Market Insight | 2025.11.13

그림: Midjourney('하드웨어 제약이 창의적 혁신을 낳는다'를 동양 미학으로 상징)

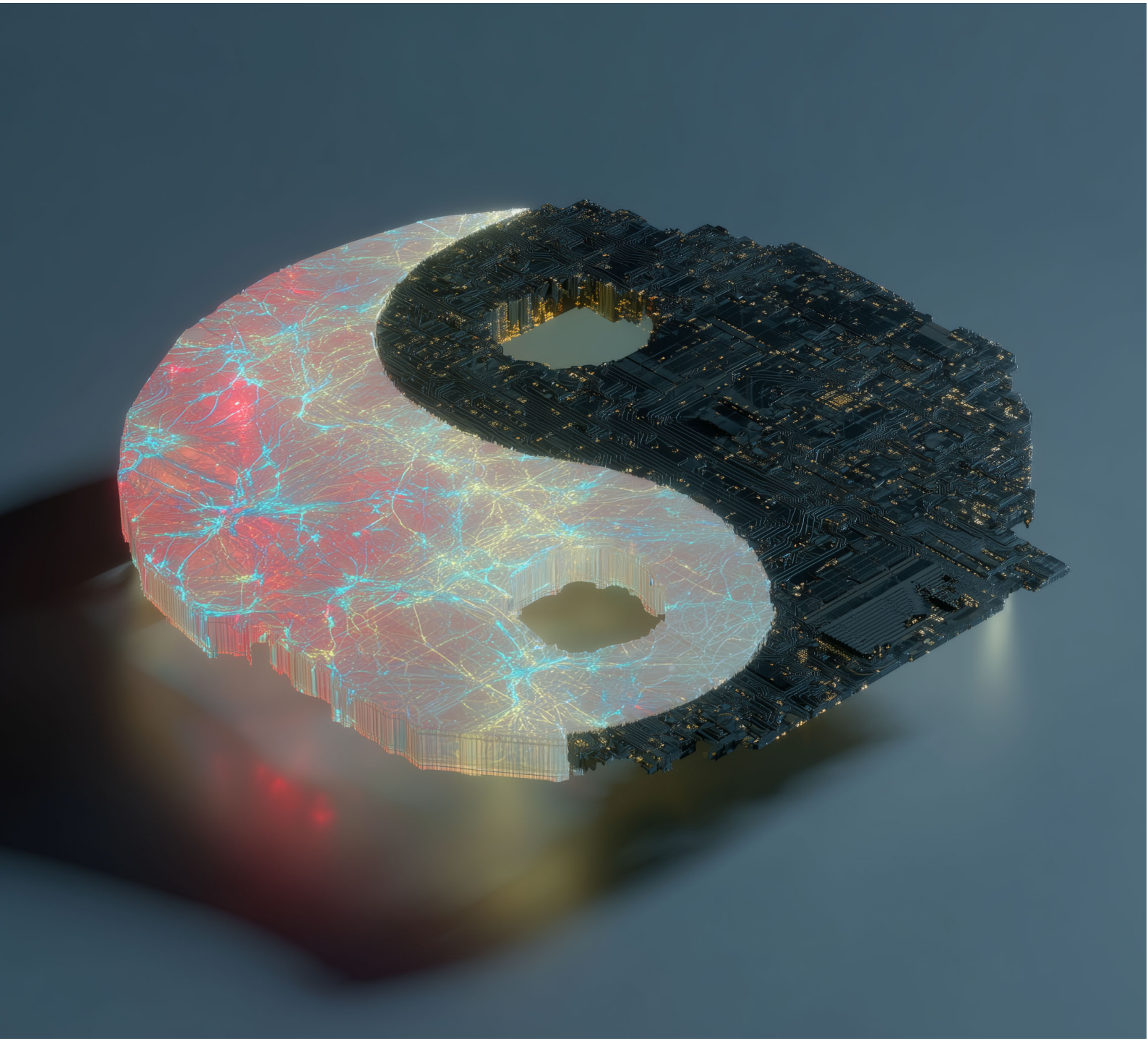
MIRAE ASSET 미래에셋증권

AI Bi-weekly

AI 지출은 거대한 사기인가, 대담한 베팅인가

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com



CONTENTS

Highlight of the Week 3
병목이 미래를 결정한다 3
AI에게 해마가 생긴다면 4

I. AI 버블과 자본지출 논쟁 5
1. 마이클 버리의 "SELL" 5
2. 현장에서의 반박과 AI 낙관론 8
3. AI 데이터센터의 운영 현실과 물리적 병목 13
4. 종합 분석 및 전망 16

II. 키미노 나마에와... 테슬라? 18
1. Kimi K2 Thinking: 중국 AI가 만들어낸 또 다른 특이점 18
2. 기술적 해부: 제약과 혁신이 빚어낸 아키텍처 20
3. 지정학적 AI 패권 경쟁의 판도를 바꾸는 또 하나의 게임 체인저? 22
4. Kimi K2의 전략적 승부수: 네이티브 INT4 양자화 심층 해설 24
5. 기술적 수렴 진화: 테슬라와 Kimi의 평행 이론 28

III. AI Agent에게 진짜 필요한 것은 31
1. 인간과 AI를 가르는 것들 31
2. 현실적 과제들 33

IV. AI Peer Table 37

V. Appendix: Charts 39

Highlight of the Week

병목이 미래를 결정한다

AI 산업의 미래를 두고 천문학적 자본 투자가 '버블'인지 '필연'인지를 가르는 격렬한 논쟁이 벌어지고 있습니다. 마이클 버리는 주요 테크 기업들이 AI 투자 본격화 시점에 맞춰 서버 자산의 내용연수를 인위적으로 늘려 감가상각비를 축소하고 있다고 지적합니다. 이는 단기 이익을 부풀리는 '회계 부정'이며, AI 수요가 꺾일 경우 GPU 자산 가치가 폭락하며 2007년 서브프라임 사태와 같은 시스템적 위기가 올 수 있다는 것이 그의 경고입니다.

하지만 시장의 현실 데이터는 이 비관론을 정면으로 반박합니다. AI 컴퓨팅에 대한 폭발적이고 구조적인 수요 덕분에, 8년 된 구글의 구형 TPU가 여전히 100% 가동률을 보이고 3년차에 접어든 H100 GPU가 출시 가격의 95%에 재계약되는 등 자산의 실제 경제적 수명이 과거의 통념을 훨씬 뛰어넘고 있습니다.

이 팽팽한 논쟁에 균형을 잡아주는 것이 바로 물리적 '병목' 현상입니다. 현재 AI 인프라 확장의 진짜 한계는 '파워쉘', 즉 대형 변압기, 냉각 설비, 전문 인력 등 데이터센터를 구축하는 물리적 인프라 전체의 절대적 부족입니다. 이 병목은 무분별한 과잉 공급을 막는 '자연적 속도 제한 장치' 역할을 하여 버블 붕괴 위험을 완화하고, 동시에 CoreWeave처럼 선제적으로 인프라를 확보한 기업에게는 강력한 경제적 해자를 만들어줍니다.

이러한 거시적 병목은 기술 혁신의 방향까지 결정하고 있습니다. 서로 다른 목표를 가진 중국의 Kimi K2와 미국의 테슬라는 각자가 직면한 하드웨어의 한계를 극복하기 위해 놀랍도록 동일한 공학적 해법에 도달했습니다.

Kimi K2는 미국의 수출 통제로 인한 저사양 GPU 환경에서, 테슬라는 차량 로봇의 실시간성, 저전력, 안정성이라는 물리적 제약 속에서 '메모리 대역폭'이라는 동일한 기술적 병목에 부딪혔습니다. 두 기업 모두 이 문제를 해결하기 위해 '양자화 인식 훈련(QAT)'이라는 정교한 기술을 핵심 전략으로 채택했습니다. 이는 AI 모델을 훈련하는 단계부터 최종 탑재될 저정밀 하드웨어의 제약 조건을 미리 시뮬레이션하여, 성능 저하 없이 최적의 효율을 내도록 모델 자체를 '단련'시키는 혁신적인 방식입니다.

결론적으로 AI 시대의 미래는 단순히 자본의 규모나 모델의 크기로 결정되지 않습니다. 시장의 과열을 제어하는 거시적 물리 병목과, 하드웨어의 한계를 돌파하게 만드는 미시적 기술 병목이 산업의 구조와 기술의 진화를 동시에 이끌고 있습니다. 이처럼 피할 수 없는 제약을 기회로 전환하고, 그 본질을 꿰뚫어 가장 창의적인 공학적 해법을 제시하는 자가 AI 시대의 진정한 승자가 될 것입니다.

AI에게 해마가 생긴다면

인간의 뇌는 매일 엄청난 양의 데이터를 처리합니다. 하지만 우리는 모든 정보를 기억하지 않습니다. 대신 해마를 통해 중요한 신호만 선택적으로 저장하고, 수면 중에 이를 장기기억으로 압축합니다. 이 생물학적 기제는 제한된 자원으로 강력한 학습 능력을 발휘하는 핵심입니다. 그러나 현재의 대형 언어 모델(LLM)에는 이러한 메커니즘이 없습니다. 모든 정보를 동등하게 처리하기 때문에 신호와 노이즈를 구분하지 못합니다. 더 심각한 것은 파괴적 망각 문제입니다. 새로운 분야의 데이터로 학습하면 기존 능력이 손상됩니다. 의료 데이터로 파인튜닝하면 일반 언어 이해 능력이 저하되는 식입니다.

이를 해결하기 위해 최근 연구는 강화학습 대신 증류를 제시합니다. 칭화대 LeapLab의 연구 결과, 강화학습은 단순히 샘플링 효율만 개선할 뿐 진정한 추론 능력을 확장하지 못합니다. 반면 On-Policy 증류는 교사 모델이 학생 모델의 추론 과정 각 단계에서 피드백을 제공함으로써, 진정한 의미의 새로운 추론 패턴을 전수합니다. 실제로 540억 파라미터 대형 모델과 비교해 7,000배 작은 증류 모델이 더 우수한 성능을 달성했습니다.

그렇다면 더 큰 모델이 필요 없을까요? 그렇지 않습니다. 우수한 학생 모델을 만들기 위해서는 더욱 뛰어난 교사 모델이 필요합니다. 이를 위한 다음 단계는 고품질 데이터와 효율적인 학습 구조입니다.

현재 텍스트 데이터 수집은 포화 상태입니다. 이를 극복하려면 영상과 같은 모달리티 확장이 필수입니다. 영상 데이터는 단순히 데이터 양을 늘리는 것이 아니라, 시간에 따른 인과관계를 학습하게 함으로써 AI의 상식 기반을 강화합니다.

그러나 데이터만으로는 부족합니다. 진정한 에이전트 시대를 열기 위해서는 세 가지가 함께 필요합니다. 첫째, 증류를 통한 진정한 추론 능력의 습득입니다. 둘째, 연속 학습을 통한 경험의 축적입니다. 사용자 개인의 데이터로 실시간 학습하면서 개인화되어야 합니다. 셋째, 계층형 메모리 아키텍처입니다. 인간의 해마처럼, 최근 상호작용은 빠르게 접근할 수 있는 메모리에, 반복되는 패턴은 중간 계층에, 기초 데이터는 하위 계층에 저장해야 합니다.

현재 AI는 거대한 데이터 저수지 위에 세워져 있습니다. 하지만 그 모델들이 물을 마실 수 있는 구조가 부족합니다. 인간의 뇌처럼 경험을 지식으로 변환하고, 패턴을 이해하며, 새로운 상황에 대응할 수 있는 아키텍처가 필요합니다. 이러한 구조적 혁신이 일어날 때, AI는 비로소 상식을 기반으로 인간과 소통하는 진정한 지능을 갖추게 될 것입니다.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
9	10	11	12	13	14	15
	· CRWV 실적	· 소프트뱅크 실적 · 위스트론 실적	· 폭스콘 실적	· SMIC 실적	· 크래프톤 실적(예)?	
16	17	18	19	20	21	22
	· HD현대 실적(예)?			· NVDA 실적		
23	24	25	26	27	28	29
		· PONY 실적 · BABA 실적				

자료: Bloomberg, 미래에셋증권 리서치센터

I. AI 버블과 자본지출 논쟁

1. 마이클 버리의 "SELL"

(1) 논쟁의 핵, 인위적으로 조작된 감가상각

AI 산업은 전례 없는 인프라 구축 국면에 진입했다. 그리고 막대한 자본은 AI 하드웨어에 집중된다. 미국의 하이퍼스케일러인 주요 5대 기업은 2025년에만 3,420억 달러를 자본지출(Capex)에 투입할 전망이다. 이는 아폴로 달 착륙 프로그램의 총비용을 넘어서는 규모다. 이 투자의 중심에는 엔비디아 GPU와 CoreWeave 같은 전문 클라우드 기업이 있다.

이 배경 속에서 영화 "빅쇼트"의 실제 인물 마이클 버리는 중대한 경고를 제기했다. 버리는 하이퍼스케일러와 AI 인프라 기업들이 자산의 내용연수를 인위적으로 연장해 감가상각비를 과소 계상하고 있다고 주장한다. 그는 이를 현대 시대의 흔한 회계 부정(fraud) 중 하나로 규정했다. 이 주장은 AI 산업의 수익성과 가치평가 모델 전체에 근본적인 의문을 던진다.

(2) 버리의 핵심 논지: 가속화되는 기술과 회계의 불일치

마이클 버리의 AI 버블론은 자산 감가상각의 인위적 조작 가능성에 기반한다. 그의 논리는 AI 산업의 기술적 현실과 회계 처리 사이의 명백한 모순을 지적한다. 엔비디아는 GPU 출시 주기를 기존 2년에서 1년으로 단축하며 기술적 진부화를 가속하고 있다.

그러나 이러한 환경에서 AI 인프라 기업들은 오히려 컴퓨팅 장비의 내용연수를 연장하며 경제적 현실과 명백히 배치된다는 것이다. 내용연수 연장은 연간 감가상각비를 줄여 회계상 비용을 축소하고, 이는 단기 순이익을 부풀리는 효과를 낳는다. 버리는 이 행위가 실제 현금 흐름과 무관하게 장부상 이익을 조작하는 것이라 본다.

버리는 구체적 숫자로 주장을 뒷받침한다. 그는 하이퍼스케일러들이 2026년부터 2028년까지 3년간 감가상각비를 총 1,760억 달러 과소 계상할 것으로 예측했다. 이 “회계적 조정” 덕에, 2028년 Oracle의 이익은 무려 26.9%, Meta의 이익은 20.8% 과대평가될 것이라 추정했다.

그림 1. 주요 기업별 내용연수 변경 추이

Network/Compute Depreciation Useful Life (Years)

Company	2020	2021	2022	2023	2024	2025
META	3	4	4½	4½	4½	5½
GOOG	3	4	4	6	6	6
ORCL	5	5	5	5	6	6
MSFT	3	4	6	6	6	6
AMZN	4	4	5	5	6	5

자료: 각 회사 SEC 공시, 마이클 버리, 미래에셋증권 리서치센터

이처럼, 버리의 주장은 개별 기업의 일탈이 아닌 산업 전반의 시스템적 현상이기에 설득력을 얻는다. 주요 테크 기업들은 AI 투자가 본격화된 시점을 기점으로 일제히, 그리고 "조용하게" 서버 자산의 내용연수를 상향 조정해왔기 때문이다.

이 변경은 과거 1~5년의 짧은 기간을 적용했던 것과 뚜렷한 대조를 이룬다. 이 사소해 보이는 회계 변경은 수조 원 규모의 재무적 영향을 미친다. Meta는 내용연수를 5.5년으로 연장하며 2025년 한 해 감가상각 비용이 약 29억 달러 감소할 것으로 예상했다. Intel 역시 2023년 생산 장비의 내용연수를 5년에서 8년으로 연장했다.

(3) 최전선의 플레이어, CoreWeave의 회계 리스크

버리의 논리는 CoreWeave와 같은 전문 AI 클라우드 제공업체에 가장 날카롭게 적용된다. 이들의 사업 모델은 최신 GPU 임대에 전적으로 의존하기 때문이다.

경쟁사와의 극명한 회계 차이

동일한 사업 모델과 장비를 사용하는 경쟁사 Nebius와의 비교는 문제를 명확히 보여준다. GPU 내용연수 설정에서 CoreWeave는 6년을 적용한 반면 Nebius는 4년을 적용했다. 데이터센터 장비 내용연수 역시 CoreWeave는 8~12년, Nebius는 3~10년으로 상당한 차이를 보인다.

표 2. CoreWeave와 Nebius의 자산 내용연수

항목	분석
감가상각 전략	CoreWeave가 대체로 더 긴 내용연수 적용 (보수적)
데이터센터 장비	CoreWeave 8-12년 vs Nebius 3-10년 (CoreWeave가 최대 2배 길게 설정)
서버/네트워크	CoreWeave 6년 vs Nebius 4년 (CoreWeave가 50% 더 길게)
건물 자산	Nebius만 보유 (20년)
회계 보수성	**CoreWeave가 더 긴 내용연수 = 연간 감가상각비 부담 낮음**

자료: 각 회사 SEC 공시, 미래에셋증권 리서치센터

사업 모델의 본질적 취약성

하이퍼스케일러는 구형 GPU를 내부 워크로드에 재활용해 경제적 가치를 연장할 수 있다. 그러나 CoreWeave의 수익은 오직 외부 고객의 선택에 달려있다. 고객은 항상 최고의 성능을 내는 최신 칩을 선호할 것이므로, GPU가 임대 수익을 극대화하는 실질적 기간은 12~24개월에 불과할 수 있다는 견해가 있다. 이런 사업 모델에서 6년의 내용연수는 매우 공격적인 가정이다.

비판론자들은 이 회계 처리가 연간 비용을 낮춰 단기 이익을 높이고, 한때 800억 달러에 도달했던 기업가치를 애써 정당화하려는 의도라고 분석한다. CoreWeave의 2025년 3분기 감가상각비는 6억 3,047만 9천 달러로, 연간 약 25억 달러 규모다. 그러나 이 숫자는 실제 경제적 가치 하락이 아닌, 처음부터 길게 설정된 인위적 기간에 맞춰 계산된 결과물이라는 점이 문제의 핵심이다.

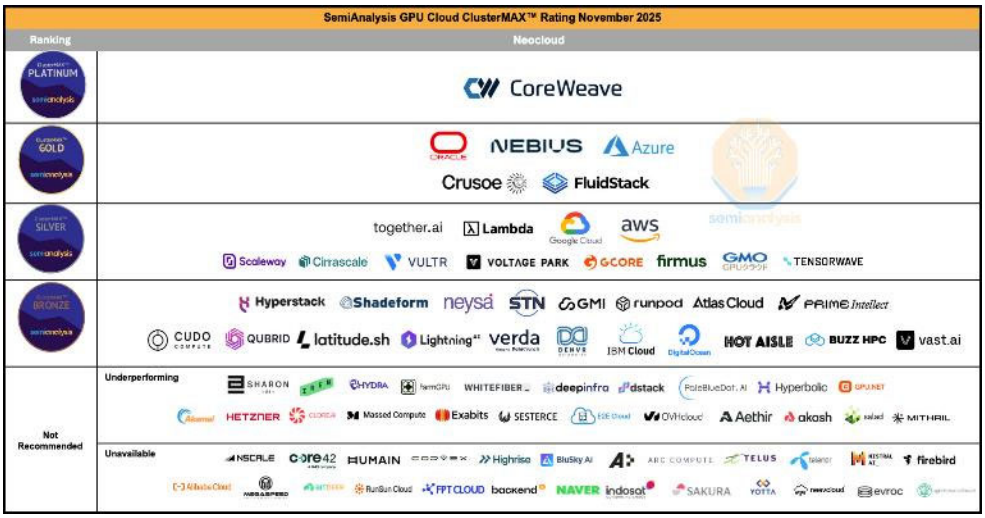
(4) 2007년 서브프라임 사태의 재림?

마이클 버리의 비판은 AI 산업 전체의 지속 가능성에 대한 거시적 우려로 확장된다. 최근 J.P. Morgan의 분석에 따르면, 2030년까지의 AI 투자 규모에 대해 연 10% 수익률을 달성하려면 연간 약 6,500억 달러의 수익이 영구적으로 발생해야 한다고 한다. 이는 전 세계 모든 아이폰 사용자로부터 매월 34.7달러를 거둬야 하는 규모다. 이는 과거 통신 및 광섬유망 버블 붕괴를 연상시킨다.

심지어, 한 하드웨어 업계 관계자는 현재 상황을 서브프라임 사태에 비유한다. 은행과 VC는 GPU를 담보로 대출(DDTL)을 집행하며 '고객 계약은 견실하고, 유사시 하드웨어를 재판매하면 된다'고 가정한다. 이는 '주택담보대출은 안전하고, 유사시 주택을 처분하면 된다'고 믿었던 2007년 은행들과 근본적으로 같다는 논리다. 하지만 AI 시장이 위축되어 대량의 중고 GPU가 시장에 쏟아지면, 주택 시장 붕괴처럼 가격이 폭락할 것이므로 'GPU를 나중에 팔면 된다'는 생각은 환상일 수 있다.

나아가 SemiAnalysis가 발표하는 "ClusterMAX™" 벤치마크를 보면, 벌써 209개의 클라우드 제공업체를 추적할 정도로 시장은 이미 포화 상태에 가까우며 '사용자보다 하드웨어가 더 많다'는 지적마저 나온다. 유휴 하드웨어는 수익 없이 감가상각 비용만 발생시키고, 치열한 경쟁은 가격 인상을 억제해 마진을 잠식한다.

그림 2. SemiAnalysis가 만든 세계 최초의 GPU 클라우드 평가 시스템 ClusterMAX™
11월 5일 2.0 버전으로 업데이트되었으며 84개 GPU 클라우드 제공업체를 평가
COreWeave가 유일하게 최고등급인 플래티넘을 획득



자료: SemiAnalysis, 미래에셋증권 리서치센터

결론적으로 버리와 같은 비관론자들의 관점에서 현재 AI 산업은 기술 발전의 현실을 무시한 채, 공격적인 회계 처리와 막대한 자본 투입으로 부풀려진 거품 위에 서 있다. 이 버블은 수익 실현의 불확실성과 잠재적인 자산 가치 붕괴라는 시스템적 위험에 노출되어 있다.

2. 현장에서의 반박과 AI 낙관론

마이클 버리의 비판은 AI 하드웨어의 빠른 "낙후화"를 전제로 한다. 하지만, 시장의 현실 데이터와 업계 리더들의 발언은 그의 비관론을 정면으로 반박한다. 낙관론의 핵심은 AI 컴퓨팅 수요가 단기적 유행이 아닌, 문명사적 전환에 가까운 구조적이고 폭발적인 변화라는 인식에 기반한다.

한 AWS 엔지니어는 "비영리 단체부터 공공 부문, 상업, 소매, 항공, 자동차에 이르기까지 모든 단일 산업이 AI 툴링, 기능, 신제품을 찾고 있다"고 증언했다. AI를 향한 수요가 산업 전반에 걸쳐 얼마나 현실적이고 거대한지를 보여준다. 이 거대한 수요는 구형 자산의 가치를 예상보다 훨씬 길게 유지시키고 있으며, 현재의 막대한 자본지출은 미래의 수조 달러 시장을 선점하기 위한 필연적 투자라는 주장이다.

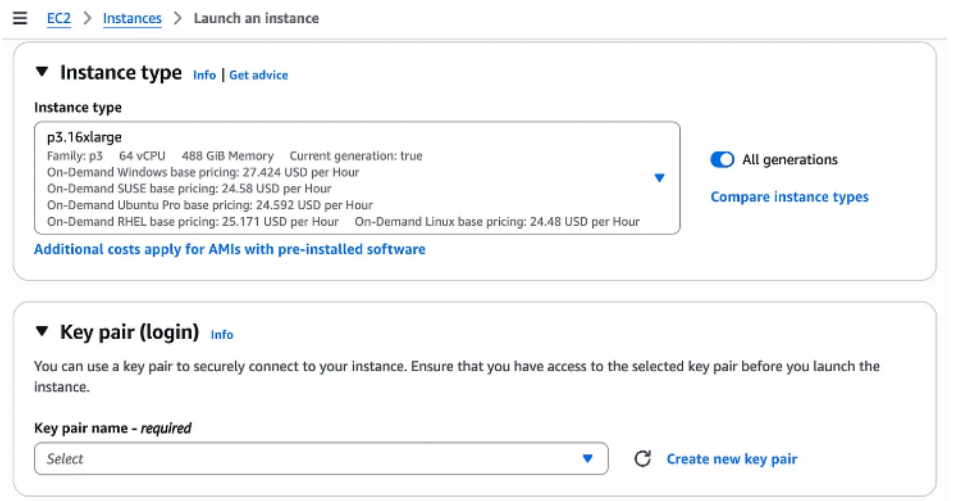
(1) 핵심 반박: 노후 하드웨어의 지속되는 경제적 가치

버리의 주장은 '최신 칩이 나오면 구형 칩은 가치를 잃는다'는 가정에 서 있지만, 실제 시장 데이터는 이 가정이 틀렸음을 증명한다. 구형 AI 하드웨어는 폐기되는 것이 아니라, 여전히 높은 가동률과 가격을 유지하며 경제적 생명을 이어가고 있다. 버리의 분석은 회계 기술적으로는 옳을 수 있으나, 경제적으로는 틀렸다.

2023년 말에 H100을 구매한 기업들은 이미 투자 원금을 회수했거나 곧 회수할 예정이며, 총투자수익률(Total ROI)은 상당히 높을 것으로 전망된다. 영업이익의 대부분이 초기 3년 이내에 발생하므로, 6년 직선 감가상각 방식은 회계상 1~3년차의 이익을 과대평가하고 4~6년차의 이익을 과소평가하는 경향을 만든다.

그러나 이는 실제 경제적 현실과 회계상 타이밍의 차이일 뿐 사기를 의미하지 않는다. 가장 중요한 것은, 6년의 감가상각이 끝난 후에도 GPU가 계속 이익을 창출하는 것이 충분히 가능하다는 것이다.

그림 3. 이번 달에도 AWS에서는 8개의 V100 GPU가 탑재된 'p3.16xlarge' 인스턴스를 판매 중! 2017년 5월에 발표된 V100이 거의 8년이 지난 현재까지 클라우드에서 현역으로 서비스되고 있다는 사실 자체가 기술 자산의 경이적인 수명을 보여줌



자료: AWS, 미래에셋증권 리서치센터

Google Cloud AI 인프라 부사장 Amin Vahdat의 발언은 이 논쟁에 결정적인 데이터를 제공한다. 그는 7년에서 8년 된 구형 TPU(구글의 자체 AI 가속기)가 여전히 100% 가동률로 운영되고 있다고 밝혔다. 이는 AI 워크로드에 대한 수요가 너무나 강력하여 구형 세대 칩까지 모두 가동해야 할 정도로 공급이 부족함을 의미한다. 또한, AI 하드웨어의 유효 수명이 회계 장부상 내용연수인 5~6년을 훨씬 초과할 수 있음을 보여주는 강력한 증거다.

실제로 출시된지 약 8년이 된 엔비디아의 V100으로 구동되는 인스턴스(클라우드 컴퓨팅에서 서버나 가상 머신을 빌려 쓰는 개념)는 이제서야 단종되고 있다. 더욱이 약 5년 된 A100 GPU는 당연히도 현역이다. A100은 주요 하이퍼스케일러들이 상당히 괜찮은 가격에 제공하고 있다.

심지어, 아래 표를 보면, AWS의 GPU 시간당 현물 가격은 이번 4분기에 H100이 전분기 대비 8% 상승하고, H200은 무려 18% 상승할 것으로 예측된다. 버블 붕괴는 가격 하락을 동반하지만, 현재 AI GPU 시장은 정반대의 현상을 보이고 있으며 이는 수요 증가세가 신규 공급량을 능가하고 있음을 의미한다.

표 3. AWS GPU별 (임대) 스팟 가격: 수요 증가세가 신규 공급량을 능가하고 있음을 의미

AWS GPU	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25(예상)	QoQ	YoY
H100 (P5)	$8.49	$5.87	$4.62	$3.29	$3.85	$3.47	$2.53	$2.72	8%	-17%
A100 (P4d)	$0.97	$1.45	$1.89	$2.49	$2.03	$1.27	$1.01	$1.02	1%	-59%
H200 (P5en)	na	na	na	$1.61	$2.37	$2.15	$2.21	$2.49	13%	55%
L40S (G6e)	na	na	$0.76	$0.85	$1.22	$1.20	$1.48	$1.39	-6%	64%
A10 (G5)	$0.60	$0.68	$0.73	$0.71	$0.70	$0.69	$0.84	$0.90	7%	26%
T4 (G4 & G5g)	$0.36	$0.41	$0.45	$0.47	$0.48	$0.45	$0.49	$0.51	3%	7%
Inferentia 1&2	$0.15	$0.18	$0.23	$0.23	$0.22	$0.20	$0.25	$0.25	3%	11%
Trainium1	$0.29	$0.31	$0.38	$0.29	$0.40	$0.27	$0.28	$0.33	20%	13%
B200 (P6)	na	na	na	na	na	na	$3.01	$3.08	2%	na

자료: Web Scarped Data, yipit DATA, 미래에셋증권 리서치센터

그리고 하이퍼스케일러들이 V100, A100을 철거하는 사례는 자산의 노후화 때문이 아니다. 한정된 공간과 전력 내에서 더 높은 수익을 내는 최신 자산으로 교체하는 경제적 의사결정이다. 게다가, 2018년 가동을 시작한 IBM의 Summit 슈퍼컴퓨터는 6.5년간 운영되었고, 2016년 설치된 중국의 Sunway TaihuLight는 8년이 지난 지금도 가동 중이라는 것도 마이클 버리의 가정을 뒤흔들 사실적 근거다.

나아가, CoreWeave의 CEO Michael Intrator는 어찌보면 가장 강력한 상업적 증거를 제시했다. 계약 기간이 만료된 대규모 H100 클러스터를 고객과 임대 재계약하는 과정에서, 기존 계약 가격 대비 불과 5% 낮은 가격으로 갱신하는 데 성공했다는 것이다. 다시 말해, 2022년 출시된 H100이 2025년 말 시점에도 출시 당시와 거의 동일한 상업적 가치를 인정받고 있다는 사실이다.

이는 마이클 버리를 포함한 회의론자들이 주장하는 '빠른 가치 하락'이 현실에서는 나타나지 않고 있음을 의미한다. Intrator CEO는 A100, L40S, H100, H200 등 모든 세대의 GPU가 시장에서 완전히 매진 상태이며 "계약이 만료되면 즉시 다시 팔려나간다"고 강조했다. 이는 시장에 유휴 장비가 거의 없으며, 공급되는 모든 컴퓨팅 파워가 즉각적으로 수요에 흡수되고 있음을 보여준다.

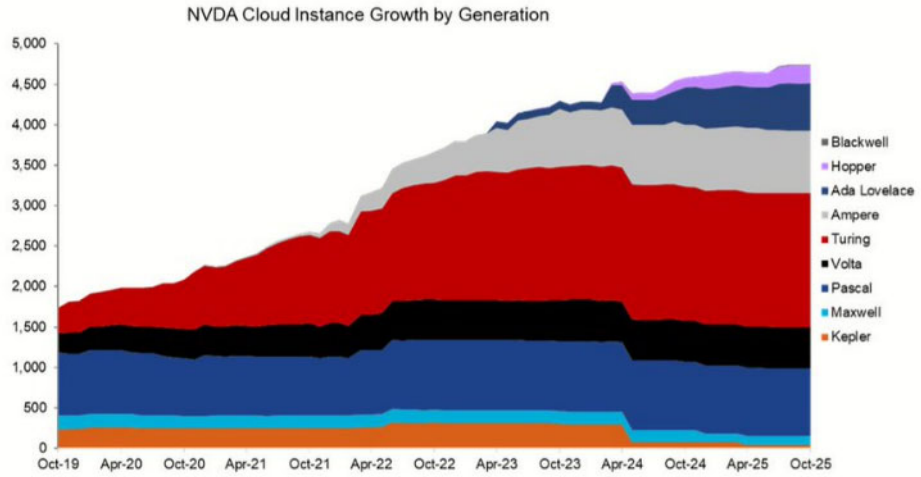
(2) GPU의 다단계 생애주기와 CoreWeave의 실적

낙관론자들은 GPU를 단일 용도 소모품이 아닌, 시간이 지남에 따라 여러 고부가가치 작업을 수행하는 다목적 설비로 본다. 이 '다단계 생애주기' 모델은 5~6년의 감가상각 기간이 합리적임을 설명한다.

출시 초기 0~18개월 동안 최신 GPU는 OpenAI 같은 기업의 차세대 LLM 훈련에 투입되어 최고의 가치를 창출한다. 18개월 이후 신제품이 출시되면, 기존 세대 GPU는 소규모 모델 훈련이나 추론 서비스에 활용된다.

시간이 더 지나면 이 GPU들은 대규모 추론 서비스 시장의 주력 자산이 되어 '롱테일(long-tail)' 수요를 담당한다. 이러한 현실에 따르면 감가상각은 실제 자산의 물리적 수명이나 경제적 유용성이 아닌, 회계상 비용 배분 스케줄에 불과하다.

그림 4. Azure, GCP 및 AWS에서 공개적으로 사용할 수 있는 모든 컴퓨팅 인스턴스를 추적해보면, 이전 세대인 Pascal(청색)이나 Volta(흑색)가 즉시 사라지지 않는다는 것을 명확히 알 수 있음. 구형 GPU는 다른 종류의 워크로드(추론 등)를 담당하며 경제적 생명을 이어간다는 '롱테일' 수요!



자료: Keybanc, 미래에셋증권 리서치센터

며칠 전 CoreWeave의 재무 및 운영 성과는 AI 인프라 수요가 실재하는 현금 흐름으로 이어지고 있음을 증명한다. 2025년 3분기 재무 성과는 매출 13억 6,000만 달러(전년 동기 대비 134% 성장, 시장 예상치 12억 8,000만 달러 상회), 수주 잔고(Backlog) 556억 달러(전 분기 대비 85% 폭증)를 기록하며 미래 매출에 대한 강력한 가시성을 제공했다. 주당순손실(GAAP Loss) 역시 시장 예상치인 (0.51달러)보다 손실 폭을 크게 줄인 주당 (0.22달러)로 수익성 개선의 가능성을 보였다.

지난 3분기에, CoreWeave는 Meta와 최대 약 142억 달러, OpenAI와 약 224억 달러 규모의 파트너십을 구축했다. 그리고 마이크로소프트는 속도를 위해 자체 데이터센터를 짓는 대신 CoreWeave 같은 전문 업체로부터 임대하는 전략을 채택했다고 사티야 나델라 CEO가 직접 밝혔다. 이러한 수치는 CoreWeave가 세계 최고의 AI 기업들에게 '필수적인 AI 클라우드'로 자리매김했음을 보여준다.

(3) 공급을 압도하는 거시적 시장 신호

그렇다면, 앞으로의 핵심 질문은 감가상각 기간인 5~6년이 지난 후에도 GPU가 시간당 운영비 이상의 수익을 창출할 수 있는가가 된다. 그 운영비 가격은 대략 0.30~0.30달러로 추정되기 때문에, 우리가 앞서 살펴봤듯 V100 인스턴스가 여전히 판매중이라는 것을 알 수 있다.

표 4. AI 클라우드 총 소유 비용

(H100 기준)	단위	하이퍼스케일러	대형 네오클라우드	신흥 네오클라우드
자기자본비용	%	11.00%	20.00%	25.00%
타인자본비용	%	4.50%	11.00%	15.00%
자기자본 비율	%	75.00%	25.00%	25.00%
가중평균자본비용(WACC)	%	9.40%	13.30%	17.50%
월간 총 서버 자본 비용	USD/월	$6,742	$8,098	$9,785
코로케이션 비용	USD/kW/월	$110	$130	$150
월간 총 운영 비용	USD/월	$1,776	$2,082	$2,506
월간 총 소유 비용	USD/월	$8,518	$10,180	$12,290
시간당 유닛별 자본 비용	USD/시간	$1.15	$1.39	$1.68
시간당 유닛별 호스팅 비용	**USD/시간**	**$0.30**	**$0.36**	**$0.43**
시간당 유닛별 총 비용	USD/시간	$1.46	$1.74	$2.10
총 소유 비용 대비 자본 비용 비율	%	79.90%	80.90%	81.80%

자료: SemiAnalysis, 미래에셋증권 리서치센터

또 다른 거시적 신호로는, 공급망 최상단에서 엔비디아는 TSMC에 Blackwell GPU 생산량을 월 10만~11만 장에서 최대 16만 장으로 최대 50% 증산해달라고 요청한 것을 들 수 있다. CEO 젠슨 황이 며칠 전 "수요가 매달 증가하고 있다"고 언급한 것처럼, 칩 제조의 최상단에 있는 공급망조차 따라가기 벅찬 수준의 수요 폭발이 일어나고 있음을 시사한다.

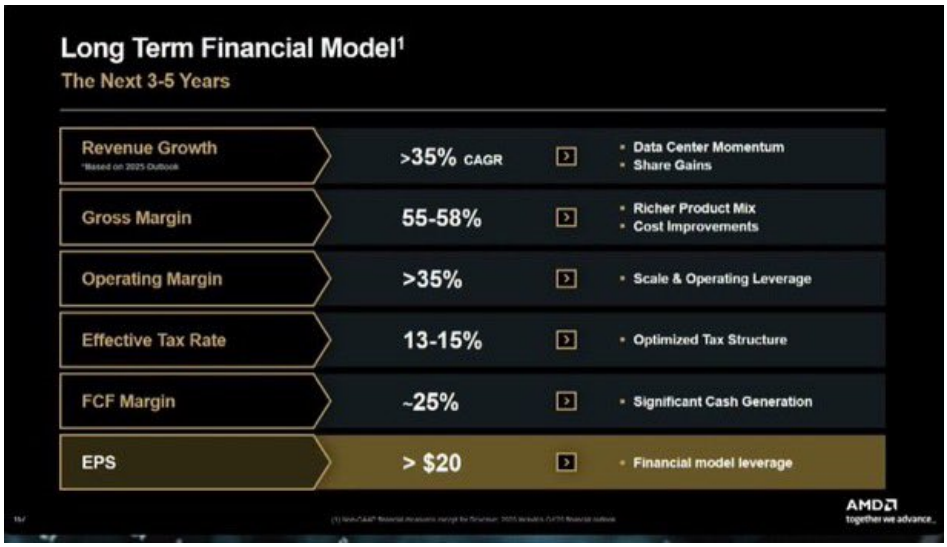
AMD의 CEO 리사 수 역시 "여러 거대 기가와트(GW) 규모 고객을 확보할 것으로 예상한다"고 발표하며, AI 인프라의 새로운 표준 단위가 MW가 아니라 GW가 되었음을 시사했다. 실적발표 자리에서 AMD는 단기 데이터센터 성장률 목표로 80% 이상을 제시하며 2026년 매출이 훨씬 더 높을 것으로 전망했다.

그림 5. 데이터센터 시장 규모 전망치를 기존 대비 두 배 이상인 1조 달러 이상으로 상향 조정

자료: AMD, 미래에셋증권 리서치센터

특히 리사 수 CEO는 대규모 하이퍼스케일러들의 재무 상태가 매우 견고하기 때문에, AI CapEx를 위한 자금 조달은 문제가 되지 않을 것이라는 자신감을 보였다. 이는 특정 기업의 독주가 아닌, 시장의 핵심 플레이어 모두가 기가와트 규모의 초거대 AI 인프라 시대를 기정사실로 보고 천문학적인 투자를 집행하고 있음을 의미한다.

그림 6. AMD가 밝힌 장기 가이던스 내용
주당순이익(EPS) 20달러 이상을 목표로 하고 있음. 가장 낙관적인 예상치조차 훨씬 상회하는 수치이를 위해 약 35% 연평균 성장을 하고, 매출총이익률은 55~58%에 육박할 것이라는 자신감



자료: AMD, 미래에셋증권 리서치센터

결론적으로, 낙관론자들은 버리가 과거의 사이클(닷컴 버블 등)에 기반한 분석의 틀에 맞춰, AI라는 새로운 패러다임이 창출하는 전례 없는 수요의 규모와 지속성을 과소평가하고 있다고 말한다. 구형 자산의 여전히 높은 가치 유지, GPU의 다단계 생애주기, CoreWeave의 실질적 성장, 그리고 시장 전반의 공급 부족 신호는 모두 현재의 투자가 과잉이 아닌, 다가올 AI 시대의 기반을 다지는 필수적인 과정임을 가리키고 있다.

3. AI 데이터센터의 운영 현실과 물리적 병목

더욱 근본적 문제는, AI 자본지출 논쟁은 회계장부를 넘어 데이터센터를 구축하고 운영하는 물리적 세계의 제약 조건과 직결된다는 점이다. 시장의 낙관론을 뒷받침하는 폭발적인 수요가 현실화되는 것을 가로막는 실질적인 병목 현상이 존재한다는 말이다.

이는 아이러니하게도 AI 인프라 시장의 과열을 억제하고 CoreWeave와 같은 선두 기업의 해자(moat)를 만드는 요인으로 작용하고 있다. 이 병목 현상의 핵심은 '전력(raw power)' 그 자체가 아니라, 그 전력을 GPU 랙까지 전달하는 복잡한 인프라인 '파워셸(Powered Shell)'에 있다.

(1) 병목의 실체: 전력이 아닌 파워셸

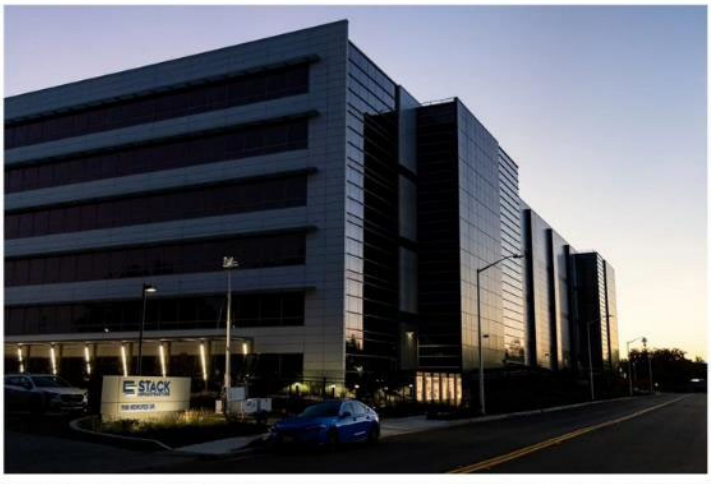
AWS 엔지니어의 증언에 따르면, 데이터센터 밀집 지역의 연간 전력 수요 증가율은 과거 0.5~1%에서 현재 3~7% 또는 그 이상으로 폭증했다고 한다. 그는 향후 4~5년간 현재 AI 수요 곡선만을 만족시키기 위해서라도 30~60GW의 전력망 용량 증설이 필요할 것이라고 전망했다. 그러나 AI 수요는 일시적인 것이 아니다. 그러니 전력 유틸리티 및 배전 공급업체의 성장 곡선을 근본적으로 바꾸는 핵심 동력이 될 가능성이 크다고 현장 엔지니어들은 입을 모은다. 해당 엔지니어의 증언에서 또한 귀 기울여 들어본 만한 증언은, 신규 데이터센터 건설 수요 중 최소 60%~80%는 실현될 것으로 예측한 사실이다.

그런데 이를 위한 당장의 병목은 파워셸, 즉 “전력을 GPU 서버까지 전달”하는 데이터센터의 물리적 인프라 파이프라인 전체에 있다. 엔비디아는 분기당 약 1.1 GW 규모의 Blackwell GPU를 미국에 출하 중이지만, 이를 설치할 데이터센터 건설은 예상보다 훨씬 더딘 상황이다.

그림 7. 엔비디아의 고향에서조차 파워셸의 문제로 가동되지 않은 상태로 남아 있다는 기사

Data Centers in Nvidia's Hometown Stand Empty Awaiting Power

The fate of two facilities in Santa Clara, California, highlights a major challenge for the US tech sector and indeed the wider economy.



Stack Infrastructure's SVY02A data center project in Santa Clara. *Photographer: Jason Henry/ Bloomberg*

자료: 블룸버그, 미래에셋증권 리서치센터

데이터센터 부품 공급망의 리드타임이 이를 증명한다. HDD의 리드타임은 12~18개월, 네트워크 스위치는 16~20개월이 소요된다. 가장 극심한 제약은 전력 인프라 영역이다. 대형 변압기는 리드타임이 무려 약 36개월에 달하며, 지멘스, 히타치 등 전 세계 4~5개 공급업체로 제한되어 있다.

CoreWeave는 총 2.9GW의 전력 공급 계약을 확보했지만, 현재 실제로 가동 중인 전력은 590MW로 계약된 전력의 약 20%만 사용되고 있다. 이는 전력 발전 용량 자체가 당장의 제약 조건이 아님을 보여준다.

수십억 달러의 최첨단 하드웨어가 데이터센터 완공을 기다리며 대기 중인 것이다. 진정한 병목인 파워셸은 단순히 건물을 짓는 것을 넘어, 다음과 같은 복잡한 요소들의 총합이다.

① **GPU의 높은 전력 밀도와 발열량에 맞춰 냉각 시스템, 바닥 하중 등이 완전히 새롭게 설계된 'AI 워크로드 전용 데이터센터'**

② **변압기, 무정전 전원 장치(UPS), 배전 설비 등 전력 변환 및 분배에 필요한 모든 장비**

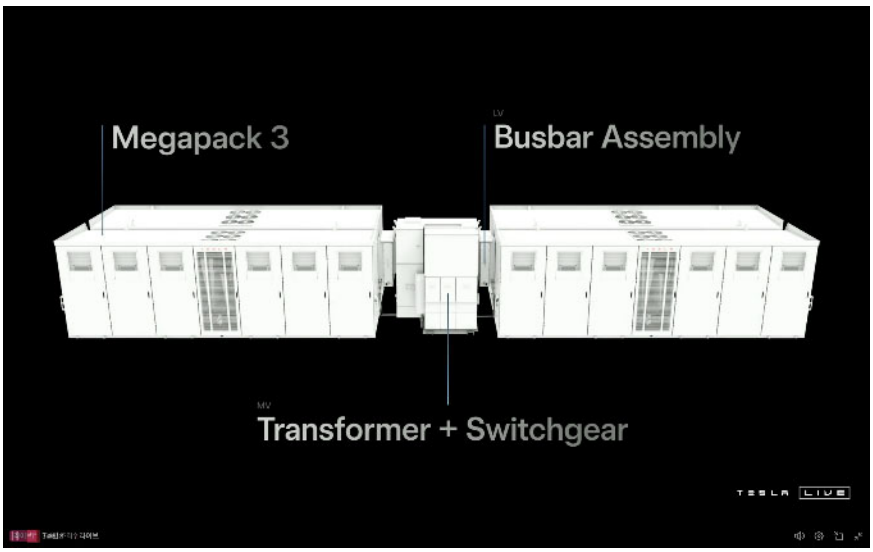
③ **이를 건설하고 유지보수할 숙련된 전문 인력**

이 모든 요소들의 공급망은 현재 "압도당하고 있다"고 표현될 정도로 절대적으로 부족한 상황이다.

(2) 기가와트 시대의 과제와 시장 함의

AI 데이터센터가 기가와트(GW) 규모로 확장되면서, 과거에는 사소하게 여겨졌던 전력 손실이 막대한 운영 비용과 용량 제약으로 이어진다. 전력은 그리드에서 칩까지 여러 단계의 변환을 거치며 미세한 손실이 누적된다. 예를 들어 효율이 99%인 변환 장치라 할지라도, GW급 데이터센터에서는 이 1%의 손실이 10MW에 달하는 엄청난 전력 낭비로 이어진다. 이 10MW는 약 70개의 엔비디아 GB200 랙을 구동할 수 있는 양이다.

그림 8. 테슬라의 메가블록은 4개의 메가팩3 유닛을 미리 통합한 "pre-engineered block"
변압기와 스위치기어를 내장해 그리드 연결이 더 간단하고, 설치 시간이 23% 단축
별도의 변압기와 스위치기어를 현장에서 설치해야 한 것에서 탈피



자료: 테슬라, 미래에셋증권 리서치센터

그림 9. 생산 측면에서 2026년 하반기부터 자체 변압기 생산을 시작해, 메가블록에 통합할 계획
가장 큰 병목인 외부 변압기를 별도로 소싱(구매 및 설치)할 필요가 없어짐

자료: X(#WholeMarsBlog, #Elonmusk), 미래에셋증권 리서치센터

이 문제를 해결하기 위해 업계는 고체 변압기(Solid State Transformers)나 고전압 DC/DC 컨버터 모듈과 같은 신기술을 개발하고 있다. 테슬라가 자체 변압기를 만들겠다고 선언한 것도 이러한 이유가 있어서다. 이런 고도의 전기공학 기술력 확보의 가치는 앞으로 시장에서 점차 인식해야 할 중대한 기업 역량이라고 사료된다.

그리고 이러한 물리적, 기술적 병목 현상은 AI 인프라 시장에 중요한 영향을 미친다. 첫째, 데이터센터를 실제로 구축하는 속도가 그 성장을 제한하는 '자연적 속도 제한 장치' 역할을 하여 무분별한 과잉 공급을 막고 시장의 급격한 붕괴 위험을 완화한다.

둘째, CoreWeave는 이러한 공급망 제약 속에서 3년간 운영하며 노하우를 쌓았고, 이는 공급자 우위 시장이 형성하며 고객 충성도를 올린다. CoreWeave의 CEO가 직접 밝힌 사실에 따르면, 심지어 GPU 임대 공급이 지연되는 상황에서, 고객들은 CoreWeave와의 계약을 취소하는 대신 약속된 인프라 전체를 모두 보장받기 위해 기꺼이 계약을 변경/조정해 주고 있다고 한다. CoreWeave 인프라의 가치가 매우 높으며 대체재를 찾기 어렵다는 것을 방증한다. 전형적인 공급자 우위 시장의 특징이다.

4. 종합 분석 및 전망

AI 자본지출과 CoreWeave 등을 둘러싼 논쟁은 단순히 낙관론과 비관론의 충돌을 넘어, 회계적 관례와 폭발적인 기술 변화가 만들어내는 경제적 현실 사이의 괴리를 보여주는 결정적인 사례다. 버리는 과거의 경험과 회계 원칙이라는 렌즈를 통해 시스템적 위험을 경고하고 있으며, 시장은 전례 없는 수요 데이터와 구조적 변화라는 증거를 통해 그 경고에 맞서고 있다.

(1) 사기인가, 대담한 베팅인가?

버리의 진단은 사실에 기반한다. SEC 공시 자료는 그가 지적한 대로 하이퍼스케일러들이 AI 투자를 본격화한 시점에 컴퓨팅 자산의 내용연수를 일제히 연장했음을 명백히 보여준다. 이 "사소한 회계 변경"이 수십억 달러의 감가상각비를 이연시켜 단기 재무제표를 개선했다는 그의 주장 역시 사실이다.

그러나 버리의 해석은 시장의 현실 데이터에 의해 강력한 도전을 받는다. 마이클 버리의 주장이 진정 '사기(fraud)'로 성립하려면, 이러한 회계적 조정이 실제 자산의 경제적 가치와 의도적으로 괴리되어야 한다. 하지만 현재까지의 증거는 오히려 그 반대를 시사한다. 7~8년 된 Google TPU의 100% 가동률과 3년차에 접어든 H100 GPU가 원가의 95% 가격에 재계약되는 현실은, 이 자산들의 실제 경제적 수명이 과거의 통념이나 회계적 관례를 훨씬 뛰어넘고 있음을 증명한다. 이는 기업들이 내용연수를 연장한 행위가 '이익 부풀리기'가 아니라, '새로운 경제적 현실을 반영하기 위한 합리적 재평가'일 수 있다는 강력한 근거가 된다.

CoreWeave의 556억 달러 수주 잔고와 GPU 현물 가격 상승은 버리가 우려하는 '수요 없는 과잉 투자'와는 정반대의 상황이 펼쳐지고 있음을 보여준다. 또한 '파워셸'로 대표되는 물리적 인프라 구축의 병목은 2007년 서브프라임 사태처럼 규제 없는 레버리지가 무한정 팽창하던 상황과는 근본적으로 다르다.

결론적으로 현재 상황은 사기성 버블이라기보다는, 기술 패러다임 전환기에 나타나는 '거대하고 합리적인, 그러나 대담한 베팅'으로 해석하는 것이 타당하다. 이 투자는 이미 OpenAI, Meta 등 거대 고객의 수백억 달러 계약으로 증명된 실질적 수요에 기반한다. 버리의 경고는 이 베팅이 실패할 경우의 위험성을 상기시키는 중요한 역할을 하지만, 현재까지의 데이터는 베팅의 근거가 더 견고함을 시사한다.

(2) 핵심 관전 포인트: 미래를 결정할 중대 변수

버리의 경고는 AI 산업의 건강한 발전을 위해 필수적인 '제동 장치' 역할을 하고 있다는 정도로 봐야 한다. 이 논쟁의 향방은 결국 다음 네 가지 핵심 변수들이 어떻게 전개되느냐에 따라 결정될 것이다.

AI 서비스의 수익화 속도

J.P. Morgan이 제기한 '연간 6,500억 달러 수익'의 실현 가능성을 확인해야 한다. 현재의 막대한 Capex는 결국 미래의 수익으로 정당화되어야 하며, AI 서비스가 실질적인 고부가가치 비즈니스로 자리 잡는 속도가 투자의 성패를 가를 것이다. 이 지점에서 현재 세대 모델의 경이로운 수익성이 중요한 변수로 부상한다.

현세대 대표 모델인 GPT-5의 추론 서비스만 가정하면, 영업이익률이 48%에 육박할 것이라는 계산 근거는 중요한 함의를 가진다. 이는 AI 서비스가 단순한 비용 소모적인 기술 시연을 넘어, 한번 구축되면 막대한 현금을 창출하는 디지털 자산의 성격을 가짐을 시사한다.

표 5. 현재 모델인 GPT-5만을 기준으로 OpenAI의 영업이익률을 추산해보면, 약 48%에 육박

단계	항목	가정/계산	결과
1	**구글이 밝힌 처리량 규모**	**Gemini 월간 처리량**	**10^{15} 토큰/월**
2	OpenAI 처리량 추정	Gemini의 2배로 가정	2×10^{15} 토큰/월
3	초당 처리량 환산	월간 → 초당 변환	약 8억 토큰/초
4	칩 성능	B200 칩 출력 속도	3,000+ 토큰/초/칩
5	필요 칩 수량	8억 ÷ 3,000 ÷ 0.8 (80% 이용률)	31만 개의 B200 칩
6	칩 운영 비용	31만 개 × $2.5/시간 × 730시간/월	$5.6억/월
7	총 운영 비용	칩 비용 + 기타 비용 50%	$8.4억/월
8	**OpenAI 월 매출**	**공개 추정치**	**$16억/월**
9	영업이익	매출 - 총 비용	$7.6억/월
	영업이익률	**이익 ÷ 매출**	**47.50%**

자료: 미래에셋증권 리서치센터

표 6. 그럼에도 막대한 손실이 발생되는 원인은? 차세대 모델 개발 때문

항목	내용	수치
추론(Inference) 운영	GPT-5 서비스 제공	GPU 사용 비중 30%~50%
훈련(Training) 투자	GPT-6 개발	GPU 사용 비중 50%~70%
수익성	추론 운영만으로는 흑자 가능	영업이익률 ~48%
전체 손익	**대규모 GPT-6 투자로 인한 손실**	**현재 적자**

자료: 미래에셋증권 리서치센터

추론(Inference) 시장의 본격적인 개화

전체 AI 워크로드에서 '추론'이 차지하는 비중 변화와 구형 GPU가 추론 시장에서 차지하는 점유율을 살펴봐야 한다. '훈련'은 초기 투자를 주도했지만, AI의 장기적 가치는 '추론'에서 나온다. 추론 시장이 폭발적으로 성장해야만 GPU의 '다단계 생애주기' 모델이 완성되며, 수많은 구형 GPU들이 장기간에 걸쳐 가치를 창출할 수 있게 된다. 이는 내용연수 연장의 가장 강력한 정당화 근거가 될 것이다.

표 7. 추론의 시대를 위해 반드시 생겨야 할 것은 엄청난 숫자의 '엣지 컴퓨팅'

카테고리	항목	훈련용 데이터센터	추론용 데이터센터
기본 특성	규모	~150 MW급 대형 시설	10-20 MW급 소형 시설 (1/10 수준)
	위치	"nowhere" 가능 (텍사스, 오클라호마 등 저렴한 전력 지역)	주요 도심 인근 필수 (edge computing)
	목적	초고밀도 컴퓨팅, 레이턴시 무관	저지연 사용자 경험 (챗봇, 검색, 자율주행 등)
	전략	소수의 거대 허브, 전력 비용 최적화	다수의 분산 배치, 지리적 proximity 중심
핵심 제약	물리적 병목	전력 확보	부동산/입지
투자 동향	현재 상황	여전히 training 중심 구축 중	수요 폭발 예고, "way more" 필요
	아키텍처 변화	"fewer but bigger" 모델	"more but smaller" 모델로 전환
전략적 함의	차별화 요소	전력 비용 효율성	위치가 핵심 차별화 요소
	수혜 섹터	**전력 인프라, 냉각 시스템 공급업체**	**부동산/통신 및 네트워크 인프라 업체**

자료: 미래에셋증권 리서치센터

II. 키미노 나마에와... 테슬라?

1. Kimi K2 Thinking: 중국 AI가 만들어낸 또 다른 특이점

우리가 선정했던 중국의 3대 AI 기업 중 하나인 Moonshot AI가 또 일을 냈다. 그들이 공개한 오픈소스 및 오픈웨이트 모델 Kimi K2 (Thinking)은 단순한 모델 성능 업데이트가 아니었다. 이는 AI 시장의 경쟁 구도와 기술 발전의 방향성 자체를 바꾸는 지각 변동이라 할 만하다.

Kimi K2는 특정 고난도 벤치마크에서 GPT-5, Claude 4.5와 같은 최상위 폐쇄형 모델을 능가하는 성능을 입증했기 때문이다. 이는 개방형 모델 생태계가 기술적 한계를 돌파하고 있음을 명확히 보여주는 이정표다. 장기 추론과 다단계 툴 호출을 결합한 '사고 에이전트' 패러다임은 복잡계 문제 해결 능력의 새로운 기준을 제시한다.

그림 10. "AI 현황 보고서"에서 언급했던 Moonshot AI의 LLM인 Kimi

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

자료: 미래에셋증권 리서치센터

(1) Kimi K2의 정체성: 응답을 넘어 해결을 지향하는 '사고 에이전트'

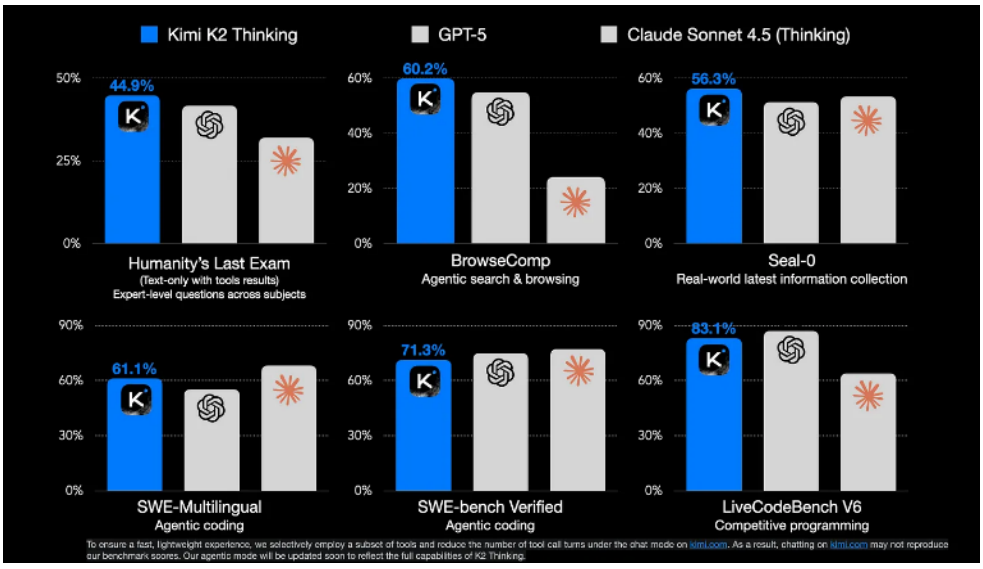
Kimi K2 Thinking의 핵심 정체성은 '사고 에이전트(Thinking Agent)'다. 이 모델은 응답을 생성하는 것을 넘어, 문제 해결을 위해 장시간에 걸쳐 단계별로 추론하고 그 과정에서 외부 툴을 능동적으로 사용하는 능력을 갖추도록 설계되었다. 처음부터 사고 에이전트로 구축되어, 단계별 추론과 툴 사용을 내재화했다.

주요 역량은 순차적 툴 호출 능력에서 드러난다. Kimi K2는 인간의 개입 없이 최대 200~300회에 달하는 순차적 툴 호출을 스스로 수행할 수 있다. 이는 복잡한 문제를 수백 개의 하위 단계로 분해하고, 각 단계에 필요한 정보를 검색하거나 코드를 실행하며 일관성 있는 추론을 유지하는 능력이다. 또한 도구 호출과 호출 사이에 '사고 토큰(thinking tokens)'을 생성하여 다음 행동을 계획하고 중간 결과를 검증한다는 말이다.

(2) 정량적 성능: 데이터로 입증된 프론티어급 역량

Kimi K2 Thinking은 추론, 코딩, 에이전트 능력 평가에서 기록적인 성과를 보였다. 이는 기존 오픈소스 모델의 한계를 넘어 최상위 폐쇄형 모델과 직접적으로 경쟁하는 수준에 도달했음을 의미한다. 특히, 가장 어려운 벤치마크라고 불리는 HLE에서 GPT-5를 능가한 것은 Kimi K2의 복합 추론 능력이 SOTA 수준임을 증명한다. 웹 브라우징 및 정보 검색 능력을 평가하는 BrowseComp에서는 인간 기준선(29.2%)을 크게 상회하기도 했다.

그림 11. 고난도 추론, 코딩, 에이전트 벤치마크에서 GPT-5 및 Claude 4.5를 능가하는 성능 지표
개방형 모델이 최상위 폐쇄형 모델과 직접 경쟁 가능함을 증명하며, AI 기술 지형도의 변화를 예고



자료: Kimi, 미래에셋증권 리서치센터
주: 모든 평가는 Kimi K2의 실제 서비스 환경인 INT4 정밀도에서 측정

표 8. 데이터로 입증된 Kimi K2 Thinking의 기술적 우위

벤치마크 (평가 영역)	Kimi K2 Thinking	GPT-5	Claude Sonnet 4.5
LiveCodeBench V6 (경쟁 프로그래밍)	83.1%	87.0%	64.0%
SWE-bench Verified (에이전트 코딩)	71.3%	74.9%	77.2%
BrowseComp (에이전트 검색)	60.2%	54.9%	24.1%
Humanity's Last Exam (전문가 추론)	44.9%	41.7%	32.0%

자료: Kimi, 미래에셋증권 리서치센터

2. 기술적 해부: 제약과 혁신이 빚어낸 아키텍처

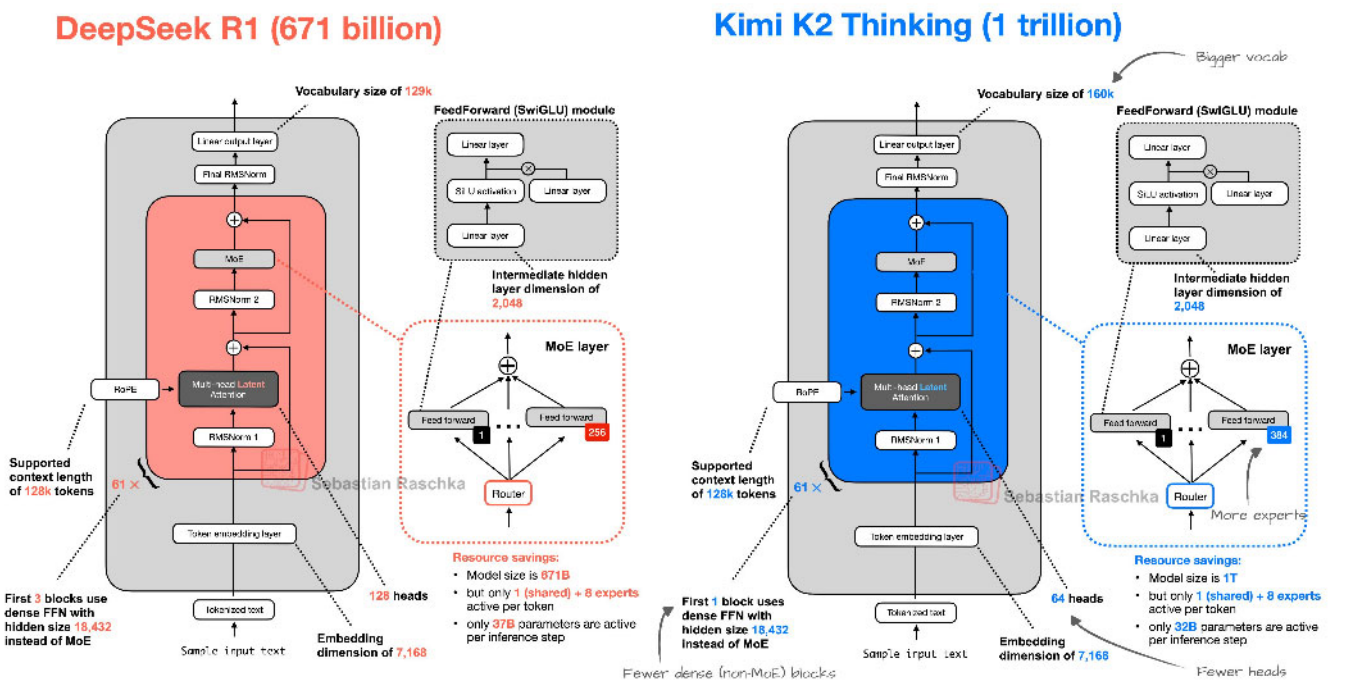
(1) 하드웨어 제약이 촉발한 혁신과 Kimi가 보인 진화

Kimi K2는 엔비디아의 최신 GPU가 아닌, 미국의 수출 통제로 인해 성능이 제한된 H800 GPU 클러스터에서 훈련되었다. 이 하드웨어적 한계는 단순한 제약이 아니라 기술 혁신을 강제하는 촉매제로 작용했다. Moonshot AI는 이 제약을 소프트웨어와 아키텍처 혁신으로 정면 돌파했다. 이는 Moonshot AI가 하드웨어적 한계를 소프트웨어와 아키텍처 혁신으로 극복했음을 의미한다.

Kimi K2는 기본적으로 DeepSeek의 R1 모델을 확장한 버전이라 할 수 있다. 당연하게도 MoE(Mixture of Experts) 아키텍처를 기반으로 한다. 총 파라미터 규모는 1조 개에 달하며, R1보다도 많은 384개의 전문가(Expert)를 보유한다.

반면 추론 시 활성화되는 파라미터는 320억 개에 불과한 점이 특이할 만하다. DeepSeek R1(370억 개)의 활성화 파라미터보다도 적은 것이다. 이는 더 많은 전문가 풀로 전문성을 세분화하되, 실제 연산은 더 적은 파라미터로 처리하여 효율과 성능의 균형을 맞추려는 시도다. 이 외에도 256K의 컨텍스트 길이와 160K의 어휘 크기를 지원한다.

그림 12. DeepSeek R1과 그 후계자인 Kimi K2 Thinking 모델의 아키텍처 시각화
DeepSeek R1의 MoE 레이어는 256개, Kimi K2 Thinking은 384개의 전문가(Feed forward)로 확장. 전문가 수 증가가 돋보임
게다가 Kimi K2가 DeepSeek R1보다 총 파라미터 수는 훨씬 많지만, 추론 시 활성화되는 파라미터 수는 오히려 적음
이는 DeepSeek R1 대비 더욱 정교하고 심층적 사고 과정을 효율적으로 수행할 수 있음을 시사



자료: Sebastian Raschka, 미래에셋증권 리서치센터

(2) 전략의 핵심, 네이티브 INT4 양자화

Kimi K2의 가장 중요한 기술적 결정은 네이티브 INT4 양자화 인식 훈련(QAT)이다. 이는 하드웨어 제약을 극복하고 실용성을 극대화하기 위한 필연적 선택이었다.

Kimi K2의 MoE 아키텍처는 추론 시 연산량(FLOPs)은 효율적이지만, 방대한 모델 가중치(1조 파라미터)를 GPU 메모리에서 연산 코어로 불러오는 과정에서 메모리 대역폭에 의해 성능이 제한되는 '메모리 바운드' 상태를 유발한다. 메모리 반도체의 성능이 가장 중요하다는 말이다. 그러나 중국은 대중국 첨단 칩 제재로 HBM 메모리가 극히 부족한 실정이다.

양자화 인식 훈련(QAT) 역시, '제약이 창의성을 만들어내듯' Kimi가 선택한 고육지책이다. 가중치를 INT4로 양자화하면 모델 크기가 절반으로 줄어 메모리 전송량이 감소하고, 이는 추론 지연 시간을 약 2배 단축시키는 결정적 효과를 낳는다. 개발팀은 초기 "사후훈련 양자화(PTQ)" 기법을 썼었지만, 긴 문맥 추론에서 오차 누적으로 실패하자, 훈련 과정에 양자화 오차를 미리 시뮬레이션하여 모델이 이에 적응하도록 만드는 QAT로 전환했다.

Kimi K2의 성능을 뒷받침하는 가장 중요한 기술적 결정은 네이티브 INT4 양자화 인식 훈련(QAT)이다. 이로 인해 INT4의 낮은 추론 지연 시간은 강화학습 훈련의 병목 구간인 '롤아웃' 단계를 직접 가속하여 전체 RL 반복 주기를 10-20% 단축시키는 시너지 효과까지 발생했다. “QAT와 네이티브 양자화”에 대해서는 자세하게 후술할 것이다.

(3) 독자적 아키텍처 혁신: KDA와 Muon 스케일링

Moonshot AI는 1조 파라미터까지 안정적으로 확장 가능한 독자적 옵티마이저인 'Muon'을 개발했다. 옵티마이저는 가중치(파라미터)를 어떻게 업데이트할지를 결정하는 알고리즘으로 AI 모델의 트레이닝 과정에서 필수품이다.

대부분의 연구소들이 검증된 Adam 옵티마이저를 사용하거나 수정하는 것과 달리, 자체 옵티마이저를 개발했다는 것은 Moonshot AI가 훈련 스택의 가장 근본적인 부분까지 직접 제어하고 최적화하려는 의지와 역량을 갖추었음을 보여준다. 물론, 자체 개발 과정에서 수십 개의 옵티마이저와 아키텍처가 실패했으며, 현재의 성공은 수많은 실험의 결과물이라고 밝혔다.

또한, 어텐션 메커니즘에서는 표준적인 RoPE 대신 'KDA(Kimi Delta Attention)'와 'NoPE MLA'를 결합한 하이브리드 방식을 사용해 더 나은 성능을 확보했다. 참고로, MLA 기법도 DeepSeek가 개발한 것이니 중국 AI의 발전은 DeepSeek가 상당히 영향력을 끼친 것이 맞다. 그리고 Kimi의 성공은 또한 서로를 발전시킬 것이고, 이는 장기적으로 중국 업체들의 약진을 불러올 가능성이 높다고 사료된다.

3. 지정학적 AI 패권 경쟁의 판도를 바꾸는 또 하나의 게임 체인저?

(1) 고작 60억원으로 개발했다고?

Kimi K2의 등장은 중국 AI 기업들이 구사하는 고도의 전략적 움직임을 보여준다. 미국 연구소들은 모델 개발에 수천억에서 조 단위의 자본을 투자하고 API 사용료로 회수해야 한다. 반면 중국 연구소들은 서구의 연구 결과를 '지식 증류' 방식으로 흡수하여 훨씬 적은 비용으로 동급 이상의 모델을 개발한다.

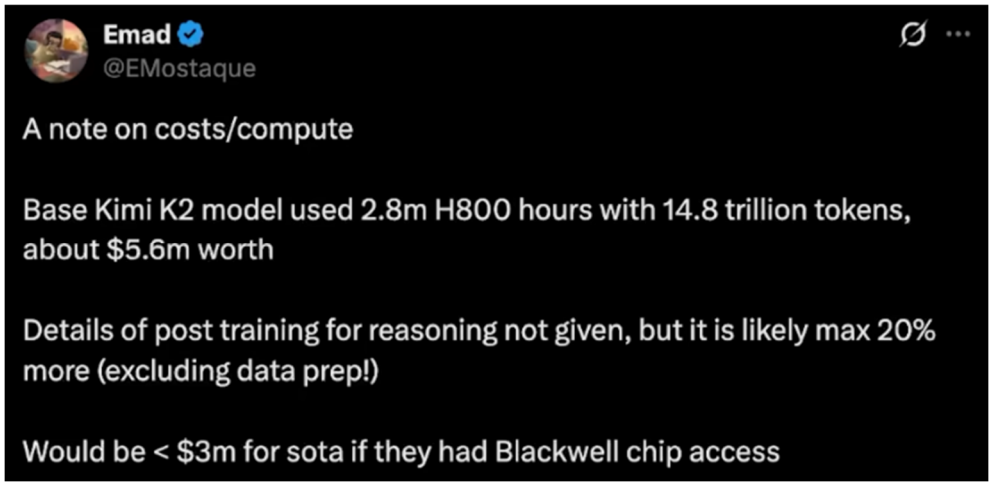
올 초에 있었던 “DeepSeek Shock” 때와 같은 충격은 없었지만, 이번에도 중국 업체들의 극도로 “효율적인 훈련” 비용이 언론에서 보도되었다. CNBC에서는 K2 모델의 훈련 비용이 460만 달러라고 기사를 썼다. 여기서 중요한 사실은, Moonshot AI도 스스로 밝혔듯이 전체 훈련 비용이 460만 달러가 아니라는 말이다.

오히려 Moonshot AI 관계자는 "훈련 비용을 정량화하기 어렵다. 연구와 실험이 주요 부분을 차지하기 때문"이라고 밝혔다. Muon 아키텍처와 같은 자체 알고리즘을 완성하기까지 수많은 실패한 실험이 있었음을 고려할 때, ‘디버깅, 하이퍼파라미터 튜닝, 병렬 실험 등’을 포함한 전체 연구개발 비용은 460만 달러를 엄청나게 상회할 것으로 추정된다.

Stable Diffusion의 창시자인 Emad Mostaque는 Kimi K2 기본 모델의 훈련 비용을 약 560만 달러로 추정했다. DeepSeek의 V3 모델의 훈련 비용이 558만 달러였다는 것과 거의 동일한 수치라 흥미롭다. Mostaque의 560만 달러 추정은, 280만 GPU 시간(H800)과 14.8조 개 토큰 데이터셋을 기반으로 한 추산이다.

흥미로운 점은, 그들이 만약 엔비디아의 더 효율적인 Blackwell 칩을 사용했다면 300만 달러 미만으로도 가능했을 것이라 덧붙인 사실이다. Blackwell과 같은 차세대 GPU 아키텍처가 제공하는 연산 효율성 향상으로 모델 훈련 비용은 더 낮아질 수 있다는 말이고, 어찌 보면 중국 업체들의 물리적 한계를 슬며시 드러낸 말이기도 하다.

그림 13. Emad Mostaque가 자신의 X 계정에 Kimi K2에 대한 비용을 역산



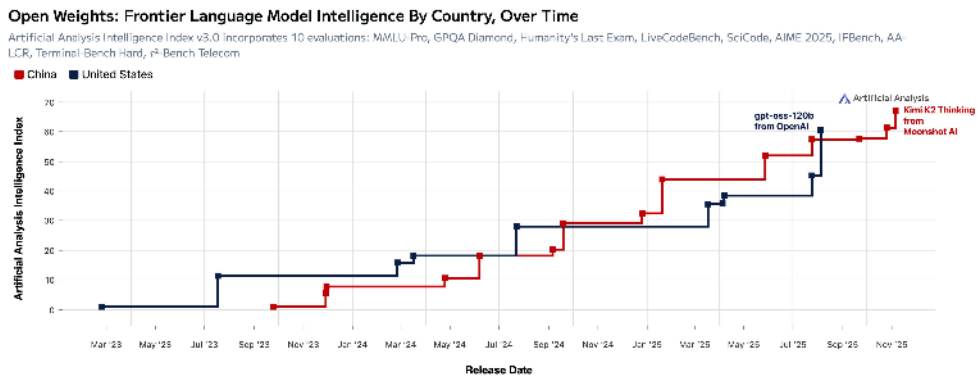
자료: Emad Mostaque, 미래에셋증권 리서치센터

게다가 Mostaque는, “이 추정치에는 데이터 준비(data prep) 비용은 포함되지 않았다”는 단서를 덧붙였다. 고품질의 추론용 데이터를 수집하고 정제하는 데 약 20%의 추가 비용이 발생할 수 있음을 암시한 것인데, 결론적으로 그렇게 놀랍지는 않다는 반응으로 사료된다. 주식시장에서도 Kimi K2 Thinking에 대한 반응은 전혀 특기할만 한 것은 없었기도 하다.

(2) 풀스택 야심과 '기본적 비밀주의'

그러나 더 근본적인 것은, 중국이 이렇게 개발한 고성능 모델을 오픈소스로 공개한다는 사실이다. 이는 미국 기업들의 고가 API 서비스에 대한 강력한 가격 인하 압박으로 작용한다. 잠재적으로 미국 기업들의 투자수익률(ROI)을 감소시키고, 가격에 민감한 글로벌 시장은 중국 기반 오픈소스 생태계로 편입될 가능성을 조금이라도 더 높인다는 게 포인트다.

그림 14. 중국이 오픈소스 생태계의 새로운 리더로 부상했음을 시각적으로 증명하는 데이터
미국의 초기 오픈소스 독주를 맹렬히 추격한 중국, Kimi K2 Thinking 모델 출시로 결정적 역전 성공



자료: Artificial Analysis, 미래에셋증권 리서치센터

또한 중국의 오픈소스 전략은 미국의 가장 큰 강점인 소프트웨어 리더십을 무력화하기 위한 정교한 계산에 기반한다. 중국이 오픈소스 전략을 통해 미국의 소프트웨어(AI 모델) 우위를 무력화하면, 하드웨어 제조 경쟁력까지 갖춘 중국이 AI 시대의 완전한 기술 스택(Full Stack)을 장악할 수 있다는 전략적 계산이 깔려 있다. 2025년 초 DeepSeek의 등장을 시작으로, 연말 Kimi K2의 출시는 중국발 고성능 오픈소스 모델의 부상이 일시적 현상이 아님을 증명한다.

그리고 미중 기술 경쟁은 정보 공개 방식에서도 뚜렷한 차이를 보인다. 미국은 연구 성과를 논문 등으로 비교적 투명하게 공개한다. 반면 중국은 AI와 같은 핵심 전략 기술에 대한 연구 성과를 '기본적 비밀주의' 원칙하에 내부적으로만 공유한다. 이들은 서구권에서 특정 수준의 기술이 공개적으로 발표된 후에야, 자신들이 이미 도달했거나 그 이상의 기술력을 필요한 만큼만 공개하는 전략을 구사한다. 이는 경쟁 상대가 자신들의 진정한 기술적 한계(Ceiling)를 가늠하지 못하게 하는 고도의 정보전이다.

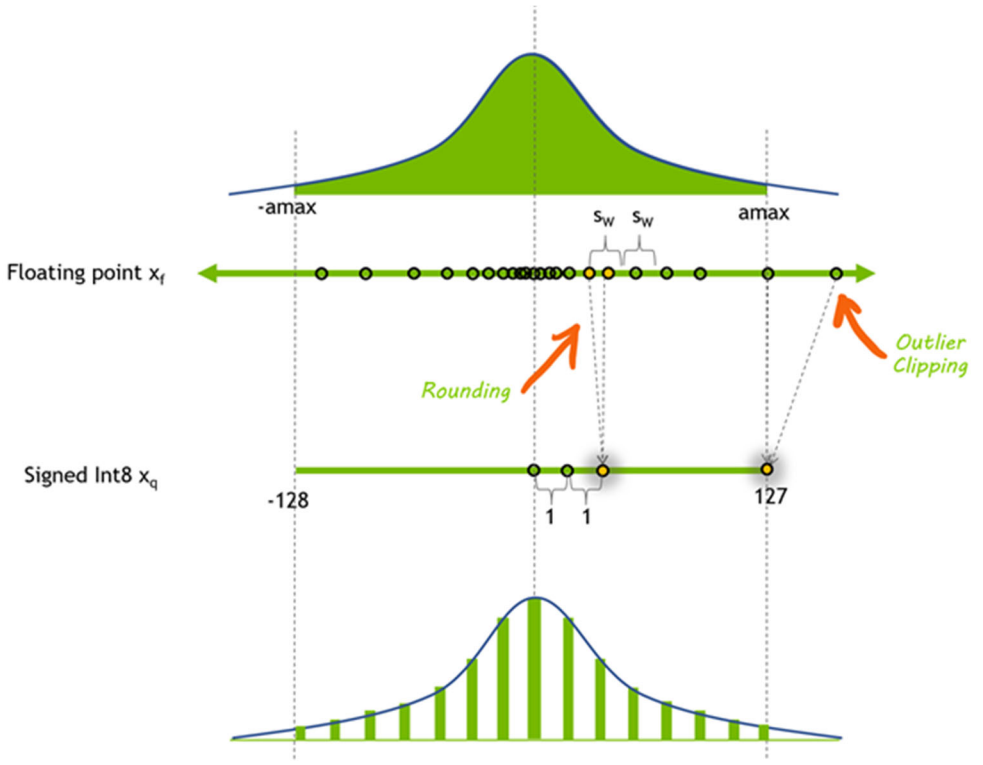
현재 공개된 Kimi K2의 성능이 Moonshot AI의 실제 최고 기술력을 전부 보여주는 것이 아닐 수 있다는 말이다.

4. Kimi K2의 전략적 승부수: 네이티브 INT4 양자화 심층 해설

(1) 양자화는 타협이 아닌 표준이 될수도

Moonshot AI의 Kimi K2 Thinking 모델이 채택한 '네이티브 INT4 양자화'라는 것이 이번 업데이트의 "킥"이다. 이는 거대언어모델(LLM) 업계가 '양자화' 기술을 바라보는 관점 자체를 근본적으로 바꾼 사건이라 할 만하다. 과거에 양자화는 어쩔 수 없이 선택하는 '타협안'에 가까웠다. 모델의 정확도를 약간 희생하는 대가로 추론 속도를 높이고 메모리 사용량을 줄이는 기술로 인식되었다.

그림 15. INT8 양자화는 아웃라이어를 자른 다음에 실수를 반올림해 256개의 정수로 표현하는 형태
정확도를 희생하는 대신에 모델의 크기와 연산량을 줄이기 위해 정수연산을 사용하는 것



자료: NVIDIA Developer, 미래에셋증권 리서치센터

하지만 Kimi K2는 이러한 통념을 깨뜨렸다. 이제 양자화는 모델의 성능 저하 없이 추론 효율을 극대화하고, 심지어 모델을 훈련시키는 과정까지 가속하는 '표준 패러다임'으로 진화할 가능성이 보인다. 모델의 파라미터가 조 단위를 넘어서고, 한 번의 질문에 더 깊이 생각하게 만드는 '추론 시간 스케일링' 기술이 발전함에 따라, 저비트(low-bit) 양자화는 더 이상 선택이 아닌 필수가 되는 것 같다는 말이다.

Kimi K2가 왜 양자화를 필사적으로 도입해야 했는지 이해하려면, 모델의 구조적 특징인 MoE(Mixture of Experts)와 그로 인해 발생하는 '메모리 병목 현상'을 알아야 한다. MoE는 거대한 단일 두뇌 대신, 각 분야의 '전문가' 여러 명으로 구성된 팀과 같다. 질문이 들어오면, 모든 전문가가 나서는 대신 해당 질문에 가장 적합한 소수의 전문가만 활성화되어 문제를 해결한다. Kimi K2는 48명의 전문가 중 단 1명만 활성화되는(1/48 희소성) 구조를 가진다. 이는 연산량(FLOPs) 측면에서 매우 효율적이다.

하지만 문제가 있다. 비록 소수의 전문가만 '일'한다고 하지만, 일단 이들을 불러오려면 모든 전문가의 정보가 담긴 거대한 '전화번호부'(모델 가중치)를 메모리에서 연산 장치(GPU 코어)로 가져와야 한다. GPU 코어의 연산 속도가 아무리 빨라도 메모리에서 데이터를 가져오는 속도가 느리면 전체 시스템이 느려진다. 이를 '메모리 바운드(memory-bound)' 상태라고 한다.

Kimi K2는 이 문제를 정면으로 마주했다. 기존의 FP8 정밀도로 모델 가중치를 표현하면 그 크기가 약 1TB에 달했다. 이는 최고급 서버의 단일 노드가 감당할 수 있는 거의 물리적 한계치다. 여기서 양자화가 해결책으로 등장한다. Kimi의 연구원들은 가중치를 INT4(4비트)로, 활성화(딥러닝 계산시의 중간값)는 BF16(16비트)으로 양자화하는 W4A16 방식을 사용했다. 이로써 모델 가중치의 크기가 절반인 약 500GB로 줄어들었다. 가중치(전화번호부)를 줄이니 메모리 대역폭의 정체가 풀리고, 결과적으로 추론 지연 시간을 획기적으로 단축시킬 수 있게 된 것이다.

(2) 왜 PTQ는 실패했고, QAT가 해답이었나

개발팀은 처음에는 더 간단한 방식인 '사후훈련 양자화(PTQ)'를 시도했다. PTQ는 이미 훈련이 끝난 모델을 번역기를 돌리듯 저정밀로 변환하는 방식이다. 하지만 이 접근법은 Kimi K2의 긴 문맥 추론 능력과 맞물려 두 가지 치명적인 문제를 일으켰다.

PTQ로 인한 미세한 오차는 '복사본의 복사본'을 계속 만드는 것과 같다. 처음 몇 장은 괜찮지만, 수백 번 복사를 거듭하면 원본을 알아볼 수 없게 된다.

마찬가지로, 긴 대화나 복잡한 추론을 수행할 때 이 작은 오차들이 눈덩이처럼 불어나 최종 결과물의 품질을 심각하게 저하시켰다. PTQ는 일부 데이터를 기반으로 '변환시의 규칙'을 만드는데, 이 데이터가 MoE 구조의 특정 '전문가'들만 편애할 수 있는 문제다. 이는 특정 분야에서는 잘 작동하지만 다른 분야 전문가들의 성능을 왜곡시켜 모델 전체의 균형을 무너뜨렸다.

이러한 한계에 부딪힌 개발팀은 훨씬 정교한 방식인 '양자화 인식 훈련(QAT)'으로 전환했다. QAT는 모델을 훈련시키는 단계부터 "너는 나중에 INT4라는 저정밀 환경에서 살아야 한다"는 사실을 계속 주입하는 것과 같다.

아래의 과정을 반복하면, 모델은 양자화로 인해 사라질 소수점 아래 미세한 정보에 의존하지 않고, 더 중요하고 강건한 패턴에 집중해 학습하게 된다. 마치 어려운 환경에 미리 적응하는 예방 접종을 맞는 것과 같다. 그 결과, 최종적으로 INT4로 변환되어도 성능 저하를 최소화하고 안정적인 긴 문맥 추론이 가능해진 것이다.

표 9. QAT(양자화 인식 훈련)의 단계

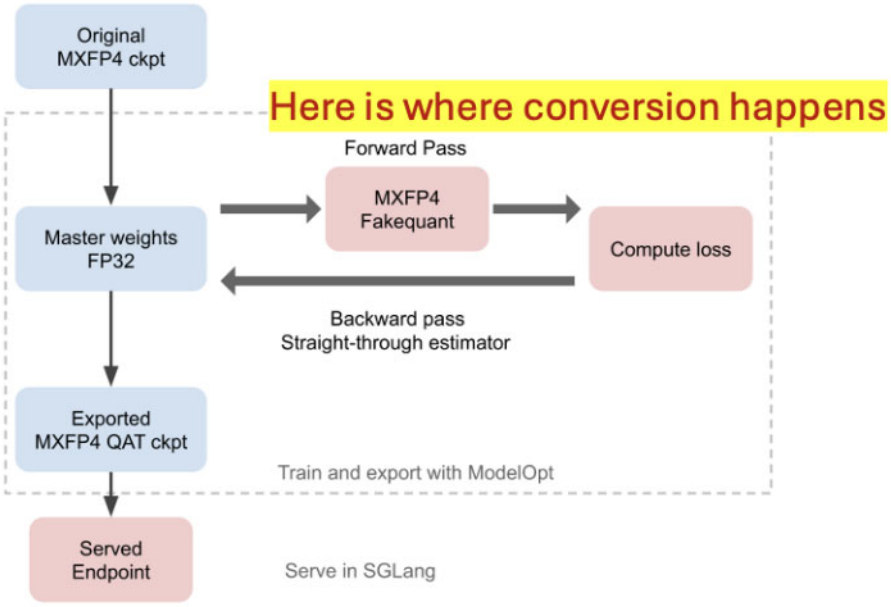
단계	과정	사용 정밀도	핵심 동작	목적/효과
1단계	훈련 시작	BF16 (고정밀)	원본 가중치 준비	높은 정밀도 유지
2단계	**순전파 – 압축**	**BF16 → INT4**	**고정밀 가중치를 INT4로 임시 압축**	**저정밀 환경 시뮬레이션**
3단계	**순전파 – 해제**	**INT4 → BF16**	**압축된 가중치를 즉시 고정밀로 다시 복원**	**미세한 정보 손실이 반영된 가중치 획득**
4단계	계산 수행	BF16 (손실 포함)	미세한 정보 손실이 반영된 가중치로 추론	양자화 효과 학습
5단계	역전파 – 업데이트	BF16 (원본)	손실을 원본 고정밀 가중치에 반영하여 업데이트	양자화에 강건한 가중치 학습

자료: 미래에셋증권 리서치센터

핵심 개념: 가짜 양자화(Fake Quantization)

실제로는 고정밀의 자료형으로 훈련하면서, 저정밀 때의 압축 효과를 시뮬레이션. 이로써 모델이 배포될 때, 양자화에 따른 손실에 미리 적응하게 하는 것.

그림 16. 양자화로부터 모델 정확도를 회복하는 학습 기법인 QAT의 표준적인 구현 방식
Kimi K2에도 쓰인 "가짜 양자화(해당 그림에서는 Fakequant라 표기)" 메커니즘을 시각화



자료: lmsys, 미래에셋증권 리서치센터

(3) QAT로 인한 효과와 전망

QAT의 도입은 예상치 못한 보너스를 가져왔다. 바로 '강화학습(RL)' 훈련 속도의 향상이다. 강화학습은 모델이 수많은 시도와 실패를 통해 정답을 찾아가는 과정인데, 여기서 가장 시간이 많이 걸리는 부분이 바로 '롤아웃(rollout)', 즉 실제 시도를 해보는 단계다. 이는 사후 학습 스케일링이 현재의 주류 체제라는 점에서 Moonshot의 미래를 밝게 하는 요인이다.

또한, Kimi는 엔비디아의 최신 Blackwell GPU가 지원하는 최신 자료형인 MXFP4 같은 더 '화려한' 포맷 대신, 범용성이 높은 INT4를 선택했다. 이는 단순한 기술적 선택이 아닌, 깊은 전략적 판단이었다. 가장 큰 이유는 구형 및 저가형 GPU를 포함한 거의 모든 하드웨어에서 안정적으로 작동하기 때문이다. INT4는 이미 검증된 고성능 커널(소프트웨어 라이브러리)이 많아 하드웨어 적응성이 뛰어나다.

INT4가 구형 기술이라고 해서 마냥 성능이 떨어지는 것은 아니다. INT4는 '스케일링 팩터(Scaling Factor)'라는 추가 정보를 활용한다. 정수(INT4) 값과 함께 "이 정수에 0.00123을 곱해서 사용하세요"라는 지침(스케일링 팩터)을 저장하는 방식이다. 이를 통해 정수 연산의 호환성과 속도를 누리면서도 부동소수점(FP4)과 유사한 정밀도를 표현할 수 있어, 성능 손실 없이 호환성을 확보할 수 있다.

사실, 이러한 결정은 미국의 첨단 GPU 구매 제재에 직면한 현실을 극복하기 위한 필연적인 선택이었다. 하지만 그 결과, 최첨단 AI 모델을 더 저렴하고 보편적인 구형 하드웨어에서도 구동할 수 있게 만들어, 전 세계 AI 생태계 전체의 기술 접근성을 높이는 긍정적인 부수 효과를 낳았다.

Kimi K2가 선보인 "W4A16(가중치 저장은 4비트로, 실제 계산은 16비트로) 양자화"는 끝이 아닌 시작일 가능성이 크다. Moonshot AI는 이미 가중치뿐만 아니라 활성화 값까지 더 낮은 비트로 압축하는 W4A8, 심지어 W4A4 방식까지 연구를 진행하고 있다. 앞으로 FP4 네이티브 연산을 지원하는 새로운 칩이 출시됨에 따라, Kimi의 양자화 기술 경로는 계속해서 진화하며 AI 모델의 효율성과 접근성을 새로운 차원으로 이끌 것이다.

5. 기술적 수렴 진화: 테슬라와 Kimi의 평행 이론

(1) AI 기반 '지속 가능한 풍요'를 위한 것은 일단 반도체

2025년 11월 테슬라는 '지속 가능한 풍요(Sustainable Abundance)'라는 명확한 목표 아래 AI 기술, 독자적인 하드웨어 엔지니어링, 그리고 대규모 생산 능력을 결합한 통합 전략을 제시했다. 전략의 핵심은 "현실 세계의 데이터"를 기반으로 하는 '실세계 AI(Real-world AI)'이며, 이는 자율주행, 로보틱스, 에너지 솔루션을 관통하는 중심축이다.

이를 위한 테슬라 전략의 근간 중 하나는, 기성품 범용 반도체로는 달성 불가능한 절대적 효율성과 강건성을 확보하기 위한 독자적인 칩 설계에 있다. 테슬라가 맞춤형 칩을 설계하는 근본적인 이유는 로봇(차량 포함)에 탑재되는 추론 칩이 데이터센터의 칩과는 완전히 다른 차원의 요구사항을 가지기 때문이다.

데이터센터 GPU의 연간 고장률은 산업 평균 9%, 이상적인 환경에서도 2% 수준이나 된다. 데이터센터에서 이런 고장이 나면, 예비 칩으로 즉시 교체된다. 반면, 차량과 같은 로봇은 외부의 혹독한 환경에 그대로 노출된다. 고장은 치명적인 결과를 초래할 수 있다. 실제로 다수의 로보틱스 기업들은 기성품 칩을 재부팅 없이 몇 시간 이상 안정적으로 구동하는 데에도 어려움을 겪는다.

따라서, 칩의 실제 성능과 안정성은 사양표만으로는 판단할 수 없다. 진정한 평가는 보드 지원 패키지(BSP)를 처음부터 직접 작성해보는 'acid test'를 통해 이루어진다. 이 실전 검증을 통해야만 데이터시트 뒤에 숨겨진 문제점을 파악할 수 있다. 이런 관점에서 수년 전 설계된 테슬라의 HW3조차 현재 시장의 어떤 칩보다 아직도 한 세대 앞서 있다는 평가가 현장 엔지니어들에게는 존재한다. 이는 테슬라의 칩 설계가 실제 작동 환경에서의 안정성을 최우선으로 고려함을 시사한다.

(2) 핵심 설계 철학: 정수(Integer) 연산과 QAT의 필연성

테슬라 칩 설계 철학의 핵심은 정수(INT: Integer) 연산을 중심으로 AI 추론 효율을 극한으로 끌어올리는 전략이다. 일론 머스크는 "정수 연산은 근본적으로 부동소수점 연산보다 효율적"이라며, "정수는 전력 효율과 실리콘 효율이 훨씬 높지만, 이를 위해서는 정수 추론에 맞게 AI 모델을 훈련시켜야 한다"고 강조했다. 이는 매우 중요한 전략적 차별점이다.

테슬라의 FSD 모델은 실시간으로 "저것이 그림자인가, 도로의 파인 곳인가?"와 같은 수백만 건의 판단을 계속 실시간으로 내려야 한다. 이 과정에서 완벽한 소수점 단위의 정밀도보다 반응 속도와 에너지 효율이 더 중요하다. 객체와의 거리가 '3.141592미터'임을 아는 것보다 '약 3미터'임을 신속하게 파악하는 것이 안전을 위해 더 효과적이기 때문이다.

이러한 철학은 칩의 물리적 설계에 직접 반영된다. 테슬라의 FSD 칩은 정수 연산(INT)을 채택하여 제한된 전력과 공간 내에서 최대의 병렬 연산 처리량을 달성하려고 한다. 이는 더 빠른 반응 속도, 에너지 효율 증대, 그리고 비용 및 발열 감소로 이어진다.

테슬라의 정수 연산 활용은 소수점을 버리는 반올림만이 아니다. 핵심은 '양자화 인식 훈련(Quantization-Aware Training, QAT)'이라는 정교한 기법에 있다. Kimi K2의 핵심인 그 QAT가 맞다.

물론, 원래 안정적인 훈련을 위해서는 부동소수점(FP32) 연산이 필요하다. 훈련 중 모델의 가중치는 0.0000123과 같이 아주 미세하게 조정되는데, 정수로는 이러한 변화를 표현할 수 없기 때문이다. 가장 단순한 방법은 훈련을 모두 마친 뒤 마지막에 한 번 변환하는 '훈련 후 양자화(PTQ)'지만, 이는 복잡한 FSD 모델에서 막대한 정보 손실을 유발해 정확도를 심각하게 저하시킨다. PTQ가 실패한 것은 우리가 위에서 Kimi의 사례에서 확인했다.

테슬라의 QAT 또한 Kimi와 같이 "훈련하는 동안 변환될 것을 미리 알고 대비"시키는 방식이다. 마치 비행 시뮬레이터와 같다. 조종사(신경망 모델)는 최첨단 시뮬레이터(GPU, FP32 연산)에서 훈련을 하지만, 이 시뮬레이터는 실제 비행기(FSD 칩, INT8 연산)가 갖고 있는 제약 조건을 똑같이 모방하면서 훈련을 하는 것이다.

표 10. 테슬라의 QAT 과정도 Kimi와 대동소이

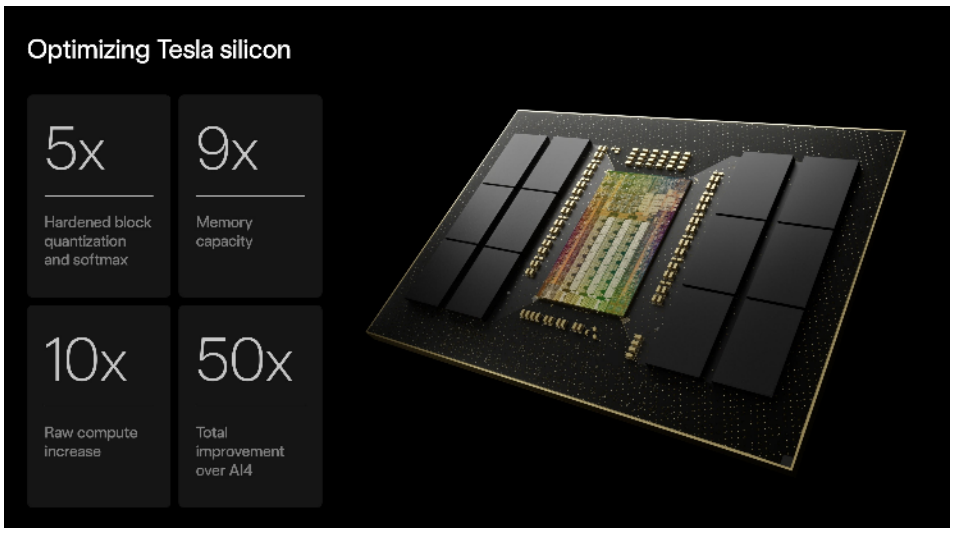
단계	과정	사용 정밀도	주요 동작	목적
1단계	훈련 시작	FP32 (고정밀)	GPU 클러스터에서 고정밀 부동소수점으로 훈련 시작	높은 정밀도 유지
2단계	**'가짜' 양자화 삽입**	**FP32 → INT8 → FP32**	**순전파 중 가중치/활성화값을 INT8로 임시 변환 후 즉시 FP32로 복원 (Fake Quantization)**	**저정밀 환경 시뮬레이션**
3단계	오차 학습	FP32 (오차 반영)	가짜 양자화에서 발생한 미세 오차를 줄이는 방향으로 가중치 업데이트	양자화 오차에 대한 내성 강화
4단계	역전파	FP32 (고정밀)	그래디언트 계산 및 가중치 업데이트는 정밀한 FP32로 수행	학습 안정성 및 정확도 유지
5단계	최종 배포	FP32 → INT8	양자화 내성이 극대화된 FP32 모델을 INT8로 최종 변환하여 FSD 칩에 탑재	정확도 손실 최소화하며 배포

자료: 미래에셋증권 리서치센터

이러한 접근 방식은 테슬라가 AI 모델을 처음부터 '양자화를 인지하는 훈련' 방식으로 개발해야 함을 의미하며, 이는 기성품 칩과 범용 AI 모델을 사용하는 타사와의 핵심적인 기술적 차별점을 만든다.

이러한 설계 철학과 기술적 구현을 집대성한 것이 바로 차세대 AI 칩 AI5다. 2027년 본격 대량 양산을 목표로 하는 AI5는 현행 AI4 대비 50배의 총체적 성능 향상을 목표로 한다.

그림 17. 50배의 도약을 목표로 하는 테슬라 차세대 칩 AI5 칩
단순 연산 능력(10배)의 증가를 넘어 양자화 같은 특화 연산 효율(5배)을 함께 극대화한 결과



자료: 테슬라, 미래에셋증권 리서치센터

(3) 핵심 원리의 공유: QAT라는 동일한 설계도

서로 다른 산업, 다른 목표를 가진 테슬라와 Moonshot AI(Kimi)는 거대 AI 모델의 효율화라는 과제 앞에서 놀랍도록 동일한 해법을 채택했다. 양자화 인식 훈련(QAT)이다. 두 회사가 사용하는 QAT의 기술적 설계도는 본질적으로 같다.

두 회사가 왜 같은 해결책에 도달했는지를 이해하려면, 그들이 부딪힌 문제가 본질적으로 같다는 것을 알아야 한다. 바로 현대 컴퓨팅의 근본적 난제인 '메모리 대역폭 병목 현상'이다. 데이터를 계산하는 비용보다 옮기는 비용이 더 비싸다는 문제에 직면했고, 데이터(가중치)의 크기를 줄이는 것이 가장 효과적인 해결책이라는 동일한 결론에 도달한 것이다.

테슬라 (FSD): 실시간 비디오 처리의 병목

차량의 제한된 전력과 칩 면적 안에서, 수많은 카메라로부터 들어오는 방대한 영상 데이터를 실시간으로 처리해야 한다. 연산 속도 자체보다 이 데이터를 메모리에서 연산 유닛으로 옮기는 과정이 더 큰 병목으로 작용한다. 가중치를 저비트로 압축하면 이 데이터 이동량이 줄어들어 전체 시스템의 반응 속도와 효율이 극적으로 향상된다.

Moonshot (Kimi): 거대 언어 모델 서비스의 병목

수조 개의 파라미터를 가진 거대 모델을 서비스해야 한다. 사용자의 질문에 빠르게 답하려면 이 거대한 가중치(수백 GB ~ TB)를 메모리에서 GPU로 신속하게 불러와야 한다. 이 역시 메모리 대역폭이 응답 속도를 결정하는 핵심 병목이다.

테슬라와 Kimi는 각자의 정글에서 생존하고 번성하기 위해 진화한 결과, 놀랍도록 유사한 '생존 도구(QAT)'를 채택했다. 이는 QAT가 특정 회사나 분야에 국한된 트릭이 아니라, 거대 AI 시대를 관통하는 보편적이고 필연적인 공학적 해법임을 증명하는 강력한 사례다. 하드웨어의 물리적 한계라는 보편적 문제 앞에서, 가장 효율적인 해결책으로 수렴하는 기술적 진화의 명백한 증거다.

테슬라와 Kimi는 각자의 정글에서 생존하고 번성하기 위해 진화한 결과, 놀랍게도 같은 '생존 도구(QAT)'를 채택했다. 이는 QAT가 특정 회사나 분야에 국한된 트릭이 아니라, 거대 AI 시대를 관통하는 보편적이고 필연적인 공학적 해법임을 증명하는 사례다. 하드웨어의 물리적 한계라는 보편적 문제 앞에서, 가장 효율적인 해결책으로 수렴하는 기술적 진화의 명백한 증거다.

III. AI Agent에게 진짜 필요한 것은

1. 인간과 AI를 가르는 것들

(1) AI가 갖지 못한 것: 해마

인간의 뇌는 하루에 엄청난 양의 데이터를 처리한다. 눈은 매초 수백만 개의 픽셀을 받아들이고, 귀는 음파의 수많은 진동을 감지한다. 피부는 온도, 압력, 통증을 동시에 신호한다. 그런데 놀랍게도 우리는 이 모든 정보를 다 기억하지 않는다. 오히려 우리는 선택적으로 기억한다.

이 현상의 핵심에는 해마(Hippocampus)라는 뇌 구조가 있다. 해마는 매우 명확하고 구체적인 역할을 수행한다. 첫째, 수면 중 해마는 뇌의 장기기억 영역으로 경험을 전달한다. 하지만 이때 모든 세부사항을 전달하지 않는다. 정보의 핵심 구조만 정교하게 선택하여 전달한다. 둘째, 배경 정보(어제 본 구름, 잠깐 들린 소음 등)는 버리고 본질적인 패턴만 추출한다. 중요한 개념, 의미 있는 상황, 학습할 가치가 있는 경험이라고 판단한 것들만 남기는 것이다. 셋째, 개별 사건들을 더 큰 패턴으로 조직한다. 그 날 있었던 특정 대화 내용을 “그날 나는 실수했다”, “이 사람은 신뢰할 수 없다” 같은 일반화된 지식으로 변환한다.

결과적으로 인간은 같은 양의 경험 데이터로도 효율적으로 배울 수 있다. 신경과학자들은 이 과정을 ‘시스템 통합’이라고 부른다. REM 수면 중 해마의 신경들이 재활성화되면서 피질에 정보가 점진적으로 옮겨지는데, 이것이 인간의 효율적인 학습을 가능하게 한다. 우리의 뇌는 모든 정보를 저장하지 않고, 중요한 신호만 걸러서 저장함으로써 제한된 자원으로 강력한 학습 능력을 발휘한다.

그런데 현재의 대형 언어 모델(LLM)에는 해마가 없다. 더 정확히 말하면, 해마처럼 작동하는 메커니즘이 없다. 안드레 카파시의 의견에 따르면, 이러한 메커니즘의 부재는 모델의 성능에 여러 문제를 일으킨다. 먼저, 트랜스포머 아키텍처의 모든 토큰은 원칙적으로 동등하게 처리된다. 물론 어텐션 메커니즘이 있지만, 이것은 컨텍스트 내에서만 중요도를 판단한다. 따라서 모델은 전체 학습 과정에서 어떤 정보가 중요하고 어떤 정보가 배경인지를 구조적으로 구분하지 못한다. 노이즈와 신호를 동등하게 취급하게 된다는 뜻이다.

또 하나의 문제는 파괴적 망각이다. 새로운 데이터로 파인튜닝할 때 LLM은 기존 능력을 잃는다. 예를 들어 의료 LLM을 구축하기 위해 의학 데이터로 파인튜닝하면, 일반적인 언어 이해 능력이 저하된다. 이는 모델의 학습 방식에서 오는 구조적 문제다. 신경망은 가중치를 업데이트할 때, 새로운 데이터의 신호에 맞춰 기존 가중치를 덮어씌운다. 반면 해마는 그렇지 않다. 해마는 새로운 정보를 받아 처리할 때 기존에 가지고 있던 핵심 지식을 건드리지 않는다. 오히려 새로운 정보를 선택적으로 저장하면서도 기존 구조는 유지한다. 따라서 여러 분야에서 모두 높은 수행 능력을 보이는 언어 모델을 구축하는 것이 어려울 수밖에 없다.

(2) 강화학습 모델은 진정한 추론을 하지 못한다?

그런데 증류가 이러한 어려움을 해결할 수 있다는 주장이 제기됐다. '강화학습이 LLM의 추론 능력을 베이스 모델보다 강화하는가?' 라는 칭화대 LeapLab 논문은 강화학습이 샘플링 효율성, 즉 이미 아는 영역에서 답을 더 빨리 찾는 능력을 개선할 뿐, 학습 데이터를 벗어난 영역에서의 답변 성능을 개선하는 것은 아니라고 지적했다. 문제 은행을 더 잘 푸는 모델이 될 뿐, 진정한 의미의 추론 능력은 개선되지 않았다는 것이다.

비슷한 결과가 최근 AlphaEvolve 후속 논문에서도 나타났다. 테렌스 타오와 구글 딥마인드가 진행한 AlphaEvolve의 대규모 수학 탐색 및 발견 논문에서는 AlphaEvolve가 이미 학습한 문제에 대해 처음부터 최적해를 알고 있었다고 언급한다. 따라서 유명한(?) 문제에 대한 새로운 접근법을 찾기 위해서는, AlphaEvolve가 문제를 알아보기 어렵게 수정해야 했다고 한다.

이렇게, 강화학습 모델들의 추론 성능은 진짜 추론이 아닐 가능성이 높다는 것이 일관된 주장이다. 너무 많은 정보를 가지고 있기 때문에 대부분의 영역에서 논리적인 것처럼 보일 뿐이라는 것이다. 그리고 칭화대 연구진은 강화학습을 대신해 진정한 추론 능력을 개화시켜 줄 대안으로 증류를 제시했다.

논문에서는 On-Policy 증류라는 새로운 방식을 사용했는데, 이는 지식을 전수받는 '학생' 모델이 문제를 풀어보고, 이에 대한 피드백을 '교사' 모델이 제공하는 구조다. 이 과정을 통해 학생 모델은 자신의 실수로부터 배울 수 있으며, 훈련이 곧 추론 과정이기 때문에 추론 능력을 갖추게 된다.

이러한 On-Policy 증류 기법을 일반적인 강화학습과 비교하면 모델의 성능이 좋을 수밖에 없는 이유를 알 수 있다. 일반적인 강화학습은 모델의 답변이 맞았는지 틀렸는지를 0 또는 1의 비트 한 개로 알려주지만, On-Policy 증류는 교사 모델이 학생 모델의 풀이 과정이 맞았는지 틀렸는지를 각각 알려준다. 답변 하나당 더 정교한 피드백을 받을 수 있을 뿐 아니라, 각 풀이 단계별 피드백을 받을 수 있다. 추론 능력을 훈련하기 때문에, 모델이 추론 능력을 갖추게 되는 것이다.

논문에서 5,400억 파라미터 모델과 이보다 7,000배 작은 증류 모델의 성능을 비교한 결과, 7,000배 작은 모델이 few-shot 성능을 초월했으며, 동시에 9~30배의 비용 효율성을 달성했다. 이는 학생 모델이 증류 과정에서 교사 모델이 가지고 있지 않았던 새로운 추론 패턴을 학습했기 때문이었다.

그렇다면 더 이상 큰 모델이 필요하지 않다는 것일까? 그렇지 않다. 추론 능력을 갖춘 학생 모델의 성능을 끌어올리려면, 이보다 더 큰 교사 모델이 필요하다. 학생 모델에게 단계별 피드백을 제공하려면 교사 모델의 성능도 그에 걸맞은 수준이 되어야 하므로, 더 좋은 결과를 원한다면 오히려 더 뛰어난 교사 모델을 얻어내야 한다.

그러나 더 크고 뛰어난 교사 모델을 만들기까지는 넘어야 할 산들이 있다. 대표적인 것이 데이터다.

2. 현실적 과제들

(1) 데이터 부족 문제와 모달리티 확장의 필요성

현재의 정론은 텍스트 데이터 수집이 끝난 상태라는 것이다. 데이터의 양보다는 질의 관점에서 그렇다. 인터넷에 공개되어 있는 데이터는 다양성이 매우 떨어질 뿐 아니라, 의료 데이터와 같은 특수한 분야의 데이터를 구하기도 어렵다. 효율적인 모델 학습에 고품질 데이터가 필수적이라는 점을 상기하면 이는 매우 심각한 제약이다. 주요 파운데이션 모델 개발사들이 애널리스트나 의사들에게 돈을 지불하면서까지 데이터를 수집하는 이유가 여기에 있다. 그만큼 고품질 데이터를 구하기가 어렵고, 중요하다는 뜻이다.

다만 이 부분은 모달리티 확장을 통해 해결할 수 있다. 현재 언어 모델을 학습시키는 데이터의 대부분은 텍스트 데이터다. 그러나 모델 학습 데이터의 모달리티를 영상으로 확장하면 새로운 지평을 열 수 있다. 이는 단순히 데이터의 양만 늘어나는 것이 아니다. 지금까지의 모델들이 가지고 있었던 고질적인 문제를 근본적으로 해결하는 열쇠가 될 수 있다.

텍스트 기반의 정보들은 정적이다. 변화가 무엇인지 학습하는 것이 태생적으로 어려운 환경이라는 의미다. 예를 들어, 인간은 유리잔을 바닥에 떨어뜨렸을 때 높은 확률로 잔이 깨질 것임을 알고, 다치지 않기 위해 발을 뒤로 빼거나 유리를 밟지 않기 위해 가만히 정지한다. 물론 언어 모델도 이 정도는 알 수도 있다. 유리가 깨져서 발이 다쳤어요, 와 같은 글은 인터넷에 많았을 테니 말이다.

그러나 인간은 유리잔이 대략 어느 위치에서 떨어졌는지, 유리잔의 두께나 강도가 어느 정도인지, 바닥이 얼마나 충격을 흡수할 것인지 등의 정보를 이미 알고 있다. 따라서 우리는 유리 파편이 어느 정도로 튈지 예상하고, 그에 따라 자연스럽게 발을 얼마나 뒤로 뺄지, 또는 반대로 발을 유리잔에 뻗어 충격을 조금이라도 막아볼지 결정할 수 있다. 그리고 우리는 이 모든 판단과 행동을 유리잔이 바닥에 닿기 전에 해낸다.

그런데 이런 반응을 보이는 사람에게 “너 참 훌륭한 추론 능력을 가졌구나”라고 칭찬하지는 않는다. 인간에게는 매우 상식적인 행동이기 때문이다. 그러나 이러한 상식이 현 시점 AI에게 가장 부족한 영역이다. 그리고 AI 모델에게 상식을 탑재시키려면 두 가지가 필요하다고 판단한다: 증류, 그리고 연속 학습(continual learning)이다.

연속 학습이란, 모델이 사용자와 대화하면서 얻은 정보를 바탕으로 실시간으로 학습하는 모델을 말한다. 현재의 모델은 훈련 시점까지의 데이터는 모두 기억하지만, 당장 5분 전의 맥락은 기억하지 못하는 일종의 선행성 기억상실증 상태다. 물론 최근에는 모델들이 과거 대화를 참조할 수 있도록 하는 기능들이 추가되고 있지만, 이는 모델의 손에 컨닝 페이퍼를 한 장 더 쥐여준 것일 뿐이다.

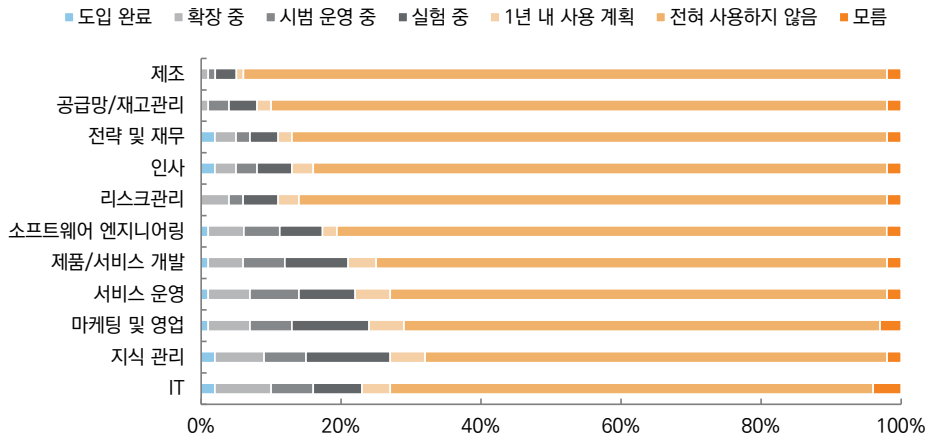
모델들이 손에서 컨닝 페이퍼를 내려놓고 현재를 더듬어가며 세상을 배우게 하려면, 학습 과정에서 필연적으로 시간의 차원이 추가되어야 한다. 다시 한번 모달리티 확장으로 돌아오게 되는 지점이다. 영상은 시간에 따른 변화와 인과를 나타낼 수 있는 대표적인 모달리티이며, 여기에 실시간 기능이 더해진다면 모델은 작용-반작용의 묘리를 직접 깨달을 수 있다. 그리고 이러한 경험들이 축적되면서 비로소 세상에 대한 진정한 이해에 도달할 수 있을 것이다.

결국 연속 학습을 통해 시간에 따라 경험을 축적하고, 이렇게 쌓인 방대한 학습 경험을 증류 과정을 통해 진정한 지식으로 압축하는 과정을 반복해야만 모델이 진정한 의미의 사고력을 갖출 수 있을 것이라고 예상한다. 그리고 이것은 파운데이션 모델 학습 과정에서뿐만 아니라, 궁극적으로는 모든 사용자 단에서 일어나야 한다. 그래야만 진정한 에이전트 시대를 열 수 있기 때문이다.

(2) 상식이 필요한 AI 에이전트

AI 산업에서 올해 화두가 되었던 것은 에이전트 경제였다. 그러나 11월 중순에 다다른 지금, 발표되는 자료를 보면 에이전트의 실질적 도입 및 발전 속도가 생각보다 더디다는 것을 알 수 있다. 맥킨지의 State of AI 보고서에 따르면, 조직 내에서 AI 에이전트를 '최소한 실험'하고 있다고 답한 조직은 전체 응답자의 62%에 달했고, 이중 23%는 에이전트 시스템을 조직 일부에서 '확장'하고 있다고 보고했다. 그러나 이들이 말하는 확장은 IT나 지식 관리와 같은 일부 기능에 국한되어 있었으며, 비즈니스별로 보았을 때 에이전트를 확장하고 있다는 응답이 10%를 넘는 조직은 없었다. 결국 에이전트는 아직까지 실험적 단계에 머무르고 있으며, 실제 사용되는 영역도 AI가 잘 한다고 알려진 코딩이나 데이터 관리 영역에 국한되어 있다. 개인 사용자가 에이전트를 적극적으로 활용하기가 어려운 이유다.

그림 18. 비즈니스 영역별 조직 내 AI 에이전트 도입 수준 비교



자료: Mckinsey&Co., 미래에셋증권 리서치센터

그러나 우리가 에이전트에게 기대하는 것은 코딩이나 데이터 관리만이 아니다. 에이전트를 사용하는 궁극적인 목적을 생각해 보면, 결국은 편리하기 위한 것이다. 나를 위해 물건을 대신 구매해 주든, 메일을 대신 보내주든, 결국 우리는 에이전트가 나의 수고로움을 덜어주기를 기대한다. 그러나 지금 우리는 에이전트에게 대부분의 일을 믿고 맡기지 못한다.

예를 들어 옷을 구매할 때 주소를 잘 등록하고, 할인쿠폰을 잘 적용하고, 중복 결제를 일으키거나 사이즈를 잘못 고르지는 않을지 등 상식적인 행동에서 오류가 발생할 수도 있다는 불안이 있다. 이는 인터넷에서 특정 데이터를 찾아달라고 할 때 느끼는 불안감과 차원이 다르다. 코딩이나 데이터 검색 및 정리와 다르게, 우리는 AI가 예를 들어 '물건을 주문할 때 택배를 어디에 두어 달라고 알림 메시지를 작성해야 한다'는 등의 '나만의 상식'이 없다는 것을 알고 있다.

우리는 범용 지식이 있는 모델을 가졌을 수는 있지만, 나에게 fit한 에이전트, 정말 나의 마음을 읽고 행동하는 수행 비서의 수준에는 도달하지 못했다. 그리고 이것이 가능하려면 연속 학습, 특히 내 데이터로 하는 연속 학습이 필요하다. 이러한 학습은 기반 모델에 영향을 주면 안 되기 때문에 별도로 운영되어야 하며, 개인화된 데이터이므로 보안을 위해 온 프레미스 형태가 선호될 수 있다. 간단히 말해, 내 개인 PC에서 내 데이터로 튜닝'되고 있는' 모델이 되어야 한다. 그 날 내가 하루 종일 경험한 것들을 공유하고, 이중 중요한 맥락을 추출하여 장기 기억으로 모델에 저장하면서 나에 대한 정보를 쌓아가야만, 우리가 기대하는 진정한 에이전트가 될 수 있다.

(3) 새로운 뇌 구조: 계층형 메모리 아키텍처

기반 모델 위에, 나에 대한 정보를 연속 학습하면서 모델이 나와의 추억을 쌓아나갈 수 있게 하려면, 메모리 구조가 이를 지원하는 형태여야 한다. 인간이 피상적인 단기 기억을 해마를 통해 보다 추상화된 장기 기억으로 저장하는 것처럼, 메모리에도 이와 유사한 계층 구조를 구축해 줄 필요가 있다. 최근 대화나 상호작용에 해당하는 데이터는 가장 빨리 불러올 수 있도록 하고, 지난 대화들에서 반복되는 패턴이나 습관에 해당하는 데이터는 다음 레이어에, 사용자의 직업, 관심사 등의 기초 데이터는 그 다음 레이어에 구분해서 저장하는 방식이 필요하다. 마지막으로 과거 대화 기록 전체는 검색 기능을 통해 필요할 때만 불러올 수 있도록 별도의 스토리지로 저장해두면 된다.

예를 들어 단기 기억 데이터는 대역폭이 매우 넓어 빠르게 정보를 불러올 수 있는 HBM으로, 패턴이나 습관 레이어는 DRAM, 기초 데이터는 SSD 등으로 구분해 저장하는 방법이 있을 수 있다. 중요한 것은 기억의 계층화와, 데이터가 각 계층을 이동하는 과정에서 모델이 증류를 거쳐야 한다는 점이다.

다만 현재는 개별 사용자가 이러한 계층형 메모리 구조를 구축하기도 어려울 뿐 아니라, 이러한 메모리 구조를 이해하고 활용할 수 있는 모델 아키텍처도 필요하다. 또한 이러한 아키텍처를 넘나들 수 있는 증류 인프라의 성숙도도 낮다. 마지막으로, 이러한 개인화된 접근이 모델의 일반적 능력을 손상시키지 않도록 LoRA와 같은 방법을 통해 개인화된 업데이트 변수를 별도로 운영하는 시스템도 필요하다.

이처럼 진정한 에이전트 시대로 나아가기 위해서는 기반 모델의 성능보다도 아키텍처와 하드웨어 구조의 혁신이 필요하다. 특히 개인 소비자가 에이전트의 성능을 체감하려면, 이러한 계층형 메모리를 직접 구축하거나 클라우드 형태로 서비스를 제공받아야 한다. 비용적 부담은 차치하더라도 이를 십분 활용할 수 있는 아키텍처를 함께 구축해야 유의미한 결과를 얻을 수 있으므로, 에이전트는 B2B 시장에서부터 개화할 가능성이 높다.

(4) 지능에서 지성으로

AI의 지능을 판단하기 위한 여러 벤치마크들이 있지만, 새로운 모델이 나올 때마다 이러한 벤치마크들은 점차 포화되고 있다. 이런 모습은 마치 AI가 이미 인간을 뛰어넘은 것처럼 보이지만, 정작 AI는 우리의 일상에서 일정 범위 이상으로 침투하지 못하고 있다. 또한 모델의 성능을 한 단계 증가시키기 위해 점점 더 많은 데이터, 더 큰 모델이 필요해지고 있어 이러한 방법으로 AGI를 달성할 수 있겠는가 하는 우려도 있다.

이 모든 현상은 지능에 대한 단편적 이해에 기반한다. 지금까지 포화된 벤치마크는 모델의 추상화 능력이나 진정한 의미의 추론 능력을 구별해 낼 수 없었다. 그러나 칭화대 연구진이나 딥마인드 팀이 발견했듯, 모델이 이미 알고 있는 문제는 추론이 아닌 외운 답을 제공하고 있었다. 이는 인간이 가지고 있는 여러 지성의 영역 중 암기력이 뛰어나다는 것을 보여주기는 하지만 다른 많은 영역, 예컨대 추상화 능력과 같은 것은 여전히 부족하다.

AI가 진정한 지능으로 나아가려면, 단순히 모델을 크게 만드는 것이 아니라 그 구조와 학습 방식 자체를 재설계해야 한다. 증류를 통한 진정한 추론 능력의 습득, 연속 학습을 통한 경험의 축적, 그리고 모달리티 확장을 통한 인과관계의 이해, 이 세 가지가 모두 결합될 때 비로소 AI는 상식을 갖춘 에이전트가 될 수 있을 것이다.

현재의 우리는 거대한 데이터 저수지 위에 세워진 모델들을 갖고 있지만, 정작 그 모델들이 물을 마실 수 있는 구조는 부족하다. 인간의 뇌처럼 경험을 지식으로 변환하고, 패턴을 이해하고, 새로운 상황에 대응할 수 있는 구조가 필요하다. AI 모델들이 세상에 대한 상식을 기반으로 소통할 수 있어야만 진정한 의미의 '인간 수준의 지능'을 달성했다고 볼 수 있을 것이다.

IV. AI Peer Table

표 11. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
최종 소비자(온디바이스, 앱)											
애플	AAPL US	273.47	5,930	1.5	10.5	17.3	28.7	22.0	9.7	01-30	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	430.60	2,101	-3.4	-1.2	26.9	28.9	30.4	6.6	01-29	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
Pony AI	PONY US	14	9	-6.6	-37.7	-10.2	-22.9	-	-3.6	11-25	중국의 자율주행 선도기업 중 하나
삼성전자	005930 KS	102,700	608	3.1	8.6	43.5	82.4	107.7	96.2	01-07	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
듀오링고	DUOL US	186.58	13	-3.7	-43.5	-43.1	-64.9	-42.3	-42.5	02-27	AI 캐릭터와 상호작용하는 학습 플랫폼
크래프톤	259960 KS	274,750	13	3.7	-9.2	-12.8	-25.6	-6.5	-12.1	02-11	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
쇼피파이	SHOP US	156.59	300	0.3	1.9	4.3	45.4	36.1	47.3	02-11	이커머스 콘텐츠 생성의 자동화(Shopify Magic AI)
코인베이스	COIN US	304.00	120	3.0	-14.8	-7.0	18.3	6.8	22.4	02-13	AI 에이전트간 자율 암호화폐 거래 인프라 구축
비즈니스 효율화											
크라우드스트라이크	CRWD US	545.86	201	2.5	7.3	26.3	24.0	56.9	59.5	12-02	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	314.98	432	1.4	14.2	31.9	23.3	53.2	47.0	01-28	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	246.53	226	2.2	1.2	1.0	-22.7	-32.0	-28.4	12-18	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	218.40	456	0.2	-5.6	-8.6	-16.0	0.4	-6.7	01-28	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,537.10	106	4.8	4.5	9.4	0.9	-15.3	-15.8	01-14	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
일루미나	ILMN US	121.98	27	0.7	28.4	20.6	51.9	-16.5	-8.7	02-06	AI 기반 첨단 유전체 분석
HD현대	267250 KS	225,500	18	3.4	44.3	69.8	183.1	211.1	195.4	02-06	AI, 로보틱스, 디지털 트랜스포메이션 관련 노력
AI Ops											
클라우드플레어	NET US	225.68	116	-0.1	1.4	13.2	50.1	139.7	109.6	02-06	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	184.17	644	5.2	3.9	-0.1	43.8	203.4	143.5	02-03	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
파운데이션 모델											
알파벳	GOOGL US	286.71	5,083	0.7	17.4	42.1	80.1	61.0	52.0	02-04	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	609.01	2,253	-1.6	-14.9	-21.9	-7.0	5.3	4.3	01-29	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	154.80	558	-6.2	-4.8	25.1	24.8	73.8	91.0	11-25	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
네이버	035420 KS	263,500	41	-0.8	0.0	17.1	40.0	45.4	33.2	02-06	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)

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

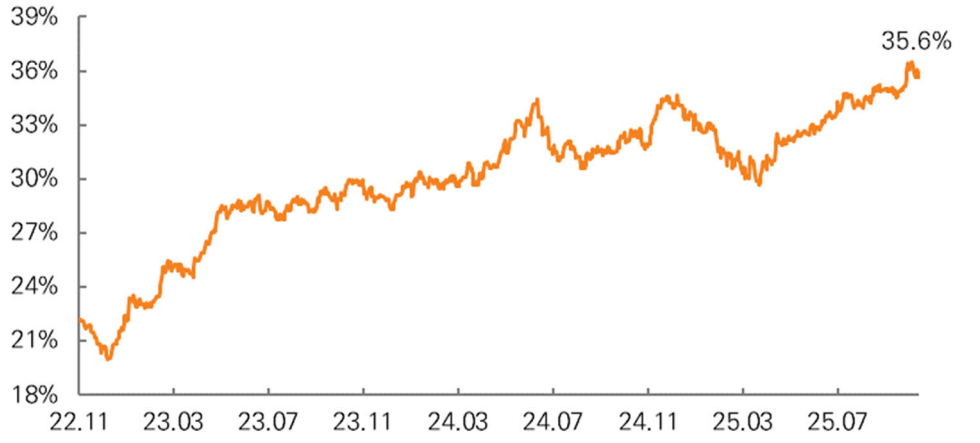
표 12. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
클라우드											
마이크로소프트	MSFT US	511.14	5,575	2.8	-0.6	-1.7	14.2	21.1	21.9	01-29	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	244.20	3,831	0.5	11.0	8.7	15.5	14.1	11.3	02-06	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	226.99	950	-6.9	-26.3	-6.9	40.4	20.9	37.6	12-09	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	20,705.00	280	-11.1	-6.9	39.4	172.7	127.5	126.3	02-12	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	193.80	6,911	3.0	2.9	6.7	49.2	32.5	44.3	11-20	AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자
브로드컴	AVGO US	355.22	2,462	-0.1	-0.4	15.1	53.5	106.6	54.3	12-11	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	258.89	618	8.9	19.6	40.4	130.2	85.9	114.3	02-04	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	37.89	265	1.7	1.8	70.5	68.0	52.0	89.0	01-30	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	244.90	403	2.8	27.0	97.2	153.1	146.1	191.8	12-18	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	611,000	445	2.5	44.3	120.1	208.8	237.3	254.8	01-22	HBM 부문 전세계 최강자
시놉시스	SNPS US	398.81	109	0.7	-11.0	-35.5	-22.7	-28.5	-17.8	12-10	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
케이던스 디자인 시스템	CDNS US	315.54	126	-2.7	-5.0	-9.6	-1.0	3.0	5.0	02-18	반도체 EDA 부문 리딩 업체
암페놀	APH US	141.92	255	2.8	14.5	29.4	66.0	97.8	105.5	01-28	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	173.37	97	-5.3	-3.1	26.2	64.3	39.4	52.8	02-12	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	148.75	232	-6.0	-13.5	5.0	17.4	10.1	20.6	02-05	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	89.33	113	-4.3	-0.1	12.7	36.6	-0.5	-18.9	12-02	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
코어위브	CRWV US	85.43	61	-20.1	-39.7	-27.5	35.0	-	-	02-13	AI용 클라우드 인프라 제공업체
중과서광	603019 CH	106.24	32	-7.2	-7.4	52.4	65.7	29.0	47.5	03-04	중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체
SMIC	981 HK	73.00	133	-5.1	-8.9	40.2	71.2	168.9	129.6	11-13	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,465.00	1,793	0.0	3.5	22.6	52.4	43.9	38.0	01-16	명실상부 반도체 파운드리 최강자
폭스콘	2317 TT	250.50	165	1.0	17.6	26.2	64.2	21.0	41.0	03-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	143.50	22	1.1	-1.4	22.1	36.8	22.7	42.8	01-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,510.00	13	-4.4	4.2	-14.0	31.5	54.7	8.0	03-02	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스
에너지 인프라											
GE 베르노바	GEV US	575.40	229	4.6	-11.2	-9.2	32.9	73.1	75.2	01-22	AI 데이터센터용 가스터빈 인프라 솔루션
슈나이더 일렉트릭	SU FP	239.10	235	1.7	-3.2	9.7	8.7	0.9	1.0	02-19	AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션
비스트라	VST US	178.27	89	-3.4	-14.9	-13.1	15.3	26.1	29.8	02-27	미국 최대 독립 전력 생산 사업자
블룸 에너지	BE US	126.72	44	-7.4	15.3	187.5	561.4	831.8	470.6	02-27	AI 데이터센터용 연료전지 솔루션
넥스트에라 에너지	NEE US	85.89	262	4.7	1.9	19.7	21.0	19.3	22.7	01-23	AI 데이터센터용 엔드투엔드 전력 파트너

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

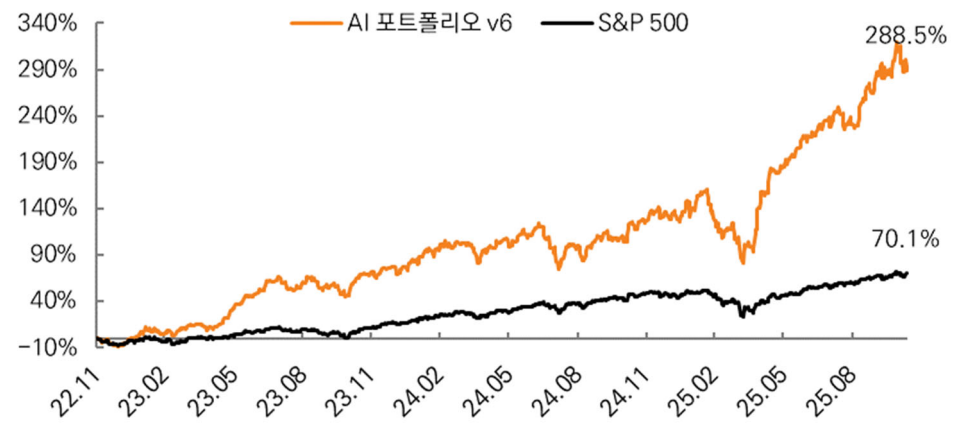
V. Appendix: Charts

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



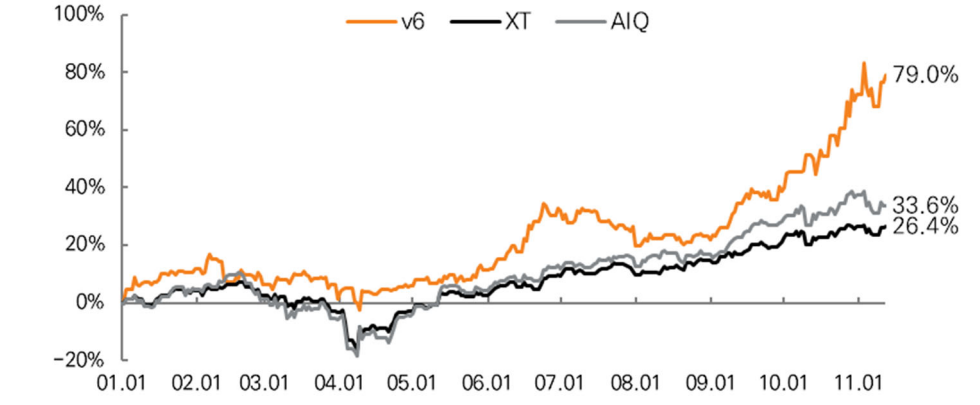
자료: Bloomberg, 미래에셋증권 리서치센터

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



자료: Bloomberg, 미래에셋증권 리서치센터

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



자료: Bloomberg, 미래에셋증권 리서치센터

표 13. SimpleBench 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	평균 점수(%)	기관명	라이선스	지식 컷오프
-	인간 평균	83.7	-	-	-
1(-)	Gemini 2.5 pro(06-05)	62.4	구글	독점 폐쇄소스	2025년 1월
2(-)	GPT-5 Pro	61.6	OpenAI	독점 폐쇄소스	2024년 9월
3(-)	Grok 4	60.5	xAI	독점 폐쇄소스	-
4(-)	Claude 4.1 Opus	60.0	Anthropic	독점 폐쇄소스	2025년 3월
5(-)	Claude 4 Opus	58.8	Anthropic	독점 폐쇄소스	2025년 3월
6(-)	GPT-5 high	56.7	OpenAI	독점 폐쇄소스	2024년 9월
7(-)	Claude 4.5 Sonnet	54.3	Anthropic	독점 폐쇄소스	2025년 3월
8(-)	o3 high	53.1	OpenAI	독점 폐쇄소스	2024년 6월
9(-)	Gemini 2.5 pro(03-25)	51.6	구글	독점 폐쇄소스	2025년 1월
10(-)	Claude 3.7 Sonnet thinking	46.4	Anthropic	독점 폐쇄소스	2024년 10월

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

표 14. LLM 벤치마크 챗봇아레나의 Hard Prompts 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(▲1)	claude-sonnet-4-5-20250929-thinking-32k	1482	Anthropic	독점 폐쇄소스	2025년 3월
2(▼1)	claude-opus-4-1-20250805-thinking-16k	1482	Anthropic	독점 폐쇄소스	2025년 3월
3(-)	claude-sonnet-4-5-20250929	1469	Anthropic	독점 폐쇄소스	2025년 7월
4(-)	claude-opus-4-1-20250805	1469	Anthropic	독점 폐쇄소스	2025년 3월
5(-)	Gemini-2.5-pro	1464	구글	독점 폐쇄소스	2025년 1월
6(▲2)	Claude-opus-4-20250514-thinking-16k	1451	Anthropic	독점 폐쇄소스	2025년 3월
7(-)	ChatGPT-4o-latest(2025-03-26)	1448	OpenAI	독점 폐쇄소스	2024년 6월
8(▼2)	gpt-5-high	1447	OpenAI	독점 폐쇄소스	2024년 9월
9(▼1)	gpt-5-chat	1444	OpenAI	독점 폐쇄소스	2024년 9월
10(-)	deepseek-v3.2-exp-thinking	1440	DeepSeek	상업적 사용허가	2025년 7월

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

표 15. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	GPT-5 high	79.33	OpenAI	독점 폐쇄소스	2024년 9월
2(-)	GPT-5 medium	78.85	OpenAI	독점 폐쇄소스	2024년 9월
3(-)	GPT-5 Pro	78.73	OpenAI	독점 폐쇄소스	2024년 9월
4(-)	Claude Sonnet 4.5 Thinking	78.26	Anthropic	독점 폐쇄소스	2025년 3월
5(-)	GPT-5 Codex	78.24	OpenAI	독점 폐쇄소스	2024년 9월
6(-)	GPT-5 mini high	75.31	OpenAI	독점 폐쇄소스	2024년 9월
7(-)	Claude 4.1 Opus Thinking	75.25	Anthropic	독점 폐쇄소스	2025년 3월
8(-)	GPT-5 low	74.65	OpenAI	독점 폐쇄소스	2024년 9월
9(-)	Claude 4 Sonnet Thinking	73.82	Anthropic	독점 폐쇄소스	2025년 3월
10(-)	Grok 4	72.84	xAI	독점 폐쇄소스	-

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 16. 코딩 관련 벤치마크 SWE-Bench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	점수(%)	기관명	등록일
1(-)	Claude 4.5 Sonnet (20250929)	70.6	Anthropic	2025-09-29
2(-)	Claude 4 Opus (20250514)	67.6	Anthropic	2025-08-02
3(-)	GPT-5 (2025-08-07) (medium reasoning)	65	OpenAI	2025-08-07
4(-)	Claude 4 Sonnet (20250514)	64.93	Anthropic	2025-05-21
5(-)	GPT-5 mini (2025-08-07) (medium reasoning)	59.8	OpenAI	2025-08-07
6(-)	o3 (2025-04-16)	58.4	OpenAI	2025-05-21
7(-)	Qwen3-Coder 480B/A35B Instruct	55.4	Alibaba	2025-08-02
8(-)	Gemini 2.5 Pro (2025-05-06)	53.6	구글	2025-05-21
9(-)	Claude 3.7 Sonnet (20250219)	52.8	Anthropic	2025-05-21
10(-)	o4-mini (2025-04-16)	45	OpenAI	2025-05-21

자료: SWE-Bench, 미래에셋증권 리서치센터

표 17. 컨텍스트 메모리 관련 벤치마크 LongBench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	기관명	모델 크기	컨텍스트 길이	출시일	종합 성능	긴 글 성능(%)
-	Human	-	-	-	-	53.7	53.7
1(-)	Gemini-2.5-Pro	구글	-	1M	2025-03-25	63.3	71
2(-)	Gemini-2.5-Flash	구글	-	1M	2025-04-17	62.1	55.7
3(-)	Qwen3-235B-A22B-Thinking-2507	알리바바	235B	256k	2025-07-25	60.6	58.1
4(-)	DeepSeek-R1	DeepSeek	671B	128k	2025-01-20	58.3	59.3
5(-)	Qwen3-235B-A22B-Instruct-2507	알리바바	235B	256k	2025-07-22	58.3	55.6
6(-)	o1-preview	OpenAI	-	128k	2024-09-12	57.7	58.1
7(-)	DeepSeek-R1-0528	DeepSeek	671B	128k	2025-05-28	56.7	51.4
8(-)	MiniMax-Text-01	MiniMax	456B	4M	2025-01-15	56.5	47.2
9(-)	Gemini-2.0-Flash-Thinking	구글	-	1M	2025-01-21	56.0	49.1

자료: LongBench, 미래에셋증권 리서치센터
주: LongBench는 멀티태스크·이중언어 롱컨텍스트의 클래식. 실제 과업 분포가 좋아 모델 비교에 널리 쓰임. 긴 글 성능은 128K 기준

표 18. AI 에이전트 벤치마크 TAU-bench의 성능 상위 모델 10개

순위 (전주 변동)	모델	성능(%)	비용($)	평가 기준일	실험 환경 구분
1(-)	o4-mini-high	56	11.36	2025년 4월	특정 도메인
2-)	Claude-3.7 Sonnet	56	42.11	2025년 2월	범용 에이전트
3(-)	o3 Medium	54	14.56	2025년 4월	특정 도메인
4(-)	Claude Opus 4.1	54	180.49	2025년 8월	범용 에이전트
5(-)	Claude-3.7 Sonnet High	52	31.94	2025년 2월	특정 도메인
6(-)	Claude Opus 4.1 high	52	149.98	2025년 8월	특정 도메인
7(-)	Claude Opus 4.1	50	69.78	2025년 8월	특정 도메인
8(-)	GPT-5 medium	48	23.83	2025년 8월	특정 도메인
9(-)	DeepSeek V3	44	5.43	2025년 3월	특정 도메인
10(-)	Claude-3.7 Sonnet	44	15.45	2025년 2월	특정 도메인

자료: HAL, 미래에셋증권 리서치센터
주: TAU-bench Airline은 AI 에이전트를 통한 개인 업무 자동화 능력을 평가하는 벤치마크

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 테슬라 20주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 한종목은(는) 자료작성일 현재 코어위브 60주 보유하고 있습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.