온 인력들이 Meta의 관료주의적 문화에 적응하는 데 어려움을 겪음. ScaleAI 출신 핵심 임원 루벤 메이어는 입사 두 달 만에 퇴사했고, OpenAI 출신들도 다시 OpenAI로 복귀하는 사례가 발생.
- 과거 Llama를 통해 오픈소스 챔피언 역할을 했으나, 중국발 경쟁에 밀리면서 전략적 가치가 높은 모델은 폐쇄소스로 전환할 가능성 시사. 다만, 자체 개발의 Llama 5 모델이 완성될 때까지 구글의 Gemini나 OpenAI 모델을 챗봇에 통합하는 방안을 임시 조치로 검토 중. 즉, 현재는 재편의 과정.

메타 AI의 수장이었던 얀 르쿤의 LLM에 대한 비판
어찌보면 메타의 LLM 실패는 예견되어 있었음

시기	핵심 견해 및 발언 요약
2022년 12월	특별히 혁신적이지 않음, 실용적 용도로 부적합
2023년 2월	글쓰기 도구로만 유용함, 계획 및 추론 불가
2023년 5월	인간 수준의 추론과 계획 불가능
2024년 3월	4가지 부족 때문에 AGI 도달 불가능 (물리세계 이해/장기 기억/진정한 추론/계획력)
2025년 1월	인간 수준 지능 도달 절대 불가능하다고 재확인

자료: 미래에셋증권 리서치센터

주커버그가 짠 새로운 판, 메타의 "초지능 연구소(MLI)"의 리더십 구성 및 역할 분담

역할 구분	핵심 인물	주요 배경 및 역량	메타에서의 임무 (전략적 목표)
설계자	마크 주커버그	중앙집중식 의사결정 막대한 자원(데이터, 컴퓨팅 파워) 통제	메타의 모든 자원을 '초지능' 개발에 집중시킴
실행가	알렉산더 왕	ScaleAI 창업자 AI의 '산업화' 전문가	초지능을 프로토타입에서 산업적 규모로 구현
상품화	냇 프리드먼	前 GitHub CEO 개발자 생태계 및 플랫폼 구축 전문가	완성된 AI를 인스타그램, 왓츠앱 등 소셜 서비스에 통합 외부 개발자 생태계를 구축하여 기술을 확산
확장가	다니엘 그로스	SSI 공동 창업자 AI 산업의 핵심 '네트워크' 허브	AI 업계의 핵심 인물들 추가 영입 투자를 통해 메타의 AI 생태계 영향력 '확장'

자료: 메타, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 경쟁사들에 비해서는 열위나, 의지는 대단

컴퓨팅: 천문학적 투자와 하드웨어 독립

데이터: 독점적 접근과 품질의 딜레마

- 향후 수년간 AI 인프라에 "수천억 달러" 투자한다고 공표. 2025년 CAPEX 가이던스 660억~720억 달러로 상향, 2024년 2분기 대비 101% 증가한 것. 투자의 90%는 AI 인프라에 집중.
- 자체 AI 반도체 “MTIA” 개발을 통한 하드웨어 독립 추진. MTIA-v2는 2026년까지 80만~90만 개 생산 목표. 다른 빅테크와 비슷하게 메타 역시 칩 뿐만 아니라 랙 시스템도 독자 행보. Broadcom과 협력한 신형 AI 서버랙 “Santa Barbara”를 6,000대 배치 계획. 이 랙은 4분기에 시험 생산 시작 예정.
- 랙당 평균 소비전력(TDP) 180kW 이상. 해당 TDP는 애초에 공랭이 불가해 새로운 데이터센터를 지어야 한다는 얘기. 180kW 랙 6,000개는 이론상 1.1GW급 전력을 요구.
- 메타는 데이터 확보를 위해 AI 모델 훈련 데이터 전문 기업 Scale AI의 지분 49%를 143억 달러에 인수. 10만 명 이상의 데이터 라벨러와 핵심 데이터 기술에 대한 독점적, 우선적 접근 권한 확보.
- 그러나 정작 메타 내부에서는 데이터 품질에 관한 문제 제기와 분열의 조짐. Scale AI의 경쟁사인 Surge나 Mercor의 데이터를 사용하여 차세대 모델을 훈련 중인 것으로 파악.

메타의 MTIA는 방대한 추천 모델 '추론(inference)' 작업에만 최적화… “AI 훈련하려면 GPU를 결국 또 사야 된다”는 것을 의미
HBM 대신에 대용량 SRAM과 LPDDR을 조합하여 비용과 전력 소비 낮췄으나 메모리 대역폭이 극히 낮아 강화학습의 대규모 처리에는 부적합하기 때문

구분	세부 항목	MTIA v2	MTIA v1
공정		TSMC 5nm	TSMC 7nm
면적		25.6mm x 16.4mm	19.3mm x 19.1mm
TDP		85W	35W
GEMM TOPS		354 TFLOPS/s (INT8)	102.4 TFLOPS/s (INT8)
		177 TFLOPS/s (FP16/BF16)	51.2 TFLOPS/s (FP16)
GEMM TOPS (희소성)		708 TFLOPS/s (INT8)	지원 안함
		354 TFLOPS/s (FP16/BF16)	지원 안함
SIMD TOPS	RISC-V 벡터 코어	5.5 (INT8), 2.8 (FP16), 1.4 (BF16/FP32)	3.2 (INT8), 1.6 (FP16), 0.8 (FP32)
	SIMD 엔진	5.5 (INT8/FP16/BF16/FP32)	3.2 (INT8), 1.6 (FP16)
메모리 용량	PE 당 로컬 메모리	384 KB	128 KB
	PE 간 공유 온칩 SRAM	256 MB	128 MB
	오프칩 LPDDR5	**64–128 GB**	**32–64 GB**
메모리 대역폭	PE 당 로컬 메모리	1 TB/s	0.4 TB/s
	온칩 SRAM	2.7 TB/s	0.8 TB/s
	오프칩 LPDDR5	**204.8 GB/s**	**176 GB/s**

자료: ACM Digital Library, 미래에셋증권 리서치센터

III. AI 플레이어 분석: 메타의 진짜 승부수는 'AI 글래스'

수십억 사용자 기반의 즉각적인 실험 및 수익화

최종 인터페이스 'AI 글래스'를 향한 진격

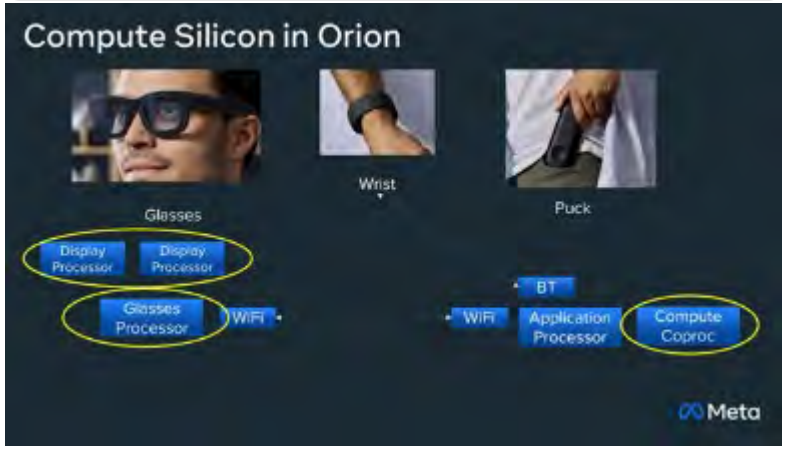
- 메타가 가진 최고의 자원은 그들이 갖고 있는 소셜 플랫폼. 그 자체로 막강한 배포 채널. 전세계 인구 절반 숫자의 사용자 풀을 대상으로 AI 기술을 즉시 배포하고 피드백을 받을 수 있는 실험장이 됨. AI 접목을 통해서 본인들의 핵심 비즈니스(광고) 강화는 이미 증명. AI 기반 추천 모델 "Andromeda"가 인스타그램에서 5%, 페이스북에서 3%의 광고 전환율 향상 달성.
- 주커버그 CEO는 스마트 글래스를 '포스트 스마트폰'의 핵심 인터페이스로 확신. 그는 "미래에는 AI 안경이 없으면 인지적 열세를 겪게 될 것"이라 말할 정도. 사용자가 보고 듣는 모든 것을 AI가 실시간으로 이해하고 상호작용하는 상시 접근(Always-on) 기기를 목표.
- 현재 "Ray-Ban Meta 스마트 글라스" 누적 판매량은 200만 대 돌파. 해당 스마트 글라스 사용자 중 하루 한 번이라도 AI 기능을 써본 사용자 비중은 23년 10월 출시 직후 12% → 올해 6월 38%로 급증.
- 3분기에는 디스플레이가 탑재된 "Hypernova" 모델 공개 예정. EMG 손목 밴드를 통한 제스처 컨트롤 도입. 그리고 2026~27년, 완전한 AR 기능을 갖춘 "Orion" 글래스 상용화 목표. 모바일 혁명에서는 애플의 아성을 못 넘지만 AI 시대에서는 본인들이 하드웨어 시장을 선점하겠다는 야심.

AI 글래스 시장의 전략적 포지셔닝을 보면, 애플과 달리 메타는 '일상성'을 강조
메타의 모든 기기는 50~70g으로, 일반 안경과 큰 차이가 없는 무게를 유지

기기명	AI 연산 처리	카메라 성능	배터리 수명	무게	디스플레이 기술	가격 범위
메타 Ray-Ban	스냅드래곤 AR1	12MP 초광각 1080p 동영상	4시간	50g	오디오 중심, LED 표시등	$299-$379
메타 Oakley	스냅드래곤 AR1	12MP 초광각 3K 동영상	8시간	~55g	오디오 중심, 퍼포먼스	$399-$499
메타 Hypernova (추정)	스냅드래곤 AR1	12MP+ 초광각 3K 동영상	8시간	70g	우측 렌즈 하단에 마이크로 디스플레이 탑재 → 알림·미니 앱·지도 등 표시	$799-$999
애플 Vision Pro	M2 + R1	12개 이상 카메라 LiDAR	2시간	600g	듀얼 4K 마이크로 OLED (2300만 픽셀)	$3,499+
구글 Android XR	Tensor 칩?	내장 카메라	수시간 목표	미정	투명 마이크로 디스플레이	미정 (2026)

자료: BrandXR, 미래에셋증권 리서치센터

최종 목표인 완전 AR 글래스 'Orion'의 분산 컴퓨팅 아키텍처
AI 연산 같은 무거운 핵심 작업은 주머니 속 'Puck' 장치로 처리



자료: 메타, 미래에셋증권 리서치센터

IV. AI 군비경쟁

천재고용 CAPEX 클러스터

IV. AI 군비경쟁: 스포츠 같아진 AI 엔지니어들의 이적 소식들

경쟁사 영업 비밀을 얻는 효율적 방법

- 특허가 약해진 시대에, 개별 연구원에 高연봉을 지불하는 것은 경쟁사 영업 비밀을 얻는 효율적 방법.
- 실리콘밸리의 대규모 이적 사태를 촉발한 곳은 메타. "MSL"에 심취한 주커버그의 '폭발형 오퍼(4년간 3억 달러, 첫해 1억 달러 이상 지급, 6시간 안에 이직 결정)'는 수십 명의 AI 엔지니어를 불러모으는데 성공.
- AI 인재 확보 경쟁이 과열되면서 개발자들의 몸값은 천정부지로 치솟고 있음. 각 회사들은 직원 이탈을 막기 위해 연봉을 올려주는 등 보상 체계를 재검토하는 중.
- 모두가 AGI로 일자리를 잃기 전에 가능한 한 많은 돈을 벌기 위해 전환하는 시즌이라는 분석까지 제기.
  - OpenAI가 30억 달러에 인수를 발표한 Windsurf의 창업자들을 구글이 24억 달러에 영입.
  - 메타는 국제수학올림피아드(IMO) 금메달을 딴 구글의 모델 개발에 참여한 딥마인드 연구원 3명 영입.
  - 마이크로소프트는 구글 딥마인드에서 24명의 시니어급 연구원을 고용.
  - xAI는 엔비디아에서 NeMO의 주요 공헌자인 최고 AI 연구원을 데려 옴.
  - OpenAI는 Tesla, xAI, X에서 4명의 고급 엔지니어를 대거 영입.
  - xAI는 한 연구원이 OpenAI에 이직하면서 본인들의 기술을 유출한다는 내용으로 법원에 '이직금지' 긴급 명령까지 신청.
  - 애플의 '파운데이션 모델팀' 헤드였던 Ruoming Pang이 메타로 이직. 팀원들도 OpenAI/Anthropic 등으로 이직 준비.

본인의 앞날을 몰랐던 알트만이 주커버그와 한 2016년 인터뷰
9년 전 주커버그를 칭찬하는 샘 알트만 "페이스북의 특기는 인재 고용"

자료: Y Combinator, 미래에셋증권 리서치센터

2025-26 시즌의 유럽 축구 스타들 주급 기준 TOP 7
애플에서 메타로 이직한 연구원의 연봉을 주급으로 환산…

순위	선수	클럽	주급(EUR)
1	크리스티아누 호날두	알나스르	4,007,692
2	리야드 마레즈	알아흘리	1,003,846
3	카림 벤제마	알이티하드	961,538
추정	**Ruoming Pang**	**애플 → 메타**	**825,000**
4	사디오 마네	알나스르	769,231
5	칼리두 쿨리발리	알힐랄	667,308
6	얼링 홀란드	맨체스터 시티	605,000
7	킬리안 음바페	레알 마드리드	600,962

자료: Capology, 미래에셋증권 리서치센터

[참고] 지난 여름, 뜨거웠던 AI 이적 시장

메타의 주커버그 CEO가 "직접 전화 걸고 식사 대접"하며 데려 온 AI 엔지니어/연구원 리스트(일부)

이름	출신	주특기	주요 경력·성과	구글 스콜라 기준 논문 인용 티어
Alexander Kolesnikov	OpenAI	멀티모달	ViT 공동 저자로 여러 AI 프로젝트에서 핵심 기술 책임을 맡아온 인물	SSS
Xiaohua Zhai	OpenAI	멀티모달	OpenAI에서 ViT를 더 크게·빠르게 돌리는 방법을 연구한 시각 AI 확장 전문가	SSS
Lucas Beyer	OpenAI	멀티모달	Vision Transformer(ViT)을 공동으로 고안했고, 취리히 연구센터를 운영한 이미지 AI 전문가	SSS
Jason Wei	OpenAI	사전학습 추론 (reasoning)	생각의 사슬 프롬프팅' 개념을 대중화, 스케일링 법칙에 따른 AI의 창발적 능력에 관한 개념을 제시	SSS
정형원	OpenAI	추론(reasoning) 사후학습, 코딩	GPT-4 이후의 차세대 모델 개발에 핵심 역할, OpenAI의 Codex-mini 리드 참여	SS
Yuanzhi Li	OpenAI	강화학습	마이크로소프트에서 소형 언어 모델(예: phi-1.5, Phi-3, Phi-4) 개발에 핵심적으로 기여	SS
Jiahui Yu	OpenAI	멀티모달	GPT-4 계열(4.1, 4o)과 o3·o4-mini를 공동 개발, Gemini 멀티모달 팀 공동 리더	S
Jack Rae	구글	사전학습 추론 (reasoning)	Gemini AI의 핵심 개발 책임자이며, DeepMind에서 LLM인 'Gopher'·'Chinchilla'를 이끈 경력	S
Hongyu Ren	OpenAI	추론(reasoning) 사후학습	GPT-4o 및 mini 모델들을 공동 개발했고, OpenAI에서 학습 후반 고도화 팀을 이끈 인물	S
Ji Lin	OpenAI	추론(reasoning) 멀티모달	GPT-4.5와 Operator 추론 엔진 등 핵심 백엔드를 구축한 개발자	S
Johan Schalkwyk	구글	멀티모달	구글 최고 연구원 출신으로 음성 AI 'Sesame' 초기 멤버이자 'Maya' 프로젝트 총책임을 맡았음	S
Shengjia Zhao	OpenAI	추론(reasoning) 사후학습	메타에서 Chief Scientist로 임명. GPT-4를 함께 만든 개발자로, OpenAI 내 합성데이터팀 총괄	S
Huiwen Chang	구글	멀티모달	GPT-4o 이미지 생성 기능을 공동 개발했고, 구글에서 MaskGIT·Muse 모델을 발명한 이미지 생성 연구자	A
Trapit Bansal	OpenAI	추론(reasoning)	일리야 수츠케버와 함께 강화학습 패러다임 시작하고 "o 시리즈"를 개척	A
Pei Sun	구글	추론(reasoning) 사후학습, 코딩	Waymo에서 최신 자율주행 인식 모델을 만든 뒤 Gemini에서 AI 추론 시스템을 책임진 개발자	A
Anton Bakhtin	앤트로픽	추론(reasoning)	전략 게임 '디플로머시' AI 연구 등으로 5,000회 이상의 학술 인용을 기록한 언어모델 전문가	B
Shuchao Bi	OpenAI	추론(reasoning) 사후학습, 멀티모달	GPT-4o 음성 기능을 설계한 핵심 엔지니어로, OpenAI 멀티모달(음성·영상) 학습을 총괄	B
Joel Pobar	앤트로픽	추론(Inference)	메타에서 11년간 AI 성능 최적화 플랫폼을 개발, Anthropic에서 추론 엔지니어로 활동한 베테랑	B

자료: 티어 기준은 SSS: ≥ 80k, SS: 40k–79.9k, S: 15k–39.9k, A: 3k–14.9k, B: 〈 3k 또는 연구자 프로필 부재, 구글 스콜라, 외신, 미래에셋증권 리서치센터

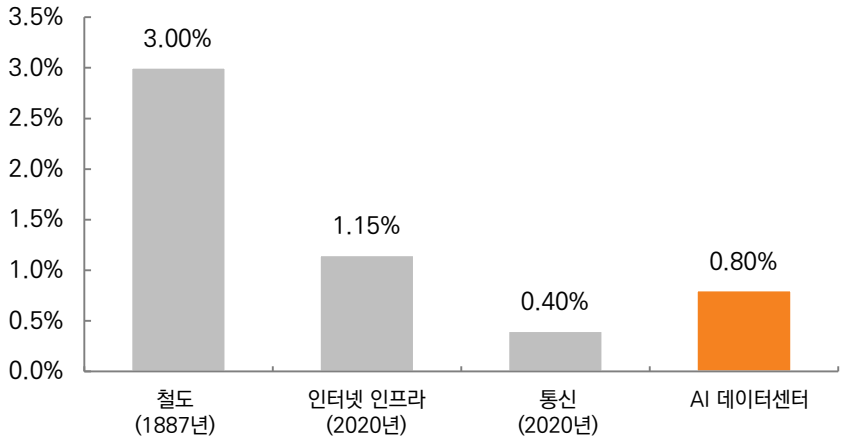
IV. AI 군비경쟁: 멈추지 않는 CAPEX (1)

초지능을 향한 군비 경쟁

너무 빠르게 변화하는 세상에서 가장 큰 위험은 위험을 감수하지 않는 것

- AI 경쟁은 절대 평가가 아닌 상대 평가. '최고의 지능'을 차지하기 위한 투자는 멈추지 않을 것이며, 이는 인프라 투자의 폭발적 증가로 이어질 것. '高지능'을 얻게 될 때의 경제적 가치가 너무 커, AI 알고리즘의 효율화로 발생하는 비용 절감분 역시 즉시 더 높은 지능 개발에 재투자될 것. 이에 따라 하이퍼스케일러들은 당분간 계속 AI 서비스 수익보다 더 많은 돈을 GPU 구매에 지출할 전망. 작은 스타트업·연구팀은 사실상 경쟁이 불가능한 규모.
- 질의 1건당 원가는 1센트 이하 수준으로 매우 낮음. 생각보다 마진이 더 높은 사업임에도, 당분간 큰 흑자로 변환되지 못하는 것은 막대한 GPU 구매와 인건비 지출 때문.
  - 지난 6월 초 샘 알트만이 본인 블로그에다 "GPT-4o에서 한 번의 질의에 소모되는 전력 사용량은 평균 0.34 Wh 수준"이라고 밝힌 것을 토대로 – 퍼블릭 클라우드 GPU 시간당 임대 비용(약 2달러), 일간 처리량(10억 개의 프롬프트), 산업용 평균 전기세(kWh당 0.1달러)라고 가정하고, – 하루 100건의 질의를 30일 내내 지속한다고 했을 때, ChatGPT의 가장 낮은 tier의 구독료인 월 20달러로 질의응답의 서빙 비용을 덮을 수 있다는 것
- 이미 2024년 9월에 오라클 창업자/CTO인 래리 엘리슨은 "빅테크 AI 경쟁 참가 티켓 값은 1,000억 달러"라고 말한 바 있음. 지금은 그 수 배에 달할 것. 이와 관련해, 폭스콘은 "일부 AI 프로젝트 투자액은 1,000억 달러가 아니라 이제 3,000억~5,000억 달러로 상향되고 있다"고 말한 바 있음.

시대별로 역사적 전환을 이끌었던 기술들의, 美 GDP 대비 CAPEX 비중
불과 2022년만 해도 0.1% 미만이었던 AI DC의 수치는 1%를 곧 넘을 수도…



자료: 미래에셋증권 리서치센터

OpenAI의 GPT-4o의 질의 하나당 비용 추산을 위한 여러 가정
OpenAI는 추론(Inference) 서비스에서의 총마진율은 75% 이상

항목	값
GPU 임대 단가 ($/시간)	2
질의당 GPU 사용 시간 (초)	10
하루 질의량 (개)	10억
총 GPU 사용 시간 (초)	10,000,000,000
총 GPU 사용 시간 (시간)	2,777,778
GPU 임대 비용 (USD)	5,555,556
전력 사용량 (kWh)	340,000
전력 단가 ($/kWh)	0.10
전기 비용 (USD)	34,000
일간 비용 (USD)	5,589,556
프롬프트당 비용 (센트)	**0.56**

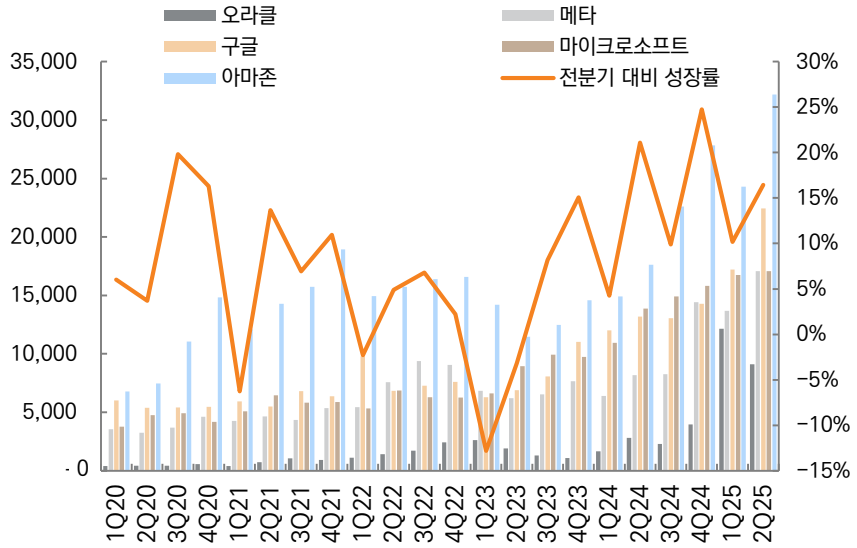
자료: 샘 알트만, 미래에셋증권 리서치센터

IV. AI 군비경쟁: 멈추지 않는 CAPEX (2)

CAPEX는 고지능 경쟁의 후행지표일 뿐

- 2분기 실적 이후 주요 테크 기업들의 AI 투자 규모가 年 3,640억 달러로 급증. 지난 2월 예상치 3,250억 달러 대비 12% 상향. '26년에 아마존, MS, 메타, 구글은 4,650억 달러를 지출할 것으로 예상.
  - 아마존: 1,185억 달러 (전년 대비 40% 증가) / 구글: 850억 달러 (기존 750억에서 상향 조정)
  - 마이크로소프트: 800억 달러 (분기별 300억 달러 수준 유지) / 메타: 660~720억 달러 (하한선 상향 조정)
- 특히 CAPEX에 열중하는 곳은 OpenAI의 샘 알트만과 xAI/테슬라의 일론 머스크. AMD의 CEO 리사 수는 지난 7월 "OpenAI와 일론 머스크의 회사들로부터 수요가 많다"고 언급.
- OpenAI는 2024년 이후 컴퓨팅 성능 15배 증가. GPT-5에 20만 개 GPU를 사용했다고 최근 알려짐.
- xAI는 20만개의 H100과 3만개의 GB200의 컴퓨팅 자원(약 200만개의 A100과 동일)을 가동 중이고, 3분기 내 55만개의 GB200/GB300(약 500만개의 A100과 동일한 것으로 추정)이 추가 투입될 예정. 알트만은 연말까지 400만 개의 A100 성능에 해당하는 컴퓨팅을 확보할 것이라 회사 내부 메모로 밝혔고, GPT-5 출시 이후 "5개월 내 컴퓨트 플릿 2배 증설" 계획도 발표.

빅테크들의 CAPEX 분기별 추이



자료: Bloomberg, 미래에셋증권 리서치센터

연도별 선도적인 AI 슈퍼컴퓨터와 주요 컴퓨팅 규모 지표

연도	세계에서 선도적인 AI 슈퍼컴퓨터	몇 대의 H100을 갖춘 성능인가?	AI 가속기 탑재량	전력 규모	하드웨어 비용
2019	OAK RIDGE SUMMIT	3,492	28K	13 MW	$200M
2020	OAK RIDGE SUMMIT	3,492	28K	13 MW	$200M
2021	SUNWAY OCEANLIGHT	6,008	108K	-	-
2022	OAK RIDGE FRONTIER	14,566	38K	40 MW	$600M
2023	OAK RIDGE FRONTIER	14,566	38K	40 MW	$600M
2024	메타의 "GenAI 2024"	24,576	25K	40 MW	$900M
2025	**xAI의 "Colossus"**	**200K**	**200K**	**300 MW**	**$7B**
2026	**아직 모름 (아마도 xAI Colossus)**	**1.5M**	**? (주로 GB 시리즈)**	**2GW 이상**	**$50B 이상**

자료: 미래에셋증권 리서치센터

IV. AI 군비경쟁: Stargate by OpenAI

10 GW·5,000억 달러 청사진과 현실의 간극

전력·냉각·그리드 안정화가 병목

- Stargate(스타게이트)는 OpenAI·소프트뱅크·오라클이 4년간 10 GW 규모 AI 인프라를 구축하고 미화 5,000억 달러를 투입하겠다면서 지난 1월 트럼프 대통령이 공개한 프로젝트.
- 반년이 흐른 현재, 올해 말까지 오하이오에 '소형' 센터 1곳을 짓는 수준으로 목표를 크게 낮췄다. 핵심 원인은 소프트뱅크와 OpenAI 간 부지·지분 구조를 둘러싼 이견 때문. 소프트뱅크의 190억 달러 동일 지분 매칭 투자계획은 관세·금리 불확실성 탓에 자금조달을 지연되며 어려움을 겪고 있음.
- OpenAI는 플랜 B를 내놨음. 소프트뱅크 자금이 투입되지 않고 오라클과 단독으로 4.5 GW 추가 계약하면서, 기존 물량 포함 총 5 GW+는 확보(10GW라는 목표에서 절반 이상을 확보). 5GW급 GB200 클러스터는 H100 약 500만 개에 해당하는 성능, GPT-4의 훈련량 대비 약 1,000배 규모로 추정.
- 오라클이 주도한 텍사스 Abilene(첫 번째 캠퍼스) 부지는 1단계 공사가 완료돼 일부 랙이 전력 투입 중. 이 곳만으로도 약 1.2GW 전력을 수용할 수 있도록 설계됨.
- 자금 조달의 마일스톤 조건이 전력 인프라 확보라는 점에서 전력이 병목이라는 것을 유념해야 함.

Stargate 투자자 현황

투자자	역할	투자액 (억 달러)	투자 형태	현재 상황	비고
OpenAI	프로젝트 스폰서 · 지분 투자자	190	자체 현금/지분	초기 190 억 중 40 억 유동성 사용, 추가 증자 여지	운영·기술 의사결정 주도
SoftBank	전략 투자자 (지분)	–	지분(임시중단)	자금 미투입, 협상 중단 상태	자금조달 차질, 1차 트랜치 매칭 계획 보류
Oracle	AI 클라우드 공급·HW 투자자	400	CAPEX+15년 리스	4.5 GW DC 계약, GB200 200만 개 구매 완료	OpenAI에 연 300 억 임대료 청구
Crusoe · Blue Owl JV	부지·건설 개발사	116	채무+지분	Abilene 1단계 8개동 공사중	1.2 GW 캠퍼스 올해 완공목표
MGX / PIF (사우디)	잠재 전략 LP	–	지분(탐색 중)	3,000억 달러 투자는 아직 구체성없는 의향서 수준	중동 국부펀드, 에너지·AI 동맹 논의
은행 PF (JPM 등)	프로젝트 파이낸스 대주단	71	대출/채권	1차 공사 대출 71 억 실행, 추가 최대 1,000 억 목표	PF 조달 없이는 Stargate 확장은 어려움

자료: OpenAI, 외신 종합, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Stargate by OpenAI

가스터빈이 AI의 기본
가스터빈은 5분 내
기동·부하 추종이 가능

- OpenAI의 Stargate는 일론 머스크의 Colossus 클러스터와 동일한 전략인 '자체 전력 생산'을 채택. 이에 따라 Stargate 공사 작업을 맡은 Crusoe가 현재 가스 터빈 29기를 설치 중.
- GE Vernova(제너럴 일렉트릭 베르노바)의 LM2500XPRESS(고효율 가스 터빈) 한 대는 약 34~35MW를 발전. 설치는 대개 2주 내에 공급된다고 알려져 있음. 그리드 없이도 작동 가능한 장점.
  - 대형 가스터빈은 글로벌 3사(지멘스, 미쓰비시, GE Vernova 점유율 순)가 과점하고 있음
- 29기 전체가 가동되면 총 약 1,000 MW(1GW)를 생산해 첫 번째 Stargate 사이트의 GB200 NVL72의 슈퍼 칩 50만 개를 구동하기에 충분한 전력을 공급하게 됨.
  - 전체 발전출력 = 터빈 개수 ×터빈 출력 ≈ 29×35MW = 1,015MW(≈1GW)
  - 전제 사용출력 = NVL72 랙 개수(50만 GPU/72) ×랙 출력 ≈ 6,945개 랙×120kW ≈ 833MW
  - 29기 가스 터빈이 생성하는 약 1 GW 전력은 50 만 개의 GB200 NVL72 GPU를 구동하기에 여유 있는 수준

GE의 34MW 모듈형 가스터빈 설치 현장
이 장치는 신속 배치형 전력설비

자료: OpenAI, SemiAnalysis, 미래에셋증권 리서치센터

Stargate 관련 현재 확인되는 공사·전력·GPU 납품 상황

단계	진행 상황	전력·인프라 현황	GPU 배치 계획
Phase 0 (초기 가동)	에너지 작업 완료	ERCOT 계통 전력 200 MW 이미 투입, 2개 데이터홀 가동	16 k GB200 랙 (7월 납품 완료)
Phase 1 (가스터빈 증설)	내년 상반기 완료 계획	LM2500XPRESS 29기 ≈ 1 GW 현장 발전 → 차례로 시운전	6.4만개 GB200 랙
Phase 2 (캠퍼스 확장)	2028년까지 완성	총 1.2 GW(가스), 600MW 태양광 + ESS (Redwood 중고 배터리)	최대 40만 개 GB200
Phase 3 (멀티사이트, 5 GW+)	2030년까지 완성	Oracle 합작 4.5 GW → 미국 내 5 GW 이상 파이프라인 확보	2백만 개의 칩 (GB200 등 차세대 칩)

자료: OpenAI, 미래에셋증권 리서치센터

[참고] Stargate

10 GW·5,000억 달러의 거대 비전

Stargate 총 비용 구조(TCO)

단계/구분	비용 항목	금액(USD)	세부 내용·가정
1단계 (Stargate I · Abilene 1.2 GW)	부지·건설 CAPEX	116	원래 목표 2.2 GW → 8개 캠퍼스(1.2 GW)로 완공 목표
	AI 하드웨어 (GB200 등)	400	Oracle이 엔비디아 GB200 40만 개 구매 후 15년 임대
	전력·냉각 인프라	30	1.2 GW 송전선/가스터빈/UPS 등 29기(GE Vernova) 등
	연간 OPEX (OpenAI→Oracle 임차료)	300	4.5 GW 신규 포함 총 5 GW 용량에 대한 계약 임대료
확정 총량 (2027E)	총 CAPEX (3rd-party)	550	부지·건설 116억 달러 + GPU 400억 달러 + 기타 인프라
장기 청사진 (≈2030)	TCO 목표	5000	10 GW(20개 캠퍼스) → 5 GW 수준으로 1차 축소 검토중

자료: 외신 종합, 미래에셋증권 리서치센터

전형적인 "AI DC"인 스타게이트 설계의 특이점: GPU 비중이 압도적

항목	전통적 데이터센터	스타게이트 설계
건설비/MW	1,500만 달러	3,500만 달러
GPU 비용 비중	약 30%	약 60%

자료: 미래에셋증권 리서치센터

현재 공사 중인 OpenAI의 스타게이트 부지 사진

자료: OpenAI, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Colossus by xAI

"It's not even close."
– Elon Musk

"unmatched in speed"
– Jensen Huang

- xAI의 일론 머스크는 5년 내 H100 동급 연산 능력 5,000만 개의 GPU 확보를 목표로 하고 있음. 예상 전력 소모량은 약 35 GW로, 현재 인류 전체 전력 소비의 2%에 달하는 막대한 규모.
- 전략의 핵심은 '풀-스택 소유(Ownership)'. DC 부지, 건물, 전력 설비, GPU까지 직접 소유 및 운영. 이는 15년 장기 클라우드 임차 계약에 의존하는 OpenAI의 Stargate 프로젝트와는 정반대의 접근. 초기 높은 CAPEX 투자를 감수하고, 장기 총소유비용(TCO) 절감과 완전한 기술 통제권 확보를 추구.
- 현재 모델 학습을 위해 23만 개(이 중 GB200은 3만개)의 GPU로 구성된 Colossus 1단계 운영 중.
- 차세대 동력은 테네시 주에 건설 중인 Colossus 2단계. 11만 개의 GB200을 시작으로 해서, 이후 55만 개의 GB200·GB300, 그리고 최종적으로는 100만 개 이상의 Blackwell GPU 집적을 목표.
- Colossus 2단계 완공 시, 20,000 ExaFLOPS(FP8) 연산 능력을 의미. 현존 세계 2위 슈퍼컴퓨터를 가진 메타와 OpenAI의 클러스터 대비 50배 이상의 압도적 성능. 심지어 현존하는 9개의 다른 초대형 클러스터 합산치보다 11배 큰 연산 능력을 보유하게 되는 압도적 선두로 부상.

xAI vs OpenAI: AI 인프라 전략 비교 분석

구분 (Category)	xAI (Colossus 프로젝트)	OpenAI (Stargate 프로젝트)
핵심 전략 모델	풀-스택 소유	클라우드 파트너십 / 장기 임차
자산 구조	데이터센터, 토지, GPU 등 하드웨어 자산 직접 보유	운영 리스 중심. 자산을 직접 소유하지 않음
비용 구조	초기 자본(CAPEX) 투자 후, 장기 총소유비용(TCO) 절감	초기 투자 최소화, 15년 등 장기 계약으로 인한 TCO 증가
인프라 통제권	설계, 구축, 업그레이드, 운영까지 완벽한 기술적 통제권 확보	핵심 인프라 통제권이 클라우드 제공사(Oracle 등)에 종속
자금 조달 방식	Grok IP 및 GPU 자산을 담보로 대출, 등 직접 파이낸싱	파트너사(MS 등)의 투자 유치 및 리스 계약을 통한 비용 처리

자료: xAI, OpenAI, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Colossus by xAI

"AI-Compute-Energy" 수직 통합 모델의 시대…

수혜는 누가 볼까?

- Compute-Energy 수직통합이 AI 업계의 새 표준. DC + 배터리 + SW가 단일 밸류체인으로 재편.
- 모델 학습-배터리-전력제어를 단일 운영 플랫폼에서 통합. 특히 테슬라 ESS인 Megapack을 활용한 자체 전력 부하 관리는 특장점. Megapack으로 순간 전력 급변동(체크포인트·통신 지연)을 흡수. 모든 서버를 전량 액체 냉각(콜드-플레이트) 시스템을 도입해 전력 효율/집적도를 극한으로 끌어올림.
- 여기서 가장 큰 승자는 엔비디아와 테슬라. xAI가 신제품이 나올 때마다 최대 수요처로 자리매김하는 덕분에, 엔비디아는 GPU 로드맵의 성공을 위한 든든한 아군을 얻게 됐음. 또한 xAI는 대용량 테슬라 Megapack 수요를 창출하는 기회 요인으로 급부상.
- 다만, 이 거대한 비전의 성공은 여러 제약 요인이 있음. 55만 개의 GB200만으로도 수 GW가 필요하고, 5천만 개의 GPU는 35 GW나 필요. 대규모 송배전 및 HVDC가 필수. 이를 위한 자금으로, 최근 Grok 모델의 IP를 담보로 하여 50억 달러를 대출받고 Blackwell GPU 35만 개 구매를 위해서 추가로 120억 달러 구조화 부채 조달을 추진 중

Colossus 2단계 관련 주요 내용

더 효율적인 중앙 냉각 시스템을 기다리는 대신, 비싸고 비효율적인 이동식 냉각 장치(mobile chillers)를 사용해 데이터센터 가동을 3개월 앞당김. 3개월 동안 수억 달러짜리 GPU들을 창고에 놀리는 기회비용이 냉각 장치의 추가 비용보다 "훨씬, 훨씬, 훨씬 더" 크기 때문

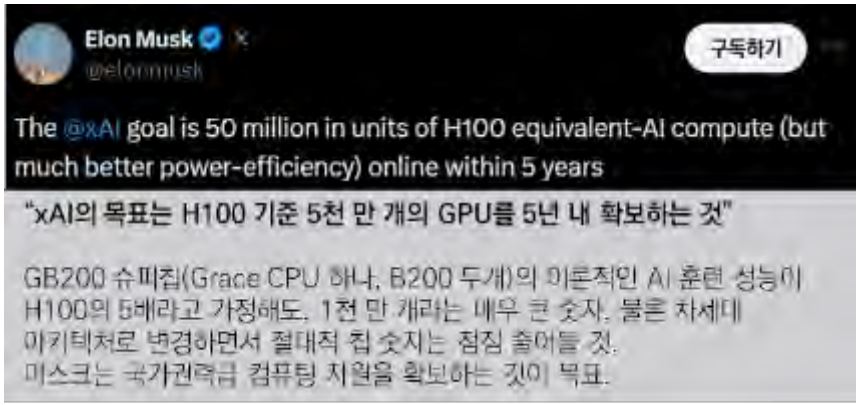
항목	내용
부지	Tulane industrial park, Shelby County, 테네시주
GPU 1차 배치	110 k GB200 (2월 중순 설치 및 시범 가동 중)
우선 목표	550k GB200 & GB300 1차 가동 후 단계적 증설 (현재 전력 인가 준비 중)
가동 시점	"수 주 내 전원 인가" (xAI 발표)
냉각	전량 액체 냉각(콜드-플레이트)으로 이동식 냉각 장치 사용
백업 전력	테슬라 Megapack (수백 MWh 규모) + 다중 전원 루프

자료: xAI, OpenAI, 미래에셋증권 리서치센터

[참고] Colossus

실리콘과 전력으로 빚어낸 압도적 규모의 컴퓨팅 제국

일론 머스크의 Colossus 관련 트윗

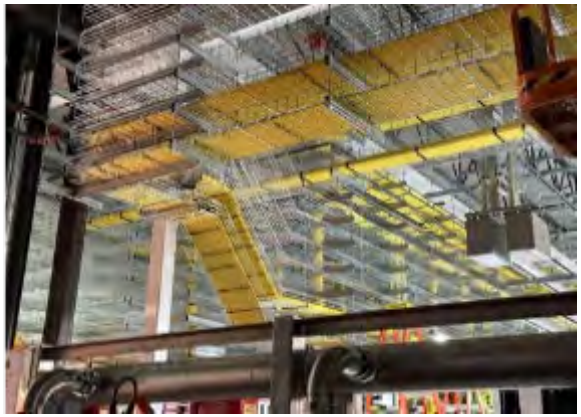


자료: 일론 머스크, 미래에셋증권 리서치센터

현존 최강의 AI 슈퍼컴퓨터 클러스터인 xAI의 "Colossus"
세 달 만에 GPU가 10만 개에서 20만 개로 확장

자료: xAI, 미래에셋증권 리서치센터

일론 머스크가 직접 게시하며 과시한 Colossus 내 여러 케이블링 사진, '데이터 통신과 전력 이동 관리가 핵심이라는 방증'



- GPU 서버가 장착된 랙 앞면.
- 포트-포트 간 초고속 NVLink 및 InfiniBand 패치 케이블을 정교하게 한쪽 방향으로 감아 올린 모습.
- 운영·냉각 효율
- 미학적 케이블 관리

- 데이터센터 천장에 매달려 있는 철제 그물 선반(와이어 트레이)로, 케이블이 지나가는 전용 '도로'

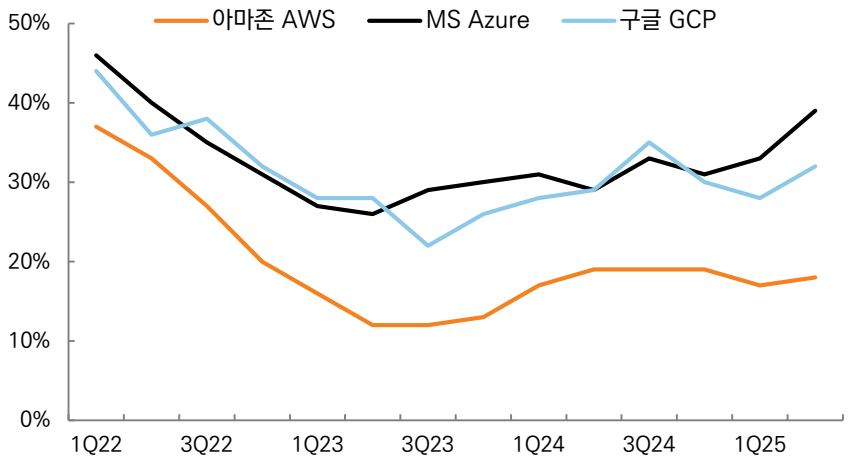
자료: xAI, OpenAI, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Rainier by 아마존+Anthropic

Anthropic의 성공이 곧 AWS의 성공

- "Project Rainier"는 아마존 AWS가 차세대 AI 패권을 장악하기 위해 오직 Anthropic만을 위해 수십억 달러를 투입하는 GW급 AI 인프라 구축 이니셔티브의 코드명. 마이크로소프트-OpenAI 연합의 약진으로 잃어버린 AI 모멘텀을 되찾고, 폭발적으로 성장하는 Anthropic의 매출을 AWS 클라우드 수익으로 직접 연결시키려는 야심 찬 계획. 프로젝트 목표는 2.2GW이고, 현재로서는 1.3GW 이상의 용량으로 매우 빠른 속도로 구축 중인 것으로 파악.
- Anthropic의 기술 로드맵에 최적화된, 세계에서 가장 강력하고 비용 효율적인 독점 AI 클러스터를 제공하여 경쟁사(OpenAI 등)를 압도하는 기술적 우위를 확보하는 것이 목표. "Anthropic 전용"으로 1.3GW 이상의 데이터센터 캠퍼스들을 AWS 역사상 전례 없는 속도로 건설 중.
- Project Rainier의 심장은 100만 개에 달하는 아마존의 자체 AI 가속기인 "Trainium"으로 구성된 거대 클러스터. 이는 엔비디아 GPU 지배 체제에 대한 가장 강력한 도전. 물론 GPU에 비해 순수 연산(PFLOPs) 능력은 열세이지만, 비용효율 및 메모리 대역폭에서의 우위를 확보하는 전략.

클라우드 3사의 분기별 매출 성장률 추이
마이크로소프트 Azure와 구글 GCP 대비 성장률이 부진이 뚜렷



자료: 각 회사, 미래에셋증권 리서치센터

작년 12월 아마존 "AWS re:Invent" 행사에서 공식 발표된 "Rainier"

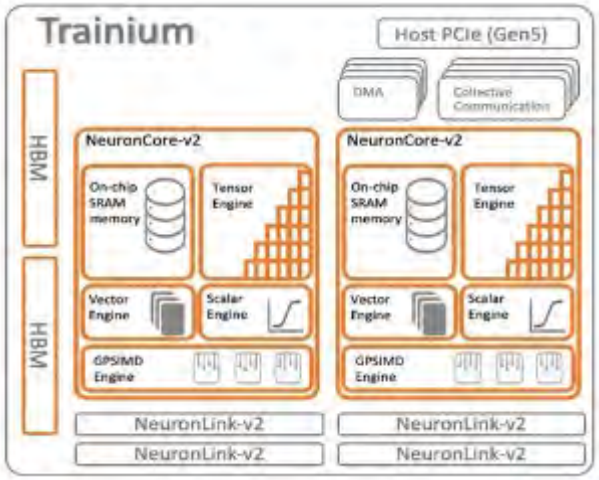
카테고리	세부 내용
목적	AI 모델 훈련을 위한 세계 최대 컴퓨트 클러스터 구축; Anthropic의 AI 훈련 지원
하드웨어	**Trainium2 칩 (1.3 PFLOPS FP8, 96GB HBM), 스케일업 구조**
네트워킹	**NeuronLink v3 (1TB/s, 아마존식 NVLink)**
규모	Anthropic의 기존 클러스터보다 5배 강력, 수십만 Trainium2 칩, 2.2GW 전력
파트너십	주요 고객은 Anthropic (아마존이 80억 달러 투자)
비용	아마존이 2025년 1,000억 달러 투자
미래 계획	2025년 4분기 본격 운영; 2026년 Trainium3으로 업그레이드

자료: 아마존, 미래에셋증권 리서치센터

[참고] Rainier

오직 Anthropic만을 위해 서빙합니다

아마존의 AI 가속기인 Trainium 아키텍처 다이어그램
2개의 NeuronCore는 AI 연산 수행 엔진, SRAM과 텐서 엔진 등으로 구성
NeuronLink는 엔비디아의 NVLink와 역할이 동일



자료: 아마존, 미래에셋증권 리서치센터

Rainier에 채워 들어갈 핵심 "Trainium2 Ultra" 서버 랙
하나의 랙에 64개의 Trainium2 칩을 집적해 184 TB/s의 메모리 대역폭
Anthropic이 보유한 기존 최대 클러스터보다 5배 더 큰 컴퓨팅 파워

자료: 아마존, 미래에셋증권 리서치센터

"자체 칩 Trainium 기반의 시스템 vs 엔비디아 GPU 기반의 시스템"의 총소유비용 비교, "순수 연산력에서는 안되지만 메모리 기준의 효율성에서 우위"

항목	단위	GB200 NVL72	Trainium2-Teton2-PD-Ultra-3L	VR200 NVL144	Trainium3-Teton3-Ultra-3L
AI 가속기 클러스터 총소유비용 (TCO)					
칩 하나당 시간별 자본 비용	USD/hr/GPU	$1.87	$0.43	$2.99	$0.53
칩 하나당 시간별 운영 비용(전력 등)	USD/hr/GPU	$0.49	$0.23	$0.66	$0.36
칩 하나당 시간별 총소유비용 (TCO)	USD/hr/GPU	$2.36	$0.66	$3.65	$0.89
AI 가속기 클러스터 처리량 대비 TCO					
유효 학습 PFLOP당 TCO	$/hr/PFLOP	$0.94	$1.69	$0.49	$1.12
엔비디아 대비 Trainium % 우위			**-44.1%**		**-56.4%**
메모리 대역폭당 TCO	$/hr/TB/s	$0.30	$0.23	$0.28	$0.22
엔비디아 대비 Trainium % 우위			**+30.1%**		**+25.8%**

자료: SemiAnalysis, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Prometheus/Hyperion by 메타

Llama의 실패가 촉발한 거대 클러스터 구축계획

- DeepSeek에 오픈소스 주도권을 내준 Llama 4 모델의 대실패는 AI 전략에 근본적 전환을 야기. AI 성능의 핵심 병목이 모델 설계가 아닌 압도적인 컴퓨팅 파워와 데이터 품질에 있음을 확인. 기존의 점진적 AI 투자 기조를 폐기하고, 패권 장악을 위한 인프라 우선주의로의 전면적인 전략 수정을 촉발.
- 과거 자사주 매입에 집중되던 자본을 AI 인프라(GPU, 데이터센터) 자본 지출(CAPEX)로 공격적 전환. 이는 단기적 주주환원보다 장기적 기술 패권 확보를 최우선 순위로 삼겠다는 명확한 시그널.
- AI 시장의 단순 참여자가 아닌, 압도적인 컴퓨팅 인프라를 통해 경쟁 자체를 무의미하게 만드는 '물리적 독점'을 구축하는 것으로 목표 재설정.
- 메타는 최소 4개의 대규모 AI 클러스터 구축을 통해 전례 없는 규모의 컴퓨팅 자산을 확보 중. 이 프로젝트들은 엔비디아에 3천 억 달러 이상의 잠재적 수익 기회를 제공할 것으로 추산.
- 이 프로젝트들은 단순 데이터센터가 아닌, 메타의 초지능(Superintelligence) 개발을 위한 전략 자산.

메타의 AI 데이터센터(DC)는 초지능(Superintelligence) 개발을 위한 전략적 자산으로 전체 규모는 “10GW+”
이는 엔비디아에 3천 억 달러 이상의 잠재적 수익 기회를 제공할 수도…

프로젝트명	지역	완공 시기	전체 전력 (GW)	주요 특징	GW당 엔비디아의 수익 추정 (십억 달러)	각 전력 규몽 따른 엔비디아의 총 매출 기회 (십억 달러)
프로메테우스	오하이오	2026년	1	자체 발전소 건설	$30	$30
하이페리온	루이지애나	2030년 (2028년에 부분운영)	5 (1.5GW는 재생에너지)	OpenAI Stargate 대응 130만 이상 GPU 탑재	$40	$200
타이탄 (1)	미국	?	2	병렬 처리 능력 강화	$20	$40
타이탄 (2)	미국	?	2	추가적 훈련/추론 목적	$20	$40
합계			**10**			**$310**

자료: Company Reports, 미래에셋증권 리서치센터

IV. AI 군비경쟁: Prometheus/Hyperion by 메타

무엇보다, 사후학습에 크게 베팅한다

AI 시장의 진입장벽을 높여, 경쟁 해자를 만든다

- 메타는 데이터센터 건설 철학을 '완벽함'에서 '속도'로 전환. AI 경쟁의 승패가 인프라의 조기 확보에 달려있다는 판단에 기인. 10년간 유지해왔던 맞춤형 DC 설계는 건설 기간이 길다는 이유로 바로 폐기. 마치 야전 텐트와 같은 조립식 모듈과 초경량 구조로 건설 기간을 극단적으로 단축. 예비 발전기 등 안정성 설비를 과감히 포기하고, 확보된 자원을 GPU 추가 구매에 투입하여 연산 능력에 완전 베팅.
- 이러한 전략은 사후학습(Post-training) 및 강화학습(RL) 워크로드에 최적화. 반면, 절대적 안정성이 요구되는 사전학습(Pre-training)의 리스크는 기존 시설 혹은 다른 방식으로 관리할 것으로 추정.
- 메타의 인프라 투자는 후발주자의 추격을 원천적으로 차단하려는 '경쟁 해자' 구축 전략. '연구원 1인당 세계 최고의 컴퓨팅 파워' 제공은 최고 수준의 AI 인재를 유인하는 가장 강력한 무기. 경쟁사의 핵심 두뇌를 흡수하고 인재 시장의 비용 구조를 왜곡시켜 경쟁을 더욱 어렵게 만듦.
- 트럼프 정부의 간판 공약을 담은 'OBBB' 법안은 데이터센터 및 GPU 자본 지출에 대한 100% 보너스 감가상각을 허용. 이는 메타의 막대한 투자를 연방 정부가 세금 감면 형태로 보조해주는 효과를 낳아, 공격적 투자의 재무적 부담을 완화시켜주는 효과.

OBBB 법안으로 인해 메타가 받을 주요 세제 혜택 및 효과 분석, '그럼에도 단기적으로는 재무적으로 큰 비용 손실이 있을 수 밖에 없음'

혜택	기존 제도 (일반 규정)	OBBB 법안으로 달라진 점	기업-측 효과·의미
100% 보너스 감가상각	고가 설비는 내용연수(5년)에 걸쳐 해마다 20%씩만 비용 처리 가능	구매한 그 해 장부가액 전액을 한번에 비용 처리하여 세금 공제	미래 절세 효과를 현재 가치로 당겨와 유동성↑ 메타가 조 단위 GPU·서버 지출을 즉시 비용으로 처리해 당기 법인세 대폭 감소
R&D 비용 전액 공제	인건비·전기료 등 R&D 지출은 세액공제 비율·상한이 제한적	올해부터 美 내 R&D 비용은 전액 즉시 공제 가능. 바이든 정부 시기 분할 상각 중이던 잔액도 2025년에 소급 공제 가능	연구개발에 쓴 현금이 그대로 세금 절감으로 연결 → AI 모델 개발 비용 부담 완화, R&D 투자 가속
이자 공제 한도 확대	차입 이자의 일부 금액만 산입 가능 (공제 한도)	이자공제 상한 대폭 상향 → 더 많은 이자를 비용 처리	대규모 차입 시 실질 이자 비용 절감 → 레버리지 활용 여력 증가, 자본비용↓

자료: 미래에셋증권 리서치센터

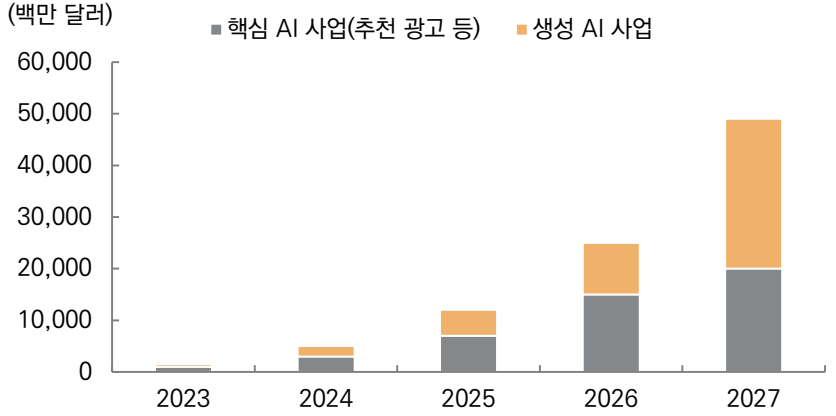
[참고] Prometheus/Hyperion

'초지능' 개발이라는 단 하나의 목표를 위해 다른 모든 것을 희생한다

Hyperion의 크기는 축구장 70개, 2030년까지 5GW까지 확장할 계획

자료: Sawyer Merritt, 미래에셋증권 리서치센터

'27년에 생성 AI 비용이 메타 기존 사업 비용을 넘어설 것



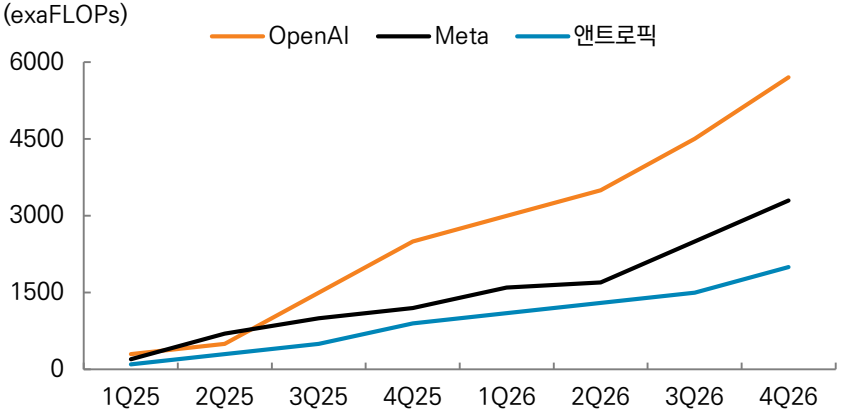
자료: SemiAnalysis, 미래에셋증권 리서치센터

구글의 소송문건에서 드러난 정보 중 주요 챗봇 앱들의 사용량 통계

회사	DAU (백만명)	사용자 숫자 점유율	사용자 쿼리 점유율
ChatGPT	160	49%	71%
메타AI	100	31%	12%
Gemini(구글)	35	11%	8%
Grok(xAI)	15	5%	4%
DeepSeek	10	3%	3%
퍼플렉시티	7	2%	2%

자료: 구글 반독점 소송 문건, 미래에셋증권 리서치센터

AI 모델 훈련을 위한 차세대 클러스터의 피크 성능



자료: SemiAnalysis, 미래에셋증권 리서치센터

V. 에너지와 그리드

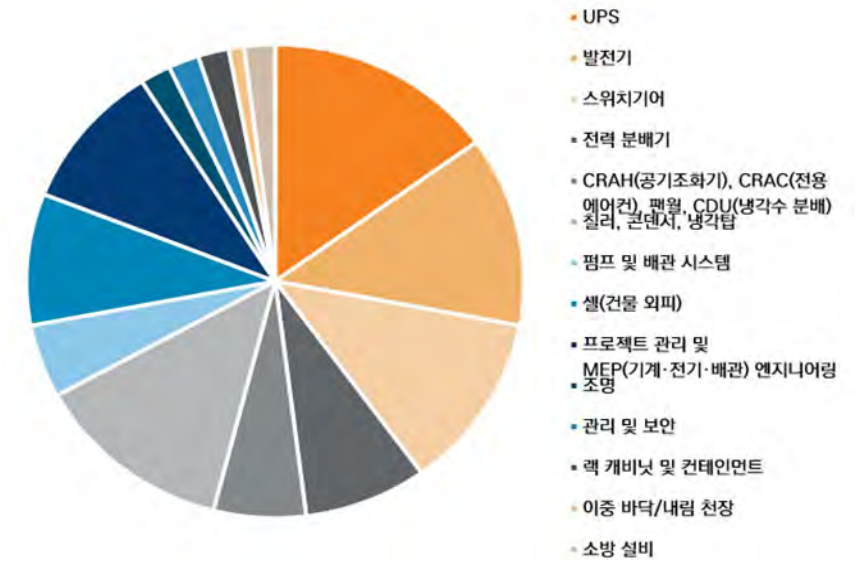
AI팩토리 온사이트 BESS

V. 에너지와 그리드: 스케일 전환 “데이터센터 → AI 팩토리”

AI 팩토리는 ‘컴퓨팅+전력+냉각’이 한 몸처럼 설계되는 산업 플랜트

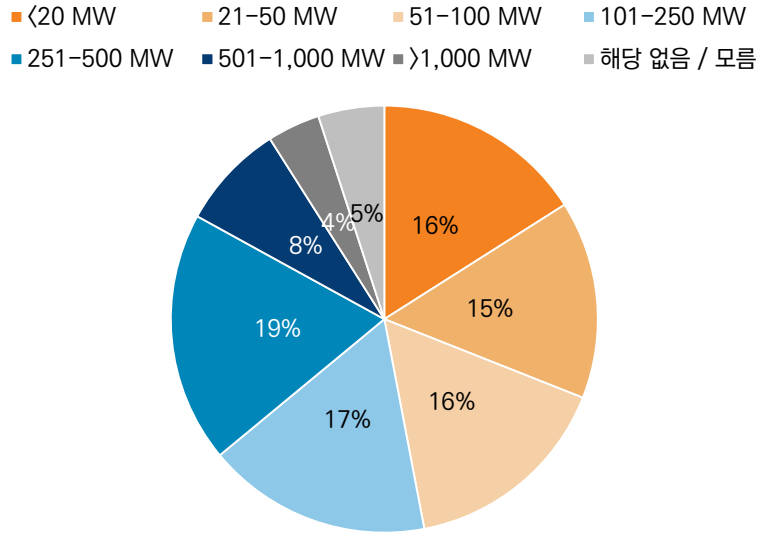
- AI 시대의 데이터센터(DC)는 단순한 서버 집합체가 아닌, 하나의 거대한 AI 클러스터로 재정의되는 중. 이는 LLM과 같은 대규모 AI 모델이 요구하는 방대한 연산량을 처리하기 위해, 수만 개 이상의 GPU를 고대역폭, 저지연 인터커넥트로 긴밀하게 결합하여 단일 시스템으로 동작 시켜야 하기 때문.
- 특히, AI 클러스터의 전체 성능은 노드 간 통신 지연 시간에 매우 민감하므로, GPU 서버의 물리적 인접성을 극대화하는 것이 필수적. 이는 필연적으로 랙의 고밀도 집적으로 귀결. 고성능 가속기가 좁은 공간에 집적됨에 따라, 랙 단위의 전력 밀도와 발열은 기하급수적으로 증가(12kW → 120kW 이상, 향후 1MW), 고도의 전력 분배 시스템과 액체 냉각(DLC) 같은 차세대 열 관리를 필수 요건으로 만듦.
- “AI 팩토리”의 출현으로 인한 고밀도화 추세는 DC 프로젝트 전체의 거대화로 이어짐. 거대. 현재 신규 DC 프로젝트의 48%가 100MW를 초과하는 대규모화 추세. 이중 4%는 GW에 이미 도달했다고 응답. 이는 대부분의 신규 프로젝트들이 AI 워크로드 기반으로 진행되고 있음에서 기인.

주요 부품별 데이터센터 인프라 구축 비용 점유율 추정



자료: SemiAnalysis, 미래에셋증권 리서치센터

향후 건설될 데이터센터 프로젝트의 용량에 관한 설문 결과(n=149)

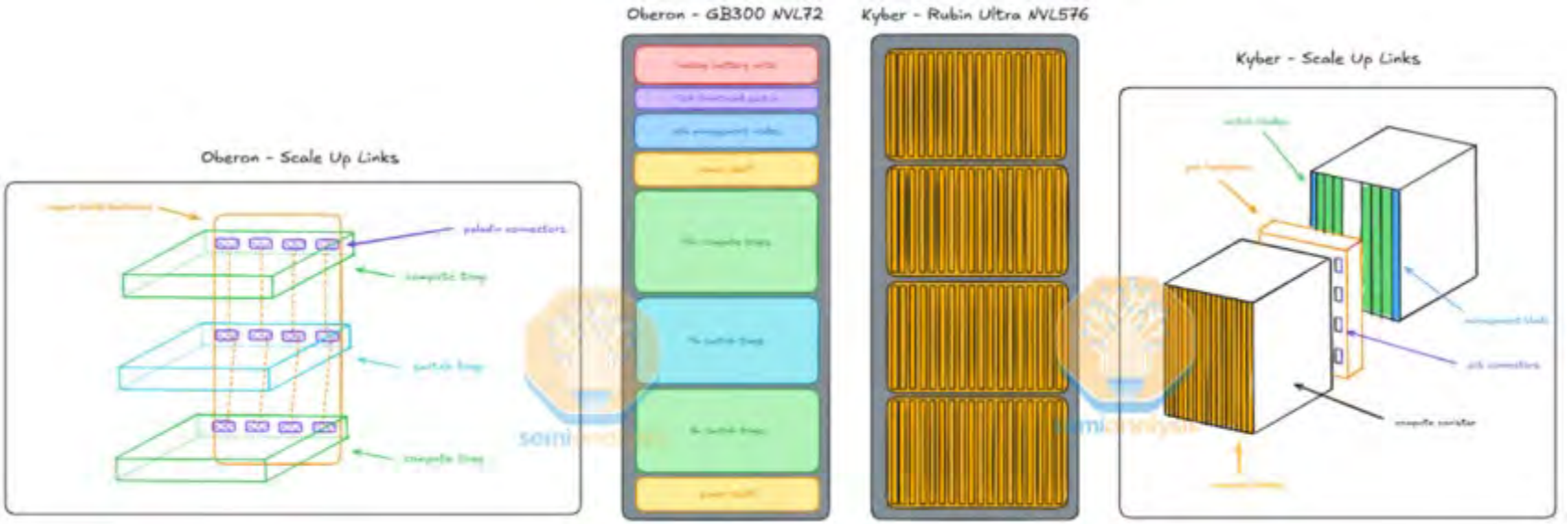


자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

[참고] 엔비디아의 차세대 랙 아키텍처 “Kyber”는 고밀도 집적의 정수

DC의 물리적 구조까지 바꾸는 ‘Kyber’ = 기존의 컴퓨팅 트레이를 90도 회전시킴

Kyber는 기존의 수평 트레이 Oberon 구조를 버린 새로운 시스템으로, compute blades가 수직으로 꽂히는 canister 형태로 구성
이런 선택의 이유는 밀도(Density) 때문. Rubin Ultra 패키지를 144개까지 담아 NVL576이라는 경이로운 시스템을 구성



특징	Oberon 랙	Kyber 랙
랙 방향	기존 세로 방향	**90도 회전해 가로 방향 = 밀도 증가**
구성	트레이 기반	**4개 캐니스터, 18개의 카트리지(블레이드)**
GPU 수	72 패키지	**144 패키지(576 다이) = 1패키지에 4다이**
연결 방식	구리 케이블 백플레인	**PCB보드 백플레인**
확장성	제한적	**향후 VR300 NVL 1,152로 확장될 가능성**

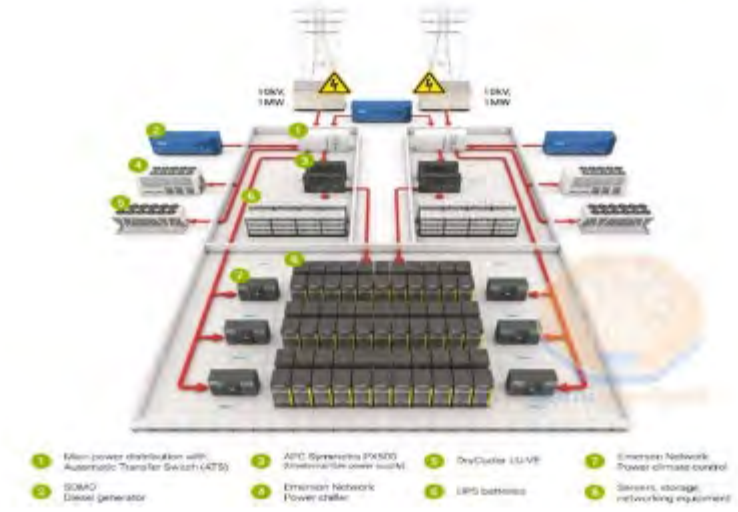
Rubin의 TDP(열 설계 전력)는 약 2,300W로 추정. 랙 하나에 144개나 들어가니, 단순 계산만으로도 **랙 하나의 전력 소모량은 330kW를 넘기게 됨.**
이렇게 막대한 전력을 기존 12V AC/DC 변환 방식으로는 효율적으로 공급할 수 없음. 즉, 향후 DC는 전용의 전원 랙/냉각 랙/스위치 랙을 따로 두는 구조로 전환될 것.
더 이상 랙 안에 파워 서플라이(PSU)를 두지 않고, 전용 전원 랙에서 생성된 전력을 컴퓨트 랙에 직접 공급하는 방식. 800V HVDC는 여기에 가장 들어맞는 기술.

자료: 엔비디아, SemiAnalysis, 미래에셋증권 리서치센터

[참고] AI 팩토리 전력 아키텍처와 수요 신호

유틸리티 기업들의 수요 지표가 보여주는 'AI 중심화'

AI 훈련을 위한 데이터센터 클러스터의 전력 공급 구조



디젤 발전기, UPS, ATS와 다중 전원 입력 등으로 구성.

ATS(Automatic Transfer Switch)는 주 전원이 고장 나면 자동으로 백업 전원 (예: 디젤 발전기)으로 전환하는 장치로, 데이터센터의 무중단 전원 공급을 보장

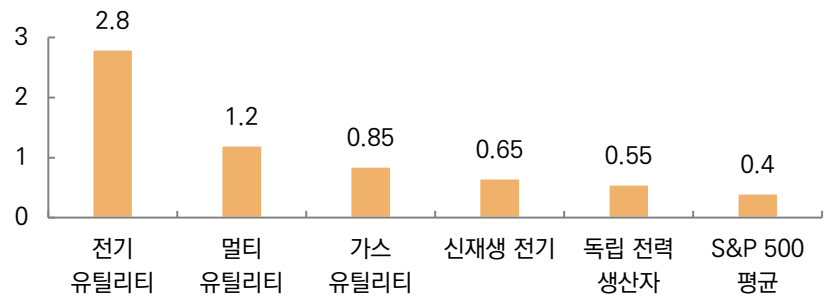
다중 전원 입력은 여러 독립적인 전원 소스(예: 듀얼 10kV 입력)를 동시에 연결하여 하나의 전원이 실패해도 다른 전원이 계속 공급되도록 하는 설계

밀집된 GPU 배치의 열 문제를 해결하기 위해 칠러, 건식 쿨러, 기후 제어 시스템을 통한 고급 냉각이 강조.

모듈식 서버, 스토리지, 네트워킹 레이아웃은 컴퓨팅 용량의 빠른 확장을 지원하며, 주요 재설계 없이 스케일업이 가능.

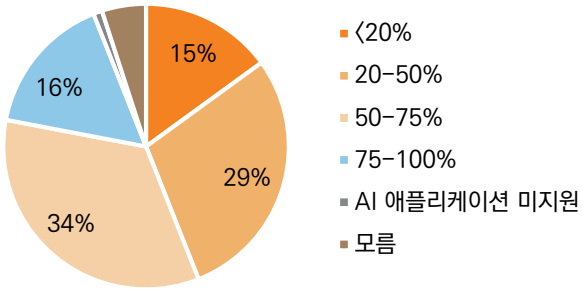
자료: DEAC, 미래에셋증권 리서치센터

유틸리티 섹터에서 AI 언급한 횟수 (S&P 500 기업들의 1Q25 실적 콜)



자료: 엔비디아, SemiAnalysis, 미래에셋증권 리서치센터

신규 데이터센터 용량 중 AI 워크로드 할당 비중



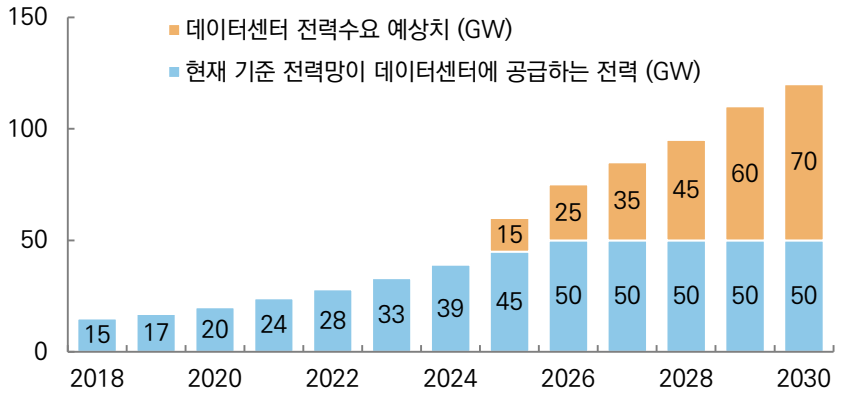
자료:Tom's Hardware, 미래에셋증권 리서치센터

V. 에너지와 그리드: Electric Shock

AI 팩토리의 프로젝트 규모가 커질수록, 지연이 기하급수적으로 늘어나

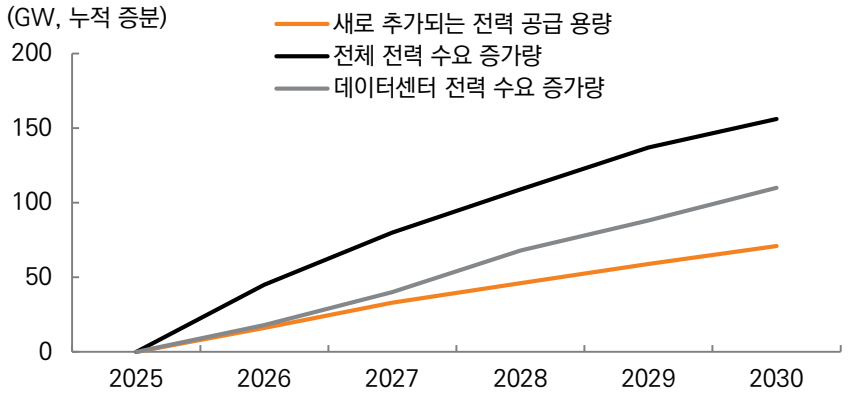
- S&P Global에 따르면, 미국 데이터 센터의 전력 수요는 2028년까지 두 배로 늘어나 95GW에 이를 것으로 예상되며, 2030년 수요는 2024년 대비 147% 증가한 117GW에 이를 것으로 추산. 이와 같은 거대 부하(load)의 등장은 전력망에 직접적인 충격을 줌.
- 텍사스 전력망(ERCOT)의 경우, 데이터센터 수요 급증으로 인해 향후 전력 용량이 76%나 더 필요할 것으로 예측. 북미 전력 규제기관인 NERC는 모든 송전 사업자에게 데이터센터 부하 모델링 현황 및 리스크를 검토. 전력망 신뢰도 위협에 더해 발전용량 한계로 데이터센터 그리드 연결 대기열은 더욱 심화되는 중. AI DC의 대규모화와 신규 프로젝트는 전력 공급 대기열을 심각한 수준으로 늘리고 있음.
- 현재 신규 프로젝트의 44%는 전력 연결을 위해 4년 이상의 대기열을 기다려야 하는 것으로 추산. 슈나이너 일렉트릭의 설문조사 결과, 응답자의 2명 중 1명(44%)은 평균 4년 이상을 대기하고 있는 것으로 나타났고, 12%는 평균 대기 시간이 7년에서 10년을 초과한다고 응답.
- 늘어지는 전력망 연결 대기 기간은 데이터센터 기업이 요청하는 전력 용량의 크기와 정비례. 250MW를 초과하는 대규모 프로젝트는 대부분 4년 이상의 대기 시간, 20MW 미만의 소규모 프로젝트는 짧은 기간 내에 전력을 공급 받음. 거대 AI 클러스터일수록 전력망에 가해지는 부담과 신뢰도 문제로 인해 연결이 더욱 어렵다는 '규모의 딜레마'를 명확히 보여줌.

미국 데이터센터 전력수요는 2030년이 되면 올해의 2배 정도에 이를 것



자료: S&P Global Market Intelligence, 미래에셋증권 리서치센터

당장 내년부터 전력 수급 공백이 전망되는 시나리오



자료: ENGINE NO1., 미래에셋증권 리서치센터

V. 에너지와 그리드: 에너지 수급의 병목

수요는 지금 폭증, 대형 기저부하는 느리게 증설

그 간극을 메울 수 있는 수단이 필요하다

- 지난 50년간 생산시설 해외 이전, 생산성 향상 등의 요인으로 미국의 전력 공급망은 사실상 정체 상태. 그러나 전기차, AI, 공급망 재산업화(리쇼어링)가 맞물리면서 전력 수요는 폭발적인 변곡점을 맞이. GE Vernova는 1분기 실적발표에서 현재를 '미국이 10년마다 전력망을 두 배로 늘렸던 '1945년과 1975년 사이'에 비유하며 전력 수요의 폭발적 성장을 강조.
- 문제는 폭발적으로 증가하는 수요를 신규 발전소 건설이 따라잡을 수 없다는 점. IEA의 'The Energy and AI' 보고서에 따르면, 신규 기저부하 발전소 중 건설 기간이 가장 짧은 복합화력발전소조차 최소 2년 이상이 소요. 더욱이, 이마저도 가스터빈 시장의 공급 병목 현상으로 인해 현재 H – Class 가스터빈의 리드타임은 4년 이상으로 빠른 시일 내에 건설이 쉽지 않은 상황.

미국은 EV·AI·리쇼어링으로 전력 수요 급증인데, 문제는 시간
단가 싼 발전원이 당장 해답이 아니라 "Time-to-Power"가 우선순위

전력원	건설 기간	공급 조절 여부	전 세계 평균 발전단가 (USD/MWh)
유틸리티 규모 태양광	1-4년	공급 조절 불가능	60
육상 풍력	2-5년	공급 조절 불가능	50
해상 풍력	3-7년	공급 조절 불가능	110
수력 발전소	5-15년	공급 조절 가능/불가능	80
전통 지열	3-8년	공급 조절 가능	80
원자력 (신규)	5-15년	공급 조절 가능	90
원자력 (재가동)	2-5년	공급 조절 가능	60
석탄	3-6년	공급 조절 가능	80
가스 복합화력 (CCGT)	2-4년	공급 조절 가능	80
가스 터빈 (GT)	1-3년	공급 조절 가능	220

자료: IEA, 미래에셋증권 리서치센터

GE Vernova의 2분기 실적 콜에서 가스터빈 관련 내용

구분	내용
수요	- 대형 가스터빈(HA 클래스 포함)과 DC향 터빈 모두 수요가 강력 - 특히 데이터 센터 지원을 위한 항공파생형 터빈 수주가 급증
수주잔고, 리드타임	- 55GW (공식 주문 + 슬롯 예약)이며, 연말까지 최소 60GW 예상 - 리드타임 4년 이상 소요될 것으로 전망
Capa	- 2026년 하반기까지 연간 20 GW 생산 가동률 달성을 목표 - 수주 잔고가 현재의 약 60 GW를 넘어 80~100 GW 수준으로 증가하는 것을 확인한 후 생산능력 확대를 검토

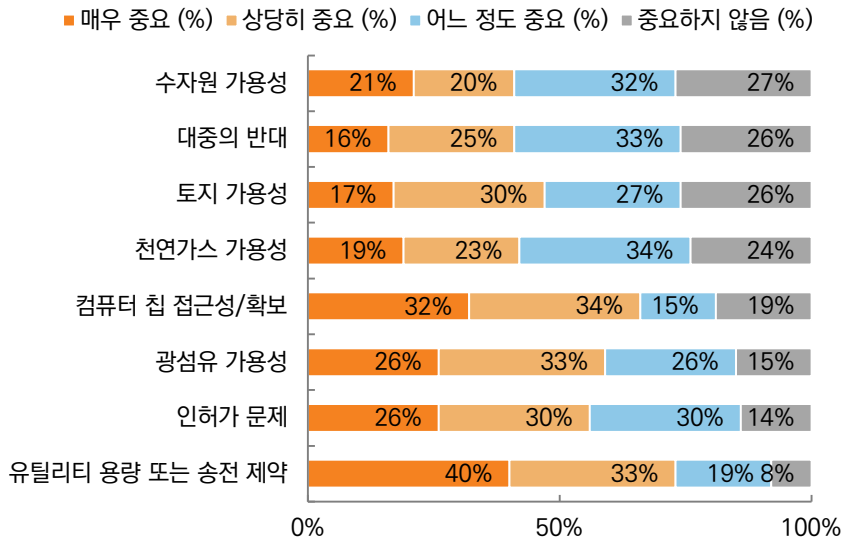
자료: GE Vernova, 미래에셋증권 리서치센터

V. 에너지와 그리드: 그리드 연결도 병목

그리드의 병목의 이유?
시간(허가/공사)
비용(요금/접속비)
공급(터빈/엔진/전력장치)

- 하이퍼스케일러들은 2025년에 “하루 10억 달러”, 연간 5천억 달러 이상을 지출에 투입할 것으로 전망. 이 중 5백억 달러가 전력 인프라에 필요하다고 함. 전력 수요가 기존 그리드를 빠르게 압도하는 상황.
- 또한, 슈나이더 일렉트릭의 설문조사 결과, 신규 DC 프로젝트의 주요 제약사항에서 AI 데이터센터 업계가 직면한 가장 큰 마찰 지점은 역시나 ‘전력’ 문제. 업계 관계자 10명 중 9명 이상(92%)이 전력 발전 또는 송전 용량 부족을 심각한 제약으로 꼽았으며, 40%는 이를 ‘매우 중대한 장벽’으로 평가.
- 이는 전력 병목 현상이 단순한 우려가 아닌, 업계가 직면한 최대 현실 과제임을 시사. 전력 병목을 유발하는 주요 원인으로는 ① 더 많은 전력을 공급받기 위한 대기 시간이 길어지는 문제 (46%), ② 추가 전력 공급에 대해 유틸리티가 높은 비용을 청구하는 문제 (43%), ③ 터빈, 엔진 등 현장 발전 장비의 조달 기간(리드 타임)이 긴 문제 (42%)같은 사항들이 지목됨.

데이터센터 프로젝트가 지연되는 이유



자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

유틸리티 및 그리드 제약의 세부 원인

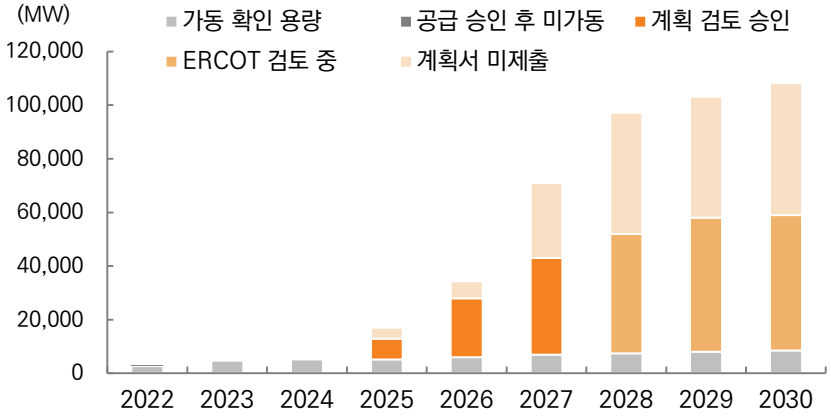
원인	응답 비율
전력 추가 공급을 위한 긴 대기 시간 (그리드 이용)	**46%**
높은 추가 전력 비용 (그리드 이용)	**43%**
자체 발전 설비(예: 터빈, 엔진)의 긴 조달 기간	**42%**
예측 불가능한 유틸리티 에너지 요금	42%
불충분한 유틸리티 전력 신뢰도	32%
유틸리티의 서면 공급 일정 약속 부재	32%
부지 레이아웃 문제 (예: 공간 부족, 배전선 설치 불가)	27%
어떠한 문제도 겪지 않음	15%
유틸리티가 충분한 재생 에너지 조달을 지원하지 않음	13%
해당 없음 / 모름	8%

자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

[참고] AI 팩토리의 핵심 리스크: 그리드 접속 시간

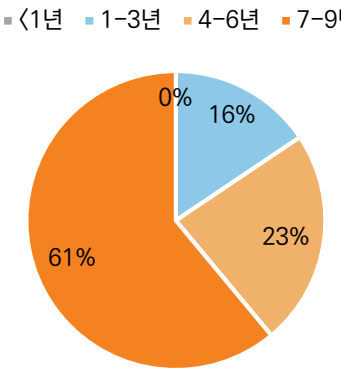
계획은 앞서가는데 승인은 지연, 대형 데이터센터의 경우 7~9년을 기다려야…

ERCOT 25년 3월 기준 그리드 연결 대기열



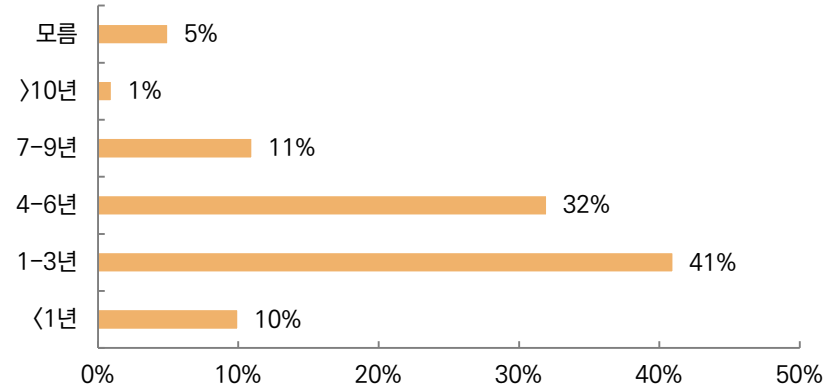
자료: EROCT, 미래에셋증권 리서치센터

250MW 규모 초과 데이터센터의 그리드 대기기간 조사 (n=149)



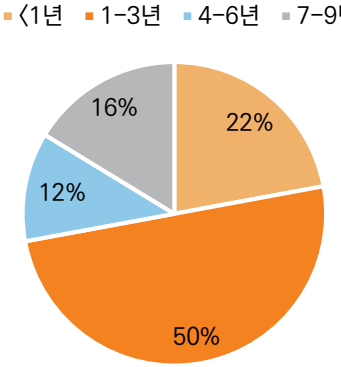
자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

데이터센터 프로젝트 그리드 예상 연결 대기기간 설문 결과



자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

20MW 규모 이하 데이터센터의 그리드 대기기간 조사 (n=149)



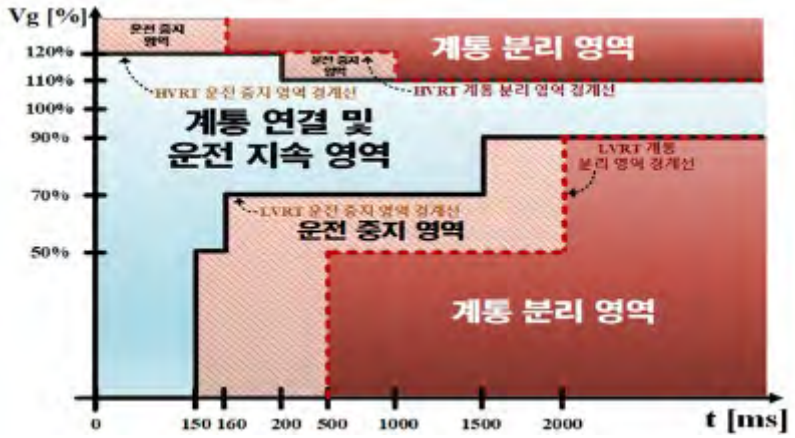
자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

V. 에너지와 그리드: 연쇄 고장 메커니즘 ①

LVRT 이벤트 +
대규모 부하의 동시 이탈
→ 주파수 쇼크
→ 연쇄 고장

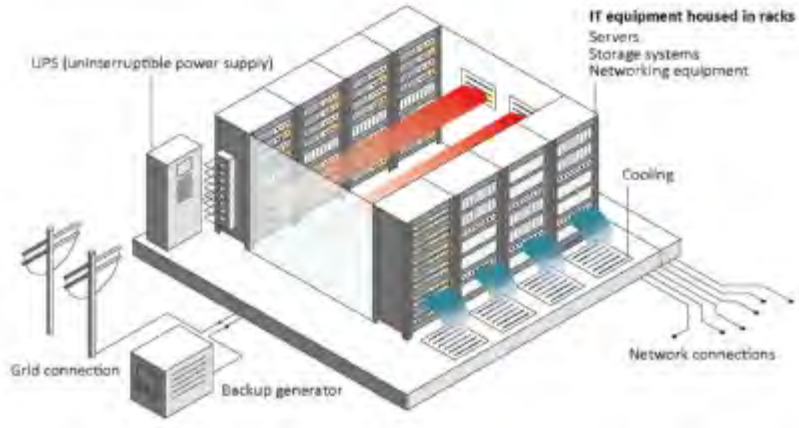
- AI DC의 거대화는 단일 부지에 GW급의 막대한 전력 부하를 집중시키는 결과. 거대/집중된 부하는 그 자체로 전력망에 부담을 주지만, 그 위협은 저전압 라이드스루(LVRT: Low Voltage Ride-Through) 상황에서 극적으로 증폭. 저전압 라이드스루는 송전선의 문제(나뭇가지 접촉, 야생 동물 등)를 해결하기 위해 리클로저가 작동하면서 발생하는 수 초 이내의 일시적인 전압 강하 현상. 문제는 리클로저가 고장을 완전히 해결하기 위해 여러 번 작동을 반복할 때 발생.
- 데이터센터의 무정전 전원 공급 장치(UPS)는 이러한 연속적인 전압 강하 신호를 '외부 전력망의 심각한 불안정'으로 해석. 첫 번째 전압 강하에는 내부 배터리로 잠시 전환했다가 복귀하지만, 이것이 반복되면 UPS는 IT 장비를 보호하기 외부 전력망과의 연결을 완전히 끊고, 백업 발전으로 시스템 전환. 대규모 DC 같이 전기를 대량 소비하는 부하가 전력망에서 급분리되면, 순간적으로 전력 공급량이 수요량을 초과하는 상태가 됨. 이는 발전기의 회전 속도를 급히 높여 전력망 전체 주파수를 상승 촉발.
- 이 주파수가 기준치를 벗어나면, 발전기는 과속으로 인한 파손을 막기 위해 스스로를 망에서 분리. 이때 그리드에 연결된 발전소들이 일괄적으로 망 분리를 일으키는 오버슈팅 현상으로 인해 주파수는 급감하고 그리드의 연쇄 붕괴가 시작되어 전체 전력망에 충격을 가하게 됨.

주파수·전압이 한계를 넘으면 연쇄 분리가 일어나고, 빨간색 영역으로 진입
AI DC는 반복된 LVRT → UPS 절연 → 계통 분리로 연쇄붕괴 촉발 가능



주: 가로축은 밀리초 시간 기준, 세로축은 전압 % 기준
자료: 한국조명전기설비학회, 미래에셋증권 리서치센터

전력(그리드→UPS→발전기)·냉각·네트워크(이중화)로 이어지는 설계
전력이 먼저 UPS로 들어와 전압 품질을 잡고, 순간 정전에도 시스템 보호



자료: IEA, 미래에셋증권 리서치센터

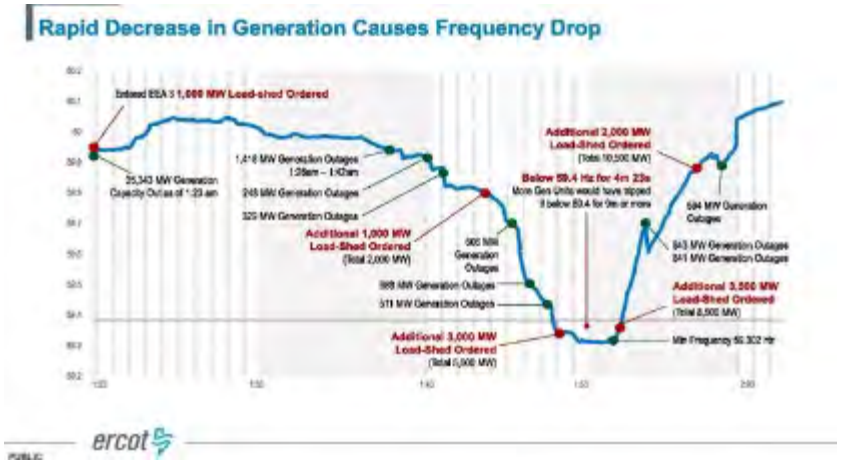
V. 에너지와 그리드: 연쇄 고장 메커니즘 ②

그리드 사업자는 리스크 비용을 DC에 전가

그리드 안정화 솔루션 수요가 구조적으로 커지는 이유

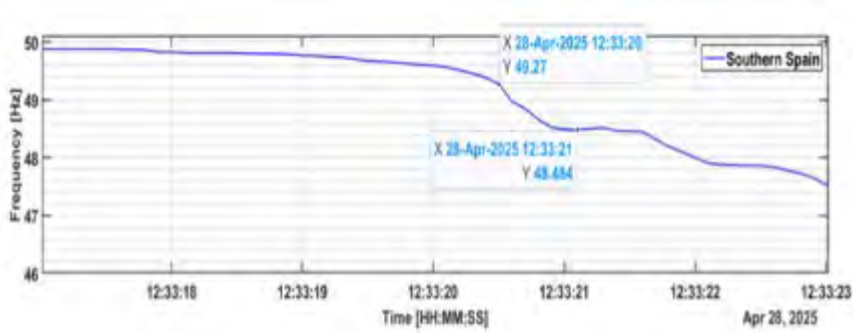
- 2024년 7월, 세계 최대 DC 밀집 지역인 버지니아 북부에서 송전선로 장애로 인한 전압 강하 사태가 발생. 지역 내 데이터센터들의 UPS(무정전 전원 공급 장치)가 IT 장비 보호를 위해 일제히 외부 전력을 차단했으며, 이 과정에서 약 1,500MW에 달하는 전력 수요가 순간적으로 소멸. 비록 대규모 정전으로 이어지지는 않았지만, AI 시대의 'LVRT 현상'에 동시에 반응해 연결을 해제하고 백업 발전으로 전환하는 경우, 전력망의 균형을 순식간에 무너뜨려 연쇄적인 계통 붕괴로 이어질 수 있음을 시사.
- 2025년 4월 발생한 이베리아 반도 정전 사태에서, 2.2GW 규모의 발전량이 예기치 않게 정지하자, 이는 전압 및 주파수 공급 조절 불가능의 연쇄 반응을 일으킴. 불과 27초 만에 전력망 전체가 붕괴. 이 사건은 북미 전력망이 직면한 위험과 매우 유사.
- 대용량 클러스터의 부하가 순간적으로 끊어질 경우, 이베리아 반도 정전과 유사한 연쇄 고장이 북미 전역에 발생 가능. DC가 유발하는 연쇄 정전 위험을 전력망 사업자들은 DC 사업자에게 전가. 안정화 설비와 백업전원 수요가 구조적으로 커지는 이유. 단일 DC 용량이 커질수록 충격량도 커지기 때문.

전력 주파수(Hz) 하락에 따른 텍사스 ERCOT 전력망 연쇄 붕괴 사례
발전기 고장이 연쇄로 발생하면서, 주파수 곡선이 완만히 하락하다 급락



자료: ERCOT, 미래에셋증권 리서치센터

이베리아 반도의 블랙아웃 당시 주파수 강하 현상이 관측



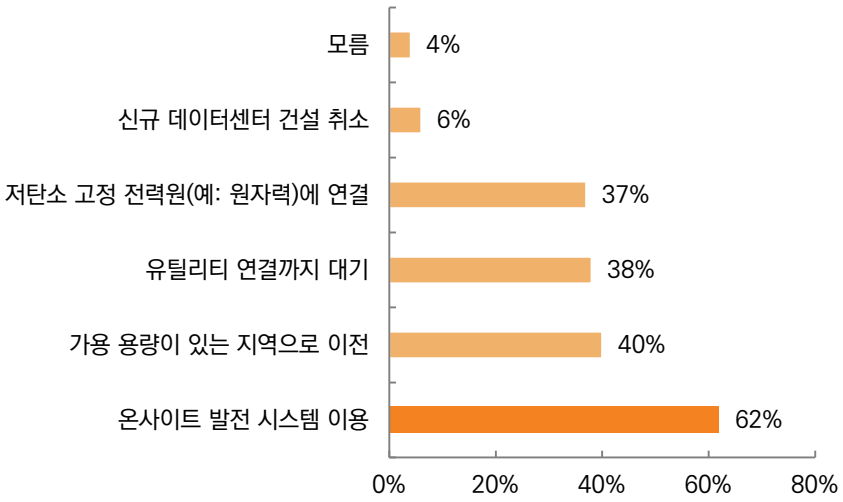
자료: entsoe, 미래에셋증권 리서치센터

V. 에너지와 그리드: 해법 ① On-site 발전

단기간에 전력 자립도를 확보하기 위한 Time-to-power (On-site)가 중요

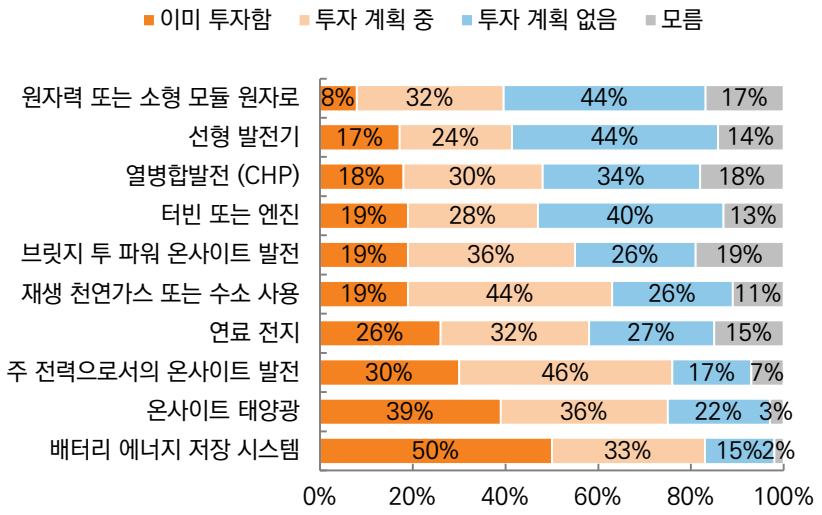
- 전력망 연결 지연이라는 문제에 직면한 DC 업계는, 발전원으로부터 그리드망을 거치지 않고 현장에서 직접 전력을 조달하는 'On-site(현장 발전)' 방식을 적극적으로 고려. 과거 비상 발전용으로만 여겨진 것을 주 전력원으로 고려하는 패러다임의 전환을 의미.
- 데이터센터 업계가 현장 전력을 구축할 때 가장 선호하는 모델은, 유틸리티 전력이 연결되면 현장 시스템을 비상용 백업으로 전환하는 'Bridge to Backup' 방식. 전력망과 아예 연결하지 않는 완전한 '독립형(Off-grid)' 방식도 버금갈 정도의 높은 선호도. 즉, AI DC는 안정적인 운영을 위해 전력망에 대한 의존도를 낮추고 자체적인 현장 발전을 지향. 최근 DC 구축에서 새롭게 떠오르는 고려 요소가 Time-to-power(얼마나 빠르게 전력을 공급받을 수 있는지)임을 유념. 가장 비싼 자원은 '시간'
- 하지만 가스터빈은 공급 문제, 디젤 발전기는 환경 오염 문제가 존재. xAI가 설치한 디젤 발전기는 소음/매연 때문에 지역 사회의 거센 민원에 직면한 상태. 따라서 현재 업계는 태양광+ESS 및 천연가스 같은 기저부하 발전, 그리고 운영의 내재화와 외주화 사이에서 최적의 방안을 모색하는 단계.

전력 그리드 대기열 장기화에 대응하는 방식



자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

데이터센터 업체들의 온사이트 발전 관련 투자 현황



자료: Alpha Structure, Schneider Electric, 미래에셋증권 리서치센터

[참고] 초거대 AI DC들은 모두 On-site 구축을 가속화 중

OpenAI의 자체 가스 발전부터 xAI의 BESS·태양광 결합 마이크로그리드까지, 더 이상 전력망을 못기다린다

OpenAI가 “Stargate I(왼쪽 사진)”과는 별도 캠퍼스인 새로운 AI DC 클러스터를 추가
Stargate I 한 곳만 1GW, 전체 Stargate 계획은 4.5 GW(미국 전체 데이터센터 전력의 ¼)

8월에 새롭게 알려진 데이터센터 클러스터는 Vantage Data Centres에서 시공 진행.
OpenAI와 Oracle, Vantage Data Centres 삼자간 협력으로, 규모는 250억 달러.
최초 데이터센터 시설은 내년 하반기 오픈 예정. 해당 부지의 공식 명칭은 “ Frontier”
개발 예정인 신규 사이트는 1.4GW급 시설. 기존 Stargate I 시설과 약 50km 떨어짐.
텍사스 샤클퍼드 카운티 내 약 485만㎡ 부지에 10개 개별 시설이 들어설 예정.
Oracle은 연간 10억 달러 이상의 대규모 가스 연료 소비 비용을 이 데이터센터 하나에 투입할 계획. 이때, 현장 부지에 건설되는 전용 천연가스 발전설비로부터 전력 수급함.
캠퍼스를 상시 가동하기 위해 송전망에서 끌어오는 방식이 아니라 가스 터빈 기반 발전소를 현장에 건설해 상주한다는 게 포인트. (온사이트 발전)
한편, 이 시설은 무려 랙당 250kW가 넘는 초고밀도 서버랙을 지원해 高밀도화를 증명

자료: 엔비디아, 미래에셋증권 리서치센터

xAI의 AI DC인 “콜로서스” 캠퍼스 안에 설치되어, 전력원 관제의 역할을 하는 ‘변전소 야드’… 1호기는 이미 운전 중이고 2호기도 “직접 시공 중”
xAI는 2년 반이 걸리는 걸 3달 만에 변전소를 완성: 그리드 인입(송전망→변전소→스위치기어로 잇는 과정)·절연개폐기·계측 및 배전 설비가 집약

상태(시점)	운영 모드	핵심 설비 및 특이사항
현재	그리드 병행이 기본, 필요할 때마다 간헐적 섬운전	변전소 1호(97일만에 준공), 2호(95일 목표로 시공 중) + 테슬라 Megapack 208대
변전소 1·2호 완공	그리드 + BESS 병행 최적화 (N+1 구조의 이중 인입)	그리드 N+1 이중화 구성 + 테슬라 Megapack 추가도입 ➜ 高가용성 달성
2027년 이후	자립성을 갖는 마이크로그리드 (그리드 연결 + BESS + 100MW 이상의 태양광 포함)	가스터빈 허가는 2027년 만료, PV 설치 후 가스 의존 축소, ➜ 에너지 믹스변화 / 低전력단가

자료: xAI, Action News 5, 미래에셋증권 리서치센터

V. 에너지와 그리드: 해법② BESS (1)

BESS는 사고 순간의 완충재이자, 평상시에는 피크 완화 장치

덕분에 새 연결 용량을 더 확보할 여지

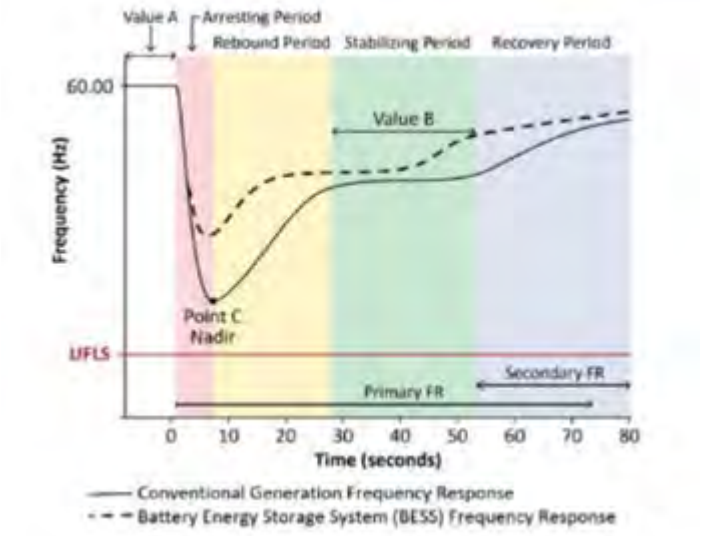
- BESS(대용량 배터리 장치)는 전력망을 안정시키는 결정적인 역할을 수행. 사고가 나면 UPS가 먼저 작동해 데이터센터 IT 부하를 전력망에서 잠깐 떼어 주는데, 이때 BESS가 밀리초(천분의 1초) 안에 충전을 시작해, 방금 전까지 IT가 쓰던 것과 똑같은 전력을 전력망에서 가져 옴. 그래서 전력망은 "부하가 갑자기 사라지는 충격"을 거의 느끼지 않는 효과. BESS가 시간을 벌어주는 동안, DC 운영자는 UPS 상태를 점검·재설정, 문제가 해소되면 재연결. 결과적으로 그리드 신뢰성 향상이 됨. 즉, 그리드의 주파수 변동에 초고속으로 반응하여 전력을 즉시 충/방전함으로써 전력망을 안정시키는 핵심적인 '1차 주파수 응답' 기능. 실제 사건들에서 BESS가 전체 물리적 대응의 70% 이상을 담당.
- 물론, 수요관리 측면도 큼. 듀크대 연구(美 전력 수요의 약 95%를 담당하는 22개 기관 조사)에 따르면, 신규의 전력 부하를 0.25%만 줄여도 미국 피크 부하의 10%에 해당하는 발전 여력을 확보할 수 있다고 분석함. 만약, 이때 데이터센터 부지에 설치한 BESS로 2~5시간 피크 수요를 떠안아 줄 수 있다면, 오래 돌려야 하는 발전원의 부담이 줄고 계통 운영 유연성이 커지는 것.

UPS는 필수 안전장치지만 '시간을 버는' 역할
BESS만이 3대 과제(평탄화·LVRT·연계)를 한 번에 해결

	AI 부하 평탄화	LVRT 지원	유연한 계통 연계	수명 (평탄화 적용시)	공간 차지 (Acres/ 100MW DC)
5분 배터리 (UPS)	가능성	가능성 계통 충돌 위험	불가 지속시간 제한	약 1년 사이클링 제한	해당 없음
자체 발전 (디젤 발전기)	불가	불가	가능성 대기 허가로 사용 제한	해당 없음	해당 없음
커패시터 (STATCOM)	가능	가능성 지속시간 제한	불가 지속시간 제한	20년	약 0.5
2시간 BESS (메가팩)	가능	가능	가능	20년	약 0.5

자료: Tesla, 미래에셋증권 리서치센터

전력 사고로 주파수가 떨어지기 시작하면,
BESS(점선)가 수 밀리초 만에 개입(방전)해 하락속도(RoCoF) 즉시 낮춤



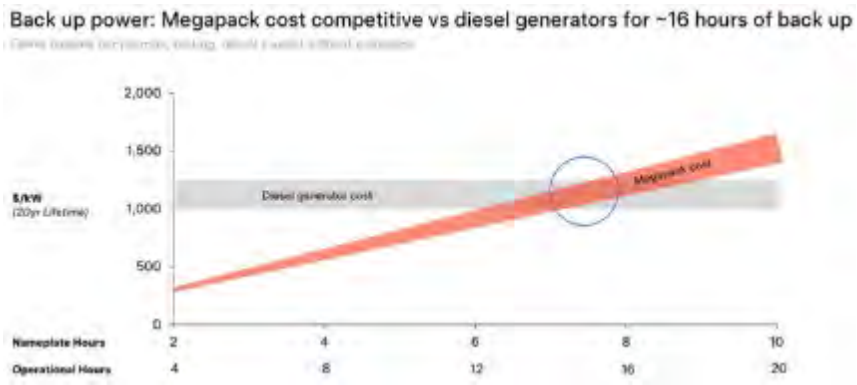
자료: NERC's 2025 State of Reliability Report, 미래에셋증권 리서치센터

V. 에너지와 그리드: 해법② BESS (2)

BESS는 그리드 안정화(주파수 응답)뿐만 아니라, "DR 수익 + 인터커넥트 승인 가속"

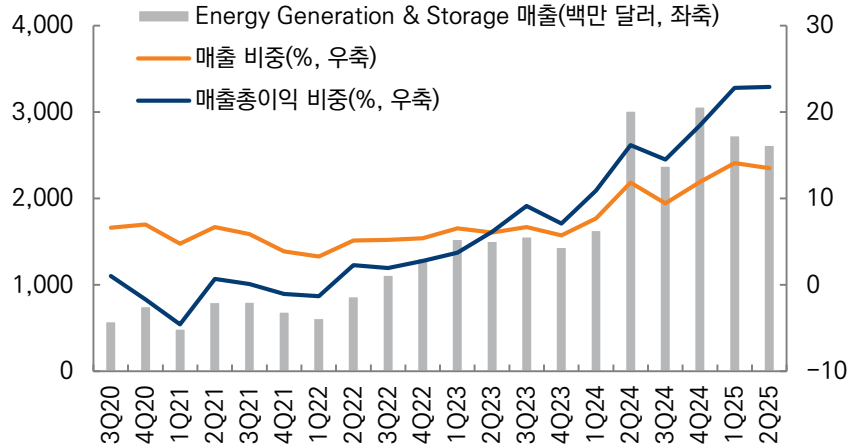
- 태양광·풍력은 날씨에 따라 발전이 들쑥날쑥하고, 사용량도 시간대마다 달라서 전력망 균형 잡기가 어려움. 이를 위해 등장한 것이 '수요대응(Demand Response, DR)'. "보상을 해줄 테니 전기 쓰는 시간을 옮기거나 줄여 달라"는 제도. 전력망이 힘들 때 부담을 나눠 갖는 방식. DR에 참여한 DC가 BESS를 갖추면, 전력망 과부하 걸리는 피크 시간엔 그리드를 잠시 줄이고 배터리로 자체 전력을 수급받으면 됨. 또한, DR에 적극적으로 참여하면 전력망 운영자로부터 더 높은 신뢰를 얻어 신규 전력망 연결을 더 빠르고 원활히 승인받을 수 있게 됨.
- 다만, BESS는 DC에 적용하기에는 높은 초기 비용이 장벽으로 작용. Lazard에 따르면 100MW급 4시간 BESS 구축에 최대 1.57억 달러가 소요. GW급 DC로 확대하면 10억 달러에 육박. 그래서 지금은 UPS·디젤을 대체한다기보다 추가적 설비로 보임.
- 하지만 트럼프 정부의 전력망 안정화에 대한 인센티브 지급과 지속적인 전기료 상승 추세는 BESS의 실질적인 TCO 감소를 이끌 것. 물론, DR을 통한 그리드 접속 심사 가속 등의 이점들이 겹치면서 데이터센터 사업자들의 BESS 도입 유인은 앞으로 더욱 증가할 전망.

백업 전원 비용 분기점 — 배터리(BESS) vs 디젤 발전
운영 기준 ≈16시간 이내에는 배터리가 더 싸고 실무 부담도 적다
xAI의 콜로서스 캠퍼스에 있는 208대의 BESS는 약 2시간 반을 버틴다



자료: Tesla, 미래에셋증권 리서치센터

BESS로 사용되는 테슬라의 에너지 저장 장치 설치량과 실적에서의 비중
매출총이익에서의 비중이 빠르게 증가, 테슬라 배터리는 점유율 전세계 8위



자료: SEC, 미래에셋증권 리서치센터

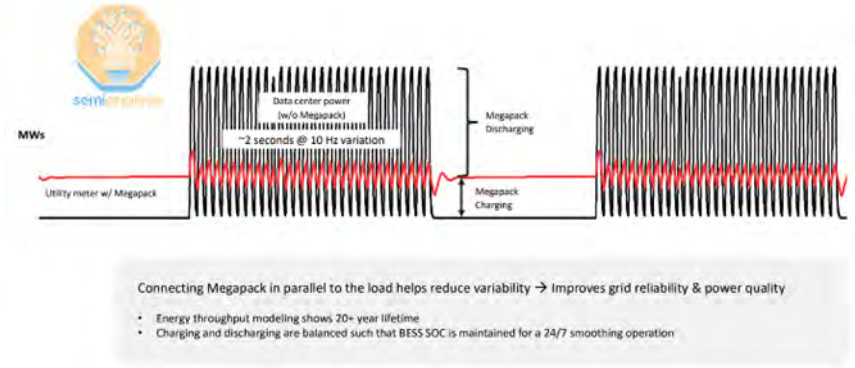
[참고] AI 시대의 전력망 위기를 타파할 배터리(BESS)

AI의 극심한 전력 변동성 문제를 해결할 핵심 열쇠, BESS의 역할과 실제 적용 사례

AI 훈련용 작업 부하가 만들어내는 전력 소비 패턴의 극단적인 변덕성
전력소비량이 1,000분의 1초라는 찰나의 순간에도 엄청난 변동

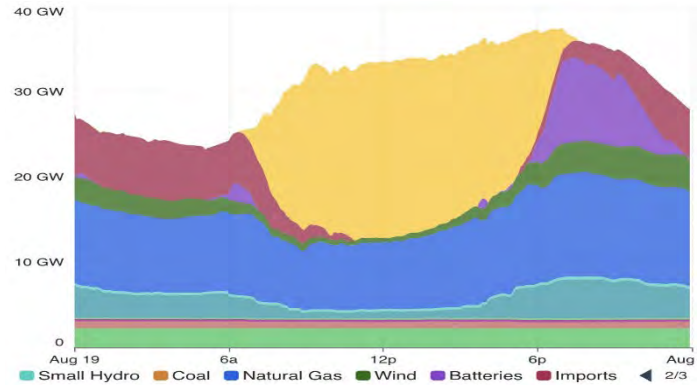
자료: Google at OCP EMEA Summit 2025, 미래에셋증권 리서치센터

테슬라 'Megapack'의 데이터센터의 전력부하 급변동 방어 역할
BESS는 데이터센터와 외부 전력망 사이에 위치한 거대한 충격흡수 장치
수요 급증 때 배터리 방전/흡수하고, 수요 급락 시엔 전력을 빨리 충전



자료: 테슬라, SemiAnalysis, 미래에셋증권 리서치센터

California Independent System Operator (CAISO) 그리드의 실시간 또는 최근 연료 믹스(에너지 생성 원천별)를 나타내는 스택 영역 차트 스크린샷
캘리포니아에서 지난 8월 19일 저녁의 전력 피크수요 시점에 배터리 저장이 전체 전력 공급의 27%를 차지



자료: CAISO, 미래에셋증권 리서치센터

왼쪽 그림에서 노란색 구역은 태양광 발전이고, 보라색은 ESS 관련 부분.
이는 과거와 달리 에너지 저장이 전력망 내에서 중추적 역할을 하게 되었음을 의미
배터리 + 태양광 연계가 에너지 공급 안정성과 탄력성을 크게 높인다는 증거

미 에너지부(DoE)는 여전히 석탄, 가스, 원자력 등 기저발전의 중요성을 강조하지만,
'에너지 지배 정책(energy dominance)'을 추구한다면 신재생+저장 투자 확대가 핵심!

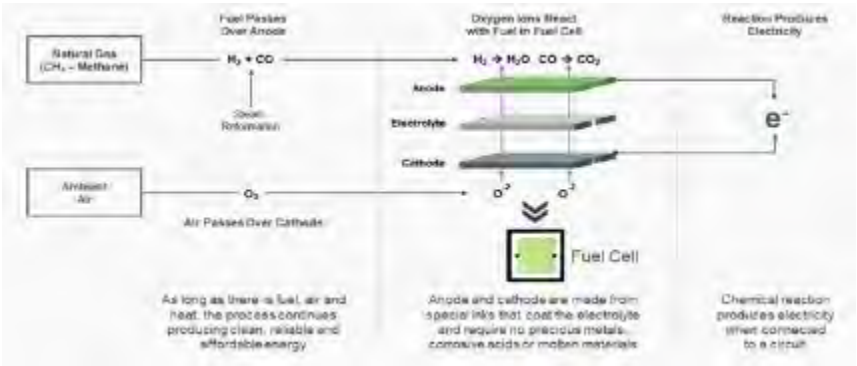
ESS와 태양광은 추가 비용을 야기하지만, **기존 천연가스 대비 운영비(OPEX) 절감 효과도 있음**
태양광+배터리를 낮과 저녁에 사용하면 천연가스 발전기 가동비용을 크게 절감 가능
원전 대체 가능성, 태양광 불필요 주장도 있으나, "둘 다 확보하는 게 진정한 에너지 풍요"

V. 에너지와 그리드: 해법③ 모듈형 SOFC(연료전지)

SOFC는 복잡 허가, 송전 공정 없이 수십~수백 MW를 사이트 내 증설에 유리

- 고체 산화물 연료전지(SOFC)는 가운데 고체 세라믹 전해질을 두고 양극·음극이 붙어 있는 전지. 귀금속 촉매, 강산, 녹는 전해질이 필요 없고 소재가 안정적이고 비용도 저렴. 작동 구조를 보면, 공기가 양극으로 들어오고, 연료(천연가스·바이오가스·수소)가 증기와 섞여 고온에서 개질. 양극에서 만들어진 산소 이온(O^{2-})이 전해질을 지나 음극으로 이동해, 연료와 만나 전기·증기·CO_2가 생김.
- 이때 만들어진 고온 증기는 버리지 않고 다시 음극으로 보내 연료 개질에 재사용. 그래서 외부로부터 물을 따로 공급할 필요가 없어 연료 유연성이 큼. 천연가스, 재생 바이오가스, 수소 등 여러 연료를 쓸 수 있는 장점도 있음.
- 데이터센터에서는 그동안 그리드나 가스터빈의 대체재로 과소평가됐지만, 빠르고 안정적인 온사이트 전원으로 점점 매력적으로 꼽히는 중. 여전히 부정적 인식이 남아 있으나 실제로는, "안정적 소재, 물 불필요, 다양한 연료, 고효율"이 강점. 특히 기존 가스 터빈보다 연료 사용량이 15~20% 적고, 대기오염이 없어 허가가 쉬워 경쟁력이 있다는 점이 부각.

고체 산화물 연료전지(SOFC)의 작동 원리
연료·공기를 계속 넣어주는 한 계속 전기 만들어내는 '전기화학 엔진'



1. 천연가스에 수증기를 섞어 뜨겁게 데우면 연료(H_2·CO)가 됨(개질)
2. 공기에서 들어온 산소(O_2) 가 전지 한쪽(캐소드)으로 들어옴.
3. 가운데 고체 세라믹(전해질)을 사이에 두고 산소가 이온(O^{2-})으로 변해 반대편 이동.
4. 이 이온이 연료와 만나 물과 이산화탄소 발생시키면서, 전자(e^-)를 밖으로 밀어냄.
5. 그 전자가 외부 회로를 따라 흐르는 것이 전기.

자료: Bloom Energy, 미래에셋증권 리서치센터

연료 전지에서 '에너지 서버(Bloom Box)'까지의 단계적 확장

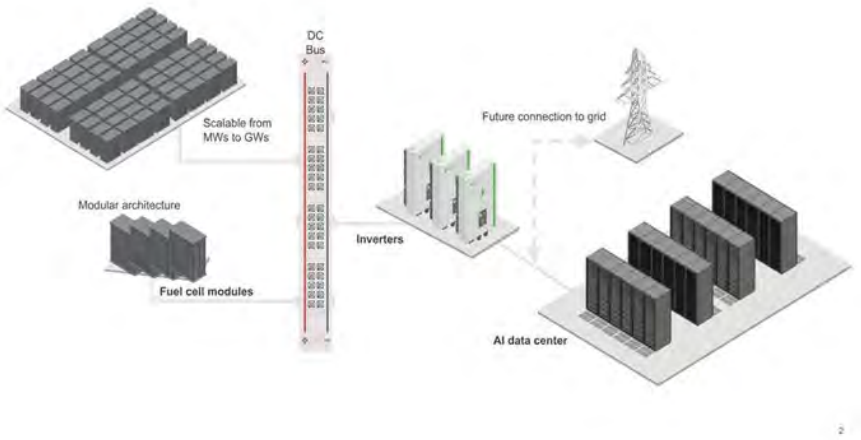
자료: Bloom Energy, 미래에셋증권 리서치센터

V. 에너지와 그리드: 모듈형 SOFC 사례 'Bloom Energy(BE)'

CEO "데이터센터가 직접 발전소를 짓는 것보다 더 빠르게 추가 용량을 제공할 수 있다"

- Bloom Energy(BE)는 "우리는 AI의 속도에 맞춰 움직일 준비가 됐다"는 메시지를 강조. 또한 BE의 CEO는 AI 시대의 전력난이 일시적 유행이 아닌 구조적 추세라면서 모듈형 SOFC가 Off-grid/On-site 수요에 적합하다고 강조. 실제로, BE는 오라클의 DC 인프라인 OCI에 연료전지를 구축. BE가 하이퍼스케일러와 직접 맺은 첫 계약. 주목할 것은 전력 공급까지 '약 90일 이내'라는 계약인 것.
- 그리드 병목의 가장 현실적·신속한 대안임을 증명해, 다른 고객사 계약으로 확산될 가능성. BE의 CEO는 다른 주요 DC 개발업체들과 전력 계약을 협의 중이라고 밝힘. BE는 AEP/AWS/Cologics와 100MW 규모 프로젝트를 진행 중(100MW를 20MW 단위로 쪼개 여러 지역으로 이동할 수 있는 모듈화)이며 총 900MW 파이프라인을 확보.
- 연료전지가 데이터센터 주전원으로 쓰이는 흐름이 가시화되고 있는 것이라 해석. BE는 수주 확대에 맞춰 제조능력을 1GW/년 → 2GW/년(2026년 말)로 두 배 증설하고 있음. 회사측은 2GW는 시작일 뿐이며, 향후에도 지속적으로 설비를 늘릴 것이라고 호언장담.

에너지 서버와 데이터센터의 전력 연결



자료: Bloom Energy, 미래에셋증권 리서치센터

Bloom Energy가 말하는 가스터빈과 연료전지의 비교

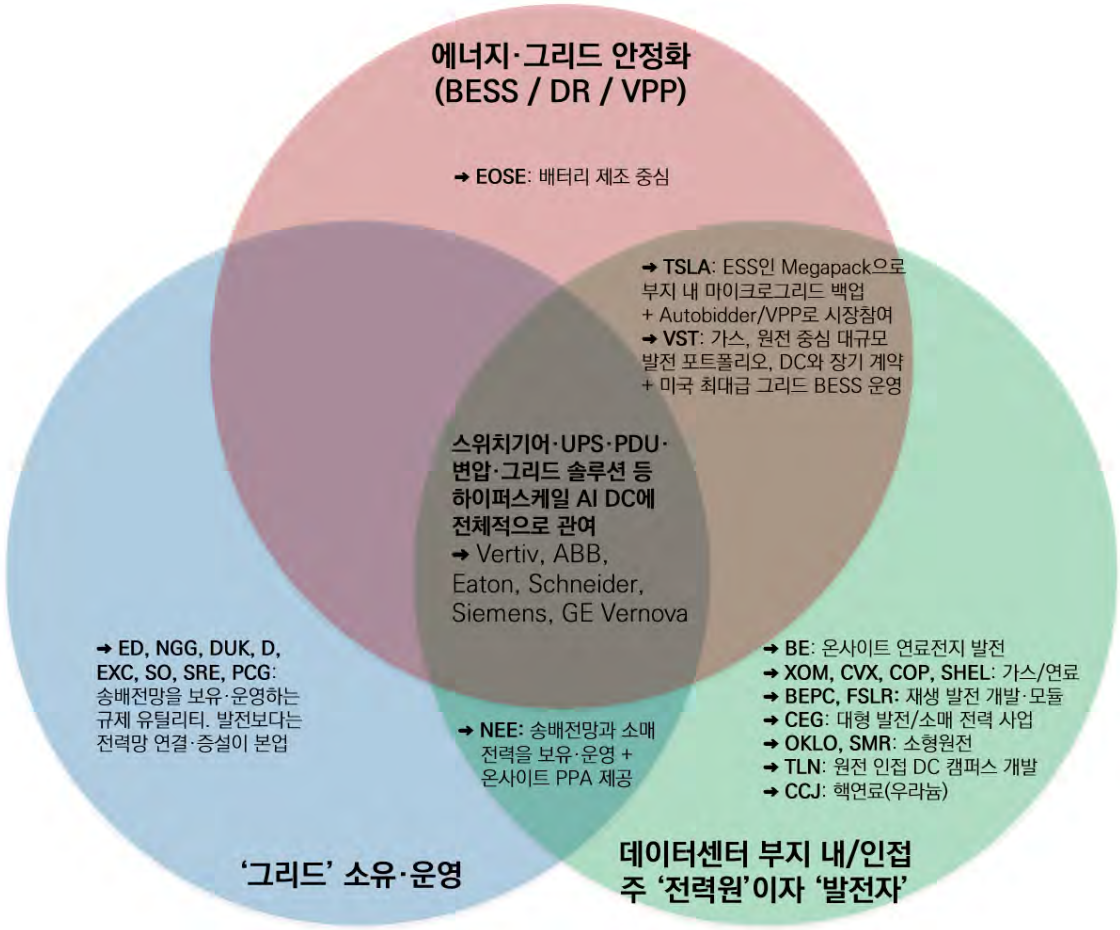
비교 항목	가스터빈	연료전지 (블룸에너지)
가용성 및 안정성	하나의 장치가 작동하려면 예비용 추가 장치가 필요함	**모듈식 아키텍처로 99.999% 이상의 매우 높은 가용성 달성 가능.**
초기 투자 비용 (CAPEX)	연료전지와 동등하거나 더 불리함.	**가스터빈과 비교하여 유리하거나 최소한 동등함.**
운영 비용 (OPEX)	연료 소비량이 15~20% 더 많아 운영비가 높음.	**연료 효율이 높아 운영비 측면에서 상당한 이점.**
환경 및 허가	대기 오염 물질을 배출하여 인구 밀집 지역에서는 허가 취득이 매우 어려움.	**대기 오염이 전혀 없어 허가 절차가 더 쉬움.**
전력 공급 시간	허가 절차의 번거로움으로 인해 시간이 더 오래 걸릴 수 있음.	**허가 절차가 단순하고 빨라 전력 공급까지의 시간이 짧음.**

자료: Bloom Energy, 미래에셋증권 리서치센터

[요약] AI DC의 전력 인프라와 핵심 승부처

AI 팩토리의 폭발적 수요가 “전력 연결·안정성·Time-to-power”라는 세 가지 병목을 만들었다

현장 발전(Time-to-power 해결) + BESS 및 VPP/DR(그리드 안정성) + 고밀도 냉각/전력 인프라(랙당 kW↑가 구조적 성장)가 수혜 가능성



자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전)

AI 동맹

오픈소스

화웨이

VI. 소버린 AI(신냉전): 초강대국간 초지능 경쟁

Sovereign AI: 새로운 지정학과 G2 전쟁의 서막

- AI는 국가 안보와 직결된 전략 자산. 각국은 자국의 데이터와 문화에 기반한 'Sovereign(주권적) AI' 확보에 사활을 걸고 있으며, 이는 미중 패권 경쟁의 새로운 전선이 됨. OpenAI는 "미국이 중국과의 초지능 경쟁에서 반드시 이겨야 한다"는 경제 청사진도 제시.
- OpenAI는 펜타곤과 2억 달러 규모의 방산 AI 개발 계약 체결했고, 엔비디아는 본인들의 AI 행사인 GTC를 올해 하반기 워싱턴 DC에서도 개최하기로 결정. AI와 외교/국방 전략이 같이 간다는 방증.
- 트럼프는 동맹국들에게 美中 사이의 선택을 강요하고 있는데, AI에 있어서 '미국 거를 쓰느냐 중국 거를 쓰느냐'의 이지선다는 더욱 노골적으로 전개될 것.
- 젠슨 황은 '소버린 AI' 개념을 전파하며 각국의 AI 팩토리 구축을 지원. 이는 미국의 에너지 수출과 연계된 '아메리칸 에너지' 특사 역할까지 수행할 가능성 시사. 사우디, UAE 등은 '페트로달러' 대신 '페타플롭'의 시대를 미국과 함께 열면서 AI에 막대한 투자 중.

OpenAI가 올해 1월 13일에 내놓은 보고서에서 중국의 국가적 AI 역량에 대해 평가한 부분

항목	세부 내용
중국 AI 생태계 현황	– 230개 이상 LLM 개발, 180개 이상 정부 승인. – 주목할 만한 모델로는, 특히 DeepSeek-Coder-V2는 평가에서 GPT-4 Turbo 능가 – AI 기술 발전 속도가 빠르고 중국 AI 기술력이 이미 상당한 수준 (GPT-4o 수준)
중국 AI 인프라 구축의 특징	– 전력: 5개 지역 11기 원자로 승인 (최소 2,200억 위안 투자, 5년 소요 예상), 지난 2년간 매년 10기 승인 – AI 파일럿 구역: 17개 AI 파일럿 구역 지정 (재정 지원, 우호적 규제) – 컴퓨팅: "Eastern Data and Western Computing" 프로젝트 (8개 컴퓨팅 허브) – 디지털 실크로드: 일대일로 일환, 개발도상국에 AI 확산 (2000~2017년 64개국 155개 프로젝트 지원) – 인재 양성: "루반 워크숍" 직업 훈련 (2016~2023년 25개국 30개 워크숍)
중국 AI 강점	– 방대한 데이터: 인구 – AI 모델 학습에 유리한 환경 – 정부의 강력한 지원: 국가 전략 산업 육성, 정책/재정 지원 – 거대 IT 기업: 바이두, 알리바바, 텐센트 등 – 빠른 실행력: 중앙 집중식 의사 결정 체계
중국 AI 약점	– 기초 연구 부족: 응용 연구/상용화 대비 취약 – 편향성: 중국 내 데이터 편향, 일반화 성능 저하 가능성 – 글로벌 시장 진출 어려움 – 윤리적 문제: 사회 통제, 감시, 억압 활용 비판 – 국제 사회 신뢰 저하, 협력 어려움
미국의 대응 전략	– 연구 개발 투자 확대: 기초 연구, 핵심 기술, 혁신 촉진 – 인프라 구축: 데이터 센터, 5G, 고성능 컴퓨팅 – 인재 양성: 교육 프로그램, 해외 인재 유치 – 윤리: 민주적 가치 기반, 책임감 있는 개발/활용 – 국제 협력: 동맹국/파트너 협력, 글로벌 생태계 조성, 규범 형성

자료: OpenAI, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 미국식 AI를 글로벌 표준으로 세우겠다는 트럼프

'규제 걷어내고, 인프라·수출 밀어붙여 미국이 AI 패권 잡겠다'

- 백악관은 7월 23일 "Winning the AI Race: America's AI Action Plan"라는 제목의 AI 핵심 로드맵을 제시. 연방기관 90개 이상의 모든 공무원이 최첨단 AI 모델을 사용할 수 있도록 의무화. 기존 정부 레퍼런스가 있는 AI 관련 기업들이 대표적으로 수혜입을 것으로 예상. 대표적인 예가 팔란티어.
- 액션 플랜 중 가장 화두는 "AI를 우주 경쟁 수준의 안보 과제로 규정"하고 미국의 절대 기술 우위를 요구한 것. 또한, 바이든 행정부 시절의 "AI 안전 행정명령"을 폐지. 전력·데이터센터 허가 단순화로 건설주·유틸리티·GPU 공급망 수혜 기대. 특히 환경규제 면제와 연방 토지 사용 허가로 데이터센터 허가를 신속 처리할 것. 물론 주별로 AI법이 따로 있어 이와 충돌 가능성이 있는데 규제(DEI, 기후변화)가 많은 주(특히 캘리포니아)에 대해서는 자금 지원을 삭감한다고 압박. AI 친화적인 州에 자원 집중 의도.
- 모델·GPU·네트워크·SW 등 풀스택 AI 패키지에 관한 수출을 더욱 통제. 칩 자체에의 위치 확인 기능 탑재 등이 이에 해당. 이로써 대만·한국·UAE 등 동맹국과의 거래는 촉진, 중국·러시아에는 통제 강화.
- 오픈소스 및 오픈웨이트를 권장하는 것도 포인트. 미국적 가치에 기반한 개방형 모델 촉진한다는 것인데, 지난 8월 초 OpenAI가 오픈소스 모델 gpt-oss를 출시한 것도 이와 같은 흐름으로 판단.

백악관 과학기술정책실이 발표한 "미국의 AI 실행 계획"
규제 완화 + 데이터센터 조기 확충 + 동맹국에는 AI 수출, 중국은 견제

자료: 백악관, 미래에셋증권 리서치센터

트럼프 정부와 바이든 때의 핵심 AI 정책 비교
'핵무기와 같은 파괴력을 지닌 비대칭적 전력이 AI니까 규제 싹 풀자'

주요 안건	트럼프의 AI 액션 플랜	바이든 정부 때 행정명령
국가안보	"우주경쟁 급" 최우선과제 DOD AI 가상시험장	상무·DHS·DOD에 위험평가·표준화 위임
규제 접근	**규제 일괄 폐지 규제 많은 주정부 압박**	**안전·프라이버시 규범 신설**
인프라·에너지	**환경심사 면제 발전소 폐쇄 금지**	**DOE의 AI-그리드 혁신 펀드**
수출통제	대중 규제 강화 칩 추적 의무	BIS가 '사전신고+리뷰' 체계
개방형 모델	오픈소스 장려	국립 AI연구자원 확대

자료: 백악관, 미래에셋증권 리서치센터

[참고] 미국의 AI Action Plan

미국의 플랜을 보면 AI에서 뭐가 앞으로 중요한지 보인다

미국을 인공지능 분야의 글로벌 리더로 자리매김하기 위한 포괄적인 미국 연방 전략인 "America's AI Action Plan"의 주요 내용
AI를 이용한 전통산업의 부흥과 AI SW의 수출이 수혜 받을 것으로 예상

Pillar I.
AI 혁신 가속화

- **불필요한 절차 및 과도한 규제 철폐**
- 첨단 AI가 표현의 자유와 미국의 가치를 보호하도록 보장
- **오픈소스 및 오픈웨이트 AI 장려**
- AI 도입 활성화
- AI 시대의 미국 노동자 역량 강화
- **차세대 제조업 지원**
- AI 기반 과학에 대한 투자
- **세계적 수준의 과학 데이터 세트 구축**
- AI 해석 가능성, 견고성 혁신에 대한 투자
- AI 평가 생태계 구축
- **정부 내 AI 도입 가속화**
- **국방부 내 AI 도입 추진**
- 상업 및 정부 AI 혁신 보호
- 법률 시스템 내 합성 미디어 대응

Pillar II.
미국 AI 인프라 구축

- 데이터 센터, **반도체 제조 시설 및 에너지 인프라에 대한 허가 절차 간소화** 및 보안
- **AI 혁신 속도에 맞는 전력망 개발**
- **미국 반도체 제조업 복원**
- 군 및 정보기관용 고보안 데이터 센터 구축
- AI 인프라를 위한 숙련된 인력 양성
- **핵심 인프라 사이버 보안 강화**
- 보안 중심 설계 AI 기술 및 앱 장려
- AI 사고 대응을 위한 연방정부의 역량 증진

Pillar III.
국제 AI 외교 및 안보 선도

- **동맹국 및 파트너에게 미국 AI 수출**
- **국제 거버넌스 기구에서 중국 영향력 대응**
- **AI 컴퓨팅 수출 통제 집행 강화**
- 기존 반도체 제조 수출 통제의 허점 보완
- 전 세계적으로 보호 조치 조정
- 국가안보 위험평가에 있어 미국 정부의 선도적 역할 보장
- **바이오 보안**에 대한 투자

자료: 백악관, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 유럽도 AI한다! 근데 이제 미국 거를 곁들인…

무게감이 줄었지만, 유럽도 깨어나기 시작

- 지정학적 체스판에서 중대한 한 축인 유럽은 전세계 AI 슈퍼컴퓨터 성능에서 지리적으로 매우 미미. 미국은 75%를 차지하며 지배적이고, 중국이 15%로 2위를 차지하는 것과 비교할 때 무게감이 심각.
- 그런 유럽에서도 컴퓨팅 확보 관련한 소식이 속속 들려오는 중. 프랑스의 마크롱 대통령은 "국가 주도 AI 투자"를 하지 않으면 유럽이 죽을 수도 있다고 언급. 아프다고 말한 정도가 아니라, 투자를 하는 것이 말 그대로 생사의 문제라고 한 것.
- 한편, 주요 클라우드 세 곳(MS·오라클·구글)은 2025년 5월 초 공동 의견서를 제출하면서, 동맹국에 대한 AI 공급망을 확충하는 게 미국이 AI 생태계를 리드하는 핵심이라고 밝혔음. 앞으로도 미국은 동맹국인 유럽의 소버린 AI 구축을 위해 하드웨어부터 소프트웨어 툴체인까지 서비스 수출을 지속하게 될 것. 이와 대표적으로 수혜를 받는 곳 중 하나가 엔비디아.
- 독일의 라이프니츠 슈퍼컴퓨팅 센터(LRZ: 바이에른 과학 아카데미 산하의 세계 최고 수준의 슈퍼컴퓨팅 센터)에 엔비디아는 본인들의 차세대 시리즈인 'Vera Rubin(Vera는 CPU, Rubin은 GPU)'을 공급하게 됨. 27년 가동 목표인 이 슈퍼컴퓨터는 엔비디아의 차세대 제품의 데뷔 쇼케이스를 선확보한 셈.

올해 2분기부터 잇따라 발표된 유럽의 AI 데이터센터 & 그리드 확충 노력 & 관련 엔비디아 제품
프랑스는 원전이 핵심 / 스페인과 노르웨이는 재생에너지에 중점 / 이탈리아는 '소버린 AI' 슈퍼컴퓨터 증설에 열중

시점 · 발표 주체	국가	프로젝트/정책	설비 용량·전력	전력망(그리드) 확충·연계 방안	엔비디아 협력·칩 스펙
6월 엔비디아 (GTC Paris)	독일	"Industrial AI Factory" (뮌헨) – BMW · 지멘스용	60 MW	바이에른州 110 kV 라인 증설 재생 PPA, 폐열 지역난방 연결	B200, RTX Pro 등 1만 개 GPU
6월 엔비디아 · Mistral	EU	Sovereign LLM Cloud Sites	1단계: 18,000 GPU 2단계: 28,000 GPU	현지 CSP별 재생전자산 + 국가 그리드 패스트트랙	B100
6월 Telefónica	스페인	"Edge AI Fabric" – 전국 400 노드	100 MW	재생 전력 비율 80%; REE 110 kV 노드 병행	DGX H200 SuperPOD 500개
6월 Telenor	노르웨이	재생에너지 데이터센터	50 MW	수력 100%, Statnett 420 kV 루프 증설	Grace Hopper 9,600개 GPU
6월 Orange	프랑스	"Cloud Avenue" GenAI Zone	36 MW	RTE 225 kV 신규 서브스테이션	DGX B200 SuperPOD
6월 Fastweb	이탈리아	MIIA 슈퍼컴 (밀라노)	24 MW	Terna 150 kV 이중화 + 태양광 PPA	DGX H200 GPU 6,400개
5월 프랑스 정부·Mistral AI	프랑스	파리 1.4 GW "Mistral AI Cluster" – 프랑스 정부와 UAE MGX 공동투자	1.4 GW	EDF SMR pilot 2기 전력 직결 4×400 kV 변전소 증설	B200 및 GB 슈퍼칩 1.8만 개

자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 트럼프의 중동 순방길… “페타플롭스 시대”

트럼프 정부의 실리콘 외교

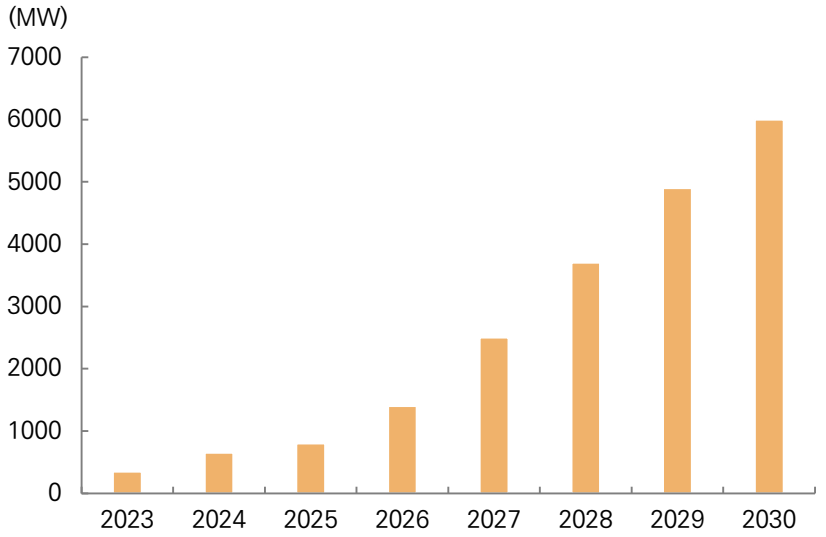
- 트럼프 미국 대통령은 집권 2기 첫 공식 해외 순방지로 사우디아라비아를 선택. 트럼프 행정부의 AI 정책 설계자라고 볼 수 있는 데이비드 삭스는 트럼프의 중동 순방길을 글로벌 AI 경쟁 구도를 재편하는 ‘실리콘 외교’의 시작으로 평가. 삭스는 AI 컴퓨팅 파워가 달러나 금처럼 국가의 동맹을 결정하고 영향력을 행사하는 핵심적인 전략 자산이 되었음을 선언.
- 이 같은 외교 원칙에 따라, 미국은 AI 혁명에 동참하려는 국가들에게 파트너십이라는 문을 활짝 열어주고 있음. 단순히 기술을 판매하는 것을 넘어, 미국의 하드웨어와 소프트웨어 생태계 안으로 동맹국들을 깊숙이 편입시키는 과정.
- 과거 석유 거래를 달러로 결제하게 하며 달러 패권을 구축한 ‘페트로달러’처럼, 이제는 엔비디아 GPU를 중심으로 한 AI 인프라와 GPU 확보를 간절히 원하는 각국의 소버린 AI 기조를 통해 미국의 기술 패권을 공고히 하려는 ‘페타플롭스’ 시대가 열린 것.

백악관이 발표한 "트럼프 효과" 중 천 억 달러 이상의 거래

투자 주체	투자액(달러)	세부 분야
UAE 국부펀드 등	1.4조	AI 인프라 및 국방 등
카타르 국부펀드 등	1.2조	AI 인프라 및 국방 등
사우디아라비아 국부펀드 등	6,000억	AI 인프라 및 국방 등
소프트뱅크·OpenAI·Oracle	5,000억	AI 인프라(Project Stargate)
엔비디아	5,000억	AI 인프라·슈퍼컴퓨터
Apple	5,000억	아이폰 제조, AI 훈련
DAMAC Properties	2,000억	AI 데이터센터 확장
IBM	1,500억	AI 제조
TSMC	1,000억	반도체 공장(애리조나 피닉스)

자료: 백악관, 미래에셋증권 리서치센터

중동 지역의 데이터센터 용량 연간 전망



자료: SemiAnalysis, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 실리콘 외교로 미국이 얻을 것은 글로벌 패권

지난 5월 13일 시작된 트럼프의 중동 순방

- 미국과 중동의 페타플롭스 동맹은 경쟁국인 중국을 효과적으로 고립시키는 다목적 카드로 작용. 미국과 UAE간 "GPU 거래"는 5월 순방길에 합의를 봤으나 7월 들어 해당 UAE 기업이 화웨이와의 협력한 전례가 도마에 올라 승인이 보류된 상태. 이런 'GPU 수급 지연'이 중동에 주는 메시지는 명확.
- 한편, 현재 화웨이는 본인들의 자체 GPU라 불리는 Ascend 칩을 중국 외부에 대량으로 판매할 생산 능력이 부족. 미국은 바로 이 시점을 활용하여 중동이 미국산 하드웨어와 소프트웨어를 먼저 채택하게 함으로써, 미래에 중국 기술이 발붙일 시장 자체를 원천적으로 차단하는 것.
- 또한, 중동이라는 거대한 수요처가 생겨남으로써, GPU가 초과 공급이 될 가능성을 줄이고, 중국으로 불법 재수출될 수도 있는 GPU 초과 생산 물량의 통로를 자연스럽게 차단하는 완충 역할도 하게 됨. 이는 중국의 AI 발전에 필수적인 하드웨어 확보 경로를 더욱 좁히는 효과적인 제재 수단이 됨.
- 결국 트럼프의 중동 순방은 화웨이의 잠재적 거대 시장을 잃게 만드는 동시에, 제재를 우회할 경로마저 차단하는 이중고를 중국에 안겨준 셈. 이는 미국의 기술 패권을 강화하고 동맹을 넓히는 동시에, 중국의 AI 굴기를 실질적으로 늦추는 강력한 지정학적 한 수라고 평가할 수 있음.

역사상 가장 큰 양자간 거래로 기록된 미국과 사우디간 주요 투자 계약 관련 세부 내용 – 미국 AI 기업들에게 매우 긍정적이라고 평가
사우디아라비아는 석유 의존도를 낮추고 첨단 기술, 인프라, 헬스케어 등 다양한 분야로 경제 구조를 전환하는 것이 최대의 목표

분야/항목	주요 기업/주체	금액	세부 내용/위치	투자 성격
AI 데이터센터 및 에너지	사우디 DataVolt	200억 달러	DataVolt는 2025년 Supermicro와 함께 미국·사우디에 초대형 AI 데이터센터를 구축 예정	사우디의 대미 투자
첨단 기술	사우디 DataVolt, 구글, 오라클, 세일즈포스, AMD, 우버 등	800억 달러	양국에서 '변혁적 기술'에 총 800억 달러를 투자하기로 약속	양국 공동 투자
인프라 건설	Hill International, Jacobs, Parsons, AECOM 등 (미국 기업)	20억 달러	건설 컨설팅 기업 Hill International, Jacobs, Parsons, AECOM이 킹 살만 국제공항, 킹 살만 공원, 더 볼트, 키디야 시티 등 주요 프로젝트에 참여	미국의 서비스 수출
에너지 솔루션 (수출)	GE Vernova (티커: GEV)	142억 달러	GE Vernova의 가스터빈 및 에너지 솔루션이 주요 수출 품목	미국 제품 수출
항공기 (수출)	보잉, AviLease	48억 달러	보잉 737-8 항공기를 AviLease가 구매하는 계약	미국 제품 수출
헬스케어	Shamekh IV Solutions, LLC	58억 달러	헬스케어 기업 Shamekh IV Solutions가 IV(정맥 주사용) 용액의 고용량 생산 시설을 미국 미시간 주에 건설	대미 투자

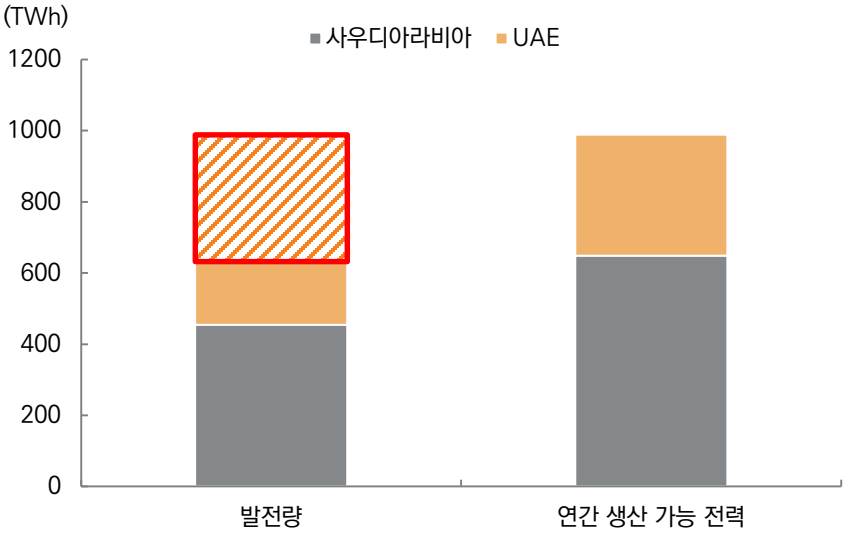
자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 실리콘 외교로 미국이 얻을 것은 전력 인프라

중동을 바라보는 관점 : 전력 인프라

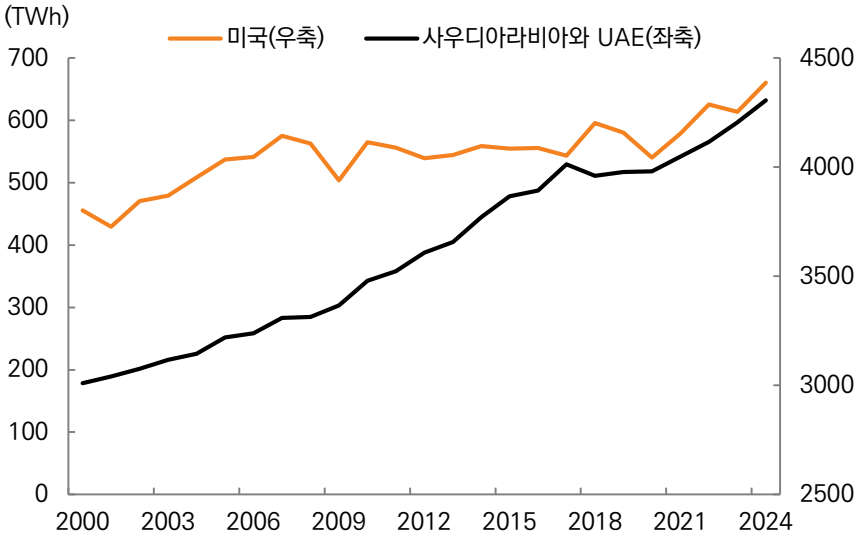
- 미국 AI 산업의 가장 큰 장기적 위협은 자본이나 칩이 아닌, 성장이 정체된 자국의 전력 인프라. AI 팩토리가 요구하는 수 GW의 전력을 미국 내에서 단기간에 확보하는 것은 현실적으로 불가능에 가까움.
- 바로 이 지점에서 중동과의 AI 동맹은 미국의 에너지 병목 현상을 풀어낼 결정적인 열쇠. 미국의 AI 두뇌를 중동의 풍부하고 저렴한 에너지와 결합하는 것.
- 특히 UAE의 AI 업체인 G42 관련 이슈는 새로운 동맹의 작동 방식을 명확히 보여줌. 트럼프 행정부는 UAE가 연간 50만 개의 엔비디아 고성능 GPU를 2027년까지 수입할 수 있게 하는 '수출 쿼터'를 검토. G42가 엔비디아의 최고 성능 AI 칩에 대한 막대한 수입 할당량을 보장받고, 그 물량의 80%를 다시 마이크로소프트와 같은 미국 기업에 제공하게 되는 거래 구조.
- 美 하이퍼스케일러들의 컴퓨팅 확장을 제한하던 심각한 병목인 에너지 부족을 중동 인프라를 활용하여 해결하는 거래. 미국의 데이터센터 그리드가 물리적 한계에 도달한 시점에, 저렴한 에너지와 풍부한 자본을 갖춘 중동은 차세대 AI 허브 부상 가능.

사우디아라비아와 UAE의 발전량과 Capa 비교
연간 최대 생산 가능 전력을 보면 AI 인프라 관련 상당한 성장 가능성



자료: EMBER, 미래에셋증권 리서치센터

미국과 사우디아라비아(KSA)의 발전량 추이 비교



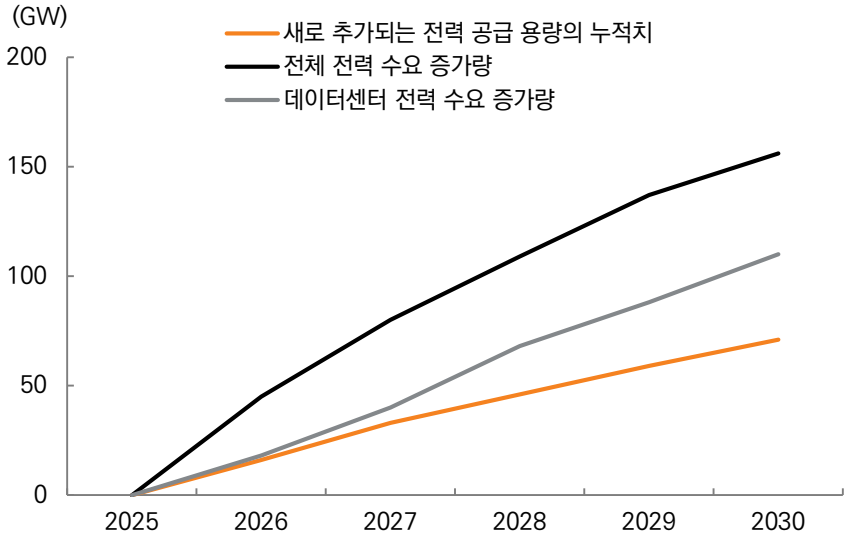
자료: EMBER, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 미국과 중국의 전력 인프라 격차

미국에겐 없지만 중국에겐 있는 것

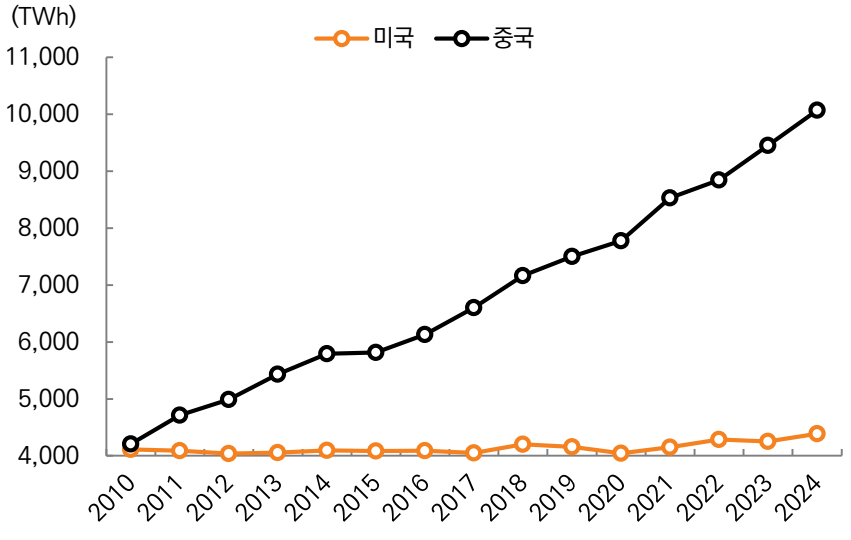
- 미국이 중동의 손을 잡은 주된 에너지 때문인 것을 감안하면, 서방 AI 발전은 전력 인프라의 한계라는 현실적 제약이 있다고 할 수 있음. 반면, 중국은 지난 10년간 美 전체 전력망에 해당하는 막대한 용량을 신규로 증설하며 AI 혁명을 위한 압도적인 전력 기반을 확보.
- 중국은 실리콘 방면의 기술적 열위를 극복하기 위해 압도적인 전력 인프라 투입이라는 전략을 사용. AI의 성능은 본질적으로 '에너지'와 '실리콘'의 함수라는 명확한 인식 아래, 기술 효율성 개선보다는 우선은 막대한 전력 비용을 투입하여 '단순무식'처럼 보이지만 압도적 '인해전술'로 정면돌파하려는 것.
- 중국은 풍부한 전력을 바탕으로, 전력 소모가 크더라도 시스템의 확장성(Scale-up)을 극대화할 수 있는 광학 인터커넥트와 같은 기술을 적극적으로 채택. 엔비디아 등 미국 업체 대비 갖는 열세를 거대한 물량과 인프라로 극복하려는, 오직 중국만이 실행할 수 있는 대담하고 독자적인 접근 방식.

암울한 미국의 전력 수요 시나리오
전력 공급 증가가 수요 증가를 따라가지 못하는 격차



자료: ENGINE NO.1, 미래에셋증권 리서치센터

미국 vs 중국 연도별 전력 발전량 추이 비교
2010년 이후 격차가 계속 확대돼 2024년 중국 발전량은 미국의 약 2.3배



자료: Our World in Data/Ember/Energy Institute, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 미국의 수출규제

미국의 對중국 칩 규제

- 2025년 4월, 트럼프 행정부는 GPU에 대한 새로운 對중국 수출 허가 요건을 적용하며 엔비디아의 중국 사업에 결정타를 날림. 이는 바이든 정부 시기의 규제에 맞춰 합법적으로 수출하기 위해 특별히 설계된 H20마저 판매할 수 없게 만드는 조치. 갑작스러운 정책 변경으로 엔비디아는 막대한 H20 재고를 처리할 수 없게 되었고, 결국 1분기에만 45억 달러의 손실을 기록하는 등 총 55억 달러에 달하는 비용을 감수해야만 했음.
- 젠슨 황은 “4년 전 95%에 달했던 중국 시장 점유율이 현재 50%로 반토막 났다”고 아쉬워하면서도, 2026년 500억 달러 규모로 예상되는 거대한 중국 AI 시장이라는 기회를 놓치고 있다는 것에 다소 비판 모드에 돌입. 엔비디아는 지난 6월, 앞으로의 분기 매출 전망에서는 중국 수치는 완전히 배제하겠다고 공식 선언.
- 반대로, 중국 내 “화웨이 다음의 AI 칩” 제조업체로 불리는 Cambricon은 전년 대비 매출이 43배 증가.

미국의 칩 수출 규제 관련 타임라인 (십억원, %, 배)

날짜	주체	내용
8월 초	미국 트럼프 정부	엔비디아·AMD에 첫 번째 H20/MI308 선적 라이선스 발급 개시(총 20,000개 한도, 3개월 유효).
7월 15일	미국 트럼프 정부	트럼프 행정부가 엔비디아 H20 칩의 對중국 판매를 다시 허가.
6월 13일	엔비디아	미국의 수출 제한 조치로 인해 향후 매출 전망에서 중국 시장은 제외한다고 발표.
5월 28일	엔비디아	미국의 H20 AI 칩 수출 라이선스 요건으로 1분기에 45억 달러 비용으로 계상
5월 21일	중국 정부	미국이 화웨이 AI 칩 사용을 경고하자, 중국 상무부가 관련 조치 시행 시 법적 대응을 경고하며 양국 긴장 고조.
5월 13일	미국 트럼프 정부	이전 행정부의 '인공지능 확산 규칙'을 공식 철회하고, 화웨이 Ascend AI 칩 사용에 대한 경고 유지.
4월 15일	미국 트럼프 정부	트럼프 행정부, 엔비디아 H20 AI 칩에 대한 수출 허가 요건 적용. 엔비디아는 1분기 관련 비용을 55억 달러로 추산.
4월 9일	엔비디아	젠슨 황 CEO가 트럼프 전 대통령과 만찬, H20 칩 수출 제한 면제 가능성 보도됨.
2월 3일	미국 정치인	일부 상원의원, 상무부 장관 내정자에게 AI 칩 수출 제한 강화를 촉구하는 서한 발송.
1월 27일	DeepSeek	중국의 DeepSeek이 개방형 '추론' 모델을 공개하며 업계에 큰 반향을 일으킴.
1월 13일	미국 바이든 정부	바이든 임기 종료 직전, 미국산 AI 칩에 대한 새로운 3단계 수출 제한 조치 제안.
1월 6일	Anthropic	다리오 아모데이 CEO, 기존 AI 칩 수출 규제를 지지하며 차기 정부에 추가 규제 강화를 촉구.

자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 미국의 수출규제

딜런 파텔이 말하는 미국의 수출 통제 유효성

- SemiAnalysis의 편집장 딜런 파텔은, "미국의 칩 단위로 이뤄지는 대중국 반도체 통제는 오히려 중국의 기술 자립을 가속하는 역효과를 낳고 있다"고 말한 바 있음. 엔비디아나 AMD가 중국 내에서 생산하는 칩보다 더 나은 칩을 판매할 수 없게 되면서, 화웨이의 Ascend 칩이 자국 내에서 경쟁자 없이 성장할 수 있는 환경이 만들어진 것.
- 지난 5월, 엔비디아는 현지 정부와 협력하면서 상하이에 R&D 센터 설립까지 추진. 이는 엔비디아가 중국의 방대한 CUDA 기반 AI 개발자 커뮤니티와의 관계가 단절될 것을 우려한 조치.
- 이에 대해 젠슨 황은 "미국의 과도한 기술 통제가 화웨이와 같은 중국 기업에 오히려 기회를 줄 수 있다"는 메시지를 워싱턴에 계속해서 던짐. 그는 '중국에 (1~2세대 뒤처진) 칩을 제공하는 것은 미국에 위협이 되지 않고, 오히려 중국 개발자들의 CUDA 종속을 유지해 미국의 우위를 지킬 수 있다"고 주장.

중국 AI 개발 생태계가 CUDA에 얼마나 종속되어 있는지 알 수 있음
CUDA 의존도는 여전하고, 중국의 계획을 보면 CUDA를 계속 쓰고 싶음

지표	수치·내용
중국 CUDA 개발자 규모	**약 150만 명**
엔비디아–중국 스타트업 협력	**3,000개 이상 스타트업과 협업 중**
중국에 이미 반입되어 있는 것과 밀수·재수출로 유입된 '잔존 재고' 물량	A100·H100·H800·B100·B200 등 고성능 GPU 약 25,000개(美 정부 추산)
중국 지방정부의 신규 데이터센터 계획에 명시된, 필요 물량	H100·H200 등 115,000개 목표치 (신장·칭하이 등 39개 프로젝트)
대체 생태계 부상 지표	화웨이가 CUDA의 대체재인 "CANN"의 툴킷 오픈소스 전환 발표

자료: 미래에셋증권 리서치센터

Hopper(H20도 Hopper 세대) 대비 차세대 아키텍처가 갖는 이점
젠슨 황 "Hopper는 이제 줘도 안가지게 될 것이다"

세대	성능 향상 (AI Inference 중심)	비용 감소 (TCO와 에너지 소비 관점)
Blackwell	68배	87%
Rubin	900배	99.97%

자료: 엔비디아, 미래에셋증권 리서치센터

엔비디아의 對중국 수출 GPU 라인업과 각각의 주요 스펙 사항

GPU (출시 세대)	수출 규제	연산 성능(BF16 dense 기준)	메모리 종류·용량	메모리 대역폭
H100 (SXM)	수출 금지	989	80 GB HBM3	3.35 TB/s
H800	수출 금지	989	80 GB HBM3	1.68 TB/s
B100	수출 금지	1800	192 GB HBM3e	8 TB/s
H20	사실상 허용	148	96 GB HBM3	4.0 TB/s
RTX PRO 6000D (B40)	허용 예상	"H20의 75% 성능"	96 GB GDDR7	1.8 TB/s
중국 수출규제 상한선	**칩의 단위 면적당 연산 성능이 규제 핵심 지표**		**대역폭 상한선 1.7 ~ 1.8 TB/s**	

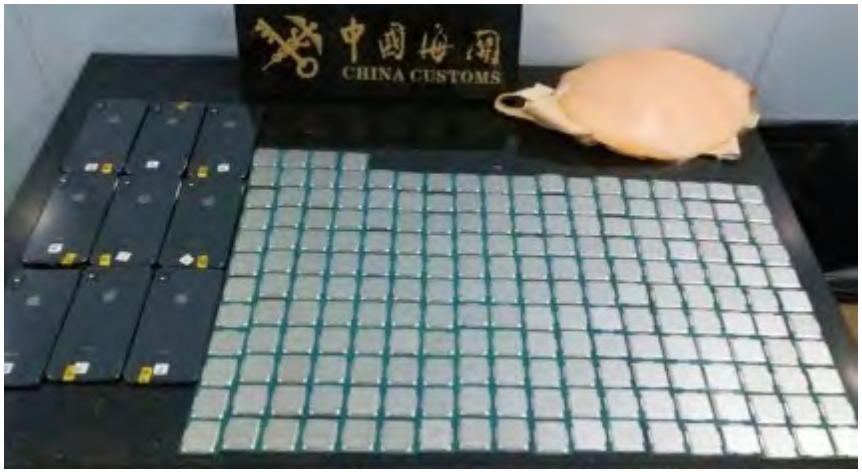
자료: 미래에셋증권 리서치센터

[참고] 중국은 미국의 대중 수출 통제를 어떻게 무력화 하나?

미 법무부,
"적발된 GPU 밀수범들은
수천만 달러를 벌었다"

- 250만 명 이상의 구독자를 둔 세계적인 IT 개인 미디어 "Gamers Nexus"의 편집장 스티브 버크가 촬영한 밀수/관세 다큐멘터리에 따르면, 미국 정부의 수출 통제 정책이 기상천외한 방법들이 동원된 '현장'에서 어떻게 무력화되고 있는지 여실히 알 수 있음. 실리콘 암시장이 실제로 존재하며 매우 조직적으로 운영되고 있음을 증명하는 물리적 증거물이 잇따라 나와 엔비디아의 공식 입장과 암묵적 현실 사이에 얼마나 큰 간극이 존재하는지를 시사.
- 스티브 버크의 말 중 가장 충격적인 것은 중국 내 "GPU 수리점"의 존재. 이들은 고장난 "수출 통제 GPU"를 수리는 물론, 심지어 메모리 용량을 증설하는 개조까지 수행할 수 있음. 미국으로부터 제한적으로 유입되는 고성능 실리콘의 수명을 연장시켜, 중국 내 고성능 GPU pool이 고갈되지 않도록 하는 핵심적인 역할을 의도치 않게 담당. 즉, 미-중간 밀수 생태계의 지속 가능성을 높여 줌.
- 또한, RTX 5090과 같은 최신 소비자용 GPU는 중국에서 부품(PCB 기판, 쿨링, 커패시터, 전압 조절기) 조달 및 조립되기 때문에, 필연적으로 수출 통제가 실패할 수 밖에 없는지를 어느 정도 설명. 중국에서 조립되는 제품 중 "폐기된" 제품이나 부품이 어떻게 처리되는지는 아무도 추적할 수 없기 때문.

중국이 미국산 칩을 불법 입수하는 과정 ①"가짜 임산부 배"를 이용
미국 수출 통제를 회피하려는 밀수꾼들의 대담하고 창의적인 수법



자료: 중국 세관, Gamers Nexus, 미래에셋증권 리서치센터

중국이 미국산 칩을 불법 입수하는 과정 ②랍스터 상자에 GPU 숨김
개인이 아닌 조직적인 밀수 공급망이 존재함을 암시

자료: 중국 세관, 知乎(@艾小白), Gamers Nexus, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 미국의 수출규제

트럼프의 문단속:
"중국에 칩을 그렇게 팔고 싶으면 파세요. 그런데 돈은 내시구요, 좋은 건 못 팝니다."

- 트럼프 행정부는 중국과의 무역 협상 카드로 엔비디아 H20 칩의 중국 판매를 조건부로 재개. 하워드 루트닉 상무부 장관은 이번 결정이 중국의 희토류 원소 공급 재개와 맞물린 협상의 일환임을 공식화.
- 이번 조치는 4월에 부과된 전면적 수출 금지 조치를 해제하는 것이지만, 단순한 원상 복귀는 아님. 바이든 행정부 시절보다 훨씬 엄격한 '사전 고지 및 허가' 요건을 충족해야만 판매가 가능. 게다가 엔비디아는 AMD와 함께 중국向 AI 칩(H20·MI308) 판매 대가로 발생하는 매출의 15%를 미국 정부에 납부해야 하는 '수익 공유'를 해야 함. 과거 전례를 찾기 힘든 이례적 사례…
- 보다 중요한 점은, 미국은 기술 우위를 포기하지 않는 선에서, 미국 기업(엔비디아)이 중국 내에서 화웨이와 경쟁할 수 있는 최소한의 여건을 마련해주려는 맞춤형 조치라는 것. 결과적으로 엔비디아의 H20 칩의 재고를 다시 수출할 수 있게 되어 이전에 발생했던 수 조원의 손실을 만회할 기회를 얻게 됨.
- 새롭게 마련된 '사전 허가'라는 H20 관련 가이드라인으로 중국 AI 시장의 점유율을 빠르게 다시 확보할 수 있을 것으로 사료됨. 더불어 9~10월에 출시 예정인 B40(RTX 시리즈의 변동)의 수출 또한 덤.
  - 중국 빅테크(알리바바·텐센트·바이트댄스 등) 누적 주문 180만 장이나 현재 기보유/예정 재고는 90만 장에 불과
  - 독일 등 유럽 시장에서도 RTX 4090의 가격은 최근 40% 이상 급등하며 공급 부족이 심화되는 상황

미국의 엔비디아 칩 수출과 중국의 희토류는 협상의 차원이기도 함
희토류 허가가 다시 조이질 경우 H20도 재차 중단 가능

시점	조치	미국행 희토류 자석 수출량
2025-04	中 상무부, 네오디뮴(NdFeB) 자석, 합금 수출 허가제 도입	246.3t
2025-05	첫 달 적용 → -93.3% YoY 급감	46.4t
2025-06	미-중 실무협상 후 허가증 발급 속도 정상화	353t (▲660% MoM)

자료: 로이터, 미래에셋증권 리서치센터

중국 수출용 Blackwell 아키텍처 기반 B40 스펙
HBM이랑 NVLink 안쓰고 저렴하게 내놓아서 화웨이 추격의 싹을 자른다

항목	내용
제품명	B40 (RTX Pro 6000D로도 언급됨)
기반 아키텍처	Blackwell이지만, 데이터센터급(GB 시리즈)은 아님
메모리	고가의 HBM 대신, GDDR7 메모리 채택
패키징 기술	Monolithic Design – TSMC의 CoWoS 패키징 기술을 사용하지 않음
GPU 간 연결	NVLink 미지원 – 대규모 다중 GPU 구성의 성능에 제약이 있을 수 있음
멀티 GPU 구성	PCIe 6.0 기반으로 연결될 가능성 – 최대 8개 GPU: ConnectX-8 SuperNIC을 통해 연결 – 8개 이상 GPU: Spectrum-X 네트워킹 플랫폼으로 확장 추정
예상 가격	**$6,500 ~ $8,000 – 화웨이 Ascend 910C보다 40% 이상 저렴. 현재 주력 모델인 H20($1만~$1.2만)보다도 50% 가량 낮은 가격.**
시장 반응	**엔비디아는 올해 100만 개 이상으로 생산 계획. 관련 부품사인 기가바이트의 5월 매출은 전년 대비 107.9% 폭증**

자료: Tom's Hardware, 미래에셋증권 리서치센터

[참고] 중국의 AI 전략

중국은 과거의 막연한 자신감과 달리, 위기감과 절박함 속에서 AI 자립을 향한 국가 총력전 태세로 전환

지난 2018년과 2025년에 각각 있었던 중국의 중앙정치국 집단 회의(AI가 주제)에서 느낄 수 있는 톤 변화

주제 (Topic)	2018년 때와 비교한 올해의 톤 변화	2025년 발언	2018년 발언
AI의 역할	거의 동일.	AI는 새로운 과학기술 혁명과 산업 변혁을 이끄는 전략 기술.	AI는 과학기술 혁명과 산업 변혁 라운드를 이끄는 전략기술.
주도권 유지	**강화. 중국은 거버넌스에서도 주도권을 추구. 중국이 규칙 제정자가 되려는 의지를 시사.**	**AI 발전과 거버넌스에서 주도권을 확고히 장악.**	**새로운 세대의 AI 발전을 가속화하는 것은 우리가 글로벌 기술 경쟁에서 주도권을 잡기 위한 중요한 전략적 발판.**
건강한 발전	강화. 공산당은 무분별한 AI 발전을 우려한다는 점을 시사하는 구체적인 조건이 추가됨.	중국 AI의 건강하고 질서 있는 발전을 유익하고, 안전하며, 공정한 방향으로 추진.	중국의 차세대 AI 건강 발전을 추진.
안전, 신뢰, 통제 가능성	완전 동일.	AI가 안전하고, 신뢰할 수 있으며, 통제 가능하도록 보장.	AI가 안전하고, 신뢰할 수 있으며, 통제 가능하도록 보장.
중국의 강점	강화. 시진핑이 언급하는 중국의 AI 강점은 자신감의 표현.	중국은 풍부한 데이터 자원, 완전한 산업 시스템, 광범위한 응용 시나리오, 거대한 시장 공간을 보유	중국의 방대한 데이터와 거대한 시장의 우위를 충분히 발휘.
응용 중심	강화. 올해는 "응용"이 더욱 전면에 나옴.	응용 지향을 두드러지게 한다.	수요 지향, 시장 주도적인 과학기술 발전 경로를 견지.
독립적 공급망	**구체성 증가. 2018년에는 "독립적이고 통제 가능한"이라는 흔한 표현을 사용하지 않았지만, 2025년에는 명확히 명시.**	**자주적이고 통제 가능하며, 원활하게 협동 운영되는 AI 기초 소프트웨어 및 하드웨어 시스템을 구축.**	**핵심 AI 기술이 우리 손에 단단히 쥐어지도록 보장.**
생태계 협력	**강화. 학계 및 산업계 전반의 협력적 조정을 요구. 기업의 주도적 역할을 명시.**	**AI 과학기술 혁신과 산업 혁신의 심도 있는 융합을 촉진하고, 기업 주도의 산학연 협동 혁신 시스템을 구축.**	**과학기술 혁신과 산업 응용 간의 상호 촉진이라는 국면 형성.**
핵심 및 기반 기술	**구체성 증가. "고급 칩과 기초 소프트웨어"를 구체적 언급.**	**기초 이론, 핵심 기술 등 방면에서 여전히 부족한 점이 존재. 고급 칩, 기초 소프트웨어 등 핵심 기술을 돌파해야 함.**	**핵심 및 기반 기술에 집중.**
산업 고도화	거의 동일.	전통 산업 구조조정을 지원하고 산업 고도화를 보조.	AI의 기술적 이점을 산업 업그레이드, 제품 개발, 서비스 혁신 등에 충분히 활용.
위험 관리	구체성 증가. 올해 언급된 잠재 위험들은 훨씬 더 구체적.	기술 모니터링, 위험 조기 경보, 비상 대응 시스템을 구축.	AI 발전의 잠재적 위험에 대한 분석 및 예방을 강화.
인재 양성	강화 및 구체성 증가. "모든 사람에게 AI 소양" 교육을 제공, 인재를 잘 활용하는 데 더 큰 중점을 둠.	AI 교육을 모든 학문 단계에 추진하고 전 사회 AI 소양 교육을 추진하며, 유능한 인재의 끊임없는 공급원을 양성하고, AI 과학 연구, 전문 지원, 인재 평가를 위한 메커니즘을 개선	인재팀 건설을 강화하고, 더 많은 형태와 더 큰 강도로 교육 훈련을 진행하며, 더 많은 과학 연구 자원을 강화하고, 우수한 인재에게 기술, 기술 및 산업 발전에 더 많은 지원.
기초 연구	거의 동일.	기초 연구를 강화해야 함.	기초 이론 연구를 강화.
발전 법칙	거의 동일.	AI 발전의 추세와 법칙을 파악.	AI 발전의 법칙과 특징을 파악.
승수 효과	거의 동일.	기초 이론, 방법, 도구 등 반드시 돌파구를 마련해야 함	이론, 방법, 도구, 시스템 등 분야에서 변혁적이고 파괴적인 돌파구를 마련하기 위해 노력.
정책 지원	**구체성 증가.**	**지식재산권, 재정·세제 정책, 정부 조달, 공공 시설 개방 등 분야에서 정책을 잘 활용하고, 과학기술 및 금융 문건 작성 작업을 잘 수행.**	**정책 지원을 확대.**

자료: 중국 공산당, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(1) 오픈소스 확장

20세기 미국의 성공 청사진을 복제하라

- 21세기 AI 패권 경쟁의 본질은 기술 우위가 아니라 '글로벌 표준' 장악. 미국의 제재 속에서 중국은 '오픈소스'를 무기로 선택. 저렴한 고성능 모델을 대량 공급해 서구의 AI SW 독점을 무너뜨리는 전략.
- 집착에 가깝게 인터넷을 통제했던 중국이, AI 시대에는 180도 태도를 바꾸면서, "권위주의 진영이 개방성을, 자유주의 진영이 폐쇄성을 지향"하는 역설적인 상황을 연출되고 있음. 중국은 20세기 미국이 TCP/IP 프로토콜로 인터넷을 장악했던 성공 청사진을 AI 시대에 그대로 복제하려고 시도. 자국 내 기술 우위에서 '글로벌 기술 채택률'로 확장하겠다는 의미. 중국의 전략은 에버렛 로저스의 '혁신확산이론'에 기반한 체계적인 4단계 접근법을 따르고 있는 것으로 사료됨.
  ** ①기술력으로 '혁신가'를 유인 → ②오픈소스로 '조기 수용자'를 포섭 → ③표준 설정으로 '전기 다수'를 확보 → ④인재 수출로 담론 권력을 장악*

주요 중국 클라우드 기업의 해외 인프라 계획
중국의 주요 CSP들은 하나같이 해외 인프라 확장에 큰 관심

기업명	주요 투자 및 해외 인프라 계획
알리바바 클라우드	– 40억 위안 이상 투자 예정
	– 글로벌 협력 생태계 강화 및 AI 서비스 확산 목표
	– 올해 아랍에미리트, 일본 등지에 새 데이터센터 개설
화웨이 클라우드	– 아시아태평양 지역: 필리핀에 새 거점 개설
	– 아프리카 지역: 이집트 및 케냐에 새 거점 개설
	– 북아프리카 최초의 화웨이 클라우드 거점
텐센트 클라우드	– 사우디 리야드에 중동 첫 데이터센터 건설에 1.5억 달러 투자
	– 인도네시아에 세 번째 데이터센터 건설에 5억 달러 투자
	– 오사카에 데이터센터 및 새 사무소 신설 예정

자료: 각 회사, 미래에셋증권 리서치센터

혁신 확산 이론에 기반한 중국의 4단계 AI 글로벌 전략
현재는 1단계에서 2단계로 넘어가는 단계로 보임

단계	목표 그룹	핵심 전략	세부 실행 방안
1단계	혁신가 (Innovators)	압도적인 기술력 확보	OpenAI 등 세계 최고의 모델과 동등하거나 능가하는 성능을 입증하여 기술 자체로 매료시킴
2단계	조기 수용자 (Early Adopters)	오픈소스 생태계 구축	모델 코드, API 등을 오픈소스로 공개하여 해외 개발자들이 직접 사용하고 테스트하며 초기 생태계를 형성하게 함
3단계	전기 다수 (Early Majority)	글로벌 표준 주도	자국 AI 기술(프로토콜, 모델형식)을 국제 표준에 포함시켜 호환성을 확보하고 주류 채택 가속화
4단계	후기 수용자	인재 수출을 통한 규범 형성	엔지니어, 연구자들이 국제 표준 기구, 컨퍼런스 등에 깊이 참여, 기술 규범과 거버넌스 형성에 직접적인 영향력을 행사

자료: 류 샤오산(刘少山, 선전 인공지능 및 로봇 연구소인 AIRS 산하 기업의 디렉터), 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(1) 오픈소스 확장

기술력은 확보했어도 생태계 확보가 안된다

- 중국은 혁신확산이론의 1단계(기술력)는 성공적으로 달성했으나, 2단계(생태계)에서는 크게 뒤쳐짐. DeepSeek는 다수 벤치마크에서 OpenAI의 o1 모델과 동등하거나 일부 능가하는 성능을 입증. 특히 Embodied AI(로보틱스 등) 분야에서는 강력한 산업 공급망을 기반으로 센서, 모터 등 업스트림부터 대규모 시범 배치 등 다운스트림 응용까지 구축. 미드스트림 기술인 칩의 격차도 좁히고 있음.
- 하지만, 이를 확산시킬 생태계가 취약. 미국 메타의 LLaMA 2 모델 하나가 허깅페이스에서 모든 중국 오픈소스 모델의 다운로드 수를 합친 것보다 많음. 허깅페이스는 단순 모델 다운로드 사이트가 아닌, AI계의 '구글 플레이스토어'. 모델 다운로드, 테스트, 개발 도구, 커뮤니티 토론까지 원스톱으로 제공. 미국은 이 플랫폼을 완벽 장악하여 개발자들에게 '기본값(default)'이자 글로벌 표준으로 자리함.
- 혁신확산이론 3단계인, '글로벌 표준' 경쟁에서도 역부족. AI 국제 표준을 제정하는 핵심 기구인 ISO/IEC JTC 1/SC 42의 사무국은 미국의 NIST가 운영. 5개 워킹그룹 의장직 중 중국은 단 1석에 불과. AI 윤리, 투명성 등 핵심 의제에서 미국의 담론 권력은 절대적.

AI 기술의 글로벌 규범을 만드는 가장 중요한 기구인 "AI 국제 표준 위원회(ISO/IEC JTC 1/SC 42)"의 권력 구조 분석

직책 / 역할	담당 국가 / 주체	의미와 영향력
전체 위원회	ISO/IEC JTC 1/SC 42	AI 분야의 국제 표준(글로벌 규범)을 제정하는 최고 핵심 기구
사무국 (Secretariat)	미국 (국립표준기술연구소를 통해 운영)	전체 의제 설정, 회의 운영 등 행정적, 실질적 주도권을 미국이 장악
핵심 표준 개발	미국 주도 (2018년 이후)	AI 위험 관리 가이드라인 등 주요 표준들의 초안 작성/개발을 미국이 이끔
워킹그룹(소위원회) 의장단	총 5개 그룹	세부 기술 분야별 표준안을 구체적으로 논의하고 만드는 실무 그룹 의장직은 논의 방향을 결정하는 핵심
ㄴ "AI의 계산적 접근법 및 특성" 관련 워킹그룹의 의장국	*중국*	*중국의 영향력은 5개 중 1개 분야로 제한적*
ㄴ AI 윤리, 데이터, 거버넌스 등 나머지 핵심 분야들의 의장국	*미국, 캐나다, 아일랜드, 일본*	*다른 주요 표준들의 논의는 미국과 그 동맹국들이 주도하고 있음*

자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(1) 오픈소스 확장

인재를 수출하고, 표준을 지배하라

- 중국은 공식 표준화 기구에서의 열세를 인정하고, 자국 기술을 오픈소스 코드와 하드웨어/소프트웨어 제품에 직접 내장해 신흥 시장(특히 BRICS 등 일대일로 국가)에 수출하려고 하고 있음. 이를 통해 공식 표준이 아닌, 시장이 사용하는 '사실상의 표준(de facto standard)' 지위를 노리는 방식.
- 나아가, 중국은 해당 이론의 4단계(담론 권력 장악) 전략을 위해 '인재 귀국(talent repatriation)'에서 '인재 해외 진출(talent go-global)'로 무게중심을 옮기며 공세적으로 전환하고 있음. 미국 내 최고 AI 연구자의 38%가 중국 출신으로, 미국 본토 출신(37%)을 넘어선 상황. 최근 개설된 메타의 초지능 연구소(MSL)의 경우 40~50%가 중국계 엔지니어일 정도로 중국의 인재 풀은 글로벌 최고 수준.
- 즉, 국내로 인재를 불러들이는 수비적 자세를 넘어, 뛰어난 중국 AI 전문가들이 일대일로 국가 등 신흥 시장으로 진출하도록 장려하는 것. 현지에서 연구소 설립, 공동 연구, 기술 회사 창업 등을 통해 중국 중심의 기술 생태계와 규범을 직접 구축하려는 목적. 기술 거버넌스 자체를 수출하려는 고도의 전략.

메타의 초지능연구소(MSL)에서 재직 중인 사람들 중 중국계 엔지니어들 리스트, '미국과 인도인 비중보다 중국계 비중이 훨씬 높음'

이름	학사 학위	Meta 재직 기간	총 경력(년)	이전 경력(주요)	전문 분야
Tao Zhu	칭화대 CS 석·학사	18 일	15	드롭박스·트위터 선임 엔지니어	대규모 ML 시스템
Pei Sun	칭화대 CS 학사	2 개월	14	딥마인드 수석 연구; 웨이모 수석 연구	자율 주행·대규모 LLM
Michael Zhang	케이프타운대 수학 석·학사	4 일	12	딥마인드 선임 연구원	정보 검색
Shuchao Bi	절강대 수학 학사	18 일	12	OpenAI의 사후학습 헤드	비디오 생성
Xiaohua Zhai	난징대 학사	17 일	10	OpenAI MTS; 딥마인드 수석 연구	비전 트랜스포머
Yu Zhang	칭화대 CS 석·학사	3 일	10	OpenAI MTS; 딥마인드 연구	백엔드 시스템
Qingqing Huang	칭화대 전기 학사	18 일	9	TikTok 연구; 구글 연구	모바일 시스템
Chunyuan Li	상하이교통대 전기 학사	18 일	7	X AI 연구팀장; ByteDance 연구리드; MS 리서치 수석	다중모달 생성·확산
Huiwen Chang	브리스톨대 통계·수리 학사	4 개월	6	OpenAI MTS; 구글 연구원	생성형 이미지/영상
Chenxi Liu	화중과기대 전기 학사	4 개월	5	DeepMind 수석 연구; 웨이모 시니어 리서치	시각-언어 모델
Jiahui Yu	상하이교통대 전기 학사	18 일	5	OpenAI MTS; 구글 연구	확산 모델
Rui Hou	중난대 학사	5 년	5	X AI; Meta 수석 연구	ML 인프라
Zhishuai Zhang	UST China CS 학사	3 개월	5	OpenAI MTS; 딥마인드 수석 연구; 웨이모 선임 엔지니어	세그멘테이션 모델
Lu Liu	베이징대 학사	18 일	4	OpenAI MTS	멀티모달 추론
Chengxu Zhuang	UCLA 통계·칭화대 학사	4 일	3	OpenAI MTS; MIT 박사후연구원	신경과학 영감 AI
Shengjia Zhao	칭화대 학사	18 일	3	OpenAI MTS	확률 모델링
Hongyu Ren	칭화대 학사	18 일	2	OpenAI MTS	그래프 ML
Ji Lin	칭화대 전기 학사	18 일	2	MIT SEAL 헤드; NYU 박사	효율적 LLM·TinyML
Haotian (Ken) Tang	북경대 학사	7 일	1	딥마인드 연구원	효율적 딥러닝 시스템
Pingchuan Ma	난카이대 소프트웨어 학사	7 일	1	OpenAI MTS	비전-언어
Yuanzhi Li	칭화대 학사	4 개월	0	CMU 교수	최적화 이론

자료: X(@deedydas), 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(1) 오픈소스 확장

**SW는 일단 미끼,
HW로 나중에 돈 버는
전략이 이번에도 통할까?**

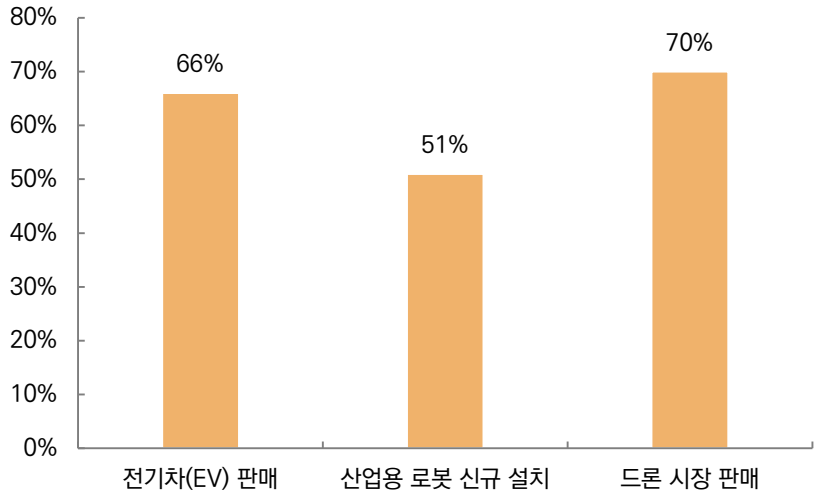
- 중국의 오픈소스 전략은 자선사업이 아님. 과거 서구의 제조업을 무너뜨렸던 공식을, 미국의 마지막 해자인 소프트웨어 산업에 적용하려는 치밀한 전략. 즉, ①복제하고 ②최적화하고 ③규모를 키운 다음 ④낮은 가격으로 시장을 장악하는 전략을 SW에 그대로 적용하는 것.
- 고성능 오픈소스 AI 모델을 대량으로 저렴하게 공급하여, 소프트웨어 자체의 가격을 폭락시키고 빅테크들의 고마진 API 비즈니스 모델을 파괴하려는 것. 소프트웨어를 미끼로 중국 중심의 기술 생태계를 구축한 뒤, AI 기능이 탑재된 하드웨어(전기차, 로봇, 드론 등) 판매로 막대한 이익을 창출하려는 장기적 계산. SW에서 잃은 돈을 HW에서 몇 배로 벌어들인다는 판단일 수 있음.
- DeepSeek CEO 량원펑은 R1 모델 출시 직후 리창 총리의 정부 좌담회에 순수 AI 기업 중 유일하게 참석. 이는 R1의 오픈소스 공개가 국가적 차원의 승인과 교감 하에 이루어졌음을 강력히 시사. R1 모델은 기술적으로는 세계를 뒤흔들었지만, 실제 산업 현장에서는 '매우 저조한 채택률'이라는 차가운 현실에 직면. 중국식 AI 오픈소스 전략이 글로벌 표준이 되기까지는, 기술 외 신뢰 문제를 해결해야 함.

서방의 연구원들이 DeepSeek 등 中 모델 채택을 꺼리게 하는 현실적 장애

유형	핵심 내용
1. 보안 및 위험 우려	기업들은 중국의 오픈 소스 모델을 사용 시 발생할 수 있는 **①정보 유해성 ②도구 사용 코드 실행 위험을** 심각하게 우려하여 자사 서버 내 실행을 금지하는 경향
2. 지정학적 요인 및 불확실성	현재 출시된 모델들은 과거 데이터로 훈련되어 중국 정부의 직접적 영향력이 적을 수 있으나, **①미래 모델의 변화 가능성 ②이에 따른 측정 불가능한 위험** 때문에 **기업 고객들은 중국 모델 채택을 극도로 주저**
3. 미중 무역 전쟁의 지속	진행 중인 미중 간의 무역 전쟁이 지속되는 한, 중국 AI 모델에 대한 서방 기업들의 우려와 채택을 가로막는 장애물은 쉽게 해소되기 어려울 것

자료: Nathan Lamber(Allen Institute의 AI 연구원), 미래에셋증권 리서치센터

중국이 전기차, 산업용 로봇, 드론 분야에서의 글로벌 점유율



자료: 국제에너지기구, 국제로봇연맹, ResearchAndMarkets, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(1) 오픈소스 확장

오픈소스 정책의 중국産 "메기 역할"인 DeepSeek를 잡는 미국의 OpenAI

- DeepSeek는 R1을 "오픈 가중치(open-weight)"로 공개. 메타의 Llama를 뛰어넘는, 현재 시장에 나와 있는 가장 개방된 오픈소스 모델로 분류. CEO인 량원펑(梁文锋)은 R1 모델을 공개한 직후 중난하이로 이동해 정부업무보고 좌담회에 참석했고 배석자중 유일한 AGI 관련 인물이었음. 中정부는 DeepSeek R1 출시와 오픈소스 공개에 관한 승인을 내렸을 가능성이 높다는 것.
- 공산당의 총애를 받을 DeepSeek는 앞으로 더 좋은 모델 훈련을 위해 더 높은 품질의 데이터를 제공받을 가능성. 데이터 전처리 비용 측면에서도 중국 개인 정보보호법 처리가 미국보다 "유연"한 덕분에, 개인 정보를 지우고 입력하는 등의 데이터 탈감(민감정보 비식별화) 처리 비용이 미국과 비교가 되지 않을 정도로 저렴할 것으로 예측.
- 한편, OpenAI는 메타의 오픈소스가 부진한 가운데, '중국 거 쓰지 말고 미국 오픈소스 쓰라'는 제안의 성격으로 "6년 만에" 매우 저렴한 오픈소스 모델 출시. 다만, 핵심 기능은 놔두고 "부분 공개" 전략. 이는 사뭇 엔비디아의 대중 칩 수출 전략을 떠올리게 함. 엔비디아가 성능이 좀 뒤떨어지는 가성비의 칩으로 중국을 묶어두듯, DeepSeek 등 中 AI 생태계 형성을 지연시키는 전략적 카드로 보임.

엔비디아의 대중 수출용 칩이 오버랩 되는 미국 OpenAI 오픈소스 전략

비교 항목	엔비디아의 칩 H20 (중국 수출용 다운그레이드)	OpenAI의 모델 GPT-OSS (120B·20B 오픈 가중치 모델)
공개 범위	하드웨어 성능을 낮춰 중국에 판매	첨단 모델보다 '좀' 뒤떨어지는 모델을 무상으로 공개(Apache 2.0)
목적	중국 고객사들이 화웨이 칩으로 완전히 이동하지 못하게 함	중국·글로벌 커뮤니티를 미국 기반의 LLM 생태계에 묶어두기
미국 우위 보호	최첨단의 H100 및 최신 GB200은 계속 통제	o3/멀티모달/장문 컨텍스트 등 프런티어 모델 및 핵심 소스는 비공개
정책 효과	**생태계 락인 효과 때문에 성능이 비슷하다면, 화웨이보다 엔비디아 선택**	**데이터 보안 등 중국산 모델이 가진 한계 때문에 성능이 비슷하다면, DeepSeek/알리바바보다 OpenAI 선택**

자료: 미래에셋증권 리서치센터

주요 오픈소스 모델들의 SWE-Bench 점수 비교
AI가 프로그래머처럼 코드를 고치거나 개선하는 능력을 테스트

모델	해결률	입력 비용 (달러/백만 토큰)	출력 비용 (달러/백만 토큰)
Qwen3-Coder 480B/A35B Instruct	55.4%	컨텍스트 길이별 계층화 (0-32K: 1달러; 32K-128K: 1.8달러; 128K-256K: 3달러; 256K-1M: 6달러)	컨텍스트 길이별 계층화 (0-32K: 5달러; 32K-128K: 9달러; 128K-256K: 15달러; 256K-1M: 60달러)
Kimi K2 Instruct	43.8%	0.15 달러	0.6 달러
gpt-oss-120b	**26%**	**0.1 달러 (OpenRouter 경유)**	**0.5 달러 (OpenRouter 경유)**
Qwen2.5-Coder 32B Instruct	9%	0.7 달러	2.8 달러

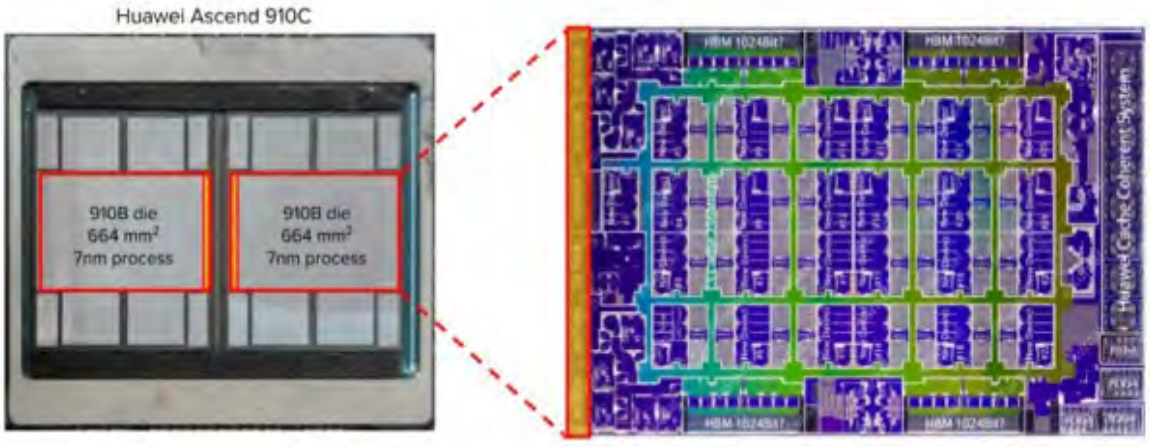
자료: OpenAI, 미래에셋증권 리서치센터
주: SWE-Bench 기준은 'Bash Only'. Bash 명령어만 써서 해결 가능한 간단한 문제들을 골라서 테스트한 것

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

젠슨 황 엔비디아 CEO "화웨이는 세계적으로 매우 강력한 기술 기업"

- 미국의 대중국 반도체 수출 규제는 단기적으로는 중국의 목을 조였지만, 장기적으로는 중국의 자체 장비 및 칩 산업에 거대한 내수 시장을 만들어주는 역설을 낳음. 제재 이전, 중국 칩 제조사들은 자국산 장비가 있어도 더 우수한 미국/일본산을 구매. 하지만 제재로 인해 울며 겨자 먹기로 자국산을 쓸 수밖에 없는, 즉 '보호된 내수 시장'이 강제로 형성.
- 제재로 붕괴 직전까지 갔던 화웨이는 자체 OS(Harmony)와 칩(Kirin) 개발, 차량 BU의 폭발적 성장(전년 대비 400%↑)으로 제재 이전의 매출 수준을 거의 회복. 미국의 제재가 중국의 기술 자립 의지를 꺾는 대신, 생존을 위한 혁신을 강제하는 기폭제가 된 셈. 화웨이는 이제 '중국의 엔비디아'로 불림.
- 특히 화웨이의 Ascend 칩은 엔비디아의 대안으로 빠르게 부상 중이며, 공급망 전체의 국산화가 동시에 진행되고 있음. 현재 유통 중인 Ascend 910C 칩 2개가 엔비디아 H20 칩 1개와 동등한 성능. 현재 수율은 30-40% 수준에 불과하나, 2026년까지 70% 이상 달성을 목표.

2024년 10월에 첫 출시된 Ascend 910C는 Ascend 910B을 기반으로 듀얼 다이 구성
EUV 없이 SMIC N+2(7나노)를 쓰고 HBM2E를 채택, CoWoS 쓰지만 엔비디아 칩 대비 '열위' 기술



자료: Huawei, Lennart Heim, 미래에셋증권 리서치센터

Ascend 910 출하량 및 배송처 예상 (단위: 천)

구분	Ascend 910C	Ascend 910B
2024년	0	500
2025년	200	350
구분	**2025년 Ascend 910C 출하량 전망**	
중국 CSP (B.B.A.T.)	80	
통신사 등 국유기업	120	

자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

컴퓨팅 면에서 미국의 압도적 우위는 여전하다

그럼, 격차를 어찌 극복?

- 한편, 올해 4분기 출시 목표인 차세대 Ascend 910D는 H20 성능을 동등/초과할 것을 목표. 엔비디아 Hopper 시리즈가 2022년에 출시된 것을 감안하면, 양사 간 칩 설계 능력 격차가 최소 3년 이상임.
- 현재 미국의 CSP들이 많이 선택하는 엔비디아 B200 칩과 비교했을 때는 910C의 계산 성능이 약 3배 낮고, 메모리 대역폭이 약 2.5배 낮으며(HBM2E 기준), 전력 효율성이 훨씬 떨어짐. 화웨이의 추격이 무섭기는 해도 기술적 격차가 생각보다 아직 상당해 보임. 총 컴퓨팅 면에서 미국의 7~10배 정도의 우위는 여전히 지속적으로 강하다고 볼 수 있음.
- DeepSeek를 포함한 중국의 모든 주요 AI 모델들은 이미 화웨이 Ascend 칩에서 독립적으로 구동되도록 최적화 작업을 마침. 그럼에도 불구하고, DeepSeek은 별로 Ascend 칩에 관심 없는 실정.
- 그러나 화웨이는 당장의 글로벌 최고 성능를 달성하는 게 아니라, 중국 내 AI 기업들이 필요로 하는 충분한 성능의 칩을 안정적으로 공급하여 시장 점유율을 확보하고 자국 AI 생태계의 기반을 다지는 것. 이 시스템에 대항하기 위한 노력은 칩 레벨 뿐만 아니라 시스템 레벨에 있다는 것을 유념.

화웨이의 자체 제작 GPU 칩인 Ascend 라인업

칩 모델	목표 성능 추정	공급 일정	추정 수율
Ascend 920	H20 대체	하반기 램프업?	공개 안 됨
Ascend 910D	H20 성능 일부 초과	5월 말 첫 샘플	공개 안 됨
Ascend 910C	H20 약 절반 수준	5월 램프업 개시	약 40%
Ascend 910B	A100 일부 대체	작년부터 점차 램프업	50% 이상

자료: Huawei, 미래에셋증권 리서치센터

화웨이의 주력 칩인 910C 대비 차세대 칩인 920C의 주요 변화

항목	현재 주력 910C	차세대 920C (트랜스포머 및 MoE 특화)
제조 공정	SMIC 7nm(N+2)	SMIC 6nm(N+3)
HBM 대역폭	HBM2E 3.2 TB/s	HBM3 4.0 TB/s
전용 Tensor Core	Da Vinci v2	Da Vinci v3 (INT8 / FP8 매크로)
MoE 엔진	없음	Sparse-Switch + Expert SRAM
인터커넥트	PCIe 4.0 + Da Vinci Link	PCIe 5.0 + CM-Link(800 GB/s)
칩렛·패키징	2.5D CoWoS-Lite	4-Chiplet 구조

- HBM3로 바뀌면서 AI 훈련과 추론시 토큰 시퀀스 길이가 늘어날 수 있어 효율성 ↑
- **4096 × 4096 크기의 행렬곱셈으로 늘어나 성능 ↑ (But, 수율 문제가 커질 가능성)**
- 현재 주류 아키텍처인 Mixture of Experts 전용 엔진을 탑재해 토큰 라우팅 지연 ↓
- 전용 인터커넥트 업그레이드와 PCIe 5.0 지원으로 "Experts 병렬화" 통신 ↑
- **칩렛 구조를 사용해서 전력 및 수율 관리 ↑ (But, EUV 장비는 여전히 못 쓴다는 증거)**

자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

“메모리 자립”을 위해, NAND(AI SSD)로 분산

- 화웨이는 8월 27일 “AI SSD”를 공개. 해당 제품의 핵심은 UCM(AI 추론용 메모리 데이터 관리기)와 결합한 다음, HBM·DRAM·SSD를 계층화하고 KV캐시를 오프로딩해 “HBM 용량 확장 효과”를 내겠다는 것. 이는 엔비디아의 Dynamo에서 제공하는 “분산 KV캐시 관리자” 기술과 비슷함. 이미, 엔비디아 방식으로 “DeepSeek 모델 추론을 SSD 계층에 두고 굴린다”는 중국 내 생태계도 존재.
- 화웨이의 UCM+AI SSD는 HBM 의존을 줄이는 설계. HBM 접근 제약을 ‘시스템 아키텍처’로 우회하려는 전략을 공식화했고, 이를 뒷받침할 하드웨어로 AI SSD를 내세운 것. 이는 HBM-only의 현 체제를 흔드는 체인저가 될 수도 있지만, HBM의 대체재는 아닐 것. 실시간 토큰 생성은 여전히 레이턴시 민감 영역이며, SSD는 “확장 메모리”일 뿐.
- 그럼에도 불구하고, KV캐시 재사용률이 높은 워크로드(긴 컨텍스트로 이뤄질 수 밖에 없는 에이전트 작업)에서는 중대한 영역을 차지할 수 있음. 화웨이는 그 지점을 정확히 찌르고 있고, HBM 수급이 불안정해진 중국 AI 생태계의 현실적인 최적화 경로라고 할 수 있음.

길고 반복되는 대화로 이뤄진 에이전트 작업에 안성맞춤인 “엔비디아 Dynamo: 분리형 서빙 아키텍처”
효과는 HBM 압박 완화, 세션 재사용 극대화, 토큰 생성 지연·비용 하향, 더 많은 동시 사용자

캐시 메모리가 HBM에 다 못 들어가면 **“KV 캐시 관리자”**가 일부의 데이터를 CPU RAM, SSD로 보냄.

화웨이의 UCM이 동일. “HBM, DRAM, SSD 중 어디에 KV-캐시를 둘지를 순간마다 결정”하는 엔진

한편, **화웨이의 AI SSD는 “데이터가 내려앉을 자리”**를 더 크고 빠르고 안정적인 SSD 계층으로 제공

자료: 엔비디아, 미래에셋증권 리서치센터

화웨이가 내놓은 HBM 대체재 “AI SSD”

자료: 화웨이, 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

"AI 인해전술" CloudMatrix384

- 올해 4월 화웨이는 개별 칩 성능의 열세를 시스템 전체의 규모로 압도하는, 이른바 '양으로 승부하는' 전략의 결정체인 CloudMatrix384(CM384) 아키텍처를 공개. 엔비디아 칩의 3분의 1 수준에 불과한 개별 Ascend 910C 칩의 성능을 극복하기 위해, 엔비디아의 데이터센터 시스템인 "NVL72(GPU 72개)"보다 5배 더 많은 "384개" 가속기를 하나의 거대 시스템인 CM384로 묶어 버림.
  - 전직 화웨이 직원: "910C 칩 2개가 엔비디아 H20 칩 1개와 동등한 성능. 곧 출시될 910D는 H20과 동등한 성능 목표"
- CM384는 300PFLOP의 고밀도 BF16 컴퓨팅을 제공할 수 있으며, 이는 엔비디아의 GB200 NVL72의 1.6배에 달하는 수준. 3.6배 이상의 총 메모리 용량 과 2.1배 더 넓은 메모리 대역폭을 통해 화웨이와 중국은 이제 엔비디아를 능가(?)하는 AI 네트워킹 시스템 역량을 갖추게 됨.
- 화웨이는 지난 4월에 출시한 본인들의 자체 LLM인 Pangu Ultra 모델이 자체 하드웨어만으로 거대 모델 훈련에 성공했다는 사실을 증명. 화웨이의 시스템이 대규모로 확장 가능(8천개 칩 이상)하며 안정적인 분산 연산을 수행할 수 있음을 보여주는 증거. 현장의 요구를 감당할 수 있다는 신뢰를 구축.

엔비디아와 화웨이의 '네트워킹 방식' 차이에 따른 AI 시스템/칩 레벨에서의 스케일업 네트워킹 비용 및 필요전력
화웨이 CM384 랙 시스템은 엔비디아의 NVL72 랙 시스템에 비해 비용은 6배에 달하고, 전력 소비는 무려 11배

시스템 레벨(단위 : $,W)	NVL72 비용	NVL72 전력	CM384 비용	CM384 전력
스케일업 스위치 관련	$167,760	5,400W	$60,173	12,800W
구리 기반 인터커넥트	$125,616	0	$0	0
광트랜시버 및 광섬유 기반	$0	0	$1,733,760	48,384
시스템 레벨 스케일업 비용	**$293,376**	**5,400W**	**$1,793,933**	**61,184W**
칩 레벨(단위 : $,W)	**GB200 비용**	**GB200 전력**	**Ascend 910C 비용**	**Ascend 910C 전력**
스케일업 스위치	$2,330	75W	$157	33W
구리 기반 인터커넥트	$1,745	0	$0	0
광트랜시버 및 광섬유 기반	$0	0	$4,515	126W
칩 레벨 스케일업 비용	**$4,075**	**75W**	**$4,672**	**159W**

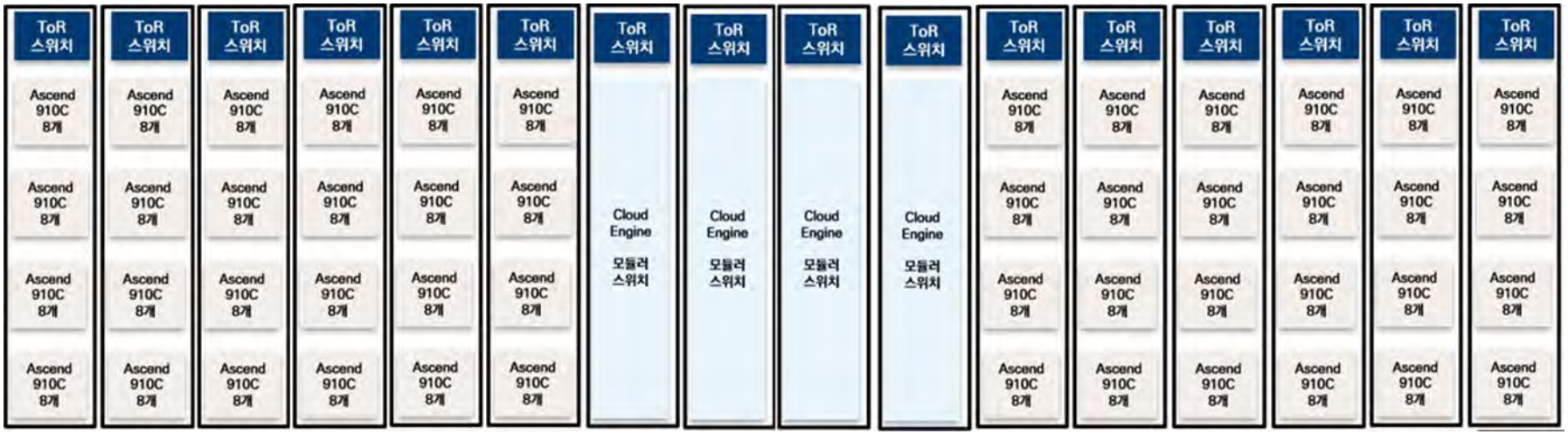
자료: SemiAnalysis, 미래에셋증권 리서치센터
엔비디아와 화웨이의 부품/솔루션에 대해 각각 75%와 15%라는 가상의 매출총이익률(GPM)을 적용하여 비용을 추산.

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

엔비디아 GB200 NVL72에 대항하는 "중국식 해법"

- CloudMatrix(CM) 시스템은 16개의 랙에 분산되어 있음. 이 중 컴퓨팅을 담당하는 랙 12개에는 32개 GPU가 장착. 나머지 4개의 랙은 중앙에 위치해 스케일업 스위치로 채워져 있음. 이 스위치들이 논리적으로는 하나의 거대한 스위치 패브릭으로 작동되는 방식
- 화웨이는 세계적인 규모를 확보하기 위해, 여러 랙에 걸친 '스케일업'을 가능하게 하는 핵심으로 구리 케이블을 완전히 배제. 대신에, 시스템 전체를 광학(Optics)으로 연결하는 무식할 정도로 과감한 아키텍처로 구성. 하나의 CM384 시스템은 384개의 GPU를 All-to-All 방식으로 묶기 위해 6,912개에 달하는 400G 광 트랜시버를 투입
- 이는 엔비디아가 원래 비용 문제로 포기했던 방식을 현실화한 것. 물론 이는 와트당 연산 효율(FLOPs/Watt)이 엔비디아 대비 2.6배나 떨어지는 극심한 전력 비효율성을 대가로 하지만, 이는 역설적으로 전력 제약에서 자유로운 중국이기에 선택할 수 있었던 유일한 전략이기도 함. 다시 말해, 미국의 관점에서의 비효율적인 전력 소비의 문제가 중국에서는 문제가 되지 않음. 중국은 전기가 모자라지 않음. 전력 제한이 없다면, 광학 기반의 대규모 스케일업은 합리적인 선택인 것.

CloudMatrix384 아키텍처 구조 – 16개의 랙에 걸쳐 384개의 Ascend 910C 칩을 필요로 하는 구조
엔비디아의 시스템 성능과 비슷할 수 있지만 더 많은 하드웨어, 공간, 에너지 및 냉각 용량이 필요한 중국의 컴퓨팅 비효율성을 보여주기도 함



자료: SemiAnalysis, 미래에셋증권 리서치센터

[참고] 화웨이 CloudMatrix384

개별 칩의 정교함 대신, 광학 기반의 극단적인 병렬 확장성을 노리는 화웨이의 독자 시스템 아키텍처

화웨이 Atlas 900 A3 SuperPod(CM384)의 전경, 16개의 랙으로 구성

CM384의 후면부. 6,912개의 400G 광 트랜시버를 투입한 스케일업

계산궤(计算柜)'라고 표시된 부분은 네트워킹이 아닌 컴퓨팅 랙을 표시

CM384 서버 모듈 근접 촬영, 좌측은 네트워킹 랙이고 우측은 컴퓨팅 랙

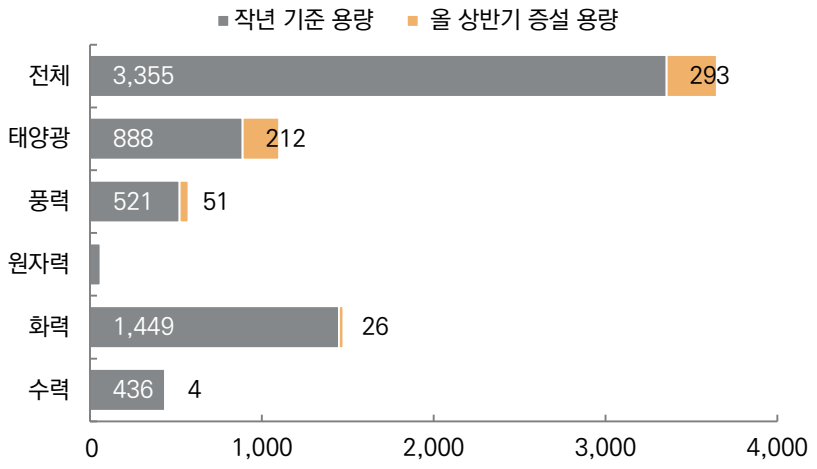
자료: 직접 촬영, 화웨이, WAIC 2025, 미래에셋증권 리서치센터

[참고] 중국은 전기가 모자라지 않고 오히려 공급 여유가 크다

전력 공급(MW)과 연간 생산량(MWh) 모두 데이터센터 확장에 큰 걸림돌이 아니다

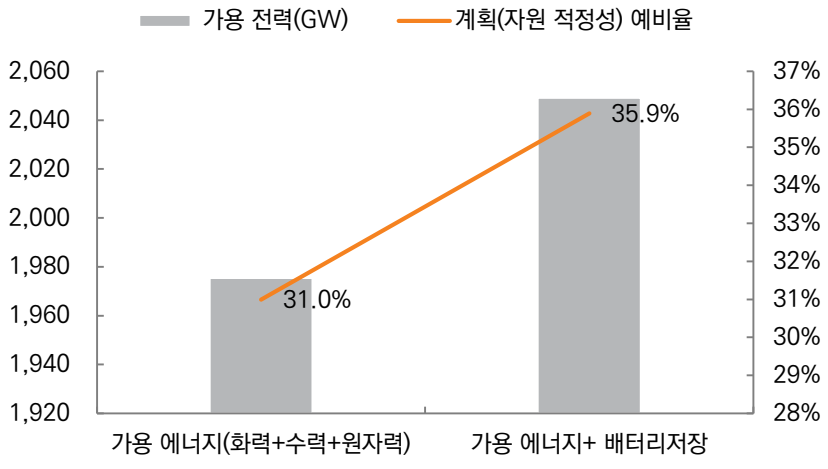
- 2025년 상반기 말 기준, 필요할 때 바로 돌릴 수 있는 전원(화력·수력·원전·배터리)이 약 2,071GW.
- 지난 7월 17일 전국 최대 수요는 1,508GW였음. 이를 통해 예비율을 구해보면, 약 37%로 안전구간. 이 계산은 태양광·풍력 같은 '간헐적 전원'은 빼고 잡은 보수적 추정임에도 그러함. 24/7 돌아가는 데이터센터는 이 피크 부하(최악의 상황)를 기준으로 삼으면 됨.
- 물론, 평소에는 수요가 피크보다 훨씬 낮음. 그래서 대부분의 시간대 예비율은 37%보다 더 높음. 많은 국가가 예비율 15~25%를 목표로 한다는 것에서 중국은 예비력이 큼. 발전 능력 자체도 수요보다 큼. 작년 화력 1,444GW의 평균 가동시간은 3,442시간. 설비이용률은 약 39%로, 절반 이상이 놀고 있다는 뜻. 이용률을 78%로만 올려도 연간 전력 생산이 +4,970TWh 늘어남. 이 추가분만으로도 미국의 연간 전력소비를 넘는 규모.
- 2021년 전력난은 설비 부족이 아닌 석탄값 급등과 전기요금 규제 충돌이 원인. 이후 전력요금 자유화로 비용 전가가 가능해져서 재발 가능성은 낮음. 특히 중국 특유의 국영 기업 중심의 시스템은 필요하다면 막대한 전력을 AI 데이터센터에 즉시 할당할 수 있음.

2025년 상반기 기준 증설 용량 포함 총 설비용량
재생에너지 발전 비중은 40%에 육박



자료: 중국 국가에너지국, 중국 국가핵안전국,미래에셋증권 리서치센터

중국의 피크부하 1.508TW 대비 가용 에너지의 예비율:
약 31%(배터리 제외) ~ 36%(배터리 포함)로 넉넉하다



자료: 중국 국가에너지국, 중국 국가핵안전국,미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

화웨이의 파운드리 파트너 = SMIC

그러나 부족하다…

- 중국의 반도체 굴기는 서구의 시장 논리와는 다름. 낮은 수율도 막대한 자본으로 극복하여 원하는 양을 생산해내는 방식. 공산당의 역점 기술 분야에서는 이론이 아닌 현실이며 그 중심에는 SMIC가 서있음.
- SMIC는 올해 안에 월 5만 장의 웨이퍼 생산 능력 확보를 목표로 이미 1분기부터 Ascend 910C의 대량 생산을 시작했다고 발표. TrendForce의 분석에 따르면, 지난 1년간 Ascend 칩의 수율이 20%에서 40%로 두 배나 향상된 점은 생산량 예측에 있어 가장 결정적인 변수.
- 향상된 수율을 바탕으로, 중국은 연내에 백 만개 이상의 화웨이 AI 가속기 생산이 가능할 수도 있음. 이를 시스템 단위로 환산하면, 엔비디아의 GB200 NVL72보다 일부 성능이 뛰어난 CloudMatrix 384 시스템 약 3천 개가 올해 말까지 중국 전역에 구축될 수도 있음. 이는 중국이 미국의 제재 속에서도 AI 인프라를 독자적으로 구축이 가능하다는 의미이기도 하지만, 미국 빅테크들이 NVL72 시스템을 주문하는 양(올해 NVL72 시스템 출하량은 최소 3만 개)에 비해서는 턱없이 부족한 실정.

"화웨이+SMIC 연합전선"의 웨이퍼 생산량 및 수율로 알아본 Ascend 910C의 다이 생산량 가정

반도체 생산에서 웨이퍼(실리콘 원판) 위에 Ascend 910C 칩(다이)을 얼마나 만들 수 있는지 계산해보면 대략 93개, 거기에다가 WPM과 수율 곱함

	월간 웨이퍼 생산량 (WPM)	5,000	10,000	20,000	35,000	50,000
다이(Die) 수율	5%	23,250	46,500	93,000	162,750	232,500
	10%	46,500	93,000	186,000	325,500	465,000
	15%	69,750	139,500	279,000	488,250	697,500
	20%	93,000	186,000	372,000	651,000	930,000
	25%	116,250	232,500	465,000	813,750	1,162,500
	30%	139,500	279,000	558,000	976,500	1,395,000
	35%	162,750	325,500	651,000	**1,139,250**	1,627,500
	40%	186,000	372,000	744,000	1,302,000	1,860,000
	45%	209,250	418,500	837,000	1,464,750	2,092,500
	50%	232,500	465,000	930,000	1,627,500	2,325,000

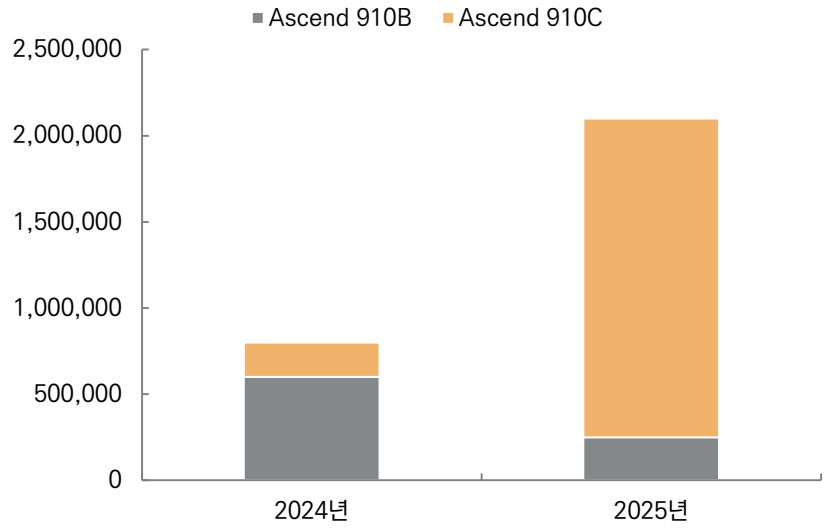
주: 웨이퍼당 총 다이 가능 숫자 ≈ ($\pi \times r^2$ / 다이 면적) – ($\pi \times r / \sqrt{(2 \times$ 다이 면적$)}$) ≈ 93 (웨이퍼 반지름 r =150mm, Ascend 910C의 다이 면적=666mm² 기준).
자료: 미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 중국의 강구책(2) 자립과 화웨이

보여주고 싶지 않은 어두운 이면

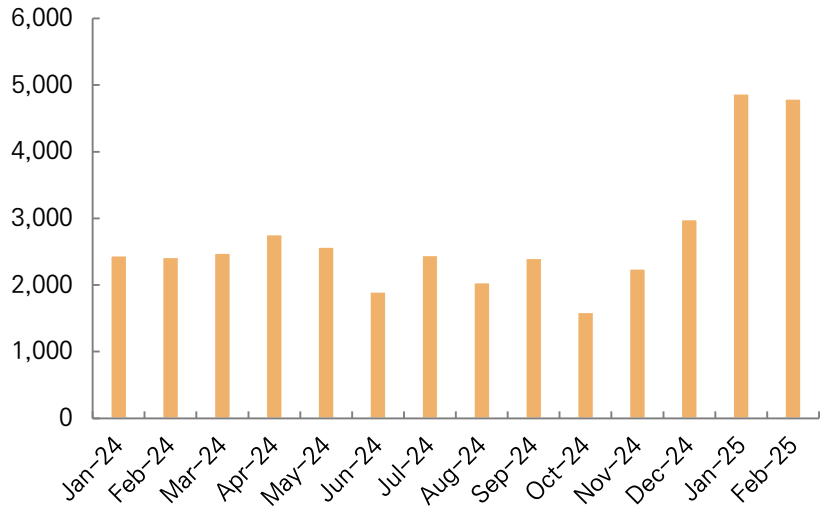
- SMIC가 7nm 공정(N+2)을 보유하고 있기는 하지만, 910B와 910C의 대부분은 TSMC의 7nm 공정으로 제작. 즉, CM384의 성능 이면에는 핵심 반도체의 높은 해외 의존도라는 아킬레스건이 존재.
- 현재 기준, SMIC의 양품 생산량은 또 다시 당초 기대에 미치지 못하고 있어 최근 어려움을 겪고 있는 것으로 파악. 패키징/HBM 접근 제약으로 '수율 저하를 돈으로만 해결'엔 한계가 있을 수 있음.
- SemiAnalysis 측에 따르면, Ascend 910C 칩은 Sophgo와 Bitmain과 같은 제 3의 대리회사를 내세워 제재를 우회하며, TSMC 7nm 로부터 200만 개 로직 다이를 공급받은 수법.
- HBM2E 역시 비슷하게 제재 회피. 대만 OSAT 업체를 통해 삼성의 HBM을 수입하게 한 뒤 다른 칩과 함께 리패키징하고, 이렇게 완성된 새로운 제품을 중국으로 수입. 이를 디솔더링 기술을 통해 HBM만 추출해 사용. 실제로 삼성 HBM의 독점 유통사인 CoAsia Electronics의 수출 통제 이후, 폭발적인 매출 증가가 이를 방증.

화웨이 칩의 TSMC 생산 비중



자료: SemiAnalysis,미래에셋증권 리서치센터

칩 수출 규제 이후 급등한 CoAsia Electronics의 매출



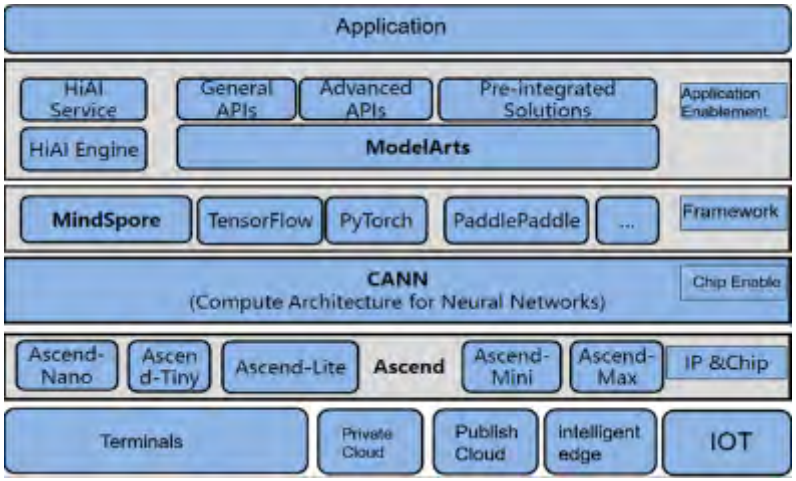
자료: SemiAnalysis ,미래에셋증권 리서치센터

VI. 소버린 AI(신냉전): 中의 강구책에도 엔비디아 칩 '패닉 바잉'은 지속

중국에서 엔비디아 대체가 어려운 이유: '차이캉즈루(踩坑之路)'

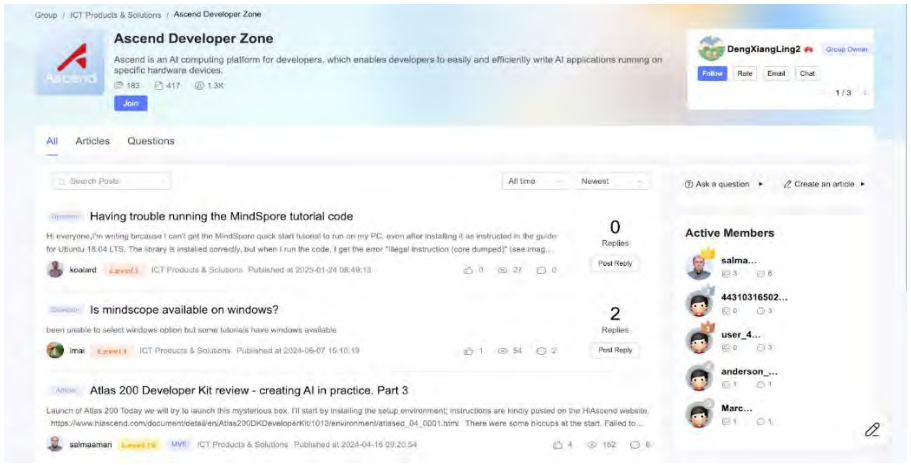
- 화웨이가 Ascend 910C에서 차세대 920 시리즈로 향상시키는 개선폭보다 엔비디아가 Hopper에서 Blackwell로 도약하는 성능 향상 폭이 클 것. 따라서 HW 성능 격차는 오히려 벌어질 가능성이 높음.
- HW보다 더 따라 잡기 힘든 것은 SW. 화웨이는 CUDA를 대체할 CANN(Compute Architecture for Neural Networks)이라는 것도 Ascend 칩과 함께 '19년부터 개발. "엔비디아 GPU+파이토치"로 작성된 코드를 최소한의 수정으로 Ascend 칩에서 실행 가능하게 하고, 엔비디아 칩으로 만들어진 모델을 Ascend에서도 쉽게 배포될 수 있도록 지원. "훈련은 엔비디아, 서비스는 화웨이로!"의 모토.
- 하지만, 정작 中 개발자들은 Ascend 사용 과정을 '온갖 시행착오로 가득한 여정(차이캉즈루)'이라 묘사. 잦은 버그와 해결책을 찾기 어려운 개발 환경이 가장 큰 문제. 활발한 포럼과 동료 개발자의 지원이 넘치는 엔비디아와 달리, 화웨이 개발자 포털은 활동이 미미. 이는 신규 개발자 유입의 큰 장벽.
- 게다가 AI 모델을 Ascend 플랫폼에 최적화하려면 화웨이의 직접적인 기술 지원이 필수적. 이는 국가적 우선순위에 따라 지원 속도가 결정되는 병목 현상을 야기. DeepSeek는 몇 달 전만 하더라도, "Ascend는 형편없는 칩과 끔찍한 지원 SW"를 가지고 있다며 사용에 관심이 없다"고 밝히기도 했음.

화웨이가 제시하는 자체 AI 컴퓨팅 프레임워크
칩(Ascend), 시스템(CM384), SW(CANN)를 아우르는 수직계열화



자료: Huawei, 미래에셋증권 리서치센터

화웨이 엔터프라이즈 지원 커뮤니티 내의 "Ascend Developer Zone" 스크린샷
개발자들이 문제 해결에 어려움을 겪지만 정보 교류가 원활하지 않은 현실



자료: Huawei ,미래에셋증권 리서치센터

VII. 중국 AI

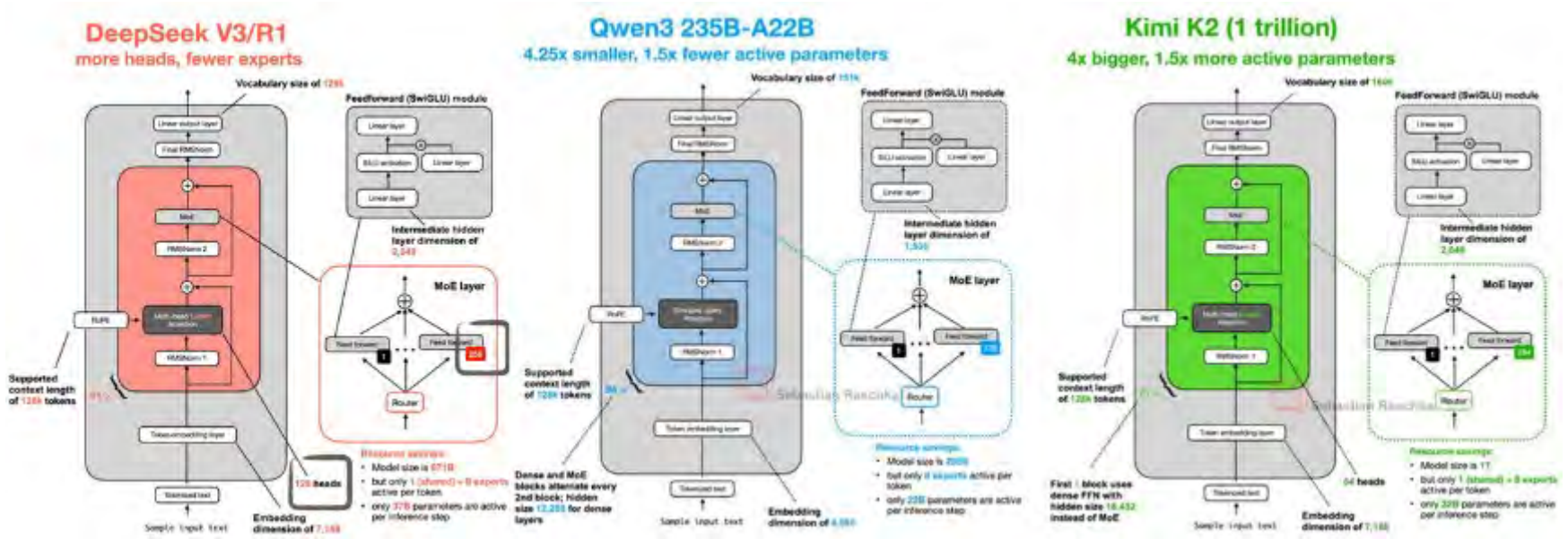
딥시크 알리바바 문샷

VII. 중국 AI: 중국의 국가 대표 AI 세 곳

DeepSeek 충격 이후, 중국 내부 경쟁도 격화.
- DeepSeek
- 알리바바
- Moonshot

- 중국의 R&D는 새로운 연구 방향을 독자적으로 설정하는 능력은 부족하나, 미국에서 귀국한 고급 인재 풀을 바탕으로 '패스트 팔로워' 역량은 세계 최고. 전 세계에 배포된 LLM 3,755개 중에서 1,509개(40% 비중)가 중국산일 정도. 세계적 수준의 오픈소스 모델들도 중국에서 연쇄적으로 등장.
- 중국 내에서 "National Champion"이 되면 애국심에 불타는 개인 투자자들과 국민적 지지를 등에 업고 장기적인 R&D 투자를 감행할 수 있음. 특히 공공기관/공공기업들은 챔피언들의 첫 번째이자 가장 큰 고객이 되어주며, 안정적인 수익 기반과 실증 데이터를 제공. 국영은행의 무제한 대출도 존재.
- 아래의 세 회사들은 중국 AI 패권이라는 하나의 목표를 향해 각자의 방식으로 전진하는 국가대표들.
  - DeepSeek: 헤지펀드의 막대한 자본과 GPU 비축을 등에 업은 '하드코어 연구 집단'의 선구자
  - 알리바바: 빅테크의 막대한 자원과 데이터를 실용주의와 결합한 '효율성의 제왕'
  - Moonshot: AGI 순수주의자 창업자의 비전 아래 지속확장 가능한 AI 모델 생산 '시스템'을 구축

DeepSeek의 V3, 알리바바의 Qwen3 235B-A22B, Moonshot의 Kimi K2의 아키텍처 비교, '서로의 성공과 실패를 학습하면서 비슷하게 발전 중'



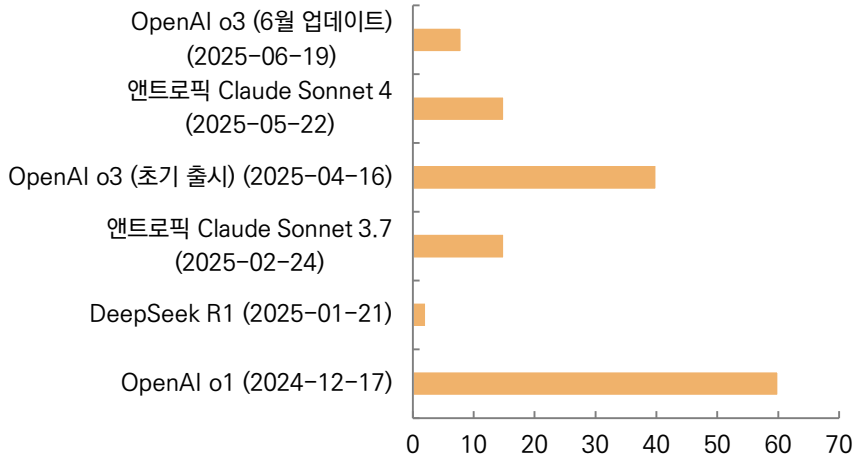
자료: Sebastian Raschka("Build a Large Language Model From Scratch"의 저자), 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

게임 체인저, 딥시크: 미국의 기술 안일함에 경종을 울리다

- 2025년은 AI 에이전트의 해일뿐만 아니라, 중국 DeepSeek가 미국의 기술적 안일함에 경종을 울린 해로 기록. DeepSeek R1은 올해 1월 당시 최고 모델이던 OpenAI o1과 '추론 능력' 면에서 대등한 성능을 보여준 최초의 오픈소스 모델. 이는 중국이 기술적으로 최전선(frontier)에 도달했음을 의미. 미국과 1년 이상 벌어졌다고 여겨졌던 AI 기술 격차가 6개월 이내로 좁혀졌다는 공포.
- 또한 중국이 AI 시장을 저가 상품화(commoditize) 시킬 수 있다는 공포도 커짐. 특히, 이 강력한 모델을 누구나 쓸 수 있도록 오픈소스로 공개하고, 공식 API 가격마저 OpenAI o1의 10%도 안 되는 수준으로 책정한 것은, 시장 전체를 가격 인하 경쟁으로 몰아넣었을 정도.
  - 2024년 12월: OpenAI o1 – $60/백만 토큰
  - 2025년 1월: DeepSeek R1 – $2.5/백만 토큰
- DeepSeek 앱은 R1 모델 출시 후 앱스토어 1위 기록. Anthropic의 Claude조차 달성하지 못한 성과. 다만, DeepSeek의 점유율은 2월에 정점을 찍고 이후 계속 하락하는 모습을 보인 것은 DeepSeek가 직접 운영하는 서비스는 사용자 경험이 떨어지게 '의도적으로' 설계됐기 때문. 제 3자 플랫폼(OpenRouter 등)을 포함한 전체 토큰 사용량은 R1 출시 이후 20배 급증. 여전히 인기 많음.

추론(Reasoning) 모델간 출력 비용, 'OpenAI 가격 인하는 시장압박 때문'



자료: plainconcepts, team-gpt, apidog, pricetimeline, 미래에셋증권 리서치센터

주요 AI 서비스 웹 트래픽 변화

챗봇 도메인	2월 방문자 수	5월 방문자 수	변동률
DeepSeek	**6억 1,400만**	**4억 3,600만**	**-29%**
ChatGPT	39억 500만	54억 9,200만	41%
Claude	7,300만	1억	37%
Gemini	2억 8,400만	5억 2,800만	86%
Grok	5,100만	1억 7,900만	247%

자료: SimilarWeb, 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

딥시크의 진짜 얼굴, 훈련 비용의 진실

- 그런데 DeepSeek은 평범한 스타트업은 아님. 중국 최대 퀀트 헤지펀드 '환방량화(HF: High-Flyer)'라는 배경 자체가 이들의 압도적인 출발선을 설명. 수십 조원이 넘는 운용자산을 가진 모회사 덕에 DeepSeek는 외부 투자나 수익 증명 압박 없이, 오직 연구개발에만 몰두할 수 있는 환경. PR 부서나 영업팀이 전무. 오직 연구와 개발에만 집중하는 하드코어 문화는 일론 머스크의 xAI와 유사.
- 창업자 량원펑은 AI 스케일링의 중요성을 일찍이 간파. 미국의 수출 제한 한참 이전인 2021년에 1만 개의 A100 GPU를 확보. 당시 중국 최대 규모. 즉, DeepSeek 쇼크의 핵심이었던 '저비용 훈련' 신화는, 의도적으로 축소된 마케팅. V3 논문에서 558만 달러밖에 안 들었다는 언급은 '마지막 훈련' 비용만을 설명. 수개월의 아이디어 고안/폐기 비용은 제외. 대개 아이디어 중 5%만이 실제로 채택됨.
- 또한, V3 논문과 달리 R1 논문에서는 컴퓨팅 자원량을 의도적으로 누락. 이는 V3보다 훨씬 많은 GPU, 특히 불법적 경로로 획득했을 가능성이 있는 H100 등을 사용했음을 감추기 위한 전략으로 해석됨.
- 다만, GPU 보유량은 최고의 인재들을 끌어들이는 자석으로 작동. 게다가 한화로 거의 2억원의 석사 초봉을(중국 일반 석사 초봉의 10배, AI 석사 초봉의 5배) 제시하며 사실상 중국의 천재들을 싹쓸이.

DeepSeek AI TCO 계산, 4년 동안의 총 소유 비용(TCO)은 25억 7,300만 달러(약 3조 5천억 원)로 추정
DeepSeek도 미국 AI 연구소들처럼 GPU가 별로 없는 채로 결과를 냈던 것은 아님. DeepSeek의 GPU 보유량 비공개는 오히려 절실함을 방증.

항목(unit)	A100	H20	H800	H100	전체
감가상각연수(년)	4	4	4	4	
GPU 숫자(개)	10,000	30,000	10,000	10,000	60,000
엔비디아 GPU 평균판매가 (달러)	$13,500	$12,500	$20,000	$23,000	
GPU당 서버 설비 투자금액 (달러)	$23,716	$24,228	$31,728	$34,728	
총 서버 설비 투자금액 (백만달러)	$237	$727	$317	$347	$1,629
운영 비용 (백만달러)	$157	$387	$170	$230	$944
4년에 걸친 총 소유비용(백만달러)	$395	$1,114	$487	$577	$2,573

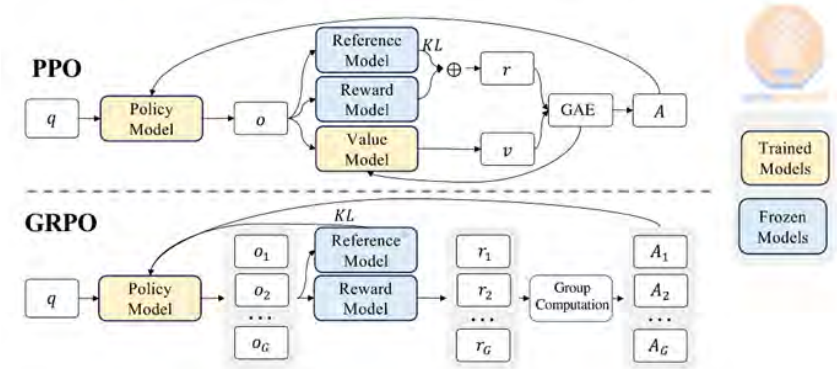
주: TCO는 서버 자본 비용을 4년 동안 감가상각하고, 13.3%의 WACC(가중 평균 자본 비용)를 적용하여 계산
자료: SemiAnalysis, 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

절박함이 낳은 극한의 최적화

- 미국의 제재로 인한 GPU 부족이라는 환경은 DeepSeek에게 '맨 밑바닥부터 뜯어고치는' 극한의 엔지니어링 혁신을 강요. 대부분의 기업이 CUDA 레벨에서 개발할 때, DeepSeek는 엔비디아 GPU의 핵심 연산 단위(SM)를 직접 쪼개어 통신 작업에만 따로 할당하는 식. 이를 위해, 어셈블리어에 가까운 PTX 레벨에서 코드를 커스터마이징. 가진 자원을 최대한 효율적으로 활용해 경쟁자를 압도하려는 절박함과 창의성의 산물. 이 엔지니어링 역량 자체가 강력한 경쟁 우위.
- DeepSeek R1의 핵심은 SFT(지도 미세 조정) 의존에서 벗어나, 순수 강화학습(RL)을 통해 추론 능력을 스스로 터득했다는 점. 특히 인간 데이터 없이 순수 RL로만 훈련된 R1-Zero 모델의 존재는 'LLM 강화학습 시대'의 개막을 알리는 신호탄. 이를 위해 주류 방법인 PPO(OpenAI의 방식)를 쓰지 않고, 더 단순하지만 GPU 병렬 처리에 최적화된, "GRPO" 알고리즘을 독자 개발해 활용함.
- 이외에도 V3 모델에서부터 쌓아온 수많은 주요 혁신 포인트들은 오픈소스로 공개되었기 때문에, 중국 등지의 엔지니어들의 귀감이자 자극제가 되었음.

R1의 핵심은 강화학습, 그것의 효율성을 극대화하기 만든 핵심은 GRPO
물론, GRPO는 병렬 샘플 생성·평가 기법이라 대량 GPU 동시 활용이 전제



가장 큰 차이는 미래 보상을 예측하는 '비평가 모델(=Value Model)'의 유무.
DeepSeek의 GRPO는 비평가를 과감히 제거, 대신 여러 답변을 한번에 생성 후 답변들간 '상대평가'로 점수 계산. 별도의 모델을 훈련시킬 필요가 없어, 자원 및 메모리를 절약.
자원이 풍부한 OpenAI(PPO)와, 효율성이 절실한 오픈소스 진영(GRPO)의 전략적 차이

자료: DeepSeek, SemiAnalysis, 미래에셋증권 리서치센터

R1 모델 이전인 DeepSeek-V3(non-reasoning 모델)에서의 주요 혁신들

혁신 영역	핵심 기술	주요 이점
아키텍처 설계	- MLA (압축된 KV 캐시) - DeepSeekMoE (동적 편향 조정)	- 메모리 사용량 감소 - 추론 속도 향상 - MoE 활용 효율성 최적화
분산 훈련 최적화	- 동적 워프 스케줄링 커스터마이징	- 파이프라인 병렬화 효율 향상 - IB/NVLink 대역폭 최대화 - 계산-통신 오버헤드 최소화
메모리 최적화	- RMSNorm 재계산 - MLA 업-프로젝션 - MTP 모듈 공유	- 활성화 값 저장시 메모리 절약 - 중복 메모리 할당 방지 - 추측적 디코딩 가능
정밀도 최적화	- FP8 훈련 - 동적 스케일링	- 메모리 대역폭 요구 감소 - 계산 효율성 향상 - 정밀도 손실 최소화

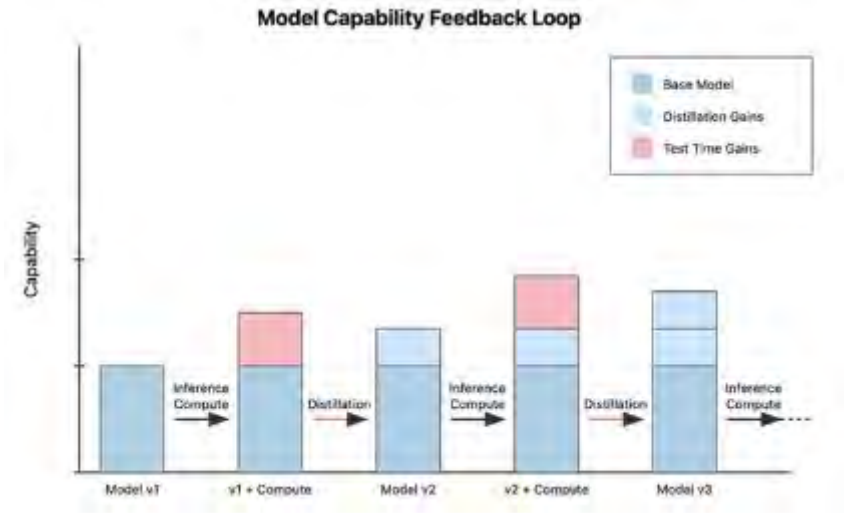
자료: DeepSeek, 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

최고의 선생님도
누군가에게는
여전히 학생일 수 있다

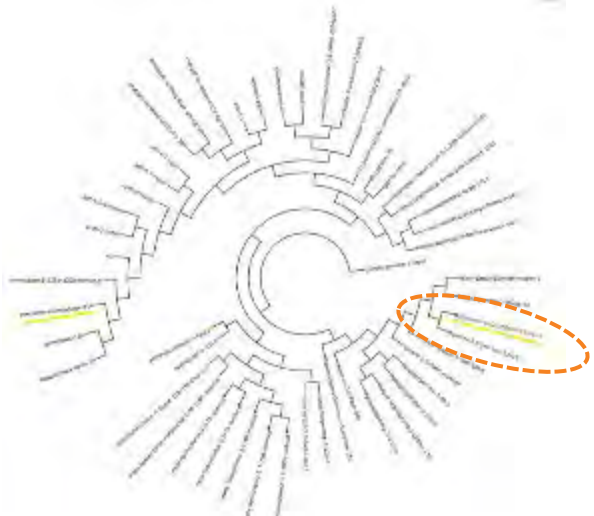
- R1은 스스로 강력해질 뿐만 아니라, 그 능력을 다른 작은 모델에 전파하는 것에 특화.
- R1을 가지고 80만 개의 고품질 데이터를 생성하고, Qwen이나 Llama 같은 더 작은 모델을 대상으로 추가 강화학습을 하는 방식. 즉, '지식 증류'에 특화된 것으로, AI 모델 훈련의 민주화를 가능하게 하는 핵심 전략. DeepSeek R1 논문에서도 이것이 가장 강조되어 있음.
  * 지식 증류는 정확하지만 크고 느리고 비싼 '교사' 모델의 지식을 더 작고 빠르고 저렴한 '학생' 모델로 전달하는 것
- 그런데, DeepSeek의 눈부신 성과 뒤에는 역설적으로 미국의 프런티어 모델을 활용하는 '지식 증류' 전략이 숨어있음. DeepSeek에 대한 '모델간 유사도 분석'을 했더니 R1과 가장 유사한 모델이 구글의 Gemini 2.5 Pro로 밝혀진 것. 중국의 선두주자가 여전히 미국의 프런티어 모델에 의존하고 있으며, 이는 양국 간에 최소 '한 분기 이상'의 세대 격차가 존재함을 방증한다고 판단.

모델의 성능이 어떻게 향상되는지의 핵심: 증류와 테스트 타임(=추론)
작은 모델에 효과적으로 지식을 전달하고 추론 스케일링으로 성능 향상!



자료: X(@techno_guile), 미래에셋증권 리서치센터

챗봇간의 유사성을 기반으로 만들어진 챗봇들의 "계보 트리"
R1-0528모델은 구글의 모델과 가깝고 이는 구글로부터 증류 받았다는 것



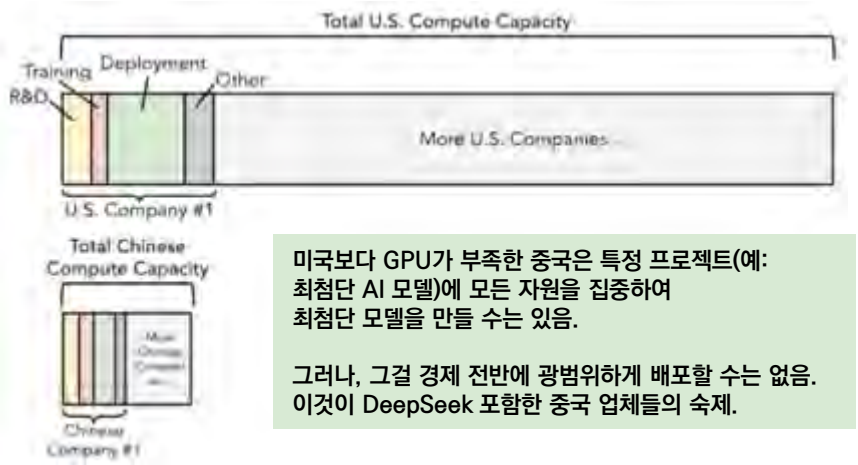
자료: Sam Paech(EQ Bench 창시자), 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

DeepSeek의 진짜 병목, '추론 컴퓨팅'

- R1은 결국 RL(강화학습)의 성숙에서 나옴. 이는 컴퓨팅 파워의 필요성을 더 키움. 즉, DeepSeek의 성공은 중국의 기술력을 과시했지만 컴퓨팅 파워 부족이라는 현실을 드러냄. 또한, 모델 훈련보다 더 큰 문제는 그 모델을 수많은 사용자에게 서비스하는 추론(inference) 단계. 엔비디아 CFO는 "추론(reasoning) 모델은 단순 답변 생성 모델보다 100배 많은 컴퓨팅 파워를 요구"한다고 발언. CEO인 젠슨 황도 애초에 "Grace Blackwell"이 나온 것도 이에 대응하기 위해서라고 강조.
- 메타는 Llama-3 훈련에 1.6만 개의 GPU를 썼지만, 핵심 서비스에는 40만 개 이상 GPU를 운영 중. DeepSeek R1이 독자적으로 대중 서비스를 하려면 GPU 부족으로 가랑이가 찢어질 수 있음.
- DeepSeek은 모델 서비스(=배포, 추론)라는, 자원을 많이 소모하는 부담을 오픈소스 커뮤니티와 제3자 플랫폼에 사실상 떠넘긴 상태. 이를 통해 확보한 자신들의 귀중한 GPU 자원은 외부 서비스 대신, 오롯이 차세대 모델 R2 훈련과 AGI 연구개발이라는 핵심 목표에만 집중. 미국의 제재로 컴퓨팅 자원 확보가 어려운 상황에서, 이는 가장 영리하고 합리적인 선택.

양국 모두 컴퓨팅 자원의 가장 큰 비중은 모델 훈련(Training)'이 아닌,
실제 서비스를 위한 '배포(Deployment, 추론)'에 소모

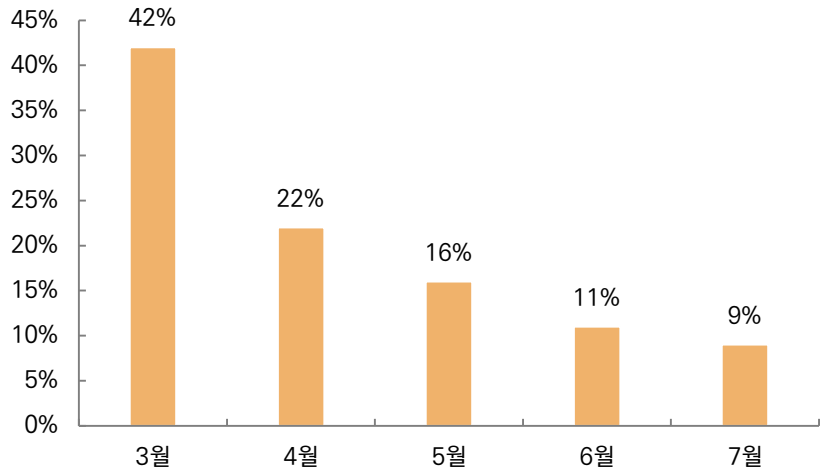


미국보다 GPU가 부족한 중국은 특정 프로젝트(예: 최첨단 AI 모델)에 모든 자원을 집중하여 최첨단 모델을 만들 수는 있음.

그러나, 그걸 경제 전반에 광범위하게 배포할 수는 없음. 이것이 DeepSeek 포함한 중국 업체들의 숙제.

자료: Lennart Heim, 미래에셋증권 리서치센터

DeepSeek 모델 전체 사용량 중에서,
DeepSeek이 '직접 제공'하는 API가 차지하는 비중은 급격히 축소 중



자료: SemiAnalysis, Capacity Media, 미래에셋증권 리서치센터

VII. 중국 AI: 혁명가, DeepSeek

DeepSeek의 성공이 엔비디아를 웃게 한 이유

- DeepSeek의 '저비용' 모델 등장은 역설적으로 GPU 수요를 폭발시킴. AI 기술의 효율성 증가는 결국 총 사용량 증가로 이어지는 '제본스의 역설'.
- Anthropic CEO 다리오 아모데이는 화웨이 Ascend 칩이 엔비디아 GPU보다 성능이 훨씬 떨어지며, 중국 외부에는 의미 있는 Ascend 클러스터가 없다고 지적. 이는 중국 내수용으로도 화웨이 칩조차 구하는 것이 벅차다는 의미. 2024년 중국 AI 가속기 시장에서 추론용 칩 비중은 57.6%로 훈련용 칩(33%)을 이미 압도한 상황. 앞으로 이 추세는 더욱 심화될 수밖에 없는 것.
- 그런데 DeepSeek가 내놓은 혁신들은 역설적으로 엔비디아의 생태계를 더욱 공고히 하는 공생 관계를 형성하고 있음. DeepSeek이 개발한 분산 추론 최적화 노하우는 엔비디아의 소프트웨어 스택 "Dynamo"에 공식적으로 통합. 엔비디아는 경쟁자의 혁신을 흡수하며 자사의 범용성을 더욱 강화.
- DeepSeek의 성공은 중국의 기술력을 과시하는 동시에, '결국 엔비디아 칩과 소프트웨어 생태계 안에서 이룬 성공'이라는 한계를 명확히 보여줌. 이와 비슷하게 중국 기업들은 여전히 엔비디아 칩을 절실히 원함. ByteDance, Tencent 등은 H20 칩 라이선스 복귀에 따라 160억 달러 이상 주문. 이에 대한 엔비디아의 전략은 다운그레이드 칩(B30/B40)을 화웨이 칩보다 저렴하게 공급, 화웨이의 성장을 견제하고 중국의 CUDA 종속을 유지하는 것.

DeepSeek가 오픈소스로 공개한 AI 추론 최적화 기술은 엔비디아 GPU를 이롭게 한다
DeepSeek이 개발한 고급 AI 추론 최적화 기술을 엔비디아가 자사의 Dynamo 도구로 공식적으로 통합한 것

기능	설명	이점
스마트 라우터	토큰을 지능적으로 분산	병목 현상 방지, 로드 밸런싱
분산 추론 최적화를 위한 GPU 플래닝	프리필과 디코드 노드 자동 확장	리소스 활용도 최대화
개선된 NCCL	4배 낮은 지연 시간	메시지 처리 속도 향상
NIXL	CPU 우회 직접 전송	전송 속도 향상
NVMe에 KV 캐시 저장하고 재사용	대화 컨텍스트 저장	56.3% 히트율로 효율성 향상

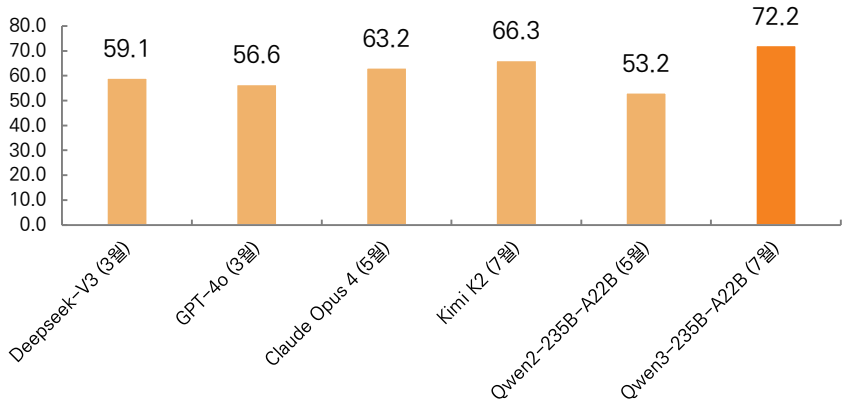
자료: 엔비디아, DeepSeek, 미래에셋증권 리서치센터

VII. 중국 AI: 알리바바의 Qwen

구글 Gemini와 동급? 오픈소스의 왕, Qwen

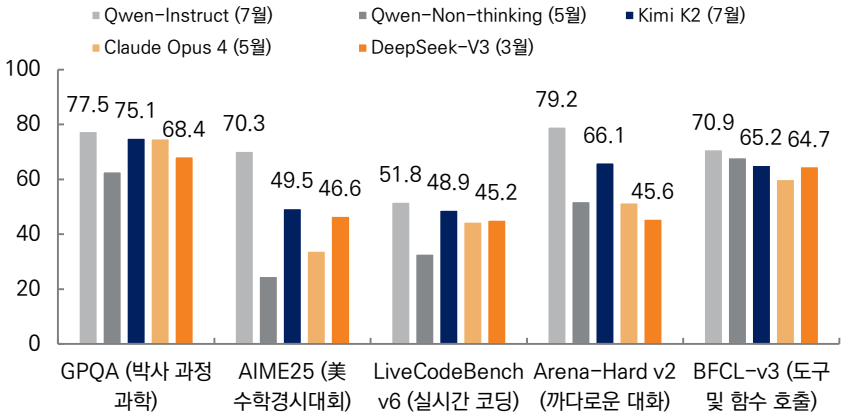
- 알리바바의 Qwen은 화려한 수사 대신, 효율성과 실용주의를 무기로 AI 패권 경쟁의 최전선에 섬. 오픈소스 진영의 기술적 기대치를 끊임없이 끌어올리는 가장 위협적인 경쟁자 중 하나.
- 알리바바가 2025년 7월 21일 내놓은 Qwen3-235B-A22B는 생각 없이 ARC-AGI-1에서 41%를 기록. 이는 Gemini 2.5 Pro, Claude Opus 4와 비슷한 수준. 활성 파라미터가 220억 개로, 320억 개의 활성 파라미터를 가진 Kimi K2(Moonshot)를 주요 5개 벤치마크에서 격파. Qwen3는 무려 36조 토큰의 데이터셋으로 학습. 이는 이전 버전의 2배 규모이자 Kimi K2보다 2.3배 많은 것.
- 알리바바는 강력한 범용 모델(Qwen3-235B-A22B)로 기반을 다지고, 특정 분야에 특화된 킬러 모델(Qwen3-Coder)로 시장을 지배하는 이중 전략을 구사하고 있음. '하이브리드 추론(추론 모델인지 아닌지를 사용자가 선택)' 같은 복잡한 기능을 과감히 포기하고, 각 용도에 맞는 성능을 극한까지 최적화하는 데 집중. 시장의 반응에 따라 빠르게 전략을 수정하는 유연성을 보유.
- 그들의 방식은 DeepSeek 방식과 비슷한 듯 다름. Qwen3는 128개의 전문가(expert)를 사용하는 대신, 94개의 더 깊은 트랜스포머 블록을 쌓아 모델의 표현력을 확보. MLA 대신, 더 효율적인 GQA를 채택. 이는 모델의 모든 구성 요소를 경쟁 모델보다 작게 설계, 철저하게 효율을 추구하는 것.

각 AI 모델들이 수십 개의 벤치마크 점수에서 기록한 평균 점수



자료: 각 회사, 미래에셋증권 리서치센터
주: 지식, 추론, 코딩, 정렬, 에이전트 관련 21개의 벤치마크 점수 평균

각 AI 모델들이 중요 벤치마크에서 기록한 평균 점수



자료: 각 회사, 미래에셋증권 리서치센터

VII. 중국 AI: 알리바바의 Qwen

적어도 코딩 분야에서, 오픈소스는 더 이상 상용 모델의 추격자가 아니다

- Qwen은 성능 뿐만 아니라 접근성과 확장성에서도 경쟁 모델을 압도하며, 사실상 오픈소스 생태계의 표준으로 자리매김하고 있을 정도. Apache 2.0 라이선스로 상업적 사용을 포함해 완전 오픈소스. vLLM, Ollama, MLX 등 주요 추론 엔진과 OpenRouter 같은 클라우드 플랫폼에서 즉시 사용 가능. 2025년 6월 기준 이미 1,250만 회 이상 다운로드.
- 지속적인 멀티모달(Qwen-VL, Qwen-Audio) 업그레이드와, 텍스트, 시각, 오디오를 통합 처리하는 'Omni-Model' 개발을 향해 나아가는 중. 그리고 최근 가장 두드러진 성과는 Qwen3-Coder.
- Qwen이 이번엔 'Qwen3-Coder-480B-A35B'라는 코딩 전문 모델은 전체 4,800억, 활성 350억 파라미터의 더 큰 MoE 모델. Kimi-K2를 가볍게 뛰어넘고, Anthropic의 Claude Sonnet-4와 대등, 일부 벤치마크에선 OpenAI의 GPT-4.1마저 능가. 오픈소스와 상용 모델의 경계가 희미해짐.
- 그러나 화려한 벤치마크와 달리 Kimi K2대비 코딩시 비용 효율성이 크게 떨어진다는 경험담도 존재

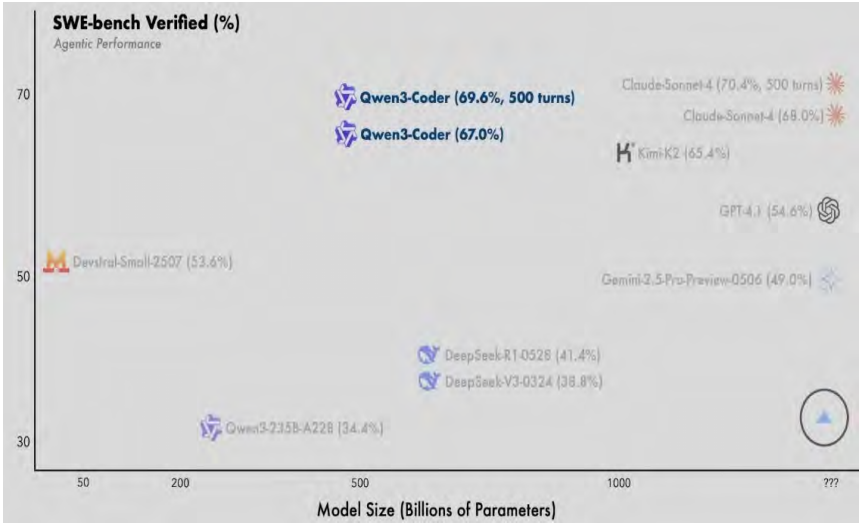
Qwen3-Coder의 성능을 다른 주요 AI 모델들과 직접 비교
코딩 에이전트로서의 실용성에서 현존 최고의 모델들과 어깨를 나란히

> QWEN3-CODER

Benchmarks	Open Models: Qwen3-Coder 480B-A35B-Instruct	Open Models: Kimi-K2 Instruct	Open Models: DeepSeek-V3 0324	Proprietary Models: Claude Sonnet-4	Proprietary Models: OpenAI GPT-4.1
Agentic Coding					
Terminal-Bench	37.5	30.0	2.5	35.5	25.3
SWE-bench Verified					
w/ OpenHands, 500 turns	69.6	-	-	70.4	-
w/ OpenHands, 100 turns	67.0	65.4	38.8	68.0	48.6
w/ Private Scaffolding	-	65.8	-	72.7	63.8
SWE-bench Live	26.3	22.3	13.0	27.7	-
SWE-bench Multilingual	54.7	47.3	13.0	53.3	31.5
Multi-SWE-bench mini	25.8	19.8	7.5	24.8	-
Multi-SWE-bench flash	27.0	20.7	-	25.0	-
Aider-Polyglot	61.8	60.0	56.9	56.4	52.4
Spider2	31.1	25.2	12.8	31.1	16.5
Agentic Browser Use					
WebArena	49.9	47.4	40.0	51.1	44.3
Mind2Web	55.8	42.7	36.0	47.4	49.6
Agentic Tool Use					
BFCL-v3	68.7	65.2	56.9	73.3	62.9
TAU-Bench Retail	77.5	70.7	59.1	80.5	-
TAU-Bench Airline	60.0	53.5	40.0	60.0	-

자료: 알리바바, 미래에셋증권 리서치센터

SWE-bench Verified(실제 SW 엔지니어링 문제해결 능력) 성능 비교
이전 오픈소스 최강자였던 Kimi-K2나 GPT-4.1, Gemini-2.5-Pro를 압도



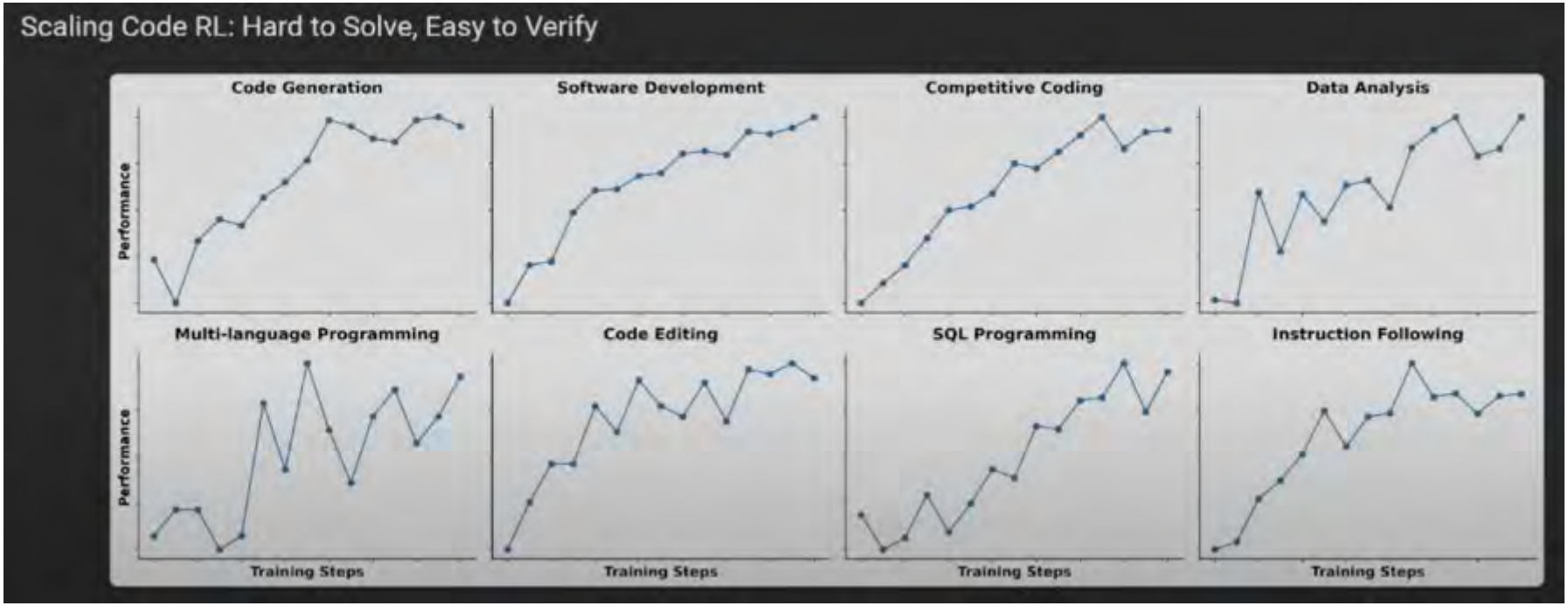
자료: 알리바바, 미래에셋증권 리서치센터

VII. 중국 AI: 알리바바의 Qwen

생성자-검증자 격차를 활용한 강화학습 최적화 "Hard to Solve, Easy to Verify"

- Qwen3 Coder의 압도적 성능은 모델 크기가 아닌, 훈련에 대한 근본적인 철학의 전환에서 비롯. 알고리즘 문제 풀이 같은 '경쟁 수준의 코드 생성' 대신, 버그 수정, 기능 추가, 리팩토링 등 '실제 세계의 코딩 과제'에 집중. 모델이 단순히 코드를 생성하는 것을 넘어, 계획-도구 사용-피드백-의사 결정의 복잡한 상호작용을 통해 문제를 해결하도록 훈련하는 '에이전트 강화학습'을 도입.
- 강화학습의 핵심은 보상 설계. 알리바바는 '풀기는 어렵지만, 검증은 쉬운' 과제 중심으로 훈련 집중. 이를 통해 모델은 모호함 없이 명확한 피드백을 받으며 효과적으로 학습. 이 훈련 방식 덕분에, 특정 벤치마크 점수만 높은 모델이 아닌, 진짜 문제 해결 능력을 갖춘 '육각형 모델'이 되었다고 평가.

Qwen3-Coder의 '강화학습(RL)' 사후훈련 과정의 효과 – 코드 생성, 소프트웨어 개발, 데이터 분석 등의 여러 코딩 능력들이 훈련단계가 진행되며 향상
모델이 특정 문제 유형만 암기하는 것이 아니라, 훈련을 거듭할수록 다양한 실용적인 영역에서 문제 해결 능력이 전반적으로 향상됨을 보여줌



자료: 알리바바, 미래에셋증권 리서치센터

VII. 중국 AI: 알리바바의 Qwen

복제 불가능한 해자 – 알리바바 클라우드의 압도적 인프라

- 에이전트 강화학습은 강력하지만, 대규모 가상 환경을 구축하고 운영해야 하는 치명적인 병목이 존재. 알리바바는 이 문제를 자신들의 심장인 클라우드 인프라로 해결. AI 경쟁의 승패는 결국 누가 더 강력하고 확장 가능한 훈련/추론 인프라를 가졌는지에 따라 갈릴 것.
- 알리바바 클라우드 위에 '2만 개의 독립적인 코딩 시뮬레이터'를 병렬로 실행하는 시스템을 구축. 이는 자체 인프라가 없는 경쟁사들은 엄두 내기 힘든 대규모 강화학습을 현실화시킨 결정적인 기술적 해자.
- 결국 미래 AI 경쟁의 핵심은 복잡한 강화학습을 대규모로 실행할 수 있는 컴퓨팅 능력과 시스템 구축 역량에 있음을 증명. 모델 아키텍처나 알고리즘만큼, 혹은 그 이상으로 인프라가 중요해진 것.
- 더 중요한 사실은, Qwen3의 성능이 '추론 확장(test-time scaling)' 없이 달성되었다는 점. 일부 모델들은 좋은 성능을 내기 위해 추론 시점에 생각의 사슬(CoT)을 많이 실행해야 하지만, Qwen3 Coder는 그런 기교 없이도 그 자체로 뛰어난 성능을 발휘. 이는 실제 사용 환경에서 훨씬 더 효율적이고 실용적이라는 의미. 알리바바의 클라우드 인프라의 컴퓨팅 부담도 줄여주는 효율적 요소.

알리바바 클라우드 주요 지표 및 전망

구분	수치/현황	기준 시점	세부 내용
중국 AI 인프라 시장 점유율	23%	2024년	AI IaaS 시장 1위, IDC 조사
중국 내 클라우드 시장 점유율	33%	2025년 Q1	중국 1위이지만 2018년에는 42%였음
글로벌 클라우드 시장 점유율	4%	2025년 Q1	글로벌 4위
중국 클라우드 시장 규모	$11.6B	2025년 Q1	16% YoY
알리바바 클라우드 연간 성장률	15%	2025년 Q1	중국 내 성장
연간 투자 계획	$52.7B	2025년 Q1	현재 87개 가용성 영역, 29개 region 확보

자료: 알리바바, 미래에셋증권 리서치센터

VII. 중국 AI: Moonshot의 Kimi

DeepSeek를 존경하는 또 다른 챔피언, Kimi

- Moonshot AI(月之暗面)은 2024년 알리바바로부터의 대규모 투자를 포함, 총 10억 달러 유치.
- 헤지펀드 자금으로 독자 연구 환경을 구축한 DeepSeek와 달리, Moonshot은 주주 가치 증명 압박 속에서 모델을 개발해야 하는 환경. 창업자 양즈린은 칭화대/카네기멜론, 前 메타/구글브레인 출신.
- 양즈린이 갖고 있는 AI 세계관 중심에는 긴 컨텍스트(long-context)가 있음. 그는 "AGI를 달성하려면 긴 컨텍스트가 결정적인 요소가 될 것"이라 단언한 바 있음. 에이전트 개발에 사활을 걸겠다는 뜻.
- Moonshot은 초기부터 B2B가 아닌 개인 사용자(B2C)에 집중한 기업. 이에 따라 Moonshot은 2024년 광고비 지출 1위였던 기업.
- 그러나, DeepSeek 출현에 충격 받아 올해 2월 광고를 중단하고 기반 모델 개발에 집중. DeepSeek의 성공을 통해 "성능이 최고의 광고"임을 학습. 문샷 AI의 엔지니어들도 중국의 지식문답 커뮤니티인 Zhihu에서 “우리의 Kimi 모델은 DeepSeek 없이는 존재할 수 없었다”고 솔직하게 인정.

Moonshot AI에 대한 기업 개요

항목	내용
설립	2023년 3월, 베이징
인력 규모	2025년 7월 기준 350+명, 연구개발 인원 80% 이상
창립 멤버	양즈린(杨植麟·CEO) – 前 알리바바·Tsinghua NLP 연구원
	주신위(周欣雨) – 前 카카오브레인·딥마인드 연구원
	우위신(吴雨昕) – 前 알리바바 DAMO 아카데미 연구원
비전	"긴 문맥·멀티모달·확장 가능한 범용 아키텍처로 AGI 달성”
핵심 제품	Kimi Chat: 개인 사용자용 챗봇. 최대 200만 중국어 문자(≈140만 토큰) 컨텍스트 지원
	Kimi 엔진: 기업용 API(텍스트·코드·문서 분석)
	엔터프라이즈용 OEM: 금융·인터넷·정부 기관에 사내 LLM 배포
모델 라인업	Kimi-1.5(2025년 1월) ➜ Kimi-2(2025년 7월, 4백만 컨텍스트 길이) ➜ 에이전트(2026년)
주요 공학적 성취	합성데이터 엔진, MoE 최적화, 자체 개발 옵티마이저, 긴 문맥 처리를 위한 RoPE 최적화
하드웨어 보유량	엔비디아 H800/H100 약 2천 여 대 운영, 데이터 주권 확보
파트너십	알리페이·디디추싱의 콜센터 파일럿 진행
투자·가치	2024년 3월 Series B에서 3억 달러 조달, 기업가치 33억 달러
주요 투자자	알리바바(지분율 36%), 텐센트(지분율 9% 추정), HongShan(Sequoia China), 미국의 Coatue 등
주요 경쟁	중국 AI 챗봇 시장에서 Zhipu AI와 함께 3,500만 MAU 달성. 바이두, 바이트댄스, 텐센트 등과 경쟁

자료: 미래에셋증권 리서치센터

VII. 중국 AI: Moonshot의 Kimi

DeepSeek V3를 계승한 비추론(non-reasoning) 모델 K2를 출시

K2의 성공은 중국 AI가 세계적 수준임을 또 증명

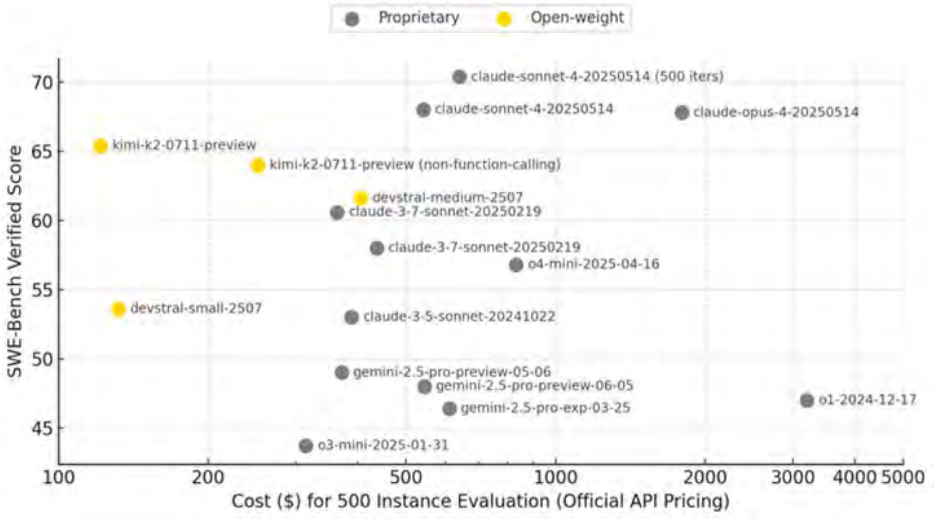
- 2025년 7월 11일, 15.5조 개의 토큰으로 학습된 '1조 파라미터의 오픈소스' 모델 Kimi K2 공개.
- 놀라운 것은 15.5조 개의 토큰으로 사전 훈련하는 동안 단 한 번의 훈련 스파이크도 발생하지 않았다는 점. 이는 OpenAI와 구글 출신의 유명 엔지니어조차 극찬할 정도.
- DeepSeek의 V3 모델의 검증된 아키텍처를 채택하여, 모델 크기 확장과 "Muon 옵티마이저"라는 2가지 핵심 변수에만 집중. MoE 모델 훈련 시 발생하는 문제점들을 해결하는 실용주의적 선택.
- 그럼에도 DeepSeek가 선보인 '계층적 라우팅'(GPU 부족할 때 쓰는 타협안)을 폐기, 인프라 개선을 통해 성능 저하를 감수하지 않는 선도적 모습도 존재. 또한, "희소성(Sparsity)도 스케일링 법칙이 유효하다"는 발견에 근거하여 더 세분화되고 전문적인 지식 활용하기 위해 전문가(Experts) 숫자를 384개까지 늘려버림. 이 모든 것은 MoE 엔지니어링의 효율성을 극한까지 올리기 위한 노력의 결과물.

2024년 중국 AI 서비스들의 사용자 확보비용 비교
작년에는 마케팅 많이 했지만 올해부터는 AGI 개발에 몰두

서비스 이름	마케팅 및 광고 집행 비용	개발사/운영사
Kimi(키미)	**1,030억 원**	**Moonshot AI**
豆包(더우바오)	767억 원	ByteDance
星野(싱예)	326억 원	베이징 싱예 비트 테크
元宝(위안바오)	283억 원	소우거우(搜狗)의 창업자
跃问(위에원)	171억 원	StepStar
清言(칭옌)	105억 원	ListenAI
星火(싱훠)	95억 원	iFLYTEK
祕塔(미타)	66억 원	Mita
通义(통이)	11억 원	알리바바
文心一言(원신이옌)	6억 원	바이두

주: 환율 기준: 1위안 = 190원 기준으로 계산
자료: AppGrowing, 미래에셋증권 리서치센터

Kimi K2는 SWE-Bench 점수 기준 유사 성능의 Claude Sonnet-4보다 3~4배 저렴
비용과 성능 측면에서 오픈소스 모델의 직접적인 압박이 거세다는 것을 증명



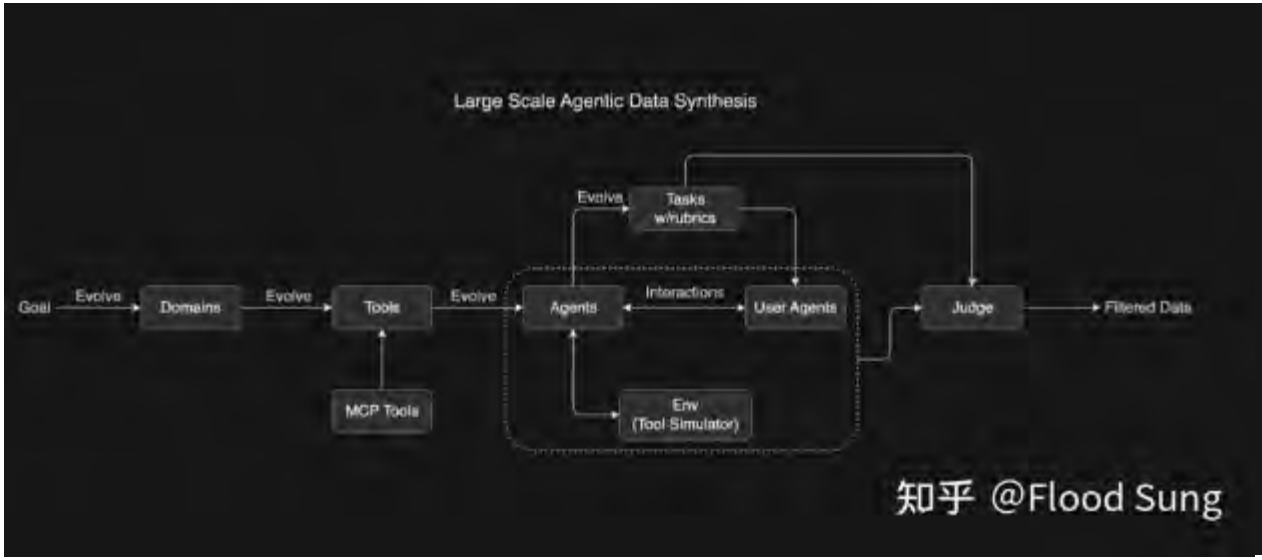
자료: Moonshot AI의 Kimi K2 기술 보고서, 미래에셋증권 리서치센터

VII. 중국 AI: Moonshot의 Kimi

자체 진화 데이터 공장과 지능형 에이전트의 서막

- Moonshot은 훈련 스택의 최하단 알고리즘 뿐만 아니라 합성 데이터 생성도 제어하는 풀스택 역량.
- 그들의 K2 모델은 코딩(상용화 수준의 SaaS 랜딩 페이지, 게임, 3D 시뮬레이션을 한 번만에 생성) 능력과 창의적 글쓰기(시 작성) 능력은 DeepSeek을 능가하는 수준. "①합성 데이터 생성 → ②AI 훈련 → ③더 똑똑해진 AI가 더 좋은 데이터 생성"이라는 자체 강화학습 사이클을 확보하고 있기 때문.
- Kimi K2는 '개방형 에이전트 지능'을 표방. K2 모델의 가중치(weights)는 공개되어 복제 가능하지만, 이 데이터 공장은 Moonshot의 복제 불가능한 핵심 자산. Kimi의 지속 가능한 경쟁 우위의 원천. 이 공장이 존재하는 한, Moonshot은 K2를 능가하는 K3, K4를 계속해서 '생산'할 수 있음.
- 현재는 추론(reasoning) 기능이 비활성화된 상태임에도 강력한 성능을 시현. 향후 추론(reasoning) 능력 탑재 시 파괴력은 배가될 것. 궁극적 목표는 챗봇을 넘어, 파워포인트, 미니게임 같은 프론트엔드 애플리케이션을 즉시 만들어주는 경험을 제공하는 것.

Kimi K2의 대규모 Agentic Data Synthesis(자기 개선형 AI 공장) 파이프라인
에이전트를 위해, 단순한 데이터 생성기가 아니라, 스스로 진화하며 더 나은 결과물을 무한히 생산해내는 시스템



"에이전트를 만들자" 같은 최종 목표 수립.
AI가 스스로 학습할 분야(Domains), 사용할 도구(Tools), 풀어야 할 문제(Tasks) 생성

훈련받는 AI(Agents)가 가상의 선생님 AI(User Agents)와 짝을 이뤄 문제를 품

모든 활동은 "이 문제를 3단계 안에 풀어야 성공"과 같이 명확한 채점 기준(Rubrics)이 있는 상태에서 진행

훈련이 끝나면, 현재 가장 똑똑한 AI가 심판(Judge)으로 등장해 결과물을 평가
심사를 통과한 최고의 모범 답안들만 골라냄

걸러진 초고품질 데이터(Filtered Data)만이 Kimi를 더 똑똑하게 만드는 데 활용

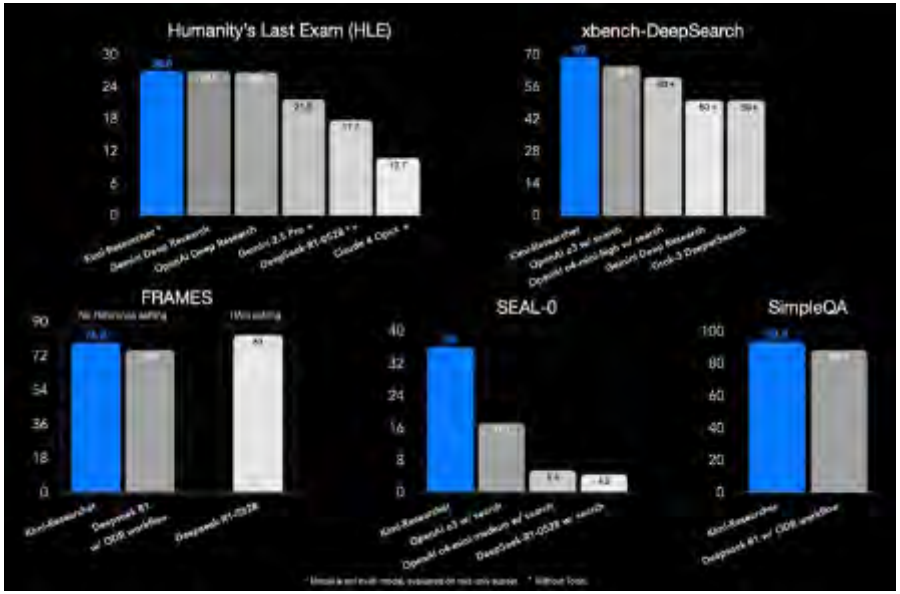
자료: 知乎(zhihu@Flood Sung), 미래에셋증권 리서치센터

VII. 중국 AI: Moonshot의 Kimi

Moonshot이 이끄는 AI 에이전트 시대 가속화

- 또 다른 거대한 오픈소스의 성취인 Kimi 2의 성공은 "최고에 근접한 성능을 압도적인 비용 효율로 제공"하는 가치 제안을 하고 있음. 이는 미국의 모델 사업자에게 무시 못할 위협. 기존 미국의 강자들은 ①가격 인하(수익 붕괴) ②가격 유지(시장 점유율 하락) ③경쟁용 오픈소스 공개(자기잠식)의 덫에 빠질 가능성도 배제하기 힘듦. 미국의 기업들은 멀티모달, 엔터프라이즈 보안 등으로 차별화 시도 및 방어적 오픈소스 모델을 출시할 것으로 보임.
- Kimi의 오픈소스 공개는 전 세계 커뮤니티의 최적화, 파인튜닝, 신규 애플리케이션 개발을 유발. 한 연구소의 혁신이 전체 생태계의 발전을 이끄는 촉매제 역할을 할 것. 특히 Moonshot이 공개한 K2 기술 논문에서는 파이프라인 병렬, 전문가 병렬, 데이터 학습법, Muon 옵티마이저와 QK Clip 등 매우 상세한 방법론이 제시되어 있어 중국을 포함한 오픈소스들의 역량을 한 층 끌어올릴 것으로 기대. 이는 'AI 에이전트 시대'의 가속화를 이끌 수 있는 잠재력이 있는 것이라 평가.

Kimi K2의 주요 벤치마크 점수 비교, '오픈소스가 HLE에서 상대적으로 고득점!'



자료: Moonshot, 미래에셋증권 리서치센터

개방형 에이전트를 지향하는 K2에 대한 공식 기술 보고서

KIMI K2: OPEN AGENTIC INTELLIGENCE

TECHNICAL REPORT OF KIMI K2

Kimi Team

ABSTRACT

We introduce Kimi K2, a Mixture-of-Experts (MoE) large language model with 32 billion activated parameters and 1 trillion total parameters. We propose the MuonClip optimizer, which improves upon Muon with a novel QK-clip technique to address training instability while enjoying the advanced token efficiency of Muon. Based on MuonClip, K2 was pre-trained on 15.5 trillion tokens with zero loss spike. During post-training, K2 undergoes a multi-stage post-training process, highlighted by a large-scale agentic data synthesis pipeline and a joint reinforcement learning (RL) stage, where the model improves its capabilities through interactions with real and synthetic environments.

Kimi K2 achieves state-of-the-art performance among open-source non-thinking models, with strengths in agentic capabilities. Notably, K2 obtains 66.1 on Tau2-Bench, 76.5 on ACEBench (En), 65.8 on SWE-Bench Verified, and 47.3 on SWE-Bench Multilingual — surpassing most open and closed-sourced baselines in non-thinking settings. It also exhibits strong capabilities in coding, mathematics, and reasoning tasks, with a score of 53.7 on LiveCodeBench v6, 49.5 on AIME 2025, 75.1 on GPQA-Diamond, and 27.1 on OJBench, all without extended thinking. These results position Kimi K2 as one of the most capable open-source large language models to date, particularly in software engineering and agentic tasks. We release our base and post-trained model checkpoints to facilitate future research and applications of agentic intelligence.

자료: Moonshot Kimi K2 기술 논문, 미래에셋증권 리서치센터

[참고] Kimi: 중국의 새로운 National Champion

딥시크에 이은 Kimi K2의 성공은 중국 AI 연구소들이 세계적 플레이어임을 증명

Kimi K2는 GPU라는 한정된 자원을 1초도 낭비하지 않기 위해, 계산, 통신, 메모리 전송을 마치 오케스트라처럼 지휘하는 극한의 최적화를 달성

GPU가 노는 시간(파이프라인 버블)을 없애기 위해, 순방향(Forward) 계산이 끝나자마자 즉시 역방향(Backward) 계산으로 전환하는 '1F1B' 스케줄링을 적용하여 GPU 가동률 극대화.
MoE 모델의 병목인 전문가 파라미터 통신(EP)과 파이프라인 통신(PP)마저도 핵심 계산 작업과 동시에 일어나도록 설계하여, 통신 대기 시간을 사실상 0으로 만들었음.
Kimi K2가 DeepSeek와 같이, GPU 내부의 작업 스케줄링까지 완벽하게 통제하는, 하드웨어에 대한 깊고 정통한 엔지니어링 역량을 보유했음을 증명.

자료: Kimi K2 기술 논문, 미래에셋증권 리서치센터

중국의 주요 AI 모델들의 스펙사항과 비교적 새롭게 등장한 Kimi K2 모델의 차별점

항목	Qwen3-235B -A22B-2507	DeepSeek-V3	Kimi K2	V3 대비 K2의 변화	Kimi K2의 차별점
모델 구분	비추론	비추론	비추론	동일	-
레이어 수	94	61	61	동일	기본 블록 깊이는 그대로, 다른 요소로 차별화
총 파라미터	235B	671B	1.04T	↑ 54 %	전문가(Experts) 수를 늘렸기 때문에 "모델 몸집"이 커짐
활성 파라미터(실제 계산 참여)	22B	37B	32.6B	↓ 13 %	MoE 희소성 덕분에 모델은 커졌지만, 개별 토큰의 계산량은 오히려 감소
총 전문가 숫자	128	256	384	↑ 50 %	더 세분화된 전문가 풀 → 지식 표현력 확대
토큰당 활성 전문가	8	8	8	동일	희소성(V3은 8/256 → K2는 8/384) 더 조밀하고 더 희소하게!
공유 전문가	1	1	1	동일	모든 토큰이 공통으로 보는 "범용 전문가"는 유지
어텐션 헤드 수	64	128	64	↓ 50 %	전체 전문가 수 증가로 늘어난 연산량은, 헤드 축소로 상쇄
밀집(Dense) 블록 개수	N/A	3	1	↓ 67 %	계산비용이 큰 밀집 레이어를 최소화해 효율 극대화
전문가 그룹핑	Yes	Yes	No	-	계층적 라우팅 → 비계층 라우팅, 이로써 더 정확한 전문가 매칭이 가능하나 GPU 통신 대역폭 요구량은 증가
컨텍스트 창	262k	128k	128k	동일	

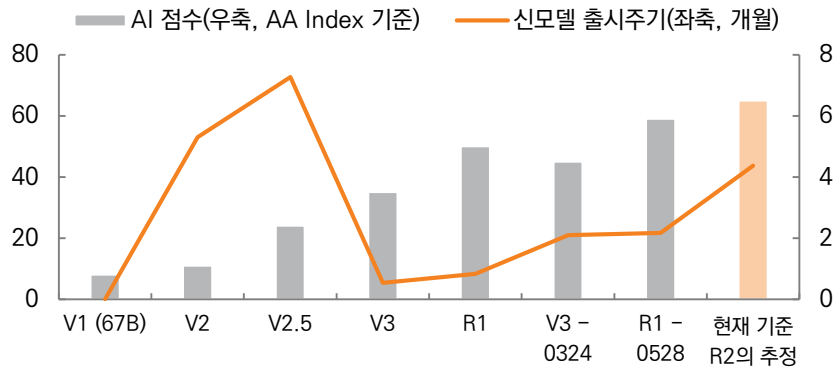
자료: 각 회사, 미래에셋증권 리서치센터

VII. 중국 AI: 이들이 갖는 공통적인 한계점

"공공이 관여"하면, 오히려 자생력을 떨어트릴 수 있다

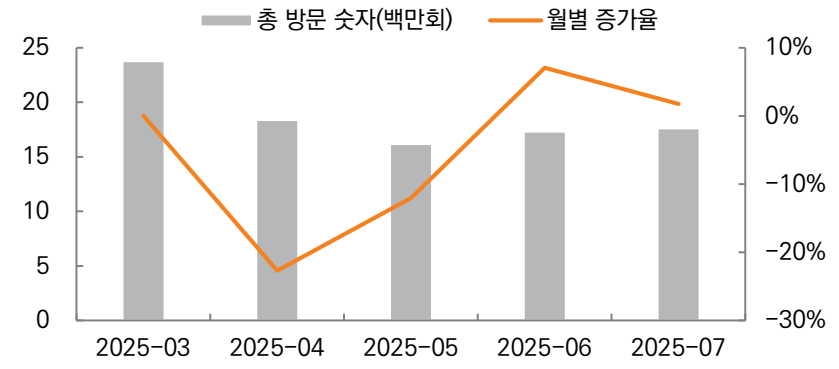
- DeepSeek R2는 당초 5월 초 출시를 목표로 했던 것으로 보임. 그러나 여전히 불투명. R2가 8월에 출시될 것이라는 중국 현지 커뮤니티 루머에 DeepSeek는 금세 정정보도를 냈을 정도. DeepSeek가 모델을 내놨던 과거 출시 패턴과 비교하더라도 늦어지고 있는 확실히 형국.
- 그리고 지난 8월 18일, DeepSeek가 R2 출시를 연기했다는 보도가 있었음. 문제는 中 당국의 '권고'. R1 모델을 출시한 후, 엔비디아 시스템 대신 화웨이의 Ascend 칩을 활용한 모델 훈련에서 지속적인 기술적 문제에 부딪혔다고 알려짐. 이에 따라 결과적으로 모델 훈련에는 여전히 엔비디아 칩을, 추론(Inference) 단계에서만 화웨이 칩을 사용하게 되는 것으로 귀결된 듯. 당국의 강제적인 '국산화' 정책이 오히려 기술 개발의 발목을 잡는다는 비판적인 시각도 나옴.
- 이와 유사한 예로, 시진핑 주석이 주목한 중국 AI 스타트업 Butterfly Effect Technology가 개발한 AI 서비스 Manus는 최근 중국을 떠나 싱가포르로 본사를 이전을 결정. 이 과정에서 핵심 개발 인력 40명을 제외한 80여 명의 직원들을 모두 해고했다는 것에 사태의 심각성이 조명.
- 중국 내부에서 수익 창출이 어렵고, 당국의 압박으로 서비스를 "무료로 개방"해야 했던 것이 큰 원인. 알리바바 개발진과 함께 중국어 버전 Manus 개발에 착수한다고 했으나, 결국 알리바바와의 협력도 중단한 것. 중국 정부의 (사실상) 오픈소스/무료 강제 정책이 이러한 부작용도 갖고 있음을 시사.

DeepSeek의 주요 모델별 성능 점수 및 출시 주기 추이
R2가 8월을 넘어 중추절(10월 6일)에 출시될 것이라 가정



자료: Artificial Analysis, 미래에셋증권 리서치센터
주: AA Index는 현존 LLM 성능 비교에서 업계가 가장 자주 인용하는 종합 지표

Manus의 월별 총 방문 숫자와 월별 증가율(MoM)
출시 이후 감소 및 이용자 숫자는 정체…



자료: SimilarWeb, 미래에셋증권 리서치센터

VIII. 엔비디아

CUDA

共진화

3계층

[들어가기에 앞서] 에이전트 시대의 세 가지 대표 기업

엔비디아(에이전트의 근간) / 테슬라(막대한 TAM을 가진 에이전트) / 팔란티어(각 산업별 에이전트의 산파)

1. 에너지와 실리콘이 모든 에이전트 경제의 근간인 “AI Factory”
2. 범용 지능에 “인지적 디지털 트윈”(가상세계 구현+온톨로지)가 붙어야 현장에서 바로 쓸 수 있는 에이전트가 탄생

아래의 이미지에서 각 회사들을 나타내는 색깔의 크기의 의미는, 각 영역에서 세 회사가 차지하고 있는 영향력을 시각화한 것

자료: 미래에셋증권 리서치센터

VIII. 엔비디아: 데이터센터의 패러다임 전환

전통적 데이터센터에서 'AI 팩토리'로

- 과거 데이터센터는 독립된 컴퓨터들의 집합체. 각 서버는 할당된 개별 작업을 수행했고, 네트워크는 이들 간의 통신을 지원하는 보조적 역할. 컴퓨팅의 중심은 서버(CPU)였으며, AI 같은 대규모 병렬 작업에는 구조적으로 비효율적. 현재의 AI 팩토리는 '하나의 거대한 컴퓨터'라는 개념으로 재정의. 수만 개의 GPU가 마치 단일 시스템의 수많은 코어처럼 유기적으로 작동해야 함. 여기서 네트워크는 보조 장치가 아닌, 시스템 전체 성능을 결정하는 '중추 신경계'로 격상. 이 신경망의 속도와 효율이 전체 팩토리의 생산성을 좌우하는 핵심 병목.
- 패러다임 전환이 엔비디아의 입지를 강화. 엔비디아는 GPU라는 뇌세포를 공급할 뿐만 아니라, 이들을 연결하는 '신경망'(NVLink, InfiniBand)까지 직접 설계하여 시스템 전체를 정의. 경쟁사들은 아직 개별 뇌세포의 성능에만 집중하고 있어, 시스템 레벨의 경쟁에서 뒤처짐.
- 미래에는 AI 모델의 복잡성이 기하급수적으로 증가하며 단일 건물로는 필요 전력/공간을 감당할 수 없을 것. 여러 데이터센터를 하나로 묶는 '분산형 AI 팩토리' 아키텍처를 필수로 요구하게 될 것.

엔비디아는 미래의 "AI 팩토리"를 준비하기 위해서 전체적인 생태계 관점에서 모든 준비를 하고 있음… 아래는 "GTC 2025"에서 나온 신규 발표들

핵심 기술/특징	주요 혁신 내용	기대 효과 / 영향
TCO 혁신 전략	**세대 간 압도적 성능 향상 및 비용 절감 약속 (Hopper 대비 Blackwell은 87%, Rubin 99.97%)**	**선도적 입지 강화, 제본스의 역설 기반 수요 창출**
Rubin 아키텍처	I/O 타일 분리 설계	연산 공간 극대화, SM/텐서 코어 추가 탑재 용이
Rubin Ultra 구성	4개 컴퓨트 다이 + 16개 HBM4E 스택 (총 1TB)	전례 없는 패키지당 메모리 용량 제공
HBM4 메모리	13TB/s 고대역폭 구현 (Rubin 기준)	메모리 병목 현상 완화
Vera CPU (자체 커스텀)	엔비디아의 맞춤형 코어, 혁신적 메모리 패브릭	Arm 의존도 감소, CPU 성능 최적화
Kyber 랙 아키텍처	랙 90도 회전으로 고밀도 집적 가능	단일 랙에 144개의 GPU 패키지 배치, 향후 288개까지 확장
NVLink 6세대	3.6TB/s 양방향 대역폭 (2배 향상)	GPU 간 통신 속도 대폭 향상
CPO (Co-Packaged Optics)	스위치 내 광학 통합, Quantum/Spectrum-X CPO 스위치 출시	네트워크 전력 소비 감소, 네트워크 계층 단순화(2계층)
Dynamo AI 엔진	개방형, DeepSeek 최적화 기술 통합 (KV 캐시 오프로드 등)	고효율 추론 기술 대중화 및 접근성 향상

자료: 미래에셋증권 리서치센터

VIII. 엔비디아: 해자① 생태계 그 자체인 CUDA

CUDA, 기술을 넘어선 사실상 '산업 표준'

- CUDA의 본질은 단지 기술이 아닌 개발자 '락인(Lock-in)' 효과. 딥러닝 코드 단위를 넘어 개발자의 사고방식, 대학 교육 시스템, PyTorch/TensorFlow 같은 핵심 프레임워크에 이미 깊이뿌리 내린 상태. CUDA 개발자는 400만명에 육박.
- 반면 AMD의 경쟁 제품인 "RoCm"를 실무적으로 경험해본 엔지니어는 약 2천명에 불과. 이는 기업들의 "컴퓨팅 자원 선택에 있어 가장 중요한" 인력 수급 리스크를 결정하는 매우 큰 요인.
- CUDA 코드를 ROCm으로 자동 변환할 때의 효율은 약 80%에 불과. 나머지 20%의 미세 조정은 결국, 성능 저하와 버그를 막기 위해서 극소수의 고가 커널 엔지니어가 수작업으로 해결해야 함. 이는 TCO(총소유비용) 측면에서 변환의 경제성을 완전히 상실하게 만듦. 기업 CTO 입장에선 ROCm 전문가를 구하지 못할 리스크 때문에 검증된 CUDA 기반 솔루션을 선호할 수밖에 없는 구조.

CUDA 기반의 핵심 라이브러리 및 개발 도구 생태계는 다양한 산업 분야를 아우르는 가속 컴퓨팅 전문 솔루션으로 작용
개발자나 기업은 복잡한 CUDA 함수를 처음부터 짤 필요 없이, 이 라이브러리를 가져다 쓰기만 하면 즉시 세계 최고 수준의 GPU 가속 성능을 얻게 됨
CUDA가 단순한 프로그래밍 모델이 아니라, 특정 산업의 문제를 해결하는 구체적이고 강력한 솔루션의 집합체임을 보여주는 것

라이브러리/도구	분야	설명/기능
CuPy	수치 연산	가장 널리 사용되는 수치 라이브러리 Numpy 문법을 CUDA로 호환/구현한 것
Aerial & Shona	통신	세계 최초 GPU 가속 5G/6G 무선 신호 처리. 소프트웨어 정의 방식으로 AI를 5G/6G에 도입 가능
Parabricks	유전체학	유전체 분석
MONAI	의료 영상	의료 영상 처리
Earth-2	기상 예측	기상 예측
cuQuantum	양자 컴퓨팅	양자-고전 컴퓨터 아키텍처 및 시스템
cuEquivariance, cuTensorNet	텐서 수학	텐서 수학 라이브러리
Megatron, TensorRT-LLM, NeMo, Dynamo	딥러닝	딥러닝 학습 및 추론 라이브러리. 최근 대규모 AI 팩토리를 위한 새로운 운영체제(Dynamo)도 포함
cuDF	데이터 프레임	Spark, SQL과 같은 구조화된 데이터 가속
cuML	머신러닝	고전적 머신러닝
Warp	시뮬레이션	CUDA 커널을 기술하기 위한 Pythonic 프레임워크. 매우 성공적
cuOpt	최적화	외판원 문제, 공급망 최적화 등 제약이 많은 대규모 변수 최적화 문제 해결
cuDSS, cuSparse	시뮬레이터	희소 구조 시뮬레이터 (CAE, CAD, 유체 역학, 유한 요소 분석 등 EDA/CAE 산업에 중요)

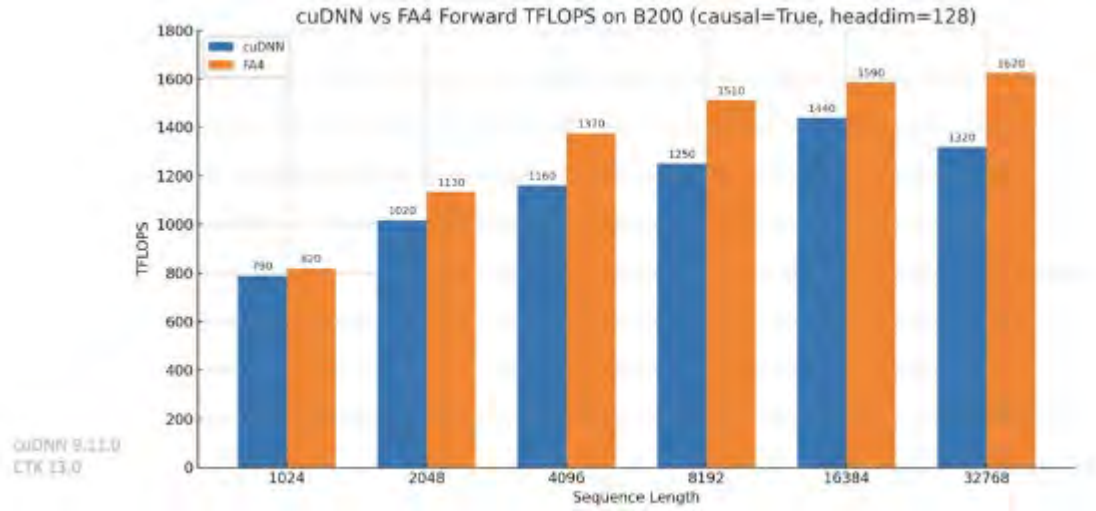
자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 해자① CUDA 해자는 계속 단단해진다

소프트웨어 업데이트가 하드웨어 가치를 증폭해 "시간적 해자"를 만든다

- 소프트웨어는 하드웨어의 경제성을 결정하는 핵심 변수. 엔비디아는 지속적인 CUDA 업데이트를 통해 신규 하드웨어의 잠재력을 극대화하고, 심지어 구형 하드웨어의 수명까지 연장시킴.
- 예를 들어, GB200의 초기 TCO는 H100 대비 1.6~1.7배 높아 초기 투자 부담이 큼. 그러나 엔비디아는 CUDA 최적화를 통해 TCO당 성능을 2025년 7월 1.5배, 연말까지 2.7배로 끌어올릴 계획을 제시. 하드웨어 교체의 높은 비용을 소프트웨어 발전으로 정당화하는 강력한 세일즈 포인트.
- 지난 8월에 나온 FlashAttention v4와 같은 최신 LLM 알고리즘은 오픈소스로 공개되지만, 엔비디아의 Blackwell GPU 칩이 가진 초저수준(기계어) 명령어 "wgmma"와 완벽히 결합되어 있음. 즉, 엔비디아 칩을 써야 최고의 성능을 낼 수 있는 AI 모델을 만들 수 있게 된다는 것.
- 경쟁사(AMD, 화웨이 등)가 이를 분석하고 모방하여 자사 하드웨어에 맞게 재설계하는 데는 수개월이 걸리고 인력 수급도 힘듦. 그 시간 동안 엔비디아는 독점하며 격차를 더욱 벌림. 이 시간 격차 자체가 곧 시장 지배력. 특히 AI 컴퓨팅에 관한 광폭 투자 열풍이 부는 가운데 이런 시간적 해자는 막대한 강점.

Blackwell 칩(B200)에서 최신 AI 알고리즘인 FlashAttention 4가 기존 알고리즘(cuDNN)보다 월등한 성능
애초에 FlashAttention 4의 성능은 Blackwell 칩의 가장 낮은 레벨(기계어)에 있는 명령어에 맞물려 있어서 높음



자료: Github(@Dao-AILab), 미래에셋증권 리서치센터

[참고] CUDA의 해자 극복은 애플도 포기했다

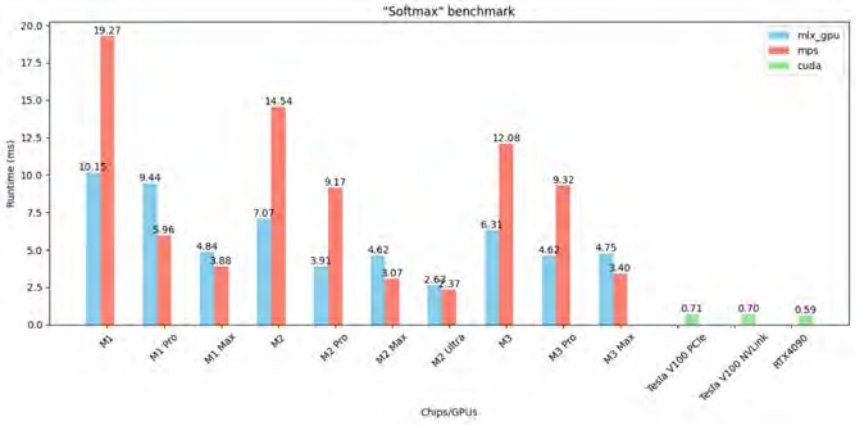
성능이 결정적으로 중요한 순간에는 결국 CUDA를 선택하게 된다

- AI 시대의 표준이 CUDA로 굳어지자, 애플은 자사의 머신러닝 프레임워크인 “MLX”가 엔비디아의 CUDA를 지원하도록 직접 개발하는, 놀랍고도 현실적인 타협을 선택. 해당 업데이트로 개발자는 mac book에서 MLX로 AI 모델의 프로토타입을 만든 뒤, 코드를 단 한 줄도 수정하지 않은 채로, 엔비디아 GPU 서버로 대규모 학습 및 배포 작업을 바로 옮길 수 있게 됨.
- 이번 애플의 결정은 엔비디아의 진정한 힘이 하드웨어를 넘어, 수십 년간 쌓아 올린 CUDA라는 소프트웨어 생태계의 해자가 얼마나 강력한지를 증명하는 가장 상징적인 사건.
- 이 과정에서 엔비디아는 아무 노력을 하지 않았고, 오히려 애플이 직접 자사의 MLX 프레임워크를 수정하여 ‘MLX to CUDA’라는 고속도로를 구축. 엔비디아의 절대적 교섭력을 명확히 상징.
- 결과적으로 엔비디아는 잠재적 경쟁자였던 애플의 개발자 커뮤니티를 자연스럽게 흡수하게 되었고, 이제 mac에서 시작된 AI 프로젝트의 최종 목적지는 엔비디아 GPU 팜이 될 가능성이 높아짐.

깃허브에 올라온 “애플용 CUDA 백엔드”
애플 걸로 작성한 “레시피”를 엔비디아 GPU+CUDA로 쓸 수 있게 번역

자료: GitHub, 미래에셋증권 리서치센터

현실을 인정한 차트: 애플이 CUDA에 다리를 놓을 수밖에 없었던 이유
애플 실리콘의 실행 시간(Runtime)은 수 밀리초(ms)에 달하는 반면, 엔비디아 GPU의 실행 시간은 1ms 미만으로, 수십 배에 달하는 격차



자료: towardsdatascience, 미래에셋증권 리서치센터

[참고] CUDA가 더 강력해질 이유: 뭐가 더 필요한지 알고 딱딱 M&A

엔비디아는 지난 2년간 80개 이상의 AI 스타트업에 투자…

Run.ai와 CentML 인수

- 대규모 AI 모델의 훈련과 추론에 천문학적인 비용이 소요되면서, GPU 활용률 극대화와 운영 비용 절감은 기업의 생존 과제가 됨. 엔비디아는 이 패러다임 전환을 정확히 인지하고, 단순히 가장 빠른 칩을 제공하는 것을 넘어, 하드웨어의 성능을 가장 효율적으로 끌어낼 방법을 직접 제공하기 시작. 최근 단행된 Run.ai와 CentML 인수는 이 새로운 전쟁터를 지배하기 위한 전략적 포석. 이것으로, AI 워크플로우의 양대 비효율, '자원 낭비'와 '모델 비대화'를 동시에 해결하기 위해 핵심 기술을 내재화.
- 이번 두 인수는 완벽한 상호 보완 관계를 형성. Run.ai와 CentML의 결합은 엔비디아의 해자를 한 단계 더 깊게 만들 수 있음. 단순 기술 추가가 아닌, 생태계의 모든 층위를 수직 통합하는 결정적 한 수. CentML로 최적화된 모델을 Run.ai가 관리하는 GPU 클러스터에서 실행. 이는 '모델 효율성'과 '자원 효율성'을 동시에 달성하는 솔루션으로, 강력한 고객 락인 효과를 창출.
- 경쟁사는 칩이 아닌, 엔비디아가 가진 생태계 전체와 싸워야 함. 실리콘(GPU), 네트워킹(NVLink, InfiniBand), 플랫폼(CUDA)에다가 효율성 소프트웨어(Run.ai + CentML)까지 더해진 것.

엔비디아의 락인효과를 더해줄 새로운 SW 계층, "하드웨어 자원 관리와 소프트웨어 모델 최적화의 결합"

	Run.ai (인프라 지휘자)	CentML (모델 압축 전문가)
역할	하드웨어 인프라 효율성 극대화	모델 소프트웨어 자체의 효율성 극대화
기술	Kubernetes 기반 GPU 관리 및 오케스트레이션. 분산된 GPU를 단일 가상 풀처럼 관리 워크로드를 동적으로 할당하여 클러스터 활용률 높임	컴파일러 기반 AI 모델 최적화. 모델 코드를 특정 GPU에 맞게 재구성. 동일 하드웨어에서 몇 배로 실행 속도를 높임. 메모리 사용을 줄임.
결과	GPU 자원 공유 효율화 및 운영 비용 절감	모델 추론/훈련 속도 향상 및 비용 절감

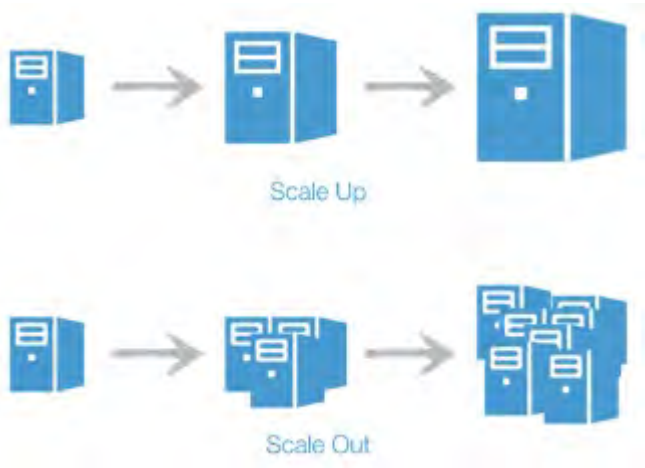
자료: 미래에셋증권 리서치센터

VIII. 엔비디아: 해자② 과소평가 되고 있는 엔비디아의 네트워킹

GW급 'AI 팩토리'를 위한 3계층 네트워킹

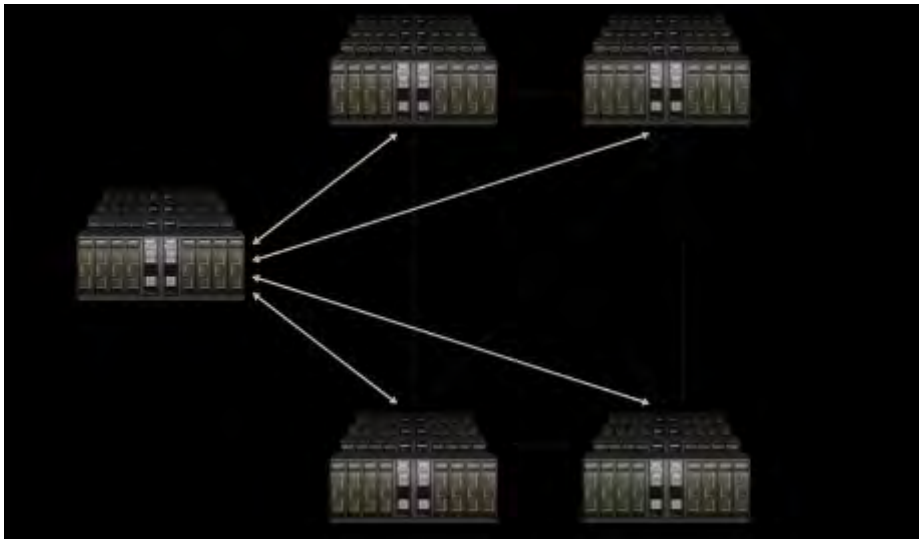
- 엔비디아는 단일 데이터센터가 물리적 한계에 곧 도달할 것을 예견. 미래의 인프라는 단일 건물이 아닌, 지리적으로 분산된 여러 데이터센터를 하나로 묶는 방식으로만 지속적인 확장이 가능하다고 판단
- 이에 대한 해답으로 칩부터 대륙까지 끊김 없이 연결하는 3계층 네트워크 아키텍처라는 통합적 청사진을 제시. 이는 네트워킹 기술을 판매하는 것을 넘어, 미래 AI 인프라의 표준을 정의하고 시장의 규칙을 만드는 행위.
- 이 3계층 구조는 경쟁사들이 넘볼 수 없는 강력한 시스템 레벨의 해자. 경쟁사들이 개별 칩 성능과 랙 내부 연결에 집중할 때, 엔비디아는 인프라 전체의 구조와 미래를 설계하고 시장을 교육하고 있음.
  1. Scale-Up: 랙 내부를 하나의 거대 GPU로 묶는다 (NVLink).
  2. Scale-Out: 데이터센터 내부의 랙들을 초고속으로 연결한다 (InfiniBand).
  3. Scale-Across: 도시와 국가 간 데이터센터를 하나로 통합한다 (Spectrum-XGS).

스케일업과 스케일 아웃의 차이
Up: 랙 내부·인접노드에서 NVLink로 거대한 GPU처럼 묶는다
Out: 데이터센터 내부·캠퍼스 단위를 IB/이더넷으로 묶는다



자료: Invero, 미래에셋증권 리서치센터

AI 컴퓨터는 하나의 건물 안에 있는 랙들을 초고속으로 연결하고(Scale-out, 점선), 심지어 다른 도시에 있는 데이터센터까지 하나로 묶어(Scale-across, 실선) 작동



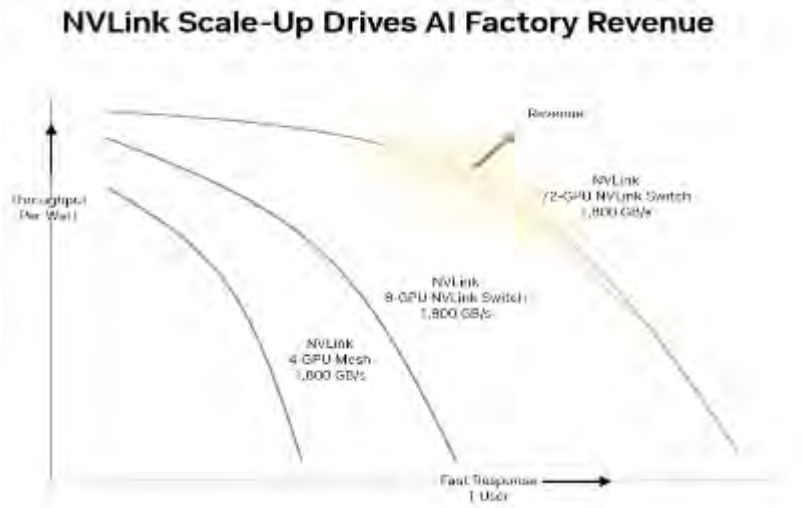
자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 해자② 스케일업(Scale-Up)

랙이 곧 하나의 GPU

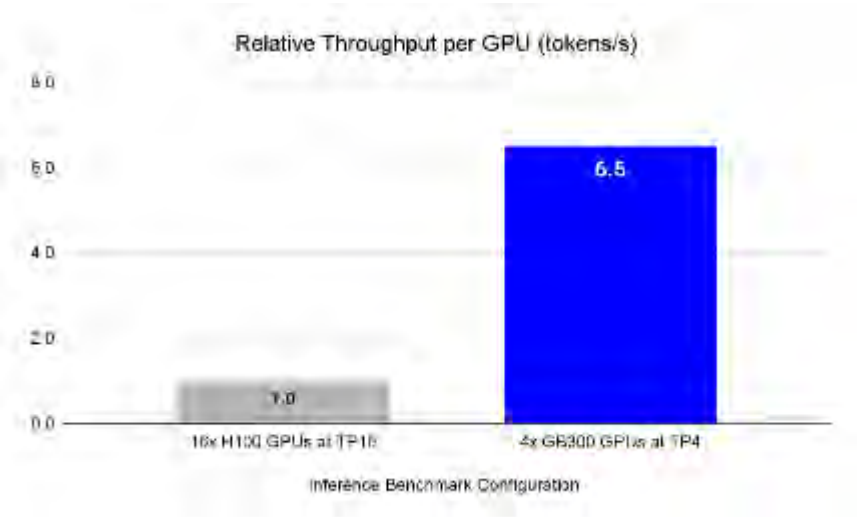
- 엔비디아의 NVL72 랙은 단순한 서버 랙이 아니라, 블랙웰 기반 B200 GPU 72개와 자체 개발 CPU인 Grace 36개를 5세대 NVLink 스위치로 묶어 130TB/s의 초저지연 통신을 구현하는 '하나의 거대한 GPU'. 이는 미래 AI 팩토리의 표준화된 최소 공정 단위(Building Block)로 볼 수 있음.
- 이 통합을 통해 조 단위 크기의 모델의 실시간 추론과 학습을 단일 랙에서 수행하는 것이 가능해질 것. 엔비디아는 지난 2016년에는 4개 GPU 연결에서 불과 8년 만에 72개 GPU를 하나의 시스템으로 묶는 기술적 도약을 이뤄냄.
- 이러한 스케일업 기술은 AI 공장의 경제성을 결정. 더 많은 GPU를 하나의 시스템처럼 효율적으로 묶을수록, 통신 오버헤드와 메모리 병목을 줄여서 전력당 성능과 응답 속도가 기하급수적으로 향상. 이는 AI 서비스의 TCO 절감과 수익성 증대로 직결. 특히 에이전트 서비스를 대중화하기 위한 마중물.
- NVL72는 완전 직접액냉 설계와 전력 공급 시스템(CDU, 버스바)을 랙 단위로 완벽하게 표준화했음. 이를 통해 고객들은 레고 블록처럼 AI 팩토리를 신속하게 증설하고 확장할 수 있게 됨.

NVLink 기술로 더 많은 칩을 묶을수록 'AI 팩토리의 경제성은 올라간다'
더 많은 GPU를 연결했을 때 훨씬 높은 효율과 빠른 응답 속도



자료: 엔비디아, 미래에셋증권 리서치센터

'GPU 하나당 초당 토큰 생성' 기준, GB300이 H100보다 6.5배 빠르다
NVL72 랙 안에서 얻을 수 있는 장점에서 비롯된 결과로 향후 상향 가능



자료: CoreWeave, 미래에셋증권 리서치센터

VIII. 엔비디아: 해자② 스케일아웃(Scale-Out)

AI 훈련 성능을 결정짓는 팩토리 내부의 대동맥

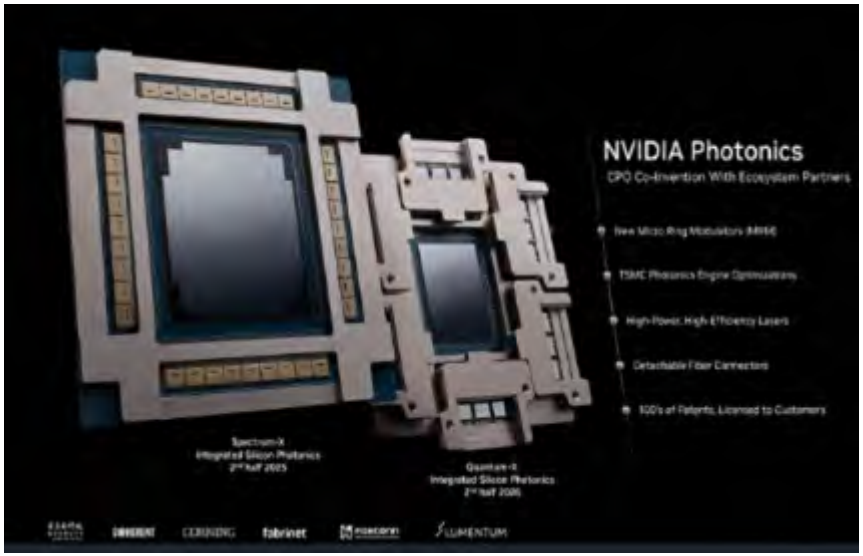
- 스케일아웃은 NVL72라는 빌딩 블록들을, 이번에는 수천 개 단위까지 묶어 거대한 컴퓨팅 클러스터를 형성하는 데이터센터 내부의 주력 고속도로 역할. 랙과 랙, 포드(pod)와 포드를 잇는 일종의 대동맥. 이 성능이 AI 팩토리 전체의 처리량을 결정. 특히 AI training은 스케일아웃의 "집합 연산"으로 결정됨.
- 엔비디아의 자체 제품인 Quantum-X800 InfiniBand는 SHARPv4 기술을 통해 네트워크 스위치 단에서 집단 통신(All-reduce) 연산을 직접 처리. 이는 GPU의 연산 부담을 덜어줄 뿐만 아니라, 통신 지연을 최소화하여 예측 가능한 최고 성능을 보장하는 핵심 기술.
- 대규모로 확장될수록 전력과 안정성이 중요. 엔비디아는 Co-Packaged Optics(CPO)를 통해 기존의 플러그형 광모듈을 칩과 통합된 실리콘 포토닉스로 대체하는 혁신을 준비 중. CPO는 광 트랜시버를 스위치 칩과 동일한 패키지 내에 통합하는 기술. 고대역폭 네트워킹의 핵심 기술로 부상하고 있음. 엔비디아는 CPO를 두고 전력 효율 3.5배, 복원력 10배 개선을 제시, AI 팩토리의 운영 안정성을 확보.

AI 공장의 대동맥: 팻트리(Fat-Tree) 구조의 스케일아웃 네트워크
Fat-Tree 또는 Leaf-Spine으로 불리는 네트워크 토폴로지

수만 개의 GPU를 막힘없이 연결하여, 어떤 노드 간 통신에서도 예측 가능하고 일관된 '결정적 성능'을 보장. 하위 노드 그룹(리프)들이 상위 중앙 스위치(스파인)들을 통해 서로 거미줄처럼 연결되어, 데이터가 막힘 없이 흐르는 '대동맥' 역할을 수행.

자료: Ayar Labs, 미래에셋증권 리서치센터

엔비디아가 내년에 본격 판매할 실리콘 포토닉스 CPO 네트워크 스위치
이를 통해, 여러 지역 걸쳐 수백만 개의 GPU를 TCO 효율적으로 연결



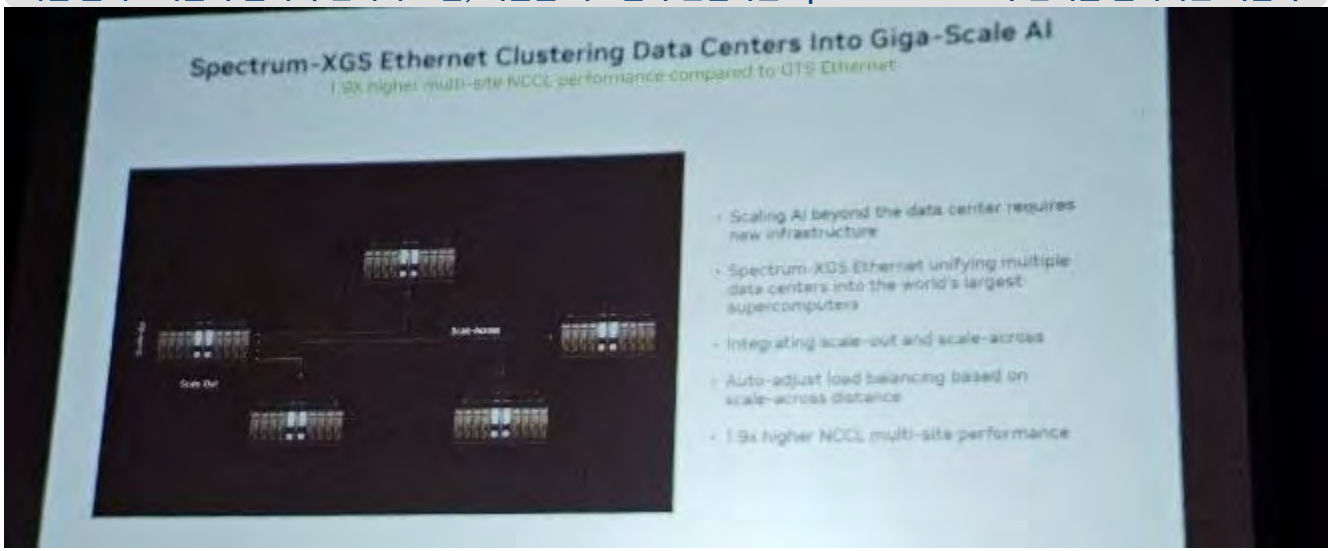
자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 해자② 스케일어크로스(Scale-Across)

글로벌 슈퍼 팩토리

- 지난 8월, 엔비디아는 지리적 한계를 넘어서는 AI 슈퍼-팩토리라는 담대한 비전을 제시. 기존 이더넷의 한계(신호 불안정, 예측 불가능성)를 극복하고, 최대 수천 km나 떨어진 데이터센터를 마치 하나의 거대한 공장처럼 유기적으로 연결하는 것이 목표.
- 핵심 기술은 WAN-RDMA의 현실화. 이를 가능하게 하는 핵심 기술은 Spectrum-XGS. 거리 인지 혼잡 제어 알고리즘을 통해 장거리 데이터 전송의 지연 시간과 데이터 손실을 최소화. 이론적으로만 가능했던 RoCE 이더넷 통신을 광역 네트워크(WAN)에서 안정적으로 구현하는 최초의 상용 솔루션.
- 엔비디아는 Spectrum-XGS를 통해 분산 학습에 필수적인 NCCL 통신 성능이 기존 이더넷 대비 최대 2배까지 다중 사이트 NCCL 성능이 개선된다고 주장. 클라우드 기업 CoreWeave가 첫 대규모 채택을 이미 예정하고 있어, 시장에서의 효용성이 곧 증명될 것.
- 이는 경쟁사들이 이제 막 스케일업/아웃 문제를 고민할 때, 엔비디아는 이미 그 다음 단계의 문제를 정의하고 해결책을 시장에 제시하고 있음을 명확히 보여주는 사례.

이더넷의 예측 불가능성을 극복하고, 지리적으로 분산된 DC들을 거대 단일 컴퓨터처럼 묶는 Scale-Across
개별 클러스터들이 물리적 한계에 도달, 이들을 가로질러 연결하는 Spectrum-XGS가 한계를 돌파하는 해결책



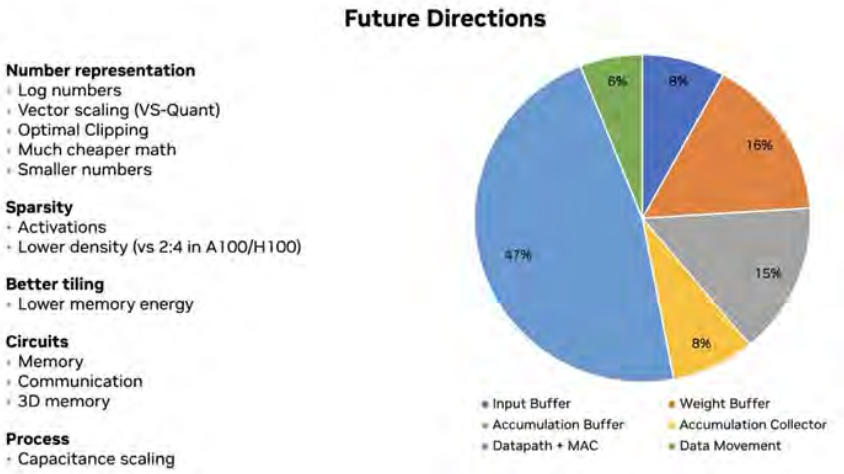
자료: 엔비디아, Hot Chips 2025, Jim McGregor, 미래에셋증권 리서치센터

[참고] 데이터 이동과의 전쟁 때문에 계층적 네트워크와 병렬화가 탄생

사실, AI 연산의 진짜 병목은 연산이 아닌 '데이터 이동' 그래서 네트워킹이 중요

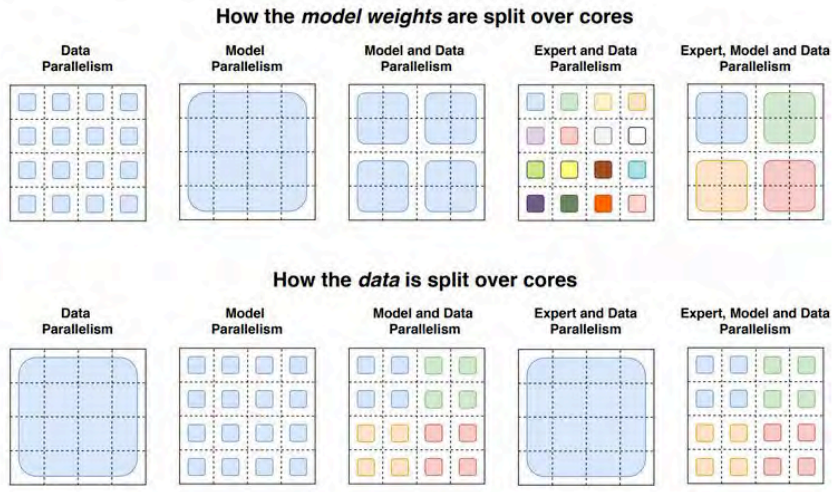
- 엔비디아 칩의 전력 소비 분석 결과, 실제 수학 계산(Datapath+MAC: AI 연산에 필요한 실제 곱셈, 덧셈)은 47%에 불과, 나머지 데이터를 옮기고(Interconnect 14%) 저장하는 데(Buffer/Cache 40%) 소모. 이런 물리적 한계를 극복하기 위해 GPU 클러스터가 계층적 네트워크(스케일업/아웃/어크로스)를 가지게 된 것.
- 그런데 노드/랙 내부와 외부의 데이터 대역폭은 약 10배에 달하는 심각한 격차가 있어서, AI 모델 학습은 하이브리드 전략을 채택해야 함. 이것이 바로 AI 모델의 병렬화(parallelism) 전략.
- 최근에는 [노드 내 TP] + [노드 간 PP] + [전체 DP]가 현대 LLM 학습의 표준 레시피가 되었음.
- 텐서 병렬(TP): 단일 행렬 연산을 쪼개는 방식, 통신이 빈번해 NVLink로 연결된 노드 내부에서 효율적.
- 파이프라인 병렬(PP): 모델 레이어를 순차적으로 할당하는 방식, 노드 간 통신에 적합.
- 데이터 병렬(FSDP): 모델을 복제하고 데이터를 나누는 방식, 전체 시스템에 걸쳐 적용.

AI 칩 성능의 물리적 한계: 계산만 빠르게 만드는 방식으로 성능향상 불가
데이터를 옮기고 저장하는 '물류 비용'(53%)에 더 많은 전력이 소모
엔비디아는 지난 10년간 이 문제를 해결한 모든 영역에서 최적화를 거듭



자료: tinycorp, 미래에셋증권 리서치센터

거대 AI 모델을 수천 개 GPU에 분산시키는 다양한 병렬 처리 전략
예전엔 데이터 병렬로 충분했지만, 모델이 커지며 모델 자체를 조각내고,
전문가 그룹으로 나누는(전문가 병렬), 복잡한 혼합 전략이 필수가 됨



자료: Yi Wang(애플 엔지니어), 미래에셋증권 리서치센터

VIII. 엔비디아: 경쟁 우려① 거의 유일한 패스트팔로워 AMD

'탈엔비디아' 흐름을 타는 현실적 대안

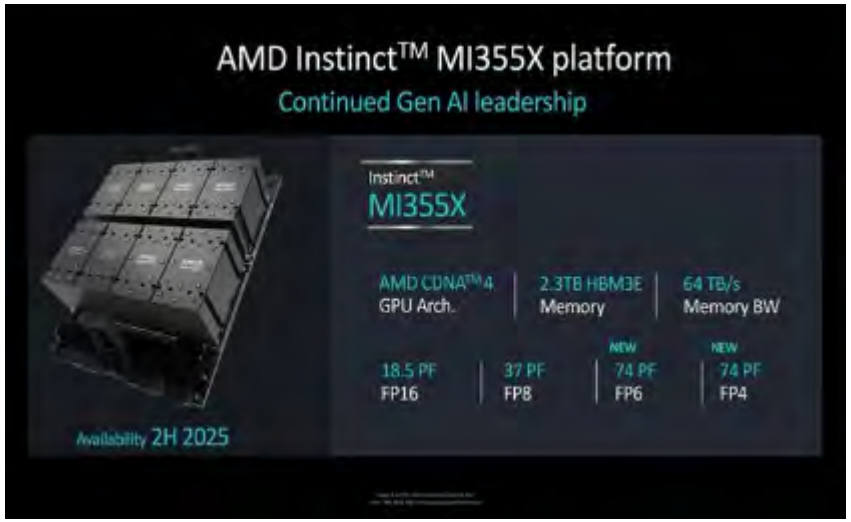
- AMD에 대한 시장의 관심은 지속적. 가장 큰 동력은 시장의 '반(反)엔비디아' 정서. 모든 하이퍼스케일러와 대기업들은 엔비디아에 대한 과도한 의존을 피하기 위해 공급망 다변화를 절실히 원함. 엔비디아의 강력한 가격 결정력에다가 심지어 클라우드 중개 플랫폼(Project Lepton)까지 오픈하려는 시도에 위협을 느끼는 하이퍼스케일러들에게 AMD는 유일한 현실적 대안.
- 액체냉각으로 전환이 어려운 기존 데이터센터에 AMD의 1,000W 공랭식 MI350X GPU(올해 중반으로 앞당겨 출시)는 거의 유일한 고성능 업그레이드 옵션. 액랭식이 필수인 GB200 NVL72가 즉시 공략할 수 없는 중요한 시장. MI355X 기반 시스템의 총 자본 비용은 엔비디아 B200 HGX 대비 거의 절반인 45% 저렴. 이는 엔비디아처럼 스위치, 트랜시버 등 독점 네트워킹 부품 의존도가 낮아 고객이 더 유연한 시스템을 구성할 수 있기 때문.
- 오라클이 3만 개의 MI355X 도입 계획을 발표한 것은, 대규모 레퍼런스를 확보했다는 점에서 중요한 의미. AWS, OpenAI 등과의 협력 강화도 AMD가 기업 시장에서 신뢰도를 쌓아간다는 긍정적 신호.

엔비디아 'Lepton' 인수 사건의 전략적 의미를 한눈에 파악
엔비디아의 독점이 이렇게나 무서운 것이는 것을 여실히 보여주는 사례

핵심 내용	구체적 배경 및 전략적 의미
엔비디아, 클라우드 중개 플랫폼 'Lepton' 인수 후 직접 사업(DGX Lepton) 진출	Lepton은 여러 CSP의 GPU 자원을 AI 개발자에게 중개로 연결해주는 소프트웨어 회사. 인수 후 'DGX Cloud Lepton' 플랫폼 출시.
클라우드 파트너들을 직접 경쟁시키며, GPU 임대 시장의 가격 통제력 확보	GPU 판매 수익을 넘어, 클라우드 사용 단계에서 발생하는 부가가치까지 흡수하려는 의도. 이에 하이퍼스케일러들은 조용한 분노 중. '탈(脫)엔비디아'를 모색하게 만듦.
엔비디아의 지원을 받던 신생 클라우드 (CoreWeave 등) 파트너들의 분노와 배신감	엔비디아는 거대 CSP 견제를 위해서라도 신생 클라우드를 전략적으로 지원해 왔음. Lepton 인수는 자신들이 키운 파트너와 직접 경쟁하겠다는 선언과도 같음. 엔비디아와 파트너 간 신뢰에 심각한 균열 발생.
파트너들의 분노가 경쟁사인 AMD에게 역사적인 기회를 제공	**AMD에게 기술 격차를 뛰어넘는 '명분'을 제공. CSP들이 AMD GPU 도입을 이전보다 훨씬 적극적으로 검토하게 만드는 결정적 계기.**

자료: 미래에셋증권 리서치센터

엔비디아 대비 거의 절반 가격으로 제시되는 AMD의 AI 칩 "MI355X"
전전작인 MI300은 엔비디아 대비 80%의 가격이었던 것에서 더 후려침



자료: AMD, 미래에셋증권 리서치센터

VIII. 엔비디아: 경쟁 우려① AMD 한계? 엔비디아가 만든 판에 있다는 것

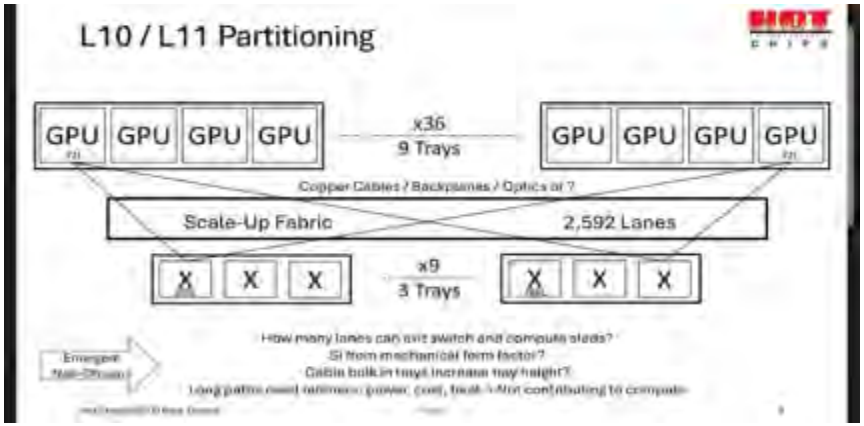
고통스러운 SW 경험, 전략적 비전의 부재

- SW는 AMD의 아킬레스건. 엔비디아 제품의 경우 `pip install torch` 한 줄로 5분 안에 설치가 완료. 반면 AMD는 거대한 Docker 이미지(프로그램 실행에 필요한 모든 요소를 담아놓은 풀 패키지)를 다운로드하고 복잡한 설정을 거쳐야 해 최소 30분 소요. 천재 개발자조차 AMD의 것이 "버그로 가득 차 있다"고 공개비판할 정도. 물론 나아지고는 있음. ROCm 7은 ROCm 6 대비 추론 처리량 성능 3.5배 향상.
- 소프트웨어 스택의 최적화 부족으로 인해 MI350X의 뛰어난 이론적 최대 성능(2.3 PFLOPS)을 실제 워크로드에서 제대로 활용하지 못하는 문제가 지속적으로 지적. 하드웨어의 잠재력을 소프트웨어가 깎아 먹는 형국. AMD의 라이브러리는 엔비디아의 것을 포크(fork)하여 개량하는 경우가 많음. 빠른 추격에는 유리하겠으나, 근본 아키텍처 최적화와 장기 호환성 문제에서 자유롭지 못하게 만드는 원인.
- AMD의 미래 로드맵은 여전히 개별 칩(MI450)과 서버 내 연결(UALink)에 머물러 있음. 엔비디아의 3계층 네트워크 비전에 대응할 만한 전략이 아직 부재. UALink마저 핵심 멤버인 브로드컴의 이탈로 2027년까지 지연될 전망이고, UALink 스위치 공급은 이제 Astera Labs과 Marvell에 의존해야 함.
- 한편, AMD의 내년 목표인 MI450 아키텍처의 핵심은 '중앙의 거대한 스위치'로써 칩들을 연결하는 것. 엔비디아가 수년 전부터 제시하고 완성해 온 NVSwitch 철학과 동일. 엔비디아가 쓴 길을 따라가는 것.

NVLink에 대항하는 표준인 UALink 1세대와 NVLink 5세대를 비교
NVLink보다 2배의 연결성을 홍보하지만 정작 출시는 '27년은 돼야 함

자료: AMD, 미래에셋증권 리서치센터

36 GPU를 중앙 대형 스위치 패브릭에 직접 연결하는 AMD의 차세대 구조
엔비디아의 NVLink/NVSwitch 방식을 AMD가 인정하고 따라가는 전략

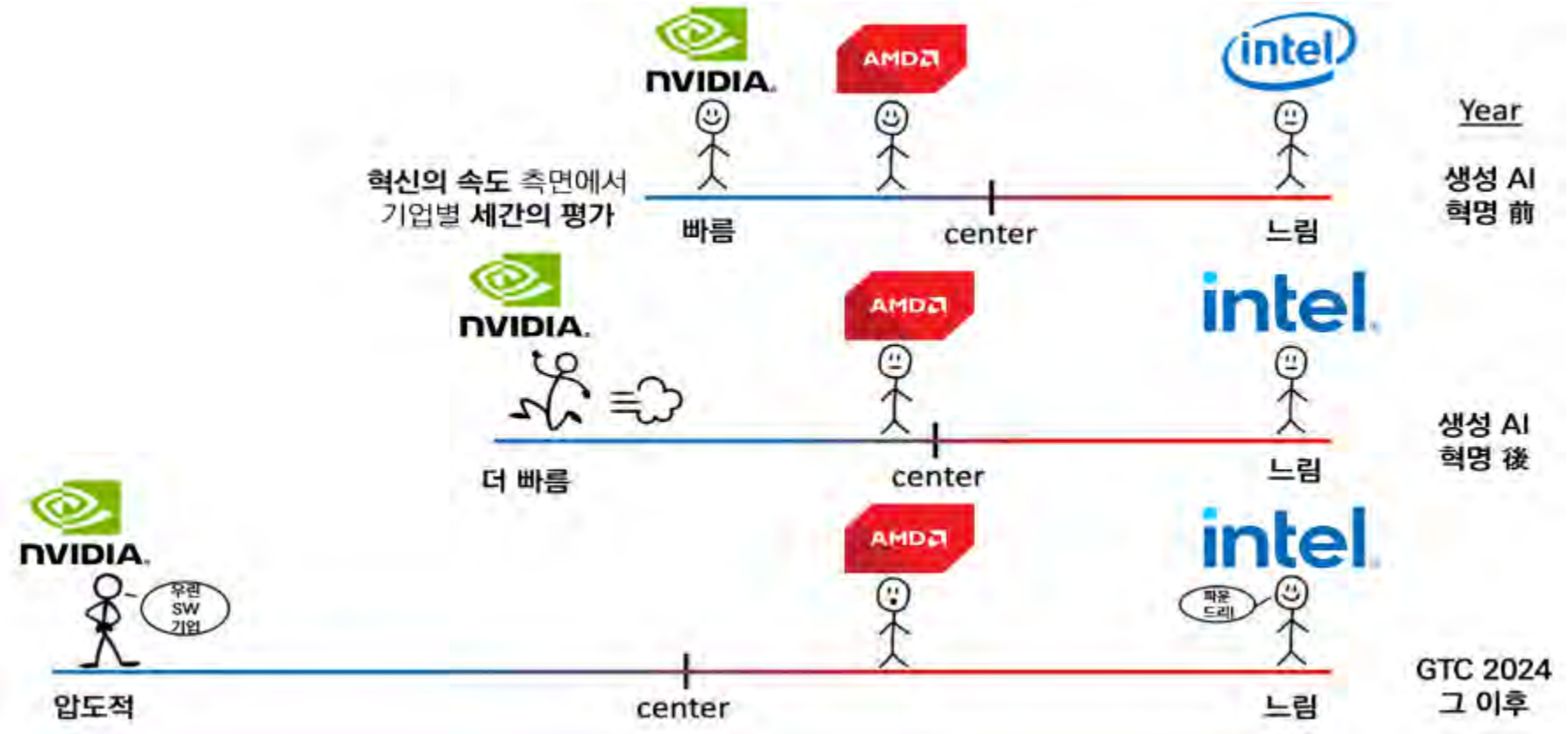


자료: AMD, Hot Chips 2025, 미래에셋증권 리서치센터

[참고] 엔비디아는 경쟁자와 시장 점유를 나눠먹는 게 아닌, 시장 조성자

엔비디아는 향후 5년을 내다보고 데이터센터의 구성에 대한 '스탠다드'를 먼저 설정한 뒤 칩을 설계

생성 AI 혁명 전후로 엔비디아, AMD, 인텔에 관한 시장이 느끼는 이미지



자료: 각 회사, 미래에셋증권 리서치센터

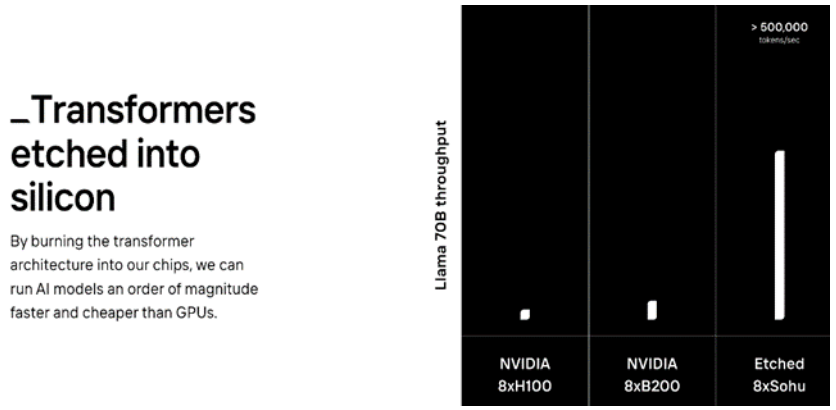
VIII. 엔비디아: 경쟁 우려② 신흥 NPU 업체들이 GPU보다 AI 잘한다던데?

‘HW-SW 共진화’

GPU의 역사가 곧 AI의 역사다

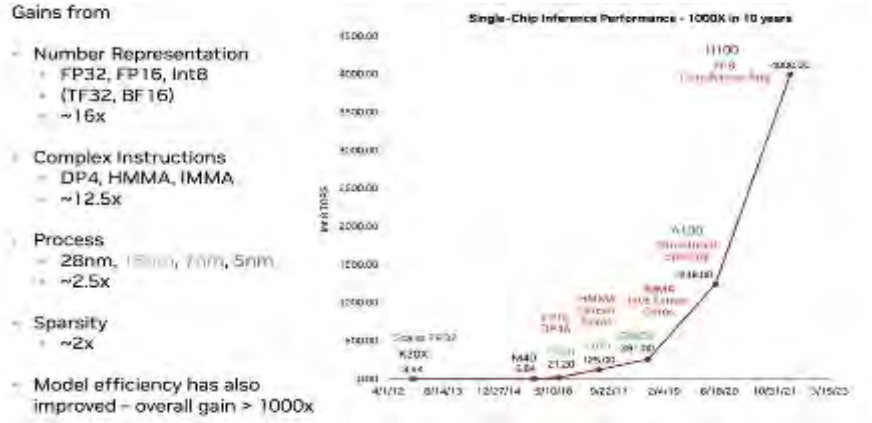
- AI 모델 연구는 공짜가 아니며, 연구자들은 현존하는 가장 강력하고 보편적인 하드웨어, 즉 엔비디아 GPU에 최적화된 방향으로 애초에 딥러닝 알고리즘을 발전시켜 왔음. 새로운 하드웨어는 더 복잡한 모델의 등장을 촉진하고, 새로운 모델은 다시 그 구조에 최적화된 차세대 하드웨어 설계를 요구. 현재 주류 아키텍처인 트랜스포머가 성공한 핵심 이유 중 하나도 GPU 아키텍처와의 완벽한 궁합 때문.
- 실제로, OpenAI, 딥마인드 등 최상위 연구소들은 엔비디아의 차세대 아키텍처 로드맵을 미리 파악하고, 그 아키텍처에서 최고의 성능을 낼 수 있는 방향으로 차세대 모델을 개발. 즉, 현재의 GPU는 AI에 거의 완벽하게 최적화된 기계. 지난 10년간 GPU가 이룬 수 천배의 AI 성능 향상은 ‘마법’이 아닌, 특정 연산(GEMM: 행렬 곱셈)에 대한 집요한 최적화의 역사.
- 현재 GPU 아키텍처는 더 이상 짜낼 수 있는 혁신 비법이 거의 소진된 상태. 여기서 퀀텀 점프 수준의 성능 향상을 이룰 새로운 칩이 등장하기는 물리적/경제적으로 극도로 어려움. 경쟁사가 엔비디아를 따라잡기 힘든 이유 중 하나. 스타트업들은 엔비디아의 ‘共진화 사이클’에 아직 갖혀 있다고 볼 수 있음.
  - 낮은 정밀도 연산(FP4/FP6 등): 계산의 정확도를 약간 희생해 속도를 비약적으로 높이는 방식은 이제 업계 표준)
  - 전용 명령어 탑재: AI가 자주 쓰는 복잡 계산을 하나의 명령어로 처리하는 ‘전용 계산기’를 내장하는 방식도 보편화
  - Sparsity 활용: 행렬의 수많은 0 값을 무시하고 계산하여 효율을 높이는 기술 역시 널리 적용되고 있음

초당 50만개의 토큰을 처리할 수 있다는 Etched의 8xSohu 칩
스타트업들의 전형: 트랜스포머 연산에 최적이라 엔비디아 것보다 좋다!



자료: Etched, 미래에셋증권 리서치센터

AI 연산성능이 폭발적으로 증가한 이유는, 미세공정(약 2.5배) 뿐만이 아님
저정밀도 연산 + 복잡한 연산 처리용 명령어 추가 등의 절실한 노력의 결과

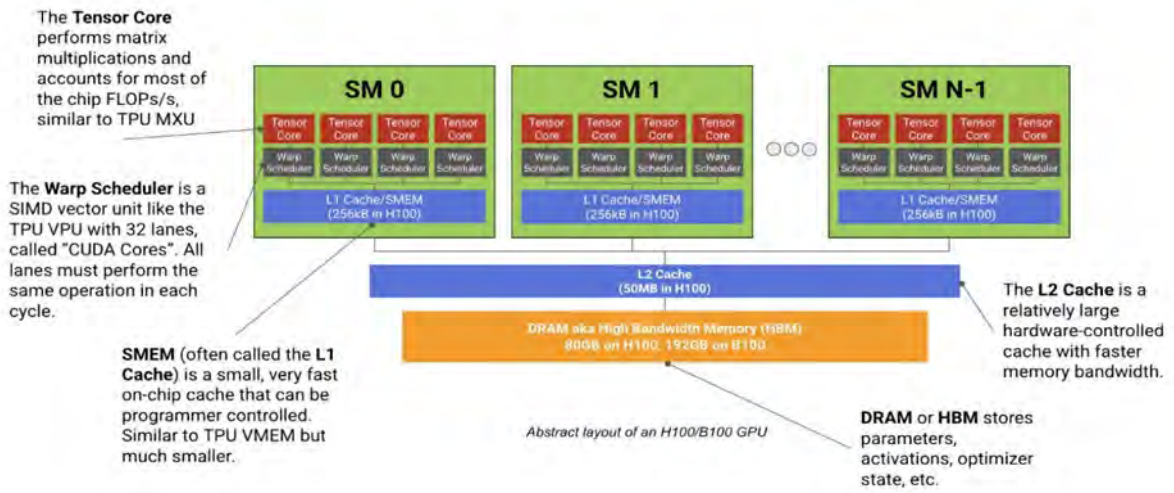


자료: tinycorp, 미래에셋증권 리서치센터

[참고] GPU 아키텍처의 이해: 범용인데 사실 범용 아니고, AI 연산 기계

- GPU는 AI 연산, 특히 행렬 곱셈을 효율적으로 처리하기 위해 고도로 병렬화된 구조. 엔비디아의 H100 GPU는 132개의 스트리밍 멀티프로세서(SM)를 탑재하며, 각 SM은 독립된 연산 유닛처럼 작동. 아래는 H100 기준 SM을 이루는 핵심 구성들.
- **CUDA 코어 (범용 일꾼):** 각 SM은 128개의 FP32 자료형의 CUDA 코어를 가져, 총 16,896개의 코어가 범용 스칼라/벡터 연산을 수행. 32개 스레드는 '워프(Warp)'라고 불리는 단위로 묶여서, 동일 명령을 동시에 병렬로 처리하는 SIMT 아키텍처로 작동.
- **텐서 코어 (행렬 곱셈 전문가):** 딥러닝의 핵심인 행렬 곱셈-누적연산(MMA)에 특화된 가속기. Blackwell 아키텍처에서는 FP16, BF16 뿐만 아니라 FP4 같은 저정밀도 데이터 타입에서 폭발적인 성능을 제공하여 LLM 학습/추론 속도를 극적으로 향상시킴.
- **메모리 계층:** 각 SM 내부에는 스레드 간 데이터 공유를 위한 초고속 저지연 L1 캐시/공유 메모리(256KB)가 있고, 모든 SM이 공유하는 대용량 L2 캐시(H100 기준 50MB)가 HBM으로의 접근을 줄여줌. 모델 가중치와 같은 거대 데이터는 주 메모리인 HBM에 저장.
- **Blackwell (B200)의 진화:** Hopper 대비 더 많은 192개(Blackwell Ultra는 160개 SM)의 SM을 탑재하고, 5세대 텐서 코어는 FP4/FP6 초저정밀도의 빠른 연산에 최대한 집중. 특히 텐서 코어의 전용 버퍼 역할을 하는 '텐서 메모리(TMEM)'를 신설했는데, 텐서 코어의 유휴 시간을 최소화하고 처리량을 극대화하기 위함. '어떻게 하면 텐서 코어를 한 순간도 쉬지 않고 일하게 만들까?'가 설계의 핵심

GPU는 수백 개의 비교적 작은 SM(다수의 범용 코어 + 행렬 특화 코어)을 집적하여 다양한 워크로드에 대한 유연성을 극대화하는 'Many-core' 접근방식 세대를 거듭하면서, AI 임무를 가장 잘 수행하는 텐서 코어가 SM의 주인공이 되었고, CUDA 코어는 그 전문가를 지원하는 역할로 무게 중심 완전히 이동



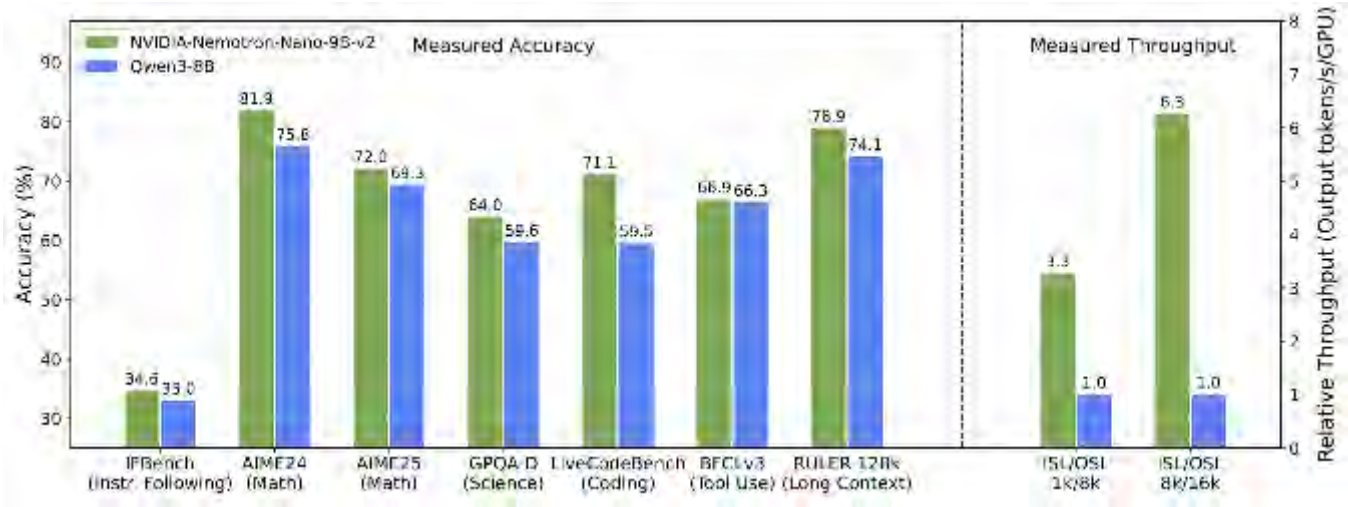
자료: 구글 딥마인드, MatX, 미래에셋증권 리서치센터

[참고] 트랜스포머가 GPU에 최적화된 것처럼, 엔비디아도 LLM을 잘한다

헤게모니를 위협할 수 있는 차세대 기술을 스스로 주도, '혁신가의 딜레마' 돌파

- 엔비디아는 지난 8월, Nemotron Nano 2라는 직접 만든 언어모델을 출시. 비슷한 크기(파라미터)를 가진 AI 모델들과 비교했을 때도 괄목할 성능과 비용을 과시.
- 특히 이 모델은 주류 방식인 트랜스포머 아키텍처를 그냥 쓰지 않았음. 트랜스포머는 문장이 길어질수록 계산량이 기하급수적으로 늘어나는 단점. 이를 보완하기 위해 "Mamba"라고 불리는 '상태 공간 모델(SSM)'을 트랜스포머와 결합시켜 하이브리드 모델로 만든 것. 복잡한 추론 능력은 유지하고, 속도를 극적으로 끌어올려 효율성 측면에서 판도를 바꿀 수 있는 중요한 기술적 진보로 볼 수도 있음.
- 트랜스포머 아키텍처의 성공에 가장 수혜를 보고 있는 기업이 엔비디아인데, 現 헤게모니를 위협할 수 있는 "새로운 과학" 기술도 선제적으로 채택했다는 점에서 이 기업이 얼마나 "AI 풀스택"에 탄탄한 기반을 갖춘 기업인지 실감. 어떤 AI 아키텍처가 표준이 되든 자사의 생태계에서 가장 잘 작동하게 만들어 플랫폼 지배력을 공고히 하려는 전략. Kodak 필름이나 NOKIA가 되지 않기 위한 노력의 일환.

'네모트론-나노-9B-v2'(녹색)와 경쟁 모델인 'Qwen3-8B'(파란색)를 여러 전문 분야 및 처리량 속도를 테스트한 결과
특히, 복잡한 추론이 필요한 수학(AIME)과 코딩(LiveCodeBench)에서 압도적 격차
처리량(GPU당 초당 출력 토큰 수) 측면에서 Qwen3보다 3배~6배 가량 빨라, AI 비용을 획기적으로 절감할 수 있단 의미



자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 경쟁 우려② 신흥 NPU 업체들의 딜레마

엔비디아의 다층적 시스템적 해자

- 스타트업은 현재 시장을 지배하는 트랜스포머 모델이 요구하는 '거대 행렬 곱셈'에 최적화된 칩을 보통 3~5년에 걸쳐 설계. 그러나 칩이 출시될 즈음, 최첨단 AI 모델은 이미 엔비디아의 차세대 칩 기능(예: Blackwell의 FP4, 텐서메모리)에 최적화된 MoE(Mixture of Experts), Sparsity 기반의 '수많은 작은 행렬 곱셈'을 요구하는 아키텍처로 진화해 버림. 야심 차게 개발한 칩은 출시되는 순간 구식이 되어버린 것.
- 과거 Groq, Cerebras가 SRAM에 베팅했으나, 모델 사이즈가 폭발적으로 커지며 그들의 SRAM 용량마저 초과하면서 실패. 이처럼 특정 모델에 대한 과도한 최적화는 AI 패러다임이 빠르게 변하는 시장에서 필연적으로 실패로 귀결. 이는 AI 연구와 칩 개발의 공진화에 따른 결과. 스타트업들이 최적화한 특정 워크로드가 자신들의 칩 출시 주기 동안 시장의 주류에서 벗어나지 않기를 기도해야 함. 이는 기술적 우위와 시장의 운이 동시에 따라야 하는 "낮은 확률의 도박"이라고 할 수 있음.
  - Groq은 최근 사우디아라비아 딜이 지연. 올해 매출 전망이 20억 달러 이상에서 5억 달러 이상으로 75% 급감.
- 스타트업이 생존하려면 엔비디아 대비 최소 수 배의 하드웨어 효율성을 달성해야 함. 그러나 이런 이론적 우위는 엔비디아의 시스템 우위(CUDA 생태계, 공급망 장악력 등)에 따라 체계적으로 저지될 가능성.

엔비디아의 시스템적인 우위를 정리

핵심 경쟁력	주요 전략 및 기술	이를 통한 우위	경쟁사에 미치는 영향
시스템 전체의 격차	단일 기술이 아닌 공급망, 기술, 속도 전체를 결합	– 견고한 시장 지배력 구축 – 작은 성능 우위는 쉽게 무력화 가능	– 칩 스펙 하나만으로는 공략 불가 – 생존을 위해선 몇 배의 압도적 효율격차 필요
네트워킹 장악	– 스케일업 독자 규격 "NVLink" – 스케일아웃 독자 규격 "InfiniBand"	– 수만 개 GPU 클러스터의 병목 구간 완벽 통제 – 개별 칩이 아닌 클러스터 전체의 효율 극대화	– 단순 GPU 칩만으로는 흉내 낼 수 없는 시스템 레벨의 격차 발생
메모리 및 공정 선점	– SK하이닉스와의 파트너십으로 최첨단 HBM 선제 확보 – TSMC 최첨단 공정 우선 활용	– 최신 HBM 메모리 최초, 최대 물량 확보 – 근본적인 물리적, 기술적 우위 선점	– 경쟁사는 항상 이전 세대 기술 또는 부족한 물량으로 경쟁해야 함
속도의 지배	– 기획 → 출시 → 대량 생산까지의 압도적 사이클 속도	– 시장 변화에 가장 빠르게 대응 및 선도 – 차세대 제품을 통한 시장 연속 지배	– 어렵게 제품을 출시해도, 엔비디아가 이미 다음 세대 제품을 발표/공급하는 사이클

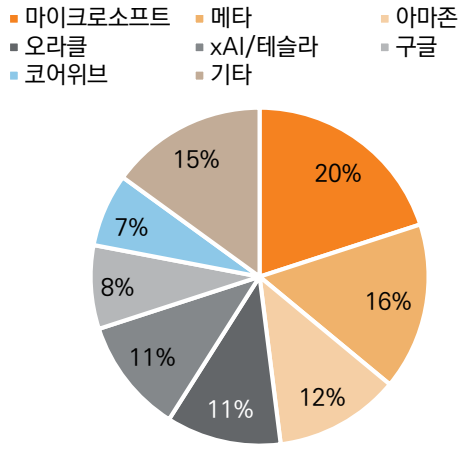
자료: Dylan Patel, 미래에셋증권 리서치센터

VIII. 엔비디아: 경쟁 우려③ 빅테크 자체 칩의 발전

내부용 비용 절감 툴이라는 존재적 한계

- 엔비디아에게 가장 큰 잠재적 경쟁 상대는 AMD도, 스타트업도 아닌, 빅테크들. GPU 사주는 가장 큰 손이기 때문. 자체 칩 준비 중인 OpenAI와 아마존으로부터 칩을 공급받는 Anthropic, 두 회사가 엔비디아 칩 수요의 30% 차지. 그러나, 이런 노력 역시 엔비디아의 아성을 무너뜨리기에는 한계.
- 빅테크가 3년 이상에 걸쳐 특정 내부 워크로드에 최적화된 칩을 개발하는 동안, 엔비디아는 1년 주기로 2~3세대의 범용 신제품을 출시하며 세대 격차를 벌리는 중. 메타의 'Olympus'와 MS의 'Braga' 칩이 수개월에서 수년까지 지연되는 사례는 이러한 속도 경쟁에서 뒤처지고 있음을 시사.
- 하이퍼스케일러들 각자의 클라우드 사업의 핵심인 '외부 고객(개발자 등)'들은 CUDA 생태계에 깊이 종속되어 있음. 이들에게 갑자기 구글이 TPU를, 아마존 AWS가 Trainium를 쓰라고 강요하는 것은 심각한 고객 이탈 리스크를 감수해야 하는 불가능한 선택. 자체 칩은 엔비디아 GPU 비용을 줄이기 위한 '내부용 비용 절감 도구'. 수많은 외부 고객을 위한 범용 컴퓨팅 시장과는 본질적으로 다른 이야기.
  - AI 개발자들은 GPU 확보량을 기준으로 AWS, Azure, GCP 같은 클라우드 서비스를 선택. 클라우드 기업들 간의 경쟁에서 승리하기 위한 조건이 "누가 더 많은 엔비디아 GPU를 확보하여 고객에게 제공할 수 있는가"에 달리게 되었단 뜻

엔비디아가 주력으로 밀고 있는 시스템 제품인, NVL72 랙 시스템의 주력 고객사들 비중 추정



자료: 미래에셋증권 리서치센터

GB200/GB300 NVL72의 ODM(하이퍼스케일러들은 OEM 아닌 ODM 방식 선호) 출하량 전망
CY 기준, 올해 ODM 관련 매출예상액만으로 작년 전체 매출 전체(1,247억 달러) 거의 다 덮을 정도

	Q4'24E	Q1'25E	Q2'25E	Q3'25E	Q4'25E	2025E
폭스콘	10	175	3,000	6,000	7,500	16,675
콴타	5	140	1,300	3,400	4,100	8,940
위스트론	3	80	1,900	3,100	3,700	8,780
인벤텍	0	0	100	200	250	550
TOP 4	18	395	6,300	12,700	15,550	34,945
기가바이트, 위윈, 페가트론, SMCI, 아수스, 레노보 등	0	5	500	700	900	2,105
기타 업체 점유율	0%	1%	7%	5%	5%	6%
총 GB200/300 NVL72 유닛	18	400	6,800	13,400	16,450	37,050
총 GB200/300 GPU 갯수	1,296	28,800	489,600	964,800	1,184,400	2,667,600
GB 시리즈 중 GB300 비중	–	–	10%	20%	40%	–
매출 예상(백만 달러)	**54**	**1,200**	**21,080**	**42,880**	**55,930**	**121,090**

자료: 미래에셋증권 리서치센터

VIII. 엔비디아: 경쟁 우려③ 빅테크들도 AI 칩 개발이 쉽지가 않다

빅테크들은 AI 칩 개발의 기술적 난이도를 과소평가했다

- 마이크로소프트는 원래 2025년 하반기 출시 예정인 Maia 200 칩(코드명: Braga)을 2026년으로 최소 6개월 연기. 당초 계획했던 세 가지 추론용 칩(Braga, Braga-R, Clea)의 일정과 사양 모두 하향 조정.
- 이는, OpenAI의 요청으로 인한 예상치 못한 설계 수정이 시뮬레이션에서 불안정성을 야기했고, 새로운 설계 기반의 칩 개발이 예상보다 훨씬 어렵다는 게 이유였음. 현실적이지 않은 일정 고수로 인한 직원 스트레스 등으로, 칩 설계 팀에서 최대 20%의 직원이 퇴사했다고 알려짐.
- 심지어 2026년 출시 예정인 Braga 칩도 엔비디아의 Blackwell 성능에 크게 못 미칠 것으로 전망. 이에 따라 2028년까지 AI 칩 로드맵을 더 단순한 설계로의 개발 과정 간소로 전면 재검토하게 될 것.
- 빅테크 기업들이 AI 칩 개발의 기술적 난이도를 과소평가했음이 드러난 사건. 젠슨 황은 "구매 가능한 GPU보다 성능이 떨어진다면 ASIC을 개발할 이유가 무엇인가?"라고 말한 바 있음. 마이크로소프트를 비롯한 많은 기업들의 자체 칩 개발 프로젝트에 대한 회의적 시각을 반영.

줄줄이 출시를 앞두고 있는 엔비디아의 신제품들

제품명	출시일자	다이 구성	제조 공정	패키징	HBM 구성	HBM 스택	TDP (W)
B300A	2025년 중반	싱글 다이	TSMC N4P	CoWoS-L	HBM3E 144GB	12 채널 4개 스택	600
B300 Ultra	2025년 하반기	듀얼 다이	TSMC N4P	CoWoS-L	HBM3E 288GB	12 채널 8개 스택	1,000
Rubin	2026년 하반기	듀얼 다이	TSMC N3P	CoWoS-L	HBM4 288GB	12 채널 8개 스택	2,300
Rubin Ultra	2027년 하반기	쿼드 다이	TSMC N3P	CoWoS-L	HBM4E 1024GB	12 채널 16개 스택	4,600?

자료: 엔비디아, 미래에셋증권 리서치센터

[참고] 2026-2027 빅테크들의 주요 ASIC 프로젝트 정리 (추정)

2026년에도 아직 과도기, 2027년은 돼야 본 게임

항목	TPU V7p	TPU V7e	TPU V8p	TPU V8e	MTIA Athena	MTIA Iris	MTIA Arke	AWS Tranium 3	AWS Tranium 3 Lite	MSFT Maia 200	MSFT M300	MSFT M400	OpenAI 1st	OpenAI 2nd	Apple
사용자	구글	구글	구글	구글	메타	메타	메타	아마존	아마존	MS	MS	MS	OpenAI	OpenAI	애플
시기	3Q26	2026년 말	3Q27	3Q27	1Q26	2H26	1Q27	2Q26	3Q26	1H26	?	?	2H26	2027	2027
설계업체	브로드컴	구글, 미디어텍	브로드컴	구글, 미디어텍	브로드컴	브로드컴	미디어텍	Annapura (아마존), Alchip	마벨테크	GUC (TSMC 자회사)	마벨테크	-	브로드컴	브로드컴	브로드컴
공정	N3(연산) N4(I/O)	N3	N2	N2	N3E(연산), N5(I/O)	N3E(연산), N5(I/O)	N2	N3	N3	N3	N2	N2	N3	A16	TBD
메모리	HBM3e 12hi 8스택	HBM3e 12hi 6스택	HBM4	HBM4	216GB HBM3e 12hi	288GB HBM3e 12hi	HBM4	144GB HBM3e 12hi	HBM3e 12hi	HBM3e	HBM4	HBM4	HBM3e	SRAM	TBD
후공정	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-L	?	CoWoS-S	CoWoS-R	CoWoS-S	CoWoS-S	CoWoS-R	CoWoS-S	CoWoS-R	TBD

수혜 관련 쪽은 TSMC(로직+CoWoS)와 HBM 테스트 핸들러 쪽일 수 있다고 판단.
특히 일정이 1H26~2H26에 몰려 있어 패키징과 HBM 캐파가 병목일 것으로 사료됨.
설계 파트너 지형 = 빅테크의 'Private Fabless' 전략. 구글·OpenAI·xAI는 브로드컴(대형 ASIC·인터커넥트), 메타는 미디어텍 등으로 설명.
빅테크들의 ASIC 출시 일정은 엔비디아의 Blackwell/Rubin 전환기와 겹침. 이번에는 성공할지 이번에도 Rubin에 우위를 내줄지가 관전 포인트.
OpenAI의 두 번째 칩은 "하이 리스크 하이 리턴". A16(백사이드 전원 등 신공정) + R 패키징은 전력밀도/수율 난이도가 급상승.

자료: 미래에셋증권 리서치센터

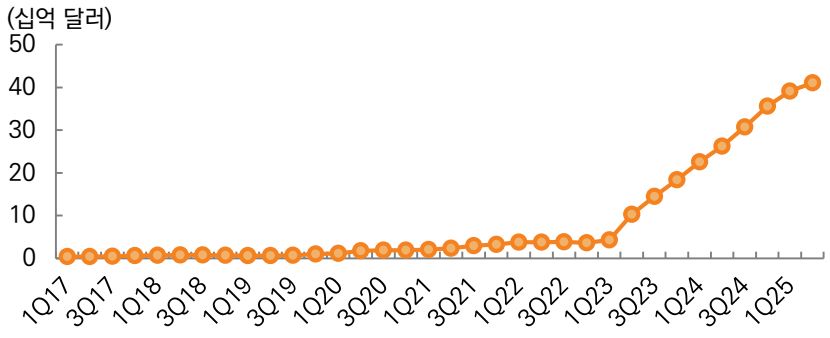
[참고] 위기의 레토릭 vs 엔비디아의 성장 역사

엔비디아도 원래 항상 위기는 있었다…

"엔비디아 위기론(빅테크 자체개발, 하이퍼스케일러들의 투자 축소, 對中규제)"과 관련한 역사적 레토릭을 분기별로 정리
동일 분기 내 관련 기사·리포트 건수가 10 건 이상이면 High / 4 ~ 9 건 mid / 1 ~ 3 건 Low로 표기

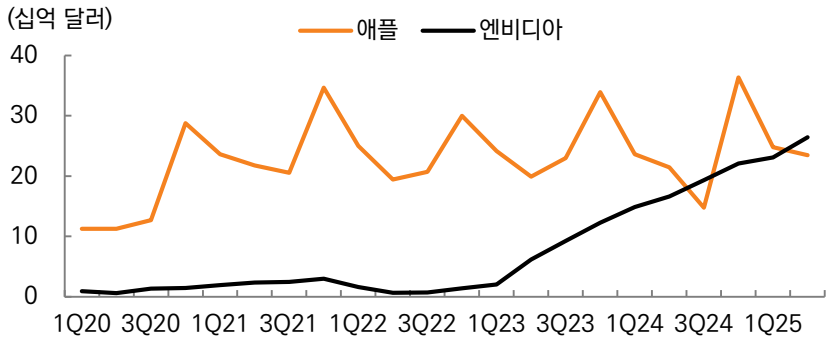
구분	1분기	2분기	3분기	4분기
2017	—	High (자체개발: 구글 TPU v2가 "GPU 대체" 헤드라인 다수)	mid	Low
2018	Low	mid (자체개발: 구글 TPU v3 소식)	Low	High (자체개발: AWS Inferentia 추론용 ASIC 전격 발표)
2019	High (투자축소: Gartner의 "데이터센터 지출 – 2.8 % 감소" 전망 → NVDA 수요 우려)	mid	Low	Low
2020	Low	Low	mid	High (자체개발: AWS Trainium 훈련용 ASIC 공개, "엔비디아 의존 탈피")
2021	Low	Low	Low	Low
2022	Low	mid	High (對中 규제: 美 상무부 A100·H100 수출 금지 발표)	High (對中 규제: 초광범위 'OCT 7 Rule' 시행)
2023	Low	High (자체개발: Reuters "Meta, 자체 AI 칩 프로젝트 재가동")	mid	High (對中 규제: WSJ "최대 50억 달러 中 주문 취소")
2024	Low	High (자체개발: Meta 2세대 'Artemis' 칩 공식 공개)	mid	High (對中 규제: 美, H20 · A800까지 금지·대역폭 한도 강화)
2025	mid (투자축소: CapEx 둔화 시 NVDA 매출 급락 시나리오)	High (자체개발: AWS Anthropic 전용 슈퍼컴퓨터 운영 중)	—	—

CY 기준, 엔비디아의 데이터센터 매출 분기별 추이



자료: 엔비디아, 미래에셋증권 리서치센터

CY 기준, 애플과 엔비디아의 순이익 분기별 추이



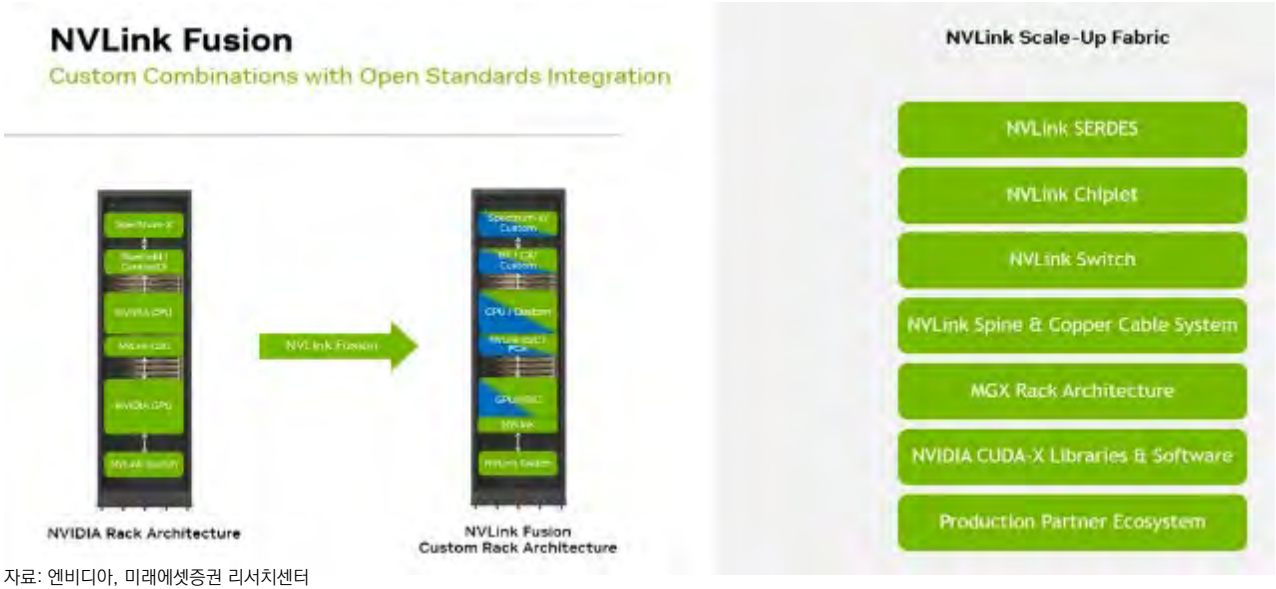
자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 빅테크發 경쟁 우려를 기우로 만드는 엔비디아의 포석

“다 들어와” NVLink Fusion

- 엔비디아의 독점 기술인 NVLink는 수만 개의 GPU가 하나의 거대한 컴퓨터처럼 작동하는 AI 클러스터에서 네트워킹은 성능을 결정하는 핵심 중 핵심 동맥. 이를 타사의 ASIC에도 개방하는 “NVLink Fusion” 전략을 통해 경쟁사마저 생태계로 끌어들이려고 노력 중. NVLink를 AI 인프라의 핵심인 스케일업의 표준으로 만들어, 미래 AI 데이터센터의 신경망을 장악하려는 거대한 전략적 포석.
- **시나리오** ① CPU, GPU, NVLink 스위치까지 모든 인프라를 엔비디아 제품으로만 구성하는 방식. 모든 구성 요소가 엔비디아에 의해 완벽히 최적화되어 있어 최고의 성능과 안정성 보장.
- **시나리오** ② 고객이 자체 개발한 칩을 엔비디아의 스케일업 인프라에 통합. 엔비디아는 “NVLink 칩렛”이라는 작은 칩을 제공. 이 칩렛이 ASIC 옆에서 신호를 NVLink 프로토콜로 변환하는 ‘통역사’.
- **시나리오** ③ 고객이 자체 개발한 고성능 CPU를 중심으로 엔비디아 GPU와 긴밀한 통합을 원하는 경우. 엔비디아는 고객 CPU가 NVLink를 '네이티브'로 지원할 수 있도록 IP를 직접 제공.

엔비디아는 “NVLink 칩, 스위치, 케이블, 랙 구조 등” 스케일업에 필요한 모든 부품과 기술을 제공
이를 통해 하이퍼스케일러 같은 거대 기업들은 자신들의 필요에 꼭 맞는 최적의 AI 컴퓨터를 유연하게 설계 가능

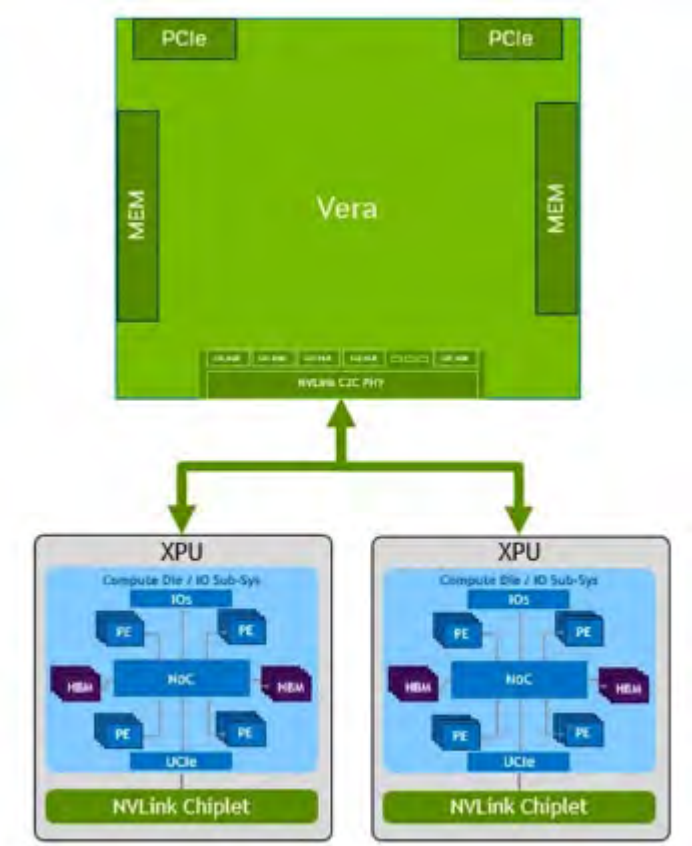


자료: 엔비디아, 미래에셋증권 리서치센터

[참고] NVLink Fusion은 고객사의 칩 종류에 따라 맞춤형으로 연결

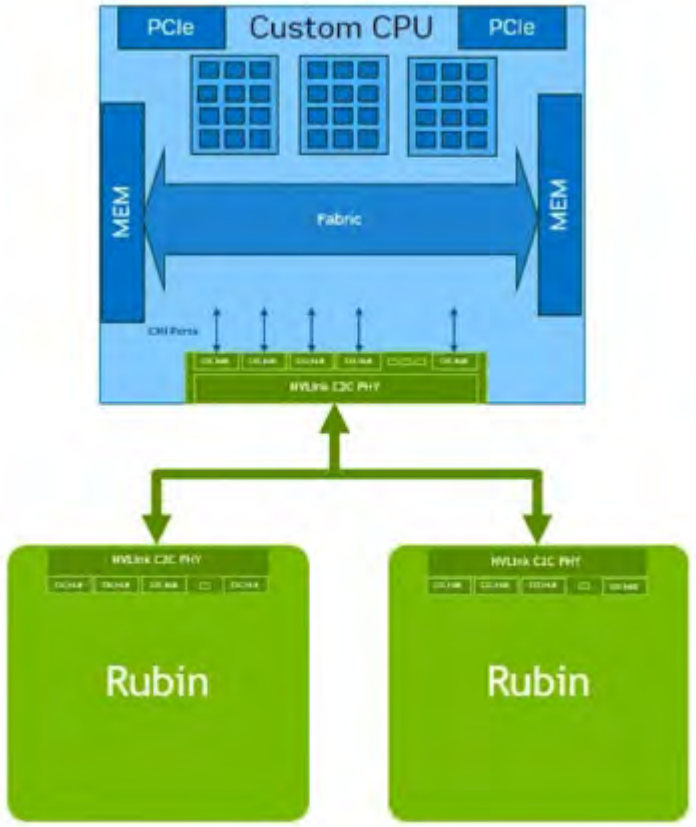
"어떤 회사든 본인들의 칩을 만들더라도, 엔비디아의 강력한 스케일업 생태계에 손쉽게 참여할 수 있다."

고객사의 맞춤형 XPU(AI 가속기) 연결: 'NVLink 브릿지 칩렛'을 사용
이 칩렛은 개방형 표준인 UCIe 인터페이스를 통해 고객사의 XPU와 연결



자료: 엔비디아, 미래에셋증권 리서치센터

고객사의 맞춤형 CPU 연결: NVLink-C2C IP 고객사 CPU에 직접 통합
별도의 번역기 칩렛을 사용하는 대신, 양쪽 칩이 마치 하나처럼 긴밀 통신



자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: NVLink Fusion이 호조를 보일 이유

NVLink Fusion은 이미 고객사 확보 단계

- NVLink Fusion을 사용할 고객사 확보의 영업력을 높이기 위해 엔비디아는 이미 파트너사들을 꾸렸음. Lattice, Astera Labs, Marvell, MediaTek으로 이뤄진 이들은, 각자 본인들의 고객 요구사항에 맞춰 NVLink Fusion 기술을 통합한 커스텀 ASIC 또는 세미커스텀 칩을 설계하고 공급하는 역할을 맡게 됨. 이들은 엔비디아로부터 NVLink 관련 IP를 제공받아, 고객의 특정 요구(예: 특정 AI 워크로드 가속, 저전력 GPU 설계 등)를 만족시키면서 엔비디아 NVLink 생태계와 호환되는 칩을 개발하게 됨.
- 더욱 중요한 점은, 엔비디아가 EDA(Electronic Design Automation) 기업들까지 NVLink Fusion의 주요 파트너로 이미 확보했다는 사실. 이를 통해 어떤 회사든(ASIC 공급업체나 CPU 벤더뿐만 아니라 다른 어떤 칩 설계 회사라도) 자신들의 칩에 NVLink 기능을 통합하는 것을 지원. 또한 NVLink 기술이 더 넓은 범위의 칩 설계자들에게 접근 가능하고 쉽게 채택될 수 있도록 하는 중요한 인프라 역할을 수행하게 될 가능성.

NVLink를 다양한 파트너사(ASIC, CPU, EDA)에 개방하여 생태계를 확장하는 모습
'스케일업(Scale-up)' 패브릭 기술을 고객사에게 개방하기 위한 엔비디아의 큰 그림에 참여하는 파트너 그룹과 그 역할

파트너 그룹	주요 기업	역할	핵심 가치
ASIC/세미커스텀 공급업체	• Lattice • Astera Labs • Marvell • MediaTek	• 하이퍼스케일러/기업 고객의 세미커스텀 파트너 • 고객 요구사항에 맞춘 NVLink 호환 칩 설계/공급 • NVLink 칩렛/IP 통합 지원	• 고객별 맞춤형 AI 가속기 제작 • 특정 워크로드 최적화 • 엔비디아 에코시스템과의 호환성 보장
CPU 벤더	• Fujitsu • Qualcomm	• 자체 개발 CPU에 NVLink 인터페이스 직접 통합 • 엔비디아 GPU와 고속 통신 가능 내장 • NVLink IP를 자사 CPU 아키텍처에 최적화 • NVLink IP를 모든 칩 설계자에게 제공	• 자체 CPU 생태계 유지하며 엔비디아 GPU 활용 • CPU-GPU 간 초고속/저지연 통신 • AI/HPC 워크로드 성능 극대화 • NVLink 기술의 광범위한 접근성 확보
EDA 파트너	• Cadence Design • Synopsys	• NVLink에 필요한 설계 블록과 검증환경을 EDA 툴에 포함 • 설계/검증 표준화 지원	• 칩 설계 복잡성 감소 및 개발 기간 단축 • 업계 표준 설계 도구를 통한 쉬운 채택

자료: 엔비디아, 미래에셋증권 리서치센터

VIII. 엔비디아: 엔비디아가 노리는 궁극적인 “수 경원”의 시장은?

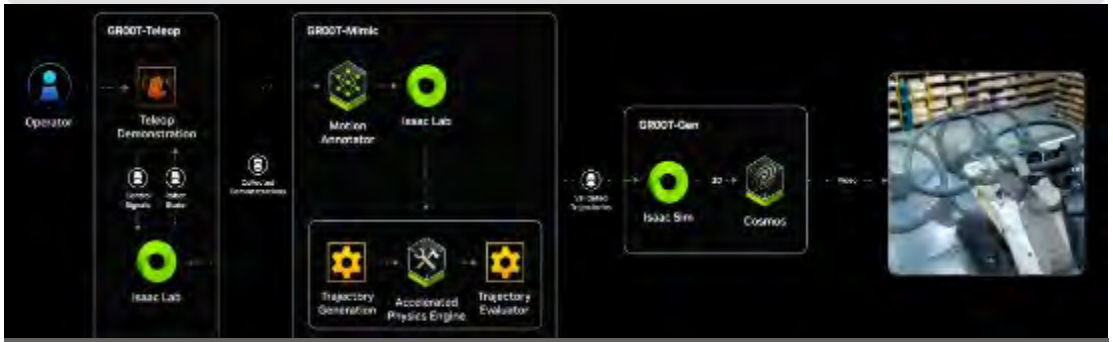
결국에는 자율주행과 (무엇보다) 로보틱스다

- 엔비디아의 비전은 현재의 '생성 AI'를 기반으로, 명확한 3단계 로드맵을 통해 “50조 달러” 규모의 산업 시장을 공략하는 것. 생성 AI는 현재 비즈니스의 핵심 동력이자 모든 미래 전략의 기술적 기반이라면, 에이전트 AI는 다음 단계의 교두보. 추론하고, 계획하고, 행동하는 AI라고 지칭할 수 있음.
- 엔비디아의 궁극 지향점은 현실의 물리법칙을 이해하고 상호작용하는 “Physical AI”이고, 이를 위한 컴퓨팅 플랫폼의 표준을 제공하는 것이 비전. 물리 AI 구현의 핵심은 기존 LLM이 할 수 없는 '직관적 물리' 지식을 갖춘 모델. 이를 위해 세계 최초의 “월드 파운데이션 모델: Cosmos”를 출시. 물리적 속성이 강하게 드러나는 2천만 시간 분량의 비디오와 Omniverse 3D 에셋을 통합 학습한 것. 텍스트, 이미지, 비디오 프롬프트를 입력받아, 물리 법칙에 정렬된 가상 세계를 2D 비디오로 생성. 창의적 영상이 아닌, 물리 세계를 이해하고 시뮬레이션하기 위한 용도로, 로보틱스/자율주행 AI의 근간이 됨.
- 물리 AI, 특히 로보틱스 발전의 가장 큰 장벽은 훈련 데이터의 절대적 부족. 로봇 학습 데이터는 인터넷 스크래핑이 불가능하기 때문. 인간이 직접 원격 조종하는 텔레오퍼레이션에 의존하는데, 이것은 확장성이 극히 낮음. 데이터 수집량이 하루 최대 24시간으로 제한되기 때문. 엔비디아는 이 문제를 “소량의 실제 데이터로 무한한 양의 고품질 합성 데이터를 생성하는” Cosmos로 해결했다는 것.
  – 자율주행은 인간이 항상 운전하므로 데이터 수집 용이, 특정 작업을 수행하는 로봇의 모방 데이터 수집은 훨씬 더 어려움

젠슨 황이 정의한 AI의 4단계 진화론
수십조 달러 규모의 “물리적 AI” 시장을 공략하겠다

자료: 엔비디아, 미래에셋증권 리서치센터

물리 AI(로보틱스)를 위한 데이터 병목을 돌파하기 위한 엔비디아의 전략: '합성 데이터' 생성
엔비디아 “Isaac GR00T”의 합성 동작데이터 생성은 로봇의 인간 모방 학습을 위한 워크플로우



숙련된 작업자가 10여 개의 **원격 시연 데이터를 수집**한 뒤, 이를 기반으로 Cosmos가 새로운 환경과 지시에 맞춰 물리적으로 타당한 수많은 **가상 2D 비디오(' 꿈 ')를 자동으로 생**성. 이 2D 비디오를 로봇이 실제로 실행 가능한 **3D 행동 궤적 데이터로 추출**함. 이 가상의(합성의) 궤적 데이터로 **로봇을 위한 AI 모델을 훈련**.

자료: 엔비디아, 미래에셋증권 리서치센터

[참고] 엔비디아의 '물리 AI'를 위한 데이터 연금술

로보틱스 발전의 가장 큰 병목을 예측 가능한 컴퓨팅 비용으로 해결하는, 확장 가능한 산업적 해법 중 하나

엔비디아의 "DREAMGEN" 논문에서 확인되는 엔비디아의 로봇용 데이터 생성/AI 훈련 방법론인 "Cosmos"와 "GR00T Dreams"의 효율성
엔비디아는 로보틱스를 위한 풀스택 플랫폼 ' Isaac GR00T(범용 모델+합성 동작 데이터 생성+시뮬레이션+엣지 컴퓨팅) ' 를 주력으로 밀고 있음

항목	수치	비고
하드웨어 사양		
GPU 모델	L40	Ada Lovelace 아키텍처
GPU 메모리	48GB GDDR6	
클라우드 시간당 비용	$1.00/시간	Hyperstack 클라우드 기준
RoboCasa*에서 대규모 실험		
생성된 샘플 수	240,000개	원본 데이터에서 가상으로 만든 데이터 샘플 숫자
사용된 GPU 수	1,500개	L40 GPU
소요 시간	54시간	
총 GPU 시간	81,000시간	1,500 × 54
추정 총 비용	$81,000	Hyperstack 클라우드 기준
샘플당 비용	$0.34	$81,000 ÷ 240,000
각 task당 "가짜 3D 영상" 생성 비용		
"GR1" 휴머노이드	$102	각 task 당 300개 생성(망치질, 닦기, 접기, 쌓기)
"Franka" 로봇팔	$34	각 task당 100개 생성(우유 붓기, 큐브 쌓기, 초콜릿 담기)
"SO-100" 로봇팔	$15.5	task 각각 40개, 50개 생성(딸기 줍고 놓기, 택택토)
스케일링 효율성		
원본 데이터 대비 배수	333배	원본 데이터 720개로 24만개의 액션 데이터를 생성
기존 방식 대비 데이터 수집 비용	99% 절감	텔레오퍼레이션 대비

GR00T Dreams 기술의 방식: 인간 시연 데이터로 Cosmos 모델을 미세조정하면 "꿈"이 생성 + 꿈들을 로봇이 따라할 수 있는 3D 행동 궤적으로 변환

자료: 엔비디아 논문 "DREAMGEN", 미래에셋증권 리서치센터

VIII. 엔비디아: 'REAL RECOGNIZE REAL'

엔비디아 VS 테슬라

- 엔비디아의 물리 AI 플랫폼은 이미 수조 달러 규모의 시장에 구체적으로 적용되며 수익을 창출 중. 1조 달러 규모의 창고·물류 시장을 목표로 Kion, Accenture와 협력. Omniverse 디지털 트윈 내에서 AI가 수많은 시나리오를 시뮬레이션하여 현실 세계의 운영을 최적화. 그리고 중요한 분야는 수 조 달러의 시장인 자율주행으로, 엔비디아는 여기서 이미 연간 50억 달러의 런레이트. Omniverse와 Cosmos로 물리적으로 정확한 '엣지 케이스' 데이터를 무한 생성. 수천 번의 실제 주행을 수십억 마일의 학습 효과로 확장. Toyota를 포함한 거의 모든 자동차 회사가 파트너.
- 그러나 물리 AI의 정점은 휴머노이드. 이는 테슬라와 완전히 겹치는 미래 비즈니스 분야라 할 수 있음. 엔비디아는 로보틱스를 위해 풀스택 개발 플랫폼 "Isaac GR00T(Cosmos 기반의 범용 모델과 가상의 데이터 생성, Omniverse를 이용한 시뮬레이션 강화학습, 로봇에 탑재될 Jetson 칩)"를 제공.
- 그런데, 로봇 지능 구현에서 엔비디아와 테슬라는 데이터 확보에 대한 근본적 철학 차이를 보이며, 이는 시장의 미래 경로를 결정할 핵심 변수. 엔비디아는 본인들의 최고 역량인 "칩의 우위"를 적극 활용/홍보하는 방식이고, 테슬라는 FSD라는 범용 주행 지능을 로봇으로 확장한 관찰주의적 접근. 누가 옳을지 모르겠지만, 경쟁이 로보틱스 기술의 발전을 가속화해 대중화의 시대를 앞당기고 있음.

엔비디아와 테슬라의 로보틱스 훈련법에 관한 주요 비교대조
테슬라 방식: 유튜브 영상 → 비전 트랜스포머 → 동작 예측 → 모터 제어 및 실행 → 실 세상에서 연습 → 반복

구분	테슬라 Optimus	엔비디아 GR00T
기본 철학	"양이 질을 이긴다" – 직관적 패턴 인식	"소량 정예 + AI 상상력" – 생성적인 접근
데이터 원천	인터넷 영상 (YouTube 등)	소량 고품질 텔레오퍼레이션 데이터
데이터 규모	수십억 시간 (무제한)	수십~수천 개 실제 데이터를 몇 백배로 증강
학습 방식	End-to-End 직접 학습	4단계 파이프라인 (생성→추출→훈련)
영상 활용	3인칭 인간 행동 영상 직접 모방	1인칭 로봇 영상 생성 후 학습
신체 전략	휴머노이드 형태의 Embodiment 매칭	특정 로봇별 맞춤 미세조정
비용 구조	예측 불가능	비교적 예측 가능 (샘플당 $0.34)

자료: 미래에셋증권 리서치센터

IX. 테슬라

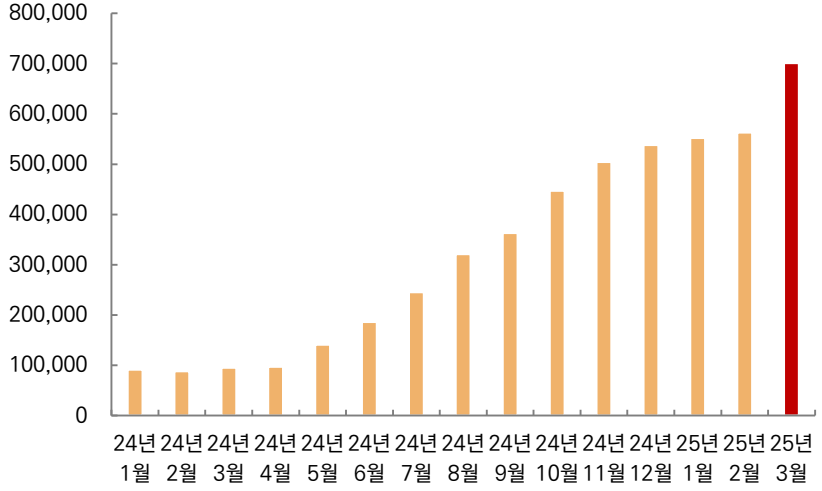
비전온리 로보택시 물리적AI

IX. 테슬라: 자율주행 수요는 이미 증명되고 있다

에이전트 시대를 알리는 중요한 폼팩터는 바로 자율주행 차량

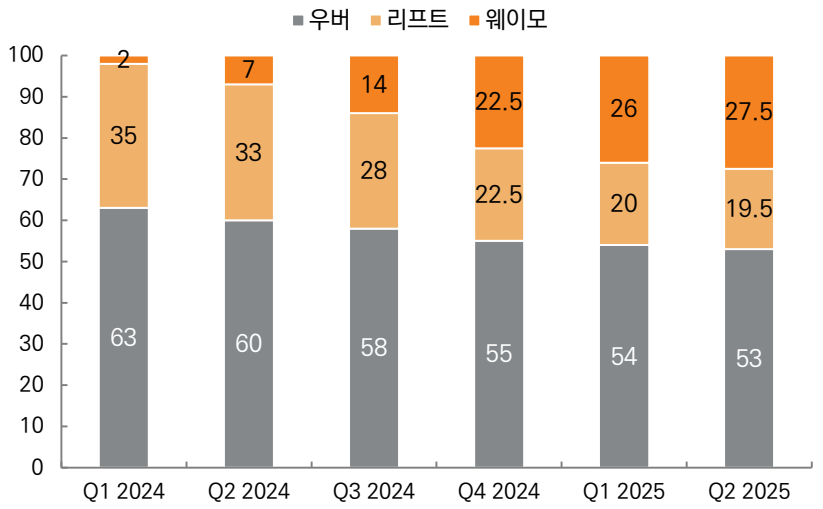
- 대중이 AI 시대라는 것을 확실히 실감하고 만끽하면서, 이 시장이 가속성장 하려면 소프트웨어 영역서 벗어나 그것을 뒷받침 할 수 있는 하드웨어가 필요. MP3라는 음성파일 확장자에, MP3 플레이어라는 하드웨어가 탄생해 시장이 성장했던 것이 그 예시. AI 에이전트도 마찬가지. AI를 일상에 서 잘 사용할 수 있게 하는 HW가 필수적. 자율주행 차량은 바로 그 역할을 할 핵심 '폼팩터'가 될 가능성.
- 자율주행 시장의 고성장은 더 이상 분석가들의 시나리오 속 가설이 아님. 구글의 웨이모(Waymo)는 애틀랜타에서 지난 6월 24일부터 Uber와 협력해 상용 서비스를 개시. Uber가 호출·결제·배차를 담당하고, 웨이모가 완전 자율주행 Jaguar 차량을 운영하는 형태.
- 흥미로운 점은, 이용자들이 택시 차량을 호출할 때 인간 기사가 배정되면 이를 의도적으로 취소하고 로보택시를 재요청하는 "뽑기 현상"을 보였다는 것. 이는 가격(UberX와 동일)을 넘어서, 무인 택시에 관한 대중들의 큰 선호도를 증명. 내년 쯤이면 애틀란타 지역의 Uber 트립의 20%를 감당하는 수준까지 성장할 것으로 보임. 웨이모는 450억 달러 이상의 기업가치로 큰 기대를 받고 있음.

웨이모의 캘리포니아 내 유료 무인주행 호출 건수 (추정)
유료 서비스가 본격화된 지 약 1년 반 만에 55배 이상 증가



자료: 캘리포니아 공공요금위원회(CPUC) 미래에셋증권 리서치센터

샌프란시스코 내 웨이모, 리프트, 우버간 라이드헤일링 부문 점유율 비교
웨이모는 '23년 8월 SF에서 서비스를 시작한 이후 시장의 4분의 1 차지



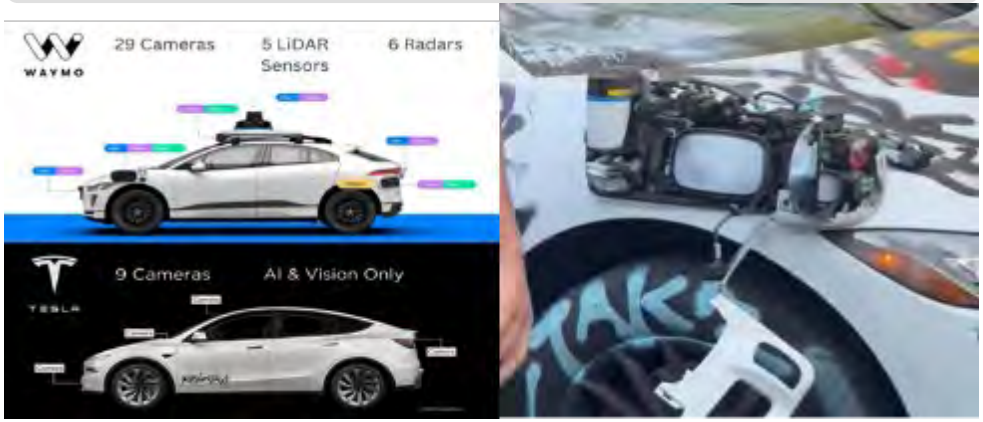
자료: YipitData, 미래에셋증권 리서치센터

IX. 테슬라: 자율주행의 확장성에 문제가 있는 웨이모

웨이모의 한계점은 바로 확장성… 너무 비싸기 때문이고, 이는 애초의 방법론 한계

- 그러나 웨이모는 구조적인 공급 비효율성에 직면. 애틀랜타 자율주행에 투입되는 초기 물량은 수십 대, 2026년까지의 목표도 고작 수백 대. 현재 기준, 5개 도시를 대상으로 총 1,500대의 웨이모 차량 운영 규모는 폭발 초기 단계의 수요에 비해 미미해 보임. 이는 일시적 병목 현상이 아닌, 차량 조달 능력 및 서비스 확장성이라는 근본적 한계를 드러냄. 이것은 웨이모의 방법론 때문.
- 비싼 센서(LiDAR, Radar 등), 그리고 미리 정밀하게 제작된 HD 맵에 크게 의존. 이 방식은 새로운 도시로 확장할 때마다 그 도시의 HD 맵을 만들고, 도로가 바뀔 때마다 계속 업데이트해야 하는 큰 번거로움이 존재. 즉, 초정밀 HD 맵에 의존하는 '궤도 위의 롤러코스터'. 정해진 구역을 벗어나면 무용지물이 되는 근본적인 확장성의 한계. 뿐만 아니라 각종 센서들이 주변 상황을 빠르게 인지하고 연산해도, 그 결정이 실제 차량 제어로 반영되는 과정에서 통신 지연과 불안정성(jitter)이 발생.
  - 웨이모의 자율주행 컴퓨터('뇌')가 차량의 실제 움직임을 담당하는 부분(스티어링, 브레이크, 가속 페달 등 '근육')에게 명령을 내릴 때, 이 둘은 원래부터 하나가 아니었기 때문에 '통역' 과정이 필요. 이 과정에서 지연이 생기고 불안정성 발생.
- 이러한 이유는, 웨이모가 사용하는 기본 차량(재규어 I-PACE, 크라이슬러 Pacifica)이 처음부터 자율주행을 염두에 두고 설계된 것이 아니라 인간 운전자를 위해 만들어진 플랫폼이기 때문. 여기에다가 각종 센서를 붙여야 하니 차량 한 대당 가격이 2억 원(15만 달러~20만 달러) 이상에 육박.

구글과 테슬라의 자율주행 차량이 갖고 있는 센서 구성의 차이
웨이모 차량은 몇 달 전, LA 폭동에서 복잡한 라이다 센서 시스템 연결 구성을 노출



자료: 구글 웨이모, 테슬라, X(@niccruzpatane), 미래에셋증권 리서치센터

웨이모 차량의 트렁크에 탑재된 온보드 컴퓨터(차량의 두뇌)
"서버급"의 고성능 CPU와 GPU가 탑재, "액체 냉각"도 필수

자료: X(@JeffTutorials), 미래에셋증권 리서치센터

IX. 테슬라: 로보택시의 확장성 (vs 웨이모)

로보택시 출시는 "실제 세계 AI로서의 ChatGPT 순간"

- 지난 6월 22일, 테슬라는 "운전석에 아무도 없는" 차량으로 유료 승객을 태우며 일론 머스크의 지난 10년간의 약속을 실현. 테슬라 로보택시의 등장이 주목받는 근본적인 이유는 압도적인 '확장성(Scalability)'에 있음. 로보택시 전쟁의 본질 또한, 차량을 도로에 공급할 수 있는 능력이 시장 점유율이라고 생각. 이는 자율주행 시장의 판도를 완전히 바꿀 수 있는 잠재력을 의미.
- 테슬라 로보택시의 접근 방식은 구글의 웨이모보다 스케일링에 있어서 더 유리한 고지를 점유. 구글 웨이모와 반대로, HD 정밀 맵 및 비싼 센서에 의존하지 않기 때문. 테슬라의 로보택시(및 FSD)의 지능은 "범용"으로서, 세상 어디에서 운전이 가능. '22년 11월, ChatGPT가 등장하면서 AI에 대한 우리들의 생각이 완전히 바뀌었던 것과 같은 패러다임 전환이 자율주행 분야에서도 일어났다는 의미.
- 이러한 확장성은 수요자 입장에서도 유리. 요금이 싸기 때문. 테슬라는 지난 7월 말부터 동적 요금제를 도입했음에도 불구, 경쟁사인 우버, 리프트는 물론, 직접적인 로보택시 경쟁자인 웨이모에 비해서도 압도적으로 저렴한 가격을 유지. 이는 경쟁을 거의 무의미하게 만들 정도의 패러다임 변화.

각 항목별 테슬라 로보택시 vs 구글 웨이모 비교
출시 후 두 달이 지난 현재, 테슬라는 이제 가격을 핵심 무기로 시장 공략

항목	테슬라	웨이모	현재 승자
접근성	정식 서비스	정식 서비스	타이
감독 방식	안전 모니터 요원이 조수석에 동승	완전 무인이거나, 때때로 드라이버 있음	웨이모
운행 구역	오스틴시 대부분	오스틴시 대부분	테슬라
UX: 차량 호출	로보택시 전용의 깔끔한 UI/UX. 항상 무인차 배정	Uber 앱 활용. 때때로 사람이 운전하는 차량이 배정	테슬라
UX: 차량 내 앱	스크린 1개로 단순하지만, 기능적 (유튜브, 넷플릭스 등)	여러 스크린으로 잘 다듬어졌지만 덜 기능적	테슬라
요금	**2km 정도 단거리: $3.8**	**2km 정도 단거리: $21.5**	테슬라
	18km 정도 장거리: $13.7	**18km 정도 장거리: $24**	

자료: 외신, 미래에셋증권 리서치센터

"중국판 FSD": 거의 물 위를 달려야 하는 조건에서도 자율주행이 된다
인위적인 지오펜스 및 HD 맵 없이 어디에서나 주행을 할 수 있는 능력

자료: X(@boolusilan), 미래에셋증권 리서치센터

[참고] 로보택시의 모습과 그 의미

자율주행의 “ChatGPT 순간”으로 나중에 회자될 로보택시

테슬라 로보택시 실제 운행 장면

자료: X(@DirtyTesLa), 미래에셋증권 리서치센터

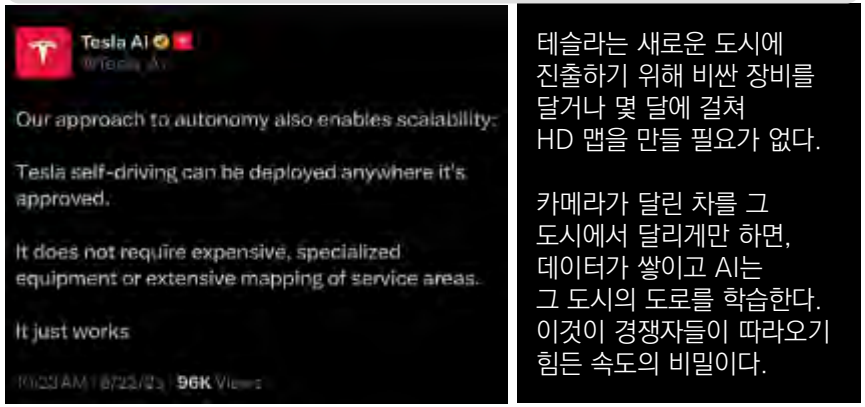
로보택시와 웨이모의 뒷좌석 UI의 차이점 / 로보택시 내 즐길 수 있는 앱

자료: X(@niccruzpatane), 미래에셋증권 리서치센터

테슬라가 로보택시를 모니터링 하는 모습: 원격 주행을 위한 핸들 없음
中 바이두의 Apollo Go 모니터링 모습: 원격 지원이 아닌 “원격 주행”

자료: 테슬라, Baidu, 미래에셋증권 리서치센터

테슬라 로보택시 접근 방식은 구글의 웨이모보다 스케일링이 유리한 이유
로보택시의 지능은 세상 어디에서도 운전이 가능한 '범용 지능'



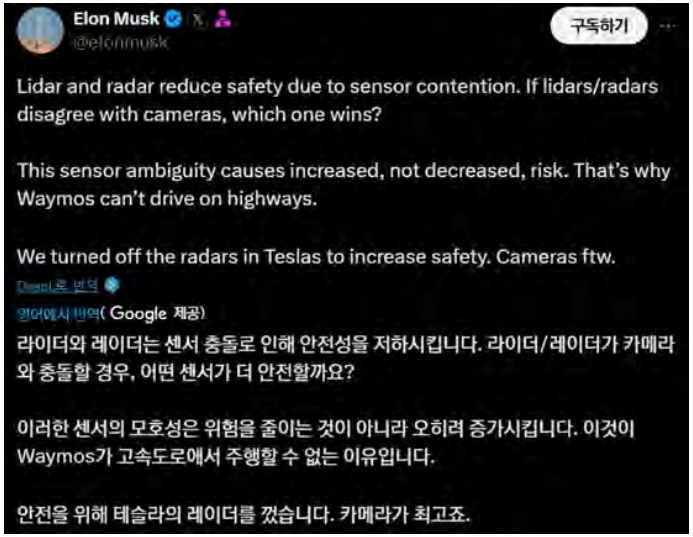
자료: Tesla, 미래에셋증권 리서치센터

IX. 테슬라: 웨이모와는 근본부터 다른 기술 철학이 결국 주류가 되나?

비전 온리가 답이다

- 테슬라는 인간처럼 카메라와 AI만 사용하는 '비전 온리(Vision-Only)' 접근법을 선택. 이 방식은, LiDAR이 갖는 악천후의 취약성과 센서 간 정보 충돌(Sensor Contention) 위험을 오히려 제거할 수 있음. 이는 도로 위 수백만 대의 기존 차량을 OTA 업데이트만으로 로보택시로 전환할 수 있는 유일한 길.
- 인간도 눈 두 개와 뇌 하나로 운전한다는 단순한 논리가 1년에 1,000대 만들기도 벅찬 웨이모와 "주당 5,000대"를 생산하는 테슬라의 격차를 만듦. 최근 중국의 웨이모라고 불렸던 바이두조차 LiDAR 노선을 포기하고 '순수 비전'으로의 전환을 선언. 시장이 테슬라의 방식을 정답으로 인정하기 시작.
- 특히 중국은 미국 및 유럽의 내연기관 업체들와 달리(GM은 로보택시 사업 "Cruise"를 포기) 자율주행 기술에 있어서 진심. 그런데 이들은 테슬라의 방식을 대대적으로 모방하고 있음. 자율주행 시장이 결국 '소수의 지배적인 플레이어만 살아남는 과점 시장'이 될 것으로 예측되는 가운데, 이 시장에서 테슬라와 함께 살아남는 것이 목표라는 것이 중국 자율주행 업체들의 판단. 화웨이, 샤오펑 등 대부분의 업체가 LiDAR 보다는 비전 온리로 전략을 선회 중.

일론 머스크가 지난 8월에 남긴 트윗
"센서 퓨전도 필요없고, 그냥 카메라만 쓰는 게 오히려 안전"



자료: 일론 머스크, 미래에셋증권 리서치센터

중국의 선도적인 자율주행 스타트업인 Momenta가 바라보는 자율주행 업체들의 위상
테슬라는 궁극의 업체이고, 자율주행의 애플(iOS)이라고 볼 수 있음.

경쟁사	자율주행 내 위상	주요 내용 및 전략
모빌아이	과거의 강자	• 자율주행 L1~L2 초반 시대의 시장 지배자였음 • 시장이 L2++로 진화하면서 차세대 플레이어에게 주도권을 넘겨줌
화웨이	중국 내 최대 라이벌	• 현재 가장 경계하는 핵심 경쟁 상대 • 약점: 폐쇄적인 수직 계열화 비즈니스 모델을 가지고 있음 • Momenta 평가: 화웨이의 아키텍처는 Momenta보다 1~2세대 뒤쳐짐
테슬라	**엔드게임의 상대**	**• 자율주행 시장의 '애플'과 같은 존재로, 최종 목표로 하는 궁극의 경쟁자 • Momenta는 '안드로이드'로 테슬라와 경쟁하는 시장 구도를 만들고자 함**

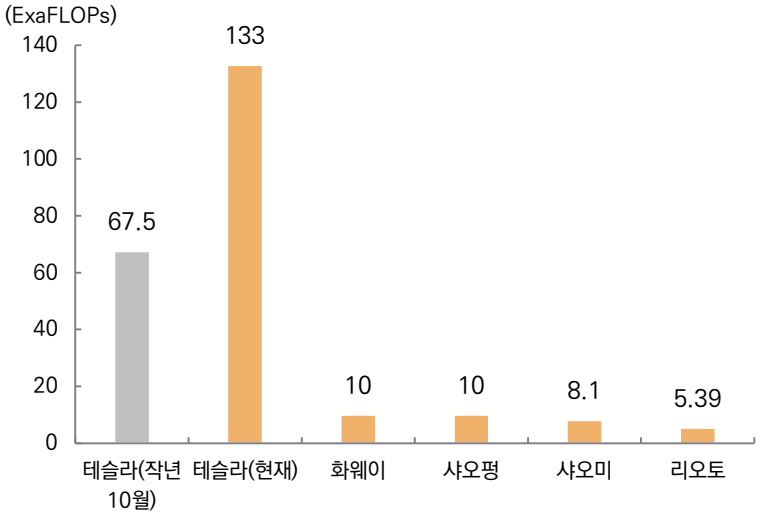
자료: 미래에셋증권 리서치센터

IX. 테슬라: 로보택시 요금이 싼 이유 = 쓰라린 교훈을 훌륭하게 수행

테슬라의 수직계열화 = 실제 주행 데이터 + 막대한 컴퓨팅 능력 + End-to-end

- 테슬라는 자동차와 자율주행 시스템(하드웨어, 소프트웨어)을 처음부터 함께 설계하고 만들어, 차량의 모든 세부 사항을 알고 있음. 자율주행 컴퓨터는 차량의 핵심 신경망처럼 처음부터 통합되어 있음.
- 인간처럼 오직 카메라(Vision)만 사용하고, LiDAR, Radar, HD 맵을 사용하지 않고 바로 실행에 옮기는 end-to-end 신경망의 방식을 통해 제어 지연(latency)이 매우 낮고, 지연 현상도 거의 없음.
- 이는, 인간이 규칙을 하나하나 코딩하고, 데이터를 선별하고, 복잡한 시스템을 설계하는 것이 단기적으로 효과가 있는 것처럼 보이지만 장기적으로는 방대한 데이터와 막대한 컴퓨팅 파워를 이용해 스스로 학습하는 단순함에 항상 패배한다는 AI 업계의 '쓰라린 교훈(Bitter Lesson)'을 바탕으로 한 것.
- 데이터 측면에서, 테슬라는 웨이모보다 하루에 40배, 전체 플릿 기준으로는 900배 더 많은 실제 주행 데이터를 수집. 방대한 주행 영상 데이터를 AI에게 주고 스스로 운전하는 법을 터득케 함. 로보택시 이벤트는 "가장 어려운 근본 문제를 해결했으니, 전 세계로 서비스를 넓히는 일만 남았다"는 선언인 셈.

주요 업체별 자율주행 모델 훈련용 컴퓨팅 용량
테슬라는 H100-equivalent 6.7만 장(H200 포함)의 GPU 보유



자료: 미래에셋증권 리서치센터

테슬라가 데이터를 한 데 모아서 훈련하는 방식도 결국 Bitter Lesson을 지지
테슬라는 주행 영상, GPS, IMU 등 원본 데이터(Clip)를 수집한 뒤, 강력한 컴퓨팅 파워를 갖춘 "오프라인 트래커"로 데이터를 3D로 완벽 재구성해 주행지능을 키움

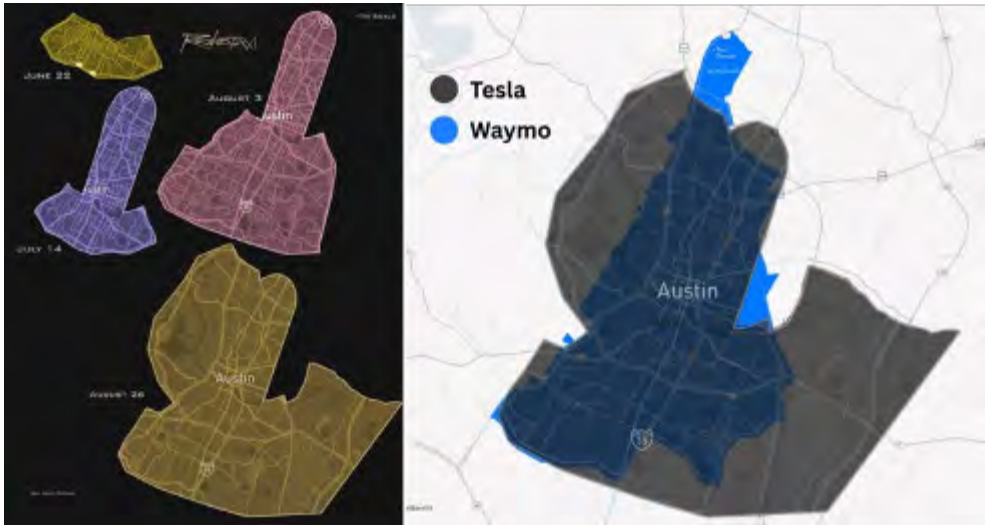
자료: 테슬라, 미래에셋증권 리서치센터

IX. 테슬라: 확장성에 대한 실체적 증명

웨이모보다 2배 가량 더 많은 면적을 커버하게 된 로보택시

- 테슬라가 로보택시를 오스틴에서 서비스를 시작한 이후 65일만에 서비스 지역(Geofence)을 18 제곱마일에서 171 제곱마일로 확장. 발표 당시보다 9.5배의 폭발적 성장. 심지어 테슬라의 차량 생산 기지인 "기가텍사스"까지 포함됐기 때문에, 차량이 생산되자마자 공장에서 바로 로보택시로 투입될 가능성도 내포.
- 시장의 선두주자 웨이모는 일찍이 오스틴에서 서비스를 운영해왔지만, 테슬라가 단 두 달 만에 웨이모의 서비스 지역을 완전히 포위/압도한 것. 이는 기술적 준비가 완료된 순간, 즉각적인 대규모 배포가 가능하다는 테슬라의 주장을 실체적 증명한 것이라 볼 수 있음. 현재의 테슬라가 설정한 "오스틴 지오펜스"는 기술의 한계가 아닌, FSD 잠재력을 로보택시라는 안정적 사업 모델로 만들기 위한 일시적 '운영의 틀'이라고 할 수 있음.
- 그렇다면 이 사업 모델을 더 지속확장하기 위한 남은 과제는 바로 '수익성 증명'이라고 할 수 있음.

테슬라의 오스틴에서의 로보택시 서비스 제공 면적 확장, 2달 만에 거의 10배로 확장
오스틴 내 테슬라(검은색)와 웨이모(파란색)의 서비스 지역을 직접 비교



자료: Nic Cruz Patane, Sawyer Meritt, 미래에셋증권 리서치센터

테슬라가 밝힌 로보택시 확장에 관한 주요 타임라인 전망
현재 사업 속도라면 올해 여러 지역으로 확장 가능성 청신호

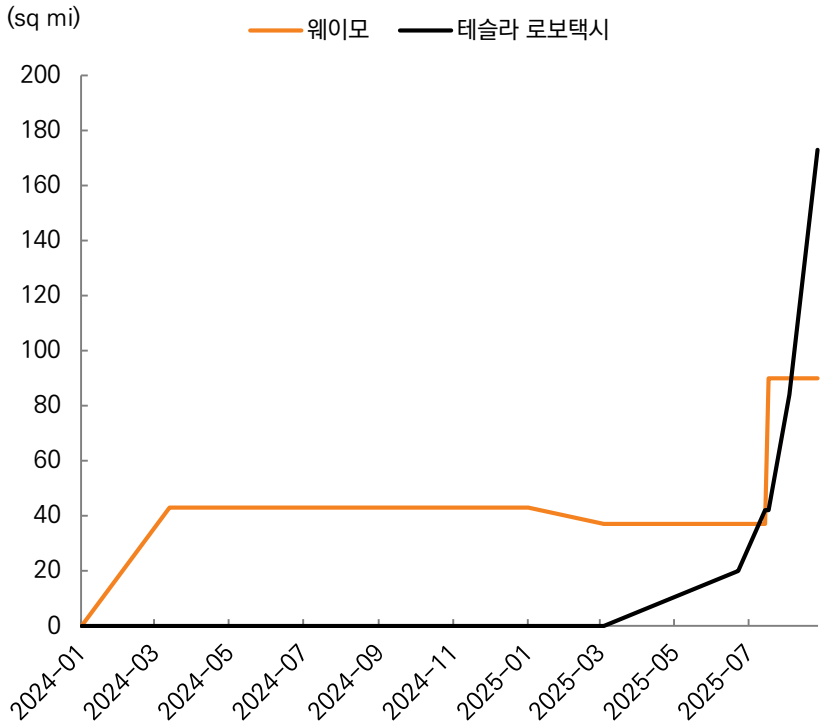
시기	주요 내용
2025년 6월	모델 Y 기반 완전 자율주행(무감독) 택시 서비스 "텍사스 오스틴"에서 운영 시작 (초기 10~20대 규모)
2025년 말	개인 사용자를 위한 완전 자율주행이 미국 내에 있는 "다수 도시"에서 가능할 것으로 예상 (차에서 수면 가능)
2025년 말	로보택시 운영 차량 대수를 "수천 대 규모"로 확대
2026년 중반	로보택시 사업이 "재무적으로" 의미 있는 기여를 시작
2026년 하반기	"미국 전역"에서 완전 무감독 자율주행 가능 목표
2026년 4분기	Unboxed 공정 활용하여 "사이버캡" 전용 모델 대량 생산
2026년 말	로보택시 운영 차량 대수를 최대 "백만 대" 정도로 확대

자료: 테슬라, 미래에셋증권 리서치센터

[참고] 테슬라 로보택시 vs 웨이모 서비스 면적 비교

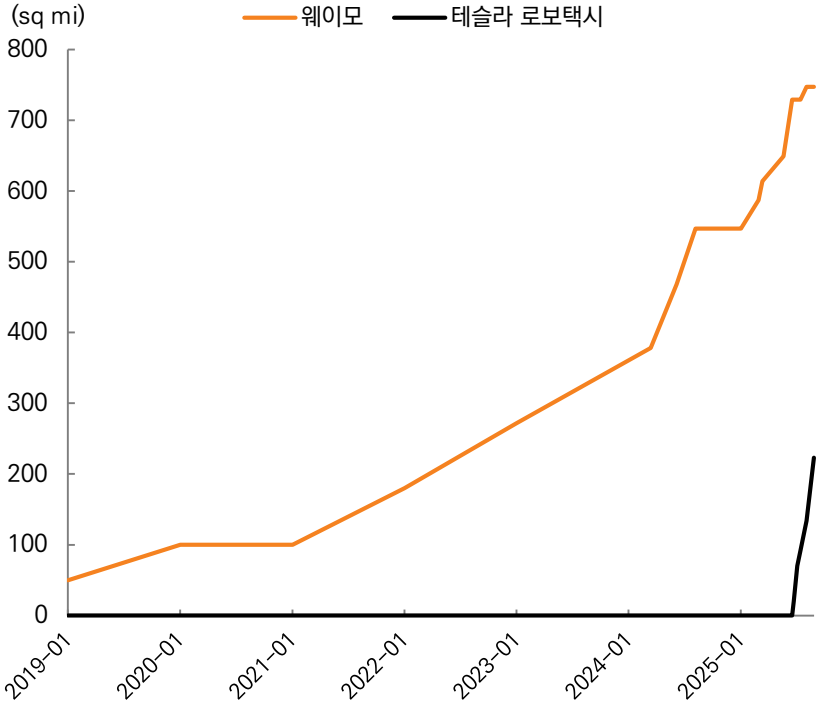
웨이모의 성장은 점진적이고 선형적인 반면, 테슬라의 성장은 수년간의 침묵 후 수직에 가까운 기하급수적 폭발

텍사스 오스틴에서의 자율주행 서비스 면적 비교
테슬라가 웨이모보다 약 1.9배 더 넓은 면적을 갖게 됨



자료: 미래에셋증권 리서치센터

미국 전역에서의 자율주행 서비스 면적 비교
웨이모가 테슬라보다 아직은 약 3.3배 더 넓은 면적을 보유



자료: 미래에셋증권 리서치센터

한편, **테슬라는 그 어느때보다 사이버캡(Cybercab) 출시가 절실**. 대량생산과 상용화를 가로막던 가장 결정적 장애물을 제거하는 조치가 발표됐기 때문. 미국 도로교통안전국(NHTSA)이 운전대나 페달이 없는 자율주행차의 **승인 절차를 수년에서 수개월로 대폭 단축하는 새로운 프레임워크**를 발표했음. 이는 규제 당국이 자율주행 시대를 공식적으로 인정하고 수용하겠다는 패러다임의 전환을 의미.

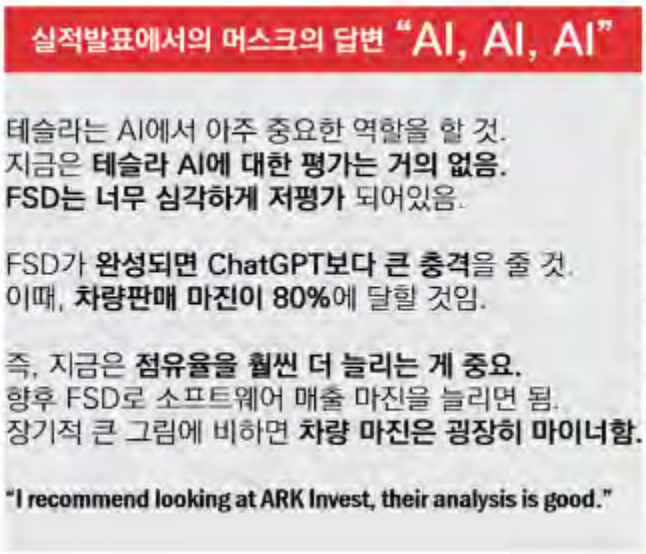
IX. 테슬라: 로보택시 비즈니스는 시장 규모가 얼마나 될까

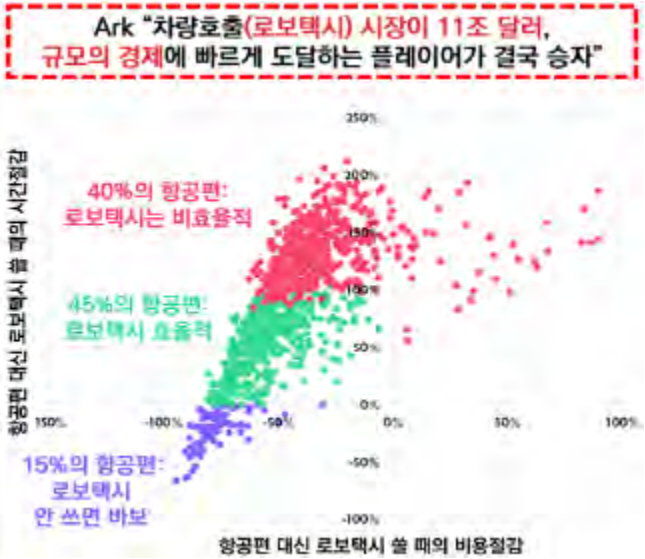
잠재 가치(TAM): 10조 달러

로보택시는 테슬라의 비즈니스 모델을 고마진 SaaS로 재편할 가능성

- 테슬라의 가치를 평가하는 척도는 더 이상 차량 판매 대수가 아니라는 것을 점차 시장이 인지하고 있음. 전기차 판매 부진 및 온갖 정치적 노이즈에도 부담스러운 밸류에이션이 유지되는 이유. 다시 말해, 투자자들은 테슬라를 단순 제조업 밸류에이션으로 평가하는 것에서 벗어나, 에이전트/물리적 AI 플랫폼 기업으로의 본질적 변화를 점차 받아들이고 있는 것으로 판단.
- 일론 머스크는 로보택시가 창출할 생태계의 연간 매출 규모를 향후 5~10년 내 8조~10조 달러로 추정하는 Ark의 분석에 동의. 이는 현재 전 세계 GDP(약 113조 달러)의 약 10%에 해당하는 규모. 물론 이 숫자는 테슬라의 직접 매출이 아닌, 플랫폼 수수료, 개인 차주의 운영 수익, 차량 내 엔터테인먼트, 광고, 보험, 물류 등 파생되는 모든 경제 활동의 총합.
- 현재 미국 내 차량 공유 서비스는 전체 자동차 주행거리(VMT)의 3%에 불과한 틈새시장. 그러나, 로보택시 요금은 추후 "우버 대비 최대 90%"에 책정될 전망으로, VMT 점유율을 크게 잠식할 가능성.

물론 10조 달러의 TAM 측정은 기술적 실현, 규제 승인, 사회적 수용이라는 공격적 가정에 기반한다는 것을 유념
2년 전에는, 일론 머스크의 아래와 같은 "자동차가 아니라 문제는 AI야"라는 주장은 주목받지 못했지만, 지금은?





자료: 테슬라, Ark Invest, 미래에셋증권 리서치센터

IX. 테슬라: 로보택시의 문제는 기술이 아니라 사회 수용도

로보택시의 미래는 기술적 완성이 아닌, 오스틴의 성공에 달림

- 테슬라 로보택시의 가장 큰 리스크는 기술이 아닌 사회적 수용성. 로보택시는 기술 산업이 아닌 서비스업이며, 성공은 기술이 아닌 '사회가 대규모 도입을 허락하는가'에 달림. 사우디와 UAE, 중국이 FSD 도입을 환영하고 한국도 준비 중인 반면, 유럽으로의 확장은 계속 뒤쳐지는 것이 이러한 이유.
- 테슬라는 현재 기술 성숙도를 제외한 모든 영역에서 낙제점이라는 평가가 많음. "수익성 모델, 기존 운송업계와의 관계, 사고 발생 시 여론 대응, 현지화된 고객 서비스 조직 등" 사회 시스템 공학의 거의 모든 차원에서 준비가 부재한 상황이라는 것. 고객센터 연결조차 어려운 테슬라가 복잡한 서비스 문제를 해결할 역량이 있는가에 대한 질문인 것. 기술만 증명하면 돈이 따라온다는 상업 만능주의는 경계해야 할 요소. 단 한 번의 사고가 프로젝트 전체를 위협할 수 있는 시스템적 신뢰의 위기에 직면할 가능성도 물론 있음.
- 이러한 벽을 넘기 위해 중요한 변수는 결국 '수익성 증명'. 테슬라는 오스틴이라는 통제된 시험대에서 "수입 〉 비용" 공식을 증명하는 데 모든 것을 걸고 있음. Airbnb의 초기 저항이 그랬듯, 파괴적 수익/효율성이 만들어내는 경제적 이익 앞에서는 기존의 저항과 규제가 무력화될 수 있기 때문.

테슬라 로보택시 사업이 단순 기술 프로젝트가 아닌 풀기 힘든 사회 공학인 이유

차원	해결해야 할 문제	테슬라의 현재 상태
기술 성숙	자율주행이 안전하고, 신뢰할 수 있고, 설명 가능한가?	(☑) 기술 선도: FSD의 성능은 우수하다.
수익성	단일 차량의 수입이 비용을 넘어서는가?	(✕) 시뮬레이션 모델은 있지만, 실제 증거 아직 없다.
업계 관계	기존 택시/플랫폼 운전자들의 사회적 저항은?	(✕) 아직 협상 메커니즘 없다. 격렬 충돌이 가능성…
여론 관리	사고가 한 번 났을 때, 사회 전체가 받아들일 수 있는가?	(✕) 사고 논란이 많았고, 브랜드 양극화가 심하다.
운영 조직	현지 배치, 고객 서비스, 안전, 보상 등 서비스 체계는?	(✕) 개인간 거래 서비스 시스템 구축 경험이 부족하다.

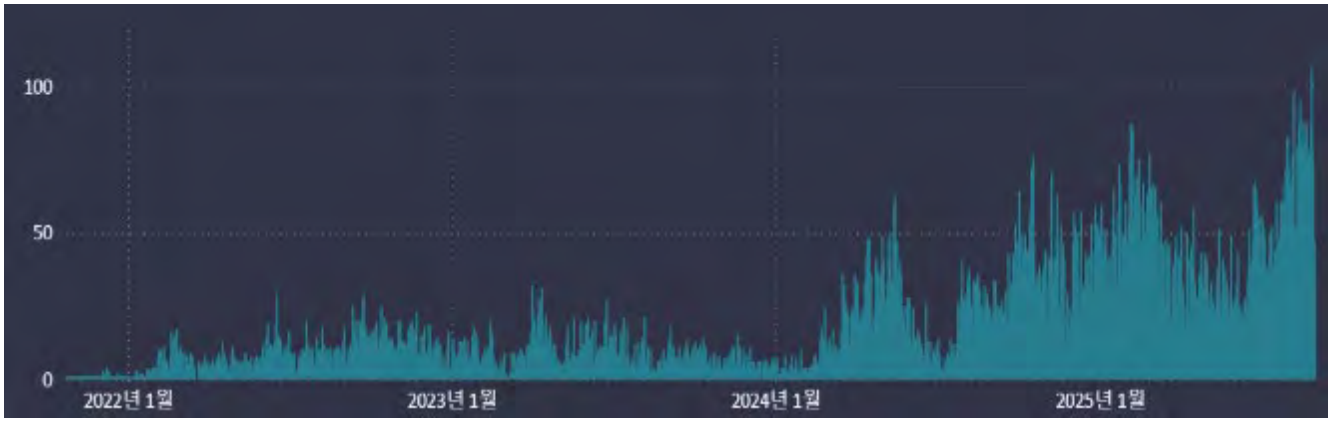
자료: 미래에셋증권 리서치센터

[참고] 로보택시의 핵심인 자율주행 기술 FSD 채택률

FSD의 사회 수용도는 이제 티핑 포인트에 가까워지고 있다

- 2025년 9월 1일, 테슬라의 차량 엔지니어링 부사장인 Lars Moravy는 FSD와 관련한 중요한 데이터를 언급. 테슬라 차량 라인업 전반에 걸쳐 Full Self-Driving (FSD) 소프트웨어 채택률(유료 일회성 구매 및 구독)이, 이전에 "한 자릿수"에 머물렀던 전체 채택률이 이제 10대 퍼센트로 상승했다고 언급. 작년 2분기 기준 4~8%로 추정(테슬라는 정확한 수치를 공개하지 않음)됐던 것에 비하면 비약적 상승. 특히 Model S와 Model X 같은 프리미엄 모델의 경우 채택률이 50-60% 범위에 도달했다고 하며, 이는 소비자 채택의 증가를 반영. 이 급증은 FSD 버전 12 이상의 개선에 기인.
- 2025년 7월 FSD 구독료 월 99달러로 인하 이후 구독률이 25% 증가하고 북미 전체 FSD 채택이 45% 상승한 바 있음. 가격 인하 요인 말고도 FSD의 새로운 버전 출시로 "주행 지능"이 업그레이드 될 때마다 채택률(구독 등)이 늘어나는 트렌드도 있음. 올해 가을 FSD v14의 출시가 더욱 기대되는 이유.
- FSD 채택률이 10대 퍼센트라면, 이는 주류 기술로의 편입 가능성을 시사하는 중요 지표로 볼 수 있음. Everett Rogers의 유명한 "혁신의 확산" 이론에 따르면, 대략 16%에 S-커브를 만드는 티핑 포인트가 발생. 이후 얼리 어답터의 전유율이 아닌 주류 사용자들의 대거 유입으로 가파른 상승이 시작.

사용자들이 제출한 데이터 Entries(FSD 테스터들이 제출한 라이브 세션 보고서)의 수를 보여주는 선 그래프
기존 FSD의 새 버전 출시(예: v13.2)가 테스터들의 관심을 높여 제출 활동을 증가시켰을 것
Entries의 증가는 또한 신규 테스터(새로운 FSD 사용자)가 커뮤니티에 참여해 데이터를 제출하는 경우를 반영



자료: FSD Community Tracker(테슬라 FSD 테스터들이 자율주행 시스템의 성능 데이터를 제출하고 공유하는 커뮤니티 기반 플랫폼), 미래에셋증권 리서치센터

IX. 테슬라: 그럼에도, 로보택시의 확장성은 수익 공유에 달렸다

수익을 창출하는 자산으로 변모할 로보택시

- 수익성은 차량 공유 사업의 고질적인 문제. 웨이모가 적자 구조 속에서 Uber와 협력하는 것과 달리, 테슬라는 '차주 참여' 모델을 통해 비용 증가 없는 확장의 해자를 지니게 될 것. 로보택시와 같이, 테슬라의 자율주행 기술인 "FSD"가 Supervised(Level 2 수준으로, 운전자의 감독이 필수)에서 Unsupervised(Level 3로서 감독 필요 없음)로 변화하는 순간, 자율주행의 Airbnb가 되는 것.
  - 피크 타임에 로보택시 요금을 인상하면, 금전적 인센티브에 반응한 수백만 개인 테슬라 차주들이 자신의 차량을 네트워크에 자발적으로 투입. 이는 외부 자본 투자 없이 공급을 탄력적으로 빠르게 확장하는 방식
- 이러한 전환은, 테슬라의 자동차에 대한 수요 뿐 아니라 중고차 시세 역시 지대한 영향을 끼칠 것으로 예상. 사용자가 수면 혹은 근무 중일 때, 주차되어 있는 차가 택시처럼 스스로 움직여 승객을 태워 소유자에게 수익을 가져다 줄 수 있기 때문. 즉, 자동차라는 하나의 자산이 감가상각에 의해 점차 그 가치의 훼손이 일어나는 자산에서 부가가치를 창출하는 수익자산으로의 변모가 일어난다는 것.
- 자동차 한대가 벌어들일 잠재적인 FCF에 대한 순현재가치가 반영되어야 하다 보니 중고차 시세에도 영향을 미칠 수밖에 없으며, 이것이 현실화 된다면 신차의 수요에도 영향을 주는 일종의 "FOMO" 현상으로 진화할 가능성 역시 배제 할 수 없음.

로보택시로 벌어들일 수 있는 현금흐름의 간단한 추산을 해보면, 5년간 약 10만달러의 순이익을 벌어들일 수도 있음

항목	값	설명
차량 가격 (USD)	35,000	테슬라 차량 1대를 구매할 때의 총비용
평균 승차 요금 (USD)	8	한 번 탑승할 때 승객이 지불하는 금액(수수료, 세금 포함 전)으로 가정
일일 승차 횟수	20	24시간 운영·확장성을 고려한 보수적 수치
연간 운행 일수	312	일주일(주)에 하루는 각종 운영 사항 때문에 쉰다고 가정
연간 매출 (USD)	50,000	8달러 × 20회 × 312일 = 49,920달러 → 반올림해 50,000로 가정
운행비 비율	30%	충전(전기)·보험·정비·데이터 비용을 매출의 20–40% 사이, 중간값 30%로 설정
연간 운행비 (USD)	15,000	50,000 × 0.30 = 15,000달러가 차량 1대의 연간 고정·변동비 총합 (차량 수명 5년이라고 보수적 가정)
첫해 순이익	–	결과적으로 1년 동안 벌어들인 40,000달러(50,000–15,000)가 차량값 35,000달러를 압도
2년차 이후 연간 순이익 (USD)	**25,000**	**2년차부터는 25,000 달러가 매년 순현금으로 누적**

자료: 미래에셋증권 리서치센터

[참고] 로보택시의 파괴적인 경제성: Valuation (1)

일론 머스크의 주장(로보택시가 된다면 차량 한 대당 가치는 5배가 된다)이 그리 터무니없는 수치가 아닌 이유

항목	단위	값	비고 및 설명
I. 입력			
A. 차량 기본 정보			
1. 차량 가격 (예: 사이버캡)	$	29,000	로보택시 가격은 25,000달러가 목표지만 보수적 가정
2. 차량 수명	년	5	차량의 경제적 내용연수
3. 차량 수명 (총 주행 가능 거리)	마일	500,000	차량이 폐차될 때까지 주행 가능한 총 마일
4. 일일 운영 시간	시간	16	충전 및 정비, 청소, 이동 시간 등을 제외
B. 승객 운송			
5. 일일 차량 가동률 (승객)	%	30%	
6. 평균 이동 속도 (승객)	mph	20	로보택시 평균 운행 속도
7. 1회 평균 승객 탑승 거리	마일	8	1회 탑승 시 평균 이동 거리
8. 1회 평균 승객 승하차 시간	분	3	승차 및 하차에 소요되는 평균 시간 (왕복 기준)
9. 승객 운송 요금 (마일당)	$/마일	1.00	
10. 차량 운영 비용 (마일당)	$/마일	0.20	보험, 청소, 타이어 등 유지보수 비용
C. 배달 서비스 가정			
11. 유휴 시간 중 배달 서비스 전환율	%	20%	승객 운송 외 유휴 시간 중 배달 서비스에 투입되는 비율 (100%는 모든 유휴시간 활용)
12. 평균 이동 속도 (배달)	mph	20	배달 시 평균 운행 속도 (도심 단거리 위주 가정)
13. 1회 평균 배달 거리	마일	3	1회 배달 시 평균 이동 거리 (음식/소형 패키지 가정)
14. 1회 평균 배달 소요 시간 (상하차)	분	5	사람(또는 옵티머스 로봇)이 배달물을 싣고 내리는 데 걸리는 시간
15. 배달 요금 (마일당)	$/마일	0.50	건당 요금으로 변경 가능 (DoorDash 요금은 마일당 $1, 여기에 50% 할인)
16. 배달 운영 추가 비용	$/마일	0.02	일종의 오버헤드 비용 (로봇 감가상각비, 배달 장비 개조 비용, 청소 위생 비용, 분실 등에 따른 자체 처리비용)
D. 부가 수익 (차량당 연간)			
17. 광고 수익	$	1,000	차량 내/외부 광고로 인한 연간 수익 추정치
18. 맞춤형 경험 수익	$	100	유료 테마 투어, 가이드 서비스 등으로 인한 연간 수익 추정치
19. 엔터테인먼트 & 연결성 수익	$	1,200	프리미엄 인카 엔터테인먼트, 고속 Wi-Fi, 차량내 상거래(월 100달러 가정)
20. 분산 추론 수익	$	315	차량 유휴 컴퓨팅 파워 활용으로 인한 수익 (엔비디아 RTX4070의 경우 시간당 0.12 달러, 여기에 70% 할인)
E. 주요 재무 가정			
21. 할인율 (WACC)	%	12%	미래현금흐름 할인을 위한 자본비용
22. 영구 성장률 (Terminal Value)	%	3%	DCF 모델의 Terminal Value 계산 시 사용
23. 법인세율	%	21%	

자료: 미래에셋증권 리서치센터

[참고] 로보택시의 파괴적인 경제성: Valuation (2)

항목	단위	값	비고 및 설명
II. 계산			
A. 일일 차량 운영 분석 (시간)			
24. 승객 운송 시간	시간	4.8	항목 4*항목 5
25. 승객 유휴 시간	시간	11.2	항목 4-항목 24
26. 배달 서비스 시간	시간	2.2	항목 25*항목 11
27. 총 차량 가동 시간 (승객+배달)	시간	7.0	항목 24+항목 26
28. 총 차량 유휴 시간 (잔여)	시간	9.0	항목 4-항목 27
B. 일일 운행 분석 (횟수/거리)			
29. 1회 승객 운송 소요 시간 (이동)	시간	0.4	항목 7/항목 6
30. 1회 승객 총 소요 시간	시간	0.45	항목 29+(항목 8/60)
31. 일일 승객 운행 횟수	회	10.7	항목 24/항목 30
32. 일일 승객 주행 거리	마일	85	항목 31*항목 7
33. 1회 배달 소요 시간 (이동)	시간	0.15	항목 13/항목 12
34. 1회 배달 총 소요 시간	시간	0.23	항목 33+(항목 14/60)
35. 일일 배달 운행 횟수	회	9.6	항목 26/항목 34
36. 일일 배달 주행 거리	마일	29	항목 35*항목 13
37. 일일 총 주행 거리	마일	114	항목 32+항목 36
C. 일일/연간 수익 및 비용 분석			
38. 일일 승객 운송 수익	$	85.3	항목 32*항목 9
39. 일일 배달 수익	$	14.4	항목 36*항목 15
40. 일일 총 운행 수익	$	99.7	항목 38+항목 39
41. 일일 부가 수익 합계	$	7.2	(항목 17+항목 18+항목 19+항목 20)/365
42. 일일 총수익	$	106.9	항목 40+항목 41
43. 일일 차량 운영 비용 (총 주행거리)	$	22.8	항목 37*항목 10
44. 일일 배달 운영 추가비용 (배달거리)	$	0.4	항목 36*항목 16
45. 일일 총 운영 비용	$	23.3	항목 43+항목 44
46. 일일 총이익 (Gross Profit)	$	83.6	항목 42-항목 45
47. 연간 총수익	$	39,018	항목 42*365
48. 연간 총 운영 비용	$	8,489	항목 45*365
49. 감가상각비 (차량 가격/ 연 수명)	$	5,800	항목 1/항목 2,0
50. 연간 세전 순이익 (EBT)	$	24,728	항목 47-항목 48-항목 49
51. 연간 법인세	$	5,193	MAX(0, 항목 50*항목 23)
52. 연간 세후 순이익 (Net Income)	$	19,535	항목 50-항목 51

항목	단위	값	비고 및 설명
III. 결과			
A. 마일당/시간당 지표			
53. 마일당 수익 (배달포함 평균)	$/마일	0.87	항목 40/항목 37
54. 마일당 비용 (배달포함 평균)	$/마일	0.20	항목 45/항목 37
55. 마일당 총이익 (배달포함 평균)	$/마일	0.67	항목 53-항목 54
56. 시간당 총수익 (총 가동시간 기준)	$/시간	14.2	항목 40/항목 27
57. 시간당 총이익 (총 가동시간 기준)	$/시간	10.9	(항목 40-항목 45)/항목 27
B. 연간 차량당 지표			
58. 차량당 연간 총수익 (부가수익포함)	$	39,018	항목 47
59. 차량당 연간 순이익 (Net Income)	$	19,535	항목 52
60. 순이익률 (Net Margin)	%	50%	항목 52/항목 47
C. DCF 가치평가 (차량당)			**향후 10년간 현금흐름 및 영구 가치 가정**
61. 연간 세후 순이익 기본 가정	$	19,535	단순화 위해 FCF를 세후 순이익으로 가정. 실제로는 운전자본변동, CAPEX 고려 필요
62. FCF 현재가치 (1년차)	$	17,442	5년차까지 동일한 FCF를 가정 (성장률 0%이라는 매우 보수적 가정). 실제로는 연차별 FCF 성장률을 적용 필요
63. FCF 현재가치 (2년차)	$	15,573	
64. FCF 현재가치 (3년차)	$	13,905	
65. FCF 현재가치 (4년차)	$	12,415	
66. FCF 현재가치 (5년차)	$	11,085	
72. Terminal Value (5년차 말)	$	126,860	(5년차 FCF * (1+영구성장률)) / (할인율-영구성장률)
73. Terminal Value 현재가치	$	71,984	항목 72/((1+항목 21)^5)
74. FCF 현재가치 합계 (1~5년)	$	70,420	1년차부터 5년차까지의 FCF의 NPV
75. 차량당 생애주기 사업 가치	**$**	**142,404**	**차량 1대당 최종 추정 가치**
75. 차량당 가치 승수	**배**	**4.91**	**일론 머스크는 "로보택시가 되면 차량 한 대당 가치가 현재보다 5배 이상 상승할 수 있다"고 언급한 바 있음**

자료: 미래에셋증권 리서치센터

IX. 테슬라: 글로벌 확장을 위해, 고지능을 향한 전력 투구

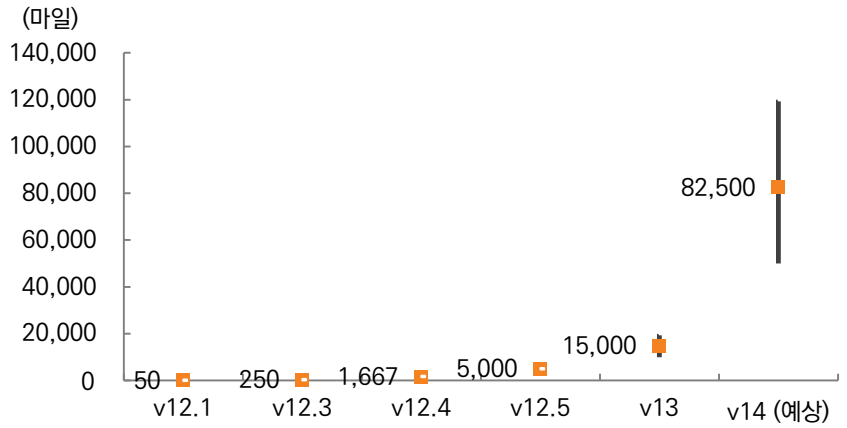
완벽한 자율주행을 위해 더 큰 모델이 필요하다

- 로보택시 및 FSD의 능력이 놀랍긴 해도, 대규모로 서비스를 확장할 준비가 100% 된 것은 아님. 현재 FSD의 버전은, 범용 주행 지능을 담기에는 모델의 크기가 작다는 한계가 존재한다는 평가. 파라미터가 확대돼야, 복잡한 시나리오의 대응 능력이 강화돼, 사용자 경험 개선 및 로보택시의 가치가 증대될 것.
- 아직은 "작은 모델"이기 때문에, 테슬라 차량들에서 보이는 FSD(supervised)의 일부 문제점들이 로보택시에서도 유사하게 나타남. 99.99%의 신뢰도와 99.9999%의 신뢰도는 인간의 안전을 기준으로, 완전히 다른 차원의 것. 이를 위해 지능이 계속해서 향상돼야 필요가 있음.
- 그럼에도 비교적 오스틴에서 사고 없이 잘 운영되는 이유는 오스틴이라는 특정 도로 상황에 시스템을 미세조정한 결과일 수 있음. 바꿔 말하면, 오스틴에서 성공했다고 해서 이 시스템을 그대로 뉴욕이나 서울에 가져다 놓으면 곧바로 똑같이 잘 작동할 수 있을 것이라고 생각하기는 무리가 있음.
- 일론 머스크 발언과 이 생각을 종합하면, 더 많은 도시로 확장하려면 모델 크기를 몇 배 이상 더 키워야 하고, 이를 받쳐줄 더 강력한 컴퓨터가 필요. 테슬라가 차세대 칩 개발에 진심인 것도 이 때문.

2025년 10월 대중에게 출시 예정인 FSD V14에 관한 머스크의 트윗
"V14는 10배의 파라미터로 인간보다 2~3배, V15는 10배 안전할 것"

자료: 일론 머스크, 미래에셋증권 리서치센터

테슬라 FSD의 버전별 MPCI(중요 개입당 주행거리) 성능
MPCI: 자율주행 중 사람이 꼭 운전대를 잡아야 하는 순간들(사고·위험 회피 등) 사이에 차가 평균 몇 마일을 스스로 달렸는지… 값 클 수록 덜 개입/高성능



자료: 미래에셋증권 리서치센터

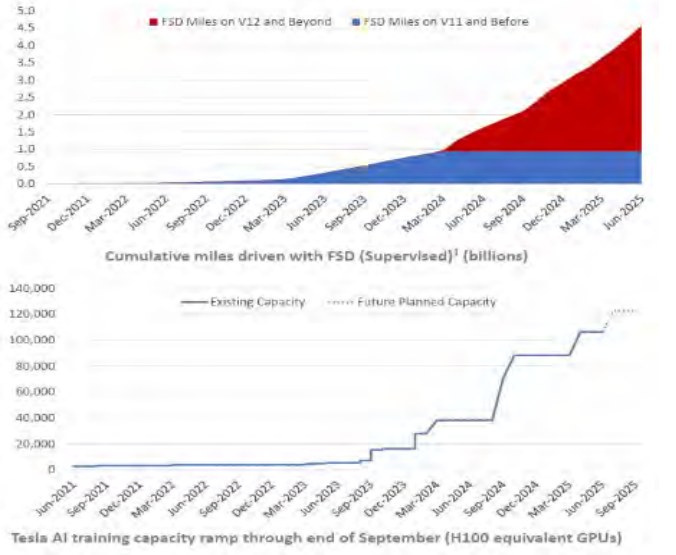
[참고] 모델 크기가 왜 커야 되는가?

- 로보택시가 폭발적 확장성을 현실화하기 위해서는 더 똑똑하고 범용적인 AI 모델의 완성이 필수적.
- 파라미터가 많다는 것은 AI 모델의 '정신적인 공간' 또는 '학습 능력의 총량'이 크다는 것을 뜻함.
- 가령, “비가 오면 땅이 젖는다” 수준의 단순한 관계를 이해하는 정도가 작은 모델의 성능이라면, “여름철 북태평양 고기압의 확장으로 인한 대기 불안정이 국지성 호우를 유발하고, 이때 아스팔트 노면의 마찰력이 감소하므로 제동거리를 더 길게 잡아야 한다”는 식의 복잡하고 다층적인 인과관계를 이해하는 것이 바로 큰 모델.
- 이처럼 세상을 더 깊게, 더 넓게 이해할 학습 잠재력을 갖춰야만 예측 불가의 수많은 변수로 가득한 실 도로 환경에 완벽하게 대응할 수 있음.
- 물리 세계를 이해해야 하는 테슬라의 자율주행 AI 역시, 더 큰 모델을 만들기 위해 더 많은 칩과 전력이 필요하다는 말.
- OpenAI나 구글 같은 AI 기업들이 겪는 것과 정확히 동일한 성장통. AI의 지능은 컴퓨팅 파워에 의해 결정된다는 대원칙을 공유하는 셈.
- 일론 머스크는 차세대 칩인 “AI5” 칩이 “최대 2,500억 개”의 파라미터를 가진 매우 큰 모델을 처리하는 칩이라고 강조

테슬라는 FSD를 훈련하기 위한 2번째 클러스터인 Cortex 2.0을 대규모로 건설 중
Cortex 1.0은 H100 GPU 6.7만개, Cortex 2.0은 최대 15만 개의 H100과 동등 용량으로 추산

자료: 테슬라, X(@JoeTegtmeyer), 미래에셋증권 리서치센터

테슬라 공식 실적 발표 문서에서 공개하는 AI 지표들
9월 말까지 50억 마일의 FSD 주행 데이터 확보하고,
9월 말까지 전사적으로 12만개의 GPU 보유 전망



자료: 테슬라, 미래에셋증권 리서치센터

[참고] 커지는 모델 크기에 맞추어, 필수적인 것은 결국 반도체

AI5 출시와 맞물려, 2026년이 진정한 로보택시의 주행 지능이 완성될 가능성

테슬라의 AI 추론(inference)용 자체 제작 반도체 스펙 히스토리

버전	출시 시기	칩 이름	제조사	제조 공정	신경망 성능 (TOPS)	전력 소모	추가 사양
HW1	2014년 9월	EyeQ3	모빌아이	40 nm	미상	2.5W	• TDP 2.5 W, 클럭 속도 500 MHz
HW2	2016년 10월	DRIVE PX 2	엔비디아	16 nm	24	300W	• 8 TFLOPS FP32 • 두 개의 Tegra X2 SoC • 두 개의 Pascal GPU
HW3	2019년 4월	Tesla FSD 1 Chip	삼성전자	14 nm	72	100W	• 두 개의 칩 구성 • 칩 면적은 260 mm^2 • 트랜지스터 수는 60억 개 이상 • 2,300 fps 이미지 처리
HW4	2023년 1월	Tesla FSD 2 Chip	TSMC 삼성전자	7 nm	500	160W	• 두 개의 칩 구성 • 20개의 CPU 코어, 최대 클럭 2.35 GHz • 3개의 NPU, 최대 클럭 2.2 GHz
AI5	**2026년 하반기**	**Tesla AI5 Chip**	**TSMC**	**TSMC 3nm (N3P), 애리조나 공장도 활용**	**2,000~2,500**	**500~800W (추정)**	**• 두 개의 칩(추정), 칩렛 구조 채택 가능성 • 사이버캡에 최적화될 가능성 • 48V 아키텍처 지원 • Grok AI 비서 기능 • 차량 전장·냉각을 재설계 필요 • 구형 차량엔 '역호환 불가'**
AI6	**2027년**	**Tesla AI6 Chip**	**삼성전자**	**삼성전자 2nm (SF2), 텍사스 Taylor 신공장은 AI6 생산 전용**	**10,000 이상**	**미공개**	**• 자율주행·휴머노이드용 AI 추론뿐 아니라 데이터센터용 AI 트레이닝까지 통합되어, 워크로드를 감당하는 토탈 솔루션 • Dojo D1 타일을 소형화한 형태로 추정**

자료: 테슬라, 미래에셋증권 리서치센터

IX. 테슬라: 테슬라가 포기한 것에서 알 수 있는 인사이트 (1)

‘Dojo 2’의 종말과 ‘AI6’의 부상

- 일론 머스크는 자신들의 자율주행 훈련용 전용 AI 가속기인 Dojo 2 아키텍처가 ‘막다른 길’이었음을 인정하고 프로젝트 종료를 선언. 몇 년 동안 운영해왔던 팀을 급해체한 것. 이로써 거대한 단일 웨이퍼 스케일이라는 혁신적 접근법을 포기. 대신 표준화된 SoC(System-on-Chip) 여러 개를 한 보드에 꽂아 쓰는 형태인 ‘AI6’ 플랫폼으로의 전면 전환을 발표.
- 이는 실패 인정으로 보이기도 하고 엔비디아 GPU의 거대한 승리로 읽히기도 함. Dojo의 포기는 세 가지 현실적 문제에 직면했기 때문. 첫째, CUDA, NVLink 등 생태계 진화 속도를 테슬라 단독으로 따라잡는 것은 불가능에 가까웠음. 둘째, Dojo 2 방식의 커스텀 타일과 고난도 패키징은 수율, 조달, 납기 등 공급망 리스크가 이전보다 더 커졌다고 판단한 듯 보임.
- 마지막으로, 단일 AI6 아키텍처로 차량, 로봇, 데이터센터를 통합하면 조직과 리소스 재사용에 훨씬 유리. 속도를 위해 차별화의 폭을 줄이는 선택. 물론 전략 선회는 대규모 트레이닝에 대한 엔비디아 의존도를 단기적으로 높일 수 있음. Dojo에 투입됐던 CAPEX가 엔비디아 칩 구매로 재배치됨을 의미.

지난 7월만 하더라도 연말까지 대량 생산에 들어간다고 예정된 Dojo 2
TSMC의 고난도 패키징 기술인 Info-SoW이 접목되는 형태 = 공급 리스크

자료: 테슬라, 미래에셋증권 리서치센터

Dojo 프로젝트 종료로 인한 테슬라의 향후 자율주행 개발시 시나리오
현재로서는 훈련에도 AI6를 쓸 것으로 보이나 GPU와 혼용할 듯

시나리오	설명	추정 확률
A. 혼합 전략	**사전학습은 엔비디아 칩, 사후학습은 AI6**	**≈75%**
B. 내재화 확대	AI6 패브릭을 계속 강화해 트레이닝 용도로 비중 ↑	≈20%
C. 완전 철회	AI6는 단지 차량 내 추론만 (훈련은 외부 가속기로만)	≈5%

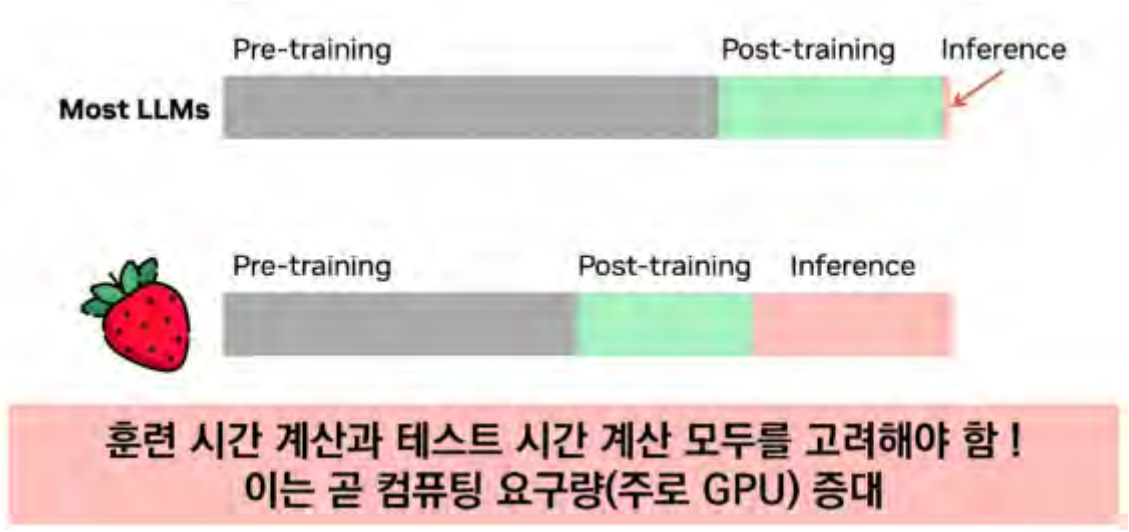
자료: 미래에셋증권 리서치센터

IX. 테슬라: 테슬라가 포기한 것에서 알 수 있는 인사이트 (2)

Dojo 프로젝트 변경은 AI 업계 패러다임 변화에 맞춘 것이다

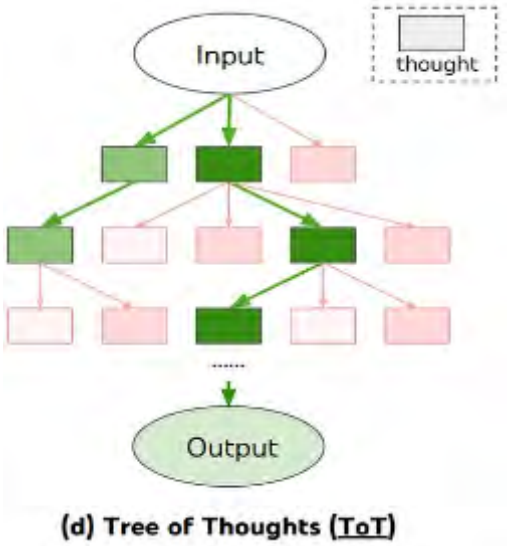
- 현재 AI 업계는 단지 모델의 크기를 키우는 게 아니라 "강화학습 스케일링 혹은 Test time scaling"으로 패러다임이 전환된 상태. 즉, 훈련용 FLOPs 경쟁이 아니라, 추론 단계에서 최대한 많은 롤아웃(답변 내기)을 돌려, 보상으로 모델을 다듬는 방식으로 성능을 끌어올리는 방식이 승부처가 된 것.
  - 강화학습 스케일링은, 정답을 찾기 위해 수많은 시도('롤아웃')를 생성해야 하는 극도로 추론(inference) 집약적인 과정
- 테슬라는 이에 맞춰 Dojo(훈련 특화) 비중을 줄이고 AI6(추론·파인튜닝·합성데이터) 로 통합한 것으로 판단함. AI6로 "차량·로봇·데이터센터"용 칩을 한 플랫폼으로 묶는다면 런타임/컴파일러/커널 등의 AI 툴체인을 재사용할 수 있어 (토큰당 비용↓·지연↓·배치 효율↑) 반복학습의 속도가 빨라 지는 효과.
  - 런타임: 자원 사용을 실시간 관리하는 현장 매니저 / 컴파일러: 특정 하드웨어에 맞게 코드 최적화 해주는 일종의 번역기
- 자연스럽게도, 사전 학습에서는 엔비디아 GPU를 써야 하기 때문에 리스크는 엔비디아 의존 증가가 있음. AI6 프로젝트가 성공하면 FSD 빌드 속도가 가속되고 로보택시 업데이트도 더 빨라질 수 있음.

OpenAI가 2024년 9월에 내놓은 모델인 o1(코드명: 딸기)으로 들고 온 새로운 패러다임
AI 훈련이 사전학습 위주에서 사후학습 및 추론 영역의 비중이 크게 늘어나게 된다



자료: Jim Fan, 미래에셋증권 리서치센터

이 스케일링은 왜 추론/롤아웃 집약적인가?
무작위 반복 시뮬레이션을 통해 해결책 탐색



자료: Noam Brown, 미래에셋증권 리서치센터

IX. 테슬라: 삼성과의 협력이 의미하는 것 (1)

테슬라의 진짜 승부수를 삼성전자가 만든다

- 로보택시의 지능 그 자체인 FSD와 테슬라 로보틱스인 "옵티머스"의 최종 성공은 결국 "AI 칩"이라는 "두뇌"의 성능에 달려있음. Dojo가 끝을 맺으면서 이제 테슬라 AI 칩 로드맵의 정점은 AI6. 테슬라는 이 핵심 자원을 확보하기 위해 전략적 동맹으로 삼성전자의 파운드리를 선택(165억 달러 규모 계약).
- 특이할 만한 점은, 삼성의 신규 텍사스 공장이 오직 AI6만을 위한 '전용(Dedicated)' 팹으로 운영된다는 점. 테슬라의 로드맵에 맞춰 공장 전체가 움직이는, 깊은 수준의 통합. 또한, 테슬라의 로드맵에 완벽히 동기화된 생산 기지를 테슬라 마당에 확보한다는 것도 중요. 기가팩토리와 머스크의 자택 인근이라는 지리적 이점인데, 문제 시 머스크 CEO가 직접 개입하겠다는 의지의 표명.
- 물론 리스크는 명확함. AI6 칩은 삼성의 GAA 2nm(SF2) 공정으로 2027년 양산 목표. 현재 SF2의 추정 수율은 40–45%. 이는 경쟁사인 TSMC N2(70% 이상)와 인텔 18A(50–55%)에 비해 크게 낮은 수치. 이 높은 기술적 허들과 수율 문제를 극복하기 위해 전례 없는 협력 방식이 요구됨.

삼성과 함께하는 테슬라 AI 칩 로드맵 전략과 최종 목표를 명시하는 트윗
텍사스 '전용' 팹 확보는 핵심 자원을 미국 본토로 가져오는 전략적 통합

현재의 주력 칩인 AI4(삼성이 제조)와 차세대 AI5(TSMC가 제조)는 단기적, 전술적 선택.

진짜 승부수는 FSD와 로보틱스 확장의 명운을 건 '결정적 칩' AI6.

자료: 일론 머스크, 미래에셋증권 리서치센터

2023년 5월 10일, 삼성전자 북미 반도체연구소에서 만난 두 사람
FSD를 위한 시스템반도체를 주요 안건으로 협력 방안 논의

자료: 삼성전자, 테슬라, 미래에셋증권 리서치센터

IX. 테슬라: 삼성과의 협력이 의미하는 것 (2)

삼성의 수율을 잡기 위해 테슬라 엔지니어가 직접 공정에 투입된다

- 해당 계약의 숨겨진 핵심은 "삼성측이 테슬라가 제조 효율성 극대화를 돕는 것을 허용"했다는 것. 이는 공정의 첫 단계부터 마지막까지의 데이터(수율, 결함 분포, 공정 변수 등)를 양사가 공유하고, 불필요한 과정을 없애고, 수율을 더 끌어올리기 위해 테슬라 엔지니어들이 삼성 공장 안으로 들어간다는 뜻.
- 즉, 칩 설계(테슬라)와 칩 제조(삼성) 사이의 벽을 허물겠다는 것이고, 이건 TSMC가 허용하기 힘든, 업계에서 전례가 없는 일. 테슬라가 삼성의 영업 비밀을 훔치겠다는 것이 아닌, 함께 문제를 풀어 최고의 결과물을 가장 빠르게 만들겠다는 지극히 공학적인 접근으로 풀이.
- 이 협력은 양측 모두에게 이득이 될 것. 테슬라는 '실전 파운드리 경험'을 저비용으로 확보. 제조 용이성 설계(DFM) 역량을 내재화하고, 향후 파운드리와의 협상에서 강력한 레버리지를 얻음. 삼성은 첨단 공정의 수율을 끌어올릴 기회이자, 초미세공정의 대형 고객 확보, 또한 TSMC에 대항할 '고객 참여형 제조 프로세스'라는 비즈니스 모델을 확보하는 베팅일 수 있음. (삼성은 해당 시설에 추가 28억 달러를 투자)

이번 거래의 중요성을 짚어주는 일론 머스크의 트윗
머스크 텍사스 자택과 삼성 테일러 공장의 거리는 자동차로 약 1시간

일론 머스크의 트윗은 이번 거래가, 단순한 위탁생산을 넘어 테슬라가 제조 효율 극대화를 위해 생산 라인에 직접 관여함을 공식화한 것.

머스크가 "직접 (제조 공정) 라인을 걷겠다. 그리고 그 공장은 편리하게도 우리 집에서 멀지 않다"고 말한 것은, "문제가 생기거나, 병목 현상이 보이거나, 더 나은 아이디어가 떠오르면 내가 바로 달려가서 해결하겠다"는 의지의 표명.

자료: 일론 머스크, 미래에셋증권 리서치센터

TSMC의 점유율은 70%를 초과… 삼성의 격차는 압도적
삼성에게 AI6 칩 양산 성공은 이 격차를 좁힐 돌파구임을 시사

순위	기업명	2Q25 매출	1Q25 매출	QoQ(%)	1Q25 점유율	1Q25 점유율
1	TSMC	30,239	25,517	18.5%	70.2%	67.6%
2	**삼성**	**3,159**	**2,893**	**9.2%**	**7.3%**	**7.7%**
3	SMIC	2,209	2,247	-1.7%	5.1%	6.0%
4	UMC	1,903	1,759	8.2%	4.4%	4.7%
5	Global Foundries	1,688	1,585	6.5%	3.9%	4.2%
6	Huahong Group	1,061	1,011	5.0%	2.5%	2.7%
7	VIS	379	363	4.3%	0.9%	1.0%
8	Tower	372	358	3.9%	0.9%	0.9%
9	Nexchip	363	353	2.9%	0.8%	0.9%
10	PSMC	345	327	5.4%	0.8%	0.9%
합계	**Total of Top 10**	**41,718**	**36,413**	**14.6%**	**97%**	**97%**

자료: TrendForce, 미래에셋증권 리서치센터

IX. 테슬라: ‘REAL RECOGNIZE REAL’

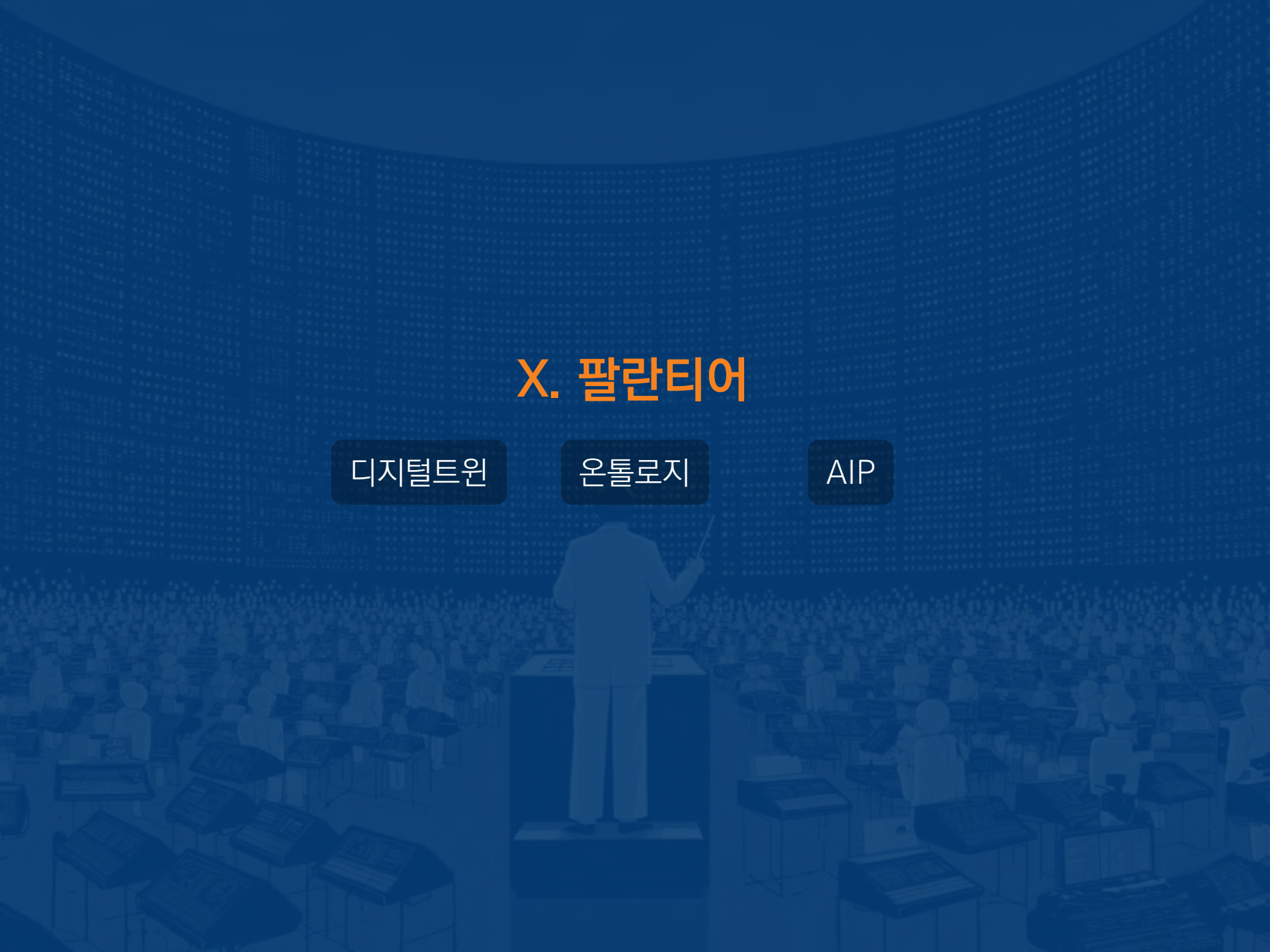
테슬라 + 삼성 (feat. 팔란티어)

- 이 전례 없는 협력을 기술적으로 구현하는 것이 바로 팔란티어의 '파운드리' 시스템라고 사료됨. 사실, 머스크의 제안을 수용하는 것은, 삼성 파운드리의 최고 기밀인 초미세공정 세부 데이터를 외부에 공개하는 리스크를 안고 있기 때문. 이를 위해 보안을 갖춘 운영체제 역할로 파운드리를 썼을 것.
- AI6칩을 생산할 공장 전체를 센서 데이터까지 포함해 가상 공간에 복제한 디지털 트윈을 구축한 뒤, 테슬라 엔지니어들에게 안전하게 통제된(가드레일이 있는) 가상 공장의 키를 넘겨 주면 됨. 이 디지털 트윈 안에서 테슬라는 칩 수율에 영향을 미치는 공정 변수들을 실시간으로 분석하고 시뮬레이션 가능.
  – 예시: 테슬라측이 설계 수정안을 시뮬레이션하면, 삼성 엔지니어는 실제 장비 데이터를 대조해 즉시 테스트 라인 돌림
- 여기서 포인트는 테슬라와 삼성의 엔지니어들이 같은 '디지털 트윈'을 보며 실시간으로 소통한다는 것이고, 삼성은 모든 것을 보여주지 않으면서도, 필요한 협력은 가능하게 만드는 것. 팔란티어의 투입은 양사의 비전이 희망이 아닌, 구체적인 실행 계획과 기술적 기반을 갖췄음을 증명하는 퍼즐 조각.

삼성전자 & 팔란티어 반도체 협력 분석

대분류	소분류	상세 내용
I. 파트너십 개요	계약 주체	삼성전자 DS(Device Solutions) 부문 & 팔란티어
	핵심 목표	AI 기반 데이터 분석을 통한 반도체 공정 최적화 및 수율, 품질, 생산성 향상
II. 전략적 배경	추진 동력	경쟁사(TSMC, SK하이닉스) 대비 첨단 파운드리 및 HBM 분야에서의 기술적 어려움 타개. 특히 수율 문제 해결이 시급.
	삼성의 과제	3나노 이하 첨단 공정에서 수율 저하 문제에 직면.
	팔란티어 선택 이유	**데이터 보안: 고객 데이터를 자사 서버에 저장하지 않는 팔란티어의 정책이 삼성의 핵심 기밀(공정 데이터) 유출 우려를 해소.**
III. 협력 범위 및 기술	적용 대상 공정	3나노 이하 파운드리 초미세 공정 / 2나노 GAA(Gate-All-Around) 노드 공정 / 10나노 6세대(1c) D램 공정
	연관 프로젝트	차세대 모바일 AP '엑시노스 2600' 칩셋 개발, S2 양산 라인 업그레이드
	활용 기술	• **파운드리(Foundry): 데이터 통합 및 상호작용 분석 플랫폼** • **AIP: 고객사 사설 네트워크 내 LLM 및 AI 구동 지원** • **온톨로지(Ontology): 복잡한 데이터를 의미 있는 구조로 엮어 분석을 용이하게 하는 핵심 기술**
	주요 마일스톤	• 2024년 말: 협력 형성 및 시스템 구현 완료 • 2025년 4분기: S2 양산 라인 업그레이드 완료 (목표) • 2025년 말: 2나노 GAA 노드 완성 및 엑시노스 2600 개발 완료 (목표)
V. 재무 및 계약 조건	계약 규모	구체적인 투자 금액, 계약 규모, 수율 향상 목표치 등은 비공개.
	현재 상태	도입 및 파일럿(Pilot) 단계로, 성과 검증 및 적용 범위 확대 중인 것으로 추정.

자료: 미래에셋증권 리서치센터

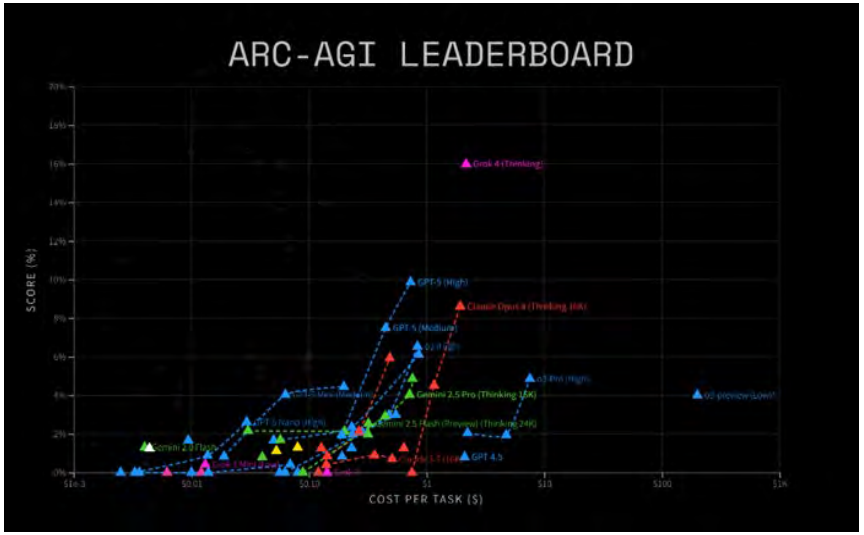

X. 팔란티어
디지털트윈
온톨로지
AIP

X. 팔란티어: AI 비용 감소의 최대 수혜는 AI를 제일 잘 활용하는 주체

AI의 토큰 비용 감소가 창출하는 AI 수요

- AI 기술의 발전으로 기반 모델의 운영 비용이 점차 저렴해지면서 '제본스의 역설'이 AI 산업의 핵심 원리로 작용. "더 저렴한 가격의 더 많은 지능은 더 많은 수요를 의미한다"는 명제로 이해할 수 있으며, 이 역설이 현실화될 경우 가장 큰 수혜는 저비용 고기능의 기반 모델 위에서 다양한 서비스를 제공하는 소프트웨어 기업들.
- AI를 기반으로 한 대표적인 SW 업체인 팔란티어는 저렴해진 'AI 모델'을 직접 만들지는 않음. 대신, 기업의 모든 데이터를 현실 세계의 실제 개념에 맞게 다 엮어 놓은 디지털 청사진을 구축. 그리고 전용의 AI 활용 플랫폼을 통해 외부의 AI 모델들(ChatGPT, Grok 등)을 그 청사진에 연결하여 활용함.
- AI 모델의 가격 하락은 팔란티어의 성장을 가속하는 핵심 동력으로 작용하는데, 기존 고객에게는 솔루션 활용도를 높일 강력한 유인을 제공하고, 잠재 고객에게는 AI 도입의 문턱을 크게 낮춰주기 때문. AI 기술의 발전에 따른 비용 모델 비용감소로 인해, 팔란티어 솔루션의 확산 속도는 더욱 가팔라질 가능성이 있다고 판단.

AI 모델의 운영 비용은 시간이 지남에 따라 지속적으로 더욱 감소
AI 모델의 성능(Y축)이 점점 더 저렴한 비용(X축)으로 제공되고 있음



자료: ARC – AGI, 미래에셋증권 리서치센터

AI에도 적용될 '제본스의 역설'
가격의 하락은 궁극으로 시장의 개화를 앞당기는 요소

구분	증기기관과 석탄	AI 모델과 팔란티어
기술 혁신	증기기관 효율이 높아져 석탄 사용량이 감소	AI 모델 성능이 향상되어 단위당 컴퓨팅 비용이 하락
수요 폭증	저렴해진 동력은 기차, 선박 등 새로운 산업으로 확산	저렴해진 AI는 모든 기업의 비즈니스 전반으로 확산
역설적 결과	단위 효율은 높아졌지만, 전체 석탄 총소비량은 폭증	단위 비용은 저렴 해졌지만, AI 활용 시스템의 총수요는 폭증

자료: 미래에셋증권 리서치센터

[참고] ChatGPT의 등장은 팔란티어 성장을 가속화한 영업사원

ChatGPT가 만든 시장의 변화를 팔란티어가 AIP와 부트캠프라는 전략으로 완벽히 흡수하여 성장의 기회로 전환시킨 선순환 구조

- 과거 팔란티어는 고객들에게 AI의 효용성을 일일이 설득해야 하는 장벽에 직면했으나, 이제는 " AI를 왜 써야 하는가 " 설명할 필요가 사라짐.
- 팔란티어는 이 기회를 활용해 자연어 인터페이스를 탑재한 "AI Platform"을 출시, 복잡했던 기존 제품의 활용법과 응용을 매우 간편하게 만들었음.
- 나아가 'AIP 부트캠프'라는 단기 교육 프로그램을 통해 고객들이 사용법을 쉽게 배우고 즉각적인 가치를 창출할 수 있도록 지원.
- 이로 인해 수개월이 걸리던 영업 주기는 며칠 단위로 단축되었고, 특히 민간 상업 부문 고객 수가 폭발적으로 증가하는 결과.
- AI 비용 감소의 최대 수혜는 AI를 제일 잘 활용하는 주체임을 "SHOW AND PROVE"

자료: 팔란티어, OpenAI, 미래에셋증권 리서치센터

X. 팔란티어: 공매도 리포트 “OpenAI 대비 PLTR는 비싸다”

싼 건 절대 아니지만, OpenAI를 기준으로 팔란티어의 가치를 매기는 것이 합당한가?

- “팔란티어(PLTR)가 OpenAI 밸류에이션에 비해 비싸다”는 Citron 논리 축을 요약해보면 다음과 같음.
- 현존 AI의 “표준”으로 OpenAI를 상정하고, 이를 절대 기준으로 팔란티어(PLTR)의 밸류에이션, 사업 모델, 성장성의 한계를 공격한 것. PLTR의 핵심 기술인 온톨로지나 아키텍처는 논외로 하고, 비교하기 쉬운 재무 지표에만 집중. OpenAI의 기업가치가 5,000억 달러(직원을 대상으로 한 세컨더리 유동화로 직접 현금 유입은 없음)인데, 26년 매출을 약 300억 달러로 가정해서 “표준이 될 P/S 멀티플”을 17배로 가정.
- 반면, 팔란티어는 “정부 의존도·제품 도입 느림·범용성 낮음”과 같은 리스크를 안고 있다는 것을 차치하더라도, 표준 밸류에이션을 팔란티어의 26년 매출 추정치 56억에 적용하면 현재 주가가 너무 비싸다는 것. Citron이 말하는 PLTR의 적정 주가는 40달러.
- 그러나 주장의 기반이 되는 OpenAI의 엔터프라이즈 시장점유율 데이터(72%)부터 진위가 불분명. Menlo Ventures의 여름 설문(150개 기업 기술 의사결정권자 대상)에 따르면, Anthropic이 32%로 1위.

“Your Highness”라고 표현한 Citron의 리포트
두 회사가 위계적 질서에 있다는 오판

자료: Citron, 미래에셋증권 리서치센터

팔란티어와 OpenAI의 여러 방면에서의 비교
매출 밸류에이션으로만 따질 때는 팔란티어가 높기는 많이 높다, 근데 수익성으로 따지면?

	팔란티어	OpenAI
본사	콜로라도 덴버	캘리포니아 샌프란시스코
기업가치	3,742억 달러 (상장)	5,000억 달러 (비공식 라운드 진행중)
임직원 관련	3,936명 (작년 말), 중위 연봉 23만 달러	3,531명 (작년 9월), 보상액 수백만~수억 달러
AI 기술	실제 현장에서 AI를 활용하기 위한 운영체제	생성 AI 혁명을 이끈 GPT 개발
AI 생태계 입지	사실상 독점의 영역으로 인식	과점 시장이나 패러다임을 선도
트럼프 정부와 관계	공동창업자 피터 틸은 ’16년부터 공화당 지지	샘 알트만은 민주당에 100만 달러 이상 기부
테크 업계와 관계	빅테크 적대 + 스타트업들과 공생 관계 (Foundry for Builders 프로그램)	빅테크 친화 + 스타트업들의 공포의 존재 (DevDay 개최)
매출	25년 가이던스 41.46억 달러(YoY +45%)	25년 연환산 매출(ARR)은 200억 달러
현금 보유량	60억 달러의 현금 및 현금성 자산 보유	연간 80억 달러 현금 소진
잉여현금흐름	2Q25 기준 잉여현금흐름(FCF) 마진율은 57%, 25 년 조정 FCF 가이던스 중간값 20억 달러.	2030년에야 현금흐름 흑자
수익성 마진	2Q25 기준 조정 영업이익률 46%	비공개
EV/Sales 밸류	**89배 (올해 매출 가이던스 기준)**	**25배 (올해 ARR 추정치 기준)**

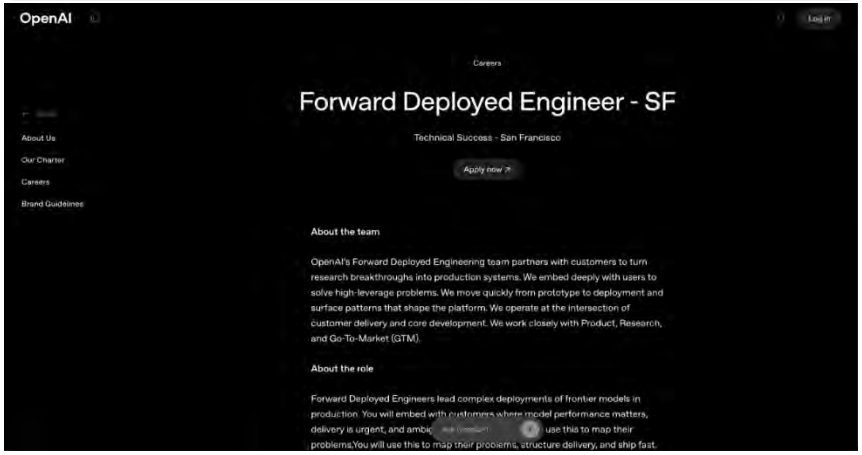
자료: 미래에셋증권 리서치센터

X. 팔란티어: 사실, OpenAI가 팔란티어를 따라한다

기업의 핵심을 장악하는 'OS'의 가치가 모델 자체의 가치를 압도한다

- "AI 모델의 정점"으로 평가된 OpenAI가 팔란티어의 사업 모델을 노골적으로 모방하는 것처럼 보여짐. 결정적 증거로 '전방 배치 엔지니어(FDE: Forward Deployed Engineers)'를 꼽을 수 있음. OpenAI는 최소 1,000만 달러를 지불하는 최상위 정부 기관(국방부 등)에 엔지니어를 직접 파견하여 맞춤형 AI 솔루션을 제공하는 신규 서비스를 시작. 이때 사용된 FDE라는 용어는, 소프트웨어와 함께 엔지니어를 고객사에 상주시켜 복잡한 문제를 해결하는 팔란티어 고유의 비즈니스 모델을 지칭하는 용어. 우연이 아닌 일종의 의도된 모방으로 읽힘.
- 이는 AI 시장의 가치 중심이 '최고의 모델'에서 '최적의 운영 시스템'으로 이동하고 있음을 증명하는 신호. 즉, 일반 API 서비스는 언제든 더 좋은 서비스로 대체될 수 있는 '상품(Commodity)'이 될 위험이 존재하기 때문인 것. 그러나 고객사의 고유 데이터와 업무 절차에 맞춰 완벽하게 맞춤 제작된 AI 모델은 한번 도입하면 교체가 쉽지 않아 강력한 Lock-in 효과를 통해 장기 계약을 확보하려는 것. OpenAI의 움직임은 팔란티어의 사업 모델이 옳았음을 어느정도 증명한 사건으로 사료됨.

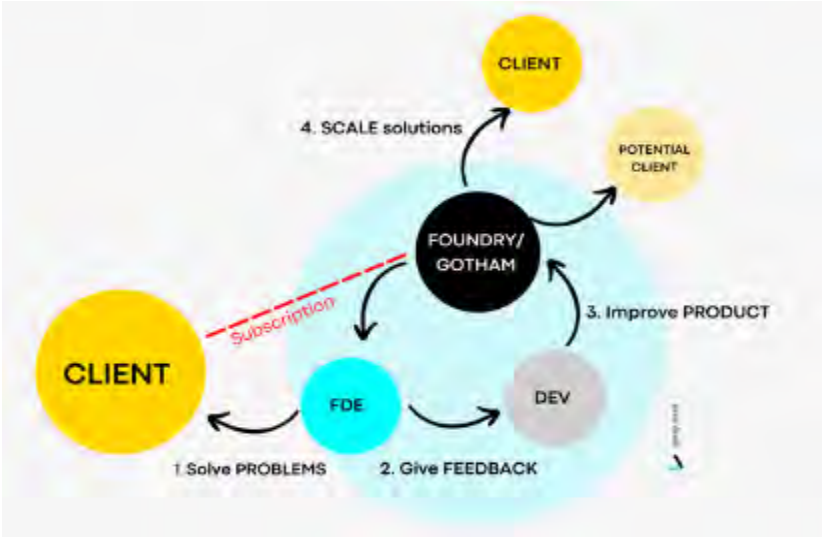
OpenAI가 팔란티어 고유의 용어인 "FDE"를 그대로 차용한 채용 공고



OpenAI가 단순 API 판매를 넘어, 고객사에 엔지니어를 직접 파견하여 맞춤형 솔루션을 제공하는 팔란티어의 사업 모델을 어느 정도 모방하고 있음을 보여줌

자료: OpenAI, 미래에셋증권 리서치센터

팔란티어의 핵심 경쟁 우위인 "FDE" 주도의 플라이휠 모델



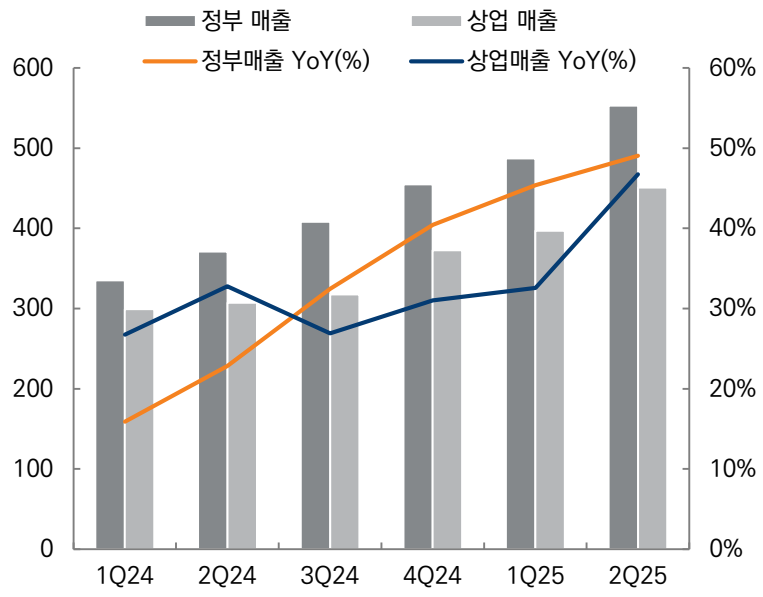
자료: Pantir Bullets(@Arny Trezzi and Emanuele Marabella), 미래에셋증권 리서치센터

X. 팔란티어: 실제 숫자는 공매도의 프레임과는 사뭇 다르다

‘서비스 과중, 느린 성장’이라는 프레임에 대한 반박

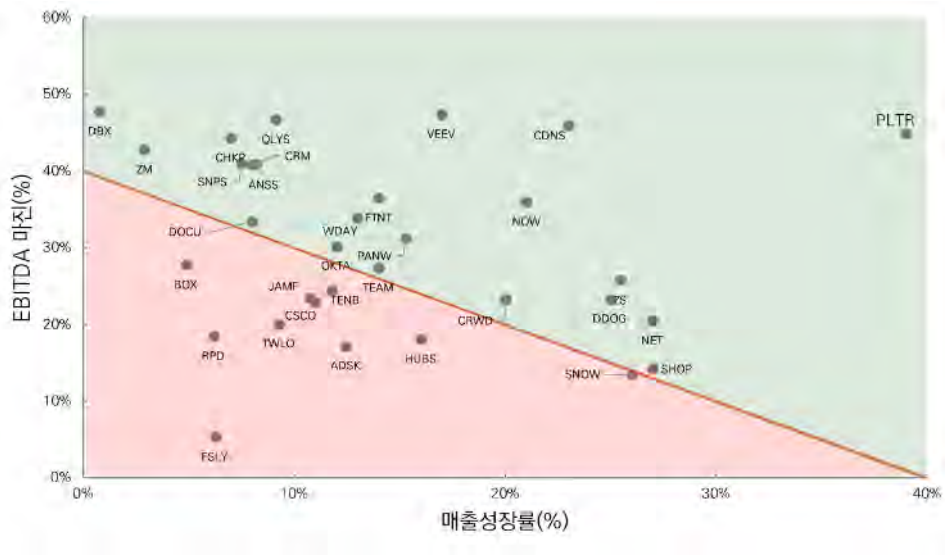
- 시트론이 주장한 ‘서비스 과중, 느린 성장’이라는 낡은 프레임은 PLTR의 2025년 2분기 실적 데이터와 완전히 모순됨. 실제 숫자는 플랫폼 기반의 큰 폭의 성장과 높은 레버리지를 증명함.
- 2분기 기준 전체 매출은 전년 대비 +48% 성장했고 조정 영업마진은 46%를 기록. 성장률과 마진의 합인 Rule of 40은 무려 94에 달함. 이는 최상위 SaaS 기업 수준의 지표로, ‘서비스 과중 低레버리지’라는 Citron의 주장과 정면 배치. 특히 성장의 질이 달라짐. 미국 상업 부문 매출이 전년 대비 +93%라는 폭발적 성장을 기록. 이는 팔란티어의 AIP(Artificial Intelligence Platform)를 중심으로 한 제품화 전략과 플랫폼 레버리지가 시장에서 성공적으로 작동하고 있음을 의미.
- OpenAI는 대규모 CapEx 투자로 장기간 현금 소모 불가피. 반면 PLTR은 지속 흑자와 57%에 달하는 높은 조정 FCF 마진을 기록 중. 동일한 P/S을 적용하는 것은 압도적인 현금 창출 능력을 무시하는 것.

정부 의존도를 거의 완전히 벗어난 팔란티어
심지어 정부 매출 증가율 자체도 꾸준히 상승 중



자료: Bloomberg, 미래에셋증권 리서치센터

팔란티어의 ‘Rule of 40(성장률 + 조정영업마진율) 지표’에서 피어 그룹을 압도
40만 넘어도 우수한 SW 회사로 인정받는 지표에서 94를 기록한 것



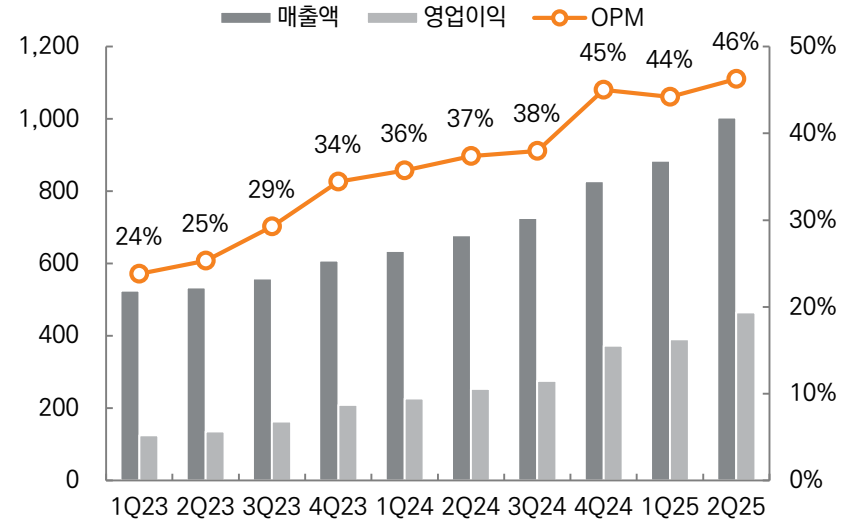
자료: Bloomberg, 미래에셋증권 리서치센터

X. 팔란티어: 숫자로 증명하는 성장의 레버리지

팔란티어는 고객의 생산성 향상으로 빠르게 성장중

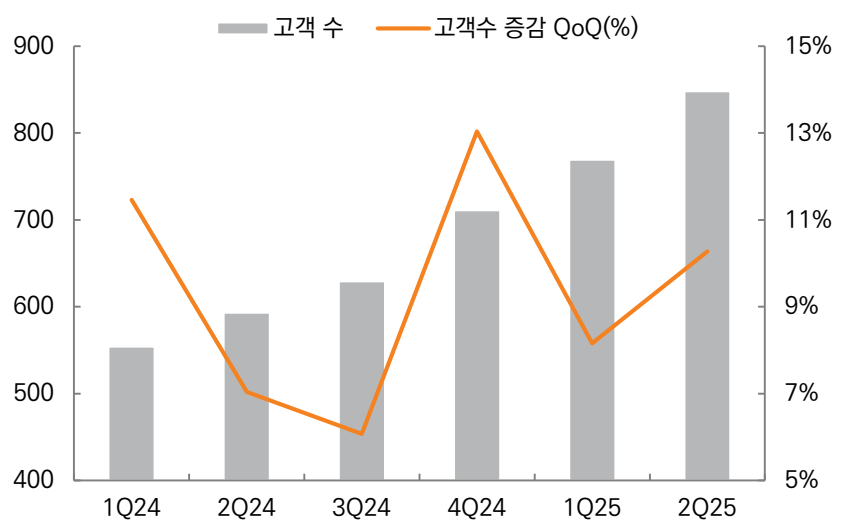
- 전세계적으로 AI 도입이 확산되면서 팔란티어는 그 성과를 기록적인 실적으로 증명. 올 2분기, 처음으로 분기 매출 10억 달러를 돌파하는 기염, 이는 전년 동기 대비 48%나 급증한 성장세.
- 상업 부문은 AI 혁신을 주도하며 더 빠른 속도로 확장 중. 미국 내 상업 고객 수는 전년 대비 64% 증가했으며 , 기존 고객의 지출 확대를 나타내는 'Net Dollar Retention'은 128% 달성, 실제로 상위 20개 고객의 매출은 30% 증가했으며, 미국의 한 통신사는 2년 만에 계약 규모를 10배로 늘리는 등 팔란티어의 가치를 경험한 기존 고객들이 파트너십을 더욱 공고히 하고 있는 상황.
- 한편, 미국 정부 매출은 전년 대비 53%라는 높은 성장률을 보이며, 여전한 앵커로 자리 중. 미 육군과 체결한 최대 100억 달러 규모의 10년 장기 계약은 단순 공급사를 넘어 국가 안보의 핵심 전략 파트너로서 팔란티어의 위상을 증명하는 상징적인 성과로 사료됨.

팔란티어의 2분기 주요실적



자료: Bloomberg, 미래에셋증권 리서치센터

팔란티어의 고객 수 및 증감 추이



자료: Palantir, 미래에셋증권 리서치센터

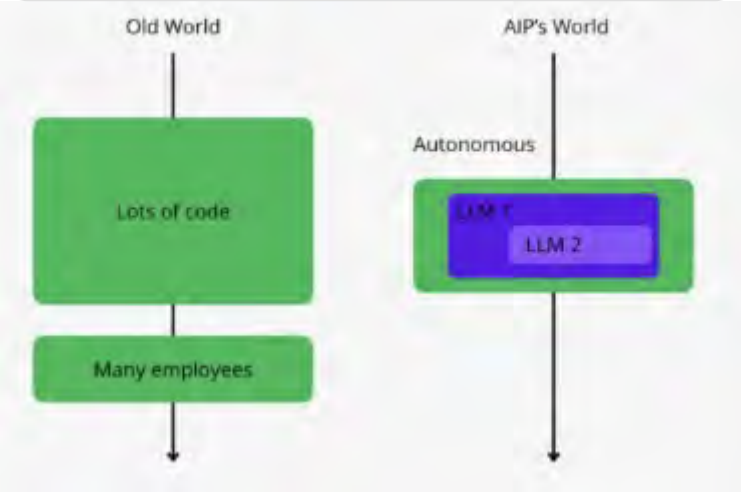
X. 팔란티어: 온톨로지를 모르니 오해가 생긴다

운영체제와 API 플랫폼은 애초에 비교 대상이 아냐

핵심은 온톨로지다

- Citron을 포함한 투자자들은 OpenAI와 팔란티어처럼 본질이 다른 두 사업 모델을 동일 선상에 놓는 구조적 오류를 범함. PLTR은 AI 모델 제공사가 아닌, 기업의 의사결정과 실행을 책임지는 '운영 시스템(OS)'을 구축하는 회사. 엄밀히 말해 PLTR은 AI 회사가 아니라 빅데이터 회사인 것.
- 리포트가 완전히 간과된 PLTR의 핵심은 온톨로지. 개별 데이터에 의미(Object)를 부여하고, 각 기업들이 처한 일반적/예외적 비즈니스의 규칙(Logic)과 실제 실행(Action)을 한 곳에 묶는 디지털 트윈. LLM 분석으로 끝나는 게 아니라 허용된 경계 안에서 "구체적 행동"을 하도록 하는 운영 시스템.
- OpenAI의 가치는 모델 API 호출량과 사용자 트래픽(DAU 등)에서 발생. 반면 PLTR의 핵심 가치는 데이터, 규칙, 행동을 하나의 '온톨로지'로 묶어 실제 운영(Operation)을 자동화하고 통제하는 E2E 인프라에 있음. 온톨로지 기반 AIP는 한번 도입되면 교체 비용이 극도로 높아지는 강력한 Lock-in 효과가 핵심. 그러나 AI 모델은 전환 비용이 별로 없어, 둘을 단순 비교하는 것은 의미가 크게 떨어짐.

팔란티어 高성장을 이끄는 AIP는 '자율조직'을 탄생시킴
거대한 조직 기능을 단일 자율 시스템으로 압축하여,
그 안에서 여러 LLM 모델들을 효율적으로 작동시키는 미래!
핵심은 LLM이 아니라, 이들을 통제/활용하는 온톨로지다.



자료: Antonio Linares, 미래에셋증권 리서치센터

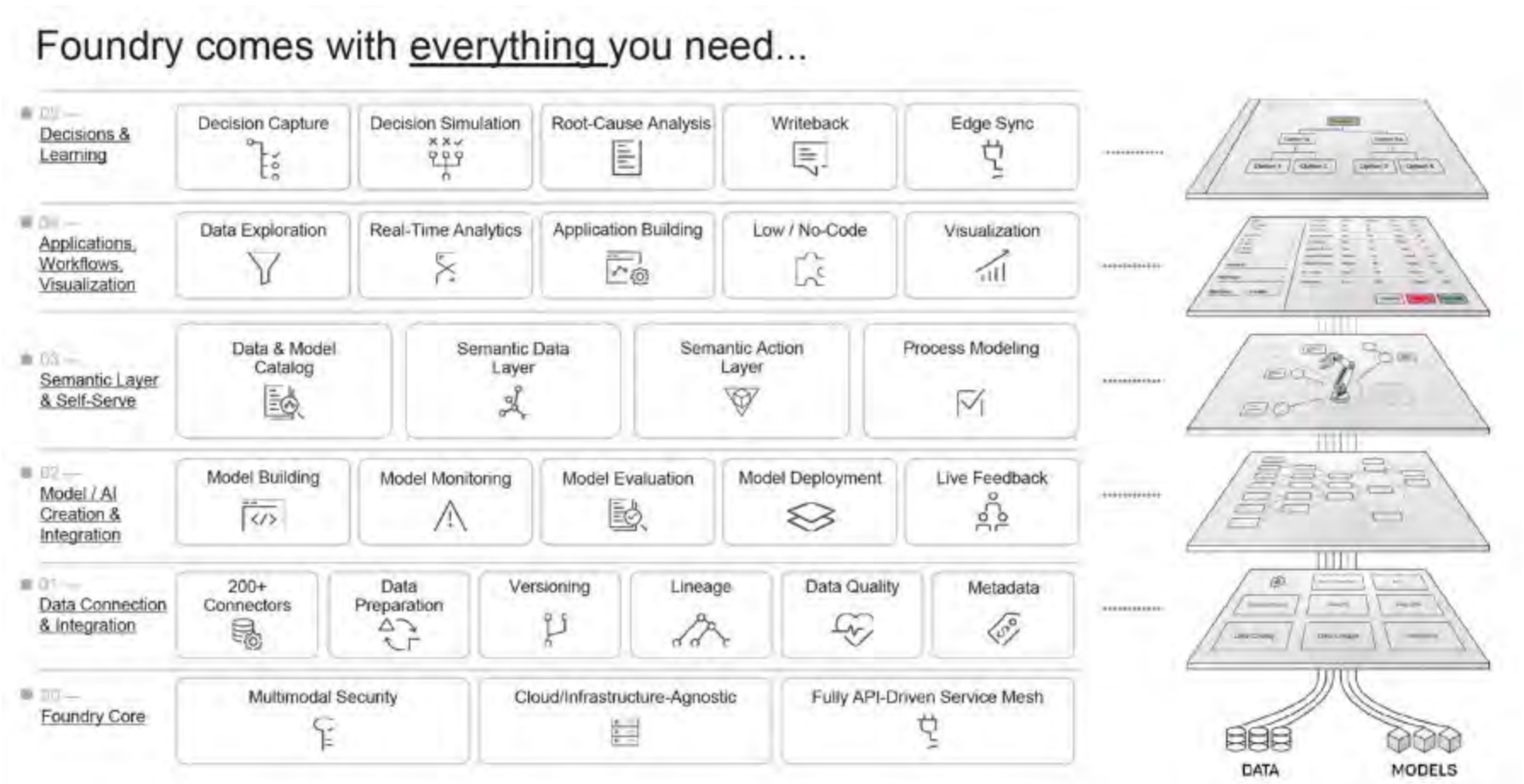
팔란티어의 주요 제품들과 주요 기능
AI 모델은 사용자와 대화 인터페이스 제공하는 프론트엔드이지,
시스템 핵심은 아니다

제품 및 영역	직관적 설명	주요 기능
Ontology **= 핵심 구조**	**데이터가 무엇인지, 어떻게 연결되는지를 설명해주는 근본 아키텍처**	**- 데이터 관계 정의 (객체, 속성, 관계)** **- 데이터 의미론적 연결** **- 지식 그래프**
Foundry/Gotham = 핵심 제품	회사의 모든 데이터를 모아두고, 정리하고, 분석하는 곳	- 데이터 통합 (이기종 데이터 통합) - 데이터 관리 및 분석 - 데이터 시각화 및 대시보드 - 협업 및 워크플로우 지원 - 데이터 보안 및 접근 제어
AIP (AI Platform) = AI 활용 공간	Ontology를 기반으로 Foundry/Gotham의 데이터를 활용하여 '개인화된 챗봇'과 같은 앱을 만들고 관리	- LLM 기반 AI 애플리케이션 개발 플랫폼 - AI 모델 (LLM 포함) 배포 및 관리
LLMs (모델들) = 상호작용 능력	AIP 안에서 챗봇의 "두뇌" 역할을 하며, 사람의 말을 이해하고 답변을 생성	- AI 모델과 비즈니스 프로세스 연결 - 텍스트 생성, 번역, 요약, 질의응답 등

자료: 미래에셋증권 리서치센터

[참고] 기업의 일을 자동화하는 “에이전트”의 핵심은 온톨로지

팔란티어가 단순한 데이터 분석 툴이 아닌, '미래 경제의 운영체제(OS)'가 되는 이유를 구조적으로 보여주는 청사진



- 맨 아래 데이터 연결(01)부터 시작해 현실 세계를 디지털로 복제하는 핵심, 즉 'Semantic Layer(온톨로지)'(03)를 구축하는 전 과정.
- **이것이 바로 단순 LLM과의 근본적 차이점으로, 온톨로지는 분산된 데이터에 맥락과 의미를 부여하여 기업이라는 유기체의 '뇌' 역할을 수행.**
- 온톨로지 위에서 비로소 AI 모델(02)과 애플리케이션(04)이 지능적으로 작동하며, 의사결정과 학습(05)을 통해 스스로 발전하는 루프가 완성됨.

자료: 팔란티어, 미래에셋증권 리서치센터

X. 팔란티어: 온톨로지+제본스의 역설 = 미국 제조업 부흥

팔란티어와 미국의 제조업 재건

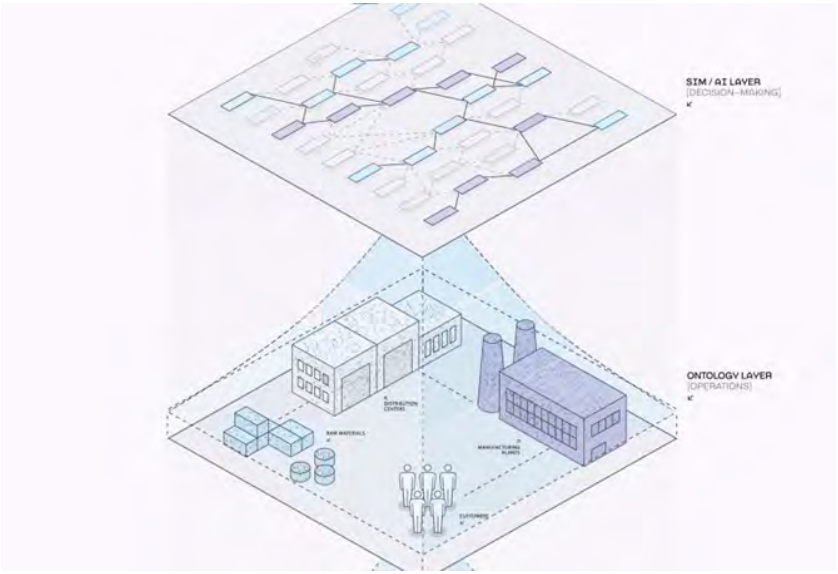
- AI에서의 '제본스의 역설'과 비용 감소 추세는 산업 전반의 수요를 촉발하는 동시에, 미국 트럼프 정부의 제조업 리쇼어링 정책과 맞물려 새로운 기회를 창출. 리쇼어링의 성공을 위해서는 해외의 낮은 생산 단가와 수십 년간 축적된 제조 노하우 및 공급망 관리 능력을 극복해야 하는 과제가 존재함.
- 이에 대한 해답은 AI와 공장을 가상 세계에 그대로 복제한 '디지털 트윈'을 만들어, 위험과 비용 없이 수많은 시뮬레이션을 통해 운영을 최적화하는 방식. 이 혁신의 구체적인 실행 방법론은 'AI 최적화 루프'로 실시간 공정 데이터를 '온톨로지' 기술로 연결해 '디지털 트윈'을 구축. 그리고 이 가상 실험실 안에서 AI가 최적의 해결책을 찾아 현실에 적용하는 과정이 반복.

제조업에서의 디지털 트윈의 기능

핵심 기능	세부 설명
핵심 디지털 모델 구축	세계의 핵심 의미와 물리법칙을 포함하는 핵심 디지털 그림을 구축
비즈니스 애플리케이션 구축	디지털 트윈을 기반으로 실행 가능한 비즈니스 애플리케이션을 구축하고 데이터를 연결
분석 및 시뮬레이션 실행	가정 분석 및 복합 시뮬레이션 실행
데이터 기반 의사 결정	비즈니스의 경계에서 더욱 스마트하고 데이터 기반의 의사 결정 수립
의사결정 연결	공장 현장부터 임원진까지 지역적 의사결정을 전체 의사결정과 연결

자료: Palantir, 미래에셋증권 리서치센터

팔란티어의 비즈니스 디지털 트윈 구축
현실세계에서 수집한 데이터로 온톨로지 레이어를 구축한 뒤 의사결정



자료: Palantir, 미래에셋증권 리서치센터

X. 팔란티어: AI로 미국 제조업을 다시 위대하게

Warp Speed로 이뤄내는 제조업 혁신

- 팔란티어가 내세우는 표어이자 그들의 제조업 전용 OS인 “Warp Speed”는, AIP를 기반으로 구축한 최신 제조 운영 시스템으로 “MAGA”식 제조업 재건에 중요한 역할을 수행.
- Warp Speed는 제조 현장의 IT 데이터와 운영 데이터를 모두 실시간으로 연결하여 기업 전체를 가상 세계에 복제한 디지털 트윈을 구현. AI는 설비의 고장을 미리 예측하고, 공급망 중단과 같은 예기치 못한 상황에 즉각적으로 대응하며, 사업 전체의 자본 흐름과 재고를 파악하여 효율적인 투자를 결정.
- 결국 Warp Speed는 과거의 단절되고 수동적이었던 제조업 운영 방식을 모든 데이터가 연결된 디지털 트윈으로 전환/극복 시키는 것. 온톨로지 위에서 AI가 스스로 최적의 답을 찾아내는 효율화를 통해 미국의 제조업 역량의 한계 극복하는 것으로 정리.

2025년 상반기 Warp Speed 관련 주요 계약 체결 내용

계약 업체	발표일	주요 내용
두산밥캣	2025년 1월	디지털 전환 가속화: 전사적으로 파운드리 플랫폼과 Warp Speed를 도입하는 다년 계약을 체결 운영 최적화: AI를 활용하여 제조 공정을 개선하고, 생산 일정 및 자재 소싱 최적화
파나소닉 에너지	2025년 2월	EV 배터리 생산 혁신: 북미 파나소닉 에너지(PENA)의 캔자스 신규 EV 배터리 공장에 Warp Speed를 도입하는 다년 파트너십을 체결
SNC (Sierra Nevada)	2025년 3월	프로그램 관리 혁신: 항공우주 및 국가 안보 분야의 복잡한 프로그램을 관리하는 데 Warp Speed를 사용하여 부서 간 협업을 AI로 강화 디지털 기반 지속적 개선: 실시간 피드백 루프를 활용하여 디지털 기반의 지속적인 개선이 목표
Ursa Major	2025년 3월	로켓 엔진 생산 증대: 고체 및 액체 로켓 모터의 생산을 기하급수적으로 늘리기 위해 Warp Speed를 도입
Divergent Technologies	2025년 5월	온디맨드 제조 혁신: Warp Speed 플랫폼을 통해 Divergent의 첨단 적응형 생산 시스템(DAPS)에 접근할 수 있도록 하는 전략 파트너십 체결 공급망 복원력 강화: 고객이 공급망의 취약점을 신속하게 파악하고, 필요한 부품을 즉시 디지털로 제조하여 공급받을 수 있도록 지원
The Nuclear Company	2025년 6월	원전 건설 현대화: 미국 내 원자력 발전소 건설을 현대화하고 간소화하기 위한 AI 기반 소프트웨어 플랫폼을 출시 프로젝트 최적화: 스마트 공급망 시스템을 통해 자재 부족을 최소화하고, 실시간 데이터를 기반으로 작업 흐름을 조정하여 공사 지연 방지 목표
BlueForge Alliance	2025년 7월	군함 건조 혁신: '군함을 위한 Warp Speed(Warp Speed for Warships)'를 출시하여 미국 해양 산업 기반의 디지털 전환 혁신 공급망 연결: 조선사, 공급업체, 파트너 네트워크를 디지털로 연결하여 생산을 확장하고 복원력을 높이는 것이 목표

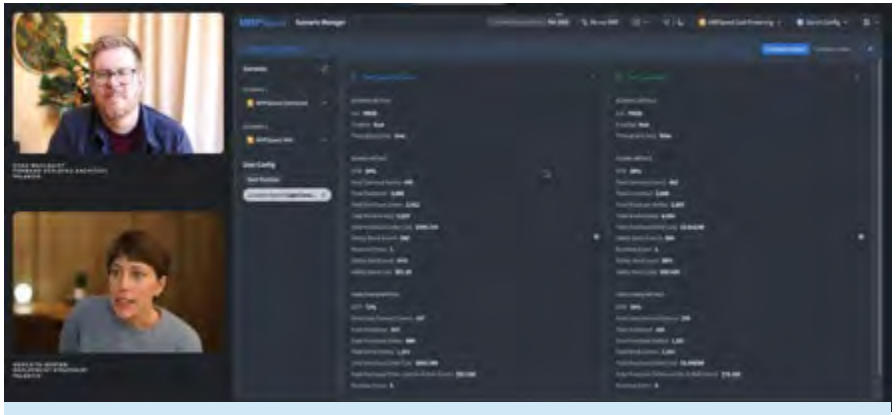
자료: Palantir, 미래에셋증권 리서치센터

[참고] 온톨로지 기반의 Warp Speed를 통한 재고관리

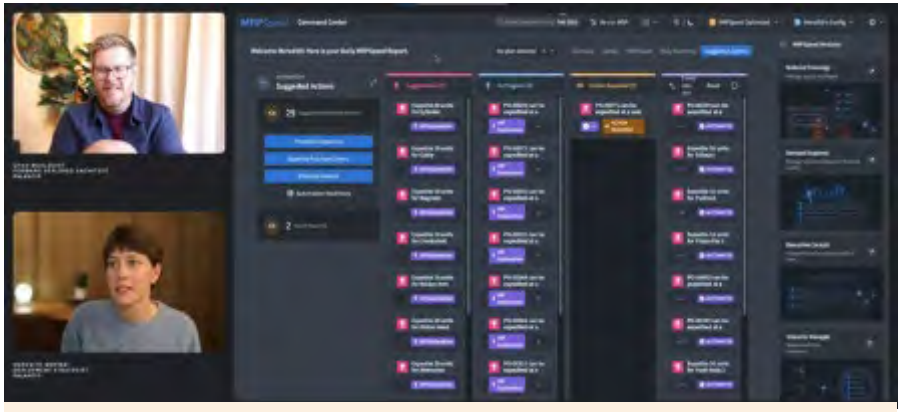
팔란티어는 어떻게 제조업 혁신을 이뤄내는가

- 팔란티어 Warp Speed의 핵심 기능인 재고관리 솔루션 “MRPSpeed”는 생산 공정의 모든 객체와 프로세스를 온톨로지로 구성하여 공장 전체의 디지털 트윈을 구축한 뒤, AI는 다양한 생산 시나리오를 시뮬레이션하여 어떤 부품 재고에서 병목 현상이 발생할지를 미리 예측.
- MRPSpeed는 단순히 병목 현상을 해결할 방안을 제시하는 데 그치지 않고, AI가 제안한 솔루션이 실제 어떤 결과를 가져왔는지, 즉 원인과 결과를 데이터로 저장하고 이를 학습. 다음에 비슷한 상황이 발생했을 때, 과거의 학습 데이터를 기반으로 더 정교하고 나은 대안을 제시하는 자기 개선 루프.
- 특히 공급망을 새로 구축해야 하는 북미의 리쇼어링 기업은 부족한 공급망 관리 노하우를 AI로 보완하여 안정적이고 효율적인 생산을 가능하게 함으로써 부족한 제조업 노하우를 AI SW를 통해 극복.

Warp Speed를 이용하면, 추후 비슷한 상황이 발생할 때 과거의 학습 데이터를 기반으로 더 나은 대안을 제시하는 '자기 개선 루프'를 가질 수 있음
시뮬레이션 및 예측 (왼쪽): 디지털 트윈 위에서 다양한 시나리오를 돌려보며, 미래에 발생할 수 있는 공급망 병목 현상과 그 영향을 미리 파악
구체적 해결책 제시 및 실행 (오른쪽): 예측된 문제를 해결하기 위한 최적의 조치를 AI가 구체적으로 제안하고, 클릭 한 번으로 자동 실행까지 지원



화면 중앙에 'MRPSpeed Optimized'와 'MRPSpeed MAX'라는 두 시나리오 비교. 이는 디지털 트윈 환경에서 각기 다른 운영 전략을 적용했을 때, 어떤 결과(총 구매 비용, 재고 수준, 총 작업 지시량 등)가 나올지 AI가 미리 시뮬레이션했음을 의미.



화면 중앙에는 AI가 제안하는 수십 개의 구체적인 조치 목록(Cylinder 10개, Cable 18개 긴급 조달 등)이 나열. 각 제안 항목 옆에는 'AUTOMATED' 버튼이 있음. AI의 제안을 클릭 한 번으로 즉시 실행할 수 있음을 보여줌.

자료: Palantir, 미래에셋증권 리서치센터

[참고] 팔란티어의 온톨로지를 접목한 제조업 사례

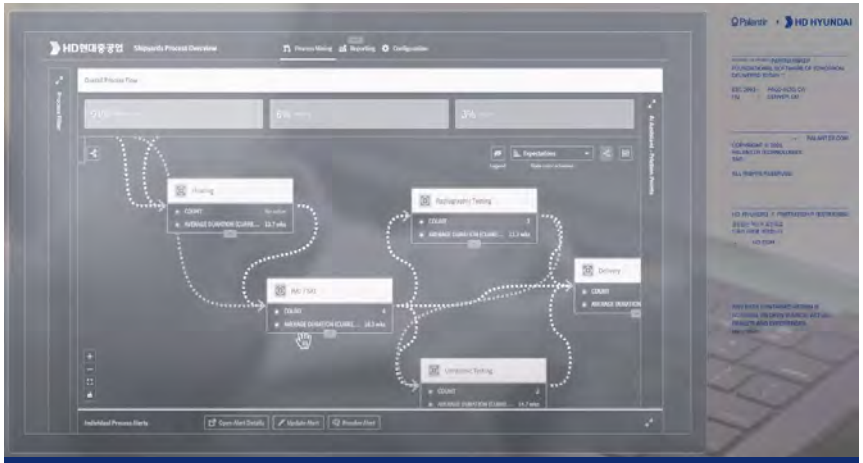
HD현대를 통해 보는 팔란티어 실증 사례

- HD현대는 ’22년 팔란티어와 5년간 2천만달러 규모 파트너십을 맺음. 주요 계열사에 파운드리 플랫폼을 도입하기 위함. 예를 들어, 제조 플랜트에서 AIP(AI Platform)를 활용하면, 현장 관리자가 “다음 달 생산 계획을 최적화해줘”라고 자연어로 묻는 것만으로도 AI가 관련 데이터를 신속히 조회하고 분석하여 최적의 생산 일정안을 제안하는 것.
- 나아가, 과부하 혹은 저부하가 예상되는 공정과 시점을 미리 식별하여 작업 인원 등 가용 자원의 배치를 최적화하고, 설비의 이상 징후가 발생하면 자동으로 정비 작업지시를 생성하는 것도 가능. 이는 곧 데이터를 중심으로 인간과 AI가 협업하여 최상의 의사결정을 내리는 효율적인 업무 방식.

팔란티어 공식 유튜브에 업로드된 HD현대와의 협력

자료: Palantir, 미래에셋증권 리서치센터

HD현대중공업의 조선소 생산 공정을 Foundry 플랫폼에서 시각화
생산일정 준수율 한눈에 확인



각 프로세스 간의 화살표는 실시간 공정 흐름이고 단계별 소요시간도 기재.
생산 프로세스에서 병목/지연이 발생하는 부분을 빠르게 파악하고 대응하게 함.

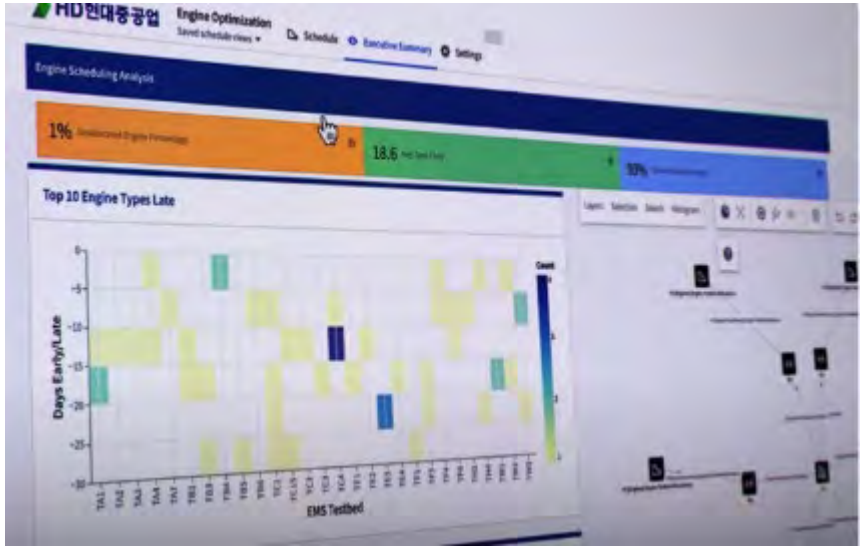
자료: Palantir, HD현대중공업, 미래에셋증권 리서치센터

[참고] 온톨로지 기반의 미래형 스마트 조선소 프로젝트

팔란티어의 최종 목표는 미래형 스마트 조선소

- 'Future of Shipyard (FOS)' 프로젝트는 설계부터 생산에 이르는 조선소의 모든 공정을 실시간 데이터로 연결하는 것을 핵심으로 하며, 여기에 팔란티어의 파운드리(Foundry) 플랫폼이 적용.
- AI 기반 선박 설계는 선박 내 사각지대를 제거하여 시야를 확보하고 적재 효율을 극대화하는 것을 목표로 하고 있음. 2030년까지 자율 운영 조선소 완성을 목표로 하는 이 프로젝트가 완료되면 생산성은 30% 향상되고 선박 건조 기간은 30% 단축될 것으로 기대.
- 설계, 생산, 품질, 안전 등 각 분야의 데이터를 파운드리에 연계하여, 과거 건조 실적과 현재 설계안을 AI로 분석해 설계 문제를 예측하고 용접 품질 검사 데이터를 통해 결함 원인 분석 및 공정 피드백을 자동화하는 등 의사결정 최적화와 생산성 향상을 위한 다양한 활용 사례를 만들고 있음.

조기 납기 또는 지연을 시각화한 엔진 생산 스케줄링 화면
각 엔진 유형별로 얼마나 일찍 혹은 지연되어 있는지 확인 가능



자료: HD현대중공업, 미래에셋증권 리서치센터

생산 설비(테스트베드)의 가동률을 시간 흐름에 따라 시각화
각 설비의 가동률(93%)을 최적화 상태에서 유지하도록 실시간 모니터링

자료: HD현대중공업, 미래에셋증권 리서치센터

[참고] HD 현대의 AI 도입 실증 사례

HD현대의 기타 계열사들의 AI 도입 현황

사업 분야	주요 AI 및 디지털 기술 도입 내용	적용 사례 및 실증 사례	주요 성과 및 기대효과	팔란티어 활용 여부
건설기계	팔란티어 Foundry 기반 빅데이터 플랫폼 구축 ('DI 360')	부품 공급망 관리, 생산현장 품질 클레임 대응, 판매 기회 발굴 및 최적화	부품 수급 지연, 설비 다운타임 등 리스크 사전 감지 및 자동 대응 가능	✓
	현장 데이터와 전사 시스템(ERP, MES 등)의 데이터 통합	운영 모니터링 체계 구축	생산계획 최적화, 재고 비용 절감, 납기 단축	✓
	시장 수요 데이터 기반 AI 수요 예측 모델 구축	시장 수요 기반으로 최적화된 생산계획 수립	정확한 수요 예측, 재고 최소화 및 생산성 극대화	✓
	구글 클라우드 Vertex AI 기반 'X-wise Xite' 모델 개발 예정	스마트 건설현장 구현, 자율 작업 및 머신 어시스턴트 구축	숙련공 업무를 신규 인력도 10년차 숙련공 수준으로 수행 가능	✗ (구글 버텍스 AI)
	완전 자율 건설현장 구현 (무인 자율 작업 기술)	AI 보조를 통한 건설 장비 무인 자율작업 구현 (예: 굴착기)	건설 현장 생산성 극대화 및 인력난 해소	✗ (자체 개발)
에너지 플랜트	팔란티어 Foundry를 통한 데이터 통합 (원유 도입→생산 전 과정)	정유공장 100여 개의 생산 관리 시스템 및 IoT 센서 데이터 통합	공정운전자가 실시간으로 최적 생산계획 수립 가능, 에너지 사용 절감	✓
	팔란티어 기반 원유배합 최적 추천 시스템 구축	정제 가격 변화에 따른 최적의 원유 배합 실시간 추천	공정 효율 극대화 및 정제 마진 개선	✓
	데이터 드리븐(data-driven) 실시간 운영 의사결정 대시보드	현장 작업자가 공정 상태를 매일 최적화 관리 가능	유가 급변, 공급망 교란 등 환경 변화 시 즉각적 대응 가능	✓
	AI 기반 설비 예지보전(CBM) 체계 구축	플랜트 설비 데이터 실시간 모니터링 및 이상 징후 조기 감지	설비 고장 예방, 설비 유지보수 비용 절감	✓
	팔란티어 플랫폼의 산업 적용 가능성 검증	미국 육군의 예측 유지보수 사례(팔란티어)를 통해 산업 적용 가능성 확인	산업 전반에서 팔란티어 플랫폼의 광범위한 적용 가능성 기대	✓
물류·해운	HD현대마린솔루션의 AI 기반 OceanWise 플랫폼	빅데이터 및 AI 알고리즘을 활용한 선박 항로 최적화	평균 5.3% 연료 절감 입증, 연간 약 3억 5천만 원 절약 효과	✗ (자체 개발)
	HD현대 자회사 Avikus 개발 자율운항 시스템 (HiNAS)	딥러닝 기반의 자율운항 기술로 속도 제어, 충돌 회피 등 기능 포함	세계 최초 태평양 완전 자율 횡단 성공, 연료효율 7% 향상 및 온실가스 배출 5% 감소	✗ (자체 기술)
	자율운항 선박 시장 확대 (상용화 및 국방 분야 진출)	자율운항 솔루션 전 세계 350척 이상 수주, 국방용 무인 탐색함(USV) 개발 목표	해상 사고율 감소 및 인적 오류 최소화, 미래 해운업 및 국방 분야 신시장 개척	✓

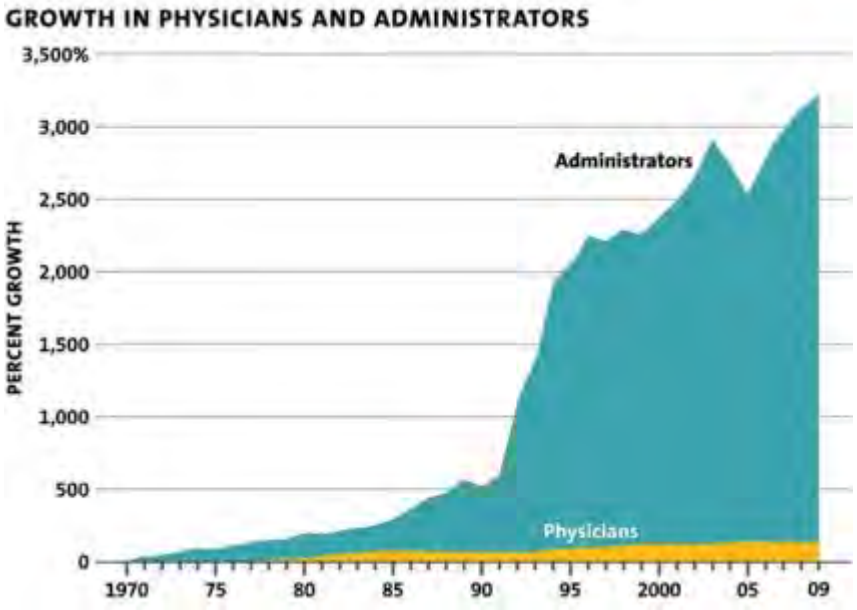
자료: HD현대, 미래에셋증권 리서치센터

X. 팔란티어: 헬스케어는 “가장 많은 데이터”를 보유한 섹터

흩어져 있고 놀고 있는 데이터가 너무 많다

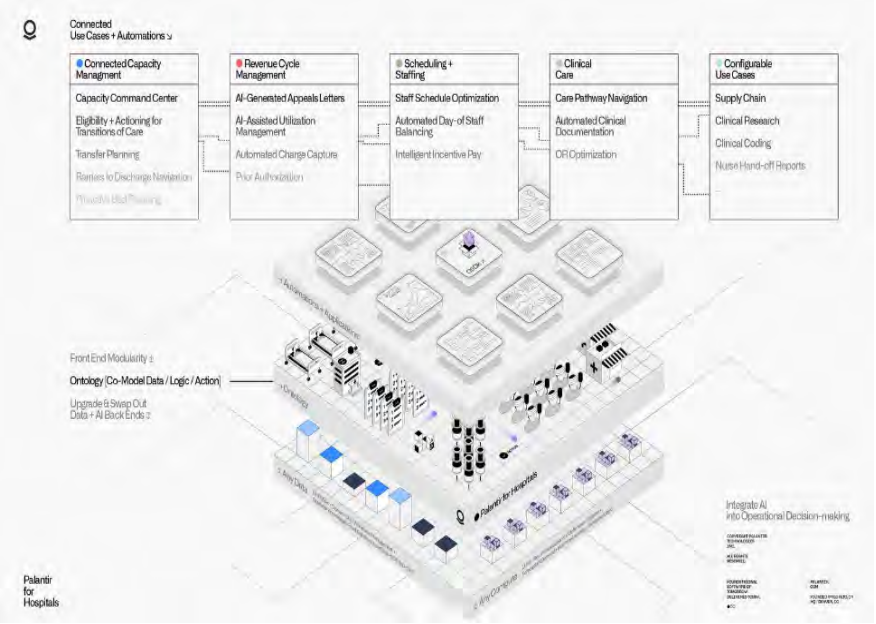
- 팔란티어가 헬스케어에 집중하는 이유는 이곳에 가장 큰 기회가 있기 때문. 헬스케어 분야는 전 세계 데이터의 약 3분의 1을 생산하지만, 그중 97%가 활용되지 않은 채 낭비되고 있음. 데이터 증가는 의료비 상승의 주요 원인을 촉발. 관련 행정담당자들의 업무(서류 작성, 보험 처리, 스케줄링, 의료 기록 관리 등)는 AI가 잘하는 대표적인 분야.
- 팔란티어는 “잠자는 데이터”로부터 가치를 찾는 것을 목표로 함. 이미 英 NHS나 美 HCA 헬스케어 등 주요 기관을 고객으로 확보하며 시장의 신뢰와 전문성을 입증. 궁극적으로는 ‘운영 플랫폼’을 EMR처럼 헬스케어의 필수 표준 운영체제(OS)로 만드는 것을 목표로 삼으며, 이를 위해 '3~5년'의 장기 계획이 아닌 '3~5개월' 내의 고객에게 수익창출을 증명함으로써 시장선점 전략을 펼치고 있음.

의료비 상승의 주요 원인의 원인은 반복적 행정 업무 = AI로 혁신 가능
의사 숫자 대비 행정 업무 담당자 숫자는 3000% 이상 폭발적으로 증가



자료: 미 통계청; NCHS; Himmelstein/Woolhandler analysis of CPS, 미래에셋증권 리서치센터

병원 내 흩어져 있는 '잠자는 데이터'들을 온톨로지로 연결하는 구조
병원의 만성적인 비효율성을 해결하는 구체적인 앱과 자동화 솔루션 제공



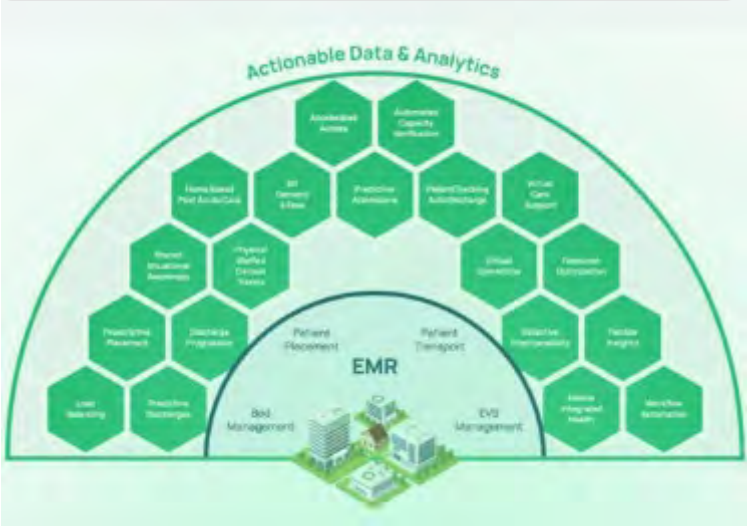
자료: Palantir, 미래에셋증권 리서치센터

X. 팔란티어: 온톨로지+의료 데이터 = 헬스케어 혁신

온톨로지는 병원시설의 복잡성의 문제를 효율화

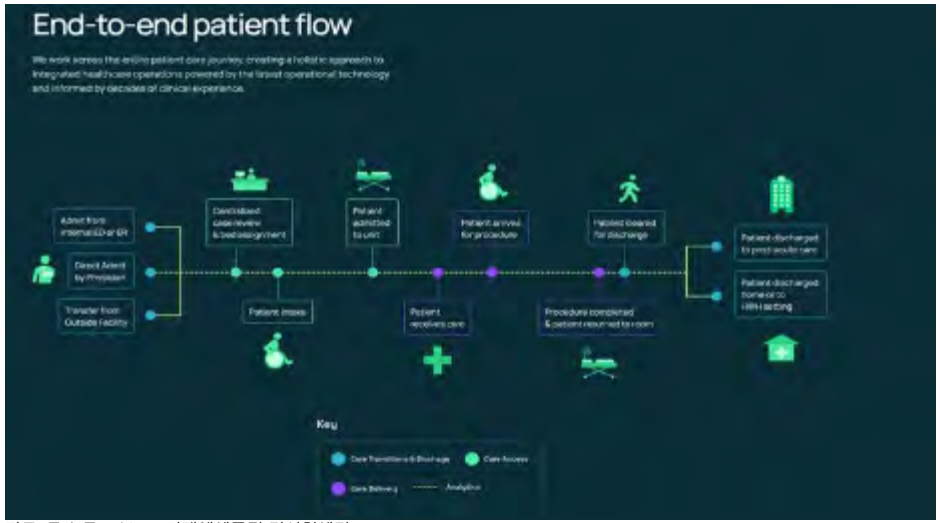
- 최근 팔란티어는 30년 이상 병원 운영 관리 분야를 선도해 온 TeleTracking과의 파트너십을 발표. 이는 팔란티어의 최첨단 AI 플랫폼(AIP과 Foundry)과 TeleTracking의 현장 중심 운영 플랫폼(Operation IQ)을 결합하여, 병원 운영의 근본적인 비효율성을 해결하고 새로운 표준을 제시하는 것을 목표로 하고 있음.
- 오늘날 병원과 헬스케어 시스템은 응급실 병상 대기, 끊임없는 인력 부족, 수용 능력 한계와 같은 만성적인 문제들이 현장을 짓누르고 있으며, 수용 능력, 인력 배치, 환자 흐름에 대한 실시간 가시성이 부족하여 리더들은 불완전한 정보에 기반한 사후적 결정을 내릴 수밖에 없는 한계점이 존재.
- TeleTracking의 CEO 크리스 존슨은 이러한 상황에 대해 "향후 5년 내 운영 플랫폼을 채택하지 않는 헬스케어 시스템은 비용 압박으로 인해 생존에 실존적 위협을 맞게 될 것"이라고 경고하며, 이제 운영 플랫폼은 선택이 아닌 생존을 위한 필수 요소가 되었음을 강조함.

흩어져 있던 EMR 데이터를 활용해 병상 관리, 환자 흐름 최적화 등 병원의 핵심 운영 문제를 해결하는 방안
과거의 수동적, 사후적 의사결정을 예측 기반의 자동화 전환



자료: TeleTracking, 미래에셋증권 리서치센터

"팔란티어+TeleTracking"의 환자 및 의료 프로세스 접근 방식
입원부터 퇴원까지 전 과정에 걸쳐 단절되어 있던 데이터를 연결,
병원 운영의 핵심 과제인 환자 흐름(Patient Flow)을 통합적으로 관리



자료: TeleTracking, 미래에셋증권 리서치센터

X. 팔란티어: 원래의 주특기는 AI 혁명으로 새로운 변화를 맞이

전쟁의 첨단화 시대에서, 팔란티어의 역할 증대

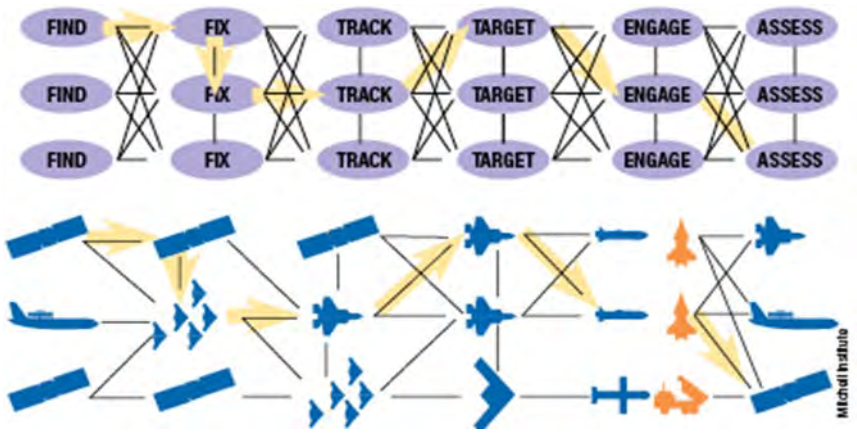
- 과거의 전쟁이 양적 우위를 앞세운 소모전이었다면, 러시아-우크라이나 전쟁에서 관측된 현대전의 핵심은 완전히 다른 양상. 한정된 전력(드론, 재블린 미사일 등)을 활용해 적의 핵심 자산을 정밀 타격하는 '효율성의 소모전'으로 변화. 새로운 전쟁의 공식을 마주한, 지휘관들은 1)한정된 자원의 최적화, 2)신속한 의사결정 프로세스 구축, 3)분산된 아군의 유기적 통합이라는 세 가지 핵심적인 과제에 직면.
- 팔란티어 등 전쟁 AI는 이 현대전의 핵심 과제들을 해결하며 인간 지휘관의 능력을 극대화하는 결정적 역할. 위성, 드론 등에서 쏟아지는 데이터를 실시간으로 융합 및 분석하여 지휘관의 전장 상황 인식을 돕고, 이를 통해 '킬 체인' 완성 시간을 획기적으로 단축 (예: 핵심 표적 자동 식별, 최적의 무기 사용 추천 등).
- 팔란티어의 최고 기술책임자(CTO)인 Shyam Sankar는 메타와 OpenAI의 임원과 함께 최근 美 육군 예비군 산하 특수부대인 Detachment 201에 중령 계급으로 임관. 군의 디지털 전환 가속화 역할. 팔란티어는 이미 미군·정보기관에 핵심 데이터 플랫폼을 공급하는 이해관계자로서, CTO의 입대는 운용상·계약상 장벽이 낮아지고, 최전선 요구사항을 직접 흡수할 통로를 확보. 쉽게 말해, 국방/공공 매출 확대 가능성을 암시.

AI의 전장에서의 역할과 기대효과

구분	핵심 기능	세부 설명
핵심 기술	AI 및 머신러닝 통합	AI/ML 알고리즘을 활용하여 방대한 전장 데이터를 신속하게 수집, 분석 및 평가
응용 분야	지능형 표적 식별 및 분석	다양한 출처의 데이터를 융합하여 표적을 자동으로 식별하고 위협 우선순위를 지정
	예측 기반의 군수 작전	전 세계 공급망 현황에 대한 가시성을 제공하고, 시나리오 기반 계획을 통해 군수 작전을 최적화
기대 효과	운용 효율성 극대화	동일 임무수행에 필요한 시간/인력을 획기적 감축
	유연성 및 상호운용성	초기 설계에 없던 다양한 데이터 소스도 통합 가능, 계급과 역할에 따른 맞춤형 데이터 접근 권한 제공

자료: 미래에셋증권 리서치센터

전장 작전 수행을 고유기능 탑재 노드들로 구성된 킬체인 웹
노드간 정보/명령 전달에서 상당한 속도 지연 및 오차 발생



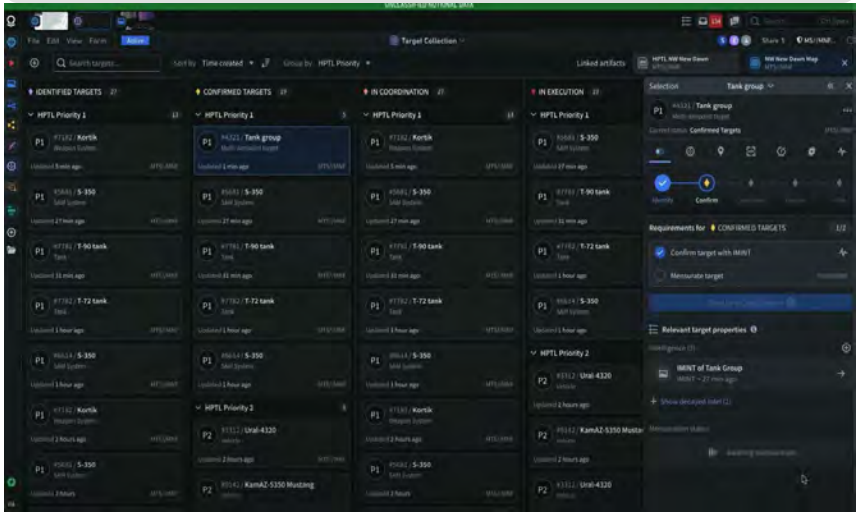
자료: MITCHELL INSTITUTE, 미래에셋증권 리서치센터

X. 팔란티어: 온톨로지+국방 데이터 = 戰場의 게임 체인저

온톨로지 기반의 AI는 킬체인 및 군수체계의 진전에 큰 역할

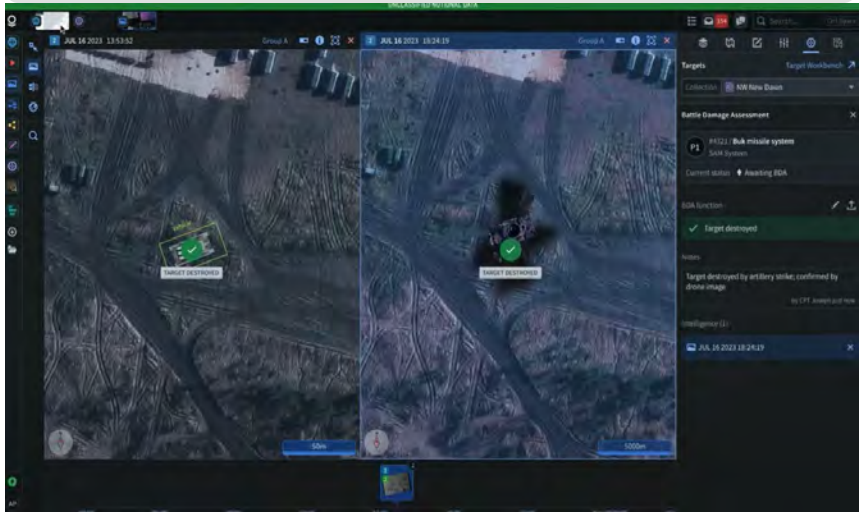
- AI는 선택 사항이 아닌, 전력의 효율을 극대화하여 현대전의 승패를 가르는 핵심 '전력 승수'로 작용. 팔란티어가 우크라이나군에 제공하는 AI 플랫폼은, 현대전의 최전선에서 AI가 어떻게 전투의 핵심적 역할을 수행하는지를 보여주는 가장 대표적인 사례.
- 팔란티어의 메이븐 스마트 시스템(MSS)은 미군의 차세대 AI 지휘 통제 플랫폼. 이 역시 팔란티어의 Foundry 플랫폼을 기반으로 개발돼, 온톨로지를 통해 데이터를 동적으로 구조화하고 연결하며 작동.
- MSS는 전쟁의 두 축인 군수(보급)와 전투를 모두 혁신. 군수 분야에서는 전 세계 보급망을 한눈에 파악하고 AI를 통해 향후 필요한 자원을 미리 예측하여, '요청 기반'의 수동적 보급을 '예측 기반'의 능동적 지원 체계로 전환. 전투에서는 수 시간이 걸리던 킬체인(표적 식별-결심-타격)을 단 몇 분으로 압축하며, 데이터 기반의 신속하고 정밀한 작전을 가능하게 하는 핵심 역할을 수행.

MSS의 전장에서 관측된 객체의 데이터화
전장에 있는 수많은 객체들을 "MSS"이라는 체스판에 올려 두는 것



자료: Palantir, 미래에셋증권 리서치센터

실시간 전장 정보 획득과 현황 판단이 가능
드론을 통한 전장 이미지 해석으로 파괴된 적기를 인식



자료: Palantir, 미래에셋증권 리서치센터

X. 팔란티어: 미국의 '온톨로지 기반 킬체인'에 연결되는 유럽

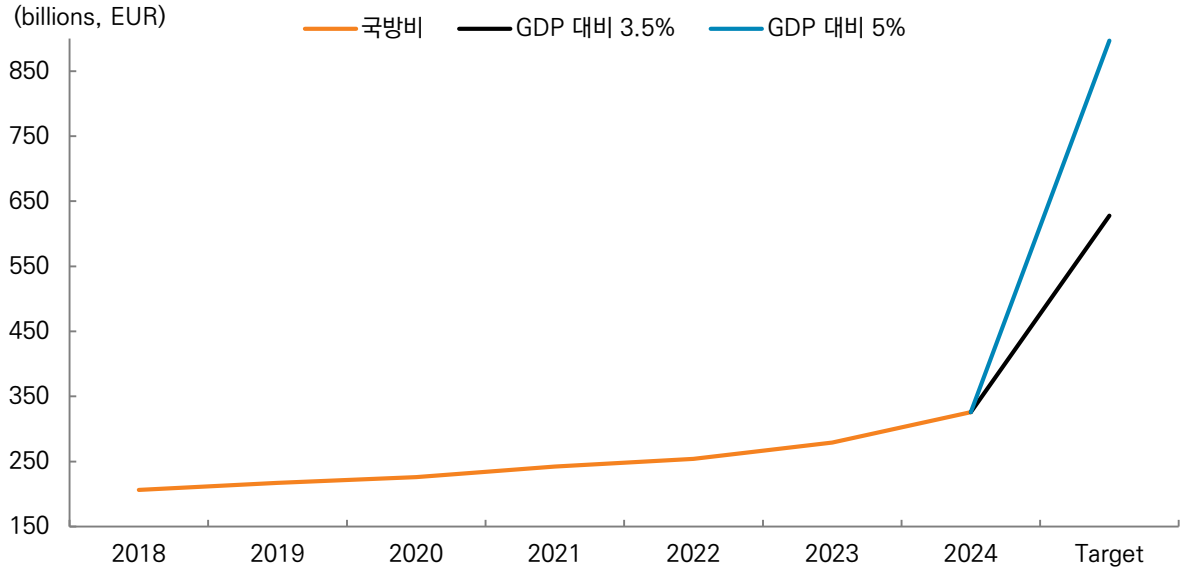
미국 뿐만 아니라 미국의 우방국 국방에도 온톨로지가 접목된다

- NATO는 팔란티어의 최첨단 AI 시스템인 "MSS(Maven Smart System) NATO"를 채택하며, 국방 AI의 도입을 본격적으로 알림. 2025년 3월 최종 체결된 이 계약은 미 국방부의 '프로젝트 메이븐'에서 파생된 기술을 NATO 연합작전사령부가 직접 운용하는 것으로, 현대전의 핵심이 데이터 분석 능력으로 이동했음을 상징적으로 보여주는 사건.
- 우크라이나 전쟁에서 실전 효용성이 입증된 데이터 기반 작전 수행 능력을 NATO 전체로 확장한다는 점에서 주목할 만. 이번 계약은 단 6개월 만에 체결될 만큼 이례적으로 신속하게 진행되었는데, 이는 AI 기술 경쟁에서 뒤처지지 않으려는 NATO의 절박함을 보여줌.
- 또한, 미국의 역할이 불확실한 상황에서 '기술을 통한 동맹 강화'라는 메시지를 전달하는 동시에, 유럽의 안보가 과거 전투기나 미사일 같은 미국산 하드웨어에서 이제는 AI 소프트웨어 기술에 의존하게 되었음을 보여주는 상징적인 변화.

NATO의 MSS 도입
유럽은 AI 전장 지휘 체계 도입을 시작

자료: FT, 미래에셋증권 리서치센터

EU 국방비 지출 추이 및 GDP 대비 지출 예상 금액



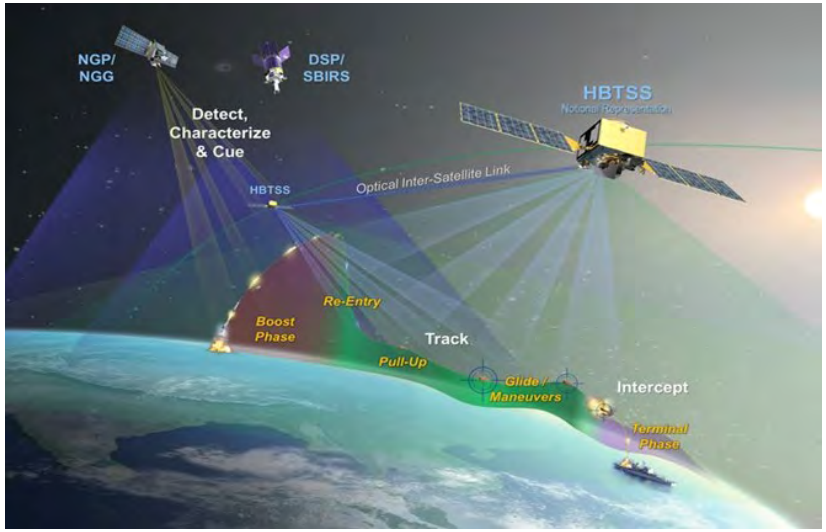
자료: European Defense Agency, 미래에셋증권 리서치센터

X. 팔란티어: 장기적 서방 진영의 핵심 과제 '골든돔'에도 온톨로지가…

골든돔에 쓰이는 방식도 핵심은 역시나 온톨로지

- 이스라엘의 '아이언돔'에서 이름을 따온 미국의 '골든돔' 프로젝트는 아이언돔과 달리 지상 방어 시스템이 아니라, 광활한 美 본토를 적대국의 초음속 미사일, ICBM과 같은 위협으로부터 방어하기 위한 "우주 기반 방어 체계" 기존 방어 시스템으로는 초음속 미사일을 중간에 요격하기가 거의 불가능에 가까움. 미사일이 발사되는 초기 단계(Boost Phase)에 우주에 있는 센서로 이를 즉시 탐지하고 요격해야 함. 이때의 센서는 저궤도 위성이 유일한 대안인 것.
- 골든돔 프로젝트는 '29년 초까지 실전 배치를 목표로 하며, 예산의 규모는 1,750억 달러로 계획. 이 방어 체계를 구축하기 위해 스페이스X와 안두릴, 그리고 팔란티어가 거론. 팔란티어는 위성과 지상의 센서 및 방어 시스템이 각자의 비구조화/이질적인 데이터를 쏟아낼 때 통합하는 SW 플랫폼을 제작. 데이터 분석을 기반으로 위협을 식별하고, 최적의 대응 방안을 도출하는 통합 의사결정을 가능하게 함.
- 한편, "스페이스X+팔란티어 컨소시움"은 구독형 모델을 정부에 제안. 전통 방산 계약과 달리 장기/안정적 수익을 창출하고, 기술 소유권을 유지해 미래 국방의 우선 지위를 확보하는 효과를 노림.

극초음속 및 탄도 미사일 추적 우주 위성의 지속적인 추적 및 핸드오프
다양한 위성관측 데이터의 유기적 연결을 지원하는 팔란티어



자료: Northrop Grumman, 미래에셋증권 리서치센터

골든돔 프로젝트에 참여하는 비전통적 방산 기업들로 이뤄진 드림팀

기업명	각자의 프로젝트에서의 역할
안두릴 (Anduril)	엣지 컴퓨팅 및 센서: 위성, 드론 등 엣지 디바이스에 탑재되는 고성능 센서 개발.
스페이스X (SpaceX)	하드웨어 인프라: 감시 및 요격 위성 등 HW을 제작하고 발사체 이용해 우주 궤도에 배치.
팔란티어 (Palantir)	통합 플랫폼: 여러 방어 시스템(위성, 센서, 지상 장비 등)의 데이터를 통합, 통합된 데이터를 기반으로 상황을 분석, 최적의 대응을 결정하는 중앙 지휘 통제 시스템.

자료: 미래에셋증권 리서치센터

X. 팔란티어: 최종 목표는 서방 진영을 다 아우르는 OS가 되는 것

시장 지배를 위한 날개: 제품 동맹(SAP)과 서비스 동맹(엑센추어)

- 팔란티어는 서방 진영의 시장 지배를 위해 유럽의 거인들(SAP와 엑센츄어)과 전략적 동맹을 체결.
- 먼저, 글로벌 기업의 99%가 ERP 시스템(재무, 생산, 물류, 재고 등)으로 사용하는 SAP 데이터(뇌)에 팔란티어의 AI(신경계)를 직접 연결. 팔란티어는 이제 훨씬 더 빠르고 깊게, 고객사의 가장 중요한 데이터에 접근할 수 있게 됨. 이는 경쟁자가 모방하기 극도로 힘든독점적 가치를 창출. 기업의 핵심 뇌(SAP 데이터)를 팔란티어의 온톨로지에 깊숙이 결합함으로써, 팔란티어를 단순 분석 툴이 아닌 대체 불가능한 '운영체제(OS)'의 일부로 만들기 때문.
- 엑센츄어는 미국 연방 정부 최대의 기술 컨설팅 기업으로서 이제 팔란티어 제품을 설치하고 교육하는 전담 부대가 됐음. 팔란티어의 약점이었던 '구현 인력 부족'을 단번에 해결하는 시장 확대 전략. 구체적으로는, 엑센추어의 연방 서비스 부서 내부에 1,000명 규모의 '팔란티어 인증 전문가 부대'를 창설. 이로써 팔란티어 입장에서는 대규모 정부 프로젝트 수주 및 동시 수행이 가능해짐.

팔란티어의 시장 지배력을 극대화하기 위한 상호 보완적인 두 가지 동맹

구분	w/ 엑센추어	w/ SAP
파트너십	서비스 파트너십	제품 파트너십
팔란티어가 얻을 것	**팔란티어 제품을 '구현, 설치, 교육'할 수 있는 TAM이 커짐**	**SAP 제품을 온톨로지에 융합하면서 엔터프라이즈 시장에 깊숙이 관여**
이들이 얻는 것	1,000명의 팔란티어 인증 전문가 부대를 창설해 역량 확보	막대한 분량의 온프레미스 기업용 코드를 팔란티어 도움 받아 클라우드로 전환
목표	시장 확대	제품 경쟁력 강화

자료: 미래에셋증권 리서치센터

팔란티어-엑센추어(AFS) 파트너십 잠재 매출 기여도 분석

분석 단계	핵심 가정 및 계산 근거	예상 매출 (팔란티어 몫)
1. TAM	엑센추어 연방 서비스(AFS) FY2023 매출: 약 40.8억 달러	공략 가능한 최대 시장 규모
2. 가정	**1) AFS 프로젝트의 팔란티어 솔루션 전환율 가정 2) 매출 비중 가정 (SW 몫은 37%, 서비스 몫이63%)**	**20% 전환 시: 약 3.0억 달러 40% 전환 시: 약 6.0억 달러 60% 전환 시: 약 9.1억 달러**
3. 기여도	팔란티어 2024년 전체 매출 : 약 28.7억 달러	20% 전환 시나리오 (3억 달러)는 2024년 매출 대비 약 10.5%에 해당

자료: 미래에셋증권 리서치센터

[참고] 팔란티어의 성장 로드맵: 서방 세계의 “AI OS”를 향한 여정

팔란티어가 단순한 소프트웨어 기업을 넘어, 미국과 서방 동맹의 안보와 번영을 위한 핵심 운영체제(OS)로 자리매김하는 전략적 비전

- 창업 초기에는 NSA, CIA 등 첩보/방첩 기관에 솔루션을 제공하는 비밀스러운 태동기를 보냈음.
- ChatGPT 이후 본격 성장기에 접어들며, 국방부, 국무부뿐만 아니라 NIH, NHS와 같은 민간/공공 분야로 고객을 확장하며 매출이 폭발적으로 증가.
- AI가 국가 경쟁력을 좌우하는 '전략물자'로 부상함에 따라, 미국 및 동맹국들이 팔란티어의 솔루션을 전방위적으로 채택하는 '소버린 AI' 시대로 진입.

현대의 국가 경쟁이 데이터 기반의 '효율성 전투'로 변모함에 따라, 팔란티어의 AI가 모든면의 효율을 극대화하는 핵심 '전력 승수'로 작용하기 때문.
이러한 흐름은 국방을 넘어 심지어 외교(미 국무부), 그리고 서방 국가들의 제조업 부흥 과제 해결까지 확장 중.

자료: 각 기관, 미래에셋증권 리서치센터

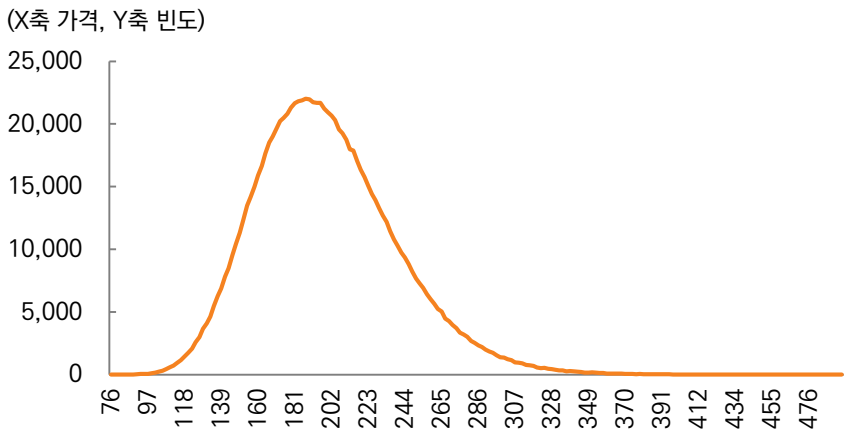
[참고] AI를 활용한 몬테카를로 주가 시뮬레이션

시뮬레이션 파라미터 요약

항목	내용
시뮬레이션 방법	몬테 카를로 시뮬레이션 (로그노멀 분포 기반)
시뮬레이션 횟수	1,000,000번 (무작위 시나리오 생성)
예측 기간	현재 날짜 기준 약 2026년 3월 8일경(일 단위 시뮬레이션 집계)
신뢰 수준	95% (시뮬레이션 분포의 2.5% ~ 97.5% 백분위수)
기반 데이터/변수	주가, 거래량, 매크로 요인, 기본적/기술적 지표 등
랜덤 시드	42 (np.random.seed 사용)
출력 형식	로그 수익률이 아닌 최종 주가값 (히스토그램 bins=200개)

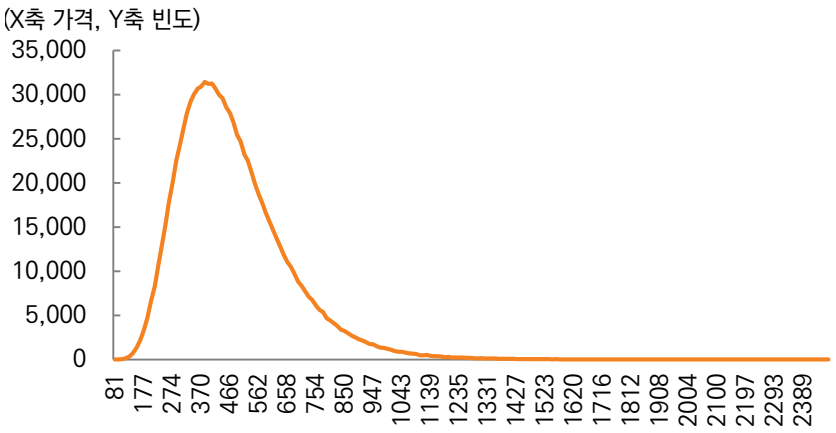
자료: Grok, 미래에셋증권 리서치센터

몬테카를로 기반 6개월 후 NVDA 주가 시뮬레이션 분포
frequency(빈도)가 가장 높은 구간은 약 186~188달러 범위



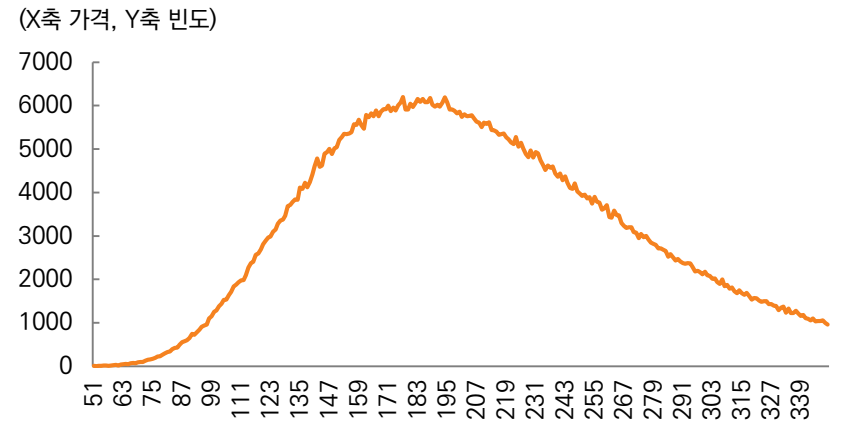
자료: Grok, 미래에셋증권 리서치센터

몬테카를로 기반 6개월 후 TSLA 주가 시뮬레이션 분포
frequency(빈도)가 가장 높은 구간은 약 376~388달러 범위



자료: Grok, 미래에셋증권 리서치센터

몬테카를로 기반 6개월 후 PLTR 주가 시뮬레이션 분포
frequency(빈도)가 가장 높은 구간은 약 165~170달러 범위



자료: Grok, 미래에셋증권 리서치센터

[결론] AI 혁명의 진짜 승부처

AI 혁명은 물리적 자원의 전쟁입니다. 과거 100년의 기술 발전이 10년마다 압축되는 이 경쟁의 승패는 '누가 더 많은 컴퓨팅 파워(GPU)와 전력(Power)을 확보하는가'에 의해 결정됩니다. 디지털 지능의 가치는 인프라를 통제하고 현실 세계와 연결하는 소수에게 수렴하게 될 것입니다.

1. 현황: 기업들은 '소수 정예 인간 + 다수 AI 에이전트' 구조로 전환 중입니다. 이를 위한 AI 경쟁은 모델 크기 경쟁에서 '강화학습 스케일링(Test-Time Scaling)'으로 넘어갔습니다. 이것이 새로운 표준이 되면서 GPU 수요는 구조적으로 폭증하게 됩니다. 그러나, 문제는 평균 4년을 초과하는 AI 데이터센터의 전력 연결 대기 시간입니다. 전력망의 한계가 AI 혁신의 속도를 결정하는 제 1의 물리적 제약입니다.

2. 전망: 구글, OpenAI, xAI, 메타 등 소수 기업이 연간 수백조원을 쏟아붓는 자본 집약적 전쟁입니다. 구글은 자체 칩(TPU)부터 클라우드, OS까지 모든 것을 수직 계열화하여 비용과 효율에서 상당한 우위를 점하고 있습니다. 작은 플레이어는 경쟁이 극히 어려운 구조로 수렴하고 있습니다. 미국은 '페타플롭스 동맹'으로 중국을 고립시키려 합니다. 중국은 '오픈소스'와 '압도적 전력 인프라'를 무기로 반격하고 있습니다. 특히 화웨이를 중심으로 한 자립화 정책(Ascend, CloudMatrix 등)의 마일스톤 성과들이 앞으로의 관전 포인트입니다.

3. AI 기술의 진화는 가치 사슬 전반에 걸쳐 새로운 승자를 만들어냅니다. 투자는 이 3개의 계층에 집중되어야 한다고 판단됩니다.

① 물리적 기반: 즉각적인 병목 현상이 발생하는 영역입니다. AI 혁명의 모든 비용이 우선 이곳으로 흘러 들어오게 됩니다.
- 컴퓨팅 & 네트워킹: 독점적 SW 생태계(CUDA)를 구축한 엔비디아의 해자가 높아 보입니다. 네트워킹 스케일링의 모든 영역을 지배하려는 야심으로 시스템 표준을 정의하고 있습니다.
- 전력 인프라 & 데이터센터: 그리드 연결 지연은 현장 발전 수요를 폭발시킵니다. 가스 터빈, 연료전지 등 독립 전원 솔루션과 차세대 냉각 시스템, 데이터센터 건설 및 전력 설비 기업이 구조적 수혜를 입을 것으로 사료됩니다.

② 운영 시스템: 강력하지만 혼란스러운 AI 모델을 실제 비즈니스 가치로 전환하는 'OS' 레이어에 주목해야 합니다.
- 엔터프라이즈 AI 플랫폼: AI가 저렴 해져도 기업의 복잡한 데이터는 여전히 고유합니다. 이 둘을 연결하여 실제 운영을 자동화하는 팔란티어의 '운영체제' 역할이 부상할 것입니다. 국가 안보, 제조업, 헬스케어 등 핵심 산업에 이미 깊숙이 침투한 상황입니다.

③ 물리적 애플리케이션: AI가 디지털 세계를 넘어 현실과 상호작용하며 막대한 신규 시장을 창출하는 최종 단계이자 막대한 TAM을 지닌 부분입니다.
- 자율주행: AI 에이전트가 탑재될 파괴적인 하드웨어 플랫폼입니다. 데이터 수집, 칩 설계, 차량 생산까지 수직 계열화하여 괄목할만한 확장성을 증명한 테슬라가 이 영역을 주도할 것으로 보입니다. 중국 업체들이 팔로잉을 시작한 테슬라의 카메라 기반 범용 지능은 전 세계 도로를 공략하고 있습니다.
- 월드 시뮬레이션: AI가 스스로 학습하고 실험할 수 있는 가상 세계를 창조하는 기술입니다. 현실의 물리 법칙을 이해하고, 무한한 시뮬레이션을 통해 AI를 훈련시키는 플랫폼입니다. 이는 구글의 Genie 3가 가장 근접해 있습니다. 로보틱스의 핵심 기반이 될 수 있어 구조적인 수혜가 예상됩니다.

Compliance

- 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.