

이슈보고서

혁신성장금융1부

VOL.2025-이슈 (2025.2)

DeepSeek의 AI 모델과 반도체산업 영향



CONTENTS

<요약>

I. DeepSeek 개요

II. AI산업 영향

III. 반도체산업 영향

작성

선임연구원 이미혜 (6252-3608)



< 요약 >

DeepSeek-V3는 AI 훈련비용이 미국기업의 훈련비용 대비 약 1/10 수준이지만 오픈AI의 GPT-4o 수준의 AI 성능을 구현해 세계의 주목을 받음

- DeepSeek-V3의 AI 훈련 비용은 약 560만 달러로 2,048개의 엔비디아 H800을 2개월간 빌린 비용 기준이며 인건비, 연구개발비, 데이터수집 비용 등은 미포함
- 2024년 12월에 출시된 DeepSeek-V3는 오픈AI의 GPT-4o('24.5 출시), Anthropic의 Claude 3.5 Sonnet('24.6 출시) 수준의 성능을 구현
- 2025년 1월에 출시된 DeepSeek-R1은 V3 기반의 추론 모델로 R1의 성능은 2024년 9월에 출시된 오픈AI의 o1 수준
- DeepSeek-R1의 사용료(API)는 오픈AI의 o1 사용료 대비 1/20 수준으로 AI 대중화의 가능성을 보여줌

DeepSeek은 AI 모델을 오픈소스로 공개하고 무료 상업적 사용을 허용해 AI 개발이 가속화될 전망

- DeepSeek은 프로그램 소스를 MIT 라이선스로 공개하여 누구나 프로그램을 수정, 개선, 재배포할 수 있으며 수정된 버전을 상업적으로 사용할 수 있음
- AI 모델 개발은 Big Tech의 전유물로 생각되었으나 DeepSeek이 저비용 AI 개발 가능성을 열어주어 다수 국가와 기업이 AI 개발을 추진할 전망

AI의 확산은 기술발전으로 자원 사용 효율성이 제고되면 해당 자원의 사용이 증가해 시장이 확대된다는 '제본스의 역설'을 유발해 반도체 수요가 증가할 전망

- AI 모델 개발이 선진국에서 개발도상국으로 확대되고 온디바이스 AI 시장이 성장하면서 반도체 수요가 증가할 전망
- AI는 텍스트 중심의 거대언어모델(LLM), 이미지·음성·영상을 생성하는 멀티모달(Multi Modal) 기반의 AI를 넘어 인간과 비슷하거나 똑똑한 범용AI(Artificial General Intelligence, AGI)로 진화·발전하면서 컴퓨팅파워 고도화가 지속될 전망

미국은 AI 패권 수성을 위해 대중국 반도체·AI모델 수출통제를 강화해 우리기업의 대중국 반도체, 반도체 장비 등의 수출이 영향을 받을 수 있음



I. Deepseek 개요

DeepSeek은 중국 퀀트 헤지펀드 운용사 High-Flyer의 자회사로 2023년 5월에 설립

- 중국 퀀트 헤지펀드¹⁾ 운용사 High-Flyer의 AI 연구소였던 DeepSeek은 2023년에 분사했으며 High-Flyer의 GPU 클러스터를 활용해 AI 모델을 개발
- High-Flyer의 창립자 Liang Wenfeng은 1985년생으로 저장대 전자정보공학 학사, 정보통신공학 석사 학위를 받았으며 2016년 High-Flyer를 설립
- High-Flyer는 2019년에 투자 기법 고도화를 위해 AI 전담 부서를 설립하고 2021년부터 엔비디아 GPU를 구매
 - 중국 언론과 인터뷰에서 High-Flyer는 2021년 미국의 대중국 고사양 반도체 수출금지 이전에 1만장 이상의 엔비디아 A100 GPU를 구매했다고 밝힘
 - DeepSeek은 중국 명문대 출신의 젊은 인재로 회사를 구성하고 AI 모델을 개발
- DeepSeek은 2024년 12월에 거대언어모델(LLM) V3, 2025년 1월 20일에 V3 기반의 추론 모델 R1와 PC에서 로컬로 실행할 수 있는 R1의 경량화 버전 6개 모델을 출시
- 클라우드에 의존하지 않고 PC에서 AI 모델을 실행할 수 있는 6개의 모델을 출시²⁾

DeepSeek의 AI모델 출시 현황

	출시 모델	비고
2023.11	DeepSeek-LLM, DeepSeek-LLM Chat, DeepSeek-Coder	
2024.4	DeepSeek-Math	
2024.5	DeepSeek-V2	2번째 LLM 모델
2024.6	DeepSeek-Coder-V2	2번째 코딩 모델
2024.9	DeepSeek-V2.5	
2024.12	DeepSeek-V3	
2025.1	DeepSeek-R1 및 R1의 경량화 버전 6개 출시	V3 기반 추론 모델

자료 : 옴디아

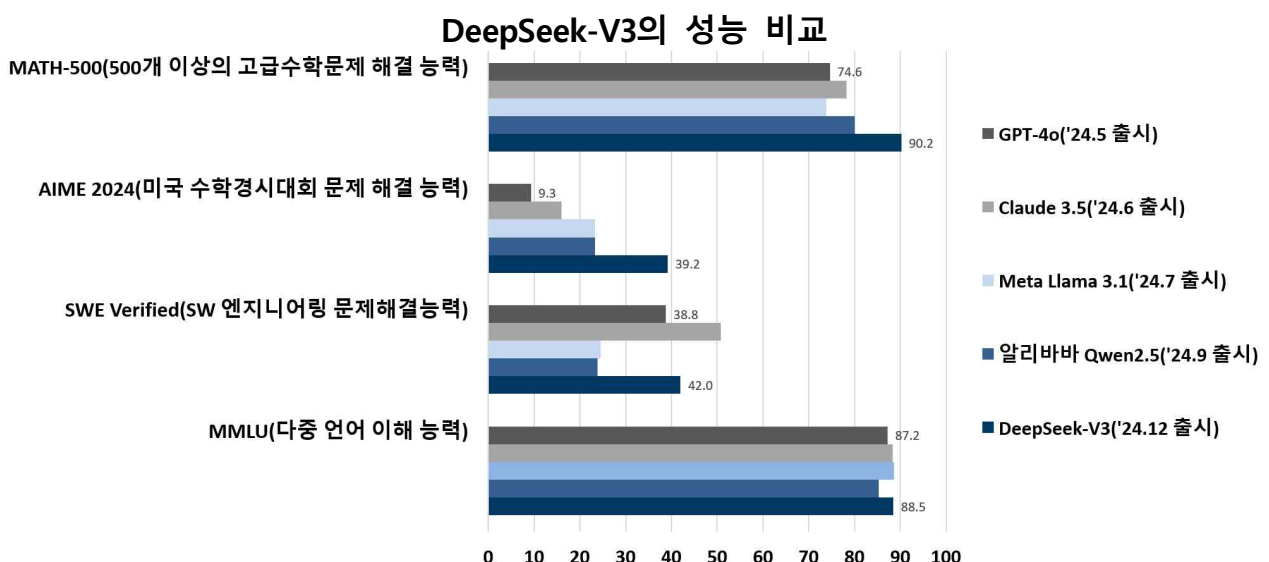
1) 수학 통계 모델과 컴퓨터 알고리즘을 이용한 투자

2) 매개변수(Parameter)는 AI 성능을 결정하며 R1의 경량화 버전은 1.5B, 7B, 8B, 14B, 32B, 70B의 매개변수를 사용



DeepSeek-V3는 AI 훈련비용이 미국기업의 훈련비용 대비 약 1/10 수준이지만 오픈AI의 GPT-4o 수준의 AI 성능을 구현해 세계의 주목을 받음

- DeepSeek-V3의 AI 훈련 비용은 약 560만 달러로 2,048개의 엔비디아 GPU H800³⁾을 2개월간 빌린 비용 기준이며 인건비, 연구개발비, 데이터수집 비용 등은 미포함
- Meta의 Llama 3.1은 16,384개의 엔비디아 H100을 54일간 훈련에 사용했으나 DeepSeek은 미국의 대중국 반도체 수출 규제로 H100 대비 대역폭이 50% 작은 저사양 H800을 사용⁴⁾
 - * 우리나라의 GPU 보유량('23)은 H100 기준으로 2,000장으로 DeepSeek의 GPU 보유량은 미국 빅테크기업보다는 적지만 단일 기업 기준으로 아주 적은 양은 아님
- 2023년 출시된 오픈AI의 GPT-4의 개발비용은 1억 달러, 훈련 비용*은 78.4백만 달러로 추정⁵⁾
 - * AI 훈련 비용은 클라우드 컴퓨팅 임대 가격 기준이며 데이터 라이선스 비용, 데이터센터 냉각비용, 인건비 등은 미포함
- 2024년 12월에 출시된 DeepSeek-V3는 오픈AI의 GPT-4o('24.5 출시), Anthropic의 Claude 3.5 Sonnet('24.6 출시) 수준의 성능을 구현
- V3는 가장 최근에 출시된 AI 모델로 언어 능력과 소프트웨어 엔지니어링 문제 해결 능력은 GPT-4o보다 소폭 우수하나 수학은 매우 우수한 벤치마크를 기록



자료 : DeepSeek.

3) H800과 대중국 수출규제 대상인 H100은 동일 연산능력을 보유하나 대역폭이 작음.

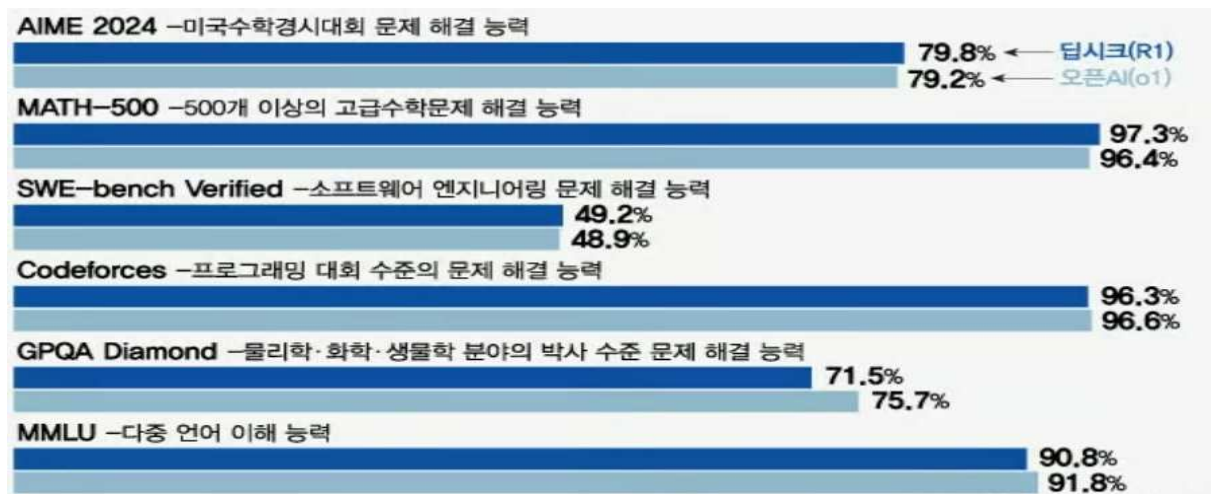
4) H800과 H100은 동일 용량의 HBM3를 사용하나 대역폭(Bandwidth)가 작음

5) 훈련비용은 Stanford University의 Artificial Intelligence Index Report 2024 기준



- 2025년 1월에 출시된 DeepSeek-R1은 V3 기반의 추론 모델로 R1의 성능은 2024년 9월에 출시된 오픈AI의 o1 수준
- 6개 평가 항목중 3개 항목(수학 등)에서 DeepSeek-R1이 오픈AI의 o1 대비 소폭 우수한 성능을 기록

DeepSeek-R1과 오픈AI의 o1의 테스트별 정확도



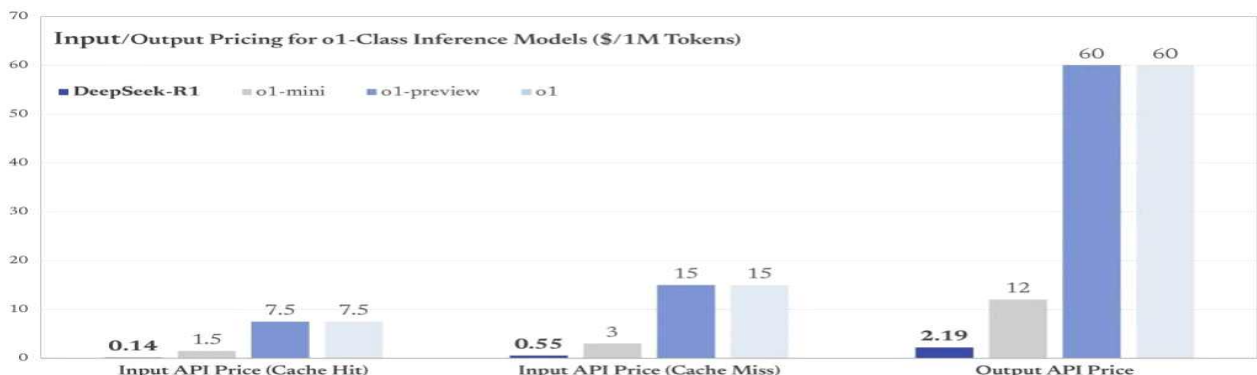
자료 : DeepSeek, 머니투데이.

DeepSeek-R1의 사용료(API, Application Programming Interface)는 오픈AI의 o1 사용료 대비 1/20 수준으로 AI 대중화의 가능성을 보여줌

- 1백만 토큰* 기준으로 DeepSeek-R1의 Input Cost는 0.55달러, Output Cost는 2.19달러로 오픈AI의 Input Cost는 15.00달러, Output Cost 60.00달러보다 매우 낮은 수준을 제시

* 토큰(Token)은 LLM(거대언어모델)이 텍스트를 이해하고 생성하는 기본 단위

DeepSeek-R1의 사용료



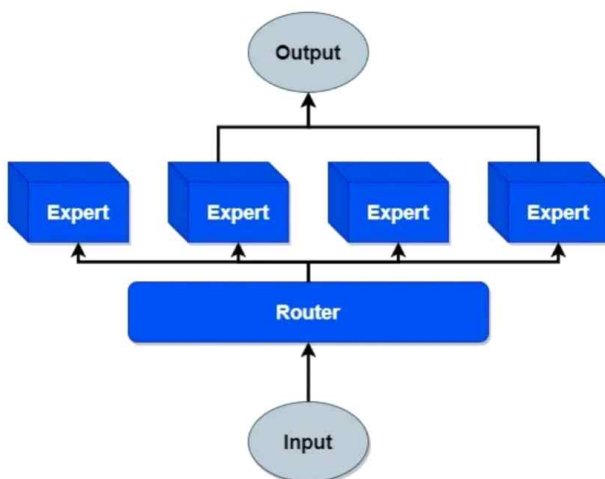
자료 : DeepSeek.



DeepSeek은 미국의 대중국 반도체 수출규제로 고성능 GPU 사용이 어려웠으나 기존에 알려진 AI 기술의 효율적 사용 등을 통해 AI의 성능을 제고

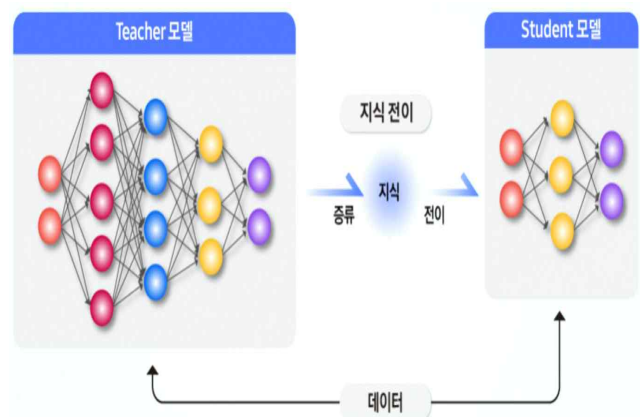
- Mixture of Experts(MoE, 전문가 혼합)는 AI 모델을 다수의 전문가 서브 모델로 나누고 입력된 작업에 따라 적합한 전문가 모델만을 활성화시켜 연산을 최적화시킴
- DeepSeek 모델의 파라미터의 수는 6,170억개이지만 실제 작동시에는 약 5%(370억개)의 파라미터만을 선택적으로 활성화해 연산 부담과 비용을 절감
- Mixture of Experts(MoE)는 구글, 프랑스 Mistral AI, 텐센트 등도 사용한 방식
- AI를 활용한 대규모 강화학습(Reinforcement Learning)으로 학습비용을 절감
- R1은 OpenAI 등이 사용하는 RLHF(Reinforcement Learning from Human Feedback, 인간의 피드백을 통한 강화학습)를 사용하지 않고 AI를 활용한 강화학습을 사용
- 지식 증류(Knowledge Distillation)를 통해 소형 AI 모델의 정확도를 제고
- 증류는 대형 AI 모델(Teacher 모델)과 소형 AI 모델(Student 모델)에 동일 데이터를 입력한 뒤 나온 출력값의 차이를 줄여나가는 방식으로 2024년에 노벨 물리학상을 수상한 토론토 대학의 제프리 힌턴 교수가 제시한 개념
- Multi-head Latent Attention(MLA, 멀티헤드 잠재 어텐션)은 입력 문장 중 유의미한 내용만을 추출하여 연산하여 추론 속도를 향상시킴
- DeepSeek은 2024년에 출시된 V2에 Multi-head Latent Attention(MLA)를 처음 적용

Mixture of Experts(MoE)



자료 : Media.

지식 증류 개념도



자료 : 삼성전자.



마이크로소프트, 아마존 등은 오픈AI에 대한 의존도를 낮추고 사용자 저변 확대를 위해 자사 서비스에 DeepSeek을 도입

- 마이크로소프트는 Azure AI 파운드리와 개발자 도구 GitHub에 DeepSeek-R1 모델을 추가했으며 Copilot+ PC에서 R1을 로컬로 실행할 수 있도록 지원할 예정
- 클라우드기업 아마존웹서비스는 자사의 거대언어모델 개발 및 배포 플랫폼인 Bedrock에서 R1 사용이 가능하다고 발표
- AI 검색 서비스를 제공하는 Perplexity AI는 DeepSeek-R1을 추가
- 엔비디아는 DeepSeek-R1을 자사 AI 소프트웨어 플랫폼인 NIM 마이크로 서비스에서 지원할 예정
- 주요국은 DeepSeek의 보안 이슈 등으로 공공부문의 사용을 금지했지만 주요 기업들은 DeepSeek에 대한 안정성 평가를 수행하고 서비스를 제공한다고 밝힘
- 클라우드를 통해 제공되는 API는 보안이나 파인튜닝(fine-tuning)⁶⁾ 등에서 장점을 보유

6) 사전 학습된 AI 모델을 특정 작업에 맞게 조정하고 정교화하는 과정



II. AI산업 영향

DeepSeek은 AI 모델을 오픈소스로 공개하고 무료 상업적 사용을 허용해 AI 개발이 가속화될 전망

- DeepSeek은 프로그램 소스를 MIT 라이선스⁷⁾로 공개하여 누구나 프로그램을 수정, 개선, 재배포할 수 있으며 수정된 버전을 상업적으로 사용할 수 있음
- 오픈AI는 AI 모델을 공개하지 않는 폐쇄형 전략, Meta는 AI 모델을 오픈소스로 공개하나 상업적 사용 등에 제약이 있음
 - 오픈AI, 구글 등은 AI 기술의 오남용 우려 등으로 소스코드를 공개하지 않고 AI 모델을 사용할 수 있는 API(Application Programming Interface)만을 제공
 - DeepSeek과 알리바바가 오픈소스 전략을 채택하고 있으며 이들은 후발주자로 집단 지성을 통해 미국 AI 기술을 추격하기 위해 오픈소스를 채택한 것으로 추정
 - * 오픈소스는 모델의 가중치, 학습 방법론 등을 공개
- DeepSeek의 여파로 오픈AI는 오픈소스로의 전환을 검토중이며 Meta는 DeepSeek의 기술 분석 등을 위해 War room을 구성하고 분석 결과를 자사 AI 모델에 적용할 계획
- AI 모델 개발은 Big Tech의 전유물로 생각되었으나 DeepSeek이 저비용 AI 개발 가능성을 열어주어 다수 국가와 기업이 AI 개발을 추진할 전망

DeepSeek의 경량화 AI 모델은 스마트폰, PC, 자동차, 로봇 등 엣지(Edge) 디바이스의 AI 탑재를 촉진할 전망

- 중국에서는 DeepSeek 열풍으로 중국 자동차회사, 금융회사, 스마트폰 제조사 등에서 DeepSeek을 탑재⁸⁾
- 중국 전기차회사 BYD는 DeepSeek을 자동차에 적용하기로 했으며 자동차 회사들은 AI 모델을 활용해 운전 중 음성명령 기능 등을 향상시킬 계획
- 주요 자동차, 스마트폰 제조사 등도 AI 탑재를 확대하는 추세
- AI PC 등이 제한적인 기능 등으로 시장 확대가 제약을 받는 상황에서 엣지 디바이스용 AI 모델의 고도화는 새로운 시장을 창출할 전망

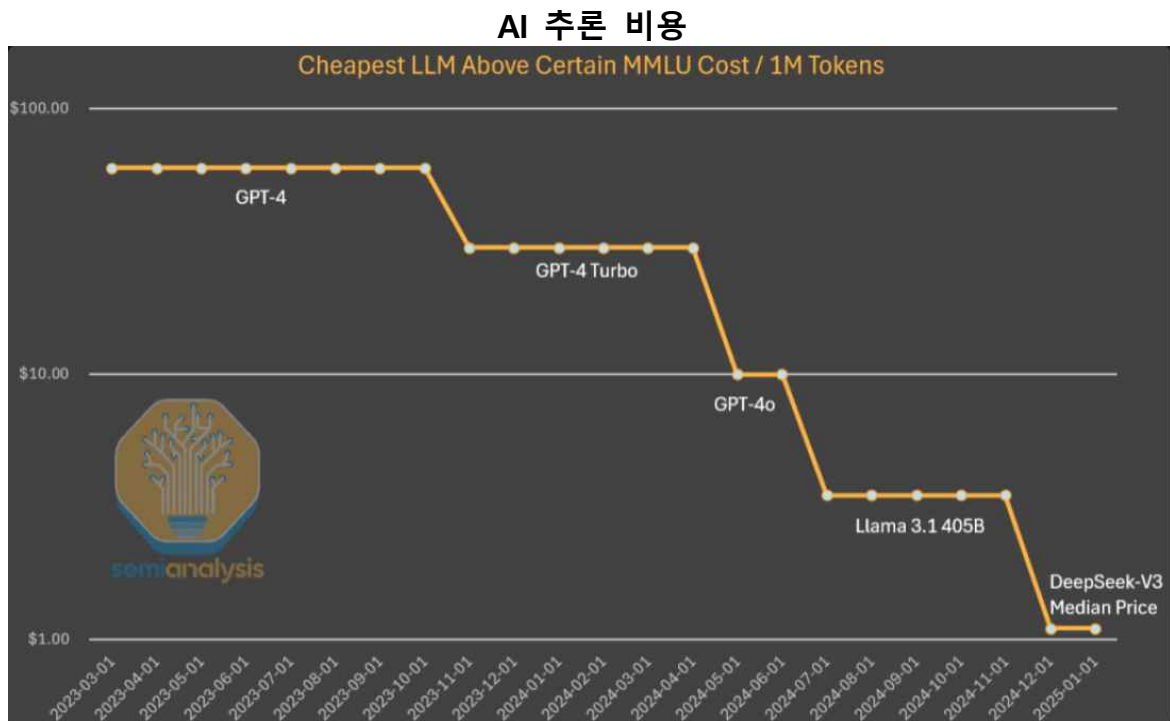
7) MIT가 MIT의 소프트웨어 공학도들을 돕기 위해 개발한 라이선스에서 출발하여 MIT 라이선스로 불림

8) 중국 국가 슈퍼컴퓨팅 네트워크는 자격요건을 갖춘개인과 기업에 DeepSeek API 3개월 무료 사용권 등을 제공



AI 추론 비용은 12개월마다 1/10 수준으로 낮아져왔으며 DeepSeek으로 인해 AI 모델 사용료 인하 등이 가속화될 가능성이 있음

- 2025년 1월, 오픈AI는 경량 추론 모델 o3-미니를 출시했으며 성능은 R1보다 높지만 사용료를 낮추어 DeepSeek-R1과의 사용료 격차를 10배에서 2~4배 수준으로 좁힘
- o3-미니는 ChatGPT 무료 버전에 탑재해 일반 사용자들의 사용을 지원
- 2025년 2월, 구글이 출시한 제미니 2.0 플래시-라이트의 사용료는 Input cost 0.075달러, Output cost 0.3달러로, Input 비용은 V3와 비슷하고 Output 비용은 1/3 이하⁹⁾



주: MMLU(Massive Multitask Language Understanding)
자료 : SemiAnalysis

9) AI타임스, '딥시크, 8일부터 'V3' 사용료 5배 인상...성능·비용 모두 우위 잃어', 2025.2.6



Ⅲ. 반도체산업 영향

AI의 확산은 기술발전으로 자원 사용 효율성이 제고되면 해당 자원의 사용이 증가해 시장이 확대된다는 '제본스의 역설'을 유발해 반도체 수요가 증가할 전망

- AI 모델 개발이 선진국에서 개발도상국으로 확대되고 온디바이스 AI 시장이 성장하면서 반도체 수요가 증가할 전망
- 데이터센터 투자는 미국과 중국 중심이었으나 투자 지역이 다변화되고 생산성 향상 등을 위해 전산업에 AI가 도입될 전망
 - 프랑스는 AI 분야에 1,090억 유로 투자를 발표('25.2)했으며 투자금의 대부분은 AI 데이터센터 건설에 사용될 예정
 - 우리나라는 국가 AI컴퓨팅 센터 구축 실행계획(안)을 발표('25.1.22)하고 민·관 합작 투자를 통해 최대 2조 원 규모의 국가 AI컴퓨팅 센터를 구축할 계획
- AI는 텍스트 중심의 거대언어모델(LLM), 이미지·음성·영상을 생성하는 멀티모달(Multi Modal) 기반의 AI를 넘어 인간과 비슷하거나 똑똑한 범용AI(Artificial General Intelligence, AGI)로 진화·발전하면서 컴퓨팅 파워 고도화가 지속될 전망
- 트럼프 대통령은 취임하자마자 미국에 AI 슈퍼컴퓨터 등 AI 인프라를 구축하는 사업인 Stargate Project를 발표했으며 투자비는 향후 4년간 5,000억 달러 규모
 - 오픈AI, 소프트뱅크, 오라클 등이 참여할 예정
- 마이크로소프트, 오픈AI, 구글 등이 범용AI(AGI)를 지향하며 오픈AI의 CEO 샘 알트먼과 구글 딥마인드의 CEO 데미스 허사비스는 5년내에 범용AI의 등장 가능성을 전망

반도체산업의 단기 영향은 제한적이나 중장기로는 AI 인프라 투자 속도 등이 조정받을 가능성이 있음

- AI 투자의 패러다임이 하드웨어에서 소프트웨어, 알고리즘으로 이동할 것이라는 우려에도 불구하고 미국 주요 기업은 2025년 AI 투자를 기존 계획보다 확대할 전망
- 아마존은 AI 인프라 투자를 2024년 830억 달러에서 2025년 1,000억 달러 이상으로 확대할 계획
- Meta는 2025년에 AI 연구개발 및 대규모 신규 데이터센터 구축 등에 최대 650억 달러를 투자할 계획
- AI 모델 개발 방향은 대형·범용 모델에서 소형, task-specific, 오픈소스 모델로 전환이 가속화될 전망



주요 기업의 DeepSeek에 대한 의견

기업	의견
Meta	DeepSeek의 사례는 오픈소스가 표준이 될 것이라는 함의를 가지며 AI 모델을 오픈소스로 제공하는 우리의 전략이 지속적이고 효과적일 것
마이크로소프트	DeepSeek은 AI의 상품화에 크게 기여하여 우리와 같은 하이퍼스케일러 및 PC 플랫폼 제공 업체에게 매우 긍정적
ASML	AI 비용이 낮아지면 더 많은 어플리케이션에 AI가 적용될 수 있으며 이는 반도체 수요 증가로 귀결됨. AI 비용 감소는 ASML에게 긍정적인 시그널임.
Anthropic	· DeepSeek-V3는 일부 중요 작업에서 최첨단 미국 AI 모델의 성능에 근접한 것으로 보이며 훈련 비용이 상당히 낮다고 평가 · AI 개발은 자원이 매우 중요해 중국이 고성능 GPU 없이 AI 모델을 개발하는 것은 한계가 있을 것으로 평가

자료 : 유진투자증권, 뉴스.

미국은 AI 패권 수성을 위해 대중국 반도체·AI모델 수출통제를 강화해 우리기업의 대중국 반도체, 반도체 장비 등의 수출이 영향을 받을 수 있음

- 2025년 1월 13일, 바이든 행정부는 중국·러시아 등 '적성국'으로 분류한 22국에 대한 초고성능 AI 모델 수출 차단 조치를 발표
- 미국은 전 세계 국가를 우방국·적성국·기타국 등 3등급으로 분류해 최상위 우방국은 AI 반도체와 초고성능 AI 모델을 추가적인 제한 없이 확보 가능
 - 오픈AI 등의 폐쇄형 초고성능 AI 모델이 대상
 - 최상위 우방국에는 한국, 일본·대만·네덜란드·영국·독일·프랑스 등 18개국이 포함됨
- 동맹국이 아닌 '기타 국가'에 대한 반도체 수출 제한 조치를 신규 도입해 기타 국가에는 미국이 수출할 수 있는 반도체의 '총연산력'에 상한을 설정
 - 엔비디아의 H100을 기준으로 '2년간 32만개'에 해당하는 연산력이 상한으로 설정
- 중국·러시아 등 22국은 모든 AI 반도체와 폐쇄형 초고성능 AI 모델에 대한 접근을 차단
- 중국을 견제하기 위해 대중국 반도체 수출 규제가 저사양 AI 반도체, 레가시 반도체 (28나노 이상 공정 사용) 등으로 확대되고 장비 등의 수출 통제 강화 가능성이 있음



참고문헌

Omdia, DeepSeek sanity check, 2025.1.31.

DeepSeek, 'DeepSeek-V3 Technical Report', 2024.12

SemiAnalysis, 'DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts', 2025.1.31.

Tom's Hardware, 'Faulty Nvidia H100 GPUs and HBM3 memory caused half of failures during LLama 3 training — one failure every three hours for Meta's 16,384 GPU training cluster', 2024.7.28.

AI Post, '6년 만에 AI 훈련 비용 '20만배' 늘었다..."구글 제미나이 울트라 훈련비 2600 억원 추정"', 2024.6.5.

AI타임스, '딥시크, 8일부터 'V3' 사용료 5배 인상...성능·비용 모두 우위 잃어', 2025.2.6