2025-02-06

반도체

AI 투자 명분 강화 국면

SK증권 리서치센터

Analyst
한동희
donghee.han@sks.co.kr
3773-8826

R.A
박제민
jeminwa@sks.co.kr
3777-8884

빅테크들의 2025F CapEx 가이던스 예상 상회 중

최근 실적 발표를 통해 빅테크들이 제시한 FY 2025 CapEx 가이던스는 시장 예상치를 상회 (알파벳 28%, 메타 23%)하고 있다. 또한 대만 Digitimes 에 따르면 TSMC 의 CoWoS 생산 능력 계획 역시 상향 (2025 년 말 65K-> 75K, 2026 년 말 80K->95K)되고 있으며, 2028 년 말까지 150K 에 이를 것으로 알려졌다. 불투명한 거시경제 상황, 지속 상승하고 있는 매출액 대비 CapEx 비중에 대한 불안, DeepSeek 영향에 따른 투자 효율화에 대한 시장의 우려와 상반되는 현상이다.

AI 성능 향상 논리의 확대는 Scaling law 의 지속을 의미

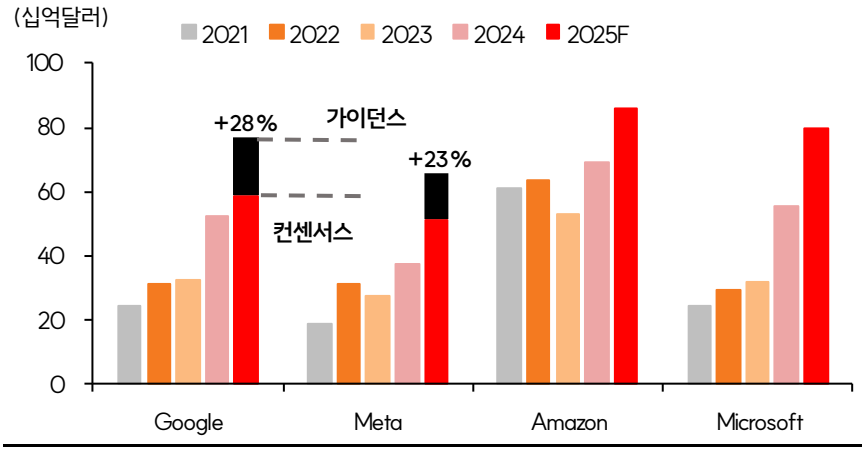
SK 증권은 알고리즘을 통한 AI 성능 향상의 논리가 오히려 Computing power 의 요구량 확대로 이어질 것이라는 기존 의견을 유지한다. AI 투자 지속 및 경쟁 심화의 명분이 결국 AI 성능 제고의 한계점에 도달하지 않았음을 기반으로 한다는 점을 감안하면, Hardware 에만 의존하던 기존의 구조 대비 Scaling law 가 더 가시적으로 지속될 수 있다는 의미로 해석 가능하다. DeepSeek-R1 논문에 따르면, AI 성능 향상을 위해서는 높은 Computing power 와 강력한 기반 모델, 거대한 규모의 강화학습을 요구한다는 점도 이를 뒷받침한다고 판단한다.

추론의 정확도 향상을 위한 Test-time scaling 역시 더 높은 Computing power 를 요구하는 개념이 될 것으로 전망한다. 정확한 답변을 위해 AI 가 고민할 수 있는 시간을 부여한다는 개념은 해당 시간 동안 더 많은 연산을 수행한다는 의미이며, 향후 지속 될 복합적 추론에 대한 수요 확대는 짧은 시간 동안 많은 연산을 감당할 수 있는 Computing power 를 요구할 것으로 판단하기 때문이다. Scaling law 가 지속되는 국면에서 AI 시장 선점 및 경쟁력 강화를 위한 투자는 당위성을 갖는다.

AI 사이클이 거시 경제 사이클보다 상위 개념

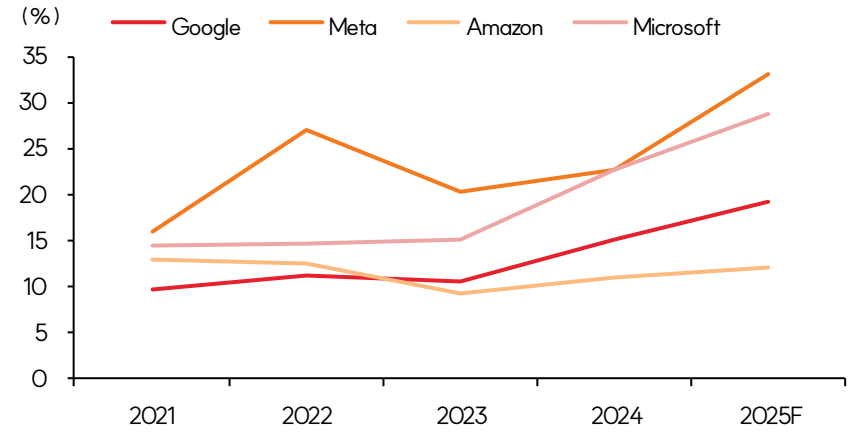
SK 증권은 AI 사이클이 거시 경제 사이클보다 더 상위 개념이라 판단한다. 우리는 거시 경제 부진 속에서도 AI 노출도가 높은 기업들의 실적 패턴은 과거와 매우 상이하다는 것을 TSMC 와 SK 하이닉스 등의 AI 를 기반으로 한 안정적 실적과 주가 연동성 등을 통해 확인했다. 효율, 효과적 AI 성능 제고에 따른 AI 투자 명분이 유지되는 한 거시 경제 부진이 AI에 대한 투자를 크게 훼손할 가능성은 제한적일 것으로 전망하며, 2Q25 부터 Commodity 재고 조정 안정화 및 2026 년 HBM 가시성 제고를 고려한 투자 전략을 권고한다. 1Q25 비수기 및 거시 경제 불확실성에 따른 주가 하락은 기회가 될 것이다. 반도체 업종에 대해 비중확대 의견을 유지한다.

북미 CSP 들의 2025 년 설비투자 가이던스 예상 상회 중



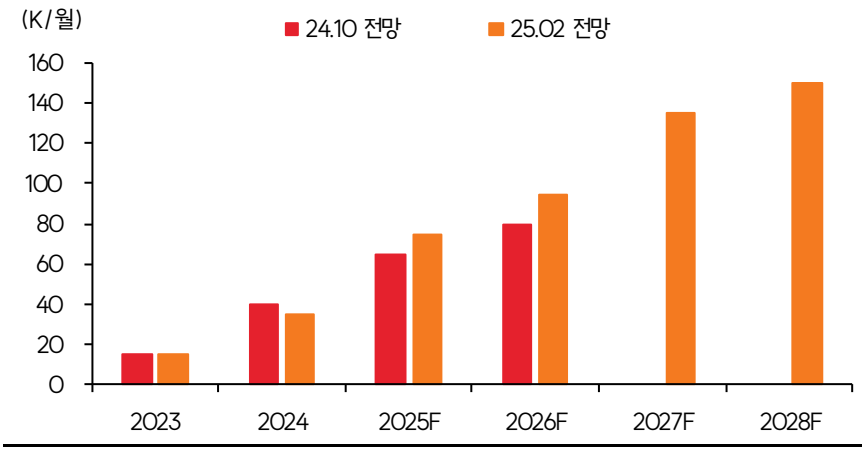
자료: Bloomberg, SK 증권

북미 CSP 들의 2025 년 매출액 대비 설비투자 비중 추이 및 전망



자료: Bloomberg, SK 증권
주: Amazon 은 실적 발표 전

TSMC 의 CoWoS 생산 능력 추이 및 전망



자료: Digitimes, SK 증권

Humanity's Last Exam 기준 AI 모델별 평가 (Accuracy, Calibration Error)

Humanity's Last Exam

Model	Accuracy (%)
GPT-4o	**3.3%**
Grok-2	3.8%
Claude 3.5 Sonnet	4.3%
Gemini Thinking	6.2%
OpenAI o1	**9.1%**
DeepSeek-R1*	9.4%
OpenAI o3-mini (medium)*	10.5%
OpenAI o3-mini (high)*	13.0%
OpenAI Deep Research pass@1**	**26.6%**

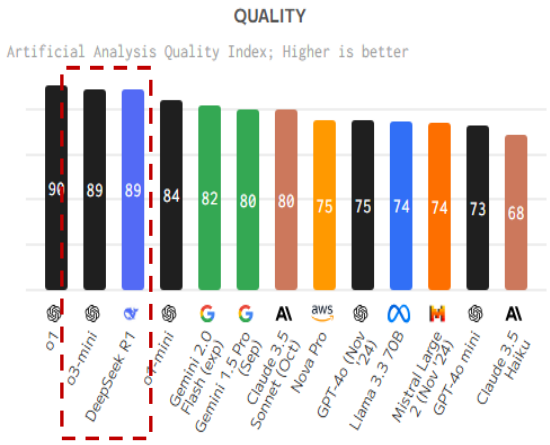
*Model is not multi-modal, evaluated on text-only subset.
**with browsing + python tools

Model	Accuracy (%) ↑	Calibration Error (%) ↓
GPT-4o	3.3	92.5
Grok-2	3.8	93.2
Claude 3.5 Sonnet	4.3	88.9
Gemini Thinking	7.7	91.2
o1	9.1	93.4
DeepSeek-R1*	9.4	81.8
o3-mini (medium)*	10.5	92.0
o3-mini (high)*	13.0	93.2

*Model is not multi-modal, evaluated on text-only subset.

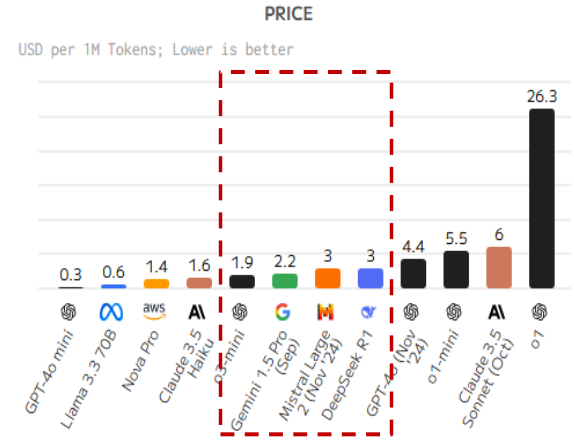
자료: Humanity's Last Exam, SK 증권

Artificial Analysis Index (Quality)



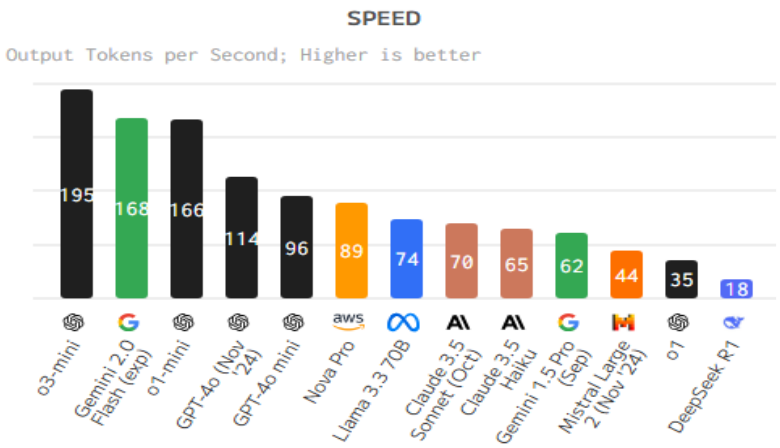
자료: Artificial Analysis, SK 증권

Artificial Analysis Index (Price)



자료: Artificial Analysis, SK 증권

Artificial Analysis Index (Speed)



자료: Artificial Analysis, SK 증권

DeepSeek-V3 훈련 비용

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Lastly, we emphasize again the economical training costs of DeepSeek-V3, summarized in Table 1, achieved through our optimized co-design of algorithms, frameworks, and hardware. During the pre-training stage, training DeepSeek-V3 on each trillion tokens requires only 180K H800 GPU hours, i.e., 3.7 days on our cluster with 2048 H800 GPUs. Consequently, our pre-training stage is completed in less than two months and costs 2664K GPU hours. Combined with 119K GPU hours for the context length extension and 5K GPU hours for post-training, DeepSeek-V3 costs only 2.788M GPU hours for its full training. Assuming the rental price of the H800 GPU is $2 per GPU hour, our total training costs amount to only $5.576M. Note that the aforementioned costs include only the official training of DeepSeek-V3, excluding the costs associated with prior research and ablation experiments on architectures, algorithms, or data.

자료: DeepSeek-V3, SK 증권

DeepSeek, 높은 컴퓨팅 파워와 강력한 거대 기반 모델에 대한 필요성 언급

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2

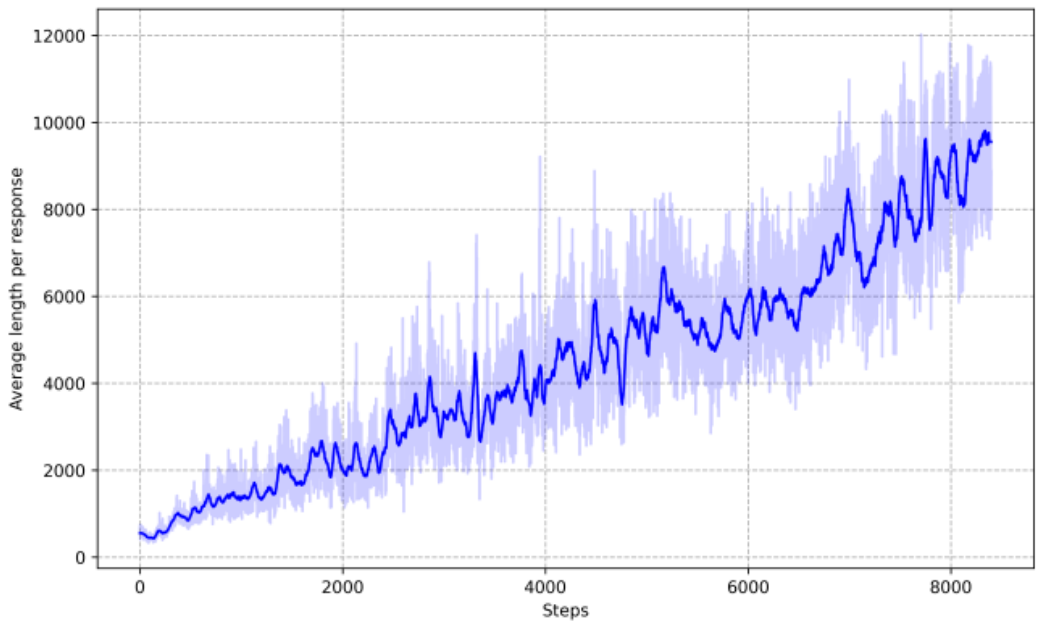
Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

RL training, achieves performance on par with QwQ-32B-Preview. However, DeepSeek-R1-Distill-Qwen-32B, which is distilled from DeepSeek-R1, performs significantly better than DeepSeek-R1-Zero-Qwen-32B across all benchmarks.

Therefore, we can draw two conclusions: First, distilling more powerful models into smaller ones yields excellent results, whereas smaller models relying on the large-scale RL mentioned in this paper require enormous computational power and may not even achieve the performance of distillation. Second, while distillation strategies are both economical and effective, advancing beyond the boundaries of intelligence may still require more powerful base models and larger-scale reinforcement learning.

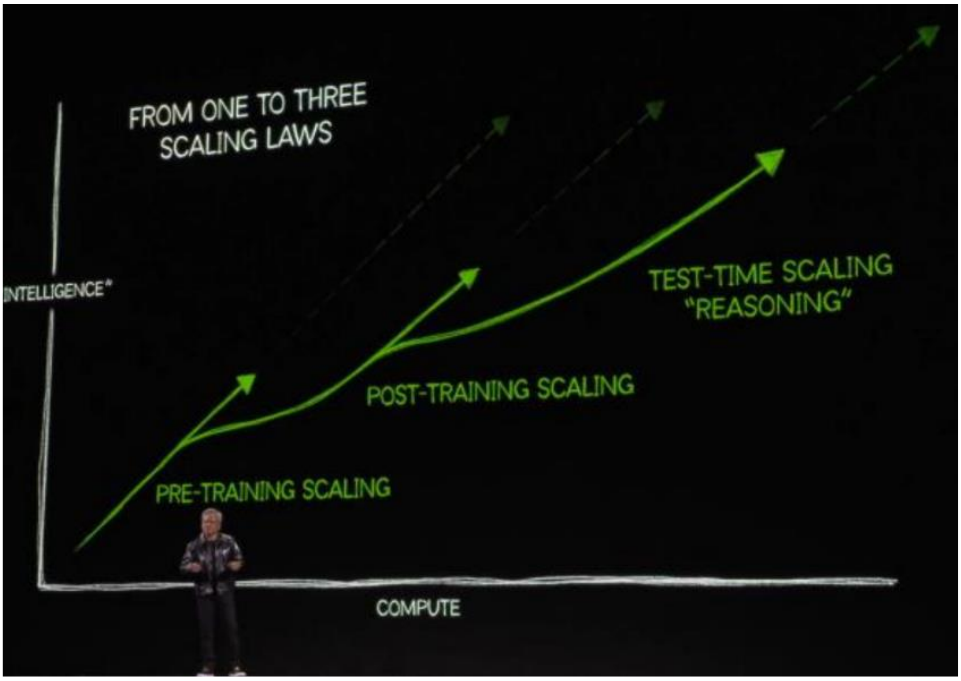
자료: DeepSeek-R1, SK 증권

DeepSeek-R1-Zero: 더 많은 사고 시간 할애를 통한 자연적인 추론 업무 학습



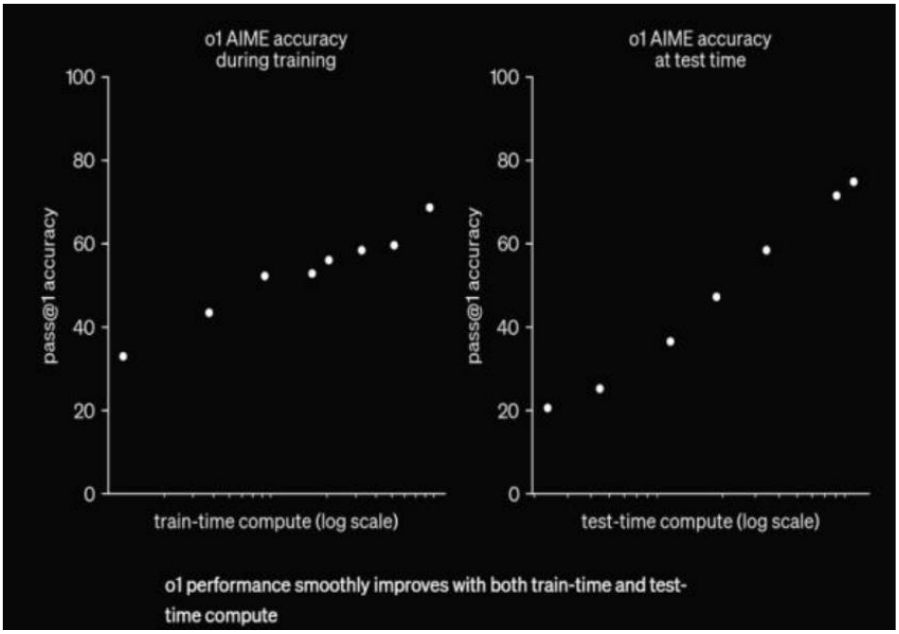
자료: DeepSeek-R1-Zero, SK 증권

추론에서의 Test-Time Scaling 을 통한 Scaling law 지속



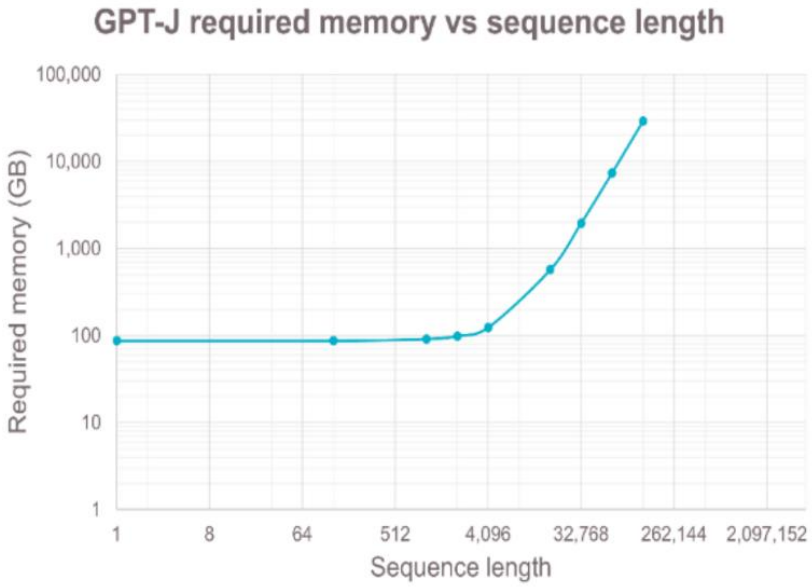
자료: CES2025 NVIDIA Keynote, SK 증권

Test time (Computing time)을 늘릴수록 모델 정확도 증가



자료: OpenAI, SK 증권

Sequence length 의 증가는 요구 메모리의 증가를 견인



자료: Cerabras Inference. SK 증권

Compliance Notice

작성자(관리자)는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.
본 보고서에 언급된 종목의 경우 당사 조사분석담당자는 본인의 담당종목을 보유하고 있지 않습니다.
본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.
당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.
종목별 투자의견은 다음과 같습니다.
투자판단 3 단계(6 개월기준) 15%이상 -> 매수 / -15%~15% -> 중립 / -15%미만 -> 매도

SK 증권 유니버스 투자등급 비율 (2025 년 02 월 06 일 기준)

매수	97.45%	중립	2.55%	매도	0.00%