미국은 지금

DeepSeek 쇼크, 위기에서 찾은 기회

키움증권 리서치센터 글로벌리서치팀
US Strategy 김승혁 ocean93@kiwoom.com

Issue Brief

DeepSeek 쇼크와 시장의 엇갈린 평가

DeepSeek 가 공개한 LLM 모델이 '저 비용'으로 GPT 수준의 '고 성능' AI 모델을 만들 수 있음을 증명하자 시장은 크게 하락했다. DeepSeek 의 V3 모델은 훈련에 약 560 만 달러의 컴퓨팅 비용을 투자해 메타 Llama 3 모델 개발 비용 대비 1/10 을 줄였다. 이후 엔비디아를 포함한 반도체 관련 기업 주가는 급락했고 에너지, 유틸리티 섹터 내 기업들 역시 빠르게 하락했다.

시장은 DeepSeek 에 대해 Open AI 데이터 무단 도용 가능성, 인건비를 누락한 비용 집계 등 여러 의혹을 제기하고 있지만, 저비용 고성능 모델을 만든 기술력에 대해서는 인정하는 모습도 보이고 있다.

DeepSeek 의 성과와 한계

DeepSeek 가 공개한 논문을 보면 **①강화 학습(RL) 중심 훈련 모델**, **②지식 증류(Distillation)**, **③MoE(Mixture of Experts)** 등을 통해 저비용 고성능 AI 모델을 만들었다고 명시되어 있다. ① 은 사람이 만든 대규모 라벨 없이 모델 스스로 학습해 데이터 비용을 절약하는 방법이고, ②는 큰 모델의 복잡한 추론 과정을 작은 모델이 흉내 내어 적은 연산으로 높은 성능을 만들어 내는 방법이다. ③은 필요한 연산만 선택적으로 수행해 불필요한 계산을 줄인 효율적 방법이다.

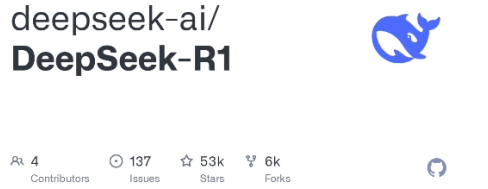
위 방법은 저비용 고성능 모델을 만들어 내는데 기술적 의의가 있다. 특히 **소형 모델의 효과적 학습 기술은 관련 전문가들의 인정**을 받고 있다. 하지만, 강화 학습(RL), 지식 증류 등은 **거대 Base 모델 없이 독립적으로 작동하기 어렵다는** 한계가 있다. 또한, **범용성과 안전성 문제로 현실 세계의 많은 변수를 처리하는 데도 어려움**이 있을 수 있다.

DeepSeek? 위기 보다는 기회

DeepSeek 가 인정받고 있는 소형 모델의 효과적 학습 기술은 단기적 관점에서는 **엣지 AI 와 온디바이스 AI 산업에 기회를 제공**할 수 있다. 현장 연산이 필요한 엣지 AI 와 네트워크 연결 없이 연산이 필요한 온디바이스 AI 에 탑재되는 소형 AI 모델이 높은 효율을 낼 수 있기 때문이다.

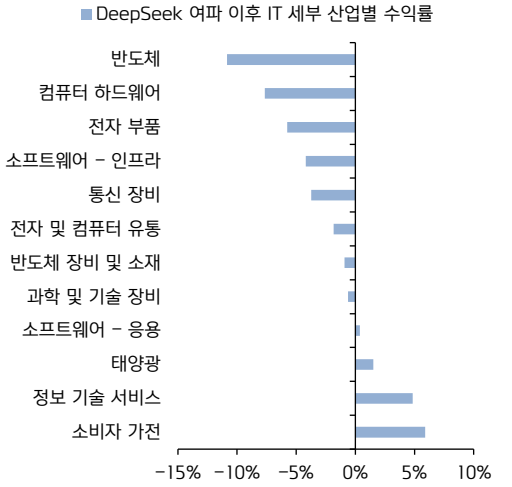
장기적 관점에서는 **AI 산업 전반의 수요 증가**로 연결될 가능성을 전망한다. Jevons 의 역설과 같이 AI 학습 효율화가 AI 모델 수요 증가로 이어져 산업 전반에서의 AI 수요를 높일 수 있기 때문이다. 나아가 소형 AI 모델 비용 절감은 진입장벽을 낮추어 **중소형 AI 기업들의 시장 진입을 촉진**할 수 있다.

DeepSeek 가 공개한 언어 모델 R1



자료: Deepseek, 키움증권 리서치센터

DeepSeek 쇼크에 따른 IT 세부 산업별 수익률



자료: Bloomberg, 키움증권 리서치센터
주) '25.01.27 ~ '25.01.31 기간 수익률

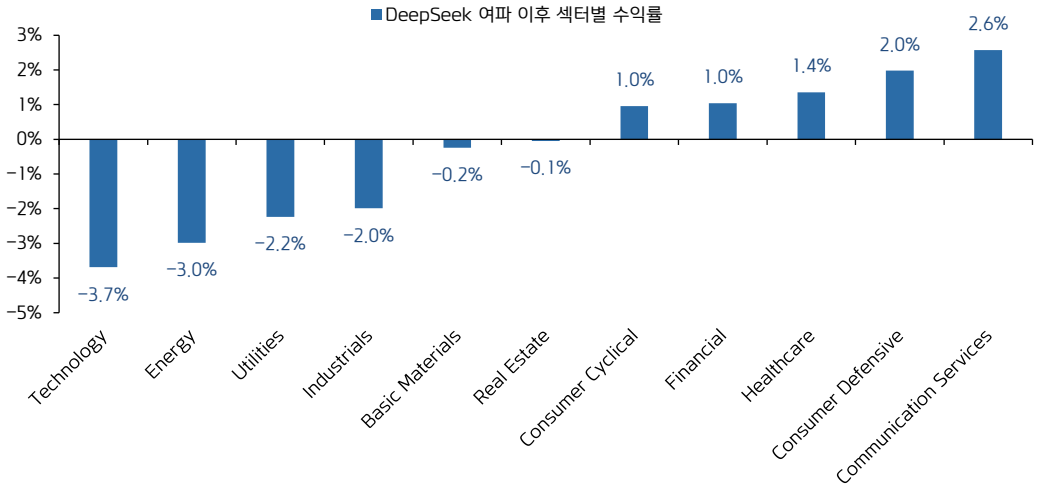
시장은 어떤 이유로 충격을 받았나?

중국 AI 모델 개발 스타트업 DeepSeek 는 25 년 1 월 20 일 R1 과 R1-zero 모델을 공개했고, 22 일 자체 논문을 공개했다. 이를 통해 DeepSeek 는 AI 업계의 전통 레거시인 Scaling Rule(많은 자원을 투입해야 모델의 성능이 좋아진다)을 정면으로 반박하며 **'낮은 비용'으로도 OpenAI 와 같이 '높은 수준의' LLM 모델을 만들 수 있다고 주장**했다.

실제로 DeepSeek 는 V3 모델 훈련에 약 560 만 달러의 컴퓨팅 비용만 투자해 메타 Llama 3 모델 개발 비용 대비 1/10 수준으로 절약했다. 또한 R1 모델 개발 비용 역시 OpenAI o1 모델 개발 비용 대비 1/27 수준으로 낮추었다. 그럼에도 불구하고 DeepSeek 의 V3/R1 모델은 여러 성능 테스트에서 GPT-4 와 유사한 레벨을 기록했다. 또한 DeepSeek 는 V3 모델 개발에 2,048 개의 H800 GPU 를 사용했다. 상대적으로 저사양 반도체인 H800 GPU 를 활용했음에도 OpenAI 모델과 비슷한 성능을 낸 것이다. 투입 비용도 낮고, 투입 반도체도 저사양 이었지만 미국 레거시 LLM 모델과 동등한 성능을 보였다는 점은 시장에 충격을 주었다.

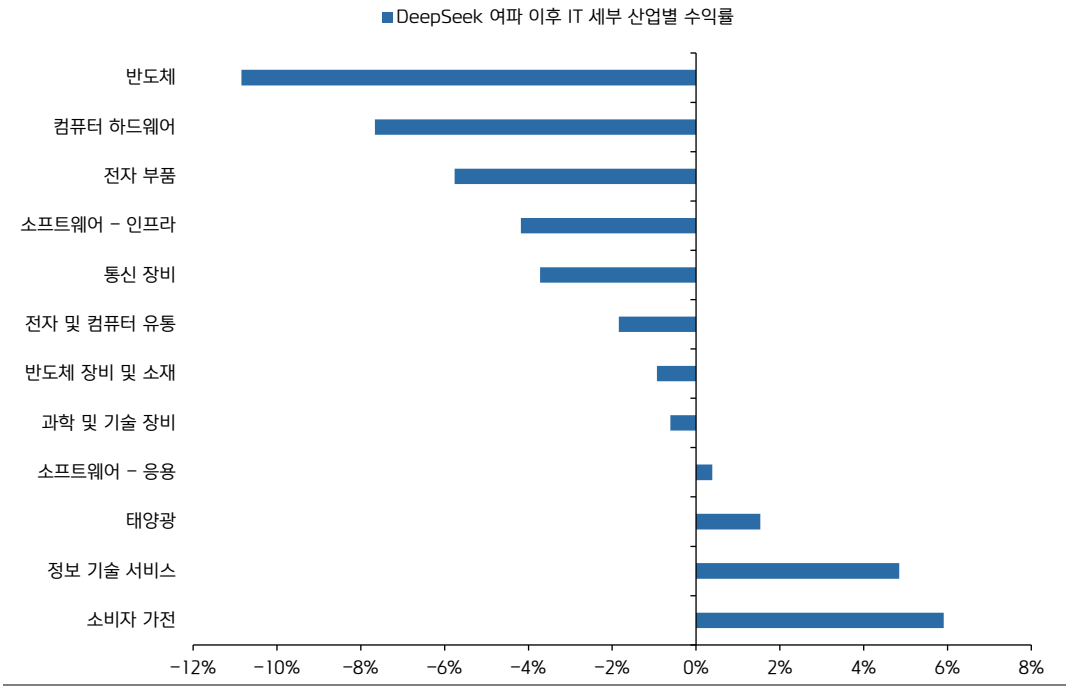
대표적으로 급락세를 보인 종목은 엔비디아(NVDA)였다. DeepSeek 發 내러티브가 확산된 직후(1 월 24 일) 엔비디아는 최대 17%까지 급락했으며, 한국 연휴 기간 동안에는(1/27~1/30) -15.3% 하락했다. 같은 기간 동안 서버 제조사인 Super Micro Computer(SMCI, -14.3%), 맞춤형 반도체 기업인 Broadcom(AVGO, -10.3%), 전력 유틸리티 기업인 GE Vernova(GEV, -12.5%), 신재생 에너지 기업인 Constellation Energy(CEG, -10.9%), AI 소프트웨어 기업 Oracle(ORCL, -8.6%) 등도 크게 하락했다. 트럼프 행정부의 스타게이트 발표에 따라 긍정적 주가를 보였던 AI 인프라 및 관련 기업들이 DeepSeek 트리거로 인해 한순간 조정 받은 것이다. 산업 기준으로 살펴보면 반도체, 컴퓨터 하드웨어, AI 전력 인프라 관련 기업들이 우선적으로 조정 받았음을 알 수 있다.

DeepSeek 쇼크에 따른 산업별 수익률



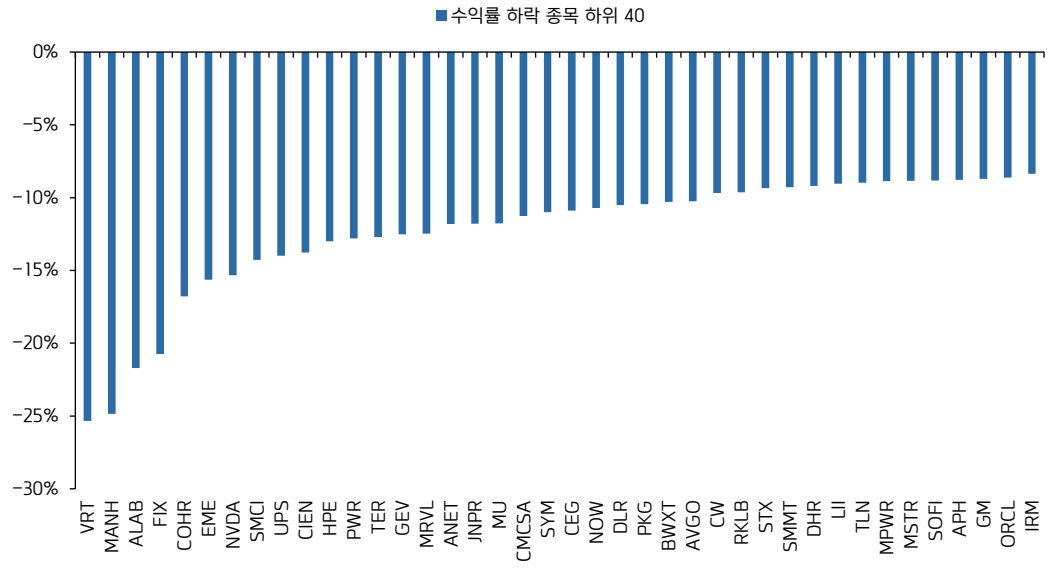
자료: Bloomberg, 키움증권 리서치센터, 주) 2025.01.27 ~ 2025.01.31 기간 수익률

DeepSeek 쇼크에 따른 IT 세부 산업별 수익률



자료: Bloomberg, 키움증권 리서치센터, 주) 2025.01.27 ~ 2025.01.31 기간 수익률

DeepSeek 쇼크에 따른 세부 종목별 수익률



자료: Bloomberg, 키움증권 리서치센터, 주) 시가총액 10Bln 이상 종목 중심 추출, 2025.01.27 ~ 2025.01.31 기간 수익률

DeepSeek 쇼크 이후 종목별 수익률 하위 40

티커	기업 명	섹터	세부 산업 분류	수익률 (%)
VRT	Vertiv Holdings Co	Industrials	Electrical Equipment & Parts	-25.3%
MANH	Manhattan Associates, Inc	Technology	Software - Application	-24.9%
ALAB	Astera Labs Inc	Technology	Semiconductors	-21.7%
FIX	Comfort Systems USA, Inc	Industrials	Engineering & Construction	-20.7%
COHR	Coherent Corp	Technology	Scientific & Technical Instruments	-16.8%
EME	Emcor Group, Inc	Industrials	Engineering & Construction	-15.6%
NVDA	NVIDIA Corp	Technology	Semiconductors	-15.3%
SMCI	Super Micro Computer Inc	Technology	Computer Hardware	-14.3%
UPS	United Parcel Service, Inc	Industrials	Integrated Freight & Logistics	-14.0%
CIEN	CIENA Corp	Technology	Communication Equipment	-13.8%
HPE	Hewlett Packard Enterprise Co	Technology	Communication Equipment	-13.0%
PWR	Quanta Services, Inc	Industrials	Engineering & Construction	-12.8%
TER	Teradyne, Inc	Technology	Semiconductor Equipment & Materials	-12.7%
GEV	GE Vernova Inc	Utilities	Utilities - Renewable	-12.5%
MRVL	Marvell Technology Inc	Technology	Semiconductors	-12.5%
ANET	Arista Networks Inc	Technology	Computer Hardware	-11.8%
JNPR	Juniper Networks Inc	Technology	Communication Equipment	-11.8%
MU	Micron Technology Inc	Technology	Semiconductors	-11.8%
CMCSA	Comcast Corp	Communication Services	Telecom Services	-11.3%
SYM	Symbotic Inc	Industrials	Specialty Industrial Machinery	-11.0%
CEG	Constellation Energy Corporation	Utilities	Utilities - Renewable	-10.9%
NOW	ServiceNow Inc	Technology	Software - Application	-10.7%
DLR	Digital Realty Trust Inc	Real Estate	REIT - Specialty	-10.5%
PKG	Packaging Corp Of America	Consumer Cyclical	Packaging & Containers	-10.5%
BWXT	BWX Technologies Inc	Industrials	Aerospace & Defense	-10.3%
AVGO	Broadcom Inc	Technology	Semiconductors	-10.3%
CW	Curtiss-Wright Corp	Industrials	Aerospace & Defense	-9.7%
RKLB	Rocket Lab USA Inc	Industrials	Aerospace & Defense	-9.6%
STX	Seagate Technology Holdings Plc	Technology	Computer Hardware	-9.4%
SMMT	Summit Therapeutics Inc	Healthcare	Biotechnology	-9.3%
DHR	Danaher Corp	Healthcare	Diagnostics & Research	-9.2%
LII	Lennox International Inc	Industrials	Building Products & Equipment	-9.1%
TLN	Talen Energy Corp	Utilities	Utilities - Independent Power Producers	-9.0%
MPWR	Monolithic Power System Inc	Technology	Semiconductors	-8.9%
MSTR	Microstrategy Inc	Technology	Software - Application	-8.9%
SOFI	SoFi Technologies Inc	Financial	Credit Services	-8.8%
APH	Amphenol Corp	Technology	Electronic Components	-8.8%
GM	General Motors Company	Consumer Cyclical	Auto Manufacturers	-8.7%
ORCL	Oracle Corp	Technology	Software - Infrastructure	-8.6%
IRM	Iron Mountain Inc	Real Estate	REIT - Specialty	-8.4%

자료: Bloomberg, 키움증권 리서치센터, 주) 시가총액 10Bln 이상 종목 중심 추출, 수익률: 2025.01.27 ~ 2025.01.31 기간

DeepSeek 쇼크 이후 종목별 수익률 상위 40

티커	기업 명	섹터	세부 산업 분류	수익률 (%)
TWLO	Twilio Inc	Technology	Software – Infrastructure	30.82%
RCL	Royal Caribbean Group	Consumer Cyclical	Travel Services	15.39%
IBM	International Business Machines Corp	Technology	Information Technology Services	14.26%
CAVA	Cava Group Inc	Consumer Cyclical	Restaurants	12.05%
CCL	Carnival Corp	Consumer Cyclical	Travel Services	11.25%
SBUX	Starbucks Corp	Consumer Cyclical	Restaurants	11.24%
LVS	Las Vegas Sands Corp	Consumer Cyclical	Resorts & Casinos	11.02%
NET	Cloudflare Inc	Technology	Software – Infrastructure	10.76%
FFIV	F5 Inc	Technology	Software – Infrastructure	10.47%
NCLH	Norwegian Cruise Line Holdings Ltd	Consumer Cyclical	Travel Services	10.03%
LAD	Lithia Motors, Inc	Consumer Cyclical	Auto & Truck Dealerships	9.85%
AFRM	Affirm Holdings Inc	Technology	Software – Infrastructure	9.41%
RDDT	Reddit Inc	Communication Services	Internet Content & Information	9.31%
ZM	Zoom Communications Inc	Technology	Software – Application	9.10%
GTLB	Gitlab Inc	Technology	Software – Infrastructure	8.51%
TMUS	T-Mobile US Inc	Communication Services	Telecom Services	8.48%
IOT	Samsara Inc	Technology	Software – Infrastructure	8.34%
DUOL	Duolingo Inc	Technology	Software – Application	8.12%
DGX	Quest Diagnostics, Inc	Healthcare	Diagnostics & Research	8.11%
META	Meta Platforms Inc	Communication Services	Internet Content & Information	7.94%
RKT	Rocket Companies Inc	Financial	Mortgage Finance	7.90%
KVYO	Klaviyo Inc	Technology	Software – Infrastructure	7.66%
CART	Maplebear Inc	Consumer Cyclical	Internet Retail	7.59%
SSB	SouthState Corporation	Financial	Banks – Regional	7.47%
LLY	Lilly(Eli) & Co	Healthcare	Drug Manufacturers – General	7.39%
SNAP	Snap Inc	Communication Services	Internet Content & Information	7.28%
CPNG	Coupang Inc	Consumer Cyclical	Internet Retail	7.27%
WMG	Warner Music Group Corp	Communication Services	Entertainment	7.25%
ROP	Roper Technologies Inc	Technology	Software – Application	7.01%
KMX	Carmax Inc	Consumer Cyclical	Auto & Truck Dealerships	6.83%
TPX	Tempur Sealy International Inc	Consumer Cyclical	Furnishings, Fixtures & Appliances	6.82%
STLD	Steel Dynamics Inc	Basic Materials	Steel	6.82%
SFM	Sprouts Farmers Market Inc	Consumer Defensive	Grocery Stores	6.71%
EL	Estee Lauder Cos., Inc	Consumer Defensive	Household & Personal Products	6.70%
T	AT&T, Inc	Communication Services	Telecom Services	6.61%
AXON	Axon Enterprise Inc	Industrials	Aerospace & Defense	6.53%
MELI	MercadoLibre Inc	Consumer Cyclical	Internet Retail	6.48%
OKTA	Okta Inc	Technology	Software – Infrastructure	6.46%
RBLX	Roblox Corporation	Communication Services	Electronic Gaming & Multimedia	6.36%
VEEV	Veeva Systems Inc	Healthcare	Health Information Services	6.32%

자료: Bloomberg, 키움증권 리서치센터, 주) 시가총액 10Bln 이상 종목 중심 추출, 수익률: 2025.01.27 ~ 2025.01.31 기간

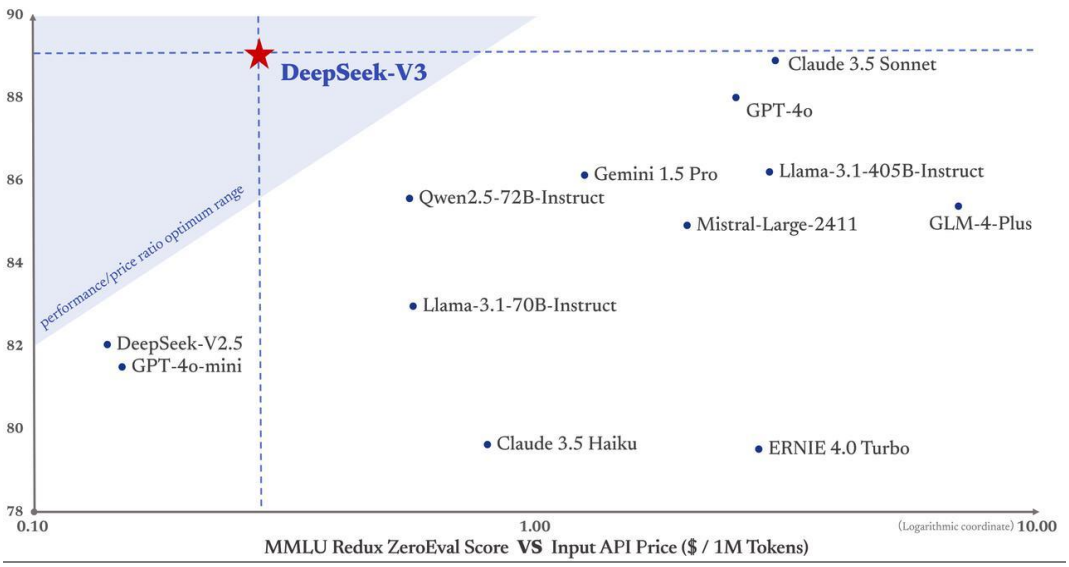
DeepSeek 에 대한 시장의 엇갈린 평가들

DeepSeek 의 성능이 공개된 뒤 여론은 여러 의혹을 제기했다. 1)OpenAI 모델에서 데이터를 무단으로 도용하여 지식 증류(Distillation)를 일으켰을 가능성, 2)투입 비용에 인건비를 누락시켰을 가능성, 3)H100 GPU 를 무단으로 입수했을 가능성 등이다. 무엇이 사실인지는 아직 확인되지 않았지만 DeepSeek 가 공개한 논문 내용이 의미 있는 결론을 도출한 것은 사실이다. 그들은 오픈 소스를 공개하여 논문 전 과정을 누구나 검증 가능하도록 만들었다. '적은 파라미터'로 고성능 추론이 가능해지는 과정을 자신 있게 선보인 것이다.

Dropbox 의 모건 브라운 AI 부사장은 DeepSeek 가 **저비용 고성능 AI 모델을 만들 수 있었던 배경에 대하여 1)메모리 효율화, 2)멀티 토큰(Multi-token) 접근, 3)전문가 시스템 설계** 등을 언급했다. **메모리 효율화**의 경우 32 비트의 메모리를 양자화 하여 8 비트만로 줄이겠다는 의미이다. 메모리를 굳이 32 비트로 표현하지 않고 8 비트만으로 표현해도 결과값이 충분히 정확하다는 주장인 것이다. 결과적으로 DeepSeek 의 메모리 사용량은 75%가 줄었다. **멀티 토큰(Multi-token)** 접근은 기존 AI 가 문장을 한 단어(토큰)씩 인식했다면 DeepSeek 모델의 경우 여러 개의 토큰(멀티 토큰)을 동시에 묶어서 인식하고 처리한다는 것이다. 병렬적으로 문장 전체를 한번에 처리하다 보니 DeepSeek 의 모델은 문맥 파악에 유리하고, 속도 역시 2 배 빨라 졌으며, 정확성은 90% 수준에 도달했다 공개했다. **전문가 시스템 설계(Expert system)** 또한 DeepSeek 가 저비용 고성능을 보이게 만드는데 일조한다. 전문가 시스템 설계는 하나의 거대 AI 가 올케어를 진행하는 것이 아니라, 다양한 전문가(부분 모델)들 중에서 필요한 경우에만 활성화 하겠다는 개념이다. 전문 분야별 소규모 모델들이 상황에 따라 유기적으로 활성화 되다 보니 파라미터 모두를 풀로 돌리지 않아도 된다.

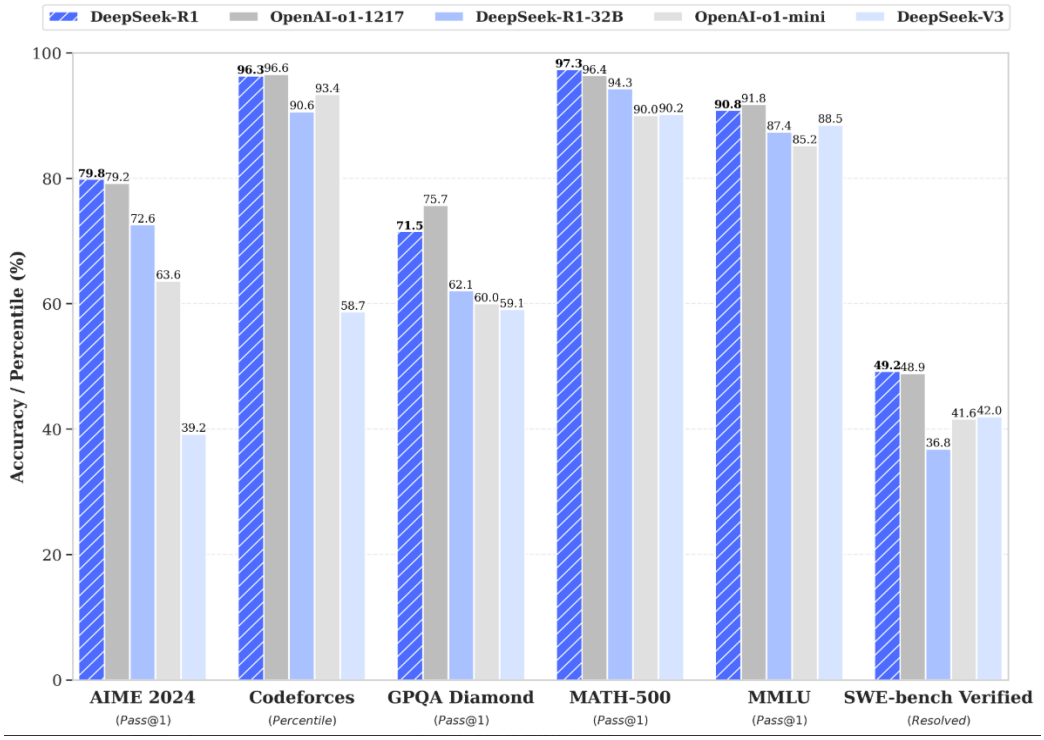
이처럼 DeepSeek 에 대한 시장의 평가는 현재까지 의구심과 새로움으로 뒤섞여 있다. 하지만 공개된 논문 내용을 살펴보면 동사가 받고 있는 의심들이 사실인지 아닌지 여부가 중요해지지 않는다. 미국 AI 기업들에게 경종을 울린 것은 맞으나, 그들의 주도권에 위협을 가할 가능성은 낮게 평가되기 때문이다.

DeepSeek 가 공개한 V3 언어 모델의 효율성



자료: DeepSeek, 키움증권 리서치센터

DeepSeek R1 과 GPT o1 의 부문별 성능 비교



자료: DeepSeek, 키움증권 리서치센터

DeepSeek V3 와 다른 대규모 LLM 모델 과의 성능 비교

Benchmark (Metric)		DeepSeek-V3	Qwen2.5 72B-Inst.	Llama3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513
	Architecture	MoE	Dense	Dense	-	-
	# Activated Params	37B	72B	405B	-	-
	# Total Params	671B	72B	405B	-	-
English	MMLU (EM)	88.5	85.3	**88.6**	88.3	87.2
	MMLU-Redux (EM)	**89.1**	85.6	86.2	88.9	88
	MMLU-Pro (EM)	75.9	71.6	73.3	**78**	72.6
	DROP (3-shot F1)	**91.6**	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	**86.5**	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	**65**	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	**38.2**
	FRAMES (Acc.)	73.3	69.8	70	72.5	**80.5**
	LongBench v2 (Acc.)	**48.7**	39.4	36.1	41	48.1
Code	HumanEval-Mul (Pass@1)	**82.6**	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	**40.5**	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	**37.6**	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	**51.6**	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	**50.8**	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	**84.2**	72.9
	Aider-Polyglot (Acc.)	**49.6**	7.6	5.8	45.3	16
Math	AIME 2024 (Pass@1)	**39.2**	23.3	23.3	16	9.3
	MATH-500 (EM)	**90.2**	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	**43.2**	15.9	6.8	13.1	10.8
Chinese	CLUEWSC (EM)	90.9	**91.4**	84.7	85.4	87.9
	C-Eval (EM)	**86.5**	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	**64.1**	48.4	50.4	51.3	59.3

자료: DeepSeek, 키움증권 리서치센터

DeepSeek 논문에서 확인한 저비용 고성능 모델의 Key

DeepSeek 가 오픈 소스를 공개하며 자신들의 방법에 대한 나름의 자신감을 보인 근거는 그들이 제공한 논문에서 찾을 수 있다. 논문의 구조는 크게 저비용 고성능 모델을 만들어낸 방법과 이에 대한 실질적 증명으로 나뉜다. 구체적으로 살펴보자.

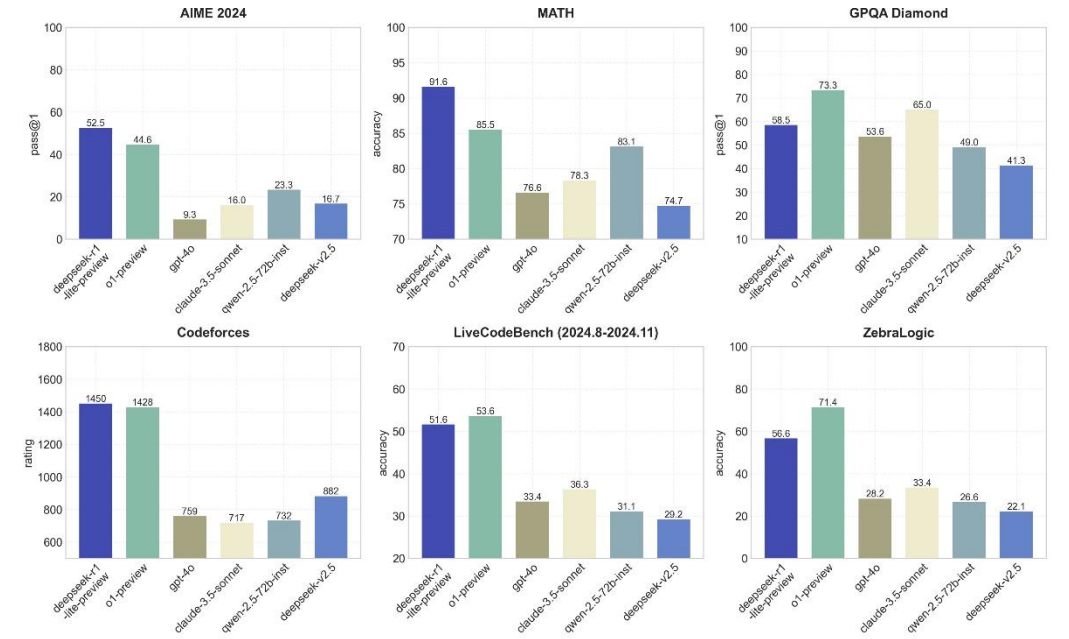
DeepSeek 의 LLM 모델 학습 과정에서 GPT 등의 다른 모델과 다른 부분은 **1)Supervised Fine-Tuning(지도학습 파인튜닝, SFT) 과정을 최대한 생략하고 강화 학습(RL)을 통해 모델을 훈련했다는 것**이다. SFT 는 사람이 직접 라벨링한 정답을 이용해 모델을 훈련하는 과정이다. 사람이 직접 질문-답 라벨을 통해 모델을 학습시키며, 일반적으로 SFT 과정에는 수십만~수백만 규모 데이터가 사용된다. 반대로 강화 학습(RL)의 경우 정답 라벨 대신 보상 함수를 설계하여 모델이 만들어내는 출력물에 따라 보상을 다르게 설정하여 스스로 학습하게 하는 방식이다. 즉 사람이 만들어낸 라벨 없이 모델 스스로 자가 발전한다. 정리해보면, **DeepSeek 는 SFT 과정을 생략해 사람이 만든 수백만 규모의 라벨링 데이터는 사용하지 않고, 모델 스스로 훈련 데이터를 만들어낸 뒤 학습하는 방식으로 모델을 발전**시킨 것이다.

위 도전은 긍정적 결과를 보여주었다. 강화 학습(RL)만을 통해 훈련시킨 DeepSeek-R1-Zero 모델이 뛰어난 추론 능력을 보인 것이다. 다만, 문장 가독성이 떨어졌고, 다양한 언어를 혼용하는 이슈가 있었다. 이를 보완하기 위해 소량의 Cold-start data 와 multi-stage training 을 결합한 파이프라인을 적용하였다. Cold-start data 는 초기 소량의 지도 데이터이다. 첫 모델을 키우기 위해 기초적 지식을 소량 유입시키는 것이다. 그 이후 진행되는 multi-stage training 은 강화 학습(RL)과 Cold-start data 학습, 작은 규모의 SFT 의 순환을 통해 단계적으로 모델을 발전시키는 것이다. 이러한 방법을 통해 만든 모델이 DeepSeek-R1 이다. **위 기술적 과정이 시사하는 바를 한 문장으로 정리하면 다른 대규모 LLM 모델과 달리 DeepSeek 는 상대적으로 적은 데이터를 활용하여 학습을 진행한다는 것이다.** 그리고 이에 대한 시장의 평가는 긍정적인 상황이다. 새로운 접근을 통해 효율적 AI 모델 학습 과정을 만들어냈기 때문이다.

DeepSeek 가 공개한 또다른 접근은 지식 증류(Distillation) 방식이다. 지식 증류를 쉽게 설명하면 선생님 모델(큰 모델)의 출력을 학생 모델(작은 모델)이 최대한 비슷하게 따라하도록 학습하는 방법을 말한다. 학생 모델의 경우 파라미터가 적기 때문에 대규모 학습을 하기 부족하다. 이에 선생님 모델이 미리 학습한 복잡한 패턴이나 추론 방식을 흉내 내며 학습한다. 이 경우 적은 연산 과정만으로 높은 성능에 근접할 수 있다. 물론 지식 증류 방식이 이전에 없던 새로운 방법론은 아니지만, 기존에는 선생 모델이 '정답'만을 제공했다면 **DeepSeek 는 선생 모델이 '중간 추론 과정'까지 가시적으로 제공하여 학생 모델에게 정답 도달 과정을 함께 공유했다.** 이 결과 학생 모델은 파라미터가 적음에도 불구하고 높은 수준의 추론 결과를 제공할 수 있게 되었다.

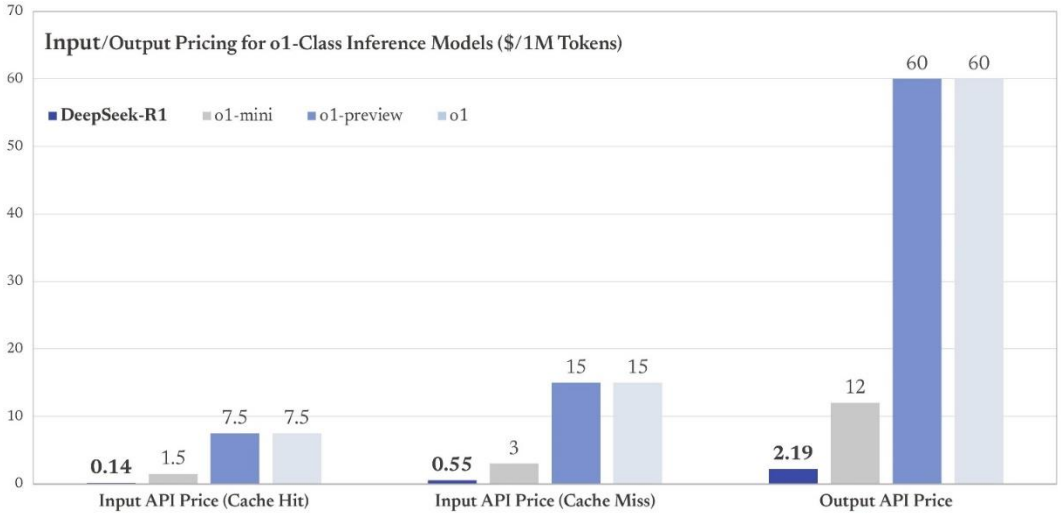
위 내용을 종합해 보면, 강화 학습(RL) 중심의 모델 학습과 중간 추론 과정을 공유하는 지식 증류(Distillation) 방식이 DeepSeek 가 적은 비용으로 고성능 모델을 만들어 낸 키이다. 여기에 더해 앞서 **전문가 시스템 설계(Expert system)로 표현된 MoE(Mixture of Experts) 아키텍처 역시 컴퓨팅 비용을 낮추고 고성능 모델을 만드는데 일조**했다.

DeepSeek LLM 모델과 다른 대규모 LLM 모델 과의 성능 비교



자료: DeepSeek, 키움증권 리서치센터

DeepSeek R1 모델과 GPT o1 모델의 가성비 비교

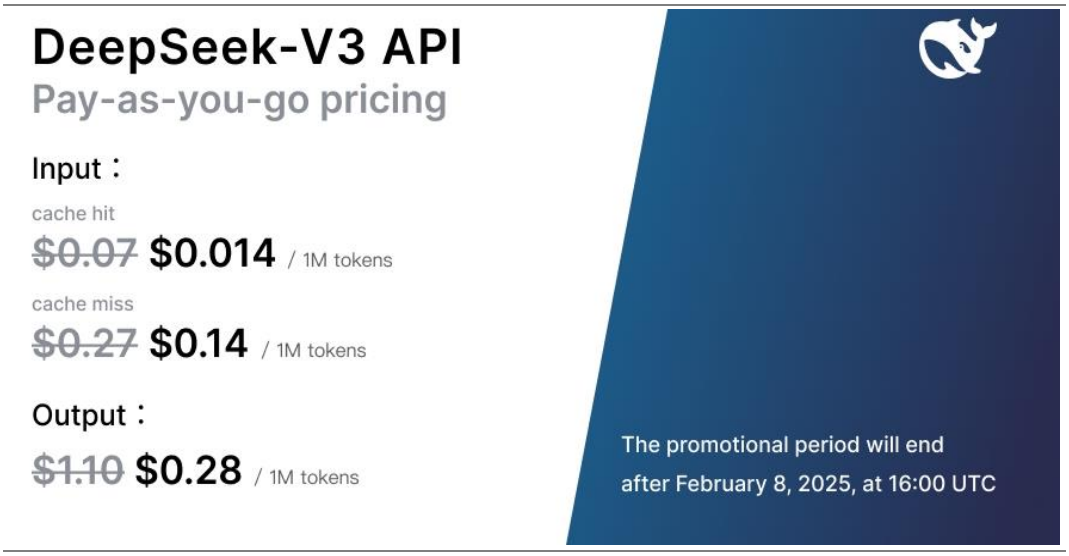


자료: DeepSeek, 키움증권 리서치센터

DeepSeek R1 모델의 Input API 가격

자료: DeepSeek, 키움증권 리서치센터

DeepSeek V3 모델의 Input API 가격



자료: DeepSeek, 키움증권 리서치센터

그럼에도 DeepSeek 가 AI 판도를 바꿀 수 없는 이유

DeepSeek 가 미국 주도 AI 시장 판도를 바꾸기 위해서는 기존 레거시인 Scaling 규칙(더 많은 자원을 투입해야 성능이 좋아진다)이 틀렸음을 보여야 한다. 위 논문 내용에서 확인된 것은 이것이 가능하다 증명하는 듯 하지만 근본적 한계점이 있다. 모든 존재는 태초의 Being 에서 비롯되는 것과 같이 모든 최초 모델은 대규모 랜덤 데이터를 학습하는 과정을 필연적으로 거쳐야 하기 때문이다. 위 논문의 내용을 다시 한번 살펴보자.

우선 **강화 학습(RL) 중심의 모델 훈련에는 대규모 사전 학습(Pre-training)이 전제되어 있다**. DeepSeek 는 6,710 억 개의 파라미터를 보유한 DeepSeek-V3-Base 모델을 밝혔다. 대규모 태초 모델인 것이다. 물론 DeepSeek 의 전문가 시스템 설계를 통해 AI 연산 과정에서 활성화 되는 파라미터는 370 억개 정도밖에 안된다고 주장했지만, 중요한 것은 거대 Base 모델이 이미 존재해야 한다는 것이다. 이것 위에서 강화 학습(RL) 방법이나, 전문가 시스템 설계 등이 적용될 수 있다. **Base 모델 없이 강화 학습(RL)으로만 모델을 만들고 훈련을 진행할 경우 성능 구현이 어렵고 오히려 비용이 커질 수 있다.**

또한, **강화 학습(RL)의 핵심인 보상 함수 설계 역시 환경이 변화함에 따라 꾸준히 파라미터 업데이트가 필요할 것**이다. 이것이 진행되지 않을 경우 범용성이 떨어지는 모델이 될 수 있다. 아무리 효율적인 방법으로 훈련한다 해도 적은 데이터 학습만을 고집할 경우 '범용성'과 '안전성'에 문제가 발생할 수 있다. DeepSeek 모델은 명확한 답이 있는 수학, 코딩 등은 잘할 수 있지만 변수가 많은 '현실 세계'에는 적용하기 어려운 모델이 될 수 있는 것이다. 지식 증류(Distillation) 방식 역시도 비슷한 문제를 내포한다. **학생 모델의 학습 벤치마크인 선생님 모델을 처음으로 만드는 경우에는 역시 대규모 파라미터를 갖춘 모델이 필요하다**. 선생님 모델이 제대로 된 추론 과정을 공유하지 못할 경우 지식 증류는 처음부터 발생할 수 없다. DeepSeek 가 받고 있는 의혹 중 하나인 OpenAI 에서 선생님 모델 데이터를 무단으로 가져왔다는 것 역시 DeepSeek 도 결국엔 대형 파라미터 기반의 모델이 필요하고, 그들이 공개한 방법론으로는 이걸 해결하지 못했음을 시사한다.

GPT 모델은 여러가지 형태의 데이터를 처리하는데 특화되어 있다. 최초 GPT 모델은 텍스트에 특화되어 있었지만, 아키텍처가 확장됨에 따라 이미지, 오디오, 비디오 등 다양한 유형의 데이터를 처리할 수 있게 발전했다. 멀티 모달 LLM 의 경우 헬스케어 부문에서 의료 영상을 분석할 수 있고, 자율주행 시스템 속 센서 데이터를 담당할 수 있다. 또한 음성 데이터를 통해 보안 시스템을 강화할 수 있다. AI 모델이 수많은 랜덤 데이터를 처리, 분석할 수 있어야 여러 산업군에 적용할 수 있다는 의미이며, **많은 랜덤 데이터 처리를 위해서는 많은 파라미터를 보유할 수 밖에 없다. 이 지점에서 DeepSeeek 가 아직은 AI 시장 질서를 뒤바꾸기에 한계가 있다고 보는 것이다.** 제한된 환경 속에서 텍스트 중심의 저비용 고성능 학습 모델 로직을 이끌어낸 DeepSeek 의 기술은 아직 적용 분야가 넓지 않다.

산업별 멀티 모달 AI 활용 사례

분야	응용 사례
헬스케어	의료 영상(CT, MRI 등)과 환자의 텍스트 기록(진단서 등)을 함께 분석하여 정확한 질병 진단을 지원
자율주행	카메라 영상, 라이다(LiDAR) 센서 데이터, GPS 정보 등을 통합하여 자율주행 차량의 정확한 환경 인식을 지원
전자상거래	제품 이미지와 사용자의 텍스트 리뷰를 동시에 분석하여 맞춤형 제품 추천을 제공
미디어	동영상에서 장면 별 주요 정보를 추출하거나 자막을 자동 생성하여 콘텐츠 소비 경험을 향상
보안 및 감시	영상과 음성 데이터를 통합하여 특정 이벤트(예: 침입자 경고)를 실시간으로 탐지하고 경고

자료: 키움증권 리서치센터

DeepSeek 는 오히려 미국 AI 산업에 긍정적일 수 있다?

위에서 서술했듯 DeepSeek 가 AI 산업 주류를 개편하기에는 역부족인 만큼, 엔비디아의 GPU 을 통한 지배력과 OpenAI 의 LLM 모델 확장성은 유지될 것으로 판단한다.
해당 관점에서 관련 주가들의 디레이팅 역시 과하다고 판단한다. 높아진 밸류에이션 부담을 털어내는 트리거로써 DeepSeek 이슈가 소화되었을 뿐, AI 주류 개편에 대한 시장의 우려가 증시 하락 원인이 아니다. 이에 AI 관련 기업들의 주가는 점차 DeepSeek 등장 이전 레벨로 복귀할 것으로 예상하며, 장기적 관점에서도 상승 여력이 높다고 판단한다.

DeepSeek 의 영향력은 오히려 긍정적일 수 있다. **수혜를 입을 것으로 예상되는 산업군은 엣지 AI 와 온디바이스(on-device) AI 중심의 산업군**이다. DeepSeek 의 논문 내용을 복기해 보면, 대규모 모델이 여전히 필요하다는 한계점은 있지만, DeepSeek 는 SFT 를 생략한 강화 학습(RL) 훈련, 지식 증류(Distillation), MoE(Mixture of Experts) 등의 방법을 통해 **'소형 모델의 효과적 학습'**을 이끌어냈다. 파라미터가 적어서 크기는 작지만, 효과적 훈련 방법을 통해 높은 성능을 낼 수 있게끔 환경을 조성한 것이다. DeepSeek 기술팀이 AI 전문가들에게 인정을 받고 있는 이유다.

엣지 AI 는 중앙 데이터센터(클라우드)에서가 아니라 실제 데이터가 만들어지는 현장 근처에서 연산 및 추론을 수행하는 것을 의미한다. 물론, 현장 데이터 근처에서 작업을 수행해야 하기에 **규모가 작은 소형 모델을 통해 추론 및 연산을 진행**해야 한다. IoT 가 점차 발전하고, 산업 현장에 AI 에 대한 요구가 커지면서 엣지 AI 에 대한 수요 역시 빠르게 높아지고 있다. 올해 CES2025 에서 공개된 구체적 사례를 보면 농업(자율주행 트랙터), 산업(스마트 공장), 해양 기술(자율주행 선박) 등 여러 부문에서 엣지 AI 에 대한 필요성이 높아지고 있다. 해당 부분에서 DeepSeek 의 소형 모델 학습 기술이 접목될 수 있다. **1)양자화를 통해 모델 규모를 더욱 작게 만들어 배치의 편리성을 높이고**, **2)MoE 를 통해 소형 모델의 메모리 사용량을 줄이며**, **3)지식 증류(Distillation)를 통해 소규모 모델의 학습 효과를 높인다**. 또한 **4)대규모 강화 학습(RL) 중심 훈련을 통해 소형 모델의 효과적 훈련을 가능케 한다**. 소형 모델 구축에 들어가는 비용은 줄고 성능은 좋아질 수 있다 보니 엣지 AI 산업 역시 긍정적 혜택을 볼 수 있다.

온디바이스 AI 역시 비슷한 논리로 수혜를 볼 수 있다. 스마트폰, 태플릿, 웨어러블 기기 등 개별 디바이스에 AI 를 직접 내장하는 추이가 확장되고 있으며, 소형 모델에 대한 필요성은 높아지고 있다. 또한 온디바이스 AI 는 환경적, 보안적 이유로 네트워크 연결이 없어도 자체 연산이 가능해야 한다. 즉 탑재되는 소형 AI 모델의 성능 자체가 높은 수준이어야 한다는 의미이다. DeepSeek 의 소형화 모델은 해당 부분을 보조한다. **양자화와 MoE 를 통해 디바이스를 작게 만들거나 모델의 배터리 사용량을 줄일 수 있고, 지식 증류(Distillation)를 통해 온디바이스 AI 의 성능을 높일 수 있다.** 엣지 AI 산업과 온디바이스 AI 산업이 큰 틀에서 '소형 AI 모델'과 맞닿아 있기에 비슷한 논리로 긍정적 혜택을 받는 것이다. 물론 DeepSeek 기술은 아직 텍스트 중심의 모델 학습에 특화되어 있어 활용 범위가 일부 제한적일 수 있지만, 효과적인 소형 모델 학습에 큰 도약을 만든 것은 사실이므로 엣지 AI 및 온디바이스 AI 기여도가 점차 높아질 것으로 판단한다.

장기적 관점의 DeepSeek 發 AI 산업 영향

마이크로소프트 CEO 인 Satya Narayana Nadella 는 DeepSeek 의 성공에 대하여 **Jevon 의 역설(Jevons Paradox)**과 연관하여 설명했다. Jevon 의 역설은 기술 발전에 따라 자원의 사용 효율성이 높아질 경우, 해당 자원의 총 사용량이 오히려 증가하는 현상을 말한다. '효율성 상승 → 수요 증가 → 총 소비량 증가' 과정이 발생하는 것이다. DeepSeek 가 공개한 기술은 AI 모델의 학습 효율화 부문의 발전을 만들었다. 산업계 전반의 변곡점이 될 정도는 아니지만 소형 AI 모델 비용 절감 가능성은 중소형 AI 기업들 역시도 경쟁에 뛰어들 수 있는 발판이 될 수 있다.
또한 DeepSeek 자체도 약 200 명의 직원으로 구성된 작은 규모의 회사인 만큼 AI 중소형 기업들에게 하나의 벤치마크 모델이 될 수 있다. 하드웨어 중심의 시장에서 소프트웨어 중심의 AI 시장으로 트렌드가 넘어가고 있는 현재, 기업들의 참여가 많아지고 AI 수요가 활성화 될 경우 **Jevons 의 역설과 같이 AI 산업 전체에 대한 수요가 오히려 높아질 가능성이 있다고 판단한다**.

엣지 AI 와 온디바이스 AI 의 수혜 부문 정리

구분	주요 산업 / 활용 사례	소형 모델 도입 수혜	기대 효과
엣지 AI	- 스마트 공장(제조 라인 불량 검출) - 자율주행(차량 내 실시간 판단) - 산업용 IoT(설비 예지보전) - 스마트시티(교통/치안 모니터링)	- **낮은 지연**: 데이터 센터 왕복 없이 현장에서 연산 - **비용 절감**: 대용량 영상·센서 데이터 업로드 최소화 - **전력/자원 효율**: 소형 모델로 엣지 디바이스에서도 무리 없는 추론	- 라인 정지/사고 방지 등 실시간 대응 가능 - 네트워크 장애에도 독립적으로 작동 - 비용 절감 및 고장 사전예방 등 생산성 향상
온디바이스 AI	- 소비자 가전(스마트 스피커, TV) - 스마트폰(사진 분류, AR, OCR) - 웨어러블(건강 모니터링) - 로봇청소기·드론 등 가정/개인용 기기	- **오프라인 동작**: 인터넷 연결 없이도 AI 기능 수행 - **프라이버시 강화**: 민감/개인 데이터 디바이스 내에서만 처리 - **저전력**: 모델 크기가 작아 모바일, 웨어러블 배터리 소모 최소화	- 사용자 경험 개선(즉각적 응답) - 인터넷 미연결 환경에서 지속적 AI 서비스 가능 - 개인 데이터 유출 리스크 감소 + 보안성 확보

자료: 키움증권 리서치센터

엣지 AI 의 세부 산업과 주요 종목

분야	세부 산업	주요 기업 및 기술
엣지 AI	1) 스마트 공장/제조업	- **Siemens**: 산업 자동화, 디지털 트윈 솔루션 제공
		- **Bosch**: IoT 센서 및 생산 라인 자동화
		- **ABB**: 로봇팔 및 자동화 시스템
		- **GE(General Electric)**: 산업 IoT 플랫폼 Predix
		- **Rockwell Automation**: 공장 제어 시스템 및 산업 소프트웨어
	2) 자율주행·자동차	- **Tesla**: 차량 내 FSD 컴퓨터 기반 자율주행
		- **Waymo**: 자율주행 전용 하드웨어 및 SW 스택
		- **Cruise**: 자율주행 택시 및 배달 서비스
		- **Mobileye**: 차량용 카메라, 라이다, 레이다 통합 인식 솔루션
	3) 산업용 IoT/에너지	- **Schneider Electric**: 에너지 관리 및 자동화 솔루션
		- **Hitachi**: IoT 플랫폼 Lumada
		- **Honeywell**: 제조·건축·항공 IoT 솔루션
		- **Siemens Energy**: 전력망 자동화 제어 시스템
	4) 스마트시티/공공 인프라	- **Cisco**: 네트워크 장비 및 엣지 데이터 처리
		- **Huawei**: 5G 인프라, 도시 영상분석 AI 솔루션
		- **NEC**: 교통, 치안 자동화 시스템
		- **LG CNS**: 스마트시티 통합 플랫폼 및 교통 관제
	5) 로컬 엣지 하드웨어	- **NVIDIA Jetson**: 임베디드 GPU 모듈
		- **Intel Movidius**: 저전력 비전 프로세서
		- **Qualcomm(Snapdragon)**: IoT 디바이스용 SoC
		- **ARM**: 엣지 디바이스용 저전력 CPU 아키텍처

자료: 키움증권 리서치센터

온디바이스 AI 의 세부 산업과 주요 종목

분야	세부 산업	주요 기업 및 기술
온디바이스 AI	1) 스마트폰/모바일	- **Apple**: A 시리즈·M 시리즈 칩에 뉴럴 엔진 내장
		- **Samsung**: 엑시노스 NPU 기반 Bixby 음성인식
		- **Google**: Pixel 텐서 칩으로 온디바이스 번역·사진 분류
		- **Xiaomi, OPPO, Vivo**: 자체 AI 칩, 카메라 최적화 등
	2) 웨어러블·헬스케어	- **Apple Watch**: 심전도(ECG) 데이터 처리
		- **Fitbit(구글)**: 심박수 및 수면 리듬 분석
		- **Garmin**: 스포츠 웨어러블로 산소포화도 모니터링
		- **Samsung Galaxy Watch**: 혈압, ECG 온디바이스 처리
	3) 스마트홈/소비자 가전	- **Amazon Echo**: 오프라인 음성 명령 인식
		- **Google Nest**: 온디바이스 음성 감지
		- **LG 전자**: ThinQ 플랫폼으로 가전기기에 AI 내장
		- **Sony**: 스마트 TV 및 이어폰에 NPU 기반 음성 처리
	4) 로봇·드론·가정용 AI	- **iRobot**: 로봇청소기 온디바이스 맵핑 및 장애물 인식
		- **DJI**: 드론의 실시간 장애물 회피
		- **Boston Dynamics**: Spot 로봇 현장 데이터 로컬 프로세싱(클라우드 연동도 일부 사용)

자료: 키움증권 리서치센터

Compliance Notice

- 당사는 동 자료를 기관투자자 또는 제 3 자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

고지사항