2024. 02.14
그림: DALL·E 3 (AI harnessing “RAG”)

AI Weekly, 특이점이 온다 #6

W-147 to A.G.I.

(AGI’s ETA: 2026.12.09)

디지털리서치팀 한종목
chongmok.han@miraeasset.com
02-3774-1498

MIRAE ASSET
미래에셋증권

Highlight of the Week

- **칩 수요전망의 천장을 뚫어버린 샘 알트만**

샘 알트만은 AI 반도체 개발에 최대 7조 달러를 투자할 계획을 발표. 엔비디아와 AMD의 CEO들도 AI 기반 데이터센터와 AI 가속기 시장의 급성장을 예측. 알트만은 반도체 확보가 AGI 개발의 핵심 장애물임을 지적한 바 있음. GPT-5(혹은 GPT-4.5) 출시를 기점으로 OpenAI가 에이전트 기능을 출시한다면 AI로 인한 생산성 향상은 상상 그 이상일 것. 물론 이를 위해서는 더 큰 규모의 반도체 확보가 필수.

- **"Week of Nvidia"**

엔비디아가 구글을 제치고 세계 4위 상장 기업이 됨. 젠슨 황은 국가의 AI 전략에 고견을 제시할 정도로 영향력이 커짐. 엔비디아는 ASIC 설계 시장 진입을 위해 새 사업부 구축과 주요 기업들과 협력을 모색함. GPU 및 인피니밴드 기술을 포함한 ASIC 공동개발에 역량을 발휘하며, TAM을 계속해서 확장중. GH200 Superchip의 성능은 인텔 CPU와 근접하며, 저렴한 가격으로 소비자용 PC 시장에서 대중화를 예고함. 엔비디아는 " Chat With RTX " 라는 로컬 PC용 맞춤형 AI 챗봇 앱을 출시함으로써 온프레미스 서버 수요 진작에 나섬. 이 앱은 개인 문서를 기반으로 한 응답으로 할루시네이션 문제를 해결. 또한, TensorRT-LLM 기술을 통해 AI 추론 성능을 최대 4배 가속해 응답의 정확성 및 속도 향상을 특징으로 함. 엔비디아는 GPU 설계와 소프트웨어 개발력 모두에서 시너지를 내며, AI를 통한 기술 혁신 속도를 가속화하고 있음.

- **AGI는 로보틱스가 필수… 로보틱스는 온디바이스 AI가 필수**

화웨이 연구진은 물리적 세계와 상호작용하는 '실체화 AI'의 필요성을 강조. AGI 달성의 핵심 단계로 로보틱스와 멀티모달을 꼽았고, 이를 위해 에너지 효율적인 반도체 혁신이 필수적. 중국은 전기차 산업에서의 헤게모니를 안드로이드 산업에서도 이어가기 위한 국가적 총력전을 펼칠 것이라 판단. 한편, 로봇공학의 새로운 연구 "Dream2Real"은 로봇이 인간처럼 '세계를 이해'하는 가능성을 제시. 제로샷 학습이 가능.

- **Paper of the week: Retrieval Augmented Generation(RAG)**

지속적으로 업데이트 되는 지식을 LLM이 반영할 수 있도록 외부 데이터베이스를 활용. LLM은 DB에 저장된 관련 정보를 검색해 응답의 정확성/효율성을 증대. 즉, 생성 AI의 고질적 문제인 환각을 완화할 수 있음. ChatGPT와 같은 대형 모델과 결합하면 모델의 재훈련 과정 없이도, 챗봇이 최신 정보에 기반한 응답을 할 수 있음. 다만 아직은 멀티모달 기능의 부재, 긴 컨텍스트에서의 정확성 하락 등이 풀어야 할 숙제.

일	월	화	수	목	금	토
11	12 **(ANET)** 실적발표	13 **(DDOG)** 실적발표 **(UAE)** AI Summit(~14) **(미국)** AI Summit(~14)	14 **(EQIX)** 실적발표	15 **(UAE)** DATA NEXT 2024 **(DDOG)** 인베스터데이 **(루닛)** 실적발표	16	17
18	19 **(4패러다임)** 실적발표	20 **(사우디)** 오픈소스 Summit **(캐나다)** 38차 AAAI(~28)	21 **엔비디아 실적** **(SNPS)** 실적발표 **(NVDA)** 실적발표 **(독일)** Generative AI for Automotive(~22)	22	23	24
25	26 MWC 2024(~29) **(U)** 실적발표	27 **(영국)** Generative AI for Marketing(~28)	28 **(바이두)** 실적발표 **(SNOW)** 실적발표 **(CRM)** 실적발표 **(미국)** Post-Industrial Summit(~29)	29 **(UAE)** 중동 Banking AI & Analytics Summit	3/1	3/2

자료: Bloomberg, 외신, 미래에셋증권 디지털리서치팀

AI issues incoming

칩 수요전망의 천장을 뚫어버린 샘 알트만

- OpenAI의 CEO **샘 알트만 "(우리와 같은) 소규모 팀은 보통의 성과에 만족하지 않고, 단호하고 집중된 배팅을 하기 때문에 혁신하기가 쉬운 것이다. 그게 전부다."** 위와 같은 본인의 말을 증명이라도 하듯, 알트만은 또 한 번 세상을 놀라게 함. 다만 이번에는 기계어로 짜여진 제품이 아니라 자연어로써 AI 반도체에 관한 광활한 야심을 드러냄. 그는 **AI 개발의 주요 총탄이 되는 칩의 생산을 늘리기 위해 최대 7조 달러(전세계 GDP의 10%)의 자금 모집 계획을 천명.** 2023년 전세계 칩 판매량이 약 0.53조 달러였으니, 샘 알트만의 배팅 규모는 작년 전체 판매액의 '디커플(10배)'이라고 할 수 있음. 한편, AI 개발에 있어 최고급 총탄인 GPU를 개발하는 **엔비디아의 젠슨황 CEO**는 AI 기반 데이터센터 규모에 대한 본인의 생각을 밝힘. 알트만의 7조 달러 전망치와는 다소 '디커플링'되어 있지만, **본인의 기존 예측치보다 2배 늘려 "2조 달러 규모의 데이터센터 시장이 앞으로 4~5년 내 AI 컴퓨팅으로 재편될 것"이라 발언**. 다시 말해, 반도체 칩 수요 규모는 **2조 달러와 7조 달러 사이 어딘가 위치해 있을 가능성이 높음**. 이와 관련해 AMD의 리사 수 CEO도 AI 가속기(시스템이 아닌 오직 칩)만이 가지는 유효시장규모가 2027년 0.4조 달러에 달할 것이라 예측. 이러한 **엄청난 수요전망 상향은 점차 컨센서스를 찾아가는 모양새.**

- 샘 알트만은 "아무도 나보다 빨리 나아갈 수 없다"라고 트윗을 남긴 이후 위와 같은 칩 생산 네트워크를 위한 청사진을 그림. 바꿔 말하면, GPU 등 '하드웨어 확보가 우선돼야 남들보다 빨리 발전할 수 있다"라는 말. 즉 **현재 기준으로 AGI 개발의 발전 경로에 발목을 잡는 것은 기술 혁신의 문제가 아니라 반도체 공급의 문제로 보고 있음을 유추할 수 있음**. 최근 OpenAI는 AGI로 가기 위한 노력의 발자취를 바삐 남기고 있음. **지난 13일, ChatGPT에다 사용자의 특정 정보를 기억해서 향후의 대화 상호작용을 개선할 수 있는 "메모리" 기능을 도입.** 물론 개인정보보호 목적으로 원치 않다면 기능을 끌 수 있음. ChatGPT가 점차 개별 사용자에게 맞춰지고 있고, **"맞춤형 AI 에이전트"를 개발하기 위한 중대한 걸음**. 이와 관련해, **OpenAI는 텍스트 입력, 마우스 클릭 등 사용자의 장치를 대신 제어할 수 있는 에이전트와, 인터넷에서 데이터를 모아주고 각종 서비스를 대신 예약해주는 에이전트를 개발 중.** AI 에이전트라는 개념은 작년에도 많이 회자됐지만 일부 코어 개발자들의 전유물에 가까웠음. 올해 공개가 예상되는 GPT-5(혹은 GPT-4.5) 출시를 기점으로 OpenAI가 에이전트 기능을 출시한다면 AI로 인한 생산성 향상은 상상 그 이상일 것. 물론 **이를 위해서는 더 큰 규모의 반도체 확보가 필수.**

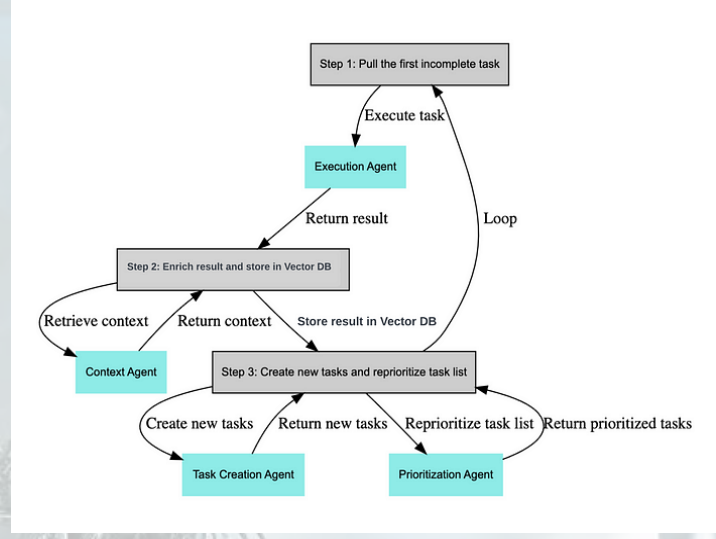
OpenAI의 샘 알트만 CEO의 트윗, 'GPU에 더 큰 배팅이 필요하다'

자료: X(@sama), 미래에셋증권 디지털리서치팀
주: 한국 시각 지난 10일 아침 8시 경에 작성한 트윗

ChatGPT의 새로운 업데이트, '사용자 정보 기억 옵션'

자료: OpenAI Blog

개발자들에게 인기인 GPT 에이전트 "BabyAGI" 작동원리



자료: Github(@yoheinakajima)

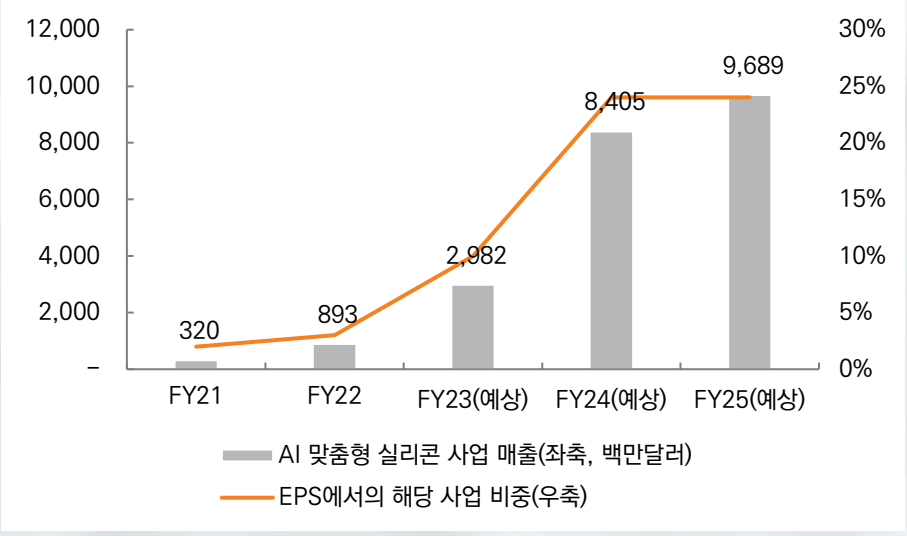
AI issues incoming

Week of Nvidia (1)

- 2월 14일 기준, **엔비디아는 구글마저 제치고 세계에서 4번째로 큰 상장 기업**이 됐음. 커진 덩치처럼 젠슨 황의 어깨에도 힘이 잔뜩 들어가고 있음. 테크 업체의 CEO로서가 아니라, 국가별 "Sovereign AI" 전략에 대해 고견을 전해줄 정도로 국제적 영향력이 확대됐기 때문. 젠슨 황은 며칠 전 두바이에서 열린 "세계 정부 정상회담"에서의 **UAE의 AI 장관과의 회담에서 "모든 국가가 자체 AI 생산 기능을 소유해야 한다면서, 데이터는 새로운 GOLD"라고 설파.** 지난주 자료에서 기술한 바 있듯, 중동은 AI에 상당히 진심. PwC는 AI로 인해 2030년까지 중동 지역경제에 3,200억 달러의 성장을 전망. 젠슨 황의 방문은, 두바이 전력수도청의 디지털 부문 자회사와 엔비디아의 "그린 데이터센터" 건립 확약을 이끌어냄. CEO가 지구촌을 누비며 영업을 하는 것과 동시에 엔비디아는 추가적 성장을 위한 내실을 쌓고 있음. **엔비디아는 하이퍼스케일러들의 AI 전용 맞춤형 반도체(ASIC) 설계 흐름에 대응하기 위해 새로운 사업부를 구축하고 있다는 정보가 유출**. 현재 하이퍼스케일러들이 가진 데이터센터의 ASIC 설계는 브로드컴(AVGO)과 마벨(MRVL)이 주도 하고 있음. 특히, **브로드컴의 경우 AI 맞춤형 실리콘 사업규모가 내년쯤 거의 100억 달러에 달할 것으로 추정**되는 만큼, AI 설계 IP의 강자. 구글의 TPU, 메타의 MTIA, 테슬라의 D1 칩을 공동설계. 그런데, **엔비디아는 바로 이 시장에 뛰어들기 위해 클라우드 3사 및 OpenAI의 대표자들과 최근 논의했다고 전해짐.** 엔비디아는 ASIC 설계 시장을 잠식할 수 있을만한 역량을 보유. 대부분의 AI 맞춤형 칩들이 AI 훈련(AI 추론은 나름 선방)에는 죽을 쑤고 있음. 빅테크들의 복잡한 병렬컴퓨팅 설계 역량에 있어 한계가 있기 때문. 엔비디아는 **자사의 GPU 및 네트워킹(인피니밴드) IP를 활용해 데이터센터 ASIC 공동개발에 역량 발휘가 가능할 것.** 빅테크들의 자체 칩 개발에 따른 위기를 기회로 바꾼다면, **상당한 TAM 추가 확보가 기대.**

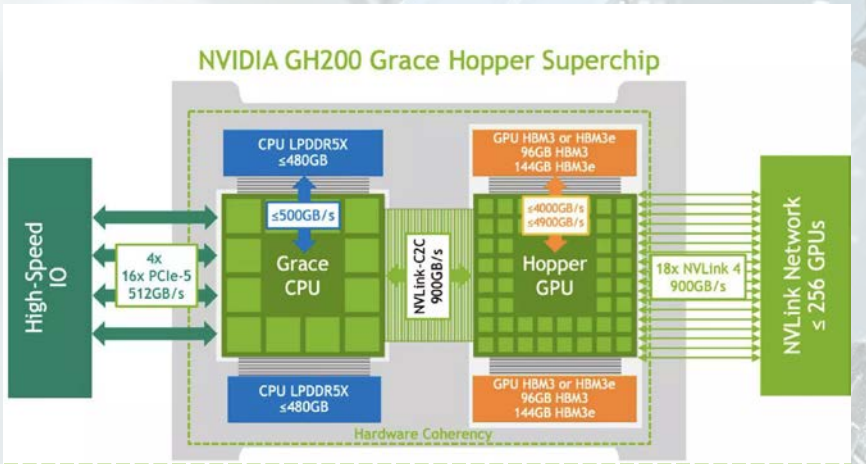
- **ASIC 설계 라이선스 사업에 있어**, 엔비디아의 역량은 이뿐만이 아님. **CPU 개발 노하우도 상당히 증가**. 실제로 차세대 실리콘인 "**GH200 Superchip"의 성능 벤치마크**가 나왔는데, **인텔의 에메랄드 래피즈 CPU 성능과 매우 근접한 결과**. RISC 기반의 CPU 설계 역량이 얼마나 발전했는지 잘 보여주는 사례. 뿐만 아니라, 올해 2분기에 출시할 이 **GH200 슈퍼칩으로 무장한 PC 시스템**의 가격이 공개. "얼리 버드" 특가이긴 하나, **4.15만 달러로 생각보다도 저렴**. 고사양 GPU 서버는 클라우드에서나 쓰이는 럭셔리 상품이었지만, (헤비) 소비자용 PC로의 곧 대중화를 시사. **데이터 유출과 같은 리스크 때문에 생성 AI 사용을 꺼리는 사용자 계층들로부터의 '온프레미스' 수요를 노리는 것**으로 판단.

現최강자 브로드컴의 AI 맞춤형 실리콘 사업 매출과 그 이익 비중 추정치



자료: Semianalysis, 미래에셋증권 디지털리서치팀
주: 브로드컴(AVGO)은 10월 결산, 해당 자료는 작년 7월 분기 실적 이후를 기준으로 추정

엔비디아의 차세대 실리콘 "그레이스호퍼(GH) 200"의 아키텍처



최대 144GB HBM3e를 탑재할 엔비디아 GH200, 올해 2분기 출시 예정
Arm v9 명령어세트를 사용, 최대 144개의 CPU 코어로 무장

자료: NVIDIA, 미래에셋증권 디지털리서치팀

GH 슈퍼칩이 탑재된 PC, '고작(?) 4만 불!'

엔비디아의 GPU 및 CPU 조합인 Grace Hopper Superchip을 탑재한 고급형 PC

576GB 모델: 480GB의 LPDDR5X, 96GB의 HBM3
624GB 모델: 480GB의 LPDDR5X, 144GB HBM3

두 모델 모두 900 GB/s의 NVLINK-C2C 인터커넥트

자료: GPTshop.ai, 미래에셋증권 디지털리서치팀
주: GPTshop.ai는 GH200를 탑재한 "최고급 데스크탑"을 구축하는 것을 목표로 하는 서버 제조사

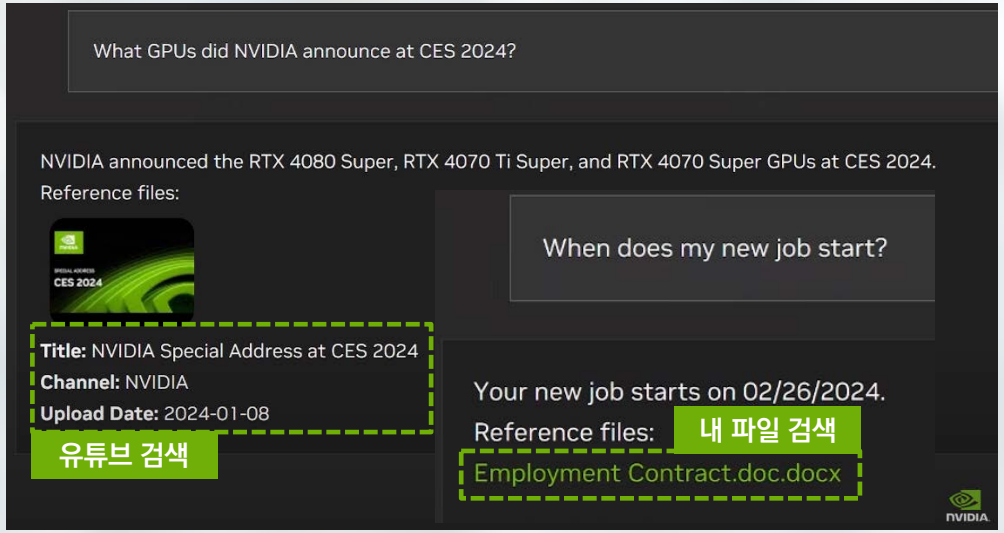
AI issues incoming

Week of Nvidia (2)

- **엔비디아는 온프레미스 서버 수요를 노리는 것에만 그치지 않고 직접 그 수요를 진작시키는 전술을 펼치고 있음**. 그 일환으로 로컬 PC에서 실행되는 '나만의 맞춤형 AI 챗봇'인 **"Chat With RTX"라는 PC 앱을 출시**. 단, 이것이 실제로 유용하려면 AI 추론에 있어 **실시간 응답의 정확성과 속도가 빨라야 함.** 13일에 출시된 이 앱은 PC에 저장된 메모장, PDF 문서, 워드 파일 등의 개인 문서의 내용을 기반으로 답변할 수 있는 챗봇. 즉, **RAG 기능을 갖춰 AI의 고질적 문제인 할루시네이션을 극복**. 뿐만 아니라, 유튜브 URL을 알려주면 링크 주소대로인 동영상을 시청해 필요한 정보를 검색해 요약하는 기능까지 있음. 무엇보다 Chat With RTX의 가장 큰 장점은 온라인 상에 개인정보를 업로드 하지 않고도 로컬로 생성 AI를 활용할 수 있다는 사실. **데이터 유출에 민감한 엔터프라이즈 수요를 이끌어낼 수 있을 것으로 판단.** 뿐만 아니라, Chat With RTX는 **사용자 쿼리에 따른 답변 속도도 빠를 것으로 기대되는데, 이는 TensorRT-LLM과 RTX 가속 기능을 지원하기 때문**. 쉽게 말해, 해당 기능들 덕분에 RTX GPU의 자원(연산을 위한 커널)을 보다 더 효율적으로 쓸 수 있게 자동/최적화하고, **최대 4배 빠른 AI 추론 성능을 뽑낸다는 점**. CUDA를 기반으로 한 TensorRT-LLM은 메타의 Llama과 Mistral 등 여러 언어모델들을 지원하는 라이브러리인데, AI 추론을 위한 전처리와 후처리 단계의 최적화를 모두 포함. 이와 관련, **젠슨 황은 "전/후처리 부분은 사람들이 언급하지 않는데, 이 부분을 효율화할 수 없다면 AI 개발 속도를 절반만 늘릴 뿐이다. 우리는 CUDA를 통해 단순히 추론 부분 가속화를 넘어 데이터 처리의 처음부터 끝까지 효율화 한다. 경쟁은 가까이 오지도 않았다"라고 언급.**

- AI 개발자 입장에서, Chat With RTX의 의의는 엔비디아 GPU를 쓰면 언어모델을 더 쉽게 튜닝해서 배포할 수 있게 된다는 것. 이 부분이 바로 엔비디아가 가진 가장 큰 해자. **GPU 설계 최강자로써 "Hardware-awareness" 알고리즘 디자인에 있어 글로벌 탑티어이기 때문에, GPU에게 임무를 하달할 소프트웨어 개발력에도 시너지가 나오는 것.** 엔비디아는 'AI로 인한 생산성 향상'을 가장 열렬히 지지하는 기업. 이것은, **AI를 사내에서 가장 많이 활용할 기업은 다름 아닌 엔비디아 본인임을 암시**한다고 생각. 실제로 엔비디아가 내놓는 연도별 신제품 라인업을 보면 기존 출시주기보다 절반으로 감소. 칩 설계 시 LLM을 활용하는 "ChipNeMo" 논문까지 발표했던 엔비디아인 만큼, AI를 통한 기술 혁신의 속도에는 가속 기어를 올리고, 불필요한 비용절감도 꾀할 것이 명백. **엔비디아는 다음주 21일에 분기 실적을 발표. 현재 시점 세계에서 가장 중요한 실적 발표**임을 기억.

엔비디아의 챗봇 앱 "Chat With RTX" 데모앱 시연, '로컬 GPU를 최적화 해서 작동!'



자료: YouTube(@NVIDIA Geforce), 미래에셋증권 디지털리서치팀

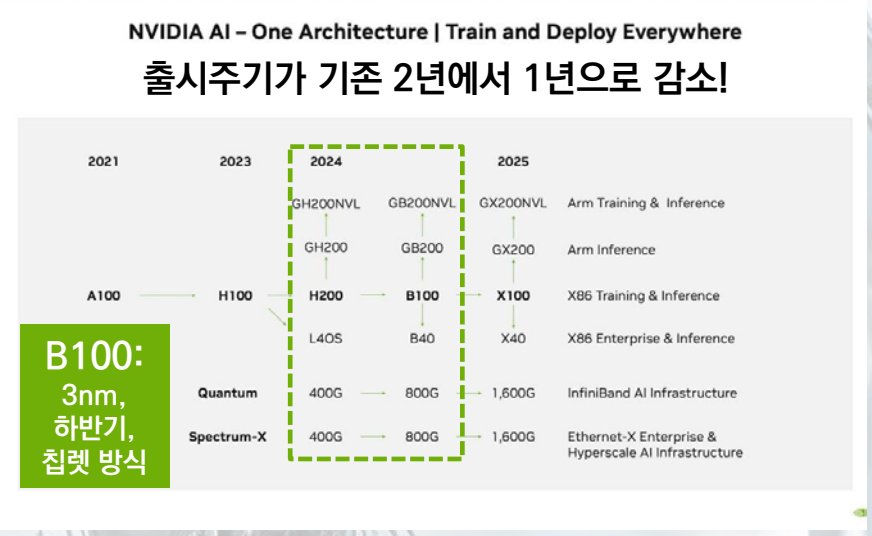
Chat With RTX의 주요 특징

최소사양	8GB VRAM을 갖춘 (권장은 16GB) NVIDIA 지포스 RTX 30 또는 40 시리즈 GPU
앱 구동	윈도우 PC 또는 워크스테이션
지원모델	메타의 Llama, Mistral
특장점	CUDA 기반의 TensorRT-LLM 라이브러리로 LLM 추론을 자동으로 최적화해 약 4배 성능 가속 RAG를 활용해서 외부 정보 (PC 내 저장된 파일 등)에 엑세스 가능한 LLM으로 작동

온디바이스 AI 수요 이끌어 내어 RTX GPU 판매 진작 전략

자료: 미래에셋증권 디지털리서치팀

엔비디아의 연도별 신제품 라인업, 'AI를 내부에서 가장 잘 활용하는 기업'



자료: NVIDIA, 미래에셋증권 디지털리서치팀

AI issues incoming

AGI는 로보틱스가 필수… 로보틱스는 온디바이스 AI가 필수

- 중국 내 AI의 선두주자인 **화웨이의 연구자들이 지난주 "Embodied AI(실체화 AI)에 대한 필요성"이라는 논문을 발표**. 화웨이는 "현재의 AI가 디지털 영역에 갖혀 있으니 물리적 세계와의 격차를 해소하는 AI 에이전트가 필요하다"고 설파. **실제 세상과 상호작용하고 그로부터 배울 수 있는 인공지능이야말로, AGI를 달성하기 위한 필수적인 진화단계**라 제안. Embodied AI는 인간의 인지 능력을 모방해 AI가 물리적 환경에서 독립적으로 행동하는 기계, **즉 로봇을 말함**. 로봇은 모빌리티 특징을 수반하고 전력을 무한정 제공받지 않음. 즉, 기계에 탑재될 '에너지 효율적' 반도체 솔루션 혁신이 반드시 필요. 해당 논문에 따르면, **화웨이 자신들의 "Ascend" 칩이 그 혁신을 위한 발전 궤적 상에 놓여 있다고 밝히고 있음**. 또한, 화웨이는 **뉴로모픽 컴퓨팅과 하드웨어-소프트웨어 공동 설계**로써 실리콘 개발 혁신을 더욱 가속화하겠다는 야심을 표출. **화웨이라는 기업이 중국 내에서 갖고 있는 정치적, 기술적, 사회문화적, 지경학적 입지가 상당하기 때문에,** AGI 개발을 향한 일종의 가이드라인을 제시했다는 점에서 의미가 있음. 현재의 전기차 생태계에서 중국의 입지가 막강하듯, **로봇 산업에서도 주도권을 끌고 나가기 위해 국가 총력전을 다할 것**. 테슬라가 "옵티머스"를 위해 Dojo와 D1칩, 그리고 자율주행 E2E 모델인 FSD에 전념하듯, **중국은 당장의 LLM 경쟁보다 그러한 기술들에 집중할 것**. 즉, 필요조건인 **미세공정 파운드리와 멀티모달리티를 위한 다양한 비정형 데이터 확보에 유리한 中 빅테크들이 더욱 부각될 것**으로 판단.

- **올해는 AI를 현실 세계로 끌어와 분명한 생산성을 올리려는 시도가 주목**을 크게 받고 있음. 올해 5월 로봇공학 국제컨퍼런스(ICRA 2024)에서 선보일 예정인 새로운 로봇 연구 **"Dream2Real"** 모델에 이목이 집중. 이 모델은 로봇이 '더 인간적인' 방식으로 세계를 이해하고 임무를 수행할 수 있는 가능성을 펼침. **주변 환경을 3D로 렌더링을 마친 로봇이 인간의 명령을 받게 되면, 마치 '상상'하는 것처럼 잠재적인 미래의 장면을 2D처럼 재구성하게 됨.** 인간의 대뇌피질에 2차원의 상이 맺히는 것과 유사. 상상했던 2D 이미지들 중에서 가장 올바른 미래가 무엇일지 평가하고, 그대로 현실 세계에서 수행하는 방식. Dream2Real의 특장점은 **이전 작업들과 달리 "제로샷" 학습이 가능**하다는 점. 즉, 어떠한 시연 작업이 필요 없음. 단, 현실 세상에 상용화되기에는 아직 갈 길이 멀지만, **분명한 팩트는 NeRF 등의 관련 알고리즘 개발속도가 가속되고 있다**는 게 미래를 긍정하게 하는 요소.

화웨이에서 강조한 "Embodied AI"는 온디바이스 AI와 로보틱스를 필요로 함

자료: Qualcomm, 미래에셋증권 디지털리서치팀

NeRF와 CLIP의 결합으로 만들어진 Dream2Real이 임무를 수행하기 위해 "꿈을 꾸고 현실로 구현하는 방식"

자료: Imperial College London "Dream2Real"
주: VLM은 Visual Language Model의 약자로, 기계가 꾼 꿈의 이미지가 임무에 얼마나 올바른 것인지(유사한지) CLIP으로 평가하는 역할

Paper of the Week

RAG(Retrieval Augmented Generation)

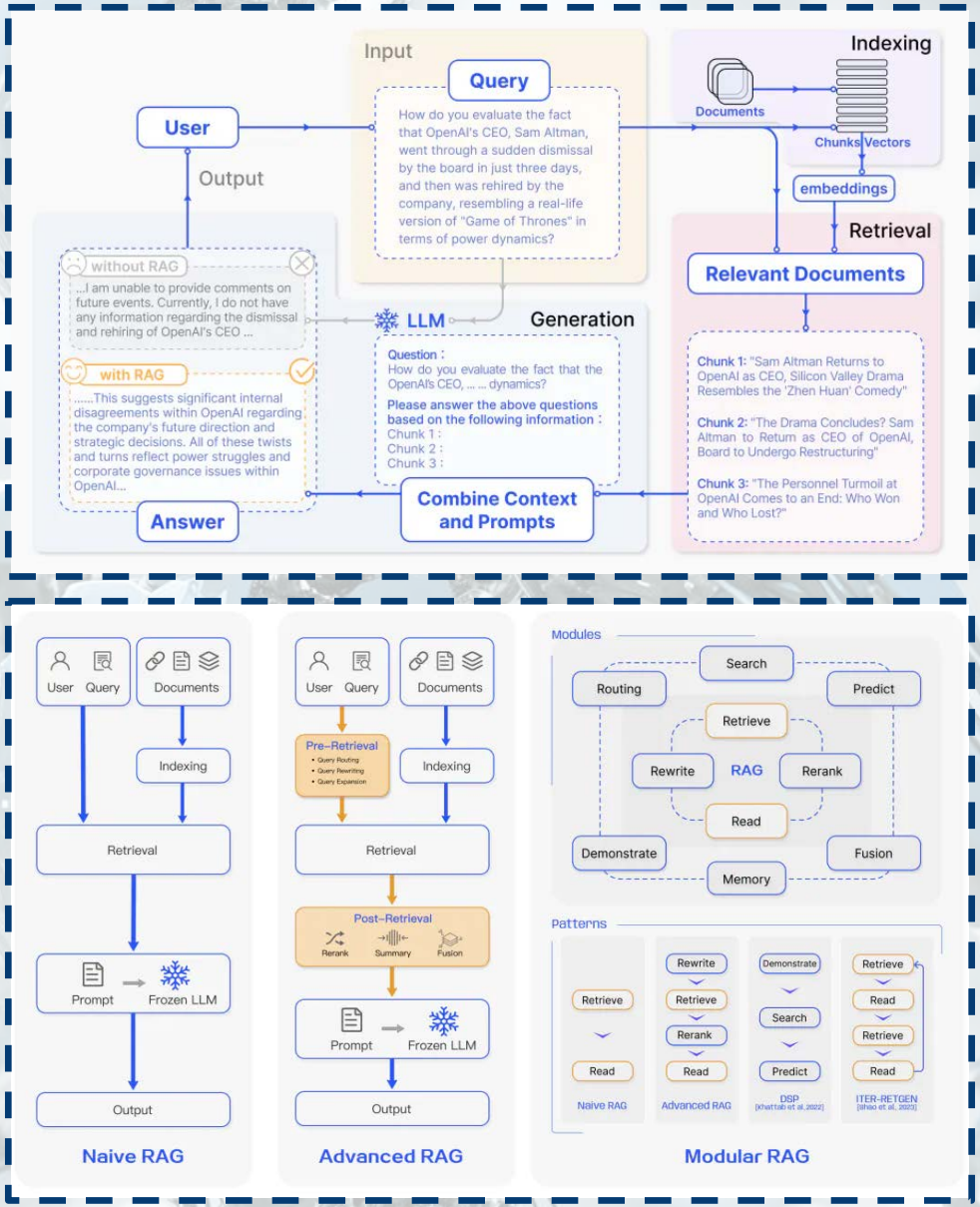
✓ Core Contents

- 검색증강생성(RAG)이란, 기존 LLM의 지식 부족, 환각 등의 문제를 보완하는 방법 중 하나로 2020년에 처음 소개됐고 최근 뜨거운 인기. **LLM의 재교육 과정 없이 외부 데이터베이스(벡터DB)를 모델에 제공**함으로써 최신 정보를 반영할 수 있는 방식.
- 입력이 주어졌을 때 데이터베이스에서 관련된 내용을 검색해서 찾아옴(Retrieval), 그 결과를 기존 입력 프롬프트와 합한 뒤(Augmentation), 텍스트 생성기(Generator)로 전달해 출력을 생성하는 것.
- 위와 같은 기본형 RAG에서 **고급 RAG**와 **모듈형 RAG**로 발전해 왔음.
  기본 구조의 경우, 1) 데이터베이스 검색 결과가 잘못 정렬되거나,
  2) 관련된 정보 중 일부만 검색되는 등의 한계가 존재했음.
- **고급(Advanced) RAG**: 검색 전, 검색, 그리고 검색 후 프로세스를 각각 최적화해 검색 품질을 개선. 특히 검색 후 컨텍스트 길이의 제한을 고려해 검색한 정보의 노이즈를 제거. **관련성이 높은 정보를 앞으로 보내거나(ReRank), 중요하지 않거나 중복되는 맥락을 압축하는(Prompt Compression) 방법을 사용.**
  - 연구에 따르면 영어는 단락이나 장 길이의 텍스트에서 약 75% 정도 중복된다는 분석이 있음
- **모듈형 RAG**: 각 단계가 모듈로 분리, 작업 요구사항에 따라 모듈의 순서 재조정 가능.

✓ Insights

- LLM의 고질적 문제인 **환각(hallucination)을 줄일 수 있다는 점이 가장 핵심.**
- 고도화된 모델이 RAG를 통해 외부 지식에 접근이 가능하게 되면 빠르게 변화하는 데이터를 반영한 맞춤형 응답이 가능해짐. **RAG의 핵심인 "벡터 데이터베이스"에 관심.**
- ChatGPT, Mixtral과 같은 LLM의 **파인튜닝에다 RAG의 강점을 결합**하는 것이 유행.
- 향후 긴 컨텍스트에서도 정확성이 유지 되는지, 여러 멀티모달에서도 확장이 되는지, 파인튜닝 모델과의 사용 최적화 방식이 추가로 개선될지 등에 집중할 필요.
  - 실제 사용자 입장에서 Prompt Compression을 RAG에 활용할시에, LLM이 정확한 응답을 생성하는 데 필요한 주요 정보를 유지하면서도 입력 토큰 길이(비용)를 6~7배 압축한다는 결과도 있음.

자료: "Retrieval Augmented Generation(RAG) for LLMs", Prompt Engineering Guide.



AI Peer Table (1)

밸류체인	종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한 줄 평
					1W	1M	3M	6M	1Y	YTD		
최종 소비자 (온디바이스, 앱 등)	애플	AAPL US	185.04	3,832.2	-2.1	-0.3	0.3	4.3	20.9	-3.8	2024-05-03	온디바이스 AI 위한 자체 경량화 언어모델 개발 중(Ajax GPT)
	테슬라	TSLA US	184.02	786.0	-0.6	-15.9	-17.7	-24.2	-5.5	-25.9	2024-04-19	자율주행 소프트웨어 FSD의 싱글스택 신경망 v12 공공출시 임박
	삼성전자	005930 KS	75,200	448.9	0.3	2.9	6.7	12.8	21.5	-4.2	2024-04-08	'24년부터는 온디바이스 AI에도 중점(Galaxy AI)
	퀄컴	QCOM US	150.04	224.6	4.1	7.0	21.9	32.7	17.7	3.7	2024-05-03	생성 AI 처리 강화용 온디바이스 AP 및 CPU 설계(X Elite)
	스냅	SNAP US	11.14	24.7	-36.2	-32.2	-1.4	11.1	4.0	-34.2	2024-04-26	이미지 생성 모델을 탑재한 증강현실 서비스(AR 필터, AR 렌즈)
	로블록스	RBLX US	42.32	35.8	4.2	3.4	9.2	41.9	22.9	-7.4	2024-05-10	생성 AI 탑재해 무료로 'Low code' 지원(캐릭터, 아이템 생성 위한 AI 도우미)
AI 비즈니스 솔루션	어도비	ADBE US	601.89	364.9	-0.9	0.9	2.0	18.4	60.5	0.9	2024-03-14	생성 AI 기반 미디어 편집 툴(Firefly)
	세일스포스	CRM US	281.15	365.0	-1.6	3.4	30.6	34.7	64.3	6.8	2024-02-28	세계 최고의 CRM 업체로 생성 AI를 탑재해 사용성 강화(Einstein GPT)
	IBM	IBM US	183.70	225.0	1.1	11.8	25.2	31.0	40.0	13.3	2024-04-24	기업 고객용 AI 어시스턴트(상담사 챗봇 구축, 업무 자동화, 코딩 지원)
	서비스나우	NOW US	773.76	212.7	-0.5	6.1	21.2	38.8	67.8	9.5	2024-04-26	워크플로우 자동화 기업으로 생성 AI를 탑재해 사용성 강화(Now Assist)
	크라우드스트라이크	CRWD US	322.44	103.9	6.7	13.8	60.4	117.5	187.5	26.3	2024-03-05	기업 고객 대상 생성 AI 챗봇 기반 클라우드 및 엔드포인트 보안(Charlotte AI)
	데이터독	DDOG US	131.68	58.0	1.9	7.1	26.3	47.2	60.1	8.5	2024-05-03	생성 AI 기반 클라우드 인프라 및 보안 모니터링(Bits AI)
	다이나트레이스	DT US	52.06	20.7	-12.6	-7.4	1.3	13.3	19.6	-4.8	2024-05-17	생성 AI 기반 어플리케이션 성능 관리 및 보안 플랫폼(Davis)
	유니티	U US	33.11	16.9	3.3	-4.3	17.6	-9.6	-11.6	-19.0	2024-02-26	올해 출시 예정인 차세대 게임엔진에 AI 기능 대거 탑재(Sentis, Muse)
	센티넬원	S US	28.99	11.6	5.6	14.4	80.6	91.7	86.7	5.6	2024-03-13	기업 고객 대상 생성 AI 챗봇 기반 클라우드 및 엔드포인트 보안(Purple AI)
	루닛	328130 KS	59,300	1.7	5.0	-22.8	-37.0	-24.2	197.6	-27.6	2024-02-16	AI 기반 암 검출/진단 보조 솔루션(인사이트)
기반 모델	알파벳	GOOGL US	145.14	2,430.3	0.7	1.7	9.9	12.0	53.4	3.9	2024-04-25	OpenAI와의 AGI 경쟁에 있어 대적할 수 있는 거의 유일한 업체(Gemini)
	메타 플랫폼스	META US	460.12	1,573.8	1.2	22.9	39.8	52.5	156.4	30.0	2024-04-26	'거의 완전한' 오픈소스 LLM 개발 및 배포를 선도(Llama)
	바이두	9888 HK	101.20	48.5	-1.8	-10.9	-2.5	-24.8	-29.6	-12.8	2024-02-28	중국 최초/최고 성능의 LLM 보유(Ernie Bot)
	네이버	035420 KS	205,000	33.3	0.2	-11.1	3.6	-6.9	-8.0	-8.5	2024-05-08	자체 개발한 한국형 LLM 개발 및 AI 서비스 출시 예정(HyperCLOVA)

자료: Bloomberg, 미래에셋증권 디지털리서치팀
주: 실적발표 예정일은 수시로 변동될 수 있음

AI Peer Table (2)

밸류체인	종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한 줄 평
					1W	1M	3M	6M	1Y	YTD		
AIOps	스노우플레이크	SNOW US	227.57	100.5	6.5	19.0	43.4	48.4	40.5	14.4	2024-02-28	AI 모델 구축, 배포, 관리를 위한 빅데이터 워크로드 플랫폼(Cortex)
	몽고DB	MDB US	472.13	45.7	6.5	20.6	24.9	32.5	117.3	15.5	2024-03-08	AI 모델을 구축할 때 필수 인프라로 떠오른 '벡터 데이터베이스' 기능 보유(Atlas)
	클라우드플레어	NET US	99.88	45.2	22.4	25.8	56.3	57.0	55.9	20.0	2024-04-26	트래픽 처리(CDN) 업체로, 실시간 AI 추론 작업에서의 역량 대두(Workers AI)
	팔란티어	PLTR US	24.01	70.1	9.8	43.3	21.8	55.8	215.5	39.8	2024-05-08	실제 조직 운영시의 의사결정 지원 및 자동화를 위한 AI 개발 플랫폼(AIP)
	포스 패러다임	6682 HK	43.00	3.4	-4.4	-20.5	-19.6	-	-	-14.3	2024-02-20	실제 조직 운영시의 의사결정 지원 및 자동화를 위한 AI 개발 플랫폼(Sage)
데이터센터	마이크로소프트	MSFT US	406.32	4,049.2	0.2	4.6	11.0	27.1	51.1	8.1	2024-04-25	OpenAI의 AI 모델 라이센스에 대한 독점권을 갖고 있는 세계 2위의 CSP
	아마존 닷컴	AMZN US	168.64	2,349.4	-0.3	9.1	18.3	21.8	69.4	11.0	2024-04-26	최고의 CSP로써 작년 9월 Anthropic에 40억 달러 투자하며 AGI 경쟁 참여
	오라클	ORCL US	113.68	419.1	-1.4	6.6	-0.0	1.3	29.1	8.2	2024-03-08	기업들의 자체 인프라 수요에 대응한 '온프레미스 서버' DB 관리 시스템의 강자
	에퀴닉스	EQIX US	832.68	104.8	-0.6	2.2	9.8	8.6	14.9	3.4	2024-02-14	물리적 공간의 데이터센터를 임대/위탁 운영하는 업계 최대 리츠 업체
반도체 인프라	엔비디아	NVDA US	721.28	2,389.4	5.7	31.8	48.4	76.6	231.2	45.6	2024-02-21	AI 훈련은 GPU가 필수, GPU 현존 최강자는 엔비디아
	브로드컴	AVGO US	1,251.65	785.9	2.4	13.0	32.8	52.4	113.0	12.1	2024-03-07	VMware 인수로 클라우드 가상화 역량도 확보한 네트워킹 반도체의 최강자
	TSMC	2330 TT	646.00	716.5	-0.3	12.2	18.3	17.8	21.7	8.9	2024-04-22	압도적인 미세공정 점유율을 가진 대만의 반도체 파운드리
	AMD	AMD US	171.54	371.7	2.2	17.0	46.9	59.5	106.4	16.4	2024-05-02	HPC CPU에서 업계리딩, GPU에서 엔비디아의 적수로 꼽히는 패스트 팔로워
	인텔	INTC US	43.16	244.7	1.0	-8.1	13.2	24.5	53.3	-13.9	2024-04-26	GPU 중심 서버 설계에는 후발주자. 파운드리 기업으로서의 지정학적 가치 부상
	마이크론	MU US	81.53	120.7	-3.6	-1.0	9.2	27.0	36.5	-4.5	2024-03-28	HBM에서 SK하이닉스를 추격하는 미국의 메모리 반도체 업체
	SK하이닉스	000660 KS	150,000	109.2	8.7	11.9	15.6	31.2	62.3	6.0	2024-04-26	GPU 병목의 키를 쥐고 있는 HBM에 있어 글로벌 선두주자
	시놉시스	SNPS US	548.53	112.2	1.1	10.9	4.9	28.0	51.6	6.5	2024-02-21	반도체 EDA 업체로, 3D 엔지니어링 시뮬레이터 업체 앤시스와 M&A 논의 중
	아리스타 네트웍스	ANET US	265.51	111.3	0.2	5.3	26.3	51.8	95.2	12.7	2024-05-01	이더넷 중심 고속 네트워킹을 위한 스위치 설계
	Arm	ARM US	119.98	165.2	64.4	71.4	132.1	-	-	59.7	2024-05-07	AI 가속기 설계 위한 다수의 IP 블록 소유한 업체로 설계 라이선스/로열티 사업
	마벨 테크놀로지스	MRVL US	68.13	79.0	1.6	3.7	27.6	19.6	50.1	13.1	2024-03-07	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
	SMIC	981 HK	14.24	30.8	0.8	-18.0	-35.4	-19.3	-16.4	-28.3	2024-05-13	최근 화웨이와 밀접히 협력 중인 중국 유일의 미세공정 가능 파운드리
	Alchip	3661 TT	4,105.00	13.7	8.0	21.1	43.8	122.5	337.6	25.3	2024-03-15	대만에 있는 세계에서 AI 반도체 관련 매출 비중이 가장 높은 ASIC 업체

자료: Bloomberg, 미래에셋증권 디지털리서치팀
주: 실적발표 예정일은 수시로 변동될 수 있음

Equity Deal Update

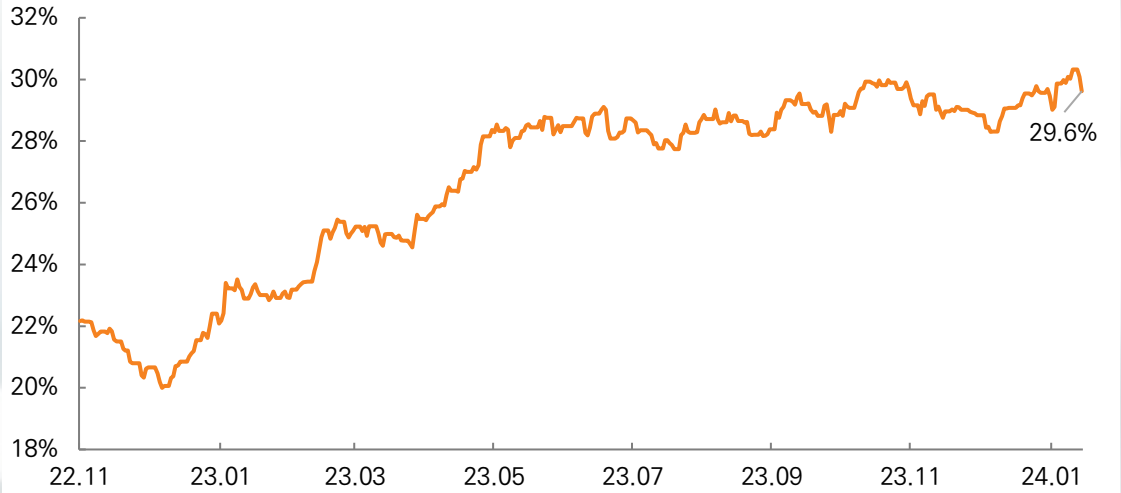
- 2월 7일~13일 기준, Top 10 비상장딜 자금조달액은 **5.5억 달러로 평균 수준 크게 상회. 헬스케어와 사이버보안에 집중.**
- **특이사항**: AI로써 헬스케어 부문을 혁신하려는 스타트업에 대한 높은 관심이 몇 주째 지속. BioAge의 플랫폼은 1만 명 이상의 환자를 포함하여 45년 이상의 인간 노화에 대한 긴 종단 연구 샘플을 포함해 **65만 개 이상의 데이터 포인트를 보유.**

회사명	소재국가	최근 투자액수 (백만 달러)	투자 라운드	펀딩 일자	신규로 진입한 주요 투자자	기업 설명
BioAge Labs	미국	170	시리즈 D	2024-02-13	Amgen Ventures, Lilly Ventures, Andreessen Horowitz	인간의 노화를 매핑하여 질병을 치료하고 건강한 수명을 연장하는 치료법을 개발하는 데 초점. 노화 질환을 새롭게 치료하거나 예방하는 다양한 신약 개발 포트폴리오 개발에 주력. 2015년 캘리포니아 리치먼드에 설립.
Bugcrowd	미국	102	시리즈 E	2024-02-12	Costanoa Ventures, General Catalyst	디지털 위협을 식별하고 완화하는 데 초점을 맞춘 크라우드소싱 사이버보안 회사. 침투 테스트, 취약점 공개, 공격 표면 관리 등의 서비스를 제공하며, AI 도구를 활용한 보안 테스트와 실행 가능한 인사이트를 제공. 2012년 캘리포니아 샌프란시스코에 설립.
Sierra	미국	85	씨드 VC - II	2024-02-13	Sequoia Capital	대화형 AI 전문업체. 공감적이고 정교한 고객 지원을 하는 비즈니스 AI 에이전트를 배치할 수 있는 플랫폼을 제공. 2023년 캘리포니아 샌프란시스코에 설립.
Unlearn	미국	50	시리즈 C	2024-02-08	Altimeter Capital, Epic Ventures, Mubadala Capital, Radical Ventures	의료 및 제약 산업을 위한 기계 학습 플랫폼을 구축하여 생물학적 디지털 트윈을 생성. 개별 건강 결과를 예측하고 임상 혁신을 가속화하는 생성 AI 방법을 개발. 2017년 캘리포니아 샌프란시스코에 설립.
Xensam	스웨덴	40	성장금융	2024-02-09	Expedition Growth Capital	기술 분야에서 소프트웨어 자산 관리에 중점. 소프트웨어 사용 최적화, 보안 결함 식별, 활용되지 않는 애플리케이션에 대한 지출 감소를 가능하게 하는 클라우드 기반, AI 구동 플랫폼을 제공. 2016년 스웨덴 스톡홀름에 설립.
Float	캐나다	37.17	신용공여	2024-02-13	Silicon Valley Bank	금융 기술 분야에서 지출 관리 소프트웨어 전문업체. 고한도 기업 카드, 비용 관리 도구, 고수익 계좌를 제공하는 플랫폼을 통해 비즈니스 지출과 재무 운영을 간소화. 2019년 온타리오 토론토에 설립.
Aizon	미국	20	시리즈 C	2024-02-07	Atlantic Bridge Capital, Crosslink Capital, Uncork Capital	생명과학 제조 분야에서 GxP-준수 SaaS 솔루션을 제공. 제조 결과 개선과 시간 및 예산 내에서 품질이 좋은 약물을 제공하기 위한 과정 최적화, 근본 원인 분석, 실시간 과정 분석, 예측 유지 관리 등의 서비스를 제공. 2013년 캘리포니아 샌프란시스코에 설립되었으며, 이전에는 Bigfinite로 알려짐.
Magenta Mobility	인도	19.06	전환사채	2024-02-13	BP Ventures, Morgan Stanley	AI를 을 활용한 전기 차량 충전 솔루션을 제공하는 회사. 직원 교통, 전자 상거래, FMCG, FMCD, 제약, 식품, 택배, 물류 회사를 위한 전기 저장 솔루션을 개발. 고객으로는 전자 상거래 사업체와 전기차 제조사가 포함. 2017년 인도 나비 뭄바이에 설립.
Super.AI	미국	12	시리즈 B	2024-02-09	Mosaic Ventures	기계 학습 프로젝트를 가속화하기 위한 AI 활성화 플랫폼 업체. 구조화되지 않은 데이터 처리, 지능형 문서 처리, 문서 처리 군중 등의 솔루션을 포함. 이전에는 Canotic으로 알려짐. 2018년 캘리포니아 샌프란시스코에 설립.
Camus Energy	미국	10	시리즈 A - II	2024-02-13	Congruent Ventures	에너지 분야에서 제로 탄소 그리드 조정을 전문으로 하는 회사. 유틸리티가 분산 에너지 자원을 관리하고 통합할 수 있는 플랫폼을 제공. 이 회사의 서비스에는 모니터링, 분석 및 예측, 지역 자원을 위한 고급 제어, 스케줄링 및 시장을 위한 인프라 등이 포함. 주로 유틸리티, 발전 사업자를 대상. 2019년 캘리포니아 샌프란시스코에 설립.

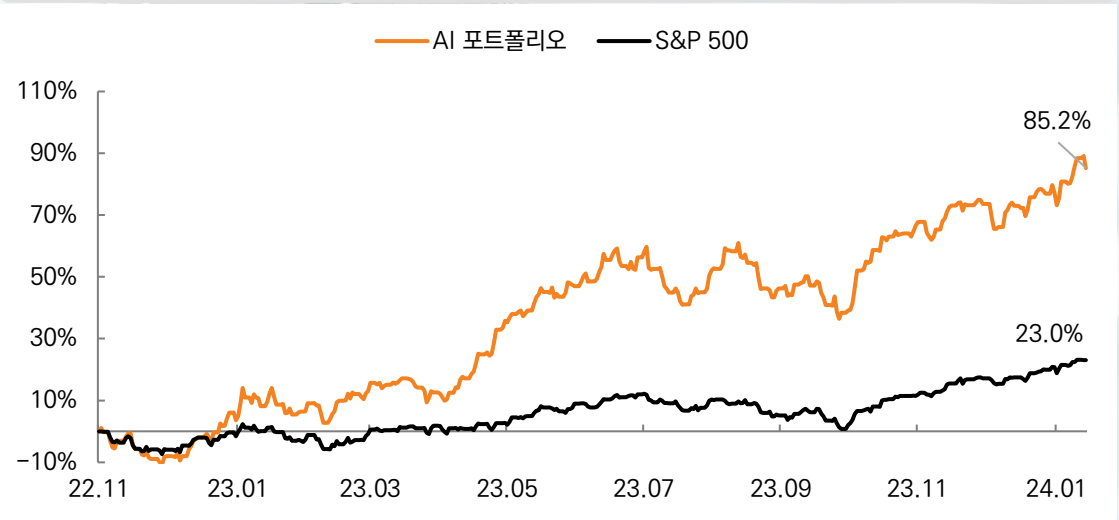
자료: CBINSIGHTS, 미래에셋증권 디지털리서치팀
* 거래 규모를 기준으로 상위 10개 항목 선정

Appendix: Charts

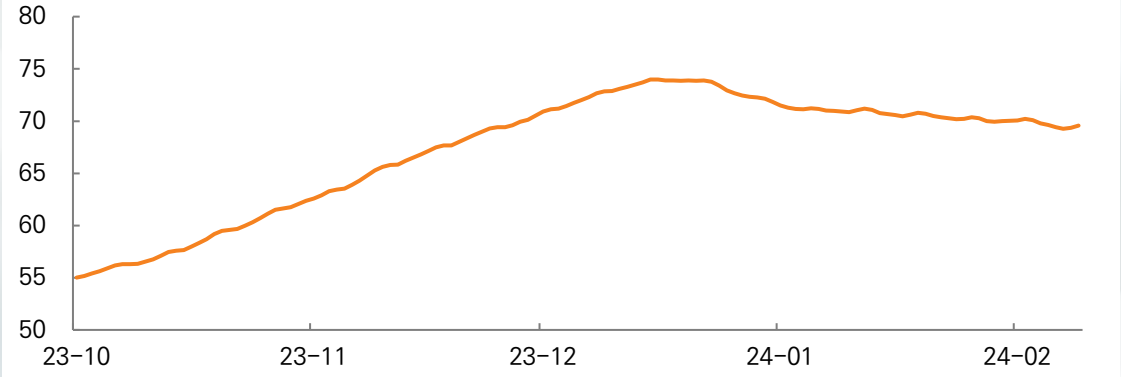
ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중 (시가총액 기준)



ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률 (동일가중 방식)

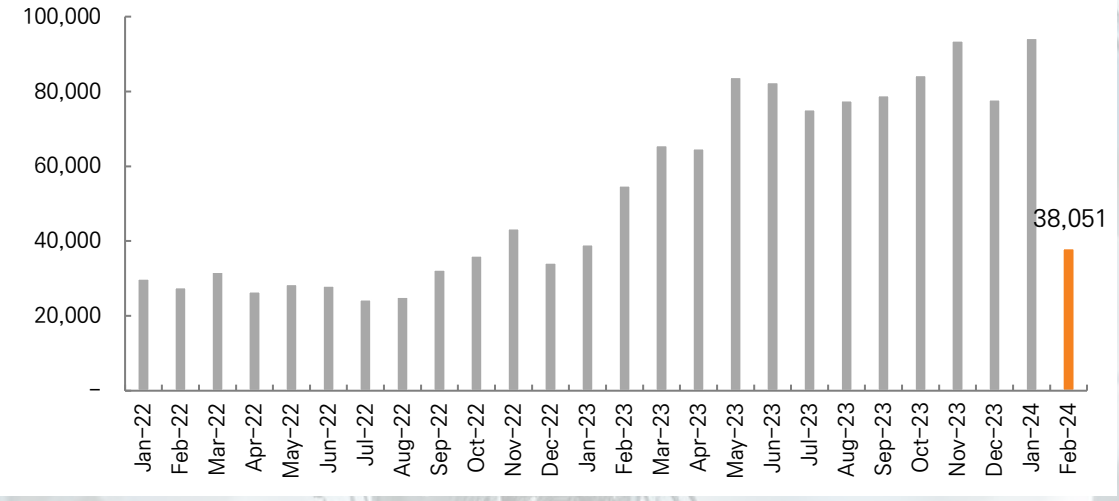


ChatGPT 키워드를 검색한 Google Trends 추이 (지난 90일 평균치 기준)



ChatGPT 관련, 지난 한 달간 “급상승 키워드” Top 3		
Copilot (마이크로소프트)	ChatGPT GPT Store	사전훈련 트랜스포머

월간 기준, 글로벌 뉴스 기사에서 AI라는 키워드가 언급된 횟수 (해당월은 여전히 진행 중)



자료: Bloomberg, CBINSIGHTS, Google Trends, 미래에셋증권 디지털리서치팀

Compliance Notice

- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.