2025-04-14

SK증권 해외주식

다음 AI 주인공은 플랫폼

해외주식. 박제민 | jeminwa@sks.co.kr

SK securities

해외주식

다음 AI 주인공은 플랫폼

SK증권 리서치센터

Analyst
박제민

jeminwa@sks.co.kr
3773-8884

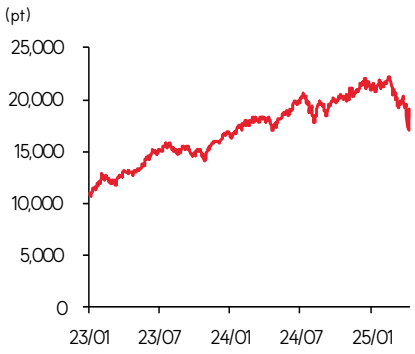
AI 제품화 시대의 도래

생성형 AI 제품화 시대 개화가 목전이다. 2024 년 12 월 기준 엔비디아 GPU 의 60%가 훈련에, 40%만 추론에 사용됐다. 추론 중 대부분은 생성형 AI 가 아닌 기존 기계 학습에 사용됐다. 생성형 AI 서비스를 위한 추론 소요는 아직 미미하다.

올해는 다르다. 1) AI 진화를 위한 기술적 가시성이 확보됐고 2) 추론 비용 하락이 예상된다. 하드웨어, 소프트웨어 발전으로 AI 제공 가격은 매년 8 배 이상 하락할 예정이다. 비용 부담으로 제한됐던 AI 서비스들의 원가가 매년 8 배 하락하는 셈이다.

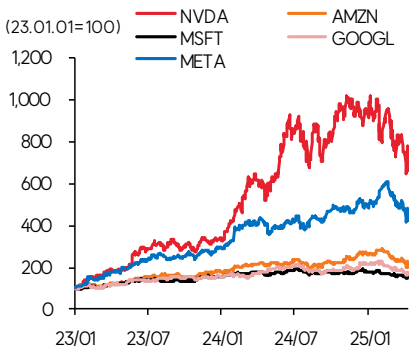
최근 OpenAI 는 ChatGPT 에 img2img 기능을 추가하며, 일명 지브리 파동을 불러 일으켰다. 1,500 만명이던 유료 사용자는 1 주일만에 500 만명 이상 늘었다. 무료 사용자를 포함하는 MAU 는 3 개월만에 2 억명 증가해 3 월에는 5 억명을 기록했다. 신기술의 발견이나 AI 모델의 진화로 성과를 낸 것이 아니다. 현재 기술을 잘 '제품화'하였고, '추론 비용 하락'으로 서비스 단가가 현실화됐다.

나스닥 주가 추이



자료: Bloomberg, SK 증권

작성 기업 주가 추이



자료: Bloomberg, SK 증권

제품화 시대, 다시 빅테크가 주도

지난 AI 사이클에서는 인프라 기업들이 매출과 주가 상승률이 가장 강했다. 그러나 앞으로의 AI 사이클은 사용자 기업 중심으로 전개될 전망이다. 제품화 초기에는 AI 모델, 인프라 조율 능력, 관련 인력, 자본력, 접근성 좋은 플랫폼을 가진 플레이어가 유리하다. 모든 것을 갖춘 빅테크가 이번에도 주역으로 떠오를 가능성이 높다.

플랫폼 사업자 주목: Meta, Alphabet, Nvidia 선호

제품화 이후 쉽게 소비자에게 노출시킬 수 있는 B2C 플랫폼 사업자들에 주목한다.

Meta 는 추론 비용 하락으로 기대되는 제품화가 많다. 개인화된 광고, 사용자 이미지 합성 광고 등을 준비 중이다. 아직 구현된 서비스의 형태는 미미하지만 Llama 4 의 Frontier 모델 달성, 매출 규모 대비 강한 Capex 집행으로 추론 비용을 적극 낮추는 중이다. 인스타그램에서 내 얼굴이 들어간 영상 광고를 볼 날이 멀지 않았다.

Alphabet 또한 5 억명 이상의 MAU 를 가진 12 개의 플랫폼, 현재 Frontier 모델인 Gemini 경쟁력, TPU 경쟁력 등을 바탕으로 많은 제품들의 출시가 기대된다.

Nvidia 는 제품화 시대와 함께 폭증하는 추론 수요의 수혜를 다시 맞이할 예정이다. 제품 출시 속도, CUDA 경쟁력 등을 고려할 때 아직 가장 안전한 선택지이다. AI 가 고도화되고 추론 비용이 하락할 수록 수요가 늘어난다는 것을, 시장은 딥시크 사태 때 확인했다. 나아가 Dynamo SW, 랙 당 GPU 탑재량 증가, 통신 기술력 증가 등으로 부품사에서 인프라 사업자로 지위가 강화되는 중이다.

Contents

들어가며	3
Key Charts	4
1. AI 산업의 3 가지 목표: 진화, 비용 감소, 제품화	7
2. AI 제품화 시대의 도래	9
3. AI 모델의 진화: 신규 Scaling Law 등장	17
4. 추론 비용 감소: 연간 8 배 속도의 하방 압력	24
5. 제품화 시대, 다시 빅테크가 주도	30
Appendix: AI 모델 기본기	35
6. 기업분석	
1) 메타 플랫폼스(META): AI 수혜를 온몸으로 받는 중	53
2) 알파벳(GOOGL): AI 강자의 저평가 구간	61
3) 엔비디아(NVDA): 영역을 확장하는 AI 인프라 제왕	69
4) 마이크로소프트(MSFT): 큰 결실을 위한 시간이 아직 필요	77
5) 아마존(AMZN): AI 활용 핵심은 비용절감	84

들어가며

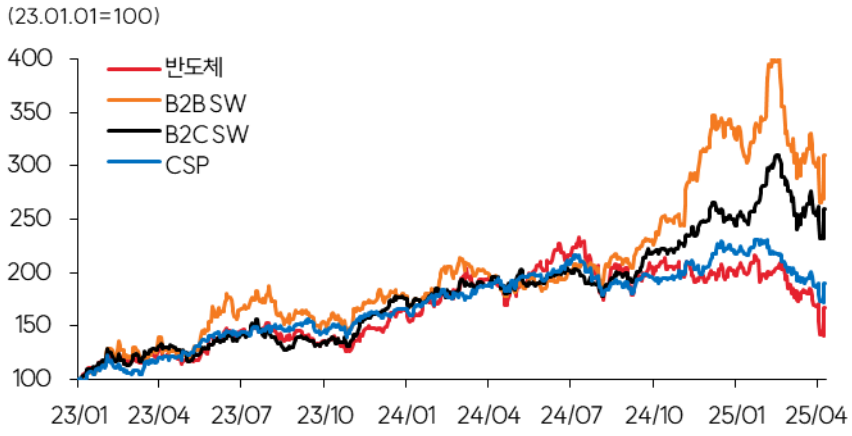
본 보고서는 Appendix 에서 'LLM 이란 무엇인가'를 시작으로, AI 제품화 시대에 접어든 현재 플랫폼 기업(META, Google)의 수혜 요인까지를 살펴봅니다. 부족한 기술 이해에도, LLM 관련 지식과 산업 흐름을 최대한 쉽게 전달하고자 했습니다.

2022 년 말부터 생성형 AI 는 첫 번째 랠리를 시작했고, 인프라 기업들이 초기 수혜를 누렸습니다. 엔비디아는 한때 미국 시가총액 1 위에 오르기도 했습니다. 그간은 AI 의 기반을 닦는 시기였다면, 올해부터는 본격적인 제품화의 시기라 판단됩니다. 이로 인해 제 2 의 AI 랠리도 가능할 것으로 전망합니다.

그 배경으로는 작년 말부터 시작된 모델 고도화 흐름(Test time scaling, 합성데이터)과 최신 인프라(Blackwell) 보급에 따른 추론 비용 감소가 있습니다. 고도화된 AI 는 1) Agent 시대의 핵심인 Coordinate AI 역할을 수행할 수 있고, 2) 증류 기술을 통해 비용 효율적인 Work Tool AI 개발도 가능합니다. 특히 추론 비용 절감은 제품화 확산에 중요한 전제입니다. 최근 지브리 사태로 GPU 부족을 언급한 샘 알트먼의 트윗처럼, 비용 문제로 인해 많은 서비스가 상용화되지 못하고 있습니다.

작년 말 젠슨황의 언급을 보면, 엔비디아 GPU 의 생성형 AI 훈련과 기존 ML 모델 고도화에 집중되어 있었습니다. 실제 서비스(추론)에 쓰인 비중은 제한적이었습니다. 하지만 올해는 다릅니다. 제품 출시가 본격화되는 한 해가 될 것입니다.

AI 관련주 주가 움직임



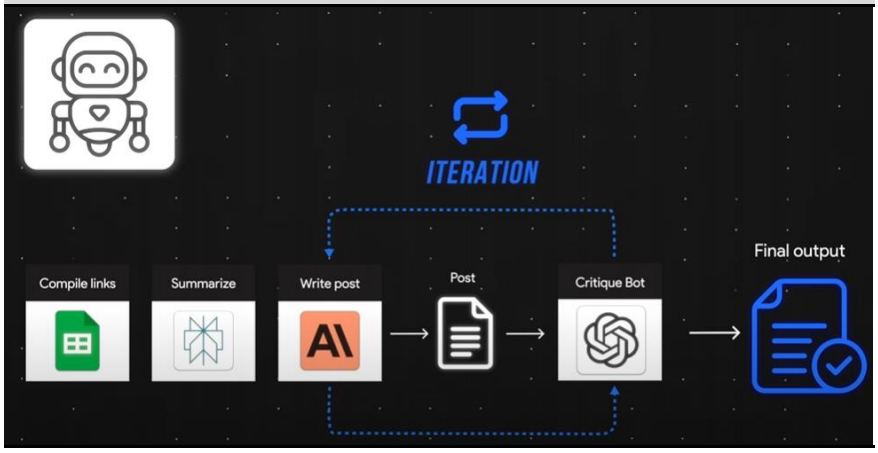
자료: Bloomberg, SK 증권 / 주: Infa = 필라델피아 반도체 지수, CSP = GOOGL, MSFT, AMZN
B2B= IBM, SAP, SNOW, TTD, NOW, CRM, IT, PLTR
B2C:= META, NFLX, SPOT, RBLX, PINS, SNAP, DUOL, MTCH, CHGG

Key Charts

단일 LLM의 작동 원리, AI 모델의 진화는 해당 성능의 강화

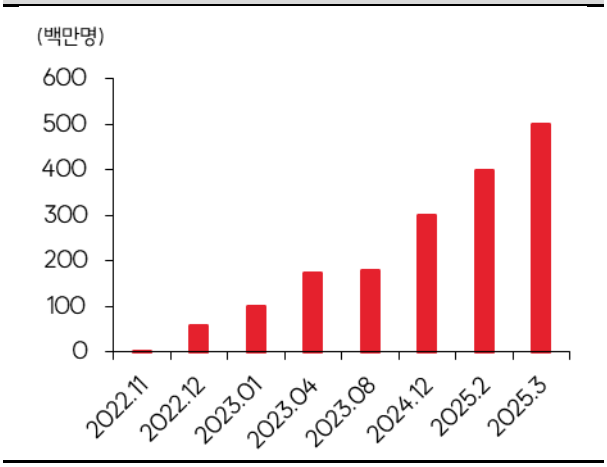
자료: Youtube(Jeff Su), SK 증권

AI Agent 구조. AI를 통해 구조를 설계, 피드백



자료: Youtube(Jeff Su), SK 증권

OpenAI 사용자 수

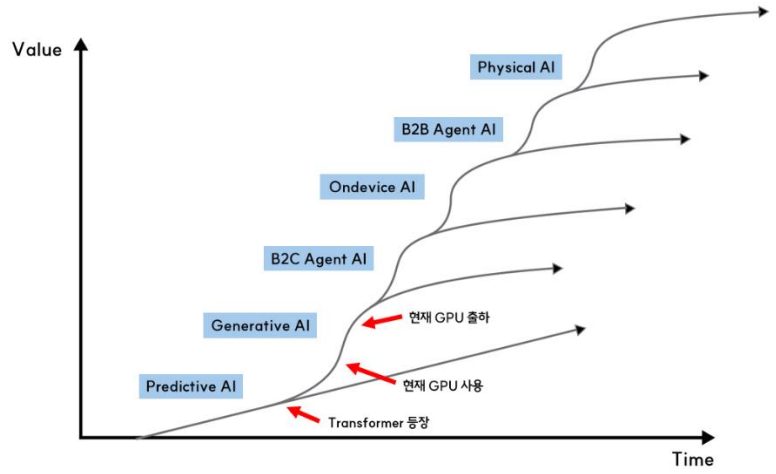


자료: 산업 자료, SK 증권

지브리 광풍으로 GPU 부족

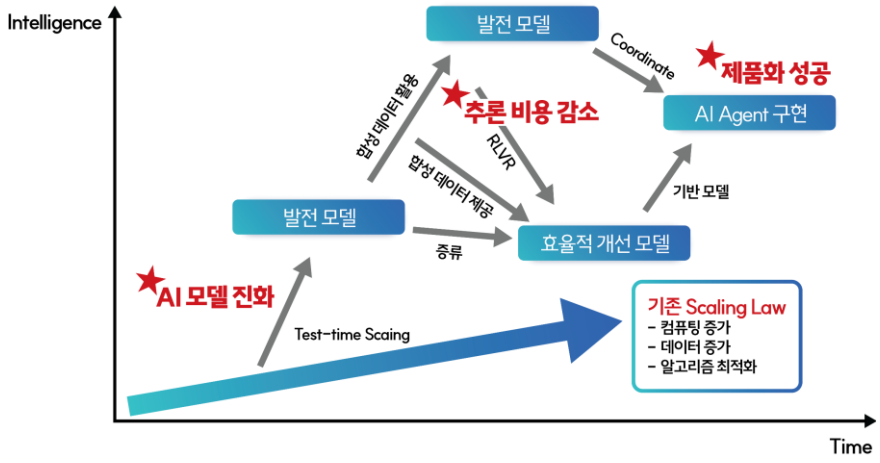
자료: X, SK 증권

AI 산업 확산 순서



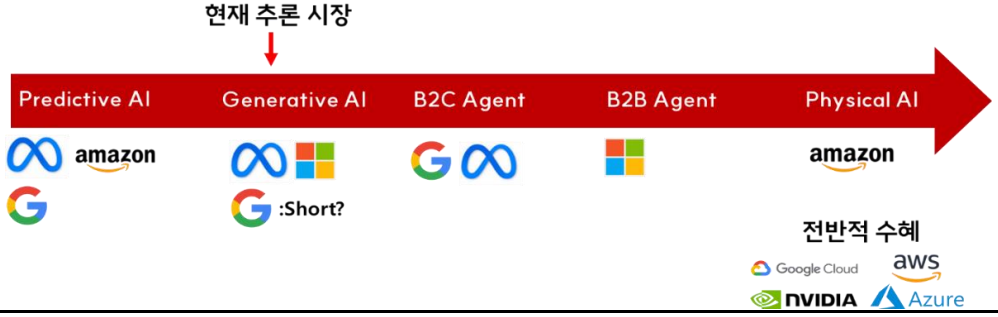
자료: SK 증권 추정
주: GPU 출하는 Blackwell, 현재 사용 GPU는 Hopper 가정

AI 발전 방향성 도식화



자료: SK 증권

향후 AI 산업 전개에 따른 빅테크 기업 수혜 정도



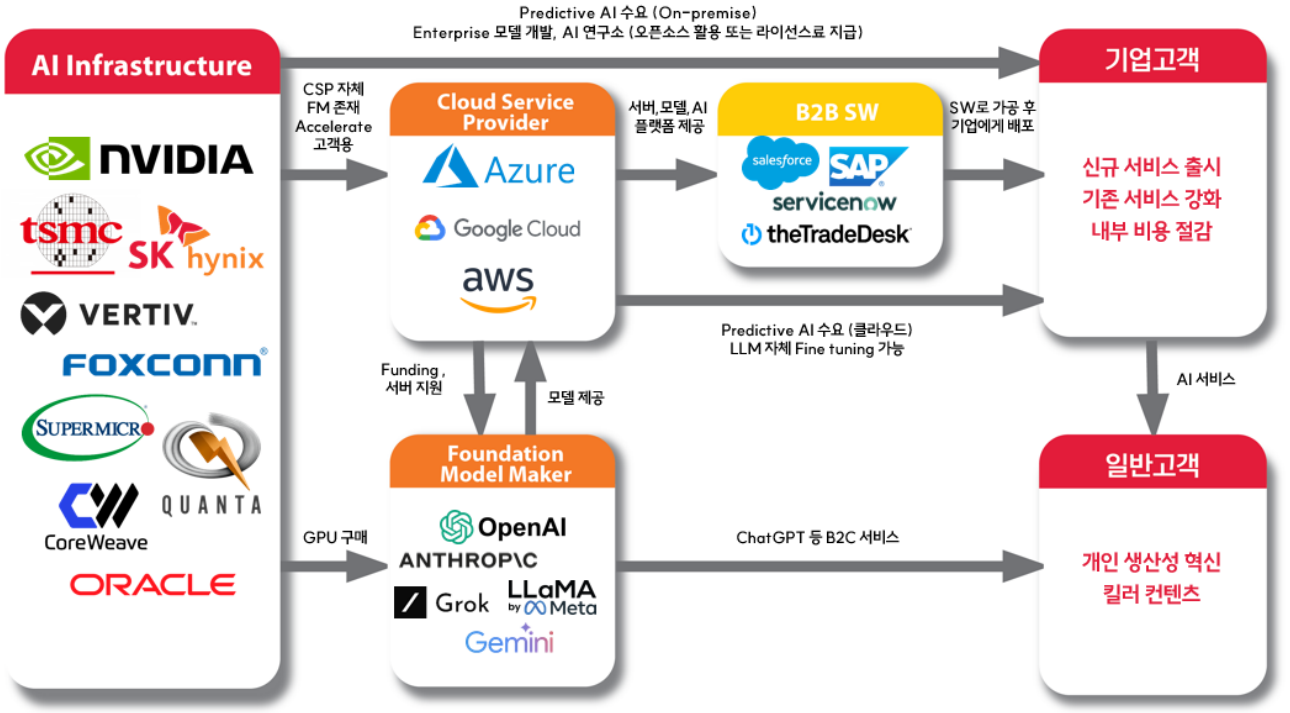
자료: SK 증권

CSP 3사 AI 관련 제품

	마이크로소프트	구글	아마존
ASIC	Maia + NVIDIA	TPU + NVIDIA	Tranium/Inferentia + NVIDIA
Foundation Model	OpenAI, PHI-4	Gemini	Claude, Amazon Nova
LLM base 플랫폼	Azure AI Foundry	vertex.ai	Amazon SageMaker, Amazon Bedrock
AI as a Service	Copilot	Gemini	Amazon Q
비즈니스 생산성 툴	Office	Google Workspace	amazon WorkDocs
ERP	Microsoft Dynamics 365	N/A	N/A

자료: 산업 자료, SK 증권

AI 산업 주요 밸류체인 그림



자료: SK 증권

1. AI 산업의 3 가지 목표: 진화, 비용 감소, 제품화

현재 생성형 AI 산업이 쏟는 노력의 방향을 다음과 같이 정리할 수 있다.
1) **AI 모델의 진화**: 더 높은 지능의 생성형 AI 모델 구현
2) **추론 비용 감소**: 서비스 비용을 감소시켜 현재 모델 확산에 기여
3) **제품화**: 소비자에게 더 접근성 있게 모델을 가공하고 배포

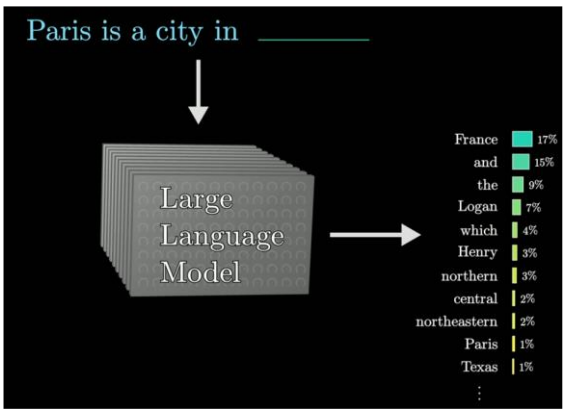
AI 는 체스, 바둑, 언어 순서로 더 많은 데이터 처리하며 진화

1-1. 언어 데이터를 다루는 AI 의 등장

생성형 AI 는 머신러닝(Machine Learning)의 진화된 형태로, 기존 ML 보다 병렬 연산에 유리해 더 방대한 데이터를 처리할 수 있다. ML 은 1997 년에는 경우의 수가 10^14 인 체스를(딥블루), 2016 년에는 10^70 인 바둑을(알파고) 다룰 수 있게 했다. 이어 2017 년 LLM(Large Language Model)의 등장으로 경우의 수가 사실상 무한한 언어까지 다루게 되며, 개연성 있는 단어 생성이 가능해졌다.

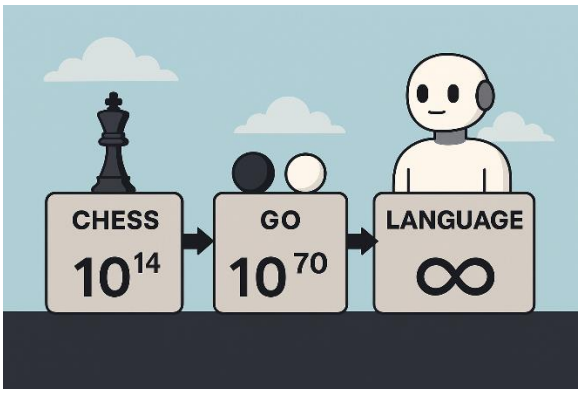
단어를 생성하는 이 기술은 산업 전반의 변화를 이끌고 있다. 많은 전문가들이 생성형 AI 를 '다음 생산성 도약'의 핵심으로 보고 있으며, 기업 ROE 와 국가 GDP 를 높일 수 있다는 전망이 쏟아진다. 오늘날 발전된 LLM 서비스를 일상에서 쉽게 접할 수 있고, 수많은 자본과 인재들이 이 기술의 진화, 확산, 제품화에 집중하고 있다.

LLM 은 Transformer 를 통해 다음 단어의 예측을 가능하게 만드는 함수



자료: 3Blue1Brown, SK 증권

점점 복잡한 데이터를 정복해온 AI



자료: SK 증권

AI 제품화와 진화

1) 진화는 단일 LLM의 강화

2) 제품화는 다수 LLM을 가공하여 사용자 이용성을 증가

1-2. 'AI 모델 진화'와 'AI 제품화'의 차이점

이 보고서에서 말하는 'AI의 진화'는 단일 생성형 AI의 성능 향상을 의미한다. 제품화는 다수 AI를 가공하여 이용성을 높인 것이다.

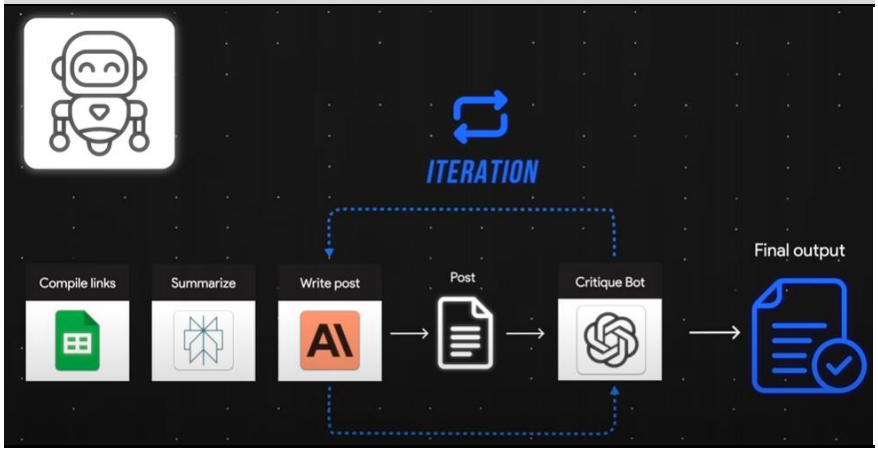
최근 여러 AI를 결합, 조율, 후처리(post-training) 하여 단일 LLM이 세운 벤치마크를 넘기는 모델이 자주 등장한다. 엔비디아의 발표에서는 Generative AI, Agentic AI, Physical AI로 이어지는 흐름이 자주 등장한다. Agent는 Generative AI 진화형으로 보이지만, 실제로는 Generative AI를 '제품화'한 형태에 더 가깝다.

AI Agent는 여러 개의 AI 기능(Worker Agent)을 하나의 똑똑한 AI(Coordinator)가 관리, 감독, 피드백하는 구조다. 예를 들어 "뉴스를 수집·요약해 매일 아침 텔레그램에 발송하라"는 명령을 수행하려면, 여러 AI가 협업해야 한다. 뉴스를 수집하는 크롤링용 AI, 요약하는 AI, 텔레그램 API로 전송하는 AI가 각각 필요하다. 이 과정을 사람이 조율하면 'AI Workflow'라고 하고, 이 Workflow를 AI가 직접 설계하고 피드백까지 수행하면 이를 'Agent'라고 부른다.

단일 LLM의 작동 원리, AI 모델의 진화는 해당 성능의 강화

자료: Youtube(Jeff Su), SK증권

AI Agent 구조. AI를 통해 구조를 설계, 피드백



자료: Youtube(Jeff Su), SK증권

2. AI 제품화 시대의 도래

AI 발전 방향은 가시적이며 유효 연산량의 증가로 추론 비용 감소도 지속될 전망이다. 이미 나와있는 생성형 AI 모델들의 확산이 용이해지고, 차세대 AI 모델 개발에 따라 신규 AI 서비스 등장이 전망된다.

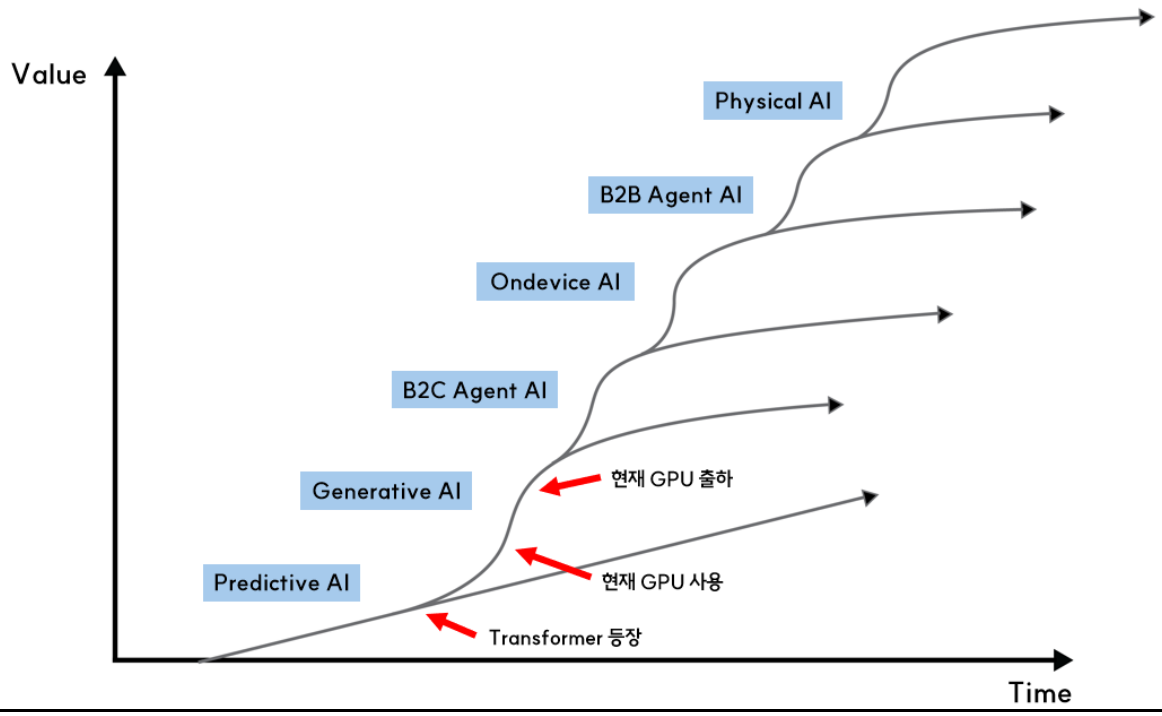
2-1. 올해가 생성형 AI 제품화 개화 시기

1) Blackwell 출하로 추론 비용 감소
2) B2C 제품화 모델 성공 사례 등장

2024 년 연말 젠슨황은 GPU 사용량의 60%가 AI 모델 훈련에 사용되고 있으며 40%의 추론 중 대부분(significant amount)이 추천 시스템에 사용되고 있다고 언급하였다. 추천 시스템은 생성형 AI 의 등장 이전에 있던 ML(Predictive AI 로 갈음)이다. 이는 GPU 중에서 실제 생성형 AI 가 제품화되어 판매되는 데 사용된 비율이 극히 일부에 불과했음을 의미한다.

올해부터는 본격적으로 생성형 AI 의 추론(제품화된 AI 판매) 이 확대될 것으로 예상된다. 첫째, 급격한 연산량 증가로 AI 추론 비용이 크게 낮아졌다. 현재 Blackwell GPU 가 출하되고 있으며, 특히 추론(FP4)에 특화된 모델인 만큼 올해의 비용 절감 폭은 더 클 것으로 전망된다. 둘째, B2C AI 모델의 제품화 성공 사례가 등장하고 있다. OpenAI 는 추론 비용 하락에 따라 img2img 기능을 ChatGPT 에 추가했다. ChatGPT 4.0 모델은 2022 년에 이미 완성되었으며, 이후 OpenAI 는 인력을 375 명(2022 년)에서 4,400 명(2025 년 5 월) 수준으로 확대했다. 올해는 추론 비용 부담이 완화된 만큼, 더 다양한 제품 출시가 기대된다. 지난 1 년간 딥시크, Claude 등 다양한 업체들이 서로 제품 컨셉을 모방하며 발전했듯, AI 제품화의 흐름은 대부분의 AI 모델로 빠르게 확산될 것이다.

AI 산업 확산 순서



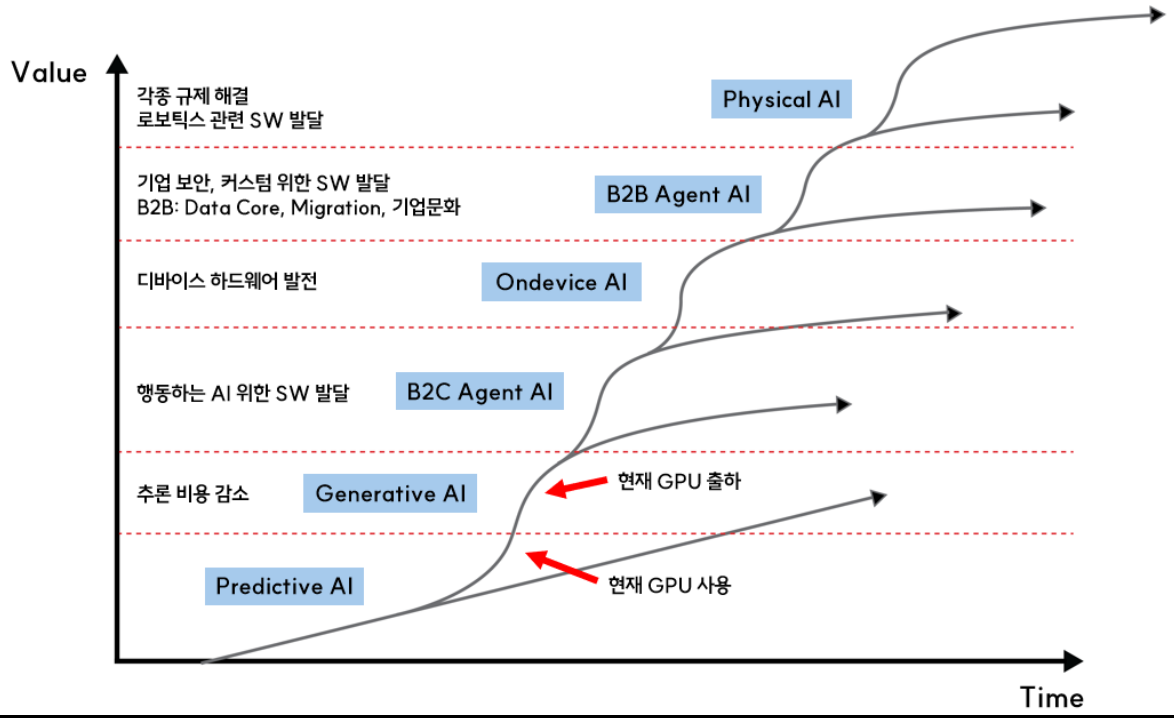
자료: SK 증권 추정
주: GPU 출하는 Blackwell, 현재 사용 GPU는 Hopper 가정

B2B SW 기업 AI 관련 주요 동향

AI 단계	정의	주요 부가가치	진입 장벽	주요 제품
Predictive AI	정형 데이터를 기반으로 미래 결과를 예측하는 전통적 AI	추천 알고리즘 기업 내부 자동화, 예측 정확도 향상		IBM Watson 추천 모델, 광고 모델
GenerativeAI	대규모 언어·비전 모델을 기반으로 새로운 콘텐츠를 생성하는 AI	언어데이터 해석 성공 콘텐츠 제작 비용 절감	Transformer 개발 추론 비용 감소 더 필요	챗봇, 이미지 생성
B2C Agent	일반 사용자를 위한 소비자용 AI 비서	개인 생산성 증가 검색 대체, 일상 업무 자동화	추론 비용 감소 필요 SW 개선 추가 필요	Copilot, ChatGPT 앱
Ondevice AI	AI 모델이 클라우드가 아닌 로컬 디바이스에서 실행	프라이버시 확보 멀티모달 input 데이터 확보 가능	추론 효율적 모델 필요 하드웨어 개선 필요	AR/VR 기기
B2B Agent	업무 자동화를 위해 기업 내에서 사용하는 AI 에이전트	기업 생산성 향상 (SaaS 확장) 인간의 Intelligence 노동 대체	기업 보안 커스텀을 위한 Unhobbling 필요	Salesforce Einstien M365 Copilot
Physical AI	현실 세계와 직접 상호작용하는 AI	제조/물류 자동화 인간의 Physical 노동 대체	각종 규제 해결 로보틱스 관련 SW 발달	휴머노이드, 드론, 자율주행

자료: 산업자료, SK 증권

AI 산업 확장 순서와 각 단계별 진입 장벽



자료: SK 증권

LLM은 그 자체로는 가치 없음
Post-training, fine tuning 등 조절하여 '제품화' 단계 필요

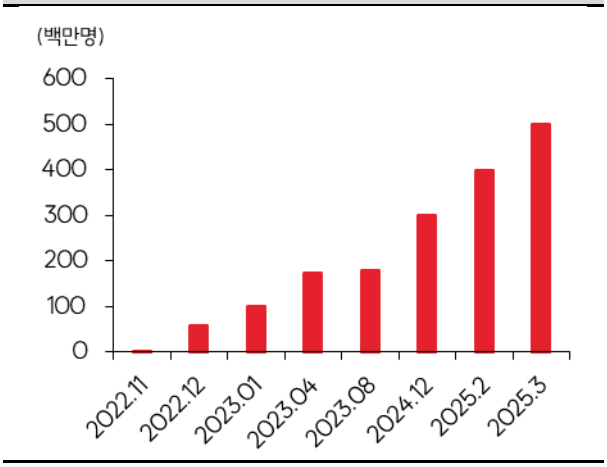
2-2. 생성형 AI도 가공돼야 가치가 있다

고성능 AI라고 그 자체로 수익이 나진 않는다. 실제 활용 가능한 수준으로 전환하기 위해서는 다양한 후처리 과정(post-training, fine-tuning 등)이 필요하다. 여기에 사용자 경험(UX), 서비스 설계, 워크플로우 통합 등의 비즈니스적 접근이 병행되어야 비로소 실질적인 제품이 된다.

가장 성공적인 LLM 제품으로 평가받는 ChatGPT 는 GPT-3.5(175B 파라미터) 기반으로 출시되었다. 동시기 Google 의 PaLM 1 은 540B 파라미터로 정량적 성능은 더 우수했지만, 실제 사용자 경험은 ChatGPT 가 더 뛰어나다는 평가를 받았다. ChatGPT 는 RLHF(Reinforcement Learning from Human Feedback) 기법을 적용하여 사용자 질문에 보다 자연스럽고 일관된 응답을 제공했다. InstructGPT 논문은 RLHF 의 효과를 100 배 더 큰 모델에 필적하는 수준의 품질 향상으로 평가한 바 있다. 초기 이용자 확보는 피드백 수집을 가능하게 했고, 이후 후속 모델의 post-training 품질을 지속적으로 끌어올리는 선순환 구조로 이어졌다.

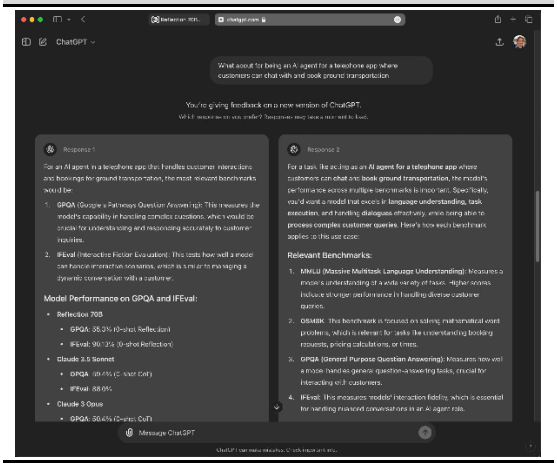
GPT 에 도입된 Deep Research 기능도 유사한 맥락에 있다. o1, o3 모델은 기존 GPT-4o 모델의 추론 방식을 개선한 모델이다. Context length 확장과 Inference 시간 증가를 통해 더 깊이 사고하는 구조를 만들었고, 이를 기반으로 고가 요금제(월 $200)를 도입했음에도 큰 수요가 확인됐다. 단순히 연산량을 확장하는 것 뿐 아니라 어떻게 모델을 만지고 사용자에게 제공하느냐가 중요해지는 흐름이다.

OpenAI 사용자 수



자료: 산업 자료, SK 증권

ChatGPT 의 피드백 구조



자료: SK 증권

2-3. 지브리 열풍, AI 제품화 성공 선례

지브리 열풍의 핵심
1) AI 설계 변경
2) 추론비용 감소 효과

최근 '지브리 이미지 생성' 열풍도 제품화의 관점에서 이해할 수 있다. GPT-4o 를 기반으로 한 이미지 생성은 모델 성능의 획기적 개선 없이도, 제품 설계 측면의 여러 임계점을 넘어서며 대중화됐다. 대표적인 변화는 다음과 같다:

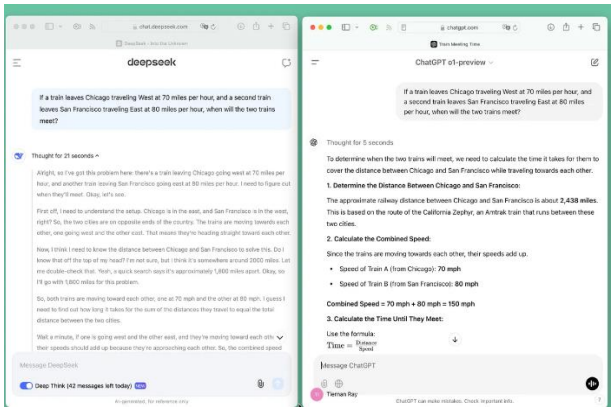
1) **img2img 기능의 탑재**: 기존 모델은 text2image 만 가능했다. 해당 기능 추가로 "내 셀카를 지브리 스타일로 바꿔줘" 같은 요청이 자연스럽게 실현되기 시작했다.
2) **접근성 증가**: ChatGPT 라는 이미 대중화된 플랫폼 내에 이미지 생성 기능이 통합됐다.
3) **텍스트 표현력 향상**: GPT-4o 는 단어 단위 텍스트 표현이 가능해지며 포스터, 배너, 명함 등 실제 제작 작업에서 유의미한 피드백 및 수정이 가능해졌다.

이 열풍의 영향으로 ChatGPT 의 유료 구독자 수는 약 1,500 만명에서 2,000 만명으로, 단 일주일 만에 급증했다. 유료 제품의 중간 가격을 $30~$40 로 가정할 경우, 연간 매출(ARR)은 $7.2B~9.6B 로 추산된다. B2B AI 도입에 가장 적극적인 마이크로소프트의 AI ARR($13B, FY2Q25 기준)과 크게 다르지 않은 규모다.

해당 기능들은 '추론 비용 감소로 인해 제공 단가가 맞아진 서비스'이기도 하다. Img 를 텍스트로 전환하여 프롬프트에 넣는 기능은 이전에도 구현이 가능했으나 긴 input 프롬프트로 단가가 맞지 않았을 것이다.

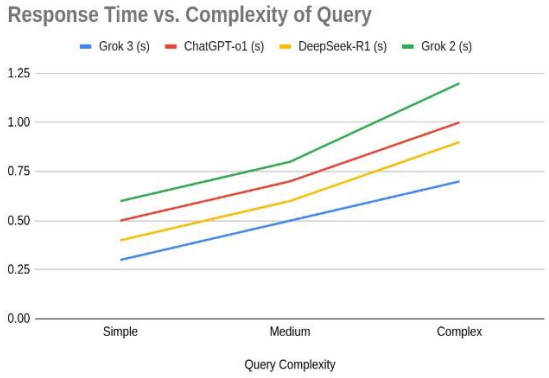
결국 LLM 의 진화는 단순한 parameter 크기 경쟁을 넘어서고 있다. 앞으로 AI 산업의 경쟁력은 base model 을 어떻게 가공(post-training)하고, 어떤 경험으로 사용자에게 전달하느냐(UI/UX, 접근성 증가, 마케팅)에 달려 있다.

Reasoning Trace 를 더 상세하게 노출시켰던 DeepSeek



자료: 산업 자료, SK 증권

Grok 3 신경망 최적화를 통해 경쟁 모델 대비 30% 빠른 응답



자료: Onedick AI, SK 증권

지브리 이미지 생성 열풍

자료:MBN, SK 증권

샘 알트먼 X: GPU 부족으로 서비스 제한

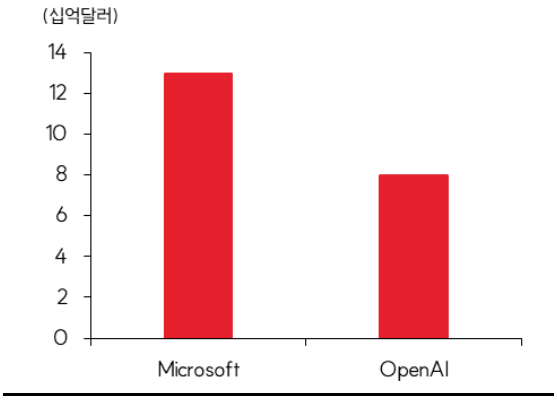
자료: X, SK 증권

Dall-E 와 GPT-4o 이미지 생성 기능 비교

구분	Dall-E 3	GPT-4o
기반 모델	Diffusion	Transformer
프롬프트 처리	텍스트 해석 능력 제한적	기존 GPT-4o 수준
기존 맥락 반영	불가능	대화 맥락 활용 가능
텍스트 렌더링	정확도 떨어짐	기존 텍스트 인식 기술 사용
img2img	불가능	가능

자료: 산업 료, SK 증권

Microsoft, OpenAI AI 관련 ARR 비교



자료: Microsoft, 산업자료, SK 증권
주: MSFT 는 Azure AI 제품, M365 Copilot, Github 매출 모두 포함

2-4. Agent 는 생성형 AI 의 제품화

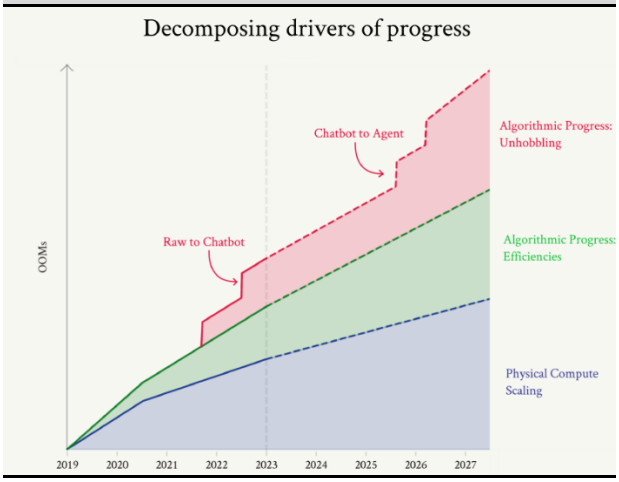
Agent 도 제품화의 일종
개선 제품이 하나씩 등장할 것

Agent 는 모델을 후처리, 합성, 가공하고 플랫폼에 붙여 제품화하는 것이다. Agent 가 필요한 이유는 생성형 AI 의 활용도가 아직 처참하기 때문이다. 2025 년 2 월 기준 미국 내 웹사이트 방문자 순위에서 OpenAI 는 여전히 하위권에 속한다. 일부 지식 노동자들에게는 유용한 도구지만, 대중에게는 AI 는 여전히 '신기한 기술' 수준에 머문다. Agent 를 가능케 하는 병목은 단순한 성능 향상으로 해결되지 않는다. 업계에서 주목하는 Agent 의 제품화 포인트는 다음과 같다

1) AI 모델의 장기 숙고*(Test time thinking 으로 일부 해결)*
2) 모델에 장기 메모리 탑재
3) AI 의 Tool Use(컴퓨터를 포함한 도구 활용)
4) 사용자 맞춤 대응, 기업 맞춤 대응 등의 유연한 변화

일부 한계는 컴퓨팅 증가를 통한 비용 감소로 해결이 가능하지만, 알고리즘적 장벽이 존재하는 문제도 있다. 가장 최근에 해결된 장기 숙고 문제 역시 Test-time thinking 이라는 새로운 접근으로 가능했다. Agent 는 완성된 형태로 한 번에 등장하지 않을 것이다. Test time thinking 과 같이 일부의 도입이 해제 포인트를 만들고 신규 제품이 등장하는 흐름이 이어질 것으로 보인다.

Unhobbling 으로 AI 모델 성능 도약 가능, 향후 Agent 로의 도약 기대

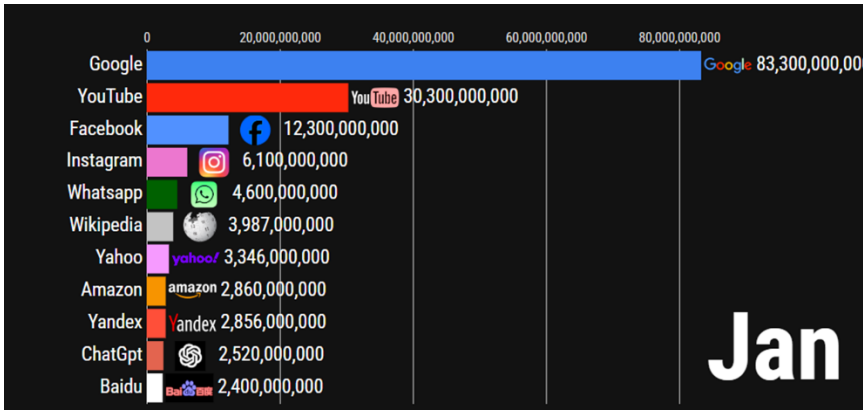


자료: Situational Awareness, SK 증권

AI Agent 에 필요한 Unhobbling

자료: Onedick AI, SK 증권

2025 년 1 월 미국에서 가장 많이 방문한 웹사이트, 아직 ChatGPT 는 비교적 하위권



자료: statisticsanddata, SK 증권

대표적인 Unhobbling 방식

방법	내용
RLHF (Reinforcement Learning from Human Feedback)	LLM 이 '자연스러운 문장 생성'이 아닌, '사용자에게 유용한 응답'을 하도록 보정하는 핵심 post-training 기법. ChatGPT 의 대중화에 결정적 역할
Chain-of-Thought (CoT)	복잡한 문제를 단일 응답이 아닌 단계별 추론 과정을 통해 해결하도록 유도. 수학, 논리, 코딩 등에서 정답률 향상에 기여
Scaffolding	복수의 LLM 이 역할을 분담해 상호작용하는 구조. 계획, 실행, 검토 등 일련의 사고 과정을 분리해 협업하는 방식으로 정밀도 향상
Tool Use (도구 사용)	계산기, 웹 검색, 코드 실행 등 외부 도구와의 연동을 통해 LLM 이 실시간 정보 활용 및 복잡한 연산을 수행할 수 있도록 지원
Long Context / Memory	입력 가능한 문맥 길이를 수천 > 수십만 토큰 단위로 확장. 문서 전체 이해, 장기 대화, 맥락 유지 등에 있어 실용성을 크게 향상시킴

자료: 산업자료, SK 증권

3. AI 모델의 진화: 신규 Scaling Law 등장

AI 모델의 진화는 Test-time scaling, 합성데이터의 사용으로 성장이 가시적이다. Scaling Law 는 작년 내내 공격의 도마에 올랐으나 이제 안정권에 돌입했다.

LLM 발전 Point
1) 컴퓨팅은 Capex, 기술력 필요
2) 알고리즘은 전문 인력 필요
3) 데이터가 꾸준한 병목

3-1. LLM 성능 향상 3 개의 축: 컴퓨팅, 알고리즘, 데이터

젠슨황 CEO 가 연설 때마다 반복적으로 강조하는 문구가 있다. '자원을 더투입하면 더 높은 지능이 나오는 구조'가 지속돼야한다는 점이다. 이는 일반적으로 Scaling Law 로 불린다. 2020 년 OpenAI 는 AI 발전의 공식을 다음 세 가지로 제시했다:

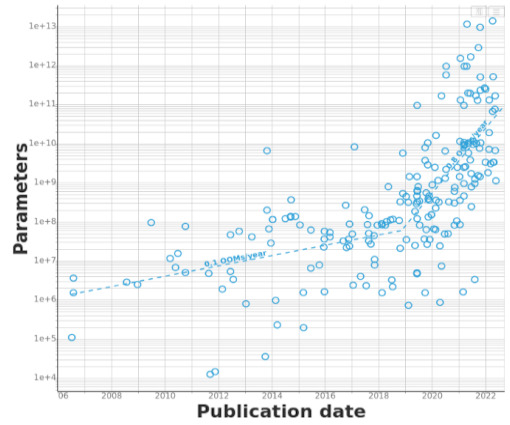
1) Compute Scaling
2) Parameter Scaling(=알고리즘 스케일링)
3) Data Scaling

LLM 발전 초기 병목은 **컴퓨팅**이었다. 특히 ChatGPT 출시 이후 AI 경쟁이 본격화되면서 컴퓨팅 리소스 확보가 산업의 핵심 이슈로 부상했다. OpenAI 는 Microsoft 의 지원을 바탕으로 컴퓨팅 인프라에서 압도적인 우위를 확보하며 시장을 선도했다. 지금도 선두권 모델에 도달하기 위해선 일정 수준 이상의 Capex 를 통한 대규모 데이터센터 확보가 필수적이다.

연구자들이 LLM 에 익숙해지면서 **알고리즘 최적화**는 지속적으로 일어났다. 예를 들어, Parameter Scaling 은 한때 성능 개선의 핵심 수단이었지만, compute 대비 효율이 낮아지는 구간(최적점)이 분명해지며 더 이상 공략 지점으로 취급되지 않는다. 알고리즘 최적화가 다시 주목받은 계기는 중국의 딥시크(DeepSeek) 등장이다. 딥시크는 순수 강화학습(RL), Mixture of Experts(MoE), 모델 증류(Distillation) 등 기법을 활용해 상대적으로 적은 컴퓨팅으로 상위권 모델과 유사한 성능을 구현했다.

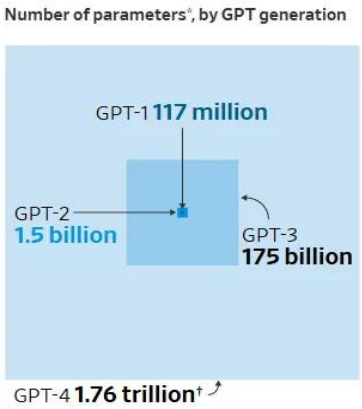
데이터는 꾸준한 병목이다. 2022 년 DeepMind 는 기존 모델 발전 방식이 데이터 대비 지나치게 많은 컴퓨팅 자원을 투입한다는 논문을 발표했다. 품질 향상을 위해서는 데이터의 양 자체를 늘릴 필요가 있었다. 하지만 인터넷 기반의 공개 데이터는 2026~2027 년이면 고갈될 것이라는 전망이 등장했고, 실제로 GPT-5 가 데이터 부족 때문에 출시되지 못하고 있다는 보도도 나왔다.

2018~2022 년까지 AI 모델은 10 만배의 Parameter 성장



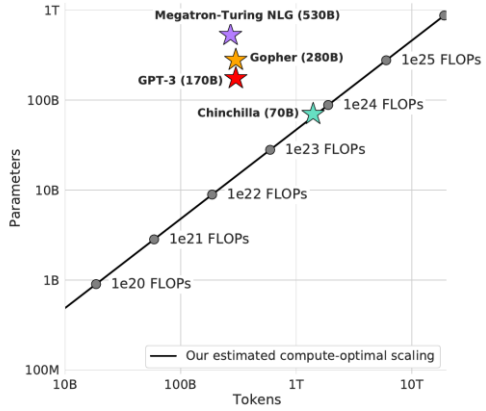
자료: Bloomberg, SK 증권

공격적으로 Parameter 를 늘려온 GPT 시리즈



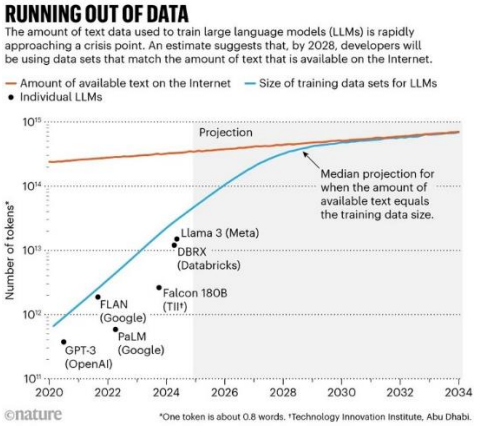
자료: Bloomberg, SK 증권

대부분의 모델들이 Optimal 대비 training token overfit 하다고 비판



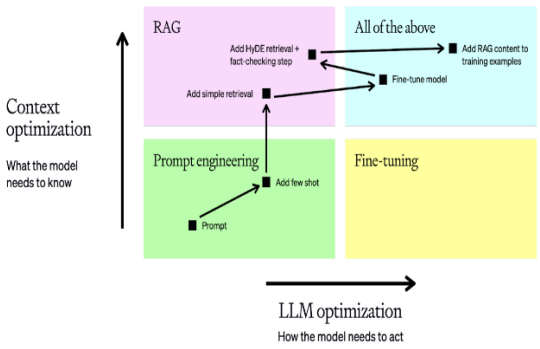
자료: 산업 자료, SK 증권

인터넷 공개 데이터는 2026~2028 년 사이 고갈 예정



자료: Nature, SK 증권

알고리즘 다양화로 LLM 최적화 움직임



자료: MediumAI, SK 증권

Data, Computing Scaling 끝으로 GPT-5 는 못나오는걸까?

자료: 한겨레신문, SK 증권

3-2. 신규 Scaling Law 등장: Test-time scaling

Test time scaling Point

1) 모델 성능 증가, 하극상 가능
2) 합성 데이터 품질 증가
3) 추론 컴퓨팅 수요 급증

2024 년 말, 새로운 Scaling 축이 등장했다. 바로 Test-time Scaling 이다. 기존의 발전 방식인 컴퓨팅, 알고리즘, 데이터는 LLM 의 Flash Think 능력을 강화하는 방향이었다. 다시 말해, LLM 이 답을 사전에 내재하고 있어야 하며, 사용자가 input 을 입력하면 1 초 이내에 답을 제공하는 구조였다. 그러나 OpenAI 의 Strawberry 모델 등장 이후 이 공식에 변화가 나타나고 있다.

Reasoning 모델은 Test-time Scaling 을 통해 AI 가 오랫동안 사고할 수 있도록 설계된 모델이다. 떠올려보면, 2016 년 이세돌과 바둑을 두던 알파고는 깊은 사고 끝에 수를 선택했다. 이런 '장고'는 기존 머신러닝에서도 자주 쓰이던 방식이다. 원리는 단순하다. AI 에게 스스로 자문자답을 반복하게 하면 된다. 이전에는 사용자가 prompt 를 반복 입력해야 했지만, Strawberry 는 이러한 사고 구조를 내재화한 모델로 볼 수 있다.

Test-time Scaling 은 AI 모델 간의 '하극상'을 가능케 한다. 훈련 단계에서 막대한 자원을 투입한 고성능 모델이 아니더라도, 추론 단계에서 반복 사고를 거치면 목표 benchmark 를 달성할 수 있다. 예를 들어, PaLM(54B) 모델에게 자문자답을 반복하게 하자 수학 벤치마크에서 57%의 정답률을 기록했는데, 이는 당시 별도 미세조정을 거친 GPT-3(175B) 모델보다 높은 수치였다. 매개변수 기준으로 3 배 이상 큰 모델을 능가한 사례다.

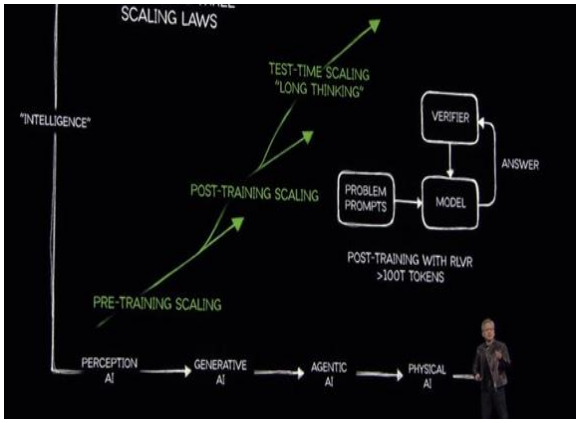
이런 변화는 Blackwell 모델이 '추론 위주 모델'로 소개되는 이유이기도 하다. GTC 2025 에서 젠슨 황은 "AI 스케일링 법칙이 초가속(hyper-accelerated)되었고, 요구 연산량이 작년 이맘때 예측보다 100 배 증가했다"고 밝혔다. 이는 Scaling Law 가 Test-time 중심으로 다변화되고 있다는 점을 반영한다.

Test-time Scaling 은 Context Length 증가 등으로 많은 비용을 초기에 요구하지만, 궁극적으로는 모델 효율성 개선과 추론 비용 절감으로 이어질 수 있다. 특히 Test-time Scaling 을 적용한 모델을 Teacher 로 활용해 Student 모델을 학습시키는 Distillation 방식을 통해 그 효과를 극대화할 수 있다.

Reasoning 모델은 기존 방식과 다르게 LLM 을 강화

자료: MediumAI, SK 증권

엔비디아 프레젠테이션에도 'Test-Time Scaling' 등장



자료: 한겨레신문, SK 증권

LLM 은 같은 질문에 대해서도 순차적 Prompt 를 통해 답변 유도 가능

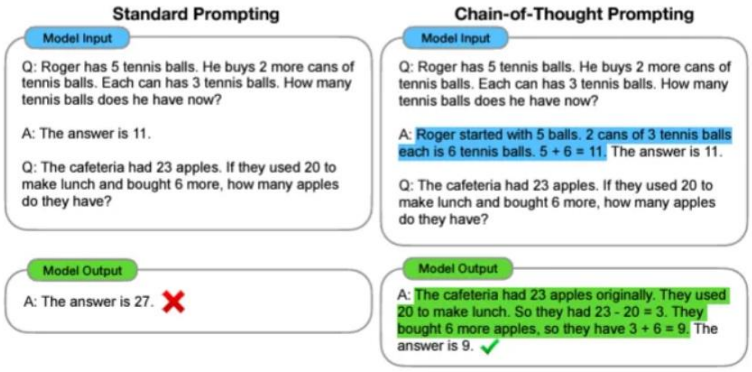
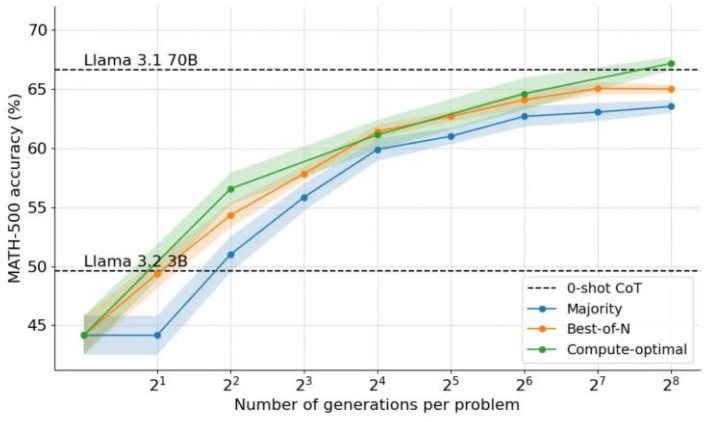


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

자료: Google, SK 증권

문제당 연산 과정을 늘리면 3B 모델로 70B 모델의 Benchmark 역전 가능



자료: Semianalysis, SK 증권

3-3. 합성 데이터 품질 증가로 기존 Scaling Law 지속 가능

합성 데이터 Point

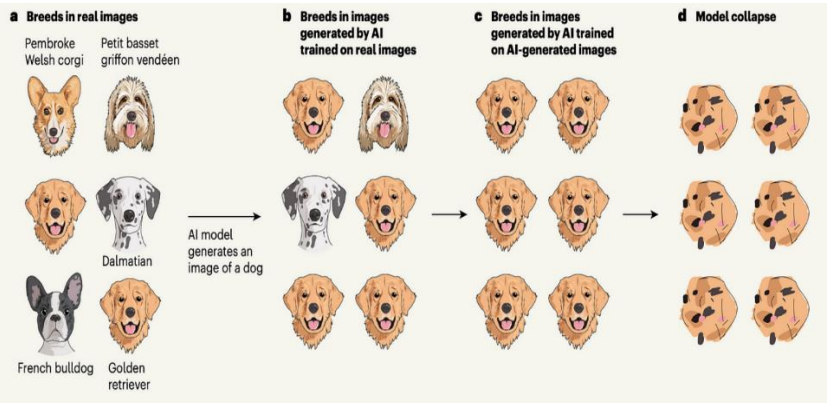
1) 모델 진화를 위한 데이터 공급
2) 효율적 소규모 모델 생산 가능

Test-time Scaling은 '100 수준의 모델'이 '1000의 성능'을 낼 수 있도록 하지만, 기본 성능인 100 자체를 끌어올리는 작업도 여전히 중요하다. 이 기본 성능 향상의 핵심은 합성 데이터(Synthetic Data)의 개선에 있다.

기존 합성 데이터 활용의 가장 큰 문제는 출력 데이터 품질의 허술함이었다. 단조로운 패턴의 데이터를 반복적으로 훈련에 사용하면, 모델이 자기참조적 순환에 빠지며 붕괴(model collapse)되는 현상이 발생했다. 그러나 최근 Reasoning 데이터가 등장하면서 데이터 품질이 크게 개선되었고, 이에 따라 합성 데이터를 활용한 모델 훈련이 다시 주목받고 있다. 이는 현재 LLM 발전의 가장 큰 병목으로 지적되는 '데이터 부족' 문제를 해결할 수 있는 유력한 방법이다.

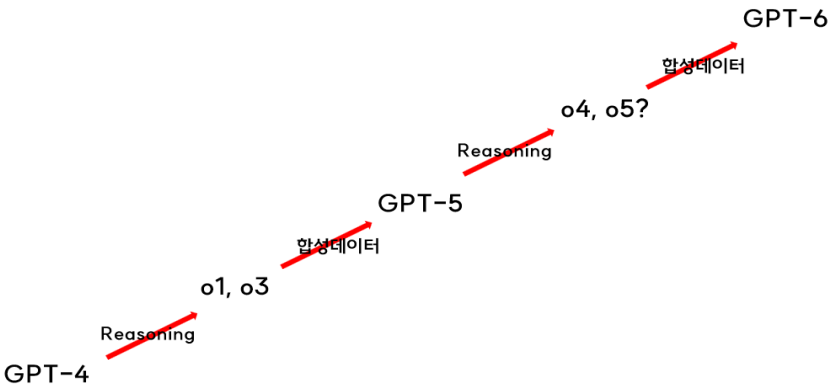
Microsoft, Anthropic, Hugging Face, Meta 등 주요 기업들은 모델이 생성한 데이터를 다음 세대 모델 훈련에 적극 활용하고 있다. GPT-5, Claude 4 역시 합성 데이터를 적극 활용하여 훈련 중인 것으로 알려져 있다. 특히 합성 데이터는 수학 및 코딩 능력 향상에 탁월한 효과를 보이는 중이다. 예를 들어, Microsoft 의 수학 특화 추론 모델인 Phi-2 는 대부분의 학습 데이터를 합성으로 구성했고, 결과적으로 자신보다 25 배 더 큰 모델보다도 높은 성능을 달성한 바 있다.

Model Collapse: 순환 학습 지속 시 모델이 과대 표준에만 편향



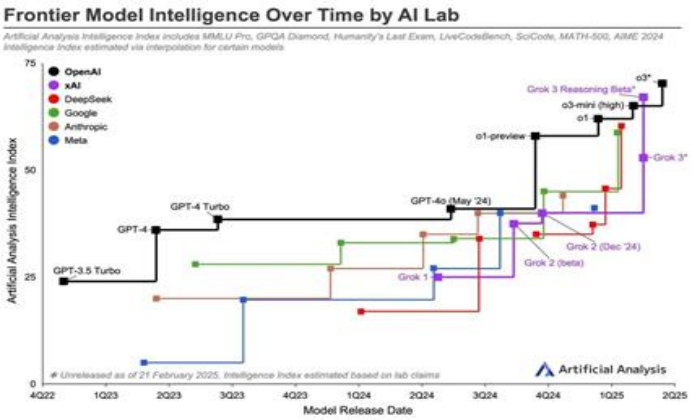
자료: medium AI, SK 증권

합성데이터를 활용한 AI 모델 발전 지속



자료: SK 증권

시간에 따른 AI 모델들의 성능 변화



자료: EpochAI, SK 증권

3-4. 합성 데이터로 Agent 구현에도 가까워진다

합성 데이터 Point

1) 차세대 모델 데이터 공급

2) 효율적 소규모 모델 비용 절감

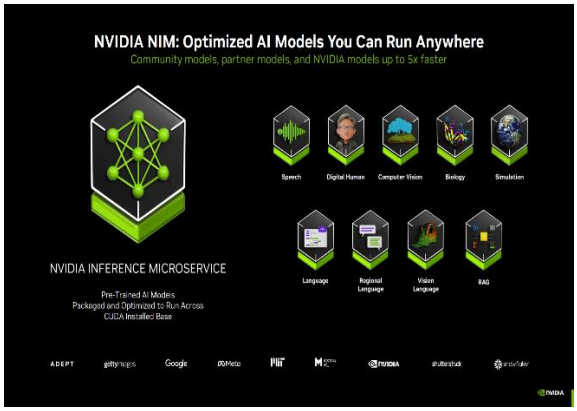
합성 데이터의 활용은 모델의 규모 확장(Scaling up)을 넘어서, Agentic AI 및 Physical AI 로의 진화에도 중요한 역할을 하고 있다.

Agentic AI 는 흔히 '행동하는 AI'로 설명된다. 작동 관점에서 보면 이 '행동'은 기존 생성형 AI 기능들이 조정(coordinate)되어 실행되는 것이다. 이러한 통합적 활용을 위해서는 똑똑한 조정자(coordinator)뿐만 아니라, 각 기능을 담당하는 효율적인 소형 모델(work tool)들도 필수적이다. 그리고 이 소형 모델들을 더 낮은 비용으로 효율적으로 학습시키는 핵심 수단이 바로 합성 데이터다.

이번 GTC 2025 에서 젠슨 황은 AI 강화 수단으로 RLVR(Reinforcement Learning with Verifier Rewards)를 소개했다. 이는 사람의 레이블 없이, AI가 문제를 스스로 생성하고 학습하며, 정답 검증(Verifier)까지 자동화하는 구조다. Verifier 까지 AI 가 수행하게 되면, 소형 LLM 을 더욱 효율적으로 생산할 수 있으며, 합성 데이터 활용을 극대화하는 방식이 된다.

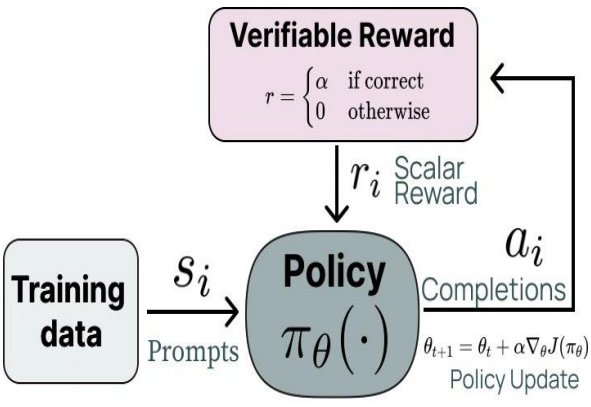
향후 Agent 등장 이후, 이를 강화하는 데에도 합성 데이터는 중요한 역할을 할 수 있다. 일반적인 LLM 이 '자연스러운 답변 생성'이라는 추상적 목표를 지닌 반면, Agent 와 Physical AI 는 구체적 목표를 향한 명확한 행동 수행이 요구된다. 이런 명확한 목적을 지닌 AI 를 훈련시키는 과정에서, 합성 데이터를 활용한 강화학습은 매우 높은 효과를 낼 수 있을 것으로 기대된다.

NVIDIA NIM(AI agent)의 구성 요소. 다양한 생성형 AI 포함



자료: NVIDIA, SK 증권

RLVR 구조, 향후 합성 데이터 생산까지 자동화될 경우 생성형 AI 양산 가능



자료: 산업 자료, SK 증권

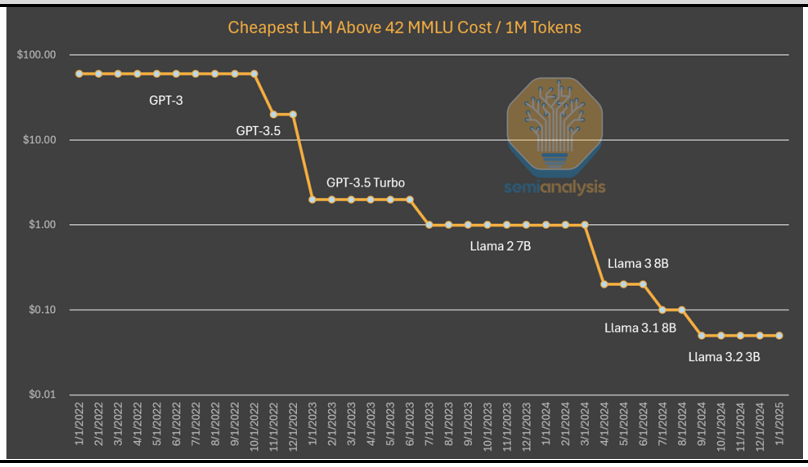
4. 추론 비용 감소: 연간 8배 속도의 하방 압력

AI 추론 비용은 2022년 이후 3,000배 이상 하락

AI는 현재 급격한 가격 하락을 경험 중이다. 특정 벤치마크(MMLU 42점)을 넘기 기위한 추론 비용은 2022년 1월 100달러에서 2025년 1월 0.03달러로 3,000배 이상 하락했다. **현재 모델간 유사한 성능으로 가격 경쟁이 치열한 상황이다. HW,SW 기술의 발전이 곧 단가 하락으로 이어질 것으로 기대할 수 있다.**

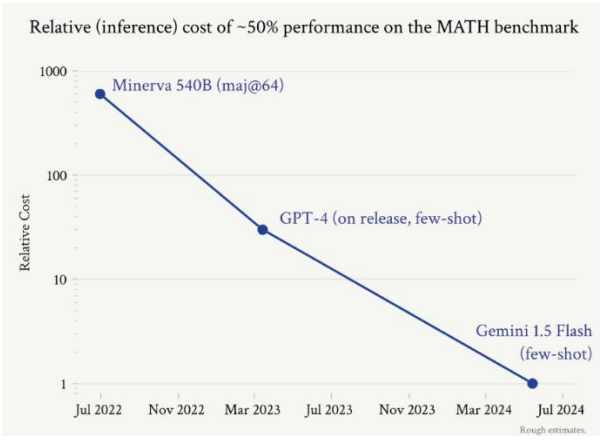
유효 연산량(Effective Compute)의 증가가 핵심이다. 유효 연산량은 HW가 구현한 연산량에 알고리즘 효율까지 포함한 '실질적인 계산 능력'을 의미한다. 유효 연산량의 증가로 전력, 자본 단위 당 처리 능력이 높아지면 추론 비용 감소로 이어진다.

벤치마크 42점 이상 LLM에서 토큰 1백만개 추론을 위한 비용 추이



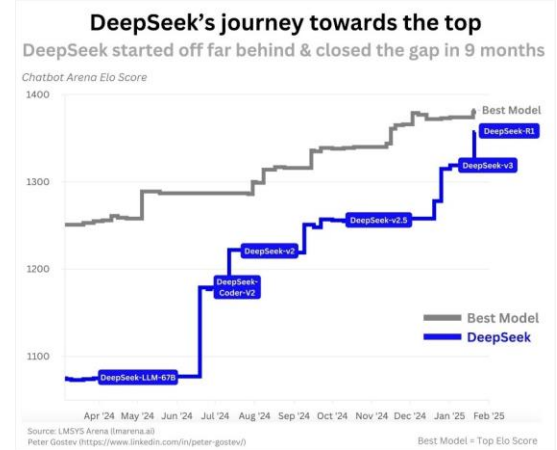
자료: semianalysis, SK 증권

MATH 벤치마크 50%를 넘기기 위해 필요한 추론 비용 1/1000 감소



자료: 산업 자료, SK 증권

Distillation을 이용한 DeepSeek의 모델 스팩 격차 좁히기



자료: 산업 자료, SK 증권

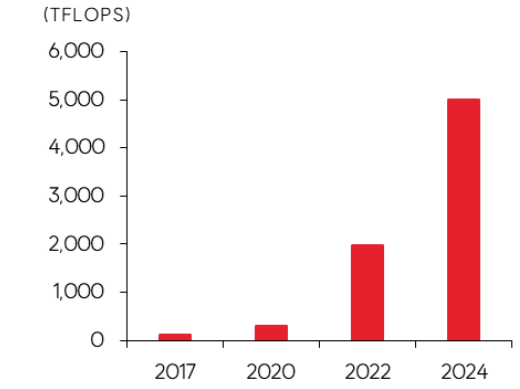
4-1. 하드웨어: 최전선에서 성능 가속 증폭 중

2017 년 대비 단일 칩 성능 720배 증가. 향후 2.7배 수준 성장 가능

현재 생성형 AI 모델의 90% 이상이 엔비디아 GPU 에서 훈련된다. 엔비디아의 성능 향상 속도로 하드웨어 연산량 증가를 추정해볼 수 있다. 2017년 V100 시리즈 대비 GB200 NVL72 단일 칩의 성능은 720 배 빠르다. Rubin 은 Blackwell 대비 13 배 성장이 기대된다. 이를 연환산할 경우 매년 2.7 배 수준의 성장이다.

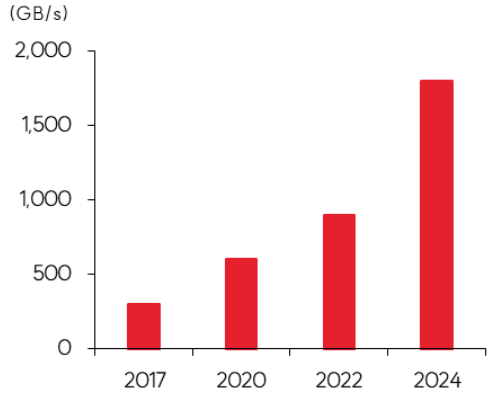
칩간 통신 속도(NVLINK, InfiniBand) 증가, 단일 랙 탑재 칩 수량 증가 (Blackwell NVL72, 최근 NVL576 로드맵 공개), 데이터센터 단위의 효율성 증가 (Liquid Cooling, Dynamo) 등 전방위적인 하드웨어 개선이 지속 중이다. 성능 증가는 총소유비용(TCO) 감소로 이어지는 중이다. 젠슨황은 Blackwell 은 Hopper 대비 TCO 가 87%, Rubin 은 97%의 비용 감소 효과가 가능하다고 언급했다.

현재까지 칩당 연산량 증가



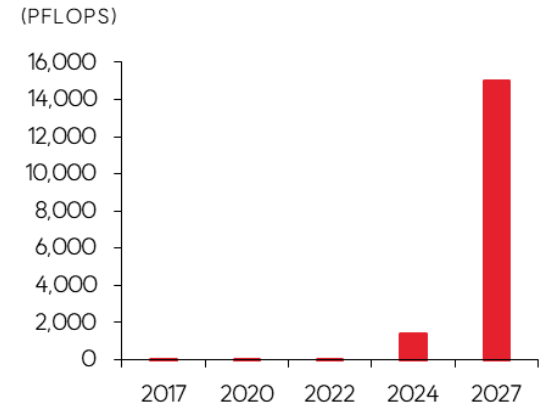
자료: NVIDIA, SK 증권

칩간 통신 속도 (NVLINK) 증가



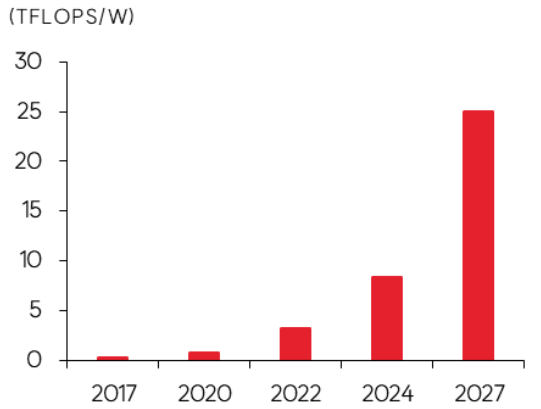
자료 NVIDIA, SK 증권

칩, 통신속도, 랙 대형화로 시스템 Throughput 상승 비약적



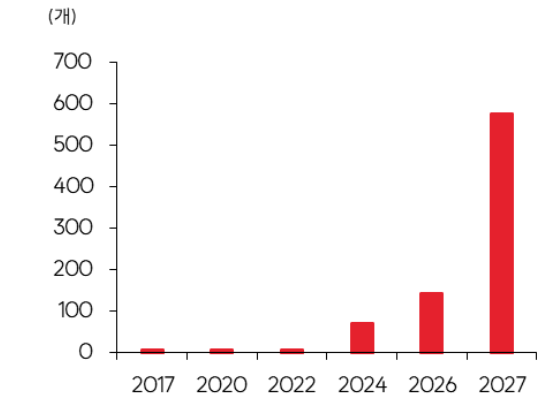
자료: NVIDIA, SK 증권
주: 2024 GB200 NVL72, 2027 Rubin Ultra NVL576 적용

시스템의 전력 1단위(W)당 연산(FLOPS) 증가



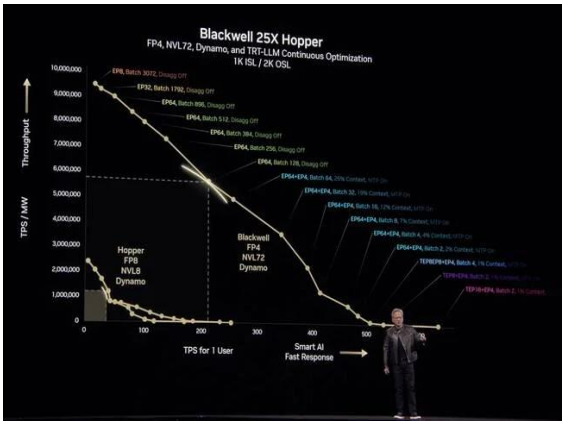
자료: NVIDIA, SK 증권
주: 2024 GB200 NVL72, 2027 Rubin Ultra NVL576 적용

단일 시스템당 탑재 GPU 개수



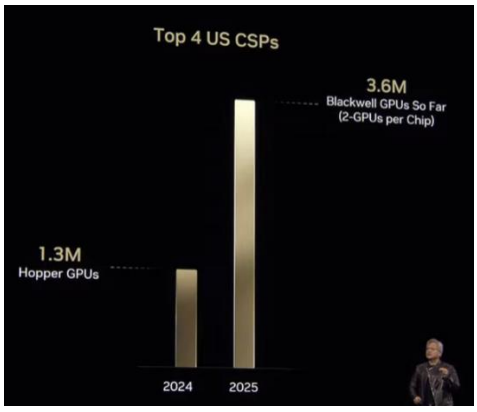
자료: NVIDIA, SK 증권

Dynamo 를 통한 모델 추론 연산 최적화



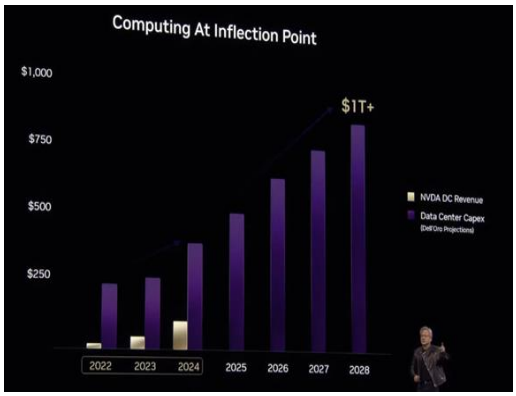
자료: NVIDIA, SK 증권

이미 Top4 CSP 들의 주문량은 최신 모델에 집중



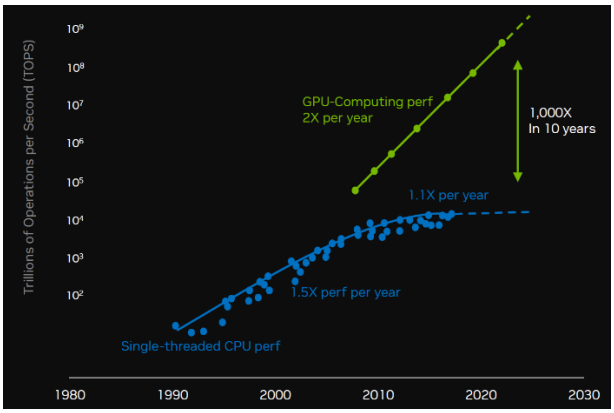
자료: NVIDIA, SK 증권

향후 Computing 부문에서 GPU 부문이 차지하는 비중 상승 전망



자료: NVIDIA, SK 증권

CPU 발전 속도는 감소하는 반면 GPU 는 가속 중



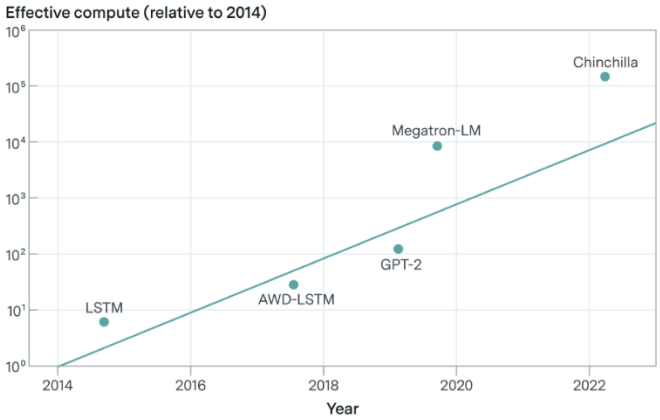
자료: NVIDIA, SK 증권

LLM 은 매해 2~4 배 효율성 증가, 향후에도 유지 가능

4-2. 알고리즘 최적화로 컴퓨팅 필요도 축소

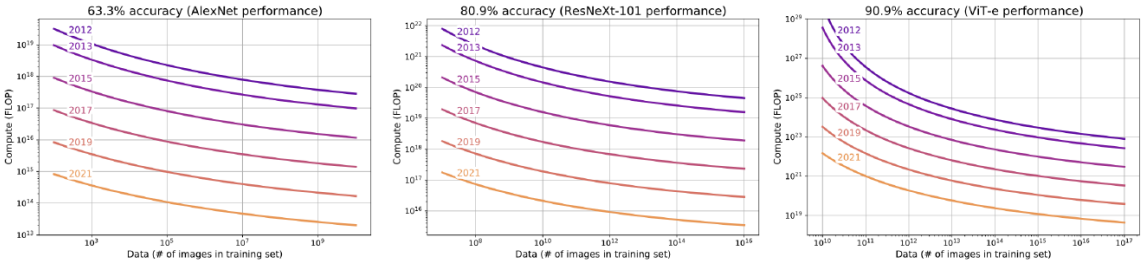
알고리즘 최적화는 같은 성능(벤치마크 달성도)을 위해 필요한 계산량을 줄여준다. 100 의 성능 도약을 위해 100 만큼 하드웨어 성능을 높일 수도 있지만, 10 만큼 성능을 높이고 10 만큼 최적화를 통해 필요 연산량을 줄일 수도 있다. 생성형 AI 모델을 개선하기 위한 모델 아키텍처 개선(Transformer+), 데이터 비중 조절(Chinchilla optimal), 추론 과정 효율화(MoE) 등 수 많은 모델 개선 방식들이 사용됐다. LLM 은 GPT-2(2019) 이후 현재까지 2~4 배의 효율성 증가를 보여주고 있다. 전문가들은 Image AI 모델이 그랬듯, LLM 도 당분간 유사한 속도의 알고리즘 효율화가 가능하다고 전망 중이다.

LLM 훈련 효율성은 하드웨어 FLOPS 증가량보다 가파르게 증가 (매해 3 배 수준)



자료: Epoch AI, SK 증권
주: Effective compute = 실제 투입된 연산량(FLOPs) * 알고리즘 효율성(multiplier)

ImageAI 에서 같은 정확도를 달성하기 위해 필요한 연산량 수가 해가 갈 수록 감소 (매해 3 배 속도)



(a) Pareto frontiers in data and compute for AlexNet performance (b) Pareto frontiers in data and compute for ResNeXt-101 performance (c) Pareto frontiers in data and compute for ViT-e performance

Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.

자료: 산업자료, SK 증권

유효 연산량 8 배 증가

1) 무어의 법칙 대비 2 배 이상
2) 추세선 대비 보수적 수치
3) AI 의 Tool 활용 시 비선형적 증가 가능성 배제

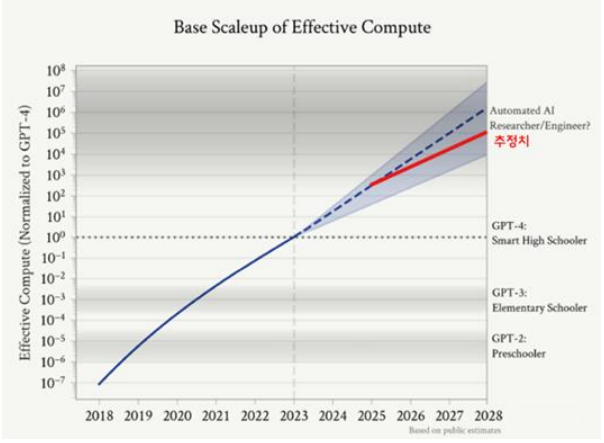
4-3. 향후에도 8 배 이상 유효 연산량 상승 기대

AI 하드웨어가 2.7 배, 소프트웨어가 3 배 수준의 성장을 지속한다면, 연간 약 8 배 규모의 AI 연산량 증가가 기대된다. 이 추세가 3 년간 지속되면 실질적 연산량(Effective Compute)은 500 배 이상 증가하게 된다. 그럼에도, 이는 지금까지 LLM 이 보여온 성장 속도(연간 14~16 배)에 비해 상대적으로 보수적인 추정이다. 따라서 AI 모델의 빠른 가격 하락은 앞으로도 충분히 지속 가능하다고 볼 수 있다.

그렇다면 연간 8 배의 연산 증가가 얼마나 빠른 것일까? 기존의 연산량 증가를 주도한 CPU 는 무어의 법칙에 따라 18~24 개월마다 트랜지스터 집적도가 2 배로 증가했고, 이에 따라 연간 약 50% 수준의 성능 향상을 이뤘다. 하지만 2010 년대 이후 무어의 법칙은 둔화됐고, 현재 싱글 코어 성능은 연 7% 증가에 그치는 수준이다. 소프트웨어 최적화를 AI 와 비슷하게 공격적으로 가정하더라도, 과거 컴퓨터의 연산량 증가는 연 4.5 배 수준이었다. 지금의 연산 증가율은 과거 컴퓨터 태동기의 공격적 시나리오보다 약 2 배 이상 빠른 셈이다.

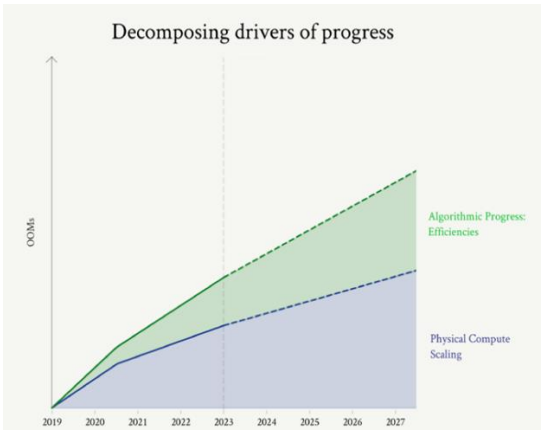
또한 연 8 배 성장은 AI 가 스스로 학습하는 '특이점(Singularity)' 없이, 인간 주도 하에서의 발전만을 전제로 한 추정이다. 향후 AI 가 도구를 활용해 환경을 조작하고, 서로 소통하며 학습을 축적하는 단계에 진입한다면, 비선형적 성장 곡선이 전개될 가능성이 크다. 실제로 DeepMind 의 CEO 데미스 하사비스는 "AGI 는 상호작용적 환경에서의 메타학습(meta-learning)을 통해 탄생할 것"이라고 밝힌 바 있다.

LLM 의 실질적 연산 능력 증가 추이 및 전망



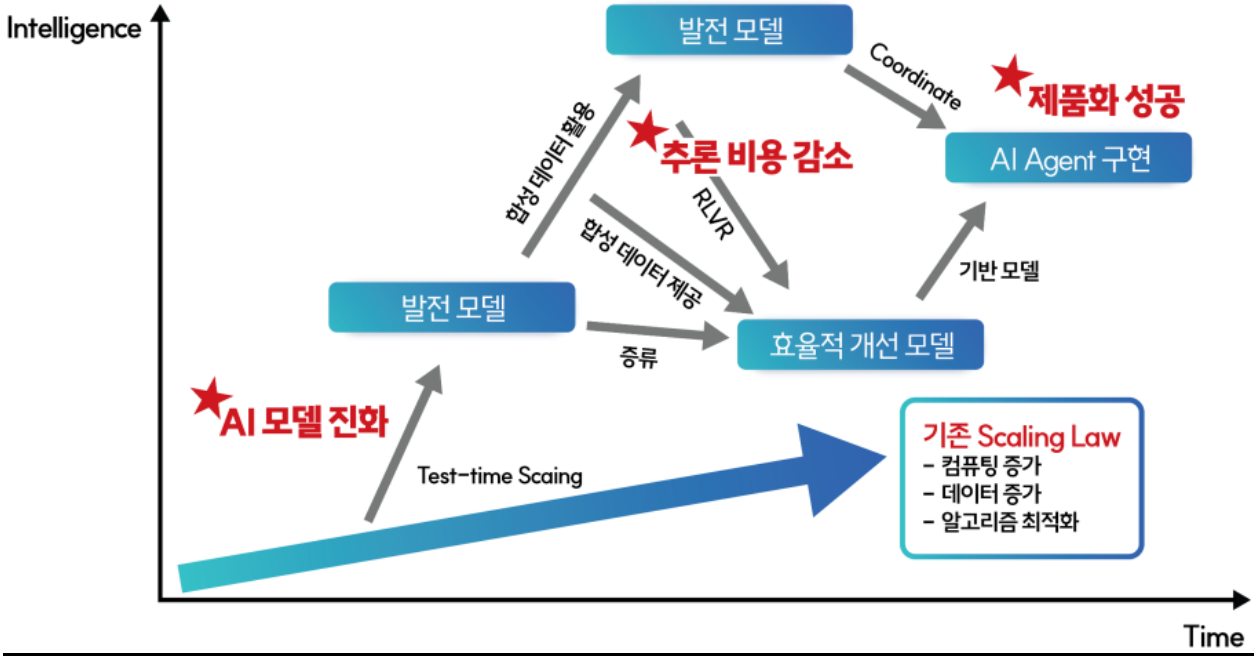
자료: Situational Awareness, SK 증권

하드웨어 성장 및 알고리즘 효율성 증가 기여



자료: Situational Awareness, SK 증권

AI 발전 방향성 도식화



자료: SK 증권

5. 제품화 시대, 다시 빅테크가 주도

인력, 자본 요구도가 아직 높은 AI 기존 빅테크들 위주로 주목 필요

Gen. AI 제품화 시기가 도래한만큼, 향후 AI를 제품화하는 기업들에 주목할 필요가 있다. AI 기업은 크게 AI 를 도입하려는 기업과 AI 를 활용한 서비스를 만드려는 스타트업(AI natives)로 분류할 수 있다. 아직 AI natives 상장사는 제한적이며 대부분의 주가 상승 및 부가가치는 AI 도입 기업들에게 인프라를 제공하는 AI 인프라 기업 및 이를 중계하는 클라우드 서비스 제공자(CSP)들로부터 나왔다.

제품화 시장도 CSP 사들의 역할이 주요할 것으로 판단된다. CSP 사업자들은 대부분 본업이 데이터센터 요구량이 많은 상황에서 이를 하나의 서비스로 발전시킨 형태이다. AWS는 아마존이 글로벌 전자상거래 플랫폼 확장에 대한 컴퓨팅 수요가 많아지면서 생겨났으며, 후발주자인 Azure 와 GCP 도 모두 내부 데이터 및 컴퓨팅 활용 노하우를 토대로 생겨난 사업이다. 따라서 미국 빅테크(CSP 사)들은 AI 산업을 중계하기도 하지만 가장 큰 수요자들이기도 하다.

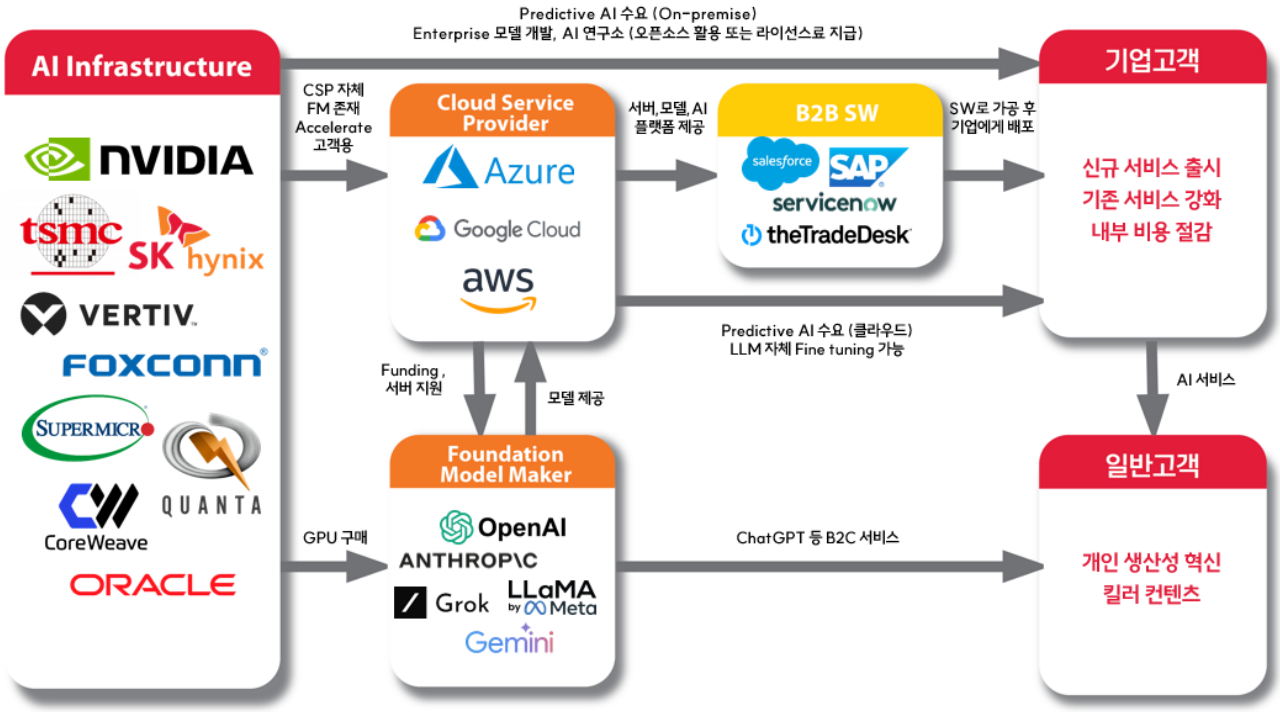
빅테크사들의 본업과 AI 활용을 간단히 정리하면 다음과 같다.
아마존: 전자상거래 서비스 구동 물류 효율화
알파벳: 플랫폼 내 노출도 효율화, 광고 엔진 효율화, 그 외 많은 AI 프로젝트 운영
마이크로소프트: 비즈니스 생산성 툴 생성, ERP 서비스
메타: 플랫폼 내 노출도 효율화, 광고 엔진 효율화

CSP 3사 AI 관련 제품

	마이크로소프트	구글	아마존
ASIC	Maia + NVIDIA	TPU + NVIDIA	Tranium/Inferentia + NVIDIA
Foundation Model	OpenAI, PHI-4	Gemini	Claude, Amazon Nova
LLM base 플랫폼	Azure AI Foundry	vertex.ai	Amazon SageMaker, Amazon Bedrock
AI as a Service	Copilot	Gemini	Amazon Q
비즈니스 생산성 툴	Office	Google Workspace	amazon WorkDocs
ERP	Microsoft Dynamics 365	N/A	N/A

자료: 산업 자료, SK 증권

AI 산업 주요 밸류체인 그림



자료: SK 증권

주요 LLM 플레이어 정보

2025년 4월 기준	누적 펀딩금액 (십억달러)	마지막 밸류에이션 (십억달러)	최근 출시 모델	차세대 모델 일정
OpenAI	17.9	157 (24년 10월)	GPT-4.5, o3 (25년 3월)	o4, o5 (25년 4월) GPT-5.0 (25년 7~8월)
ANTHROP\C	14.3	61.5 (25년 3월)	Claude 3.7 Sonnet Claude 3.5 Haiku (25년 2월)	Claude 4 (25년 중순)
deepseek	N/A	110 (25년 2월) *블룸버그 추정	DeepSeek-R1 (25년 1월)	DeepSeek-R2 (25년 5월)
xAI	12	75 (25년 2월)	Grok 3 (25년 2월)	Grok 4 (25년 말)
Gemini	N/A	N/A	Gemini 2.5 Pro Experimental (25년 4월)	미정
LLaMA	N/A	N/A	Llama 4 (25년 4월)	Llama 4 (수 개월 내 추가 출시)

자료: 산업자료, SK 증권

5-1. 단기적 B2C AI 시장 우위 전망

히트 제품이 이미 등장하는 B2C
낮은 생산성 AI(B2B) 침투율

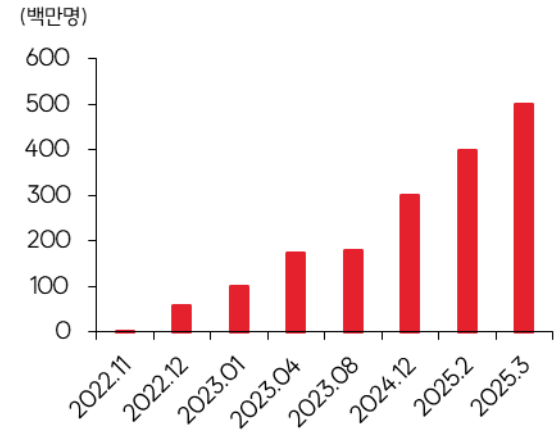
AI 시장에서 B2B 보다 B2C 의 AI 확산이 더 빠르게, 더 높은 부가가치의 형태로 나타날 것으로 전망한다.

이는 Microsoft 와 OpenAI 의 고객 확산 차이로 파악이 가능하다. 양사는 같은 AI 모델(GPT)을 기반으로 각각 B2B, B2C 위주의 서비스를 제공한다. OpenAI 는 가파른 속도로 5 억명의 MAU 를 확보한 반면 Microsoft 의 Copilot 계열 제품(M365, GitHub, Dynamics 등)의 유료 사용자는 대부분 수백만명으로 추정된다. 이는 Copilot 과, GitHub 전체 이용자 수 대비 2% 이내의 비율이다. AI ARR(Annual Recurring Revenue)를 분석해볼 경우 OpenAI 의 API 호스팅이 70% 이상을 차지한다.

주요 B2B 소프트웨어 기업들이 실적 및 전망 부진도 이를 뒷받침한다. 대표적인 B2B 소프트웨어 기업 6 개 중, 2023 년 대비 2024 년에 매출 성장률이 유의미하게 개선된 기업은 SAP 이 유일하다. 나머지 기업들은 성장률이 정체되거나 둔화된 모습을 보였으며, 향후 성장률 전망 역시 대체로 보수적이다.

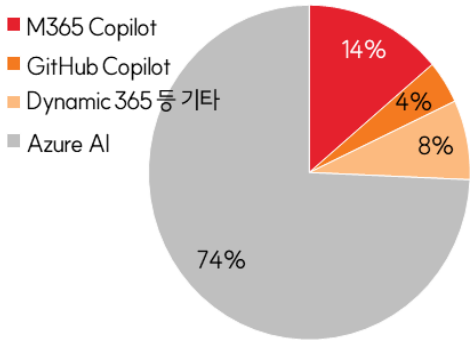
Microsoft AI ARR 의 대부분을 차지하는 B2B API 호스팅은 빠르게 성장하는 비즈니스가 맞지만 높은 부가가치의 비즈니스라고 생각하기 어렵다. AI 모델을 중계해주고 일부 SW 를 붙여서 편의성을 늘려준 서비스에 불과하기 때문이다. 또 관련 서비스 제공사들의 경쟁이 매우 심하여 API 순위가 신규 모델 출시 직후 빠르게 바뀌는 모습을 보이는 중이다. B2B AI 는 향후 기업형 AI Agent 가 도래한 이후 큰 부가가치를 창출할 수 있을 것으로 판단된다.

ChatGPT MAU 추이



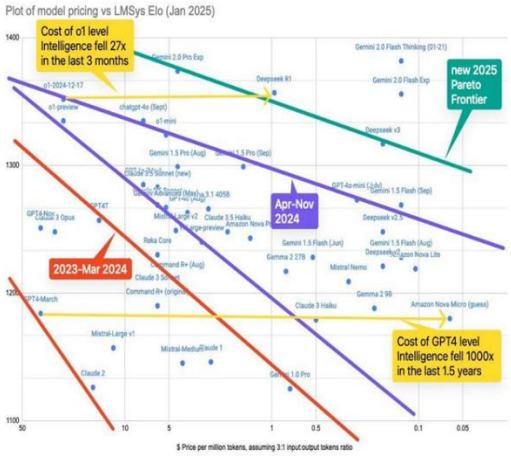
자료: 산업 자료, SK 증권

FY2Q25 Microsoft AI ARR 구성, Azure AI 호스팅이 70% 상회



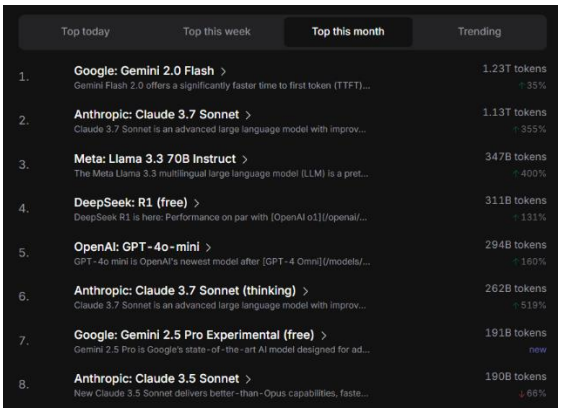
자료: SK 증권 추정
주: M365 Copilot 월 평균 가격 $30, 사용자 수 5M 가정
Github Copilot $15, 사용자 수 3M 가정

시간에 따라 Pareto 가 지속적으로 경신되는 B2B AI 모델 시장



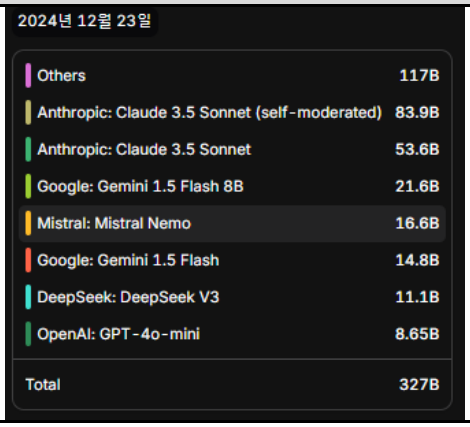
자료: 산업 자료, SK 증권

B2B API 호스팅 모델 월간 순위, 신규 모델 위주 다양한 모델들 포진



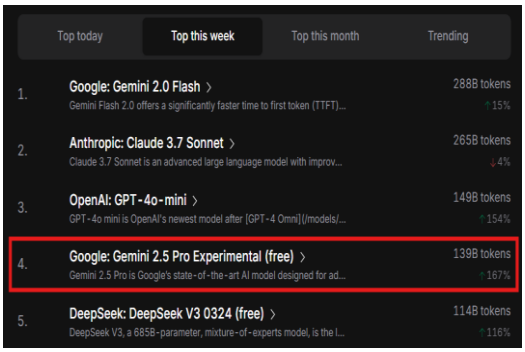
자료: Openrouter, SK 증권

API 호스팅 순위, 불과 3 개월 이전인 연말과 순위권 모델이 매우 다른 모습



자료: Openrouter, SK 증권

2025 년 3 월 25 일 출시된 Gemini 2.5 가 주간 순위에 등재

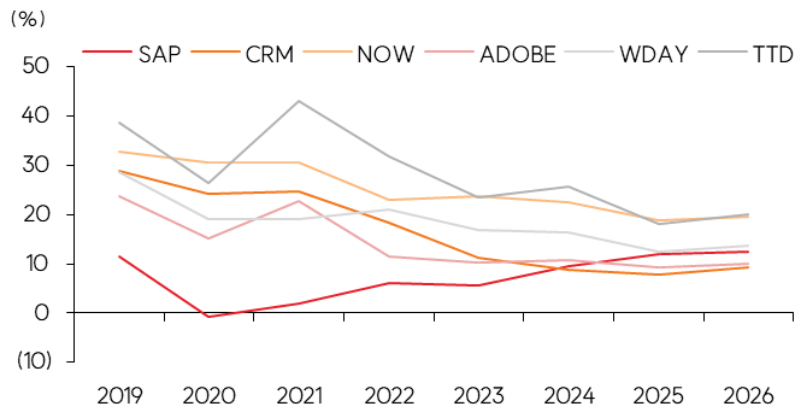


자료: Openrouter, SK 증권

AI Coding 보조 AiaaS 만 하여도 스타트업과 CSP 들간 경쟁이 치열

자료: 산업 자료, SK 증권

대표 B2B SW 매출액 추이 및 전망. 2024 년 성장률이 가속화된 기업은 SAP 이 유일



자료: Bloomberg, SK 증권

B2B SW 기업 AI 관련 주요 동향

기업 (주요 사업)	내용
SAP (ERP)	클라우드 계약의 50% 이상이 AI 사용 사례 포함 2024 년 동안 130 개 이상의 생성형 AI (Gen AI) 기능 출시 AI 기반 가상 비서 'Joule' 출시, 30,000 개 이상의 고객이 사용 중
Salesforce (CRM)	Agentforce 도입 90 일 만에 3,000 개 이상 고객 확보 Salesforce 내부에서도 Agentforce 활용, 고객 지원 요청 38 만 건 중 84% 자동 처리 Data Cloud & Agentforce 는 AI 소비량 기반 요금제로 적용 ($2 per conversation)
ServiceNow (ITSM)	AI 기반 업무 자동화가 서비스 요청의 85%를 해결, 개발자 생산성 20% 이상 향상 AI 활용 워크플로우 자동화 > 연간 40 만 노동 시간 절감 AI Agent 활용량 기반 소비 모델(consumption-based pricing) 추가
Adobe (컨텐츠 제작)	Firefly 앱 구독 모델 도입, 신규 구독 모델을 통한 추가 매출 창출 기대 Acrobat AI Assistant 에 추가 언어 지원, 법률 문서 및 계약서 분석 기능 추가 FY25 말까지 AI 매출 두 배 성장 목표
Workday (HR 및 재무관리)	AI Agent 신제품 올해 하반기 다수 출시 계획 채용 기반 AI (Recruiting Agent) 전분기 대비 신규 계약 건수 2 배 성장 앱 개발 플랫폼 (Extend pro) AI 기반 앱 개발 효율성 50% 이상 향상, 전분기 대비 계약 2 배 성장
TTD (디지털 광고 최적화)	AI 기반 UID2 (유저 데이터 보호 시스템) 확산 AI 기반 광고 최적화 엔진 Kokai 도입, 광고 성과 개선 AI 기반 광고 공급망 정화 Sincera 인수, 광고 효율성 증가

자료: 산업자료, SK 증권

5-2. 선결 과제 해소가 필요한 B2B

B2B AI 도입 지연 원인
1) 높은 기술적 요구도
2) 고객사 준비 미흡

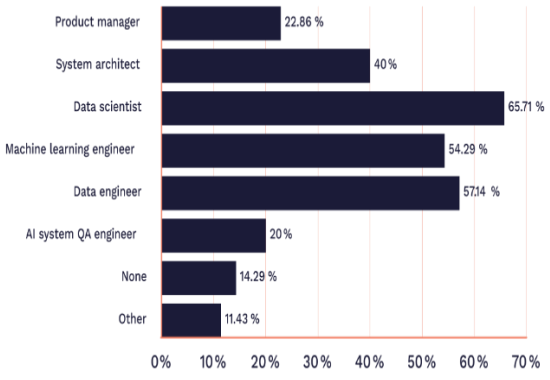
B2B AI 서비스의 본격적인 확산을 위해서는 두 가지 기술적·산업적 과제가 선결되어야 한다. 첫째, AI 모델의 고도화가 필요하다. 기업 데이터를 무리 없이 통합할 수 있는 RAG 기술의 개선과 함께, 모델의 신뢰도(Reliability) 제고가 요구된다. 특히, AI 가 단순 응답형 챗봇을 넘어 Agent 형태로 진화할 경우, 모델의 판단 오류 가능성은 기업 운영에 직접적인 리스크로 작용할 수 있다. Agent 는 현재 Reasoning 기반 모델보다 긴 context length 를 토대로 반복적인 Chain of Thought 사고를 이어갈 것으로 보인다. 합성 데이터, 시뮬레이션 상황으로 이내 개선될 것으로 판단되나 단기적인 시간 내에서는 어렵다.

둘째, 고객사 측의 도입 준비 부족이다. 대표적인 문제가 Data Silo 와 Migration 복잡성이다. IT 시스템 도입 컨설팅사인 IBM 은 2025 년 3 월 Morgan Stanley 컨퍼런스콜에서 "기업 데이터의 99%가 아직 AI 에 활용되지 못하고 있으며, AI 전환을 위한 노력의 80% 이상이 Data 를 준비하는데 사용 중"이라 진단했다. IBM 고객사 중 실제 생성형 AI 프로젝트를 운영하는 기업은 26%에 그친다. Accenture 역시 최근 실적발표에서 클라우드 마이그레이션과 데이터 코어 미비는 명확한 구조적 병목 요인으로 꼬집었으며 기업들마다 준비된 정도가 천차만별이라고 언급하였다. AI 관련 인력 부족 문제까지 고려하면, 해당 병목은 단기간 내 해소되기 어려울 전망이다.

AI 도입을 위해선 Silo 로 분산된 데이터를 통합할 수 있어야

자료: 산업 자료, SK 증권

AI 도입 기업들의 실제 채용 직군, 데이터를 다루는 인력 위주 채용



자료: 산업 자료, SK 증권

5-3. CPU 에서는 CSP 가 독점, GPU 에서는 지형이 바뀐다

CSP 비즈니스 변화
1) AI 로 전체 시장 증가
2) 엔비디아의 IaaS 가치 침투
3) AI 기업들의 SaaS 가치 침투

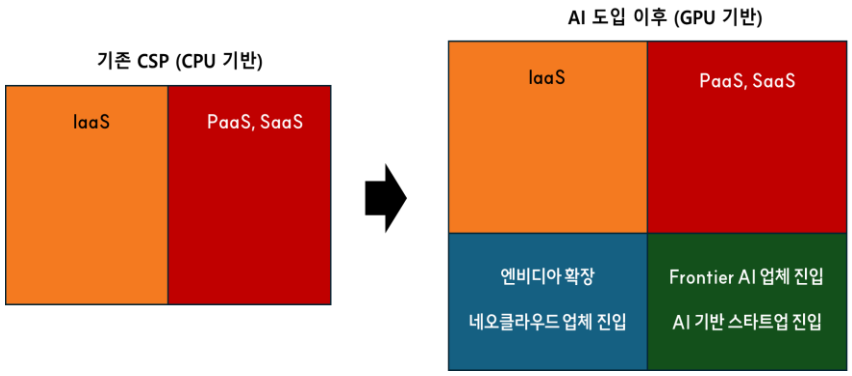
AI 모델 등장으로 기업들이 더 많은 컴퓨팅을 수요하게 됐다. 기존 On-premise 를 운영하던 기업들이 서비스 필요를 느껴 Migration 도 늘어나는 추세다. CSP 시장은 현재 호황이다.

그러나 GPU 기반으로 데이터센터가 전환되면서 CSP 산업에도 구조적 변화가 나타나고 있다. 기존 CSP 산업은 두 가지 주요 부가가치를 중심으로 성장해왔다. 첫째, 기업들이 자체적으로 구축해야 했던 서버 인프라를 대신 제공함으로써 고정비를 변동비로 전환시켜주는 역할이다. 둘째, 서버 기반 서비스를 통해 각종 데이터 관리, 보안, 생산성 향상 등을 포함한 비즈니스용 SW 를 추가로 제공한다.

전통적 CSP 사들은 부품사와 서비스사 양측에서 모두 영역을 침범 받는 중이다. 엔비디아는 전통적인 부품 공급자에서 벗어나 IaaS 로의 역할로 확대하려는 움직임을 보인다. 최근 Blackwell 세대부터는 랙 단위 제품인 NVL72 의 사양이 HGX 와의 격차를 보이기 시작했다. 차세대 Rubin 플랫폼에서는 500 개 이상의 GPU 가 탑재된 랙 단위 시스템이 등장할 예정이며, 이는 기판 단위와 랙 단위의 성능 편차를 심화한다. 또한, 엔비디아가 새롭게 선보인 Dynamo 는 데이터센터 전체의 연산 효율을 향상시키는 역할을 수행하며, 기존 CSP 의 핵심 역량 중 하나였던 전력·냉각 최적화까지 대체하는 모습이다. 엔비디아로부터 완성형 인프라를 도입하는 구조로 전환된다면 밸류체인 내 일부 수익이 이탈할 수 있다.

AI 모델 호스팅으로 인해 SaaS, PaaS 사업의 부가가치도 감소한다. CSP 들의 AI 경쟁력은 어떤 고성능 AI 모델과 계약되어 있느냐에 집중된다. GCP 는 자사 모델인 Gemini 를 기반으로 독립적인 경쟁력을 확보했으나, Azure 와 AWS 는 각각 OpenAI, Anthropic 에 의존하고 있는 상황이다. 최근 빠르게 늘어나고 있는 AI 모델 스타트업들의 성장세를 고려할 때 이러한 흐름은 더욱 심화될 가능성이 높다.

AI 도입으로 인한 CSP 산업 구조 변화



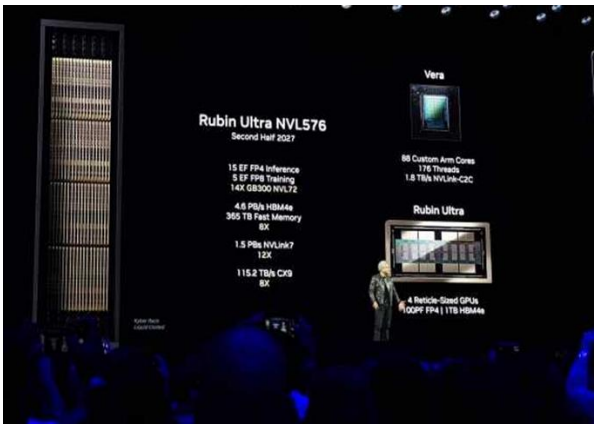
자료: SK 증권

같은 Blackwell 도 냉각 방식, 연결 방식으로 NVL72 성능이 더 우위

GPU Model	NVL72 Blackwell GPU	HGX Blackwell GPU
FP4 Tensor Core (FLOPS)	20	18
FP8 Tensor Core (FLOPS)	10	9
메모리 대역폭	8 TB/s	7.7 TB/s
메모리 용량	186GB	180GB
TDP	1200W	1000W

자료: 산업 료, SK 증권

GTC2025 를 통해 NVL 576 단까지 공개



자료: NVIDIA, SK 증권

CSP 4 사 주요 경쟁력

기업명	시장 점유율	주요 고객	주요 경쟁력	AI 경쟁력
AWS	33%	전 산업군 대상. 스타트업 부터 대기업까지 광범위	글로벌 최대 커버리지 및 긴업력 다양한 서비스 포트폴리오 및 고객으로 높은 안정성	Anthropic·Stability 등 다수 파트너 Trainium·Inferentia 등 자체 칩 활성화 지속
Azure	20%	대기업, 정부기관, Microsoft 제품 사용 기업 위주	Office, Windows, GitHub 등과 통합된 SaaS 연계 력	OpenAI 와 독점 파트너십 Copilot 제품군과의 추가 연계 기대 가능
GCP	10%	개발자, AI 스타트업, 연구 기관, 데이터 사이언스 중심	오픈모델 다양성, Workspace 및 구글 기존 서비스와 연계 위즈 인수로 보안 위주 멀티클라우드 시장 겨냥	PaLM2, Gemini, Anthropic, Cohere 접근권 TPU 가 가장 활성화된 ASIC 으로 평가
OCI	3%	Oracle DB, ERP 고객	AI 용 설계, 점유율 확장 정책으로 낮은 단가 제시 중	NVIDIA DGX Cloud 직접 호스팅 OpenAI 와 Stargate 진행

자료: 산업자료, SK 증권

네오클라우드와 CSP 차이점

	전통 CSP	네오클라우드
주요 기업	미국 CSP 4 사, 중국 CSP 사 등 대형 기업 위주	Coreweave, Lambda DB 등 스타트업 위주
GPU 제공 방식	HGX 기반 인프라 구성, 자체 서버 커스터마이징 중심 기존 정립된 부품에 GPU 장착, 범용성 위주	DGX 위주 서비스 NVLink, NVSwitch, Dynamo 등 엔비디아 최신 기술 적용 가능
서비스 범위	IaaS + PaaS + SaaS 까지 아우르는 통합 서비스 제공 (예: SageMaker, Azure ML, Vertex AI, BigQuery 등)	GPU 인프라(IaaS) 중심, 고성능 컴퓨팅 능력 위주 제공 Coreweave 의 경우 DB 없이 GPU 만 제공 (GPUaaS)
주요 고객	대형 기업, 인터넷 기업 위주	초대형 생성형 AI 스타트업(LLM 중심) 단독 제공보다는 멀티클라우드 개념으로 일부 제공
특징	리전 기반으로 정해진 할당량, 대형 고객 사전 예약 위주 정해진 인스턴스 스펙 으로 제한적 커스터마이징	단기 임대 위주, 커스텀 구성 가능, AI 훈련 최적화 구조

자료: 산업자료, SK 증권

향후 AI 산업 전개에 따른 빅테크 기업 수혜 정도

자료: SK 증권
주: 좌측 기업이 수혜강도 더 강함, Ondevice 는 B2C Agent 에 포함

Appendix: AI 모델 기본기

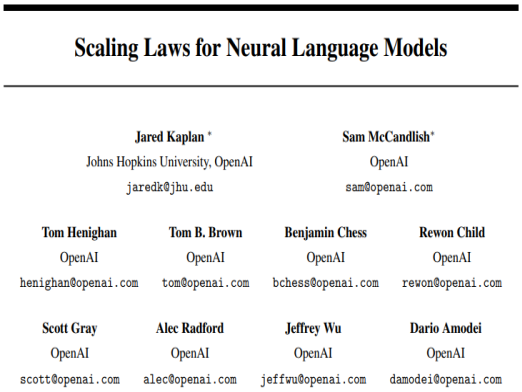
현재 자본 시장의 중심인 AI는 무엇인가?

AI는 꽤 오래된 개념이지만, 생성형 AI(Generative AI)는 비교적 최근에 알려졌다. 알파고 등장 이후 AI 는 언론에 자주 등장했지만 생성형 AI 는 ChatGPT 의 등장 이후로 주로 언급됐다. 현재 시장성을 인정받고 사람들이 열광하는 AI 는 생성형 AI 라고 할 수 있다. 생성형 AI 는 '토큰(Token)'을 생성하는 AI를 의미한다. Nvidia 가 지속적으로 자신들을 AI Factory 라고 칭하는 것은 이 '토큰'을 생성하는 공장이 되겠다는 설명이다.

생성형 AI 는 Transformer 아키텍처로 시작한다. 생성형 AI 모델들은 대부분 Transformer 구조를 기반으로 한 파생 모델들이다. 따라서 해당 아키텍처의 기본적인 사항들을 이해해야 AI 확장의 방향성과 속도를 가늠해볼 수 있다. 해당 구조가 처음 소개된 것은 2017년 Google Brain 팀에서 발표한 'Attention is all you need' 논문이었다. 해당 논문은 2025년 현재 기준 17만회 이상 인용된 논문으로 AI 분야에서 가장 인용 횟수가 많은 논문이다. 가장 유명한 챗봇인 Open AI 의 GPT 는 Generative Pre-trained Transformer 의 약자이다.

논문 발표 당시까지만 해도 Transformer 는 딥러닝 모델의 한 유형일 뿐이었다. 그러나 병렬 학습이라는 특징 때문에 GPU 가속과 결합되면서 성능이 개선됐다. 2022년 ChatGPT-3.5 는 가장 복잡한 데이터로 여겨졌던 '언어 데이터'를 해석할 수 있어지면서 모델이 문맥을 이해하는 것이(이해하는 것처럼 보이는 것이) 가능해졌다. Transformer 의 등장은 AI 가 체스, 바둑에 이어 언어까지 공략하기 시작하는 지점이었다. OpenAI 의 연구원진은 2020년 'Scaling Laws for Neural Language Models' 논문을 통해 더 많은 컴퓨팅과 데이터로 언어 모델을 공략해나간다는 방향성을 잡았다

Transformer, LLM, 생성형 AI의 시작을 알린 논문



Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

자료: Archive, SK 증권

이상적인 모델은 모델 사이즈를 강하게 늘려야한다

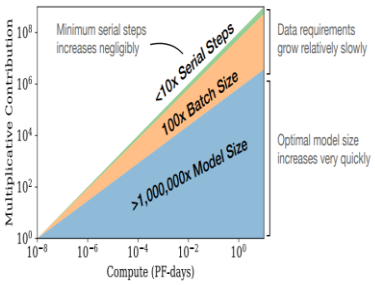
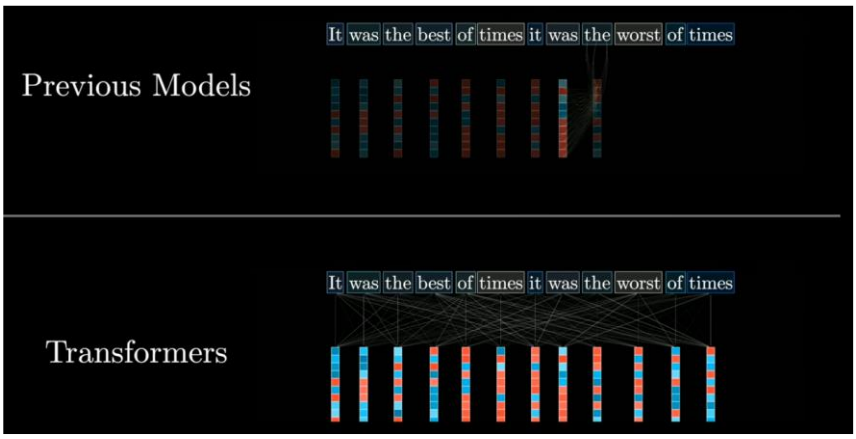


Figure 3 As more compute becomes available, we can choose how much to allocate towards training larger models, using larger batches, and training for more steps. We illustrate this for a billion-fold increase in compute. For optimally compute-efficient training, most of the increase should go towards increased model size. A relatively small increase in data is needed to avoid reuse. Of the increase in data, most can be used to increase parallelism through larger batch sizes, with only a very small increase in serial training time required.

자료: Archive: Scaling Laws for Nueral Language Models, SK 증권

Transformers는 이전 ML 모델과 다르게 병렬 처리 가능



자료: 3Blue1Brown, SK 증권

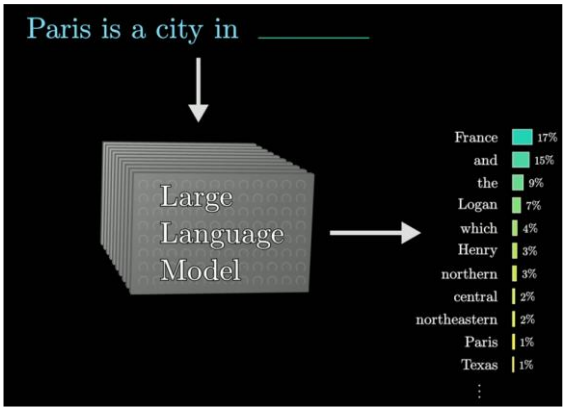
Transformer의 기본 구조

Transfromer 아키텍처는 Vectorize 과 Self-Attention 두 가지 주요한 특성을 가진다. Transformer 는 1) Vectorize 로 input data 를 Tensor 로 치환하고 2) 치환된 Tensor 를 Attention 모델을 통해 계산하여 다음 단어를 예측하는 행위를 반복하여 문장을 만들어 내는 함수로 단순화할 수 있다.

AI 모델에서는 입력 토큰(단어)를 계산 가능한 단위로 치환해주어야 한다. Transformer 는 단어를 행렬 값(Tensor)으로, 단어의 위치를 Position Encoding 값으로 바꾼다. 이 과정을 Embedding 이라고 부른다. 고차원 행렬 값에 단어 정보를 담아 한 단어에 많은 정보를 담을 수 있게 된다. 기존 인터넷 포털 검색은 Vlookup 엔진을 기반으로 하고있어 1 대 1, 또는 1 대 N 수준의 대응이 이루어진다. 같은 검색을 하더라도 생성형 AI 에서는 문맥을 이해한다는 느낌을 주는 이유가 기본적으로 더 많은 정보량을 다루기 때문이다.

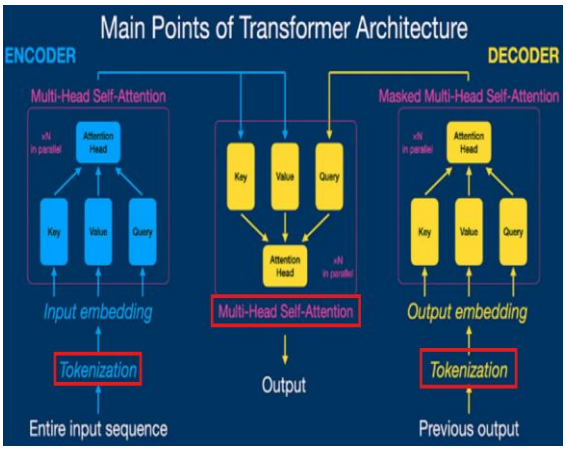
Vectorize 된 단어들은 Self-Attention 을 통해 문맥과 의미를 반영한 벡터로 변환된다. 같은 단어라도 문장 내 위치와 전후 관계에 따라 벡터값이 달라진다. 이러한 벡터들을 병렬적으로 연산하면서 AI 모델은 모든 단어 간의 관계를 학습하고, weight 와 bias(모델의 가중치)를 최적화해 '언어의 지도'를 만들어낸다. 이후 추론 과정에서는 훈련된 weight 를 활용해 Self-Attention 과 피드포워드 네트워크를 빠르게 계산하여 결과를 생성한다. 훈련 과정에서는 높은 정밀도의 FP16 을 사용해 weight 를 미세 조정하지만, 추론 과정에서는 이미 학습된 weight 를 사용하기 때문에 FP8 이나 FP4 같은 낮은 정밀도로도 계산이 가능한 이유다.

LLM 은 Transformer 를 통해 다음 단어의 예측을 가능하게 만드는 함수



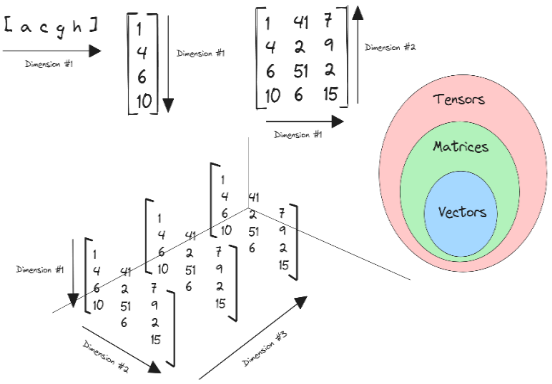
자료: 3Blue1Brown, SK 증권

Transformer 는 tockenize + self attention 구조



자료: 산업자료, SK 증권

N곱하기 N의 Tensor 구조

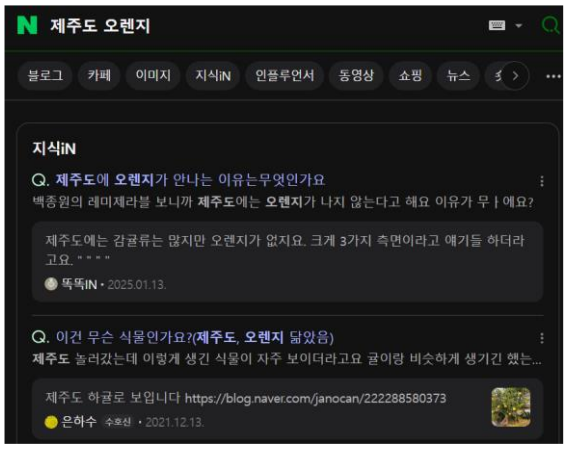


자료: MediumAI, SK 증권

Attention은 각 단어의 벡터값을 조정하는 일련의 과정

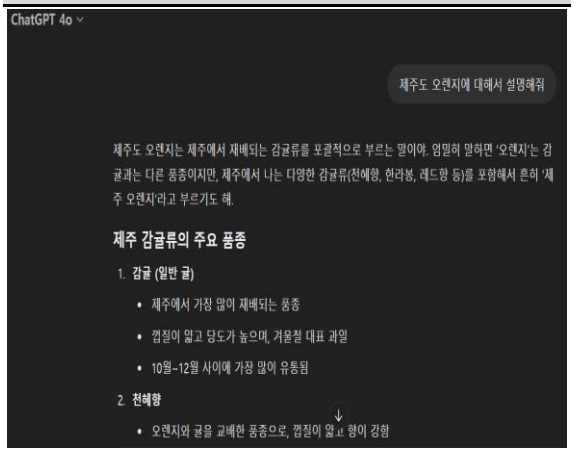
자료: 3Blue1Brown, SK 증권

네이버 검색은 Vlookup 기반 키워드 매칭



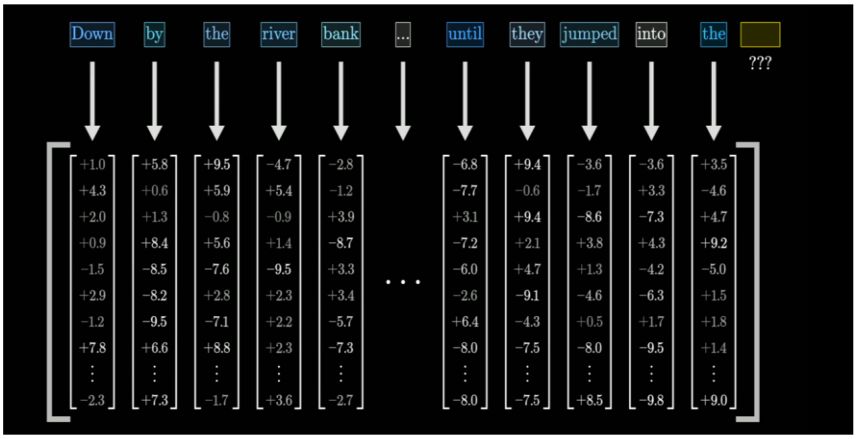
자료: 네이버, SK 증권

생성형 AI는 제주도라는 맥락, 유사한 과일류라는 개념까지 고려하여 답변



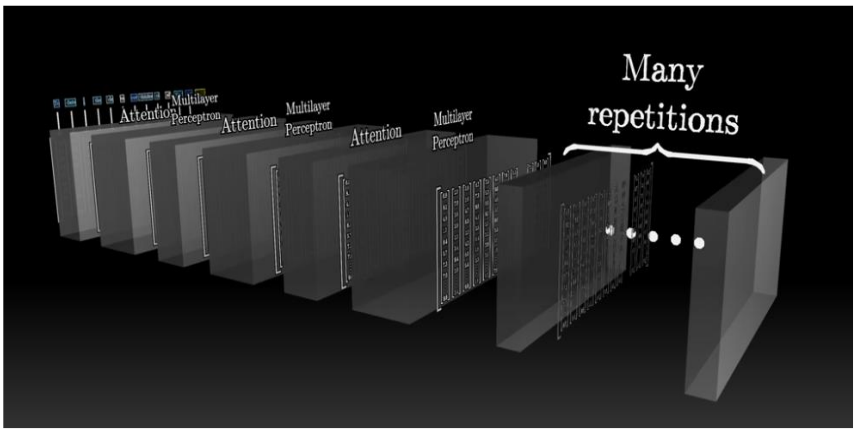
자료: ChatGPT, SK 증권

다차원 Vectorize 를 통해 단어에 많은 값을 부여



자료: 3Blue1Brown, SK 증권

Vectorize 된 단어를 여러번 Attention 하여 다음 값을 계산



자료: 3Blue1Brown, SK 증권

AI 모델의 고도화

훈련(Pre-training) 고도화

1. Parameter 증가

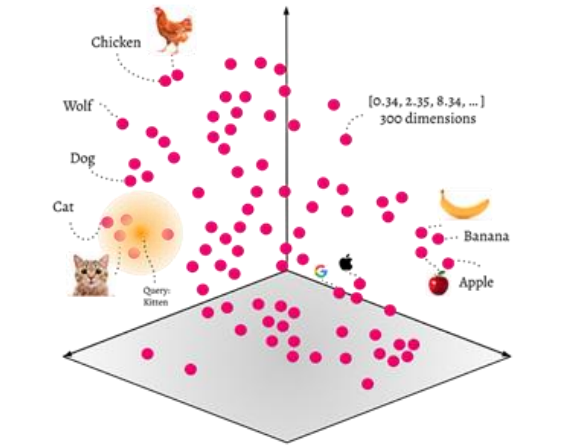
구조에서 볼 수 있듯, Vectorize 을 정교하게할 수록 모델 정확도에 유리하다. 한 단어에, 한 관계에 더 많은 정보를 담을 수 있게되기 때문이다. 최초의 Transformer 인 BERT 모델은 768 차원으로 단어에 값을 부여했으나 처음 상용화된 ChatGPT 의 기반 모델인 GPT-3.5 는 1 만차원 이상으로 임베딩됐다. 현재 주요 모델들은 이의 수 배 수준의 vectorize 가 이루어졌을 것으로 추정된다. LLM(Large Language Model)의 성능 지표로 거론되는 Parameter 는 Self-attention 과정에서 생성된 Weight 와 Bias 들이 저장된 용량이 대부분을 차지한다. 따라서 Vectorize 가 자세히 될 수록 Parameter 가 커지고, 미리 계산해놓은게 많은 크고 무거운 AI 모델이 된다.

Vectorize 뿐 아니라 모델의 Layer 수 증가도 모델 정확도를 높인다. 깊은 신경망일수록 복잡한 패턴을 학습할 수 있어진다. 이 외에도 Attention Head 증가, FFN 크기 증가 등 Transformer 구조 내에서 많은 부분들의 역량을 키움으로써 모델의 크기와 연산량을 키우는 대신 더 복잡하고 정교한 모델을 지향할 수 있다. 다만 모델 크기에 적합한 수준의 데이터 양이 받쳐줘야하고 연산량의 급증은 결국 추론 비용의 증가로 이어지므로 효율화를 위한 연구가 많다.

2. Attention 효율화

Attention 효율화는 Sparse Attention(불필요한 연산을 줄이는 방식)부터 Flash Attention(GPU 메모리 최적화), Multi-Query Attention(응답 속도 개선) 등의 다양한 기법을 포함한다. 특히 모델 크기가 커질수록 Attention 연산량이 기하급수적으로 증가하는만큼 최적화 기법의 중요성이 커지고 있다. Attention 효율화는 모델의 응답속도와 관련된 개선으로 특히 최근 Grok-3 가 Flash Attention, Multi-Query Attention 으로 응답 속도를 키워 소비자에게 긍정적 피드백을 받는 중이다. 딥시크는 Multi-Head Latent Attention 이라는 새로운 방식으로 모델을 효율화하였다. 해당 방식은 기존의 방식 대비 출력 컨텍스트가 긴 상황에서 80~90%의 메모리를 절약할 수 있다.

Vector 의 차원(dimension)이 증가함에 따라 더 정확한 군집화 가능

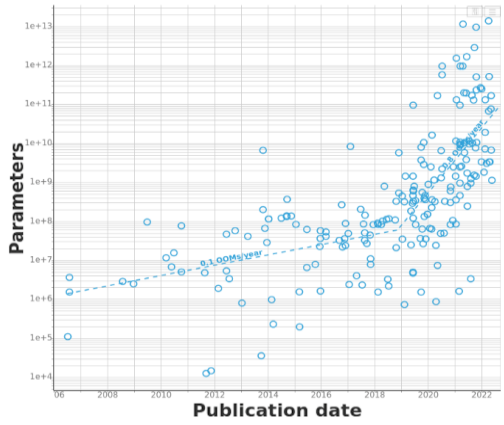


자료: medium.com, SK 증권

다양한 쿼리 처리 방식, MQA 는 응답속도 개선에 유리

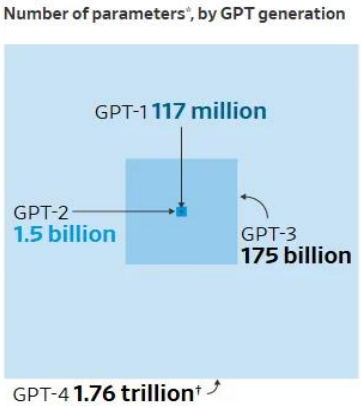
자료: medium.com, SK 증권

주요 LLM Parameter 수 증가



자료: 산업 자료, SK 증권

GPT Parameter 증가량



자료: 산업 자료, SK 증권

3. Mixture of Experts (MoE)

여러 개의 전문가(Experts) 네트워크 중 일부만 활성화하여 연산량을 줄이는 구조를 의미한다. 기존 Transformer 모델은 GPT-3 의 경우 175B 의 Weight 와 Bias 로 모델 규모가 결정됐고, 이들을 추론 과정에서 모두 검토해야했다. MoE 는 이 파라미터를 수십개의 구획으로 나누고, 추론할 때 필요한 구획만 활성화하는 알고리즘 기법이다. 기존 Parameter 를 몇개의 전문가 집단(Experts)로 나누는 기술과 시퀀스에 따라 어떤 전문가를 활성화할지 결정하는 Gating Network 기술을 필요로 한다. 예를 들어 전문성 별로 30 개의 Expert 를 나눠놓은 모델에서 의학 관련 질문이 들어오면 Gating Network 가 의학, 약학 2 개의 구역만 활성화하고 나머지는 활용하지 않는 구조다. 전체 모델 크기는 커도, 실제로 사용되는 연산량은 적어져서 비용이 절감시킬 수 있다.

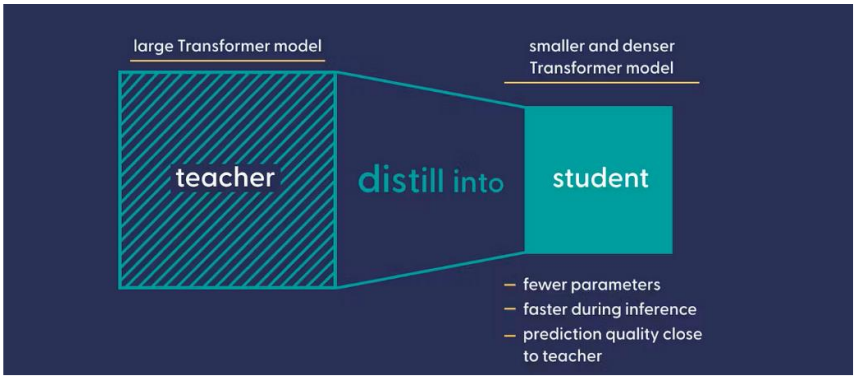
MoE 구조는 일반 Transformer 보다 복잡한 구조라서 훈련이 더 까다롭고, 전문가 간 불균형 문제(load balancing issue)가 발생할 수 있다. 다중 전문가 호출이 발생할 경우 일반 LLM 대비 추론 속도가 오히려 느려지는 문제도 발생 가능하다.

4. Distillation (모델 증류)

대형 모델의 지식을 보다 작은 모델로 압축 이전하는 기술이다. 교사(teacher) 모델로부터 나온 풍부한 예측 분포를 학생(student) 모델이 학습하도록 하여, 작은 모델이 큰 모델의 성능을 모방하도록 만드는 방법이다. 예를 들어 GPT-3 175B 모델이 생성한 답변들을 사용해 10 억 규모 모델을 학습시키면, 해당 작은 모델이 GPT-3 의 지식을 상당 부분 내재화하여 추론 속도는 훨씬 빠르지만 정확도는 준하게 나올 수 있다.

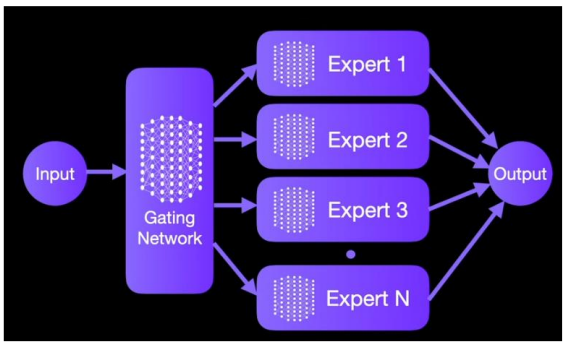
지식 증류를 거친 학생 모델은 매개변수 수가 크게 줄어들어 훨씬 경량이므로 운영 비용 또는 지연 시간 면에서 유리하다. 또 엣지 환경처럼 제한된 자원에서 LLM 을 구동하거나, 대량 트래픽 처리시 비용을 줄이는 데 유용하다. 예시로, Stanford 의 Alpaca 모델은 텍스트다빈치 003 (GPT-3.5 급)의 출력으로 70 억 파라미터 LLaMA 를 미세조정하여, GPT-3.5 의 상당한 능력을 작은 모델에 구현해 주목받은 바 있다. Google 도 LaMDA 등을 증류해 모바일 기기용 Bard 의 경량 버전을 만드는 등 실제 응용에 활용하고 있다. 모델 증류는 기술 개념상 신규 소형 모델의 훈련 과정에서 사용되는 기술이지만, 사용 목적은 추론 비용의 감소에 있으므로 추론 관련 기술로 분류하였다.

Distillation 을 통해 적은 더 가벼운 성능의 모델 달성 가능



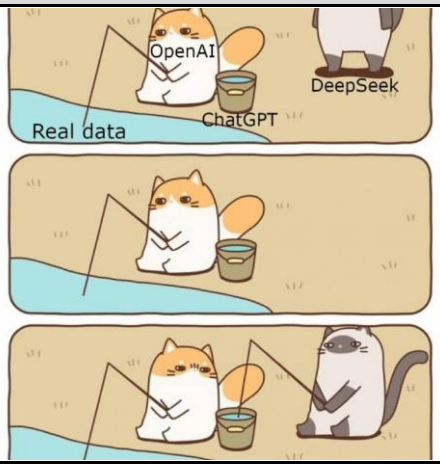
자료: TowardsAI, SK 증권

훈련 데이터를 분산하는 Mixture of Expert 구조



자료: TowardsAI, SK 증권

Distillation 을 통한 Teacher 모델 기생



자료: X, SK 증권

후처리(Post-training) 고도화

Post training 은 Pretraining 이후에 모델의 행동을 조정하는 과정을 의미한다. 대표적으로 Fine-tuning 이 있는데, 데이터셋으로 모델을 다시 학습시켜서 특정 도메인이나 작업(Task) 성능을 높이는 방식이다. 예를들면 법률 문서에 특화된 챗봇을 만들기 위해 법률 문서로만 추가 학습하는 방식이 있을 수 있다.

1. RLHF (Reinforcement Learning with Human Feedback)

사람이 모델의 응답을 평가하거나 순위를 매긴 피드백을 활용해 모델을 개선하는 방식이다. 모델이 여러 응답을 생성하고 사람이 어떤 응답이 좋은 지를 순위를 주는 형태로 피드백을 한 후 그 보상을 기준으로 강화학습을 수행한다.

2. RLVR (Reinforcement Learning with Verifier Rewards)

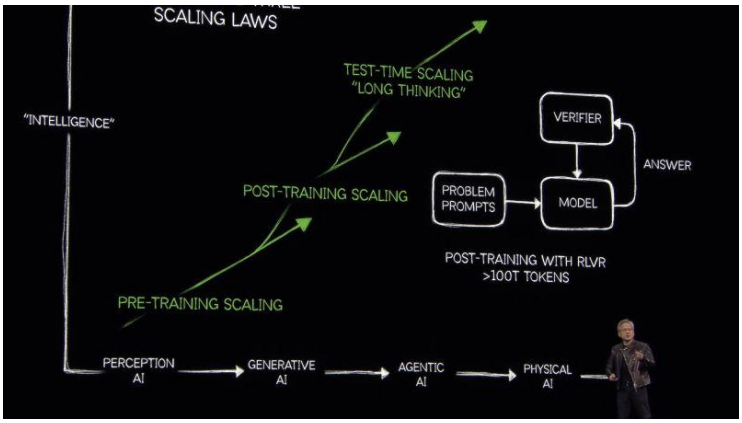
RLHF 와 유사한 개념이나, 사람 대신 자동화된 검증기(Verifier)가 보상을 주는 방식이다. 딥시크가 LLM 에 적용한 순수 강화학습의 경우 사람의 피드백(HF) 없이 특정 규칙을 미리 만들어 두고 이런 환경 하에서 강화학습을 이룬 경우로, 피드백을 줄 Teacher 모델이 있는 경우에 가능한 형태이다. 이는 상위 모델의 정답을 복사하는 Distillation 과는 구분되는 개념으로, Distillation 은 단순히 지식의 압축 이전을 통해 정답을 복사하는 것이라면 해당 개념은 정답 학습에 상위 모델을 활용한다. 딥시크 R1-Zero 의 경우 RLVR 의 컨셉을, Distill 모델들은 증류의 컨셉을 활용한다.

3. Retrieval-Augmented Generation (RAG, 검색 증강 모델)

모델 자체 지식에만 의존하지 않고, 추론 중에 외부 지식 소스나 도구를 활용하는 방식을 의미한다. RAG 을 통해 Fine tuning 을 이룬 특화 모델을 생성할 수 있다. 특정 분야에 특화된 DB 제공으로 소형 모델을 활용한 법, 금융, 수학과 같은 특화된 모델을 만드는 것이 가능하다. B2B AI Agent 로 하여금 회사 데이터를 연결하여 특정 기업 데이터에 최적화된 추론 기능 제공도 RAG 로 가능해진 기술이다. 수 많은 유형의 AI agent 가 출시되고 있는데, 이들 모두 데이터셋을 달리하는 것이 핵심이다. Gemini 2.0 에서는 지도, 검색, 유투브 등과 연결하여 분석을 특화하는 Agent 를 선보이고 있다. B2C 로 개인화된 비서 서비스 제공에도 RAG 는 필수적이다. 개인 이메일, 연락처, 사진 등을 참고하여 모델이 질문에 답변할 수 있게끔 하여야 진정한 비서 기능이 가능하다.

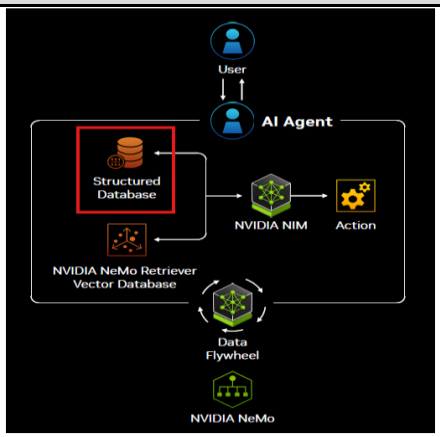
RAG 역시 모델의 성능을 크게 개선시켜주지만 모델의 운영 비용을 늘린다. 1) 추가적인 DB 검색 2) 검색된 데이터 처리 3) 기존 LLM과 처리 데이터의 결합 과정 에서 메모리와 연산에 대한 필요가 증대된다.

RAG 를 활용하여 유저 데이터 기반의 답변 가능



자료: NVIDIA, SK 증권

외부 데이터를 연결하는 구조를 가진 엔비디아 Agent 알고리즘 NIM

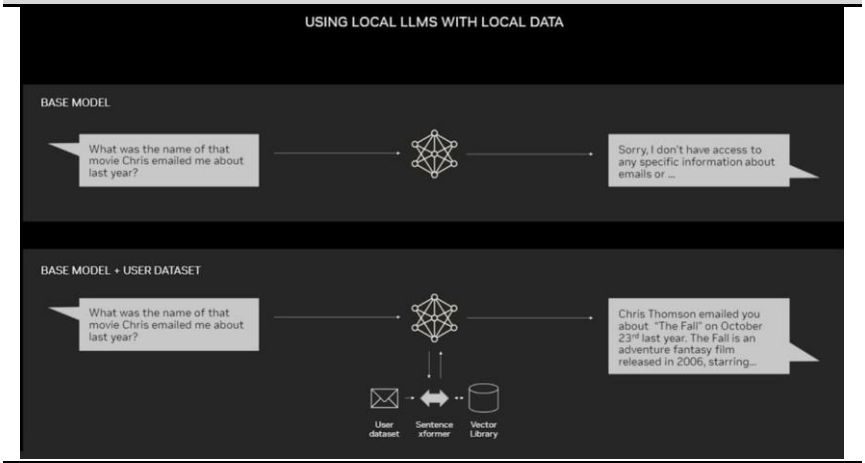


자료: NVIDIA, SK 증권

Agent 로의 발전

자료: 산업 자료, SK 증권

RAG 를 활용하여 유저 데이터 기반의 답변 가능



자료: NVIDIA, SK 증권

추론(Inference) 과정

Reasoning Model (추론 모델)

모델이 충분한 고민(Reasoning) 이후에 답변을 하게끔 모델을 설계하는 방식이다. CoT (Chain of Thought), Test time scaling, Self Consistency decoding 등의 기법을 합쳐 구현한다.

CoT 는 모델이 문제를 풀 때 중간 추론 단계를 여러 단계로 나눠서 표현하도록 유도하는 프롬프트를 내제화하는 방식이다. 초기 LLM 부터 몇 가지 연쇄 사고 시범 사례를 프롬프트에 제공하고 나면 같은 질문에 대한 output 이 개선되는 현상이 존재했다. 이를 모델 자체에 내제화한 첫 시도가 o1 이다. Test-time Scaling 은 모델 추론 시 특정 조건이 충족되면 연산 자원을 추가 투입하게끔 설계한 모델을 의미한다. 모델의 검증 과정을 길게하고 CoT 기법을 활용해 답을 다듬는 방식을 도입한다. OpenAI 는 향후 모델들에서는 모델이 스스로 질문의 추론 단계를 판단하고 사고하여 답변하는 모델(compute optimal)을 구상 중이라고 언급한 바 있다. Consistency decoding 은 CoT 방식을 다양한 경로로 여러번 시행하고 가장 일관적으로 나오는 답을 최적의 답변으로 제시하는 모델로 CoT 방식의 심화 적용이라고 생각할 수 있다.

Reasoning 의 도입은 AI 모델의 하극상을 가능하게 한다. 훈련 단계에서 많은 자원을 투입한 고성능 AI 모델이 아니더라도 추론 단계에서 여러번의 사고를 거치게 되면 목표 benchmark 지수를 달성할 수 있다. PaLM(54B) 파라미터 모델에 CoT 프롬프트를 적용했더니 GSM8K 수학 문제에서 57%의 정답률로 당시 SOTA 를 달성했고, 이는 별도 미세조정 과정을 거친 GPT-3(175B) 모델보다도 높은 성능이었다. o1 모델을 처음 시도한 openAI 의 Noam Brown 연구원은 바둑 AI 인 알파고 역시 사고 시간을 주지 않았을 때와 사고 시간을 주어졌을 때 모델 성능의 차이가 매우 컸다고 얘기한다. 컴퓨팅할 시간을 늘리면 AI 는 성능이 좋아지는 셈이다.

반대급부도 존재한다. CoT 적용 모델은 답변 생성 과정에서 많은 토큰을 생성하고 output 은 요약해서 줄여주는 방식인만큼 컴퓨팅이 많이 소요된다. 이는 모델 운영 비용을 크게 증가시킬 수 있다. 특히 모델이 기억해야하는 Context Legnth 의 증가로 Batch size(한번에 몇개의 서버에 구동할 수 있는지)가 감소하여 토큰 당 비용을 증가시켜 비용 문제를 중첩시킨다. Batch 뿐 아니라 기억해야하는 KV-Cache 도 증가하게 되는데, 이는 메모리 비용 증가로 이어진다.

LLM은 같은 질문에 대해서도 순차적 Prompt를 통해 답변 유도 가능

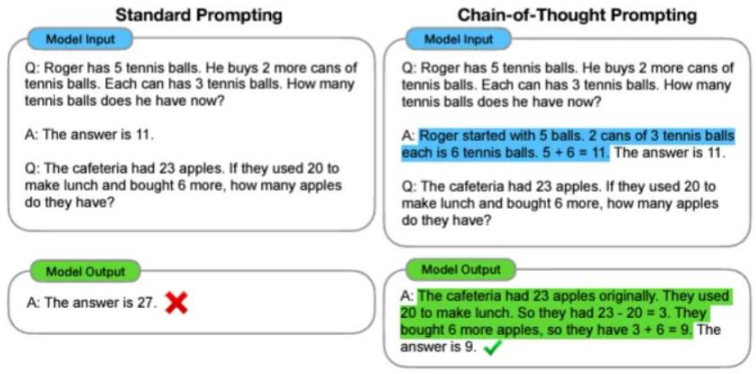
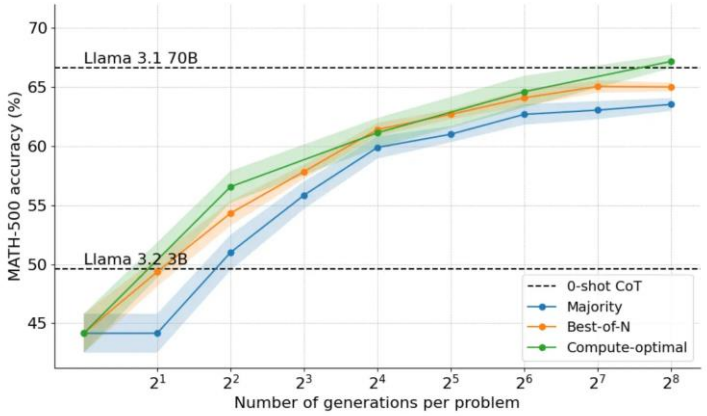


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

자료: Google, SK 증권

문제당 연산 과정을 늘리면 3B 모델로 70B 모델의 Benchmark 역전 가능



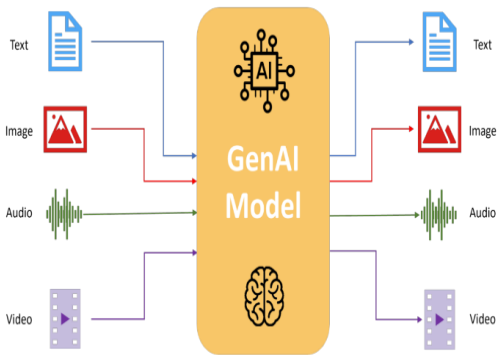
자료: Semianalysis, SK 증권

AI 모델의 확산 - Multi modal 부터 Physical AI 까지

Transformer 는 자연어 처리를 위해 개발됐지만 Multi-Modal 모델들도 Transformer 구조에 기반한 모델들이 많다. 이미지 생성 모델인 Dall-E, 영상 생성 모델인 Sora 모두 이미지, 영상을 벡터화하여 Transformer 기반으로 작동된다. 단백질 구조 예측 모델로 노벨상을 수상한 AlphaFold 의 경우에도 CNN 기반이었던 과거 모델과 다르게 최신 AlphaFold 3 모델에는 Transformer 구조 추가된 것으로 평가 받는다. Transformer 로 인해 벡터로 치환할 수 있는 모든 단위가 더 수월하게 계산 가능해진 것이다. Multi-modal 은 Transformer 모델의 상품화에 큰 도움을 줄 것으로 보인다. 현재 LLM 은 챗봇과 일부 생산성 증가 수준에서 상품화가 이루어지고 있는 것에 비해 사진과 영상 관련된 Agent AI 가 다수 나타나고 있다.

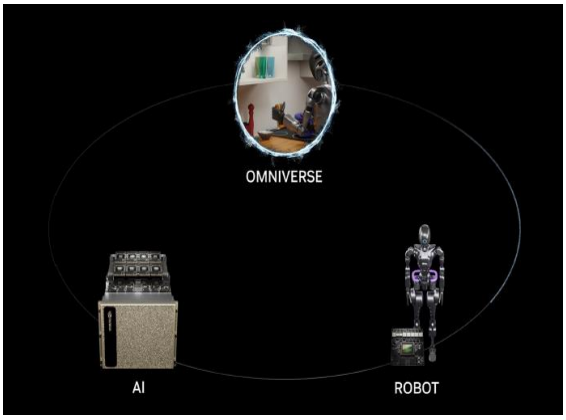
Nvidia 에서 얘기하는 Physical AI 개념도 움직임과 주위 환경을 벡터화하는 것에서 출발한다. 나아가 Physical AI 시스템은 현실 세계의 센서 데이터가 가지는 애매함, 물리 환경의 예측 불가능성과 이로 인한 행동 결과의 불확실성을 처리해야 한다는 점에서 Multi-modal 중에서도 난이도가 높다. 치환에 성공한 가상 시뮬레이션 환경에서 고도로 정밀한 물리 시뮬레이터가 구동된다는 점, 강화 학습 등의 추가 알고리즘이 필수적이라는 점에서 종합적 고도화가 요구되는 생성형 AI 의 차세대 모델으로 볼 수 있다. Physical AI 는 자율주행차부터 산업용 로봇, 휴머노이드, 물류창고 및 공장에 이르기까지 로봇 시스템에 구현될 것으로 기대된다.

다양한 아날로그 데이터를 AI 모델에 맞게 치환하는 Multimodal



자료: 산업자료, SK 증권

시뮬레이션 모델, Nvidia 의 Physical AI



자료: NVIDIA, SK 증권

2025-04-14

메타 플랫폼스(META)

AI 수혜를 온몸으로 받는 중

SK증권 리서치센터

Analyst
박제민

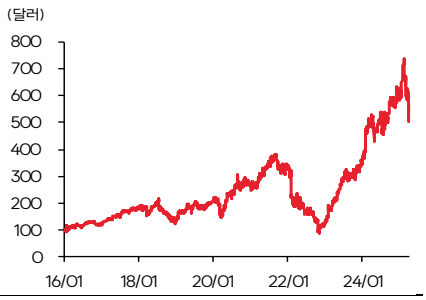
jeminwa@sks.co.kr
3773-8884

기본 정보	
국가	미국
상장거래소	NASDAQ
결산 기준월	12 월
시가총액 (십억달러)	1,347
시가총액 (조원)	1,932.1
현재주가 (달러)	532

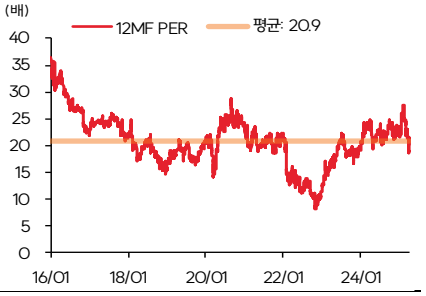
기업 개요 (Bloomberg)

메타 플랫폼스(Meta Platforms, Inc.)는 소셜 기술회사. 동사는 사람들 간의 연결, 커뮤니티 탐색 및 사업 성장을 지원하는 응용프로그램과 기술을 구축한다. 동사는 또한 광고, 증강 및 가상 현실 사업도 운영한다.

주가 추이



12MF PER 추이 및 평균



생성형 AI 활용한 '광고 개인화' 가시권

사용자의 사진, 주변 관계, 선호도 등을 반영한 개인화 광고 제품의 도입을 전망한다. 이에 따라 META 는 광고 제작 회사로 확장된다. META 의 플랫폼 데이터로 경쟁사 대비 차별적인 서비스가 될 것으로 예상된다. Frontier 급 성능으로 출시된 Llama 4 와 DAU 30 억 명 이상의 플랫폼 접근성으로 큰 시너지가 가능하다.

META 의 광고 수익은 CPM(1,000 회 노출 당) 5~20 달러로, 한번 노출에 15 원 전후의 수익이 발생한다. 현재 오픈소스에서 합성 사진 생성은 50~80 원, 영상은 최대 400 원까지 추론 비용이 발생한다. 광고 생성 부가가치 추가, 내부 모델 최적화, 공격적 Capex 로 하드웨어 확보 등을 고려할 때 1~2 년 내 출시가 가능하다.

광고 엔진 수혜 지속

AI 개선으로 광고 엔진 강화 수혜로 인한 수혜가 지속될 전망이다. 광고 전환율이 지속적으로 높아지면서 ASP 상승이 가능하다. 이미 2022 년부터 GPU 기반 피드 개선, 광고 엔진 개선으로 매출이 지속 증가 중이다. 작년 하반기 Applovin 이 GPU 도입을 통해 광고 효율성을 비약적으로 높인 바 있다. META 는 더 풍부한 데이터와 대규모 투자를 통해 장기간 이어갈 가능성이 크다. 현재 광고에 완전한 AI 추천 시스템(Advantage+)을 이용하는 광고 비중은 10%에 불과하다.

침체 우려 유효하나 단기적 매수 기회로 이용 가능

META 는 2025 년 연간 매출 가이던스를 제시하지 않은 가운데 높은 비용 증가(+21% YoY)를 발표했고, 1 분기가 계절적 비수기라는 점에서 단기적으로 영업이익 성장률 둔화가 예상된다. 이에 매크로 불확실성이 더해지자 최근 큰 주가 하락이 발생했다. 그러나 올해 광고 엔진 강화에 따른 매출 및 영업이익 서프라이즈 가능성이 존재하며, 생성형 AI 를 활용한 신규 광고 제품의 확산을 통해 멀티플 상승도 기대할 수 있다. 미국 빅테크 내에서 '막내' 위치에 머물던 META 가 제품화 시대에 '큰형'으로 도약할 것으로 전망한다.

영업실적 및 투자지표 (FY기준)

구분	단위	2021	2022	2023	2024	2025E	2026E
매출액	십억달러	118	117	135	165	188	213
영업이익	십억달러	47	29	47	69	74	85
순이익(지배주주)	십억달러	39	23	39	62	64	73
EPS	달러	14.0	8.6	15.2	24.6	25.0	28.6
PER	배	40	49	32	23	21	18
PBR	배	12	11	9	8	6	5
EV/EBITDA	배	15	10	16	19	14	12
ROE	%	31	19	28	37	30	27

자료: Bloomberg, Consensus(25.04.11)

SK 증권

12 년째 꾸준히 이용자가 늘어나는 플랫폼 강자

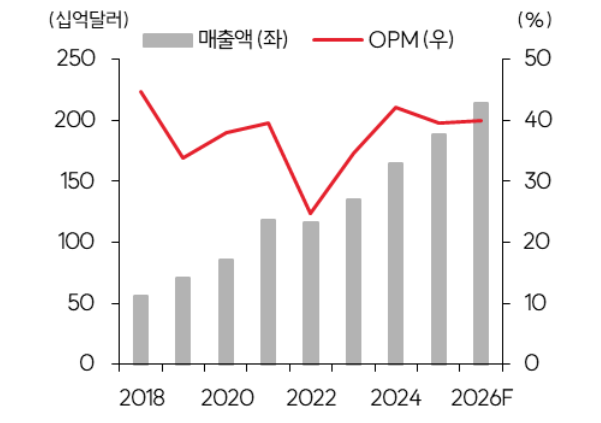
Meta 는 Facebook, Instagram, WhatsApp 을 중심으로 광고 기반 플랫폼을 운영하며, 전체 매출의 98% 이상을 광고에서 창출한다. 광고 단가는 광고 엔진의 정밀도, 사용 편의성, 전환율에 따라 결정된다. 광고 엔진의 성능 증가로 광고주들 선호가 증가하거나 수요가 몰리는 대선, 연말 기간 등에 단가가 오른다.

META 의 광고 노출창구(Inventory)는 지속적으로 변화해왔는데, Facebook Feeds 부터 Instagram 의 Reels, Stories 등 쇼트폼 영상 포맷, 최근에는 Threads 로 이어지는 중이다. WhatsApp 은 아직 본격적인 수익원은 아니지만, Business API 기반의 커머스·결제 확장을 통해 잠재 성장동력으로 키우고 있다. SNS 에는 수명이 있다는 것이 널리 알려진 사실이지만 META 는 공격적인 플랫폼 UI/UX 변경 및 인수합병으로 12 년째 꾸준한 DAU 상승을 보여주는 중이다.

비용 측면에서는 대규모 사용자 기반을 처리하기 위한 데이터센터 인프라가 비용의 대부분을 차지하며, AI 기반 콘텐츠 추천·광고 엔진 고도화를 위한 R&D 투자도 지속된다. Reality Labs 는 AR/VR 하드웨어와 메타버스 플랫폼(Horizon 등)을 개발하는 사업부로, 장기적 기술 주도권 확보를 목표로 하고 있다. 다만 현재는 수익보다 전략적 투자의 성격이 강하며, 2024 년 $17B 이상의 영업 손실을 기록했다.

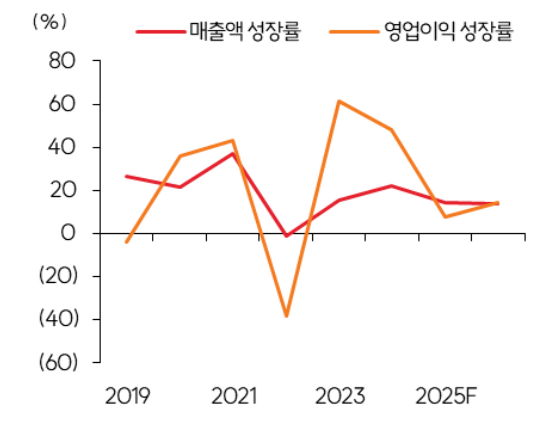
Llama 는 Meta 가 공개한 오픈 소스 기반의 대형 언어 모델(LLM)로, 수익보단 플랫폼 영향력 확대를 위한 전략적 자산이다. 외부 개발자와 기업들이 자유롭게 모델을 fine-tune 하고 활용할 수 있게 해, Meta 중심의 AI 생태계 형성을 유도한다.

매출액 영업이익 추이 및 전망



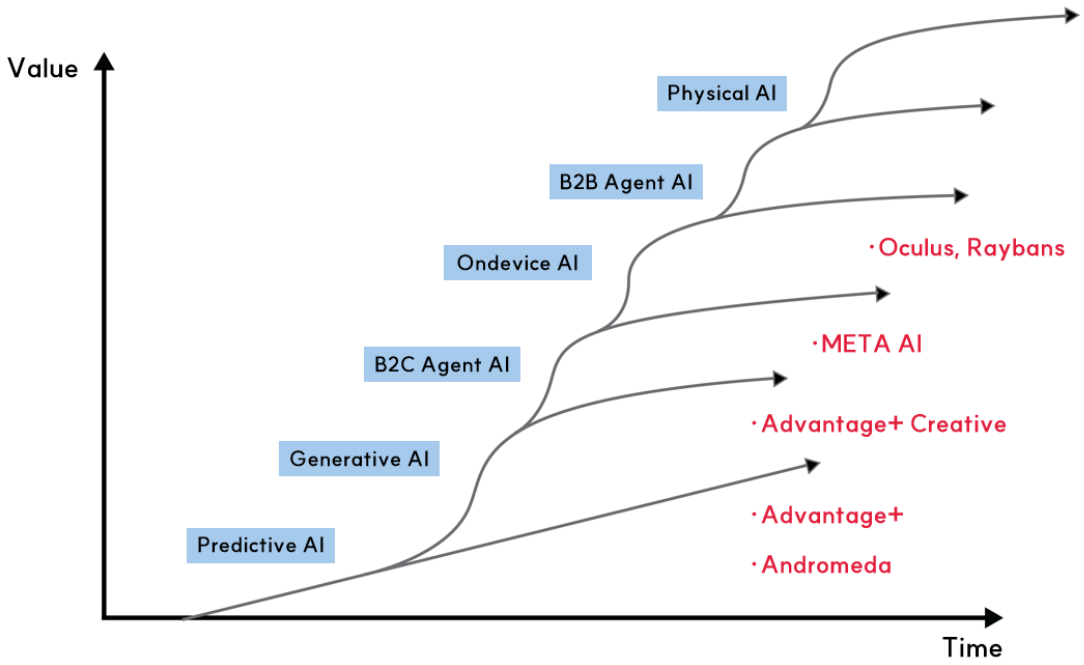
자료: Bloomberg, SK 증권

영업이익 성장률 추이 및 전망



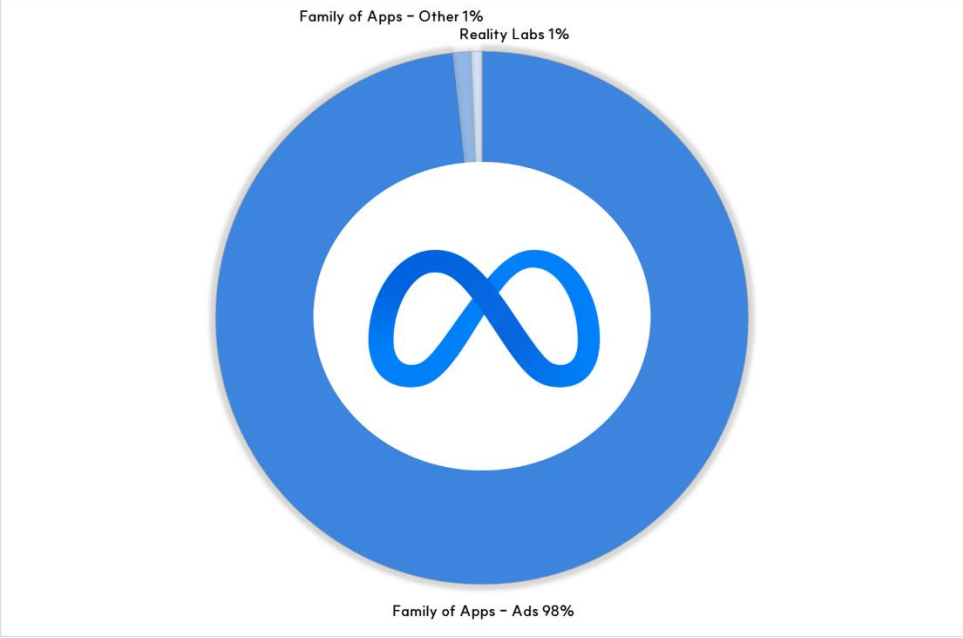
자료: Bloomberg SK 증권

사업부별 AI 단계 포지션



자료: SK 증권

FY2024 사업부별 매출 비중



자료: META, SK 증권

광고 엔진, 피드 엔진 강화 수혜

AI 연산량 증가에 따른 Predictive AI 수혜가 기대된다. 1) 피드 추천 개선으로 인한 체류 시간 증가 2) 광고 엔진 성능 증가로 인한 광고 단가 상승이 기대된다.

이미 2021 년부터 기존 CPU 기반 광고엔진을 GPU 기반으로 전환하기 시작했다. Meta 의 3Q21 실적 발표에서 AI 및 ML 투자가 처음 언급되었으며, 이후 2022 년에는 Capex 를 전년 대비 67% 증가시키며 AI 투자에 집중했다. 2022~204 년 기간동안 해당 투자는 릴스(Reels) DAU 급증, P(광고 단가)와 Q(광고 수량) 동반 상승이라는 성과로 나타났다. 향후 AI 성능 증가에 따라 해당 효과는 지속될 것으로 기대되며, 이는 META 의 광고 시장 내 입지 강화에 기여할 것으로 전망된다.

2024 년말 메타가 확보한 GPU 수는 H100 기준 60 만개 수준이며 2025 년 Capex 투자 금액을 60~65bil 수준으로 제시했다. 해당 GPU 들로는 엔비디아와 협력 중인 신규 광고 강화 시스템 'Andromeda'에 투자 중이다. 2024 년 동안 Andromeda 로 광고 품질이 8% 향상됐다. 광고주들이 AI 에 전적으로 광고 매칭을 맡기게되는 Advantage+ 캠페인의 경우 2024 년 20B 수준의 매출을 기록하며 광고 매출 비중의 12% 를 차지 중이다. 향후 Advantage+ 비중 증가와 함께 광고 단가 상승이 기대된다.

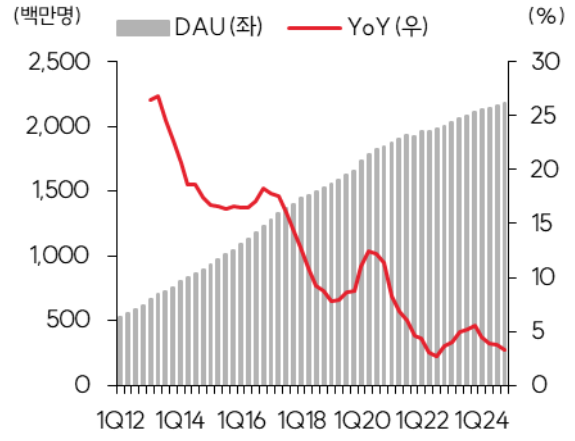
2024 년 설치 어플 Top 5 중 3 개가 META platforms

App	Downloads (mm)
TikTok	773
Instagram	759
Facebook	571
WhatsApp	527
Temu	438
Telegram	409
CapCut	361
Threads	322
Snapchat	302
ChatGPT	278
WhatsApp Buiness	271
Messenger	265
Spotify	239
Shein	211

Sources: Appfigures, AppMagic

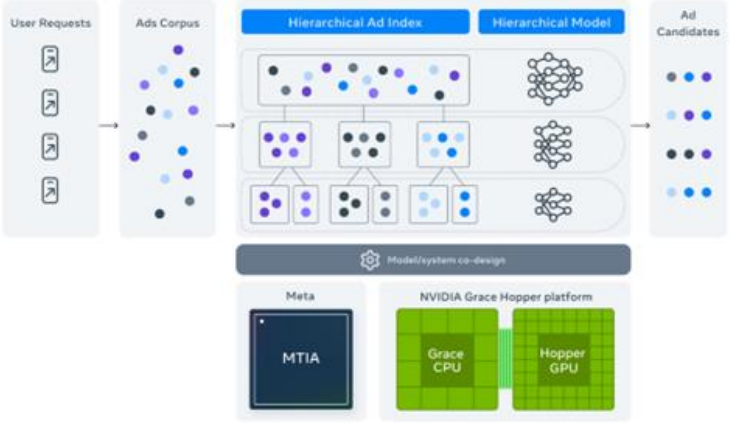
자료: 산업자료, SK 증권

META DAU 추이 및 YoY



자료: Bloomberg, SK 증권

Andromeda 광고 모델 구조, Hopper GPU 탑재로 고려 Data 수 증가

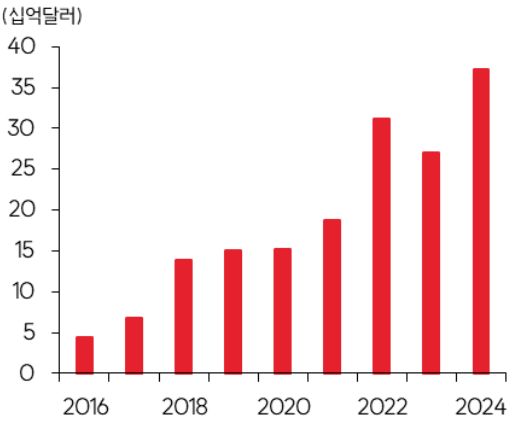


자료: META, SK 증권

ATT 정책 도입에 따른 메타 주가 및 매출액 성장률 급락

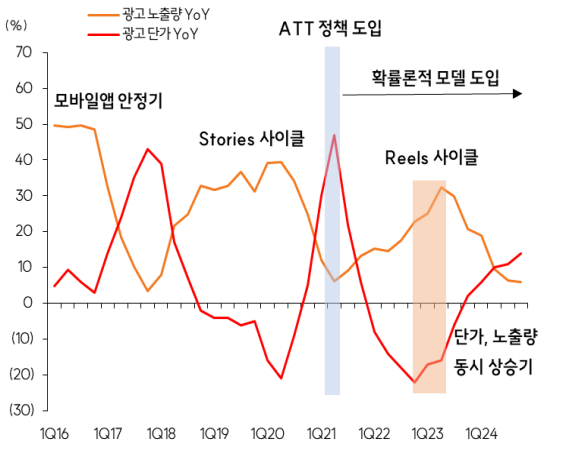
자료: Bloomberg, SK 증권

광고 엔진 투자한 2022 년부터 Capex 급증



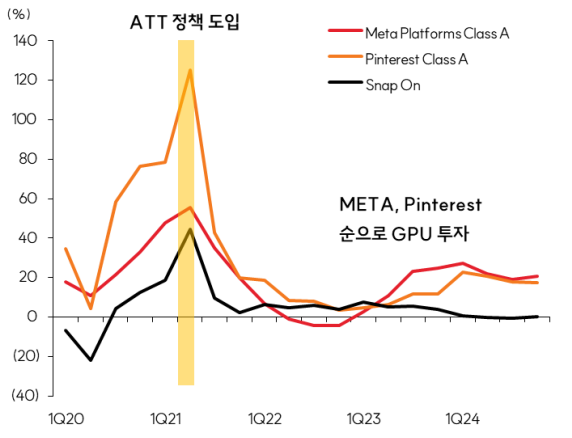
자료: Bloomberg, SK 증권

광고 노출량, 광고 단가 YoY



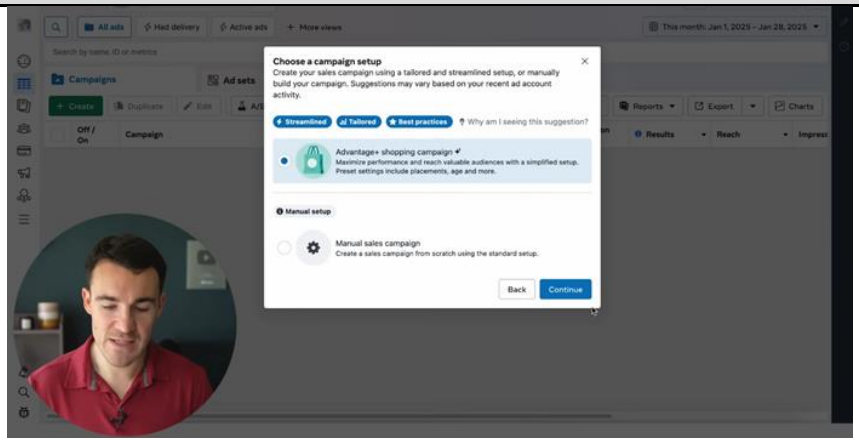
자료: META, SK 증권

GPU 투자 순서로 매출액 성장률 증가



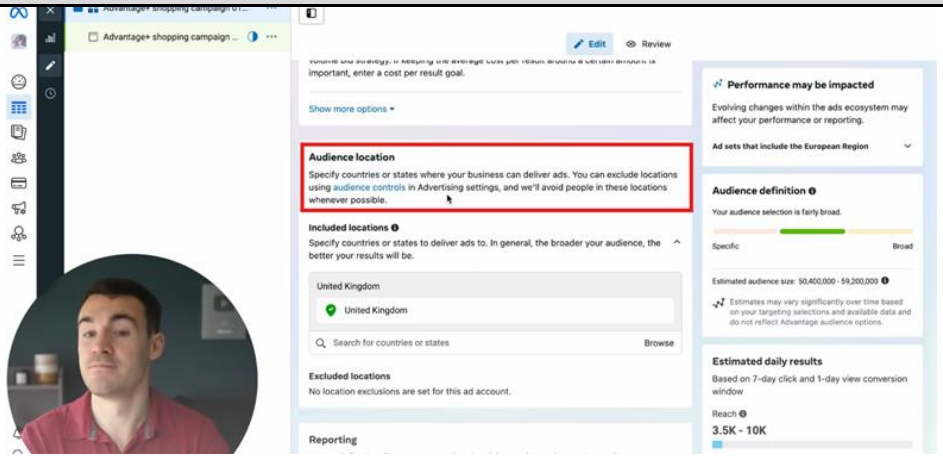
자료: Bloomberg, SK 증권

Advantage+ 선택 화면, 광고주가 Manual(기존 모델)과 Advantage+ 중에 선택 가능



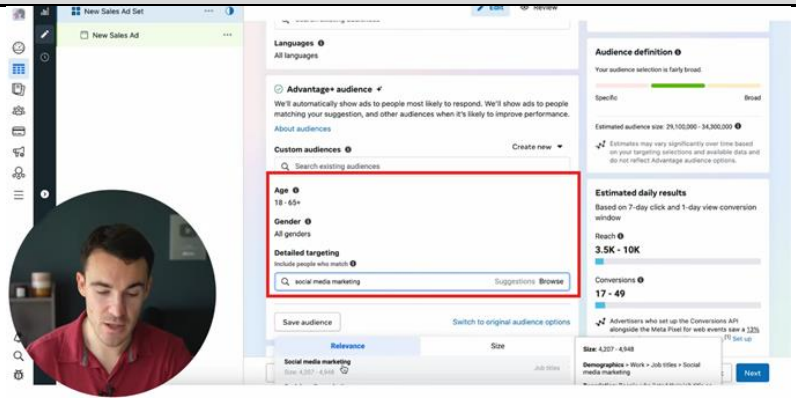
자료: Youtube, SK 증권

Advantage+를 선택할 경우 광고주는 Location 만 선택, 나머지는 AI 엔진이 해결



자료: Youtube, SK 증권

Manual 을 선택할 경우 다른 여러 조건 삽입



자료: Youtube, SK 증권

생성형 AI: 광고와 컨텐츠가 합성되는 시대

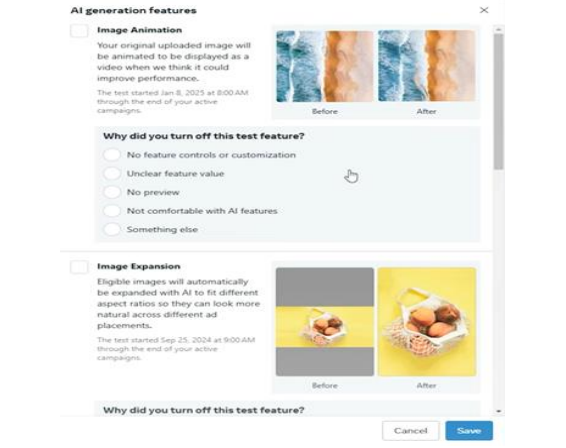
생성형 AI 의 성능은 현재 급속도로 빨라지고 있으며, 이는 생성되는 컨텐츠(텍스트, 동영상, 이미지 등)의 질을 높이고 가격을 낮추는 중이다. 최근 지브리 열풍에서 볼 수 있듯, 일부 제품화 요소가 합쳐지면 현재 수준의 생성형 AI 도 성공적인 AI 서비스를 만들 수 있다.

지브리 사진이 남들에게 보여지기 위한 카카오톡 프로필 사진 교체로 이어졌던 것처럼, 컨텐츠 다양성의 증가는 곧 SNS 의 이용 증가로 연결된다. META 입장에서 이는 방문객 및 체류 시간 증가로 인한 광고 노출도 증가가 기대된다. META.AI 가 인스타그램, 페이스북 등 더 접근성이 높은 플랫폼에서 서비스를 제공한다면 OpenAI 보다 더 큰 파급효과를 예상할 수 있다.

추후 생성형 AI 로 인한 컨텐츠 증가는 광고의 결합으로 이어질 가능성이 높다. Advantage+ creative 서비스는 광고 생성에 집중한다. 아직까지 META.AI 가 제공하는 기능은 이미지를 애니메이션으로 변환, 동영상 배경 변환 등의 미미한 수준이다. 향후에는 광고를 위한 동영상 생성을 목표로 한다. 메타가 갖고 있는 사용자 데이터를 고려할 때, 사용자 동의가 있다면 사용자가 직접 나타나는 형태의 광고가도 생성이 가능하다 이는 추론 비용의 하락 속도를 볼 때, META 가 Blackwell 을 수급하는 단계인 1년 이내로 서비스 가능한 수준이 될 것으로 보인다.

Advantage+ creative 까지 장착한 META 의 향후 목표는 비즈니스에 대한 가이던스만 받고 마케팅과 타게팅은 전부 META 가 담당하는 것이다. 현재 플랫폼 광고 비즈니스는 광고 중계에 그친다면 생성형 AI 의 성장으로 광고 생성, 즉 광고 대행사의 부가가치를 흡수하게 된다.

현재 Advantage+ creative 는 보정 서비스 제공 수준



자료: META, SK 증권

광고에 유저 사진 삽입

자료: AI Boost, SK 증권

Compliance Notice

작성자는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

- 본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.
- 투자판단 3 단계 (6 개월 기준) 15%이상 → 매수 / -15%~15% → 중립 / -15%미만 → 매도

2025-04-14

알파벳(GOOGL)

AI 강자의 저평가 구간

SK증권 리서치센터

Analyst
박제민

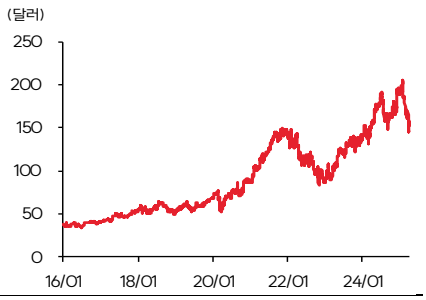
jeminwa@sks.co.kr
3773-8884

기본 정보	
국가	미국
상장거래소	NASDAQ
결산 기준월	12 월
시가총액 (십억달러)	1,849
시가총액 (조원)	2,651.6
현재주가 (달러)	151

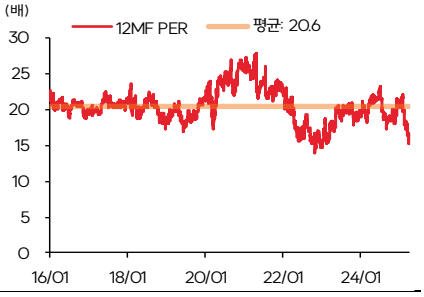
기업 개요

알파벳(Alphabet Inc.)은 지주 회사. 동사는 자회사를 통해 웹기반 검색, 광고, 지도, 소프트웨어 애플리케이션, 모바일 운영시스템, 소비자 콘텐츠, 기업 솔루션, 상거래 및 하드웨어 제품을 제공하고 있다.

주가 추이



12MF PER 추이 및 평균



SK 증권

AI 업계의 원조

구글은 AI 업계에서 가장 긴 역사를 가진 기업이다. 초기 추천 엔진 기반 AI 부터, 현재 LLM 의 핵심 구조인 Transformer, LLM 에 강화 학습 적용, 모델 증류 개념 등 주요 기술 모두 구글에서 최초로 제시됐다. 현재 AI 알고리즘 대부분이 구글로부터 시작했다고 해도 과언이 아니다. 하드웨어 측면에서도 강력한 경쟁력을 보유하고 있다. 2016 년부터 자체 AI 반도체인 TPU 를 개발해 AI 행렬 연산에 최적화된 인프라를 구축했으며, Nvidia 를 제외하면 의미 있는 AI 가속기 시장 점유율을 가진 유일한 기업이다.

본업 리스크 노출, 저평가 이유 확실

그러나 구글의 주가는 2022 년 말 ChatGPT 출시 이후 12MF P/E 가 횡보하고 있다. 같은 기간 매출 성장률은 반등했다. 이는 전사 영업이익의 80% 이상을 차지하는 검색(Search) 부문의 경쟁 리스크 때문이다. 과거 90% 이상 점유율을 차지하던 검색 시장에 OpenAI(ChatGPT), Grok 등 AI 기반 스타트업이 본격 진입하면서 경쟁 우려가 부각됐다. 특히 ChatGPT는 최근 MAU가 5억 명까지 급증하며 구글의 핵심 사업에 직접적인 위협으로 자리 잡았다.

제품화 시대대는 분명한 기회, 승리한다면 큰 Upside 가능

반면 AI 제품화 시대가 본격화되면서 새로운 기회도 열리고 있다. 이미 구글은 5 억 명 이상 MAU 를 가진 앱이 12 개 에 달한다. 해당 앱의 데이터가 Gemini 모델에 활용되고 있고 향후 제품들이 서비스될 소비자 접점이 많다. GCP 역시 Gemini 성능 개선을 통해 AI 호스팅 시장 점유율 상승 효과가 나타나고 있다. 검색 사업의 불확실성은 존재하지만 AI 기술 발전으로 인해 검색 시장 자체가 확장되고 있다. 아직까지 '전체 검색 쿼리 중 광고가 연결되는 쿼리'의 비중은 20%에 불과하다.

영업실적 및 투자지표

구분	단위	2021	2022	2023	2024	2025E	2026E
매출액	십억달러	258	283	307	350	343	366
영업이익	십억달러	79	75	84	112	128	141
순이익(지배주주)	십억달러	76	60	74	100	109	122
EPS	달러	5.7	4.6	5.8	8.1	8.9	10.1
PER	배	30	31	25	19	17	15
PBR	배	8	8	7	6	5	4
EV/EBITDA	배	18	14	18	19	11	10
ROE	%	32	24	27	33	29	27

자료: Bloomberg, Consensus(25.04.01)

다양한 플랫폼 기반 SW 사업 전개, 주요 수익원은 검색 광고

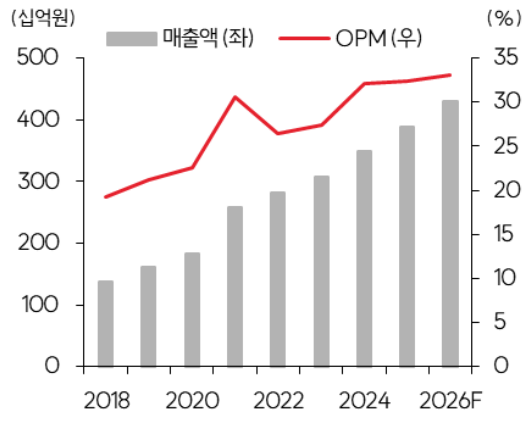
Alphabet 의 매출 구조는 검색 광고 중심의 플랫폼 사업에 기반하며, 클라우드 및 구독 서비스, 광고 네트워크와 디바이스까지 다각화되어 있다.

Google Search & others 사업부는 구글 검색엔진 및 지도에 노출되는 광고 매출이 포함된다. 대부분의 수익은 검색 상단 노출을 원하는 광고주가 CPC(Cost per Click) 기반으로 지불하는 광고비에서 발생한다. 해당 사업부는 전체 매출의 절반 이상을, 영업이익의 80% 이상을 차지한다.

Google Cloud 사업부는 두 축으로 나뉜다. GCP(Google Cloud Platform)는 AI/ML 기능 중심의 IaaS·PaaS 클라우드로, Vertex AI 및 Gemini AI를 통해 사용량 기반의 과금 체계를 운영한다. 데이터센터 인프라 및 R&D 비용이 주요 원가 요소다. Workspace는 Gmail, Docs, Meet 등 SaaS 기반 협업 도구로, 사용자당 구독료 기반의 수익을 창출하며 Microsoft365, Zoom 등과 경쟁하고 있다. 최근에는 AI 기능 결합이 중요한 경쟁 요소로 부상 중이다.

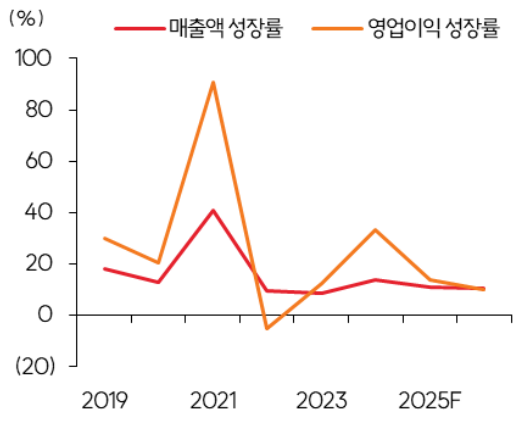
Google Network 는 구글이 직접 소유하지 않은 외부 웹사이트·앱에 자사의 광고 엔진을 공급하는 사업이다. 대표적으로 웹용은 AdSense, 앱용은 AdMob 이 있으며, 광고 경매 플랫폼인 Ad Exchange 도 포함된다. 수익은 광고주 CPC 중 약 30%를 구글이 수취하는 구조이며, 주요 비용은 파트너에게 지급하는 TAC 와 광고엔진 인프라 유지 비용이다. Meta, Amazon 등과의 경쟁에서 광고 타겟팅 정밀도와 네트워크 확장성(Inventory 확보)이 핵심 차별화 요소다.

매출액 영업이익 추이 및 전망



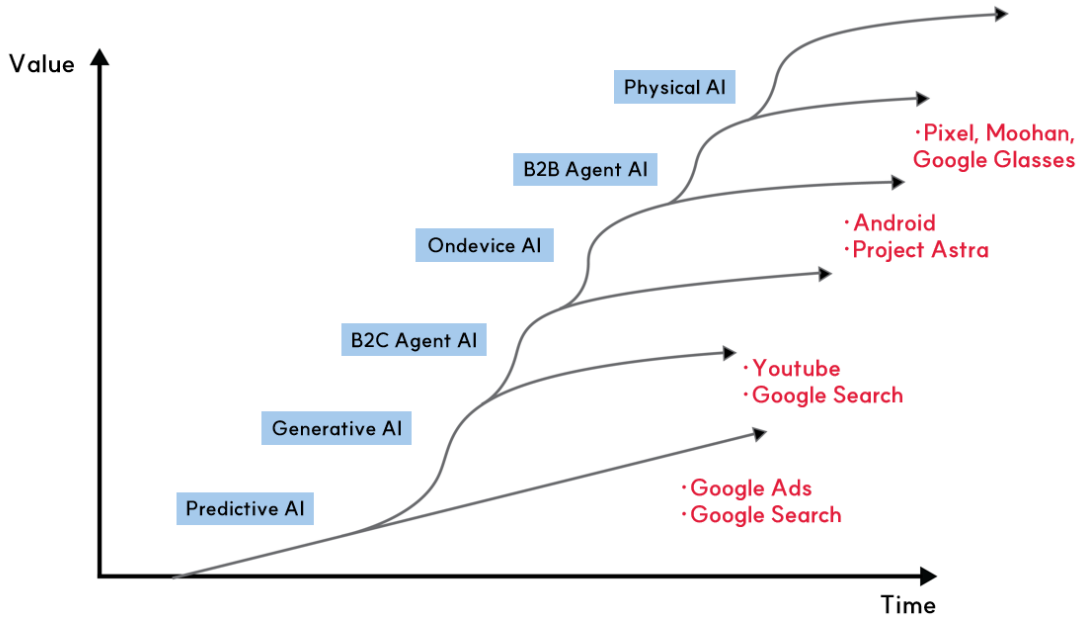
자료: Bloomberg, SK 증권

사업부별 매출액 추이



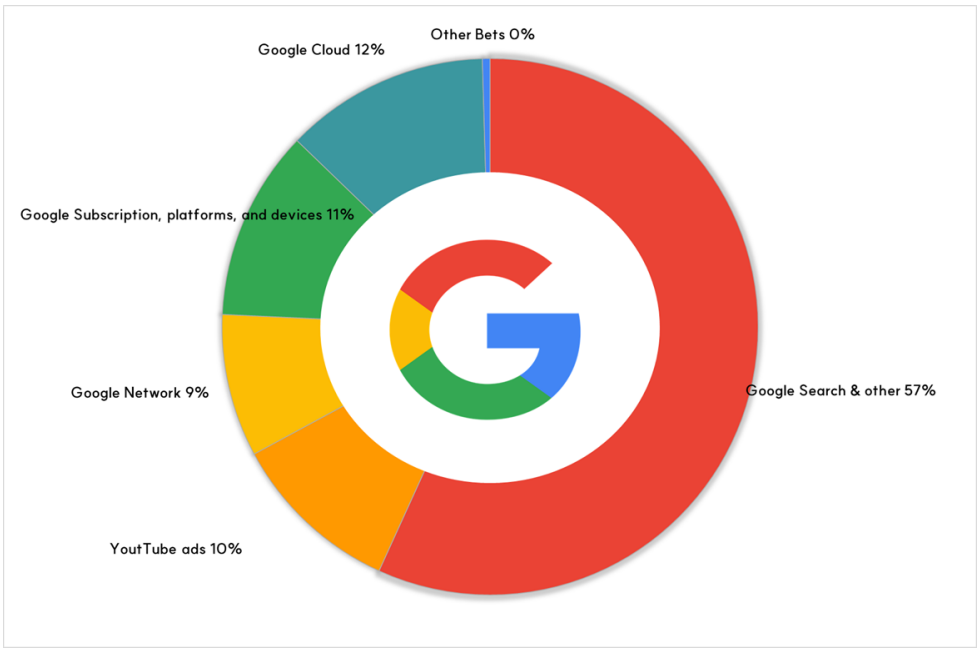
자료: Bloomberg SK 증권

사업부별 AI 단계 포지션



자료: SK증권 / 주: CSP 사업은 전 AI 단계에서 수혜

FY2024 사업부별 매출 비중



자료: Google, SK증권

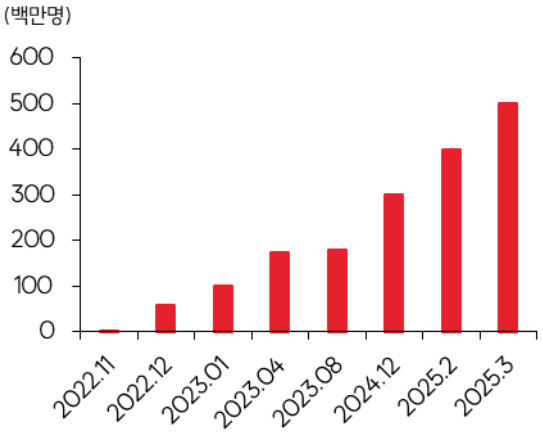
검색 광고 부문: 불안한 90%와 희망의 20%

최근 Sam Altman은 인터뷰를 통해, OpenAI가 향후 5년 안에 SOTA 모델 보유만큼 10억 유저를 확보하는 것을 중시한다고 밝혔다. 이는 OpenAI가 단순히 기술 기업이 아닌, 소비자 중심의 기술 플랫폼으로의 진입을 모색하고 있음을 시사한다. Altman 은 아직 광고 산업 진출에 대한 구체적인 계획은 없다고 언급했지만, ChatGPT 는 광고 시장에 영향을 줄 수 있는 모든 요소를 내포하고 있다. Google의 검색 광고 사업 점유율 90%가 불안하게 비쳐지는 이유이다.

경쟁 위험도 있지만 AI 는 검색 시장 전체를 키우는 중이다. 구글 반독점 조사 결과 보고서에 의하면 '전체 검색 쿼리 중 광고가 연결되는 쿼리'의 비중은 20%에 불과하다. 이 수치는 검색과 광고 간 연결 고리를 늘릴 여지가 충분함을 시사한다. AI 는 1) 멀티모달 능력으로 input 검색 시도 증가 2) 타겟팅 정밀도 증가로 전환율 증가 등을 통해 검색을 통한 광고 시장 전체의 크기를 늘릴 수 있다.

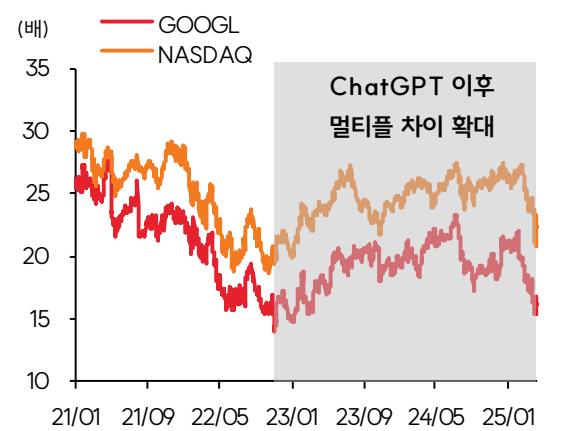
최근 AI 를 접목한 검색 기능의 진화가 이러한 연결을 강화할 수 있는 모멘텀으로 작용하고 있다. 대표적으로 AI Overview 기능은 초기 사용자 반응이 매우 긍정적이며, 2024 년 한 해 동안 100 개국 이상에 출시됐다. 2025 년에는 이 기능에 Gemini 2.0 모델이 탑재될 예정으로, 응답 정밀도 및 만족도 향상이 기대된다. 이외에도 Circle to Search 는 시각 기반 검색 경험을 제공하며, 한번이라도 이 기능을 사용한 사용자는 전체 검색의 10% 이상을 이 기능으로 사용한다. 이는 새로운 유형의 쿼리를 생성하며, 기존 텍스트 기반 검색 대비 광고 적합성이 더 높은 검색 구조를 만들어낼 수 있는 기반이 될 수 있다.

ChatGPT MAU 추이



자료: 언론 종합, SK 증권

ChatGPT 이후 구글 12MF P/E vs Nasdaq P/E



자료: Bloomberg, SK 증권

B2C Agent 시대는 기회

구글은 5 억명 이상의 사용자를 확보한 어플리케이션만 12 개로 소비자 접점이 가장 많은 기술 기업이다. B2C Agent 제품화 단계에서 사용자 접근성과 데이터가 매우 중요 요소로 작용할 예정이다.

Gemini 는 이미 캘린더, 지도 앱과의 연계는 물론, 화면 인식 기능까지 탑재되며 점차 사용자에게 디지털 에이전트 역할을 수행하고 있다. 아직까지는 ChatGPT 가 대답의 품질과 Deep Research 기능의 우위로 점유율이 높지만 향후 LLM 이 Agent 로 제품화될 때 이런 데이터들의 유무는 점유율을 크게 가를 수 있다.

판매량이 미미하지만 Pixel Phone, AI Glasses 을 통해 Ondevice 시대를 대비한 하드웨어 사업도 전개 중이다. 이 모든 요소와 맞물려 멀티모달 능력 탑재를 향해 진행 중인 Astra 프로젝트는 구글의 차세대 AI agent 개발의 정점이라 할 수 있다. 구글이 해당 프로젝트로 AI 시장에서 소비자들의 점유율을 되찾아온다면, 기업가치의 큰 성장이 가능하며 현재의 낮은 멀티플을 충분히 역전시킬 수 있다.

구글 어플리케이션 MAU

플랫폼	MAU	플랫폼 설명	AI 활용
Google 검색	50 억명+	세계 최대 규모의 인터넷 검색 엔진	검색 엔진 강화, Overview 로 요약 제공
YouTube	25 억명+	동영상 공유 및 스트리밍 플랫폼	추천 엔진 강화, 크리에이터 보조 AI 제공
Google Play	25 억명+	모바일 앱/콘텐츠 유통 플랫폼 (안드로이드 앱 마켓)	AI 활용 컨텐츠 정책 유지
Google Chrome	30 억명+	웹 브라우저	Google Lens 로 이미지 검색 강화
Gmail	20 억명+	이메일 서비스	AI 요약 기능 제공
Google Maps	20 억명+	디지털 지도 및 내비게이션 서비스	AI 요약 기능 제공, 광고 추천 피드 강화
Google Drive	20 억명+	클라우드 파일 저장 및 공유 서비스	AI 요약 기능 제공
Google Photos	10 억명+	사진/동영상 백업 및 관리 서비스	AI 편집 기능 지속 강화 중
Google Docs	10 억명+	온라인 문서 편집 도구	AI 요약 기능 제공, NoteBookLM 등의 생산성 도구 출시
Google SpreadSheet	9 억명+	온라인 스프레드시트 도구	데이터 시각화, 각종 자동화 (Copilot 과 유사)
Google Slides	8 억명+	온라인 프레젠테이션 도구	텍스트 설명 기반 맞춤형 이미지 제공 (Copilot 과 유사)
Google Calender	5 억명+	온라인 일정 관리 서비스	일정 관리 효율화

자료: 산업자료, SK 증권

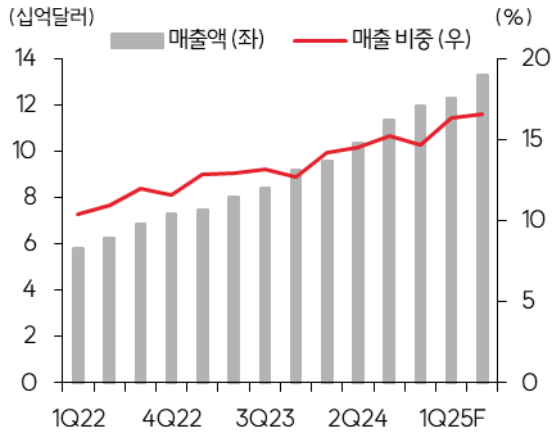
GCP: Gemini 경쟁력 증가, 전 층위 AI 서비스 제공

GCP(Google Cloud Platforms)는 Gemini 모델을 중심으로 AI 경쟁력을 빠르게 강화하고 있으며, Gemini 2.5 출시로 현재 시점에서는 새로운 기술 프론티어를 개척한 것으로 평가된다. 현재는 주요 AI 선도 기업 간에 알고리즘을 상호 참고하는 속도가 빨라지고 있어, LLM API 경쟁의 우열이 자주 바뀌고 있는 상황이다. 이에 따라 B2B AI 모델 시장은 락인이 어렵고, 점점 commoditize 되고 있다.

결국 AI 모델 자체보다는 AI 전반에 걸친 수직 통합형 인프라 경쟁력이 핵심이 될 전망이다. GCP 는 1) 자체 기반 모델(Gemini), 2) 고성능 TPU, 3) Vertex AI 및 TensorFlow 를 포함한 AIaaS, 4) 5 억 명 이상의 MAU 를 보유한 앱을 15 개 이상 보유한 사용자 접점까지 완비한 구조로, AI 생태계 전반을 갖춘 CSP 중 하나다. 또한 전통적으로 다양한 AI 알고리즘을 개발해온 경험은 AI 기술이 보편화될수록 더욱 부각될 수 있는 경쟁 포인트다. 최근 보안 기업 Wiz 인수를 통해 멀티클라우드 보안 역량과 전략적 유연성도 확보하고 있다.

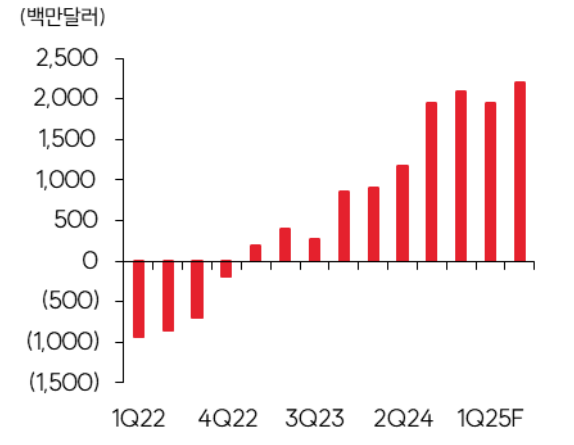
다만, 한계도 명확하다. GCP 는 CSP 3 사 중 매출에서 클라우드 비중이 가장 낮고, 시장 진입도 가장 늦어 고객 기반이 제한적이다. 향후 AI 혁신이 스타트업 및 연구기관 중심의 산업 재편으로 이어질 경우에는 GCP 의 입지가 유리할 수 있으나, 대기업 중심의 AI 도입이 자본수익률(ROE) 개선으로 귀결되는 방향이라면, 이미 강력한 엔터프라이즈 고객을 확보한 경쟁사들이 유리하다.

GCP 매출 추이 및 전사 매출 비중



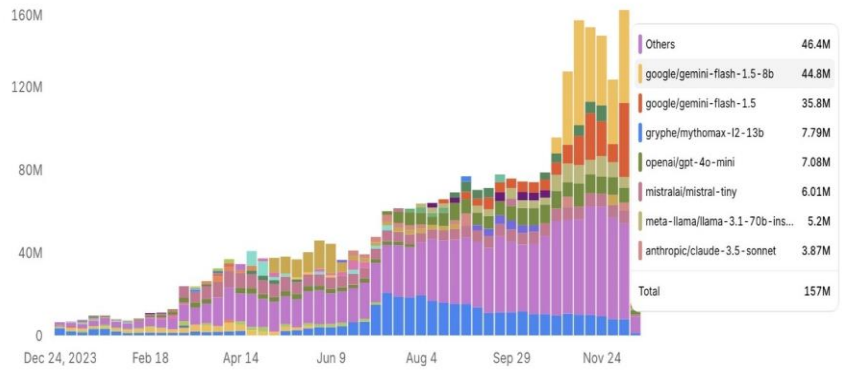
자료: Bloomberg, SK 증권

GCP 영업이익 추이 및 컨센서스



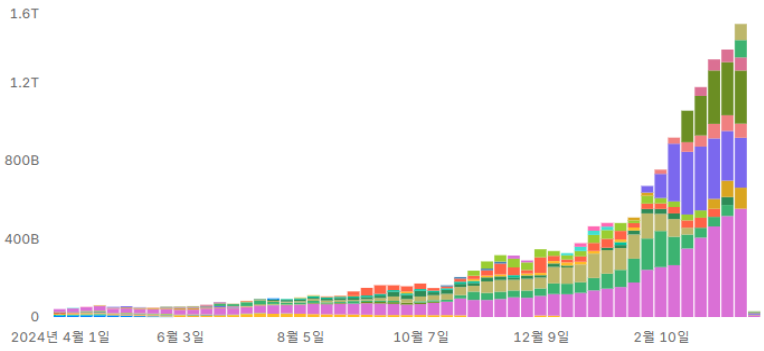
자료: Bloomberg, SK 증권

2023 년 후반기에는 5% 수준의 Gemini 점유율이 50%대로 증가



자료: OpenRouter, SK 증권

2025 년 3 월 누적 점유율은 60%대

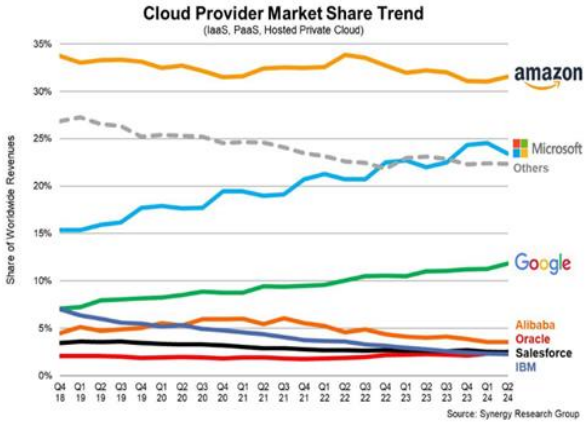


자료: OpenRouter, SK 증권

Wiz 인수를 통해 보안을 앞세운 멀티클라우드 공략

자료: 언론보도, SK 증권

CSP 3 사 점유율 추이



자료: Synergy Research, SK 증권

CSP 3 사 AI 서비스 현황, AI 훈련 가담 중인 ASIC, 자체 프론티어 모델 갖춘 유일한 기업

	마이크로소프트	구글	아마존
ASIC	Maia + NVIDIA	TPU + NVIDIA	Tranium/Inferentia + NVIDIA
Foundation Model	OpenAI PHI-4	Gemini	Claude Amazon Nova
LLM base 플랫폼	Azure AI Foundry	vertex.ai	Amazon SageMaker Amazon Bedrock
AI as a Service	Copilot	Gemini	Amazon Q

자료: SK 증권

Gemini 모델 변화

세대	출시 시기	모델 구분	주요 기술 변화
Gemin 1.0	2023.12	Ultra, Pro, Nano	최초 멀티모달 기반, 32k 컨텍스트
Gemini 1.5	2024.02	Pro, Flash	MoE 구조 도입, 100 만 토큰 장문 컨텍스트 지원
Gemini 2.0	2024.12 ~ 2025.02	Flash, Flash Thinking, Pro	도구 사용 능력 내장, 멀티모달 출력(이미지/TTS), Agent 지향
Gemin 2.5	2025.03	Pro (Experimental)	CoT 내제화, 코딩 및 추론 능력 강화

자료: 산업자료, SK 증권

2025-04-14

엔비디아(NVDA)

영역을 확장하는 AI 인프라 제왕

SK증권 리서치센터

Analyst
박제민

jeminwa@sks.co.kr
3773-8884

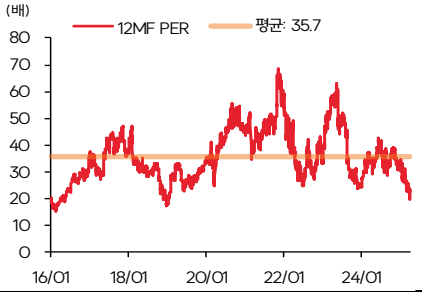
기본 정보	
국가	미국
상장거래소	NASDAQ
결산 기준월	01월
시가총액 (십억달러)	2,484
시가총액 (조원)	3,563
현재주가 (달러)	102

기업 개요 (Bloomberg)

엔비디아(NVIDIA Corporation)는 IT 기업. 동사는 과학 컴퓨팅, 인공지능(AI), 데이터 과학, 자율 자동차, 로봇 공학, 메타버스 및 3D 인터넷 애플리케이션에 필요한 플랫폼을 개발하며 PC 그래픽에 중점을 두고 전 세계 고객에게 서비스를 제공한다.

주가 추이

12MF PER 추이 및 평균



AI 제품화 시대에 따른 수요 증가 전망

AI 제품화 시대가 본격화되면서 GPU 수요가 지속 증가할 전망이다. Test-time scaling 과 합성 데이터의 활용이 확대되며 AI 모델의 발전은 당분간 명확한 성장세를 유지할 것으로 보인다. 특히 Test-time scaling은 추론 비용 절감을 성능 향상으로 전환하는 구조이기 때문에 고성능 AI 에 대한 수요가 높아질수록 GPU 기반 추론 수요 역시 지속 증가할 것이다. Agent 시대가 도래하면 Context Length, 반응 속도, 신뢰도 등으로 더욱 높은 컴퓨팅 성능이 요구된다.

해자 구축으로 인프라 사업자까지

엔비디아는 단순한 칩 제조업체에서 CSP 와 유사한 인프라 사업자로 진화하고 있다. 데이터센터의 효율성을 크게 높이는 Dynamo SW, NVLINK, Ethernet 등 네트워킹 기술 고도화와 단일 랙에 탑재되는 GPU 수량 증가를 통해 인프라 단위의 경쟁력을 강화하고 있다. 주요 고객사인 CSP 들은 과거 기판 제품인 HGX 위주의 구매에서 최근 서버 단위 제품인 NVL 에 대한 수요가 늘어나고 있다.

칩체와 관세에도 경쟁력은 빛난다

관세 우려에도 엔비디아의 경쟁력은 유지될 전망이다. 엔비디아 매출의 36%를 차지하는 미국 CSP 시장은 현재 주요 3 사가 치열하게 경쟁 중이며, 이로 인해 투자를 줄이기 어렵다. AWS 가 지난 10 년간 선제적 투자로 영업이익 성장을 누려온 사례를 감안하면, 경쟁이 심화된 현 상황에서 CSP 들의 투자 축소 가능성은 낮다. 2024 년 2 분기 AI ROE 논란 당시에도 CSP 업체들의 투자 의지는 컨퍼런스 콜을 통해 굳건히 확인됐다. AI 를 제품화하는 기업들 또한 성공적인 AI 서비스가 쉽게 순위를 바꾸지 않는 ChatGPT 사례를 보며 투자 의지를 유지할 가능성이 크다.

영업실적 및 투자지표 (FY기준)

구분	단위	2022	2023	2024	2025	2026E	2027E
매출액	십억달러	27	27	61	130	205	254
영업이익	십억달러	10	4	33	81	130	165
순이익(지배주주)	십억달러	10	4	30	73	109	138
EPS	달러	0.4	0.2	1.2	3.0	4.5	5.7
PER	배	272	419	90	37	24	22
PBR	배	102	121	62	33	15	10
EV/EBITDA	배	57	85	48	38	21	17
ROE	%	45	18	91	119	85	66

자료: Bloomberg, Consensus(25.04.11)

SK 증권

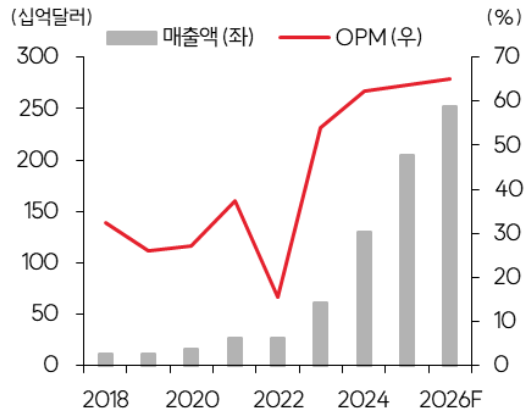
엔비디아는 단순 GPU 설계사가 아니다

엔비디아의 핵심 경쟁력은 3 가지로 나눌 수 있다.
1) 6 종의 반도체를 설계하는 능력
2) 시스템을 밸류체인 사이에서 빠르게 제조, 출시하는 추진력
3) CUDA 라이브러리의 베타성과 확장력

엔비디아는 TSMC, SK 하이닉스, ASML, Cisco 등의 핵심 관계사만 20~30 개, 그 외 광의의 파트너 생태계를 합치면 500 여개의 밸류체인을 가진다. 밸류체인들과 같이 조율하여 제품을 추진하는 능력은 1.5 년에 불과한 신규 제품 출시 시기로 나타난다. 이번 Blackwell 시리즈의 경우에도 랙 단위의 NVL72 시리즈는 일부 공급 제약이 있었으나 이전 버전인 H200 의 대량생산 시기였던 3Q24 에 이어 반년만에 공급이 진행됐다. 향후 로드맵의 경우에도 GB300 시리즈가 하반기에 램프업, Rubin 은 내년 하반기, Rubin Ultra 를 내후년에 출시하면서 출시 로드맵 속도전을 지속 중이다. Blackwell 의 경우에도 랙 단위 공급에 일부 제약이 있었으나 대량 생산 계획까지 6 개월 이내로 해결 됐다.

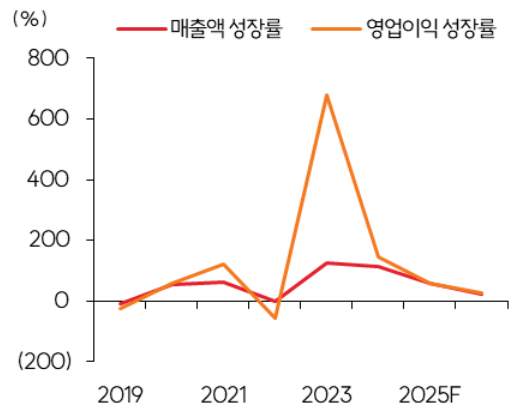
현재 엔비디아 생산에 있어 가장 병목은 TSMC 의 CoWoS Capa 로 평가받는다. Blackwell 전환에 있어 CoWoS-S 에서 CoWoS-L 로의 전환에 문제가 있었다. 그러나 TSMC 는 2025 년 대부분의 CoWoS 생산량을 L 로 배치하고 엔비디아에게 공급할 예정이다. 향후 Blackwell 시리즈와 유사한 공급 차질이 빚어질 가능성은 낮다고 판단된다.

엔비디아 매출액 영업이익 추이



자료: Bloomberg, SK 증권

매출액 영업이익 성장률 추이 및 전망



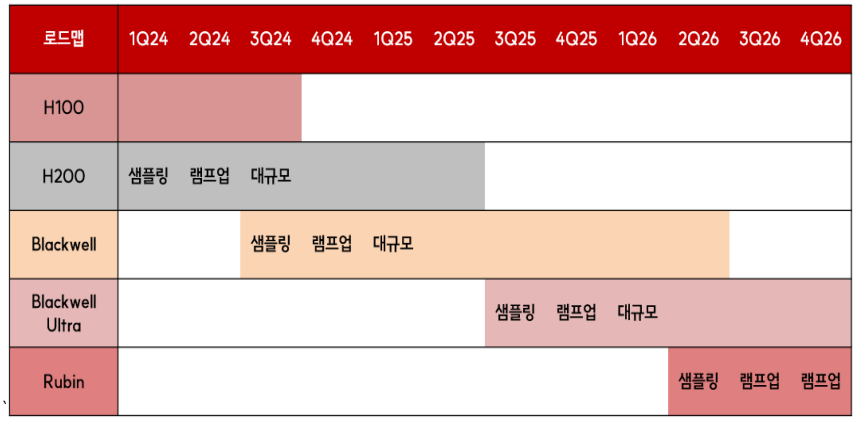
자료: Bloomberg SK 증권

엔비디아 주요 반도체 제품

반도체	제품명	기능	시장 지위 및 경쟁사
GPU	Ampere, Hopper, Blackwell, Rubin	AI 연산, 그래픽 컴퓨팅	AI 가속기 점유율 90% 이상 경쟁사: AMD(Instinct), Google TPU, Amazon Trainium 등
CPU	Grace	범용 컴퓨팅, GPU 와 통합 사용	신규 진입자 지위, GPU 점유율에 따라 상승 전망 경쟁사: Intel, AMD(EPYC), Ampere, AWS Graviton
DPU	BlueField	네트워크, 보안, 스토리지	선도적 지위 경쟁사: Intel(IPU), Marvell, AMD/Xilinx SmartNIC, Broadcom
NVSwitch	NVLink System	GPU 간 초고속 연결 (Scale-up)	사실상 독점, 타사는 PCIe 기반 구성
Ethernet	Spectrum	데이터센터 이더넷 네트워크 연결	중상위권 경쟁사: Broadcom(시장 1위), Cisco, Arista 등
Infiniband	Quantum	초고속 저지연 네트워크 (HPC 용)	Mellanox 인수로 사실상 독점

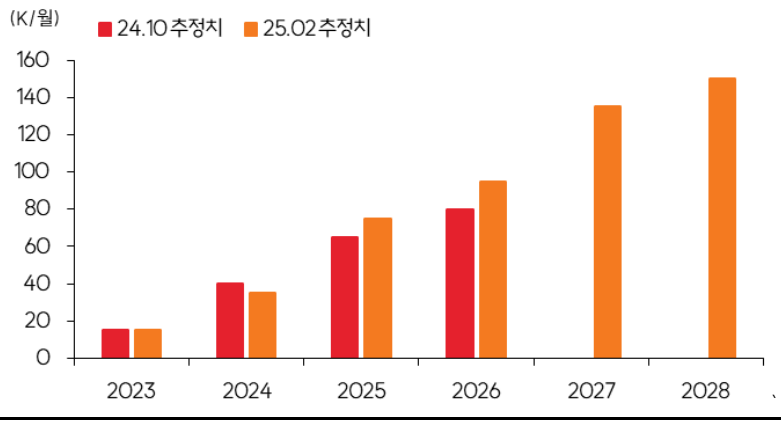
자료: NVIDIA, 산업자료, SK 증권

엔비디아 제품별 타임라인



자료: NVIDIA, SK 증권

CoWoS Capa 추정치 증가



자료: Digitimes, SK 증권

GPU 가 확장될수록 CUDA 경쟁력은 강화된다

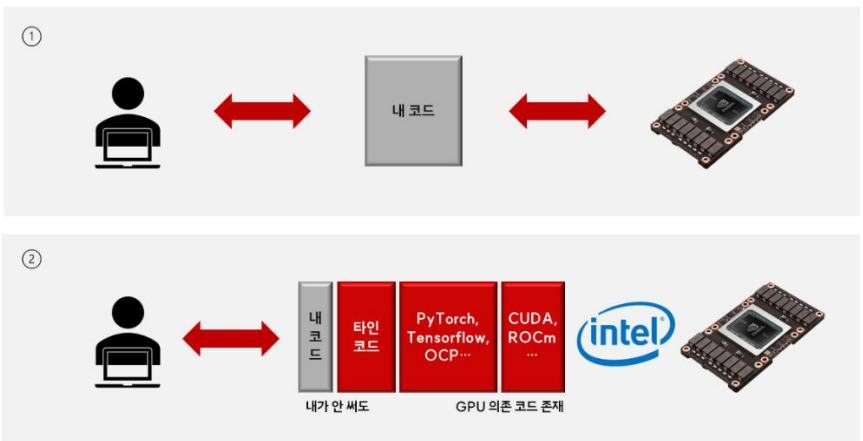
CUDA 는 GPU 위에서 작동하는 C-언어라고 생각하면 쉽다. 머신러닝, AI 알고리즘들이 전방위로 확대되면서 대부분의 산업들에 CUDA 로 코딩된 라이브러리들이 많다. 개발자들이 CUDA 기반 코딩에 익숙하며 신규 코드들의 호환도 더 용이하다.

코딩을 배우는 사람 입장에서 C-언어(로우 레벨) 보다 Phython(하이 레벨)을 더 배우기 쉽다. AI 연구자들 또한 PyTorch(META), TensorFlow(Google), cuDNN(NVIDIA) 등의 하이레벨 프레임 워크를 주로 사용한다. 이들 대부분이 CUDA 기반이며, 따라서 AI 연구자들은 자신도 모르게 CUDA 에 락인되고 있다.

AMD 플랫폼인 ROCm 이나 ASIC 의 개별 플랫폼 위에서 한 코딩들은 CUDA 와 호환이 안되는 경우가 많다. CUDA 코드를 ROCm 에서 실행할 수 있게 하는 코드가 있으나 전환율이 낮고 따로 커스텀이 필요하다.

CUDA 는 확장될수록 대체되기 힘든 성격을 가진다. 이번 GTC2025 에 'cuLitho' 라이브러리가 소개됐다. 반도체 제조 공정에서 필수적인 컴퓨팅 리소스를 GPU 에서 가속화하는 라이브러리다. TSMC, 삼성, ASML 와 같은 주요 반도체 기업들이 채택한다고 소개했다. 이렇게 기존 사업에서 CUDA 기반 코딩들이 확산될 수록, 후발주자가 AI 를 통한 부가가치를 창출할 영역이 줄어든다고 볼 수 있다. 고객사 입장에서 CUDA 가 돌아가는데 굳이 다시 품을 들여 신규 코딩을 짤 이유가 없다.

PyTorch, Tensorflow, 타인 코드 사용 시에도 모두 CUDA 에 종속



자료: 와이스트릿, SK 증권

AI는 더 Compute Demanding 해지는 중

AI 가 능동적으로 행동하면서 컴퓨팅 수요가 늘어나는 중이다. 생성형 AI 에서는 Reasoning 으로 인한 step-by-step reasoning, chain-of-thought, consistency checking 으로 서비스가 강화되는 방향이다. 모두 output tocken 의 폭증을 불러오며 컴퓨팅 수요가 증가한다. 향후 Agentic AI 에서는 output tocken 이 더 길어질 예정이다. 토큰의 양도 중요하지만 서비스 개선을 위한 응답 속도 개선도 중요하다. 응답 속도를 줄일 수록 서비스 가능한 batch 수가 줄어들면서 요구되는 컴퓨팅 양이 급증한다.

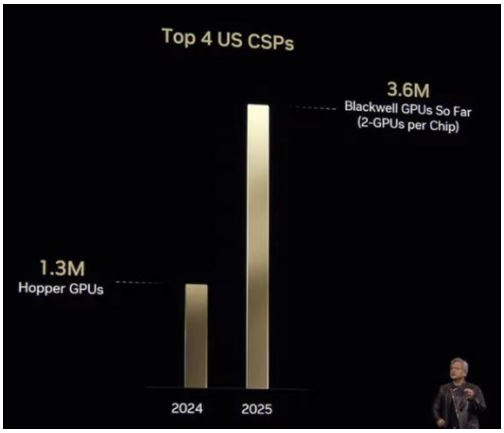
GTC2025 에서 젠슨황은 AI 스케일링 법칙이 초가속(hyper accelerated) 됐으며 요구 연산량이 작년 이맘 때 예상했던 것보다 100 배 늘었다고 언급했다. 처리 속도가 빨라진 Blackwell 모델의 2025 년 주문량이 작년 Hopper 수준 보다 높다고 보여준 표에서도 이를 알 수 있다.

AI 산업에서 제본스의 역설이 성립된다는 것은 이미 확인됐다. 알고리즘 혁신을 불러온 DeepSeek 출시 이후 중국 클라우드 서비스 제공업체들의 H20 GPU 수요가 급증했다는 사실이 언론을 통해 보도되었으며, 바이두와 텐센트 등의 대형 CSP 들은 CapEx(설비 투자) 가이던스를 상향 조정했다. 중국뿐만 아니라 글로벌 시장에서도 알고리즘 최적화가 컴퓨팅 수요 증가로 직결되고 있음이 기업들의 실적 발표에서 확인된다. 러시아 클라우드 서비스 기업 Nebius 는 DeepSeek R1 출시 이후 H200 GPU 수요가 급증했다고 발표했으며, GPU 서버 제공 기업 Lambda Labs 역시 동일한 현상을 확인했다. 데이터센터 HVAC 기업 AAON 은 실적 발표에서 DeepSeek 관련 질문에 대해 "더 효과적인 AI 모델을 만들 수 있는 능력이 곧 수요를 창출한다"며 AI 산업의 구조적 성장 가능성에 대한 확신을 표명했다.

Straberry 는 기존 모델들과 추론 구조 상이

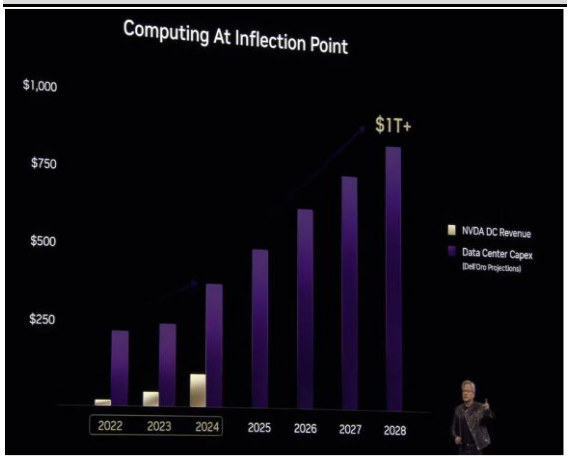
자료: MediumAI, SK 증권

이미 Top4 CSP 들의 주문량은 최신 모델에 집중



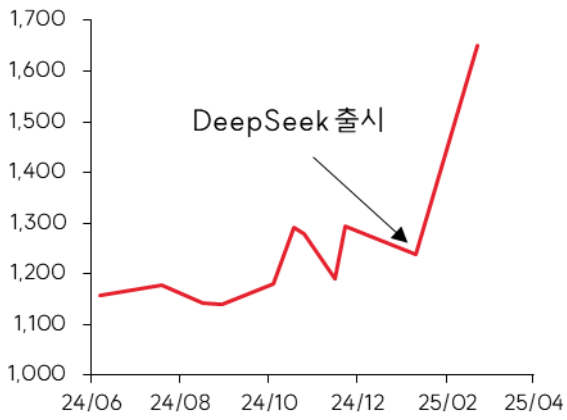
자료: NVIDIA, SK 증권

기존 CPU 서버들의 GPU 전환으로 IT Capex 내 비중 상승



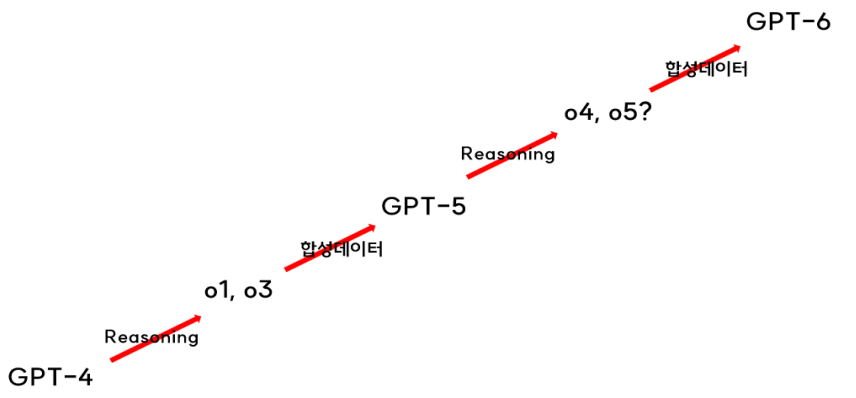
자료: MediumAI, SK 증권

Deepseek 등장 이후 중국 CSP 3 사 Capex 전망치 급증



자료: NVIDIA, SK 증권

합성데이터를 활용한 AI 모델 발전 지속



자료: SK 증권

부품사가 아닌 인프라 사업자로 변화

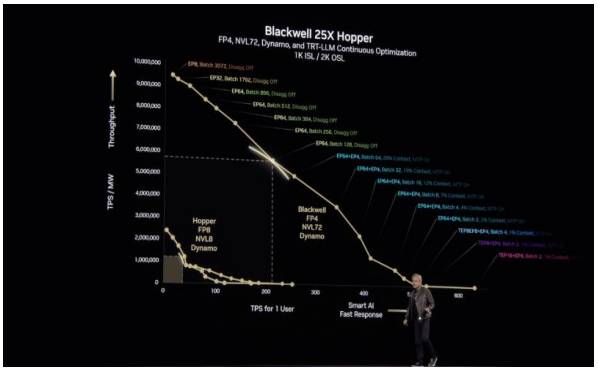
구글 TPU가 유투브 추천 알고리즘 연산을 담당하듯, AI 도 향후에는 연산하는 영역이 나누어질 수 있다. 특히 동일 업무에 대한 연산 요구량이 많을 것으로 추정되는 빅테크들 위주로 주요 컴퓨팅 처리를 ASIC 으로 전환할 것이라는 우려가 있다.

그러나 이는 기우로 판단된다. ASIC 대비 엔비디아의 강점은 단순 컴퓨팅 능력(GPU 스팩)과 CUDA SW 가 아니다. 엔비디아는 Ethernet, Inviniband, NVLINK 등의 통신칩 개발을 통해 랙 단위의 성능 증가에 집중하는 모습이다. 이는 아직 GPU 능력이 Hopper 단위에도 못 미친 ASIC 연구개발 부서를 크게 따돌리는 움직임으로 판단된다.

GTC 2025 에서 언급된 Dynamo 와 Homogenious cluster 컨셉은 경쟁 우위를 더 공고히 한다. Dynamo 는 데이터센터 단위로 GPU 의 리소스를 최적화해주는 소프트웨어 플랫폼이다. 엔비디아 GPU 의 효율성을 늘리는 외부장치(네트워크 장치, 냉각 장치, SW)가 많아질 수록 ASIC 의 교섭력은 떨어지게 된다. GPU 의 수명(3~4 년)보다 최신 시스템의 출시 속도(1 년)가 훨씬 빨라진 상황에서 이종칩(Heterogenious Cluster) 데이터센터는 워크로드 변환이 어렵다. Homogenious cluster 로 유지하는 것이 데이터센터 운영 유연성, 장기적 TCO 에 유리하다.

엔비디아의 랙 단위 움직임은 엔비디아를 단순 부품 공급사가 아닌 인프라 공급자로 바꾸는 중이다. CSP 사들은 기존 HGX 위주로 받아가던 부품들을 이제 NVL 로 변환하고 있는 모습이다. 엔비디아의 곧 밸류체인 내의 교섭력 증가가 예상된다.

Dynamo 는 다양한 스팩으로 GPU 최적화 가능하도록 관장



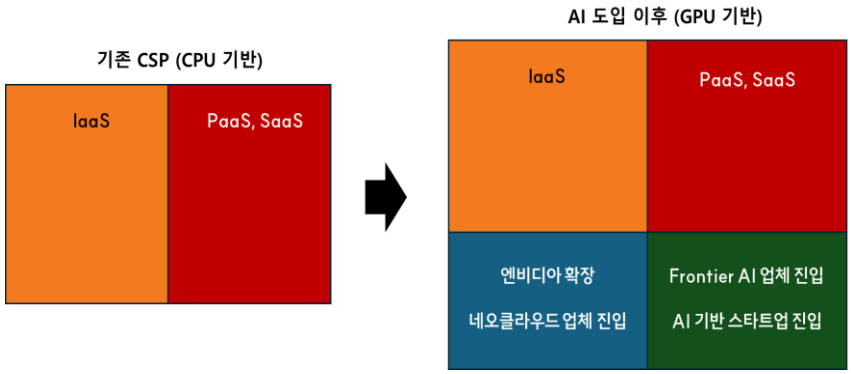
자료: NVIDIA, SK 증권

Homogenious, Heterogenious Cluster 비교

구분	Homogenious	Heterogenious
구성	동일한 HW, SW 구성의 노드 구성	다양한 HW, SW 구성의 노드 혼합 (ASIC 활용)
유연성	모든 노드 유연성 증가	특정 노드에 특정 활동
자원 활용률	클러스터 전체 사용률 최적화 용이 (Dynamo)	워크로드 불균형 발생 시 컴퓨팅 낭비 발생
운영 복잡도	SW 활용 운영 최적화 가능	처리 방식 상이하여 하드웨어 매핑 복잡

자료: 언론 종합, SK 증권

AI 도입으로 인한 CSP 산업 구조 변화



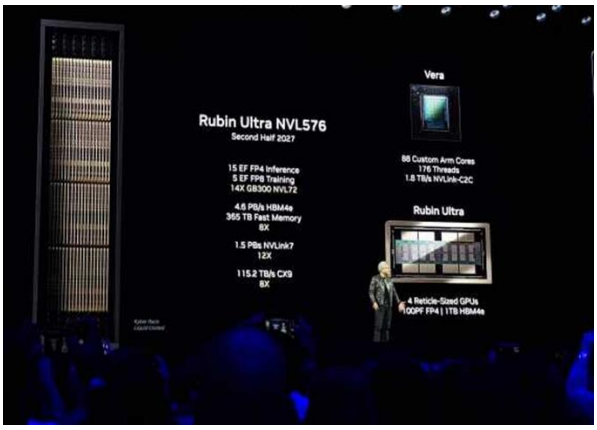
자료: SK 증권

같은 Blackwell 도 냉각 방식, 연결 방식으로 NVL72 성능이 더 우위

GPU Model	NVL72 Blackwell GPU	HGX Blackwell GPU
FP4 Tensor Core (FLOPS)	20	18
FP8 Tensor Core (FLOPS)	10	9
메모리 대역폭	8 TB/s	7.7 TB/s
메모리 용량	186GB	180GB
TDP	1200W	1000W

자료: 산업 료, SK 증권

GTC2025 를 통해 NVL 576 단까지 공개



자료: NVIDIA, SK 증권

Compliance Notice

작성자는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

- 본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.
- 투자판단 3 단계 (6 개월 기준) 15%이상 → 매수 / -15%~15% → 중립 / -15%미만 → 매도

2025-04-14

마이크로소프트(MSFT)

큰 결실을 위한 시간이 아직 필요

SK증권 리서치센터

박제민

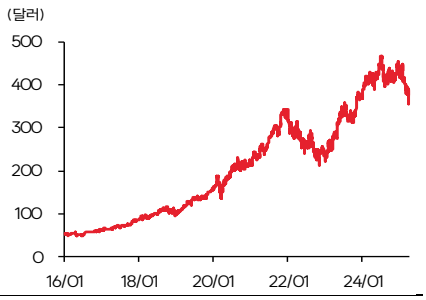
jeminwa@sks.co.kr
3773-8884

기본 정보	
국가	미국
상장거래소	NASDAQ
결산 기준월	06 월
시가총액 (십억달러)	2,774
시가총액 (조원)	3,978.6
현재주가 (달러)	373

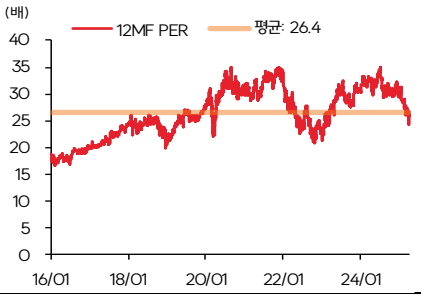
기업 개요 (Bloomberg)

마이크로소프트(Microsoft Corporation)는 소프트웨어 회사. 동사는 애플리케이션, 추가 클라우드 스토리지 및 첨단 보안 솔루션을 제공한다. 동사는 전 세계 고객을 대상으로 서비스를 제공한다.

12MF PER 추이 및 평균



12MF PER 추이 및 평균



SK 증권

AI는 생산성 혁명, 비즈니스 강자에 수혜 기대

Microsoft는 생산성 중심의 서비스로 매출의 대부분을 창출하고 있다. Windows와 Office는 안정적인 캐시카우이며, 비즈니스 생산성 소프트웨어 경쟁력을 바탕으로 Dynamics(ERP)와 Azure(CSP)로 사업 영역을 확장해왔다. AI 시대 도래로 Microsoft는 비즈니스 생산성 시장에서 Upsell 가능성이 크게 높아지고 있다. AI는 생산성 향상이라는 측면에서 Microsoft 에 구조적 수혜 요인으로 작용하며, 이미 차별화된 AI 기반 서비스를 다수 출시하고 있다. 최근 실적 발표(FY2Q25)에 따르면 Microsoft 의 AI ARR(연환산 매출)은 130억 달러(+175% YoY)로, FY2025 예상 매출의 약 4.6%를 차지하고 있다.

B2B AI 시장은 아직 초기

우려되는 부분은 아직까지 AI 관련 가시적 성과가 제한적이라는 점이다. AI ARR 130억 달러 중 생산성 소프트웨어 Upsell 비중은 최대 20억 달러 수준으로 추정되며, 나머지는 대부분 AI API 호스팅 매출일 가능성이 높다. 지난 4 개 분기 동안 해당 사업부의 매출액을 고려하면 실질적 Upsell 효과는 1%대에 불과하다. 최근 생성형 AI 시기의 매출 증가는 AI 도입 자체의 성과라기보다는, 기업들의 E5 전환과 클라우드 Migration 등 AI 도입 준비 과정에 따른 효과로 보는 것이 합리적이다. AI의 직접적인 수혜 효과는 아직 본격적으로 검증되지 않았다.

B2B 방향성은 명확, 중장기적 관점 필요

B2B AI 시장의 중장기 방향성은 명확하므로 지속적인 모니터링이 필요하다. 현재 기업들의 클라우드 Migration 확대를 감안하면 AI 기반 사업의 장기적 방향성은 긍정적이다. 다만 거시경제 불확실성 확대 국면에서 단기적 기대감 형성은 어렵다. GPT-5 출시로 인한 클라우드 점유율 상승 효과가 일시적 상승 요인이 될 수 있으나, OpenAI 와의 훈련 인프라 분리를 수혜의 지속성은 제한적이다.

영업실적 및 투자지표 (FY기준)

구분	단위	2021	2022	2023	2024	2025E	2026E
매출액	십억달러	168	198	212	245	277	314
영업이익	십억달러	70	83	89	109	124	140
순이익(지배주주)	십억달러	61	73	72	88	98	112
EPS	달러	8.1	9.7	9.7	11.9	13.1	15.0
PER	배	48	42	39	32	29	25
PBR	배	20	17	14	11	8	7
EV/EBITDA	배	25	19	24	23	19	16
ROE	%	47	47	39	37	31	29

자료: Bloomberg, Consensus(25.04.11)

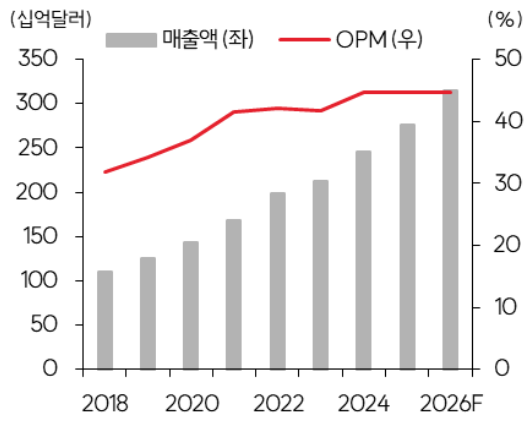
생산성 소프트웨어, Cloud 쌍두마차

Microsoft 의 성장을 이끄는 주요 매출 사업부는 365 Commercial products(매출 비중 30% 수준, LTM 성장률 10% 중반)와 Intelligence Cloud(매출 비중 40% 수준, LTM 성장률 20% 초반)가 있다.

Microsoft 365 는 Word, Excel, Outlook, Teams 등 생산성 소프트웨어 제품군을 포함하며, 대부분 기업 고객을 대상으로 월 10~50 달러 수준의 요금제를 운영한다. 최근에는 Copilot 등의 AI 기능을 추가하며 사용 단가를 기존 대비 약 2 배 수준(월 30 달러 추가)까지 끌어올릴 수 있는 업셀링 여력이 생겼다. Fortune 500 기업 중 90% 이상이 이미 Microsoft 365 를 사용하는 만큼 강력한 네트워크 효과와 진입 장벽이 존재한다. Google Workspace, Slack, Zoom 등과의 경쟁은 있으나, 윈도우와 Azure 의 연계성과 시장 장악력을 고려하면 대체재 위협은 제한적이다.

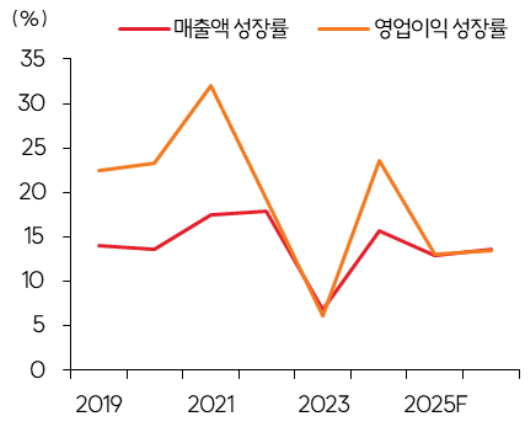
Intelligence Cloud 사업부는 Azure 기반의 IaaS, PaaS, 그리고 온프레미스용 소프트웨어와 SaaS 제품군을 포함한다. 고성능 AI 워크로드나 데이터 활용 수요가 증가할수록 인프라 사용량과 단가 모두 상승할 수 있어 구조적으로 업셀 여력이 크다. 클라우드 전환 추세와 함께 AI 및 데이터 기반 업무 증가가 수요를 자극하며, 경기 확장기에 IT 예산 확대가 동반될 경우 성장 모멘텀이 강하다. 비용 구조상 클라우드 인프라 운영비가 상당 부분(서버 50%, 스토리지 20%)을 차지하며, 나머지는 R&D 와 고객지원이 주를 이룬다. Azure 는 Dynamics 와 연계된 통합 솔루션 제공 역량이 강점이며, AI 관련 PaaS 경쟁력도 점차 강화 중이다.

매출액 영업이익 추이 및 전망



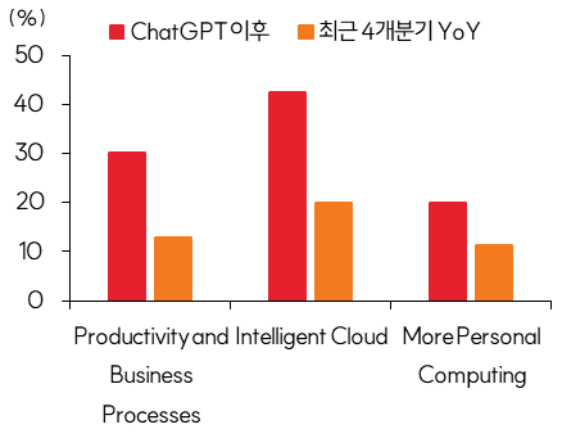
자료: Bloomberg, SK 증권

영업이익 성장률 추이 및 전망



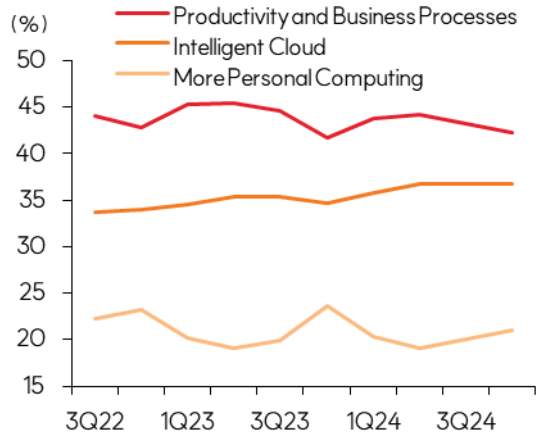
자료: Bloomberg SK 증권

사업부별 매출액 성장률



자료: Microsoft, SK 증권
주: Microsoft 365 사업부는 Productivity and Business Processes 부문
주: ChatGPT 이후는 CY4Q24 매출액 대비 최근 분기 매출액 과의 성장률

사업부별 매출액 비중 추이



자료: Microsoft, SK 증권
주: Microsoft 365 사업부는 Productivity and Business Processes 부문

Microsoft 365 요금제 및 특징 비교

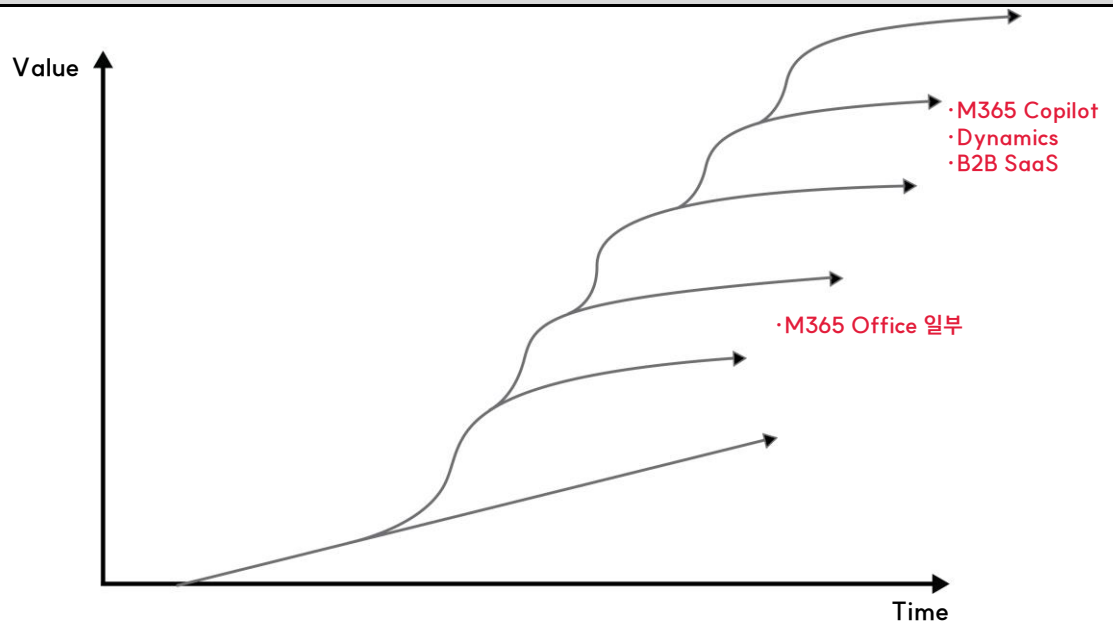
구분	요금	구성
M365 E3	$36	Office, Outlook, Teams 등 기본 엔터프라이즈
M365 E5	$57	보안, 컴플라이언스, 통화 기능 추가
Copilot	+$30	Office 강화 (요약, 생성) Teams 내 회의 내용 요약

자료: Microsoft, SK 증권

M365 Copilot Key Features

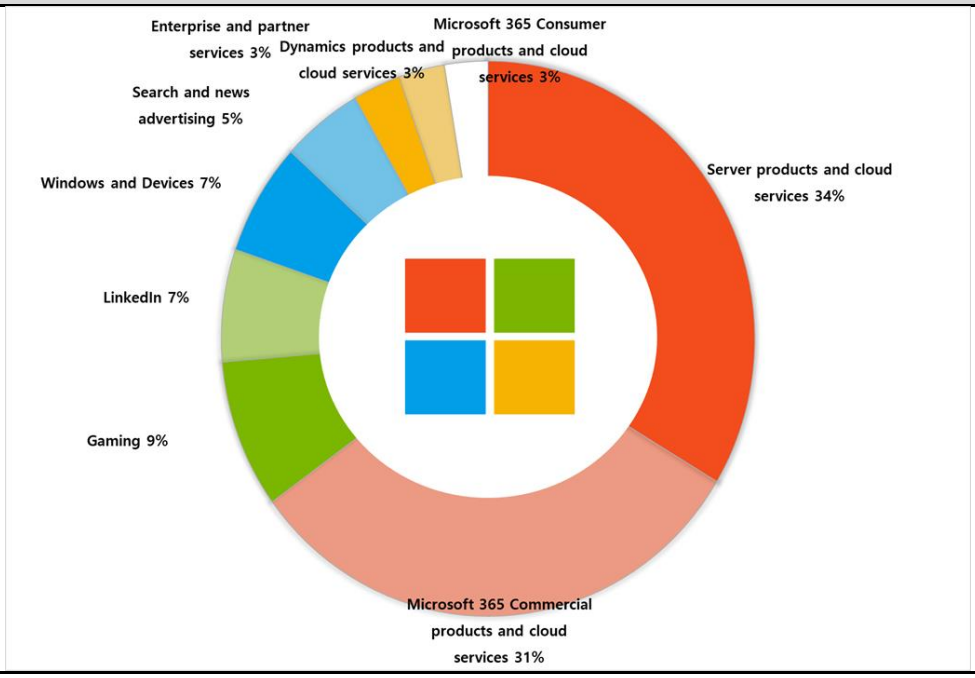
자료: Microsoft, SK 증권

사업부별 AI 단계 포지션



자료: SK증권 / 주: CSP 사업은 전 AI 단계에서 수혜

MSFT 최근 4개분기 합산 사업부별 매출 비중



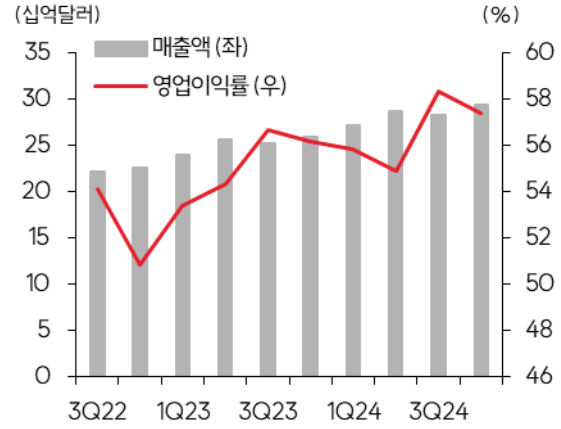
자료: Microsoft, SK증권

선결 과제 필요한 시장

B2B AI 서비스의 본격적인 확산을 위해서는 두 가지 기술적·산업적 과제가 선결되어야 한다. 첫째, AI 모델의 고도화가 필요하다. 기업 데이터를 무리 없이 통합할 수 있는 RAG 기술의 개선과 함께, 모델의 신뢰도(Reliability) 제고가 요구된다. 특히, AI 가 단순 응답형 챗봇을 넘어 Agent 형태로 진화할 경우, 모델의 판단 오류 가능성은 기업 운영에 직접적인 리스크로 작용할 수 있다. Agent 는 Reasoning 기반 모델 이상 수준의 긴 context length 와 높은 정확도를 필요로 한다. 공정이 길어지는 가운데 수율을 증가시켜야하는 상황과 유사하다고 볼 수 있다.

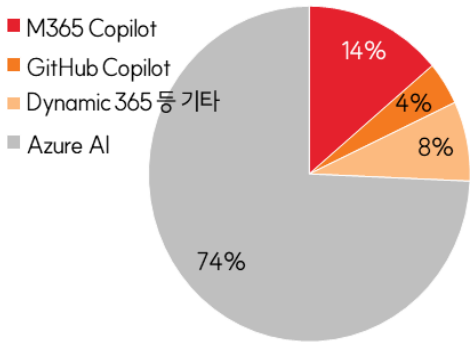
둘째, 고객사 측의 도입 준비 부족이다. 대표적인 문제가 Data Silo 와 Migration 복잡성이다. IT 시스템 도입 컨설팅사인 IBM 은 2025 년 3 월 Morgan Stanley 컨퍼런스콜에서 "기업 데이터의 99%가 아직 AI 에 활용되지 못하고 있으며, AI 전환을 위한 노력의 80% 이상이 Data 를 준비하는데 사용 중"이라 진단했다. Fortune 500 기업들의 90% 이상은 하이브리드 클라우드 전략을 사용 중이나 전체 미국 시장의 Cloud 사용률은 아직 50% 남짓이다. 이러한 한계들로 실제 생성형 AI 프로젝트를 운영하는 기업은 현재 IBM 고객사의 26%에 그친다. Accenture 역시 최근 실적발표에서 클라우드 마이그레이션과 데이터 코어 미비는 명확한 구조적 병목 요인으로 꼬집었으며 기업들마다 준비된 정도가 천차만별이라고 언급하였다. AI 관련 인력 부족 문제까지 고려하면, 해당 병목은 단기간 내 해소되기 어려울 전망이다.

OpenAI 사용자 수



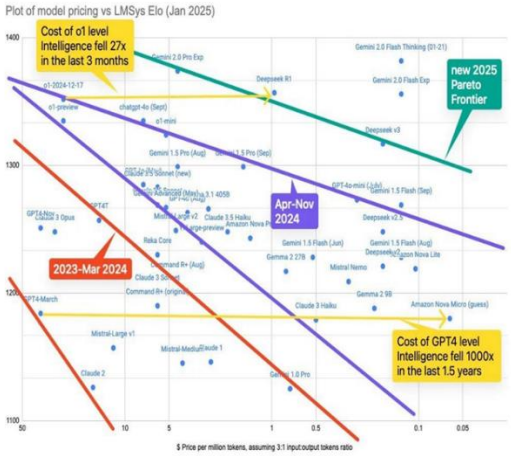
자료: 산업 자료, SK 증권

Microsoft AI ARR 구성 (추정), Azure AI 호스팅이 70% 상회



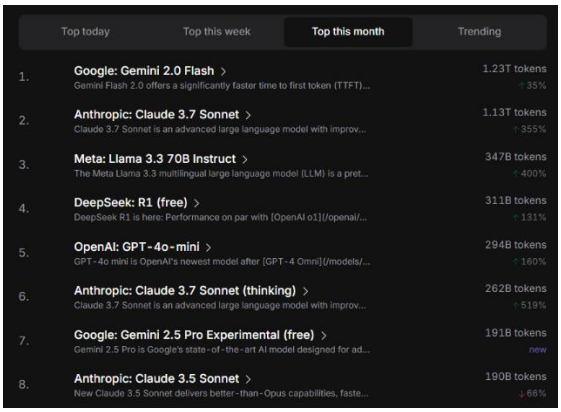
자료: SK 증권
주: M365 Copilot 월 평균 가격 $30, 사용자 수 5M 가정
Github Copilot $15, 사용자 수 3M 가정

시간에 따라 Pareto 가 지속적으로 경신되는 B2B AI 모델 시장



자료: 산업 자료, SK 증권

B2B API 호스팅 모델 월간 순위, 신규 모델 위주 다양한 모델들 포진

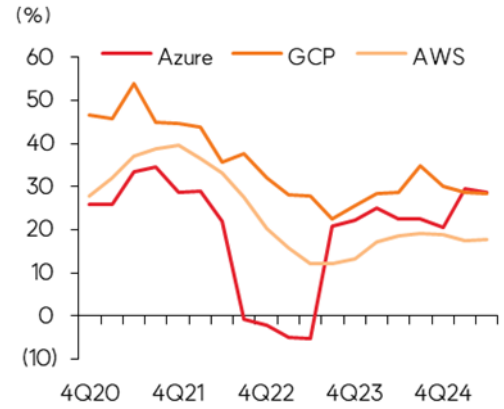


자료: Openrouter, SK 증권

API 호스팅 순위, 불과 3 개월 이전인 연말과 순위권 모델이 매우 다른 모습

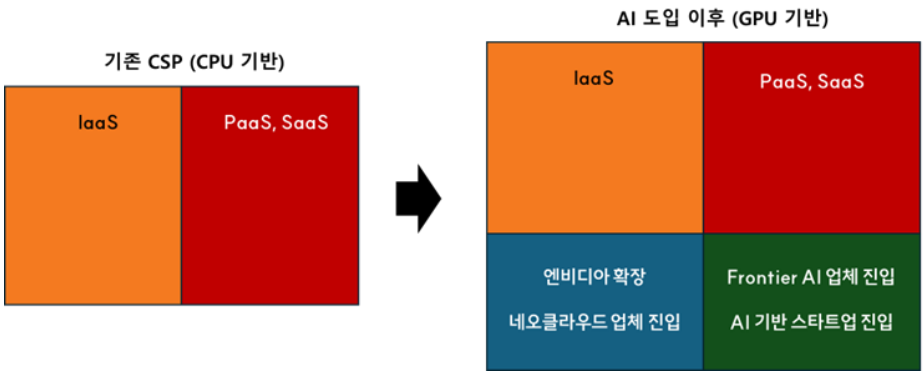
자료: Openrouter, SK 증권

클라우드 3 사 매출액 성장률 추이



자료: Bloomberg, SK 증권 / 주: CY 기준

AI 도입으로 인한 CSP 산업 구조 변화



자료: SK 증권

Compliance Notice

작성자는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

- 본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.
- 투자판단 3 단계 (6 개월 기준) 15%이상 → 매수 / -15%~15% → 중립 / -15%미만 → 매도

2025-04-14

아마존(AMZN)

AI 활용 핵심은 비용절감

SK증권 리서치센터

Analyst
박제민

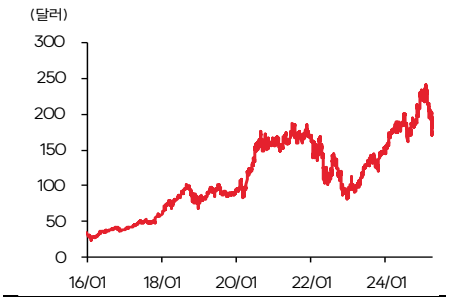
jeminwa@sks.co.kr
3773-8884

기본 정보	
국가	미국
상장거래소	NASDAQ
결산 기준월	12 월
시가총액 (십억달러)	1,891
시가총액 (조원)	2,712.1
현재주가 (달러)	178

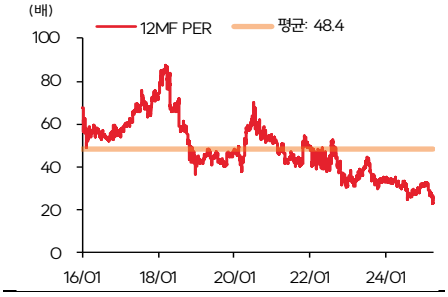
기업 개요 (Bloomberg)

아마존 닷컴(Amazon.com, Inc.)은 다양한 상품을 제공하는 온라인 소매업체. 동사는 도서, 음악, 컴퓨터, 전자기기, 기타 다양한 상품을 제공한다.

주가 추이



12MF PER 추이 및 평균



전사가 AI 수혜 가능

아마존은 글로벌 1 위 CSP 사업자인 AWS 를 통해 AI 밸류체인에서 핵심 역할을 하고 있다. AWS 는 아마존 전체 영업이익의 58%를 차지하며, 최근 AI 수요 증가로 성장률 재가속 국면에 진입했다. 본업에 해당하는 물류 사업 역시 AI 의 수혜가 가능하다. 트랜스포머 기반 AI 를 활용한 수요 예측, 재고 배치, 노드 간 경로 최적화를 진행 중으로, GPU 성능 향상과 알고리즘 개선의 직접적인 수혜가 기대된다. 소비자 접점에서도 AI 활용이 증가하고 있다. 아마존 홈페이지는 AI 기반 피드와 광고 엔진을 고도화 중이며, 셀러들을 위한 AI 제품 등록 자동화 툴(Seller Tools)은 2024 년 기준 셀러 4 명 중 1 명이 사용하고 있다.

핵심 효과는 PhysicalAI 를 통한 물류 비용 절감

중장기적으로 가장 큰 효과는 'Physical AI'를 통한 물류 비용 절감이다. 아마존은 시각 지능, 경로 계획, 디지털 트윈 시뮬레이션 등 다양한 로보틱스 영역에서 엔비디아의 AI 기술을 폭넓게 도입하고 있다. 최근 도입된 물류 로봇 Sparrow, Cardinal, Proteus 는 개발 단계부터 Omniverse 를 통한 합성 데이터 생성 및 시뮬레이션을 활용했다. 특히 Proteus 는 로보틱스 시뮬레이션 플랫폼 Isaac Sim 기반의 강화학습 시뮬레이션을 사용했다. 로봇 도입으로 효율적인 물류 네트워크 구축이 가능해져 Same-day 배송 비중도 늘어나고 있다.

아마존의 전체 고용 인원은 2024 년 기준 156 만 명, 연간 Fulfillment 비용은 980 억 달러에 달한다. 인건비 비중을 보수적으로 절반으로 가정하면 로봇 도입으로 최대 500 억 달러의 비용 효율화 여력이 존재한다. 아마존은 Sparrow 로봇의 본격 도입만으로도 약 2.8 만 명의 인력 절감 효과가 발생해 연간 10 억 달러 규모의 비용 절감이 가능할 것으로 추정했다. 이처럼 Physical AI 는 아마존의 서비스 전반에서 Q(Quantity)와 C(Cost)를 동시에 개선할 수 있는 핵심적인 성장 동력이다.

영업실적 및 투자지표 (FY기준)

구분	단위	2021	2022	2023	2024	2025E	2026E
매출액	십억달러	470	514	575	638	697	768
영업이익	십억달러	25	12	37	69	80	97
순이익(지배주주)	십억달러	33	3	30	59	68	82
EPS	달러	3.3	0.3	3.0	5.7	6.3	7.6
PER	배	78	246	62	32	26	21
PBR	배	13	13	9	7	5	4
EV/EBITDA	배	22	20	18	19	12	11
ROE	%	29	2	17	24	20	19

자료: Bloomberg, Consensus(25.04.11)

SK 증권

미국 소비 내 전자상거래 비중 16%, 아직 갈 길이 멀다

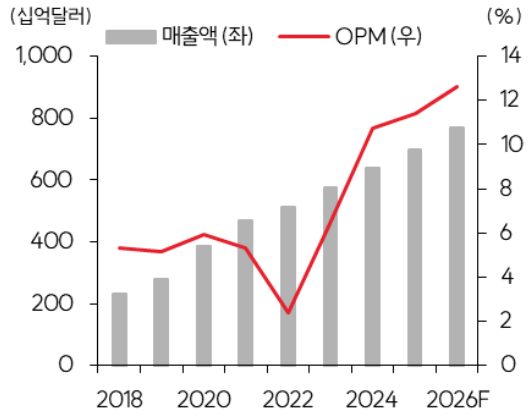
Amazon의 주요 사업은 AWS, 아마존 물류 배송 (1P는 Online Stores로, 3P는 Third-Party Seller Services, 오프라인은 Offline 사업부로 구분), Amazon Prime에 포함된 플랫폼 사업 및 이에 포함되는 광고 서비스로 구분된다.

AWS는 CSP 1위 사업자로, EC2, S3, RDS 같은 IaaS/PaaS가 핵심이고, AI SaaS로는 SageMaker, Bedrock을 제공한다. 가장 오래된 사업자로 공공·금융 부문 고객사를 가장 폭 넓게 보유 중이다. 그러나 대형 고객사들은 간단한 분석 위주의 수요로 SW 위주의 고객사들의 단가가 더 높다. 향후 AI 전환으로 해당 고객사들의 ASP 상승이 기대된다.

Online Stores는 Amazon.com 기반 1P 유통 비즈니스다. 다양한 브랜드 및 PB 상품을 판매하며, Prime 중심의 회원 기반으로 운영된다. 이커머스 분야는 가격 경쟁이 심하여 가격 상승을 통한 이익률 개선이 제한된다. 이에 아마존은 Prime 고객사를 통한 락인 효과 및 대규모 집행을 통한 비용 효율화에 집중 중이다. Third-Party Seller Services는 아마존 플랫폼을 외부 판매자에게 개방해 수수료를 받는 사업이다. 수익은 판매 수수료(8~15%), 물류 대행비, 광고비로 구성되며 광고는 Advertising 부문에 귀속된다.

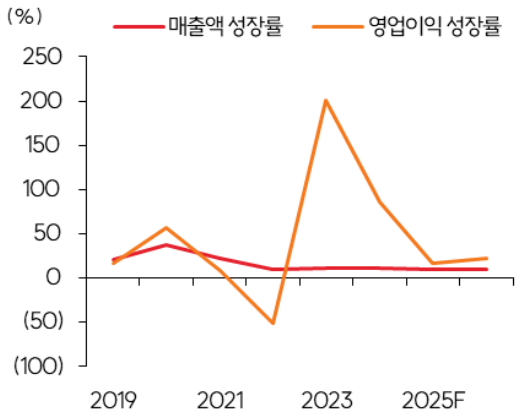
2024년 미국 전체 소매 판매에서 전자상거래 비중은 약 16%로 전년 대비 1%p 증가하였다. 여전히 전체 소매판매의 84%는 오프라인에서 이루어지는 중이다.

매출액 영업이익 추이 및 전망



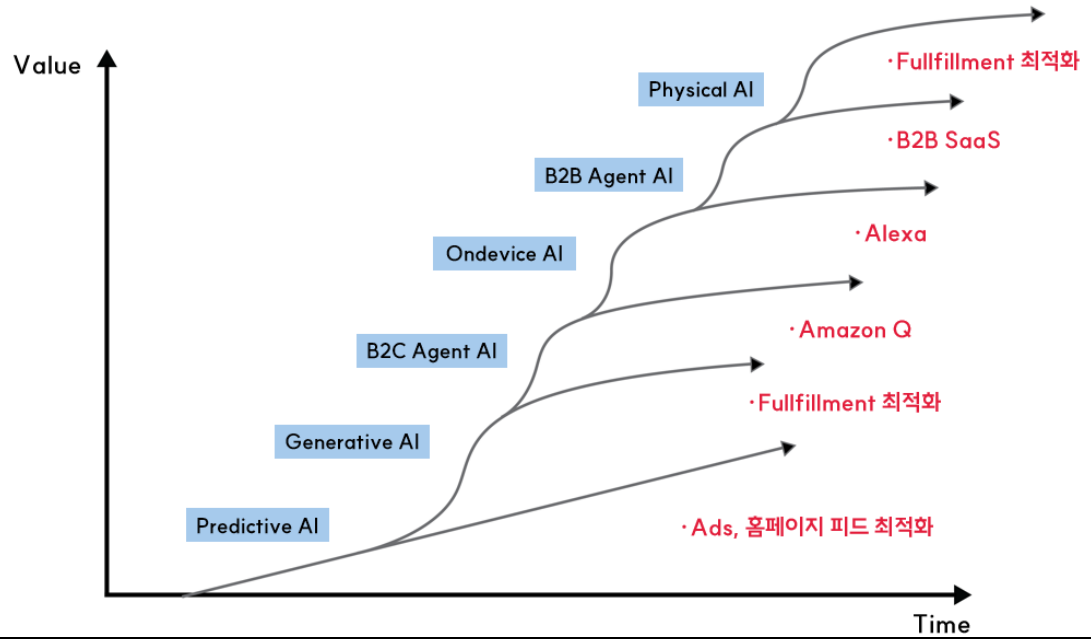
자료: Bloomberg, SK 증권

영업이익 성장률 추이 및 전망



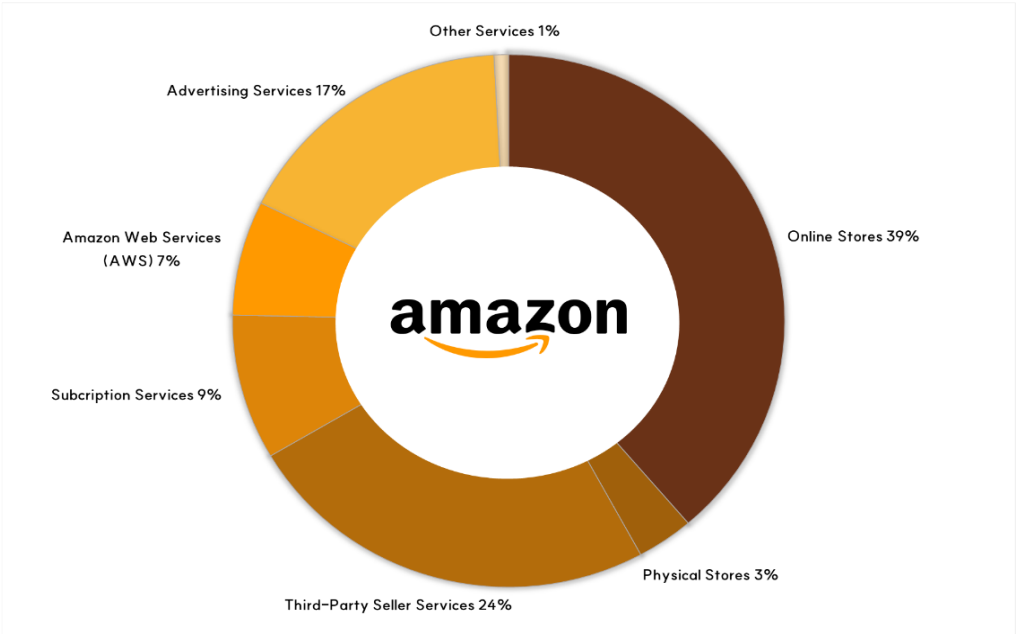
자료: Bloomberg SK 증권

사업부별 AI 단계 포지션



자료: SK증권 / 주: CSP 사업은 전 AI 단계에서 수혜

FY2024 사업부별 매출 비중



자료: Amazon, SK증권

3 단계로 물류 혁신, 로봇 도입으로 원가 절감

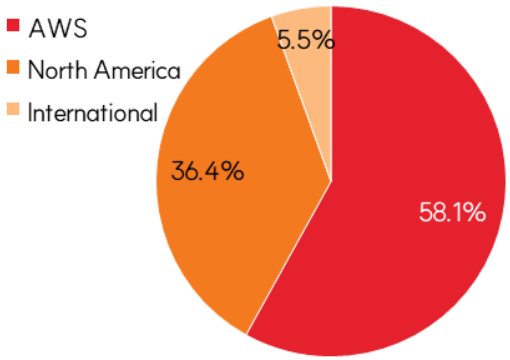
아마존은 물류 혁신을 세 단계에 걸쳐 진행 중이다. 1 단계는 아웃바운드 배송 효율화이다. 포장·분류·적재 등 배송 전 과정의 자동화를 강화하고 차량 적재 최적화 등을 통해 고객까지의 배송을 효율화했다. 2024 년 Same-Day 및 Next-Day 배송 건수는 90 억 건으로 전년 대비 28% 증가, 관련 배송 센터 수도 60% 증가했다.

2 단계는 인바운드 배송 혁신이다. 2H24 부터 공급업체에서 창고까지의 운송 과정에 AI 기반 수요 예측을 통한 재고 배치를 도입 중이다. 자체적으로 평가한 이상적 위치의 재고 비중이 2024 년 2 분기 25%, 4 분기 40%까지 증가했다.

3 단계는 로보틱스 아키텍처의 대규모 전환이다. 2026 년부터 로봇 활용도를 10 배 늘린 물류센터의 도입을 확대할 계획이며, 현재 Sequoia, Robin, AVI 등이 도입 중이고 Sparrow(픽킹), Cardinal(분류), Proteus(자율주행) 등도 추가 통합될 예정이다. 로봇 도입으로 물류 처리 속도는 최대 25% 향상, 배송 단가는 최대 25% 절감이 기대된다.

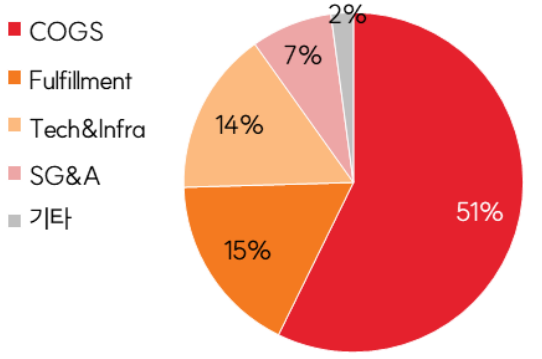
2024 년 블랙프라이데이 시즌에는 루이지애나 센터에 신규 로봇을 전면 적용해 역대 최대 하루 매출인 108 억 달러를 소화하며 효과를 입증했다. 향후 기존 시설을 대체하고 신규 센터 확장을 통해 Same-Day 및 Next-Day 배송 가능한 품목 범위를 넓혀갈 계획이다.

2024 사업부별 영업이익 비중



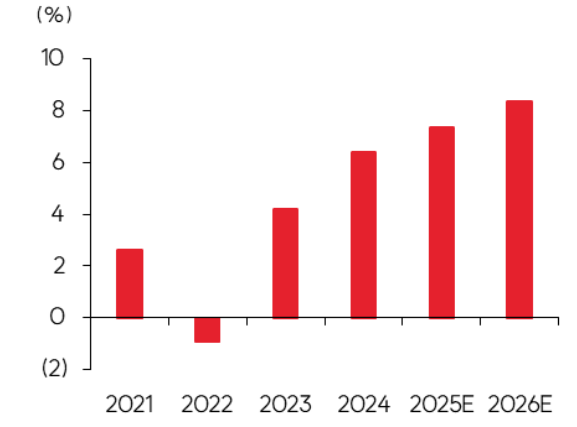
자료: Bloomberg, SK 증권

전체 운영비에서 COGS 제외, 인건비는 보수적으로 절반 이상



자료: Amazon, SK 증권

North America 사업부 영업이익률 추이 및 전망

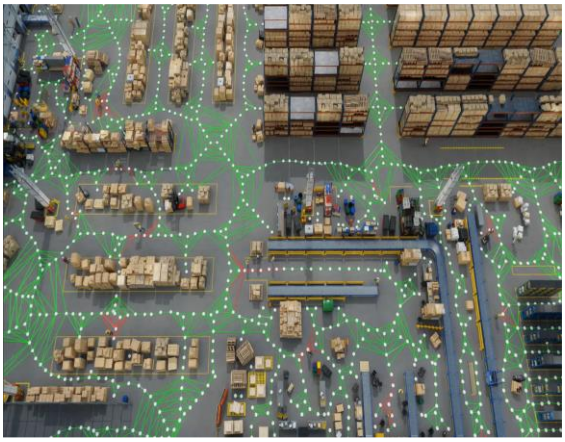


자료: Bloomberg, SK 증권

아웃바운드 혁신, Center 수 증가로 Same-day 배송 증가

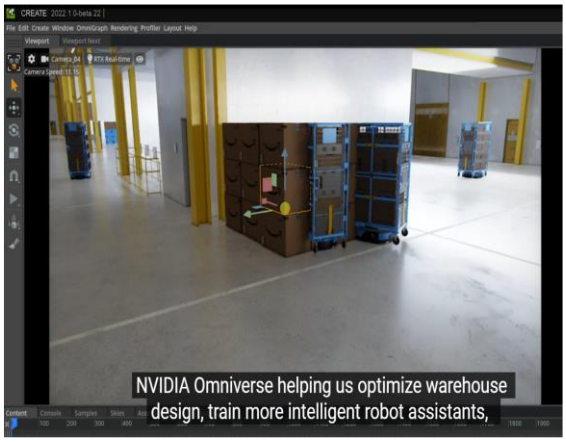
자료: Amazon, SK 증권

AI 를 활용한 물류 동선 최적화



자료: NVIDIA, SK 증권

물류센터 디자인을 엔비디아 SW 를 통해 최적화



자료: NVIDIA, SK 증권

12 세대 물류센터 적용 Sparrow Picking Robot 를 digital twin 을 통해 학습

자료: NVIDIA, SK 증권

안정적 점유율을 바탕으로 효율화에 진심

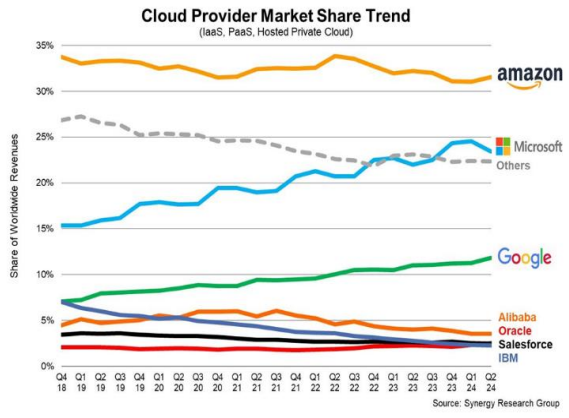
AWS 는 가장 높은 시장점유율을 보유한 CSP 로, 대기업 및 공공기관을 포함한 폭넓은 고객층을 기반을 가졌다. ChatGPT 등장 이후 Microsoft 의 성장률 급증 시기에 성장률이 비교적 뒤쳐졌으나 경쟁력을 회복하여 순항 중으로 보인다.

OpenAI 독주 이후 Anthropic 에 대한 투자 강화로 Claude 기반 API 호출 수요가 빠르게 증가 중이다. Anthropic 에 의존 중이던 LLM 또한 자체 LLM 인 Nova 를 통해 기존 모델 대비 75% 비용이 절감된 형태의 AI 서비스를 제공 중이다.

AI 인프라 확장을 통한 비용 효율화도 적극 진행 중이다. Azure 의 Maia 는 아직 양산 계획이 불투명하고, 구글은 TPU 를 활용한 Gemini 가 이미 출시된 반면 AWS 는 Tranium 은 이번 2 시리즈부터 본격 확장 국면이다. 주요 고객사는 Anthropic, Apple, Databricks 등으로 파악된다. AWS 에 의하면 Tranium 2 는 출시 이후 수요가 급증했고, 제조사로부터 생산량 2 회 증액 요청이 있었다.

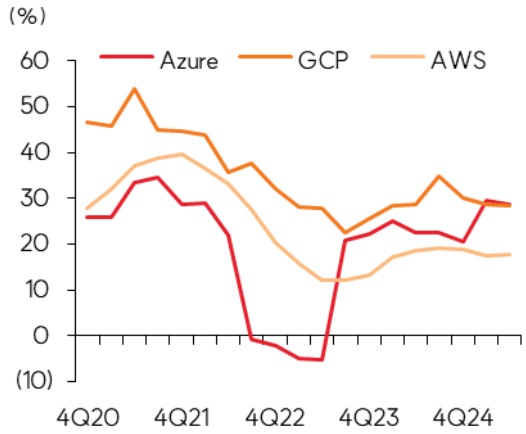
최신 버전인 Tranium 2 는 Anthropic 의 차세대 Frontier 모델 학습에 활용 중(Project Rainier)이며, 최근 NVIDIA Blackwell 대응 차원에서 Tranium 기반 서버 가격을 대폭 인하하는 전략적 조정도 단행했다. CPU ASIC 인 Graviton 또한 기존 x86 대비 효율적인 CPU 로 현재 90% 이상의 대형 고객이 채택 중이다.

CSP 점유율 압도적 1 위 지위를 유지 중



자료: Synergy Research, SK 증권

클라우드 3 사 매출액 성장률



자료: Theinformation, SK 증권

Compliance Notice

작성자는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

- 본 보고서는 기관투자가 또는 제 3 자에게 사전 제공된 사실이 없습니다.
- 투자판단 3 단계 (6 개월 기준) 15%이상 → 매수 / -15%~15% → 중립 / -15%미만 → 매도