중국은 지금

Deep seek發 스푸트니크 쇼크?

키움증권 리서치센터 글로벌리서치팀
I China Strategist 홍록기 hongluckiee@kiwoom.com
I RA 조호준 hojun.cho@kiwoom.com

SUMMARY

최근 DeepSeek V3, R1모델이 연속으로 출시되었다. 미국의 기술 제재에도 불구, 소프트웨어 혁신을 통해, 사양이 낮은 칩으로도 저비용고효율의 AI 모델을 구축하는데 성공하면서, 시장이 이목이 집중되고 있다. 미국의 제재를 극복한 사례로써, 중국 첨단산업(특히 소프트웨어) 기업 성장성에 대한 의구심을 일부 해소하는데 기여할 것으로 기대된다. 그러나 중장기 관점에서 불안한 대외환경 등을 감안하면, 중국 반도체 국산화를 위한 하드웨어에 대한 투자는 계속 늘어날 수 밖에 없는 환경이 계속될 것으로 판단한다.

DeepSeek, 절반의 성공

중국 소프트웨어 기업에 대한 새로운 인식 기대. 중장기적으로는 여전히 하드웨어 투자 증가 환경이 계속될 수 밖에 없을 듯

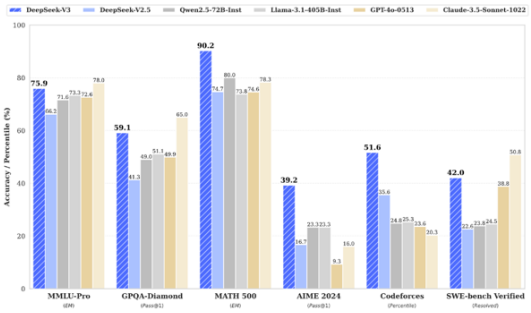
중국 AI 스타트 기업 딥시크는(Deep seek) 인공 지능 일반화를(AGI) 목표로 기초 기술 개발에 집중하고 있었다. 2023년 항저우에 회사를 설립한 이후, DeepSeek-LLM, DeepSeek-MoE 등 여러 모델을 발표했다. 특히 2024년 12월에 DeepSeek-V3 모델 발표 이후, 1/20일 대규모 언어모델 R1까지 효율성을 극도로 높여서, 서구권의 LLM 모델과 경쟁할 수 있는 수준의 성능을 공개하면서, 시장 이목이 집중되었다.

G2 갈등 국면에서 특히 AI, 반도체 핵심 산업에 대하여 미국의 제재가 강력하게 유지되고 있었던 만큼, 중국 AI 기업의 약진에 미국 AI 기업뿐만 아니라 시장 또한 당황스러운 모습을 보였다. 일각에서는 AI 산업의 G2갈등에서 스푸트니크 쇼크를 연상케 한다는 언급도 있었다.

미국의 Tech 기업들은 고성능 AI 모델을 구축하고 위해 엔비디아의 첨단 칩을 최대한 많이 확보하여 학습해야한다는 인식이 팽배했었다. 또한 이러한 인식 아래, 중국 AI 산업을 견제하기 위한 수단으로써, 고성능 칩을 중국으로 수출하는 것을 엄격하게 제한해왔다.

그러나, DeepSeek는 상대적으로 성능이 낮은 엔비디아 H800 칩만 사용, 소프트웨어의 혁신을 통해 Open AI와 같이 수준 높은 LLM 모델을 훨씬 낮은 비용으로 구현함으로써, 시장의 충격을 주었다.

DeepSeek V3 모델 및 여타 AI 모델과 성능 비교



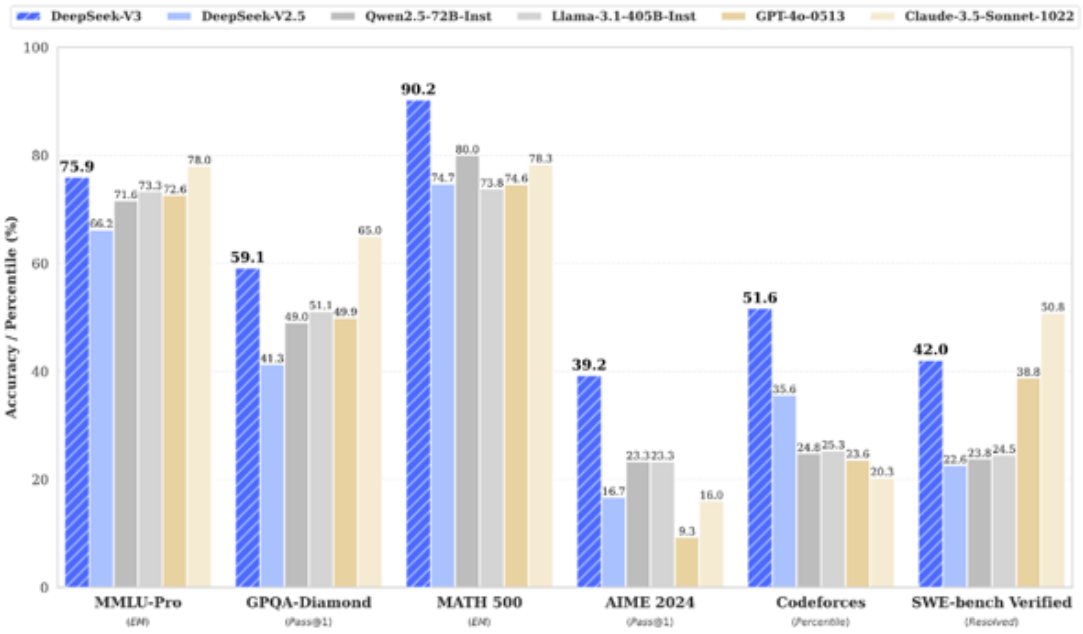
자료: 국가통계국, Bloomberg, 키움증권 리서치

DeepSeek V3 모델 개발 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

자료: DeepSeek, 키움증권 리서치센터

DeepSeek V3 모델 및 여타 AI 모델과 성능 비교



자료: DeepSeek, 키움증권 리서치센터

DeepSeek V3 모델 및 여타 AI 모델과 성능 비교

	Benchmark (Metric)	DeepSeek V3	DeepSeek V2.5 0905	Qwen2.5 72B-Inst	Llama3.1 405B-Inst	Claude-3.5 Sonnet-1022	GPT-4o 0513
	Architecture	MoE	MoE	Dense	Dense	-	-
	# Activated Params	37B	21B	72B	405B	-	-
	# Total Params	671B	236B	72B	405B	-	-
English	MMLU (EM)	88.5	80.6	85.3	**88.6**	88.3	87.2
	MMLU-Redux (EM)	**89.1**	80.3	85.6	86.2	88.9	88.0
	MMLU-Pro (EM)	75.9	66.2	71.6	73.3	**78.0**	72.6
	DROP (3-shot F1)	**91.6**	87.8	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	80.6	84.1	86.0	**86.5**	84.3
	GPQA-Diamond (Pass@1)	59.1	41.3	49.0	51.1	**65.0**	49.9
	SimpleQA (Correct)	24.9	10.2	9.1	17.1	28.4	**38.2**
	FRAMES (Acc.)	73.3	65.4	69.8	70.0	72.5	**80.5**
	LongBench v2 (Acc.)	**48.7**	35.4	39.4	36.1	41.0	48.1
Code	HumanEval-Mul (Pass@1)	**82.6**	77.4	77.3	77.2	81.7	80.5
	LiveCodeBench (Pass@1-COT)	**40.5**	29.2	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	**37.6**	28.4	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	**51.6**	35.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42.0	22.6	23.8	24.5	**50.8**	38.8
	Aider-Edit (Acc.)	79.7	71.6	65.4	63.9	**84.2**	72.9
	Aider-Polyglot (Acc.)	**49.6**	18.2	7.6	5.8	45.3	16.0
Math	AIME 2024 (Pass@1)	**39.2**	16.7	23.3	23.3	16.0	9.3
	MATH-500 (EM)	**90.2**	74.7	80.0	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	**43.2**	10.8	15.9	6.8	13.1	10.8
Chinese	CLUEWSC (EM)	90.9	90.4	**91.4**	84.7	85.4	87.9
	C-Eval (EM)	**86.5**	79.5	86.1	61.5	76.7	76.0
	C-SimpleQA (Correct)	**64.1**	54.1	48.4	50.4	51.3	59.3

자료: DeepSeek, 키움증권 리서치센터

DeepSeek가 공개한 논문에 의하면, 아키텍처 개선을 통해 효율성을 극도로 높였다.

MLA(Multi-Head Latent Attention), MoE(Mixture of Experts) 아키텍쳐를 결합하여, 효율성을 극도로 높였다. 먼저, MLA를 통해, 세부 사항을 반복적으로 추출, 압축하는 과정을 통해, 결과값의 정확성을 유지하면서, 속도는 높이고, 사용되는 메모리 용량을 줄였다. 그리고 MOE로 작업을 분해, 가장 적합한 신경망으로 요청을 보내는 방식을 통해 추론 비용을 낮췄다. 6,710억 개 매개 변수 중 340억 개만 활성화하여 GPU 의존도를 낮추고, 관련 비용도 함께 낮아졌다.

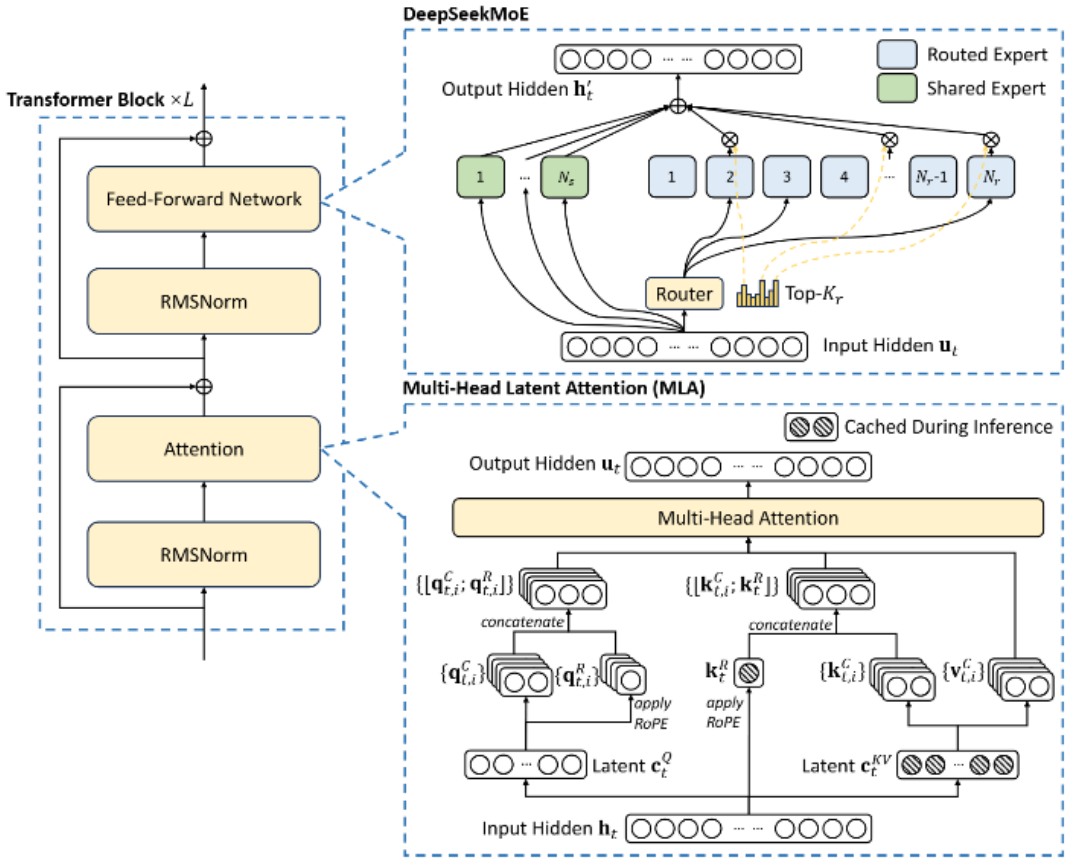
또한 Multi-token 접근을 통해, 여러 개의 토큰을 동시에 묶여서 인식하고 처리한다. 병렬적으로 문장을 해석하여 문맥 파악 능력을 높이는 동시에 속도와 정확도 또한 함께 높였다.

R1 모델에서는 인간(RLFH(**R**einforcement **L**earning form **H**uma **F**eedback, 인간 피드백을 통한 강화학습)) 혹은 인공지능(RLAF(**R**einforcement **L**earning form **AI F**eedback, AI 피드백을 통한 강화학습)) 개입 없이, 규칙 기반의 강화 학습(RL)을 통해 스스로 배우고 진화하는 방식을 택하면서, 노동력 및 비용을 절감했다.

덕분에, V3 발표 당시 논문에 의하면, 12월에 출시된 LLM 모델에서 2,048개의 H800 GPU를 사용, 훈련비용으로 557만 달러를 사용했다. GPT4 훈련비용의 10% 수준에 그치면서, 경제적인 효율성이 더욱 부각되었다.

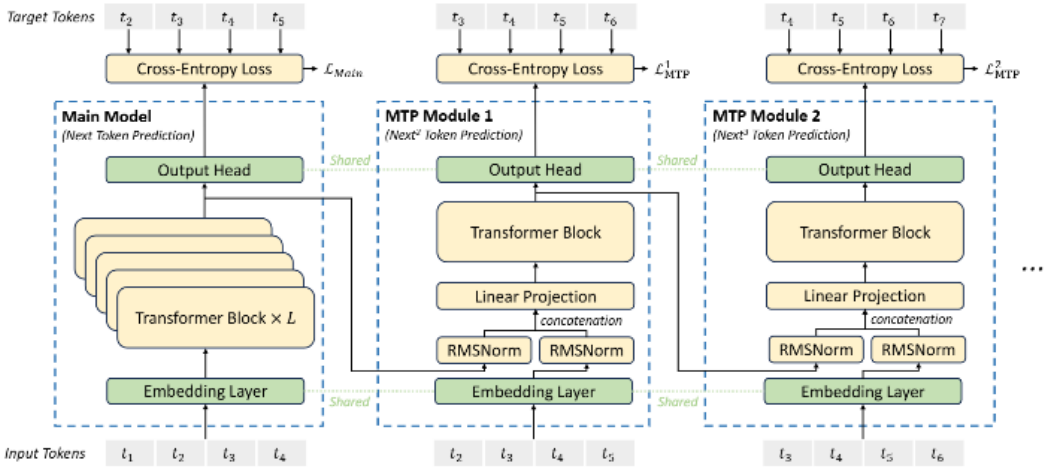
이를 통해, 파격적인 가격 정책을 펼치고 있다. DeepSeek R1 모델 요금은 입력토큰 100만개 당 0.55달러, 출력 토큰 100만개 당 2.15달러이다. 오픈 AI o1 모델 요금은 입력 토큰 100만개 당 15달러, 출력 토큰 100만개 당 60달러이다. 즉 R1 모델은 90% 넘게 할인된 요금으로 서비스를 제공한다. DeepSeek發 가격 경쟁이 중국 본토를 넘어서, 전세계적으로 확장될 가능성도 높아 보인다.

MLA와 MOE 아키텍처 개선을 통해 효율성 제고



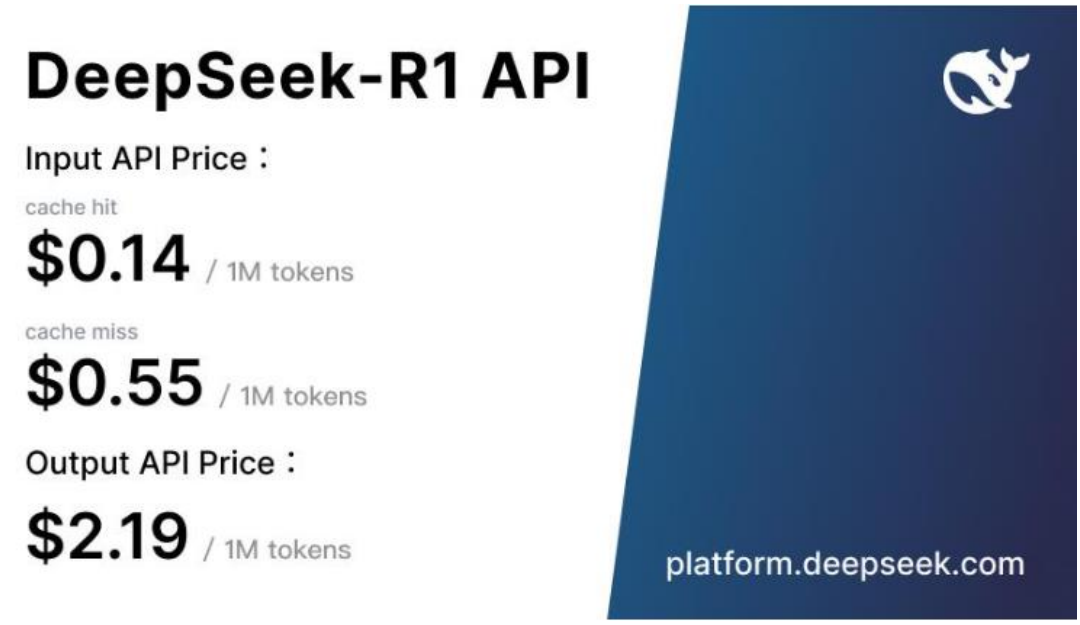
자료: DeepSeek, 키움증권 리서치센터

멀티토큰 방식을 통해 해석 정확성 및 효율성 제고



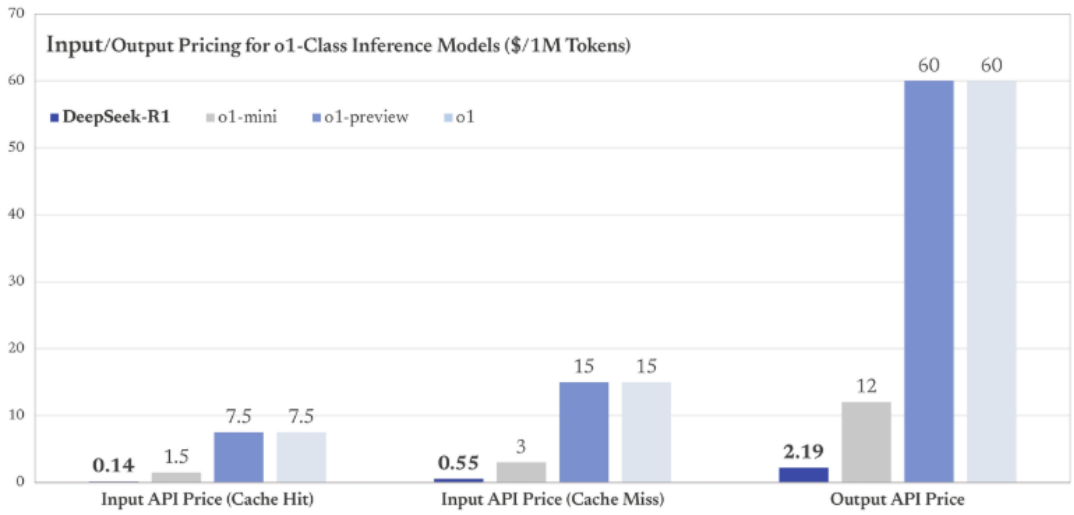
자료: DeepSeek, 키움증권 리서치센터

중국 DeepSeek R1의 높은 가격 경쟁력



자료: DeepSeek, 키움증권 리서치센터

중국 DeepSeek R1의 높은 가격 경쟁력



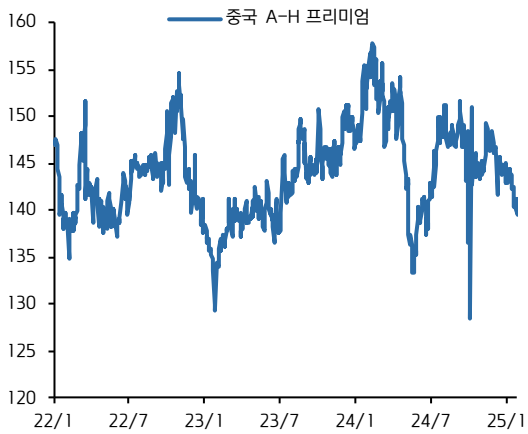
자료: DeepSeek, 키움증권 리서치센터

DeepSeek가 미국 견제에 의한 하드웨어의 한계를 스프트웨어 혁신을 통해 극복하면서, 중국증시 또한 소프트웨어 혁신이 기대되는 대형 플랫폼 기업 중심으로 투자심리가 집중되는 모습을 보인 바 있다.

중국에서 첨단산업을 이끌 수 있는 기업들은(특히 플랫폼 등 소프트웨어 기업) G2 갈등 국면이 시작된 이후, 장기 성장성에 대한 의문이 제기되면서, 꾸준히 De-rating이 되어 왔다. 이는 소프트웨어 섹터 비중이 높고, 외국인 투자 심리에 민감도가 높은 역외증시에서 더욱 선명하게 확인되어 왔다. DeepSeek의 미국 견제 극복 사례를 통해, 중국 소프트웨어 기업에 대한 투자심리가 유의미한 회복세를 보일 수 있다고 판단된다.

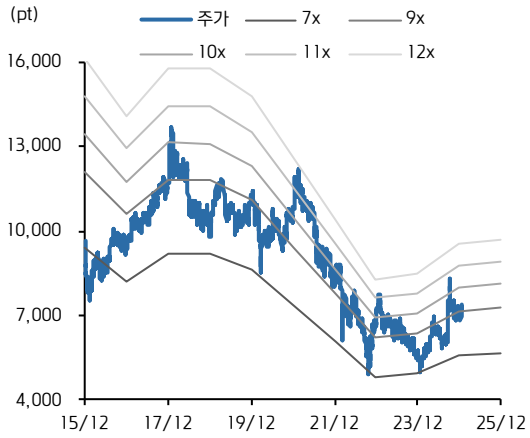
다만, 중장기 관점에서 살펴보면, 미국의 빅테크들의 천문학적인 대규모 투자 속에서, 중국의 AI 스타트 기업 혹은 어떤 중국의 플랫폼 기업들이 최종적으로 수혜를 입을 수 있을지 여전히 예단하기 어렵다고 판단한다. 금번의 성과는 소프트웨어 혁신이지만, 중국 AI 산업의 발전 성과는 중국 국산 반도체 칩과 데이터센터 인프라 투자에 따라 결정될 것으로 생각한다. 중국에 대한 기술 제재 등 대외 환경을 감안하면, 반도체 국산화의 중요성은 여전히 높은데, 상대적으로 사양이 낮은 중국산 반도체를 활용하여 소기의 성과를 거둔 사례를 통해, 중국 AI 및 반도체 등 인프라 투자에 대한 기대감이 더욱 높아질 것으로 기대된다.

중국 A-H 프리미엄 추이



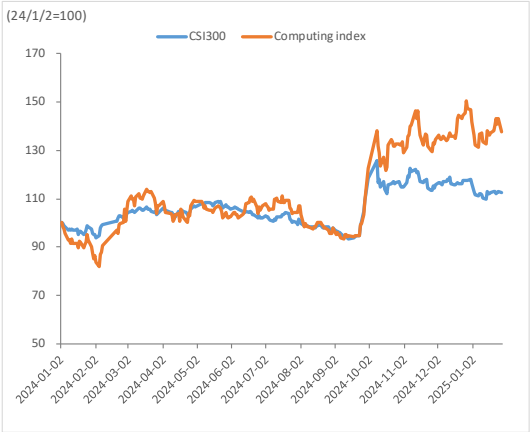
자료: Bloomberg, 키움증권 리서치

중국 항셍중국기업지수 PER밴드



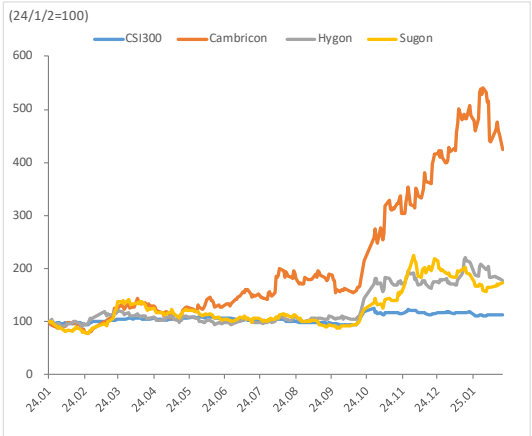
자료: Bloomberg, 키움증권 리서치

중국 컴퓨팅 인프라 관련 지수 추이



자료: Wind, 키움증권 리서치

중국 컴퓨팅 인프라 관련 기업 주가 추이



자료: Wind, 키움증권 리서치

Compliance Notice

- 당사는 동 자료를 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

고지사항

- 본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없고, 통지 없이 의견이 변경될 수 있습니다.
- 본 조사분석자료는 유가증권 투자를 위한 정보제공을 목적으로 당사 고객에게 배포되는 참고자료로서, 유가증권의 종류, 종목, 매매의 구분과 방법 등에 관한 의사결정은 전적으로 투자자 자신의 판단과 책임하에 이루어져야 하며, 당사는 본 자료의 내용에 의거하여 행해진 일체의 투자행위 결과에 대하여 어떠한 책임도 지지 않으며 법적 분쟁에서 증거로 사용 될 수 없습니다.
- 본 조사 분석자료를 무단으로 인용, 복제, 전시, 배포, 전송, 편집, 번역, 출판하는 등의 방법으로 저작권을 침해하는 경우에는 관련법에 의하여 민·형사상 책임을 지게 됩니다.