클라우드에서 기기로: 온디바이스가 바꾸는 AI 산업지형

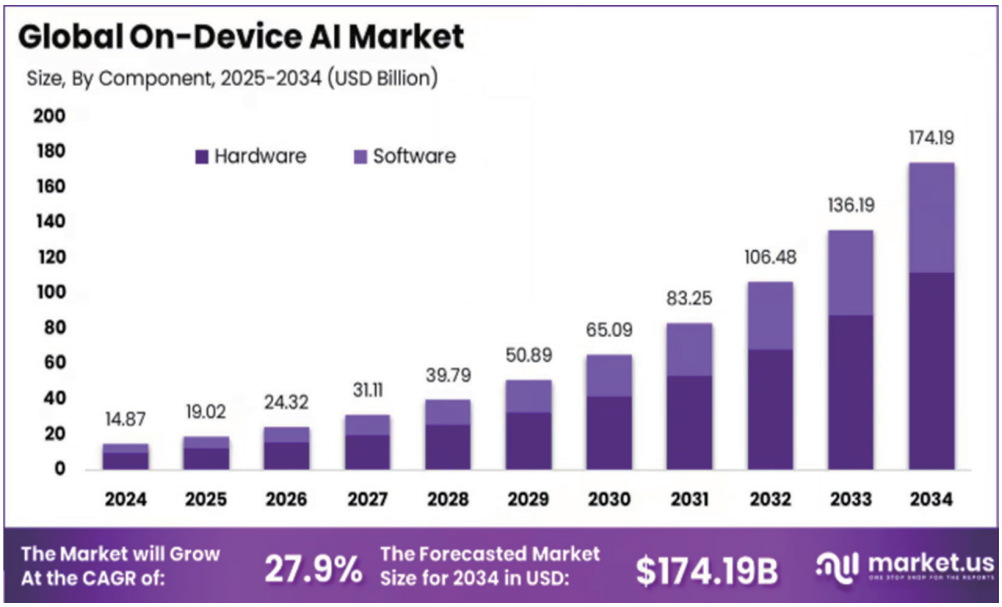
클라우드에서 기기 속으로 AI 운영 패러다임의 이동

오늘날 우리가 사용하는 문서 번역, 이미지 인식 등 대부분의 AI 서비스는 대규모 데이터센터를 거쳐 작동한다. 스마트폰에서 명령을 내려도 실제 연산은 수천 킬로미터 떨어진 서버에서 이뤄지고 결과만 전송되는 형태다. 이 방식은 고성능 모델을 활용할 수 있다는 장점이 있지만, 지연시간(latency), 네트워크 의존성, 그리고 개인정보 유출 위험이라는 구조적 한계를 갖는다.

반면 온디바이스 AI(On-device AI)는 단말기 내부에서 AI 연산을 직접 처리하는 방식이다. 네트워크 연결이 불안정해도 오프라인 상태에서 작동할 수 있고, 데이터가 기기 밖으로 나가지 않기 때문에 프라이버시 보호에도 유리하다. 응답 속도도 거의 실시간에 가깝다. 또한 저전력, 저비용의 장점도 있어 온디바이스 AI환경에서는 개인 맞춤형 서비스에도 유리하다.

이러한 기술적·운용상의 이점으로 인해 AI의 '무게중심'이 클라우드에서 기기로 빠르게 이동하고 있으며, 그 결과 글로벌 시장에서 온디바이스 AI는 폭발적인 성장세를 보이고 있다. 2024년 148억 달러 규모였던 시장은 2034년 1,741억 달러를 돌파할 것으로 예상된다. 이는 연평균 27.9%가 넘는 고성장률이다.

온디바이스 AI 시장규모



출처: market.us (2025.8월)

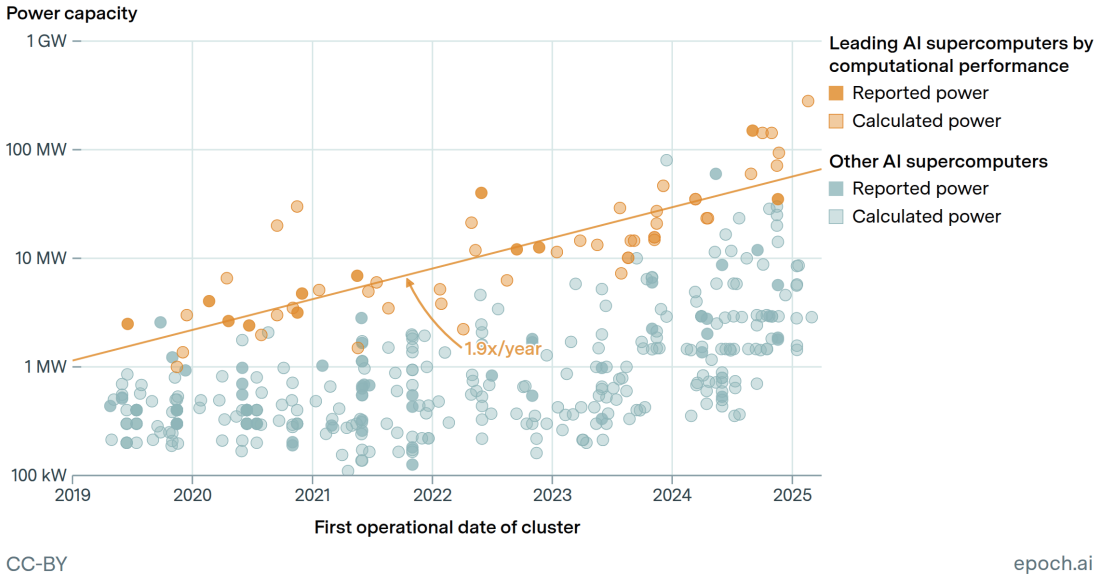
데이터센터 운영비 증가가 부른 '탈클라우드'화

클라우드 기반 AI 모델의 규모가 커질수록 필요한 데이터센터 운영 비용도 기하급수적으로 증가하고 있다. EPOCH AI 연구에 따르면 전세계 AI 데이터센터의 하드웨어 비용은 매년 약 1.9배씩, 전력 수요는 2배씩 증가하고 있다. 이로 인해 운영비가 수익을 잠식하는 현상이 심화되고 있다. 오픈AI의 연간 매출이 약 120억 달러에 달할 것으로 추정*되지만 전력 및 GPU 운영 비용 때문에 단기간 흑자 전환은 어렵다는 소식도 나온다.

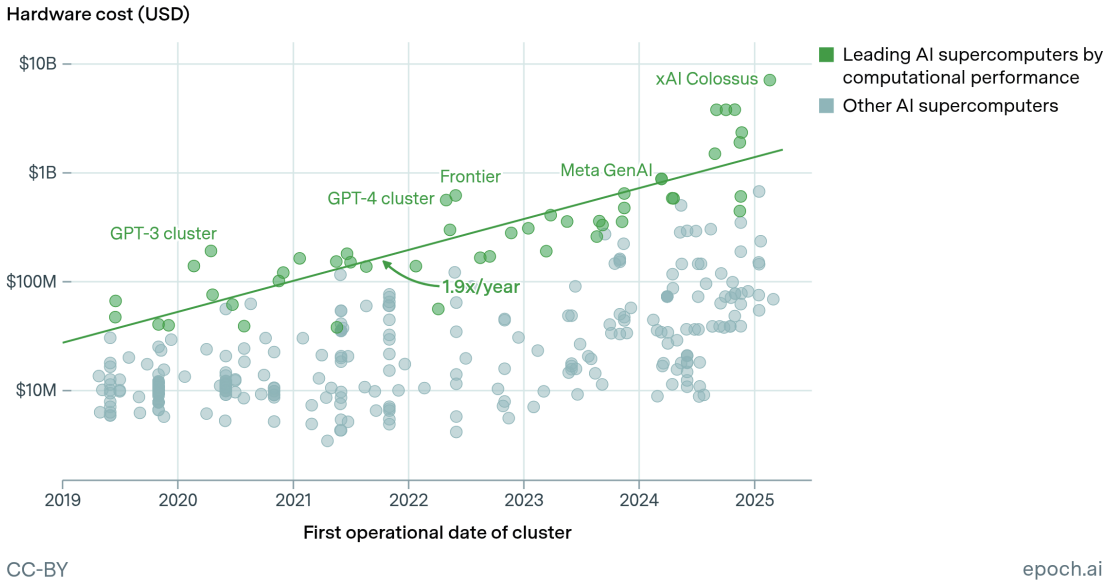
* 로이터(Reuters)는 OpenAI가 올해 첫 7개월 동안 매출이 약 2배 증가해 연환산 수익이 약 120억 달러에 이른다고 보도함(7.31일자). 반면 OpenAI는 2029년까지 약 1,150억 달러를 소진할 것으로 전망하고, 이러한 대규모 지출은 AI 모델 및 데이터센터, 서버 칩 개발 등에 막대한 투자를 하고 있기 때문이라고 보도함(9.6일자)

AI 슈퍼컴퓨터 하드웨어 비용, 전력수요 전망

1.Power capacity of AI supercomputers



2.The hardware cost of leading AI supercomputers has doubled every year



출처: EPOCH AI, 'Trends in AI Supercomputers' (2025.4월)

더 심각한 문제는 탄소배출이다. AI 학습용 데이터센터는 막대한 전력을 소비하며, 에너지 공급의 상당 부분이 여전히 화석연료에 의존하고 있다. 이에 따라 ESG 경영을 추구하는 기업입장에서는 친환경적 대안으로서 온디바이스 AI 전환이 매력적으로 부상하고 있다.

작지만 강한 AI, 소형 언어모델(SLM)의 부상

온디바이스 AI의 핵심 도전은 '작은 기기에 큰 AI를 담는 일'이다. 많게는 수조 개의 매개변수를 가진 대형 언어모델(LLM)을 스마트폰이나 차량 칩 내에 탑재하는 것은 물리적으로 불가능하다. 이 한계를 극복하기 위해 업계는 소형 언어모델(SLM) 전략으로 전환했다. 이는 모델규모를 줄이되 성능은 유지하거나 특정 업무에 최적화하는 방식이다.

구분	정의 및 예시
대형언어모델 LLM (Large Language Model)	수백억~수조 개의 매개변수(Parameter)를 가진 대규모 인공지능 언어모델로, 광범위한 데이터 학습을 통해 복잡한 추론·창작·요약 등 범용적 자연어 이해 및 생성 능력을 보유함 * 예: GPT-4(오픈AI), Claude 3(앤트로픽), Gemini 1.5(구글 딥마인드), Llama 3(메타 플랫폼스) 등
소형언어모델 SLM (Small Language Model)	수억~수십억 개의 매개변수로 구성된 경량화 언어모델로, 문서요약, 질의응답, 번역 등 특정 과업에 최적화되어 온디바이스 환경에서도 작동 가능한 효율적 모델을 의미함 * 예: Phi-3(마이크로소프트), MiniLM(마이크로소프트), Gemma 2B/7B(구글 딥마인드) 등

마이크로소프트의 연구진은 논문을 통해 데이터의 양보다 질이 성능을 좌우함을 증명했다. 고품질 교과서급 데이터로 학습한 소형 모델은 적은 데이터로도 대형 모델에 근접한 성능을 보였다. 예컨대 마이크로소프트의 소형 언어모델 'Phi-3.5'는 메타의 대형 언어모델 'Llama 3.1'에 근접한 수준의 일반상식 추론 성능을 보이고 있다.

인공지능LLM과 SLM 모델간 성능 비교

모델명	매개변수 수(억개)	주요 용도	Llama 3.1대비 성능(%)
메타 Llama 3.1	700	범용 언어	100
마이크로소프트 Phi-3.5	35	일반상식 추론	95
마이크로소프트 MiniLM	33	번역 특화	90

출처: Microsoft Research, 'Textbooks Are All You Need' (2023.6월)

미국 스탠퍼드대 HAI 연구소는 '2025년 AI 인덱스 보고서'에서 경량 모델이 초대형 모델과 거의 유사한 성능을 내는 최근의 흐름을 'AI 시대의 전환점'으로 평가했다. 2024년 MS의 경량모델 'Phi-3-Mini'는 약 38억 파라미터로 MMLU 벤치마크에서 62.9%의 점수를 받았다. 이는 2022년 구글이 공개한 5,400억 파라미터 초대형모델인 '팜(PaLM)'과 거의 비슷한 기록(60% 이상)이다.

* MMLU (Massive Multitask Language Understanding, 대규모 다중과제 언어이해) : 여러 학문 분야(과학, 수학, 역사, 법학, 의학 등 57개 분야)에 걸친 지식 기반 문제를 통해 언어모델(LLM)의 종합적 사고력과 지식 이해도를 평가하는 벤치마크

이와 함께 모델 경량화 기법인 '양자화(Quantization)', '프루닝(Pruning)', 증류(Distillation)도 AI에 활발히 적용되고 있다. 예를 들면, 모델 파라미터의 정밀도를 낮춰 크기를 줄이는 기법(양자화), 중요도가 낮은 연결을 제거해 연산량을 줄이는 방법(프루닝), 고성능의 대형모델에서 지식을 전달받아 소형모델을 학습시키는 방법(증류)이다.

AI를 담는 하드웨어 혁명: 칩과 기기의 재편 경쟁

온디바이스 AI 시장의 성장은 AI 전용 하드웨어 수요를 폭발적으로 늘리고 있다. 대형 GPU를 탑재하면 높은 비용에 비해 효율성이 떨어지기 때문에, 그 대안으로 저전력, 저비용으로 온디바이스 AI에서 작동할 수 있는 NPU(Neural Processing Unit)가 대두되고 있다. NPU는 AI 연산에 특화된 프로세서로 전력 소모를 최소화할 수 있어 최근 주목받고 있다.

글로벌 기업들은 이 기술 전환을 하드웨어 경쟁의 기회로 활용하고 있다. 애플은 개인정보 보호를 핵심 가치로 삼아 'Apple Intelligence' 전략을 전면화하고, 아이폰과 맥북에 온디바이스 AI를 적용하고 있다. 구글은 자체 하드웨어인 크롬북에 'Gemini Nano'를 탑재하여 녹음 요약, 통화 메모 등을 오프라인으로 처리하고 있다. 메타는 XR과 스마트안경 등 확장현실 디바이스에, 마이크로소프트는 AI PC 시장에 집중하고 있다. 이처럼 온디바이스 AI는 단순한 기술 혁신을 넘어, 하드웨어·칩 패권 재편을 동반한 새로운 산업 질서를 만들고 있다.

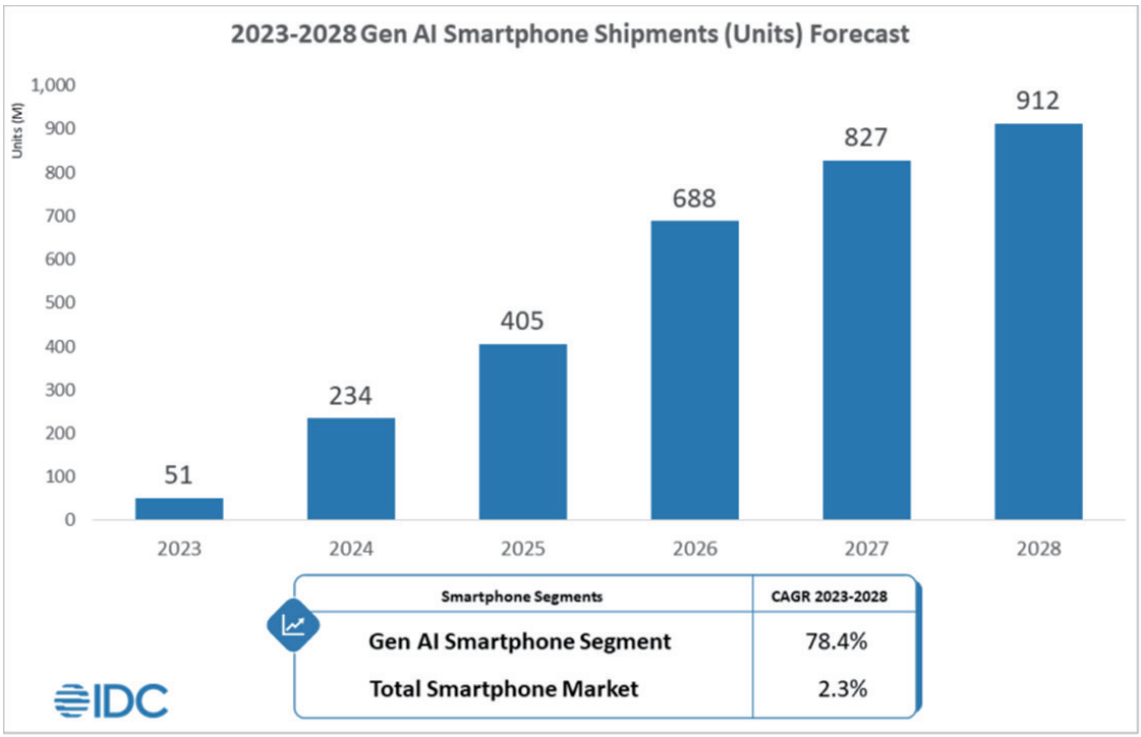
우리 기업들은 스마트폰, XR(확장현실), 스마트가전 등 하드웨어 제작 경험이 풍부하다. 예컨대, 삼성전자 'Galaxy AI'는 온디바이스 AI 기능을 자사 디바이스에 적극 적용하고 있다. 이처럼 우리 역시 강점을 살려 '제품-칩-소프트웨어'가 통합된 AI 생태계를 구축할 기회로 삼아야 한다.

한국형 온디바이스 AI 전략과 '소버린 AI'의 확보

이러한 세계적 흐름 속에서 한국은 온디바이스 AI 상용화를 가장 빠르게 추진하고 있는 대표적 국가 중 하나다. 삼성은 2024년 세계 최초로 온디바이스 AI 스마트폰 'Galaxy S24'를 출시했으며, 네트워크 없이도 통화 번역, 노트 요약, 문장 수정 등이 가능한 'Galaxy AI'를 탑재했다. 2025년 언팩 행사에서는 "4억 대 기기에 갤럭시 AI를 적용하겠다"는 목표를 공언했다.

시장조사기관 IDC에 따르면 2023년에서 2028년까지 온디바이스 생성형 AI 스마트폰 출하량은 연평균 78.4% 성장해 9억 대를 넘어설 것으로 예상된다. 이는 스마트폰이 곧 AI 단말기인 시대의 개막을 의미한다.

생성형 AI 스마트폰 출하량 전망



출처: IDC, "The Rise of Gen AI Smartphones," IDC Insights Blog, 5 July 2024.

더 주목할 점은 중소기업과 스타트업의 약진이다. 2025년 CES에서 한국 기업들은 총 208개 혁신상을 수상해 역대 최다 기록을 세웠고, 그중 131개(63%)가 중소, 스타트업의 성과였다. 예비 유니콘 기업인 '페르소나 AI'는 온디바이스 교육 플랫폼으로 혁신상을 받으며 "인터넷 없이 작동하는 AI 모델"로 주목받았다.

한국은 대형 모델 경쟁에서는 미국과 중국에 뒤처졌지만, 경량 모델과 온디바이스 분야에서는 기술 응용력과 하드웨어 강점을 바탕으로 새로운 기회를 찾고 있다. 현지화된 경량 모델과 신경망처리장치(NPU) 결합 서비스 등으로 틈새시장을 공략할 수 있다. AI 서비스 운영비용이 낮아지면 새로운 형태의 AI 서비스도 다양하게 등장할 수 있다. 이미 스퀴즈비츠(메모리 줄이고 성능 유지), 노타(칩 경량화 솔루션), 클리카(AI 모델 압축) 등 경량화 기술을 갖춘 국내 스타트업이 글로벌 시장에서 주목받고 있다. 딥엑스, 모빌린트 등 온디바이스용 NPU 개발사도 반도체 영역에서 활약하고 있다. 이러한 성과는 대기업 뿐만 아니라 중소·스타트업까지 포괄하는 국가적 AI 생태계의 확장을 의미한다. AI 기술의 주권(Sovereign AI)을 확보하기 위해서는 이 생태계가 지속적으로 자립하고 성장해야 한다.

분산형 AI로 향하는 길, 한국은 온디바이스 AI 발전모델 제시

AI 모델이 클라우드에서만 작동하던 시대를 넘어, 개별 기기 속으로 내재화되는 '분산형 AI 시대'가 열리고 있다. 이는 데이터 주권, 전력 효율, 보안 신뢰성을 동시에 달성할 수 있는 전환점이기도 하다. 한국은 세계적 수준의 하드웨어 기술력과 빠른 시장 적응력을 갖춘 몇 안 되는 국가다. 그렇기에 한국은 클라우드 AI 시장에서도 물론 선전해야 하겠지만, 최근 부각되고 있는 온디바이스 AI 시장만큼은 주도권을 확보해야 한다.

한국형 온디바이스 AI 전략은 기술개발 중심의 접근을 넘어, 제품개발 뿰주기 지원체계가 필요하다. 다양한 수요발굴 단계부터 제품개발, 표준화, 최적화, 성능 인·검증, 실증, 사업화, 글로벌시장 진출까지 연결하는 통합지원이 이뤄져야 한다. 특히, NPU 기업의 경우 레퍼런스 부족으로 시장개척에 애로가 많았으므로 성능 검인증을 통한 레퍼런스 확보 지원을 강화해야 한다.

기업·연구기관이 공용으로 활용할 수 있는 오픈 테스트베드형 'AI 디바이스 실증 플랫폼'을 마련하고, 대기업-중소기업-스타트업-지원기관-연구기관 등이 참여하는 '온디바이스 AI 얼라이언스'도 생태계 구축을 위해 필요하다. 자율주행, 로봇, 가전 등 분야별 실증을 통한 응용부문(Application Sector)에서의 확산전략을 구체화해 나갈 뿐만 아니라 휴머노이드 등 피지컬 AI 활용분야를 온디바이스의 대표 품목으로 전략 육성하는 것도 바람직하다. 이러한 실증과 생태계 구축을 통해, 글로벌 기업들이 AI를 단순 서비스로 다루는 것과 달리 한국은 온디바이스 AI를 산업 인프라 수준으로 발전시키는 모델을 제시해야 한다. FKI