Ⅱ 주요 동향(2) : ICT

1 온디바이스 AI, 작아진 모델이 여는 가까운 AI

AI 실행 경쟁, 모델 성능 고도화에서 사용자 단말 기반 실행으로 확장

AI 실행 위치, 중앙 데이터센터에서 사용자 접점 단말로 재배치

- '26년 MWC·COMPUTEX·Build·WWDC 등에서 글로벌 기업이 단말 내부 AI 실행전략을 전면에 배치하면서, 모델 성능 중심이던 AI 경쟁에 사용자 접점 기반 실행 경쟁이 본격적으로 가세하는 흐름
- 이는 AI 경쟁의 초점이 초거대 모델을 누가 더 잘 만드는가의 문제에서, AI 기능을 사용자 접점에 얼마나 빠르고 안전하며 비용 효율적으로 내장하는가의 문제로 확장되고 있음을 의미
- 특히 AI 기능이 검색, 문서, 통화, 사진 편집, 개인비서 등 일상 서비스로 확산되면서, 모든 추론을 중앙 클라우드에 집중시키는 구조는 비용·전력·응답 지연 측면의 부담이 확대되는 추세
- 이에 따라 AI 추론의 처리 지점은 중앙 데이터센터에서 PC, 스마트폰, 웨어러블 등 사용자 단말과 엣지 환경으로 분산되고, 기기 안에서 사용자 맥락을 상시 처리하는 기능도 본격화
- 결국 온디바이스 AI는 단말에 작은 모델을 탑재하는 기술 이슈를 넘어, AI 추론의 실행 위치가 사용자 접점으로 재배치되는 구조적 변화로 해석 필요

온디바이스 AI, 경량 모델·반도체·플랫폼을 묶는 결합 경쟁으로 확대

- 온디바이스 AI는 AI 추론을 클라우드가 아닌 사용자 단말에서 처리하는 구조인 만큼, 제한된 전력·메모리 안에서도 품질을 유지할 수 있는 경량 모델이 우선적으로 요구
- 그러나 경량 모델만으로는 기능 구현이 어렵기 때문에, 이를 단말 내부에서 구동할 NPU·CPU·GPU와 모델 실행을 앱으로 연결하는 OS·런타임·개발도구가 함께 필요
- 특히 AI PC·스마트폰·스마트 안경처럼 사용 맥락이 다른 단말이 동시에 부상하면서, 단말별로 요구되는 모델 크기, 연산 장치, 메모리 구조, OS 통합 방식도 달라지는 양상
- 과거 PC·모바일 시기에는 하드웨어와 OS가 각 계층 안에서 우위를 겨루는 구조였으나, 온디바이스 AI에서는 모델·칩·OS가 동시에 맞물려야 기능이 작동한다는 점에서 경쟁의 구조 자체가 변화

- 결국 온디바이스 AI 전환은 모델 경량화 자체에 그치지 않고, 경량 모델과 단말·반도체, OS·개발도구가 함께 맞물려야 구현되는 실행 구조의 변화로 이해할 필요

AI 추론 실행, 중앙 집중에서 단말·엣지 분담 구조로 이동

클라우드 집중 추론, 비용·지연·전력 측면에서 한계 노출

- 생성형 AI 이용이 확대되면서 데이터센터의 GPU 수요, 전력 소비, 추론 비용이 동시에 상승하고 있으며, 중앙 클라우드에 추론을 집중시키는 구조의 운영 부담이 가시화
- 클라우드 추론은 요청 건수에 비례해 비용이 늘어나는 종량 구조이므로, AI 기능이 일상 서비스에 상시 내장될수록 운영 비용 부담이 커지는 양상
- 특히 음성 비서, 실시간 번역, 화면 맥락 인식처럼 즉각적인 응답이 필요한 기능에서는 네트워크 왕복에 따른 지연이 사용자 경험을 저하시키는 요인으로 작용
- 또한 개인정보와 업무 문서처럼 민감한 데이터를 외부 서버로 전송해야 하는 구조도 개인화 AI 확산 과정에서 보안 및 신뢰 부담으로 부각
- 이에 따라 모든 AI 기능을 중앙 클라우드에서 처리하는 방식은 비용, 지연, 전력, 데이터 보호 측면에서 한계가 커지고 있으며, 추론 실행 위치를 분산하려는 필요성이 확대

단말 실행, 저지연·오프라인·프라이버시·비용에서 보완 수단으로 부상

- (저지연) 단말 기반 실행은 네트워크 왕복 없이 기기 내부에서 추론을 처리하므로, 즉시 응답이 필요한 기능에서 클라우드 대비 지연시간을 줄이는 데 유리
- (오프라인) 또한 네트워크 연결이 불안정하거나 클라우드 접속이 제한되는 환경에서도 기기에 탑재된 모델을 통해 일부 AI 기능을 지속할 수 있어 서비스 연속성 확보 가능
- (프라이버시) 사진, 메시지, 일정, 위치 등 민감한 개인 데이터를 외부 서버로 전송하지 않고 기기 내부에서 처리할 수 있어, 개인정보 보호와 개인화 기능을 동시에 뒷받침
- (비용 효율) 비용 측면에서도 이용 빈도가 높은 반복 작업은 건별 클라우드 호출을 줄일 수 있어, 대규모 서비스 운영에서 비용 효율성을 높일 가능성
- 다만 단말은 연산 성능, 메모리, 배터리, 발열 등 물리적 제약이 크기 때문에 모든 AI 기능을 기기 안에서 처리하기는 어렵다는 한계가 존재

하이브리드 실행, 작업 배치가 서비스 비용과 품질을 좌우

- 온디바이스 AI는 모든 기능을 기기 안에서 처리하는 방식이 아니라, 작업

특성에 따라 단말과 클라우드가 추론 역할을 나누는 실행 구조이며 이는 각각의 강점과 한계가 다르기 때문

- 단말은 민감 데이터 처리, 즉시 응답, 오프라인 기능, 반복 업무에 적합하고, 클라우드는 장문맥 처리와 대규모 검색, 복잡 추론, 최신 정보 활용을 보완하는 역할을 수행
- 이에 따라 실제 AI 서비스는 간단하고 빈번한 작업은 단말에서 처리하고, 복잡하거나 최신 정보가 필요한 작업은 클라우드로 연결하는 방식으로 전개
- 이 구조에서는 개별 모델의 성능뿐 아니라, 어떤 작업을 어느 위치에서 실행할지 결정하는 작업 배치 전략이 서비스 비용과 응답 품질을 좌우
- 결국 하이브리드 실행의 핵심은 단말과 클라우드 중 하나를 선택하는 것이 아니라, 비용·지연 시간·보안·품질 조건에 따라 추론 처리 지점을 유연하게 조정하는 데 있음

모델·반도체·단말·플랫폼, 온디바이스 AI 실행 스택의 계층별 경쟁

(1) (모델) 경량 모델, 크기 경쟁을 넘어 단말 실행형 AI로 진화

○ 경량 모델 기준, 크기 축소에서 단말 실행 안정성 확보로 이동

- 최근 경량 모델은 초거대 모델을 단순히 작게 줄인 형태를 넘어, 특정 업무에서 충분한 품질을 낮은 전력과 비용으로 구현하는 목적형 모델로 진화
- 이에 따라 경량화의 핵심도 “얼마나 작은 모델인가?”보다 “단말 제약 안에서 충분한 품질과 낮은 지연, 안정적 실행을 함께 달성할 수 있는가?”로 이동
- 특히 온디바이스 AI가 확산되면서 모델은 크기뿐 아니라 어떤 입력을 다루고, 단말과 클라우드 중 어디에서 실행되며, 어떤 칩 환경에서 구동될 것인지까지 설계 단계에서 함께 고려하는 방향으로 전환
- 이러한 변화는 △용도별 특화, △멀티모달 확장, △하이브리드 실행 등 세 가지 축을 중심으로 구체화

① 범용 성능보다 특정 업무에 집중하는 용도별 특화: Microsoft Phi

- 최근 경량 모델은 범용 성능을 겨루기보다 추론·코딩·업무 보조 등 특정 기능에서 충분한 품질을 내도록 설계되는 방향으로 이동하는 추세로, Microsoft Phi 계열이 대표 사례
- Microsoft Phi 계열은 15B 규모의 소형 모델로도 수학·과학 추론, 화면 인식, 에이전트 작업에서 대형 모델에 근접하는 성능을 달성한 사례를 제시 (Microsoft Research, ’26.6.)

- 이는 모든 AI 기능에 대형 모델이 필요한 것이 아니라, 업무 성격에 따라 소형 모델로도 비용·전력·지연 측면의 효율을 확보할 수 있다는 점에서 모델 선택의 기준이 크기에서 용도로 이동하고 있음을 시사
- 특히 Phi-4-reasoning-vision은 학습 데이터를 대폭 줄이면서도 정확도와 연산 비용 간 균형을 개선한 점에서 소형 모델 경쟁이 데이터 품질과 목적형 학습 설계로 이동하고 있음을 보여줌

② 텍스트를 넘어 이미지·음성까지 다루는 멀티모달 확장: Google Gemma

- Phi가 특정 업무의 추론·코딩 성능 효율화에 초점을 맞췄다면, Google Gemma는 단말에서 활용되는 카메라·마이크·센서 등 다양한 입력을 경량 모델 안에서 함께 처리하는 방향을 제시
- Gemma 4 12B는 시각·음성 처리를 별도 인코더 없이 언어 모델 내부에서 수행하는 구조를 채택해, 16GB 메모리 환경의 일반 노트북에서도 실행이 가능한 경량 멀티모달 모델 사례로 부각 (Google, '26.6.)
- 이러한 구조는 입력 유형별 별도 모델을 조합하는 방식보다 구성 요소와 데이터 이동을 줄일 수 있어, 단말 환경에서 지연시간과 메모리 사용을 낮추는 방향과 연결
- 아울러 Gemma 4 QAT는 학습 과정에 양자화를 통합해 모바일 환경에서 E2B 모델의 메모리 사용량을 1GB 이하로 줄이면서도 품질 저하를 최소화하는 방향을 제시 (Google, '26.6.)

③ 단독 실행이 아닌 단말·클라우드 분담형 하이브리드 설계: Apple Intelligence

- Phi와 Gemma가 경량 모델의 성능 효율과 입력 확장 가능성을 보여준 사례라면, Apple Intelligence는 경량 모델이 서비스 구조 안에서 어떤 역할을 맡는가를 보여주는 사례
- Apple은 온디바이스 모델과 서버 모델을 구분하고, 기기 내 처리와 보안이 강화된 클라우드인 Private Cloud Compute를 결합하는 하이브리드 전략을 제시 (Apple, '26.6.)
- 이 구조에서 개인 맥락 이해, 화면 인식, 반복 업무 등 민감 데이터와 즉시 응답이 중요한 기능은 기기에서 처리하고, 최신 정보 확인이나 복잡한 추론은 클라우드가 보완
- 이는 경량 모델이 클라우드 대형 모델을 전면 대체하는 것이 아니라, 단말과 클라우드의 역할 분담을 전제로 서비스 품질과 개인정보 보호를 함께 조정하는 사례

(2) (반도체) 단말 AI 실행, 단일 성능 수치를 넘어 칩별 실행 효율이 관건

- 단말 AI 실행 구조가 클라우드 호출에서 칩 최적화로 이동
  - 경량 모델이 마련되더라도 그 자체만으로 곧바로 서비스가 구현되는 것은 아니며, 이를 기기 내부에서 구동할 반도체 기반이 갖춰져야 실제 사용자 기능으로 작동
  - 이에 따라 온디바이스 AI 경쟁은 모델 경량화 단계를 넘어, 기기별 NPU·GPU·CPU와 메모리 조건에 맞춰 모델을 변환하고 안정적으로 구동하는 실행 경쟁으로 확대
  - 단말은 데이터센터와 달리 전력·발열·배터리·메모리 용량이 모두 제약으로 작용하는 만큼, 칩의 역할도 단순 연산 성능 제공을 넘어 제한된 환경 안에서 실행 효율을 관리하는 방식으로 확대
  - 즉 AI 반도체 경쟁의 핵심은 TOPS 같은 단일 성능 수치를 넘어, 실제 기기 안에서 지연시간, 전력 소모, 발열, 메모리 병목을 얼마나 안정적으로 제어하느냐로 이동
- NPU, 상시 작동 AI 기능을 저전력으로 처리하는 기본 장치로 자리매김
  - 화면 자막, 음성 인식처럼 반복적으로 작동하는 AI 기능을 범용 장치인 CPU·GPU로 처리하면 전력 소모와 발열이 커지므로, 이를 낮은 전력으로 지속 처리할 전용 장치의 필요성이 확대
  - NPU는 신경망 추론 연산에 특화된 저전력 연산 장치로, 처리에 필요한 데이터 이동을 줄여 반복 작업을 더 적은 전력과 짧은 지연으로 수행하는 데 유리
  - 이러한 효율 덕분에 NPU는 배터리로 구동되는 단말에서도 화면 자막, 음성 인식과 같은 AI 기능을 상시 유지하는 기본 연산 장치로 확산
  - 특히 AI PC와 모바일 기기에서 NPU 탑재가 기본 요건으로 자리 잡으면서, 저전력 추론 성능은 고급 단말의 차별 요소를 넘어 온디바이스 AI를 구현하는 필수 조건으로 부상
  - 다만 NPU는 단말의 모든 AI 작업을 대체하기보다, 저전력으로 반복·실시간 추론을 안정적으로 처리하는 실행 축에 특화되어 있으며, 고부하 작업은 GPU가 보완하는 역할 분담이 전제
- GPU, 고부하 로컬 AI를 구동하는 고성능 축으로 부상
  - NPU가 소규모·반복 상시 작업에 적합하다면, 대형 모델을 기기에서 직접 구동하거나 영상· 이미지를 생성하는 고부하 작업은 더 높은 연산력과 대용량 메모리를 요구
  - 이에 따라 고성능 GPU 기반 단말은 외부 클라우드 호출 없이 로컬에서 대규모

모델을 자체 실행하려는 개발자·창작자·기업 수요를 중심으로 부상

- NVIDIA는 '26년 6월 RTX Spark를 발표하며 고성능 GPU에 대용량 메모리를 결합해, 대형 모델을 외부 클라우드 호출 없이 단말에서 직접 구동하는 사례를 제시(NVIDIA, '26.6.)
- 기존 PC에서는 GPU 전용 메모리와 시스템 메모리가 분리돼 데이터 복사와 이동 지연이 발생했으나, 통합 메모리는 GPU와 CPU가 같은 메모리 공간을 활용하도록 해 병목을 완화
- 이러한 연산·메모리 요구 차이로 단말 AI 시장은 저전력 NPU 중심 단말과 고성능 GPU 중심 단말이 병존하고, 작업 부담과 사용 목적에 따라 실행 장치가 나뉘는 방향으로 분화
- 다만 높은 가격과 메모리 공급 제약, 수요 검증의 불확실성으로 고성능 GPU 기반 단말은 단기적으로 개발자·창작자 등 전문 사용자층을 중심으로 제한적으로 확산될 전망 (IDC·Reuters, '26.6.)

● 단말 AI 반도체, 단일 칩이 아닌 CPU·GPU·NPU의 역할 분담 구조로 정착

- 앞서 살핀 NPU와 GPU는 서로 다른 AI 작업에 특화된 장치이지만, 실제 단말 AI 작업은 저전력 추론, 대규모 연산, 앱·OS 작업 조율이 함께 작동하므로 단일 장치만으로 처리하기 어려운 구조
- 이에 따라 CPU는 OS·앱 실행과 전체 작업 조율을 맡고, GPU는 대규모 병렬 연산이 필요한 생성 작업을, NPU는 저전력 반복 추론을 담당하는 방식으로 동일 기기 안에서 역할이 분화
- 이처럼 AI 작업이 CPU, GPU, NPU로 나뉘어 배정되면서, 한 장치의 성능만으로 전체 서비스 품질이 결정되지 않으며 여러 장치가 함께 작동하는 이종 연산 구조가 중요
- 특히 AI 에이전트처럼 여러 단계의 판단과 도구 호출, 앱 실행을 연계하는 작업에서는 연산 장치 간 작업 배분과 메모리·전력·발열 조율이 실제 체감 성능을 좌우하는 핵심 요소로 부상
- 이로써 반도체 경쟁은 단일 칩의 연산 성능 비교를 넘어, 기기별 하드웨어 제약 내에서 모델을 효율적으로 배분하고 안정적으로 실행하는 이종 연산 최적화 역량으로 확장

(3) (단말) 온디바이스 AI, 생산성·개인화·실시간 인식으로 단말별 영역 분화

○ 입력 데이터 차이가 단말별 AI 역할을 가르는 출발점으로 작용

- 반도체 실행 역량이 경량 모델을 기기 내부에서 구동하는 기반이라면, 단말 확산은 그 실행 결과가 여러 사용자 접점에서 구체적인 제품 및 기능으로 구현되는 단계
- 단말마다 주로 다루는 데이터와 입력 방식이 다르기 때문에, 같은 온디바이스 AI라도 각 기기에 적합한 기능이 달라지며 활용 영역이 다르게 형성 (IDC, '26.3.)
- 이에 PC는 업무 문서와 회의·코드 등 생산성 데이터를, 스마트폰은 사진·위치·연락처·일정 등 개인 맥락 데이터, 스마트 안경은 실시간 시야·음성·공간 맥락 정보를 각각 중심으로 활용하며 특화 (Forrester, '26.3.)
- 이처럼 온디바이스 AI는 하나의 단말로 수렴하기보다 AI PC, AI 스마트폰, 스마트 안경 등 여러 접점으로 확산되며, 단말별 사용 맥락에 따라 서로 다른 초기 시장을 형성

① AI PC, 생산성 서비스와 직결되는 온디바이스 AI의 1차 승부처로 부상

- AI PC는 문서, 회의 정리, 코딩, 협업 도구 등 일상 업무 기능과 직접 연결되는 단말로, 온디바이스 AI가 실제 업무 환경에서 곧바로 활용되는 대표 채널로 부각
- 기존에는 업무 자료를 클라우드로 전송해 처리해야 했으나, AI PC는 문서 요약이나 검색 보조 등 일부 작업을 기기 내부에서 바로 처리할 수 있어 빠른 응답성과 민감한 업무 자료 보호가 핵심 강점
- 이러한 강점을 바탕으로 초기 AI PC 시장은 낮은 전력으로 반복적인 AI 작업을 처리하는 NPU 탑재 프로세서를 중심으로 형성됐으며, 주요 기업들도 해당 기능을 핵심 차별점으로 제시
- 다만 AI PC가 처리하는 기능이 단순 보조 작업을 넘어 문서 이해, 이미지 생성, 코딩, 로컬 에이전트 실행 등으로 넓어지면서, 더 높은 연산 성능과 메모리 용량을 요구하는 영역도 확대
- 이러한 흐름 속에서 '26년 6월 NVIDIA가 Arm 기반 CPU와 고성능 GPU를 결합한 PC용 플랫폼으로 진입하며, AI PC 경쟁은 고성능 AI 작업까지 포괄하는 방향으로 확장
- 그 결과 AI PC 시장은 가벼운 보조 기능을 맡는 전력 효율 중심 단말과 복잡한 AI 작업을 맡는 고성능 연산 중심 단말로 분화하며, 단말 AI 경쟁이 가장 먼저 본격화되는 1차 승부처로 부상

② AI 스마트폰, 개인 맥락의 OS 통합 수준이 개인화 경쟁의 관건으로 부상

- 스마트폰은 사용자가 가장 자주 휴대하는 단말로, 카메라, 위치, 연락처, 일정, 메시지 등 개인 맥락 데이터가 지속적으로 축적돼 개인화 AI가 가장 직접적으로 구현되는 핵심 채널로 기능
- 이러한 데이터를 바탕으로 사진·영상 편집, 통화·문자 요약, 실시간 번역, 화면 맥락 이해 등 소비자가 일상에서 곧바로 활용할 수 있는 기능이 여러 서비스 장면으로 확산
- 이때 경쟁의 초점은 AI 기능의 수를 늘리는 데 그치지 않고, 개인 맥락 데이터를 OS와 기본 서비스에 얼마나 자연스럽게 연결해 별도 조작 없이 활용하게 만드느냐로 이동
- Apple은 Siri AI를 고도화하면서 AI를 별도 앱이 아니라 사진·메시지·Safari 등 기본 서비스에 흡수해, OS 전반을 가로지르는 개인화 인터페이스로 확장하는 방향을 제시 (Apple, '26.6.)
- 이러한 통합은 민감 데이터는 기기에서 처리하고 복잡한 추론은 보안 클라우드가 맡는 하이브리드 실행을 전제로, 단말 성능 제약을 경량 모델과 클라우드 연동으로 함께 보완하는 구조
- Android 진영에서도 개인 맥락 데이터를 OS에 결합하는 흐름이 나타나면서, AI 스마트폰 경쟁은 개별 앱 기능을 겨루는 단계를 넘어 OS 차원의 개인화 경험 경쟁으로 전개

③ 스마트 안경, 실시간 시야 및 음성 인식으로 차세대 AI 접점 형성

- 스마트폰이 필요할 때 손에 들고 화면을 조작하는 단말이라면, 스마트 안경은 착용 상태에서 카메라, 마이크, 디스플레이, 센서를 통해 시야와 음성을 인식하는 상시형 인터페이스로 부상
- 이러한 착용형 구조는 길 안내, 통역 등 실시간 맥락을 활용한 AI 기능을 별도 화면 조작 없이 음성으로 수행하도록 해, 손과 시선을 자유롭게 유지하는 핸즈프리 경험으로 차별화
- 다만 배터리, 발열, 무게 같은 하드웨어 제약과 함께, 카메라가 상시 작동할 수 있다는 점에서 프라이버시 우려와 사회적 수용성 문제가 스마트 안경 확산의 핵심 병목으로 작용 (WIRED, '26.6.)
- 이러한 우려를 해소하기 위해 민감한 시각·음성 데이터를 기기 안에서 처리하는 온디바이스 실행과, 데이터 접근 범위를 제한하는 OS·앱 권한 체계가 확산의 전제 조건으로 부각

- 따라서 스마트 안경은 단기간에 대중화될 단말이라기보다, 온디바이스 AI가 스마트폰과 PC를 넘어 공간컴퓨팅 및 웨어러블 영역으로 확장될 가능성을 보여주는 차세대 접점으로 주목

(4) (생태계) 플랫폼 경쟁, 모델·칩·단말을 하나로 묶는 실행 생태계 경쟁으로 수렴

○ 개발·실행 도구, 모델·칩·단말을 실제 서비스로 연결하는 실행 기반으로 부상

- 온디바이스 AI는 모델 성능이나 칩 사양만으로 구현되지 않으며, 기기 안의 데이터, 앱 기능, 연산 장치, 보안 권한이 OS와 런타임 안에서 함께 연결되어야 실제 사용자 경험으로 작동
- 이에 따라 플랫폼 경쟁의 초점은 개별 AI 기능을 추가하는 단계를 넘어, AI 가 사진·문서·일정·메시지 등 기기 내 데이터와 기본 앱을 얼마나 자연스럽게 활용하도록 설계하느냐로 이동
- Apple은 운영체제·칩·기본 앱을 긴밀히 결합해 개인 맥락 기반 AI 경험을 자사 기기 안에서 통합하고 Google은 Android와 Gemini를 결합해 외부 앱까지 포함하는 AI 경험 확장을 추진
- Microsoft는 Windows AI API와 Copilot+ PC를 중심으로 업무용 PC에서 앱·파일·네트워크 접근을 조율하는 방식으로, 개인용 단말과 다른 기업용 AI 플랫폼 전략을 전개
- 결국 OS·런타임·개발도구는 온디바이스 AI가 일부 기업의 자체 기능에 머물지, 다양한 앱과 서비스로 확산될 수 있을지를 좌우하는 핵심 실행 기반으로 작용

○ 권한·보안·클라우드 연동, 개인 데이터 활용과 신뢰 확보의 전제 조건으로 정착

- 온디바이스 AI는 사진, 문서, 일정, 메시지 등 기기 안의 민감 데이터를 활용해 개인화·업무 보조 기능을 구현하므로, 데이터 접근 범위와 실행 권한을 통제하는 체계가 필수
- AI가 여러 앱의 기능을 대신 수행할수록 권한 오남용, 정보 유출, 에이전트 오작동을 통제할 수 있는 보안 설계가 중요해지며, 권한 밖의 파일·네트워크 접근을 차단하는 구조가 요구
  · (Apple) 개인용 단말에서 운영체제·칩·기본 앱을 자사가 직접 설계해 하나로 묶고, 외부 개방을 제한한 폐쇄형 구조 안에서 개인 데이터 접근을 일원적으로 통제 (Apple, '26.6.)
  · (Google) 개인용 단말에서 외부 앱과의 폭넓은 연결을 허용하는 개방형 구조를 택하되, 데이터 접근 시점마다 사용자 동의를 요구해 최종 통제권을 사용자에게 부여 (Google I/O, '26.6.)

· (Microsoft) 업무용 PC에서 에이전트의 파일·네트워크 접근을 운영체제 차원에서 제한하되, 허용 범위를 기업 IT 관리자가 중앙에서 설정하는 관리자 중심 통제 방식을 적용 (Microsoft, '26.6.)

- 종합하면 온디바이스 AI 생태계 경쟁의 핵심은 모델·칩·단말의 개별 성능을 넘어, 이를 안전하고 효율적인 사용자 경험으로 연결하는 실행 생태계 장악력에 의해 좌우

출처 : IDC 외(2026.6.)

https://www.idc.com/resource-center/blog/intelligent-devices-mwc-2026/
https://www.forrester.com/blogs/mobile-world-congress-2026-from-5g-hype-to-ai-hope/
https://www.forrester.com/blogs/from-endpoints-to-orchestration-hps-plan-to-redefine-the-future-of-work/
https://newsroom.arm.com/blog/five-takeaways-from-the-moor-insights-and-strategy-report
https://www.couchbase.com/blog/on-device-ai/
https://www.spheron.network/blog/hybrid-cloud-edge-ai-inference-guide/
https://www.mindstudio.ai/blog/on-device-ai-vs-cloud-ai-economics
https://www.ibm.com/think/topics/edge-vs-cloud-ai
https://blog.google/innovation-and-ai/technology/developers-tools/introducing-gemma-4-12b/
https://developers.googleblog.com/blazing-fast-on-device-genai-with-litert-lm/
https://android-developers.googleblog.com/2026/05/android-ai-intelligence-system.html
https://learn.microsoft.com/en-us/windows/ai/apis/
https://security.apple.com/blog/expanding-pcc/
https://developer.apple.com/apple-intelligence/whats-new/
https://www.wired.com/story/meta-removes-face-recognition-code-meta-ai-app-smart-glasses/
https://counterpointresearch.com/en/insights/counterpoint-conversations-agi-on-device-ai-agents-replace-apps-on-smartphone
https://www.reuters.com/world/china/nvidias-ai-pc-push-banks-unproven-demand-beyond-niche-users-2026-06-08/
https://nvidianews.nvidia.com/news/nvidia-dgx-station-for-windows-puts-a-trillion-parameter-ai-supercomputer-on-every-enterprise-desk