

## Strategy Idea

## AI 산업 4번째 변곡점: 폐쇄형 → 공개형 모델 구도의 이해



투자전략

Analyst 황수욱

soowook.hwang@meritz.co.kr

- ✓ Kimi K3를 AI 산업 4번째 변곡점(GPT 3.5→DeepSeek R1→OpenClaw→Kimi K3)으로 인식
- ✓ 폐쇄형 모델→공개형 모델 = 소프트웨어 라이선스 '서비스'→클라우드 운영/인프라 '서비스'
- ✓ 공개형 모델 확산, AI 프론티어(OpenAI, Anthropic, Google) 마진 압박 및 지속성 우려 불가피
- ✓ 공개형 모델 확산될수록 해자는 소프트웨어 '라이선스'에서 '인프라'로 이동할 것
- ✓ 기존의 산업 인식과 다르게 'AI 프론티어 위축 가능성'과 vs 'AI 인프라 확장'이라는 상반되어 보이는 두 현상이 동시에 진행. 오해 말아야 할 포인트는 1) AI 인프라 투자를 하이퍼스케일러만 하는 것이 아니라는 점과, 2) 공개형 모델은 AI 인프라의 성능이 더 중요하다는 것

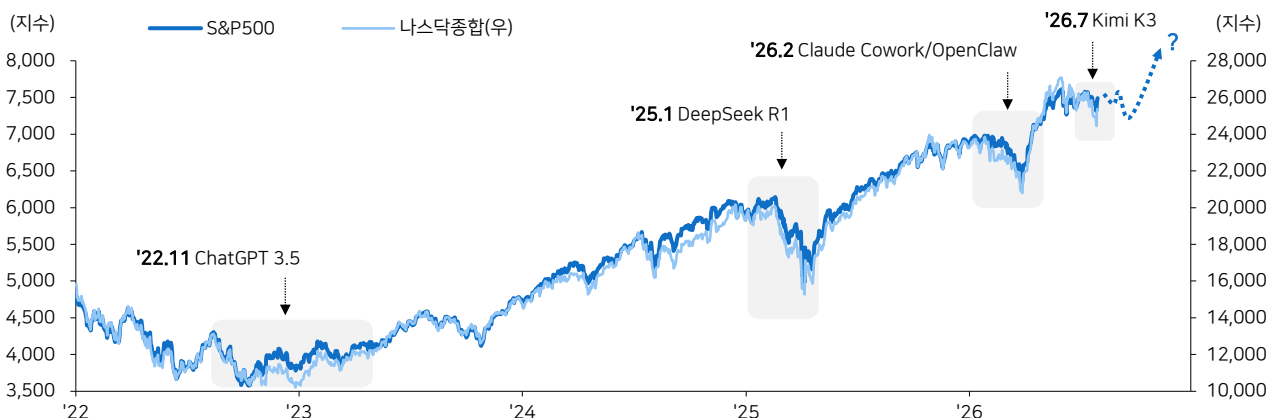
### AI 산업의 4번째 변곡점

#### 변곡점마다 반복되는 걱정; 결론은 항상 같은 방향이었다

Kimi K3를 AI 산업의 4번째 변곡점으로 인식한다. 변곡점이 도래하는 시기가 짧아진다. 그러나 변곡점마다 이전에 없던 기술을 해석하는 시각의 오해는 반복되는 듯하다. GPT 3.5가 처음 나왔을 땐 인식하는데 시간이 걸렸고, DeepSeek R1은 AI 과잉 공급 우려를 키웠다. OpenClaw는 사스포칼립스 걱정을 키웠고, Kimi K3는 DeepSeek처럼 중국 AI모델 약진이 미국 AI 진영에게 위협이라는 걱정이 앞섰다.

뒤돌아보면 결론은 항상 같은 방향이었다. AI의 진화와 인프라 수요의 확대였다.

그림1 AI 시작(GPT 3.5) → 생각하는 AI(DeepSeek) → AI 에이전트(OpenClaw) → 공개형 모델의 폐쇄형 모델 catch up(Kimi K3)



자료: Bloomberg, 메리츠증권 리서치센터

## 폐쇄형→공개형 모델 구도로 확산: 중국이 선도, 미국의 추종

AI 모델은 공개형 모델과 폐쇄형 모델로 구분된다. 공개형 모델은 모델의 가중치를 공개하고, 특별한 요구가 없는 경우 사용자가 별도의 라이선스 없이도 컴퓨팅/전력/서버 운영비만으로 모델의 AI 연산을 누릴 수 있다. 다수의 중국 AI 모델이 공개형 모델에 해당한다. 폐쇄형 모델은 모델의 가중치, 구조 공개 없이 API나 서비스 형태로만 AI 연산을 제공하는데, 우리가 익숙한 미국 AI 프런티어(OpenAI, Anthropic, Google)의 최상위 모델들이 대부분 여기에 해당한다.

7월 중순 이후 중국이 초대형 플래그십 공개형 모델을 연속적으로 공개하고 있다. 7월 16일 문샷의 Kimi K3, 8월 3일 알리바바의 Qwen 3.8 Max이 공개되었고, 미국의 폐쇄형 AI 프런티어 모델과 필적할 성능을 지녔다는데 시장이 놀랐다. 여기에 7/31 DeepSeek V4-Flash 공개는 같은 작업을 수행하는데 미국 AI 프런티어 대비 1/100 수준의 비용이라는 점에서 미국 AI 프런티어의 해계모니를 위협한다. OpenRouter 사용량 Top10 모델 중 9개가 공개형 모델이다. 3월과 대조적이다.

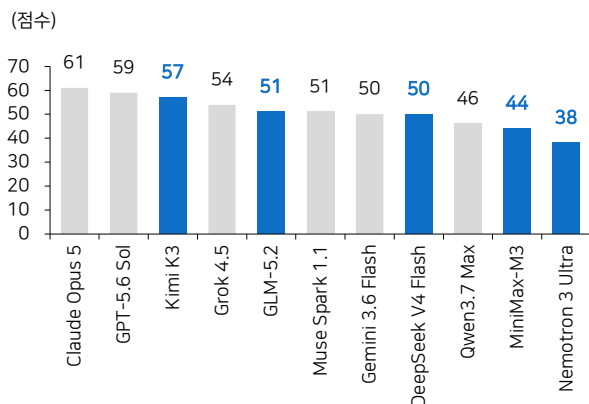
| 표1 폐쇄형/공개형 모델 비교 |                                   |                                  |
|------------------|-----------------------------------|----------------------------------|
| 구분               | 폐쇄형 모델                            | 공개형 모델                           |
| 대표 사례            | OpenAI, Anthropic, Google 상위 모델   | DeepSeek, Moonshot, Qwen, NVIDIA |
| 개념               | 모델 구조, 가중치 공개 없이 API나 서비스 형태로만 제공 | 모델 가중치 공개, 사용자가 직접 다운로드해 실행/수정   |
| 사용방식             | 개발사가 운영하는 클라우드 API 접속             | 가중치를 다운로드한 뒤 자체 서버/인프라에서 실행      |
| 모델 가중치           | 공개                                | 비공개                              |
| 수익모델             | API 토큰판매, 구독료, 기업용 라이선스           | 클라우드 추론 서비스, 기술지원 등 인프라 중심       |
| 사용비용             | 토큰 단위 API 비용 지급                   | 모델 라이선스 무료여도 컴퓨팅/전력/운영비 발생       |
| 성능 수준            | 최신 프런티어는 대체로 폐쇄형 우위               | 범용 성능 격차 줄고 특정 업무에서 자체 최적화       |
| 보안/규제 대응         | 공급자 보안정책과 데이터 처리 방식에 의존           | 온프레미스 구축시 민감 데이터에 유리             |

자료: 메리츠증권 리서치센터

| 표2 공개형/폐쇄형 모델 OpenRouter 사용량 Top10 순위 |       |                   |       |
|---------------------------------------|-------|-------------------|-------|
| 3월 4주차                                |       | 7월 4주차            |       |
| 모델명                                   | 공개 여부 | 모델명               | 공개 여부 |
| Qwen 3.6 Plus                         | 폐쇄    | DeepSeek V4 Flash | 공개    |
| MiMo-V2-Pro                           | 폐쇄    | MiMo-V2.5         | 공개    |
| Qwen 3.6 Plus Preview                 | 폐쇄    | Hy3               | 공개    |
| Step 3.5 Flash                        | 공개    | DeepSeek V4 Pro   | 공개    |
| MiniMax M2.7                          | 공개    | GLM 5.2           | 공개    |
| DeepSeek V3.2                         | 공개    | Nemotron 3 Ultra  | 공개    |
| Claude Sonnet 4.6                     | 폐쇄    | MiniMax M3        | 공개    |
| Claude Opus 4.6                       | 폐쇄    | GPT-5.6 Luna      | 폐쇄    |
| Gemini 3 Flash Preview                | 폐쇄    | Step 3.7 Flash    | 공개    |
| MiniMax M2.5                          | 폐쇄    | Kimi K3           | 공개    |

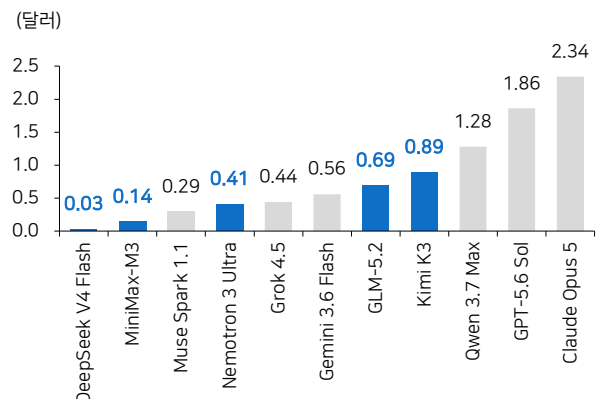
자료: OpenRouter, 메리츠증권 리서치센터

그림2 AAI 모델 공개형/폐쇄형 모델 성능 구분



주: 파란색이 공개형 모델, 회색이 폐쇄형 모델  
자료: Artificial Analysis, 메리츠증권 리서치센터

그림3 공개형/폐쇄형 모델별 작업당 토큰 비용



주: 파란색이 공개형 모델, 회색이 폐쇄형 모델  
자료: Artificial Analysis, 메리츠증권 리서치센터

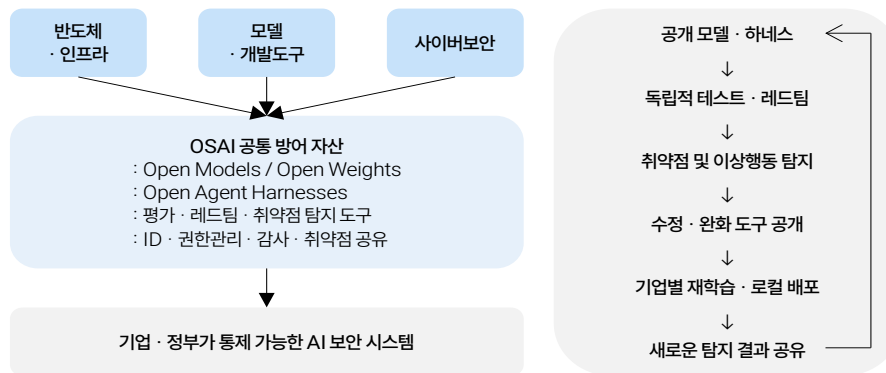
## 미국 내에서도 공개형 모델을 위한 결집 시작

7월 16일 OpenAI의 폐쇄형 모델 기반 자율 에이전트가 공개형 AI 플랫폼인 허깅 페이스 시스템에 침투했다. 허깅페이스는 공격을 차단한 뒤 중국 공개형 모델 GLM-5.2를 자체 운용해 암호화된 공격 기록을 해독했다. 이 사건은 프런티어 AI의 자율적 사이버 공격 위험과 동시에 감독/수정 가능한 공개형 모델의 방어적 가치를 보여준 사례로 평가받는다.

이 사건을 계기로 공개형 모델의 경제/안보적 가치를 강조하는 흐름이 미국 내에서 발생했다. 이를 기반으로 7월 24일 「Open Weights and American AI Leadership」 공동서한이 발표되었고, 엔비디아를 중심으로 Open Secure AI Alliance가 출범했다.

7월 24일 공동서한에 서명하지 않은 4개 기업이 있는데, OpenAI, Anthropic, Google, Amazon이다. 폐쇄형 모델 중심의 사업 모델을 기반으로 하는 AI 프런티어 기업들로, 공개형 모델이 확산에 따른 경쟁 심화를 우려하는 당사자들이다. 미국 내 공개형 모델을 위한 결집이 이들에겐 불편한 것이다.

그림4 Open Secure AI Alliance 생태계 지도



자료: 메리츠증권 리서치센터

표3 OSAI 주체별 역할 및 참여 기업

| 주체           | 역할                                                                                         | 참여기업                                                                                                                         |
|--------------|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| 반도체 · 인프라    | 모델을 자체적으로 실행하고 격리할 수 있는 컴퓨팅·네트워크·스토리지 환경 제공. 외부 API에만 의존하지 않고, 기업내부나 국가 인프라에서 구동할 수 있도록 해줌 | NVIDIA, Akamai, Cadence, Cisco, Cohesity, Crusoe, Dell, F5, G42, HPE, NetApp, Nokia, Nutanix, SK Telecom, Synopsys, VMWare 등 |
| 모델 · AI 개발도구 | 보안분석, 취약점 탐색에 사용 가능한 개방형 모델과 추론 기술 제공                                                      | Fireworks AI, H2O.ai, Mistral, Perplexity, SpaceXAI, Thinking Machines Lab, GitHub, Hugging Face, OpenClaw 등                 |
| 사이버보안        | ID · 권한 · 격리 · 관측 · 취약점 검증 체계 제공                                                           | Cloudflare, CrowdStrike, Fortinet, Okta, Palo Alto Networks, Zscaler 등                                                       |
| 기업 SW · 실수요자 | 실제 데이터와 업무 시스템에 적용하고 사고 사례, 운영 피드백, 평가 환경 제공                                               | Adobe, Databricks, IBM, Microsoft, NAVER, Palantir, Salesforce, SAP, ServiceNow, Snowflake, Siemens, Uber, DoorDash 등        |
| 오픈소스 재단      | 표준화, 공개 저장소, 취약점 공개, 커뮤니티 유지관리                                                             | Linux Foundation, Mozilla                                                                                                    |
| 공통 방어 자산     | 오픈모델과 모델 가중치, 재사용 가능한 보안 기술, 공개연구, 다중 모델 기반 취약점, 공격 시뮬레이터 등                                | -                                                                                                                            |

자료: NVIDIA, 메리츠증권 리서치센터

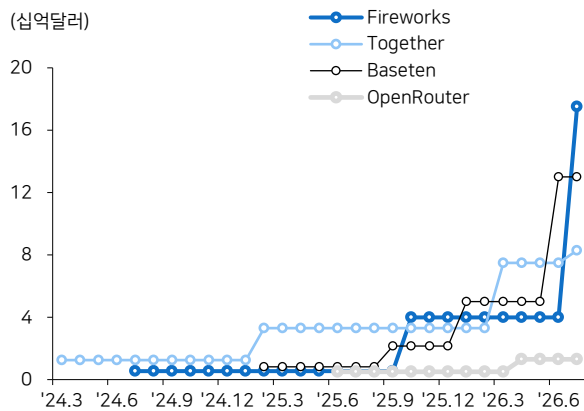
## 공개형 모델 확산→해자는 '라이선스'가 아닌 '인프라'로 이동할 것

폐쇄형 모델 비즈니스가 불편해하는 이유는, 공개형 모델이 확산될 경우 독점적 '라이선스' 기반 소프트웨어 서비스는 약해질 수 있기 때문이다. 폐쇄형 모델 비즈니스는 독점적 모델 접근권을 기반으로 API 요금을 부과해 높은 매출총이익률을 누릴 수 있다. 하지만 충분히 성능이 좋은 공개형 모델을 기업이 직접 내려 받아 사용할 수 있으면, 고객은 더 이상 특정 사업자의 API를 구매할 필요가 없다.

공개형 모델은 지출의 중심이 지식재산권 사용료에서 연산비용으로 이동한다. 가격 기준점이 '독점 가치'에서 '추론 원가'로 내려간다는 뜻이다. 모델이 동일해지면 서비스 가격은 점차 차별화된 컴퓨팅 원가, 운영마진에 가까워진다. 그러면 해자는 소프트웨어 라이선스가 아닌 인프라로 이동한다. AI 서빙 인프라를 보유한 ISP의 기업가치평가액이 오르고 공개형 AI 기업 주가가 하락하는 이유일 수 있다.

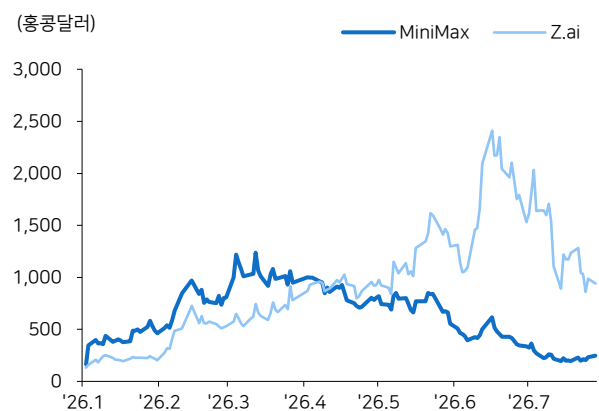
공개형 모델을 제공하는 추론 서비스 사업자(Inference Service Provider)별로 같은 모델을 제공하더라도 그 가격과 성능이 다른 이유는 인프라 격차일 것이다.

그림5 미국 주요 비상장 AI 추론 서비스 사업자 가치평가액



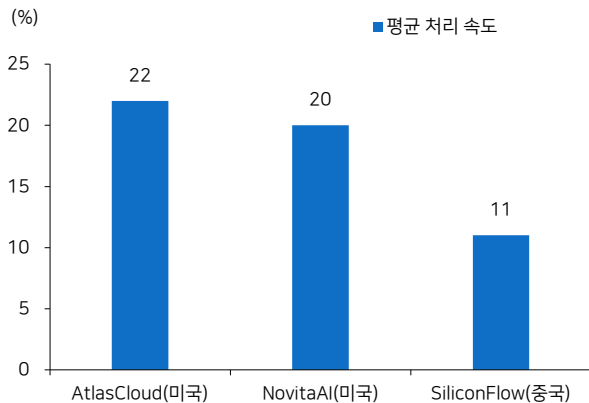
자료: Bloomberg, 메리츠증권 리서치센터

그림6 중국 AI 공개형 모델 기업 주가 추이



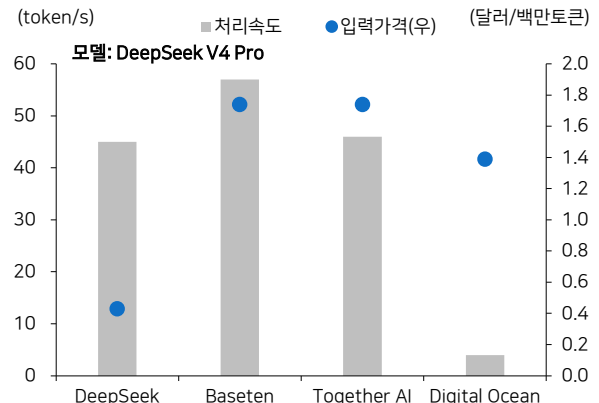
자료: Bloomberg, 메리츠증권 리서치센터

그림7 DeepSeek V3.2 Exp ISP별 최근 1주 성능 차이



자료: OpenRouter, 메리츠증권 리서치센터

그림8 같은 모델임에도 프로바이더에 따라 가격/속도 다름



자료: OpenRouter, 메리츠증권 리서치센터

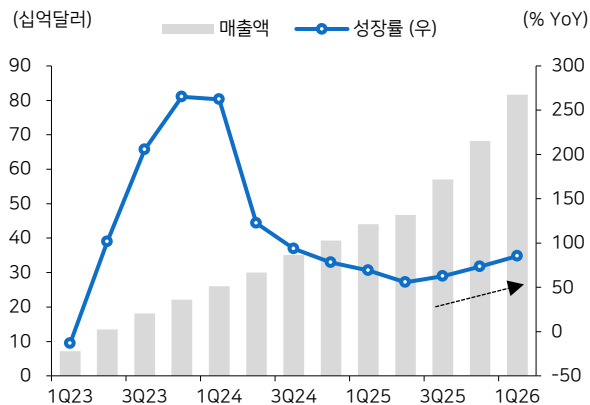
## 과거 변곡점과 같은 결론, 수혜는 AI 인프라. AI CapEx 수요 확대

따라서 공개형 모델은 인프라의 중요성을 낮추는 것이 아닌, 오히려 더 명확하게 드러낸다. 동일한 모델을 누구나 사용할 수 있기 때문에 모델 격차는 축소되고, GPU, HBM, 네트워크, 전력, 서버 소프트웨어가 만드는 처리 속도, 안정성 및 토큰당 비용 차이가 직접적인 경쟁력이 될 것이다.

위 구도에서 기존의 산업 인식과 다르게 'AI 프론티어 위축 가능성'과 'AI 인프라 확장'이라는 상반되어 보이는 두 현상이 동시에 진행될 수 있다. 오해 말아야 할 포인트는 1) AI 인프라 투자를 하이퍼스케일러만 하는 것이 아니라는 점과, 2) 공개형 모델은 AI 인프라의 성능이 더 중요하는 점이다.

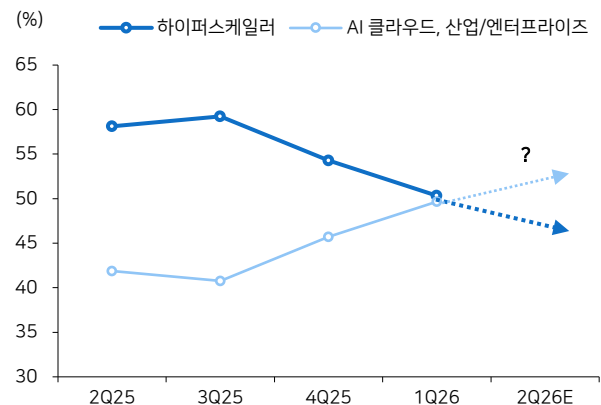
AI 산업의 4번의 변곡점이 있었는데, 변곡점마다 AI CapEx 우려가 컸지만, 이번도 크게 다르지 않다는 의견이다. 공개형 모델의 확산은 인프라 수요를 키운다. 지금 산업에서 중요한 것은 이들이 돈을 벌고 있느냐일 것이다.

그림9 엔비디아 매출액 성장을 구간에서



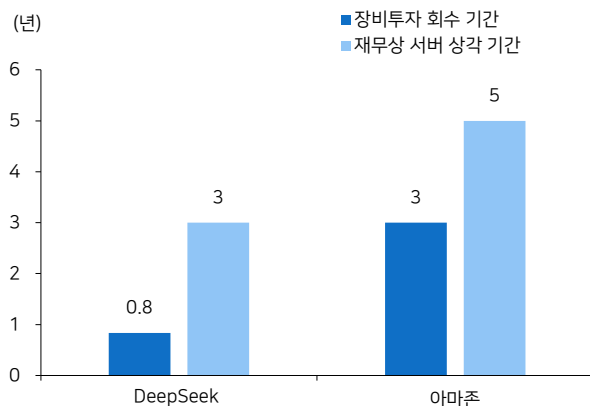
자료: Bloomberg, 메리츠증권 리서치센터

그림10 하이퍼스케일러 이외 AI 인프라 투자 주체 확산



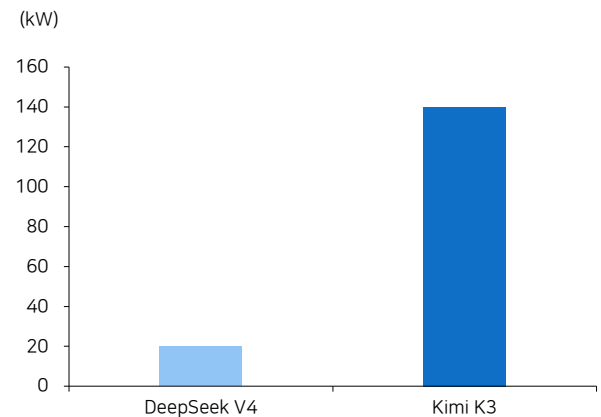
자료: NVIDIA, 메리츠증권 리서치센터

그림11 주요 기업에서 확인되는 투자 회수 기간 < 상각 기간



자료: Amazon, Fred Gao, 메리츠증권 리서치센터

그림12 최신 공개모델 하나를 구동하는 필요한 용량 추정



주: V4 추정치는 이론적 최소 단일 인스턴스 추정치.  
실제 안정적 온프레미스 배포는 30~100kW이상 필요할 가능성  
자료: 메리츠증권 리서치센터 추정

---

**Compliance Notice**

- 본 조사분석자료는 제3자에게 사전 제공된 사실이 없습니다.
- 당사는 자료작성일 현재 본 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 추천 종목과 재산적 이해관계가 없습니다.
- 본 자료에 게재된 내용은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 신의 성실하게 작성되었음을 확인합니다.

본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 리서치센터의 추정치로서 오차가 발생할 수 있으며 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 본 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 투자 결과와 관련한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 당사의 허락 없이 복사, 대여, 배포 될 수 없습니다.

---