

THE
AI
REPORT

글로벌 AI 가치 사슬
분석을 통한
AI 경쟁력 강화 제언

2024

한국지능정보사회진흥원

「The AI Report」는 인공지능 기술·산업·정책의 글로벌 이슈와 동향, 시사점을 적시에 분석, 인공지능 현안에 빠르게 대응하고 관련 정책을 지원하기 위해 한국지능정보사회진흥원(NIA)에서 기획·발간하고 있습니다.

1. 본 보고서는 방송통신발전기금으로 수행하는 정보통신·방송 연구개발 사업의 결과물이므로, 보고서 내용을 발표할 때는 반드시 과학기술정보통신부 정보통신·방송 연구개발 사업의 연구 결과임을 밝혀야 합니다.
2. 한국지능정보사회진흥원(NIA)의 승인 없이 본 보고서의 무단전재를 금하며, 가공·인용할 때는 반드시 출처를 「한국지능정보사회진흥원(NIA)」이라고 밝혀 주시기 바랍니다.
3. 본 보고서의 내용은 한국지능정보사회진흥원(NIA)의 공식 견해와 다를 수 있습니다.

▶ 발행인 : 황 중 성

▶ 작 성

- 알서포트 전략기획팀 신동형 팀장(donghyung.shin@gmail.com)
- 한국지능정보사회진흥원 인공지능정책본부 미래전략팀 정지선 수석(jjs@nia.or.kr)

목 차

I. AI 시대의 도래와 대한민국의 전략적 위치	1
1. AI 혁신과 글로벌 경쟁력의 재편	1
2. 대한민국 AI 경쟁력 현황과 도전 과제	3
3. 본 보고서의 목적과 구성	6
II. 가치 사슬 관점에서 본 AI 생태계	7
1. 생성형 AI 가치 사슬에 대한 선행 연구	7
2. 본 보고서에서의 생성형 AI 가치 사슬 개요	9
3. 생성형 AI 가치 사슬의 동향과 전망	16
III. AI 가치 사슬로 본 글로벌 기업의 경쟁력	18
1. 빅테크들의 All Round Player 화: AI 가치 사슬 수직 계열화	18
2. AI 컴퓨팅 인프라	31
3. AI 모델 개발	42
4. AI 서비스 개발·배포	49
IV. 대한민국 AI 경쟁력 강화를 위한 제언	59
1. 거대 기업들이 주도하는 AI산업	58
2. 스마트폰 시대의 교훈과 AI 산업에의 적용	60
3. 대한민국의 강점과 전략적 선택	62
4. 결론 : 강점을 살린 전략적 접근의 필요성	68
참고 자료	70

글로벌 AI 가치 사슬 분석을 통한 AI 경쟁력 강화 제언

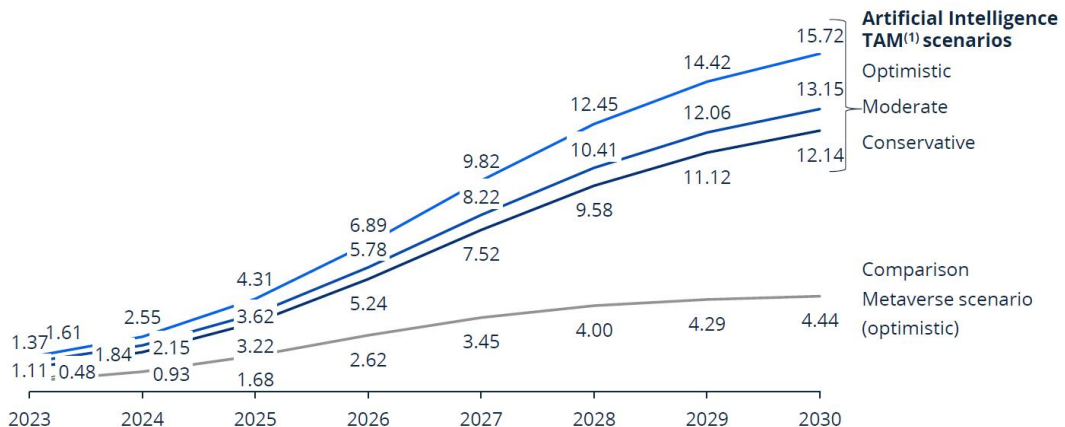
I. AI 시대의 도래와 대한민국의 전략적 위치

1. AI 혁신과 글로벌 경쟁력의 재편

AI의 발전은 단순히 기술 혁신에 그치지 않고, 전 세계 경제와 산업 구조의 근본적인 변화를 이끌고 있다. 글로벌 시장조사기관 Statista의 “Insights Compass 2023¹⁾” 보고서에 따르면, AI의 글로벌 시장 규모는 낙관적 시나리오 기준으로 2030년 약 15.72조 달러(2경 2,569조 원)²⁾에 이를 것으로 전망된다. 이는 차세대 유망 시장 중 하나로 주목받았던 메타버스의 시장 전망치(4.44조 달러)보다 3.5배 큰 규모이다.

[그림 1] 글로벌 AI 잠재 시장(TAM) 규모 전망 시나리오 (2023-2030)³⁾

Comparison of AI total addressable market scenarios and Metaverse (optimistic) scenario in trillion US\$



※ 출처 : Insights Compass 2023(Statista, 2023.07)

1) Statista(2023.7), ‘Insights Compass 2023 – Unleashing Artificial Intelligence’s true potential’

2) 2024.12.16 환율 기준(1달러 = 1,435.52원)으로 계산함.

3) TAM(Total Addressable Market)은 특정 제품이나 서비스가 잠재적으로 도달할 수 있는 최대 시장 규모 또는 이론적으로 가능한 최대 시장 기회의 크기를 의미한다. AI 시장의 경우, 이는 AI가 모든 산업에 완전히 도입되었을 때 창출할 수 있는 전체 경제적 가치를 의미한다.

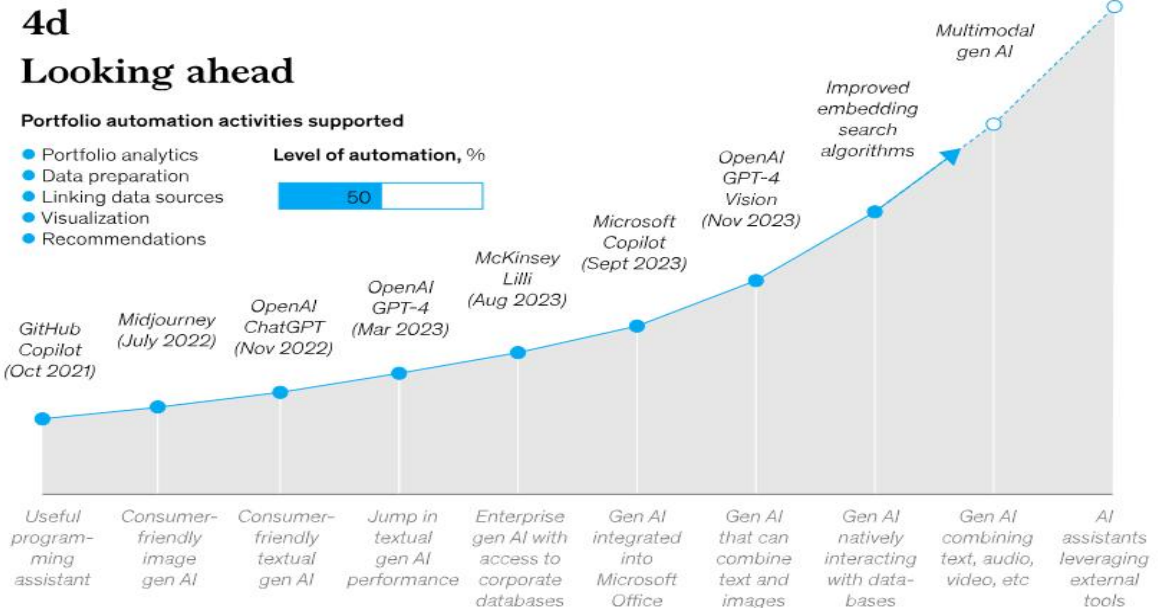
AI 시장은 보수적 시나리오에서도 2023년 1.11조 달러에서 2030년 12.14조 달러로 약 11배 성장할 것으로 예상된다. 이는 AI가 산업의 생산성을 높이고, 기존에 없던 새로운 비즈니스 모델을 창출하며, 경제 구조를 근본적으로 변화시키는 데 기여하고 있기 때문이다. AI가 국가의 미래 경쟁력을 결정짓는 핵심 요소로 부상하고 있음을 명확히 보여준다.

최근 몇 년 사이 GPT-4, Claude, Gemini, Llama 3.3, Mistral과 같은 생성형 AI 모델의 등장은 AI 기술의 가능성을 새로운 차원으로 끌어올리고 있다. 생성형 AI는 단순한 도구를 넘어 복잡한 업무를 수행하는 AI 에이전트(AI Agent)로 발전하며 기업의 업무 자동화를 가속화하고 있다. McKinsey & Company의 분석에 따르면⁴⁾, 2023년에는 기업 업무의 약 10%가 생성형 AI를 통해 자동화되었으며, 2024년에는 이 비율이 30%로 증가했다. 맥킨지는 2025년에는 자동화 비율이 50%를 초과할 것으로 전망하였다. 이러한 생성형 AI는 자동화뿐만 아니라 자율주행, 의료 진단, 금융 서비스 등 다양한 분야에서 혁신을 주도하며 산업 경쟁력을 재편하고 있다.

[그림 2] 생성형 AI의 발전과 기업 업무의 자동화 수준의 변화

Generative AI is developing rapidly, further increasing its usefulness in portfolio optimization.

Summary of global generative AI (gen AI) headlines, % level of automation



※ 출처 : McKinsey & Company(2024), 'Supercharging product portfolio performance with generative AI'

4) McKinsey & Company(2024.11), 'Supercharging product portfolio performance with generative AI'

2. 대한민국 AI 경쟁력 현황과 도전 과제

글로벌 AI 시장에서는 구글, 엔비디아, 마이크로소프트, 아마존, 메타, 애플과 같은 빅테크 기업들이 AI 파운데이션 모델 개발, AI 가속 컴퓨팅 반도체(예: GPU, TPU 등), AI 서비스 분야에서 주도권을 확보해 가고 있다. 또한, OpenAI, Anthropic, Stability AI, Hugging Face와 같은 AI 스타트업들은 생성형 AI 모델 개발과 오픈소스 생태계 지원 등 다양한 영역에서 기술 혁신을 주도하고 있다. 이들 글로벌 AI 선도 기업의 기술력과 자본력은 다른 기업들이 따라잡기 어려운 수준에 도달했으며, AI 생태계 전반에 걸쳐 영향력을 확대하고 있다.

GPT 시리즈를 개발한 OpenAI는 2024년 10월 투자 유치(Funding Round)를 통해 기업 가치 1,570억 달러(약 225조 3,766억 원)를 달성했다⁵⁾. 이는 삼성전자 시가총액(약 332조원)⁶⁾의 약 70% 수준이며, SK 하이닉스의 시가총액(약 131조원)보다 약 1.72배 정도 높다. Claude AI 모델 시리즈를 개발한 앤스로픽(Anthropic)의 기업 가치는 184억 달러(약 26조 5,571억원) 정도로 알려졌다.⁷⁾ 참고로, 한국의 대표적인 IT 기업인 네이버의 시가총액은 약 33.91조원 정도이다. 이처럼 AI 스타트업들은 짧은 시간 안에 기존 IT 강자들과 어깨를 나란히 하며, 글로벌 기술 생태계에서 ‘새로운 거인’으로 부상하고 있다.

AI 컴퓨팅 인프라 분야에서도 빅테크 기업들의 영향력이 커지고 있다. AI 가속 컴퓨팅 반도체 시장에서는 엔비디아(NVIDIA)의 GPU가 여전히 시장을 주도하고 있고, 구글의 TPU, 메타의 AI 칩 등 빅테크 기업들의 자체 AI 칩 개발도 활발히 이루어지고 있다. 또한, 이들 기업은 클라우드 AI 서비스를 통해 AI 기술의 접근성을 높이고 있다. 아마존의 AWS, MS의 애저(Azure), 구글의 GCP 등은 기업들이 손쉽게 AI 기술을 활용할 수 있는 환경을 제공하고 있다.

한편, 생성형 AI 모델의 개발에는 막대한 물리적 인프라 구축, 데이터 확보, AI 훈련 비용, 그리고 전문 인력 유치를 위한 투자가 필수적이다. OpenAI는 2023년 기준, 연간 약 85억 달러(약 12조 2,019억 원)를 인프라 및 인건비로 지출하는 것으로 알려졌다.⁸⁾ 구글 CEO 순다르 피차이(Sundar Pichai)는 2024년 7월 Alphabet 실적 발표에서 “AI와 같은 혁신적 기술에서는 과소투자의 위험이 과잉투자의 위험보다 훨씬 크다.”⁹⁾고 언급하며, 지속적인 투자의 중요성을 강조했다.

5) ‘OpenAI scoops up \$6.6B in funding round at \$157B valuation’, Mobile World Live, 2024.10.3.

<https://www.mobileworldlive.com/ai-cloud/openai-scoops-up-6-6b-in-funding-round-at-157b-valuation/>

6) 국내 기업의 시가총액 정보는 ‘한국거래소 정보데이터시스템’에서 확인(2024년 12월 16일 기준), <http://data.krx.co.kr>

7) Anthropic is expanding to Europe and raising more money, TechCrunch, 2024.5.13.

<https://techcrunch.com/2024/05/13/anthropic-is-expanding-to-europe-and-raising-more-money/>

8) 빅테크는 AI에 얼마를 쏟아붓고 있을까, 디지털투데이, 2024.07.31.

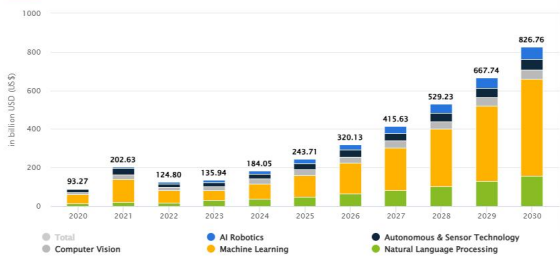
9) Alphabet(2024. 7. 23), ‘2024 Q2 Earnings Call’, <https://abc.xyz/2024-q2-earnings-call>

현재 주도권을 확보한 글로벌 AI 기업들조차 천문학적인 비용을 계속 투입해야 하는 상황에서, 우리는 자본의 한계와 치열한 경쟁이라는 두터운 벽에 직면해 있다. 이러한 현실 속에서 대한민국은 어떤 전략을 취해야 할까?

글로벌 AI 산업의 급격한 변화 속에서 대한민국은 중대한 기로에 서 있다. Statista에 따르면 2024년 전 세계 인공지능 시장 규모는 1,840억 달러로 예상된다. 한편, 국내 인공지능 시장 규모는 2024년에 32억 천만 달러로 추정되며, 이는 세계 시장 규모의 약 1.78%에 해당한다.

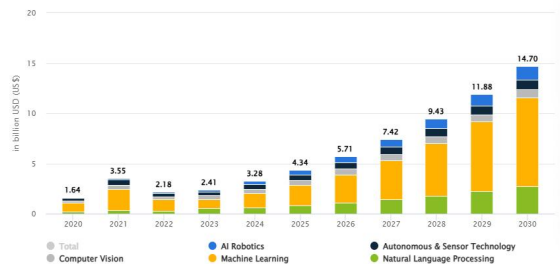
[그림 3] 국내외 AI 시장 규모 전망

전 세계 AI 시장 규모 전망(단위: 10억 달러)¹⁰⁾



※ 출처 : Statista Market Insights(2024.03)

국내 AI 시장 규모 전망(단위: 10억 달러)¹¹⁾



※ 출처 : Statista Market Insights(2024.03)

글로벌화를 통한 경쟁력 강화를 논외로 하더라도, AI 경쟁에서 뒤처진다면 국내 산업에 미치는 영향은 더욱 클 것이다. AI 기술과 산업의 발전은 앞으로 국내 산업 구조와 일자리에 큰 변화를 가져올 것으로 예상된다.

한국직업능력연구원에 따르면 2030년에 전체 일자리의 9.5%가 AI에 대체될 위험에 노출되고, 48.6%는 일자리 자체가 사라지지는 않지만 그 일자리에 있는 사람이 일하는 방식을 바꿔야 할 것으로 나타났다¹²⁾. 반면 AI가 사람의 일자리를 뺏기만 하는 것이 아니라, AI 도입에 따른 생산성 향상이 더 많은 부가가치와 더 많은 일자리를 낳고, 지금껏 없었던 새로운 직업도 만들어낼 수 있을 것이다.

예를 들어 1970년대 ATM 기계 도입으로 은행 출납 사무원 일자리가 줄어든 것으로 예상했으나, 실제로 출납 사무원의 일자리는 더 증가하였다. 자동화 기계 도입이 가져온 편의성과 생산성 향상이 은행 서비스에 대한 수요 증가로 이어졌기 때문이다¹³⁾.

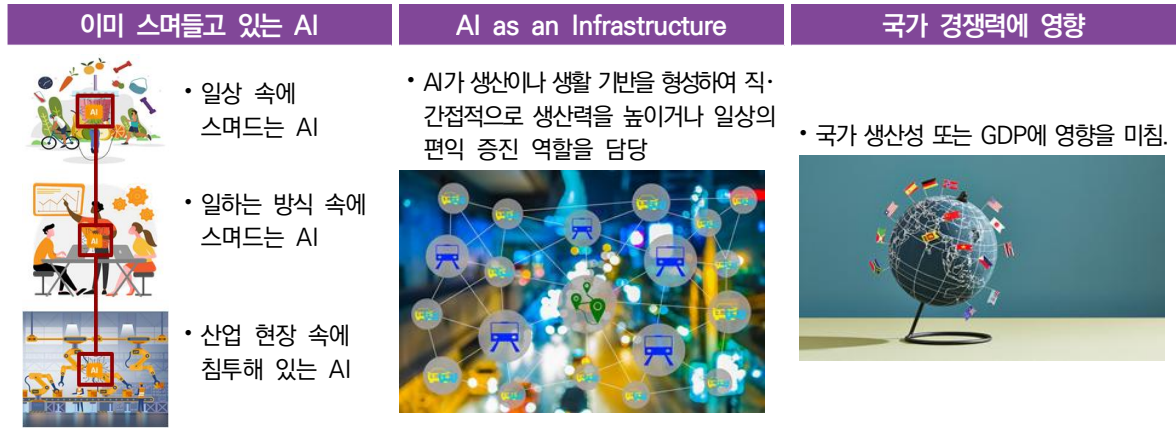
10) <https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide>

11) <https://www.statista.com/outlook/tmo/artificial-intelligence/south-korea>

12) 데이터 기반 미래숙련 전망체계 구축(한국직업능력연구원, 2022)

13) [단독] AI에게 '한국 일자리 미래' 예측시켜봤더니', 조선일보, 2023.03.08.

[그림 4] 국가 경쟁력에 영향을 미치는 AI



AI는 일하는 방식을 변화시키고 산업 현장에 침투하여 생산성을 향상시키는 등 다양한 형태로 우리 삶에 영향을 이미 미치고 있다. 더 나아가 AI는 직·간접적으로 국가 생산성을 높이고, 이를 통해 GDP 성장에 기여하는 핵심 인프라로 발전하고 있다. 이러한 AI의 중요성을 고려할 때, 국가적 차원의 전략적 접근이 필요하다.

우리나라는 AI 파운데이션 모델, AI 컴퓨팅 인프라와 같은 일부 영역에서는 글로벌 선두 기업들과의 격차가 존재한다. 그러나 다른 관점에서 보면, 우리나라는 여전히 IT 강국으로서 세계적 위상을 가지고 있으며, AI 분야에서도 글로벌 경쟁력을 강화할 잠재력을 보유하고 있다. 이런 면에서 AI시대는 우리에게 새로운 기회가 될 수도 있다.

우리는 과거에도 그런 경험을 한 적이 있다. 스마트폰 시대에 한국은 운영체제(OS)와 애플리케이션 프로세서(AP) 시장을 주도하지 않고도 글로벌 시장에서 성공을 거둘 수 있었다. 이는 우리가 모든 분야에서 선두를 달리지 않더라도, 강점을 살릴 수 있는 특정 영역에 선택과 집중을 통해 경쟁력을 확보할 수 있음을 보여준다. 이러한 경험은 AI 시대에도 중요한 시사점을 제공한다.

AI 시대에는 우리의 강점을 살릴 수 있는 특정 영역에 집중하는 전략이 필요하다. 이를 통해 글로벌 경쟁력을 확보하고, 새로운 도약의 기회를 만들 수 있다.

3. 본 보고서의 목적과 구성

우리 실정에 맞는 AI 경쟁력 강화 전략을 모색하기 위해서는 글로벌 AI 산업 생태계의 발전 현황과 핵심 기술 및 서비스 동향을 체계적으로 진단하고 분석할 필요가 있다.

본 보고서는 급변하는 글로벌 AI 생태계를 종합적으로 분석하고, 이를 바탕으로 국내 AI 산업의 경쟁력 제고를 위한 정책 방향을 검토하고 제언하는 것을 목적으로 한다. 특히, 생성형 AI 기술 가치 사슬을 구성하는 주요 영역별 동향을 심층적으로 살펴보고, 국내 AI 산업의 현황을 진단하여 다양한 개선 방안을 도출하고자 한다.

보고서는 크게 세 부분으로 구성된다.

첫째, 생성형 AI를 중심으로 AI 가치 사슬의 전반적인 구조와 동향을 파악한다. 보고서의 작성 배경과 필요성을 설명한 뒤, 가치 사슬 관점에서 AI 산업을 분석하며 데이터, 알고리즘, 하드웨어, 서비스 등 핵심 요소를 검토한다.

둘째, 생성형 AI 가치 사슬의 핵심 영역별 동향을 심층적으로 분석한다. 글로벌 빅테크 기업들의 AI 가치 사슬 계열화 전략을 조명하고, AI 컴퓨팅 인프라, 모델 개발, 서비스 개발 및 배포 등 주요 분야별 최신 트렌드와 주요 기업들의 전략을 상세히 다룬다. 이를 통해 AI 기술과 산업의 발전 방향을 종합적으로 이해할 수 있을 것이다.

마지막으로, 과거 스마트폰 시대의 성공 사례를 기반으로 대한민국의 강점을 극대화할 수 있는 AI 경쟁력 강화를 위한 세 가지 정책 시나리오를 제시한다.

본 보고서는 이러한 구성을 통해 글로벌 AI 생태계의 거시적 관점에서부터 세부 영역별 동향, 그리고 국내 AI 산업의 발전 방향까지 체계적이고 종합적인 분석을 제공한다. 이는 정책 입안자, 기업 관계자, 연구자들이 생성형 AI를 포함한 AI 산업의 현재와 미래를 이해하고 효과적인 전략을 수립하는 데 유용한 기초 자료로 활용될 것으로 기대한다.

II. 가치 사슬 관점에서 본 AI 생태계

1. 생성형 AI 가치 사슬에 대한 선행 연구

AI 기술은 마치 거대한 레고 세트와 같다. 단순히 블록 하나만으로는 아무것도 만들 수 없지만, 여러 종류의 블록들이 정교하게 조립될 때 비로소 멋진 작품이 탄생한다. AI 가치 사슬도 이와 비슷한 원리로 작동한다. 데이터 수집부터 최종 서비스 제공까지, 여러 단계가 유기적으로 연결되어 있어야 진정한 AI의 가치가 실현된다.

McKinsey & Company는 ‘생성형 AI 가치 사슬에서의 기회 탐색’ 보고서¹⁴⁾에서 급변하는 시장에서 투자 기회를 평가하는 데 기본이 되는 통찰을 제공하기 위해 생성형 AI의 가치 사슬을 제시하였다. 맥킨지는 생성형 AI의 가치 사슬을 컴퓨터 하드웨어, 클라우드 플랫폼, 파운데이션 모델, 모델 허브 및 머신러닝 운영(MLOps), 애플리케이션, 서비스라는 6가지 범주로 구분하였다.

[그림 5] 맥킨지가 제시한 생성형 AI 가치 사슬

There are opportunities across the generative AI value chain, but the most significant is building end-user applications.

Generative AI value chain

Opportunity size for new entrants
in next 3–5 years, scale of 1–5

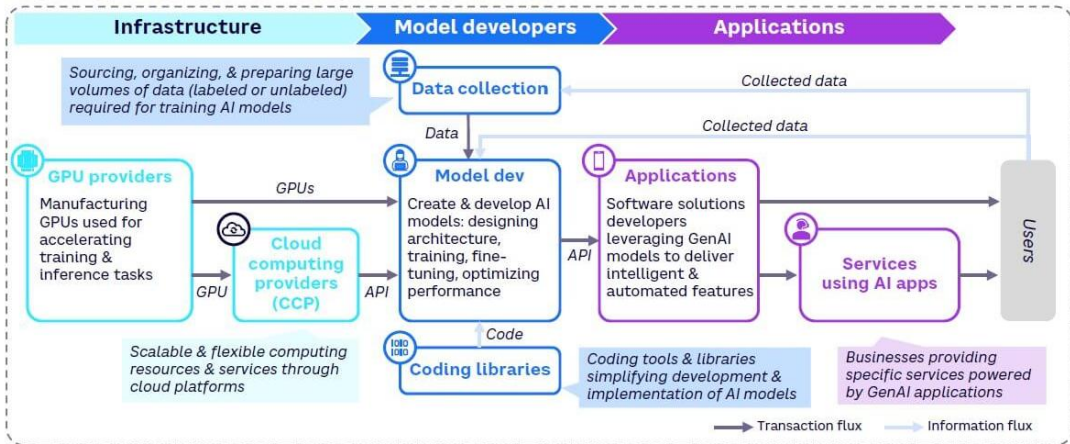


※ 출처 : McKinsey & Company(2023.4.26.), ‘Exploring opportunities in the generative AI value chain’

14) McKinsey & Company(2023.4.26.), ‘Exploring opportunities in the generative AI value chain’

글로벌 경영 컨설팅 기업인 Arthur D. Little(ADL)은 생성형 AI 가치 사슬을 인프라(컴퓨팅), 모델 개발, 애플리케이션의 세 계층으로 구분하였다.¹⁵⁾ ADL은 구글, 아마존, 메타, 애플, Microsoft 등 빅테크 기업들이 모델 개발과 애플리케이션을 포함한 여러 계층에서 생성형 AI 시장에 활발히 참여하고 있다고 분석했다.

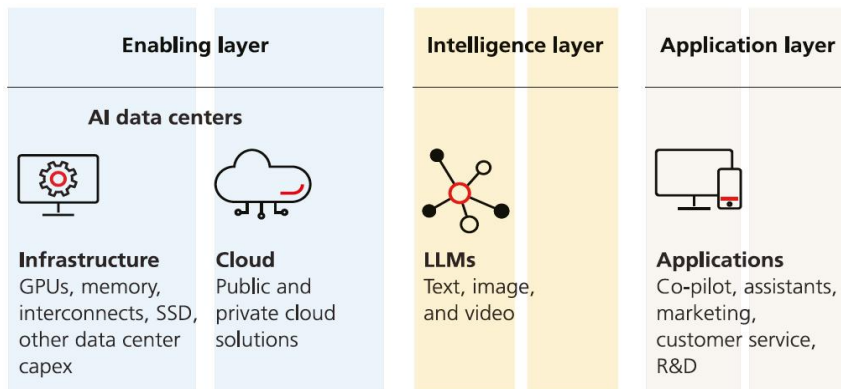
[그림 6] ADL이 제시한 생성형 AI 가치 사슬



※ 출처 : Arthur D. Little(2023.10.), 'Generative artificial intelligence: Toward a new civilization?'

글로벌 금융 기업인 UBS는 생성형 AI의 작동 방식을 조명하고 투자 기회를 파악하기 위한 보고서에서 생성형 AI 가치 사슬을 인에이블링 레이어(enabling layer; 인프라), 인텔리전스 레이어(intelligence layer; 모델), 애플리케이션 레이어(application layer; 서비스)로 세 가지 계층으로 구분하였다.

[그림 7] UBS 제시한 생성형 AI 가치 사슬



※ 출처 : UBS(2024.6.10.), 'Artificial intelligence: Sizing and seizing the investment opportunity'

15) <https://www.adlittle.com/en/insights/report/generative-artificial-intelligence-toward-new-civilization>

2. 본 보고서에서의 생성형 AI 가치 사슬 개요

본 보고서에서는 선행연구 결과를 참고하여 AI 가치 사슬은 크게 세 단계로 구분하였다: AI 컴퓨팅 인프라, AI 모델 개발, 그리고 AI 서비스 개발 및 배포이다. 이와 함께, 데이터 수집과 처리는 이러한 모든 단계에 걸쳐 필수적인 기반 역할을 수행한다. 데이터는 AI 시스템이 작동하고 진화하기 위해 필요한 연료이자 각 단계의 성공을 좌우하는 중요한 요소다.

첫째, AI 컴퓨팅 인프라 단계에서는 AI 가속기(GPU, NPU 등), AI 메모리, AI 데이터 센터와 같은 요소들이 핵심이다. 고성능 컴퓨팅 자원은 복잡한 연산을 빠르게 처리하여 AI 모델의 학습과 추론을 지원하며, AI 데이터 센터에서 다양한 GPU, CPU, NPU 등이 결합되어 최고의 성능을 만들어 낸다.

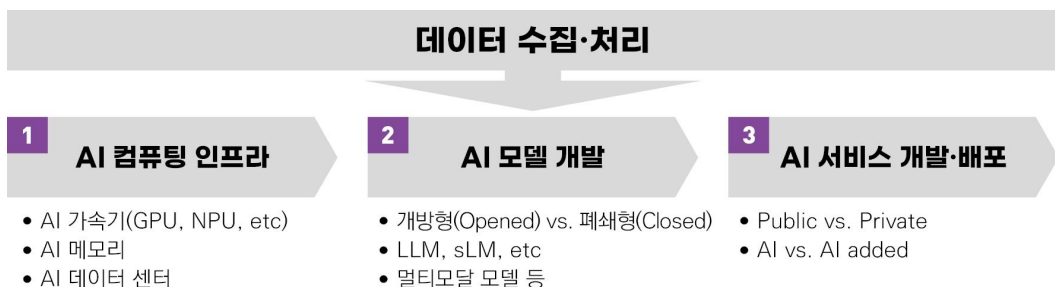
둘째, AI 모델 개발 단계는 개방형(Open) 또는 폐쇄형(Closed) 모델, 대규모 언어 모델(LLM), 멀티모달 모델 등의 다양한 AI 모델을 개발하는 과정이다. 각기 다른 모델들은 다양한 AI 서비스의 기반이 되며, 이를 통해 여러 산업 분야에 적용할 수 있는 혁신적인 AI 솔루션이 탄생한다.

셋째, AI 서비스 개발 및 배포 단계에서는 최종적으로 개발된 AI 모델을 기반으로 다양한 형태의 서비스가 실생활에 적용된다. 공용(Public)과 상용(Private) 서비스, AI 기반 솔루션과 AI부가형(AI-added) 서비스 등 다양한 형태로 AI가 사용자들에게 제공된다. 또 챗GPT와 같은 대중화된 AI 서비스부터 기업 내부용 AI 솔루션 까지, 다양한 방식으로 AI가 활용된다.

데이터 수집과 처리는 이 모든 단계에 걸쳐 지속적으로 이루어지는 과정이다. 데이터는 AI 모델의 학습과 추론에 필요한 필수 자원으로, 각 단계에서 데이터의 수집, 가공, 관리가 효율적으로 이루어져야만 AI의 성능이 최적화 된다. 이 과정은 AI의 성능을 좌우하는 가장 중요한 요소로, 가치 사슬 전체를 지탱하는 핵심 인프라 역할을 한다.

이러한 세 가지 주요 단계와 데이터를 기반으로 한 AI 가치 사슬은 서로 긴밀하게 연결되어 있다. 더 강력한 컴퓨팅 인프라는 더 복잡한 AI 모델 개발을 가능하게 하고, 혁신적인 AI 모델은 새로운 서비스의 출현을 촉진한다. 또한, 다양한 서비스에서 생성된 데이터는 다시 AI 모델의 성능 향상에 기여하는 선순환 구조를 형성한다.

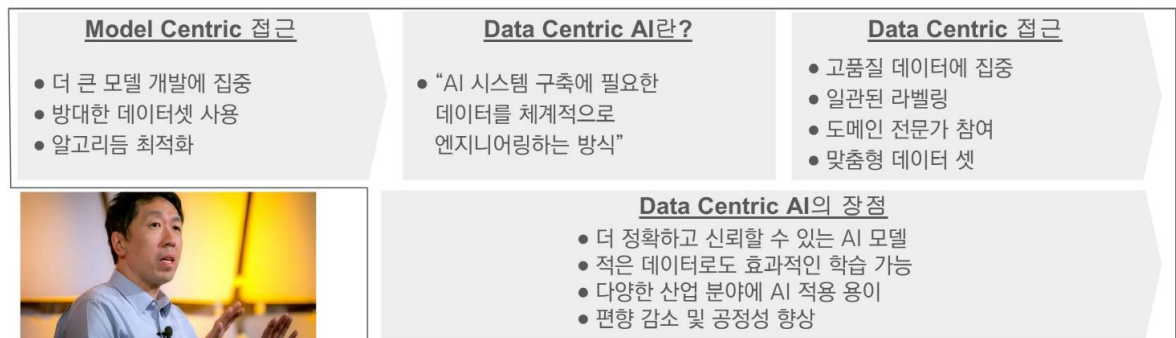
[그림 8] AI 가치 사슬 구조



1.1 데이터 수집 및 처리

AI의 핵심 원료는 단연 데이터다. 방대하고 질 좋은 데이터 없이는 강력한 AI 알고리즘도 무용지물이 되고 만다. 딥러닝의 대가 앤드류 응(Andrew Ng) 역시 “AI의 발전에 있어 큰 진전을 이루고 싶다면 알고리즘을 만드는 것보다 데이터를 수집하는 데 주력하라”고 조언¹⁶⁾한 바 있다.

[그림 9] 앤드류 응의 「데이터 중심 AI(Data Centric AI)」¹⁷⁾



☞ AI의 성공은 알고리즘보다 데이터의 품질에 달려 있음.

실제로, 최근 AI 기술을 선도하는 글로벌 빅테크 기업들은 자사 서비스를 통해 수집한 데이터를 기반으로 AI 모델을 고도화하는 데 주력하고 있다. 예를 들어, 구글은 검색과 유튜브 같은 서비스를 통해 확보한 방대한 데이터를 활용하여 자연어 처리 분야의 AI 모델을 개발하고 있다. 메타 역시 고객 정보를 AI 학습 데이터로 활용하기 위해서 양식을 제출하지 않으면 자동 동의로 유도하는 등 고객 데이터를 확보하기 위해 노력하고 있다¹⁸⁾. 이처럼 AI 개발에 필요한 대규모 데이터를 확보하기 위해서는 체계적인 수집 체계와 가공 역량이 중요하다.

스케일AI(Scale AI)는 데이터 수집과 처리에서 중요한 역할을 맡고 있는 대표적인 기업으로, 2016년 설립 이후 ‘데이터 라벨링’ 사업을 중심으로 성장했다. 이 회사는 인간의 주석 작업과 기계 학습 알고리즘을 결합해 AI 학습에 최적화된 고품질 데이터를 제공하고 있다.

데이터 라벨링은 AI가 학습할 수 있도록 데이터를 분류하고 정리하는 작업이다. 예를 들어, 자율주행차에 필요한 데이터는 도로의 차선, 신호등, 보행자 등을 구분하는 작업이 필요하다. 스케일AI는 수작업 라벨링과 알고리즘 기반 자동화를 결합해 대규모 데이터를 처리하면서도 높은 정확도를 유지한다. 즉, 사람과 AI가 협업해 라벨링 작업을 수행하며, AI 모델이 실제 환경에서 신뢰성 있게 작동하도록 돕는다.

16) AI 학습용 데이터 플랫폼과 표준화 동향 (박영진, 2022)

17) Why it's time for 'data-centric artificial intelligence' (BrownSara, 2022)

18) 메타, AI 학습 데이터 안전장치 마련...양식 제출 안 하면 '자동 동의' (임대준, 2023)

또한, 데이터의 품질은 전처리 작업을 통해 결정된다. 앤드류 응 교수가 강조했듯이, 어쩌면 좋은 알고리즘보다 더 중요한 것은 데이터의 품질일 수 있다. 데이터 레이블링(labeling), 클렌징(cleansing) 같은 전처리 작업이 필요하며, 이는 마치 재료의 껍질을 벗기고 불순물을 제거하는 과정과 같다. 이러한 과정을 통해 고품질의 학습 데이터가 확보될 수 있으며, 이 데이터가 곧 AI의 성능을 좌우한다. 스케일AI는 이러한 데이터 전처리 작업에서 탁월한 성과를 내며, AI 모델의 성능을 극대화하는 데 기여하고 있다.

스케일AI의 주요 고객으로는 오픈AI, 핀터레스트(Pinterest), 에어비엔비(AirBnB), 도어대시(DoorDash) 등 다양한 산업 분야의 기업들이 포함된다. 이들은 자율주행, 자연어 처리, 컴퓨터 비전, 로봇틱스 등에서 스케일AI의 라벨링 데이터를 활용해 AI 모델을 학습시키고 있다. 이는 데이터의 품질과 처리 방식이 AI 성능에 얼마나 중요한지를 잘 보여준다.

AI 데이터 수집과 처리의 단계는 단순히 많은 데이터를 모으는 것을 넘어, 이 데이터를 AI가 이해하고 학습할 수 있도록 정제하는 작업이 필요하다. 사물인터넷(IoT), 스마트폰 등 다양한 경로로 수집된 데이터를 AI 학습에 적합한 형태로 가공하고, 개인정보 보호와 보안 문제를 해결하기 위한 데이터 거버넌스 체계를 수립하는 것이 중요한 과제이다. 스케일AI와 같은 기업들은 이러한 데이터 처리의 중요성을 잘 알고 있으며, 이를 통해 AI 산업에서 중요한 역할을 담당하고 있다.

[그림 10] AI 기반 라벨링 기업 스케일AI(Scale AI)¹⁹⁾²⁰⁾

회사 개요	사업 모델								
 <p>SCALE VALUED AT NEARLY \$14 BILLION</p> <ul style="list-style-type: none"> • '16년 샌프란시스코 설립됨. • 기업가치: '24년 140억\$ '21년 73억\$. AI 용 데이터 라벨링 사업이 중심이며, 사람의 수작업과 ML 알고리즘을 결합해 고품질 데이터를 저비용으로 구현하는 것이 핵심 	<table border="1"> <tr> <td>데이터 라벨링 사업모델</td><td> <ul style="list-style-type: none"> • 주요 서비스 : Scale API를 통한 데이터 주석 및 라벨링 • 가격 책정 : 데이터 복잡성, 양, 요구 사항에 따른 맞춤형 견적 • 엔터프라이즈 솔루션: 대규모 고객을 위한 맞춤형 </td></tr> <tr> <td>차별화 요소</td><td> <ul style="list-style-type: none"> • 인간 전문가와 AI 알고리즘의 결합으로 높은 정확도 제공 • 다양한 데이터 유형 처리 능력 • 지속적인 R&D 투자로 기술 혁신 </td></tr> <tr> <td>주요 고객·분야</td><td> <ul style="list-style-type: none"> • 고객: 오픈AI, 핀터레스트, AirBnB, 도어대쉬, 리프트 등 • 분야: 자율주행, 자연어 처리, 컴퓨터 비전, 로봇틱스, 이커머스 </td></tr> <tr> <td>성장 요인 / 도전 과제</td><td> <ul style="list-style-type: none"> • 성장요인: AI 데이터 수요 증가, 기술 기업 파트너십, 혁신기술 • 도전과제: 윤리적 문제, 데이터 보안, 경쟁 심화 </td></tr> </table>	데이터 라벨링 사업모델	<ul style="list-style-type: none"> • 주요 서비스 : Scale API를 통한 데이터 주석 및 라벨링 • 가격 책정 : 데이터 복잡성, 양, 요구 사항에 따른 맞춤형 견적 • 엔터프라이즈 솔루션: 대규모 고객을 위한 맞춤형 	차별화 요소	<ul style="list-style-type: none"> • 인간 전문가와 AI 알고리즘의 결합으로 높은 정확도 제공 • 다양한 데이터 유형 처리 능력 • 지속적인 R&D 투자로 기술 혁신 	주요 고객·분야	<ul style="list-style-type: none"> • 고객: 오픈AI, 핀터레스트, AirBnB, 도어대쉬, 리프트 등 • 분야: 자율주행, 자연어 처리, 컴퓨터 비전, 로봇틱스, 이커머스 	성장 요인 / 도전 과제	<ul style="list-style-type: none"> • 성장요인: AI 데이터 수요 증가, 기술 기업 파트너십, 혁신기술 • 도전과제: 윤리적 문제, 데이터 보안, 경쟁 심화
데이터 라벨링 사업모델	<ul style="list-style-type: none"> • 주요 서비스 : Scale API를 통한 데이터 주석 및 라벨링 • 가격 책정 : 데이터 복잡성, 양, 요구 사항에 따른 맞춤형 견적 • 엔터프라이즈 솔루션: 대규모 고객을 위한 맞춤형 								
차별화 요소	<ul style="list-style-type: none"> • 인간 전문가와 AI 알고리즘의 결합으로 높은 정확도 제공 • 다양한 데이터 유형 처리 능력 • 지속적인 R&D 투자로 기술 혁신 								
주요 고객·분야	<ul style="list-style-type: none"> • 고객: 오픈AI, 핀터레스트, AirBnB, 도어대쉬, 리프트 등 • 분야: 자율주행, 자연어 처리, 컴퓨터 비전, 로봇틱스, 이커머스 								
성장 요인 / 도전 과제	<ul style="list-style-type: none"> • 성장요인: AI 데이터 수요 증가, 기술 기업 파트너십, 혁신기술 • 도전과제: 윤리적 문제, 데이터 보안, 경쟁 심화 								

19) How Scale AI Became a \$7 Billion AI Data Powerhouse: Business Model Breakdown (Page 21 Team, 2023)

20) Scale AI - Founding Story, Features, Business Model and Growth (Team TBH, 2023)

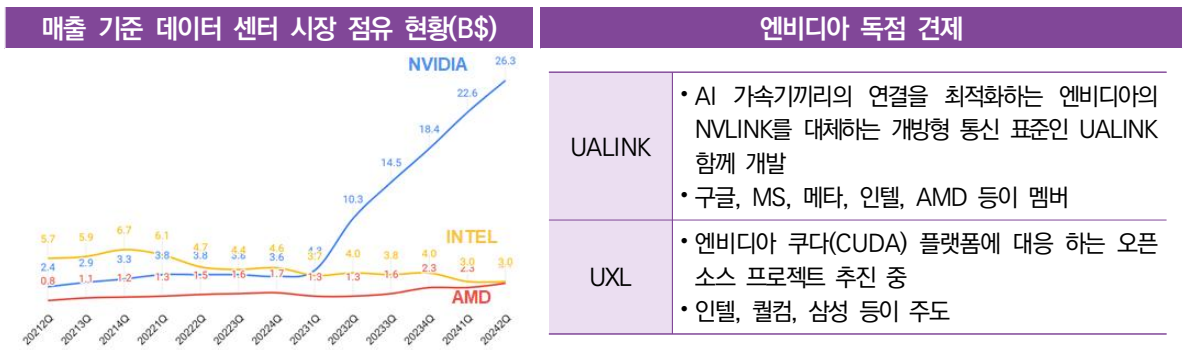
1.2. AI 컴퓨팅 인프라

AI 모델의 학습과 추론에는 방대한 연산량이 소요되므로, 이를 뒷받침할 고성능 하드웨어 인프라가 필수적이다. CPU와 같은 범용 프로세서로는 처리 속도와 효율성 면에서 한계가 있기 때문에, 딥러닝에 특화된 AI 가속기(AI Accelerator) 기술이 각광받고 있다.

구글의 TPU(Tensor Processing Unit), 테슬라의 도조(Dojo)와 같은 AI 전용 하드웨어는 대규모 행렬 연산을 고속으로 처리하도록 설계되어 기존 대비 수십 배의 성능 향상을 이뤘다. 메타와 MS도 이러한 AI 인프라에 대한 투자를 강화하고 있다. 예를 들어, 메타는 2024년 상반기에 두 개의 24,576 GPU 클러스터를 구축하여 라마3(Llama 3) 훈련에 사용하고 있으며, 2024년 말까지 600,000개의 엔비디아(NVIDIA) H100 GPU에 해당하는 컴퓨팅 파워를 갖춘 인프라 구축을 목표로 하고 있다²¹⁾. 뿐만 아니라 이들은 엔비디아의 새로운 GPU인 블랙웰(Black Well)이 출시됨에 따라 그에 따른 최신 컴퓨팅 인프라 확보도 진행할 것으로 예상된다. 또 MS와 오픈AI는 최대 1,000억 달러 규모의 데이터 센터 프로젝트를 계획 중으로, AI 슈퍼컴퓨터인 “스타게이트(Stargate)”를 2028년에 출시할 예정이다.²²⁾ 이처럼 주요 글로벌 기업들은 AI 컴퓨팅 인프라에 대한 막대한 투자를 통해 초거대 AI 모델의 학습을 지원하고 있다.

AI 데이터 센터 시장은 이러한 가속 컴퓨팅 인프라를 기반으로 빠르게 성장하고 있다. 엔비디아는 이 시장에서 독점적인 지위를 확보하고 있으며, 자사의 GPU와 쿠다(CUDA), NVLink를 통해 AI 학습에 필수적인 컴퓨팅 파워와 연결성을 제공하고 있다. 하지만 이러한 엔비디아의 독점을 견제하려는 움직임도 활발하다. 구글, MS, 메타, 인텔, AMD 등은 개방형 통신 표준인 UALINK를 통해 NVLink를 대체하려 하고 있으며, 인텔, 퀄컴, 삼성 등은 오픈소스 프로젝트 UXL을 추진하여 엔비디아의 쿠다(CUDA) 플랫폼을 대체하고자 하고 있다.

【 그림 11 】 엔비디아 중심의 AI 컴퓨팅 인프라와 이를 견제하려는 시도²³⁾



※ 출처 : Data center segment revenue of Nvidia, AMD, and Intel from 2021 to 2024, by quarter(Thomas Alsop, 2024)

21) Building Meta's GenAI Infrastructure (Kevin LeeAdi, 2024)

22) Microsoft and OpenAI planning \$100 billion data center project: report (The Economic Times Tech, 2024)

23) Data center segment revenue of Nvidia, AMD, and Intel from 2021 to 2024, by quarter (AlsopThomas, 2024)

1.3. AI 모델 개발

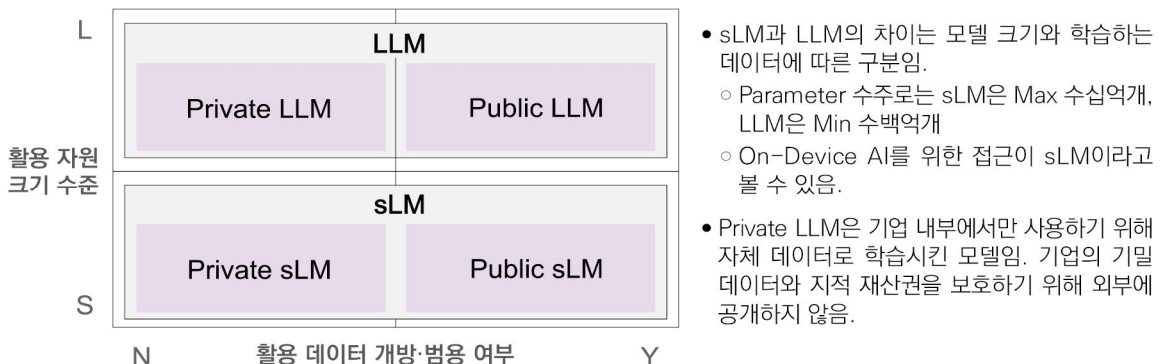
AI 알고리즘은 데이터에서 유의미한 특징과 패턴을 학습하여 모델로 구현된다. 이러한 모델의 성능은 데이터를 어떻게 해석하고 모델링하느냐에 따라 좌우된다. 최근에는 사전 학습된 언어 모델(Pre-trained Language Model)을 기반으로 한 전이 학습(Transfer Learning)이 각광받고 있으며, 대표적으로 GPT와 라마(Llama)와 같은 초거대 AI 모델이 있다. 이들은 대규모 코퍼스(Corpus)²⁴⁾를 통해 사전 학습되어 기존의 기계 학습보다 뛰어난 언어 이해력과 문제 해결 능력을 보여주며, 자연어 처리(NLP) 분야에서 새로운 지평을 열었다.

멀티모달화(Multimodal)도 중요한 트렌드로 자리 잡고 있다. 오픈AI의 소라(Sora)는 2024년 2월 출시되었을 당시 일반인들에게 개방되지 않았으나, 12월에는 유료 가입자는 누구나 활용할 수 있게 공개되었다. 또한 런웨이(Runway) Gen2는 일반인들이 쉽게 영상 제작을 시도할 수 있게 만들었다. 중국의 클링AI(Kling AI)는 초기에는 물리 법칙이 적용되지 않고 파스타를 먹을 때 포크가 입을 뚫고 나가는 월 스미스 영상으로 유명했지만 이제는 생동감 있는 영상 제작이 가능 할 만큼 기술이 발전했다. 이러한 멀티모달 AI는 텍스트뿐만 아니라 이미지, 영상, 음성 등 다양한 데이터를 학습하고 생성할 수 있는 능력을 갖추고 있다.

초거대 AI 모델의 발전과 더불어 소형 언어 모델(sLM) 개발도 주목받고 있다. sLM은 모델 압축, 가지치기(Pruning), 양자화(Quantization) 등의 기술을 활용해 제한된 하드웨어 환경에서도 효율적으로 구동될 수 있다. 이는 비용 절감과 에너지 효율성 측면에서 중요한 의미를 가진다.

LLM과 sLM은 모델 크기와 학습 데이터 규모로 구분된다. LLM은 방대한 데이터와 고성능 자원을 활용하여 학습되며, 기업들은 민감한 데이터를 보호하기 위해 프라이빗 LLM을 채택하고 있다. 반면, sLM은 작은 데이터셋과 제한된 자원으로 학습되며, 온디바이스 AI(On-Device AI)에 탑재될 수 있을 정도로 경량화된 형태로 제공된다. 이러한 구분을 통해 기업들은 작업 환경과 요구에 맞는 모델을 선택해 활용할 수 있다.

[그림 12] AI 모델 분류






24) 자연어 처리(NLP) 분야에서 대규모 텍스트 모음을 의미

LLM의 경량화도 지속적으로 진행되고 있다. 예를 들어, 앤스로픽(Anthropic)은 클로드3(Claude3) 이후 고성능과 경량화를 동시에 달성한 소넷3.5(Sonnet3.5)를 출시했다. 오픈AI 또한 챗GPT-4에 이어 성능을 유지하면서 자원 소비를 줄인 챗GPT-4o mini를 선보이며 이러한 흐름에 동참하고 있다. 이러한 노력은 대규모 모델의 우수한 성능을 유지하면서도 효율적인 자원 활용을 목표로 한다.

이처럼 AI 분야에서는 초거대 모델과 소형 모델이 각각 확장 및 최적화되며 동시에 발전하고 있다. 더불어, 멀티모달 AI의 등장으로 다양한 데이터 형식을 처리하는 능력이 강화되고 있다. 기업들은 이러한 모델들을 활용하여 필요에 맞는 솔루션을 구현하고 있으며, 이는 AI 기술의 발전과 실용화를 더욱 가속화하고 있다.

[그림 13] 최근 AI 모델 트렌드

멀티모달화		경량화	
	<ul style="list-style-type: none">• 오픈AI의 소라(Sora)는 2024년 2월 출시되었을 당시 일반인들에게 개방되지 않았으나, 12월에는 유료 가입자는 누구나 활용할 수 있게 공개됨	앤스로픽	<ul style="list-style-type: none">• Claude3 이후 버전인 Claude3.5는 Opus가 아닌 Sonnet 버전부터 출시함.
	<ul style="list-style-type: none">• 런웨이(Runway) Gen2는 일반인들에게도 개방된 영상 제작 AI임.		오픈AI
	<ul style="list-style-type: none">• 중국에서 개발된 클링AI(Kling AI)은 초기에 물리법칙이 적용되지 않는 월 스미스 영상으로 유명했었는데 이제는 생동감 있는 영상 제작 가능	기존에는 성능 향상에만 집중한 반면, 높은 비용과 에너지 소모로 자원 소비를 줄이는 경량화가 트렌드	

여기서 간과하지 말아야 할 것이 AI 모델 개발 및 운영 비용이다. '23년 기준 오픈AI는 AI 훈련 및 추론에 70억 달러를 투자하고 있으며, 인건비로도 15억 달러를 지출하고 있다. 이에 반해 2023년 매출은 16억 달러를 기록했으며, 2024년에는 34억 달러를 목표로 하는 등 사실상 지속적인 적자를 감수하고 투자되고 있다.²⁵⁾ 이처럼 AI 모델의 성능 향상을 위해서는 막대한 투자가 필요한 상황이다.

AI 모델 개발에는 세 가지 주요 접근 방식이 있다. ① 직접 개발(in-house)하는 경우, ② 공개된 사전학습 모델을 활용하는 경우, ③ 클라우드 API를 통해 필요한 모델을 호출하여 사용하는 경우이다. 특히, 학습에 막대한 비용이 소요되는 초대형 AI 모델은 기업 간 협력으로 개발하거나 우수한 오픈소스 모델을 도입하는 사례가 증가하고 있다. 이러한 추세는 비용 절감과 개발 효율성을 동시에 달성하기 위한 전략으로 자리잡고 있다.

25) Report claims that OpenAI has burned through \$8.5 billion on AI training and staffing, and could be on track to make a \$5 billion loss (EvansonNick, 2024)

1.4. AI 서비스 개발 및 배포

최근 클라우드 기반의 사전 학습된 고성능 AI 모델이 API 형태로 제공되면서, AI 서비스를 손쉽게 개발하고 배포할 수 있는 환경이 조성되어 AI 서비스의 대중화가 이루어지고 있다. MS의 에저AI(Azure AI) 서비스는 음성인식, 이미지 분석, 자연어 처리 등 다양한 AI 기능을 API로 제공하며, 아마존의 세이지메이커(SageMaker)는 데이터 준비부터 모델 학습, 배포까지 전 과정을 지원하는 머신러닝 플랫폼이다. 한편, 구글의 버텍스AI(Vertex AI)는 AutoML 기능을 통해 코딩 없이 AI 모델을 만들 수 있는 것이 특징이다. AutoML은 데이터 전처리, 모델 선택, 하이퍼파라미터 튜닝 등을 자동화하여 사용자가 머신러닝 프로세스를 쉽게 수행할 수 있도록 지원한다. 이처럼 각 기업은 저마다의 강점을 바탕으로 차별화된 AI 플랫폼 서비스를 제공하고 있다.

이러한 플랫폼들은 자연어 이해, 음성인식, 이미지 분류 등 범용 AI 기능을 제공하는 것은 물론, 개별 기업의 데이터로 추가 학습시키는 커스터마이징도 지원하고 있다. 이를 통해 기업들은 별도의 AI 모델 개발 없이도 자사의 비즈니스에 AI 기술을 손쉽게 접목할 수 있게 되었다. 실제로 금융, 의료, 제조 등 다양한 산업 분야에서 AI 도입이 빠르게 확산되고 있다.

한편, AI 모델을 서비스 형태로 구현하는 과정에서 MLOps(Machine Learning Operations)의 중요성이 부각되고 있다. MLOps는 소프트웨어 개발의 DevOps와 유사하게 AI 모델의 개발, 배포, 운영을 자동화하는 체계이다. 지속적 통합/배포(CI/CD), 모델 버전 관리, 모델 모니터링 등을 통해 안정적이고 효율적인 AI 서비스 제공을 가능하게 한다.

서비스 관점에서 AI는 버티컬AI(Vertical AI)와 일반 소비자용 AI로 구분해 접근할 수 있다. 최근 AI를 활용한 생산성 증대를 목표로, 버티컬AI가 각 산업별로 활발히 도입되고 있다. 버티컬AI는 특정 산업의 문제를 해결하기 위해 최적화된 AI 솔루션으로, 기존의 일반적인 AI가 해결하기 어려운 산업 특화 문제들을 보다 효과적으로 다룬다. 예를 들어, 통신 시장이나 새로운 산업 영역에서는 데이터 확보와 전문성 강화를 통해 새로운 가치를 창출하고 있다. 산업 특화 데이터로 학습된 AI는 높은 정확성과 효율성을 제공하며, 이를 통해 각 산업 내에서 기존에 불가능했던 서비스를 가능하게 한다.

아울러 사용자와의 상호작용이 중요한 AI 서비스의 특성상 AIE이전트(AI Agent)²⁶⁾ 혁신도 주목할 만한 트렌드다. 챗GPT의 음성UX 도입 등 음성 기반 대화형 AI가 그 예시가 될 수 있고, 또 퍼플렉시티(Perplexity)와 같이 다양한 AI 모델을 기반으로 검색 기능을 함께 제공하며 고객 접점을 잡아가는 서비스도 출시 및 확대되고 있다.

26) 엔비디아의 수석 연구과학자인 짐 팬은 AI 에이전트를 “역동적인 세상에서 자율적으로 의사 결정을 내릴 수 있는 AI 모델과 알고리즘”이라고 정의한다.

3. 생성형 AI 가치 사슬의 동향과 전망

AI 기술은 데이터 수집부터 컴퓨팅 인프라, 모델 개발, 서비스 개발 및 배포에 이르는 전 영역에 걸쳐 혁신을 동반하며 발전하고 있다. 이러한 기술적 진전이 선순환 구조를 이루며 AI 산업 발전을 가속화하는 가운데, 향후에는 다음과 같은 주요 트렌드가 주목된다.

기술적으로는 첫째, AI 모델의 대형화와 복잡화가 가속되고 있다. 초거대 언어 모델은 파라미터 수가 수백억에서 수천억에 이르는 등 그 규모가 기하급수적으로 증가하고 있다. 또한 텍스트를 넘어 음성, 이미지, 영상 등 멀티모달로의 발전이 진행되고 있어, 모델의 성능 향상과 활용 범위 확대로 이어지고 있다. 하지만 동시에 학습에 막대한 연산량과 비용이 소요되는 문제와 함께 경량화도 시도되고 있다.

둘째, 데이터의 대규모화와 품질 고도화가 뚜렷한 추세다. 웹 크롤링, IoT 센서, 모바일 기기 등 다양한 경로를 통해 수집되는 데이터의 양이 폭발적으로 증가하는 한편, 데이터 레이블링과 클렌징 등 품질 관리에도 많은 공을 들이고 있다. 이는 AI의 학습 효과를 극대화하기 위한 필수적인 과정으로, 데이터의 품질이 곧 AI의 성능을 좌우하기 때문이다.

셋째, AI 인프라 환경의 클라우드화도 빠르게 진행되고 있다. AWS, GCP 등 주요 클라우드 사업자들은 GPU 기반 고성능 컴퓨팅 자원을 제공하는 동시에, AI 개발 플랫폼과 API 서비스를 통해 기업들의 AI 도입을 적극 지원하고 있다. 이에 따라 자체 인프라 구축 없이도 클라우드 기반으로 AI 기술을 손쉽게 활용할 수 있게 되었다.

산업적으로도 여러 변화가 감지되고 있다. 무엇보다 빅테크 기업들의 대규모 AI 모델 개발이 시장 주도권 경쟁을 불러일으키고 있다. 구글, 메타, MS 등은 웹 검색, 소셜미디어 등 자사 서비스를 통해 확보한 방대한 데이터와 컴퓨팅 자원을 바탕으로 초거대 AI 모델을 개발하고 있다. 이들은 또한 AI 스타트업 인수, 개방형 생태계 조성 등을 통해 다각도로 영향력을 확대하고 있다.

반면, 오픈AI, 앤스로픽(Anthropic) 등 AI 스타트업들은 틈새 분야를 공략하며 빠르게 성장하고 있다. 이들은 대화형 AI, 이미지 생성 AI 등 특화 모델을 개발하고 API 형태로 제공하는 한편, 오픈소스 프로젝트에 참여하여 개발자 커뮤니티와의 협력도 강화하고 있다. 챗GPT의 성공에서 보듯, 독창적인 서비스로 단기간에 시장을 선점하는 사례도 늘어나고 있다.

한편, AI 반도체 기업들의 행보도 주목할 만하다. 엔비디아, 구글 등은 GPU, TPU 등 전용 칩을 앞세워 AI 모델 개발을 위한 컴퓨팅 인프라 시장을 선점하고 있다. 더불어 쿠다(CUDA), 텐서플로우(TensorFlow) 등 개발 프레임워크와 클라우드 서비스까지 제공하며 AI 생태계 전반에서 입지를 굳건히 하고 있다.

이러한 가운데 데이터 확보 경쟁과 AI 칩 개발 경쟁도 그 어느 때보다 치열해지고 있다. 방대한 데이터 확보가 AI 경쟁력의 핵심 요소로 부상하면서, 데이터의 양적 확대뿐만 아니라 질적 고도화를 위한 노력도 한층 강화되고 있다. 아울러 초거대 모델의 기하급수적 연산량 증가에 대응하기 위해, 엔비디아, 구글 등이 고성능 AI 칩 출시 경쟁을 가속화하고 있다.

하지만 최근 들어 AI 분야의 수익성에 대한 의문도 제기되고 있다. 빅테크 기업들이 AI에 연간 약 6,000억 달러를 투자하고 있지만, AI 서비스 애플리케이션을 통한 수익은 아직 기대에 미치지 못하고 있다²⁷⁾. 이는 현재의 AI 기술로는 사람 수준의 지능을 구현하기 어렵고, 실질적인 수익 창출이 쉽지 않기 때문이다. 이에 따라 합성 데이터 생성, 새로운 데이터 소스 발굴, 학습 효율 개선 등 다양한 노력이 진행되고 있다.

또한, 단일 대형 모델보다는 도메인 특화 모델이나 모델 조합 기술이 주목받고 있다. 예를 들어 전문가 믹스(Mixture of Experts) 방식을 통해 여러 전문 모델을 상황에 맞게 호출하는 기술이 GPT-4를 비롯한 최신 대형 모델에 널리 채택되고 있다.

이러한 변화 속에서 우리 기업들은 자사의 강점과 특성에 맞는 차별화된 전략을 수립하고 실행해야 할 것이다. 데이터, 알고리즘, 컴퓨팅 등 AI 핵심 자원을 효과적으로 확보하고 내부 AI 역량을 지속적으로 강화해 나가는 한편, 오픈 이노베이션과 협업을 통해 부족한 부분을 보완하고 신속한 기술 혁신을 도모하는 개방적 자세가 요구된다. 무엇보다 AI가 가져올 사회경제적 변화에 선제적으로 대비하고, 책임감 있는 AI 개발과 활용을 위한 윤리 원칙을 수립해 나가야 할 것이다.

AI는 이제 단순한 기술 트렌드를 넘어 산업과 사회 전반의 판도를 바꿀 핵심 동인으로 자리매김하고 있다. 기술적 잠재력과 파괴력을 정확히 인식하고 국가적 차원의 전략과 실행 로드맵을 마련해 나가는 지혜가 그 어느 때보다 절실한 시점이다. 민관 협력을 바탕으로 AI 경쟁력 확보와 활용에 속도를 내는 동시에, 기술 혁신이 가져올 폭넓은 변화에 대한 사회적 공감대 형성에도 힘써야 할 것이다.

27) “2년 내 LLM 학습 데이터 고갈...데이터 문제로 AI 발전 중단될 것” (임대준, 2024)

Ⅲ. AI 가치 사슬로 본 글로벌 기업의 경쟁력

1. 빅테크들의 All Round Player 화: AI 가치 사슬 수직 계열화

1.1. 구글(Google)

구글은 AI 시대의 핵심 기술 전반을 아우르는 'All Round Player'로서의 입지를 굳건히 하고 있다. 구글은 데이터 수집과 처리, AI 컴퓨팅 인프라, AI 모델 개발, 그리고 AI 서비스 개발 및 배포에 이르는 AI 가치 사슬의 모든 영역에 걸쳐 기술을 개발하고 있다. 이러한 구글의 전략은 단순히 AI 기술을 개발하는 것에 그치지 않고, 이를 전 세계에 걸쳐 서비스 형태로 확장하여 제공하는 데 주력하고 있다는 점에서 두드러진다.

AI 가치 사슬내 구글의 활동을 살펴보면, 구글은 AI 모델 개발부터 인프라 구축, 그리고 AI 서비스 배포에 이르기까지 다양한 활동을 전개하고 있다. 예를 들어, 구글은 초거대 AI 모델의 기반이 된 트랜스포머(Transformer)를 개발하여 언어 모델 연구를 이끌어 GPT와 같은 모델의 토대를 마련했다. 또한 AI 컴퓨팅 인프라를 위해 알파칩과 같은 칩 설계 AI와 텐서(Tensor) 칩과 같은 AI특화 칩을 개발하고, GCP(Google Cloud Platform)를 통해 AI 인프라를 제공하고 있다. 생성형 AI 모델 개발 측면에서는 제미니나 시리즈(Gemini Series)와 같은 다양한 모델을 개발해 AI 서비스와 제품에 적용하고 있다.

구글의 AI 서비스는 일반 사용자가 쉽게 접근할 수 있도록 설계되었으며, 유튜브 추천과 같은 일상적인 서비스에도 AI 기술이 활용되고 있다. 이는 AI 기술이 단순히 연구실에 머무르지 않고, 실제 사용자의 삶에 직접적인 가치를 제공하는 방식으로 확장되고 있음을 보여준다.

[그림 14] 구글의 AI 개요

All Round Player	AI Value Chain 내 구글의 활동	
<p>구글은 AI 시대의 기반이 되는 핵심 기술은 모두 개발했으나 시장화에 실패 그 과실은 '24년(현재) 많은 경우 오픈AI에 빼앗긴 상태임.</p>	AI 컴퓨팅 인프라	<ul style="list-style-type: none"> 추론에 특화된 텐서(Tensor) 칩 개발 GCP를 통해 AI 인프라 제공
	AI 모델 개발	<ul style="list-style-type: none"> Gemini Series(Ultra, Pro, Nano, Flash), PaLM2, Imagen, Codey, Chirp, Veo, MedLM, LearnLM, SecLM, Gemma Series(Code, Recurrent, Pali)
	AI 서비스 개발·배포	<ul style="list-style-type: none"> 검색, 유튜브 등에서 AI를 활용하고 있음. Gemini 서비스 제공

그러나 최근 구글은 오픈AI와 같은 경쟁사의 빠른 움직임에 밀려 일부 시장에서 뒤처졌다는 평가도 받고 있다. MWC 2024에서 딥마인드(DeepMind) CEO인 데미스 허사비스(Demis Hassabis)가 밝힌 바와 같이, 구글은 “100배 더 정확해야 한다”는 목표 아래 AI 시스템의 완성도에 집착한 나머지, 시장에서의 선택에서는 뒤처지게 되었다.²⁸⁾ 반면 오픈AI는 대중에게 결함이 있을 수 있는 모델이라도 빠르게 내놓고, 실제 사용자들의 피드백을 통해 개선해 나가는 전략을 취했다. 이로 인해 챗GPT와 같은 모델이 빠르게 대중화되며 시장에서 큰 가치를 인정받고 오픈AI가 새로운 AI 산업의 리더로 급부상했다.

구글은 이러한 경험을 통해 기술 혁신뿐 아니라 실제 사용자 경험을 통한 학습의 중요성을 인식하게 되었다. 구글의 AI 전략은 이제 기술의 정확성뿐만 아니라, 시장에서 얼마나 빠르게 적용되고 실제로 사용될 수 있는지를 중시하는 방향으로 전환되고 있다. 이는 “AI 기술은 연구실의 성과가 아니라, 실제 사용자들이 느끼는 가치를 통해 평가된다”는 교훈을 구글이 다시금 인식하게 만든 중요한 전환점이 된 것 같다.

결론적으로, 구글은 AI 기술의 모든 가치를 아우르는 종합적인 접근 방식을 취하고 있다. 구글은 AI 원천 기술부터 인프라, 모델 개발, 서비스 배포까지 전방위적으로 AI 생태계를 계열화하고 있으며, 이를 통해 AI 기술의 상업적 가치를 극대화하고 있다.

[그림 15] 기술혁신은 빨랐으나 시장 혁신에 못 다다른 구글²⁹⁾³⁰⁾³¹⁾

MWC 2024 DeepMind CEO	구글이 뒤쳐진 이유
	<ul style="list-style-type: none"> 구글은 AI 기술 혁신에 집중했으나, OPEN AI는 먼저 시장에 서비스를 출시하고 빠른 속도로 확장했음. 일반 대중은 챗GPT 사용할 때, 분명히 결함이 있음에도 불구하고 이러한 시스템을 사용할 준비가 되어 있어 있고 환각이 있고 사실이 아닐 수 있음. 구글은 이러한 시스템을 출시하기 전에 “100배 더 정확해야” 한다고 했지만, OPEN AI가 출시하면서 수백만 명의 사람들이 그로 부터 가치를 찾음.

28) 'MWC: Google Admits Being Blitzed By OpenAi In The Gen Ai Race', channelnews, 2024.2.28.

<https://www.channelnews.com.au/google-admits-being-blitzed-by-openai-in-the-generative-ai-market/>

29) Keynote 3: Our AI Future (HassabisDemis, 2024)

30) Google DeepMind CEO on AGI, OpenAI and Beyond - MWC 2024 (Ben WodeckiDeborah, 2024)

31) [MWC 24] '알파고의 아버지' 데미스 허사비스 “5년 후 AI 기기는 모바일 아닐 것” (허준, 2024)

1.1.1. AI 컴퓨팅 인프라

구글은 AI 컴퓨팅 인프라 분야에서도 선도적인 역할을 하고 있다. 구글이 개발한 TPU(Tensor Processing Unit)는 AI 모델 학습과 추론 속도를 극대화하기 위해 설계된 전용 하드웨어로, 딥러닝을 위한 최적의 컴퓨팅 성능을 제공한다. TPU는 특히 ASIC(Application Specific IC), 즉 특정 용도에 맞게 설계된 칩으로, 일반 용도인 GPU와 달리 AI 연산에 특화된 구조를 가지고 있다.

TPU와 GPU의 가장 큰 차이는 용도와 효율성에 있다. GPU는 범용 프로세서로서 여러 역할을 수행할 수 있지만, TPU는 AI 학습에 필요한 대규모 병렬 연산을 처리하는 데 특화되어 있다.

구글의 TPU는 AI 모델 학습을 위해 대량의 데이터를 병렬로 처리할 수 있는 능력을 갖추고 있다. 예를 들어, 최신 TPU v6는 이전 세대에 비해 최대 4.7배의 컴퓨팅 성능 향상과 67%의 에너지 효율 개선을 이루었다. 이는 대규모 AI 모델의 학습과 추론에서 중요한 의미를 가진다. 또, 자율주행차와 같은 시스템은 실시간으로 방대한 데이터를 처리해야 하므로, 이러한 컴퓨팅 성능이 안전과 직결된다. 구글의 TPU는 이러한 높은 연산 요구를 충족시켜, AI가 더 빠르고 정확하게 작동할 수 있도록 돕는다. 추가적으로 TPU v6 개발에는 또 다른 구글의 비밀 병기가 숨겨져 있는데 알파칩(AlphaChip)이라는 AI 반도체 설계용 AI이다³²⁾. 알파칩으로 인해 구글이 출시하는 TPU 개발 속도는 더 빨라지고 성능과 에너지 효율성은 더 강화될 것으로 기대할 수 있다. 또한, 구글은 TPU를 통해 자사의 클라우드 서비스인 GCP(Google Cloud Platform)에서 고성능 AI 컴퓨팅 인프라를 제공하고 있다. 이를 통해 기업들은 자체적인 고성능 하드웨어를 구축할 필요 없이, 구글의 인프라를 활용하여 AI 모델을 학습시키고 배포할 수 있게 되었다. 이는 특히 스타트업이나 중소기업이 AI 기술을 도입하는 데 큰 도움이 된다.

[그림 16] 구글의 TPU(Tensor Processing Unit)³³⁾³⁴⁾



Google TPU(Tensor Processing Unit)은

- 구글이 자체 개발한 딥러닝 모델 학습 및 추론 가속 위해 특별히 설계한 ASIC(Application Specific IC)임.

Google TPU가 GPU에 비해서는

- GPU는 범용 프로세서인 반면, TPU는 AI에 특화됨
- GPU는 병렬 처리로 다양한 연산을 수행하고, TPU는 대규모 행렬 연산에 특화된 ML 최적화 아키텍처를 가짐

'24년 출시된 6세대는

- 이전 세대 TPU v5e 대비 칩당 최대 컴퓨팅 성능 4.7배 향상 및 에너지 효율 67% 향상됨.

32) 구글의 AlphaChip과 TPU 전략 분석 (신동형)

33) Google Cloud TPU로 AI 개발 가속화 (Google Cloud, 2024)

34) Announcing Trillium, the sixth generation of Google Cloud TPU (Google Cloud, 2024)

1.1.2. AI 모델 개발

구글은 AI 모델 개발에 있어서도 선두 주자로서의 입지를 굳건히 하고 있다. 구글의 AI 모델 개발 전략은 크게 상용 모델과 공개(Open-source) 모델로 나눌 수 있다. 상용 모델은 기업들이 특정 용도로 사용할 수 있도록 최적화된 모델이고, 공개 모델은 누구나 사용하고 개선할 수 있는 오픈소스 형태로 제공된다.

구글의 대표적인 상용 AI 모델로는 제미니(Gemini) 시리즈가 있다. 활용처에 따라 다양한 기능을 갖추는 등 여러 종류의 모델을 포함하고 있다. 예를 들어, 제미니 울트라(Gemini Ultra)는 최고의 성능을 자랑하며, 제미니 프로(Gemini Pro)는 대용량 작업과 광범위한 활용에 적합하게 설계되었다. 또한 제미니 나노(Gemini Nano)는 경량화된 모델로, 주로 자원 제약이 있는 환경에서 활용된다. 이러한 다양한 모델들은 각각의 용도에 맞게 최적화되어 있어, 기업들은 자신들의 필요에 맞는 AI 모델을 선택할 수 있다.

또한 구글은 이메진(Imagen)과 같은 모델을 통해 이미지 생성의 가능성을 보여주고 있다. 이메진(Imagen)은 사용자가 입력한 텍스트를 바탕으로 이미지를 생성하는 모델로, 예술 작품을 창작하는 것과 같은 창의적인 작업에 널리 활용될 수 있다.

구글은 상용 모델 외에도 오픈소스 AI 모델도 제공하고 있다. 젤마(Gemma) 시리즈가 대표적이며 그 중 팔리젤마(PaliGemma)는 다목적 경량화된 비전-언어 모델로 동시에 이미지와 텍스트를 이해하며 광범위한 비전 언어 처리 역량을 갖추고 있다. 예를 들어, 글로벌 고객을 대상으로 하는 전자상거래 회사는 팔리젤마를 활용하여 자동으로 다양한 언어로 제품 설명을 제공할 수 있다. 이는 마치 한 번의 버튼 클릭으로 여러 언어로 번역된 제품 카탈로그를 손에 쥐게 되는 것과 같은 효과를 제공한다.

구글은 또한 처프(Chirp)와 같은 모델을 통해 음성 인식 기술도 제공하고 있다. 처프는 100개 이상 언어의 음성을 처리할 수 있다. 이는 글로벌 사용자들이 각자의 언어로 AI와 소통할 수 있는 환경을 제공하며, AI의 접근성을 높이는 중요한 역할을 한다.

구글 클라우드는 머신러닝(ML) 시스템을 위한 지속적 통합(CI), 지속적 전달(CD), 지속적 학습(CT)을 구현하고 자동화하는 MLOps를 적용하고 있다. 이를 기반으로 개발된 구글 AI 모델들은 MLOps가 적용되어 모델 개발 과정에서 발생할 수 있는 여러 문제를 효율적으로 해결할 수 있다.

[그림 17] Google의 AI 모델³⁵⁾

상용 (Closed)	범용	멀티모달	일반 작업	• Gemini Ultra(최고 성능), Gemini Pro(대용량·광범위 작업)
		텍스트	일반 작업	• PaLM : 추론 및 코딩 능력 향상
		경량화	온디바이스	• Gemini Nano(경량 모델), Gemini Flash(최적화 모델)
	특화	이미지 생성	시각 콘텐츠	• Imagen(이미지 생성, 텍스트 렌더링 지원)
		음성 인식	음성 처리	• Chirp(100개 이상 언어 지원, 정확도)
		비디오 생성	영상 콘텐츠	• Veo(영화적 효과, 상세한 톤 이해)
		텍스트	의료	• MedLM(의료 워크플로우 최적화, 맞춤형 솔루션)
			교육	• LearnLM(학습자 적응형, 호기심 자극)
			사이버 보안	• SecLM(데이터, 보안 사용 사례 강화)
공개 (Open)	오픈소스	텍스트	일반 작업	• Gemma(책임있는 설계), RecurrentGemma(높은 처리량)
		코드 생성	개발	• CodeGemma(지능형 코드 완성 및 생성, 다중 언어 지원)
		멀티 모달	시각-언어 작업	• PaliGemma(강력한 미세 조정, 다양한 언어 지원)

※ 출처 : Learn about our leading AI models(Google, 2024)

1.2. 메타(Meta)

메타는 AI 가치 사슬의 모든 단계에서 두각을 나타내고 있는 대표적인 ‘All Round Player’이다. 메타는 AI 원천 기술 연구에서부터 AI 컴퓨팅 인프라, AI 모델 개발, 그리고 AI 서비스 개발 및 배포에 이르기까지 전반적인 분야를 아우르고 있다. 이러한 전방위적인 참여를 통해 메타는 AI 가치 사슬 내 확장과 고도화의 혜택을 자사 서비스로 효과적으로 연결하고 있다.

메타는 2013년부터 AI 연구소인 FAIR(Facebook AI Research)를 설립하여 AI 연구와 모델 개발에 지속적으로 투자해 왔다. 이를 통해 라마(Llama), SAM과 같은 다양한 AI 모델을 개발하고, 그 성과를 개방하여 AI 생태계 발전에 기여하고 있다. 예를 들어, 메타가 개발한 라마(Llama)는 대규모 언어 모델로, 다양한 언어 작업에서 뛰어난 성능을 발휘한다. 이는 마치 글로벌 사회에서 여러 언어를 구사할 수 있는 번역가를 키워내는 것과 같다. 이러한 연구 개발 노력은 메타가 AI 산업과 가치 사슬에서 주도적인 역할을 하도록 만들어주고 있다.


AI 컴퓨팅 인프라에 있어서도 메타는 AI 반도체를 직접 개발하여 활용하는 것은 물론, 엔비디아 GPU를 통해 고성능 AI 컴퓨팅 인프라를 구성하고 있다. 메타는 특히 자사의 대규모 서비스, 예를 들어 페이스북과 인스타그램 같은 소셜 네트워크 서비스(SNS)를 통해 수집되는 방대한 데이터에서 중요한 정보를 찾아내어 AI 모델의 학습에 활용해 AI 성능을 극대화하고 있는 것이다.

메타는 오픈소스 AI 모델 개발에도 가장 적극적으로 참여하고 있다. 메타가 개발한 오픈소스 AI 모델들은 다양한 산업에서 활용되고 있다. 이러한 오픈소스 전략은 AI 기술을 개방하고, 더 많은 개발자와 연구자들이 메타의 AI 기술을 활용할 수 있도록 함으로써 AI 기술의 발전을 가속화하는 역할을 하고 있다.

35) Learn about our leading AI models (Google AI, 2024)

그리고 메타는 자사에서 개발한 AI 기술을 자사의 주요 서비스에 통합하고 있다. AI 에이전트는 페이스북, 인스타그램, 왓츠앱 등의 플랫폼에서 사용자 경험을 개선하는 데 중요한 역할을 하고 있다. 예를 들어, 인스타그램의 콘텐츠 추천 시스템은 AI 모델을 활용해 사용자가 좋아할 만한 콘텐츠를 정교하게 추천한다. 이는 마치 사용자의 취향을 잘 아는 큐레이터가 맞춤형 전시회를 열어주는 것과 같은 느낌을 준다. 이러한 AI 기술 통합은 사용자 경험을 크게 향상시키고, 사용자들이 플랫폼에 더 오랜 시간 머무르게 함으로써 메타의 경쟁력을 강화하는 데 기여하고 있다.

[그림 18] 메타의 AI 개요

All Round Player	AI Value Chain 내 메타의 활동						
 <p>데이터 수집·처리</p> <p>AI 컴퓨팅 인프라 → AI 모델 개발 → AI 서비스 개발·배포</p> <ul style="list-style-type: none"> • 메타는 AI 가치 사슬 내 개방형을 선택 • AI 가치 사슬 확대·고도화의 혜택은 메타 내 서비스로 향하게 	<table> <tr> <td>AI 컴퓨팅 인프라</td><td>• AI 반도체를 직접 개발하여 활용할 뿐만 아니라, NVIDIA GPU도 활용해 AI 컴퓨팅 인프라를 구성</td></tr> <tr> <td>AI 모델 개발</td><td>• 메타의 서비스 강화에 영향을 줄 수 있는 오픈소스 AI 모델을 개발</td></tr> <tr> <td>AI 서비스 개발·배포</td><td>• 메타AI는 메타에서 개발한 AI 에이전트로, 페이스북, 인스타그램, 왓츠앱 등 메타의 다양한 서비스의 기반이 됨.</td></tr> </table>	AI 컴퓨팅 인프라	• AI 반도체를 직접 개발하여 활용할 뿐만 아니라, NVIDIA GPU도 활용해 AI 컴퓨팅 인프라를 구성	AI 모델 개발	• 메타의 서비스 강화에 영향을 줄 수 있는 오픈소스 AI 모델을 개발	AI 서비스 개발·배포	• 메타AI는 메타에서 개발한 AI 에이전트로, 페이스북, 인스타그램, 왓츠앱 등 메타의 다양한 서비스의 기반이 됨.
AI 컴퓨팅 인프라	• AI 반도체를 직접 개발하여 활용할 뿐만 아니라, NVIDIA GPU도 활용해 AI 컴퓨팅 인프라를 구성						
AI 모델 개발	• 메타의 서비스 강화에 영향을 줄 수 있는 오픈소스 AI 모델을 개발						
AI 서비스 개발·배포	• 메타AI는 메타에서 개발한 AI 에이전트로, 페이스북, 인스타그램, 왓츠앱 등 메타의 다양한 서비스의 기반이 됨.						

1.2.1. AI 컴퓨팅 인프라

AI 모델의 학습과 추론에는 방대한 연산량이 필요하다. 이러한 대규모 연산을 지원하기 위해서는 고성능의 하드웨어 인프라가 필수적이다. 일반적인 CPU나 GPU로는 최적화에 한계가 있기 때문에, AI에 특화된 가속기(Accelerator) 기술이 점점 더 중요해지고 있다.

메타는 AI 컴퓨팅 인프라를 강화하기 위해 세 가지 주요 기술적 접근을 하고 있다. 반도체, 슈퍼 컴퓨터, 그리고 데이터 센터다.

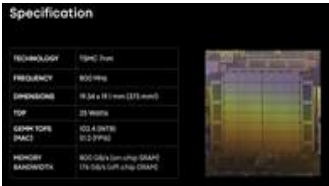



먼저, 메타는 자체적으로 개발한 MTIA(Meta Training and Inference Accelerator)라는 반도체를 통해 AI 모델의 학습과 추론 속도를 높이고 있다. MTIA는 AI 전용 가속기 칩으로, 일반 GPU와 협력해 데이터 센터에서 높은 성능을 발휘한다. 쉽게 말해, MTIA는 수많은 연산 작업을 빠르게 처리할 수 있도록 돕는 특화된 도구라고 할 수 있다. 또, 비디오 데이터를 효율적으로 처리하기 위해 메타는 MSVP(Meta's Scalable Video Processor)라는 ASIC(Application Specific Integrated Circuit)을 개발했다. 이는 비디오 데이터를 빠르고 효율적으로 처리해, 많은 사람들이 동시에 영상을 시청할 때도 끊김 없이 매끄럽게 제공될 수 있도록 한다.

다음으로, 메타는 슈퍼 컴퓨터 인프라를 구축하여 AI 모델의 학습에 필요한 연산 능력을 강화하고 있다. 2022년에 발표된 RSCC(Research Super Cluster)는 엔비디아(NVIDIA)의 DGX A100 GPU 6,080개로

구성된 연구용 슈퍼 컴퓨터다. 2023년에는 이보다 더 많은 GPU를 추가로 도입해, 현재는 약 16,000개의 GPU를 사용하고 있다. 이 슈퍼 컴퓨터는 초거대 AI 모델을 훈련시키기 위해 설계되었다. 이 강력한 컴퓨팅 파워 덕분에 AI 모델의 학습 속도와 성능이 크게 향상되었다.

마지막으로, 메타는 차세대 데이터 센터를 설계하고 있다. 이 데이터 센터는 MTIA와 MSVP를 활용해 AI 작업의 효율성을 극대화하며, 액체 냉각 시스템을 사용해 에너지 소비를 줄이고 있다. 마치 컴퓨터가 뜨거워질 때 팬으로 식히는 것처럼, 데이터 센터에서도 열을 효과적으로 식히기 위해 강력한 냉각 시스템을 사용한다. 이렇게 하면 데이터 센터가 더 적은 에너지로 더 많은 작업을 처리할 수 있다.

[그림 19] 메타의 AI 컴퓨팅 인프라³⁶⁾³⁷⁾

반도체	슈퍼 컴퓨터	데이터 센터
 <ul style="list-style-type: none"> • META의 학습·추론 가속기인 MTIA가 데이터센터 내에서 GPU와 워크로드 배분  <ul style="list-style-type: none"> • META의 비디오트랜스코딩ASIC:MSVP 	 <ul style="list-style-type: none"> • '22.01. 메타는 RSC(Research Super Cluster)를 발표하고 NVIDIA DGX A100 760개, NVIDIA A100 GPU 총 6,080개 • '23.05. NVIDIA DGX A100 2,000개, NVIDIA A100 GPU 16,000개 	 <ul style="list-style-type: none"> • 메타의 차세대 데이터 센터 설계는 현재 제품을 지원하는 동시에 MTIA 및 MSVP를 활용해 비디오를 포함한 AI 최적화 설계, 액체 냉각 AI 하드웨어 포함

1.2.2. AI 모델 개발

메타는 다양한 AI 모델 포트폴리오를 통해 AI 가치 사슬 또는 산업을 확장하고 있다. 이러한 모델들은 텍스트 생성, 이미지 처리, 음성 분석 등 다방면에서 활용될 수 있도록 설계되었으며, 개발자와 기업들이 손쉽게 AI 기술을 활용할 수 있도록 지원한다.

메타의 주요 AI 모델로는 라마(Llama) 시리즈(Llama 3.1, 3.2 등), SAM(Segment Anything Model); SAM, SAM2 등), 심리스엠포티(SeamlessM4T), 코드라마(CodeLlama) 등이 있다. 이들은 각기 다른 분야에서 특화된 기능을 제공하며, AI 기술의 대중화와 활용성을 높이는 데 중요한 역할을 하고 있다.

36) Transforming our infrastructure for the next generation of AI (META AI, 2024)

37) Reimagining Meta's infrastructure for the AI age (Meta AI, 2023)


[그림 20] 메타의 AI 모델 포트폴리오³⁸⁾

공개 (Open)	언어 모델	텍스트 생성	일반 작업	• Llama(오픈소스 LLM(경량 가능)), CodeLlama(코드특화)
	컴퓨터 비전	이미지 분할	시각 인식	• SAM(모든 객체 분할), DinoV2(이미지 자기지도학습)
		이미지 생성	시각 콘텐츠	• CM3Leon(이미지 생성)
	음성 처리	음성 생성·편집	오디오 처리	• Voicebox(음성합성·편집), Audiobox(음성·효과음), Audioraft(오디오 생성)
	다중 모달	번역	언어 처리	• SeamlessM4T(다국어 번역), Seamless(표현력 보존)
	범용	로봇 제어	로보틱스	• Adaptive Skill Coordination(사전 훈련 정책을 교정)
	3D 모델링	3D 생성	XR	• I-JEPA(자기지도학습 기반 3D 표현 학습)
	렌더링	아바타 생성	VR	• Avatar RSC(실시간 아바타 렌더링)

※ 출처 : Celebrating 10 years of FAIR: A decade of advancing the state-of-the-art through open research(Meta, 2023)

먼저, 라마3시리즈는 메타가 공개한 최대 규모의 오픈소스 언어 모델로, 라마 3.1은 GPT-3의 두 배 이상인 4050억 개의 파라미터를 가지고 있다. 학습 데이터 또한 15조 개 이상의 토큰을 포함하여 이전 모델 대비 10배 이상 증가했다. 이를 통해 라마 3.1은 더 복잡하고 미묘한 언어 표현을 이해하고 생성할 수 있게 되었다. 또한, AWS, MS 에저, GCP와 같은 주요 클라우드 플랫폼과의 연동을 통해 더 많은 기업들이 손쉽게 라마 3.1을 활용하고 있다. 예를 들어 미국 통신사업자인 AT&T는 이 모델을 이용해 업무 자동화와 데이터 분석에서 혁신을 이루고 있다.

[그림 21] 메타의 대표적 LLM인 라마(Llama)³⁹⁾⁴⁰⁾

 <p>• Llama 3.1은 최대규모 오픈소스 AI LLM 모델 중 하나임. ◦ 모델 크기 : 최대 405B로 GPT3의 2배 이상 큼. ◦ 학습 데이터 : 15조개 토큰 (이전 버전 대비 10배 이상 증가)</p>	개발자 커뮤니티 확산	<ul style="list-style-type: none"> • 라마 모델은 HuggingFace에서만 '24.08.까지 약 3.5억건의 다운로드 진행됨(이는 약 1년 전 보다 10배 이상 증가함.) • 라마3.1이 출시된 후 라마 모델은 1달동안만 2천만 건 이상 다운로드 진행됨. • 라마 파생모델만 6만건 이상이 있음.
	대형 클라우드 사업자 제휴	<ul style="list-style-type: none"> • 메타는 Llama 3.1 확산을 위해 AWS, MS Azure, Google Cloud 등 주요 클라우드 사업자와 제휴 진행함. ◦ 이들 플랫폼을 통한 라마 모델 사용량은 출시 3개월만에 2배 이상 증가함. 일부 사업자의 경우 '24.01 대비 07 사용량이 10배 이상 증가함.
	다양한 산업 분야 활용	<ul style="list-style-type: none"> • AT&T(고객응대 정확도 33% ↑) 액센추어(ESG 리포팅 업무 생산성 70% ↑), 도어대쉬(소프트웨어 개발 효율화), 쇼피파이(상품 메타 데이터 처리 후 하루 4K~6K건의 추천 작업 진행 중) 등 다양하게 활용


38) Celebrating 10 years of FAIR: A decade of advancing the state-of-the-art through open research (META AI, 2023)

39) 메타 라마3.1(Llama 3.1) 공개로 보는 오픈소스 AI 미래 (신동형)

40) 전년동기 대비 10배, 총3.5억 D/L의 라마(Llama):메타의 오픈소스 모델 혁신을 가속화하다. (신동형)

다음으로, SAM(Segment Anything Model)과 그 후속 버전 SAM2는 이미지와 비디오 처리에 특화된 AI 모델이다. SAM은 이미지나 비디오에서 특정 객체를 정확하게 구분할 수 있도록 설계된 모델로, 예를 들어 사진 속 특정 인물만 골라내거나 영상에서 움직이는 물체를 추적하는 작업이 가능하다. 특히 SAM2는 기존 SAM 대비 비디오 분할 성능과 처리 속도를 크게 개선해, 실시간으로 필요한 정보를 선택적으로 저장하고 장기 기억하는 능력을 갖추고 있다. 즉 SAM2는 여러 개의 사진 앨범 중 필요한 사진만 빠르게 골라내는 것과 같은 기능에 강해 특히 영상 분석과 같은 실시간 처리에 매우 유용하다.

[그림 22] 객체 분할 AI 모델: SAM⁴¹⁾42)

객체 분할 기술인 SAM	SAM 개요와 '24년 출시된 SAM2
	<p>SAM</p> <ul style="list-style-type: none"> • SAM(Segment Anything Model)은 이미지나 비디오에서 특정 객체를 정확하게 구분하는 모델로, 이미지·비디오를 이해하는 능력 향상에 필요함.
	<p>SAM2</p> <ul style="list-style-type: none"> • 통합 아키텍처 : 이미지·비디오 동시 처리 • 스트리밍 메모리 기술 : 실시간으로 필요한 정보만 선택적으로 저장해 효율성 및 장기 기억이 가능함. • 기존 SAM 대비 SAM2는 비디오 분할 성능 및 처리 속도 개선됨.

또한, 메타는 심리스엠포티(SeamlessM4T)와 코드라마(CodeLlama) 같은 모델들을 통해 다국어 번역과 코드 생성 등 특정 목적에 맞는 AI 모델을 제공하고 있다. 심리스엠포티는 다국어 번역에서 표현력을 보존하는 기능을 갖춘 모델로, 여러 언어 간의 원활한 소통을 가능하게 한다. 이는 외국어로 작성된 문서를 쉽게 이해하거나 번역하는 데 큰 도움을 줄 수 있다. 코드라마는 코드를 생성하고 수정하는 데 도움을 주는 모델로, 개발자들이 더 효율적으로 작업할 수 있도록 돕는다. 이러한 모델들은 다양한 분야에서 활용되며, 메타의 AI 생태계를 한층 더 확장시키고 있다.

결론적으로, 메타는 라마 시리즈, SAM 시리즈, 심리스엠포티와 같은 다양한 AI 모델들을 공개하고 개발함으로써 AI 기술의 발전에 기여하고 있다. 이러한 모델들은 텍스트, 이미지, 음성 등 다양한 데이터를 처리하고 생성하는 개발자와 기업들이 AI 기술을 보다 쉽게 활용할 수 있도록 지원하고 있다. 메타의 이러한 접근은 AI 기술의 대중화와 활용성을 높이는 역할을 하고 있다.

41) SAM 2: 이미지와 비디오의 경계를 넘는 혁신적 AI 분할 모델 (신동형)

42) Meta Releases Open Source Segment Anything Model (SAM) (TrujilloCarlos, 2023)

1.2.3. 메타의 서비스

메타는 AI 모델을 활용하여 다양한 서비스를 개발하고 이를 자사의 여러 플랫폼에 적용함으로써 AI의 활용성을 극대화하고 있다. 메타AI(Meta AI)는 메타에서 개발한 AI 에이전트로, 페이스북, 인스타그램, 왓츠앱 등 메타의 다양한 서비스에 적용되어 사용자들과의 상호작용을 혁신하고 있다.

이 AI 에이전트는 단순히 텍스트로 대화를 주고받는 것을 넘어, 멀티 모달(multimodal) 기술을 통해 다양한 유형의 데이터를 실시간으로 처리할 수 있는 능력을 가지고 있다. 이를 통해 사용자들은 음성 인터페이스로 질문을 하거나 이미지를 업로드하고 그에 대한 설명을 들을 수 있으며, 자동 번역 기능을 통해 외국어 대화를 원활하게 이어갈 수 있다. 예를 들어, 여행 중에 현지어로 된 간판을 AI에게 사진으로 보여주면 즉시 번역된 결과를 얻을 수 있는 식이다. 이는 마치 여행 가이드가 곁에서 항상 필요한 정보를 제공해 주는 것과 같은 역할을 한다.

또한, 메타AI는 AI 콘텐츠 생성 기능을 활용해 사용자들이 사진이나 동영상을 편집하고, 그 위에 설명을 추가하거나 필요한 정보를 삽입하는 등의 작업을 간편하게 할 수 있도록 돕는다. 예를 들어, 친구와의 대화에서 우주 테마의 사진을 만들어보자는 요청이 들어왔을 때, 메타AI는 사용자가 원하는 대로 이미지를 생성해주어 대화의 재미를 더할 수 있다. 이러한 기능들은 메타의 LLM(Large Language Model), 특히 라마를 기반으로 운영되며, 이를 통해 복잡한 요청도 자연스럽게 빠르게 처리할 수 있다.

[그림 23] 메타AI⁴³⁾

	<table border="1"> <tr> <td>Meta AI</td><td> <ul style="list-style-type: none"> • Meta에서 개발한 AI Agent로 메타의 Facebook, Instagram, Whatsapp 등 다양한 서비스들에 적용 </td></tr> <tr> <td>특징 (’24 현재)</td><td> <ul style="list-style-type: none"> • 멀티 모달과 실시간 처리 능력 : ①음성 인터페이스, ②이미지 인식 및 편집, ③자동 번역 및 더빙, ④AI 콘텐츠 생성, ⑤개인화, ⑥기업용 • Llama를 포함한 메타의 LLM(Large Language Model)을 기반으로 운영됨. </td></tr> </table>	Meta AI	<ul style="list-style-type: none"> • Meta에서 개발한 AI Agent로 메타의 Facebook, Instagram, Whatsapp 등 다양한 서비스들에 적용 	특징 (’24 현재)	<ul style="list-style-type: none"> • 멀티 모달과 실시간 처리 능력 : ①음성 인터페이스, ②이미지 인식 및 편집, ③자동 번역 및 더빙, ④AI 콘텐츠 생성, ⑤개인화, ⑥기업용 • Llama를 포함한 메타의 LLM(Large Language Model)을 기반으로 운영됨.
Meta AI	<ul style="list-style-type: none"> • Meta에서 개발한 AI Agent로 메타의 Facebook, Instagram, Whatsapp 등 다양한 서비스들에 적용 				
특징 (’24 현재)	<ul style="list-style-type: none"> • 멀티 모달과 실시간 처리 능력 : ①음성 인터페이스, ②이미지 인식 및 편집, ③자동 번역 및 더빙, ④AI 콘텐츠 생성, ⑤개인화, ⑥기업용 • Llama를 포함한 메타의 LLM(Large Language Model)을 기반으로 운영됨. 				

1.3. MS(마이크로소프트)

MS는 AI 시대에 빠르게 대응하기 위해 협력을 통한 전략적 접근을 하고 있다. MS는 AI 가치 사슬 내에서 전 영역을 아우르는 All Round Player로서의 위치를 강화하고 있으며, 이를 통해 AI 컴퓨팅 인프라부터 AI 모델 개발, 서비스 개발 및 배포에 이르는 전 과정을 포괄적으로 지원하고 있다. 특히, 오픈AI와의 협력은 MS의 AI 역량을 크게 강화하는 데 중요한 역할을 하고 있다.

43) Meta AI:일상을 혁신하는 지능형 비서의 진화와 Meta의 전략 (신동형)

먼저, MS는 AI 원천기술 연구에서 개발·응용에 이르기까지 오픈AI와 협력하여 AI 연구, 개발 시장을 선도하고 있다. MS는 오픈AI의 최대 주주로 이를 자산화해 자사의 AI에 적용 및 발전시키고 있다. 이 협력은 MS가 최신 AI 기술을 빠르게 도입하고 발전시키는 데 큰 도움이 되고 있다.


또한, MS는 AI 컴퓨팅 인프라의 강화에도 힘쓰고 있다. 오픈AI뿐만 아니라 엔비디아(NVIDIA)와의 협력을 통해 고성능 AI 데이터 센터를 준비하고, 이를 바탕으로 대규모 AI 모델의 학습과 추론을 지원하고 있다. 이를 통해 MS는 AI 컴퓨팅의 기반을 튼튼히 하여, 더 많은 연산 작업을 빠르고 효율적으로 처리할 수 있도록 하고 있다. 예를 들어, AI 모델이 복잡한 문제를 해결하기 위해 수많은 계산을 동시에 해야 하는데, 이는 마치 여러 명의 학생이 각자 다른 문제를 푸는 상황에서 모두가 동시에 답을 도출해야 하는 것과 같다. MS의 AI 컴퓨팅 인프라는 이러한 복잡한 연산 작업을 신속하게 처리할 수 있도록 돕는다.

AI 모델 개발 분야에서 MS는 대형 언어 모델(LLM)은 오픈AI와 협력하여 제공하며, 소형 언어 모델(sLM)은 자체 개발로 구축해 다양한 AI 모델 포트폴리오를 형성하고 있다. MS는 선도적인 AI 모델과 서비스를 빠르게 고객이 활용할 수 있도록 지원하며, sLM을 통해 고객이 필요로 하는 맞춤형 AI 서비스를 선택할 수 있는 유연성을 제공하고 있다.

마지막으로, AI 서비스 개발 및 배포에 있어서 MS는 자사의 클라우드 플랫폼인 에저(Azure)와 오피스(Office) 등 다양한 서비스에 AI를 빠르게 적용하고 있다. 에저는 기업들이 AI 기술을 손쉽게 활용할 수 있도록 지원하는 클라우드 서비스로, AI 모델의 학습, 배포, 관리에 이르는 모든 과정을 포괄적으로 지원한다. 예를 들어, 기업이 자사의 데이터를 활용해 맞춤형 AI 모델을 만들고자 할 때, 에저는 필요한 컴퓨팅 자원을 제공하고, 학습 과정을 자동화해 주는 역할을 한다.

또한, MS의 오피스 제품군에도 AI 기술이 적용되어 사용자들이 문서 작성, 데이터 분석, 프레젠테이션 준비 등 다양한 작업을 더 효율적으로 수행할 수 있도록 돕고 있다. 이는 마치 개인 비서가 옆에서 사용자가 필요한 모든 작업을 도와주는 것과 같은 경험을 제공한다.

[그림 24] MS의 AI 개요

협력을 통한 AI Round Player	AI Value Chain 내 MS의 활동						
 <p>• MS는 협력을 통해 AI 시대를 빠르게 대응하고 있음.</p>	<table border="1"> <tr> <td>AI 컴퓨팅 인프라</td><td>• 오픈AI 및 엔비디아와 협력을 통해 빠르게 AI 데이터 센터 준비 및 사업 확대</td></tr> <tr> <td>AI 모델 개발</td><td>• 오픈AI와 협력을 통해 LLM 대응 및 sLM은 자체 개발하며 포트폴리오 충족</td></tr> <tr> <td>AI 서비스 개발·배포</td><td>• MS의 AZURE 및 OFFICE 등 다양한 서비스에 빠르게 적용함.</td></tr> </table>	AI 컴퓨팅 인프라	• 오픈AI 및 엔비디아와 협력을 통해 빠르게 AI 데이터 센터 준비 및 사업 확대	AI 모델 개발	• 오픈AI와 협력을 통해 LLM 대응 및 sLM은 자체 개발하며 포트폴리오 충족	AI 서비스 개발·배포	• MS의 AZURE 및 OFFICE 등 다양한 서비스에 빠르게 적용함.
AI 컴퓨팅 인프라	• 오픈AI 및 엔비디아와 협력을 통해 빠르게 AI 데이터 센터 준비 및 사업 확대						
AI 모델 개발	• 오픈AI와 협력을 통해 LLM 대응 및 sLM은 자체 개발하며 포트폴리오 충족						
AI 서비스 개발·배포	• MS의 AZURE 및 OFFICE 등 다양한 서비스에 빠르게 적용함.						

1.3.1. AI 모델 개발

MS는 다양한 AI 모델 개발을 통해 대형 언어 모델(LLM)과 소형 언어 모델(sLM)로 이원화된 포트폴리오를 구축하고 있다. 이를 통해 사용자의 다양한 요구에 부응하는 맞춤형 솔루션을 제공하며 AI의 적용 범위를 더욱 확대하고 있다.

LLM(대형 언어 모델)은 방대한 데이터를 학습하여 복잡한 질문에 대한 답변을 할 수 있는 모델이다. 대표적으로 오픈AI(OpenAI)와의 협력으로 탄생한 챗GPT가 있다. MS는 오픈AI에 전략적 투자를 하고, 이와 협력을 통해 MS의 클라우드 플랫폼인 애저(Azure)에 챗GPT를 통합하여 제공하고 있다. 예를 들어, 사용자가 애저(Azure)를 통해 챗GPT를 활용하면 복잡한 텍스트 생성 작업도 간편하게 처리할 수 있으며, 이를 통해 업무 효율성을 높일 수 있다. 이는 마치 모든 지식을 갖춘 도서관 사서가 사용자가 원하는 책을 즉시 꺼내주는 것과 같은 편리함을 제공하는 것이다. 그리고 GPT-4o와 같은 최신 AI 모델도 애저(Azure) 플랫폼에서 빠르게 제공하여 사용자들이 최신 기술을 활용할 수 있도록 지원하고 있다.

또 sLM(소형 언어 모델)인 MS의 파이-3(Phi-3) 시리즈는 온디바이스(On-device) 환경, 즉 스마트폰이나 IoT 기기와 같이 제한된 메모리와 연산 능력을 가진 장치에서 작동하기를 원하는 고객에게 제공된다. Phi-3는 작은 자원에서도 높은 성능을 낼 수 있도록 설계되었으며, 특정 작업에 최적화되어 있다. 예를 들어, 스마트 스피커에서 음성 명령을 처리할 때처럼 단순하고 빠른 응답이 중요한 경우에 유용하다.

파이-3모델은 다양한 버전으로 제공되며, Phi-3-mini부터 Phi-3-Large까지 사용자 환경에 맞춰 선택할 수 있다. 제한된 환경에서도 높은 성능을 발휘할 수 있도록 설계된 Phi-3는 비용 절감과 저전력 소비가 중요한 분야에 특히 유리하다. 예를 들어, 사물인터넷(IoT) 기기에서 실시간 데이터 분석을 수행할 때 Phi-3는 매우 유용하다.

[그림 25] MS의 AI 모델 포트폴리오

LLM	SLM
<div data-bbox="159 1433 351 1483">  OpenAI </div> <div data-bbox="432 1433 649 1483">  Microsoft </div> <div data-bbox="197 1546 337 1582">ChatGPT</div> <div data-bbox="418 1546 658 1582">Microsoft Azure</div> <ul style="list-style-type: none"> • MS는 LLM 모델 분야에서 OpenAI에 전략적으로 투자하며 협력을 통해 모델을 소싱 • MS Azure는 OpenAI와의 협업을 통해서 사업을 확장. 예를 들어 GPT-4o 출시와 동시에 MS Azure에서 'GPT-4o on Azure' 서비스를 제공함. 	<div data-bbox="799 1367 1210 1598">  </div> <ul style="list-style-type: none"> • 제한된 자원 환경에서의 추론 • 온디바이스 등 저지연성이 중요한 분야 적용 • 비용 제약이 있는 분야 적용 • 다양한 포트폴리오 : Phi-3-mini 부터 Phi-3-Large (Specialized LM)

1.3.2. AI 컴퓨팅 인프라

AI 모델의 학습과 추론에는 방대한 양의 데이터를 신속하게 처리할 수 있는 연산 능력이 필요하다. MS는 AI 컴퓨팅 인프라의 중요성을 인지하고, 이를 대규모로 확장하기 위해 오픈AI와 협력하여 강력한 슈퍼 컴퓨팅 인프라를 구축하고 있다.

오픈AI와의 협력은 MS의 AI 컴퓨팅 인프라 강화에 중요한 전환점이 되었다. 2018년, 오픈AI는 기존의 컴퓨팅 리소스가 AI 모델 학습을 위해 필요한 대규모 연산을 충분히 지원하지 못한다는 점을 인지하고, MS에게 슈퍼 컴퓨팅 인프라의 확장을 제안했다. 이 제안은 기존의 한계를 넘어 더 많은 데이터를 학습할 수 있는 환경을 마련하는 데 중요한 계기가 되었다.

MS는 오픈AI와 협력하여 엔비디아 GPU 등 최첨단 슈퍼 컴퓨팅 리소스를 구축했다. 이 슈퍼 컴퓨터는 높은 연산 파워와 저지연 네트워크를 통해 방대한 데이터를 빠르고 정확하게 처리할 수 있도록 돕는다. 이는 AI 모델이 더 복잡한 패턴을 학습하고, 더 높은 정확도로 결과를 도출하는 데 매우 중요한 역할을 한다.

또한, MS의 애저(Azure) 클라우드 플랫폼은 이러한 슈퍼 컴퓨팅 자원을 활용해 사용자들이 필요로 하는 컴퓨팅 능력을 쉽게 활용할 수 있도록 지원한다. 예를 들어, 기업이 대규모 AI 모델을 학습하고자 할 때, 필요한 컴퓨팅 자원을 Azure를 통해 손쉽게 확보할 수 있다.

결론적으로, MS는 AI 컴퓨팅 인프라를 대폭 강화하여 AI 기술 발전을 위한 강력한 토대를 마련하고 있다. 엔비디아와 오픈AI와의 협력을 통해 구축한 슈퍼 컴퓨팅 자원은 AI 모델 학습의 효율성을 극대화하고, 이를 애저를 통해 많은 사용자들이 활용할 수 있게 하여 AI 기술의 대중화를 이끌고 있다.

[그림 26] MS의 AI 컴퓨팅 인프라⁴⁴⁾

OpenAI와의 협력의 출발

- '18년 OpenAI에서 사람들이 컴퓨터와 상호작용하는 방식을 변화시킬 AI 시스템 구축에 대한 아이디어를 MS에 제시함.
- AI 모델과 GPU 결합으로 다양한 언어 작업을 한 번에 처리할 수 있는 잠재력 인지
- 기존 컴퓨팅 리소스의 한계에 부딪히며, MS와 OpenAI간 서로 요구하는 슈퍼컴퓨팅 인프라의 종류와 필요에 대한 규모 이해 시작함.

거대 자본이 필요한 AI 이해

- AI는 대용량 데이터를 보유, 모델 처리를 위한 슈퍼 컴퓨팅 인프라 보유 및 오랜 기간동안 훈련이 필요한 자본 집약적 사업 특징 보유
 - “연구를 통해 배운 것 중 하나가 모델이 클수록 더 많은 데이터를 보유하고 더 오래 훈련할 수록 모델의 정확도가 향상된다는 것임.”
- MS Azure 고성능 컴퓨팅 및 AI 제품 책임자 Nidhi Chappell -

‘19.07.
MS와
OpenAI간
전략적 협력

OpenAI와 함께 최적화 인프라

- MS는 OpenAI의 요구사항에 대응하기 위해서 고성능 슈퍼 컴퓨팅 리소스 구축
 - NVIDIA 쿼텀 인피니밴드를 기반으로 높은 컴퓨팅 파워와 저지연 네트워크와 함께 연결된 수천 개 NVIDIA AI 최적화 GPU 포함.



44) How Microsoft's bet on Azure unlocked an AI revolution (RoachJohn, 2023)

2. AI 컴퓨팅 인프라

2.1. AI 반도체

AI 반도체는 AI 모델이 막대한 데이터를 빠르게 처리하고 학습하는 데 필요한 핵심적인 부품이다. AI 반도체 중에서도 특히 GPU(그래픽 처리 장치)는 AI 연산에 필수적인 역할을 한다. GPU는 원래 게임 그래픽을 처리하기 위해 만들어졌지만, 다수의 연산을 동시에 수행하는 특성 덕분에 AI 학습에도 적합하다는 점이 밝혀졌다. 마치 수백 명의 사람이 동시에 한 작업을 분담해 해결하는 것처럼, GPU는 많은 계산을 병렬로 처리할 수 있다. 그래서 현재 AI 모델의 학습에는 GPU가 주로 사용되고 있다.

대표적인 AI 반도체 기업으로는 엔비디아, AMD, 인텔, 그리고 ARM 등이 있다. 우리나라의 SK 하이닉스, 삼성전자도 AI 메모리 반도체 분야에서 높은 경쟁력을 보유하고 있다. 이들 기업은 AI 모델의 학습과 추론을 위한 다양한 기술과 제품을 개발하고 있다. 예를 들어 엔비디아는 자사의 GPU를 통해 AI 모델 학습에 필요한 연산 속도를 획기적으로 높였으며, 이를 통해 AI의 발전을 견인하고 있다. 인텔 역시 CPU와 더불어 NPU(신경망 처리 장치)라는 AI 연산 특화 반도체를 개발 및 통합하며 시장에 진입하고 있다. NPU는 AI 모델의 특정한 연산을 더 빠르게 처리하기 위해 만들어진 특수 칩으로, CPU와 GPU의 역할을 보완해준다.

AMD와 ARM은 각각 고성능 GPU와 저전력 CPU 기반의 반도체를 공급하며 AI 시장에서 중요한 역할을 맡고 있다. AMD의 GPU는 엔비디아와 경쟁하며 AI 학습에 필요한 컴퓨팅 파워를 제공하고 있으며, ARM은 저전력 프로세서 기술을 통해 모바일 AI와 엣지 컴퓨팅에서 중요한 역할을 하고 있다.

AI 반도체는 AI 시대가 계속되는 한 꾸준히 발전할 것이다. 특히, GPU 공급이 부족한 환경에서 NPU나 ASIC(특정 용도에 최적화된 반도체)과 같은 특화 칩에 대한 수요가 늘어나면서, AI 모델의 학습 속도와 효율성도 더욱 향상될 것으로 보인다. 한편, 엔비디아는 추론 작업이 앞으로 더 중요해질 것임을 인식하고, GPU에 NPU나 추론에 특화된 ASIC의 기능을 포함(Convergence)할 가능성도 크다. 이는 마치 PC나 스마트폰에서 CPU나 AP가 다양한 ASIC 반도체 칩셋을 통합한 것과 유사한 접근 방식이다. 이러한 상황에서, NPU나 ASIC과 같은 특화 반도체를 개발하는 기업들도 시장성이 높은 영역을 빠르게 찾아야 한다. 이를 통해 GPU 중심의 시장을 보완하거나, 통합 칩셋으로 전환하는 전략을 선제적으로 마련할 필요가 있다.

[그림 27] AI 반도체 주요 플레이어 개요



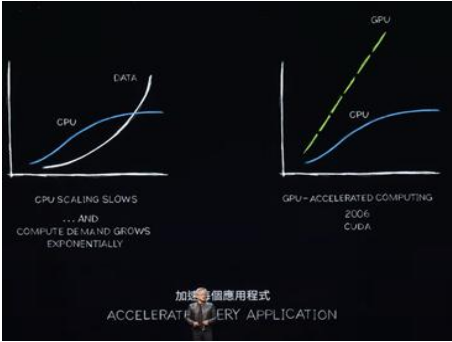
2.1.1. 엔비디아(NVIDIA)

엔비디아는 AI 반도체 시장에서 가장 중요한 플레이어 중 하나로, GPU를 통해 AI 연산의 혁신을 이끌어가고 있다. 엔비디아의 시장 주도 전략은 크게 3가지 관점인데, 첫째 엔비디아는 GPU 아키텍처 혁신을 지속하고 있다. 엔비디아의 GPU는 매 세대별로 딥러닝 성능을 대폭 높이고 있으며, 전력 소모량도 함께 줄여가고 있다. 이러한 지속적인 GPU 아키텍처 혁신을 통해 엔비디아는 AI 학습에 필요한 컴퓨팅 성능을 크게 향상시켰다.

둘째 엔비디아는 소프트웨어 스택의 강화를 통해 독점적 지위를 강화하고 있다. 엔비디아는 CUDA (Compute Unified Device Architecture)와 cuDNN(CUDA Deep Neural Network) 등의 GPU 가속 라이브러리를 제공하여 딥러닝 프레임워크들이 엔비디아의 GPU를 최적화하여 사용할 수 있도록 지원하고 있다. 이러한 소프트웨어 지원을 통해 엔비디아는 하드웨어뿐만 아니라 AI 생태계 전반에서 중요한 위치를 차지하고 있다.

셋째 엔비디아는 AI 생태계 조성에도 많은 노력을 기울이고 있다. 자사의 기술을 적용한 레퍼런스 시스템을 파트너에게 제공하고, 고객사와 협력해 다양한 AI 활용 사례를 창출하고 있다. 예를 들어, 엔비디아의 DGX 시스템은 NVIDIA GPU로 가속화된 워크스테이션 및 서버 제품군으로, AI 연구와 개발을 위한 강력한 컴퓨팅 파워를 제공한다.


[그림 28] 엔비디아의 AI 반도체 전략⁴⁵⁾

가속 컴퓨팅 시대	엔비디아의 전략						
	<table><tr><td>지속적인 GPU 아키텍처 혁신</td><td>• GPU 매 세대 별로 답러닝 성능을 대폭 높여 왔음. 뿐만 아니라 전력 소모량도 함께 줄여가고 있음.</td></tr><tr><td>소프트웨어 Stack 강화</td><td>• CUDA(Compute Unified Device Architecture), cuDNN(CUDA Deep Neural Network) 등 GPU 가속 라이브러리 뿐만 아니라 텐서플로, 파이토치 등 주요 답러닝 프레임워크 지원</td></tr><tr><td>AI 가치 사슬 조성</td><td>• 자사 기술을 적용한 레퍼런스 시스템을 파트너에게 제공하고 고객사와 협력해 다양한 AI 활용 사례 생성 중임. DGX는 NVIDIA GPU로 가속화된 AI 워크스테이션 및 서버 제품군임.</td></tr></table>	지속적인 GPU 아키텍처 혁신	• GPU 매 세대 별로 답러닝 성능을 대폭 높여 왔음. 뿐만 아니라 전력 소모량도 함께 줄여가고 있음.	소프트웨어 Stack 강화	• CUDA(Compute Unified Device Architecture), cuDNN(CUDA Deep Neural Network) 등 GPU 가속 라이브러리 뿐만 아니라 텐서플로, 파이토치 등 주요 답러닝 프레임워크 지원	AI 가치 사슬 조성	• 자사 기술을 적용한 레퍼런스 시스템을 파트너에게 제공하고 고객사와 협력해 다양한 AI 활용 사례 생성 중임. DGX는 NVIDIA GPU로 가속화된 AI 워크스테이션 및 서버 제품군임.
지속적인 GPU 아키텍처 혁신	• GPU 매 세대 별로 답러닝 성능을 대폭 높여 왔음. 뿐만 아니라 전력 소모량도 함께 줄여가고 있음.						
소프트웨어 Stack 강화	• CUDA(Compute Unified Device Architecture), cuDNN(CUDA Deep Neural Network) 등 GPU 가속 라이브러리 뿐만 아니라 텐서플로, 파이토치 등 주요 답러닝 프레임워크 지원						
AI 가치 사슬 조성	• 자사 기술을 적용한 레퍼런스 시스템을 파트너에게 제공하고 고객사와 협력해 다양한 AI 활용 사례 생성 중임. DGX는 NVIDIA GPU로 가속화된 AI 워크스테이션 및 서버 제품군임.						

엔비디아의 최신 GPU 아키텍처인 ‘블랙웰(Blackwell)’은 차세대 AI 컴퓨팅의 핵심으로 평가받고 있다. 블랙웰은 최대 10조 개의 파라미터로 확장되는 모델에 대한 AI 훈련과 실시간 대규모 언어 모델(LLM) 추론을 지원한다. 블랙웰 아키텍처는 이전 세대인 호퍼 아키텍처 또는 그 전인 암페어 아키텍처에 비해 트랜지스터 수가 크게 증가했고, 데이터 처리 속도와 효율성 또한 대폭 개선되었다.

또한 엔비디아는 NVLINK라는 기술을 통해 서버 클러스터 내의 모든 GPU 간 빠르고 안정적인 통신을 가능하게 하고 있다. 이를 통해 최대 576개의 GPU를 연결하여 하나의 거대한 컴퓨팅 네트워크를 구축할 수 있다. 이는 여러 대의 컴퓨터가 한 팀처럼 움직여 복잡한 문제를 해결하는 모습과 비슷하다. 이를 통해 엔비디아는 AI 모델의 학습과 추론에서 최적의 성능을 발휘할 수 있는 환경을 제공하고 있다.

[그림 29] 차세대 GPU 아키텍처 : 블랙웰⁴⁶⁾⁴⁷⁾

차세대 GPU 아키텍처 블랙웰	세부 내용						
	<table><tr><td>세계에서 가장 강력한 칩</td><td>• 블랙웰 아키텍처 GPU는 2,080억 개의 트랜지스터를 탑재함. GPU 다이(die) 간 초당 10TB의 칩 투 칩 링크로 연결된 단일 통합</td></tr><tr><td>2ND 트랜스포머 엔진</td><td>• LLM, MoE(Mixture of Experts) 모델에 대한 추론 및 학습을 가속화함. • 4비트 부동 소수점(FP4) AI를 구현. 이를 통해 메모리가 지원할 수 있는 차세대 모델의 성능과 크기를 두 배로 늘리면서도 높은 정확도를 유지</td></tr><tr><td>NVLINK</td><td>• 서버 클러스터 내의 모든 GPU 간 빠르고 원활한 통신이 필요하며 5세대 NVIDIA NVLINK는 최대 576개 GPU 확장 가능함.</td></tr></table>	세계에서 가장 강력한 칩	• 블랙웰 아키텍처 GPU는 2,080억 개의 트랜지스터를 탑재함. GPU 다이(die) 간 초당 10TB의 칩 투 칩 링크로 연결된 단일 통합	2ND 트랜스포머 엔진	• LLM, MoE(Mixture of Experts) 모델에 대한 추론 및 학습을 가속화함. • 4비트 부동 소수점(FP4) AI를 구현. 이를 통해 메모리가 지원할 수 있는 차세대 모델의 성능과 크기를 두 배로 늘리면서도 높은 정확도를 유지	NVLINK	• 서버 클러스터 내의 모든 GPU 간 빠르고 원활한 통신이 필요하며 5세대 NVIDIA NVLINK는 최대 576개 GPU 확장 가능함.
세계에서 가장 강력한 칩	• 블랙웰 아키텍처 GPU는 2,080억 개의 트랜지스터를 탑재함. GPU 다이(die) 간 초당 10TB의 칩 투 칩 링크로 연결된 단일 통합						
2ND 트랜스포머 엔진	• LLM, MoE(Mixture of Experts) 모델에 대한 추론 및 학습을 가속화함. • 4비트 부동 소수점(FP4) AI를 구현. 이를 통해 메모리가 지원할 수 있는 차세대 모델의 성능과 크기를 두 배로 늘리면서도 높은 정확도를 유지						
NVLINK	• 서버 클러스터 내의 모든 GPU 간 빠르고 원활한 통신이 필요하며 5세대 NVIDIA NVLINK는 최대 576개 GPU 확장 가능함.						

45) 2024 컴퓨텍스 기조연설로 본 엔비디아의 미래 비전과 전략, 「엔비디아, AI 시대를 이끄는 ‘게임 체인저’로 부상」(신동형)
46) 2024 컴퓨텍스 기조연설로 본 엔비디아의 미래 비전과 전략, 「엔비디아, AI 시대를 이끄는 ‘게임 체인저’로 부상」(신동형)
47) ‘NVIDIA Blackwell 아키텍처’(NVIDIA, 2024), [nvidia.com/ko-kr/data-center/technologies/blackwell-architecture/](https://www.nvidia.com/ko-kr/data-center/technologies/blackwell-architecture/)

엔비디아는 아키텍처 기반의 컴퓨팅 칩 자체를 넘어 데이터 센터 관점에서 에너지와 효율성을 통합적으로 고려한 솔루션도 제공하고 있다. 크게 2가지 관점으로 제공되고 있는데, 첫째 엔비디아의 GB200 NVL72이다. GB200 NVL72는 랙 스케일 설계로 36개의 Grace CPU와 72개의 블랙웰 GPU를 연결하여 고성능을 제공한다. 액체 냉각 기술을 통해 컴퓨팅 밀도를 높이고 사용되는 바닥 공간을 줄였으며, 이는 AI 데이터 센터의 효율성을 크게 향상시킨다. H100 공랭식 인프라와 비교했을 때 동일한 전력으로 25배 더 많은 성능을 제공한다.

둘째 GB200 Grace Blackwell Superchip이다. GB 브레이스 블랙웰 슈퍼칩은 엔비디아의 최신 기술로, NVLINK를 통해 2개의 블랙웰 텐서 코어 GPU와 Grace CPU를 연결하여 일반 CPU 대비 18배 많은 데이터 처리가 가능하다. 이는 대규모 데이터와 복잡한 연산을 처리하는 데 최적화된 설계로, AI 모델의 학습과 실시간 추론을 위해 높은 성능을 발휘한다. 이러한 기술은 마치 여러 개의 강력한 엔진이 동시에 작동하여 큰 배를 빠르게 전진시키는 것과 같은 원리로, AI 시스템 전반의 성능을 극대화한다.

[그림 30] 데이터 센터를 위한 엔비디아의 포트폴리오⁴⁸⁾



GB 200 NVL72	<ul style="list-style-type: none"> • 랙 스케일 설계로 36개 Grace CPU, 72개의 블랙웰 GPU를 연결함. • 액체 냉각으로 컴퓨팅 밀도를 높이고 사용되는 바닥 공간을 줄임 • H100 공랭식 인프라 비교시 GB200 동일한 전력으로 25배 더 많은 성능
GB 200 Grace Blackwell Superchip	<ul style="list-style-type: none"> • NVIDIA GB200 NVL72의 핵심 구성 요소로, NVIDIA NVLINK를 통해 2개의 블랙웰 텐서 코어 GPU와 NVIDIA Grace CPU를 2개의 블랙웰 GPU에 연결함. • CPU 대비 18배 많은 데이터 처리 가능함.

2.1.2. 인텔(INTEL)

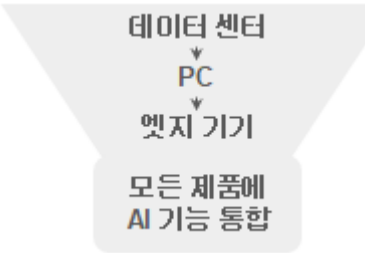
인텔은 AI 반도체 시장에서 CPU와 AI 가속기 기술을 결합하는 전략을 펼치고 있다. 인텔의 서버용 CPU는 지속적으로 성능을 향상시키고 있다. 예를 들어 최신 제온(Xeon) 프로세서는 엔비디아 H100 대비 학습 속도는 40% 더 빠르고 추론 성능도 1.5배 뛰어나면서 가격은 2.3배 더 좋은 강점이 있다고 인텔은 강조한다.

인텔은 AI Everywhere 전략을 통해 AI를 모든 컴퓨팅 기기에 필수 요소로 통합하고 있다. 이는 전기가 모든 가전 제품에 필수적인 요소가 된 것처럼, AI도 모든 컴퓨팅 기기의 필수 요소로 자리 잡도록 하겠다는 목표이다. 이에 인텔은 데이터 센터, PC, 엣지 기기까지 AI 기술을 확대 적용하고 있다. 예를 들어, 인텔의 루나 레이크(Lunar Lake) 칩은 AI PC를 주도하기 위해 출시된 제품으로, 기존의 CPU 코어와 함께 GPU 및 NPU를 추가하여 더 높은 연산 능력을 제공한다.

48) NVIDIA GB200 NVL72 (NVIDIA, 2024)

또한, 인텔은 엣지 컴퓨팅을 통해 데이터가 생성되는 곳 가까이에서 AI를 활용할 수 있도록 하고 있다. 예를 들어, 삼성 메디슨과 협력한 AI 기반 초음파 기기인 엣지 장비에서 태아 심장 영상 10개 단면을 실시간으로 캡처하며 AI 처리 속도를 20% 향상시켰다.

[그림 31] 인텔의 AI 반도체 전략⁴⁹⁾⁵⁰⁾

AI EVERYWHERE 전략	세부 내용
 <ul style="list-style-type: none"> 전기가 모든 가전 제품의 필수적인 요소가 된 것처럼, AI를 모든 컴퓨팅 기기의 필수 요소로 만들겠다는 INTEL 의지 	데이터 센터 <ul style="list-style-type: none"> CPU와 AI 가속기를 결합하여 AI 시장 공략 인텔의 서버용 CPU의 성능을 지속 향상시키고 있음.(Xeon6 등) 엔비디아보다 더 빠르고 더 저렴하게(가우디3은 엔비디아 H100 대비 학습속도 40% 더 빠르고 추론 성능도 1.5배 뛰어나면서 가격은 2.3배 더 좋다고 인텔은 강조하고 있음.
	PC <ul style="list-style-type: none"> AI PC를 주도하기 위해 루나 레이크(Lunar Lake) 출시 CPU 코어와 캐시만 수용했던 이전 버전인 Meteor Lake와 달리 CPU 코어와 캐시, GPU, NPU를 수용하고 있으며 모바일을 타겟
	엣지기기 <ul style="list-style-type: none"> 데이터가 생성되는 곳 가까이에서도 AI를 활용할 수 있도록 엣지 컴퓨팅 분야도 확대하려고 하고 있음. 예시로 삼성 메디슨과 협력한 AI 기반 초음파 기기임. 이를 통해 태아 심장 영상 10개 단면을 실시간 캡처하며 AI처리속도 20% ↑

2.1.3. AMD

AMD는 AI 반도체 시장에서 GPU와 NPU를 결합해 인텔과 엔비디아에 대항하고 있다. AMD는 인텔과 유사하게 PC, 데이터 센터, 엣지기기에 걸친 포트폴리오에 모두 진출하고 있는데, PC 시장에서는 AMD의 라이젠(Ryzen) AI 프로세서(Zen5 CPU와 XDNA 2 NPU 결합)로 강력한 성능을 제공하고 있다. XDNA 2 NPU는 AI 연산에 특화된 프로세서로, 정교하면서도 효율적인 연산이 가능하다. AMD는 이 프로세서를 탑재한 프리미엄 노트북을 HP, 레노버(Lenovo), ASUS와 같은 주요 제조사를 통해 출시하였다. 이는 마치 자동차에 최신 하이브리드 엔진을 탑재하여 더 높은 연비와 성능을 제공하는 것과 유사하다.

데이터센터 분야에서는 AMD의 CPU인 EPYC 시리즈가 성능 향상과 함께 에너지 효율을 추구하고 있다. AMD의 APU(Accelerated Processing Unit)인 Instinct MI300 시리즈는 GPU 가속기를 통해 대규모 AI 학습과 추론 작업을 진행하며, 높은 성능과 효율성을 동시에 제공한다. 이는 마치 더 적은 연료로 더 많은 거리를 주행할 수 있는 효율적인 엔진을 탑재한 대형 트럭과 같다.

49) AI(Claude3)가 작성한 인텔, AI 시대를 선도하는 기술 혁신과 비전 (신동형)

50) AI(Claude3)가 작성한 「Intel의 AI 시대 도전과 전략」보고서 (신동형)

엣지 기기에서 AMD의 버살AI(Versal AI)는 프로그래밍 가능한 로직, AI 엔진, 디지털 신호 처리 등을 통합하여 산업용 AI 애플리케이션에 적합한 제품이다. 예를 들어, 스마트 제조 공정에서 실시간으로 제품의 품질을 검사하고 결함을 자동으로 탐지하는 데 활용될 수 있다.

그리고 AMD는 개방형 표준을 추구하며 엔비디아 GPU가 AI 산업에서 둘러싸 놓은 철옹성을 깨뜨리려 노력하고 있다. UALink는 AMD의 인피니티 패브릭 기술을 기반으로 한 것으로, 엔비디아의 NVLINK에 대응하며, GPU 간의 빠르고 안정적인 통신을 가능하게 한다. ROCm(Radeon Open Compute)은 AMD의 GPU 컴퓨팅을 위한 개방형 소프트웨어 플랫폼으로, 이는 엔비디아의 CUDA와 경쟁하는 소프트웨어 스택이다. 이러한 개방형 접근은 마치 모든 사람들이 동일한 언어를 사용해 소통할 수 있게 하는 국제 공용어와 같다. 이를 통해 더 많은 개발자와 기업이 AMD의 기술을 쉽게 활용할 수 있도록 지원하고 있다.

[그림 32] AMD의 AI 반도체 전략⁵¹⁾



2.1.4. ARM

ARM은 AI 반도체 시장에서 CPU, GPU, 그리고 NPU를 통합한 플랫폼을 통해 경쟁력을 강화하고 있다. ARM의 컴퓨팅 플랫폼은 다양한 환경에서 활용될 수 있도록 설계되었으며, CPU, GPU, NPU를 하나의 네트워크로 연결해 효율적인 데이터 처리가 가능하다고 한다. 예를 들어, ARM의 Mali GPU는 스마트폰과 같은 모바일 기기에서 높은 그래픽 성능을 제공하며, AI 연산에 필요한 NPU인 Ethos 시리즈는 스마트폰과 같은 소형 기기에서도 복잡한 AI 연산을 가능하게 하여, 다양한 환경에서 적절한 성능을 발휘할 수 있게 해준다.

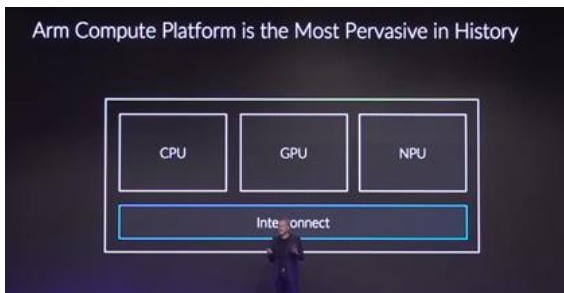
ARM은 클라이드AI(Kleide AI)라는 컴퓨팅 라이브러리를 제공하여 개발자들이 ARM의 AI 기능을 쉽게 활용할 수 있도록 돕고 있다. 클라이드AI는 텐서플로, 파이토치, 라마와 같은 주요 AI 프레임워크 및 모델에 최적화된 라이브러리이다. 또 그 중 ACL(Arm Compute Library)는 클라이드AI 중 핵심 모듈로 ARM의

51) AMD, AI 시대 컴퓨팅 혁신으로 지능화 가속화 (신동형)

칩에서 AI 가속 기능에 접근할 수 있는 컴퓨팅 라이브러리이다. 이를 통해 개발자들이 ARM의 기술을 손쉽게 활용하여 AI 애플리케이션을 개발할 수 있다.

ARM의 전략은 엣지 기기부터 데이터센터까지 모든 환경에서 활용 가능한 플랫폼을 구축하는 것이다. 이를 통해 ARM은 컴퓨팅 자원을 효과적으로 활용하고, 다양한 기기에서의 AI 응용을 가능하게 하여 AI 산업을 확대시키고 있다. ARM의 이러한 플랫폼은 에너지 효율이 높고, 적은 전력으로도 높은 성능을 발휘할 수 있어 다양한 분야에서 AI 기술의 적용을 촉진하고 있다.

[그림 33] ARM의 IP와 라이브러리⁵²⁾



- CPU : 스마트폰용AP, 노트북용 윈도우온ARM, 데이터 센터용 서버칩 등
- GPU : 스마트폰용 Mali GPU 등
- NPU : AI 연산에 특화된 하드웨어로 Arm의 Ethos



- Arm은 Cy AI와 Kleide AI라는 컴퓨팅 라이브러리 제공
- Cy(Compute Library) AI는 Arm 칩의 AI 가속 기능 접근 가능함
- Kleide AI는 TensorFlow, PyTorch, Llama 등 주요 AI 프레임워크 및 모델과 호환되는 라이브러리임.

2.1.5. AI 메모리 반도체 : SK 하이닉스

SK 하이닉스는 AI 반도체 시장에서 특히 HBM(High Bandwidth Memory) 분야에서 강력한 경쟁력을 보유하고 있다. HBM은 데이터 처리 속도를 획기적으로 높일 수 있는 고대역폭 메모리로, AI 연산과 같은 대규모 데이터 처리 작업에 최적화되어 있다. 이는 더 빠른 물 흐름을 위해 넓은 관을 사용하는 것과 같다. SK 하이닉스는 2023년 HBM 시장에서 54%의 점유율로 1위를 차지하며, 글로벌 AI 인프라의 핵심 공급자로 자리매김하고 있다. 삼성전자와 마이크론이 뒤를 잇고 있지만, SK 하이닉스의 HBM 기술력은 지속적으로 시장을 선도하고 있다.

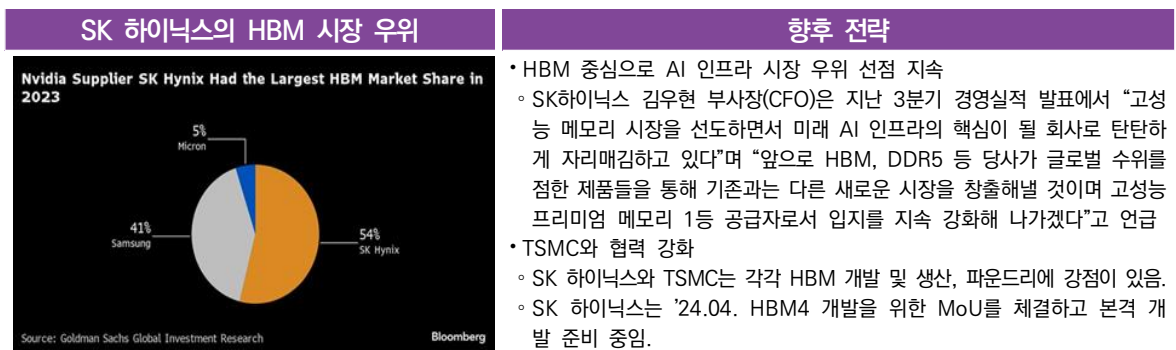
SK 하이닉스는 고성능 메모리 시장에서 선도적 위치를 유지하기 위해 다양한 전략을 추진하고 있다. SK 하이닉스의 김우현 부사장은 “고성능 메모리 시장을 선도하면서 미래 AI 인프라의 핵심이 될 회사로 자리매김하고 있다”고 언급했다. 이를 위해 SK 하이닉스는 HBM뿐만 아니라 DDR5와 같은 차세대 메모리 제품군을

52) Arm, AI 컴퓨팅의 미래를 향한 비상(飛上) (신동형)

통해 새로운 시장 창출에 주력하고 있다. 예를 들어, DDR5는 기존 DDR4보다 데이터 전송 속도가 두 배 이상 빨라, AI 및 클라우드 컴퓨팅과 같은 고성능 컴퓨팅 환경에 최적화되어 있다.

또한, SK 하이닉스는 TSMC와의 협력을 강화하여 HBM 개발 및 생산 파트너십을 통해 경쟁력을 더욱 높이고 있다. SK 하이닉스는 2024년, HBM4 개발을 위한 양해각서(MoU)를 체결하고 본격적인 협력 체계를 준비하고 있다. 이러한 협력은 마치 서로 다른 전문 기술을 가진 두 회사가 함께 새로운 제품을 만들어 시장에서 경쟁력을 확보하는 것과 같다. SK 하이닉스의 이러한 전략은 AI 반도체 시장에서 지속적인 성장을 위한 중요한 발판이 될 것으로 보인다.

[그림 34] SK 하이닉스의 HBM 사업 중심 전략⁵³⁾⁵⁴⁾



2.1.6. AI 메모리 반도체 : 삼성전자

삼성전자는 AI 반도체 시장에서 GAA(Gate-All-Around) 공정과 첨단 패키징 기술을 통해 혁신을 통해 시장 주도권을 가져가고자 한다. GAA 공정은 반도체 소자의 성능과 전력 효율을 동시에 개선하는 기술로, 3nm GAA 공정은 2022년부터 양산이 시작되었으며, 2nm 공정은 2027년 양산을 목표로 하고 있다.

삼성전자는 차세대 패키징 기술인 2.5D 인터포저 및 3D 적층 기술을 활용하여 AMD 등 AI 가속기를 개발하는 기업들에게 최적화된 솔루션을 제공하고 있다. 이러한 패키징 기술은 메모리, 로직 칩, 광학 소자 등을 하나의 패키지에 집적해 성능과 효율성을 극대화한다. 이는 여러 부품을 하나의 통합된 기기로 만들어 공간을 절약하고 성능을 높이는 것과 같다.

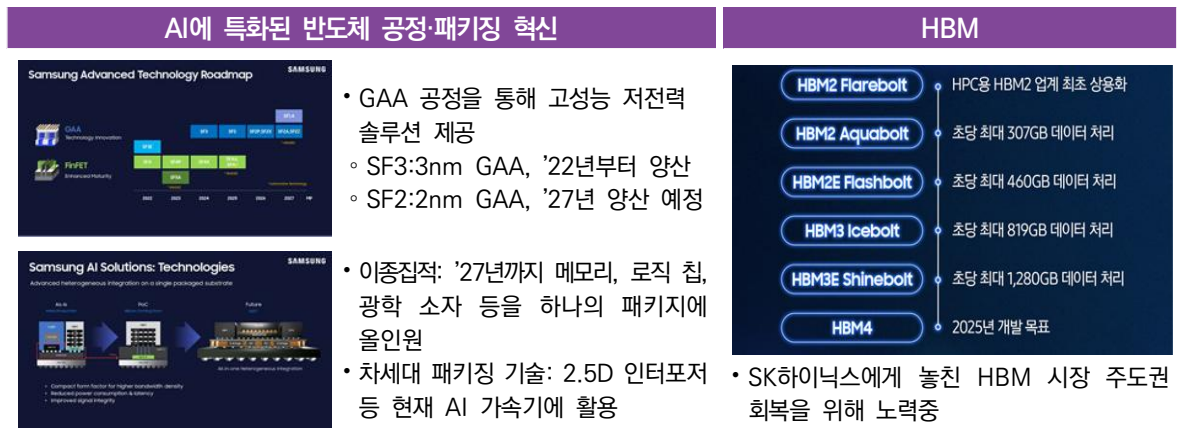
HBM 분야에서도 삼성전자는 HBM2부터 HBM3 그리고 HBM4 개발까지 다양한 제품을 통해 AI 메모리 시장에서 SK 하이닉스와의 경쟁을 이어가고 있다. 예를 들어, HBM3는 초당 최대 1,280GB의 데이터를 처리할 수 있으며, 이는 기존 메모리 대비 훨씬 더 빠른 속도로 데이터를 처리할 수 있는 능력을 제공한다.

53) Korea's SK Hynix investing over \$1 Billion to improve its high-bandwidth memory capacity (DIGITIMES Asia, 2024)

54) “우리에게 ‘이것’이 있다”...삼성전자의 차세대 ‘비밀 병기’ [황정수의 반도체 이슈 짚어보기] (황정수, 2024)

삼성전자는 AI에 특화된 반도체 공정과 패키징 기술을 지속적으로 혁신하고 있으며, 이를 통해 고성능 및 에너지 효율을 동시에 만족하는 반도체 솔루션을 제공하고 있다. 또한, 삼성전자는 AI 솔루션 통합 패키지로 스마트폰, 노트북, 데이터센터 서버 등 다양한 제품에 적용할 수 있는 기술을 개발하고 있다. 이러한 기술은 모든 용도에 맞는 다목적 도구를 개발하는 것과 같아, AI 산업 전반에서 활용 가능성을 높이고 있다.

[그림 35] 삼성전자의 AI 특화 반도체 전략⁽⁵⁵⁾⁽⁵⁶⁾

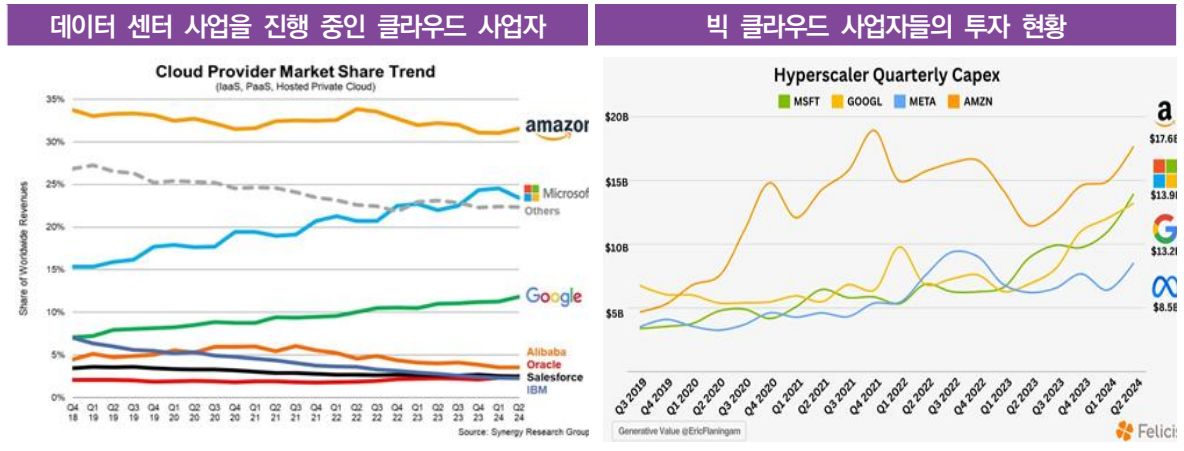


2.2. AI 데이터 센터

AI 데이터 센터는 대규모 AI 모델을 학습하고 추론하는 데 필수적인 고성능 컴퓨팅 자원과 인프라를 제공하는 시설이다. 특히 클라우드 사업자들이 AI 데이터 센터 인프라를 주도하며, 전 세계 곳곳에서 데이터 센터를 구축하고 운영하고 있다. 대표적인 클라우드 서비스 제공자로는 아마존 AWS, MS 애저, 구글 클라우드(GCP), 그리고 알리바바 클라우드 등이 있다. 빅 클라우드 사업자들은 AI 데이터 센터를 위한 대규모 투자를 지속하고 있으며, 특히 GPU 구매와 같은 고성능 하드웨어 인프라에 대한 자본 지출(CAPEX)이 크게 증가하고 있다. 아마존, MS, 구글 등은 매 분기 수십억 달러를 투자하여 AI 시대의 요구를 충족시키기 위해 데이터 센터를 확장하고 있다. 예를 들어, 2024년 2분기 기준 아마존은 176억 달러, MS는 139억 달러, 구글은 132억 달러를 데이터 센터 인프라에 투자하였다. 이러한 대규모 투자는 AI 모델의 학습과 추론에 필요한 고성능 컴퓨팅 자원을 안정적으로 제공하기 위한 필수적인 조치이다.

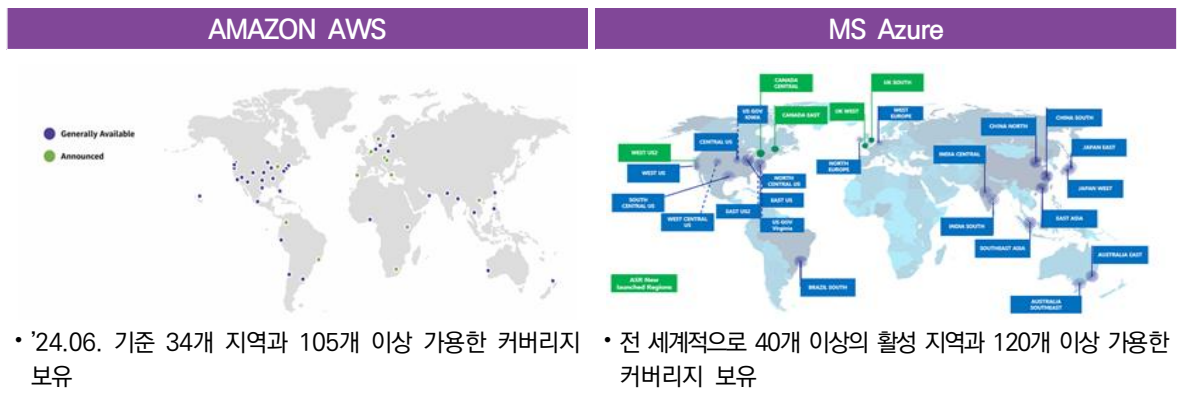
55) 삼성전자, 파운드리 포럼 2024 개최 AI 시대 파운드리 비전 제시 (삼성전자, 2024)

56) [인터뷰] AI 시대 삼성전자 HBM이 만들어 내는 완벽한 하모니 (삼성전자, 2024)

[그림 36] 글로벌 AI 데이터 센터 경쟁 우위를 위한 빅 클라우드 사업자들의 현황⁵⁷⁾⁵⁸⁾

아마존 AWS는 전 세계적으로 가장 큰 클라우드 인프라를 구축하고 있다. 현재 기준 34개 지역과 105개 이상의 가용한 커버리지를 보유하고 있으며, 이는 전 세계에 흩어진 수많은 전초기지처럼, 사용자가 어디에 있든지 필요한 컴퓨팅 자원을 제공할 수 있는 체계를 갖추고 있다는 것을 의미한다.

MS Azure는 전 세계적으로 40개 이상의 활성 지역과 120개 이상의 가용한 커버리지를 보유하고 있다. 이를 통해 전 세계 어디서나 빠르게 서비스를 제공할 수 있는 인프라를 갖추고 있으며, 이는 각 나라에 지사를 둔 대형 기업이 고객들에게 신속한 서비스를 제공하는 것과 같은 원리이다.

[그림 37] TOP2 플레이어의 글로벌 확장⁵⁹⁾

57) The Current State of AI Markets (FlaninamEric, 2024)

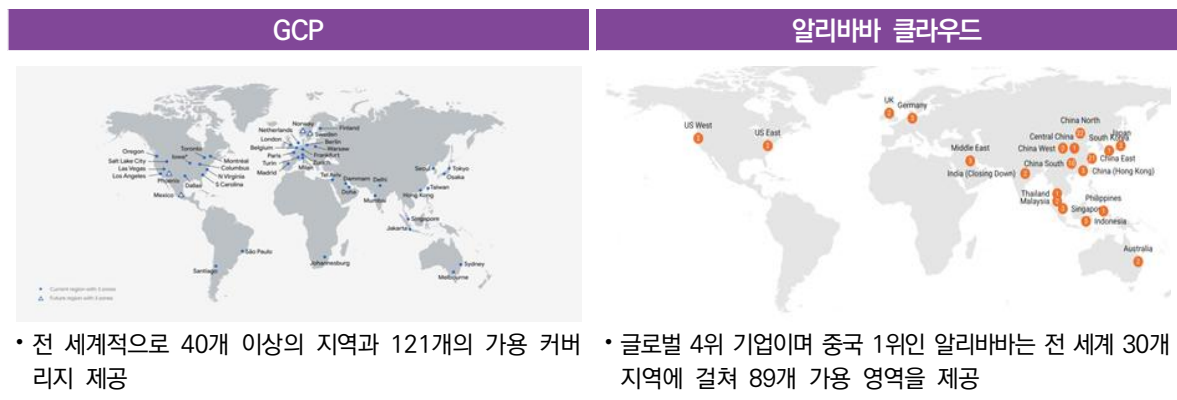
58) Cloud market share 2024 - AWS, Azure, GCP growth fueled by AI (Alexandre, 2024)

59) Top Cloud Providers in 2024 - Hyperscalers and Alternative vendors (Alexandre, 2024)

구글 클라우드(GCP) 역시 40개 이상의 지역과 121개의 가용 커버리지를 제공하며, 특히 데이터 분석 및 AI 관련 서비스를 강점으로 삼고 있다. 이는 도서관이 책뿐만 아니라 그 책의 내용을 쉽게 찾아볼 수 있는 다양한 도구들을 제공하는 것과 비슷하다. 구글은 AI와 빅데이터 처리에 최적화된 인프라를 통해 데이터 분석과 AI 모델 학습에서 경쟁력을 갖추고 있다.

알리바바 클라우드는 글로벌 4위의 클라우드 서비스 제공자로, 주로 중국 및 아시아 지역에 강점을 보유하고 있다. 30개 지역에 걸쳐 89개의 가용 영역을 제공하며, 특히 중국 시장에서의 우위를 바탕으로 빠르게 성장하고 있다. 이는 특정 지역에서 강한 영향력을 가진 기업이 그 지역의 수요를 효율적으로 충족시키는 것과 같다.

[그림 38] TOP 3~4 플레이어의 글로벌 확장⁶⁰⁾



글로벌 클라우드 사업자들이 AI 데이터 센터에 엄청난 투자를 감행하는 주된 이유는 AI 모델 학습과 실시간 추론을 위한 고성능 컴퓨팅 환경을 조성하기 위함이다. 그리고 특히 아마존, MS, 구글이 더 경쟁력 있는 AI 컴퓨팅 환경 확보를 통해 시장 리더십을 확보하기 위한 치열한 경쟁을 하고 있기 때문이기도 하다. 이로 인해 클라우드 사업자들은 데이터 센터를 점차 글로벌화하며 사용자의 요구를 신속히 충족시키기 위한 노력을 기울이고 있다. 이는 세계 여러 지역에 연결된 고속도로망을 구축해 어디서나 빠르고 효율적으로 이동할 수 있게 하는 것과 같은 원리로, AI 서비스의 접근성과 효율성을 극대화하는 것이다.

결론적으로, AI 데이터 센터는 AI의 발전에 있어 중요한 역할을 담당하고 있으며, 글로벌 클라우드 사업자들이 이를 주도하고 있다. 이들은 각자 자사의 강점을 기반으로 인프라를 구축하고 있으며, 지속적인 투자와 기술 개발을 통해 AI 시대에 최적화된 컴퓨팅 자원을 제공하고 있다.

60) Top Cloud Providers in 2024 – Hyperscalers and Alternative vendors (Alexandre, 2024)

3. AI 모델 개발

AI 모델 개발은 데이터 수집부터 AI 파운데이션(Foundation) 모델의 설계, 훈련, 배포에 이르는 과정으로, AI 가치를 창출하는 핵심 단계 중 하나다. 이 분야에서 활약 중인 대표적인 기업으로는 오픈AI(OpenAI), 미스트랄AI(Mistral AI), 앤스로픽(Anthropic), 코히어(Cohere), 스테빌리티AI(Stability AI) 등이 있다. 또한, 오픈소스 AI 모델을 활용해 혁신적이고 맞춤형 모델을 개발하는 기업들도 두각을 나타내고 있다. 이들 기업들은 각자의 독창적인 접근법으로 AI 모델 개발을 주도하며, AI 생태계를 확장하고 있다.

예를 들어 오픈AI는 GPT-4와 같은 대규모 언어 모델을 개발하여 자연어 처리와 생성 분야에서 큰 영향을 미치고 있다. 이 모델은 사람이 대화를 나누듯이 정보를 제공하거나 질문에 답변하는 등 다양한 작업을 수행할 수 있다. 이러한 대규모 언어 모델은 다양한 산업에서 활용될 수 있으며, 예를 들어 고객 지원 자동화, 콘텐츠 생성, 교육 등에서 큰 가치를 제공하고 있다. 앤스로픽은 AI의 실용성과 안전성 및 윤리적 사용을 중시한 방향으로 AI를 개발하는 데 중점을 두고 있다. 즉 이들은 사람들에게 해를 끼치지 않도록 안전하게 그리고 더 실질적으로 활용되는 AI 기술을 개발하는 것이 목표이다.

다양한 AI 모델 개발사들이 경쟁하며 새로운 모델을 개발하는 이유 중 하나는 모든 상황에 완벽히 대응할 수 있는 AI 모델이 없기 때문이다. 각 모델은 특정 용도나 환경에 최적화되어 있어, 상황에 따라 적합한 모델을 선택해 사용하는 것이 중요하다. 예를 들어, 코히어는 언어 이해와 생성에 특화된 모델을 개발하고 있으며, 스테빌리티AI는 이미지 생성과 같은 멀티미디어 콘텐츠 생성에 강점을 가지고 있다. 이러한 모델들은 각각의 강점을 바탕으로 특정한 문제를 해결하는 데 최적화되어 있다.

또한, 최근에는 규제와 같은 외부 환경에 따라 다양한 AI 모델들이 등장하고 있다. 이러한 환경에 미스트랄 AI가 혜택을 받고 있다. 예를 들어 EU의 AI ACT와 같은 규제 환경에 대응하기 위해 윤리적이고 투명한 AI 모델 개발에 대한 요구가 증가하고 있다. 이는 교통법규가 자동차의 설계와 운행 방식을 규제하는 것처럼, AI 기술도 사회적 책임과 규제를 준수하면서 발전해야 한다는 점을 보여준다.

[그림 39] AI 모델 전문 사업자 현황



3.1. 오픈AI(OpenAI)

오픈AI는 AI 모델 개발 분야에서 가장 주목받는 기업 중 하나로, GPT 시리즈와 같은 대규모 언어 모델을 통해 AI 기술의 혁신을 주도하고 있다. 최근 오픈AI는 비영리법인에서 영리 법인으로 전환하면서 투자자들로 부터 더 많은 자금을 유치할 수 있게 되었다고 한다. 이 전환은 연구 개발에 필요한 자금을 확보하고 더 많은 자원을 투입할 수 있는 발판을 마련한 것이다. 향후 오픈AI는 영리 법인으로의 전환을 통해 기술적 관점에서 AI 산업의 주도권을 더 강화해 나갈 것으로 예상된다.

[그림 40] 오픈AI 확장을 위한 영리 법인화⁶¹⁾⁶²⁾⁶³⁾

비영리에서 영리 법인으로 전환	추가 편익
<ul style="list-style-type: none"> • 오픈AI가 비영리법인이 회사 주요 사항을 결정하는 현재의 지배구조를 벗어나 영리 법인이 관할하는 형태로 전환을 추진 중 이라고 로이터 통신이 익명의 소식통을 통해 '24.09.25. 보도함. • 비영리법인이 자회사 영리법인의 모든 주요 사업을 통제하며 영리법인의 투자자에 대한 이익 배분에 상한선이 설정되어 있음. • 오픈AI가 핵심 사업을 비영리법인 이사회가 통제하지 않는 영리 공익법인으로 재편하면, 이익 상당 부분을 주주들에게 돌려 줄 수 있어 투자자들에게 좋은 환경 제공함. • '15년 비영리 AI 연구 단체로 설립한 오픈AI는 '19. 영리 법인인 오픈AI LP를 자회사로 설립했음. 이 자회사를 통해 MS로 부터 자본 투자를 받았음. • 현재 MS가 영리법인 지분 49%를 보유하고 있음. 	<ul style="list-style-type: none"> • '24.10. 오픈AI는 1,570억\$ 기업 가치로 투자 라운딩을 마감함. 이는 1년도 안되어 약 2배의 기업 가치가 상승한 오픈AI의 저력을 알려 주는 것임. • 이번 라운딩에서는 약 66억\$의 투자가 진행되었 으며, MS는 지금까지 130억\$ 투자에 추가로 7.5 억\$를 투자한 것으로 알려져 있음. • 뿐만 아니라 엔비디아와 소프트뱅크(5억\$)도 투자한 것으로 알려짐. • 오픈AI가 높은 운영 비용으로 자금 소진에 다다랐다는 뉴스에도 불구하고 오픈AI의 시장성과 AI 시장 확대에 대한 가능성을 본 것이라 보임.

그러나 이러한 변화는 AI의 윤리적 측면보다는 이윤 추구에 초점을 맞춘 방향으로 전환되고 있다는 점에서 주목할 필요가 있다. 오픈AI는 본래 인류에 유익한 AI 개발을 목표로 비영리법인으로 출발했지만, 영리 법인 으로의 전환 이후에는 회사 이익과 투자자들에게 더 큰 이익을 제공하기 위한 사업 전략을 펼칠 것으로 예상 된다. 이는 AI 기술의 발전에 있어 긍정적인 측면이 있는 동시에, 윤리적 문제와 상업적 이해 사이의 균형을 어떻게 유지할 것인지에 대한 도전 과제를 내포하고 있다.

오픈AI의 대표적인 모델로는 멀티 모달 GPT-4o, 텍스트 기반의 비디오 생성 모델 Sora, 그리고 추론 중심의 o1이 있다. GPT-4o는 텍스트, 이미지, 음성 등 다양한 형태의 데이터를 이해하고 생성할 수 있는 멀티 모달 모델로, 그리고 기존 GPT-4 대비 속도를 두 배 향상시키고 비용을 50% 절감하는 등 AI의 효율성을 높이는 데 중점을 둔 모델이다.

61) 오픈AI, 영리법인 관할 형태로 전환 추진 (현대인, 2024)

62) OpenAI scoops up \$6.6B in funding round at \$157B valuation (RobuckMike, 2024)

63) OpenAI Raises \$6.6 Billion in Funds at \$157 Billion Value (Shirin GhaffaryKatie, 2004)


추론 중심의 'o1' 프로젝트는 학습 중심의 기존 접근 방식을 변화시켜 문제 해결 능력을 향상시켰다. o1은 특히 STEM(과학, 기술, 공학, 수학) 분야에서의 복잡한 문제 해결에 강점을 가지고 있으며, AI의 적용 범위를 확장하고 있다.

소라(Sora)는 오픈AI가 새롭게 선보인 텍스트 기반의 비디오 생성 모델로, 영상 콘텐츠 AI 시대를 열고 있다. 소라는 텍스트 입력만으로도 고품질의 비디오 콘텐츠를 생성할 수 있다. 이는 마케팅, 교육, 엔터테인먼트 등 다양한 분야에서 활용될 수 있다. 아직은 제한된 기업들만이 사용 가능하다.

2024년 10월 기준 오픈AI는 1,570억 달러의 기업 가치로 투자 라운딩을 마감하였으며, 이는 '24년 상반기에 800억\$ 기업가치를 인정 받은지 1년도 안 돼 거의 두 배에 달하는 등 가파른 기업 가치 상승을 보여준다. 이번 라운딩에서는 약 66억 달러의 투자가 진행되었으며, MS가 그중 7.5억 달러의 추가 투자를 통해 오픈 AI의 성장을 지원하고 있다. 엔비디아와 소프트뱅크도 각각 5억 달러를 투자하며 AI 시장 확대에 대한 기대감을 보여주고 있다.

이와 같은 자금 조달과 투자 유치는 오픈AI가 AI 기술의 발전뿐만 아니라 상업적 성공을 위한 발판을 마련하고 있음을 보여준다. 높은 운영 비용에도 불구하고 지속적인 투자를 통해 오픈AI는 AI 시장에서의 리더십을 공고히 할 것으로 기대되고 있다.

[그림 41] 오픈AI의 포트폴리오⁶⁴⁾⁶⁵⁾

GPT-4o	o1	Sora
<ul style="list-style-type: none"> GPT-4o의 핵심 특징 <ul style="list-style-type: none"> 멀티 모달(텍스트, 이미지, 음성)의 통합처리 50개 이상 언어 지원 및 토큰 압축률 향상 효율적 구현(GPT-4 대비 속도 2배, 비용 50% 감소) GPT-4o mini라는 비용 효율 모델도 출시 <ul style="list-style-type: none"> 낮은 비용과 지연 시간으로 AI 활용성 확대 GPT-3.5 Turbo 보다 60% 이상 저렴하게 활용 가능함. GPT-4o와 동일한 범위의 언어 처리 가능 	<ul style="list-style-type: none"> 접근 변화: 모델 크기 키우기 → 추론·해결 <ul style="list-style-type: none"> 책을 많이 읽으면 더 똑똑해 질 것이라는 생각으로 더 많은 정보를 암기해 왔음. 이제 주어진 문제를 차근차근 풀어가는 능력 즉, 사고가 가능한 시스템으로 전환 o1의 핵심 기술 <ul style="list-style-type: none"> 강화학습을 통한 추론 능력 향상 단계별 문제 해결 방식 o1은 특히 STEM 분야의 복잡한 문제, 즉 수학과 코딩 분야에 강점 o1 mini라는 비용 효율적인 모델도 함께 출시 	<ul style="list-style-type: none"> Text to Video 모델로 영상 콘텐츠 AI 시대를 열. 현재 제휴된 기업들만 사용 가능 

64) AI(Claude3)가 작성한, OpenAI의 GPT-4o 공개, 멀티 모달 AI 혁명의 신호탄 (신동형)

65) OpenAI o1:AI의 새로운 패러다임, 추론 중심 접근의 혁명 (신동형)

3.2. 앤스로픽(Anthropic)

앤스로픽은 AI의 안전성과 윤리적인 개발을 중심에 두고 있는 AI 연구 기업으로, AI가 인류에 긍정적인 영향을 미칠 수 있도록 연구를 이어가고 있다. 기존 AI 벤치마크가 실제 상황에서의 활용에 한계를 보였던 것과는 달리, 앤스로픽은 보다 현실적인 문제 해결을 목표로 하고 있다. 앤스로픽은 AI 평가를 단순한 암기나 패턴 인식이 아니라 복잡한 상황에서의 이해와 추론 능력을 평가하는 방향으로 발전시키고 있다. 이는 학교에서 학생들이 단순히 암기한 내용을 시험 보는 것이 아니라, 실제 생활에서 문제를 해결할 수 있는 능력을 평가받는 것과 같다.

앤스로픽의 대표적인 AI 모델로는 클로드(Claude) 시리즈가 있다. 클로드 3.0에서는 오퍼스(Opus), 소넷(Sonnet), 하이쿠(Haiku)와 같은 하부 모델을 통해 다양한 문화적, 언어적 요구에 대응하고 있다. 초기에 오퍼스는 복잡한 보고서와 분석 작업에 특화되어 있으며, 소넷은 예술 및 문학 작업에 특화된 모델로 유럽 정형시와 같은 문학 형식을 이해하고 생성할 수 있다. 하이쿠는 간결하면서도 깊이 있는 표현을 가능하게 한다. 최근에는 클로드3.5 소넷을 출시하면서 클로드3.0의 오퍼스 에 상응하는 성능과 함께 속도와 비용 효율성이 더 뛰어난 모델로 고도화했다.

앤스로픽은 실용적이지 않은 기존의 AI 평가를 바꾸려고 하고 있다. 기계적인 평가를 넘어서 전문가와 실제 사용자들이 직접 참여하는 평가 방식을 강화하고 있다. 이는 AI 모델이 실제로 사람들에게 얼마나 유용하게 활용될 수 있는지를 더 정확히 평가할 수 있게 한다. 예를 들어, MMLU(추론, 역사 등 전문분야) 테스트에서 단순히 책에서 배운 내용을 확인하는 것을 넘어 실전에서 그 지식을 활용하고 응용하는 능력에 중점을 두고 평가가 바뀌어야 한다고 한다. 이에 앤스로픽은 실생활에서의 응용 가능성을 더욱 높이기 위해 노력하고 있다.

[그림 42] 앤스로픽 AI 모델 현황⁶⁶⁾

실생활로 스며들려는 앤스로픽

기존 AI 벤치마크의 한계와 실제 활용과의 괴리 발생

AI 벤치마크

GPQA
(박사급 과학 문제)

MMLU
(수학, 역사 등 전문분야)

고어체 문장 판별

실제 일상

일상 대화, 업무 보조
(이메일 작성, 자기 소개서 첨삭 등)

MMLU는 대다수 사용자에게는 불필요한 기능

고어체는 일상적 대화에는 거의 등장하지 않음.

앤스로픽은 보다 효과적이고 신뢰할만한 AI 평가 노력중

평가의 난이도와 복잡성 향상(단순 암기나 패턴 인식으로 불가)

기계적 평가를 넘어 전문가와 실사용자의 직접 참여 확대 강조

Claude 서비스

3.0

Opus

(유명 작곡가 번호 매겨진) 작품

복잡한 보고서 및 분석 작업 탁월

Sonnet

(유럽 정형시 중 하나) 작은 노래

예술 및 문학 작업에 특화

Haiku

일본의 전통 단시

단답형 응답에 특화

3.5

범용 고성능 모델로

속도·비용 효율성 특화

66) AI 평가 체계 대전환을 향한 앤트로픽의 도전:한계 극복과 신뢰 확보의 과제 (신동형)

3.3. 미스트랄AI(Mistral AI)

2023년 4월 설립된 미스트랄AI는 설립 초기부터 ‘유럽 AI’의 강화를 목표로 하며, 유럽 내에서의 경쟁력을 높여 입지를 굳히고자 노력하고 있다. CEO인 딘스는 유럽 언어에 고집해 유럽 시장, 즉 상대적으로 미국 기업들의 입지가 부족한 상황을 역으로 활용하고 있다. 이는 유럽에서 재배된 지역 특산물을 강조해 그 지역 소비자에게 더 큰 가치를 제공하는 전략과 유사하다. 이러한 목표와 함께 미스트랄 AI는 구글과 메타 출신 창업자들이 설립했다는 측면에서 그 역량 가능성을 인정받아 설립 당시 이미 1억 5백만 달러의 자금을 유치해 시장에 큰 관심을 받았다.

최근 EU AI Act 등 규제를 감안해 MS 등으로부터도 투자를 유치하고 있다. 오픈AI를 통해 AI 시장을 좌지우지하고 있는 MS 입장에서는 미스트랄을 놓치는 것은 유럽을 놓치는 것과 같다. 미스트랄은 유럽의 법적, 윤리적 기준을 준수하는 AI 모델 개발에 집중하고 있다. 여기에 소버린 AI를 주창하는 네이버도 앵커투자자로 출자한 코렐리아캐피탈의 유럽 특화 K-펀드2를 통해 미스트랄AI에 투자했다.

미스트랄은 주로 유럽 지향인 고성능 AI 모델 개발에 집중하고 있으며, 특히 다양한 분야에서 활용될 수 있는 범용 AI 모델 개발을 목표로 하고 있다. 구체적으로 미스트랄의 주요 제품 포트폴리오로는 믹스트랄(Mixtral) 8×22B, 미스트랄 라니2(Mistral Large2), 그리고 픽스트랄(Pixtral) 12B가 있다. 믹스트랄 8×22B는 MoE (Mixture of Experts) 구조를 활용해 141억 개의 파라미터 중 39억 개를 활성화시키는 효율적인 모델이다. 이 모델은 다국어 처리와 수학적 코딩 작업에서 높은 성능을 발휘하며, 다양한 문제를 효율적으로 해결할 수 있다. 마치 여러 전문가들이 각자의 전문성을 바탕으로 문제를 나누어 해결하는 것처럼, 믹스트랄 8×22B는 필요한 경우에만 전문가를 활성화해 성능을 최적화한다.

미스트랄 라지2는 1230억 개의 매개변수를 통해 긴 맥락을 이해하고 복잡한 추론을 수행할 수 있는 모델이다. 이는 주로 전문적 업무에 사용되며, 긴 텍스트나 복잡한 데이터를 처리하는 데 강점을 보인다. 예를 들어, 법률 문서의 분석이나 금융 데이터의 추론과 같은 작업에서 미스트랄 라지2는 여러 장의 복잡한 책을 동시에 읽고 그 내용을 이해하는 것처럼 높은 수준의 분석을 가능하게 한다.

픽스트랄(Pixtral) 12B는 멀티모달 모델로, 이미지와 텍스트를 동시에 이해하고 설명할 수 있다. 이는 시각적 AI 작업에 사용되며, 예를 들어 이미지를 보고 그 내용을 설명하거나 이미지와 관련된 질문에 답할 수 있다.

[그림 43] 미스트랄AI 기업 현황 및 제품 포트폴리오⁶⁷⁾⁶⁸⁾

미스트랄 AI		제품 포트폴리오			
<ul style="list-style-type: none">• 미스트랄AI는 '23.04. Google 및 Meta 출신 창업자들이 설립했으며, 설립시 이미 1.05억\$의 자금 유치◦ 구글 딥마인드 출신인 아르튀르 멘슈, 메타 FAIR에서 라마 모델을 개발했던 티모테 라크루아, 기욤 램플이 설립했음.◦ '23년 말에는 20억\$의 기업 가치를 인정받음.◦ 네이버가 앵커투자자로 출자한 코렐리아캐피탈의 유럽 특화 K-펀드2에서 미스트랄AI에 투자 진행• 미스트랄AI는 시작부터 '유럽의 AI혁명'을 추구함.◦ CEO인 멘슈는 유럽 언어에 고집하고 있으며, 현재 미국 AI 기업들이 강력한 편집권을 가졌기 때문에, 유럽 고객들이 더 큰 통제권을 갖도록 힘을 강조함.◦ EU AI Act 등 규제를 감안해 MS 등도 투자		모델명	주요 특징	강점	주요 용도
		Mixtral 8×22B	MoE 구조, 141B 중 39B 활성화	효율성, 다국어, 수학/코딩	범용AI
		Mistral Large2	123B 매개변수, 긴 문맥 이해	복잡한 추론, 전문 분야	전문적 업무
		Pixtral 12B	멀티모달 (이미지&텍스트)	이미지 이해 및 설명	시각적 AI 작업

3.4. 코히어(Cohere)

코히어는 2019년에 캐나다 토론토에서 설립된 AI 기업으로, 데이터 보안 및 기업 맞춤형 AI 솔루션을 개발하는 데 중점을 두고 있다. 공동 창립자인 에이단 고메즈(Aidan Gomez)는 생성형 AI의 근간이 되는 Transformer 모델을 다룬 “Attention is All You Need” 논문의 공동 저자이기도 하다. 코히어는 이를 바탕으로 언어 이해와 생성에 특화된 기술을 개발하고 있다.

코히어는 기업용 AI라는 명확한 목표를 가지고 지속적으로 투자를 유치했다. 2023년 6월에는 2억 7천만 달러의 투자를 유치했으며, 2024년에는 기업 가치가 약 55억 달러에 달했다. 이는 AI 기술이 상업적인 성공을 거두기 위해서는 기술력뿐만 아니라 이를 지원할 수 있는 자금 조달이 필수적이라는 점을 보여준다.

코히어의 주요 서비스로는 코히어 커맨드 R+(Command R+)가 있다. 이 솔루션은 기업들이 효율적으로 검색을 강화하고 신뢰성을 높일 수 있도록 설계된 RAG(검색 강화 생성형 인공지능) 모델이다. RAG는 대규모 언어 모델(LLM)을 사용하여 기업의 데이터로 특정 질문에 대한 답변을 제공하는 방식으로, Command R+는 이를 기업 환경에 맞춰 조정한 고급 솔루션이다. 예를 들어, 대형 보험회사가 고객의 복잡한 질문에 신속하고 정확하게 답변하기 위해 Command R+를 사용하는 모습은 마치 경험 많은 상담원이 고객의 문의에 적절하게 답변하는 것과 비슷하다.

67) 다윗 ‘미스트랄’, 골리앗 오픈AI에 도전 (김은광, 2024)
68) [단독] 유럽 AI스타트업 네이버, 지분투자 (황순민, 2024)

Command R+의 또 다른 강점은 클라우드 선택의 자유와 데이터 보안 기능이다. 기업들은 AWS, GCP, Azure와 같은 다양한 클라우드 플랫폼을 선택해 AI를 사용할 수 있으며, 오픈프라이버시 옵션을 통해 기업 내부 서버에 AI를 설치하여 데이터 유출을 방지할 수 있다. 또한 데이터 암호화를 통해 데이터 전송 및 저장 시에도 보안을 유지할 수 있다. 이는 중요한 서류를 안전한 금고에 보관하고, 필요한 경우에만 철저한 검증을 거쳐 열람하는 것과 같다.

코히어는 AI 솔루션이 실제 기업 현장에서 활용될 수 있도록 AI 모델을 최적화하는 데 집중하고 있다. 이는 맞춤형 정장을 제작해 고객의 몸치수는 감춰주고 고객의 몸에 딱 맞도록 최적화하는 과정과도 유사하다. 앞으로 코히어는 AI의 기업용 솔루션으로 비즈니스 적용 가능성을 높일 것으로 예상된다.

[그림 44] 코히어 기업 개요 및 제품⁶⁹⁾

코히어	코히어 코맨드 R+
<ul style="list-style-type: none"> • 코히어는 에이단 고메즈(Aidan Gomez), 닉 프로스트(Nick Frosst), 이반 장(Ivan Zhang)이 '19년 캐나다 토론토에 설립한 데이터 보안 및 맞춤형에 특화된 「기업용 AI」 전문 기업임. ◦ 에이단 고메즈는 최근 생성형 AI의 근간이 되고 있는 Transformer를 담은 “Attention is All You Need” 논문의 공동 저자이기도 함. • 기업용이라는 명확한 목표와 기술력 인정을 통해 지속 투자를 받고 있음. ◦ '23.06. 2.7억 달러 투자를 유치했고, '24.07. 5억 달러를 약 55억 기업 가치로 유치했음. 	<ul style="list-style-type: none"> • 코맨드 R+(Command R+) 특징 <ul style="list-style-type: none"> ◦ 환각을 줄이기 위한 인용을 통한 검색 강화 생성(RAG) <ul style="list-style-type: none"> ▪ RAG는 기업이 LLM과 자체 독점 데이터를 결합하여 특정 작업에 최적화된 AI 응답을 생성하는 기반 기술 ▪ Command R+는 기업 준비, 높은 신뢰성, 검증 가능한 솔루션을 제공하도록 고급 RAG에 최적화 ◦ 글로벌 사업 운영 지원을 위한 10개 주요 언어 지원 ◦ 복잡한 비즈니스 프로세스를 자동화하는 도구 사용 • 기업 특화 <ul style="list-style-type: none"> ◦ 클라우드 선택 자유 : AWS, GCP, Azure 등 선택 가능 ◦ 온프레미스 옵션: 기업 내부 서버에 AI를 설치하여 데이터 유출 위험 최소화 ◦ 데이터 암호화: 데이터 전송 및 저장시 강력한 암호화

3.5. 스테빌리티AI(StabilityAI)

스테빌리티AI는 이미지 생성형 AI 기술에 집중하고 있는 기업 중 하나로, 최근에는 멀티 차원과 비디오 생성 분야로 확대하고 있다. 최근 스테빌리티AI는 ‘스테이블 패스트3D(Stable Fast 3D)’와 ‘스테이블 비디오 4D(Stable Video 4D)’라는 혁신적인 기술을 선보이며 업계를 선도하고 있다.

‘스테이블 패스트3D’는 단일 이미지를 입력받아 고품질의 3D 자산을 0.5초 만에 생성하는 기술이다. 이는 어린아이가 그림 한 장을 보고 그 모양을 머릿속에서 입체적으로 상상해내는 것처럼, 컴퓨터가 단일 이미지에서 입체적 구조를 빠르게 만들어내는 것이다. 이러한 기술은 게임 개발, 애니메이션, 디자인 등 다양한 분야에서 사용될 수 있으며, 특히 빠른 제작 속도가 필요한 상황에서 유용하다.

69) GN+: Cohere의 Command R+ - 비즈니스를 위해 구축된 확장 가능한 LLM (권정혁, 2024)

‘스테이블 비디오 4D’는 단일 객체 비디오를 여러 각도와 시점에서 새로운 관점의 비디오로 변환하는 기술이다. 이는 영화 촬영 시 하나의 장면을 여러 카메라로 촬영해 다양한 각도에서 보는 것과 유사하다. 예를 들어, 제품 광고에서 하나의 제품을 여러 시점에서 보여줄 때 이 기술을 사용하면 간단하게 여러 각도의 비디오를 제작할 수 있다. 이로 인해 콘텐츠 제작의 효율성을 크게 높일 수 있으며, 마케팅 및 광고 분야에서 매우 유용하게 활용될 수 있다.

스테빌리티AI는 이러한 기술로 콘텐츠 생성의 속도와 품질을 더욱 향상시킬 것으로 기대된다. 특히 3D 및 4D 콘텐츠 제작에 있어 빠른 처리 속도와 높은 품질을 동시에 제공함으로써, 다양한 산업에서 스테빌리티AI의 기술을 채택할 것이다.

[그림 45] 3D, 4D로 확대하는 스테빌리티AI⁷⁰⁾⁷¹⁾

Stable Fast 3D	Stable Video 4D
	
<ul style="list-style-type: none"> • 단, 0.5초 만에 단일 이미지에서 고품질 3D 자산 생성 	<ul style="list-style-type: none"> • 단일 객체 비디오를 8개의 다른 각도/시점의 여러 개의 새로운 관점 비디오로 변환

4. AI 서비스 개발·배포

AI 서비스 개발·배포 분야는 AI의 실질적인 가치를 사용자에게 전달하는 중요한 역할을 한다. 초기에는 주로 ‘보편적 AI’ 형태의 챗GPT와 같은 대화형 AI가 중심이었지만, 시간이 지나면서 AI 서비스는 점점 더 다변화되고 있다. 이제는 AI가 단순한 질문 답변 기능을 넘어 플랫폼, 버티컬 AI, 소비자용 AI로 고도화되어 다양한 분야에서 구체적인 서비스를 제공하고 있다.

예를 들어, 챗GPT 같은 AI는 사용자가 질문을 하면 답변을 제공하는 데 그쳤지만, 이제는 특정 산업에 특화된 AI, 즉 ‘버티컬 AI’가 등장하여 금융, 의료, 제조 등 각기 다른 산업에서 맞춤형 솔루션을 제공하고 있다. 이러한 변화는 AI가 더 깊이 있는 지식과 기술을 바탕으로 특정 문제를 해결하는 데 적합하도록 발전하고 있음을 보여준다.

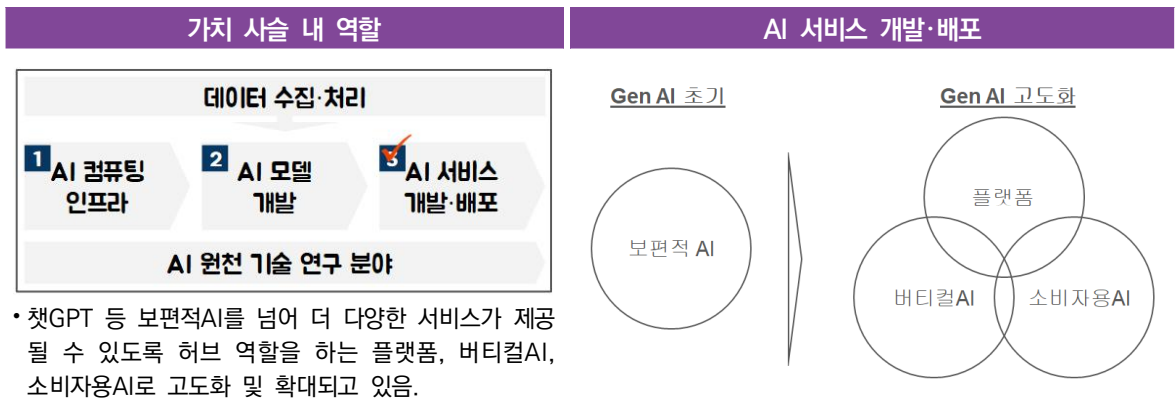
70) Introducing Stable Video 4D, Our Latest AI Model for Dynamic Multi-Angle Video Generation (stability.ai, 2024)

71) Introducing Stable Fast 3D: Rapid 3D Asset Generation From Single Images (stability.ai, 2024)

또한, 소비자용 AI는 스마트홈, 개인 비서 등 일상생활에서 사용자가 AI의 혜택을 누릴 수 있도록 돕고 있다. 예를 들어, AI 기반의 가전제품은 사용자의 생활 패턴을 학습해 최적의 조건으로 자동 조절하는 기능을 제공하며, 이는 사용자의 생활 습관을 잘 이해하는 개인 비서가 일상적인 일들을 대신 처리해주는 것과 같다.

이처럼 AI 서비스 개발·배포는 보편적 AI에서 시작하여 버티컬 AI와 소비자용 AI로 확장되면서 플랫폼으로서의 역할을 점점 더 강화하고 있다. AI는 더 이상 단순한 도구가 아닌, 다양한 서비스의 중심에 서서 사용자 경험을 개선하고 비즈니스 가치를 창출하는 핵심 요소로 자리매김하고 있다.

[그림 46] AI 서비스 개발·배포 개요



4.1. 플랫폼: 허깅페이스

허깅페이스는 “머신러닝계의 깃허브(GitHub)”를 목표로 하는 플랫폼으로, 개발자와 연구자들이 머신러닝 모델과 데이터 세트를 쉽고 빠르게 공유하고 협력할 수 있는 환경을 제공하고 있다. 이는 개발자들이 필요한 AI 모델과 데이터를 한 곳에서 구할 수 있도록 하여, AI 기술의 민주화를 이루겠다는 목표를 갖고 있다. 오픈소스 소프트웨어의 집합체인 깃허브(GitHub)이 전 세계 개발자들에게 필수적인 도구가 된 것처럼, 허깅페이스도 머신러닝 개발자들에게 필수적인 공간으로 자리잡아가고 있다.

허깅페이스는 다양한 기업과 연구자들이 자연어 처리(NLP), 컴퓨터 비전, 강화 학습 등의 머신러닝(ML) 기술을 활용할 수 있도록 데이터 세트, 라이브러리, API 및 커뮤니티를 지원하는 AI 플랫폼이다. 허깅페이스의 주요 서비스로는 트랜스포머 라이브러리, 모델 허브(Hub), 스페이스(Spaces), API, 자동 학습(AutoTrain) 등이 있다.

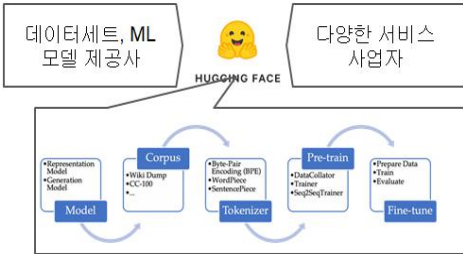
트랜스포머 라이브러리는 다양한 AI 모델을 개발하고 훈련할 수 있는 오픈소스 라이브러리로, BERT, GPT-2, T5와 같은 모델을 쉽게 활용할 수 있도록 설계되어 있다. 모델 허브는 사전 훈련된 모델과 데이터 세트를 공유할 수 있는 저장소 역할을 하며, 이를 통해 사용자들은 필요한 모델을 빠르게 찾고 활용할 수 있다.

메타(Meta)의 라마(Llama) 모델도 허깅페이스를 통해 사용자가 다운로드하여 활용할 수 있으며, 라마를 기반으로 한 다양한 파생 AI 모델들도 허깅페이스의 플랫폼에서 확산되고 있다. 허깅페이스는 이러한 도구와 커뮤니티를 통해 AI 및 머신러닝 생태계에서 필수적인 역할을 하고 있다.

스페이스(Spaces)는 개발자와 연구자들이 AI 모델을 구축하고 이를 호스팅하여 다른 사용자와 공유할 수 있는 플랫폼이다. 이는 각자의 작업실에서 만든 작품을 전시하고 서로의 작품을 참고하는 미술관과 같다. 허깅페이스의 API는 다양한 AI 모델을 쉽게 사용할 수 있도록 지원하며, 자동 학습(AutoTrain)은 사용자가 자신의 데이터를 업로드하면 그에 맞는 최적의 모델을 학습하고 평가하여 배포할 수 있게 도와주는 기능을 제공한다.

허깅페이스는 이러한 다양한 서비스를 통해 머신러닝 모델 개발의 진입 장벽을 낮추고, 누구나 쉽게 AI 기술을 활용할 수 있는 환경을 조성하고 있다. 뿐만 아니라, 모델과 데이터 세트 등 AI 서비스 개발에 필요한 모든 요소를 한 번에 구할 수 있는 플랫폼을 제공하며, “머신러닝계의 깃허브”로 자리매김해 나가고 있다.

[그림 49] ML계의 깃허브를 지향하는 허깅페이스⁷²⁾⁷³⁾

허깅페이스 ML계의 GitHub	제공 서비스										
<ul style="list-style-type: none"> 허깅 페이스는 다양한 기업, 연구자들이 NLP, 컴퓨터 비전 및 강화 학습과 같은 다양한 ML 활용 할 수 있도록 데이터 세트, 라이브러리, API 및 커뮤니티와 공유할 수 있도록 하는 호스팅 플랫폼 	<table border="1"> <tr> <td>트랜스포머 라이브러리</td><td>• 기업·연구자들이 트랜스포머를 활용할 수 있도록 라이브러리를 제공하는 오픈소스 저장소</td></tr> <tr> <td>허브</td><td>• 사전 훈련된 ML모델, 데이터 세트 및 ‘Space’ 제공</td></tr> <tr> <td>공간(Spaces)</td><td>• 개발자가 모델을 빌드, 호스팅 및 ML 커뮤니티와 공유할 수 있는 Spaces라는 플랫폼 제공</td></tr> <tr> <td>API</td><td>• 허깅 페이스 다양한 서비스를 API를 통해서 다양한 ML 모델들을 통합 사용할 수 있음.</td></tr> <tr> <td>자동 학습 (AutoTrain)</td><td>• 사용자가 데이터를 업로드하면, 그에 가장 적합한 모델을 찾아서 학습, 평가 배포함.</td></tr> </table>	트랜스포머 라이브러리	• 기업·연구자들이 트랜스포머를 활용할 수 있도록 라이브러리를 제공하는 오픈소스 저장소	허브	• 사전 훈련된 ML모델, 데이터 세트 및 ‘Space’ 제공	공간(Spaces)	• 개발자가 모델을 빌드, 호스팅 및 ML 커뮤니티와 공유할 수 있는 Spaces라는 플랫폼 제공	API	• 허깅 페이스 다양한 서비스를 API를 통해서 다양한 ML 모델들을 통합 사용할 수 있음.	자동 학습 (AutoTrain)	• 사용자가 데이터를 업로드하면, 그에 가장 적합한 모델을 찾아서 학습, 평가 배포함.
트랜스포머 라이브러리	• 기업·연구자들이 트랜스포머를 활용할 수 있도록 라이브러리를 제공하는 오픈소스 저장소										
허브	• 사전 훈련된 ML모델, 데이터 세트 및 ‘Space’ 제공										
공간(Spaces)	• 개발자가 모델을 빌드, 호스팅 및 ML 커뮤니티와 공유할 수 있는 Spaces라는 플랫폼 제공										
API	• 허깅 페이스 다양한 서비스를 API를 통해서 다양한 ML 모델들을 통합 사용할 수 있음.										
자동 학습 (AutoTrain)	• 사용자가 데이터를 업로드하면, 그에 가장 적합한 모델을 찾아서 학습, 평가 배포함.										

72) Hugging Face 홈페이지(<https://huggingface.co/>)

73) Pre-train and Fine-tune Language Model with Hugging Face and Gaudi HPU. (Congress.gov, 2022)

4.2. 버티컬 AI(Vertical AI) 서비스

버티컬 AI는 특정 산업 또는 분야에 특화된 AI 솔루션으로, 보편적인 AI와 달리 특정한 목적을 가지고 설계된 것이 특징이다. 초기의 보편적 AI인 챗GPT와 같은 모델들은 누구나 접근할 수 있는 형태였지만, 환각(Hallucination)과 같은 문제로 인해 기업용이나 산업용으로 활용하기에 한계가 있었다. 이러한 이슈 때문에 더욱 신뢰성 있고 구체적인 산업 문제를 해결할 수 있는 버티컬 AI가 등장하게 되었다. 예를 들어, 의료, 법률, 금융, 제조, 교육 등 각기 다른 분야에서 맞춤형 AI 솔루션을 제공하는 것이다.

버티컬 AI의 핵심은 각 산업의 특수성을 반영하여 구체적인 문제를 해결하는 데 있다.

첫째 의료 분야에서는 ‘에이브릿지(Abridge)’라는 AI 솔루션이 의사와 환자 간 대화 내용을 자동으로 기록하고 분석하여 의료 효율성을 크게 향상시킨다. 에이브릿지는 의사들이 환자와의 대화를 보다 쉽게 기록하고 정리할 수 있도록 지원하는 의료 전문 AI 플랫폼으로, 일상적인 업무인 의료 기록 작성에서 큰 변화를 이끌어 내고 있다⁷⁴⁾.

의료진은 많은 시간을 진료 기록을 작성하고 이를 전자의무기록(EHR) 시스템에 업로드하는 데 할애한다. 피츠버그에 기반을 둔 Abridge는 이 과정을 자동화하는 플랫폼을 제공한다. 의사들은 모바일 앱을 통해 대화를 녹음하고, 그 대화를 자동으로 기록한 후 이를 기반으로 의학 노트를 생성하는 기능을 사용할 수 있다. 이 과정은 모든 기록을 자동화하지는 않지만, 생성된 초안을 의사들이 검토하고 수정할 수 있는 기능을 제공한다. 특히, 에이브릿지의 ‘Linked Evidence’ 기능은 AI가 생성한 각 문장을 대화에서 추출된 구체적인 발췌본과 연결시켜 오류를 쉽게 찾아낼 수 있도록 한다.

이 플랫폼은 전자의무기록 시스템인 에픽(Epic)과 통합되어 있으며, 의료진은 여러 탭을 전환하지 않고도 AI가 생성한 의료 노트를 빠르게 편집할 수 있다. 에이브릿지는 이 시스템을 통해 의료 기록 작성 작업의 91%를 자동화할 수 있다고 주장하며, 이를 통해 의료진이 매달 약 70시간을 절약할 수 있다고 보고했다. 이는 의료진이 환자 진료에 더 많은 시간을 할애할 수 있도록 하며, 더 세부적인 의료 기록을 작성할 수 있게 해준다.

에이브릿지의 플랫폼은 자체 개발한 AI 모델로 구동되며, 14개 이상의 언어와 50개 이상의 의료 전문 용어를 이해할 수 있다. 또한, 경쟁사보다 월등한 성능을 자랑하며 의료 기록 자동화 작업에서 다른 AI 모델들을 크게 능가한다고 한다. 이 회사는 2024년 10월 25억\$ 기업가치로 약 2.5억\$의 투자 유치를 추진 중이라 알려져 있다.

둘째, 법률 분야에서는 ‘하비(Harvey)’라는 AI가 계약서 검토, 리스크 평가, 및 문서 분석을 빠르게 수행하여 법률 전문가들의 업무 효율성을 크게 높이고 있다. 하비는 AI를 기반으로 한 “법률 보조” 역할을 수행하는

74) Healthcare AI startup Abridge reportedly raising \$250M at \$2.5B valuation (DeutscherMaria, 2024)

솔루션으로, 주요 로펌인 Allen & Overy, Macfarlanes, Ashurst, CMS, Reed Smith, PwC와 같은 곳에서 일상적으로 사용되고 있다. 이 AI는 계약서나 법률 문서의 방대한 데이터를 분석하여 법적 위험 요소를 빠르게 식별하고 검토 과정을 단축해주는 역할을 한다.

하비는 오픈AI의 GPT-4 모델을 기반으로 하며, 법률 문서에서 중요한 정보를 추출하고, 리스크를 자동으로 평가하며, 필요한 경우 조항을 재작성하여 법적 문제를 해결한다. 예를 들어, 변호사는 “이 조항이 캘리포니아 법을 위반하는지 확인하고, 그렇다면 수정하라”는 자연어로 질문을 할 수 있고, 하비는 이를 분석하여 답변과 수정된 문구를 제공한다. 또한 하비는 재판 기록에서 중요한 정보를 자동으로 추출하거나 법률 데이터베이스에서 인용 자료를 찾아 초안을 작성하는 등의 기능을 제공해 법률 작업의 많은 부분을 자동화하고 있다.⁷⁵⁾

2024년 7월, 하비는 구글 벤처스(Google Ventures)가 주도한 시리즈C 펀딩에서 1억 달러 규모의 투자 유치에 성공했으며, 이로써 기업 가치는 15억 달러로 평가받고 있다. 이 자금은 새로운 도메인 특화 AI 모델 개발과 인력 확장, 글로벌 서비스 확장을 위한 기반이 될 것이다. 하비는 새로운 AI 모델을 지속적으로 개발하고 이를 법률 분야의 복잡한 지식 업무에 적용할 계획이다. 이 AI는 현재 수만 명의 변호사들이 매일 사용하고 있으며, 연간 매출이 세 배로 증가하는 등 빠르게 성장하고 있다.

셋째, 제조 분야에서는 ‘악시온 레이(Axion Ray)’가 AI 기반 솔루션을 통해 제조 데이터를 분석하고 품질 관리 및 생산성을 최적화하고 있다. 2024년 3월 시리즈 A 라운드에서 Bessemer Venture Partners가 주도한 1,750만 달러의 투자를 유치하며 기업 가치를 1억 달러로 평가받았다.⁷⁶⁾ 특히 악시온 레이는 제품 결함을 조기에 탐지하여 리콜을 방지하는 데 중점을 두고 있으며, 이는 마치 생산 라인에서 발생하는 모든 문제를 AI가 실시간으로 모니터링하고, 최적의 해결책을 제시하는 것과 같다.

악시온 레이는 다양한 데이터 소스, 예를 들어 현장 보고서, 센서 데이터, 지리적 위치 정보 등을 결합하여 제품 결함의 초기 신호를 포착하고, 이를 분석해 리콜로 이어질 수 있는 문제를 사전에 해결할 수 있도록 돕는다. 이 플랫폼은 자동차, 항공우주, 소비자 전자제품, 의료기기 등의 분야에서 사용되며, 고객으로는 Boeing과 Denso 같은 대형 제조업체가 포함되어 있다. 악시온 레이의 CEO인 다니엘 퍼스트(Daniel First)에 따르면, AI 솔루션을 통해 제조업체는 더 이상 개별 부서 간에 중복된 분석을 수행하는 대신, 통합된 데이터 플랫폼을 사용해 협업하며 문제를 해결할 수 있다.

악시온 레이의 AI 알고리즘은 예를 들어 특정 차량 모델에서 발생한 ABS(안티록 브레이크 시스템)의 결함을 먼저 현장 기술자의 보고서에서 탐지한 후, 콜센터의 고객 불만, 정비소 방문 기록, 차량 텔레메트리 데이터를 결합하여 유사한 문제를 신속하게 파악한다. 이렇게 통합된 데이터 분석을 통해 제조업체는 문제를 조기에 인식하고, 적절한 수정 조치를 취해 리콜을 방지하며 비용을 절감할 수 있다.

75) OpenAI-backed legal tech startup Harvey raises \$100M (WiggersKyle, 2024)

76) Axion Ray's AI attempts to detect product flaws to prevent recalls (WiggersKyle, 2024)

이러한 버티컬 AI의 확장은 보편적 AI가 해결할 수 없는 구체적인 산업 문제에 대한 솔루션을 제공하며, 산업 내 효율성을 극대화하고 있다. 특히, 생산성과 효율성 같은 명확한 가치가 요구되는 제조, 금융 등의 분야에서 버티컬 AI는 큰 역할을 하고 있다. 이는 마치 각 분야에 특화된 전문가가 등장하여 해당 문제를 보다 깊이 있게 해결해주는 것과 같은 방식으로, AI의 활용 범위를 더욱 넓히고 있다.

[그림 48] 버티컬AI 확대⁷⁷⁾

버티컬 AI 확대 도래		예시				
<div><div>1 버티컬AI</div><div>• 특정 산업이나 분야에 특화된 AI 솔루션</div><div>2 챗GPT 등 보편적인 AI 확산</div><div>• 챗GPT로 인해 보편적 AI 활용이 늘어가는 가운데, 특정 산업 또는 영역 문제를 해결 못하는 사례 등장</div><div>◦ 생산성·효율성과 같은 명확한 가치 요구 증대</div><div>◦ 환각 등의 피해가 커지면서 정확성 향상에 대한 요구 증대</div></div>		산업 분야	기업명	AI솔루션	주요 기능	기대 효과
		의료	Abridge	의료 대화 분석	진료 내용 자동 기록	의사의 업무 효율성 증대, 환자 이해도 향상
		법률	Harvey	법률 문서 분석	계약서 검토, 리스크 평가	법률 업무 처리 속도 및 정확성 향상
		제조	Axion Ray	제조 데이터 분석	품질 관리, 생산성 향상	제조 과정 최적화, 비용 절감

뿐만 아니라 한 산업 전반에 적용되어 변화시킬 수도 있다. 글로벌 컨설팅 회사인 KPMG는 AI를 통한 금융 산업 내 프론트-미들-백 오피스 전반의 변화가 도래할 것이라 전망한 바 있다.

프론트 오피스에서는 AI를 통해 고객 맞춤형 서비스가 자동화되고, 챗봇과 음성 비서 등을 활용해 고객과의 상호작용을 실시간으로 처리할 수 있다. 이러한 기술들은 고객의 요구에 빠르게 대응할 뿐 아니라, 거래 및 상품 추천에서 더 높은 정확도를 제공함으로써 고객 만족도를 크게 향상시킨다.

미들 오피스에서는 AI가 리스크 관리와 준법 감시를 강화한다. 특히, 방대한 데이터 분석을 통해 실시간으로 잠재적 위험을 탐지하고, 위험 요소를 조기에 제거하는 것이 가능해진다. 이로 인해 금융 기관은 위험을 보다 효율적으로 관리하고, 규제 준수 비용을 절감할 수 있다.

백 오피스에서는 AI를 통한 자동화가 급진적인 운영 비용 절감을 가져올 것으로 기대된다. 데이터 입력, 회계, 보고서 작성 등 반복적이고 수작업에 의존했던 업무들은 AI 알고리즘을 통해 자동화되어 더욱 정확하고 신속하게 처리된다. 이는 금융 기관이 더 적은 인력으로도 더 많은 업무를 처리할 수 있도록 돕는 동시에, 오류 발생률을 줄여 비용을 절감하는 데 기여할 것이다.

77) Vertical AI :SaaS의 미래와 산업 혁신의 새로운 물결 (신동형)

[그림 49] 금융사 조직 내 전반에 걸친 AI 적용⁷⁸⁾

프런트 오피스 (고객, 영업마케팅, 상품 등)		미들 오피스 (리스크, 컴플라이언스)	백 오피스 (계약관리, 심사, 운영 등)	
AI-Powered Customer Service (고객 경험 혁신)		AI-Powered Insight & Action (인사이트 및 실행)	AI-Powered Compliance (리스크 관리, 컴플라이언스 준수)	
AI-Powered Operational Efficiency (운영 최적화)		AI-Powered Compliance (리스크 관리, 컴플라이언스 준수)		AI-Powered Operational Efficiency (운영 최적화)
AI 챗봇, 가상비서	AI 기반 신용평가	이상거래탐지(FDS)	AI 기반 심사(대출, 언더라이팅 등)	
AI 컨택 센터	AI 기반 상품 개발	자금세탁방지(AML)	AI 기반 디지털 결산	
신원인식(얼굴, 음성)	AI 알고리즘 트레이딩	AI 기반 품질관리(불완전판매 등)	지능형 서류 관리(계약 관리 등)	
초개인화 서비스	지능형 예측	AI 기반 리스크 관리(운영/시장/신용)	합성 데이터	소프트웨어 코딩

※ 출처 : 혁신의 부스터 AI에 물드는 금융(KPMG, 2024)

4.3. 소비자용 AI

2024년, 소비자용 AI 애플리케이션이 급격히 확산되고 있다고 a16z가 발표했다. 자료에 따르면, 월간 활성 사용자를 기준으로 가장 많이 사용되는 50개의 AI 모바일 앱 중 다수가 소비자용 AI 서비스로 구성되어 있다. 특히 챗GPT, MS 엣지 브라우저, 포토매스(Photomath)와 같은 앱들이 상위를 차지하고 있으며, 이는 소비자들이 AI를 활용해 일상에서 다양한 편의와 창의적인 도구를 활용하고 있음을 보여준다.

특히, 웹 기반 AI 앱의 52%가 콘텐츠 생성과 편집에 집중하고 있다. 이는 소비자들이 AI를 단순한 정보 검색 도구를 넘어 창의적인 결과물을 만들어 내는 데 활용하고 있음을 보여준다. 예를 들어, 레미니(Remini) 같은 앱은 사용자의 옛 사진을 고화질로 복원하여, 오래된 앨범 속 희미한 기억을 선명하게 되살리는 기능을 제공한다. 이처럼 AI는 개인의 창의적 활동을 지원하며, 삶의 질을 향상시키는 도구로 자리잡고 있다.

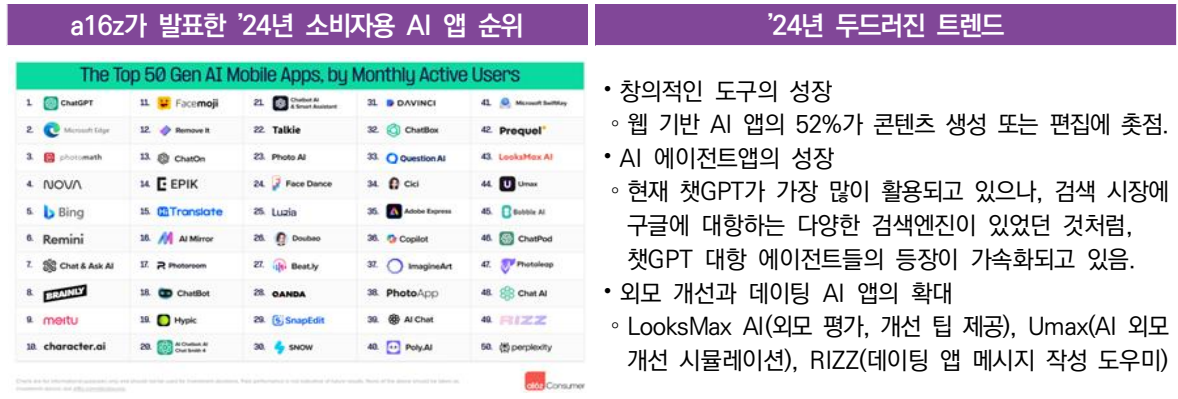
또한, AI 에이전트 앱의 성장도 두드러지고 있다. 기존의 챗GPT가 가장 널리 활용되는 가운데, 검색 시장에서는 구글에 도전하는 다양한 검색 에이전트가 잇따라 출시되고 있다. 이러한 에이전트는 사용자가 정보를 더욱 다양하고 효율적인 방식으로 검색하고 활용할 수 있도록 돕는다. 예를 들어, 챗GPT 대신 특정 목적에 특화된 AI 에이전트를 통해 더욱 정교하고 관련성 높은 결과를 얻을 수 있다.

외모 개선과 데이터 기반의 메시지 작성 도우미와 같은 앱들도 주목받고 있다. 룩스맥스AI (LooksMax) AI는 외모 평가 및 개선 팁을 제공하며, 사용자가 자신의 모습을 더 잘 이해하고 개선할 수 있도록 돕는다. 이 외에도, 데이터 기반의 메시지 작성 도구인 리프(RIZZ)는 사용자들이 보다 효과적인 의사소통을 할 수 있도록 돕고 있다. 예를 들어, 중요한 이메일을 작성할 때 리프를 활용하면 AI가 문맥에 맞는 적절한 표현을 추천해주는 식이다.

78) 혁신의 부스터 AI에 물드는 금융 (삼성KPMG 경제연구원, 2024)

이처럼 소비자용 AI는 소비자의 생활 속에서 점점 더 중요한 역할을 하며, 단순한 검색에서 창의적 활동과 개인 맞춤형 서비스까지 다양한 용도로 활용되고 있다. 앞으로도 이러한 소비자용 AI 서비스의 발전은 사용자 경험을 더욱 풍부하게 하고, AI가 우리 생활 속에서 필수적인 도구로 자리잡아 갈 것으로 예상된다.

[그림 50] 확대되고 있는 소비자용 AI 앱 현황⁷⁹⁾



4.4. AI 에이전트 : 퍼플렉시티 AI

퍼플렉시티(Perplexity) AI는 2022년에 오픈AI 출신 스리니바스(Srinivas)와 메타 AI 연구원 출신인 데니스 야라츠(Denis Yarats) 등이 설립한 AI 기반 검색 엔진이다. 퍼플렉시티는 차별화된 검색 경험을 제공하기 위해 검색 결과와 실시간 질문 응답을 결합한 방식을 도입하였다. 이러한 방식은 사용자가 원하는 정보를 보다 빠르고 정확하게 얻을 수 있도록 돕는다. 예를 들어, 일반적인 검색 엔진에서 사용자는 여러 링크를 클릭해 필요한 정보를 찾아야 하지만, 퍼플렉시티는 질문을 입력하면 직접 필요한 답변을 대화형으로 제공한다.

퍼플렉시티 AI는 2024년 10월 기준으로 80억\$(약 11조원) 이상의 기업 가치를 받는 신규 자금 유치를 진행 중이다. 2024년 초 5.2억\$ 기업 가치가 기준이었는데 자금 유치가 완성된다면 1년도 안되어 15배 이상으로 빠르게 성장하고 있다고 볼 수 있다. 투자자로는 아마존 창업자인 제프 베조스, 엔비디아, SKT 등이 포함 되어 있으며, 이는 퍼플렉시티의 기술력과 성장 가능성을 시장에서 인정받고 있다는 것을 보여준다.

퍼플렉시티의 주요 기술은 RAG(Retrieval-Augmented Generation) 시스템이다. 퍼플렉시티는 RAG를 통해 실시간 웹 검색과 생성 AI의 응답을 결합하여 복잡한 질문에 대한 실시간 검색과 신속한 응답을 제공한다. 예를 들어 여행지를 검색할 때 단순히 장소 정보만을 보여주는 것이 아니라, 여행 계획, 추천 루트까지도 AI가 제안해준다. 이는 전문 여행 가이드가 사용자의 기호와 목적에 맞는 최적의 여행 계획을 실시간으로 제공해주는 것과 같다.


79) '24년 소비자 AI앱 시장 동향:성장 동인과 미래 전망 (신동형)

퍼플렉시티의 수익 모델은 ‘프로(Pro) 등 구독모델(Subscription) 버전’과 광고 기반 모델로 구성된다. 프로 사용자들은 무료로 비해서 더 높은 성능의 AI 모델을 활용할 수 있으며 사용 횟수의 제약이 거의 없는 환경에서 AI의 모든 기능을 사용할 수 있다. 그리고 광고는 응답에 직접 통합되어 자연스러운 형태로 제공된다. 예를 들어, 도쿄 여행지 정보를 검색하면 관련된 스폰서의 호텔 광고가 함께 표시되는 식이다. 이는 사용자가 필요한 정보를 얻는 동시에 관련된 서비스도 참고할 수 있게 하여 사용자 경험을 개선하고 광고 효과도 극대화하는 방식이다.

퍼플렉시티는 검색 AI의 새로운 가능성을 보여주고 있다. 일반적인 검색 엔진이 링크 중심의 정보를 제공하는 반면, 퍼플렉시티는 복잡한 질문에도 직접적인 답변을 제공하여 사용자의 시간을 절약하고 검색 경험을 향상시킨다. 이러한 방식은 AI가 정보를 단순히 제공하는 것을 넘어, 사용자의 의도를 파악하고 그에 맞는 최적의 솔루션을 제시할 수 있는 진정한 ‘지식 도우미’ 역할을 하게 하는 것이다.

이처럼 퍼플렉시티 AI는 검색의 패러다임을 변화시키고 있으며, 실시간 검색과 자연어 처리 기술의 결합을 통해 사용자가 더 나은 결정을 내릴 수 있도록 돕겠다는 의지를 갖고 발전하고 있다.

[그림 51] 퍼플렉시티AI 기업 개요 및 수익모델⁸⁰⁾

퍼플렉시티 개요	수익 모델
<ul style="list-style-type: none"> • 퍼플렉시티 AI는 '22년 오픈AI출신 스리니바스(Srinivas)와 메타 AI 연구원 출신인 데니스 야랏츠(Denis Yarats) 등이 설립한 AI 기반 검색 엔진임. • '24년 기업가치 10억\$ 이상으로 약 1.65억\$의 자금을 조달했음. 투자자로서는 아마존 창업자인 제프 베조스, 엔비디아 및 SKT도 있음. <div data-bbox="135 1242 696 1327"> <div>검색어·질문입력</div> <div>검색 (RAG)</div> <div>자연어처리 (다양한 시도·실패 반복)</div> <div>다양한 작업수행</div> </div> <ul style="list-style-type: none"> • RAG시스템: 실시간 웹 검색과 AI생성 응답의 결합 • 자연어처리: 복잡한 질문을 이해하고 응답 제공 • 정확성 중시: 출처 명시로 신뢰성 확보 • 다양한 작업 수행: 코드 작성, 수학문제 해결, 표생성 등 	<ul style="list-style-type: none"> • Pro 등 Subscription 버전 • 광고 : 광고가 답변에 직접 통합됨. 광고 유형으로는 스폰서드 미디어, 관련 질문, 답변 페이지 전면 광고 가능 <div data-bbox="728 1123 1278 1471"> <p>We've designed our product from the ground up so AI will read relevant information from the web and synthesize it for you in a conversational way.</p>  </div>

80) Perplexity (Perplexity, 2024)

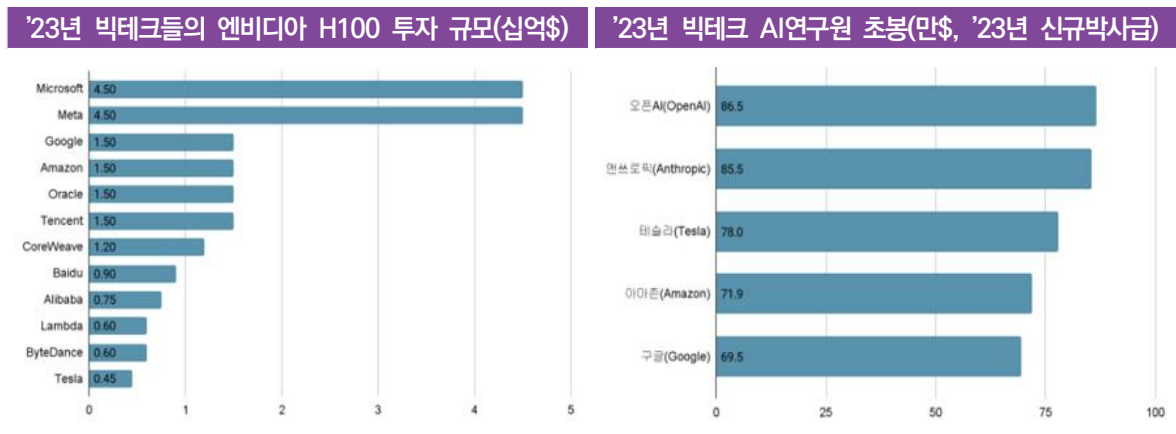
Ⅳ. 대한민국 AI 경쟁력 강화를 위한 제언

1. 거대 기업들이 주도하는 AI산업

AI 산업이 소수의 거대 기업들에 의해 주도되는 현상이 뚜렷해지고 있다. MS, 메타, 구글, 엔비디아와 같은 전통적인 빅테크 기업들뿐만 아니라, 오픈AI, 앤쓰로픽(Anthropic), 미스트랄AI(Mistral AI)와 같은 기업들은 불과 몇 년 만에 글로벌 기업으로 성장했다.

이들 기업의 급속한 성장은 AI 기술의 핵심인 고성능 GPU와 대규모 데이터 처리 능력, 그리고 뛰어난 연구 인력을 확보했기 때문이다. 이들은 수만 대의 GPU를 갖춘 대규모 AI 데이터센터를 활용하며, 연간 수십억 달러를 AI 연구개발에 투자한다. 또한 파격적인 연봉 조건으로 세계 최고의 AI 인재들을 영입하고 있다.

[그림 52] 머니 게임 중심인 AI 인프라, 모델 개발⁸¹⁾⁸²⁾



대규모 투자와 자원 집중은 AI 기술의 급속한 발전을 촉진하고 있다. GPT, 클로드(Claude), 미스트랄(Mistral)과 같은 대규모 언어 모델은 이러한 노력의 결과물이다. 이들 기반 기술은 AI 산업의 발전 방향을 결정하며, 이를 바탕으로 다양한 AI 서비스와 애플리케이션이 개발되고 있다.

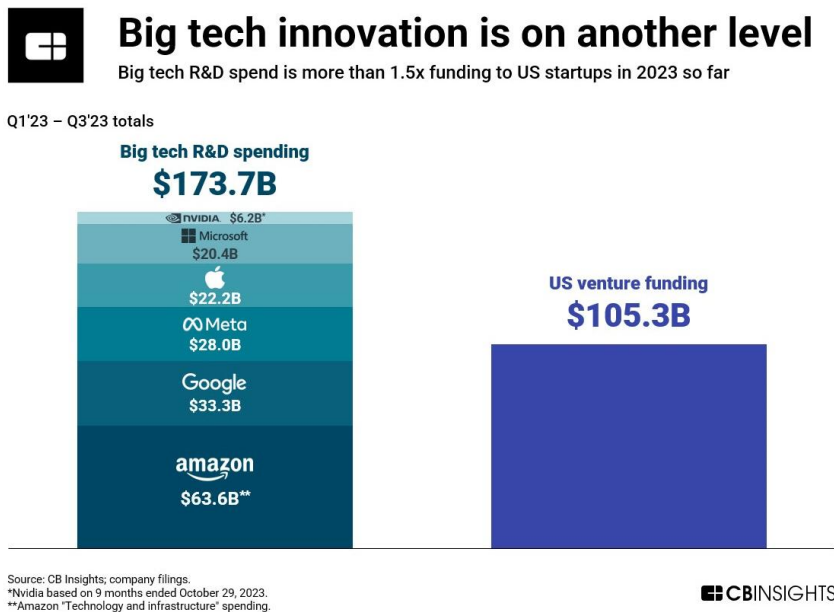
AI 산업에서는 소수의 거대 AI 기업들이 막대한 자본을 바탕으로 기반 기술을 선점하고, 나머지 기업들은 이들이 제공하는 API나 모델을 활용한 서비스 개발에 주력하는 구조로 발전할 가능성이 있다.

81) Nvidia's biggest customers are also the AI chip maker's biggest threat (ChengMichelle, 2024)

82) 오픈AI 초봉 12억 vs 한국 2억...머스크 "가장 미친 인재 전쟁" (고은이, 황동진, 2024)

CB Insights에 따르면, 2023년 1분기(Q1)부터 3분기(Q3)까지 빅테크 기업들은 총 1,737억 달러를 연구개발(R&D)에 지출한 반면, 같은 기간 미국 스타트업들이 유치한 벤처 펀딩은 1,053억 달러에 그쳤다. 이는 빅테크의 R&D 지출이 벤처 펀딩 규모를 약 1.65배 초과하며, 두 그룹 간의 투자 격차를 명확히 보여준다. 이러한 투자로 인해 빅테크는 스타트업 생태계보다 훨씬 강력한 혁신 기반을 구축하고 있으며, AI 산업의 속도와 방향을 결정짓는 핵심 주체로 자리 잡고 있다. 당분간 AI 산업은 이들 빅테크 기업의 주도로 빠르게 성장할 것으로 전망된다.

[그림 53] 빅테크와 스타트업 간 혁신 투자 격차⁸³⁾



※ 출처 : CB Insights(2023.12.), 'The future of big tech in 10 charts'

그러나 시장이 더욱 성장할수록 새로운 기회도 나타날 것이다. 과거 컴퓨팅 시장에서는 IBM이 초기 시장을 장악했지만, 시장이 확장되면서 인텔의 CPU와 MS의 OS 중심으로 재편되었다. 이후 모바일 컴퓨팅 시장에서는 퀄컴과 다양한 AP, 안드로이드OS와 iOS가 부상했으며, 현재 AI 시대에 들어서는 엔비디아의 GPU와 다양한 AI 파운데이션 모델이 시장을 이끌고 있다. 여전히 AI 산업에서 엔비디아의 독점을 깨기 위한 다양한 시도가 이뤄지고 있듯, AI도 분명 시장이 확대되고 고도화되며 새로운 기회가 생길 것이다.

따라서 우리는 시장 변화에 빠르게 적응하고, AI 산업에서 경쟁력을 강화하며, 새로운 기회를 선도하기 위해 지속적으로 노력해야 한다.

83) <https://www.cbinsights.com/research/report/big-tech-future-charts/>

2. 스마트폰 시대의 교훈과 AI 산업에의 적용

2.1. 스마트폰 시대 대한민국의 성공 전략

스마트폰 혁명은 21세기 초반 전 세계 산업 구조를 근본적으로 변화시킨 중요한 사건이었다. 이 시기에 대한민국은 매우 성공적인 전략을 펼쳤다. 우리나라는 운영체제(OS)나 애플리케이션 프로세서(AP) 주도권에 집착하지 않았다. 대신, 스마트폰 제조, 메모리 및 디스플레이 등 핵심 부품 산업 강화, 그리고 안드로이드 OS와 iOS 기반의 앱 서비스 개발에 주력했다. 그 결과, 다음과 같은 성과를 이루며 글로벌 시장에서의 경쟁력을 확보할 수 있었다.

첫째로, 스마트폰 제조에 집중하여 삼성전자와 LG전자를 중심으로 한 국내 기업들은 뛰어난 하드웨어 제조 능력을 바탕으로 고품질의 스마트폰을 생산했다. 이들은 빠르게 변화하는 소비자의 요구를 민첩하게 반영하여 다양한 모델의 스마트폰을 출시했고, 이는 글로벌 시장에서 큰 호응을 얻었다.

둘째로, 메모리 및 디스플레이 등 핵심 부품 산업을 강화하여 삼성전자와 SK하이닉스가 메모리 반도체 분야에서 세계 시장 점유율 69% 수준을 차지하며 압도적인 경쟁력을 보여주었다⁸⁴⁾. 또한 삼성디스플레이와 LG디스플레이는 OLED 등 첨단 디스플레이 기술을 선도하며 스마트폰 산업의 발전에 크게 기여했다.

셋째로, 안드로이드 OS와 iOS 기반의 앱 서비스 개발에 주력한 카카오, 네이버 등 국내 IT 기업들은 글로벌 OS 플랫폼을 기반으로 혁신적인 모바일 서비스를 개발하여 국내 시장을 석권하고, 나아가 해외 진출의 발판을 마련했다.

이러한 전략적 선택은 대한민국이 글로벌 스마트폰 시장에서 선도적인 위치를 차지하는 데 결정적인 역할을 했다. 특히 삼성전자의 사례는 주목할 만하다. 삼성전자는 초기에 자체 OS인 '바다OS'를 개발했지만, 시장의 변화를 빠르게 인지하고 과감히 안드로이드를 채택했다. 이는 글로벌 시장에 빠르게 적응할 수 있게 해주었고, 결과적으로 애플과 함께 글로벌 스마트폰 시장을 양분하는 성과를 이루었다. 더 나아가 삼성전자는 2019년 세계 최초로 5G 스마트폰을 상용화하는 등 기술 혁신을 선도했다. 이는 스마트폰 제조뿐만 아니라 네트워크 장비, 반도체 등 관련 산업 전반에서의 경쟁력이 시너지 효과를 낸 결과였다.

84) Home Industry insights Strategy Insights Gen AI, HPC to fuel HBM market growth Gen AI, HPC to fuel HBM market growth (Yole Group, 2024)

2.2. 노키아와 일본 기업들의 실패 사례

반면, 한때 글로벌 휴대폰 시장을 석권했던 노키아와 일본 기업들의 사례는 우리에게 중요한 교훈을 준다. 노키아는 자체 OS인 심비안(Symbian)에 집중하다가 시장 변화에 적응하지 못했다. 심비안은 한때 세계 스마트폰 OS 시장의 절반 이상을 차지하며 절대적인 입지였지만, 애플의 iOS와 구글의 안드로이드가 등장하면서 급격히 시장 점유율을 잃었다. 노키아는 이러한 변화에 신속하게 대응하지 못했고, 결국 스마트폰 시장에서의 주도권을 완전히 상실했다.

일본 기업들의 경우, 자국 시장에 최적화된 피쳐폰에 안주하여 글로벌 트렌드를 놓쳤다. 일본의 휴대폰은 이메일, 모바일 결제, 디지털 TV 등 다양한 기능을 갖춘 고성능 제품이었지만, 이는 일본 시장에 특화된 것이었다. 이른바 ‘갈라파고스 증후군’으로 불리는 이 현상은 일본 기업들이 글로벌 시장의 요구를 제대로 파악하지 못하고 자국 시장에만 집중한 결과였다. 이러한 사례들은 폐쇄적 전략과 글로벌 표준 수용 실패가 얼마나 큰 위험을 초래할 수 있는지를 잘 보여준다. 특히 빠르게 변화하는 기술 산업에서는 글로벌 트렌드를 정확히 파악하고 이에 신속하게 대응하는 것이 매우 중요하다.

2.3. AI 시대에의 적응

AI 혁명은 스마트폰 혁명을 뛰어넘는 더 큰 변화를 가져올 것으로 예상된다. AI는 단순히 하나의 제품이나 서비스를 넘어 모든 산업과 일상생활에 깊이 침투하여 우리의 삶을 근본적으로 변화시킬 것이다. 스마트폰 시대의 성공 전략이 IT 산업의 성공 공식을 보여주었듯, AI 시대에도 중요한 시사점을 제시한다.

첫째, 글로벌 표준의 신속한 수용이 중요하다. 그 이유는 IT 산업의 가장 큰 특징인 승자독식(勝者獨食, Winner takes it all)이 반영되기 때문이다. 글로벌 표준을 바탕으로 생태계, 즉 가치 사슬이 전개되기에 표준을 벗어나서는 생존 자체가 힘들 수밖에 없다. 이는 스마트폰 시대에 노키아가 자체OS와 생태계에 집중하며 퇴출된 것과 반대로 삼성전자가 빠르게 안드로이드OS를 채택하고 이를 기반으로 빠르게 스마트폰 기기 산업을 장악한 것이 예시가 될 것이다. AI 시대에도 글로벌 표준 기술과 플랫폼을 창출하는 것도 중요하지만, 그렇지 않은 경우, 신속하게 수용하고 이를 바탕으로 혁신적인 서비스와 제품을 개발하는 전략이 필요하다.

둘째, 우리 강점을 바탕으로 부족한 부분은 협력을 통해 보완하며 경쟁력을 향상시킨다. 스마트폰 사업을 철수한 LG그룹이 여전히 스마트폰 부품사업에서 높은 경쟁력을 확보할 수 있는 것은 애플과의 협업 때문이다. LG그룹은 LG전자 스마트폰에만 집착하지 않고, LG의 부품 기술 경쟁력을 바탕으로 애플과 협업을 통해 더 나은 카메라, 디스플레이를 개발했기 때문에 스마트폰 시대에 경쟁력 키울 수 있었다. 이처럼 AI 시대에도 우리가 경쟁력을 가질 수 있는 특정 분야에 자원을 집중하며 글로벌 협력을 함께 모색해야 한다.

셋째, 빠른 시장 적용과 소비자 경험을 통해 새로운 시장 기회를 찾고 장악해 나가야 할 것이다. 2012년 글로벌 스마트폰의 침투율이 15% 수준⁸⁵⁾이었을 당시, 우리나라 성인의 스마트폰 사용율은 약 70%에 육박하는 수준⁸⁶⁾으로 빠른 스마트폰 활용이 스마트폰 제품과 부품, 스마트폰 앱과 서비스 확대를 가져왔었다.

AI 신산업 육성도 중요하겠지만, 내부에만 집중하다가 일본과 같이 갈라파고스 섬과 같이 될 수 있으니, 상황에 따라 AI 시대 빠른 도입과 활용을 통해 부가 산업 확대에도 더 관심을 갖고 지원을 하는 방향도 고려할 수 있을 것이다.

예를 들어, 우리나라의 발달된 IT 인프라와 높은 스마트폰 보급률, 그리고 새로운 기술에 대한 높은 수용성을 바탕으로 혁신적인 AI 서비스를 개발하고 이를 글로벌 시장으로 확장하는 전략을 고려할 수 있다. 또한, 이미 강점을 가진 AI 메모리를 바탕으로 어떻게 AI 반도체 및 데이터 센터로 그 영향력을 확대할 수 있을까 고민이 필요할 것이다.

3. 대한민국의 강점과 전략적 선택

앞서 언급한 도전 과제들에도 불구하고, 대한민국은 AI 시대에 경쟁력을 가질 수 있는 여러 가지 강점을 보유하고 있다. 이러한 강점을 바탕으로 전략적 선택을 해야 한다. 스마트폰 시대의 교훈을 살려, AI 시대에도 우리가 잘하는 분야에 집중할 필요가 있다.

3.1. AI 서비스에서 찾는 기회

AI 파운데이션 모델과 AI 가속 컴퓨팅 반도체 개발 뿐만 아니라 가치 사슬 전반에 걸쳐 더 넓은 시각에서 기회 요인을 찾아야 한다. 대표적으로 AI 서비스 개발에 주력하여 경쟁력을 확보하는 것이 현실적인 접근이 될 수 있다. AI 서비스 모델에 초점을 맞춰, 빠르게 변화하며 수용하는 우리의 강점을 살린 특화된 AI 서비스를 개발하고 이를 글로벌 시장으로 확장하는 전략이 필요하다.

글로벌 시장조사 전문 기업인 IoT Analytics의 분석에 따르면, 생성형 AI 시장에서 ‘데이터 센터 GPU’ 부문은 NVIDIA가 전체 시장의 92%를 차지하며 압도적인 점유율을 기록하고 있다. ‘모델 및 플랫폼’ 분야에서는 OpenAI와 Microsoft 등 글로벌 기업들이 우위를 점하고 있는 반면, ‘서비스’ 시장은 다양한 기업들이 경쟁하며 시장을 공략하고 있는 상황으로, 여전히 많은 기회가 남아 있는 영역이다.

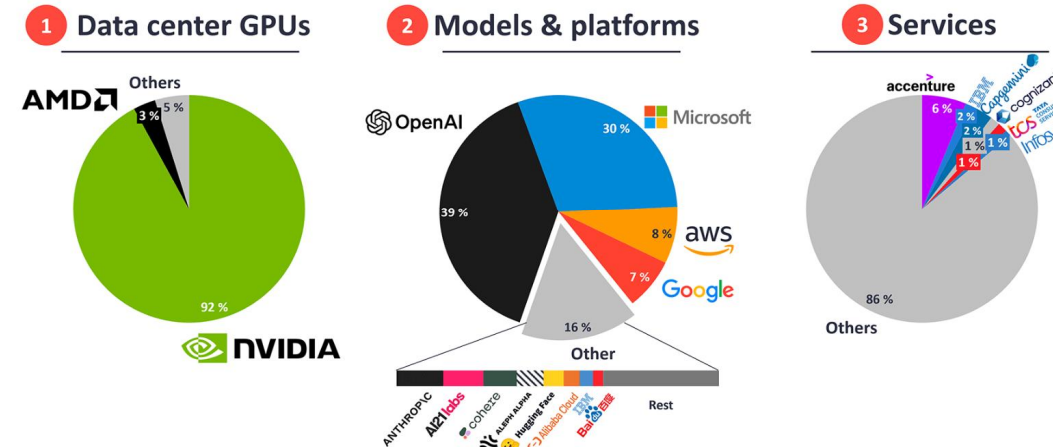
85) 스마트폰 보급률 PC 첫 추월 (임근호, 2015) <https://www.hankyung.com/article/2015012069551>

86) 2012-2018 스마트폰 사용률, 현재 사용 & 향후 구입 예정 브랜드 (한국갤럽, 2018)

<https://www.gallup.co.kr/gallupdb/reportContent.asp?seqNo=943>

[그림 54] 2023년 생성형 AI 주요 기업들의 시장 점유율⁸⁷⁾

Generative AI: Market share of leading vendors 2023



Note: Numbers are rounded and might not add up to 100%; Market share is based on 2023 market sizes (based on revenue).
Source: IoT Analytics Research 2023-Generative AI Market Report 2023-2030. We welcome republishing of images but ask for source citation with a link to the original post and company website.

※ 출처 : IoT Analytics(2023.12.), 'The leading generative AI companies'

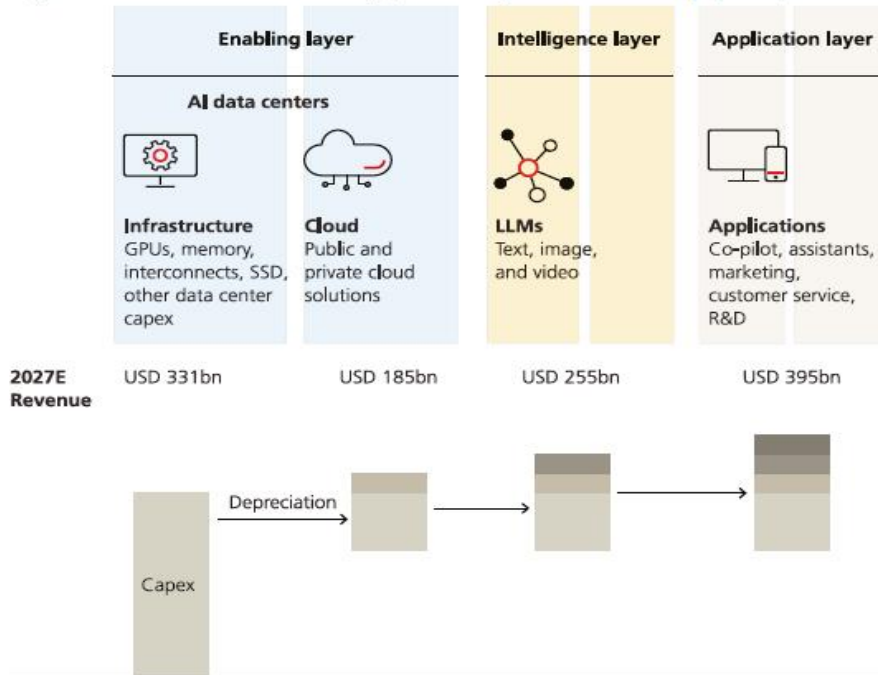
AI 서비스 부문은 시간이 지날수록 가장 큰 수익을 창출할 것으로 전망된다. 글로벌 금융기업 UBS는 '인공지능의 투자 기회 규모 파악 및 포착'이라는 보고서⁸⁸⁾에서, AI 투자 초기에는 대규모 컴퓨팅 리소스를 제공하는 AI 데이터 센터가 주요 수익원이 되지만, 시간이 지남에 따라 AI 서비스(Application Layer)가 높은 부가가치를 창출하며 가장 큰 경제적 가치를 제공할 것으로 예상했다. UBS는 2027년까지 AI 시장의 총 매출에서 애플리케이션 계층(AI 서비스)이 3,950억 달러로 가장 큰 비중을 차지할 것이라고 전망했다.

또한, UBS는 기술 노하우와 훈련 비용이 높은 장벽으로 작용하기 때문에, 기업들이 자체적으로 LLM을 구축하기보다는 이미 구축된 선도적인 LLM을 활용할 가능성이 높다고 분석했다. 이로 인해 대부분의 기회와 가능성은 인에이블링 계층(AI 데이터 센터 등) 또는 애플리케이션 계층에서 발생할 것으로 예상했다.

87) <https://iot-analytics.com/leading-generative-ai-companies/>

88) UBS(2024.6.10.), 'Artificial intelligence: Sizing and seizing the investment opportunity'

[그림 55] AI 시장의 기회



※ 출처 : UBS(2024.6.10.), 'Artificial intelligence: Sizing and seizing the investment opportunity'

현재 AI 모델 개발은 글로벌 빅테크 기업들이 선도하고 있다. 이에 따라 사용자 경험 중심의 소비자용 AI 서비스나 특정 산업에 특화된 버티컬 AI(Vertical AI) 서비스 개발에서 경쟁력을 확보할 필요가 있다. 특히, 오픈소스 AI 모델을 활용해 자연어 처리(NLP)와 컴퓨터 비전 같은 기능을 개발하고, 전이학습(Transfer Learning)을 통해 한국어와 한국 문화에 최적화된 AI 모델을 효율적으로 구축할 수 있을 것이다.

또한, AI 서비스 개발 및 배포는 클라우드 기반 플랫폼과 MLOps, 버티컬 AI(Vertical AI) 도입을 통해 점차 쉬워지고 있다. 각 산업에 최적화된 AI 솔루션이 제공되면서, AI 기술이 실질적인 가치를 창출하는 시대가 열리고 있다.

[그림 56] 특화용 버티컬AI(Vertical AI)

VERTICAL AI의 기회/과제		VERTICAL AI의 중요성	
틈새 시장 공략	• 기존에는 너무 작아서 개발사들이 관심 갖지 않았던 틈새 시장을 공략할 수 있음.	산업 특화 문제 해결	• 각 산업별 고유한 문제와 요구사항 존재 • 명확한 문제정의와 함께 가치 창출 가능
새로운 산업 영역 개척	• 기존 일반 AI로 해결하기 어렵거나 비용이 너무 높았던 분야의 문제와 가치를 다룰 수 있음.	효율성 증대	• 프로세스를 명확히 정의하여 반복적이고 시간 소모적인 자동을 자동화하여 생산성 향상 가능(VERTICAL 내 확대 가능)
데이터 확보	• AI 성능은 해당 산업의 고품질 데이터에 의존함. 하지만 많은 경우 데이터가 파편화되어 있거나 품질이 고르지 못함.	정확성 향상	• 산업 특화 데이터로 훈련된 AI는 해당 분야에서 더 높은 정확도를 보임. 금융 분야에서는 금융 사기 탐지 등 활용 가능
T자형 전문성 확보	• AI분야도 알아야 하지만, VERTICAL 산업의 전문성도 동시에 가져야 가능	새로운 가치 창출	• 일반 AI로 불가능했던 서비스 가능함.

AI 서비스들은 산업, 사회, 경제 전반에 걸쳐서 또 글로벌로 확대될 것이기 때문에, AI 서비스의 범위와 시장은 무궁무진하다. 뿐만 아니라 우리가 활용하고 있는 모바일 서비스 등 IT 서비스 전반에 AI가 적용될 것이기에, AI가 지금까지 우리가 상대적으로 약했던 소프트웨어 서비스 경쟁력을 단번에 끌어올릴 기회를 만들어 줄 수 있을 것이다.

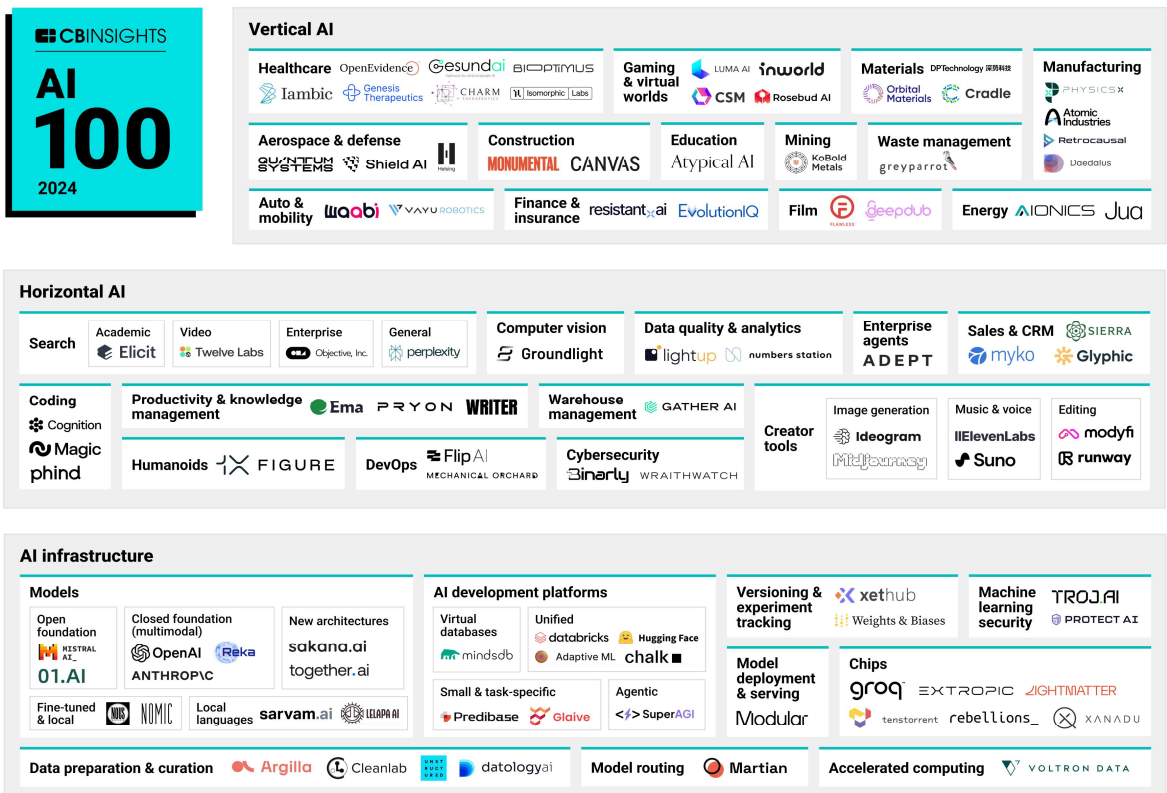
이를 위해 첫째, 무엇보다도 빠르게 시장에 출시하며 한국 내 변화를 좋아하는 초기 수용자(Early Adopter) 소비자뿐만 아니라 글로벌 소비자들과 만나며 더 고도화되어야 할 것이다. 둘째, 만약 우리가 AI (파운데이션) 모델에서 주도권을 갖지 못하더라도 AI 서비스에 주도권을 갖고 AI 모델은 필요에 따라서 선택할 수 있도록 하는 준비되어야 할 것이다. 이는 모바일 서비스가 안드로이드OS와 iOS 상관없이 두 OS 기반의 서비스를 제공하는 것과 같은 맥락이다.

3.2. AI 하드웨어 경쟁력 강화

2024년 CB Insights가 발표한 '전 세계 유망 AI 스타트업 100대 기업'(AI 100) 보고서⁸⁹⁾에 따르면, 100개 기업 중 69개가 미국에 본사를 두고 있다. 또한, 100개 기업 중 미국 기업이 총 펀딩 금액의 90.3%, 기업 평가 가치의 93.0%를 차지하며 압도적인 우위를 보이고 있다. 그러나 미국 외 국가들은 전체 스타트업의 3% 미만을 보유하며, 글로벌 AI 생태계에서의 비중은 낮지만, 기술 차별화와 특화 전략을 통해 경쟁력을 점차 높이고 있다.

우리나라는 리벨리온(Rebellions)이 AI 100대 기업 중 하나로 선정되었다. AI 반도체와 컴파일러 설계로 주목받는 리벨리온은 총 펀딩 금액 2억 1,931만 달러를 유치했으며, 기업 가치는 5억 9,884만 달러에 달해 글로벌 기업으로 도약하고 있다. 이는 한국 AI 스타트업이 AI 반도체와 같은 첨단 기술 분야에서 경쟁력을 발휘할 수 있음을 보여주는 사례이다.

[그림 57] AI 100: 2024년 가장 유망한 인공지능 스타트업



※ 출처 : CB Insights(2024.4.), 'AI 100'

89) <https://www.cbinsights.com/research/report/artificial-intelligence-top-startups-2024/>

대한민국은 이미 메모리 반도체와 디스플레이, 카메라 등 부품 산업에서 세계적인 경쟁력을 보유하고 있다. 특히 반도체용 메모리 분야에서는 삼성전자와 SK하이닉스가 세계 시장 점유율 약 69%를 차지하며 압도적인 우위⁹⁰⁾를 점하고 있다. AI 분야에서는 AI 특화 메모리로 알려진 HBM(High Bandwidth Memory)에서 거의 2023년 기준 약 90% 수준의 시장 독보적인 경쟁력을 보유⁹¹⁾하고 있으며, 스마트폰 시대에 확보한 하드웨어 경쟁력을 AI 시대에서도 유지할 수 있는 방안을 찾는 것이 관건이다.

이를 위해서는 더 크고 넓게 살펴봐야 할 것이다. 현재 AI 가속 컴퓨팅 반도체 분야는 엔비디아가 독점적인 시장 지배력과 주도권을 행사하는 상황이다. 그러나 그 외 분야에서 대한민국이 잘 하는 고객 접점의 기기에서 AI와 관련된 부품, 예를 들면 카메라, 센서 등이 있을 것이고, 또 데이터 센터까지 영역을 넓혀서 전력 기기 등까지 포함한다면 오히려 전략적 또는 정책적 선택지도 늘고 성과도 더 늘어날 수 있을 것이다.

3.3. 변화에 열광하는 국내 시장 특성 활용

대한민국 국민은 새로운 기술과 서비스에 대한 수용도가 매우 높다. 이는 혁신적인 AI 서비스를 개발하고 테스트하기에 최적의 환경을 제공한다. 예를 들어, 스마트폰, 모바일 결제, 온라인 쇼핑, 배달 서비스 등 다양한 분야에서 새로운 서비스가 빠른 속도로 확산된 것은 이러한 국민적 특성 덕분이다. 이러한 특성은 AI 서비스 혁신의 테스트베드로서 활용될 수 있다. 국내 시장에서 빠르게 검증된 AI 서비스는 글로벌 시장 진출의 발판이 될 수 있다. 실제로 이러한 전략은 이미 K-POP과 K-드라마 등의 문화 콘텐츠 분야에서 성공을 거두고 있다.

따라서 우리는 이러한 강점을 활용하여 다양한 AI 서비스를 국내에서 먼저 개발하고 테스트한 후, 이를 바탕으로 글로벌 시장에 진출하는 전략을 고려해볼 수 있다. 예를 들어, AI 기반의 개인화된 교육 서비스, AI를 활용한 헬스케어 서비스, AI 기반의 스마트시티 솔루션 등을 국내에서 먼저 상용화하고 이를 글로벌 시장으로 확장할 수 있다.

90) Home Industry insights Strategy Insights Gen AI, HPC to fuel HBM market growth Gen AI, HPC to fuel HBM market growth (Yole Group, 2024)

91) Global High Bandwidth Memory (HBM) Market Share [2024-2032] | Top Key-Players in the Industry are - SK Hynix, Samsung, Micron (Discover Global Insights, 2024)

4. 결론 : 강점을 살린 전략적 접근의 필요성

현시점에서 한국의 AI 산업은 몇 가지 주요 도전과 기회를 마주하고 있다.

첫째, AI 산업 환경은 매우 빠르게 변화하고 있으며 글로벌 빅테크 기업들은 AI 인프라, 모델, 반도체 등에서 막대한 투자를 통해 기술 격차를 벌리고 있다. 둘째, 국내 기업들은 기술력과 자본력에서 상대적으로 불리한 위치에 있지만, 스마트폰 시대에서 보여준 대한민국 국민이 기반이 된 빠른 적응력과 혁신성으로 일부 강점을 발휘할 수 있는 여지가 있다. 셋째, 국내 AI 산업이 글로벌 시장에서 자생력을 확보하기 위해서는 보다 경쟁적이고 혁신적인 환경이 조성되어야 한다.

이러한 요소들을 고려할 때, 단기적인 협력과 경쟁 촉진을 통해 장기적인 독자 생태계를 구축하는 전략이 필요하다. 이에 따라 다음 세 가지 시나리오를 제시한다.

첫 번째 시나리오 방안은 글로벌 협력을 통해 성장 속도를 가속화하는 것이다. 예를 들어, 글로벌 선도 AI 모델과 오픈소스 AI 모델을 활용하여 한국 시장의 특성과 요구에 맞는 고유한 AI 모델과 서비스를 개발하는 방안도 그 중 하나이다. 이 전략은 국내 기업들이 이미 공개된 고성능 AI 모델을 기반으로 전이학습(Transfer Learning)과 맞춤형 데이터 적용을 통해 차별화된 경쟁력을 확보할 수 있도록 지원하는 데 초점을 맞추고 있다.

우리나라 기업인 업스테이지(Upstage)가 개발한 소형 언어 모델 '솔라(Solar)'는 허깅페이스의 오픈 LLM 리더보드에서 상위권을 차지하며 세계적인 기술력을 입증했다. 업스테이지는 솔라를 기반으로 금융, 법률, 헬스케어 등 다양한 산업에 특화된 AI 솔루션을 개발하고 있으며, 이를 통해 국내외 기업들과 협력을 확대하고 있다. 특히, 아마존웹서비스(AWS)와의 파트너십을 통해 글로벌 시장에 솔라 프로(Solar Pro)를 출시⁹²⁾하며 해외 진출에도 성과를 내고 있다. 이러한 노력으로 업스테이지는 글로벌 AI 시장에서의 입지를 강화하며 성공 사례를 만들어가고 있다.

두 번째 시나리오는 국내 AI 기업들 간의 경쟁 촉발을 통한 자생적 성장 달성이다. 두 번째 시나리오는 국내 AI 기업들 간의 경쟁을 촉발함으로써 기술 혁신과 성장을 유도하는 방안이다. AI 인프라와 서비스 경쟁력은 국내 기업들 간의 자발적인 혁신과 경쟁을 통해 더욱 강화될 수 있다. 특히, AI 데이터센터와 클라우드 인프라에서 경쟁을 촉진하여 인프라의 성능을 개선하고, AI 서비스의 품질을 높이는 것이 필요하다.

92) '업스테이지, AWS에서 차세대 '솔라 프로' 제너레이티브 AI LLM 출시', 업스테이지, 2024.12.4.

<https://ko.upstage.ai/blog/press/solar-pro-aws>

예를 들어, 국내 주요 클라우드 기업들이 AI 인프라 확충에 적극적으로 참여할 수 있도록, 정부는 정책 지원을 통해 기업들의 투자 의지를 높이고 경쟁을 촉진해야 한다. 또한, AI 반도체 등 하드웨어 부문에서 국내 스타트업들이 혁신적 기술을 개발하고 시장에 진입할 수 있도록, 규제 완화와 투자 지원을 통해 글로벌 진출 기반을 탄탄히 마련하는 것이 중요하다. 이를 통해 국내 AI 산업은 자생적인 성장 동력을 확보하고, 글로벌 수준의 기술력을 갖출 수 있을 것이다.

세 번째 시나리오는 중장기적으로 경쟁력 있는 ‘소버린 AI’를 목표로 자립적인 AI 산업 기반을 구축하는 방안이다. AI 인프라, 모델, 서비스 전반에 걸쳐 독자적인 AI 산업과 가치 사슬을 형성하는 것을 목표로 한다. 단기적으로는 협력과 경쟁을 통해 성과를 도출하고, 장기적으로는 국내 AI 산업이 지속 성장할 수 있는 토대를 마련하는 것이 중요하다. 이를 위해 정부와 기업이 협력하여 R&D 투자와 정책적 지원을 강화하고, 공공 데이터 활용과 현지화된 AI 모델 구축을 장려해야 한다.

또한, AI 하드웨어 분야에서도 단순히 반도체에 국한되지 않고 엣지 디바이스와 데이터 센터용 부품 등 다양한 AI 기술로 확장하여 글로벌 빅테크의 독점적 지위를 벗어나 다각적인 경쟁력을 확보할 필요가 있다. 이를 통해 한국은 독자적인 AI 생태계를 구축하고, 글로벌 AI 시장에서의 경쟁력을 확보할 수 있을 것이다.

결론적으로, 한국이 AI 산업에서 경쟁력을 강화하고 글로벌 시장에서 선도적인 위치를 차지하기 위해서는 단기적으로 글로벌 기업들과의 협력과 경쟁을 통해 성장을 도모하고, 중장기적으로는 독자적인 소버린 AI 생태계를 구축하는 방향으로 나아가야 한다. 이를 통해 한국은 글로벌 AI 시장에서 자생력을 확보하고, 미래 산업의 선도 국가로 자리매김할 수 있을 것이다.

【 그림 58 】 AI 산업 경쟁력 강화 시나리오

시나리오 ①	【 글로벌 협력 강화를 통한 성장 가속화 】 • 글로벌 AI 모델과 기술을 신속히 도입하고, 이를 통해 국내 AI 산업이 글로벌 AI 서비스 시장을 선점하고 초기 기반을 확립할 수 있도록 지원하는 방안
시나리오 ②	【 국내 AI 기업의 자생적 성장 환경 구축 】 • 국내 AI 기업들 간의 자생적 경쟁력을 강화하고 산업 생태계를 확장하기 위해, 국내 데이터 센터 기업, 반도체 스타트업, 파운데이션 모델 기업들의 경쟁을 촉진하여 기술 혁신과 성장을 유도하는 방안
시나리오 ③	【 중장기적 ‘소버린 AI’ 달성을 위한 자립적 AI 산업 경쟁력 구축 】 • AI 인프라, AI 모델, AI 서비스를 모두 아우르는 독자적인 소버린 AI 산업과 가치 사슬을 구축하여 대한민국이 AI 주권을 확보하는 방안

〈참고 자료〉

1. Aditya RameshPavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark ChenMikhail. (2021). “Zero-Shot Text-to-Image Generation.” arXiv.
2. AI云原生智能算力架构. (2024년 04월 02일). “国内主流AI 大模型架构及应用场景深度分析 2024.” 53ai: <https://www.53ai.com/news/qianyanjishu/382.html>에서 검색됨
3. Alexandre. (2024년 06월 25일). “Cloud market share 2024 - AWS, Azure, GCP growth fueled by AI.” holori: <https://holori.com/cloud-market-share-2024-aws-azure-gcp/>에서 검색됨
4. Alexandre. (2024). “Top Cloud Providers in 2024 - Hyperscalers and Alternative vendors.” holori: <https://holori.com/top-cloud-providers-in-2024/>에서 검색됨
5. AlsopThomas. (2024년 08월 29일). “Data center segment revenue of Nvidia, AMD, and Intel from 2021 to 2024, by quarter.” Statista: <https://www.statista.com/statistics/1425087/data-center-segment-revenue-nvidia-amd-intel/>에서 검색됨
6. APPLE. (2024년 06월 10일). “Introducing Apple’s On-Device and Server Foundation Models”. Machine Learning Research: <https://machinelearning.apple.com/research/introducing-apple-foundation-models>에서 검색됨
7. Artificial Intelligence (AI). (2023년 12월 11일). “Meta and Microsoft Lead Demand for NVIDIA’s Powerful H100 AI Chips.” Spearhead: <https://spearhead.so/meta-and-microsoft-lead-demand-for-nvidias-powerful-h100-ai-chips/>에서 검색됨
8. Ashish VaswaniShazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia PolosukhinNoam. (2017). “Attention Is All You Need.” ARXIV.
9. Bay Area Times. (2023년 09월 27일). “OpenAI in talks to raise at \$80B - \$90B valuation, expecting \$1B in revenue in 2023”. Bay Area Times: <https://www.bayareatimes.com/p/openai-talks-raise-80b-90b-valuation-expecting-1b-revenue-2023>에서 검색됨
10. Ben WodeckiYaoDeborah. (2024년 02월 27일). “Google DeepMind CEO on AGI, OpenAI and Beyond - MWC 2024”. AIBUSINESS: <https://aibusiness.com/nlp/google-deepmind-ceo-on-agi-openai-and-beyond-mwc-2024#close-modal>에서 검색됨
11. BrownSara. (2022년 06월 07일). “Why it’s time for ‘data-centric artificial intelligence’”. MIT Management: <https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence>에서 검색됨
12. Bruce D. SoklerHecht, Christian Tamotsu Fjeld, Raj GambhirAlexander. (2023년 06월 08일). “National Priorities for Artificial Intelligence — AI: The Washington Report.” MINTZ: <https://www.mintz.com/insights-center/viewpoints/2191/2023-06-07-national-priorities-artificial-intelligence-ai>에서 검색됨
13. ChengMichelle. (2024년 01월 30일). “Nvidia’s biggest customers are also the AI chip maker’s biggest threat”. QUARTZ: <https://qz.com/nvidia-generative-ai-google-microsoft-meta-1851206854>에서 검색됨
14. congress.gov. (2019년 04월 10일). “Algorithmic Accountability Act of 2019”. congress.gov: <https://www.congress.gov/bill/116th-congress/house-bill/2231>에서 검색됨
15. Congress.gov. (2022년 12월 30일). “American Data Privacy and Protection Act”. Congress.gov: <https://www.congress.gov/bill/117th-congress/house-bill/8152>에서 검색됨
16. Congress.gov. (2022년 12월 15일). “Children and Teens’ Online Privacy Protection Act”. Congress.gov: <https://www.congress.gov/bill/117th-congress/senate-bill/1628/text>에서 검색됨
17. DeutscherMaria. (2024년 10월 11일). “Healthcare AI startup Abridge reportedly raising \$250M at \$2.5B valuation.” siliconANGLE: <https://siliconangle.com/2024/10/11/healthcare-ai-startup-abridge-reportedly-raising-250m-2-5b-valuation/>에서 검색됨
18. DIGITIMES Asia. (2024년 03월 07일). “SK Hynix invests US\$1 Billion in key AI memory chip technology.” DIGITIMES Asia: <https://www.digitimes.com/news/a20240307VL206.html&chid=9>에서 검색됨

19. Discover Global Insights. (2024년 07월 01일). “Global High Bandwidth Memory (HBM) Market Share [2024–2032] | Top Key-Players in the Industry are – SK Hynix, Samsung, Micron.” LinkedIn: <https://www.linkedin.com/pulse/global-high-bandwidth-memory-hbm-market-share-dyvbc/>에서 검색됨
20. etc.Schuhmann and Christoph. (2022). “LAION-5B: An open large-scale dataset for training next generation image-text models.” NeurIPS 2022.
21. EvansonNick. (2024년 07월 29일). “Report claims that OpenAI has burned through \$8.5 billion on AI training and staffing, and could be on track to make a \$5 billion loss.” PCGAMER: <https://finance.yahoo.com/news/report-claims-openai-burned-8-105046378.html>에서 검색됨
22. FlaningamEric. (2024년 09월 08일). “The Current State of AI Markets.” Generative Value: <https://www.generativevalue.com/p/the-current-state-of-ai-markets>에서 검색됨
23. Google AI. (2024). “Learn about our leading AI models.” Google AI: <https://ai.google/discover/our-models/>에서 검색됨
24. Google Cloud. (2024년 05월 15일). “Announcing Trillium, the sixth generation of Google Cloud TPU.” Google Cloud: <https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus?hl=en>에서 검색됨
25. Google Cloud. (2024년 10월). “Google Cloud TPU로 AI 개발 가속화.” Google Cloud: <https://cloud.google.com/tpu?hl=ko#features>에서 검색됨
26. GoswamiRohan. (2024년 06월 21일). “Apple Intelligence won’t launch in EU in 2024 due to antitrust regulation, company says.” CNBC: <https://www.cnbc.com/2024/06/21/apple-ai-europe-dma-macos.html>에서 검색됨
27. HAI. (2024). “Artificial Intelligence Index Report 2024.” Stanford.
28. HassabisDemis. (2024년 02월 26일). “Keynote 3: Our AI Future”. MWC: <https://www.mwcbarcelona.com/agenda/sessions/4665-keynote-3-our-ai-future>에서 검색됨
29. HeavenDouglasWill. (2023년 03월 14일). “GPT-4 is bigger and better than ChatGPT—but OpenAI won’t say why.” MIT Technology Review: <https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/>에서 검색됨
30. Hugging Face. (2024). “Hugging Face Llama variation”. Hugging Face: <https://huggingface.co/models?other=llama>에서 검색됨
31. Kevin LeeGangidi, Mathew OldhamAdi. (2024년 03월 12일). “Building Meta’s GenAI Infrastructure”. Engineering at Meta: <https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/>에서 검색됨
32. KosinskiMatt. (2024년 04월 08일). “EU AI 법이란 무엇인가요?” IBM: <https://www.ibm.com/kr-ko/topics/eu-ai-act>에서 검색됨
33. KroetCynthia. (2024년 07월 18일). “Meta stops EU roll-out of AI model due to regulatory concerns.” euro news: <https://www.euronews.com/next/2024/07/18/meta-stops-eu-roll-out-of-ai-model-due-to-regulatory-concerns>에서 검색됨
34. LundenIngrid. (2024년 05월 13일). “Anthropic is expanding to Europe and raising more money”. Techcrunch: https://techcrunch.com/2024/05/13/anthropic-is-expanding-to-europe-and-raising-more-money/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAADmODcQrmckbIUquej1bDLApMG1wgnRt8SViVtwlY-gsQ0E1IKV9T4BhZV30J2O9GfVgOQxx9umlX1xaGnA에서 검색됨
35. META AI. (2023년 11월 30일). “Celebrating 10 years of FAIR: A decade of advancing the state-of-the-art through open research.” META: <https://ai.meta.com/blog/fair-10-year-anniversary-open-science-meta/>에서 검색됨
36. Meta AI. (2023년 05월 18일). “Reimagining Meta’s infrastructure for the AI age.” Meta AI: <https://ai.meta.com/blog/meta-ai-infrastructure-overview/>에서 검색됨
37. META AI. (2024년 07월 31일). “Transforming our infrastructure for the next generation of AI.” META AI: <https://ai.meta.com/infrastructure/>에서 검색됨
38. NVIDIA. (2024). “NVIDIA GB200 NVL72.” NVIDIA: <https://www.nvidia.com/en-us/data-center/gb200-nvl72/>에서 검색됨

39. NVIDIA Korea. (2024년 03월 19일). “NVIDIA Blackwell 플랫폼, 새로운 컴퓨팅 시대를 열다.” NVIDIA Korea: <https://blogs.nvidia.co.kr/blog/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing/>에서 검색됨
40. Omar SansevieroTunstall, Philipp Schmid, Sourab Mangrulkar, Younes Belkada, Pedro CuencaLewis. (2023년 12월 11일). “Mixture of Experts Explained.” Hugging Face: <https://huggingface.co/blog/moe>에서 검색됨
41. Page 21 Team. (2023년 08월 04일). “How Scale AI Became a \$7 Billion AI Data Powerhouse: Business Model Breakdown.” Page21: https://www.pagetwentyone.com/post/how-scale-ai-became-a-7-billion-ai-data-powerhouse-business-model-breakdown#google_vignette에서 검색됨
42. pageO'DonnellarchiveJames. (2024년 03월 11일). “LLMs become more covertly racist with human intervention.” MIT Technology Review: <https://www.technologyreview.com/2024/03/11/1089683/llms-become-more-covertly-racist-with-human-intervention/>에서 검색됨
43. ParkerLynne. (2022). “National Artificial Intelligence Initiative.” USPTO.
44. Perplexity. (2024). “Perplexity.” adweek.
45. Reuter. (2024년 02월 17일). “OpenAI valued at \$80 billion after deal, NYT reports”. Reuter: <https://www.reuters.com/technology/openai-valued-80-billion-after-deal-nyt-reports-2024-02-16/>에서 검색됨
46. RoachJohn. (2023년 03월 13일). “How Microsoft’s bet on Azure unlocked an AI revolution.” Microsoft: <https://news.microsoft.com/source/features/ai/how-microsofts-bet-on-azure-unlocked-an-ai-revolution/>에서 검색됨
47. RobuckMike. (2024년 10월 03일). “OpenAI scoops up \$6.6B in funding round at \$157B valuation.” mobileworldlive: https://www.mobileworldlive.com/ai-cloud/openai-scoops-up-6-6b-in-funding-round-at-157b-valuation/?ID=a6g1r000000yCNEAA2&JobID=2033213&utm_source=sfmc&utm_medium=email&utm_campaign=MWL_20241003&utm_content=https%3a%2f%2fwww.mobileworldlive.com%2fai-cloud%2f에서 검색됨
48. SatarianoAdam. (2024년 06월 11일). “Mistral, a French A.I. Start-Up, Is Valued at \$6.2 Billion”. The New York Times: <https://www.nytimes.com/2024/06/11/business/mistral-artificial-intelligence-fundraising.html>에서 검색됨
49. Shirin GhaffaryRoof, Rachel Metz and Dina BassKatie. (2004년 10월 04일). “OpenAI Raises \$6.6 Billion in Funds at \$157 Billion Value.” Bloomberg: <https://finance.yahoo.com/news/openai-closed-funding-round-raising-161842157.html>에서 검색됨
50. stability.ai. (2024년 08월 01일). “Introducing Stable Fast 3D: Rapid 3D Asset Generation From Single Images.” stability.ai: <https://stability.ai/news/introducing-stable-fast-3d>에서 검색됨
51. stability.ai. (2024년 07월 24일). “Introducing Stable Video 4D, Our Latest AI Model for Dynamic Multi-Angle Video Generation.” stability.ai: <https://stability.ai/news/stable-video-4d>에서 검색됨
52. Statista Market Insights. (2024년 05월). “Artificial Intelligence – Worldwide.” Statista: <https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide#market-size>에서 검색됨
53. Team Counterpoint. (2022년 07월 29일). “US Chips Act Takes New Form Before August Recess, Leaves Some Unhappy.” Counterpoint Research: <https://www.counterpointresearch.com/insights/us-chips-act-takes-new-form-before-august-recess-leaves-some-unhappy/>에서 검색됨
54. Team TBH. (2023년 07월 08일). “Scale AI – Founding Story, Features, Business Model and Growth.” The Brand Hopper: https://thebrandhopper.com/2023/07/08/scale-ai-founding-story-features-business-model-and-growth/#google_vignette에서 검색됨
55. The Economic Times Tech. (2024년 03월 30일). “Microsoft and OpenAI planning \$100 billion data center project: report”. The Economic Times Tech:

- <https://economictimes.indiatimes.com/tech/technology/microsoft-and-openai-planning-100-billion-data-center-project-report/articleshow/108883808.cms?from=mdr>에서 검색됨
56. TrujilloArreolaCarlos. (2023년 04월 20일). "Meta Releases Open Source Segment Anything Model (SAM)." LinkedIn: <https://www.linkedin.com/pulse/meta-releases-open-source-segment-anything-model-sam-carlos/>에서 검색됨
 57. White House. (2022). "Blueprint for an AI Bill of Rights." White House.
 58. WHITE HOUSE. (2023). "NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN 2023UPDATE." WHITE HOUSE.
 59. WiggersKyle. (2024년 03월 12일). "Axion Ray's AI attempts to detect product flaws to prevent recalls." TECHCRUNCH: <https://techcrunch.com/2024/03/12/axion-rays-ai-attempts-to-detect-product-flaws-to-prevent-recalls/>에서 검색됨
 60. WiggersKyle. (2024년 07월 23일). "OpenAI-backed legal tech startup Harvey raises \$100M." TECHCRUNCH: <https://techcrunch.com/2024/07/23/openai-backed-legaltech-startup-harvey-raises-100m/>에서 검색됨
 61. Xi | JiangLiu, Aaron Gember-Jacobson, Arjun Nitin Bhagoji, PaulShinan. (2024). "NetDiffusion: Network Data Augmentation Through." arXiv.
 62. Yole Group. (2024년 02월 07일). "Home Industry insights Strategy Insights Gen AI, HPC to fuel HBM market growth Gen AI, HPC to fuel HBM market growth." Yole Group: <https://www.yolegroup.com/strategy-insights/gen-ai-hpc-to-fuel-hbm-market-growth/>에서 검색됨
 63. 강경주. (2024년 05월 17일). "엔비디아 GPU보다 2배 빠르다...‘과물칩’ NPU 베풀한 리벨리온." 한경: <https://www.hankyung.com/article/2024051761381>에서 검색됨
 64. 고은이황동진. (2024년 05월 02일). "오픈AI 초봉 12억 vs 한국 2억...머스크 "가장 미친 인재 전쟁"". 한국경제신문: <https://www.hankyung.com/article/2024050231001>에서 검색됨
 65. 과학기술정보통신부. (2019년 12월 17일). "인공지능(AI) 국가전략 발표". 과학기술정보통신부: <https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=112&pageIndex=1&bbsSeqNo=94&nttSeqNo=2405727&searchOpt=ALL&searchTxt=%EA%B5%AD%EA%B0%80%EC%A0%84%EB%9E%B5>에서 검색됨
 66. 권정혁. (2024년 04월). "GN+: Cohere의 Command R+ - 비즈니스를 위해 구축된 확장 가능한 LLM." Geeknews: <https://news.hada.io/topic?id=14169>에서 검색됨
 67. 김경필. (2023년 03월 08일). "[단독] AI에게 '한국 일자리 미래' 예측시켜봤더니." 조선일보: <https://www.chosun.com/national/labor/2023/03/08/BYJR3HJ47ZDLHIYXZLORYBYJJM/>에서 검색됨
 68. 김상윤. (2024년 07월 25일). "과잉투자가 과소투자보다 낫다?...커져가는 AI 투자회의론". NATE: <https://news.nate.com/view/20240725n31297>에서 검색됨
 69. 김우용. (2023년 03월 27일). "올해 국내 AI 시장 규모 2조6천억원." zdnet: <https://zdnet.co.kr/view/?no=20230427153507>에서 검색됨
 70. 김은광. (2024년 02월 28일). "다윗 '미스트랄', 골리앗 오픈AI에 도전." 내일 신문: <https://www.naeil.com/news/read/502688>에서 검색됨
 71. 박영진. (2022). "AI 학습용 데이터 플랫폼과 표준화 동향." 한국정보통신기술협회.
 72. 박의명. (2024년 06월 03일). "'AI 혁명의 관문은 컴퓨팅...대만이 시장 선도.'" 한경: <https://www.hankyung.com/article/2024060328221>에서 검색됨
 73. 박초화. (2023). "중국 반도체 국산화." 대신증권.
 74. 삼성KPMG 경제연구원. (2024). "혁신의 부스터 AI에 물드는 금융."
 75. 삼성전자. (2023년 11월 08일). "삼성전자, '삼성 AI 포럼'서 자체 개발 생성형 AI '삼성 가우스' 공개." 삼성전자 뉴스룸: <https://news.samsung.com/kr/%EC%82%BC%EC%84%B1%EC%A0%84%EC%9E%90-%EC%82%BC%EC%84%B1-ai-%ED%8F%AC%EB%9F%BC%EC%84%9C-%EC%9E%90%EC%B2%B4-%EA%B0%9C%EB%B0%9C-%EC%83%9D%EC%84%B1%ED%98%95-ai-%EC%82%BC%EC%84%B1-%EA%B0%80>에서 검색됨

76. 삼성전자. (2024년 04월 18일). “[인터뷰] AI 시대 삼성전자 HBM이 만들어 내는 완벽한 하모니.” 삼성전자:
<https://semiconductor.samsung.com/kr/news-events/tech-blog/the-perfect-harmony-created-by-samsung-hbm-powering-the-ai-era/>에서 검색됨
77. 삼성전자. (2024년 06월 13일). “삼성전자, 파운드리 포럼 2024 개최 AI 시대 파운드리 비전 제시.” 삼성전자:
<https://semiconductor.samsung.com/kr/news-events/news/samsung-showcases-ai-era-vision-and-latest-foundry-technologies-at-sff-2024/>에서 검색됨
78. 소프트웨어정책연구소. (2024). “2023년 소프트웨어산업 연간보고서.” 소프트웨어정책연구소.
79. 신동형. (날짜 정보 없음). “‘24년 소비자 AI앱 시장 동향:성장 동인과 미래 전망.” 네이버 블로그:
<https://blog.naver.com/jack0604/223581159322>에서 검색됨
80. 신동형. (2024년 06월 25일). “AI(Claude3)가 작성한 「EU의 AI 규제 강화와 빅테크의 대응:Meta와 Apple 중심으로」보고서”. 네이버:
<https://blog.naver.com/jack0604/223490305572>에서 검색됨
81. 신동형. (2024년 07월 11일). “갤럭시 언팩 2024」보고서-폴더블과 AI 기술의 융합으로 모바일 경험의 새 지평을 열다.” NAVER BLOG:
<https://blog.naver.com/jack0604/223508740038>에서 검색됨
82. 신동형. (2024년 08월 29일). “메타의 유럽 AI 혁신 가속화:규제 개선과 오픈소스 AI의 잠재력 활용”. 네이버:
<https://blog.naver.com/jack0604/223563717112>에서 검색됨
83. 신동형. (날짜 정보 없음). “2024 컴퓨텍스 기조연설로 본 엔비디아의 미래 비전과 전략, 「엔비디아, AI 시대를 이끄는 ‘게임 체인저’로 부상,」” 네이버 블로그: <https://blog.naver.com/jack0604/223478992442>에서 검색됨
84. 신동형. (날짜 정보 없음). “AI 평가 체계 대전환을 향한 엔트로픽의 도전:한계 극복과 신뢰 확보의 과제.” 네이버 블로그:
<https://blog.naver.com/jack0604/223499501997>에서 검색됨
85. 신동형. (날짜 정보 없음). “AI(Claude3)가 작성한 「Intel의 AI 시대 도전과 전략」보고서.” 네이버 블로그:
<https://blog.naver.com/jack0604/223489055131>에서 검색됨
86. 신동형. (날짜 정보 없음). “AI(Claude3)가 작성한 인텔, AI 시대를 선도하는 기술 혁신과 비전.” 네이버 블로그:
<https://blog.naver.com/jack0604/223413081950>에서 검색됨
87. 신동형. (날짜 정보 없음). “AI(Claude3)가 작성한, OpenAI의 GPT-4o 공개, 멀티 모달 AI 혁명의 신호탄.” 네이버 블로그:
<https://blog.naver.com/jack0604/223446062901>에서 검색됨
88. 신동형. (날짜 정보 없음). “AMD, AI 시대 컴퓨팅 혁신으로 지능화 가속화.” 네이버 블로그:
<https://blog.naver.com/jack0604/223482899387>에서 검색됨
89. 신동형. (날짜 정보 없음). “Arm, AI 컴퓨팅의 미래를 향한 비상(飛上).” 네이버 블로그:
<https://blog.naver.com/jack0604/223484036699>에서 검색됨
90. 신동형. (날짜 정보 없음). “Meta AI:일상을 혁신하는 지능형 비서의 진화와 Meta의 전략.” 네이버 블로그:
<https://blog.naver.com/jack0604/223600891442>에서 검색됨
91. 신동형. (날짜 정보 없음). “OpenAI o1:AI의 새로운 패러다임, 추론 중심 접근의 혁명.” 네이버 블로그:
<https://blog.naver.com/jack0604/223582644398>에서 검색됨
92. 신동형. (날짜 정보 없음). “SAM 2: 이미지와 비디오의 경계를 넘는 혁신적 AI 분할 모델.” 네이버 블로그:
<https://blog.naver.com/jack0604/223533671731>에서 검색됨
93. 신동형. (날짜 정보 없음). “Vertical AI :SaaS의 미래와 산업 혁신의 새로운 물결.” 네이버 블로그:
<https://blog.naver.com/jack0604/223580529910>에서 검색됨
94. 신동형. (날짜 정보 없음). “구글의 AlphaChip과 TPU 전략 분석.” 네이버 블로그:
<https://blog.naver.com/jack0604/223622005954>에서 검색됨
95. 신동형. (날짜 정보 없음). “메타 라마3.1(Llama 3.1) 공개로 보는 오픈소스 AI 미래.” 네이버 블로그:
<https://blog.naver.com/jack0604/223523414984>에서 검색됨

96. 신동형. (날짜 정보 없음). “전년동기 대비 10배, 총3.5억 D/L의 라마(Llama):메타의 오픈소스 모델 혁신을 가속화하다.” 네이버: <https://blog.naver.com/jack0604/223567946580>에서 검색됨
97. 연합뉴스. (2024년 06월 13일). “파운드리 1위” TSMC 1분기 점유율 61.7%…11% 삼성과 격차 확대.” 한국무역협회: https://www.kita.net/board/totalTradeNews/totalTradeNewsDetail.do?SESSIONID_KITA=EEC88EA1FC91C8F491335326FA2B6AEB.Hyper?no=84282&siteId=2에서 검색됨
98. 이덕주. (2024년 06월 06일). “스탠퍼드대 AI연구소 “GPU 부족한 대학에 AI연구 기회 줘야.”. 미라클아이: <https://www.mk.co.kr/news/society/11034698>에서 검색됨
99. 이종호. (2023년 09월 14일). “대한민국 인공지능 도약방안 발표”. 정책브리핑: <https://www.korea.kr/briefing/policyBriefingView.do?newsId=156590066>에서 검색됨
100. 임대준. (2023년 08월 31일). “메타, AI 학습 데이터 안전장치 마련…양식 제출 안 하면 ‘자동 동역.’” 시타임즈: <https://www.aitimes.com/news/articleView.html?idxno=153266>에서 검색됨
101. 임대준. (2024년 04월 02일). “‘2년 내 LLM 학습 데이터 고갈…데이터 문제로 AI 발전 중단될 것’.” 시타임즈: <https://www.aitimes.com/news/articleView.html?idxno=158463>에서 검색됨
102. 전남혁. (2024년 05월 03일). “‘시칩’ 살 돈 없어… 구형 게임칩으로 연구하는 대학들.” 동아일보: <https://www.donga.com/news/Society/article/all/20240503/124768210/1>에서 검색됨
103. 정성진. (2024년 07월 19일). “네이버 최수연, ‘AI 주권’ 강조…“소버린 AI 확산 위해 노력.” SBS 뉴스: https://news.sbs.co.kr/news/endPage.do?news_id=N1007729276에서 검색됨
104. 하준. (2024년 02월 27일). “[MWC 24] ‘알파고의 아버지’ 데미스 하사비스 “5년 후 AI 기기는 모바일 아닐 것””. TECH M: <https://www.techm.kr/news/articleView.html?idxno=120836>에서 검색됨
105. 현대인. (2024년 09월 26일). “오픈AI, 영리법인 관할 형태로 전환 추진.” 전자신문: <https://www.etnews.com/20240926000405>에서 검색됨
106. 황순민. (2024년 07월 04일). “[단독] 유럽 AI스타트업 네이버, 자본투자.” 매일경제: <https://www.mk.co.kr/news/it/11059112>에서 검색됨
107. 황정수. (2024년 07월 13일). “‘우리엔겐 ‘이것’이 있다’…삼성전자의 차세대 ‘비밀 병기’ [황정수의 반도체 이슈 짚어보기].” 한경: <https://www.hankyung.com/article/202407131340i>에서 검색됨
108. Statista(2023.7), ‘Insights Compass 2023 – Unleashing Artificial Intelligence’s true potential’
109. McKinsey & Company(2024.11), ‘Supercharging product portfolio performance with generative AI’
110. ‘OpenAI scoops up \$6.6B in funding round at \$157B valuation’, Mobile World Live, 2024.10.3. <https://www.mobileworldlive.com/ai-cloud/openai-scoops-up-6-6b-in-funding-round-at-157b-valuation/>
111. Anthropic is expanding to Europe and raising more money, TechCrunch, 2024.5.13. <https://techcrunch.com/2024/05/13/anthropic-is-expanding-to-europe-and-raising-more-money/>
112. 빅테크는 AI에 얼마를 쏟아붓고 있을까, 디지털투데이, 2024.07.31.
113. Alphabet(2024. 7. 23), ‘2024 Q2 Earnings Call’, <https://abc.xyz/2024-q2-earnings-call>
114. <https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide>
115. <https://www.statista.com/outlook/tmo/artificial-intelligence/south-korea>
116. IoT Analytics(2023.12.), ‘The leading generative AI companies’
117. UBS(2024.6.10.), ‘Artificial intelligence: Sizing and seizing the investment opportunity’
118. CB Insights(2024.4.), ‘AI 100’ <https://www.cbinsights.com/research/report/artificial-intelligence-top-startups-2024/>
119. ‘업스테이지, AWS에서 차세대 ‘솔라 프로’ 제너레이티브 AI LLM 출시’, 업스테이지, 2024.12.4. <https://ko.upstage.ai/blog/press/solar-pro-aws>

**THE
AI
REPORT
2024**

NIA 한국지능정보사회진흥원