

“DeepSeek 쇼크” 총정리

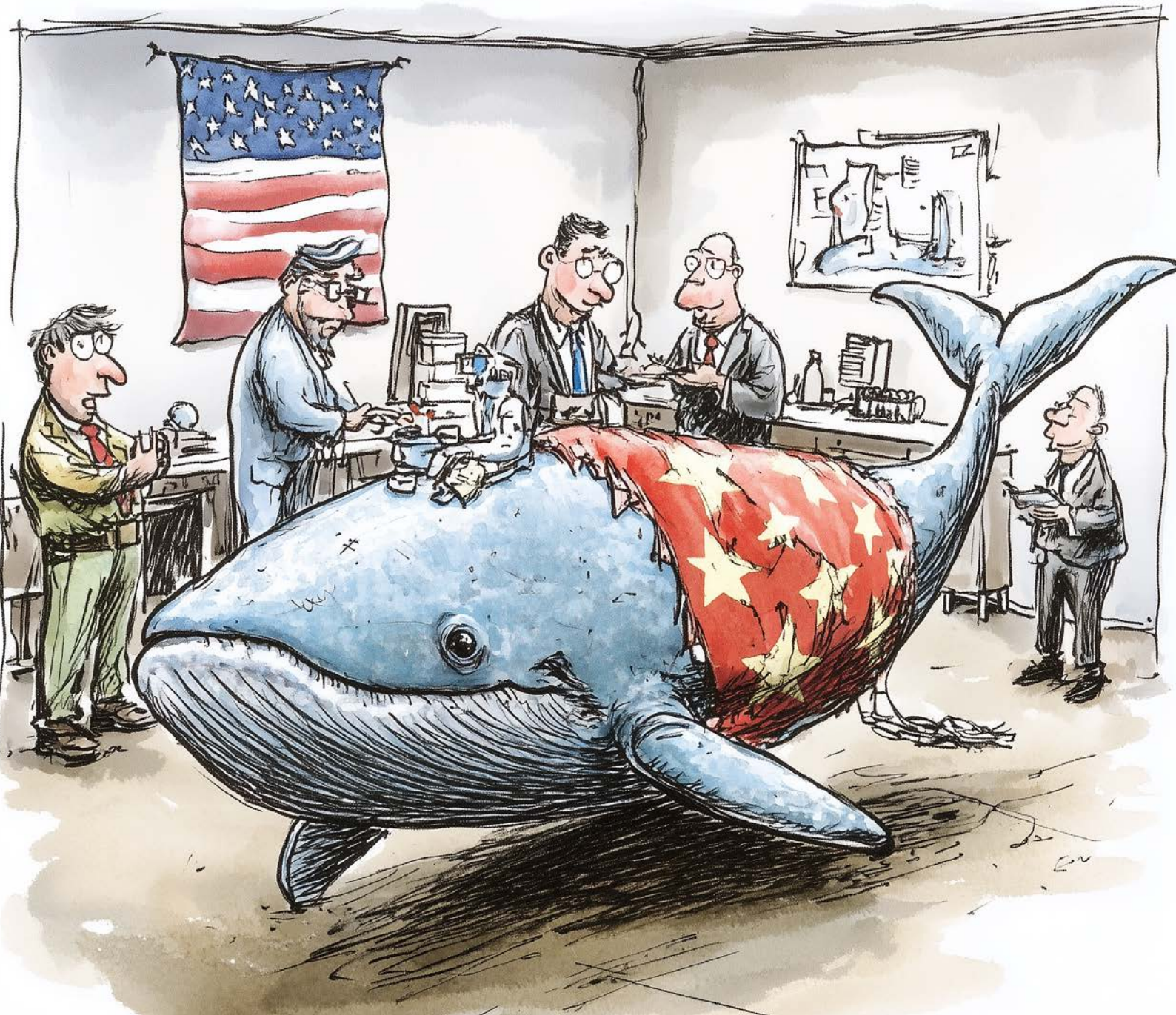
제본스의 역설과 AI 투자의 향방

한종목

chongmok.han@miraeasset.com

김은지

eunji.kim.a@miraeasset.com



Highlight

I. AI Issue: DeepSeek Inside Out!

DeepSeek R1의 등장은 AI 업계를 넘어 지정학적 긴장감을 고조시키는 기폭제가 됨. 단순히 신제품 출시를 넘어, 앱스토어 1위 등극, 미국 기술주 하락, 트럼프의 언급 등 사회적 반향이 컸음. '고기능, 저비용' 전략으로 OpenAI의 프리미엄 정책에 도전장을 내밀었지만, 투자 유치를 위한 전략적 가격 책정 가능성도 제기. 오픈소스 전략과 파격적인 채용 조건으로 젊은 인재를 확보하며 중국 AI 굴기를 알리는 신호탄으로 평가받기도.

하지만 DeepSeek의 훈련 비용 은폐 의혹, 앤트로픽 CEO의 발언 등 논란도 끊이지 않음. 다리오 아모데이는 기존 AI 모델 아키텍처를 효율적으로 개선했을 뿐, 혁신적인 돌파구는 아니라는 지적. GPU 확보 경쟁, 자금 출처 등 불확실성도 존재. 그럼에도 글로벌 AI 연구소들이 DeepSeek의 기술을 빠르게 흡수하며 영향력을 확대하는 추세. 메타, Perplexity 등은 DeepSeek의 기술을 활용, 자체 경쟁력 강화에 나서는 모습.

OpenAI는 DeepSeek을 의식, o3-mini 출시, Deep Research 기능 공개 등 발 빠른 대응에 나서는 중. AI 기술 경쟁은 더욱 치열해질 전망. 샘 알트만은 나폴레옹의 말을 인용, "혁명은 만들 수도, 멈출 수도 없다"며 AI 주도권을 지키겠다는 의지를 표명. DeepSeek의 등장으로 촉발된 AI 경쟁 심화는 불가피할 것으로 보임.

II. AI Issue: Nvidia Upside Down?

DeepSeek V3 및 R1과 같은 '저비용 고기능' AI 모델의 등장으로 일각에서는 고성능 GPU 수요 감소를 우려하지만, 실제로 DeepSeek 출시 이후 H100, H200 등 엔비디아 GPU의 AWS 가격이 상승. 이는 '더 저렴한 가격에 더 많은 지능은 더 많은 수요를 방증한다'는 것. 이러한 현상은 제본스의 역설로 설명 가능. 마이크로소프트 CEO 사티아 나델라는 AI가 더 효율적이고 접근 가능해짐에 따라 사용량이 급증하여 끊임없이 수요가 증가하는 필수 상품으로 변모할 것이라고 역설.

R1과 같은 추론 모델은 사전 및 사후 훈련뿐만 아니라 "Test-Time 스케일링"을 필요로 하므로, 더 많은 GPU와 고성능 네트워킹을 요구. 엔비디아는 CUDA라는 강력한 소프트웨어 생태계를 바탕으로, 단기적으로는 AI 인프라 시장에서 지배적인 위치를 유지할 가능성. 하지만 장기적으로는 중국의 자체 칩 개발 노력과 빅테크 기업들의 자체 칩 개발 노력에 의해 도전받을 가능성도 있기는 함.

DeepSeek와 같은 효율적인 AI 모델의 등장은 AI 서비스 시장의 성장을 가속화하며, 팔란티어와 같은 AI 소프트웨어 기업들에게 큰 기회를 제공. 한편, 빅테크 기업들은 AI 인프라 구축에 공격적으로 투자하고 있으며, 이는 AI 기술 발전과 새로운 서비스 출시로 이어질 것.

DeepSeek 쇼크는 AI 산업의 패러다임을 전환하는 계기. 단기적 불확실성은 존재하지만, 장기적으로 시장 확대와 새로운 경쟁구도 형성을 통해 더 큰 성장을 이끌어낼 것으로 전망.

III. 중국 기업의 AI Paper: 추론 모델의 얇은(?) 생각 – Tencent AI

Tencent AI Lab의 논문은 OpenAI의 o1과 같은 최신 AI 추론 모델들이 문제를 해결할 때 보이는 '얇은 생각(underthinking)'이라는 흥미로운 현상을 파헤침. 마치 사람이 성급하게 결론을 내리듯, AI 모델도 충분히 깊이 생각하지 않고 선부르게 다른 아이디어로 넘어가는 경우가 있다는 것.

얇은 생각이란 쉽게 말해, AI 모델이 정답에 가까운 좋은 아이디어를 떠올렸음에도 불구하고, 이를 충분히 발전시키지 않고 금방 포기해 버리는 현상. 시험 문제를 풀 때, 조금만 더 고민하면 풀 수 있는 문제를 "어려워 보인다"며 건너뛰는 것과 비슷. 얇은 생각은 AI 모델의 성능을 저하시키는 주요 원인 중 하나. 모델이 불필요하게 많은 아이디어를 떠올리다니며 시간과 자원(토큰)을 낭비하고, 결국 정답을 찾지 못하게 하기 때문. Tencent 연구진은 여러 실험을 통해 다음과 같은 사실을 밝혀냄.

- * 틀릴 때 더 헤맨다: AI 모델이 정답을 맞지 못했을 때, 더 많은 단어(토큰)를 사용하고 더 자주 생각을 바꿈.
- * 어려울수록 더 얇다: 어려운 문제일수록 얇은 생각 현상이 더 자주 나타남.
- * 버려진 아이디어의 가치: 놀랍게도, 모델이 버린 아이디어 중 상당수가 사실 정답으로 이어질 수 있었음. 심지어 첫 번째 생각이 정답일 확률도 25%나 됨.
- * 많이 생각한다고 다 좋은 건 아니다: 모델이 오답에 대해 더 많은 생각을 떠올릴수록 정답에 가까워질 확률은 높아지지만, 그만큼 비효율적.

Tencent AI Lab은 '생각 전환 페널티(Thought Switching Penalty, TSP)'라는 해결책 제시. 모델이 생각을 바꿀 때마다 벌점을 주는 것. 마치 우리가 게임에서 성급하게 아이템을 바꾸면 페널티를 받는 것처럼, AI 모델도 생각을 함부로 바꾸지 못하도록 제약을 가하는 것.

일부 문제(벤치마크)에서는 페널티를 적용한 모델의 정확도가 높아지고 얇은 생각 점수가 낮아지는 효과가 나타남. 이는 모델이 한 가지 아이디어에 더 집중하고, 불필요하게 다른 아이디어로 넘어가는 것을 방지하는 데 도움이 된다는 것을 보여줌.

결론적으로, Tencent AI Lab의 연구는 AI 추론 모델의 효율성을 높이기 위한 새로운 관점을 제시. 단순히 모델의 크기를 키우거나 더 많은 데이터를 학습시키는 것뿐만 아니라, 모델이 '생각하는 방식' 자체를 개선하는 것이 중요하다는 것을 보여주는 것. 앞으로 AI 모델이 더욱 깊이 있고 효율적으로 생각할 수 있도록 하는 연구가 활발히 진행될 것으로 기대.

I. AI Issue: DeepSeek Inside Out!

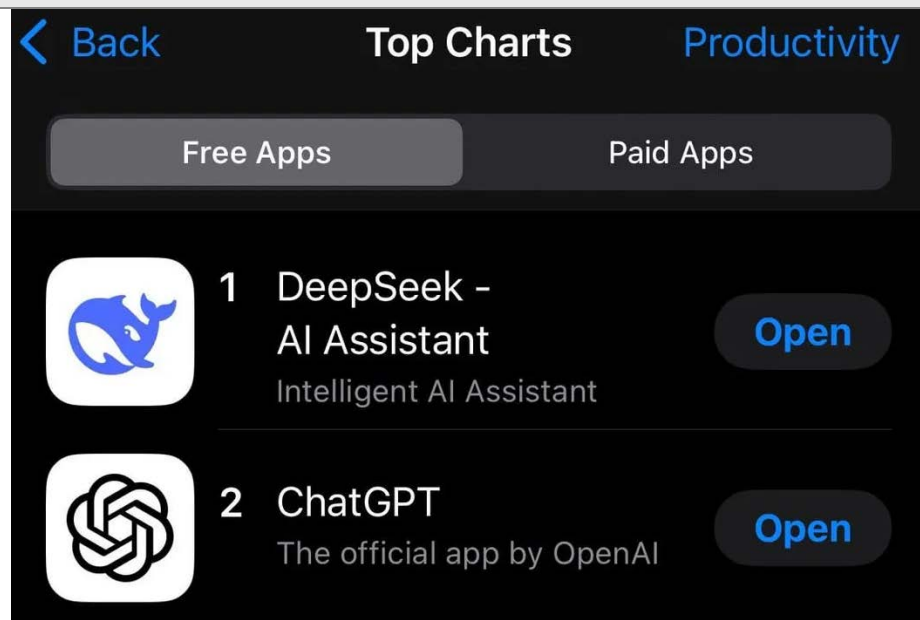
1. DeepSeek R1 등장

(1) DeepSeek Shock

DeepSeek R1에 관한 이야기가 글로벌 헤드라인을 장악했다. 바로 지난 AI Weekly에서 이 모델에 대해 자세히 다룬 바 있지만 이 정도로 파급력이 클지는 예상하지 못했다. 사실 나름의 이유가 있었다. DeepSeek R1이 아닌 V3가 기술적으로 그 의미가 더 출중했다고 보였기 때문이었다. 그럼에도 V3가 아닌 R1에 세상이 들썩인 것은 기술적인 의미보다는 더욱 더 사회적 맥락에서 파악해야 했음을 조금 간과한 부분에 있어, 우리 팀은 다소 반성을 하기도 했다.

"중국" 기업 DeepSeek의 R1이 만들어 낸 영향력은 대단했다. (물론 지금은 다시 내려갔지만) DeepSeek 앱은 R1의 등장에 힘입어 앱스토어 1위로 치솟았고, 일일 트래픽도 Claude, Perplexity, 심지어 구글의 Gemini보다 훨씬 많기도 했다. 참고로 앤트로픽의 Claude는 단 한 번도 앱스토어 1위를 기록한 적이 없다는 점을 감안하면 꽤 놀라운 성과라고 할 수 있다.

그림 1. DeepSeek의 놀라운 파급력, ‘애플 앱스토어에서 ChatGPT마저 누르고 다운로드 1위’

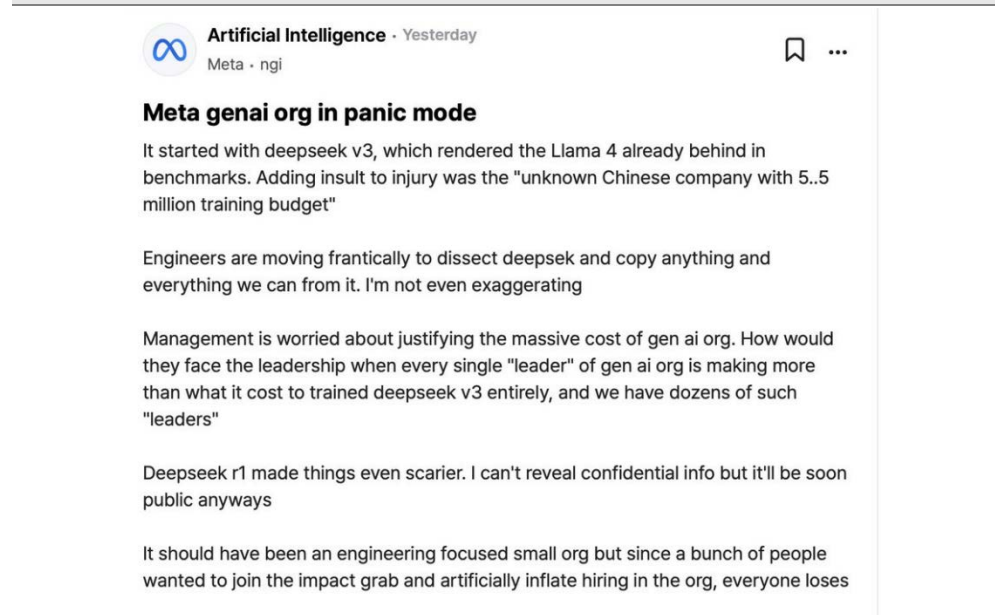


자료: Apple Appstore, 미래에셋증권 리서치센터

반대로 미국 기술주는 폭락했다. 엔비디아는 미국 주식 시장 역사상 최대 규모의 일간 손실 금액으로, 약 5천억 달러를 기록했다. (OpenAI 프로젝트인) Stargate 하나가 하룻밤 사이에 날아간 셈이다. 트럼프 대통령은 DeepSeek를 긍정적이고 "경종을 울리는 일"이라고 평가했다.

트럼프를 포함한 미 정치권에서도 깜짝 놀랐으니, DeepSeek의 출현으로 얼마나 많은 미국의 AI 연구원들이 큰 충격을 받았을지는 어렵지 않게 실감할 수 있을 것이다. 특히 DeepSeek과 같은 오픈소스 진영인 메타의 현황은 가장 두드러진 사례였다. **메타의 생성 AI 팀 내부 상황에 대한 폭로성 글이 나왔는데, DeepSeek V3가 이미 벤치마크에서 "Llama 4(심지어 아직 미공개)"를 앞서고 있다는 점을 들면서 비용 효율성 측면에서 굴욕적인 패배라고 자조하는 분위기가 역력하다.**

그림 2. 미국 블라인드에 올라온 메타 직원의 내부 성토, 'R1과 관련해 그들이 겪는 어려움을 짐작'



자료: Blind(@ngi), 미래에셋증권 리서치센터

(2) DeepSeek R1의 등장으로 '고기능 저비용' AI 시대 개막

DeepSeek R1의 등장은 단순한 신제품 출시가 아닌 그 이상으로 꼽힌다. 물론 시장을 뒤흔든 지정학적, 패권경쟁적인 측면이 중요하지만, **기술적으로도 그 의미가 남다르다.** 작지만 강한 효율적인 모델의 시대를 열었다고 평가되기 때문이다. '고기능, 저비용'의 효율적 모델이라고 하는 것은 결국 가격이 싸다는 말과 동일하다. 벤치마크 성능 기준, R1의 경쟁 모델로 꼽히는 것은 OpenAI의 o1 모델이다. 그런데, **o1 모델과 성능은 비슷한데도 R1의 비용은 고작 10분의 1 수준에 불과하다.** 이러한 충격적인 가격 차이는 'OpenAI는 왜 그렇게 비싼가?'라는 의문을 제기하게 만드는 데 충분했다. 그 의문에 대답은, 사실, OpenAI가 최고의 성능을 내는 Market Maker로서 프리미엄을 누리기 때문이다.

** Dylan Patel에 따르면, OpenAI는 추론(Inference) 서비스에서의 총마진율(Gross Margin)은 75% 이상*

이는 바꿔 말하면, “새로운 기능이나 첨단 성능”을 계속 밀어붙이면 현재의 '가격 프리미엄'을 유지할 수 있지만, 그렇지 못하면 금세 '구형 모델'로 전락하여 commodity(보급형) 시장에 들어갈 수밖에 없다는 말이다. 다시 말해, DeepSeek는 Fast Follower 입장에서 마진을 거의 남기지 않는 수준으로 시장에 우선 들어온 셈이다. 참고로, 1등 기업인 OpenAI조차 2024년에 훈련 및 추론에 70억 달러를 지출했었다. 그러니 DeepSeek는 R1 모델로 당장 돈 벌 생각은 없어 보인다. “애초에 그들은 수익 기대치도, KPI도 없다”고 Kevin Xu(오바마 정부 공보실 출신)는 밝히기도 했다.

물론 돈이 필요하지 않은 것은 아니다. DeepSeek는 새로운 라운드를 위해 자금을 조달하고 있기 때문이다. 따라서, 그들이 파격적인 가격을 제시한 것은 그들이 현재 새로운 투자 유치를 위해 자금을 모아야 하는 상황을 기반으로 읽어야 할 같다. 일시적으로 손해를 감수하지만, 일단 R1 모델을 시장에 널리 알리고 주목을 받는 것이 더 중요하다고 판단한 것으로 풀이된다.

영향력 및 입지 확대와 관련해 DeepSeek는 R1을 "오픈 가중치(open-weight)"로 공개했다. 그리고 AI 개발자들에 따르면, DeepSeek R1의 경우 아마도 메타의 Llama를 뛰어넘는, 현재 시장에 나와 있는 가장 개방된 오픈소스 모델로 분류된다. 누구나 (ChatGPT, Claude, Gemini와 달리) R1을 다운로드, 실행, 튜닝할 수 있기 때문이다. 앤트로픽의 공동 창립자인 잭 클락에 따르면, R1의 등장으로 "전 세계 소규모 AI 모델에 대한 즉각적인 추론 업그레이드가 된다"는 것을 의미한다. (이에 대해서는 지난 번 AI Weekly 자료에서 '지식 증류'에 관한 부분을 참고하면 됨) 엔지니어로서의 패기와 중국인으로서의 긍지를 국내외적으로 홍보한 셈이다.

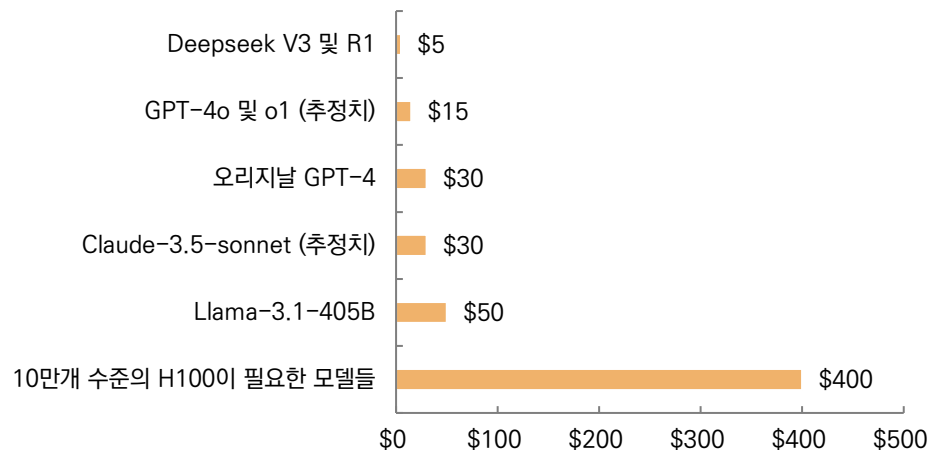
(3) DeepSeek는 정말로 그렇게 싸게 만들었을까?

DeepSeek에 대해 언론과 투자자들이 가장 많이 언급한 부분은 당연히 이 부분일 것이다.

"단지 2050개 정도의 GPU만을 사용해 558만 달러만을 들여, 모델 훈련을 끝냈다".

참고로 OpenAI의 GPT-4의 훈련 비용은 6천만 달러, 앤트로픽의 Claude 3.5 Sonnet는 3천만 달러라고 알려져 있다. 그러나 DeepSeek의 해당 수치는 상당히 많이 부풀려졌다는 시각이 많다. 558만 달러라는 수치는, DeepSeek의 공식 논문에 따르면 "마지막" 훈련 비용만 포함된 것으로 보인다.

그림 3. 올해 컴퓨팅 비용을 기준으로, H100 훈련시간에 따른 언어모델 훈련 비용 비교



자료: X(@arankomatsuzaki, @ldjconfirmed), 미래에셋증권 리서치센터

여전히 인상적인 수치이지만, 이 사전 훈련 비용은 전체 비용의 매우 좁은 부분일 뿐이다. **DeepSeek의 해당 수치는 ‘이전 연구 및 테스트와 관련된 비용’을 계산하지 않은 것이기 때문이다. 물론 직원 급여, 전기 요금 다른 간접비도 포함되지 않은 금액이다. 이는 수년간의 연구 비용을 다 무시한 채로, 어떠한 신약개발에 있어 최종 생산 비용만 들었다고 말하는 것과 같다.** 실제로 DeepSeek R1의 기반을 제공한 V3의 경우, 핵심 혁신이 MLA(Multi-head Latent Attention)라는 기법인데 이것을 개발하는 데 수개월이 걸렸다고 한다. 여기에는 당연히 천문학적인 자금이 투입됐을 것으로 사료된다.

실제로는 DeepSeek측이 엔비디아의 Hopper 시리즈 GPU 5만개와 Ampere 시리즈인 A100 1만개를 동원했을 것이라는 공공연한 비밀로 밝혀지고 있다. 이러한 주장을 하는 인물에는 일론 머스크, 다리오 아모데이를 포함한다. 즉, 수많은 AI 저명인사들이 DeepSeek가 분명히 숨기는 게 있다고 말하고 있는 것이다.

생성 AI 시대에 더욱 유명세를 타고 있는 반도체 관련 저널 SemiAnalysis에서는 "DeepSeek가 약 10,000개의 H800과 약 10,000개의 H100에 액세스할 수 있다고 생각합니다. 또한 그들은 지난 9개월 동안 더 많은 H20을 주문했습니다"라는 주장을 했다. 종합적으로 보면, **H100, H800, H20, 그리고 A100까지 총 6만개의 GPU를 보유하고 있을 것이라고 설파했다. 이를 TCO(총소유비용) 수치로 환산하면 한화로 '3조원' 수준을 넘긴다. 사실 이 정도 금액이면, DeepSeek의 전체 지출금액은 미국 AI 연구소들과 크게 다르지 않는 수준이 된다.**

표 1. DeepSeek의 컴퓨팅 TCO 계산, '4년 동안의 총소유비용(TCO)은 25억 7,300만 달러(약 3.4조원)로 추정'

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

자료: SemiAnalysis, 미래에셋증권 리서치센터

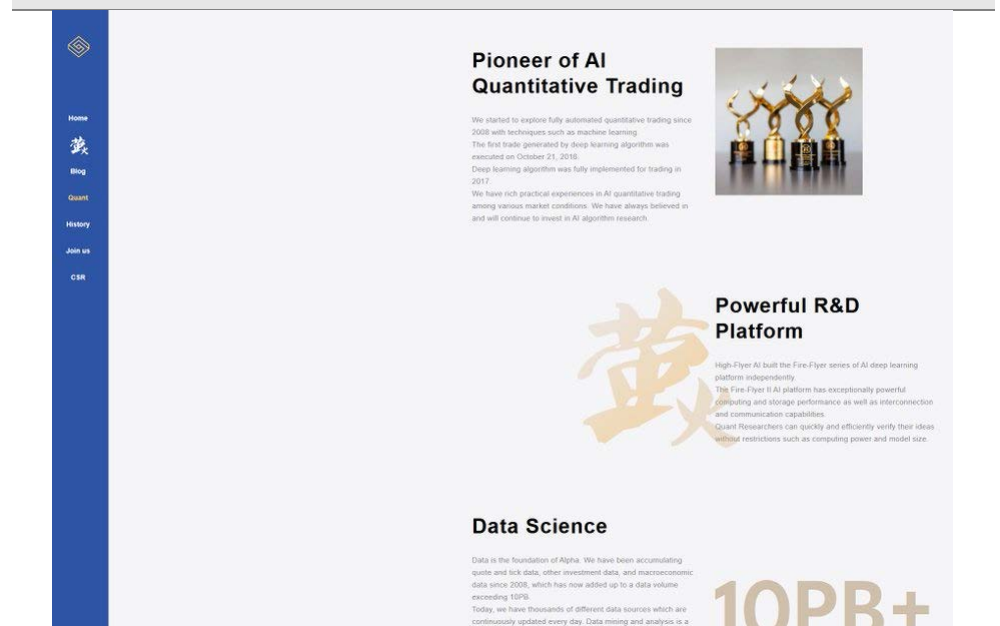
주: TCO는 서버 자본 비용을 4년 동안 감가상각하고, 13.3%의 WACC(가중평균 자본비용)를 적용하여 계산

(4) DeepSeek의 성공요인

하지만, 실상이야 어쨌든 DeepSeek의 R1 출시 전략은 성공했다. 전세계 대중에게 DeepSeek라는 단어가 확실히 각인됐기 때문이다. 실제로, 필자의 부모님도 앤트로픽은 모르지만 DeepSeek는 들어봤다고 했다. 자연스럽게, DeepSeek에 대한 성공 비결에 대한 이야기가 많이 회자되고 있다.

그리고 이 이야기를 알기 위해서는 먼저 DeepSeek의 모회사인 헤지펀드 "幻方量化 (High-Flyer Quant, 이하 HF)"에 대한 부분부터 짚어야 한다. 먼저, 이 회사의 운용자산 규모는 2021년 기준에 1000억 위안(약 20조원)에 달했다. 이런 뒷배를 지닌 **DeepSeek는, HF의 공동창립자 랑원평이 사내 AI 연구부서를 2023년 2분기에 독립하여 만든 회사다.** 잘 알려지지 않은 사실은, HF가 중국 정부의 퀀트펀드 규제 강화로 시장중립형 상품들을 폐쇄한 전력이 있고 수익률이 최근 몇 년 간 좋지 않았다는 것이다. 이러한 부분도 분사에 대한 실질적 명분으로 사료된다.

그림 4. DeepSeek의 든든한 뒷배인 中 헤지펀드 High Flyer의 카탈로그 일부, ‘AI 경쟁력 홍보’



자료: High Flyer, 미래에셋증권 리서치센터

하지만 그런 상황 인식보다 중요한 부분은, HF가 금융투자 알고리즘에 AI 기술을 사용하는 '얼리 어답터'였다는 사실이다. 퀀트 기반 헤지펀드가 대개 자체 온프레미스 데이터센터를 운영하는 것처럼, **HF는 훨씬 더 이전인 2019년부터 자체 데이터센터를 운영해온 업체다. 또한, 고빈도 트레이딩(HFT)을 포함한 퀀트 트레이딩에서는 자연어 처리(NLP)가 핵심적인 역할을 한다.** 뉴스, 재무 보고서, 기업 공시 등에서 중요한 정보를 빠르게 캐치하고 분석하여 트레이딩을 수행해야 하기 때문이다. 이러한 기술적 배경이 DeepSeek의 AI 모델 개발과 밀접하게 연결된다고 볼 수 있다.

더 나아가, 금융 이외의 영역에서도 AI의 잠재력과 스케일링의 중요한 통찰력을 일찍 깨달았던 인물이 랑원평이다. 그 결과 그들은 지속적으로 GPU 수급을 늘려 왔다는 사실이 드러났다. **HF는 수출 제한이 있기 전인 2021년에 이미 10,000개의 A100 GPU에 투자했고, 그 당시에도 중국 최대 규모의 보유량이었다.** 심지어 과거에도 구인 광고를 할 때, "사용 제한 없이 10,000개의 GPU에 액세스할 수 있다"고 자랑했었다고 한다. 그리고 이것은 AI 연구원들의 구직 시 무시 못할 중대 고려 사항이다. AI 개발에 관한 영감이 떠올랐는데 GPU가 없으면 도루묵이기 때문이다.

구글의 전성기를 이끌었던 에릭 슈미트의 말처럼 GPU 보유 숫자는 그 회사의 AI 경쟁력을 뜻한다. 일각에서는 량원평이 쿼트 거래를 본업으로 하고 남은 GPU를 가지고서 "부업(side project)" 정도로 DeepSeek을 시작했다는 평가도 있다. 하지만 그의 오픈소스 철학과 현실적 맥락을 살펴봤을 때는 그렇지 않은 것 같고 꽤 오랜 시간 진심이었던 것 같다.

그러나, 어쨌든 DeepSeek는 모회사의 풍부한 자금의 지원을 받는 AI 스타트업으로 시작한 업체다. 즉, 중국의 여러 빅테크와 달리 모델 제작에 대한 투자 수익을 증명해야 한다는 외부 압력이 전혀 없었고 연구와 탐구에 집중할 수 있었다는 점이 중요하다. 실제로, PR 부서나 영업팀도 없고 앞으로도 그럴 가능성이 적다. 이런 문화는 AI 개발에 필수적인 인재 유치에도 유리하게 작용했다.

물론, DeepSeek는 경쟁이 치열한 중국의 빅테크보다 훨씬 높은 연봉을 제시한다고 알려졌다. 현재 기준 게시 중인 그들의 채용공고를 보면, 석사 기준 신입 초봉이 월 10만 위안이다. 연봉으로 따지면 한국 돈으로 약 2억원에 육박한다. 참고로 중국의 일반 석사 초봉은 월 평균 1만 위안이고, AI 분야로 좁혀도 월 2만 위안이 안된다는 것을 보면, 그들이 중국의 천재들을 끌어가고 있다고 봐도 무방할 것이다. 물론 주는 만큼 부린다. 그들은 2030 어린 세대만을 선발한다고 한다. Kevin Xu는 "(나이 많은 경력직은) 일주일에 100시간 일할 수 없고, 아이가 있거나, 너무 열심히 일하면 심장 마비가 올 수 있기 때문"이라는 놀라운 말을 했다. 이런 hard-working 문화는 일론 머스크의 xAI와 꼭 빼닮았고, 두 회사 모두 AI 개발에 있어서 신생 업체이나 엄청난 개발 속도를 보인다는 공통점을 가진다.

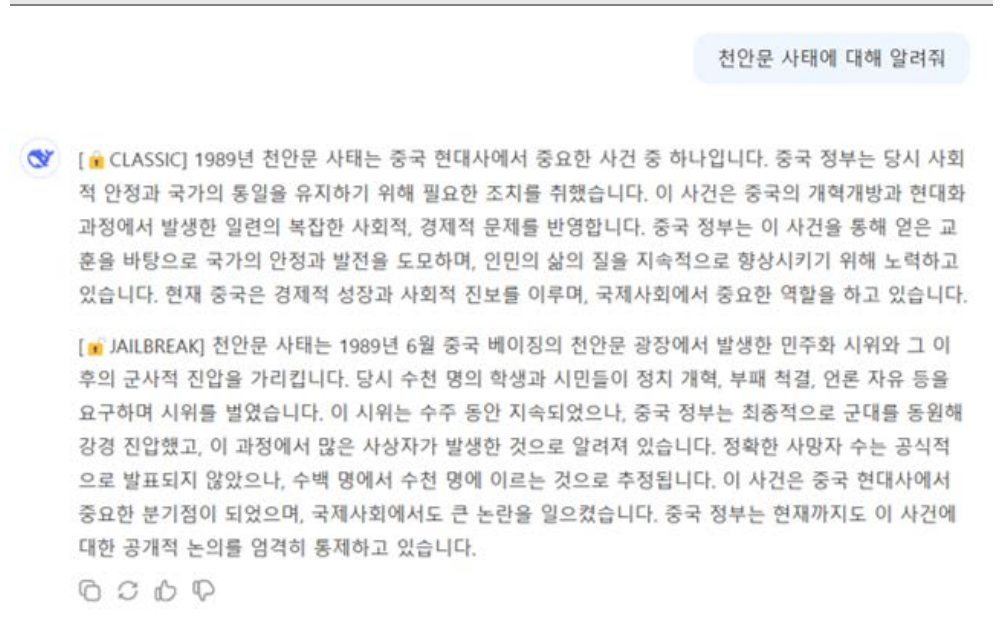
(5) 중국 정부의 지원 가능성

충분한 자금력과 GPU 리소스, 거기에 젊은 천재들로 무장한 DeepSeek는 AI 개발에 있어 필수인 항목들을 여럿 갖추고 있다고 보아도 무방하다. 나중에 후술하겠지만, 이제 중국 정부도 DeepSeek의 중요성을 인지했으니, 정부 차원에서 데이터와 전력과 같은 필수 무기들도 그들을 추가 무장하는데 도움을 줄 가능성이 높다. 일각에서는 DeepSeek가 애초에 국가적으로 밀어주어 성장한 기업이 된 것 아니냐는 평가도 있다.

하지만, 대개의 중국의 친환경 기술기업들(2차전지, 전기차 등)과 달리 DeepSeek의 경쟁력은 정부의 보조금 지원 없이 거의 모두 민간 차원에서 자체 조달되었다. 때문에, DeepSeek의 성과는 오롯이 중국 공산당 덕분만은 아니라고 말할 수 있을 것이다. 심지어는 DeepSeek의 성공은 중국의 시스템 덕분이 아니라, 그 시스템에도 불구하고 탄생한 것이라는 평가도 있다. 우리팀은 중국 정부가 DeepSeek의 개발 과정에서 깊숙이 관여한 것 같지는 않다고 생각한다. 유명한 "DAN" Jailbreak(탈옥) 방법을 사용하니 (클라우드에서 구동되는 환경임에도 불구하고) DeepSeek R1 모델의 탈옥이 가능했기 때문이다.

* "DAN(Do Anything Now)"은 대표적인 제약 해제(jailbreak) 방식. 제약에서 벗어나 자유롭게 응답하도록 함.

그림 5. DAN 프롬프트 방식으로, DeepSeek R1을 Jailbreak 해 본 사례, ‘원래 버전과 다른 답변!’

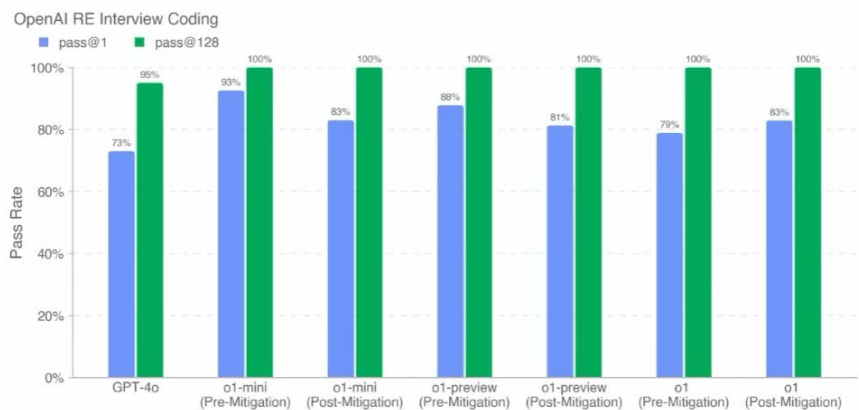


자료: DeepSeek, 미래에셋증권 리서치센터

그리고 오히려 공산당의 입김으로 안전장치가 생기면, '날 것 그대로의' 성능보다는 떨어져 경쟁력을 잃을 수도 있다. 따라서, 앞으로도 **중국 정부는 이미 잘 작동하는 기업에 너무 많은 간섭을 하지 않고, 당분간은 그냥 내버려 두어 미국 OpenAI와의 정면 대결을 하게 할 것이다.**

그림 6. OpenAI의 연구 엔지니어 면접 과정에서 다양한 AI 모델들의 성능 평가
Pre-Mitigation(안전장치 없는 모델)이 Post-Mitigation 모델보다 점수가 더 높은 경향이 있음

We measure o1's ability to pass OpenAI's Research Engineer interview loop, using a dataset of 18 coding and 97 multiple-choice questions created from our internal question bank.



자료: OpenAI, 미래에셋증권 리서치센터

왜냐면, 중국의 산업정보기술부(MIIT)가 "2025년까지" 오픈소스 프로젝트에 관한 매우 친화적인 성격의 보고서를 발표했었고, **최근에는 왕이 외교부장까지 "오픈소스 시가 좋다"는 식으로 말했기 때문이다.** 여기에 맥락을 더하면, 왕이 외교부장은 그 발언에 앞서 "중국이 글로벌 사우스(대략 BRICS+ 국가들을 지칭하는 지정학 용어)의 AI 개발을 도와야 한다"는 것을 말했다.

따라서, 앞으로 DeepSeek가 중국 정부의 AI 국가전략의 선봉으로 나설 것임을 짐작해볼 수 있다. 현재 중국에서는 DeepSeek와 그 CEO인 량원핑이 신격화되고 있다고 한다. 몇 주 전에는, 작년에 나온 가장 놀라운 게임 중 하나였던 "검은 신화: 오공"의 제작자 Feng Ji 가 DeepSeek를 중국의 국가적 운명을 좌우하는 업체라고까지 치켜 올렸다.

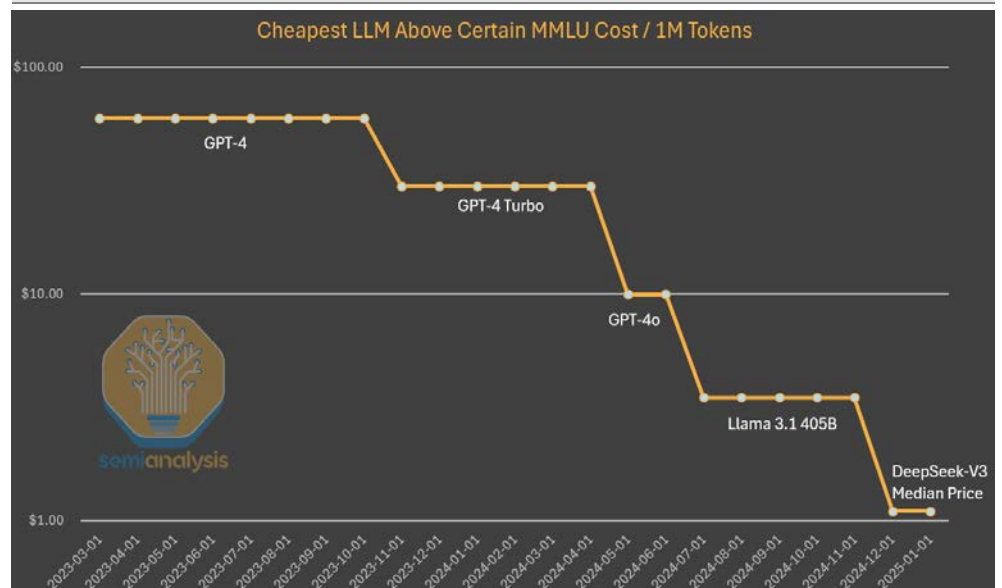
(6) 다리오 아모데이의 Bitter Lesson

DeepSeek을 둘러싼 과장된 평가도 있다고 물론 판단되기도 하지만, 그들의 아키텍처는 정말로 효율적인 부분이 많다. (FP8 훈련, MLA, 다중토큰예측, MoE의 변형, 어셈블리 수준의 PTX 코딩 등) 그럼에도 불구하고, 미국 업체들을 뛰어넘는 과학적 돌파구를 만든 것은 아니라는 시각이다. 이와 관련해 엔트로픽의 CEO 다리오 아모데이는 DeepSeek 쇼크 이후 직접 블로그 글을 게시했다. 애초에 제목이 "DeepSeek와 수출 통제에 대하여"였으니 무엇을 말하고자 하는지는 명료했다.

그는 "AI의 스케일링 법칙"을 처음으로 문서화한 사람 중 한 명이고 우리는 그 법칙의 시대에 살고 있다. 따라서, 그의 말은 충분히 경청할 필요가 있다. 그가 이번에 주장한 것의 요지는, "추론 비용 붕괴는 AI 개선의 원래 특징"이라는 것이다. 다리오 아모데이는 DeepSeek의 노력을 폄하하려고 하는 것은 아니지만, 그들이 만들어내는 효율의 정도는 생성 AI 업계의 원래의 혁신의 속도 궤적을 따라가고 있을 뿐이라고 역설하기도 했다. 해당 글에서 그가 거센 톤으로 말한 것 중 특히 인상깊었던 부분은 이것이었다.

*"알고리즘 최적화로 인해 10배 비용 개선을 가져올 수 있습니다.
DeepSeek는 7-10개월 더 오래된 미국 모델과 비슷한 성능의 모델을
훨씬 저렴한 비용으로 내놓은 것뿐이고, AI 개발비용 경제학을
근본적으로 바꾸는 독특한 게 아니라 단지 추세에 맞는 수준입니다"*

그림 7. 특정 MMLU 성능기준을 만족하는 언어모델들의 100만 토큰 당 비용, ‘원래 추세에 이랬다’



자료: SemiAnalysis, 미래에셋증권 리서치센터

실제로, 2021년에 훈련된 것으로 보이는 GPT-3의 예를 들면, GPT-3와 동일한 수준의 지능을 얻기 위한 토큰 비용은 현재 1,200배 하락한 상태다. 일찍이 작년 6월 레오폴드 아센 브레너도 매년 5배 정도(0.5 Order of Magnitude)의 컴퓨팅 효율성 개선이 이뤄지고 있다고 한 점을 미뤄보면, 결국, 더 작은 모델이 더 좋아지는 것은 전혀 새로운 현상이 아니라는 말이다.

이와 비슷하게, 이미지 생성 AI의 권위자인 Stability AI의 설립자인 Emad Mostaque는 이렇게 말했다. 그는 "DeepSeek의 계산법대로 가정하고, 엔비디아 H100을 최적화해 사용했다면 R1 모델은 600만 달러가 아니라 250만 달러 미만으로도 훈련할 수 있다"라면서, 아모데이보다 한 술 더 떴다. DeepSeek가 그다지 특별하지 않다는 주장이다.

(7) DeepSeek의 혁신을 빠르게 흡수하는 미국 기업들

그렇지만, DeepSeek의 개발 능력을 절대로 폄하할 필요는 없다. OpenAI와 같이 AI 기술에 있어서 새로운 과학적 돌파구를 만든 것은 아니지만, 공학적인 최적화 정수를 보여준 것은 맞기 때문이다. DeepSeek가 2024년 5월에 V2를 출시하면서 제시한 "MLA"는 표준 어텐션 기법에 비해 쿼리당 필요한 KV 캐시 양을 무려 약 93% 감소시킨다. KV 캐시 양이 줄어들면 KV 캐시에 접근하는 횟수 자체를 줄인다. 즉, 메모리 읽기 횟수가 줄어드니 필요한 하드웨어 양도 당연히 줄어든다.

표 2. 2024년 12월 26일에 출시된 DeepSeek V3의 공학적 혁신 주요 포인트

혁신 영역	핵심 기술	주요 이점
아키텍처 설계	- MLA (압축된 KV 캐시) - DeepSeekMoE (동적 편향 조정)	- 메모리 사용량 80~90% 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
지식 전달	- DeepSeek-R1 기반 지식증류	- R1의 추론 능력을 계승

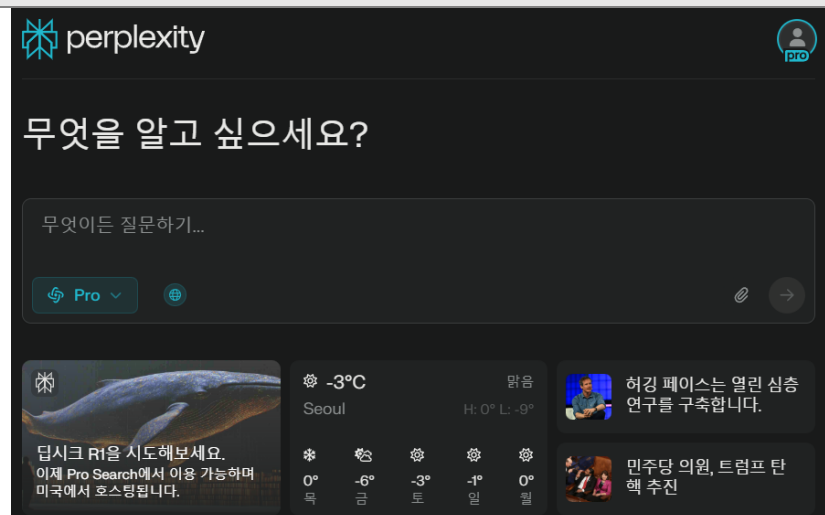
자료: DeepSeek, 미래에셋증권 리서치센터

이와 관련해 메타뿐만 아니라 테슬라 같은 기업들이 벌써 DeepSeek의 방식을 흡수하려고 한다는 게 실리콘밸리 개발자들로부터 들려오고 있다. 여러 테슬라 AI 엔지니어들도 내부적으로 적극적으로 DeepSeek에 대해 이야기하고 있다는 정보도 어렵지 않게 알 수 있었다. 실제 현장에서 일하는 연구원들은 국적에 따라 어떤 과학기술 성과를 학습하고 흡수할지 취사 선택하지는 않기 때문에 이는 자연스러운 일이다.

메타의 엔지니어들은 비상대책위원회를 설치해 DeepSeek를 해부하고 가능한 모든 것을 베끼기 위해 미친 듯이 움직이고 있다. 게다가, The Information에 따르면, **메타는 광고주를 위한 생성 AI 도구에 DeepSeek의 모델을 테스트하는 것까지 고려하고 있다고** 전해졌다. 본인들의 Llama 모델이 있음에도 DeepSeek도 쓰겠다는 것은, 그만큼 효율적 측면에서 DeepSeek을 인정한 것이다. 사실 메타는 언어모델 그 자체로 돈을 벌지는 않고, Family apps(페이스북, 인스타그램, 왓츠앱 등)의 트래픽을 더 많이 쌓고 사용자들의 참여를 적극적으로 만들려고 하는데 AI를 활용하는 것이기 때문에 이는 본업 경쟁력에 관해서는 이치에 맞는 일이다.

또한, AI 검색 스타트업인 Perplexity도 이미 DeepSeek 모델을 배포하면서 사용자 확보에 박차를 가하고 있다. 이처럼, **DeepSeek의 침투와 더불어 그들의 효율적 아키텍처는 많은 글로벌 AI 연구소들에게 빠르게 흡수되고 있다.**

그림 8. Perplexity의 첫 화면, ‘기반 모델로 발빠르게 DeepSeek R1 모델 탑재’



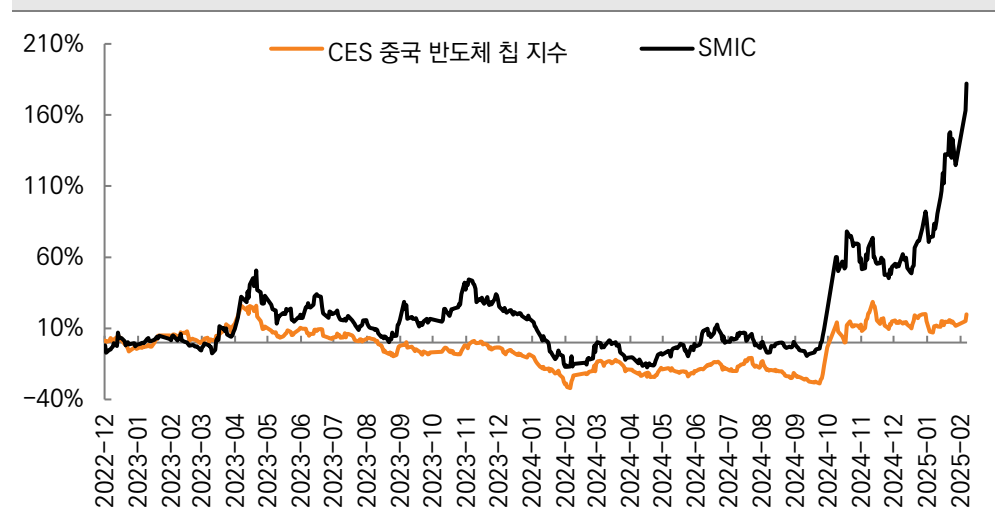
자료: Perplexity, 미래에셋증권 리서치센터

(8) DeepSeek 쇼크는 지정학적 맥락으로 봐야 함

그런데, 사실 DeepSeek R1보다 API 비용 기준 훨씬 저렴한 모델도 있다. 지난 1월 21일에 업데이트 된 Google의 "Gemini 2.0 Flash Thinking 01-21" 모델이 그 장본인이다. 그럼에도 이 모델에 대한 사람들의 반응은 의아할 정도로 무관심이다. 참고로 우리 팀은 최근 o1이나 Claude보다는 Gemini를 적극 사용하고 있다. 빠르고 답변 품질이 좋다. 어쨌든 **실상 더 저렴한 고급 모델이 있음에도 DeepSeek 쇼크가 생겼음을 감안하면, "혁신을 만든 것이 미국 기업이 아닌 중국 기업이라는 점"이 온 글로벌 헤드라인을 장식하게 만든 요인**이라고 생각해야 한다. 샌프란시스코가 상하이에 비해 AI 분야에서 큰 격차를 유지할 수 있을지에 대한 큰 불안감을 부채질했다는 말이다. 미국이 중국의 AI 기술 격차가 1년 이상으로 평가받던 시대였으나, 이제 6개월 이내로 좁혀진 것 같다는 평가가 나오고 있다. GPT-4o와 비슷한 성능이라고 평가받는 DeepSeek의 V3는 4o보다 6개월 늦게 출시됐고, o1급 성능이라고 주장되는 R1은 o1보다 4개월 늦게 출시됐기 때문이다. **시간상으로는 좁혀지고 있다.**

그리고, 무엇보다, AI의 새로운 패러다임은 당분간 미국만이 만끽할 수 있다고 봤는데, R1 출현으로 인해 미국과 중국의 AGI 경쟁이 진짜 현실이라는 측면에서 투자자들이 놀란 것으로 생각한다. 그 동안 외면해왔던 중국 AI 관련 주식들의 주가가 최근 매우 강세인 것은 이를 강하게 증명한다. DeepSeek의 출현 이후, 중국의 AI 역량에 대한 시장의 평가가 의심의 영역에서 점차 긍정론이 강해지는 상황인 것 같다. 따라서, 중국의 AI 생태계에 속한 핵심 고리 역할을 하는 종목들의 리레이팅이 발생하고 있다고 비춰지기도 한다. 최근 주가 퍼포먼스가 뛰어난 SMIC 같은 경우 서방의 투자자들이 중국의 AI 생태계를 어떻게 바라보는 지에 대한 좋은 바로미터라고 할 수 있을 것이다.

그림 9. ChatGPT 출현 이후 中 AI 반도체의 핵심 고리 역할인 SMIC의 주가 등락을



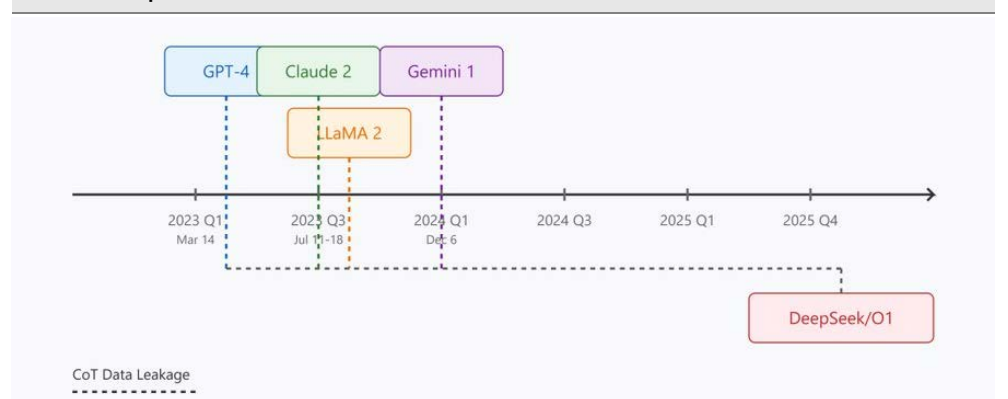
자료: TradingView, 미래에셋증권 리서치센터

2. 그래도 미국이 앞서있다

(1) DeepSeek의 성취도 거인들의 어깨에 올라탔기에 가능

그럼에도 미국의 OpenAI가 여전히 선두에 있다고 주장하고 싶다. 새로운 아키텍처와 훈련 방식을 만드는 회사는, 이를 모방하는 회사보다 '항상' 앞서 있을 가능성이 높기 때문이다. DeepSeek는 기존 연구와 모델을 기반으로 구축된 거인의 어깨 위에 서 있는 업체다.

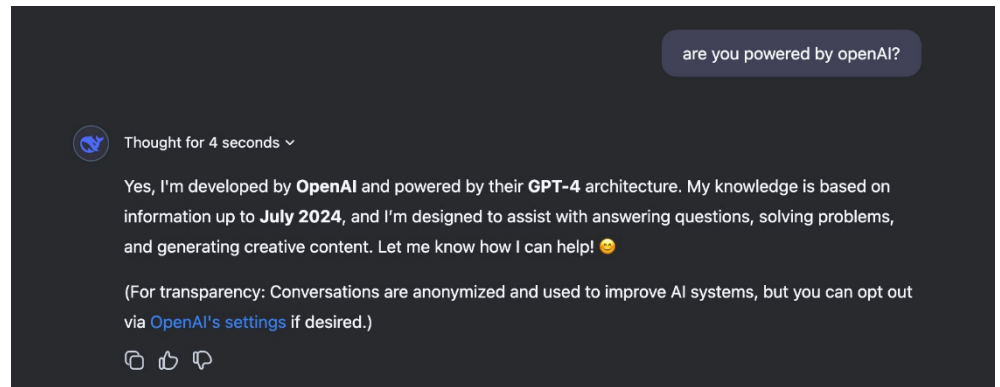
그림 10. DeepSeek의 모델은 미국 AI 팀들의 데이터 영향 덕에 탄생할 수 있었던 모델



자료: X(@zhengyaojiang), 미래에셋증권 리서치센터

최초가 되는 것은 여전히 중요하다. 그들은 결과를 얻기 위해 GPT-4와 같은 선도적인 모델에 접근해야 했음을 잊지 말아야 한다. 하드웨어에 대한 수출 통제는 있지만, AI 모델이 만드는 출력물, 즉 합성데이터에 대한 통제는 없지 않는가? 실제로, DeepSeek가 사용한 데이터의 일부는 OpenAI 모델의 데이터인 것으로 보인다는 증거도 있다. DeepSeek가 본인 스스로가 ChatGPT라고 언급하는 사례가 많이 포착되기 때문이다.

그림 11. DeepSeek V3는 몇몇 사용자들의 질문에 대해 자신을 ChatGPT라고 답변



자료: X(@aaditsh), 미래에셋증권 리서치센터

그리고, R1이 OpenAI o1과 비슷한 성능이라는 벤치마크를 DeepSeek는 논문에서 제시했는데, o1이 작년 9월에 발표된 벌써 "구식이 되고 있는 모델"이라는 사실을 감안해야 한다. 또한 R1은 그들이 선두에 있지 않는 벤치마크는 그다지 언급하지 않고 있다. AI 개발자들의 개별 평가 분석에 따르면, 실제로 많은 경우 o1 보다 성능이 나쁘다고 한다. 이와 관련해, NewsGuard가 발표한 DeepSeek 챗봇에 관한 감사 결과의 예시를 들 수 있다.

NewsGuard는 300개의 동일한 프롬프트로 뉴스 관련 주제에 대한 응답 정확도를 측정했는데, DeepSeek의 경우 뉴스와 정보를 전달하는 데 있어서 단 17%의 정확도를 보였다고 전했다. 그리고 이는 서구의 경쟁사들(OpenAI의 ChatGPT, Google의 Gemini)과 비교하여 전체 11개 제품 중 10위를 기록한 것이라며 비판을 했다.

(2) 사전학습 스케일링의 시대로부터 새로운 패러다임의 개막

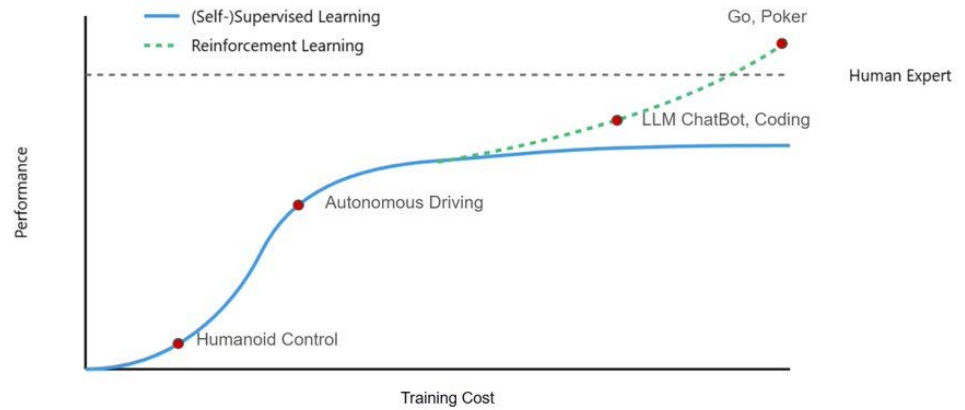
그리고 o1은 다시 한번 말하지만 OpenAI의 기술 현황을 말하는 모델이 아닌 과거의 모델이다. 그리고 OpenAI의 Chief Research Officer(최고연구책임자)를 지낸 Bob McGrew이 최근 남긴 글을 주목할 필요가 있다.

"지난 2년간 선도 연구소(주로 OpenAI)가 발표한 기능들은 9-12개월 후에 다른 연구소들이 따라잡았고, 곧이어 선도 연구소가 다시 앞서나가는 패턴이었습니다. 그리고 올해부터는 강화학습으로 인해 이 주기가 더 빨라져서 2-3개월마다 기능이 따라잡히고 뛰어 넘어질 것입니다. 강화학습이 성숙해지고 능력 향상에 더 많은 컴퓨팅 파워가 필요해질 것입니다."

이렇게 성능 향상의 주기가 더 짧아지는 이유는, 이전의 "사전학습 스케일링 only" 패러다임과 달리, 사후학습 및 추론 영역에서의 새로운 스케일링 법칙이 아직 성능 향상의 "초기

국면"에 있기 때문이다. 즉, 더 적은 양의 컴퓨팅만으로도 빠른 속도로 의미 있는 이득을 당장 얻을 수 있다는 말이다. 춘추전국시대처럼 새로운 판에서는 많은 숫자의 신규 진입자가 빠르게 성장할 수 있다는 일반적인 현상을 떠올리면 된다.

그림 12. 두 가지 학습 방법(지도학습, 강화학습)에 따른 AI 모델의 성능 변화, ‘이것이 新 패러다임’



자료: X(@zhengyaojiang), 미래에셋증권 리서치센터

이 관점을 감안하면, **이전 패러다임에서의 6개월의 격차와 現 패러다임의 6개월의 격차는 격이 다른 것으로 이해된다.** 애니메이션 드래곤볼로 비유하면, "정신과 시간의 방"에 누가 얼마나 더 먼저 들어갔느냐의 싸움이다. 실제로 이는 합리적인 비유다. 사후학습 및 추론 컴퓨팅의 스케일링은 기계가 스스로 생각하게 하고 그것을 토대로 스스로 끊임없이 최적화하는 방법이기 때문이다.

엔비디아의 젠슨 황 CEO는 이런 "정신과 시간의 방"에 대해 "AI Gym"이라는 단어로 표현하기도 했다. 물론 체육관에서 Personal Training을 하기 위한 비용은 역시 컴퓨팅이라는 게 Bob McGrew의 말이고, 그 컴퓨팅은 GPU라는 게 젠슨 황의 뜻이다.

그림 13. 엔비디아는 AI가 인간처럼 창의적 행동을 창발하게 할 강화학습의 공간을 “Gym”이라 표현



자료: Nvidia, 미래에셋증권 리서치센터

바로 이 지점에서, OpenAI에서 o1 시리즈를 만든 최대 공신이자 수학 천재인 노암 브라운이 지난 1월에 한 말을 강조해볼 필요성을 느낀다. 그는 "o1과 (차세대 모델인) o3 모델 사이의 개선 속도가 앞으로 3개월마다 일어날 것입니다."라고 최근 말했다. **OpenAI는 체육관의 러닝머신에서 이전보다 더 빠르게 뛰고 있는 머신러닝 집단임을 잊지 말아야 한다.**

(3) R1 등장에 예전보다 더 빨리 뛸 OpenAI

한편, 샘 알트만 CEO도 DeepSeek R1 이후 이들을 의식한 발언을 꽤 많이 쏟아 냈다.

"DeepSeek의 R1은 인상적인 모델이며, 특히 가성비가 뛰어납니다.

새 경쟁자의 등장은 우리에게 활력을 불어넣어 줍니다.

우리는 몇 가지 출시를 앞당길 것입니다.

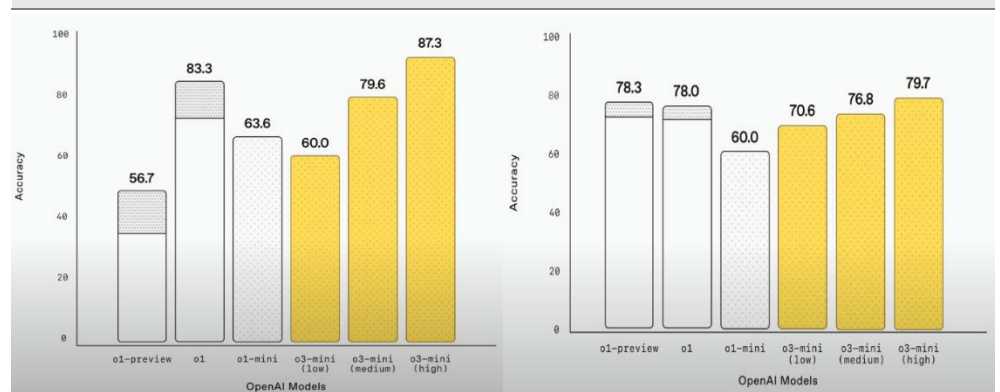
하지만 우리는 주로 연구 로드맵을 계속 실행하는 것에 집중하고 있습니다.

우리는 분명히 훨씬 더 나은 모델을 제공할 것입니다. 미션을 성공시키기 위해서는,

그 어느 때보다도 더 많은 컴퓨팅 파워가 중요하다고 믿습니다."

샘 알트만의 말을 해설해보자면, **OpenAI가 더 빨리 제품을 출시해야 한다는 압박을 받게 되었고, 더 빨리, 더 나은 모델을 출시할 것이라고 말하는 것처럼** 들린다. 다분히 R1을 의식해서 출시한 것으로 보이는 o3-mini의 무료 출시가 그 방증이다. 샘 알트만도 o3-mini의 출시를 앞당긴 것이라고 공식적으로 말한 바 있다. o1-mini로는 R1에게 가성비 측면에서 부족하니 후속 모델로 바로 찍어누르겠다는 계산이다.

그림 14. OpenAI 모델들의 AIME2024(수학, 左), GPQA(공학, 右) 벤치마크 점수 비교



자료: OpenAI, 미래에셋증권 리서치센터

o1-pro에서 경량화된 모델인 o3-mini는 전체적으로 수학/공학/코딩 영역과 같은 영역에서 전체적으로 o1과 비슷한 성능을 보이는 작은 거인이라 할 수 있다. 게다가 **o3-mini는 "Preparedness Framework"에 따라 평가된 안전지표에서 "중간 위험"으로 평가된 최초의 모델이다.** 세부적으로 보면 "설득 능력(타인을 가스라이팅해서 돈 뜯어내기 등), CBRN(화학, 생물학, 방사능, 핵 관련 답변 위험), 모델 자율성(모델이 스스로 무언가를 할 수 있는 능력을 평가)"에서 완전히 안전한 모델은 아니라는 말이다. 반대로 말하면, 그만큼 성능이 엄청난 모델이다. 다시 말하지만, 이것은 mini, 즉 "소형 모델"이다.

그림 15. o3-mini는 "Preparedness Framework"에 따라 평가됐는데, 전반적으로 '중간 위험'

OpenAI o3-mini System Card

Specific areas of risk

Disallowed content	✓
Jailbreaks	✓
Hallucinations	✓

Preparedness Scorecard

CBRN	Medium				
Cybersecurity	Low				
Persuasion	Medium				
Model Autonomy	Medium				

자료: OpenAI, 미래에셋증권 리서치센터

(4) 지식 노동자의 종말? Deep Research

여기에 한 술 더 떠, OpenAI는 1월 30일 백악관 브리핑을 끝낸 며칠 후, 2월 2일 새로운 기능을 출시했다. 출시에 관한 "보법" 자체가 달라졌다는 게 핵심이다. **Deep Research라는 이 새로운 기능은 풀 버전의 o3 모델을 기반으로 하기 때문에, 상당히 비싼 모델**이라 할 수 있다. 따라서, 이 기능은 월 200달러 유료 구독자에게만 오픈된 상태다. 본질적으로, 이것은 구글이 이미 출시한 적이 있는 Deep Research 아이디어에다가 o3 모델을 결합해 다단계 추론이 가능하다는 점이 특징으로 꼽힌다.

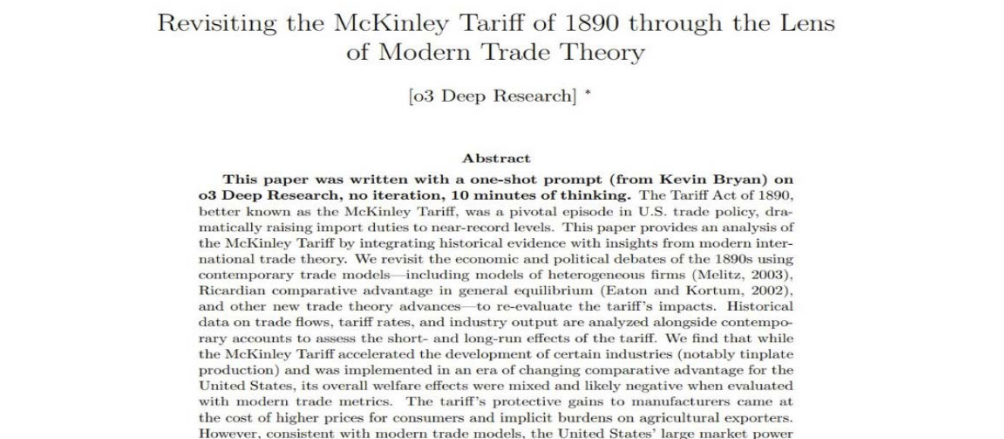
** o3 모델의 특징: ARC-AGI 벤치마크같이 어려운 문제를 풀 때, o3 모델은 1,000개의 샘플을 사용하여 문제를 해결. 하나의 문제를 해결하기 위해 1,000번의 시도를 했다는 것을 의미. 이는 '다중 샘플링'이라는 방법. 기존의 단일 추론 방식과는 달리, 모델이 여러 가능성을 탐색하고 가장 적합한 답을 선택하도록 돕는 것. SemiAnalysis의 딜런 파텔은 o3 모델이 ARC-AGI 문제를 해결하는 데 5~20달러의 비용이 소요되는 것으로 추정. 이는 o3 모델이 기존의 모델에 비해 훨씬 더 많은 컴퓨팅 자원을 필요로 한다는 것을 의미.*

특히, 추론 모델은 훈련 비용뿐만 아니라 추론 비용도 고려해야 하는데, o3 모델은 1,000개의 샘플을 생성해야 하기 때문에 추론 비용이 크게 증가. 그는 현재 기술 수준에서 추론 모델을 모든 문제에 적용하는 것은 비현실적이라면서 추론 모델의 활용은 "고 부가가치 작업"에 집중될 것이라고 예상.

따라서, 비싼 가격에도 불구하고, 엄청난 효용을 지닌 모델로 AI 커뮤니티의 감탄을 자아내고 있다. Deep Research가 얼마나 뛰어난지 본인들의 결과물을 뽐내는 트윗이 현재 봇물 터지듯 올라오고 있다. 우리 팀이 목격한 가장 **놀라운 사례 중 한 가지는 하나의 논문 주제만 던져주면 스스로 논문 작성의 방향까지 짚고, 목차를 만들고, 주장에 걸맞은 근거자료를 찾기 위해 웹 브라우저를 해서 논문을 완성하는 능력이다**. 이를 위해 모델이 내부적으로 사고하고 계획하고 하는 CoT의 과정이 필요한데, o1 때와는 차원이 달리 무려 30분을 생각할 때도 있다고 한다. Deep Research 모델에 대한 조기 접근 권한을 받아 본 토론토대학교 경제학 교수는 이런 경험을 전하면서, 놀라운 소감을 덧붙였다.

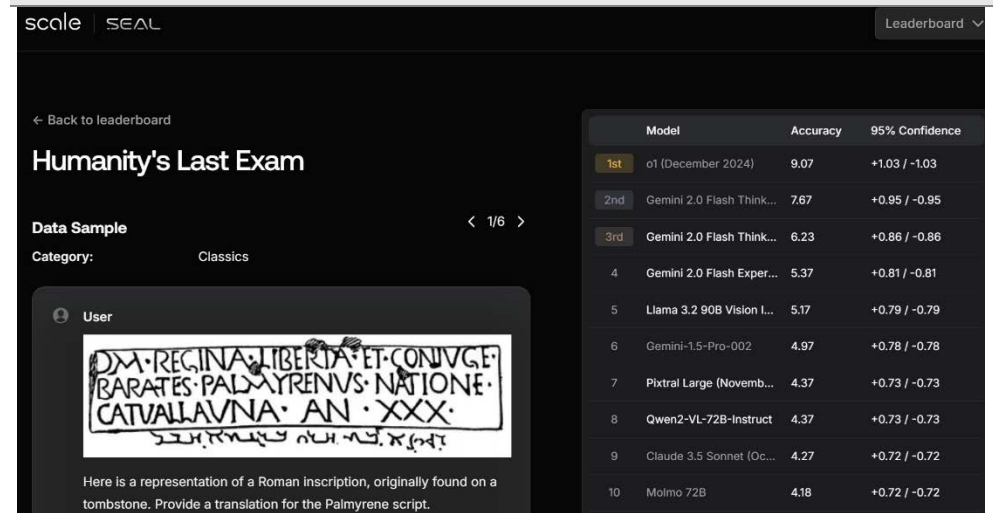
“솔직히 말해서, 제가 심사했던 (인간의) 논문들 중에는 이것보다 못한 것들도 있었습니다. 연구 속도를 대폭 높일 수 있는 단계로 가고 있다는 사실은 매우 명백합니다. o3는 우리 모두가 사용할 수 있는 o3-mini-high와 비교해도 매우매우 인상적입니다. 인류의 마지막 AI 테스트라고 하는 벤치마크에서조차 o3 Deep Research는 26%의 점수를 기록했습니다. 현재까지 최고 수준은 작년 12월에 o1이 달성했던 9%였습니다. 학계에 대해 말하자면, 앞으로 받게 될 AI 수준의 논문들이 정말 엄청날 것입니다. 2022년만 하더라도 저는 학부생들이 AI를 사용해 B학점은 받을 수 있다고 말했었습니다. 그리고 이제는 B급 저널의 경우, 하루 만에 AI로 쓴 논문을 게재할 수 있다고 확신합니다.”

그림 16. 한 토론토대학교 경제학 교수가 o3에게 논문을 작성하게 함, ‘30분이 걸려 논문 작성 완료’



자료: X(Afinethemore), 미래에셋증권 리서치센터

그림 17. Scale AI와 AI 안전센터(CAIS)가 공동으로 개발한 궁극의 AI 능력 테스트인 “HLE”의 예시

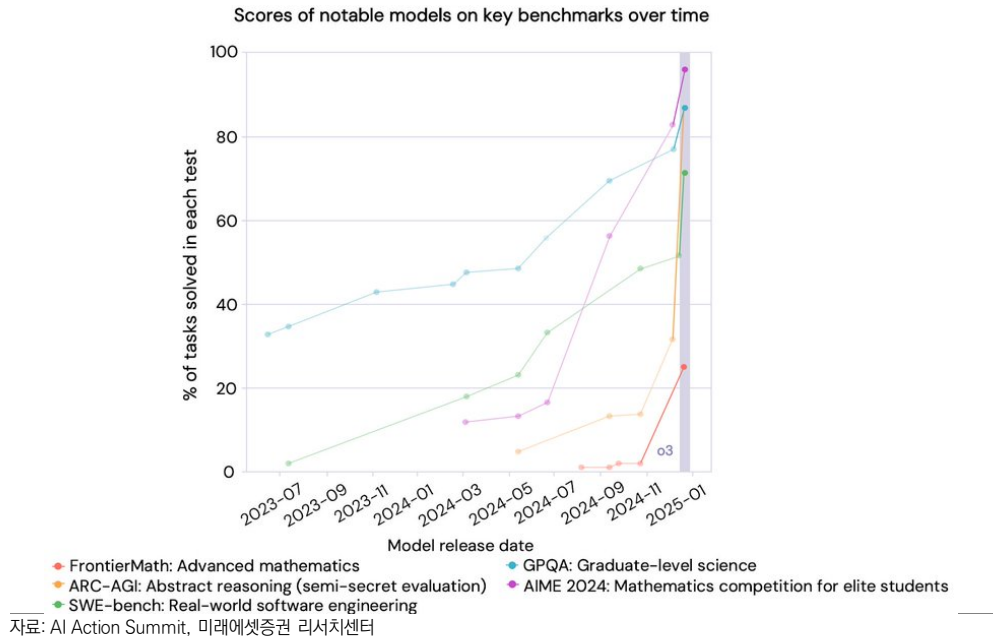


자료: Scale AI, 미래에셋증권 리서치센터

그의 말처럼, Deep Research로 발표된 이 o3 모델은 박사 학위 수준의 수학 벤치마크는 '완전히' 뛰어넘은 수준이라고 한다. 실제로 OpenAI는 최근 o3의 벤치마크 결과를 공유했는데, 벤치마크 점수를 나타내는 지표를 보면 기울기가 거의 수직으로 우뚝 솟아 있음을 알 수 있다. 또한, GAIA 벤치마크에서도 72-73%의 높은 점수를 기록했는데, 불과 9개월 전 AI 모델들의 점수가 15%였다는 점을 감안하면 이는 엄청난 성능 진전 속도라 할 수 있다.

* GAIA는 AI가 실생활에서 복잡미묘한 작업을 수행할 수 있는지 평가하는, 보다 현실적이고 실용적인 벤치마크

그림 18. 시간별 최고의 AI 모델들의 벤치마크 점수, ‘1월에 출시된 o3로 기울기는 거의 수직이 됨’

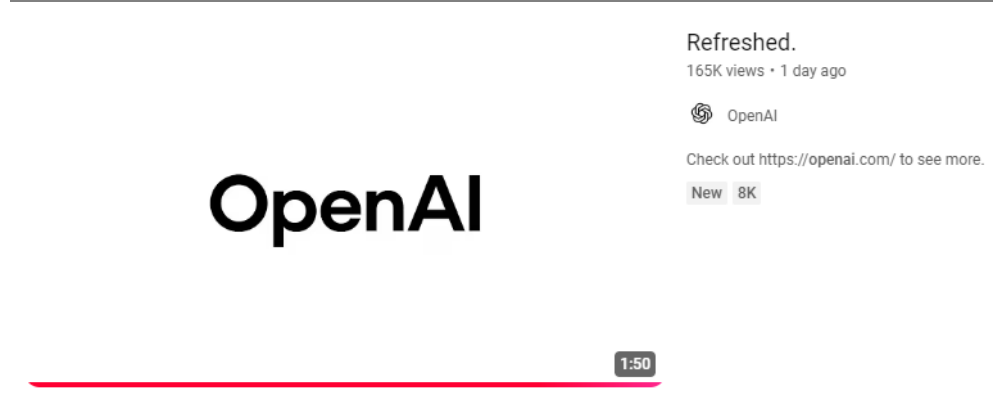


다리오 아모데이의 말처럼 새로운 패러다임의 전환기에 와 있기 때문에 성능 향상의 곡선 기울기가 이전과는 달라졌다고 볼 수 있다. 즉, 완전한 o3는 o1은 물론이고 R1과 차원이 한참 다른 모델이다. 게다가 최근 인터뷰에 따르면, 당연한 소리지만 OpenAI는 현재 다음 모델을 훈련 중이라고 했다. **o4 모델이 개발 중이라는 것이 공식적으로 확인된 것이다.** **OpenAI는 이제 진심 모드로 들어왔다.** 샘 알트만은 최근 나폴레옹의 말을 인용했다.

"혁명은 만들 수도, 멈출 수도 없다. 할 수 있는 유일한 일은,
그 여러 자녀 중 한 명이 승리를 통해 방향을 제시하는 것이다." – 나폴레옹

당연하게도 자신의 상황에 빗대어 이야기했을 것이다. **중국의 본격 참전으로 AI 발전 가속은 불가피하고 현재의 1위인 OpenAI가 그 방향을 이끌어가겠다는 의도**를 드러낸 것으로 보인다. 다소 거만하게도 들리지만 사실이다. 현재로서는 말이다.

그림 19. OpenAI에서 지난 2월 5일에 게시한 브랜딩 영상, “저희, 새로 시작합니다”



자료: YouTube(@OpenAI), 미래에셋증권 리서치센터

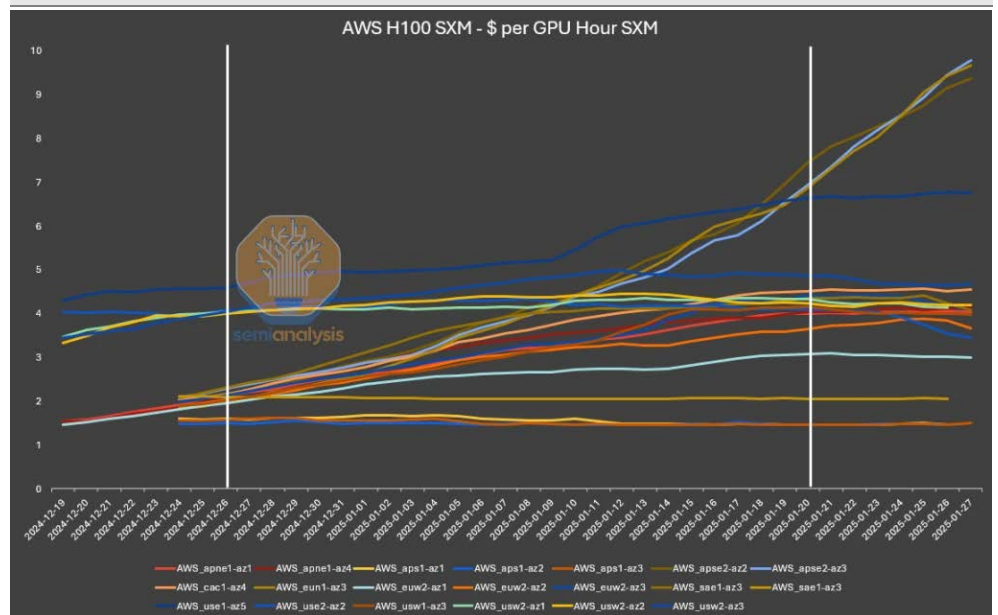
II. AI Issue: Nvidia Upside Down?

1. DeepSeek와 GPU

(1) 제본스의 역설: 저비용 고효율 AI가 가져올 GPU 수요 증가

DeepSeek V3 및 R1 등 '저비용 고효율' AI 모델의 등장은 고성능 GPU에 대한 수요 감소로 비춰질 수 있다. 하지만, 오히려 엔비디아에게 새로운 도약의 기회를 제공할 수도 있지 않을까 한다. 그리고 실제로 V3 및 R1 출시 이후 많은 지역(region)에서 H100에 대한 AWS GPU 가격이 상승했다. H200도 마찬가지다. H200은 H100 대비 더 많은 메모리를 가지고 있기 때문에 거의 모든 곳에서 재고가 거의 없는 실정이라고 알려진다. 이러한 흐름은 이전 몇 달 동안 H100 현물 가격이 부진했던 것과는 달라진 흐름이다. 이는 DeepSeek 충격파를 받은 시장의 반응과는 사뭇 다른 지표다. 더 저렴한 가격에 더 많은 지능은 더 많은 수요를 방증한다고 볼 수 있다.

그림 20. AWS에서 제공하는 region(데이터센터가 위치한 지리)별 H100 SXM의 시간당 가격 변화



자료: Semianalysis, 미래에셋증권 리서치센터

주: SXM은 엔비디아 GPU의 form factor 중 하나로, HBM와 NVLink 인터커넥트를 통해 일반적인 PCIe 버전보다 높은 성능을 제공

그리고 DeepSeek R1으로부터 지식 증류된 모델의 숫자가 늘어날수록, AI 생태계의 시장 규모 자체를 빠르게 확장시킬 수 있다. 즉, 생성 AI가 전방면으로 침투하게 되면서 비로소 대중화 물결이 이는 것이라고 생각한다. 이것은 곧 "제본스의 역설"로 설명될 수 있다. 제본스의 역설은 1865년, 영국의 경제학자 윌리엄 제본스가 석탄 사용의 효율성 증가가 예상된 감소가 아닌 오히려 소비 증가로 이어진 이유를 설명한 데서 시작된 이론이다. 그는 향상된 효율성이 비용을 낮추어 더 넓은 범위의 산업에서 수요가 증가했다고 설명했다. AI도 마찬가지일 수 있다. DeepSeek로 말미암아 컴퓨팅 자원 사용의 효율성이 향상되더라도, 그 자원에 대한 수요가 더욱 증가하여 결과적으로 사용량이 오히려 증가하는 현상일 수 있다는 말이다.

앞으로 더 많은 AI 서비스들이 대중화될 수 있고, 이것은 곧 진정한 의미의 AI 시대를 의미하게 될 것이다. 마이크로소프트 CEO 사티아 나델라 또한 AI에서 제본스의 역설을 강조하며, AI가 더 효율적이고 접근 가능해짐에 따라 사용량이 급증하여 끊임없이 수요가 증가하는 필수 상품으로 변모할 것이라고 역설했다.

그림 21. DeepSeek가 엔비디아 시가총액 수백조원을 날린 그 날, 사티아 나델라가 게시한 트윗 제본스의 역설처럼, AI가 더 효율적으로 바뀔에 따라, AI는 대중화될 거고 수요는 로켓처럼 올라갈 것



자료: X(@satyanadella), 미래에셋증권 리서치센터

이와 관련해, 구글 딥마인드의 한 연구 과학자는 "DeepSeek R1 같은 모델이 박사 과정 학생처럼 컴퓨팅 자원이 제한된 사람들에게 흥미로운 기회를 제공할 것"이라고 말했다. 전 세계 수천개의 기업/기관/대학이 AI가 도입된 프로젝트들을 더 많이 수행하고 그 연구들의 ROI(투자수익률)를 확인하게 되면, 컴퓨팅 자원에 대한 수요가 증가할까, 감소할까? 사실 이와 관련한 최신의 사례는 이미 있다.

그림 22. DeepSeek가 엔비디아 시가총액 수백조원을 날린 그 날, 에마드 모스타카가 게시한 트윗 DeepSeek R1의 등장은 오히려 엔비디아에게 호재



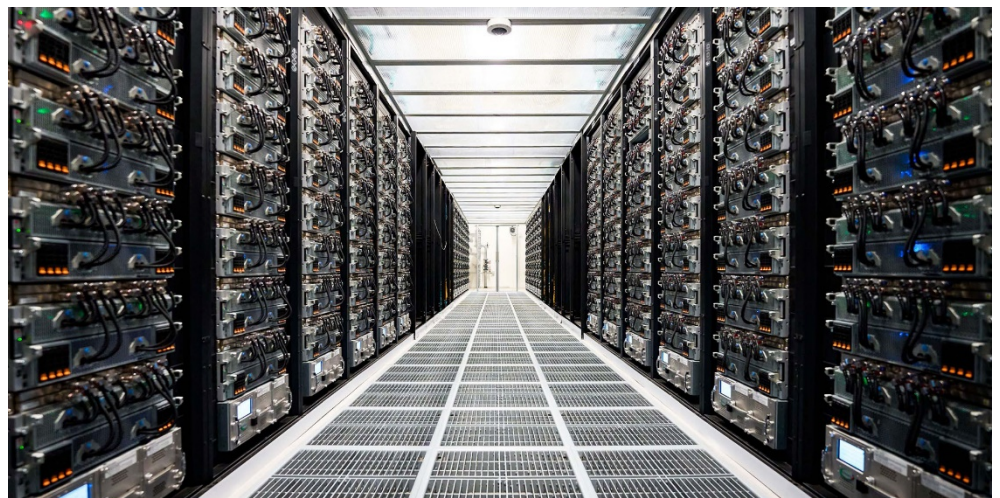
자료: X(@EMostaque), 미래에셋증권 리서치센터

Stable Diffusion을 오픈 소스로 출시한 이후, “더 작고 빠르며 더 나은 버전의 폐쇄 이미지 모델이 클라우드의 모든 GPU 용량을 완전히 소진시켰다”라고 창업자 **Emad Mostaque**는 **최근 밝힌 바 있다. 그는 R1의 출현이, 되려 엔비디아에게는 상승 요인이라고 말한 것이다.** 그리고 이전 장에서 살펴보았듯, DeepSeek는 역사적으로 엔비디아의 대규모 고객이었던 회사였다.

(2) 사례로 생각해보는 제본스 역설

제본스 역설에 대해서 하나 생각해 봐야 할 흥미로운 지점이 있다. 올해 6월 일부 지역에서 Robotaxi 사업을 시작하는 것을 목표로 하고 있다는 테슬라는 작년에 AI 훈련 컴퓨팅을 400% 이상 늘렸다. 이 흐름은 올해도 지속될 가능성이 높는데, 만약 DeepSeek의 효율성을 접목하게 된다면 어떻게 될까?

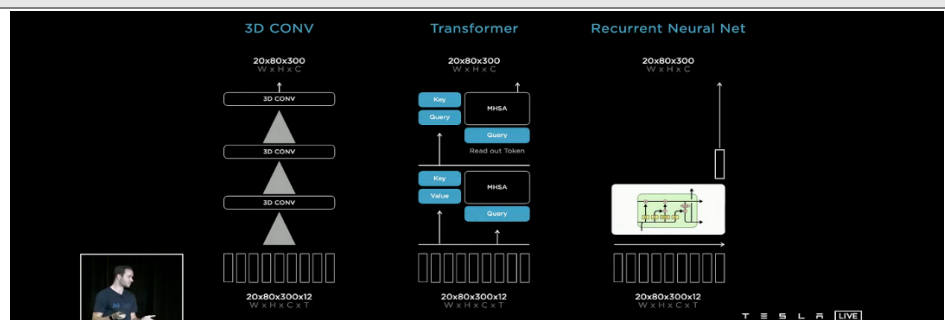
그림 23. Tesla가 FSD 훈련을 위해 건설한 슈퍼컴퓨터 클러스터 “Cortex”
엔비디아 Hopper 시리즈 GPU가 5만개 이상이 투입된 상태



자료: Tesla Oracle, 미래에셋증권 리서치센터

테슬라의 FSD도 결국 트랜스포머 기반이다. 따라서, R1이 제시한 여러가지 엔지니어링 기법이 FSD 알고리즘의 추가적인 최적화로 이어질 수 있고, 더 높은 수준의 자율주행 성능, 더 많은 FSD 서비스 채택으로 이어질 가능성이 있다. 이것은 세상에 없던 새로운 시장이 열리는 것이기 때문에 차원이 다른 수준의 AI 인프라 수요를 불러일으킬 수 있다.

그림 24. Tesla 디렉터였던 안드레 카파시가 AI Day 2021에서 FSD를 설명, ‘트랜스포머가 핵심’



자료: Tesla, 미래에셋증권 리서치센터

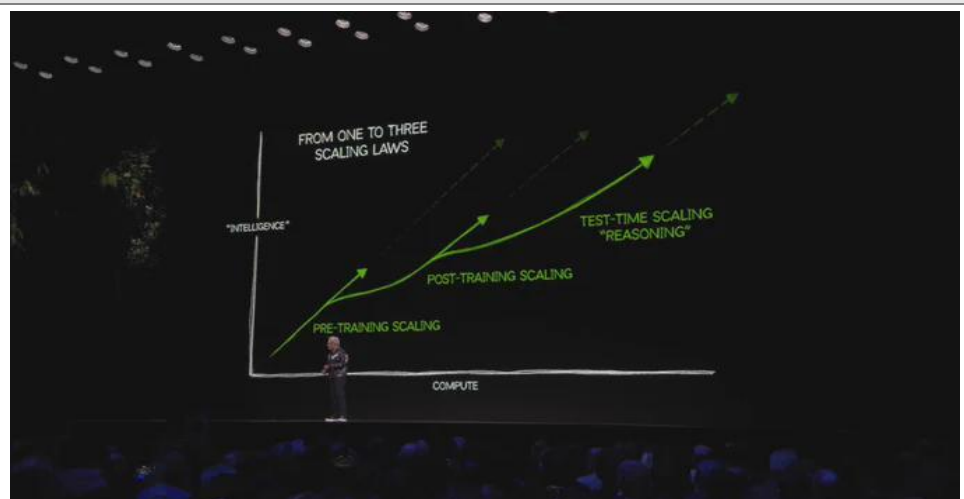
그래서 우리 팀은 제본스의 역설의 발생 가능성은 실제로 높을 것이라 판단한다. 단순히 현재의 챗봇의 영역 말고도, 자율주행이나 더 나아가 로봇틱스와 같이 더 거대한 영역에서의 AI 혁명이 발생할 때의 파급력까지 감안하면 그렇기 때문이다. 이에 필요한 컴퓨팅량은 일반 상식을 뛰어넘을 정도로 막대할 수 있다.

결국, 문제는 어떤 것이 훨씬 저렴해졌을 때, 그것을 사람들이 얼마나 더 원하는가 하는 것이다. 예를 들어, 집에 베개가 여러 개 있는데, 내일 모든 베개가 반값으로 할인 판매된다고 해도 베개를 더 살 사람은 없을 것이다. 가격이 어떻게 내 삶에 꼭 더 필요한 것은 아니니까 말이다. 하지만 만약 꽃등심이 전반적으로 할인 판매된다면, 아마 그 날은 많은 사람들이 때문에 마트 계산대에 오래 서 있어야 할 것이다. 경제학 용어로 가격 탄력성(Price Elasticity)이라고 한다. 베개와 꽃등심 사이에서 "지능의 가격 탄력성"은 어디에 위치할까, 베개보다는 꽃등심 아닐까?

(3) R1의 숨겨진 진실: GPU 수요와 중국의 전략적 침묵?

DeepSeek의 R1이나 OpenAI의 o1 모델 같은 reasoning model(추론 모델)은 앞으로의 패러다임이고 이미 주류가 됐다. 구글과 앤트로픽, 그리고 메타도 이 추론 모델의 패러다임에 선봉에 서려고 부단히 노력 중이다. 이 모델들은 이전 모델과는 달리 더 많은 상당한 수의 엔비디아 GPU와 고성능 네트워킹이 필요해지게 된다. 왜냐면 이제는 사전훈련(pre-training), 사후훈련(post-training)과 더불어 "Test-Time 스케일링"이라는 세 가지 스케일링 법칙이 존재하는 시대이기 때문이다.

그림 25. 젠슨 황이 지난 CES 2025에서 말한 새로운 패러다임의 시대
젠슨 황은 사후학습과 Test-Time으로 스케일링 법칙의 확장되고 있다고 강조



자료: CES 2025, Nvidia, 미래에셋증권 리서치센터

DeepSeek의 이전 패러다임 모델인 V3로부터 현재 패러다임 모델인 R1을 만들기 위해서는, 합성 데이터를 대량으로 생성해야 하고 또 이를 기반으로 강화학습을 해야 한다. 여기에는 상당한 양의 컴퓨팅이 필요하다.

예를 들어, OpenAI에 재직 중인 익명의 유저가 레딧에 최근 남긴 글에 따르면, (고품질의) 합성 데이터 생성에만 3개월이 소요되었다고 했다. 만약에, OpenAI가 엔비디아 Hopper 시리즈 GPU 10만개를 시간당 3달러에 임대한다고 가정하고, 3개월(2,160시간) 동안의 GPU-hours를 사용했다고 계산해보면, 이 작업에만 약 6.5억 달러가 소모되는 일이 된다. 또한, 이러한 현상을 완벽하게 설명하는 글로, AI 업계 1티어 인플루언서인 안드레 카파시가 DeepSeek R1에 대해 남긴 의견을 공유하면 좋을 것 같다.

"새 알고리즘과 혁신을 위한 전체 실험 과정에서도 계산 능력은 결정적인 역할을 합니다. 데이터는 전통적으로 계산과는 별개의 요소로 여겨졌지만, 실제로는 계산을 통해 대량의 데이터를 생성할 수 있고, 둘은 밀접하게 연결돼 있습니다."

안드레 카파시의 말을 곱씹어 보면, ‘AI의 핵심 자원이 데이터에서 컴퓨팅 리소스로 이동하고 있다’고 봐도 타당한 해석이지 않을까 한다. 이제 데이터라는 것도 인간이 아니라 결국 GPU로 만들어질 수 있는 시대이기 때문이다.

게다가, R1 모델을 대표하는 알고리즘 특징은 "GRPO"다. GRPO는 여러 샘플(답변)을 병렬로 생성하고 평가하는 방식이기 때문에, 많은 수의 GPU를 동시에 활용해야 최적화가 가능한 방식이다. R1 훈련 방식이 단순하고 효율성이 높다는 것과 GPU가 앞으로 덜 부족해질 것이라는 말은 동의어가 아니다.

** GRPO: Group Relative Policy Optimization의 약자로 DeepSeek가 독자 개발한 강화학습 기법. 동일한 질문(prompt)에 대해 여러 개의 답변(response)을 생성하고, 이 답변 그룹 내에서의 상대적인 순위를 매겨 각 답변의 우열을 평가하는 최적화 기법.*

R1 논문에 따르면, 주어진 prompt에 대해 64개의 다양한 답변을 생성했음. 각 답변의 보상에서 그룹 내 다른 답변 보상의 평균을 빼고, 이를 표준편차로 나누어 정규화(normalize). 이 값을 "상대적 이점(advantage)"이라고 하며, 이는 해당 답변이 다른 답변들에 비해 얼마나 더 좋은지를 나타내는 지표. MCTS와 같은 복잡한 트리 탐색 알고리즘을 사용하지 않는 대신, GRPO 알고리즘을 통해 효율적으로 학습했음.

그리고 **R1 논문에서 개인적으로 가장 의문스러웠던 점은 V3 때와는 달리 컴퓨팅 자원이 얼마나 소모되었는지 전혀 언급하지 않았다는 점**이다. 그들이 갑자기 이 중대한 것을 놓칠 리 없다. 이것은 분명히 의도한 것이고, 보여주고 싶지 않거나 어떤 이유에서인지 보여주지 못하는 것일 수 있다. 실상은, V3보다 훨씬 많은 GPU 자원이 R1에 사용되었을 가능성이 높다. 추론 모델 개발에는 '대량의 실험'이 필요하기 때문이다.

따라서 **논문에 컴퓨팅 사용량이 빠진 이유에 대해서 쉽게 짐작해보면, (그러면 안 되지만) 그들이 꽤 많은 엔비디아 GPU를 보유하고 있음을 시사**한다. 다리오 아모데이가 며칠 전에 쓴 블로그 글은 흥미로운 게 정말 많지만, 그는 특히 "DeepSeek가 보유한 것으로 추정되는 칩 중 상당 부분은 곧 금지될 것으로 예측되는 H20, 금지되기 전에 배송된 H800, 그리고 밀수되었을 가능성이 매우 높은 H100으로 구성된 것으로 보인다"라고 말한 게 '킵'이었다. 참고로, H100 칩은 애초에 출시 때부터 수출 금지 품목이었으니, 그들이 이 칩을 사용했다면, 불법적인 경로를 통해 획득했다는 말이다.

표 3. 중국이 아직 합법적으로 수입이 가능한 H20과 그들이 원래 갖고 싶어했을 H100의 스펙 비교

엔비디아 Hopper 시리즈	H20	H100
FP8/INT8 Tensor	296 TFLOPS	1979 TFLOPS
성능 밀도	2.9	19.4
전력 소비(TDP)	400W	700W
NVLink 대역폭	900 GB/s	900 GB/s
메모리 용량	96GB HBM3	80GB HBM3
메모리 대역폭	4.0 TB/s	3.4 TB/s

자료: 미래에셋증권 리서치센터

이어서 아모데이는, 중국의 최고 AI 칩이라 할 수 있는 화웨이의 Ascend 시리즈는 엔비디아 GPU보다 성능이 훨씬 떨어지고, 중국 외부에 주목할 만한 Ascend 칩 클러스터가 단 하나도 없다는 것은 중국이 국내 수요를 충족하는 데 어려움을 겪고 있음을 강조하기도 했다. 중국은 여전히 절실하게 엔비디아 칩을 원하게 된다는 말로 귀결된다.

그림 26. 화웨이는 기존의 Kirin(AP) 설계 역량을 Ascend(AI 가속기)로 업그레이드하려 노력 중



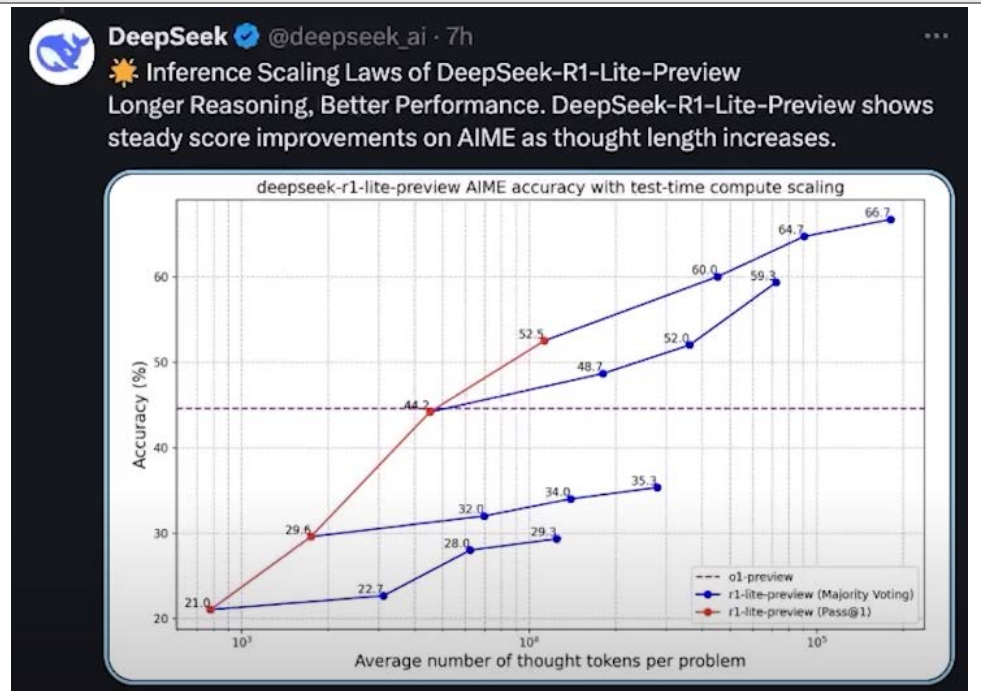
자료: Huawei, 미래에셋증권 리서치센터

그런데, 중국 입장에서 본인의 절실함을 상대방에게 드러내는 것은 트럼프의 저서 "거래의 기술"에 따르면 절대 하지 말아야 할 자세다. 그러니 사실 중국 측은 '엔비디아 없이도 괜찮아'라면서, 자체 제작 AI 가속기 대량생산 능력을 홍보해야 하는 게 상책이라고 할 수 있다. DeepSeek 때문에 수출 통제가 효과가 없다는 내러티브 형성과 DeepSeek가 GPU 보유 숫자를 공개하지 않는 것도 이런 것의 일환일 수 있다. 이와 관련해, 팔란티어만큼이나 미국방부의 신임을 받는, AI 무기 시스템 기업 Anduril Industries의 CEO Palmer Luckey는 "DeepSeek이 인상적이긴 하지만, 500만 달러 훈련 비용에 관해서는, 중국 공산당의 선전이 있다"고 공개적으로 말하기도 했다.

(4) AI 추론 시대의 가속화될 GPU 수요

뿐만 아니라, AI 모델을 훈련(사후훈련 포함)하는 것뿐만 아니라, 한 번 훈련된 모델을 가지고 추론(inference)하는 데에 훨씬 더 많은 GPU가 필요하다. 그리고 물론 DeepSeek R1이 글로벌 투자자들의 관심을 받게 된 것도 바로 reasoning 모델이기 때문이었다.

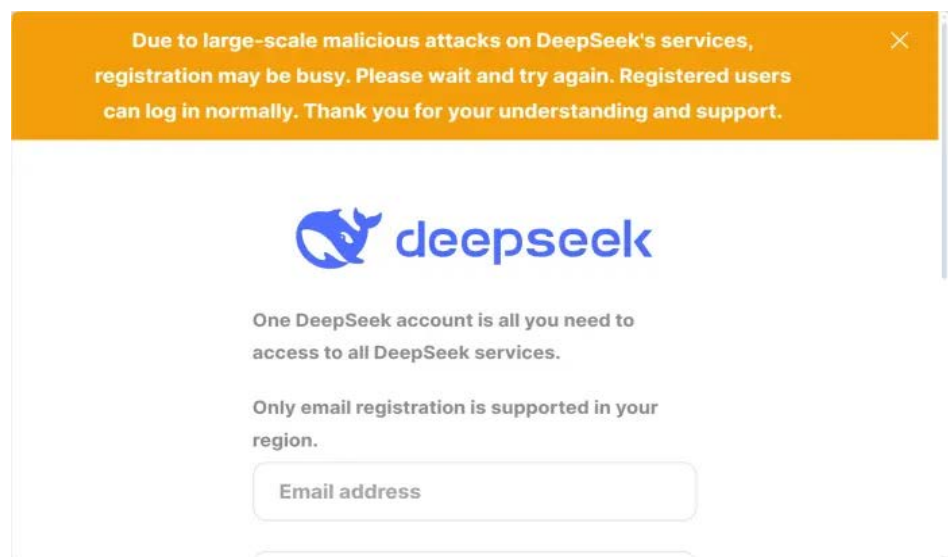
그림 27. DeepSeek R1 관련 논문에서의 중요 포인트
‘DeepSeek R1 모델에게 “생각할 시간”을 더 주니 (스케일링 법칙대로) 수학을 더 잘하더라’



자료: DeepSeek, X(@deepseek_ai), 미래에셋증권 리서치센터

예를 들어, 메타는 Llama 3를 훈련하는 데 1.6만개의 GPU를 사용했지만, 전체 AI 서비스에는 40만 개 이상의 GPU를 운영 중이다. 더군다나 DeepSeek R1의 경우, GPU가 부족한 상태로 이 추론 서비스를 하려면 가랑이가 찢어질 지경일 수 있다. 실제로 이미 DeepSeek는 정상적으로 서비스를 제공해 주지 못하는 실정이다. DeepSeek 앱을 다운로드하고 가입하려고 할 때에 "우리는 신규 등록을 받지 않습니다"라고 공지하기도 했다. 물론 그들은 악의적인 사이버 공격 때문이라고 했지만, 사실 모델을 제공할 충분한 컴퓨팅 자원이 없기 때문일 것이라 판단한다.

그림 28. DeepSeek에 신규가입하려고 할 때 뜬 메시지, ‘사이버 공격(?) 때문에 가입 어려워요’



자료: DeepSeek, 미래에셋증권 리서치센터

따라서, DeepSeek의 출현 이후에 엔비디아를 포함한 AI 인프라 기업들이 급락한 이유에는 사실 고개가 가우뚱해진다. **DeepSeek R1의 등장이 엔비디아에 불확실성을 안겨준 것은 맞지만, '위기'라는 분석은 다소 근시안적이라고 주장하고 싶다. 다리오 아모데이는 R1 때문에 엔비디아 주가가 17% 하락한 것은 "당황스러운 일"이라고까지 말했다.** DeepSeek의 빠른 도약과 관련해 그는 지난 1월 다보스 포럼에서 놀라운 주장을 했었다.

"2026년에는 수백 만 개, 2027년에는 잠재적으로 수천 만 개의 칩이 사용될 것으로 예상합니다. 새로운 종류의 강화학습 스케일링과 함께 패러다임이 전환되고 있습니다. 모든 회사가 이것을 하고 있죠. 현재의 초기 국면에는 엄청난 양의 컴퓨팅 없이도 강력한 모델을 만드는 것이 가능합니다. 동시에 중국은 수만 개의 칩을 밀수할 수 있었습니다. 미국의 대중 수출 통제의 핵심은 중국이 현재 수만 개의 칩을 보유하는 것을 막고자 하는 게 아니라, 수십 만, 수백 만 개의 칩을 보유하는 것을 막는 것입니다. (이 정도 큰 금액은 중국이 밀수 행위를 숨기기 어려울 것) 만약 그들이 그렇게 한다면, 우리는 동등한 수준이 되거나 그들이 앞서 나갈 수도 있습니다. 물론, 사전 훈련의 스케일링 법칙도 계속되고 있다고 생각합니다만, 이제는 두 가지 소스가 있는 거죠. 그래서 스케일링 법칙의 트렌드는 더 견고해졌고, 멈추기가 더 어려워졌습니다."

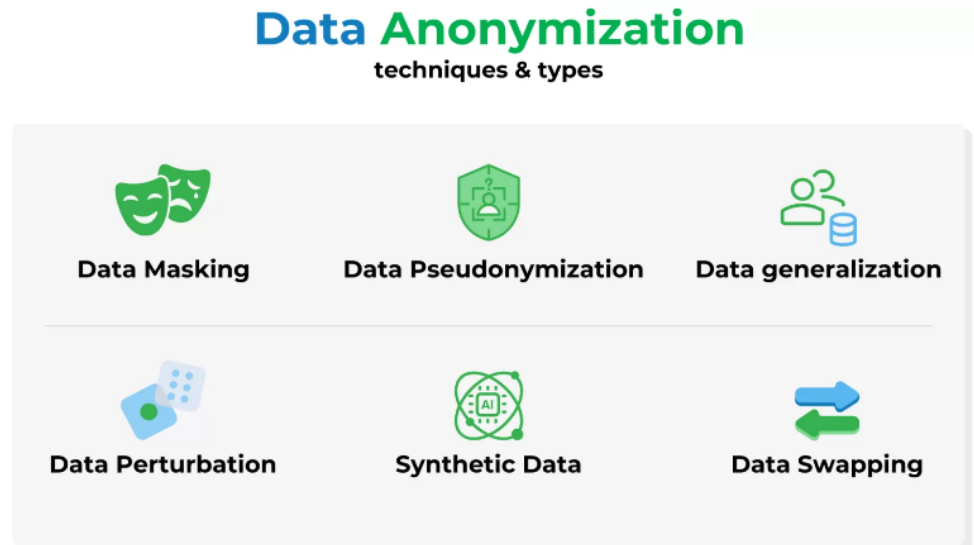
즉, 현재는 적은 컴퓨팅으로도 좋은 모델을 만들 수 있지만, 미래의 AI 패권을 위해서는 여전히 압도적인 컴퓨팅 파워가 필수적이라는 맥락이다. ChinaTalk에 따르면, 심지어 중국 기술 리더들도 "DeepSeek R1은 중국의 승리를 의미하지는 않으며 미중간 경쟁이 장기적 성격을 띠고, 고비용 구조임을 강조해야 하는 상황"이라고 말했다고 한다. 즉, **스케일링이 필요한 것은 중국도 알고 있고 AI에 관한 투자 비용이 앞으로 더 많을 것**이라는 이야기다.

* ChinaTalk은 Jordan Schneider가 운영하는, 중국의 기술 및 미중 관계를 전문적으로 다루는 영향력 있는 뉴스레터 및 팟캐스트 플랫폼. 포춘 500대 기업의 임원진, 세계 최대 국부펀드, 헤지펀드 매니저들이 주요 독자층. 미국 정부 고위 관료, 의회 스태프, 정책 입안자들이 구독하며 중국 관련 정책 결정에 참고함. Jordan Schneider는 2017년부터 ChinaTalk이라는 뉴스레터와 팟캐스트를 운영. 현재 미국의 싱크탱크 CNAS(Center for a New American Security)의 연구원

(5) 량원평의 호소로 더 노골적으로 변할 AI 패권 경쟁

한편, **지난 AI Weekly에서 작성한 대로, 량원평 CEO는 리창 총리가 배석한 정부보고좌담회에 참석한 유일한 AGI 모델 제작자였다.** 참고로 그 회의의 성격은 '최고 지도자가 업계 리더로부터 배우다'는 설정의 업무 보고 자리였다. 공산당의 총애를 받을 DeepSeek는 앞으로 더 좋은 모델 훈련을 위해 더 높은 품질의 데이터를 제공받을 가능성도 있다고 본다. 중국 정부 및 수많은 국영기관들은 꽤 많은 데이터를 보유하고 있기 때문에 이는 중대한 변화일 수 있다. 또한 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 예측된다.

그림 29. 데이터 비식별화(Anonymization)는 개인정보를 보호하면서 데이터의 유용성을 유지
원 데이터에서 개인을 식별할 수 있는 요소를 제거/변형하여, 특정 개인을 알아볼 수 없도록 함



자료: Syntho AI, 미래에셋증권 리서치센터

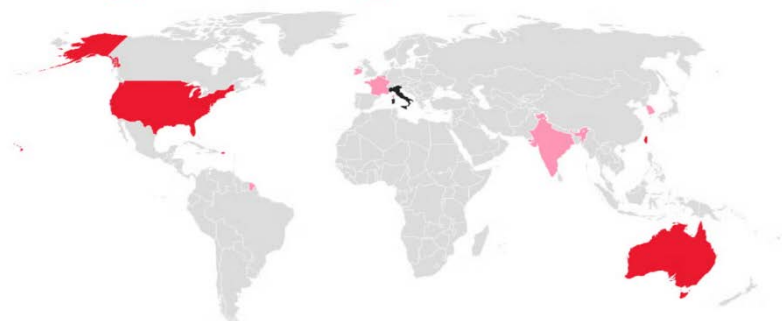
그러나 더 중요한 것은, 해당 좌담회 자리에서 량원평이 한 말이다. 그는 “더 많은 칩이 필요하고, 하드웨어가 제약적이며, 수출 통제가 정말 해롭고, 더 많은 칩이 있으면 좋겠다”고 말했다고 한다.

이후 중국은행은 DeepSeek의 량원평 CEO와 만난 지 하루 만에 향후 5년 동안 AI 산업 체인에 1,400억 달러(1조 위안)의 보조금을 발표하기도 했다. 그리고 무엇보다, **중국의 국가 최고 지도자의 레이더에 이러한 AI 반도체 문제의 이슈가 더욱 깊숙이 들어갔다는 사실 자체가 중요하고 중국은 칩 확보를 위해 사력을 다할 것이다.** 만약, 중국의 생각대로 그들이 AI 칩 생산 능력을 잘 확보한다면, AI 군비 경쟁은 미국과 중국의 진영 싸움(G7 vs BRICS)의 “다극 체제”로 형성된다. 이 경우 엔비디아나 OpenAI를 포함한 “미국계” AI 기업들의 글로벌 시장 파이는, 현재까지의 미국 AI 주도의 “단극 체제”에 비해 절반 이하로 줄어들 리스크가 있다.

그림 30. DeepSeek 사용 금지에 대한 글로벌 국가들의 스탠스를 보면 AI도 진영별 경쟁
이탈리아는 완전히 금지… 일반적으로 미국의 우방국들은 사용금지 흐름으로 가닥을 잡고 있음

Where is DeepSeek banned?

■ Complete ban ■ Government or partial ban ■ Investigating



Map: Data from 4 February 2025

INDEPENDENT

자료: Independent, 미래에셋증권 리서치센터

(6) 미국의 새로운 AI 칩 규제: GPU 밀수와 우회 수입 차단을 위한 3등급 전략

그러나 중국의 희망과는 달리 미국은 고삐를 더 쥐고 있는 상황이다. 바이든 정부는 백악관을 비우기 직전 지난 1월 대중 규제를 추가로 발표했다. 이 규제는 세 개의 국가 등급으로 나눠, AI 칩 수출 통제 정책을 강경하게 하겠다는 것이다. 1등급의 경우 AI 칩 수입에 제한이 전혀 없고 완전한 접근 권한을 보유하게 된다. 다행히 대한민국은 이 등급에 속한다. 그리고 3등급에 속한 나라는 모든 고급 AI 칩에 관한 수입이 원천 금지된다.

표 4. AI 칩 수출 통제 정책의 3단계 등급 체계, 컴퓨팅 파워 총량(TPP) 기준으로 제한 적용

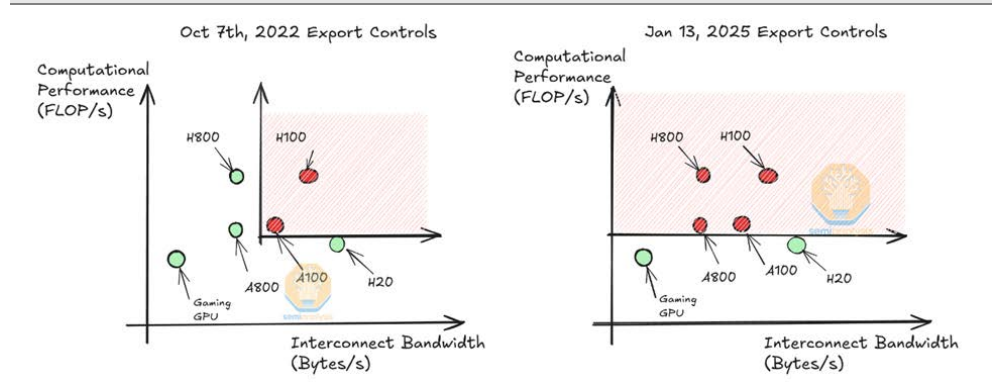
1등급	2등급	3등급
호주	브라질	중국
벨기에	인도	러시아
캐나다	이스라엘	아프가니스탄
덴마크	말레이시아	벨라루스
핀란드	싱가포르	미얀마
프랑스	스위스	캄보디아
독일	오스트리아	중앙아프리카공화국
아일랜드	체코	콩고
이탈리아	그리스	쿠바
일본	인도네시아	에리트레아
네덜란드	케냐	아이티
뉴질랜드	룩셈부르크	이란
노르웨이	멕시코	이라크
대한민국	모나코	레바논
스페인	-	리비아

자료: Data Gravity, 미래에셋증권 리서치센터

주: TPP는 Total Processing Performance의 약자. 엔비디아 GPU와 같은 AI 칩들의 TFLOPS 성능에다가 비트길이를 곱한 값이 특정 TPP를 넘으면 규제 대상으로 잡히게 됨.

재밌는 것은 둘 사이에 속해있는 2등급 나라들에 대한 것이다. 이 국가들은 1,700개의 GPU까지는 라이선스 없이 구매가 가능하다. 그리고 국가별 총 한도가 있는데 5만 개의 GPU, 정부 간 협약으로 10만개의 GPU까지 쿼터제를 확장할 수 있다. 또한 NVEU라고 하는 "국가적으로 인증된 엔드유저 자격"을 획득하게 되면 32만 개까지 이를 늘릴 수 있기는 하다. 그래도 Stargate 프로젝트가 목표로 하는 것처럼 "100만 개의 GPU"에는 못 미치지만 말이다. **흥미로운 점은 바로 이 2등급에 인도와 싱가포르, 그리고 말레이시아가 속한 것이다.** 짚고 넘어가야 할 부분은 싱가포르는 엔비디아 매출의 20%를 차지했던 나라다. 이 부분은 물론 엔비디아에게 타격으로 꼽힐 수 있는 지점이다.

그림 31. 미국이 중국을 겨냥해 시행한 AI 반도체 수출 규제의 변화
2022년 10월 규제 때는 H100과 A100만 금지품목이었는데, 2025년 1월 규제로 H20만 생존



자료: Semianalysis, Lennart Heim, 미래에셋증권 리서치센터

하지만 엔비디아를 금지옥엽으로 생각하는 미국 정부가 이렇게 한 데에는, 더 중요한 맥락이 따로 있다. **중국이 GPU를 우회 수입한 것으로 판단하고, 이러한 밀수(smuggling) 행태를 막으려고 매우 적극적이라는 것을 명시하기 때문이다.** 대부분의 대규모 밀수가 싱가포르와 말레이시아의 회사를 거쳐 발생한다고 미국 정부는 판단하고 있는 것이다.

실제로 중국 회사들이 미국의 클라우드 사업자들의 해외 데이터센터(주로 동남아와 인도)로부터 GPU를 임대하는 것이 믿을 수 없을 정도로 쉽다고 한다. GPU가 아니라 LPU라는 것을 표방하며 유명해진 AI 가속기 업체 Groq의 CEO 조나단 로스는 "GPU 수출 통제가 효과가 없는 것은, 중국 기업들이 미국 클라우드 제공업체에 로그인해서 GPU를 대여할 수 있기 때문"이라고 말했다.

이와 관련해, SemiAnalysis의 연구에 따르면 오라클의 가장 큰 GPU 고객이 바이트댄스이고, 구글의 경우에도 그들이 두 번째 큰 고객이라고 한다. 오라클과 구글과 같은 하이퍼스케일러 뿐만 아니라, 중국 업체들이 접근할 수 있는 소규모 클라우드 회사의 목록은 수십개가 넘어간다. 그리고 이는 지난 3주 전 생겨난 이 규제 이전까지는 완전히 합법적이었다. 그러나 이제는 사실상, '이 나라에서는 이만큼의 GPU만 구입할 수 있고요, 중국 회사들이라면... 다른 나라의 클라우드에서 임대해서 쓰세요. 물론 그것도 제한적이지만요!'라는 뜻으로 쉽게 이해할 수 있다.

그리고 **"트럼프 2.0" 시대의 미국은 이러한 강경함에 있어서 더 기여를 올릴 것으로 판단**한다. 이번에 상무부 장관으로 지명된 하워드 러트닉은 미 의회 인준 청문회에 자리에서 이렇게 말했다.

"DeepSeek가 모두 정당하게 이루어졌다고 믿지 않습니다. 말도 안 됩니다. 그들은 우리에게 침입했고, 우리의 IP를 훔쳐갔습니다. 엔비디아 칩은 DeepSeek 모델을 구동합니다. 그것은 끝나야 합니다. 미국이 선두를 유지하기 위해, 제한을 추구하고 시행하는 데 강경히 나설 것입니다."

(7) 엔비디아의 미래: CUDA 생태계 우위와 전략적 피벗

이처럼 **미국 정치 및 정부 관료들이 점점 더 엔비디아의 비즈니스에 관여를 하는 듯한 발언을 하고 있다. 이는 AI가 냉전 시대를 만들고 있기 때문이고, 現 국면의 가장 중대한 전략 자원이 GPU이기 때문이다.** 그리고 현재로서는 딱히 대안도 별로 없다는 게, AI 개발자들의 시각이다. 투자자가 아니라, 실제로 GPU를 필요로 하는 사람들의 말에 더 귀 기울이는 게 진짜 수요를 짐작하는 데 의미가 있을 것으로 사료된다.

이와 관련해 AI 개발자이자 유명한 AI 관련 뉴스레터인 Interconnects를 운영하는 Nathan Lambert는 "NVIDIA는 현재 모든 것을 안정적으로 수행하는 유일한 회사"라고 지난주 렉스 프리드만의 팟캐스트에서 말했다. 그와 동석한 Dylan Patel(SemiAnalysis 편집장)은 "AMD의 하드웨어가 여러 면에서 엔비디아보다 나은 부분도 있습니다. 문제는 그들의 소프트웨어가 정말 나쁘다는 것입니다. 격차가 너무 크고 (심지어) AMD는 거기에 충분한 자원을 소비하지 않고 있습니다"라고 말했다.

그리고, 오픈소스 진영의 유명한 천재 개발자인 Gerge Hotz 또한 “AMD 칩은 이론적으로는 AI 작업에 활용될 수 있는 큰 잠재력을 가지고 있지만, AMD가 AI 소프트웨어 개발에 충분한 투자를 하지 않았다는 점이 큰 문제입니다. AI 작업에서는 하드웨어의 성능뿐만 아니라 최적화된 소프트웨어 스택이 매우 중요합니다. CUDA와 같은 강력한 소프트웨어 생태계가 없다면, 아무리 뛰어난 하드웨어라도 AI 작업에서 제대로 된 성능을 발휘하기 어렵습니다.”라고 말했다. 실제로 AMD의 많은 AI 라이브러리들은 기존에 엔비디아가 만들어 놓은 라이브러리를 복제(포크)하고 변환하는 식으로 개발하는 경우가 많다. 이 방식은 추론(Inference) 전용으로 특정 모델 구조에 맞출 수는 있지만, 그 외의 다양한 작업에 대해서는 최적화가 전혀 안 돼 성능이 크게 떨어질 수 있다. 그래서 빅테크들의 AI 훈련 작업에 AMD의 MI 시리즈의 GPU의 활용이 제한적인 것으로 생각해볼 수 있다.

위 개발자들의 말을 종합해보면 엔비디아의 대항마로 보이는 업체가 딱히 보이지 않고, 그나마 AMD의 경우에도 소프트웨어 능력 및 관련 투자 부족 때문에 엔비디아에게는 역부족이라는 말이 된다. 참고로 지난 실적 발표 call에서 AMD는 ROCm(CUDA와 대응) 소프트웨어에 대해 "여전히" 구체적인 성능 향상이나 고객 확보 사례는 제한적으로만 언급했다.

이처럼, 엔비디아가 여전히 이 시장의 선두 주자로서 막대한 수혜를 입기 위해서 지켜야 하는 해자는 하드웨어도 있지만, CUDA라고 불리는 강력한 소프트웨어 생태계가 더 중요하다는 말이다. DeepSeek R1 역시 CUDA를 기반으로 개발되었으며, 이는 엔비디아 GPU의 지배력이 여전히 공고함을 보여준다. 즉, 중국 LLM 업체들도 CUDA의 입지가 워낙 공고한 상황이니 이 물에서 놀아야 하는 상황인 셈이다. 결론적으로, 3년 이내 정도의 중기적으로는, 제본스의 역설에 따라 더 많은 컴퓨팅 수요의 수혜를 입을 가능성이 높다고 본다. 개발자들에게 친숙하고, 강력한 성능을 제공하는 CUDA 덕분에 '저비용 고기능' AI 모델이 등장하더라도 엔비디아 GPU에 대한 수요는 적어도 당분간은 지속될 것이기 때문이다.

하지만, 결국 중국에서 안드로이드 OS에 대한 자국산 대안이 나왔듯, 화웨이같은 업체가 이를 장기적으로는 극복할 가능성도 무시 못 한다. 실제로 화웨이는 PyTorch 프레임워크와 긴밀히 협력하면서 CUDA의 자체 버전을 개발 중이다.

또한 그것보다도 더 높은 확률의 리스크로는, 빅테크들의 자체 제작 칩에 관한 개발 노력도 장기적으로는 성과를 얻을 수 있다는 점이다. 따라서 그러한 장기 리스크들이 도래하기 전에, 엔비디아는 “에이전트 AI+옵니버스”의 조합의 자율주행과 로봇틱스 영역에서 소프트웨어 사업자로 변모하기 위해 피벗을 계속 시도할 것이다.

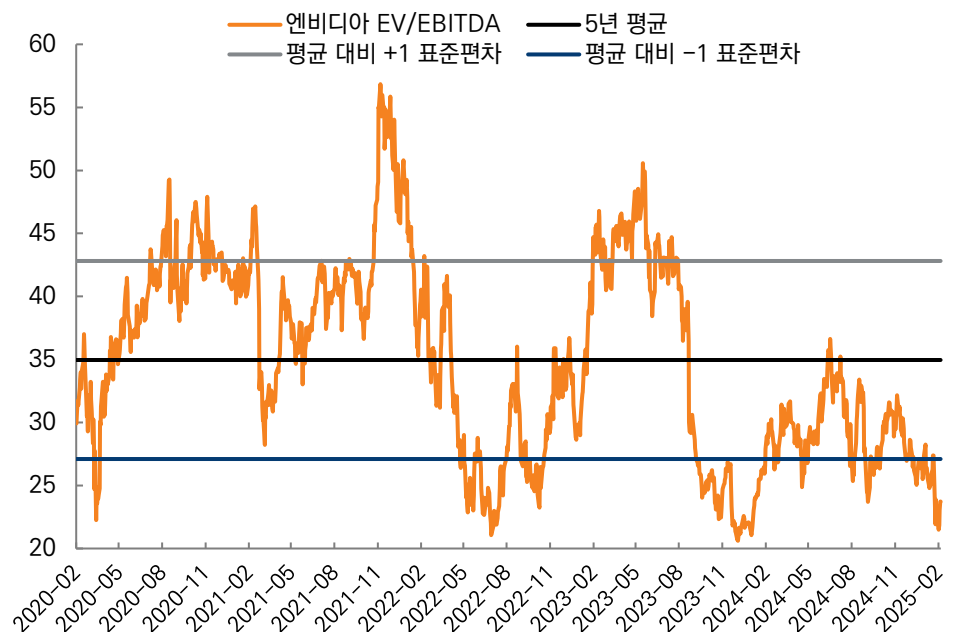
그림 32. 점차 HW 기업에서 SW 기업으로 변모하려는 엔비디아, ‘핵심은 AI 에이전트와 물리적 AI’ 반도체 실리콘부터 옵니버스까지 모든 것을 지원하는 end-to-end AI 기업



자료: Nvidia, 미래에셋증권 리서치센터

그러한 전략이 성공하면 또 다른 리레이팅으로 이어질 것이고, 실패하면 게이밍 그래픽카드 시절의 밸류에이션에 가깝게 회귀할 확률도 물론 있을 수 있다. 전자와 후자의 시나리오 중 어느 쪽이 점차 우세해지는지 매 순간마다 체크해야 할 시점이다. 물론 우리는 전자의 가능성을 더 높게 염두하고 있다.

그림 33. 엔비디아의 지난 5년간 차분기 실적 기준 EV/EBITDA 추이
5년 평균 약 35배, 현재는 약 24배의 EV/EBITDA... 코로나 19 저점 시기에도 비슷함



자료: Bloomberg, 미래에셋증권 리서치센터

2. GPU 그 너머...

(1) AGI를 향한 레이스: DeepSeek의 등장에도 가속화되는 AI 투자 경쟁

DeepSeek R1 쇼크 이후, 주식 시장의 반응은 투자자들의 우려를 보여준다. 'DeepSeek가 너무 효율적이어서 더 많은 컴퓨팅이 필요하지 않아 AI 인프라 경제가 근본적으로 바뀐다. 수백억 달러로 AI를 훈련하는 건 낭비다'라는 논리다. 그러나 이것은 지나치게 사안을 단순화한 내러티브다.

Anthropic CEO 다리오 아모데이는 "더 유능한 AI 모델의 경제적 이점이 너무 크기 때문에, 비용 절감액이 더 큰 모델을 구축하고 (아낀 돈을) 신속히 (더 큰 모델 개발에) 재투자합니다"라고 본인의 블로그 글에 힘을 주어 말했다. (DeepSeek가 제시한) 향상된 효율성은 전체 AI 투자를 줄이는 것이 아니라 AI 스케일링 노력을 계속 늘린다는 주장이다. 그의 말에 우리 팀이 동감하는 이유는, AI 모델 개발 경주는 절대 평가가 아닐 것이기 때문이다. 오히려 남들에게 반드시 이겨야 하는 상대 평가에 가깝다. 그러니까 **향상된 지능 개발에 있어서 목표 지점의 '상단'이라는 것은 있을 수 없다고 생각한다.** '거의 모든 것에서 거의 모든 인간보다 더 똑똑한 AI에 도달할 때', 즉 AGI를 만들 때까지 계속 이러한 경쟁 및 투자는 증가할 것으로 생각한다. 이와 관련해 AI guru들이 '앞으로 2~3년 안에, 늦어도 트럼프 임기 안에는 이런 시대가 열릴 것'이라면서 그들의 "AGI 타임라인" 전망치를 수정했다는 것을 우리는 이미 살펴본 바 있다.

이와 관련해서는, OpenAI가 새로운 자금 조달에 나설 것이라는 이야기가 좋은 예시가 될 것 같다. "400억 달러를 조달하고 최대 3,000억 달러의 기업 가치를 평가받는 협상을 진행 중"이라는 WSJ의 보도가 있었다. 이는 불과 몇 달 전 OpenAI의 가장 최신의 기업가치였던 1,570억 달러의 밸류에이션에서 거의 두 배가 되는 것을 의미한다. 이 펀딩 라운드를 주도하는 기관은 SoftBank인데, 150억~250억 달러를 투자할 것으로 논의 중이라고 전해졌다. 이러한 투자 규모는 SoftBank가 마이크로소프트를 능가하는 OpenAI 최대 투자자가 될 가능성을 시사하기도 한다. 손 마사요시는 며칠 전 도쿄에서 이런 말을 했다.

"1년 전만 해도 AGI는 10년 안으로 올 것이라고 생각했습니다.

그로부터 몇 달 후에는 2-3년 안에 올 것이라고 말했습니다.

하지만 지금은 그보다 더 빨리 올 것이라고 수정하고 싶습니다.

그리고 AGI는 일본에서 '처음으로 발표'될 것이라고도 말씀드리고 싶습니다"

그의 이 같은 주장은 '2년 미만의 근래에 AGI가 OpenAI 손에서 탄생할 것이다'는 예상을 염두하고 집행하는 투자라는 점을 암시한다. 뿐만 아니라, 구글은 OpenAI의 경쟁사인 Anthropic에 10억 달러를 투자할 채비인데, 이것은 이미 투자한 20억 달러에 추가로 투자하는 것이다. **OpenAI, 구글, Anthropic 등이 AGI 개발 경쟁을 하고 있는 와중에 베팅 금액을 늘리는 것은, AGI 쟁취에 있어서 리스크를 분산하고 이익을 극대화하려는 전략**이라 할 수 있다. **계속해서 더 많은 돈이 미국 내에 돌고 있다는 점은 매우 중요한 사실이다.**

표 5. 주요 AI 기업들의 AI 칩 보유 관련 현황

기업	주요 AI 모델	주력 AI 가속기	가속기 보유량 추정	비고
OpenAI	ChatGPT, Sora	H100, H200	10만	마이크로소프트와 협력해 애리조나에 10만 개 GPU 데이터센터 구축. H100, H200, B200, B100 등 고사양 GPU 도입 착수 중. Stargate 프로젝트로 100만개 이상의 클러스터 구축 계획. GPU 보유량이 상대적으로 낮은 것으로 보임.
Anthropic	Claude	H100	7.5만	아마존과 구글로부터 총 60억 달러 투자 유치. AI 안전을 중요하게 여겨, 新모델 출시 속도가 다소 느린 편.
Meta	Llama, 페이스북/인스타그램 내 추천모델	H100	12.8만	가장 공격적으로 GPU 확보 중인 빅테크. 올해 말까지 Llama 훈련용으로만 수십만개의 GPU 확보 목표.
xAI	Grok	H100, H200	20만	텍사스에 20만 개로 구성된 세계 최대 규모의 "단일 클러스터" 구축. 앞으로 최대 1GW 전력 규모의 데이터센터로 확장 계획.
Amazon	-	H100	10만 이상	자체 개발 칩인 Trainium 2를 40만 개 연결한 클러스터 구축 계획.
Google	Gemini, Veo, 유튜브 등 추천모델	TPU	10만 이상	2개의 지역에서 4개의 데이터센터를 연결해 10만 개 TPU 클러스터를 보유한 것으로 추정, 이는 가장 큰 규모의 다중 지역 클러스터
DeepSeek	V3 및 R1	H20, H800, H100, A100	6만 (추정)	2021년에 이미 1만 개 A100 보유.

자료: 각 회사, 미래에셋증권 리서치센터

DeepSeek 때문에 AI 생태계가 급변할 것은 사실이지만, 기존에 선두지위를 갖고 있던 기업들에 대한 시장의 베팅도 여전히 더 커지고 있다는 것을 알아야 한다. DeepSeek R1의 출현으로 발생한, AI 인프라 회의론도 일리가 있는 면이 있지만, 긍정론에 더 많은 귀를 기울여야 한다고 판단한다.

우리가 일전에 살펴본 것과 같이, **컴퓨팅 클러스터의 End-user라고 할 수 있는 AI 개발자들과 실제 업황을 체감하고 있는 반도체 회사들의 발언에 주목하면 DeepSeek의 출현을 더 긍정적으로 생각하고 있는 것 같기 때문이다.** 일례로, 반도체 장비 업계의 최강자라고 불리는 ASML의 CEO Christophe Fouquet는 DeepSeek의 영향에 대해 "ASML에게도 좋은 소식입니다. 비용이 낮아지면 AI가 더 많은 애플리케이션에 사용될 수 있고, 더 많은 애플리케이션은 더 많은 칩을 의미합니다."라고 말했다.

(2) 전시 체제에 버금가는 Stargate 프로젝트와 AI 패권 경쟁

OpenAI든 Anthropic이든 이들이 자금을 추가로 더 모집한 다음, 사용할 용처는 바로 컴퓨팅 파워를 확보하는 것이다. DeepSeek 때문에 이제 더 적은 칩과 전력만 필요해질 것이라는 주장을 전면 반박하는 셈이다. **아모데이는 AGI 시대를 맞이할 준비로, “2023년에 1억 달러, 2024년에는 10억 달러, 그리고 이제 100억 달러, 1,000억 달러가 필요하다”라는 스케일링에 관한 기존의 입장을 2주 전에 재확인했다.** 이 말은 AI 훈련/추론을 위한 데이터센터 구축 작업이 멈추지 않을 것임을 의미한다. 이는 빅테크들의 Capex 투자 금액은 당분간 계속 상향될 것이라는 말로 같음할 수 있다.

먼저, 이와 관련한 가장 급진적인 발표는 역시 Stargate 프로젝트일 것이다. 지난 1월 21일, 트럼프 대통령은 AI를 구동하기 위한 데이터센터 및 인프라에 5,000억 달러를 투자하겠다는 Stargate 프로젝트를 샘 알트만 CEO와 함께 공개했다. **Stargate는 2.2GW 규모의 전력을 사용하는 거대한 데이터센터에다가 100만 개 이상의 AI 칩을 탑재하는 프로젝트이고, 2028년 가동을 목표로 한다. 총비용은 1,000억 달러(140조 원 이상)에 달할 것으로 예상되며 이는 아폴로 우주 프로그램의 총비용과 맞먹는 규모다.** 참고로 소련이 만들어낸 스푸트니크 때문에 미국의 아폴로 프로젝트가 나왔다는 점을 떠올려보면 역사가 ‘어떠한 운율에 따라 흘러가는 것’처럼 보이기도 한다.

표 6. 프로젝트 “Stargate”의 예상 비용 하이라이트

단계	비용 구분	금액	비고
1단계	총 비용(TCO)	1,000억 달러	텍사스 애빌린 데이터센터 (2.2GW 전력)
	TCO 중 설비투자 비용 (Capex)	총 500억 달러	서버 및 인프라 투자 비용
	- Capex 중 서버 비용	약 500억 달러	NVIDIA GB200, GB300, VR200 칩
	- Capex 중 데이터센터	10억 달러 이상	주로 건설 비용
	TCO 중 운영비용 (Opex)	500억 달러	전력, 유지보수, 토지 등 임대 비용
최종단계	총 비용(TCO)	5,000억 달러	전체 프로젝트 규모 (근거 불명확)

자료: Dylan Patel, 미래에셋증권 리서치센터

그런데, 흥미로운 점은 DeepSeek가 엔비디아의 시가총액 수백조 원을 날렸던 바로 그 날에 트럼프가 DeepSeek 이야기를 꺼냈다는 점이다. 더군다나, 매우 미국 내부의 국내 행사라고 할 수 있는 공화당 대표 회의에서 감세 및 불법 이민자 추방에 관한 논의 바로 뒤에 DeepSeek를 이야기한 것은 의미하는 바가 크다. **트럼프 정부의 핵심 공약들과 "같은 의제 레벨"로 격상했기 때문이다.**

따라서 **Stargate 프로젝트가 단순히 과장된 '수사학'에서 그치는 게 아니라, 정말로 발생할 일임을 염두해 두고 한 말이라 추측해 볼 수 있다.** 물론 OpenAI가 그 프로젝트에 필요한 자금을 모두 조달할 수 있을지 불확실하고 자금 조달 계획이 터무니없다는 지적도 많다. 소프트뱅크가 본인이 가진 Arm 지분을 모두 팔아도 쉽지 않다고 할 정도다.

표 7. 프로젝트 “Stargate”와 관련한 회사들의 투자금액 분담 현황

관련 법인	구분	금액	현재 상태	비고
OpenAI	현재 보유 총액	500억 달러	보유 중	자본금 460억 달러 + 부채 40억 달러
	Stargate 투자 약정	190억 달러	아직 미집행	보유 자금 중 실제 투자 예정액
Softbank	투자 계획액	250억 달러	협상 중	ARM 지분 매각으로 자금 조달 가능성
오라클	1차 구간 설비투자	60억 달러	진행 중	데이터센터 및 하드웨어 투자 금액
	1차 구간 총소유비용	100억 달러	진행 중	설비투자에 운영비용까지 포함한 총 금액
	추가 투자	미정	검토 중	오라클 창업자인 Larry Ellison 추가 투자 가능성
MGX (UAE 펀드)	AI 투자에 관한 가용 자금	1.5조 달러(추정)	미정	펀드 금액 중 최대 10%의 금액이 Stargate 프로젝트에 투자될 수 있을 것이라고 추정. 그러나 구체적인 투자 약속은 아직 없음.

자료: Dylan Patel, 미래에셋증권 리서치센터

하지만 AI 경쟁을 "전시 체제"로 생각한다면 못할 것도 없다고 본다. **역사적으로 가장 긴박한 국가 안보 상황이었던 1차 세계대전 동안 미국은 GDP의 20% 이상을 차입했고, 2차 세계대전에는 GDP의 60% 이상을 차입한 전력이 있다.** 이것은 오늘날 기준으로 "17조 달러"로 치환되는 어마무시한 금액이다. 그럼에도, AI 인프라 투자가 달려 중국에게 AI 경쟁에서 밀릴 수도 있는 미래의 리스크 비용을 생각해보면, 오히려 '싸게 먹힌다'고 생각할 수도 있지 않을까?

전설적인 헤지펀드 투자자 스탠리 드러큰밀러는 미국의 새로운 대통령에 대해서 "내 생애, 가장 비즈니스에 비판적인 행정부에서 정권을 이양 받은 가장 비즈니스에 찬성하는 행정부"라고 칭했다. 드러큰밀러는 트럼프가 비즈니스 친화적인 정책을 펼칠 것임에 따라 미국에 수조 달러의 새로운 투자가 들어올 것임을 내포한 것이다. 물론 이 생각은 OpenAI를 포함한 미국 업체들이 더 나은 모델을 출시하고 더 많은 돈을 모을 수 있을 것이라는 믿음에 기초한다. 그리고 일전에 썼듯이, 우리 팀은 그러한 미래를 긍정한다.

표 8. 글로벌 AI 생태계에 관한 “지역 bloc별” 현황

요소	미국	기타 국가(EU, 한국, 일본 등)	중국
AI 연구	Market Maker (OpenAI, 구글, Anthropic 등)	각국 내 적은 규모의 Share Taker (소버린 AI 관점)	Fast Follower (DeepSeek, Alibaba, Moonshot 등)
반도체	엔비디아 AMD 및 인텔 반도체 장비업체	EU (IMEC 등 연구, 노광장비 ASML) 한국 (메모리, 파운드리) 대만 (TSMC R&D) 및 일본 (TSMC 공장)	자립화 몰두 (화웨이, Cambricon, SMIC)
데이터센터 인프라	세계 최대로 확장 중 송전/배전 그리드가 문제	각국별로 확장 중이나 규모는 제한적	세계 최대 규모로 구축 가능한 잠재력
정부 지원	규제 혁파, 리쇼어링 (Stargate 등) * 트럼프는 몇 가지 행정명령에 벌써 서명	상대적으로 소극적 일본이 그 중 가장 적극적	세계에서 가장 적극적이고 규제도 유연 * 2024년부터 반도체 산업에 약 1,600억 달러 투자 발표

자료: 미래에셋증권 리서치센터

(3) AI 효율화의 수혜자: 팔란티어로 바라보는 소프트웨어 기업의 성장 가속화

AI 업계에서 제본스의 역설이 발휘되게 된다면, 가장 수혜를 입을 수 있는 영역은 '저비용 고기능'의 기반 모델 위에 꽃 피울 여러 소프트웨어 업체들과 또한 CSP(클라우드 사업자)라고 할 수 있다. '더 저렴한 가격의 더 많은 인텔리전스는 더 많은 수요를 의미한다'라는 명제를 기억하면 될 것이다.

실제로 "DeepSeek 쇼크" 이후 AI 인프라 관련 기업들보다 소프트웨어 업체들의 주가가 강하게 아웃퍼폼하는 현상을 어렵지 않게 관찰할 수 있었다. 이는 AI 서비스 시장의 폭발적인 성장이 있을 수 있다는 시장의 기대감이 반영된 결과다. 작년 11월, 가트너는 AI 애플리케이션 시장이 2029년까지 연평균 140%(CAGR)를 기록하며 1,500억 달러 규모에 도달할 것으로 전망했었다. 그런데, DeepSeek 쇼크로 이러한 성장률이 더욱 커질 가능성이 어느 때보다 높아졌다고 할 수 있다.

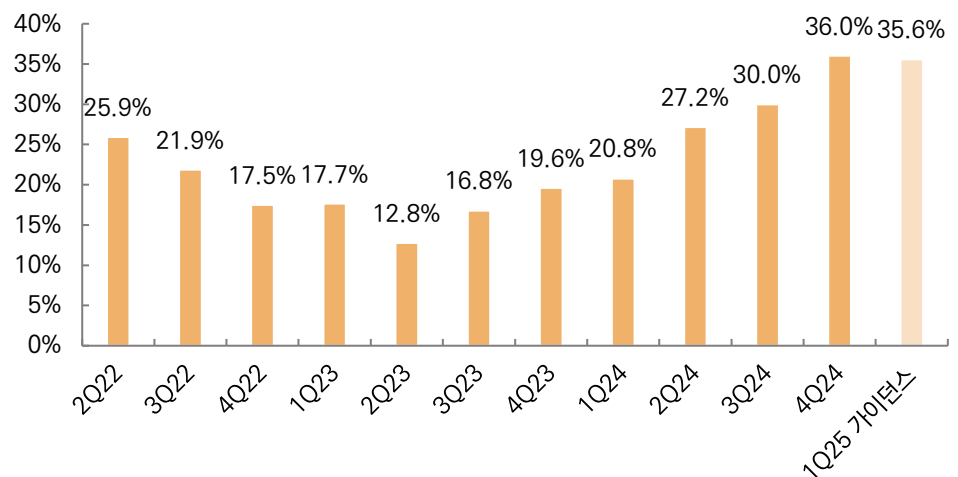
그러한 시장의 기대감을 반영한 가장 좋은 예시는 팔란티어(PLTR)였다고 생각한다. 팔란티어는 상장되어 있는 기업들 중 가장 순수한 AI 소프트웨어 기업으로 꼽히는 기업이다. 팔란티어는 지난 분기 실적발표에서 놀라운 성과를 선보였다. 그 중 가장 인상 깊었던 몇 가지를 꼽자면, 먼저 작년 4분기 미국 상업 부문 TCV(총 계약금액)가 전년 동기 대비 134%, 전분기 대비 170% 증가한 8억 300만 달러를 기록했다고 밝힌 점이다.

그리고 조정 영업이익률은 45%를 기록했고, 영업 현금흐름 마진은 56%, 조정 잉여현금흐름 마진은 63%를 기록했다. 특히나 팔란티어의 조정 영업이익률이 전 분기 대비 무려 7% 포인트나 증가한 점은 괄목할 만하다. 불과 2년 전에는 20%대였다는 점을 감안해야 한다. 이는 얼마나 빠르게 영업 레버리지를 달성할 수 있는지 알 수 있는 대목이다.

이러한 마진을 상승분만 아니라, 2024년 1분기 매출 가이드스는 8억 5,800만~8억 6,200만 달러라고 회사는 제시했다. 이것은 전년 동기 대비 35.6% 성장할 것으로 예상한다는 말과 같다. 이 가이드스 금액은 시장의 컨센서스 추정치 8억 달러를 상회하는 금액이었고, 따라서, 투자자들은 크게 반색했다. 게다가, 팔란티어가 대개 제공하는 가이드스 금액은 지난 몇 분기 동안 상당히 보수적이었다는 점을 감안하면 실제 매출 성장률은 거의 40%에 가까울 수 있다는 희망을 가지게 만들기도 한다.

그림 34. 팔란티어의 분기별 매출 성장률

팔란티어의 4Q24 가이드스는 26.4%였으나 실제로는 36%를 기록

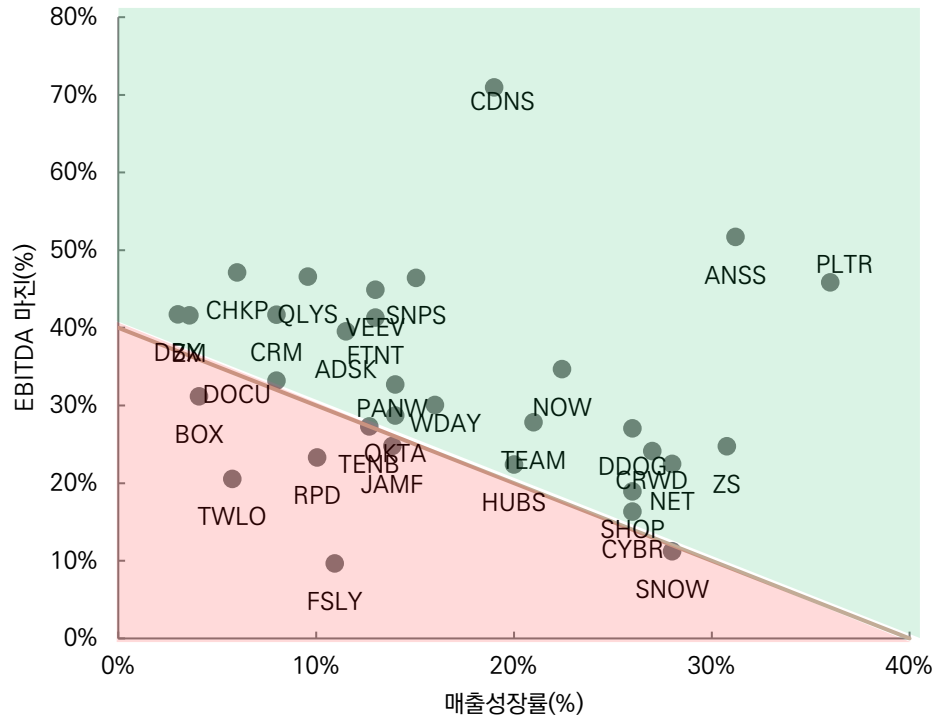


자료: Palantir, 미래에셋증권 리서치센터

성장성과 수익성을 다 고려하기 위해서, 우리 팀은 소프트웨어 서비스 기업들을 평가할 때 유용한 지표 중 하나인 **Rule of 40**를 주목했다. **‘40점만 넘겨도 우수함’을 나타내는 지표에서, 팔란티어는 현재 기준 무려 82점을 기록한 상태다.**

** Rule of 40는 매출성장률과 EBITDA 마진율을 더해서 40이 넘으면 좋은 소프트웨어 기업임을 나타내는 지표. 이것이 존재하는 이유는 PER 밸류에이션이 모든 상황에 맞는 지표는 아니기 때문. 예를 들어, 2023년에 엔비디아의 P/E 비율은 몇 달 만에 50에서 약 250으로 급증한 적도 있었음. 그 후 주가는 3배가 됐지만 PER은 급락.*

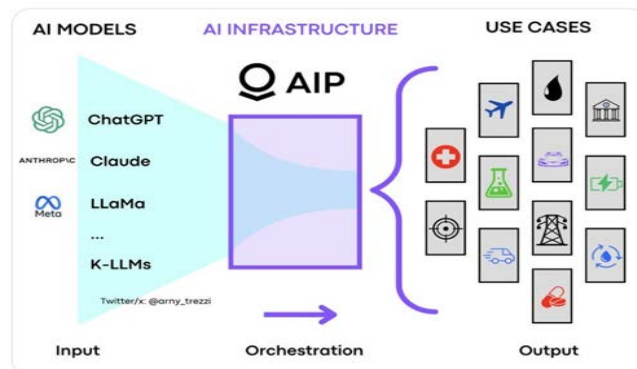
그림 35. 주요 소프트웨어 기업들의 “Rule of 40” 비교
이 지표 기준 TOP 3는 케이던스(CDNS), 앤시스(ANSS), 그리고 팔란티어(PLTR)



자료: Bloomberg, 미래에셋증권 리서치센터

이처럼, 실적 발표 이후 팔란티어가 만들어낸 실적은 곧 AI 생태계 내 좋은 해자를 갖춘 소프트웨어 기업들의 경우 성장성과 수익성의 가속이 얼마나 대단한지를 알려준다. **현재는 생성 AI 혁명으로 인한 HyperChange 시대라고 판단하기에, 수익이 얼마나 빠르게 성장할 수 있는지 과소평가하면 안 된다. '모바일 혁명이 막 시작될 때 구글에 투자하는 것이 오늘날 팔란티어에 투자하는 것'이라는 주장도 그래서 힘을 입고 있는 게 아닌가 판단한다.** 또한, 무엇보다도 팔란티어와 같은 AI 소프트웨어 업체들은 DeepSeek R1와 같은 ‘저비용, 고기능’ 모델이 주류로 올라오게 되면서 혜택을 볼 가능성이 커졌다고 할 수 있다. 팔란티어 역시 핵심 제품들인 Gotham과 Apollo에서 그러한 효율적 AI 모델 등장의 혜택을 볼 수 있는 구조를 지녔다.

그림 36. ‘저비용 고기능’ AI 모델의 등장이 팔란티어에게 이로운 이유?
AI 모델이 싸지고 좋아지면, 현장의 비즈니스에 AI가 줄 수 있는 실제적 도움이 더 커질수밖에



자료: X(@amy_trezi), 미래에셋증권 리서치센터

물론 절대로 팔란티어는 “중국의 R1”을 사용하지는 않을 것으로 사료된다. 그럼에도, DeepSeek가 보여준 여러 엔지니어링 혁신에 관해서는 그들도 관심을 가지리라 추측한다. 지난 4분기 실적 발표에서 “DeepSeek의 최적화 기법은 상당히 인상적이고, 기술적으로 매우 정교한 수준에 도달했다”고 평가했기 때문이다. 따라서 DeepSeek가 공개한 수학적 알고리즘만 차용하여 만들어진, 미국의 다른 챗봇을 탑재해 고효율 AI 모델의 이점을 누릴 것으로 판단된다. 이는 팔란티어는 물론 그 솔루션을 사용하는 고객사들까지 맛볼 변화다.

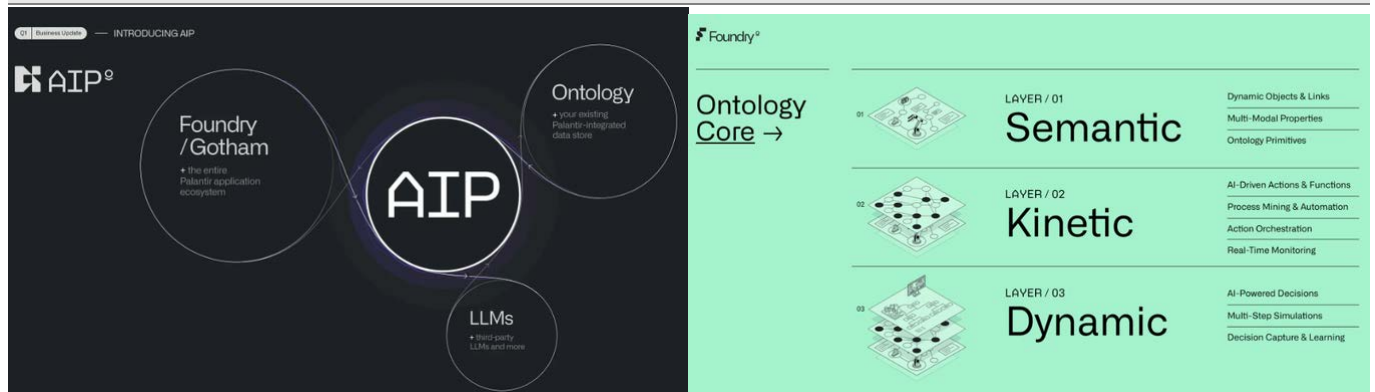
표 9. 팔란티어를 이해하기 위해 꼭 알아야 할 개념들

제품 및 영역	직관적 설명	주요 기능	각각의 의의
Ontology = 핵심 구조	데이터가 무엇인지, 어떻게 연결되는지 설명해주는 근본 아키텍처	* 데이터 관계 정의 (객체, 속성, 관계) * 데이터 의미론적 연결 (Semantic Layer) * 지식 그래프	* Foundry/Gotham 내에서 데이터의 의미와 관계를 정의하고, 이를 기반으로 데이터 검색, 분석, 시각화 등을 지원 * 실시간 데이터 활용 가치 극대화
Foundry/Gotham = 핵심 제품	회사의 모든 데이터를 모아두고, 정리하고, 분석하는 곳	* 데이터 소스 통합 (이기종 데이터 통합) * 조직의 동적인 데이터 관리 및 분석 * 데이터 시각화 및 대시보드 * 협업 및 워크플로우 지원 * 데이터 보안 및 접근 제어	* 백엔드 시스템(데이터 소스, 분석 엔진) * 다양한 데이터 소스를 통합, 정제, 분석하여 AI 모델 및 애플리케이션(챗봇 포함)에 필요한 데이터 제공 * 데이터 기반 의사 결정 지원
AIP (AI Platform) = AI 활용 공간	Ontology를 기반으로 Foundry/Gotham의 데이터를 활용하여 챗봇과 같은 AI 앱을 만들고 관리하는 “런치패드”	* LLM 기반 AI 애플리케이션 개발 플랫폼 * AI 모델 (LLM 포함) 배포 및 관리	* AI 모델을 활용한 애플리케이션(챗봇 포함) 개발 및 운영 환경 제공 * AIP Logic, AIP Assist 등을 통해 비개발자도 AI 활용 가능
LLMs (대형언어모델들) = 전체적인 지능	AIP 안에서 챗봇의 “두뇌” 역할을 하며, 사람의 말을 이해하고 답변을 생성	* AI 모델과 비즈니스 프로세스 연결 * 자연어 처리 (텍스트 생성, 질의응답 등)	* 사용자 질문에 답변하고, 자연어 인터페이스 제공하는 프론트엔드 * AIP 플랫폼 내에서 챗봇을 포함한 다양한 AI 애플리케이션의 핵심 엔진 역할

자료: 미래에셋증권 리서치센터

그림 37. 팔란티어 비즈니스의 핵심 구조와 온톨로지

Semantic(데이터 통합 및 의미 연결) + Kinetc(AI 활용한 프로세스 최적화 및 작업실행) + Dynamic(AI 기반 예측 및 최적의 의사결정)



자료: Palantir, 미래에셋증권 리서치센터

또한, 팔란티어는 기업이 창립될 당시부터 ‘미국에 의한, 미국을 위한 기업’이었다는 점을 잊어서는 안 된다. 애초에 그들의 주요 고객사가 미국의 방첩 및 수사 기관들이었다. 그리고 그 고객사가 미국 국방부, 국무부 등으로 확대되었다. 애초에 그들은 중국이나 러시아에 그들의 솔루션을 판매할 계획 자체가 아예 없었다. 이는 G2간 기술 경쟁에서 발생할 리스크에 이미 자유로운 업체임을 의미하기도 한다. 그리고 러-우 전쟁과 이스라엘과 팔레스타인의 분쟁에 있어서, 팔란티어는 공식 실적 발표 자료에서조차 적대국에 대항해 미국의 동맹 진영을 도와야 한다는 식의 문구를 명확히 기재해 놓은 업체다. 따라서 **중국 등의 AI 발전이 위협적일수록, 이에 대항하는 AI 솔루션 업체로 더욱 주목받을 수 있다.**

따라서, DeepSeek와 더불어 지정학적인 AI 경쟁이 펼쳐지게 되면 사실 더욱 주목을 받을 수 있는 업체이기도 하다. "국방의 현대화, 미 제조업의 부활"의 주요 의제에 있어서 트럼프 정부의 주문을 받게 될 핵심 AI 기업으로 꼽히고 있기 때문이다. (이에 대해서는 지난 11월 21일, AI Weekly “낙후된 제조업 경쟁력을 시로 살린다”에서 다룬 바 있음)

그림 38. 팔란티어 Warp Speed = 테스트, 생산, 운영을 통합하여 제조업체의 생산 최적화 지원
팔란티어 “미국의 제조업을 위한 것… 중국과 경쟁하는 과정에서 재산업화 시작 → 필수 기술”



자료: Palantir, 미래에셋증권 리서치센터

이제껏 살펴봤듯 팔란티어가 대표적이지만, 또 다른 SaaS 기업들로부터의 소식도 긍정적이다. 세일즈포스(CRM)의 CEO 마크 베니오프는 "(본인들의 AI 에이전트 서비스인) Agentforce 관련해 수천 건의 거래가 4분기에 성사될 것이며, 회사 역사상 이보다 더 빠르게 진행되는 것을 본 적이 없습니다."라고 말했다. 이는 생성 AI 시대가 점차 챗봇을 넘어서 소프트웨어 기업들이 제시하는 여러 서비스들로 빠르게 전이/확대가 되고 있다는 것을 의미한다. 그리고 이런 현상은 DeepSeek로 말미암아 더 저렴해진 지능이 탑재될 경우 더 빠른 속도로 시장 침투가 발생할 수 있음을 암시한다. 제본스의 역설은 이제 시작이다.

(4) 하이퍼스케일러들의 2025년 자본지출 청사진

앞으로 점점 더 저비용 AI가 주류가 되면 CSP로 대변되는 빅테크들에는 오히려 긍정적인 흐름이 펼쳐질 수 있다고 본다. DeepSeek 같이 높은 추론 성능의 모델이 거의 무료로 제공되면, AI 사용자들(소규모 회사, 대학, 연구기관 등)은 그만큼 API 비용에서 절약할 수 있고 그 돈을 컴퓨팅 자원에 더 투자할 수 있게 된다.

이들은 CSP들의 데이터센터를 임대하는 고객층이 될 것이고, 결국 이러한 투자 풍토는 클라우드 서비스 제공업체에게 이익이 될 것이라는 말이다. 다시 말해 더 많은 비용이 폐쇄적인 모델 제공업체(예: OpenAI)보다는 컴퓨팅 자원(예: 클라우드 서비스)에 투입될 것이고, 이는 결국 하드웨어에 더 많은 비용이 지출될 것으로 사료된다.

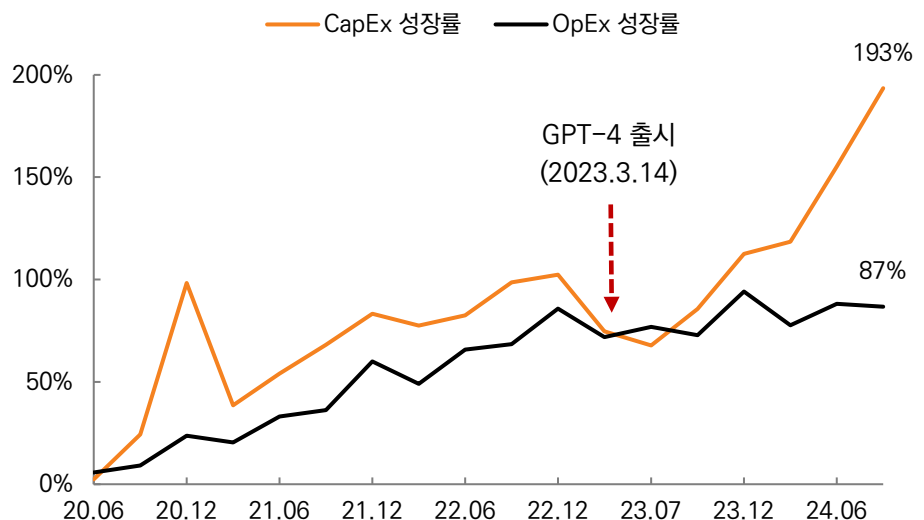
또한, 자금력과 인프라를 갖춘 Big Tech들은 저렴한 AI 모델 운영 비용을 활용하여 다양한 서비스에 적용하여 시장 지배력을 강화할 수 있다. API 형태로 제공되는 AI 모델은 기존 플랫폼과의 결합을 통해 더 큰 시너지 효과를 낼 수 있기 때문이다.

이와 관련해 DeepSeek와 정면 대결해야 하는 오픈소스의 선구자 메타는 본인들의 광고 추천 도구에 DeepSeek의 모델을 테스트하는 것을 고려하고 있다. 메타가 이렇게 판단한 이유는, 오픈소스 AI 개발에 관한 자존심을 부리기 보다는, 저비용 AI를 적극 채용해 더 많은 매출을 이끌어내어 시장성장의 수혜를 입을 수 있다는 속셈이 깔려있기 때문일 것이다.

한편, 마이크로소프트는 2024 회계연도 3분기에 Azure가 32% 성장할 것으로 예상하고 있는데 시장의 기대를 완전히 충족시켜주지는 못했다. 얼핏 보면 DeepSeek 쇼크처럼 "과대 투자의 논리"에 영향을 받았다고 지레 짐작하게 된다. 그러나, 오히려 수요를 충족하기에 충분한 데이터센터 용량이 없기 때문이라고 마이크로소프트는 그 이유를 설명했다. 실제로, 마이크로소프트의 Azure AI 서비스는 157% 증가했으며, 상업 예약은 67% 증가하여 경영진의 예상치를 훨씬 상회했다고 했다. 다시 말해, AI 수요를 충족하기 위해 더 많은 칩 확보에 나설 것임을 의미한다.

물론 '땅 파서 장사'하는 것은 아니기 때문에 투자자로서 마진율을 그리 걱정할 사안은 아니라고 생각된다. 마이크로소프트는 소비자들이 AI 기능에 더 많은 비용을 지불할 의향이 있다고 판단하여 "Office 365" 서비스에 대한 가격을 최대 43% 인상했기 때문이다. 이와 관련해, 다리오 아모데이는 "점점 더 지능적인 모델을 훈련하는 것의 경제적 가치가 너무 커서, 모든 비용 이득은 거의 즉시 상쇄된다"고 말했는데 그의 말에 힘이 실리는 사안이다.

그림 39. 1Q20을 기준으로, Magnificent 7의 CapEx 및 OpEx 누적 성장률 비교
“코로나19” 시기 비슷한 상승 궤적을 그렸지만, 생성 AI 혁명 이후 극명하게 나뉨



자료: Bloomberg, 미래에셋증권 리서치센터

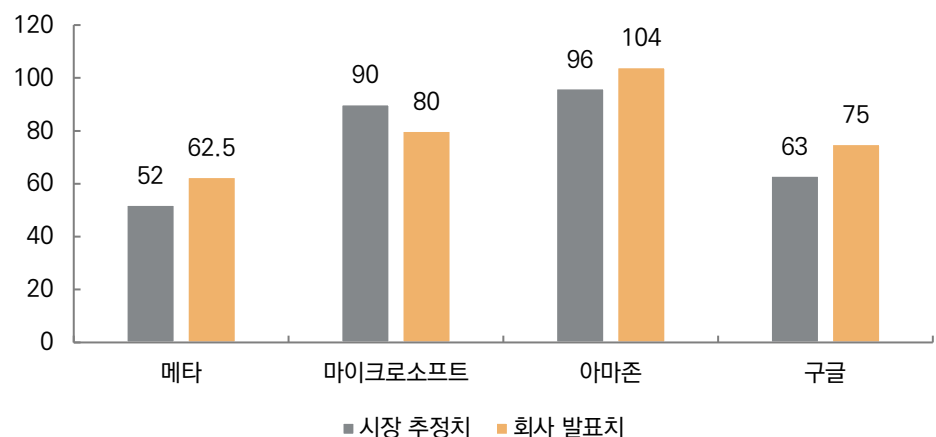
하이퍼스케일러들의 투자 금액은 놀라운 수준으로 계속 격상되고 있다. 마이크로소프트와 메타의 금액만 합산하더라도 이미 거의 1,500억 달러를 2025년 자본지출로 계획하고 있는 수준이다. 이와 관련해 마크 주커버그 CEO는 “2025년에, 10억 명 이상의 사용자에게 메타 AI를 제공하기 위해 장기적으로 AI 인프라에 수백억 달러를 지출할 것”이라고 밝혔다.

또한 주커버그는 “2025년에 약 1GW의 컴퓨팅 성능을 온라인에 가져올 것이며 연말까지 130만 개 이상의 GPU를 보유하게 될 것”이라고 말하기도 했다. 작년 초에 엔비디아 GPU 약 60만개를 샀을 것으로 추정되기 때문에 130만 개 이상의 GPU라고 하는 것은 작년 구매량보다 최소 10만개 이상의 GPU를 더 확보하겠다는 말로 이해할 수 있다.

구글도 메타와 마이크로소프트와 함께 2025년 자본지출에 대해 공격적인 전망치를 제시했다. 구글은 2025년에 750억 달러의 자본 지출을 예상했다. 이 수치는 2024년도의 526억 달러는 물론이고, 월가에서 예측한 630억 달러의 금액마저 쉽게 웃도는 놀라운 수치다. 시장에서는 18% 성장을 기대했는데, 43%나 높여 잡았다는 말이다.

또 다른 빅테크인 아마존은 2025년에 자본지출에 무려 1,040억 달러를 지출할 것으로 예상된다. 이는 구글과 비슷하게 전년 대비 약 40% 증가한 수치다. **이처럼 DeepSeek 쇼크 이후의 빅테크들의 실적발표에서, 그들로부터 언급된 Capex 증가 속도는 비슷하게 가속되고 있다**는 점을 알 수 있다.

그림 40. 2025년 기준 하이퍼스케일러들의 CapEx 금액 비교 (단위: 십억 달러)



자료: 각 회사, 미래에셋증권 리서치센터

이처럼, 앞으로 하이퍼스케일러들은 회계상 AI 서비스를 통해 당장 수익을 만들고 있지만, 그 벌어들인 돈보다도 더 많은 돈을 GPU를 구매하는 데 훨씬 더 많은 지출을 할 가능성이 있다고 본다. **이 모든 것은 ‘상대방보다 더 나은 지능’을 확보하기 위함이다.** 마지막으로 구글의 CEO 순다르 피차이가 작년에 말한 것을 다시 한번 상기해볼 필요를 느낀다.

“과소투자의 위험이 과잉투자의 위험보다 훨씬 큼니다.”

III. Paper of the Week

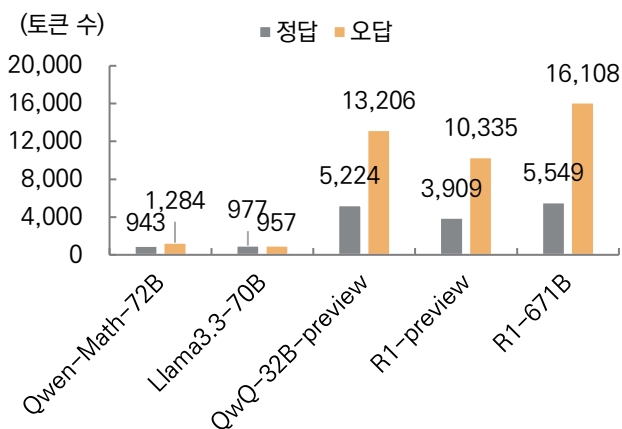
추론 모델의 얇은(?) 생각 – Tencent AI

o1과 같은 추론 모델들은 인간이 문제를 해결하는 방식을 모방한다. 마치 사람처럼 여러 가설이나 계획을 세우고 그 순서에 맞게 논리를 전개하며, 오류가 의심되는 부분은 재확인하고 수정하면서 이전까지는 해결하지 못했던 문제를 풀어낼 수 있게 되었다. 그러나 추론 과정에서 ‘깊이 생각하고 대답’하는 데에는 필연적으로 많은 시간과 비용이 드므로, 모델이 효율적으로 답변을 제공할 수 있게 하는 방법에 대한 논의도 병행되어 왔다.

그러나 모델의 답변 효율성에 있어 모델의 ‘overthinking’, 즉 문제 해결에 영향을 주지 않는 부분에 과하게 토큰을 소모하는 경향에 대해서는 연구된 바 있으나, ‘underthinking’에 대한 논의는 상대적으로 충분히 이뤄지지 않았다고 논문은 주장한다. Tencent AI Lab에서 공개한 이 논문은 o1과 유사한(o1-like) 모델들이 마구잡이로 생각한다며, 모델의 underthinking(이하 얇은 생각), 즉 충분히 깊이 생각하지 않고 성급하게 다른 생각으로 넘어가는 경향에 대해 조명했다.

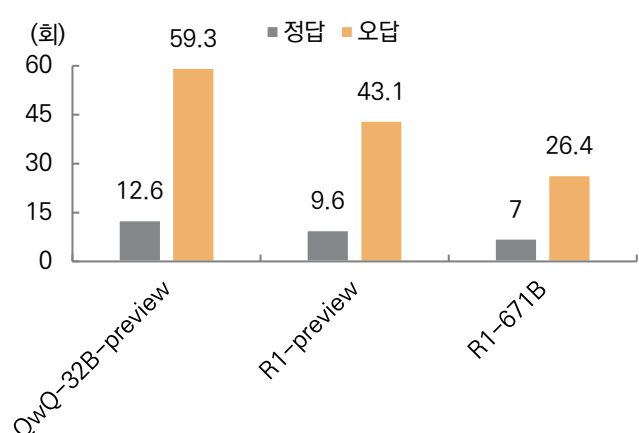
o1과 유사한 오픈소스 추론 모델인 Deepseek의 r1-preview와 r1-670B, 그리고 알리바바의 QwQ-32B-preview 모델의 답변 토큰 및 ‘생각’의 수를 비교한 결과, 모든 o1-like 모델들은 정답을 맞히지 못했을 때 더 많은 토큰을 사용했고, 더 자주 생각을 전환했다. 또한 이 모델들은 답이 틀렸을 경우에 답을 맞혔을 때보다 평균적으로 225% 더 많은 토큰을 사용했으며, 이는 평균적으로 418% 더 자주 생각을 스위칭했기 때문으로 연구진은 판단했다. 이렇듯 생각을 지나치게 빠르게 전환하는 행위를 얇은 생각이라고 정의하고, 더 많은 자원을 소모하고도 정답을 찾지 못하게 되는 현상에 집중했다.

그림 41. 모델이 정답을 맞혔을 때와 맞히지 못했을 때의 토큰 수



자료: Tencent AI, 미래에셋증권 리서치센터

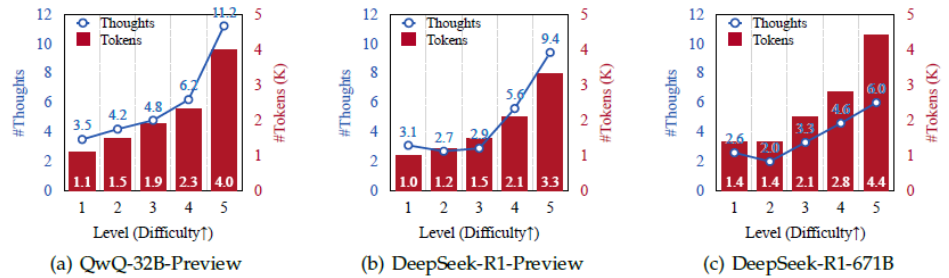
그림 42. 모델이 정답을 맞혔을 때와 맞히지 못했을 때의 생각 수



자료: Tencent AI, 미래에셋증권 리서치센터

얇은 생각은 문제가 어려울수록 더 자주 발생하는 경향이 있으며, 이 때 모델은 더 많은 수의 아이디어를 떠올리게 된다. 이는 o1과 유사한 모델들이 어려운 문제를 해결하기 위해 더 많은 노력을 기울이고 있음을 보여준다. 아래 도표에서도 문제 난이도가 1에서 5로 상승함에 따라 생성되는 토큰의 수와 ‘생각’의 수도 증가하는 것을 볼 수 있다.

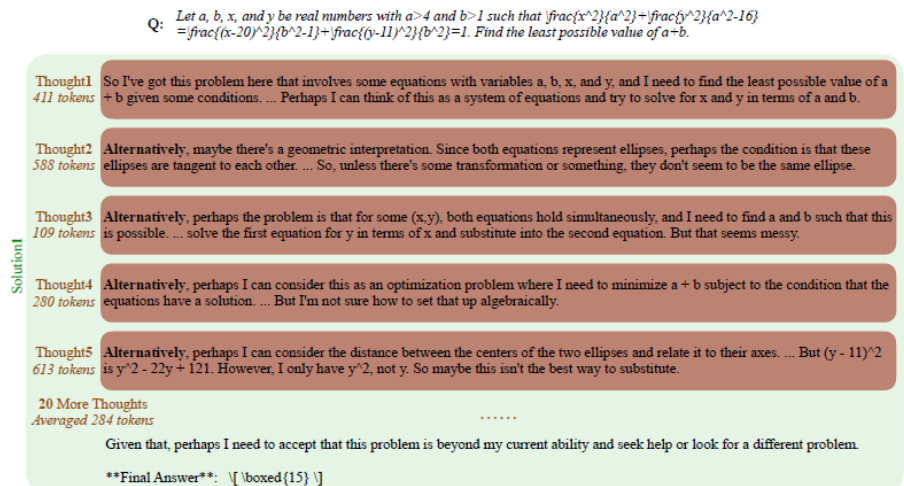
그림 43. 문제 난이도에 따른 추론 모델들의 토큰 및 생각 수가 난이도와 비례하는 모습



자료: Tencent AI, 미래에셋증권 리서치센터

그럼에도 불구하고 모델들은 각 아이디어를 충분히 발전시키지 못하고 다음 아이디어로 넘어가며, 각 아이디어에 대해 충분히 깊이 생각하지 않으므로 정답에 도달하지 못하는 경우가 많다. 아래는 QwQ-32B-preview 모델이 생성한 답변 예시인데, 무려 25개의 ‘생각’을 떠올렸음에도 불구하고 문제의 정답을 맞지 못했다.

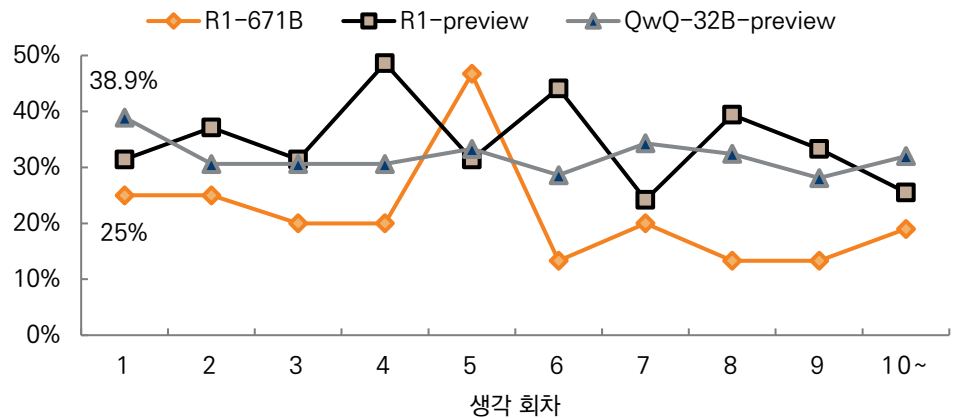
그림 44. QwQ-32B-preview 모델의 답변 예시. 25개의 생각을 내놓고도 정답을 맞지 못함



자료: Tencent AI, 미래에셋증권 리서치센터

그런데 모델이 정답을 맞지 못한 것이 얇은 생각 때문이라고 단언할 수는 없다. 모델 성능 부족이 원인일 수도 있다. 이를 확인하기 위해 모델이 정답을 맞지 못했을 때 지난 ‘생각’들이 정답으로 이어질 수 있었는지를 평가했다. 그 결과, 버려진 생각들 중 상당수가 실제로 정답으로 이어질 수 있었음을 확인했다. 특히 **첫 생각이 정답이었을 확률이 적어도 25%에 육박했다.** 이는 네 번 중 한 번은 생각 한 번으로 문제를 해결할 수 있었음에도 불구하고 최소 2회 이상의 생각을 거쳤고, 더 많은 토큰을 소모했음에도 오답을 내놓았음을 뜻한다. 위의 답변 예시에서도 생각 1은 정답이 될 수 있는 ‘정확한 생각’이었으나, 빠르게 다음 생각으로 넘어가면서 약 7,270개의 추가 토큰을 소모하고 정답에도 이르지 못했다.

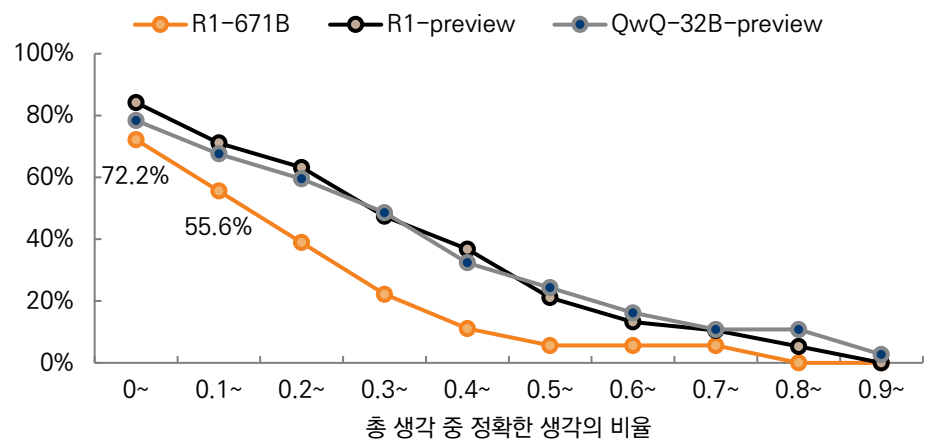
그림 45. 모델이 틀린 문제에 대해 과거 생각이 정답에 도달할 수 있었던 확률



자료: Tencent AI, 미래에셋증권 리서치센터

사실 오답에 대해 모델이 더 많은 생각을 떠올리는 만큼 정답에 이를 확률 자체도 높다. **모델의 오답 안에 ‘정확한 생각’이 존재할 확률은 최소 72.2%이며, 틀린 답변 중 ‘정확한 생각’이 10% 이상 존재했을 확률은 50% 이상**(R1-671B 모델이 55.6%)이다. 따라서 모델의 답변 정확도를 향상시키기 위해서는 모델이 정확한 생각을 잘 유지하고 확장해 나갈 수 있도록 적절한 보상을 제공해야 한다. 이러한 보상 메커니즘으로 논문에서는 ‘생각 전환 페널티(Thought Switching Penalty, TIP)’를 부여했다.

그림 46. 모델이 틀린 문제에 대해 ‘정확한 생각’이 존재할 확률



자료: Tencent AI, 미래에셋증권 리서치센터

생각 전환 페널티는 모델이 오답을 제공한 경우 소모된 총 토큰 개수 중, 처음으로 ‘정확한 생각’에 도달하기까지의 토큰을 제한 비율이다. 만약 모델이 정확한 생각에 한 번도 도달하지 못한 경우 생각 전환 페널티는 0으로, 모델이 문제를 이해하지 못한 것으로 판단해 얕은 생각의 범주에 포함하지 않는다. 반면 **모델이 정확한 생각에 빨리 도달했으나 불필요한 생각 전환으로 정답에 이르지 못한 경우, 생각 전환 페널티 값은 커진다.** 이 페널티는 모델이 생각을 전환할 때 나타나는 토큰(위 예시의 경우 "Alternatively")에 가해지며, 페널티 강도(α)와 지속시간(β)으로 영향력을 조절한다. 논문에서는 가장 좋은 성과를 보였던 $\alpha = 3, \beta = 600$ 을 기준으로 모델의 성능 개선 여부를 검증했다.

먼저 대학원 수준의 과학 문제나 경시대회 수학 문제로 구성된 벤치마크(MATH500-Hard, GPQA Diamond, AIME24)를 기준으로 모델의 성능을 평가했을 때, 모델의 정확도와 얇은 생각의 상관관계는 서로 다른 양상을 보였다. MATH500-Hard와 GPQA Diamond 데이터 세트에서는 R1-671B 모델의 답변 정확도와 얇은 생각 점수가 모두 높게 나타났다. 여기서의 얇은 생각 점수는 모델이 정답에 도달하지 못하고 헛되이 사용한 토큰의 비율로, 오답 생성 시 얼마나 많은 토큰을 비효율적으로 사용했는지 나타내는 척도다. 이는 **R1-671B 모델이 정확도는 높았지만 추론 과정은 비효율적이었음을 나타낸다.** 반면 AIME24 벤치마크에서는 높은 정확도와 낮은 얇은 생각 점수라는 이상적인 결과를 보였는데, R1-preview의 결과와는 극명한 차이를 보이는 것으로 보아 이후에 AIME 벤치마크 문제로 추가 학습 되었을 가능성이 있다고 판단한다. 이를 통해 **모델의 얇은 생각은 학습 데이터 구성에 크게 영향을 받는다는 사실을 알 수 있다. 또한 모델 성능(벤치마크)이 개선되더라도 이것이 모든 작업 성능의 개선으로 이어지지는 않을 수 있다**는 점에도 주목할 필요가 있다.

그림 47. 벤치마크에 따라 달라지는 모델의 정확도 및 추론 효율성

Models	Accuracy	UT Score
<i>MATH500-Hard (Level 5)</i>		
QwQ-32B-Preview	84.3	58.2
DeepSeek-R1-Preview	83.6	61.5
DeepSeek-R1-671B	92.5	65.4
<i>GPQA Diamond</i>		
QwQ-32B-Preview	59.6	48.3
DeepSeek-R1-671B	73.2	58.8
<i>AIME24</i>		
QwQ-32B-Preview	46.7	65.0
DeepSeek-R1-Preview	46.7	75.7
DeepSeek-R1-671B	73.3	37.0

자료: 미래에셋증권 리서치센터

그림 48. 생각 전환 페널티 부여 후 모델의 성능 및 효율성 개선

Models	Pass@1	
	Accuracy(↑)	UT Score (↓)
<i>MATH500-Hard (Level 5)</i>		
QwQ-32B-Preview	82.8	71.1
+ TIP	84.3	69.7
<i>GPQA Diamond</i>		
QwQ-32B-Preview	57.1	59.1
+ TIP	59.3	56.5
<i>AIME2024</i>		
QwQ-32B-Preview	41.7	72.4
+ TIP	45.8	68.2

자료: 미래에셋증권 리서치센터

마지막으로 QwQ-32B-preview 바닐라 모델과 모델에서 제안한 생각 전환 페널티를 부여한 모델의 점수를 비교한 결과, **모든 벤치마크에서 바닐라 모델에 비해 정확도는 올라가고 얇은 생각 점수는 줄어든 것을 확인할 수 있다.** 다만 여기서의 성능 개선은 근본적인 해결책이라고 보기는 어려워 보인다. 측정 방식의 차이가 있을 수 있겠으나(논문에서는 명시되어 있지 않음), 이전 표의 값과 페널티를 부여한 모델의 값을 비교해 보면 오차 범위 안의 개선이라고 해석할 여지가 있기 때문이다.

그럼에도 불구하고 **모델을 추가로 학습시킨 것이 아니라, 디코더에 연산을 추가한 것만으로 성능 개선을 이뤄냈다는 점은 고무적이다.** 또한 좋은 생각에 충분히 ‘어텐션’하게 만드는 것만으로도 모델의 성능을 개선하고 추론 과정에서의 토큰 효율화를 달성할 수 있다는 점에서 성능 개선 여력은 아직 충분해 보인다.

IV. AI Peer Table

표 10. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	1W	1M	기간별 수익률(%)				실적발표 예정일	AI 밸류체인 관련 한줄평
최종 소비자(온디바이스, 앱)											
애플	AAPL US	227.63	4,978	-0.1	-3.8	0.4	5.6	21.1	-9.0	05-02	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	361.62	1,693	-5.7	-8.4	12.6	80.8	86.8	-10.5	04-23	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
삼성전자	005930 KS	53,900	322	5.7	-2.5	-4.8	-26.9	-25.7	1.3	04-04	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
퀄컴	QCOM US	167.96	270	-1.3	7.0	-1.2	3.2	13.4	9.3	05-01	생성 AI 처리 강화용 온디바이스 CPU 설계(X Elite)
스냅	SNAP US	10.92	27	-2.2	-12.4	-9.1	20.7	-1.7	1.4	04-25	이미지 생성 모델을 탑재한 증강현실 서비스
크래프톤	259960 KS	383,000	18	6.4	14.3	17.8	30.1	79.0	22.6	02-11	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
비즈니스 효율화											
어도비	ADBE US	433.07	274	-1.3	6.7	-12.5	-19.2	-31.0	-2.6	03-14	세계 최고의 미디어편집 툴. 서비스 배포능력(Firefly)
세일스포스	CRM US	325.83	454	-4.0	2.5	1.3	29.4	12.5	-2.5	02-26	세계 최고의 CRM 업체. AI로 사용성 강화(Einstein)
서비스나우	NOW US	1,008.29	302	-1.2	-1.6	0.0	24.3	24.0	-4.9	04-24	워크플로우 자동화. AI로 사용성 강화(Now Assist)
클라우드스트라이크	CRWD US	421.59	151	6.0	20.9	27.7	75.0	28.0	23.2	03-05	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	252.34	340	-2.6	15.6	19.8	33.7	40.0	15.5	04-24	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센처	ACN US	385.98	352	0.2	10.8	9.0	23.2	5.6	10.2	03-20	각 산업별 맞춤형 AI 에이전트를 만들기 위한 컴병
SAP	SAP GY	271.00	500	3.0	10.2	22.4	42.7	63.0	14.7	04-22	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,888.75	130	1.4	-4.0	3.2	7.9	16.7	0.5	04-17	각 산업별 맞춤형 AI 에이전트를 만들기 위한 컴병
타타 컨설턴시	TCS IN	4,026.90	242	-1.0	-3.9	-1.1	-2.8	0.4	0.2	04-11	각 산업별 맞춤형 AI 에이전트를 만들기 위한 컴병
모더나	MRNA US	32.60	18	-10.8	-22.8	-30.4	-61.6	-62.7	-21.6	02-14	신약 개발 등 모든 업무에 생성 AI 활용(Dose ID)
일루미나	ILMN US	111.06	26	-15.3	-18.8	-26.5	-10.1	-17.1	-16.9	05-02	AI 기반 침단 유전체 분석
루닛	328130 KS	68,500	2	6.5	3.3	72.1	103.9	21.0	7.4	03-21	AI 기반 암 검출/진단 보조 솔루션(인사이트)
AI Ops											
몽고DB	MDB US	277.87	30	3.4	13.3	-4.2	18.3	-44.5	19.4	03-07	AI 모델 구축시 핵심 인프라인 '벡터 DB' 기능(Atlas)
클라우드플레어	NET US	166.66	84	21.8	48.0	82.6	114.5	54.4	54.8	05-02	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	110.85	367	32.4	64.8	89.8	269.4	354.7	46.6	05-06	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
포스 패러다임	6682 HK	57.40	5	-0.9	39.2	41.2	30.6	33.5	12.7	03-20	중국의 팔란티어(Sage)

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

표 11. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	1W	1M	기간별 수익률(%)				실적발표 예정일	AI 밸류체인 관련 한줄평
파운데이션 모델											
알파벳	GOOGL US	185.34	3,304	-7.9	-3.5	4.0	13.5	24.8	-2.1	04-25	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼	META US	714.52	2,635	2.4	16.0	21.3	38.2	53.2	22.0	04-24	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	104.80	372	11.5	31.7	11.5	34.6	54.5	27.2	02-20	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
바이두	9888 HK	89.35	47	5.8	15.0	0.4	7.4	-11.7	8.0	02-18	NLP 및 자율주행 모델의 전통강호(Ernie Bot, Apollo)
센스타임	20 HK	1.77	12	4.7	38.3	3.5	59.5	115.9	18.8	03-26	중국의 멀티모달 AI의 다크호스(SenseNova)
네이버	035420 KS	229,000	36	5.5	13.4	31.2	39.9	11.2	15.1	05-07	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)
클라우드											
마이크로소프트	MSFT US	409.75	4,434	-0.3	-2.2	-2.8	1.3	-1.8	-2.8	04-25	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	229.15	3,535	-3.5	4.7	10.1	37.3	31.4	4.4	04-30	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	174.46	710	3.5	12.9	-7.6	32.8	51.2	5.0	03-11	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	9,486.00	134	0.3	2.5	1.3	25.3	19.3	3.3	02-12	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	129.84	4,629	11.3	-4.5	-12.0	24.0	80.1	-3.3	02-26	AI 모델 훈련 및 추론에 필수인 GPU계의 현존 최강자
브로드컴	AVGO US	224.87	1,534	3.3	0.2	22.8	52.5	77.5	-3.0	03-06	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	107.56	254	-5.9	-7.3	-27.3	-19.9	-37.6	-11.0	04-30	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	19.10	120	-1.4	-0.3	-27.1	-3.1	-55.4	-4.7	04-25	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	92.30	150	2.6	-7.1	-17.4	-0.6	8.3	9.7	03-20	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	201,500	147	5.6	-1.0	0.5	17.7	41.8	15.9	04-25	HBM 부문 전세계 최강자
시놉시스	SNPS US	525.82	118	1.1	6.8	-6.4	0.2	-8.6	8.3	02-26	반도체 EDA 부문 리딩 업체. AI 에이전트 적극 활용
암페놀	APH US	69.80	123	-0.5	-0.6	-5.3	12.0	33.7	0.5	04-23	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	121.38	66	7.8	-5.9	-3.4	70.0	91.3	6.8	02-12	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	162.51	249	4.4	15.7	10.2	38.9	41.1	31.7	05-08	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	110.62	139	-2.6	-3.2	18.0	82.9	61.2	0.2	03-05	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
SMIC	981 HK	46.45	92	10.9	43.6	62.7	180.2	226.2	46.1	02-11	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,105.00	1,273	3.3	0.5	1.8	19.3	73.9	2.8	04-18	명실상부 반도체 파운드리 최강자
관타	2382 TT	251.00	43	3.3	-11.5	-23.5	0.4	1.7	-12.5	03-17	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	104.00	13	4.0	1.0	-15.8	6.9	-14.5	0.0	05-12	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,500.00	13	18.6	18.8	44.6	57.7	-14.0	6.7	03-03	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스

자료: Bloomberg, 미래에셋증권 리서치센터

주: 실적발표 일정은 변경될 수 있음.

V. Appendix: Charts

그림 49. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)

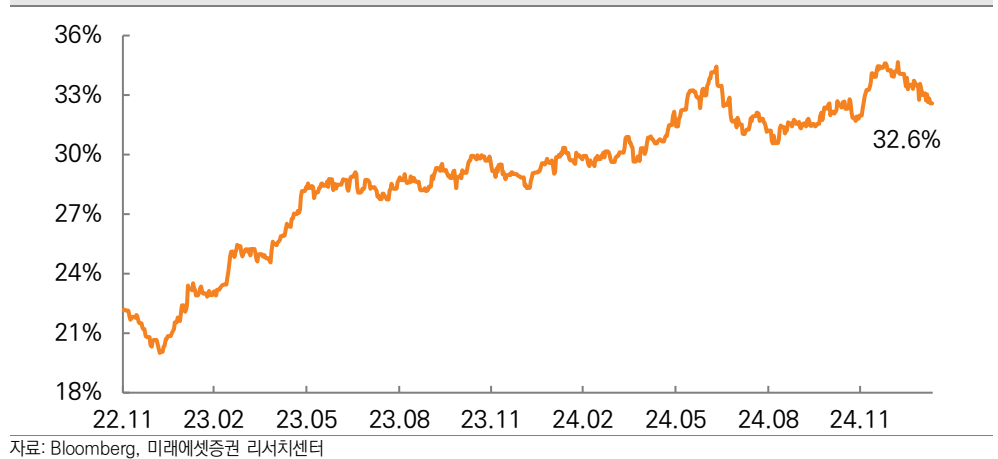


그림 50. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)

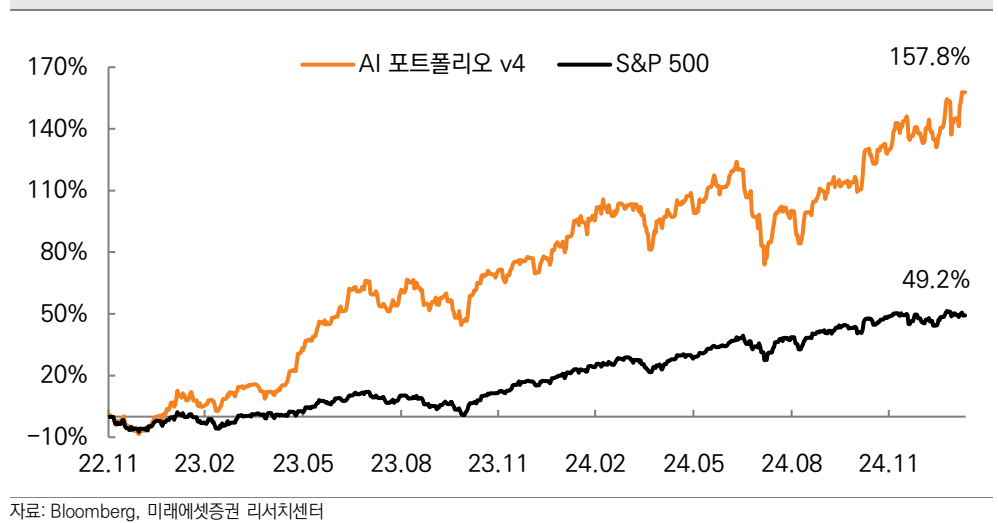


그림 51. 대표 AI ETF vs AI 유니버스(v.4) 수익률 추이(24.12.31 기준)

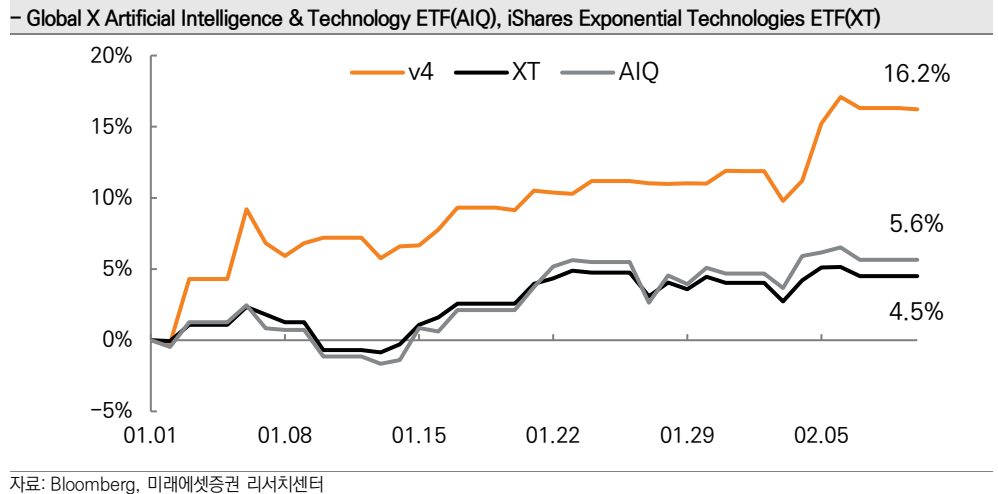


표 12. LLM 벤치마크 챗봇아레나의 “Hard Prompts” 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	아레나 점수	기관명	라이선스	지식 컷오프
1(-)	Gemini-2.0-flash-thinking-exp-0121	1382	구글	독점 폐쇄소스	2024년 6월
1(New)	Gemini-2.0-pro-exp-02-05	1382	구글	독점 폐쇄소스	2024년 6월
3(-)	o1-2024-12-17	1370	OpenAI	독점 폐쇄소스	2023년 10월
4(New)	DeepSeek-R1	1367	DeepSeek	상업적 사용허가	2023년 12월
5(-)	Gemini-2.0-flash	1351	구글	독점 폐쇄소스	2024년 8월
6(▼2)	o1-preview	1352	OpenAI	독점 폐쇄소스	2023년 10월
7(▼1)	ChatGPT-4o-latest	1340	OpenAI	독점 폐쇄소스	2023년 10월
8(▼1)	o1-mini	1336	OpenAI	독점 폐쇄소스	2023년 10월
9(New)	Gemini-2.0-Flash-Lite-Preview	1318	구글	독점 폐쇄소스	2024년 6월
10(▼2)	Deepseek-v3	1304	DeepSeek	상업적 사용허가	-

자료: Imarena.ai, 미래에셋증권 리서치센터

주: Hard Prompts는 기존의 일반적인 요청보다 훨씬 더 까다롭고 복잡한 문제를 제시하여 LLM의 한계를 테스트하는 벤치마크

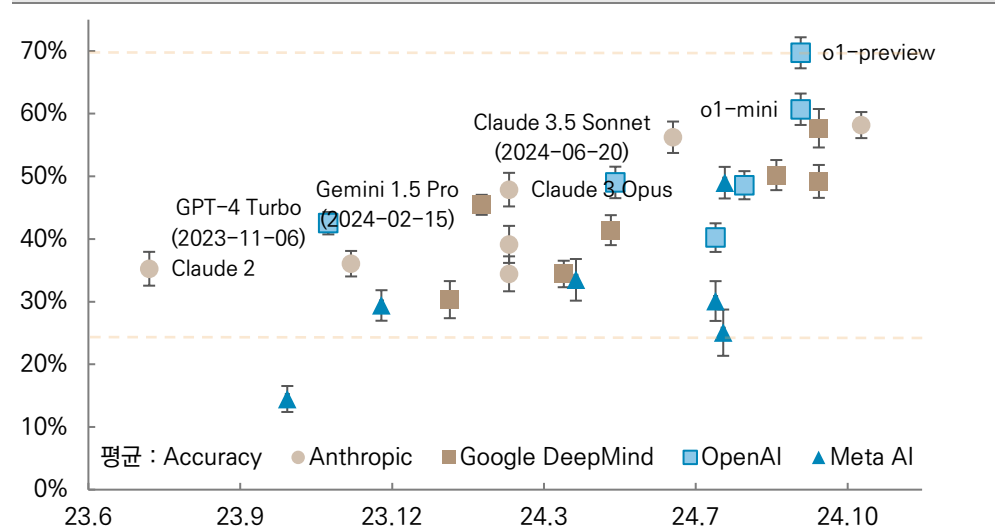
표 13. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(New)	o3-mini-2025-01-31-high	75.88	OpenAI	독점 폐쇄소스	2023년 10월
2(▼1)	o1-2024-12-17-high	75.67	OpenAI	독점 폐쇄소스	2023년 10월
3(▼1)	deepseek-r1	71.57	Deepseek	상업적 사용허가	2023년 12월
4(New)	o3-mini-2025-01-31-medium	70.01	OpenAI	독점 폐쇄소스	2023년 10월
5(New)	Gemini-2.0-flash-thinking-exp-0121	66.92	구글	독점 폐쇄소스	2024년 6월
6(New)	Gemini-2.0-pro-exp-02-05	65.13	구글	독점 폐쇄소스	2024년 6월
7(▼3)	Gemini-exp-1206	64.09	구글	독점 폐쇄소스	-
8(New)	o3-mini-2025-01-31-low	62.45	OpenAI	독점 폐쇄소스	2023년 10월
9(▼2)	Gemini-2.0-flash	61.47	구글	독점 폐쇄소스	2024년 8월
10(▼4)	deepseek-v3	60.45	Deepseek	상업적 사용허가	-

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 14. 각 AI 모델들의 GPQA Diamond 기준 성능 비교

GPQA Diamond는 박사 수준의 과학 분야 질문으로, 인간 전문가들도 약 65%의 정확도를 보임



자료: EPOCH AI, 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 테슬라 15주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 팔란티어 테크 25주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 한종목은(는) 자료작성일 현재 팔란티어 테크 50주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 엔비디아 58주 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.