삼정 KPMG

# Issue Monitor

## 생성형 AI에게 펼쳐진 새로운 무대, 온디바이스 AI

June 2024 | 제165호


삼정KPMG 경제연구원

home.kpmg/kr

## 생성형 AI에게 펼쳐진 새로운 무대, 온디바이스 AI

Issue Monitor | June 2024

# Contacts


**삼정KPMG 경제연구원**

**최창환**
책임연구원
T 02-2112-7438
E changhwanchoi@kr.kpmg.com

**이종민**
선임연구원
T 02-2112-7815
E jlee547@kr.kpmg.com

**이효정**
상무
T 02-2112-6744
E hyojunglee@kr.kpmg.com

**류승희**
선임연구원
T 02-2112-7469
E seungheeryu@kr.kpmg.com

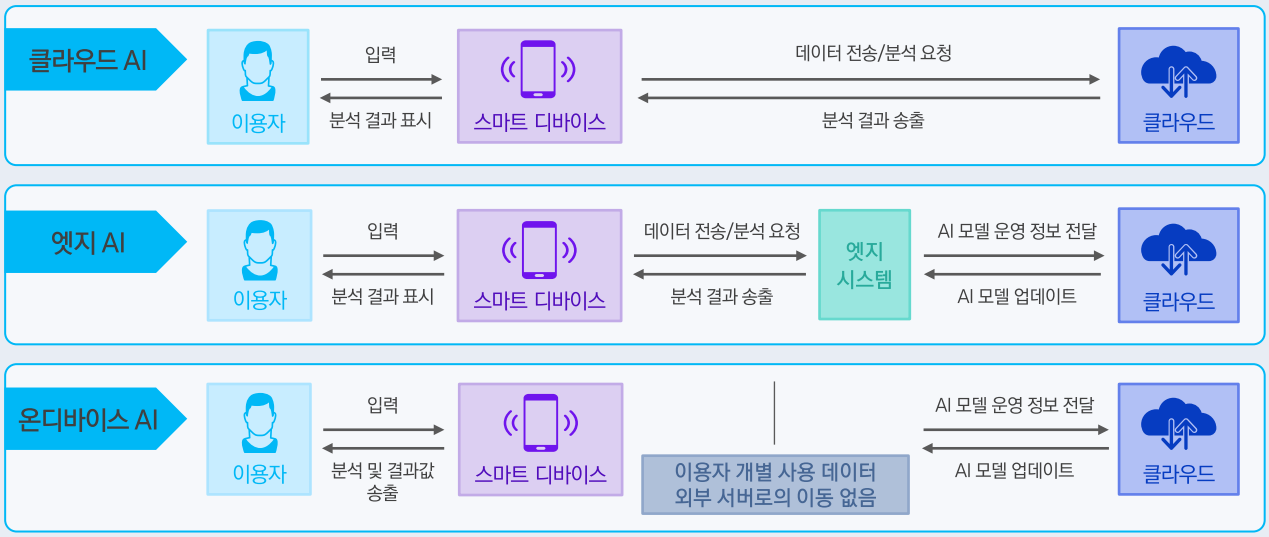
# Contents

생성형 AI의 기술력이 발전하고 다양한 서비스가 등장하여 AI 시대의 가속화가 촉진되고 있다. 클라우드와 데이터센터, AI 반도체, 대규모 언어 모델 등 생성형 AI 시대의 성공을 이끌고 있는 AI 산업 인프라의 생태계에 새로운 변화가 전망되고 있다. 우수한 컴퓨팅 인프라를 갖춘 디바이스와 경량화를 통해 효율성을 높인 소형 AI 모델을 중심으로 관심도가 높아지고 있는 온디바이스 AI 기술의 특징과 활용 사례를 분석하였다. 생성형 AI의 새로운 플랫폼이 될 온디바이스 AI 시대에 Scale-up을 위한 기업의 여정에 도움이 되기를 기대한다.


| | Page |
|---|---|
| **Infographic Summary** | **3** |
| **AI의 넥스트 스테이지, 온디바이스 AI** | **4** |
| 곁으로 다가온 AI(인공지능)의 시대 | 4 |
| AI의 발달과 생성형 AI 시대의 개막 | 5 |
| 생성형 AI 시스템의 구성 요소 | 7 |
| [ Issue Brief ] 국내외 주요 대형 AI 모델 개발 동향 | 11 |
| **왜 온디바이스 AI인가** | **13** |
| 대형 AI 모델과 클라우드 기반 AI의 한계 | 13 |
| 온디바이스 AI의 부상 | 15 |
| 온디바이스 AI의 구조 | 16 |
| 온디바이스 AI의 활용 방식 | 17 |
| 온디바이스 AI의 운영 요소 | 18 |
| [ Issue Brief ] 국내외 주요 소형 AI 모델 개발 동향 | 20 |
| 온디바이스 AI가 탑재된 제품 및 Use case | 23 |
| **온디바이스 AI 시대 Scale-up을 위한 전략** | **28** |
| 온디바이스 AI 시대의 주목해야 할 키워드 'SCALE' | 28 |
| S(Semiconductor): 저전력 AI 반도체 성장 속 창출되는 기회 발굴 | 29 |
| C(Cloud): 클라우드의 역할 변화로 인한 인프라 시장 변화 대응 | 30 |
| A(Ambient Computing): 앰비언트 컴퓨팅 가속화 속 새로운 시장 모색 | 31 |
| L(Language Model): AI 모델 역할 변화 속 빅테크·스타트업 움직임 주목 | 32 |
| E(Explainable AI): 기술적 데이터 통제 방안 마련으로 설명 가능한 AI 시스템 구축 | 33 |

# Infographic Summary

## AI 운영 방식의 변화와 온디바이스 AI의 장점




### 1. AI 모델 분석 속도 향상

- 온디바이스 AI는 외부 통신 없이 이용자가 사용하는 스마트 디바이스 내부에서 분석 진행
- 분석 속도 향상 기대

### 2. 외부 시스템 비용 감소

- 외부 클라우드 및 데이터센터를 모델 업데이트 용도로만 사용
- 클라우드 및 데이터센터 이용 비용 감소 효과 기대

### 3. 데이터 보안 우려 감소

- 이용자의 개별 사용 데이터가 디바이스 외부로 나가지 않음
- 외부 전송 및 저장 과정에서 발생할 수 있는 데이터 보안 우려 적음

Source: 언론보도 종합, 삼정KPMG 경제연구원

## 온디바이스 AI 시대 Scale-up을 위한 전략


온디바이스 AI 시대에는 고성능 반도체 중심이던 AI 반도체 시장에서 저전력 반도체 중심의 반도체 성장세의 주목해야 함

클라우드는 AI 서비스 운영에 소모되는 역할 비중을 줄이고, 대형 AI 모델의 성능 향상을 지원하는 역할에 집중해야 함

온디바이스 AI의 높은 지능과 우수한 데이터 보안 능력이 촉발하는 앰비언트 컴퓨팅 시장 활성화 속 기회를 발굴해야 함

고지능을 위한 학습 중심의 대형 AI 모델과 효율적 서비스 운영을 위한 소형 AI 모델로 양분화된 시장 속 적합한 활용 모델을 모색해야 함

이용자 밀착형인 온디바이스 AI 모델은 안정적 운영을 위한 기술 통제 방안을 필수적으로 마련해야 함

Source: 삼정KPMG 경제연구원

# AI의 넥스트 스테이지, 온디바이스 AI

> 온디바이스 AI는 이용자가 사용하는 디바이스 자체에서 데이터를 처리하고 AI 모델을 구동할 수 있도록 하는 더 작은 개념의 운용 방식을 의미

챗GPT가 촉발한
초거대 AI 비즈니스 혁신


보다 자세한 내용을
원하시면 보고서
이미지를 클릭하시거나
QR코드를 스캔해주세요.

## 곁으로 다가온 AI(인공지능)의 시대

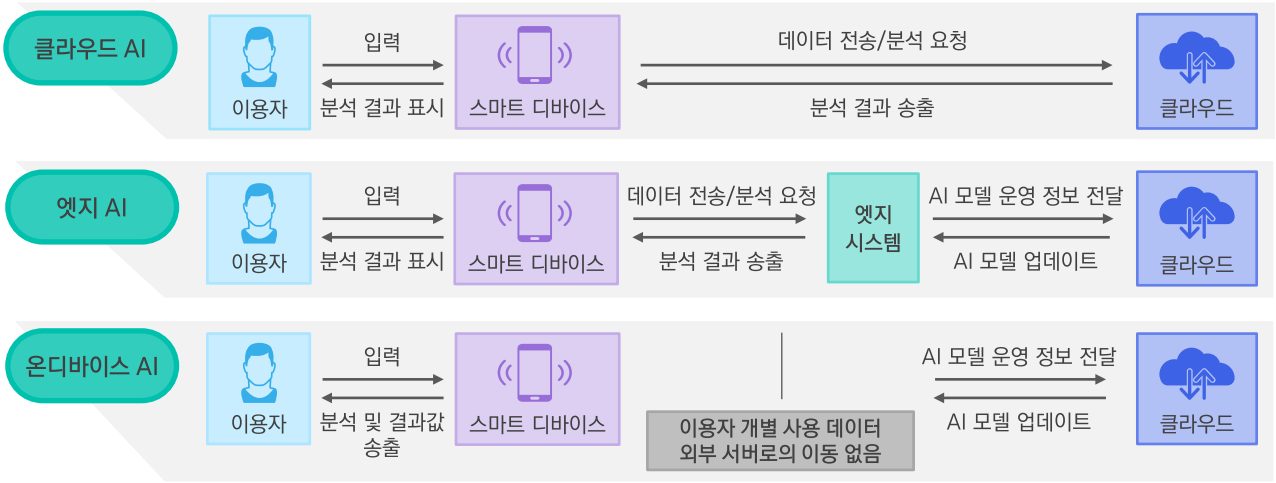
오픈AI의 챗GPT, 구글의 Gemini(구 Bard), 마이크로소프트의 Copilot 등 현재 많은 화제를 모으고 있는 생성형 AI 기반의 AI 서비스는 대규모 언어 모델(Large Language Model)을 기반으로 운영되며, 클라우드 시스템을 활용하고 있다. 클라우드를 기반으로 운영되는 AI는 이용자가 사용하는 스마트 디바이스(노트북, 스마트폰 등)를 통해 원하는 명령어를 입력하면 스마트 디바이스가 해당 데이터를 클라우드로 전달한다. 클라우드는 전달 받은 데이터를 클라우드 내의 프로세싱 역량을 활용하여 AI 모델의 분석 결과를 도출하고 이를 스마트 디바이스로 다시 전달하여 이용자가 원하는 결과값을 확인할 수 있도록 제공한다.

클라우드 AI와 함께 기존 AI 산업에서 많은 관심을 받은 AI 운영 방식으로 엣지(Edge) AI가 있다. 엣지 컴퓨팅에서 파생된 의미로 먼 거리에 위치한 클라우드 서버를 통한 데이터 통신보다 이용자와 더 가까운 곳에서 데이터를 처리하는 방식인 엣지 AI는 이용자와 인접한 장소에 엣지 서버를 두어 데이터를 처리하고 AI 모델의 운영을 통한 데이터를 송출한다.

엣지 AI는 데이터 이동 과정이 짧아지고 단순화되는 특성을 가지고 있어 안정적인 시스템 운영과 빠른 데이터 처리가 장점으로 주목되고 있다. 보안 등 민간성이 높은 데이터를 처리하는 경우에도 개별 엣지 서버를 통한 데이터를 기록 및 처리한다는 특성 덕분에 클라우드 기반 방식보다 높은 보안 안정성을 보인다고 평가받는다.

최근에는 이러한 엣지 AI의 장점을 더욱 극대화할 수 있는 방식인 온디바이스 AI(On-Device AI)가 새롭게 주목 받고 있다. 엣지 AI가 이용자와 가까운 곳에 위치한 엣지 서버를 통한 AI 모델의 구동까지 포함하는 개념인 반면, 엣지 AI의 일종인 온디바이스 AI는 이용자가 사용하는 디바이스 자체에서 데이터를 처리하고 AI 모델을 구동할 수 있도록 하는 더 작은 개념의 운용 방식을 의미한다.

[ AI 운영 방식의 변화 ]




Source: 언론보도 종합, 삼정KPMG 경제연구원

## AI의 발달과 생성형 AI 시대의 개막

> 빅테크 기업 등을 포함한 다양한 기업들의 AI 시장 진입으로 AI 생태계가 빠른 속도로 확대

1956년 최초로 인공지능이라는 개념이 등장한 이후, 반세기가 조금 넘는 기간 동안 AI 기술의 발전은 놀라운 속도로 이루어져왔다. 1996년 세계 체스 챔피언과의 대국에서 승리한 IBM의 AI '딥블루', 2016년 이세돌 9단과의 바둑 시합을 이긴 구글 딥마인드의 AI '알파고' 등 AI가 세상을 놀라게 하는 일이 종종 있었지만, 최근 AI가 가장 큰 충격을 가져온 이슈를 꼽으라면 챗GPT와 함께 나타난 생성형 AI(Generative AI)의 등장을 꼽을 수 있다.

2022년 11월 챗GPT로 큰 관심을 받기 시작한 생성형 AI는 기존에 화제가 된 AI 기술이 바둑, 체스 등 정해진 틀 안에서 특정한 태스크를 수행하며 인간을 넘어서는 놀라운 가치를 보여줬던 것과 달리, 정해진 틀을 넘어서 이용자가 원하는 다양한 태스크를 수행할 수 있다는 확장성 면에서 많은 화제를 모았다. 생성형 AI도 AI라는 개념이 생긴 초창기에서 오래 지나지 않은 1960년대부터 정의된 개념이지만, 2014년 생성형 AI의 핵심 기술이라고 불리는 GAN(Generative Adversarial Network) 등의 머신러닝(Machine Learning) 알고리즘 기술이 개발된 이후 빠른 속도로 진화하기 시작했다.

2017년 구글이 공개한 Transformer는 단어의 의미와 문장 내 단어 간 관계를 분석하는데 특화된 딥러닝 구조로 개발되어 생성형 AI 시대가 다가오는데 큰 역할을 한 것으로 평가된다. 언어를 생성하는데 특화된 Transformer를 활용하여 오픈AI의 GPT 모델 등 생성형 AI 서비스의 기반이 되는 모델이 등장하였다.

오픈AI가 만든 챗GPT, 구글이 만든 Gemini 등의 생성형 AI 서비스가 텍스트 기반으로 문서 요약, 창작 등의 다양한 콘텐츠를 생성해내며, 지난 2023년은 생성형 AI의 가능성을 확인하는 기회가 되었다. 빅테크 기업 등을 포함한 다양한 기업들의 AI 시장 진입으로 AI 생태계가 빠른 속도로 확대되고 있다. AI 모델은 이 기간 동안 우수한 기술력을 가진 기업들의 AI 시장 진입이 확대되며 더 큰 규모의 모델로 진화하였다.

[ 딥러닝의 발전 단계 ]

| 딥러닝(Deep-Learning) | | | |
|---|---|---|---|
| **CNN** (Convolutional Neural Network) | **RNN** (Recurrent Neural Network) | **GAN** (Generative Adversarial Network) | **Transformer** |
| • 1980년대 등장한 개념으로 Convolutional Layer를 거치며 부분별 이미지가 가지는 패턴을 파악하고 데이터를 분석함<br>• 패턴분석을 통한 영상과 이미지를 인식하는 데 특화 | • 1980년대 등장한 개념으로 데이터의 입력과 출력을 순환적 구조로 해결<br>• 문장의 흐름을 이해하고 단어별 분류 기능을 탑재하여 자연어 처리에 특화 | • 2014년에 발표된 개념으로 정보를 생성하는 기능과 판별하는 기능이 서로 대립하여 기능을 향상시키는 모델<br>• 사전 훈련된 데이터를 기반으로 주어진 데이터를 판별 및 실제와 가장 유사한 정보를 생성 | • 2017년 구글이 공개한 모델로 단어의 의미와 문장 내 단어 간 관계를 분석<br>• 언어의 구조적 특성을 이해하고 언어를 생성하는 데 특화 |

Source: 언론보도 종합, 삼정KPMG 경제연구원

> 많은 인프라 자원을 활용하는 클라우드 AI와 달리, 온디바이스 AI는 디바이스에서 AI 모델이 직접 구현되어 이용자 맞춤형으로 사용 가능

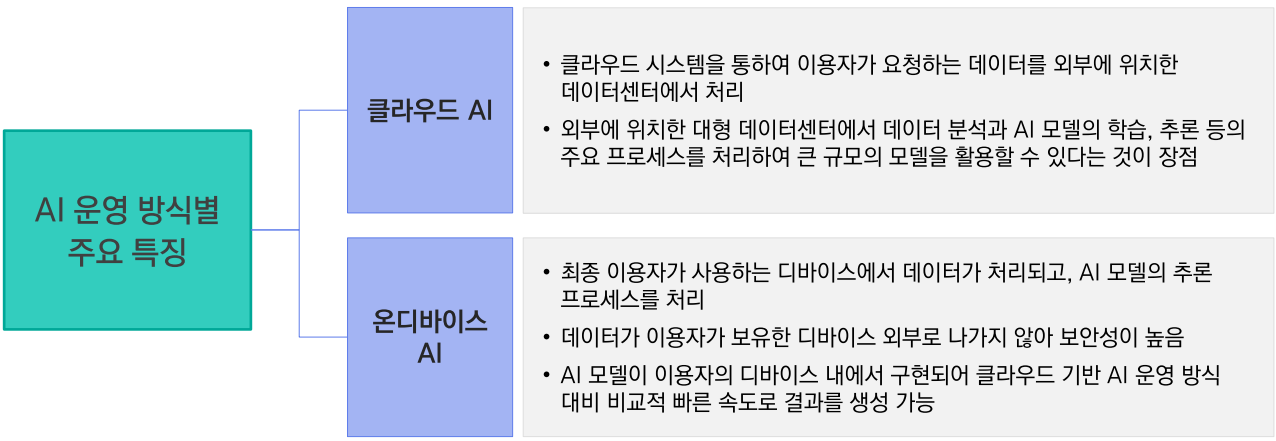
챗GPT(오픈AI), Gemini(구글), Copilot(마이크로소프트) 등 많은 이용자를 확보하여 인지도가 높은 주요 생성형 AI 서비스의 대부분은 클라우드를 기반으로 구현되고 있다. 이용자가 해당 서비스를 이용하여 생성형 AI가 수행하기를 원하는 태스크를 입력하면, 해당 데이터가 클라우드를 통하여 외부 서버에 위치한 AI 모델로 접수되어 분석되는 과정을 거친다. 이후, AI 모델이 분석하여 도출된 결과물이 다시 클라우드 서버를 통하여 이용자에게 송출된다.

네트워크 기술의 발전을 통하여 빠른 데이터 처리가 가능해지고, 고성능 반도체를 기반으로 구현된 서버 기술을 통하여 빠른 분석이 가능해짐에 따라 생성형 AI를 이용하는 사용자가 이러한 데이터의 이동과정을 체감하기 어려울 정도로 빠른 처리가 가능하다. 하지만 클라우드를 기반으로 한 생성형 AI 서비스를 운영하기 위해서는 AI 모델을 구동하기 위한 다량의 데이터 처리가 요구되며 하나의 결과물을 만들어내기 위하여 많은 파라미터(Parameter)를 거치는 분석이 필요한 AI 모델의 특성상 많은 비용과 자원의 사용이 필요하다.

높은 효용 가치를 제공할 수 있지만 많은 인프라 자원을 활용하는 클라우드 기반의 AI 서비스와 달리, 온디바이스 AI는 이용자가 사용하는 디바이스에서 AI 모델이 직접 구현되어 이용자 맞춤형으로 사용이 가능하다는 장점이 있다. 온디바이스 AI는 이용자의 요청에 대한 답을 생성하기 위한 추론 과정이 클라우드를 거치지 않고 이용자의 디바이스 내부에 탑재된 AI 모델에서 이루어져 데이터의 전달이 필요하지 않고 빠른 결과 생성이 가능하다는 특징을 가진다.

AI 기능이 탑재된 스마트폰 '갤럭시 S24'와 구글의 '픽셀8' 등 온디바이스 AI의 빠른 확산이 이루어지고 있는 스마트폰 외에도, TV, 냉장고, 세탁기 등 생활에 밀접한 분야의 다양한 기기가 온디바이스 AI 기능을 통하여 새로운 가치를 제공하는 혁신이 이루어지고 있다.

[ 클라우드 AI와 온디바이스 AI 운영 방식별 특징 및 장점 ]




Source: 언론보도 종합, 삼정KPMG 경제연구원

AI 반도체를 탑재하여 고성능을 보이는 데이터센터에 대한 수요가 확대 … 빅테크 기업 중심으로 데이터센터와 클라우드 시스템 투자 지속

## 생성형 AI 시스템의 구성 요소

### 클라우드: 대용량의 데이터를 처리할 수 있는 고성능 클라우드의 수요 증가

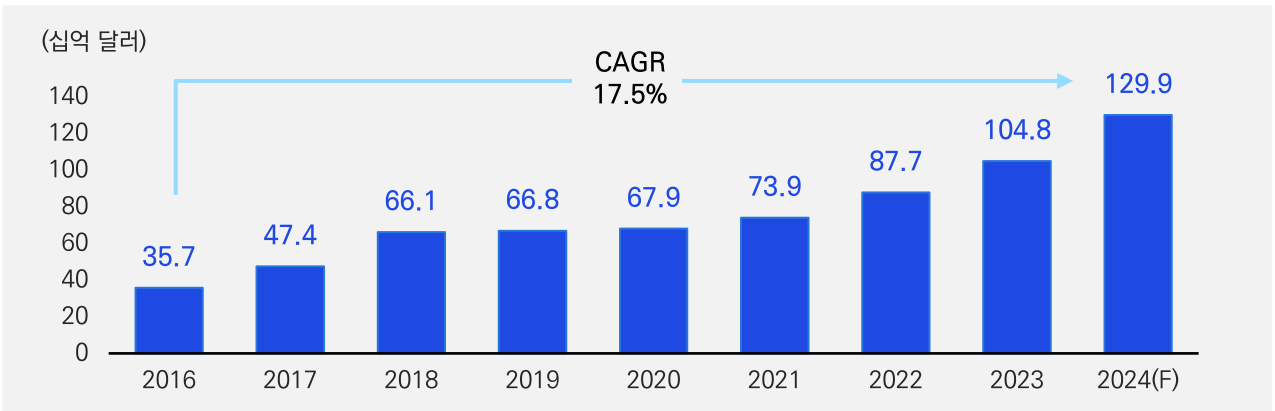
대규모 언어 모델을 기반으로 한 AI 시대가 확장됨에 따라, AI 모델을 학습시키고, 추론의 과정을 통하여 결과물을 낼 수 있도록 하는 인프라의 확보 차원에서 클라우드의 역할이 더욱 강조되고 있다. 클라우드는 데이터를 저장하는 서버의 개념뿐 아니라 데이터를 처리하여 전송하는 시스템으로, 회사에 위치해 있던 중앙 서버가 마치 구름 위에 위치한 듯 하다고 여겨진다는 의미로 클라우드(Cloud)라는 이름으로 불리고 있다.

클라우드 시스템을 활용한 컴퓨팅 방식은 클라우드라는 인터넷 기반의 가상의 컴퓨팅 공간에서 네트워크망을 통해 실시간으로 데이터센터 등 외부에 위치한 컴퓨팅 역량과 연동하는 특징을 지닌다. 이를 통해 자체 보유한 컴퓨팅 자원 이상을 활용할 수 있도록 하는 서비스 체계이다.

현재 다수의 대형 AI 모델은 클라우드를 기반으로 운영되고 있으며, 클라우드는 AI 모델의 학습을 위한 데이터의 저장 및 AI 모델 운영 과정에서 필요한 추론 등의 대규모 연산 과정을 수행하여 산출물을 송출하는 역할까지 수행 중이다. 클라우드 컴퓨팅은 기존 컴퓨팅 시스템 대비 개별 이용자가 컴퓨팅 역량을 충분히 갖추지 않아도 외부의 자원을 활용하여 처리할 수 있다는 점에서 다량의 데이터 처리를 요하는 분야에서 높은 관심을 받고 있다.

시장조사기관 IDC에 따르면, 2023년 글로벌 클라우드 인프라 투자 규모는 1,048억 달러로 전년 대비 19.5% 성장한 수준을 기록한 것으로 분석됐다. 2024년에도 클라우드 분야의 인프라 투자가 지속 증가하여 2023년 대비 24.0% 늘어난 1,299억 달러를 기록할 것으로 전망되었다.

[ 글로벌 클라우드 인프라 투자 규모 전망 ]




Source: Statista, IDC, 삼정KPMG 경제연구원 재구성
Note: (F)는 전망치

> 대용량의 데이터를 처리할 수 있는 역량을 갖춘 데이터센터 보유 … 빅테크 기업에서 운영하는 클라우드 서비스의 경쟁력 확보를 위한 주요 수단

특히, AI 서비스가 확대되며 대용량의 데이터를 빠르게 처리할 수 있는 클라우드 컴퓨팅의 중요성이 강화되었다. 이에 더하여, AI 반도체를 탑재하여 고성능을 보일 수 있는 데이터센터에 대한 수요도 확대됨에 따라 빅테크 기업 등을 중심으로 데이터센터와 클라우드 시스템 운영을 위한 투자가 지속 이루어지고 있다.

AI 반도체를 탑재하는 등 대용량의 데이터를 처리할 수 있는 역량을 갖춘 데이터센터를 보유하는 것이 빅테크 기업에서 운영하는 클라우드 서비스의 경쟁력 확보를 위한 주요 수단으로 주목 받았다. 이에 따라 데이터센터를 확충하고, 데이터 처리 역량을 늘이기 위한 경쟁이 치열하게 이루어지고 있다.

클라우드 비즈니스를 이끌어가는 빅테크 기업의 데이터센터 신규 투자를 지속 발표하며 많은 관심을 받고 있다. 아마존웹서비스는 2024년 1월 미국에서 100억 달러를 투자하여 두 곳의 신규 데이터센터 건립 계획을 공개하였으며, 일본에서는 약 2조 엔, 인도에서는 37억 달러를 투입하여 신규 데이터센터를 건립하는 등  글로벌 시장에서의 영향력을 확대하고 있다. 구글도 미국에서 6억 달러, 영국에서 10억 달러 규모의 투자를 통하여 신규 데이터센터를 확충하고 있으며, 마이크로소프트도 32억 달러를 투자하여 영국의 신규 데이터센터 건립을 발표하여 데이터센터 관련 빅테크 간의 경쟁은 더욱 치열하게 펼쳐지는 중이다.

특히, 글로벌 빅테크 기업인 구글과 마이크로소프트는 자체 대형 AI 모델을 지속 공개하고 있으며, 다양한 생성형 AI 서비스를 공개하여 많은 주목을 받고 있는 기업이다.

구글은 2022년 당시 최대 규모인 5,400억 개의 파라미터로 구성된 대규모 언어 모델인 PaLM(Pathway to Language Model)을 공개한 후, 2023년에는 PaLM 2와 Gemini 등의 대규모 언어 모델을 지속 출시하며 AI 모델의 기술력을 지속적으로 강화했다. 구글은 자사의 검색 플랫폼과 연계한 생성형 AI 기반의 검색 서비스 ‘Bard’를 2023년 출시하였으며, 2024년에는 ‘Bard’를 개선한 신규 대규모 언어 모델인 Gemini 기반으로 업데이트하여 서비스명을 ‘Gemini’로 변경하였다.

마이크로소프트도 자사의 클라우드 플랫폼인 Azure에 외부 대규모 언어 모델인 오픈AI의 GPT와 Meta의 LLaMA(Large Language Model Meta AI)를 기반으로 한 생성형 AI 서비스가 운영될 수 있도록 구성하였다. 또한 마이크로소프트는 윈도우 플랫폼에 생성형 AI 기반 오피스 프로그램 서비스인 ‘Copilot’ 서비스를 Azure 기반으로 운영 중이다.

구글, 마이크로소프트, 아마존웹서비스 등 글로벌 클라우드 시스템 보유 기업 외에도, 네이버, KT, 카카오 등의 국내 클라우드 업체도 자사의 클라우드 시스템을 주요 자원으로 활용하여 자체 대형 AI 모델을 구성하여 서비스를 출시하고 있다.

## 대형 AI 모델: 모델의 사이즈 확대를 통해 활용 가치를 높임

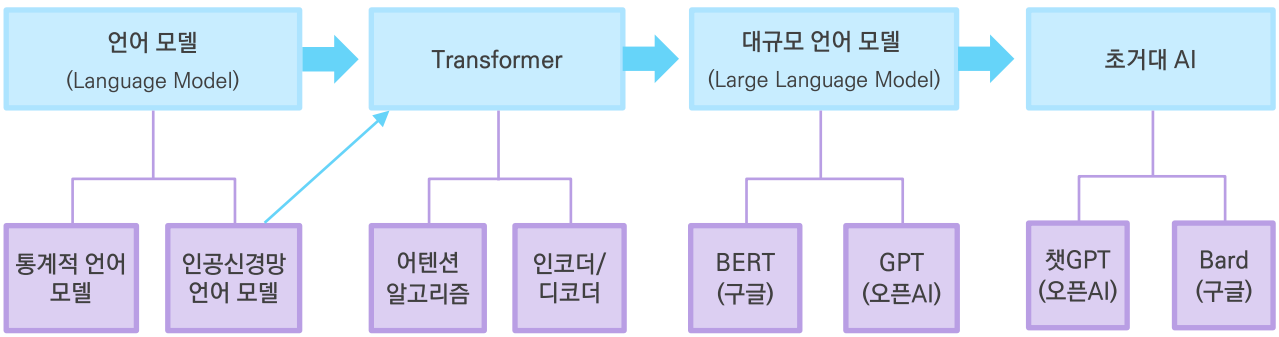
> ‘파라미터’는 크기가 커지면 일반적으로 기능이 더욱 강화되는 것으로 판단 … 수천 억 개의 파라미터를 보유한 AI 언어 모델을 대규모 언어 모델로 분류

2022년 말 챗GPT가 등장한 이후 맞이한 2023년은 생성형 AI가 큰 주목을 받은 시간이었다. 생성형 AI는 이용자의 다양한 질문에 답변을 하는 문자 기반의 커뮤니케이션 뿐 아니라, 주어진 콘셉트에 맞는 음악을 창작하기도 하며, 이미지에 담긴 상황을 판단하여 적절한 해석을 제공하고, 이미지를 그려주는 멀티모달(Multi-Modal) 기반의 다양한 커뮤니케이션 능력을 기반으로 활용 가치가 다양해지며 많은 화제를 모았다.

AI 서비스의 운영을 위하여 필수적으로 구축되어야 하는 AI 파운데이션 모델은 다량의 데이터화 된 자료를 학습하고, GAN, Transformer 등의 구조를 활용하여 딥러닝 기반으로 구성되는 신경망을 의미한다. 생성형 AI와 함께 부상한 언어 모델(Language Model)은 AI 파운데이션 모델 중 자연어 처리(NLP, Natural Language Processing) 기술에 특화된 방식으로 학습된 모델을 칭한다. 언어 모델은 NLP 역량 강화를 위하여 언어별 특성에 따른 문장의 구조를 학습하고, 문장 내 단어를 데이터화하여 상황별 각 문장 속 구성된 단어 등의 요소 간의 관계와 각 구성요소 간의 관계성을 패턴화하는 과정을 거친다. 이러한 NLP 역량을 강화하는데 크게 기여한 것으로 평가받는 딥러닝 방식이 구글이 개발한 Transformer 방식으로 현재 생성형 AI 서비스 운영을 위하여 활용되고 있는 주요 대규모 언어 모델은 Transformer 방식을 활용하는 경우가 다수다.

AI 서비스의 기능이 다양화되고 효용 가치가 높아지는 데는 활용 가능한 데이터 학습량을 확대하고, 더 많은 파라미터를 기반으로 연산 능력을 강화한 대규모 언어 모델(Large Language Model) 등의 대형 AI 모델이 만들어낸 역할에 주목하여야 한다. 대규모 언어 모델은 NLP 기반의 딥러닝 학습 패턴을 가지며, 모델의 크기, 모델을 구성하는 파라미터의 수, 모델의 훈련 과정에서 활용되는 데이터의 크기를 극대화하여 가치를 향상시킨다. AI 모델의 내부에서 AI 모델에 적합한 결과물을 도출할 수 있도록 변수를 통제하는 값인 ‘파라미터’는 그 크기가 커지면 일반적으로 기능이 더욱 강화되는 것으로 여겨진다. 일반적으로 수천 억 개의 파라미터를 보유한 AI 언어 모델을 대규모 언어 모델로 분류하고 있다.

**[ 대규모 언어 모델로의 발전 단계 ]**




Source: 소프트웨어정책연구소 ‘초거대언어 모델의 부상과 주요이슈’, 삼정KPMG 경제연구원 재구성

> 생성형 AI 모델이 본격적으로 주목을 받고, 대규모 AI 모델을 기반으로 한 서비스의 출시가 이어지며 대규모 언어 모델 출시 및 개편하는 노력 지속

생성형 AI 시대의 시작을 알린 챗GPT를 구현한 오픈AI는 2020년 출시한 GPT-3 모델부터 모델의 파라미터 수를 1,750억 개의 달하는 수준으로 구현하였으며, 2021년에는 NVIDIA가 마이크로소프트와 협업하여 5,300억 개의 파라미터를 보유한 'Megatron-Turing NLG'라는 이름의 대규모 언어 모델을 출시하여 모델의 파라미터 크기가 빠른 속도로 증가할 수 있음을 증명하였다. AI 시장의 주요 플레이어인 구글은 2022년 'Megatron-Turing NLG'보다도 파라미터 수를 100억 개 이상 추가하여 5,400억 개의 파라미터를 탑재한 대규모 언어 모델 'PaLM'을 공개하였다.

생성형 AI 모델이 본격적으로 주목을 받고, 대형 AI 모델을 기반으로 한 서비스의 출시가 이어지며 기능이 강화된 새로운 대규모 언어 모델을 공개하는 기업의 노력도 지속됐다. 2023년 다수의 글로벌 빅테크 기업은 생성형 AI 서비스를 제공하기 위한 대규모 언어 모델을 공개하고, AI 언어 모델을 활용한 생성형 AI 서비스의 출시를 통하여 화제를 모았다.

오픈AI는 2022년 11월 'GPT-3.5' 모델을 기반으로 출시한 서비스 '챗GPT'를 출시한 이후, 2023년 3월에 멀티모달 기능을 도입하고 언어 처리 능력을 강화한 'GPT-4' 모델을 출시하여 대규모 언어 모델의 능력을 강화하였다. 또한 GPT-4의 학습 데이터를 새롭게 추가하여 기존 모델 대비 최신 정보에 대한 이해도를 높이고, 한번에 처리 가능한 데이터의 양을 늘린 'GPT-4 터보' 모델을 2023년 11월에 출시하였다.

2022년 대규모 언어 모델 'OPT-175B'을 출시한 바 있는 미국 빅테크 기업 메타(Meta)는 2023년 2월 생성형 AI 서비스 구현을 위한 대규모 언어 모델 'LLaMA'를 공개했다. 같은 해 7월 메타는 마이크로소프트와 협업하여 'LLaMA 2'를 마이크로소프트의 클라우드 서비스 'Azure'를 기반으로 구현하도록 공개하며 대규모 언어 모델 기반의 생성형 AI 서비스 개선을 추진하고 있다.

2022년 대규모 언어 모델 'PaLM'을 출시했던 구글은 기존 PaLM 모델이 텍스트 기반의 태스크 만을 수행할 수 있었던 것과 달리 이미지, 음성 데이터 등을 생성 및 분석할 수 있는 멀티모달 기능이 포함된 신규 대규모 언어 모델 'Gemini'를 2023년 12월에 출시하였다. 구글의 대표적인 생성형 AI 서비스인 검색 솔루션 'Bard'에도 신규 대규모 언어 모델인 'Gemini' 기반으로 업그레이드 하고 서비스명을 'Gemini'로 변경하는 등 더 강화된 서비스를 제공하기 위한 노력을 지속하고 있다.

[ 빅테크 기업의 주요 대규모 언어 모델 ]

| 개발 기업 | 대규모 언어 모델 | 출시일 | 파라미터 수 |
|---|---|---|---|
| NVIDIA, Microsoft | Megatron-Turing NLG | 2021.10 | 5,300억 개 |
| OpenAI | GPT-3.5 | 2022.11 | 1,750억 개 |
| Meta | LLaMA | 2023.02 | 6,500억 개 |
| Google | PaLM | 2022.04 | 5,400억 개 |
| | Gemini | 2023.12 | 1조 6,000억 개 |

Source: 언론보도 종합, 삼정KPMG 경제연구원

# [ Issue Brief ] 국내외 주요 대형 AI 모델 개발 동향

다수의 글로벌 빅테크 기업이 주도하는 대규모 언어 모델(LLM) 등 대형 AI 모델의 출시가 이어지고 있다. 대형 AI 모델은 다량의 파라미터를 탑재하여 복잡한 연산을 처리할 수 있는 능력을 확보하고 있으며, 클라우드와 데이터센터 등을 통해 챗봇, 이미지 분석 서비스 등을 운영하고 있다.

| 기업명 | | 주요 제공 기능 | | |
|---|---|---|---|---|
| | | 대형 AI 모델 | 파라미터 수(개) | 주요 서비스 |
| 해외 | 구글 | Gemini Ultra | 약 1조 6천 억 | • Gemini(검색엔진 기반 챗봇 서비스) |
| | 오픈AI | GPT-4 | 미공개 | • 챗GPT(검색엔진 기반 챗봇 서비스) |
| | 앤트로픽 | Claude 3 | 미공개 | • Claude3(검색엔진 기반 챗봇 서비스) |
| | 미스트랄AI | Mixtral 8x22B | 1,410억 | • 개방형 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용 |
| | 화웨이 | PanGu-$\alpha$ | 2,000억 | • 중국어 자연어 처리에 특화(인공지능 챗봇, 맞춤형 교육 등 다양한 분야에 적용 계획) |
| | 텐센트 | Hunyuan | 1,000억 이상 | • 클라우드 기반 생성형 AI 서비스 제공 예정 |
| | 바이두 | PLATO-3 | 최대 2,600억 | • 어니봇(검색엔진 기반 챗봇 서비스) |
| 국내 | 네이버 | HyperCLOVA X | 미공개 | • Cue(검색엔진 기반 챗봇 서비스) |
| | 삼성 | Samsung Gauss Language | 미공개 | • AI Playground(이메일, 요약, 번역)<br>• code.i(코드 분석, 생성, 수정)<br>• FabriX(기업 내 데이터, 지식자산, 업무 시스템 등에 생성형 AI 서비스 접목)<br>• Brity Copilot(기업 내 이메일, 메신저, 영상회의 등을 지원하는 협업 솔루션) |
| | LG | 엑사원 2.0 | 최대 3,000억 | • Universe(AI 및 머신러닝 분야 학술문헌 분석, 분야 확대 예정)<br>• Discovery(신소재/신약의 분자 구조 설계 및 합성 예측)<br>• Atelier(이미지 분석 및 창작) |
| | KT | 믿음 | 최대 2,000억 | • KT 믿음 스튜디오(파운데이션 모델 개방 전용 포털) |

Note: 1) SK텔레콤은 통신 서비스에 특화된 멀티LLM '텔코LLM' 개발 중(에이닷엑스, GPT-4, Claude 기반),
2) 카카오는 한국어 특화 초거대AI모델 '코GPT 2.0' 사업화 검토 중
Source: 각 사, 언론보도 종합, 삼정KPMG 경제연구원 재구성

> AI 반도체는 데이터를 처리하는 속도(Latency), 처리할 수 있는 데이터의 양(Throughput), 전달하는 데이터의 양(Bandwidth) 측면에서 높은 성능

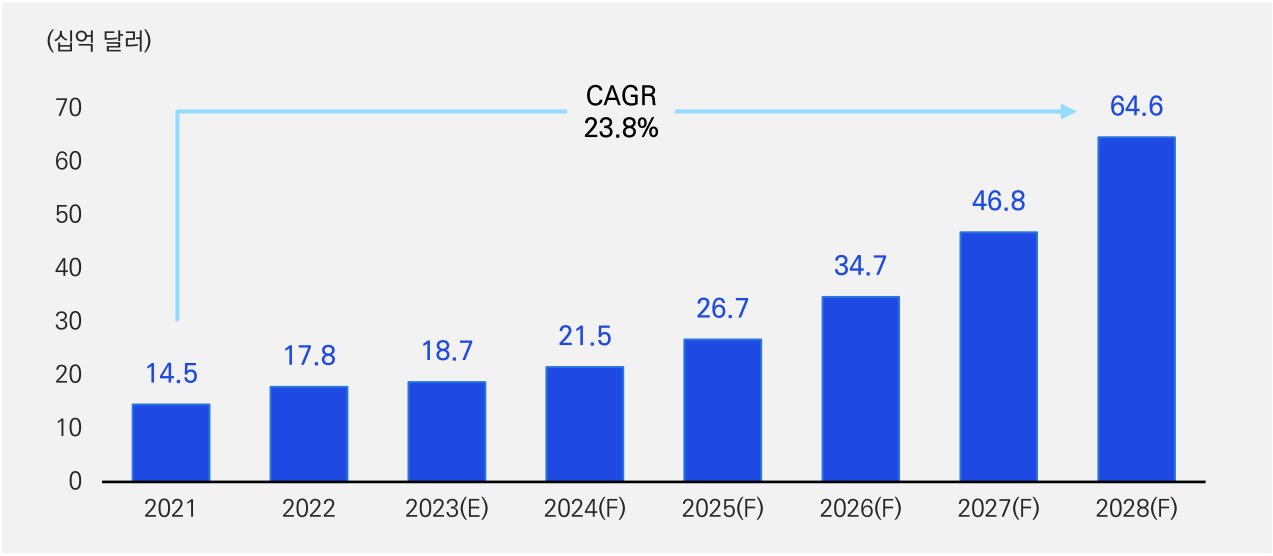
## AI 반도체: 고성능 반도체 중심으로 AI 반도체 시장이 형성

생성형 AI의 활용이 확대되는 데는 AI 모델을 만들어내는 기술력의 발달과 함께, AI 모델이 다양한 환경에서 활용될 수 있도록 하는 관련 인프라의 발달이 주요했다. 대규모 언어 모델은 다량의 데이터를 학습하는 과정을 필요로 하고, 학습된 데이터를 기반으로 다량의 연산을 빠르게 처리할 수 있는 추론 과정을 통하여 효용성을 갖춘 서비스를 운용할 수 있다.

생성형 AI의 활용을 위해서 빠르게 많은 데이터를 처리할 수 있는 AI 반도체, 다량의 데이터를 처리하여 결과물을 송출하기 위한 클라우드 서비스 등의 중요성이 강조되었다. AI 반도체는 데이터를 순차적으로 처리하는 기존 반도체 구조와 달리 데이터를 병렬 형태로 처리할 수 있어 대용량 데이터 연산에 용이하다는 장점을 기반으로 대규모 언어 모델을 활용하는 AI 시장에서 활용성이 크게 주목받고 있다. AI 반도체는 기존 반도체 대비 데이터를 처리하는 속도(Latency), 빠르게 처리할 수 있는 데이터의 양(Throughput), 유효하게 전달할 수 있는 데이터의 양(Bandwidth) 측면에서 높은 성능을 가지고 AI 시대의 핵심 자원으로 평가된다.

시장조사기관 MarketsandMarkets에 따르면, 글로벌 AI 반도체 시장은 2023년 187억 달러 수준에서 2028년까지 646억 달러 수준까지 큰 성장을 거듭할 것으로 전망된다. 미국 반도체 기업 엔비디아의 반도체 제품이 그래픽 분야의 기술력을 활용하여 확보한 병렬형 데이터 처리 역량이 우수하다고 알려지며 엔비디아가 AI 반도체 시장을 선도하고 있다. 국내 반도체 기업 삼성전자와 SK 하이닉스도 고대역폭의 메모리 성능을 구현하는 구조인 PIM(Processing In Memory) 기반의 HBM 반도체 기술을 활용하여 AI 반도체 시장에서 적극적인 움직임을 가져가고 있다.

[ 글로벌 AI 반도체 시장 규모 전망 ]




Source: MarketsandMarkets, 삼정KPMG 경제연구원 재구성
Note: (E)는 추정치, (F)는 전망치

# 왜 온디바이스 AI인가

AI 모델에 대한 관심은 지속 확장 … 대형 AI 모델의 활용이 증가하며 대형 AI 모델이 가지는 한계점도 나타남

## 대형 AI 모델과 클라우드 기반 AI의 한계

대형 AI 모델을 활용하는 다양한 서비스가 지속 출시되고, 새로운 대형 AI 모델의 등장으로 AI의 기능이 강화되는 추세가 나타남과 함께, 대규모 언어 모델을 기반으로 한 AI 모델에 대한 관심은 확장되고 있다. 관심이 커짐과 함께, 대형 AI 모델의 활용이 지속 증가함에 따라 대형 AI 모델이 가지는 한계점도 나타나고 있다.

AI 모델이 점차 대형화 됨에 따라 대규모 언어 모델을 운영하는 데 따른 문제도 거론되고 있다. 생성형 AI로 대표되는 AI 기술의 새로운 혁신은 과거 AI 모델 대비 보다 많은 영역에서 서비스를 대응할 수 있고, 기존 모델 대비 더 우수한 결과값이 도출된다는 장점이 있지만, 이를 위하여 대규모 언어 모델은 더욱 복잡한 구조를 갖게 됐다.

대형 AI 모델은 더 많은 자원을 사용해야 한다는 단점을 지닌다. 대규모 언어 모델은 모델의 숙련도를 강화하기 위한 학습을 진행하는 태스크와 모델을 운영하여 추론되는 결과값을 산출하기 위한 태스크 등 운영 과정에서 다량의 데이터가 소모된다. 이 과정에서 많은 전력과 반도체 등의 컴퓨팅 역량 자원이 소모된다.

[ 대형 AI 모델과 클라우드 기반 AI의 한계 ]

**한계점 1. 비용**

- 대규모 언어 모델 등의 대형 AI 모델 운영을 위해서는 클라우드 시스템과 데이터센터에서 방대한 양의 데이터 처리가 요구되며, 데이터 처리를 위해 막대한 시스템 운영 비용이 발생함
- 대규모 언어 모델 기반 서비스가 확대되며 대형 AI 모델 사용 수요가 늘어나는 경우, 데이터센터와 클라우드 통신 시스템 등의 추가 구축을 위한 인프라 투자 부담이 증가함

**한계점 2. 에너지**

- 대형 AI 모델 운영에 필수적으로 요구되는 인프라 자원인 클라우드와 데이터센터 운영을 위해서는 AI 반도체 등의 인프라 구축이 필요하며, 대량의 데이터 처리를 위한 과정에서 과다한 전력 소모 등의 환경적 비용이 발생함
  - AI 반도체 등의 장비 운용을 위한 전력 소모 및 반도체 등 장비 운용 과정에서 발생하는 열을 내리기 위한 쿨링 비용 등

**한계점 3. 정보 보안**

- 클라우드 기반 생성형 AI 운영을 위해 다양한 데이터가 클라우드 시스템을 통해 데이터센터로 이동
- 이용자의 편의성 개선을 위해 이용자 개인에 대한 다양한 데이터가 데이터센터에 저장되어 이용자 식별 데이터 등 민감한 사생활 문제에 정보보안 이슈 존재
- 기업내 생성형 AI 활용을 저해하는 요인 중 하나로 기업 내부 기밀 자료가 데이터센터 등 외부로 유출되는 것을 방지하기 위한 체계 마련 부담이 지속적으로 거론되고 있음

Source: 언론보도 종합, 삼정KPMG 경제연구원

많은 전력이 소모되는 대형 AI 모델을 구성하기 위해서는 데이터센터와 같은 인프라 구축이 필수적이다. 데이터센터는 대량의 데이터를 처리하기에 적합한 인프라를 갖추고 있지만, 다량의 데이터를 처리하기 위하여 많은 전력을 소모한다. 또한 데이터센터의 AI 반도체 등 장비를 운용하는 과정에서 발생하는 열을 내리기 위해 쿨링 비용도 발생하는 등, 데이터센터 운영 과정에서 비용이 크게 발생한다.

데이터센터는 대형 AI 모델의 학습 과정뿐 아니라 모델의 추론 과정에서도 주요 프로세스를 직접 수행한 후, 산출물을 디바이스로 송출하는 역할을 수행한다. 대규모 언어 모델의 많은 데이터를 필요로 한다는 점과 무거운 사양은 데이터센터에 대한 AI 모델의 과한 의존도를 보일 수 밖에 없는 요건을 구성한다.

> 데이터 보안성에 대한 우려가 지속되며 국내외 기업에서는 업무 중 생성형 AI 사용을 제한하는 사례도 나타남

대형 AI 모델을 운영하기 위하여 필요한 데이터센터, AI 반도체 등의 대량의 데이터 처리 역량 확보를 위한 노력은 AI 모델을 운용하기 위하여 필요한 부담이 과해지는 결과로 나타냈다. 대표적인 생성형 AI 서비스인 챗GPT는 이용자 1명이 검색 1회를 하기 위해 발생하는 서버 운용 비용 등은 약 26원 수준에 달하는 것으로 알려져 있으며, 1억 명의 이용자가 10번씩 챗GPT 서비스를 이용한다고 가정하면 약 260억 원의 비용이 발생한다는 점에서 대형 AI 모델을 기반으로 서비스를 운영하는데 확장성에 제약이 있다는 점을 알 수 있다.

대형 AI 모델을 활용하는 클라우드 기반 AI에서 거론되는 다른 문제점은 데이터 보안 우려이다. 클라우드를 기반으로 AI 시스템을 운영하기 위해서는 다량의 데이터의 이동이 필요하다는 점에서 많은 기업에서 데이터 안정성 확보를 위한 고심이 이어지고 있다. 대형 AI 모델은 다량의 데이터를 학습하여 구성된다는 특징이 있다. 이를 위하여, 다양한 소스의 데이터를 취합하여 모델이 학습함과 함께, 지속적으로 이용자가 입력하는 데이터를 학습하여 모델의 성능을 강화하기 위한 과정을 진행하기도 한다. 이때, 클라우드 시스템을 통하여 이용자의 다양한 정보가 외부 데이터센터로 유입된다는 우려가 존재한다.

생성형 AI를 활용하는 클라우드 AI의 역량이 다양화되고 기업에서의 활용성이 강화되어 가며, 기업의 민감한 정보를 생성형 AI에 입력할 때 발생할 수 있는 데이터 유출의 리스크와 개인이 생성형 AI 서비스에 민감한 정보를 입력하여 발생할 수 있는 개인정보 문제 등도 지속적으로 제기되고 있다.

데이터 보안성에 대한 우려가 지속되며 국내외 기업에서는 업무 중 생성형 AI 사용을 제한하는 사례도 나타나고 있다. 다량의 데이터가 외부로 지속 전송되며 발생할 수 있는 데이터 처리 역량 확보에 대한 부담과 데이터 보안성에 대한 문제를 해소하기 위해 일부 기업에서는 생성형 AI 서비스를 기업 내부에서만 사용할 수 있도록 자체 서비스를 구축하고 있다.

> 온디바이스 AI는 디바이스 외부로 이용자가 입력하는 정보를 내보내지 않는다는 장점 보유 … 민감한 데이터를 처리하기에 적합한 방식으로 주목

## 온디바이스 AI의 부상

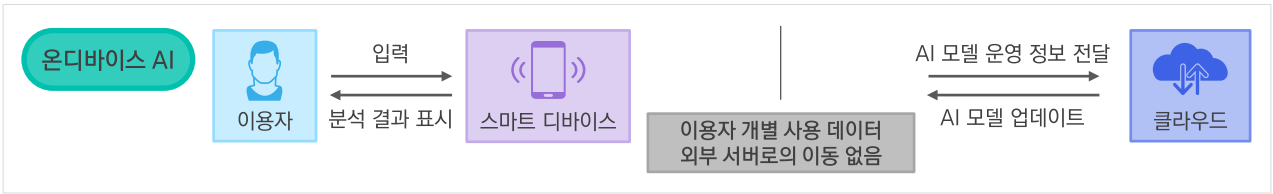
대형 AI 모델과 클라우드를 기반으로 하는 AI 시스템의 운영이 반도체, 클라우드 시스템 등의 높은 인프라 구축 부담과 데이터 보안 관리 측면에서의 문제점이 지속 거론되는 단점이 부상함과 함께, 온디바이스 AI에 대한 관심도는 지속 확대되고 있다. 온디바이스 AI는 클라우드 기반의 AI 시스템 운영 방식과 달리, 이용자가 사용하는 디바이스 자체에서 AI 시스템을 구동한다는 특징 덕분에 클라우드 기반의 AI에 대한 우려 요소가 해결 될 수 있다는 기대가 높다.

온디바이스 AI는 최종적으로 이용자가 사용하는 디바이스 내에 AI 서비스를 운영할 수 있는 모델을 탑재 및 자체적으로 구동할 수 있는 시스템을 구축하여 운영된다. 클라우드 기반의 AI는 서비스에 접속되어 있는 이용자의 디바이스에서 받은 정보를 네트워크를 통해 클라우드 시스템으로 전달하고, 클라우드 시스템을 통하여 데이터를 전달 받은 외부의 서버에서 AI 모델을 구동 및 산출된 결과값을 다시 클라우드 및 네트워크 시스템을 통해 이용자에게 전달되는 플로우를 가진다.

이와 달리, 온디바이스 AI를 기반으로 한 AI 서비스 운영 과정에서는 이용자가 AI를 통하여 결과물을 얻고자 하는 요청사항을 디바이스에 입력하면 데이터의 외부 이동 없이 이용자가 사용하는 디바이스 내에 탑재된 AI 모델과 반도체 등의 데이터 처리 시스템이 구동되어 이용자에게 AI 서비스 구현을 통한 결과값을 제공한다.

위와 같이 온디바이스 AI 기반 디바이스는 AI 서비스를 통한 결과값을 얻기 위하여 디바이스 외부로 이용자가 입력하는 정보를 내보내지 않는다는 점에서 개인정보 등의 민감한 데이터를 처리하기에 적합도가 높은 AI 서비스 구현 방식으로 주목받고 있다. 데이터 처리를 위하여 외부 네트워크 및 클라우드로의 전송을 위한 인프라 확보 부담이 적다는 점 또한 클라우드 AI 대비 높은 평가를 받고 있다.

[ 온디바이스 AI의 특징 ]




**1 AI 모델 분석 속도 향상**

- 온디바이스 AI는 외부 통신 없이 이용자가 사용하는 스마트 디바이스 내부에서 분석 진행
- 분석 속도 향상 기대

**2 외부 시스템 비용 감소**

- 외부 클라우드 및 데이터센터를 모델 업데이트 용도로만 사용
- 클라우드 및 데이터센터 이용 비용 감소 효과 기대

**3 데이터 보안 우려 감소**

- 이용자의 개별 사용 데이터가 디바이스 외부로 나가지 않음
- 외부 전송 및 저장 과정에서 발생할 수 있는 데이터 보안 우려 적음

Source: 언론보도 종합, 삼정KPMG 경제연구원

## 온디바이스 AI의 구조

온디바이스 AI 운영을 위해 클라우드 기반 AI와 달라지는 분야는 AI 모델이 하는 역할이 대형과 소형으로 나뉘어지는 점과 클라우드의 컴퓨팅 역할이 클라우드와 디바이스로 나뉘어진다는 점이다.

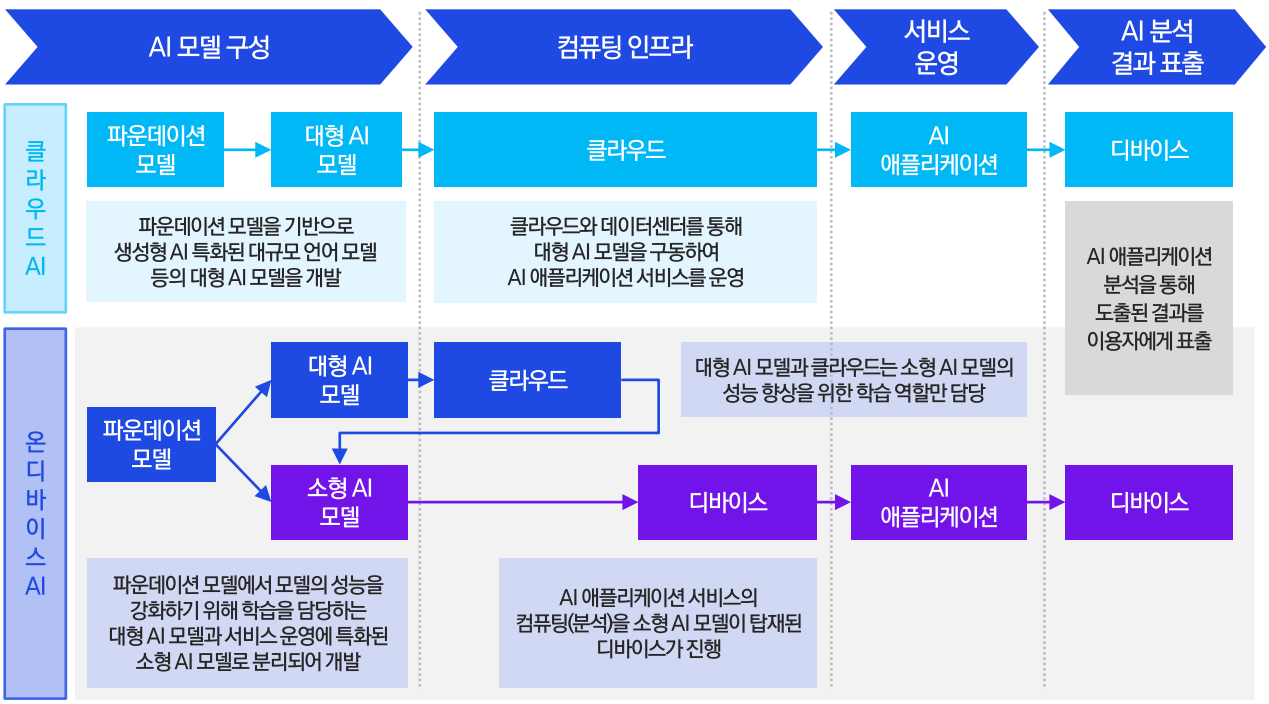
> 온디바이스 AI는 AI 모델과 컴퓨팅 인프라의 역할을 세분화하여 각각의 구성 요소가 가진 장점을 융합하여 활용

클라우드 기반 AI에서 대형 AI 모델을 중심으로 운영되는 것과 달리, 온디바이스 AI에서는 소형 AI 모델이 서비스 특화된 형태로 구성되어 대형 AI 모델과 함께 운영된다. 온디바이스 AI에서 소형 AI 모델은 이용자가 사용하는 서비스의 운영을 담당하여 결과물을 도출하기 위한 분석을 진행한다. 대형 AI 모델은 전반적인 AI 모델의 성능을 강화하는 역할을 하며, 소형 AI 모델은 대형 AI 모델을 통한 업데이트를 지속적으로 받으며 성능을 향상시키는 역할을 맡는다.

클라우드 기반 AI에서 클라우드와 데이터센터는 서비스 운영과 AI 모델 개발을 위한 컴퓨팅 인프라 역할을 한다. 온디바이스 AI에서는 대형 AI 모델의 개발과 소형 AI 모델의 성능 향상을 위한 역할을 담당하며 서비스 운영을 위한 역할은 온디바이스 AI가 탑재된 디바이스가 대신하게 된다.

온디바이스 AI는 AI 모델과 컴퓨팅 인프라의 역할을 세분화하여 각각의 구성 요소가 가진 장점을 융합하여 활용한다는 점에서 생성형 AI의 향상된 성능을 누릴 수 있는 플랫폼으로 주목받고 있다.

[ 클라우드 기반 AI와 온디바이스 AI의 구조 ]




Source: 언론보도 종합, 삼정KPMG 경제연구원

> 텍스트, 음성 외에도 온디바이스 AI의 높은 보안성을 활용할 수 있는 생체 인식 데이터를 활용하는 온디바이스 AI 서비스도 높은 성장률을 보일 것으로 전망

## 온디바이스 AI의 활용 방식

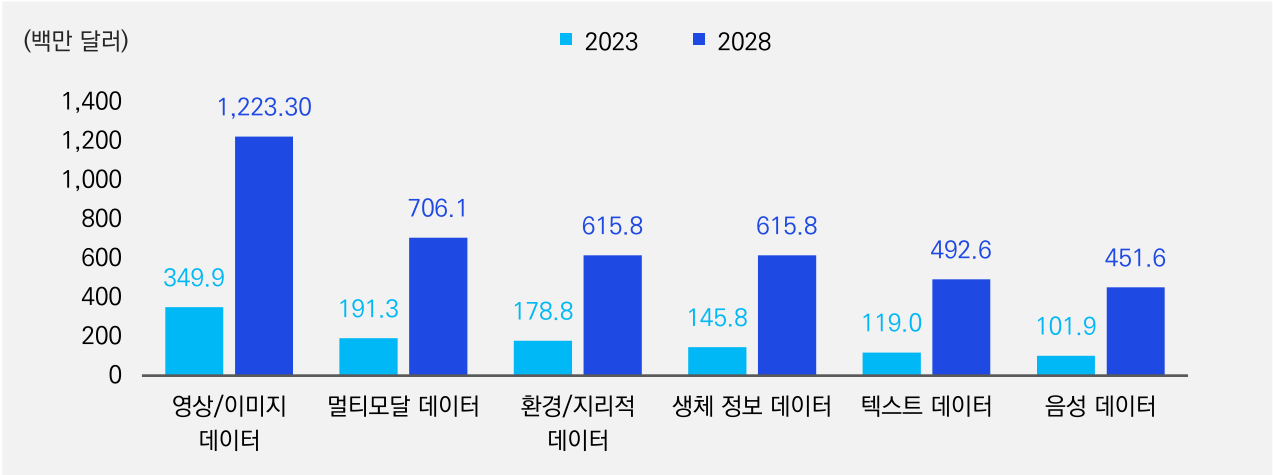
온디바이스 AI는 최종적으로 결과물이 활용되는 디바이스를 중심으로 활용 가치가 높아질 것으로 보인다. 이용자가 최종적으로 직접 디바이스를 활용하고, 이용자가 원하는 태스크를 직접적으로 수행할 수 있는 디바이스에 탑재된 온디바이스 AI는 별도 데이터의 외부 이동 없이 AI 모델의 운영을 통한 결과를 이용자에게 전달할 수 있다.

온디바이스 AI는 이용자가 원하는 AI 모델을 통해 분석하고자 하는 인풋(Input) 데이터를 확보할 수 있어야 하므로 스마트폰, 노트북과 같이 데이터를 입력할 수 있는 수단이 있는 디바이스와 센서 등의 데이터 입력 기능을 가진 디바이스가 활용되는 경우가 많을 것으로 보인다. 온디바이스 AI 기능을 활용하기 위하여 사용되는 데이터는 텍스트 외에도 음성, 영상, 장소 등의 다양한 데이터가 활용되어 AI 모델의 분석을 지원한다.

시장조사기관 MarketsandMarkets에 따르면, 온디바이스 AI를 포함한 엣지 AI의 소프트웨어가 활용하는 데이터는 텍스트, 영상, 음성 등 다양한 영역에서 모두 큰 폭의 성장이 이뤄질 것으로 전망됐다. 활용 가능한 주요 데이터 중 영상과 이미지 분석 관련 AI 서비스 시장이 가장 큰 규모를 형성할 것으로 보인다. 온디바이스 AI 기능을 탑재한 삼성전자의 스마트폰 갤럭시 S 시리즈 제품에서 AI 기반의 이미지 편집, 보정 기능 등이 활용되고 있다. 온디바이스 AI 기능을 탑재한 TV에서는 과거 저화질로 촬영된 영상을 고화질 영상으로 보정하여 화면을 송출하는 등의 영상 및 이미지 기반 온디바이스 AI 솔루션이 출시되었다.

텍스트, 음성 등 생성형 AI 시장에서 대규모 언어 모델을 활용하여 서비스 출시가 지속된 영역 뿐 아니라 온디바이스 AI의 높은 보안성을 활용할 수 있는 생체 인식 데이터를 활용하는 온디바이스 AI 서비스도 높은 성장률을 보일 것으로 전망되고 있다.

[ 온디바이스 AI 포함한 엣지 AI 활용 데이터 종류별 글로벌 시장 규모 전망 ]




Source: MarketsandMarkets, 삼정KPMG 경제연구원 재구성

## 온디바이스 AI의 운영 요소

### 소형 AI 모델: 테스크 수행에 특화된 목적 기반형 AI 모델

대형 AI 모델이 창출하는 결과물의 수준이 많은 이용자들이 AI 모델의 능력에 감탄할 수 있는 결과를 만들어 냈다. 그러나 AI 모델 운영을 위해 필요한 과다한 전력 소비, AI 반도체, 데이터센터 등의 인프라 확보 부담이 부각되며 대형 AI 모델을 활용한 AI 서비스가 확대되는데 어려움이 부각되고 있다. 이를 통해, AI 서비스가 다양한 환경에서 활용될 수 있도록 하기 위하여 대규모 언어 모델 외에 다른 구조의 AI 모델을 확보하는 것에 대한 관심도가 증가하였다.

> "소형 AI 모델은 AI 모델의 주요 역할인 학습과 추론 중 이용자의 요청을 받아 추론 과정을 통하여 결과물을 표출해주는 역할에 집중"

대형 AI 모델을 운영하기 어렵게 하는 요소를 해소하고 다양한 서비스로 AI 모델의 활용 가치를 확대하기 위해 부각되고 있는 것이 경량화된 소형 AI 모델이다. 소형 언어 모델(small Language Model, sLM)은 대규모 언어 모델이 지속적으로 파라미터 수를 확대하며 기능을 강화했던 것과 달리 모델의 크기를 상징하는 파라미터 수를 일반적으로 340억 개 이하 수준으로 구성하는 모델을 칭한다.

소형 언어 모델은 작은 사이즈의 AI 모델로 대규모 언어 모델보다는 제한된 태스크를 수행하지만 사전 학습된 데이터를 기반으로 생성형 AI의 주요 기능을 구현할 수 있도록 특화되었다. 소형 언어 모델은 대규모 언어 모델 대비 결과물을 창출하기 위한 데이터 처리 양이 적어지며 대형 인프라를 필요로 하지 않는다. 이러한 소형 언어 모델의 특징은 작은 디바이스 내에서 AI 모델을 자체적으로 구동해야 하는 온디바이스 AI 구현이 가능하게 하는 필수 요소이다.

소형 AI 모델은 AI 모델 구현을 위한 모든 기능을 자체적으로 소화하기보다는 AI 모델의 주요 역할인 학습과 추론 중 이용자의 요청을 받아 추론 과정을 통하여 결과물을 표출해주는 역할에 집중한다. 클라우드를 기반으로 한 AI 모델은 학습과 추론을 모두 클라우드에서 수행하고, 이용자가 사용하는 디바이스는 이용자의 요청을 클라우드로 전달 및 분석된 결과값을 공유 받아 이용자에게 표출하는 역할을 한다. 이러한 구조적 특성에 기인하여, 소형 AI 모델을 탑재하는 디바이스는 AI 모델 구현을 위한 시스템적 요구가 비교적 줄어들어 온디바이스 AI 서비스를 구현하기 위한 환경이 갖춰질 수 있다.

[ 대형 및 소형 AI 모델 주요 사항 비교 ]

| 구분 | 대형 AI 모델 | 소형 AI 모델 |
|---|---|---|
| 주요 형태 | 대규모 언어 모델(Large Language Model) | 소형 언어 모델 [sLM(small Language Model), sLLM(small Large Language Model)] |
| Parameter 수 | 1천 억 개 이상 | 10억 ~ 수백 억 개 수준 |
| 사용 목적 | 다양한 태스크를 수행하는 범용 AI | 특정 태스크에 특화된 목적 기반형 AI |
| 주요 특징 | • 대용량의 데이터를 학습하고 다수의 Parameter를 기반으로 운영됨<br>• 모델의 규모가 크기 때문에 태스크 수행을 위해 필요한 인프라의 수준이 높음<br>– AI 서비스 운영 과정에서 발생하는 비용이 크며, 결과물 도출 속도가 비교적 오래 걸림 | • 대형 AI 모델이 학습한 데이터를 기반으로 특정 태스크에 특화된 형태로 조정함<br>• 작은 규모의 모델로 운영되어 개별 태스크 수행을 위해 필요한 인프라의 수준이 낮음<br>– AI 서비스 운영 비용이 비교적 저렴하며, 결과물 도출 시간이 빠른 편임 |

Source: 언론보도 종합, 삼정KPMG 경제연구원

> 소형 AI 모델을 활용한 온디바이스 AI는 추론 과정의 필수적인 태스크만을 수행하여 효율성을 강화 … 대형 AI 모델 방식 대비 인프라에 대한 부담 적음

클라우드를 통하여 추론과 학습 과정이 모두 수행되는 대형 AI 모델은 우수한 네트워크 기술을 기반으로 이용자가 크게 기다리지 않아도 회신을 얻을 수 있을 정도로 빠른 결과를 보여주는 수준까지 기술의 발전을 이뤄냈다. 그러나 많은 데이터를 주고 받고, 여러 파라미터를 거치며 결과물을 생성하는 대형 AI 모델의 특성상 일정 수준의 데이터 송수신을 위한 지연과 지연 시간을 최소화하기 위한 반도체, 클라우드 시스템 등의 우수한 인프라 구축 부담이 존재한다.

소형 AI 모델을 활용한 온디바이스 AI는 내부에서 추론 과정의 필수적인 태스크 만을 수행하여 효율성을 강화한 덕분에 대형 AI 모델을 사용하는 클라우드 기반의 방식 대비 인프라에 대한 부담이 적다. 온디바이스 AI도 우수한 성능을 유지하기 위해서는 학습 부분을 담당하는 클라우드 기반의 대형 AI 모델의 지원이 필수적임에 따라, 클라우드 시스템을 확보하기 위한 노력이 이루어져야 하지만, 이용자가 원하는 태스크를 수행하기 위하여 매번 클라우드를 거치는 프로세스가 필요하지 않기 때문에 인프라 부담이 적은 편이다.

이러한 온디바이스 AI의 특징을 활용하는 소형 AI 모델의 개발에는 기존 대형 AI 모델을 개발 중인 빅테크 기업과 비교적 적은 개발 부담이 들어가는 소형 AI 모델의 특성을 활용한 스타트업의 진입 등이 주목 받고 있다. 대형 AI 모델 시장에서 빠르게 영역을 넓혀가고 있는 빅테크 기업 구글, 마이크로소프트 등은 기존 대형 AI 모델의 구조를 활용하고, 대형 AI 모델과의 연결성이 높은 형태로 구현한 소형 AI 모델을 출시하고 있다.

구글은 대규모 언어 모델로도 출시된 'Gemini'의 소형 버전인 'Gemini Nano'를 공개하였으며, 마이크로소프트는 자체 연구소에서 개발한 소형 AI 모델 'Phi-3'를 출시하여 노트북, 모바일 등 온디바이스 AI에 활용 계획을 공개하였다.

국내 스타트업 업스테이지는 소형 AI 모델인 '솔라 미니'를 자체 개발하였다. 오픈소스 형태로 구성된 업스테이지의 '솔라 미니'는 AI 모델의 성능을 평가하여 순위를 정리한 글로벌 포털 허깅스페이스에서 2023년 12월 1위에 오르며 화제가 된 바 있다.

# [ Issue Brief ] 국내외 주요 소형 AI 모델 개발 동향

Gemini(구글), LLaMA(메타) 등 글로벌 빅테크 기업은 대형 AI 모델을 기반으로 파라미터 등의 모델의 크기를 줄인 소형 AI 모델(sLM 등)을 출시하였다. 국내외 다수의 기업에서도 대형 AI 모델을 활용하여 경량화한 소형 AI 모델의 출시가 이뤄지고 있다.

| 기업명 | | 주요 제공 기능 | | |
|---|---|---|---|---|
| | | 소형 AI 모델 | 파라미터 수(개) | 주요 서비스 |
| 해외 | 마이크로소프트 | Phi-3 Medium / Small / Mini | 최대 14억 | • 개방형 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용 |
| | 애플 | ReALM-3B | 최대 30억 | • 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용 |
| | 구글 | Gemini Nano | 최대 33억 | • 개방형 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용 |
| | | Gemma | 최대 70억 | |
| | 메타 | LLaMA(Large Language Model Meta AI) 3 – 8B | 최대 80억 | • 메타AI (SNS 플랫폼 기반 검색 챗봇 서비스, 이미지 생성 기능, 가상개인비서 업무 – 페이스북, 인스타그램, 왓츠앱, 메신저 기반) |
| | 미스트랄 | 미스트랄 7B | 70억 | • 미스트랄 7B 인스트럭트 (개방형 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용) |
| | 스태빌리티 | Stable LM Zephyr | 최대 70억 | • 개방형 파운데이션 모델로서 챗봇, 가상개인비서 등 다양한 플랫폼에 활용 |
| | 알리바바 | Tongyi Qianwen -7B | 최대 70억 | • Qwen-Audio(다양한 종류의 오디오 데이터 분석 및 해석 후 텍스트로 결과 출력)<br>• Qwen-Audio-Chat(오디오 데이터 분석 및 해석 후 대화를 통해 결과 출력, 대화의 톤 감지 가능) |
| 국내 | LG 유플러스 | 익시젠 | 미공개 | • 챗 에이전트(통신 산업 특화 고객 대응 서비스) |
| | 업스테이지 | 솔라 미니 | 107억 | • 아마존웹서비스(AWS) 기반 플랫폼에 적용되어 다양한 솔루션 상에서 맥락 이해 및 텍스트 생성에 도움을 주는 생성형 AI 서비스 |

Source: 각 사, 언론보도 종합, 삼정KPMG 경제연구원 재구성

## AI 반도체: 온디바이스에서 활용 가능성이 높은 저전력 반도체 시장에 주목

> 온디바이스 AI에 탑재되는 반도체 등의 프로세서 … 디바이스의 낮은 전력에도 효율적으로 활용될 수 있는 저전력 반도체 중심 시장 구성

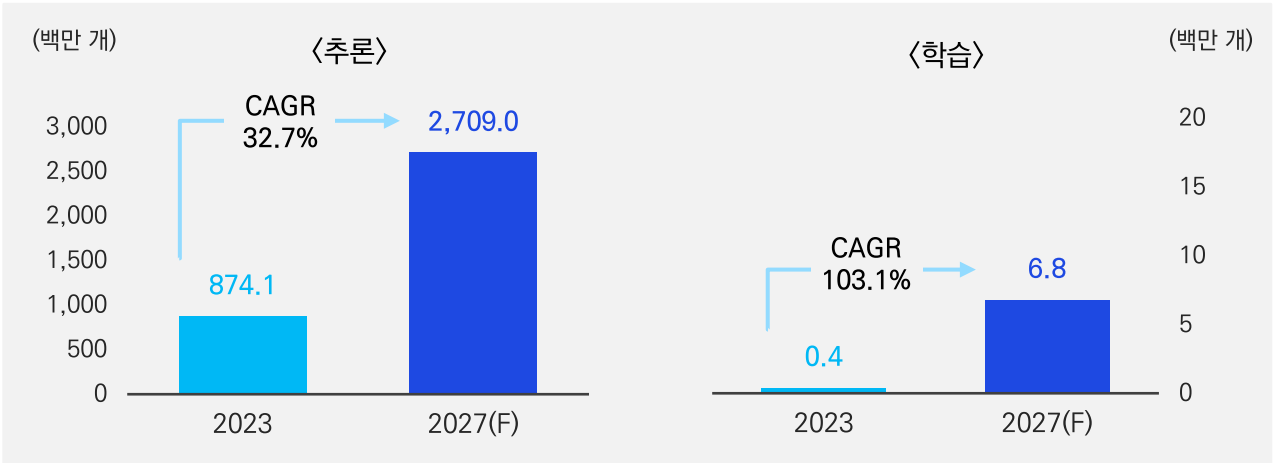
AI 모델은 일반적으로 학습과 추론 과정을 지속적으로 수행한다. AI 모델의 학습 과정은 온라인에 존재하는 다양한 데이터를 모델에 저장 및 분석하도록 하여 모델이 다양한 정보를 획득할 수 있도록 하는 과정이다. 이어지는 추론 과정에서는 AI 모델이 사전 학습한 데이터와 모델 설계자가 제공한 Parameter 값 등의 사전 학습된 정보를 기반으로 이용자가 만드는 요청에 가장 적합하다고 판단하는 결과물을 생성하여 제공하는 태스크가 이루어진다.

온디바이스 AI는 클라우드에서 추론과 학습의 모든 AI 분석 과정이 진행된 후 결과물을 송출 받아 표출하는 기존 AI 서비스 운영 방식과 달리 추론 과정을 별도 분리하여 디바이스 내에서 운영한다. AI 모델의 주요 과정 중 하나인 학습 과정은 클라우드에서 처리된다. 이는, 온디바이스 AI 구현을 위하여 주로 사용되는 AI 모델인 소형 언어 모델이 가진 한정된 자원을 활용하여 서비스를 구현하기 위한 목적을 가진다.

글로벌 시장조사기관 MarketsandMarkets에 따르면, 온디바이스 AI를 포함한 엣지 AI용 하드웨어 시장에서 추론 역할을 담당할 하드웨어 시장은 2022년 8.7억 개 수준에서 27억 개 수준에 달하며 엣지 AI 하드웨어 시장에서 절대적인 비중을 차지할 것으로 전망됐다. 반면, 학습 역할을 담당하는 엣지 AI용 하드웨어 시장은 2022년 40만 개에서 2027년 680만 개 수준으로 연평균 100% 이상의 큰 성장이 기록될 것으로 전망되나, 전체 시장에서 차지하는 비중은 미미할 것으로 분석됐다.

엣지 AI용 하드웨어 시장은 엣지 서버와 온디바이스 AI를 구현하는 디바이스 등 AI 서비스의 프로세스를 담당하는 디바이스와 함께, 온디바이스 AI 등의 구현을 위하여 디바이스 내에 탑재되는 프로세서를 포함한다. 온디바이스 AI에 탑재되는 반도체 등의 주요 프로세서는 스마트 디바이스의 낮은 전력에도 효율적으로 활용될 수 있는 저전력 반도체를 중심으로 시장이 구성될 것으로 전망된다.

[ 온디바이스 AI 포함 엣지 AI 하드웨어의 활용 분야별 글로벌 출하량 성장 전망 ]




Source: MarketsandMarkets, 삼정KPMG 경제연구원 재구성
Note: (F)는 전망치

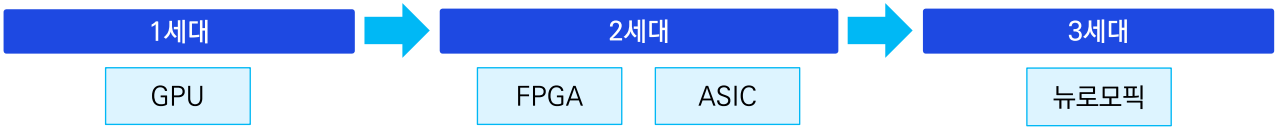
> 높은 성능을 발휘하며 낮은 전력 소모를 보일 수 있는 저전력 반도체의 구조적 특성 … 2세대 AI 반도체 기술인 FPGA와 ASIC 주목

다량의 데이터를 빠른 속도로 처리하여 AI 모델을 통한 결과를 도출하여야 하는 클라우드, 데이터센터 등의 AI 시대의 인프라 구성에서 고성능 반도체의 역량이 크게 강조되었다. 이러한 흐름 속에서 다량의 데이터를 병렬형으로 처리할 수 있는 구조적 특성을 가진 그래픽 처리 기반의 GPU가 각광 받았으며, GPU 분야에서 우수한 기술력을 보유한 미국의 반도체 기업 엔비디아가 AI 반도체 시대의 주요 기업으로 큰 주목을 받았다.

온디바이스 AI를 운영하기 위한 반도체 인프라에는 고성능뿐 아니라 저전력이 주요한 키워드로 부상하고 있다. 저전력 반도체는 디바이스 자체에 내장된 반도체 등의 프로세서 역량을 활용하여 AI 기능을 운영하는 온디바이스 AI 분야에서 영향력이 더욱 커질 것으로 주목되는 분야이다. 저전력 반도체는 반도체가 정보를 처리하는 과정에서 소비되는 전력이 적도록 설계하는 방식으로 AI 시대에는 낮은 소비 전력으로도 높은 성능을 낼 수 있는 고성능·저전력 반도체 기술이 핵심 경쟁력으로 지목되고 있다.

높은 성능을 발휘하며 낮은 전력 소모를 보일 수 있는 반도체의 구조적 특성으로 2세대 AI 반도체 기술인 FPGA와 ASIC가 주목 받고 있다. 논리회로형 반도체 기술인 FPGA와 특정 목적에 맞게 구성이 가능한 주문형 반도체 구조인 ASIC는 최적화된 제어 및 산술 논리 구성 요소를 갖춘 NPU(Neural Network Processing Unit)를 활용하여 AI 반도체 분야에서 요구되는 성능 수준을 보인다. 또한 FPGA와 ASIC는 온디바이스 AI를 구현하기 위하여 특화된 형태로 구성되어 낮은 전력으로도 구동이 가능하게 설계될 수 있다.

[ AI 반도체의 발달 과정 ]




| 타입 | 주요 형태 | 내용 |
|---|---|---|
| 1세대 | GPU<br>(Graphics Processing Unit) | • 1세대 AI 반도체는 CPU와 GPU를 유기적으로 연계하여 대량의 데이터를 활용한 연산에 효과적인 형태로 구성하는 방식으로 구성됨<br>• 동시에 데이터 처리량이 많은 그래픽 데이터 처리를 위해 고안된 그래픽 프로세서를 활용하여 AI 연산 소프트웨어 구성<br>• 데이터 병렬 처리 효과에 따른 동시다발적인 학습 및 추론 기능 강화 |
| 2세대 | FPGA<br>(Field Programmable Gate Array)<br><br>ASIC<br>(Application Specific Integrated Circuit) | • 2세대 AI 반도체는 기계 학습을 위해 최적화된 제어 및 산술 논리 구성 요소를 갖춘 NPU(Neural Network Processing Unit)를 활용하여 AI 반도체의 성능 개선을 도모함<br>[FPGA]<br>• 논리회로형 **반도체로** 회로 변경이 가능한 형태로 제작되어 이용 목적에 따라 반도체 형태를 비교적 자유롭게 재구성할 수 있다는 장점이 있음<br>• 사용중인 반도체에 필요에 따라 새로운 형태를 재입력하여 새롭게 최적화하는 것이 가능하여 양산형 반도체 개발 방식으로 활용됨<br>[ASIC]<br>• NPU셀에 메모리가 병합된 형태로 특정한 용도에 맞게 사전에 설계된 주문형 반도체<br>• 정해진 용도에 맞춤형 설계를 통하여 높은 효율과 안정적인 성능이 장점임 |
| 3세대 | 뉴로모픽 | • 기존 반도체의 연산 구조인 폰 노이만 방식을 따르지 않고, 인간의 뇌신경구조인 뉴런과 시냅스의 형태를 모방한 형태의 뉴로모픽 형태의 AI 반도체<br>* 폰 노이만 방식: 입력되는 데이터를 순차적으로 처리하는 방식 |

Source: 언론보도 종합, 삼정KPMG 경제연구원

> 온디바이스 AI 기술과 소형 AI 모델, AI 반도체 등의 인프라 자원 발달 … 온디바이스 AI가 활용될 수 있는 영역이 다양화될 전망

## 온디바이스 AI가 탑재된 제품 및 Use case

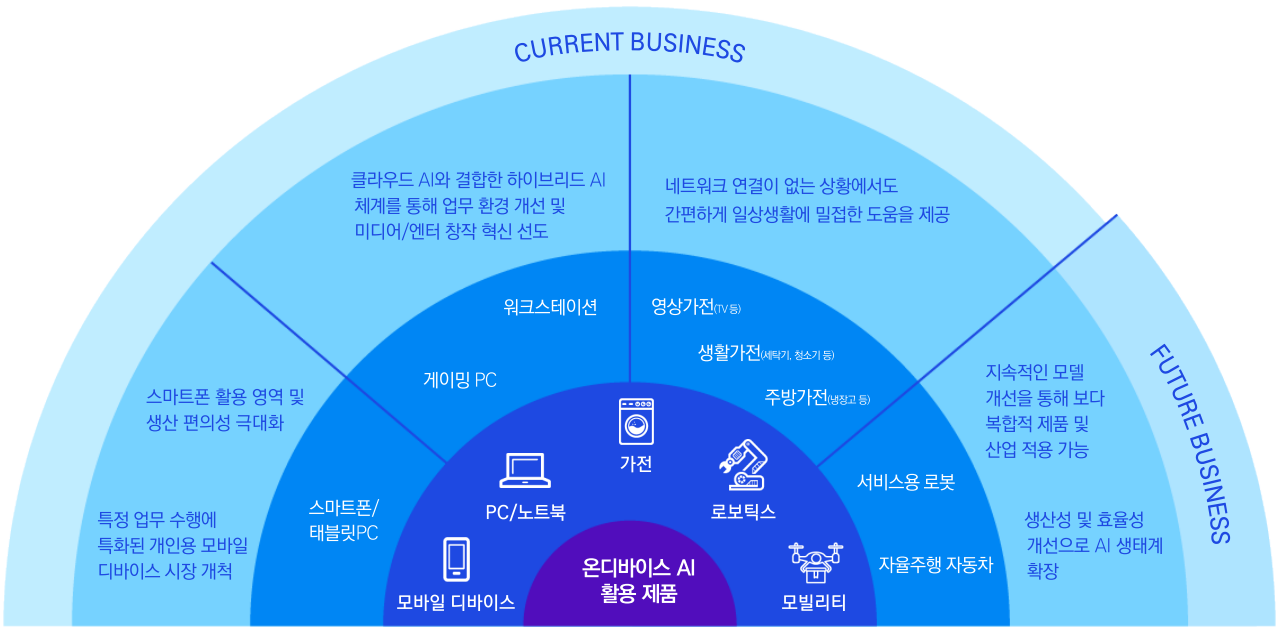
삼성전자, 애플, HP 등 IT 분야의 글로벌 선도 기업을 중심으로 온디바이스 AI 가 탑재된 기기가 출시되며 시장의 관심도를 높이고 있다. 온디바이스 AI 기술과 온디바이스 AI의 구현을 가능하게 하는 소형 AI 모델, AI 반도체 등의 인프라 자원이 발달되며 온디바이스 AI가 활용될 수 있는 영역이 다양화될 것으로 전망된다.

2024년 상반기 출시된 온디바이스 AI 제품은 주로 모바일 디바이스와 PC/노트북, 가전 분야의 제품으로 구성되어 있다. 이미지 편집, 문서 요약 등의 기능을 제공하는 온디바이스 AI 기능을 탑재하여 화제를 모은 삼성전자의 '갤럭시 S24 시리즈'가 출시되며 모바일 디바이스 분야에서 본격적으로 온디바이스 AI의 활용이 주목받고 있다. 애플의 온디바이스 AI 기능이 포함된 생성형 AI 서비스 'Apple Intelligence'는 모바일 디바이스뿐 아니라 애플 OS(운영체계)를 사용하는 PC/노트북, 태블릿 PC 제품군에서도 활용이 가능하게 제공될 예정이다.

가전 분야에서도 온디바이스 AI 기능을 구현하여 이미지 퀄리티를 자동으로 개선하는 영상가전(TV 등) 제품과 AI를 기반으로 기기가 자체적으로 판단 및 운영되어 가사 노동의 편의성을 개선한 생활 및 주방가전(세탁기, 냉장고 등) 제품이 출시되어 관심을 받고 있다.

온디바이스 AI는 높은 보안성과 빠른 분석 속도를 활용하여 활용 영역을 다양화할 것으로 기대된다. 향후에는 서비스용 로봇, 자율주행 자동차 등 생활 밀착형 분야로의 확대가 이루어져 AI 생태계 확대에 기여할 것으로 보인다.

[ 온디바이스 AI 활용 비즈니스 생태계 ]




Source: 언론보도 종합, 삼정KPMG 경제연구원

## 모바일 디바이스 – 삼성전자 '갤럭시' · 애플 'Apple Intelligence'

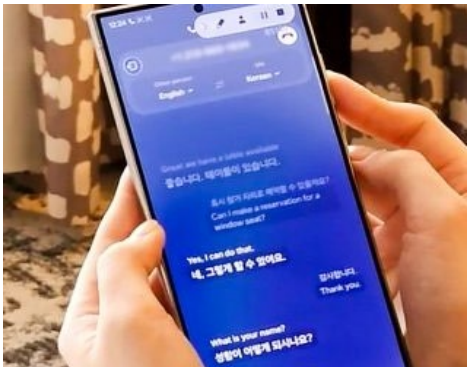
"'갤럭시 S24' 시리즈 외에도 '갤럭시 S23', '갤럭시 Z5' 시리즈에도 온디바이스 AI 기능이 구현 … 삼성전자는 자체 개발한 생성형 AI 모델 '가우스'와 함께 구글의 소형 AI 모델 'Gemini Nano' 활용"

모바일 디바이스는 온디바이스 AI가 가지는 효용성이 빠른 속도로 나타나고 있는 분야로 주목 받고 있다. 일상 생활 속에서 이용자와 많은 시간을 함께 하고, 이용자의 다양한 정보가 문자, 통화, 인터넷 등의 수단을 통하여 얻어지게 되는 모바일 디바이스의 특성상 높은 보안성을 가진 온디바이스 AI 기술이 AI 기능 보급의 주요한 요소로 자리하고 있다.

국내 스마트폰 제조기업 삼성전자는 2024년 초 온디바이스 AI 기능이 탑재된 신규 플래그십 스마트폰 '갤럭시 S24' 시리즈를 공개하며 본격적인 모바일 디바이스의 온디바이스 AI 시장 개척을 알렸다. '갤럭시 S24'에 탑재된 주요 온디바이스 AI 기능으로는 '실시간 통/번역', '텍스트 변환 어시스트' 등이 주목 받았다. 온디바이스 AI를 기반으로 구현되는 '실시간 통/번역' 기능은 이용자가 '갤럭시 S' 디바이스를 활용하여 다른 언어를 사용하는 상대방과 통화, 문자 등의 커뮤니케이션을 할 때 사용할 수 있다. 다른 언어를 구사하는 상대방과의 커뮤니케이션 과정에서 음성 및 문자를 한국어 등 사용자가 사용하는 언어로 실시간 변환해주고, 사용자의 언어를 다시 상대방의 언어로 변환하는 기능이다. 빠른 속도로 결과물을 변환하여 끊김 없는 커뮤니케이션을 지원해야 하고, 민감한 정보가 오고갈 수 있는 커뮤니케이션의 특성상 온디바이스 AI가 해당 기능을 구현하는 주요 요소로 자리하고 있다.

삼성전자는 '갤럭시 S24' 시리즈 외에도 '갤럭시 S23', '갤럭시 Z5' 시리즈에도 온디바이스 AI 기능이 구현될 수 있도록 업그레이드를 지원하였다. 온디바이스 AI 구현을 위해 삼성전자는 자체 개발한 생성형 AI 모델 '가우스(Gauss)'와 함께 구글의 대형 AI 모델 'Gemini Pro', 소형 AI 모델 'Gemini Nano' 등을 활용하고 있다. 온디바이스 AI 모델을 구현하는 삼성전자의 주요 모바일 디바이스 모델에는 모바일 디바이스용 반도체 분야에서 기술력이 높은 것으로 평가되는 퀄컴의 '스냅드래곤 8' 제품과 삼성전자의 모바일 특화 반도체 '엑시노스' 제품군 등이 탑재되어 있다.

[ 온디바이스 AI 기능이 탑재된 갤럭시 S24 ]



[ 갤럭시 S24의 온디바이스 AI를 통한 '실시간 통번역' 주요 처리 기술 ]

| 자동 음성인식 | 인공신경망 기반 기계번역 | 텍스트 음성변환 |
|---|---|---|
| – 신조어 | – 관용어 | – 음성 데이터베이스 |
| – 숫자 | – 동음이의어 | – 말뭉치 |
| – 대명사/고유명사 | – 방언/사투리 | – 사전 데이터베이스 |
| – 동음어 | – 집중 요점 | – 스크립트 |
| – 노이즈 데이터 | – 문맥/테마 | |

Source: 삼성전자 보도자료, 삼정KPMG 경제연구원 재구성
Photograph Source: 삼성전자 보도자료

> 애플은 온디바이스 AI와 클라우드 AI를 결합한 하이브리드(Hybrid) AI 형태로 운영 … 프라이빗 클라우드가 수신한 정보는 대형 AI 모델 분석 종료된 후 즉시 삭제

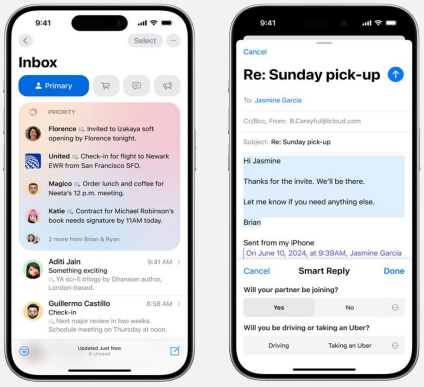
애플도 2024년 6월 개최한 자체 연례행사WWDC(세계개발자회의) 2024에서 온디바이스 AI 기능을 포함한 AI 서비스 'Apple Intelligence'를 공개하고 신규 OS 배포와 함께 스마트폰, 태블릿PC, PC 제품군에 서비스를 도입할 계획을 공개하였다. 'Apple Intelligence'는 애플의 기존 제품 중 모바일용 반도체 'A17 Pro'가 탑재된 아이폰15 프로 이상의 제품과 PC용 반도체 'M1' 이상이 탑재된 아이패드와 Mac 제품에 보급될 것으로 발표되었다.

'Apple Intelligence'는 오픈AI에서 2024년 새롭게 발표한 대형 AI 모델인 GPT-4o를 활용하여 생성형 AI 기능을 구현한다. 'Apple Intelligence'에는 이용자의 요청에 맞는 이미지와 이모티콘을 생성하는 '이미지 플레이그라운드', '젠모지(Genmoji)' 등의 콘텐츠 생성 기능과 기기 내 저장된 이미지와 영상의 주요 특징을 문장으로 검색하여 찾을 수 있는 이미지 분석 기능 등의 멀티모달(Multi-Modal) 기능이 탑재되었다. 애플은 또한 스마트폰에 수신되는 문자, 이메일 등의 문서를 자동 정리하여 문서 내용 요약, 회신용 샘플 메일 작성, 빠른 대응이 요구되는 우선 처리 사항 알림 등의 분석 기능도 제공한다.

애플의 생성형 AI 서비스는 온디바이스 AI와 클라우드 AI를 결합한 하이브리드(Hybrid) AI 형태로 운영되는 특징을 가진다. 예를 들어, 애플의 음성 비서 서비스 시리(Siri)는 대형 AI 모델의 처리가 필요한 복잡한 요청이 들어오면 클라우드 시스템을 통해 외부 서버로 데이터를 송출하여 분석 결과를 전달 받아 이용자에게 응답한다. 애플의 하이브리드 AI는 클라우드 AI 운영을 위해 프라이빗 클라우드(Private Cloud) 형태를 도입하여 클라우드 AI의 단점으로 지적 받고 있는 보안 우려를 줄이도록 설계되었다. 애플의 프라이빗 클라우드는 AI 서비스 운영을 위한 기능만 구현되는 형태로 운영된다. 이용자의 요청을 처리하기 위해 수신한 정보는 대형 AI 모델의 분석이 종료된 후 즉시 삭제된다. 이를 통해, 이용자 개인의 정보가 클라우드에 남지 않도록 구현하여 개인정보 보호 등 보안성이 높다는 특징이 있다.

[ 온디바이스 AI 기능이 탑재된 애플의 AI 서비스 ]

**애플–'Apple Intelligence'**




- **온디바이스 AI와 클라우드 AI를 융합한 하이브리드 AI 시스템 도입**
  - 클라우드 AI는 'Apple Intelligence' 를 위해 별도 구현한 프라이빗 클라우드를 활용하여 운영
  - 이용자의 개별 요청 데이터는 클라우드 AI를 통한 분석이 완료되어 결과를 디바이스로 송출한 직후 시스템에서 바로 삭제되어 개인정보 보안 우려 해소
- **오픈AI의 대형 AI 모델인 GPT-4o 기반으로 구현**
- **iOS 18로 업그레이드 된 기기 중 모바일용 반도체 'A17 Pro', PC용 반도체 'M1' 이상 탑재한 디바이스에서 구동 가능**

Source: 애플 보도자료, 삼정KPMG 경제연구원
Photograph Source: 애플 보도자료

> PC 분야에서도 온디바이스 AI를 탑재하여 이용자의 편의성과 생산성을 강화하기 위한 제품의 출시가 이어짐

## PC/노트북 – 마이크로소프트 온디바이스 AI 솔루션 'Copilot+PC' 공개 삼성전자와 LG전자는 온디바이스 AI 탑재 노트북 개발

모바일 디바이스 외에도 이용자의 휴대가 용이하고 다양한 정보의 커뮤니케이션이 이뤄지는 PC/노트북 분야에서도 온디바이스 AI 탑재 노력이 지속되고 있다.

PC/노트북에서 활용되는 운영체계(OS)를 개발하는 대표적인 기업 마이크로소프트는 2024년 5월 21일 MS 연례 최대 개발자 컨퍼런스 'MS 빌드 2024'에서 'Copilot+PC'를 발표하며 AI PC 시장으로의 참여를 본격화했다. MS의 자사 노트북 'Surface Pro'의 11번째 에디션에 온디바이스 AI 기반으로 'Copilot+PC' 기능을 구현하여 공개하였다. 'Surface Pro'는 온디바이스 AI 기능 구현을 위해 고성능의 반도체로 평가되는 퀄컴의 '스냅드래곤 X 엘리트'칩을 탑재하였다. 마이크로소프트는 자사 모델인 'Surface Pro' 외에도 삼성, 에이수스, 델 등의 글로벌 노트북 제조 기업의 신규 모델에서도 온디바이스 AI 기능을 도입한 'Copilot+PC' 기능을 제공할 계획임을 공개하였다.

삼성전자는 2023년 12월 출시한 '갤럭시 북4' 시리즈부터 PC/노트북 분야에서 온디바이스 AI 기능을 구현할 계획을 공개하였다. 그리고 2024년 5월 삼성전자는 마이크로소프트의 AI 기능인 'Copilot+PC' 등이 탑재된 하이브리드 AI(온디바이스 AI·클라우드 AI 동시 지원) 노트북 '갤럭시북4 엣지' 제품을 공개하여 온디바이스 AI 기능을 구현하였다.

LG전자는 2024년 3월에 자사 노트북 제품인 'LG 그램'에 온디바이스 AI 기술을 구현하기 위한 일환으로 중소벤처기업부 등과 함께 온디바이스 AI 분야 기술력을 보유한 스타트업 발굴을 위한 '온디바이스 AI 챌린지' 출범식을 진행하였다. 해당 챌린지를 통하여 발굴된 보안, 엔터테인먼트, 생산성 강화 기능 등을 포함하는 온디바이스 AI 기능을 추후 노트북 제품에 탑재하여 공개할 예정이다.

[ 온디바이스 AI가 탑재된 노트북 ]

**마이크로소프트 – 'Copilot+PC'**


- 윈도우 OS 제품에서 활용 가능한 온디바이스 AI 솔루션 'Copilot+PC' 공개
- 'Surface Pro', '갤럭시북 4 엣지' 등에 탑재

**삼성전자 – '갤럭시 북4 엣지'**

- 마이크로소프트의 'Copilot+PC' 솔루션 기반 온디바이스 AI 구현
- 퀄컴 '스냅드래곤 X 엘리트' 반도체 탑재

**LG전자 – 온디바이스 AI 챌린지**


- 온디바이스 AI를 자사 노트북 'LG 그램'에 적용 방안 모색을 위한 챌린지 진행
- 보안, 생산성 강화 기능 등 도입 기대

Source: 언론보도 종합, 삼정KPMG 경제연구원
Photograph Source: 각 사 보도자료

삼성전자와 LG전자는 가전 제품에 온디바이스 AI 기능을 활용할 수 있는 AI 프로세서를 선보이는 등 생태계 확대 노력 지속

## 가전 – 삼성전자 · LG전자의 온디바이스 AI용 반도체와 가전 제품 공개

영상·생활·주방 가전 등 이용자의 가정 내 다양한 업무 환경을 지원하는 가전 분야에서도 온디바이스 AI를 기반으로 한 지능화된 솔루션 도입이 이어지고 있다. 국내 대형 가전 제품 제조 기업인 삼성전자와 LG전자는 다양한 가전 분야 제품에 온디바이스 AI 기능을 구현하기 위해 가전 분야에서 활용될 수 있는 AI 프로세서를 선보이는 등 온디바이스 AI 생태계 확대를 위한 노력을 지속 기울이고 있다.

삼성전자는 온디바이스 AI 기능을 구현한 TV·냉장고 제품을 공개하였다. 삼성전자에서 선보인 AI 프로세서 'NQ8 AI 3세대'는 지난 세대 제품 대비 8배 이상 많은 512개의 뉴럴 네트워크를 통해 AI 성능을 강화하여 온디바이스 AI 기능 구현을 위한 주요 요소로 자리하고 있다.

CES 2024에서 삼성전자가 전시한 TV 제품인 'Neo QLED 8K'는 온디바이스 AI 기능을 통해 저해상도 영상을 고해상도로 업스케일링(Upscaling)하는 기능과 TV 외부 환경의 소리를 분석하여 이용자가 편히 들을 수 있는 소리 수준으로 맞춤 조정하는 기능 등을 공개하였다. 삼성전자에서 2024년 3월 출시한 비스포크 냉장고는 온디바이스 AI와 비전 AI 기술을 기반으로 내부 카메라를 활용하여 신선식품 관리 용의성을 개선한 기능을 선보였다.

LG전자도 온디바이스 AI 기능을 자사의 주요 가전 제품에 접목하기 위해 온디바이스 AI 기능을 구현하도록 최적화된 가전 전용 반도체 'DQ-C'를 공개하였다. 'DQ-C' 반도체는 LG전자에서 2024년 3월 공개한 올인원 세탁건조기 'LG 트롬 오브제컬렉션 워시콤보'(이하 워시콤보) 제품에 탑재되어 온디바이스 AI 기능을 구현하였다. '워시콤보'는 투입된 옷감 등 세탁물의 환경을 디바이스가 자체 분석하여 세탁, 탈수 등의 적정한 세탁 방식을 자동 결정한다.

[ 온디바이스 AI 기능이 탑재된 주요 가전 제품 ]

**삼성전자 – 'Neo QLED 8K'**


- 온디바이스 AI 기반 화질 및 음질 개선 기능 제공
  - 저화질 영상의 고화질 변환
  - 이용자 특화 소리 맞춤 조정 기능

**삼성전자 – 비스포크 냉장고**

- 온디바이스 AI와 비전 AI 기술 기반 식재료 관리 기능 제공
  - 내부 탑재 카메라로 식자재 종류를 인식하여 신선식품 보관 리스트 제공

**LG전자 – 'LG 트롬 오브제컬렉션 워시콤보'**

- 온디바이스 AI용 반도체 'DQ-C' 탑재
- 세탁 작업 효율화 솔루션 제공
  - 옷감의 상태 및 세탁 환경을 세탁기가 자체 판단하여 적정한 세탁 방식 결정

Source: 언론보도 종합, 삼정KPMG 경제연구원
Photograph Source: 삼정KPMG 경제연구원, 삼성전자 뉴스룸, LG전자 뉴스룸

# 온디바이스 AI 시대 Scale-up을 위한 전략

> 온디바이스 AI 시장 규모(Scale) 확대 예상 … 온디바이스 AI 기술을 구현하고 서비스의 다양화를 이끌어 낼 수 있는 'SCALE' 전략 주목

## 온디바이스 AI 시대의 주목해야 할 키워드 'SCALE'

온디바이스 AI 시장 규모(Scale)의 확대가 예상됨에 따라, 온디바이스 AI 기술을 구현하고 서비스의 다양화를 이끌어 낼 수 있는 'SCALE[S(Semiconductor)-C(Cloud)-A(Ambient Computing)-L(Language Model)-E(Explainable AI)]' 전략에 주목해야 한다.

먼저 온디바이스 AI를 구성하는 인프라 측면에서 온디바이스 AI 시대의 핵심 요소인 반도체, 클라우드에 주목해야 한다. 저전력·고성능 기반의 AI 반도체 시장이 확대되며 온디바이스 AI 기능이 탑재되는 디바이스가 확대될 수 있을 것으로 전망된다. 클라우드는 온디바이스 AI 시장의 확대와 함께 기존 AI 모델의 학습 및 추론 전과정을 담당하던 것과 달리, 학습에 더 특화된 형태로 발달될 것으로 보인다. AI 반도체와 클라우드 시스템의 역할 변화 속 새로운 기회를 모색해야 한다.

온디바이스 AI 시대가 일반 이용자의 가정에 AI를 구동하는 기기의 다양화를 유도하는 효과가 나타나면 일상 속에서 AI가 이용자의 삶 속에 더욱 다양하게 자리매김하고 가속화되는 앰비언트 컴퓨팅 시장 속 기회를 모색해야 한다.

AI 모델은 대형 AI 모델과 소형 AI 모델의 역할이 클라우드를 기반으로 운영되며 다양한 기능을 수행하는 대형 AI 모델과 특정 태스크에 집중되도록 분할하여 서비스에 특화된 sLM 등의 소형 AI 모델 시장이 함께 성장하는 시장이 펼쳐질 것으로 보인다. 온디바이스 AI가 이용자에게 즉각적으로 AI 분석 결과를 제공하는 형태로 이용자와 AI 모델 간의 거리가 더욱 가까워지며 이용자의 혼란을 줄이고 AI 모델의 운영 안정성을 강화할 수 있는 '설명 가능한 AI(Explainable AI)' 기술 통제 방안을 마련해야 한다.

[ 온디바이스 AI 시대 Scale-up을 위한 전략 ]


| S | C | A | L | E |
|---|---|---|---|---|
| 온디바이스 AI 시대에는 고성능 반도체 중심이던 AI 반도체 시장에서 저전력 반도체 중심의 반도체 성장세의 주목해야 함 | 클라우드는 AI 서비스 운영에 소모되는 역할 비중을 줄이고, 대형 AI 모델의 성능 향상을 지원하는 역할에 집중해야 함 | 온디바이스 AI의 높은 지능과 우수한 데이터 보안 능력이 촉발하는 앰비언트 컴퓨팅 시장 활성화 속 기회를 발굴해야 함 | 고지능을 위한 학습 중심의 대형 AI 모델과 효율적 서비스 운영을 위한 소형 AI 모델로 양분화된 시장 속 적합한 활용 모델을 모색해야 함 | 이용자 밀착형인 온디바이스 AI 모델은 안정적 운영을 위한 기술 통제 방안을 필수적으로 마련해야 함 |

Source: 삼정KPMG 경제연구원

## S(Semiconductor): 저전력 AI 반도체 성장 속 창출되는 기회 발굴

온디바이스 AI를 포함한 생성형 AI 기반 서비스 시장의 확대는 다량의 데이터를 처리할 수 있는 고성능 반도체를 중심으로 많은 관심을 받고 있다. 다량의 데이터를 빠른 속도로 처리하여 AI 모델의 분석 결과를 제공하여야 하는 생성형 AI 서비스의 특성상 대용량 데이터를 빠르게 처리할 수 있는 고성능 반도체의 수요는 지속적으로 확대되고 있다.

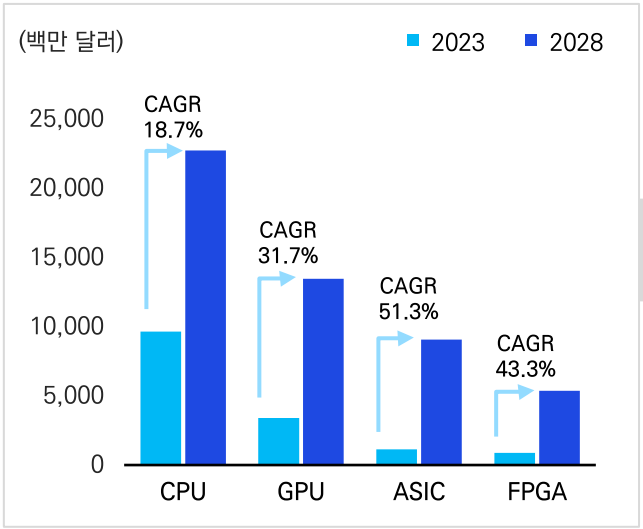
> 고성능 및 저전력 성능을 구현할 수 있는 AI 반도체의 2세대 기술을 탑재한 형태인 FPGA와 ASIC에 대한 주목도가 지속 커질 것

온디바이스 AI 시대가 확대되면 AI 반도체 시장에서도 새로운 트렌드가 자리할 것으로 전망된다. 온디바이스 AI는 개별 디바이스에 AI 모델의 분석을 처리할 수 있는 반도체가 탑재되어야 함에 따라 높은 성능을 보이는 것뿐 아니라 스마트폰 등의 휴대용 디바이스에서도 AI 모델 분석을 처리할 수 있도록 하는 저전력 시장이 각광 받을 것으로 전망된다. 이를 위하여 고성능 및 저전력 성능을 구현할 수 있는 AI 반도체의 2세대 기술을 탑재한 형태인 FPGA와 ASIC에 대한 주목도가 지속적으로 커질 것으로 전망된다. FPGA와 ASIC는 논리회로형 구조와 개별 논리 구성에 최적화된 구조를 기반으로 온디바이스 AI 구현을 위한 저전력 설계에 유리한 형태로 평가받는다.

AI 반도체 분야에서 현재 가장 많은 관심을 받고 있는 반도체 기업 엔비디아가 강점을 보이고 있는 고성능 반도체 기술인 GPU 제품도 FPGA, ASIC보다는 성장률이 낮지만 연평균 30%대 이상의 높은 성장률을 기록할 것으로 보인다. GPU 등 고성능을 발휘하는데 집중된 AI 반도체 제품도 클라우드 시스템과 데이터센터 운영을 위한 사용이 지속 확대되어 시장 규모가 증가할 것으로 예측된다.

향후 AI 반도체 시장에서는 고성능 반도체 중심이던 기존 시장에서 저전력 반도체 중심의 온디바이스 AI가 창출하는 새로운 AI 반도체 시장의 성장세를 주목하여 기회를 모색하여야 한다.

**[ 저전력 반도체 성장에 주목 ]**




**저전력 반도체 성장에 주목해야 함**

- 온디바이스 AI는 디바이스 내 전력을 활용하여 AI 모델을 운영하기 때문에 AI 서비스 운영 과정에서 소비되는 데이터를 낮게 유지해야 함
- 저전력 반도체 기술인 FPGA, ASIC는 2028년까지 연평균 40% 이상의 성장률이 전망되는 등 높은 시장 성장성에 주목해야 함

**고성능 반도체는 클라우드와 데이터센터 중심으로 시장 확대 전망**

- GPU 등의 고성능을 강조하는 AI 반도체 시장도 큰 성장을 지속할 것으로 전망됨
- 클라우드, 데이터센터 등 전력의 제약이 덜한 AI 서비스 분야에서 고성능 반도체 시장도 지속 확대될 것임

Source: MarketsandMarkets, 삼정KPMG 경제연구원
Note: CPU(Central Processing Unit), GPU(Graphics Processing Unit), ASIC(Application Specific Integrated Circuit), FPGA(Field Programmable Gate Array)

> 온디바이스 AI 시대에는 대형 AI 모델의 성능 강화를 위한 프로세스를 빠르게 진행하는 것이 경쟁력 확보를 위한 주요 요소

## C(Cloud): 클라우드의 역할 변화로 인한 인프라 시장 변화 대응

온디바이스 AI 보급의 확대는 클라우드와 데이터센터 시장에도 변화를 줄 것으로 전망된다. 생성형 AI의 보급화와 본격화되는 AI 서비스 출시의 영향으로 데이터센터의 수요가 지속적으로 확대되고 있다. 대용량 데이터 처리를 위하여 구축되는 데이터센터의 수요 증가로 인해 AI 반도체 등의 고성능 반도체 등의 반도체 부족 이슈, 대규모 데이터 처리 시설 운영을 위해 필요한 전력 수요의 과다와 같은 이슈가 지속 부상하고 있다.

클라우드 시스템을 기반으로 운영하는 AI 서비스는 클라우드 시스템에서 AI 모델의 학습과 추론 과정이 모두 이뤄지도록 하기 위하여 클라우드, 데이터센터의 수요가 높게 요구된다는 우려가 있다. 온디바이스 AI 시장이 확대되면 클라우드는 학습과 추론 과정을 모두 담당하였던 기존 AI 시스템 운영 방식에서 학습 과정에 집중하는 방식으로 역할이 변동될 것으로 보인다.

온디바이스 AI 시대의 클라우드와 데이터센터의 역할이 온디바이스 AI 모델이 탑재된 디바이스와 나눠질 것으로 보이나, 클라우드 및 데이터센터 시장의 수요의 성장은 지속될 것으로 전망된다.

현재 클라우드와 데이터센터의 경쟁력을 확보하기 위해서는 AI 모델의 학습과 추론 전과정에서 더 많은 데이터를 처리할 수 있는 역량과 보다 빠르게 데이터를 교류할 수 있는데 있는 반면, 온디바이스 AI 시대에는 AI 모델의 학습 역량에 집중하여 많은 학습용 데이터를 확보하고, 대형 AI 모델의 성능 강화를 위한 학습 프로세스를 빠르게 진행할 수 있도록 구성하는 것이 경쟁력 확보를 위한 주요 요소가 될 것이다.

이러한 시장 변화에 맞춰, 클라우드 기업은 AI 서비스 운영에 소모되는 역할 비중을 줄이고, 초거대 AI 모델의 성능 향상을 지원하는 역할에 집중해야 하며, 클라우드의 역할 변화에 따른 반도체 등 인프라 시장의 변화에 주목하여야 한다.

[ 클라우드와 데이터센터의 역할 전망 ]

**대형 AI 모델 성능 향상 지원 역할**

- 현재 생성형 AI 기반 서비스 운영을 위하여 추론과 학습 과정을 모두 부담하던 역할에서 학습 과정에 집중하는 형태로 변화할 것
- 대형 AI 모델이 범용적인 성능을 더 우수하게 갖출 수 있도록 학습 데이터 확보 및 더 많은 파라미터 기반으로 운용 가능하도록 성능 확보 노력이 이어질 것


**반도체 등 인프라 확대 부담은 줄어들 것**

- 현재 클라우드와 데이터센터는 대형 AI 모델을 기반으로 운영되는 서비스 수요 감당을 위해 반도체 부족, 전력 과다 소모 등의 인프라 운용 부담이 큼
- 온디바이스 AI 시대로 대형 AI 모델과 클라우드 시스템이 서비스 운영 관련 역할이 줄어들며 인프라 확충을 위한 부담이 감소할 것으로 전망됨

Source: 삼정KPMG 경제연구원

## A(Ambient Computing): 앰비언트 컴퓨팅 가속화 속 새로운 시장 모색

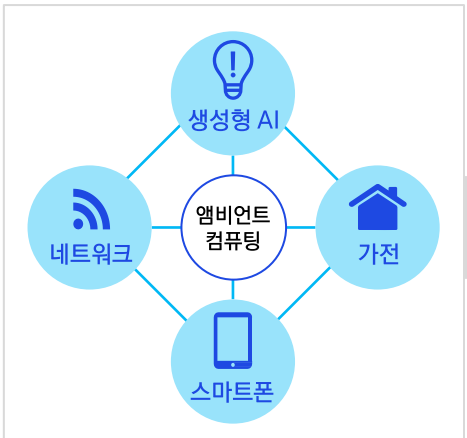
온디바이스 AI는 이용자가 직접 사용하는 디바이스에서 AI 모델이 작동하여 AI 시스템에 기반한 결과물을 제공하여 일상 속에서 사용하는 IT 디바이스가 현재보다 더 능동적이고 효율적으로 진행될 수 있게 하는 효과가 기대되고 있다. 온디바이스 AI의 보급이 확대되고 일상 속 다양한 영역에서 온디바이스 AI 기능을 이용자가 체감할 수 있게 될수록 부각될 개념으로 '앰비언트 컴퓨팅'에 주목하여야 한다.

> 온디바이스 AI는 일상 생활 속 이용자가 사용하는 데이터를 외부로 유출하지 않고 분석 가능 … 앰비언트 컴퓨팅 환경을 구축하는 데 높은 활용 가치를 가질 것

앰비언트 컴퓨팅은 주변에 있다는 의미의 단어 Ambient와 컴퓨팅이 결합된 단어로 주변에 항상 존재하는 컴퓨팅 시스템을 표현한다. 앰비언트 컴퓨팅은 이용자 주변에 있는 IT 디바이스가 이용자의 패턴을 자체적으로 학습 및 분석하여 이용자가 디바이스를 조작하지 않더라도 디바이스가 알아서 이용자가 필요한 서비스를 제공하는 환경이 갖춰진 상태를 의미한다. 앰비언트 컴퓨팅은 현재 스마트홈 분야에서 인공지능 스피커와 스마트 가전 등 가정 내 와이파이와 연계된 디바이스를 활용하여 관련 서비스가 운영되고 있다. 다만, 앰비언트 컴퓨팅이 이용자의 주변에서 이용자가 직접 조작하지 않더라도 맞춤형으로 서비스를 제공하는 수준으로 고도화되기 위해서는 온디바이스 AI에 역할이 클 것으로 전망된다.

앰비언트 컴퓨팅 환경이 구축되기 위해서는 일상 생활 속 사용하는 디바이스가 높은 지능을 기반으로 한 분석력을 가지는 것과 함께, 이용자 개인별 일상 생활 속 다양한 데이터를 확보하여 분석할 수 있는 환경이 갖춰지는 것이 필요하다. 개인별 데이터를 확보하고 분석하기 위한 보안성에서 강점이 있는 온디바이스 AI는 일상 생활 속 이용자가 사용하는 데이터를 외부로 유출하지 않고 분석이 가능하다는 점에서 앰비언트 컴퓨팅 환경을 구축하는데 높은 활용 가치를 가질 것으로 보인다. 온디바이스 AI의 높은 지능과 우수한 데이터 보안 능력이 촉발하는 앰비언트 컴퓨팅 시장 활성화 속 기회를 발굴할 수 있는 전략 마련이 필요하다.

[ 온디바이스 AI를 통한 앰비언트 컴퓨팅의 확대 ]




### 앰비언트 컴퓨팅 시장의 확대를 위한 온디바이스 AI의 역할

**1 외부 유출 우려가 덜한 데이터 활용으로 보안 강화**

- 이용자의 데이터가 외부로 유출될 우려가 적은 온디바이스 AI의 특성상 앰비언트 컴퓨팅 기능을 발휘하는 디바이스에서 이용객 맞춤형으로 특화되도록 다양한 데이터의 확보 및 분석이 가능해지는 효과가 나타날 것으로 보임

**2 강화된 AI 모델의 지능을 활용한 능동적 대응 시스템 구축**

- AI 모델의 고도화를 통한 지능 강화로 앰비언트 컴퓨팅에 속한 디바이스가 능동적으로 이용자의 니즈를 파악하고 대응할 수 있도록 강화될 수 있음

Source: 삼정KPMG 경제연구원

## L(Language Model): AI 모델 역할 변화 속 빅테크·스타트업 움직임 주목

AI 서비스 구현의 중추 역할을 하는 파운데이션 모델은 언어 모델의 기반이 되는 Transformer 방식이 등장한 2010년대 후반부터 지속적으로 더 커지고 다양한 태스크를 수행할 수 있는 대형 AI 모델을 중심으로 산업의 성장이 이뤄지고 있다. 대규모 언어 모델을 통한 AI 서비스 운영이 처리해야 할 데이터 양이 과다하게 증가하여 AI 반도체에 대한 과다 수요 문제와 데이터센터 운영을 위한 높은 전력 수요 문제로 인한 문제가 지속적으로 거론되고 있다.

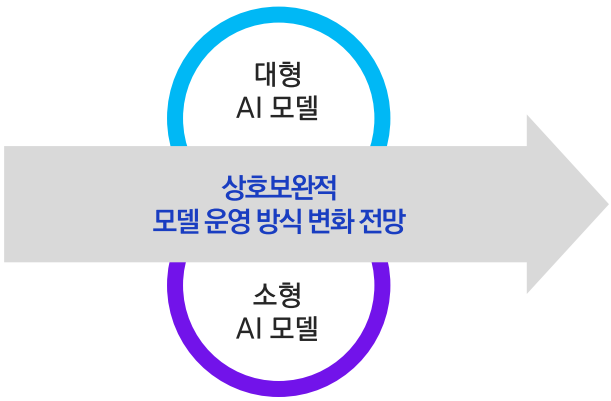
> "파운데이션 모델을 개발하는 빅테크 기업 … 특정 분야에 특화된 기능을 보여주는 경량화 모델을 개발하는 빅테크 및 스타트업의 움직임 주목"

대규모 언어 모델 등의 대형 AI 모델을 활용하는 기존 AI 서비스와 달리, 온디바이스 AI는 소형 AI 모델을 기반으로 서비스를 운영한다는 점에서 앞선 단점을 해소할 수 있는 방안으로 주목받고 있다. 온디바이스 AI 시장이 본격 활성화되면 온디바이스 AI 기능을 활용할 수 있도록 특화되는 소형 AI 모델 시장이 본격적으로 확대될 것으로 전망된다. 소형 AI 모델은 개별 디바이스에 탑재된 데이터 처리 역량과 적은 전력 소모로 구현이 가능한 AI 서비스를 운용하기 위해 온디바이스 AI 기능이 탑재된 각각의 디바이스에서 활용할 수 있는 기능에 특화된 형태로 다변화될 것으로 예상된다.

AI 모델은 온디바이스 AI 시대가 본격화 됨에 따라 대형 AI 모델과 소형 AI 모델이 세분화된 영역으로 기술의 발전이 지속될 것으로 보인다. 대규모 언어 모델 등을 포함한 대형 AI 모델은 소형 AI 모델이 세부적으로 발달 될 수 있도록 충분한 학습 과정과 모델의 기초 설계를 강화하여 파운데이션 모델로서 더욱 고도화된 방향으로 설계될 것이다. 소형 AI 모델은 스마트폰, 가전 등 AI 모델이 탑재되는 개별 온디바이스 AI용 디바이스에서 요구되는 결과값의 퀄리티를 높일 수 있도록 개발이 이루어질 것으로 전망된다.

AI 모델 개발 시장에서는 특히 다양한 기능을 갖춘 대형 AI 파운데이션 모델을 개발하고 있는 구글, 마이크로소프트 등의 빅테크 기업과 특정 분야에 특화된 기능을 보여주는 경량화 모델을 개발하는 빅테크와 스타트업 기업의 움직임에 주목하여야 한다.

[ 대형 AI 모델과 소형 AI 모델의 상호 보완적 방향으로 발달할 것 ]




**대형 AI 모델: 범용적 기능 강화를 위한 학습 중심 강화**

- 대형 AI 모델은 학습 데이터의 확대 및 기초 설계의 기술적 강화를 통해 소형 AI 모델이 만들어지는 파운데이션 모델로서의 역량 강화를 중심으로 개발될 것으로 보임

**소형 AI 모델: 특화된 태스크를 효율적으로 수행하는 기능 강화**

- 소형 AI 모델은 온디바이스 AI용 디바이스 등의 탑재되어 세부적인 태스크를 전력 소비 등을 최소화하여 효율적으로 수행할 수 있도록 효율성을 강화하는 방향으로 발달될 전망임

Source: 삼정KPMG 경제연구원

> 효율적이고 안전한 생성형 AI 기술을 활용하기 위해서는 AI 모델이 안정적으로 운영될 수 있는 기술적 통제 방안을 마련하여야 함

## E(Explainable AI): 기술적 데이터 통제 방안 마련으로 설명 가능한 AI 시스템 구축

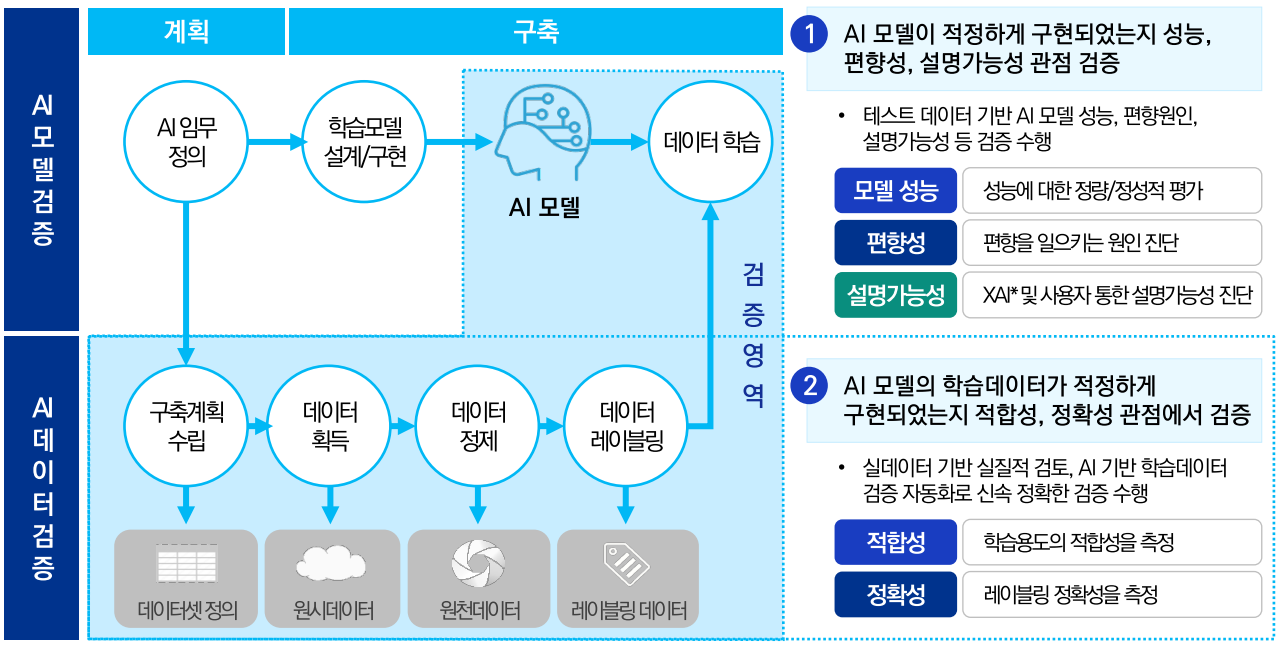
온디바이스 AI는 이용자가 직접적으로 사용하는 가장 가까운 디바이스에서 이용자에게 AI 모델의 분석을 통한 결과값을 제시하기 때문에 이용자의 생활과 밀착된 분야에서 활용되는 경우가 많다. 이로 인해, 온디바이스 AI를 기반으로 운영되는 AI 서비스가 분석 및 도출하는 결과물은 이용자의 일상 생활에 많은 영향을 줄 수 있다는 점에서 AI가 분석하여 도출하는 결과물이 올바른 방향으로 분석되도록 통제하는 방안을 마련하는 것이 중요하다.

생성형 AI 모델의 기술적 통제 방안을 마련하기 위하여 온디바이스 AI용 모델을 개발하고 AI 서비스를 운영하는 기업은 AI 모델 및 서비스에 대한 성능, 편향성 및 설명가능성에 대한 검증과 생성형 AI 모델에 적용되는 데이터 등이 적합, 정확한 지에 대한 검증 및 통제 방안을 마련하여야 한다.

대형 AI 모델과 소형 AI 모델 모두 검증 시 AI에게 주어진 임무가 적절하게 정의되었는지, AI 모델이 적정하게 구축되었는지를 검증하는 과정과 AI 모델이 도출하는 결과물에 타당성을 검증할 수 있는 과정을 마련하여야 한다.

데이터를 통하여 많은 것을 학습하는 생성형 AI 모델에서는 적절한 모델을 구축하는 것과 함께 적절한 데이터가 학습될 수 있도록 통제하는 것이 주요하다. 이를 위하여, 생성형 AI가 학습하는 데이터의 구축, 획득, 정제 및 레이블링 등의 데이터 학습 과정에서 데이터가 가지는 적합성과 정확성을 분석하여야 한다.

[ AI 모델 및 데이터의 기술적 통제 방안 마련 ]




* XAI(Explainable Artificial Intelligence) – 인공지능의 행위와 도출한 결과를 사람이 이해할 수 있는 형태로 설명하는 방법론

Source: 삼정KPMG

# [ Glossary ] AI(인공지능) 관련 용어 설명

| 용어 | | 설명 |
|---|---|---|
| GPT | Generative Pre-trained Transformer | 사전 훈련된 생성 변환기. GPT 모델의 뿌리는 2017년 구글이 발표한 트랜스포머(Transformer)로, 자연어 처리 모델로서, 다음에 오는 단어나 문장을 예측하며 맞추는 방식으로 문장의 생성을 위한 모델 |
| 온디바이스 AI | On-Device AI | 디바이스 내 탑재된 AI 모델을 활용하여 AI 서비스를 운영하는 방식으로 AI 서비스 운영 과정에서 인터넷 등의 외부 연결이 필요하지 않다는 특징을 지님 |
| 생성형 AI | Generative AI | 기존 대규모 데이터의 패턴을 자기지도 학습하여 명령어(Prompt)에 따라 새로운 이미지, 영상, 음악, 텍스트, 코드 등의 콘텐츠를 생성하는 인공지능 기술 |
| 멀티모달 | Multi-Modal | 시각, 청각을 비롯한 여러 인터페이스를 통해 정보를 주고받는 것을 말하는 개념으로, 다양한 채널의 정보를 받아들여 학습하고 사고하는 AI를 '멀티모달 AI'라고 지칭 |
| LLM | Large Language Model | 사람의 언어를 분석하여 사람들이 대화하는 방식과 문장이 구성되는 방식을 이해하여 초거대 AI가 답을 할 수 있도록 하는 모델 |
| sLM | small Language Model | 상대적으로 작은 양의 데이터로 훈련되거나, 간단한 구조를 가진 AI 언어 모델. 이러한 모델은 대규모 언어 모델(LLM)보다 훨씬 적은 매개변수를 가짐 |
| NLP(자연어처리) | Natural Language Processing | 인간의 언어를 컴퓨터가 이해하고 해석하여 그에 맞는 반응을 할 수 있도록 하는 컴퓨터 과학의 한 분야로, 언어 데이터를 분석하고 처리하는 데 필요한 방법론과 알고리즘을 포함 |
| 클라우드 | Cloud | 데이터센터 등의 외부 자원을 활용하여 인터넷 컴퓨팅 시스템을 운영하는 방식으로 네트워크 등의 ICT 인프라를 기반으로 운영됨 |
| GPU | Graphics Processing Unit | 영상, 이미지 등의 그래픽 데이터 처리를 위해 고안된 고성능 처리장치로 다량의 데이터 처리에 유리한 병렬형 데이터 처리 구조를 가지고 있음 |
| 앰비언트 컴퓨팅 | Ambient Computing | 이용자 주변에 일상적으로 존재하는 컴퓨팅 시스템을 의미하는 용어로 이용자가 인지하지 못할 때에도 주변에서 알아서 일을 처리해주는 시스템 |

Source: 언론보도 종합, 삼정KPMG 경제연구원

# How KPMG can help

삼정KPMG AI 센터는 고객의 AI Transformation 실현을 위해 고객 니즈와 기회를 선제적으로 파악하여 AI 도입 전략, Use Case 발굴부터 적용까지 End-to-End 서비스를 제공합니다. 산업별 경험과 AI 기술을 결합하여 기업의 비즈니스 혁신을 위한 전략적 파트너 역할을 수행하고, 단발성 프로젝트가 아닌 지속적인 가치 창출을 위한 관점으로 접근하고 있습니다.

## 삼정KPMG AI 센터 역할 및 주요 상품

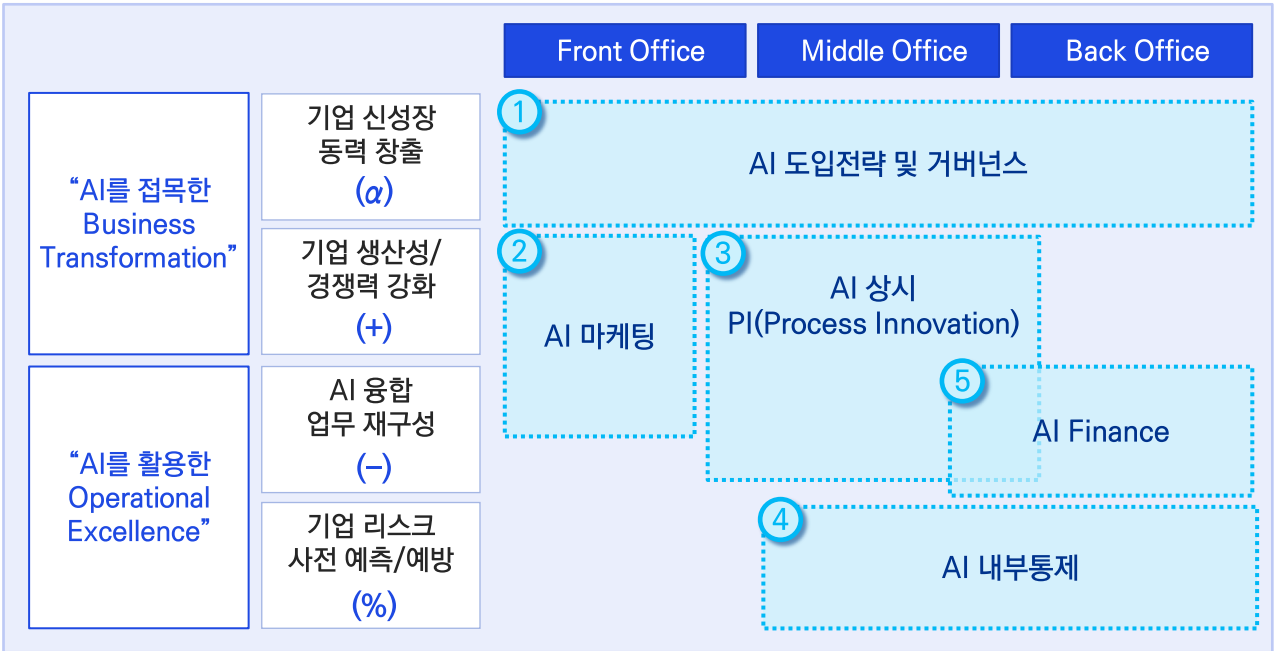
### AI 관련 고객 니즈

- ☑ AI 출현 및 확대에 따른 우리의 대응 방향은?
- ☑ AI 도입을 위해 무엇부터 어떻게 해야 하는가?
- ☑ AI로 고객에게 제공할 수 있는 가치는?
- ☑ AI로 일하는 방식을 혁신하려면?
- ☑ 지속가능한 AI 활용을 위한 고려사항은?

### 삼정KPMG AI 센터 역할

- 업의 미래 재정의 기반 AI 비전 수립
- 단계별 AI 도입전략 및 Use Case 도출
- 상품/서비스 내 AI 내재화
- 업무 프로세스, 조직 구조 재정의
- AI 리스크 관리 및 거버넌스 수립

### 삼정KPMG 주요 AI 상품



삼정 KPMG

# Business Contacts

## AI 센터


조재박
부대표
T 02-2112-7514
E jaeparkjo@kr.kpmg.com

이동근
전무
T 02-2112-7587
E tongkeunlee@kr.kpmg.com

이준기
상무
T 02-2112-0615
E jlee199@kr.kpmg.com


## 전자정보통신엔터미디어산업 전문팀


염승훈 Industry Leader
부대표
T 02-2112-0533
E syeom@kr.kpmg.com

전철희
부대표
T 02-2112-0355
E cjun@kr.kpmg.com

박성배
부대표
T 02-2112-0304
E sungbaepark@kr.kpmg.com

한상현
부대표
T 02-2112-0387
E sanghyunhan@kr.kpmg.com

민성진
전무
T 02-2112-0852
E smin@kr.kpmg.com

장현민
전무
T 02-2112-0546
E hyunminjang@kr.kpmg.com

노원
전무
T 02-2112-0313
E wroh@kr.kpmg.com

강인혜
전무
T 02-2112-0363
E ikang@kr.kpmg.com

최이현
전무
T 02-2112-0505
E yeehyunchoi@kr.kpmg.com

안창범
전무
T 02-2112-0312
E cahn@kr.kpmg.com


**home.kpmg/kr**