2024. 11.06
그림: Midjourney(AI의 MAGA: Make AI Great Agents)

AI Weekly, 특이점이 온다 #38

W-159 to A.G.I.

(AGI's ETA: 2027.11.27)

디지털리서치팀 한종목, 김은지
chongmok.han@miraeasset.com
eunji.kim.a@miraeasset.com

MIRAE ASSET
미래에셋증권

Highlight of the Week

- **중국은 사실 다 갖췄다(칩 + 병렬 컴퓨팅 엔지니어링 + 알고리즘 + 전력)**

중국 상하이 AI 연구소가 o1 모델과 유사한 LLaMA-Berry 논문 발표. 이는 LLaMA-3.1-8B Instruct를 기반으로 수학적 추론 능력을 대폭 향상시킨 모델. 핵심 기술은 AlphaGo Zero 스타일의 SRMCTS(자가개선 몬테카를로 트리 서치)와 PPRM(페어와이즈 보상모델)을 결합. 이는 o1 모델의 방법론과 매우 유사. 11월 말까지 모델 훈련 완료 후 허깅페이스 공개 예정. 주목할 점은 1저자 Di Zhang이 중국과학원, 화웨이 노아방주 연구소 출신이며, 6월에도 유사 연구를 발표했다는 것. 중국이 국가 주도로 AI 개발을 진행 중임을 시사. 미국 오픈소스 모델 Llama를 활용한 점도 주목. 중국군이 Llama-2로 군사용 ChatBIT 개발한 사례도 존재. 이는 Meta의 오픈소스 정책에 대한 논란을 야기할 것으로 전망. 미 행정부의 NSM 발표와 함께 오픈소스 진영의 긴장이 고조될 전망.

- **OpenAI의 진전사항: SearchGPT 출시와 "무엇이든 물어보세요"**

OpenAI가 ChatGPT의 검색 기능인 SearchGPT를 공개. Perplexity와 유사하게 LLM 기반 웹 검색 및 출처 링크 제공. 구글 검색 시대를 종식시키려는 의도로 해석됨. 구글의 AI overview와 달리 실제 활용성이 높으나, 환각 현상은 여전히 과제. 주목할 점은 o1-preview를 활용한 SearchGPT의 개발 방식. 프론티어급 모델이 고품질 합성 데이터를 생성하고, 이를 통해 미세조정하는 구조. 이는 프론티어 모델의 중요성을 재확인. 대규모 모델 개발에는 막대한 비용 소요. 오라클 CTO와 AWS 고위 임원이 언급한 1,000억 달러 규모의 투자, 5GW 전력 소비 전망은 스케일링의 현주소.
OpenAI가 레딧 AMA를 통해 주요 개발 방향성 공개. GPT-5는 연내 출시 계획 없음을 확인했으며, o1 및 후속 모델 개발에 우선순위 부여. DALL-E 3 업데이트나 비디오 채팅 기능은 후순위로 밝힘. 핵심 전략은 'AI 에이전트' 개발에 집중. 하드웨어 제약과 모델 복잡성 증가로 인해 병렬 개발이 제한적인 상황이기 때문. o1 모델의 정식 버전은 곧 출시 예정이며, AGI는 현재 하드웨어로도 달성 가능하다는 입장 표명. GPU 헤게모니는 계속될 것. 다만, 주요 과제는 환각 현상. OpenAI 연구진도 "제로 환각" 달성의 명확한 해법을 찾지 못한 상태. 이는 AI 에이전트의 실생활 도입 시기에 영향을 미칠 수 있음. 한편 AGI 개발 속도는 가속화 중. ARC-AGI 벤치마크 점수가 5개월 만에 34%에서 55.5%로 상승. 현 추세 지속 시 내년 9월경 AGI 수준(85%) 도달 전망.

- **Paper of the week: 사람을 앞지른 손끝 – Meta FAIR**

메타, 인간의 촉각을 디지털화하기 위한 인공 손가락 시스템 Digit 360을 공개. Digit 360, 탄성 중합체로 이루어진 손끝이 물체와 접촉할 때 발생하는 변형을 내부의 빛 반사 패턴으로 감지하는 시각 기반의 촉각 센서를 중심으로 진동, 압력, 온도, 가스 감지 등 다양한 센서를 통합한 멀티모달 센싱 시스템을 구현. 특히 온디바이스 AI를 통한 1.2밀리초의 초고속 응답 속도를 보이며 인간의 척수반사 속도를 앞질렀을 뿐 아니라, 공간 해상도, 힘 감지, 진동 감지 등 대부분의 측정 항목에서 인간의 능력을 크게 상회했음. 이는 AI가 에이전트로서 실제 세계와 상호작용하는 능력을 발전시킨 것.

일	월	화	수	목	금	토
3	4 (PLTR) 실적발표	5	6 (ARM) 실적발표 (QCOM) 실적발표 FT Live: Future of AI(~7)	7 (MRNA) 실적발표 (DDOG) 실적발표 (NET) 실적발표 (U) 실적발표 (SMIC) 실적발표	8 (네이버) 실적발표	9
10	11 (위스트론) 실적발표 (콴타 컴퓨터) 실적발표(예)	12 (소프트뱅크) 실적발표	13 (루닛) 실적발표(예) Post-Industrial Summit Fall 2024(~14)	14	15 (BABA) 실적발표	16
17	18	19	20 (NVDA) 실적발표	21 (BIDU) 실적발표	22	23

주: 컨퍼런스 일정은 다음과 같은 기준으로 선별함: 1) 주요 AI 기업이 연사로 참여하거나 스폰서인 경우, 2) 주요 대학 및 기관이 연사로 참여하는 경우 등.
자료: Bloomberg, 외신, 미래에셋증권 디지털리서치팀

AI Issue

중국은 사실 다 갖췄다(칩 + 병렬 컴퓨팅 엔지니어링 + 알고리즘 + 전력)

지난 10월 3일, **o1 모델을 Llama 모델로 복제한 듯한 논문이 중국의 AI 연구소로부터 나왔다. 상하이 AI 연구소에서 작성한 "LLaMA-Berry"**(Pairwise Optimization for o1-like Olympiad-Level Mathematical Reasoning)**라는 논문이다**. 이름에서 보듯, LLaMA-3.1-8B Instruct 모델의 성능을 대폭 개선하기 위한 방법론으로, 특히 수학 문제 해결에 강점을 보인다. (수학 벤치마크인 MATH: 모델의 성능이 47.2%에서 75.3%로 향상. 훨씬 어려운 수학 벤치마크인 AIME2024: 성능이 6.7%에서 26.7%로 크게 향상) **reasoning 능력 향상에 중점**을 뒀다는 말이다. 이 논문의 핵심은 Meta의 LLaMA 모델을 기반으로 사용하고, 여기에 AlphaGo Zero와 유사한 구조를 도입했다는 점이다. **AlphaGo Zero의 가장 큰 특징은 Tree Search를 활용한다는 것이고, 이는 OpenAI의 o1 모델에 사용된 것으로 파악**되었다. **Llama-Berry의 핵심 요소는 SRMCTS(자가개선 몬테 카를로 트리 서치)와 PPRM(모델이 생성한 두 답변을 서로 비교하여 어느 것이 더 나은지 판단하는 보상모델)이다.** 먼저 몬테카를로 트리 탐색은 불확실성이 있는 의사 결정 문제에서 사용할 수 있는 탐색 알고리즘이다. 무작위로 선택된 경로를 따라가 보고, 좋은 결과를 낸 경로를 더 탐색하면서도 새로운 경로도 탐험하는 방식이다. 거기에다가 Self-Refined는 모델이 자신의 출력을 바탕으로 스스로 답을 개선하는 과정을 반복하는 것이니, 인간의 주석이나 피드백 없이도 모델이 스스로 학습하고 개선한다는 것을 의미한다. 즉, **종합하자면 MCTS에 자기 개선을 결합한 것으로, 모델이 다양한 추론 경로를 효율적으로 탐색하고 스스로 답을 개선하도록 하는 것이다. 이 방식은 우리 팀이 지난 9월에 발간했던 o1 모델에 관한 인덱스 페이퍼에서 분석했던 방법론과 정확히 일치한다**.

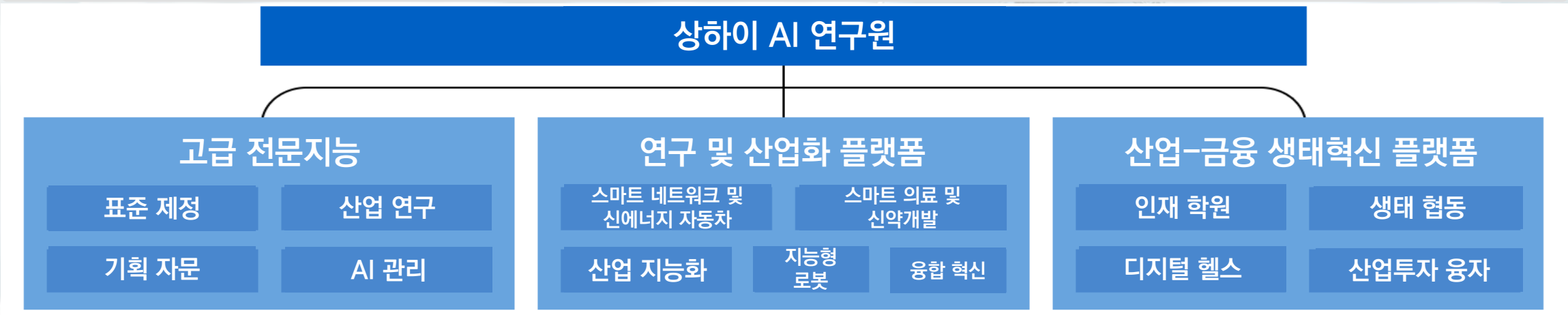
해당 논문의 **주 저자인 상하이 AI 연구원은 11월 말까지 모델의 첫 번째 훈련과 평가를 완료할 계획이며, 이후 모든 데이터를 허깅페이스에서 공개할 예정**이라고 밝혔다. 이는 함의하는 바가 매우 크다. 우선 이 논문의 1저자인 Di Zhang은 중국 국무원 중국과학원과 화웨이 노아방주 연구소를 경험한 사람이다. 즉, **중국은 미국과 달리, 민영 빅테크들이 아닌 국영 기관에서 패러다임을 바꿀만한 모델 연구를 선도하고 있다고 봐야 하고, 국가 경쟁력의 사명을 가지고 모델을 개발하고 있다**는 말이다. 또 하나의 통찰은, Di Zhang의 연구팀이 이미 지난 6월에 MCTS 및 자가 개선의 방식을 전면에 내세운 논문(제목: Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B)을 내놓은 적이 있다는 점이다. 즉, **이번에 나온 연구는 후속 논문이고, 중국 입장에서 o1 모델이 제시한 새로운 패러다임 전환을 어느정도는 예상했다고 볼 수 있는 대목**이다. OpenAI를 떠나 엄청난 폭로를 했던 레오폴드 아센브레너가 말했듯이 **미국의 AI 스타트업들이 가진 비밀의 레시피는 유출에 너무 취약했다**는 점도 어느 정도 방증되었다고 판단한다. 또한 무엇보다 중국은 AI 국가 경쟁력에 있어서 전혀 무시할 만한 단계가 아니라는 점을 유념해야 한다. 더군다나 중국은 미국과는 달리 엄청난 규모의 대규모 전력 인프라 및 그 확장성 또한 무시무시하다. 레오폴드 아센브레너, 딜런 파텔 등 AI 업계의 인플루언서들이 입을 모아 말한 내용이다. 그리고 그 중심에 있는 업체인 화웨이의 동향을 항시 주시해야 할 것이다.

또 하나 우리가 알아야 할 것은, 이들이 Llama라는 미국의 오픈소스 모델을 활용했다는 점이다. **최근에는 중국 인민해방군 산하 군사과학원이 Llama-2-13B 모델을 개조해서 군사 응용을 위한 AI 도구인 "ChatBIT"을 개발했다는 소식**도 있었다. ChatBIT은 향후에 군대 전략 계획, 시뮬레이션 훈련 및 지휘 결정으로 확장될 수 있다. 군사 목적으로 오픈소스를 활용하려는 중국의 체계적인 야심을 읽을 수 있다. 이에 따라서, **Meta를 포함한 미국의 오픈소스 정책에 관한 뭇매가 발생할 수 있을 것**이다. Meta는 명시적으로 군사 목적으로 모델을 사용하는 것을 금지하고 있지만, 현실적으로 이러한 제한이 얼마나 어려운지를 보여주기 때문이다. 심지어, 지난 10월 24일, 바이든-해리스 행정부가 국가안보를 위한 AI 활용에 대한 접근방식인 "국가 안보 메모랜덤(NSM)"을 발표했었다는 점을 감안해야 한다. 또한 대표적인 폐쇄소스 업체들인 OpenAI와 앤트로픽의 CEO를 백악관으로 불러 모은 것까지 감안하면 **오픈소스 진영의 주장은 그 힘을 다소 잃을 수 있을 것이라 생각한다**. 한편, **마크 주커버그는 Meta의 3분기 실적 발표에서 10만개 '이상의' H100 GPU를 갖춘 클러스터에서 Llama-4를 훈련하고 있다면서 내년 초에 기대할 수 있다고 발언**한 바 있다. 내년에는 오픈소스 진영을 포함해 **G2간의 AI 모델 개발 암투가 더 활발해질 것**으로 본다.

자료: 미래에셋증권 디지털리서치팀

관련 자료: o1을 모방한 Llama-Berry

상하이 AI 연구원(상해 교통대학교, 상하이 민항구 지방정부, 국유기업인 린강그룹, 센스타임이 공동투자)의 조직도, '중국이 중점적으로 보는 AI 혁신 분야는 자율주행/헬스케어/로보틱스'



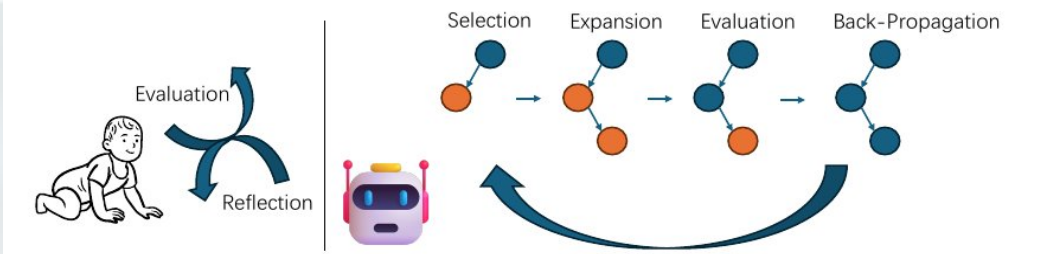
자료: 상하이 AI 연구원, 미래에셋증권 디지털리서치팀

중국의 연구소가 o1을 모방해 만든 모델 Llama-Berry의 수학 및 생물/물리/화학 대학원 문제풀이 성능

모델	MATH	AIME2024	GPQA Diamond
Claude 3 Opus	60.1	6.7	50.4
GPT 4o	60.3	13.4	56.1
OpenAI o1 Preview	85.5	56.7	78.3
OpenAI o1	94.8	83.3	78
Gemini 1.5 Pro	67.7	6.7	46.2
Gemini Math-Specialized 1.5 Pro	80.6	23.3	-
Meta-LLaMA-3.1-8B-Instruct	47.2	6.7	30.4
Qwen2-Math-7B-Instruct	75.1	13.3	-
Qwen2-Math-72B-Instruct	84	20	-
LLaMA-3.1-8B-Instruct + 자가정제 몬테카를로 트리 서치 = **Llama-Berry**	54.8@major8 69.4@rm8 75.3@rm16	13.3@major8 16.7@rm8 26.7@rm16	39.4@major8 77.3@rm8 92.4@rm16

자료: 상하이 AI 연구소 논문 "Llama-Berry", 미래에셋증권 디지털리서치팀
주1: major8: "majority voting with 8 samples"의 약자. 8번의 샘플링을 수행하고 그 중에서 가장 많이 나온 결과를 선택.
주2: @rm8: "running mean with 8 samples"의 약자. 8번의 반복 시행에서 얻은 결과들의 평균값을 선택.

AI 에이전트의 학습 과정, 'MCTS와 LLM의 결합이 인간 추론 능력에 더 가까워지는 길'



인간의 학습 과정은 시행착오를 통해 평가하고 반성하는 과정이라 할 수 있음.

몬테카를로 트리 검색(MCTS)은 "Selection(선택), Expansion(확장), Evaluation(평가), Back-Propagation(역전파)"로 이뤄짐.

시행착오를 통한 학습은 인간의 학습 과정과 AI의 학습 과정 사이의 유사성을 강조.
복잡한 수학 해결에서 자가 개선 MCTS의 성공은 인간 유사 학습 방식의 효과성을 입증.

자료: 상하이 AI 연구소 논문 "Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B", 미래에셋증권 디지털리서치팀

AI Issue

SearchGPT의 출현은 구글의 지위를 위협

드디어, **ChatGPT의 검색 기능이 ChatGPT 사용자들에게 공개**되었다. 유료 사용자들만 이용할 수 있지만 곧 모든 사용자에게 제공될 예정이다. 이 기능은 Perplexity와 크게 다르지 않다. 즉, **ChatGPT같은 LLM을 사용해 웹을 검색하고 출처 링크를 얻을 수 있는 형태다.** 샘 알트만은 검색이 ChatGPT 출시 이후 자신이 가장 좋아하는 기능이라고 말했다. 검색 기능이 이끌 수 있는 변화가 생각보다 크기 때문이다. 이어서, 샘 알트만은 단순히 정보를 보여주는 것을 넘어 "사용자의 쿼리에 맞춰 실시간으로 맞춤형 웹페이지를 생성"하는 비전을 제시했다. 다시 말해, 이러한 **미래에는, 사용자의 의도를 정확히 이해해서 필요한 여러 소스의 정보를 하나의 통합된 페이지로 맞춤형 재구성 가능하기 때문에, 단순한 UI/UX 개선을 넘어 정보 소비의 패러다임을 바꾸는 혁신이 될 수 있다**고 본다. 우리 팀이 보기에 해당 주장은, **구글이 만든 검색 시대를 저물게 하겠다는 확실한 청사진**이다. 우리 팀이 실제로 SearchGPT를 써본 결과, 구글 검색처럼 복잡한 레이아웃도 아니고 당연히 광고 링크도 없어 매우 깔끔한 느낌을 받았다. 물론 OpenAI는 구독료를 통해 수조원의 자금을 벌어들이고 있기 때문에 우선은 광고 수익이 필요하지 않기에 가능한 일이다. 참고로 SearchGPT는 구글의 "크롬" 웹 브라우저의 플러그인으로도 사용이 가능하다. 그런데 어째서인지 구글 검색에서는 해당 설치 링크를 찾기 힘들다. 구글도 이를 의식하고 있는 것이다. (그래서 샘 알트만이 트윗으로 직접 다운로드 링크 주소를 공유하기도 했다)

물론 구글도 몇 달 전부터 "AI overview"라는 기능으로 본인들의 검색 엔진에다가 생성 AI 기능을 붙였다. 그러나 관련 수치 정보를 따로 공개하지는 않고, 전반적인 검색 사용량 및 사용자 만족도 개선을 "확인"했다고만 매분기 실적발표에서 주장하고 있다. 우리 팀이 보기에 실제 활용성은 아직 형편없다. AI가 잘못된 정보를 생성해주기 때문이고, 심지어 그것을 검색 결과의 최상단에 노출시키기 때문에 사용자 경험을 기존보다 더 떨어트리는 문제가 있기 때문이다. 이와 관련해, **SearchGPT도 문제가 없지는 않다. 환각이 숨어 있다는 점을 반드시 유의해야 한다. 출처를 달고 관련 사실을 찾아 주기는 하지만, 그 과정에서도 그럴 듯한 단어 생성을 하기 때문에 간혹 거짓말을 하기도 한다.** 앞으로 이러한 점이 풀어야 할 숙제가 아닌가 싶다. 그럼에도 실제의 활용성은 충분하기 때문에 구글이 누린 어마어마한 TAM(총 유효시장)을 어느 정도 잠식할 수 있다. 다시 말해, **Perplexity에 더해 이제는 SearchGPT까지 등장하면서 구글의 검색 시장 지배력이 흔들리고 있다는 점은 명백하다. 그리고 하나 짚고 싶은 것은, Perplexity보다 SearchGPT의 출현은 구글에게는 더욱 압박이 될 수 있다**. 신기술 도입에 있어서 역사적인 큰 교훈 중 하나는, 성능 그 자체도 중요하지만 대규모 배포능력이자 시장 침투력이 더 중요할 수 있다는 것이다. "챗봇"이라는 일반명사보다 "ChatGPT"라는 고유명사가 더 많이 쓰이고 있는 현재 상황을 감안하면, ChatGPT 사용자들이 이제 SearchGPT로 크로스플레이 하는 것에 구글은 긴장하지 않을 수 없을 것이다. 게다가, 단순히 링크 페이지를 제공하는 구글 방식이 아니라, OpenAI는 FT, 로이터 등 많은 업체들과 (데이터 사용에 관한) 거래를 해놓았기 때문에 그런 콘텐츠에 원활하게 접근할 수 있는 장점까지 있다고 본다.

한편, **OpenAI는 ChatGPT를 Search 기능으로 더 적합하게 만들기 위해 o1-preview의 힘을 빌렸다. o1-preview가 내놓는 출력물이 고품질의 (합성) 데이터가 되는 것이다.** 그리고 이를 통해 GPT-4를 미세조정해서 SearchGPT로 탄생시킨 점이 포인트다. 즉, 경량화되어 각종 영역에 쓰일 수 있는 여러 버전의 언어모델의 산파 역할을 하는 것은 역시 프론티어급 언어모델이라는 점을 유념해야 한다. **큰 모델을 잘 만들어야 작은 모델도 잘 만드는 것이다**. 이러한 **프론티어 모델 개발에는 당연히 엄청난 돈이 수반된다. 오라클의 창업자이자 CTO인 래리 엘리슨은 이 비용이 1,000억 달러에 육박한다고 주장한 바 있다**. 그리고 최근 아마존 AWS의 매트 가먼 또한, "최근 생성 AI 모델들이 최대 5GW 전력을 요구한다"고 말했다. 5GW는 300만 개의 블랙웰 GPU와 동일하고, 이 비용이 약 1,000억 달러로 계산된다는 점은 의미심장한 일치다. **업체들이 스케일링할 유인은 아직도 너무 많다**.

자료: 미래에셋증권 디지털리서치팀

AI Issue

OpenAI가 진행한 “무엇이든 물어보세요”

사실, **SearchGPT보다 더 뜨거웠던 이슈는 OpenAI가 레딧에서 AMA라는 이름의 Q&A를 진행하면서 알려진 여러 놀라운 정보들이었다**. 참고로 AMA는 “Ask Me Anything”의 줄임말로 한국어로 의역하면 “무물(무엇이든 물어보세요)”이란 뜻이다. 먼저 사람들이 가장 궁금한 질문은 이것이었다. “GPT-5의 출시일이 언제인가”라는 질문에 샘 알트만은 "올해 말까지 GPT-5라고 부를 만한 것은 없을 것”이라고 답했다. 실망스럽기도 하지만 o1-preview가 이미 출시(o1과 GPT-5는 다르다!)되었기에 충분히 예상 가능한 답변이었다. 그는 **"o1와 그 후속 모델들의 출시를 우선시하고 있다."라면서 오히려 추론 컴퓨팅 기반 모델 개발에 더 정진할 것임을 밝혔다**. "언제 새로운 텍스트-이미지 모델이 제공되나? DALL-E 3는 좀 구식이다"라는 질문에 대해서도 "아직 출시 계획은 없다”고 했다. 그리고 OpenAI가 몇 달 전에 시연했던 ChatGPT와의 ‘비디오 채팅(영화 her에서 보는 것과 같은 비디오+음성 에이전트)’에 대한 짧은 기술적 업데이트도 했는데, “작업중이고 정확한 날짜는 아직 없다”고 했다. 두 가지에 대해서 확실히 OpenAI는 우선순위를 낮게 배정한 것 같다.

반면, **그들은 “AI 에이전트”에 극도로 집중하는 것처럼 보인다. “GPT 제품 라인의 다음으로 올 획기적인 발전은 무엇인가?"라는 질문에 샘 알트만은 “에이전트가 될 것"이라고 분명히 짚었기 때문이다**. 에이전트라는 것에 전략적 선택과 집중을 하는 이유에 대해서는, "모든 모델들이 꽤 복잡해져서 원하는 만큼 많은 것들을 병렬로 출시할 수 없다”는 말로 그 의중을 읽을 수 있다. **모델이 점차 커지고 복잡해지는 것과 동시에, 엔비디아 GPU 칩 수급의 한계 때문에 여러 모델을 동시에 개발/출시하기 어렵다**는 말이다. 이를 바꿔 말하면, 이런 **제약 조건 속에서 OpenAI가 가장 중요하게 생각하는 모델이 바로 o1 모델이라고 추측해볼 수 있다**. 왜냐면 이 모델의 성능이 곧 대중이 AI 에이전트를 얼마나 수용하여 사용할지 그 침투율을 결정할 것이기 때문이다. 이어서 “o1-preview가 아니라 완전한 o1 출시는 언제 받을 수 있나?라는 질문에는 최고 제품 책임자인 Kevin Wild가 “soon”이라고 답했고 알트만도 오래 걸리지 않을 것이라고 말한 점은 반가운 부분이다. 또한, 가장 흥미로운 답변은 바로 이것이었다. “**AGI는 현재의 하드웨어로 달성 가능한가? 아니면 완전히 다른 것이 필요한가?”라는 질문에 대해 샘 알트만은 “우리는 현재의 하드웨어로 달성 가능하다고 믿는다”라고 했다. 이는 엄청난 의미를 지닌다. 자체 제작 칩 개발에 나선 OpenAI조차 GPU가 AGI로 가는 열쇠라고 못을 박은 셈이기 때문이다**. 추론 영역에서 토큰 생성 속도가 빠르다며 각종 NPU 업체들이 자랑하듯 떠들어도, 선도적인 업체들이 가지는 GPU에 대한 갈증은 계속되고 있다는 것이다. 그리고 추론 컴퓨팅이라는 또 다른 스케일링 법칙의 등장은 그 욕구를 한층 더 돋구었다고도 볼 수 있다.

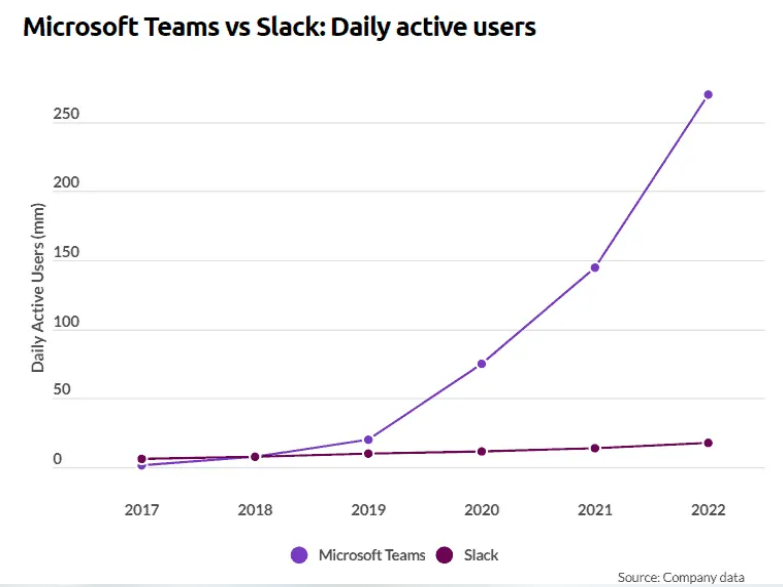
다만 부정적인 부분도 있었다. “환각 현상이 영구적인 특징이 될 것인가? o1-preview는 추론의 끝 부분에서 왜 점점 더 많은 환각을 보이나?"라는 질문에 대해서 Mark Chen(연구 수석 부사장)은 근본적으로 어려운 문제라고 답했다. 즉, **현재로서는 "제로 환각”을 달성할 명확한 경로를 보지 못하고 있는 것으로 해석된다. 에이전트로 이뤄질 新경제의 완전한 변혁을 이루려면 무엇보다 신뢰성을 확보해야 하는데, 이 해답이 풀리지 않는다면 AI 에이전트들이 실생활에 투입되는 대중화 시기는 생각보다 더 늦춰질 수 있다. 이것은 아주 큰 리스크다**. 다만, OpenAI에서는 2-3개월마다 새로운 제품 및 업데이트들을 내놓고 있다. AI 개발 속도가 오히려 더 빨라지고 있다는 점을 들면 나름 미래를 낙관해보게 된다.

샘 알트만은 내년에 모든 벤치마크를 완전히 압도하고 싶다고 말했다. 또한 그는 (o1가 아니라) o2는 GPQA-다이아몬드 테스트에서 105%를 얻었다고 트윗하기도 했다. 이것이 “트롤링”일 지도 모르겠지만 **o2에 대해 꽤 자주 언급한다는 것은, 적어도 o1은 거의 완성단계라는 것을 의미할 것이고 내년에는 o2가 나오리라 본다**. 한편, 이와 관련해 구글 AI의 Francis Cholle가 만든 AGI 테스트인 **ARC-AGI의 벤치마크**(800개의 논리 퍼즐 모음)에 따르면, AI의 달성 점수가 매달 향상되고 있다. **85% 이상을 얻으면 AGI 수준의 시스템이라고 말할 수 있는데, 현재는 55.5%**다. 아직 낮다고 낙담할 수는 있지만 불과 5개월 전에는 34%에 불과했다는 점을 알아야 한다. 추세를 따라간다면 **내년 9월에는 AGI를 달성할 것이다.**

자료: 미래에셋증권 디지털리서치팀

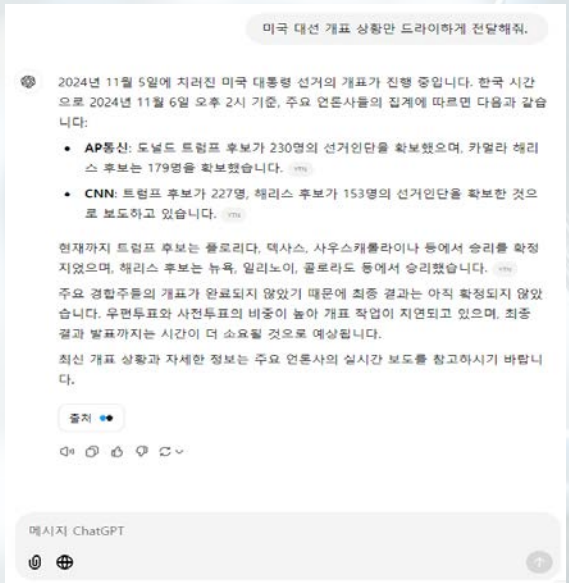
관련 자료: OpenAI의 신기능 Search와 “A.M.A.”

팀즈와 슬랙의 일간활성화 사용자 수 비교, ‘중요한 건 배포능력’



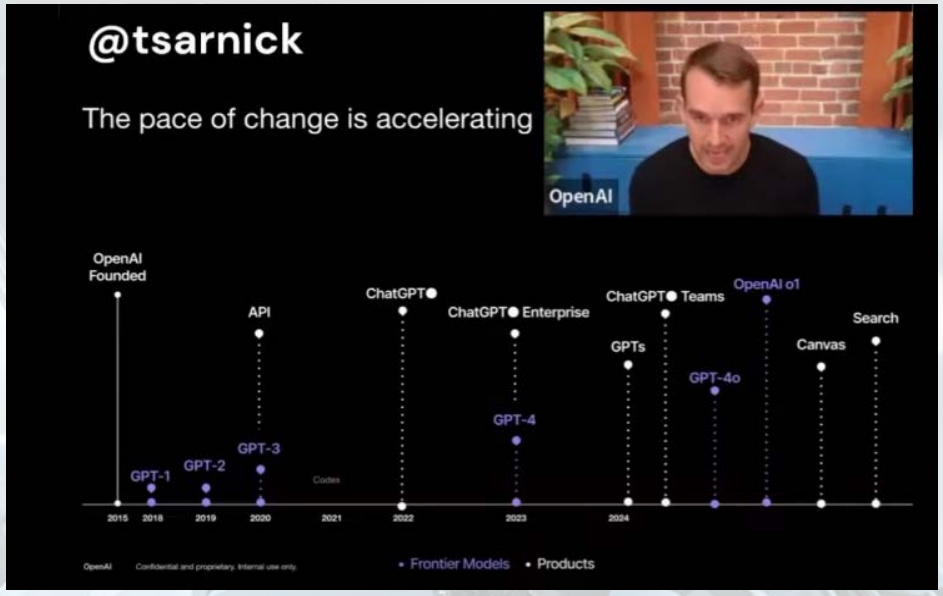
자료: X(@kwharrison13), 미래에셋증권 디지털리서치팀

ChatGPT에서 Search 기능을 시연한 모습



자료: OpenAI, 미래에셋증권 디지털리서치팀

며칠 전 OpenAI의 비공개 웨비나 자료 발췌, ‘OpenAI의 속도는 가속 중…’



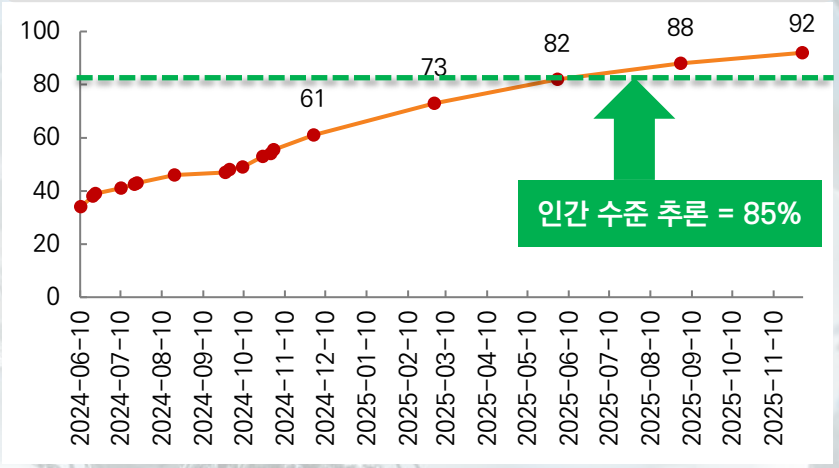
자료: OpenAI, X(@tsarnick)

OpenAI가 진행한 “무엇이든 물어보세요”에서 공개된 여러 정보들 중 일부를 정리

질문	답변
GPT-5의 출시일이 언제인가?	올해 말에 매우 좋은 릴리스를 내놓을 예정. GPT-5라고 부를 만한 것은 없을 것이다. o1과 후속 모델들의 출시를 우선시하고 있음. 완전한 o1은 “곧” 출시될 것이다. 모델들이 꽤 복잡해져서 원하는 만큼 많은 것들을 병렬로 출시할 수 없다. (o1 모델은 거의 다 완성이 되었고 출시 시점만 조율중. 그런데 GPU 수급이 문제)
GPT 제품 라인의 다음 획기적인 발전은?	다음 거대한 획기적인 발전처럼 느껴질 것은 에이전트가 될 것이다. (OpenAI가 선택과 집중으로 선택한 것은 AI 에이전트)
AGI는 달성 가능한가?	우리는 현재의 하드웨어로 달성 가능하다고 믿는다. (엔비디아의 헤게모니는 계속될 것으로 판단) 환각 문제는 단순한 기술적 버그가 아닌 근본적인 도전 과제다. (현재로서는 '제로 환각' 달성 경로를 찾지 못했다는 말)
2025년에 대한 대담한 예측?	모든 벤치마크를 포화시키고 싶다. (내년이면 AGI를 테스트하는 현재의 벤치마크 점수를 100% 뛰어넘을 거라는 말)

자료: OpenAI, 미래에셋증권 디지털리서치팀

추론력 테스트하는 ARC-AGI 벤치마크에서 AI들이 달성한 점수 추이



자료: ARC-AGI, 미래에셋증권 디지털리서치팀
주: 2024년 12월의 수치부터는 예상치

AI Insight

새삼 엔비디아가 대단하다고 느껴지네

지난 10월 24일, 인도 뭄바이에 있었던 AI Summit 행사 중 젠슨 황 CEO의 키노트 연설

1964년 IBM의 시스템/360은 범용 컴퓨팅(HW와 SW를 분리하는 아키텍처 표준화)의 표준이 됐고, 지금까지 60년 동안 존재해왔다. 그 덕에 HW가 업그레이드 되더라도 SW가 호환되는 방식으로 계속 개선될 수 있었고, 거기다 무어의 법칙의 현상 때문에 성능은 매년 2배씩 개선되었다. 하지만 **이제 CPU의 스케일링이 한계에 도달했고, 무어의 법칙이 끝났다. 다른 일을 하지 않으면 컴퓨팅 "인플레이션"을 경험하게 된다**. 그런 측면에서 엔비디아는 범용 컴퓨팅을 강화해 SW를 가속화하기 위해 시작됐다. **응용 프로그램을 엄청나게 가속화하는 것을 가능하게 한 것이 엔비디아의 프로그래밍 모델인 CUDA다**. 실제로 GPU는 CUDA를 실행하고 컴퓨터 그래픽을 실행하는 최초의 가속 컴퓨팅 아키텍처였다. 그리고 지난 30년 동안 다른 영역으로 가속화를 확장하는 여정을 계속했다. 하지만, **세상의 모든 것을 가속화할 수 있는 마법 같은 프로세서는 존재하지 않는다. 알고리즘, 하부 아키텍처, 상부 어플리케이션까지 연결되는 모든 스택을 재발명해야 한다**. 산업계, 개발자들, 고객들과 협력해야 하고, 오랜 시간이 걸린다. (우리 팀의 생각: 이게 해자다) 엔비디아는 cuDNN 등을 포함해 각 산업을 위한 가속 라이브러리를 출시했고, 최대 50배의 속도로 가속화했다. (엔비디아가 강조하는 CUDA 라이브러리 영역에는 반도체 노광, 5G, 양자컴퓨팅, 유전자 시퀀싱, SQL, 물리학 등 매우 다양) **CUDA는 티핑 포인트에 도달했다**. IT 산업은 SW가 만들어지는 방식에 좌우된다. 이제껏 우리는 SW 1.0이라고 불리는 방식이었다. 프로그래머들이 알고리즘을 코딩하여 컴퓨터에서 실행되는 함수를 만들고, 그것을 입력 정보에 적용하여 출력을 예측하는 것이다. 하지만 이제 접근 방식이 바뀌었다. 코딩이 아니라 머신러닝이다. 대규모로 관찰된 데이터의 패턴을 연구하고, 그로부터 학습한 다음, 이를 예측하는 근사 함수를 만드는 것이다. 이게 **SW 2.0이고, 이제 SW은 (사람이 아닌) 컴퓨터가 작성한다. SW 1.0이 CPU에서 최적으로 실행되었다면, SW 2.0 기반 신경망은 GPU에서 가장 잘 실행된다.** (우리 팀의 생각: 젠슨 황이 TAM을 이야기 할 때 가속 컴퓨팅으로의 전환이라는 거대 담론으로 매번 답변하는 이유)

블랙웰 시스템을 구축하기 위해서는 물론 GPU가 필요하지만, 다른 칩(CPU, DPU, NIC, 각종 스위치)도 필요하다. **TSMC가 이 모든 칩을 제조하고 있고, 블랙웰은 4분기에 대량 생산으로 납품할 것으로 예상한다**. (우리 팀의 생각: TSMC와 엔비디아의 불화설 일축) 엔비디아가 하려는 것은 더 크고 더 똑똑한 모델을 학습하는 것이다. 이것을 스케일링 법칙(LLM을 훈련하는데 사용되는 데이터가 많을 수록, 그에 상응하는 모델 크기가 커져야 하고 그 반대도 마찬가지라는 귀납적 법칙)이라고 부른다. **현재 업계는 데이터의 양과 모델 크기를 각각 약 2배씩 매년 증가시키고 있다. 이 둘의 곱인 "컴퓨팅이 매년 4배씩 증가해야 한다"는 것을 의미한다**. (우리 팀의 생각: 레오폴드 아센브레너는 2년에 10배씩 컴퓨팅이 증가한다고 주장. 그 대담한 예측과 젠슨의 생각이 거의 똑같다는 것을 유념) 과거 무어의 법칙 때는 1년 반마다 2배씩 기술이 발전했지만, 스케일링 법칙 시대에는 매년 4배의 기술 발전이 있다. **스케일링 법칙에 있어 둘째로 발견된 것이 있다. (OpenAI의) "Strawberry"로 개시된 미래에는 기계의 지능이 단 한 번의 시도가 아니라, thinking를 필요로 한다는 것을 깨달은 것이다**. Thinking이란 reasoning이며, 마음 속으로 경로 계획 및 시뮬레이션을 하거나, 자신의 답변에 대해 성찰 중이라는 개념이다. 결과적으로 더 많이 계산할수록 더 높은 품질의 답변이 나온다. **이게 2nd 스케일링 법칙이고, 또한 "추론 시간"의 스케일링 법칙**(Test-Time Compute와 동의어. 자세한 내용은 OpenAI o1에 관한 우리 팀의 인뎁스 보고서를 참고. 추론 시간 스케일링에서는 GPU가 더 많이 쓰일 것)이라고 부른다.

종합하자면, **점점 더 많은 기업들이 자신들만의 파운데이션 모델 기술을 구축해야 한다는 것을 인식하고 있는 상황에서, 두 가지의 스케일링 법칙 때문에 필요한 컴퓨팅이 엄청나게 증가했다**는 점이다. 그렇기 때문에, 어떤 산업, 국가이든 먼저 AI 인프라부터 구축해야 한다. 예를 들어 올해 말까지 가동될 **인도의 컴퓨팅 파워는 1년 전보다 약 20배나 늘어나게 된다.** AI 인프라를 구축한 인도는 더 이상 SW 백오피스 운영의 국가가 아니라, AI(엔비디아의 플랫폼을 활용해 만들어진 각종 기업용 AI 에이전트 등)를 생산하고 여러 국가들에 배포하는 일종의 수출국가 역할을 할 것이다. (우리 팀의 생각: 그 과정에서 핵심 파트너가 엔비디아라는 것. 엔비디아는 삼국지의 '군사(軍師)'와 같은 역할을 맡아 한 나라를 글로벌 AI 허브로 성장하게 해줌)

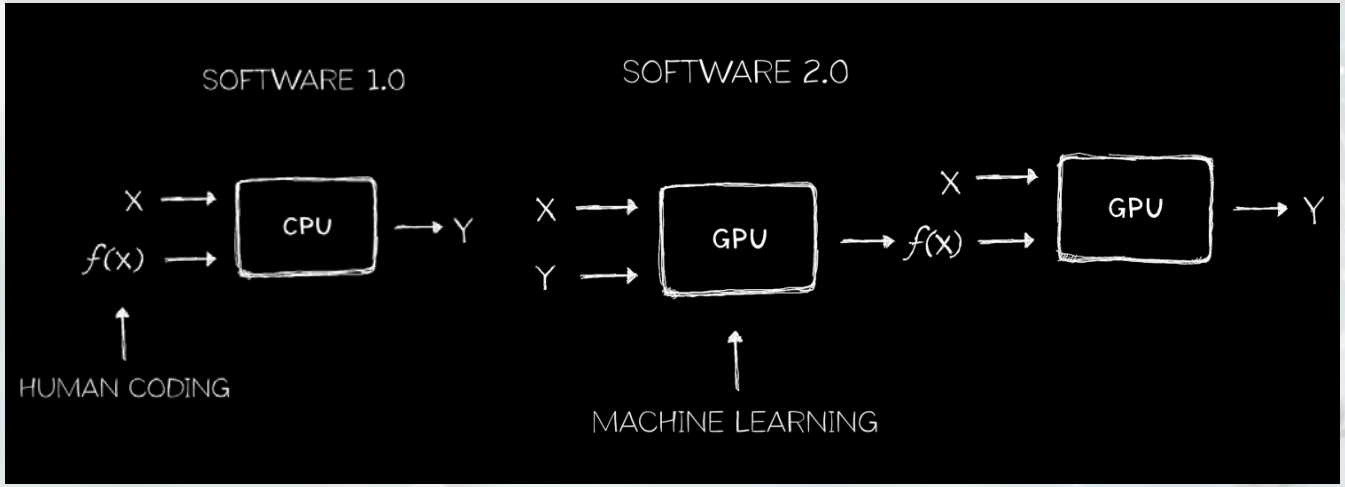
자료: 미래에셋증권 디지털리서치팀

관련 자료: 엔비디아와 그 기술 스택들

엔비디아의 CUDA 가속화 라이브러리 예시, 'HW인 GPU를 기반으로 SW은 해당 라이브러리를 통해 각 산업 영역의 응용 프로그램들의 발전이 가속화될 수 있음'

라이브러리	영역	주요 기능	파트너사 및 활용 사례
cuLitho	반도체 제조	노광 장비 지원 엔지니어링. 전산 리소그래피 시뮬레이션 기능	• ASML: 컴퓨테이셔널 리소그래피 소프트웨어 제품에 GPU 지원 통합 • TSMC: 리소그래피 솔루션 및 딥러닝 기술 도입 • Synopsys: OPC 소프트웨어를 cuLitho 플랫폼에서 실행
AI Aerial	5G	5G vRAN(가상화된 무선 접속망) 개발과 배포를 위한 SW/HW 통합 솔루션	• NTT 도코모: 세계 최초 GPU 기반 상용 5G Open RAN 출시. 기존 네트워크 대비 TCO 30% 절감
cuQuantum	양자 컴퓨팅	고전컴퓨터와 양자컴퓨터의 하이브리드 컴퓨팅	• Zapata Computing: Orquestra 플랫폼과 cuQuantum SDK 통합 • QMware: Oracle과 협력하여 하이브리드 양자 컴퓨팅 개발
Parabricks	헬스케어	유전자 게놈 시퀀싱 소프트웨어 스택	• Broad Institute: Terra 클라우드 플랫폼에서 게놈 분석 가속화. GATK(게놈 분석 툴킷) 팀과 협력
cuVS	클라우드	AI 데이터베이스 구축 및 모든 데이터 벡터화 지원	• 오라클: 벡터 서치 인덱스 생성 가속화. 데이터베이스 벡터 검색 최적화 협력
cuDF	클라우드	구조화된 데이터 처리 및 SQL 가속화	• Google: Colab에서 기본적으로 RAPIDS cuDF 제공
cuOPT	알고리즘	조합 최적화 문제 해결 특화 라이브러리, 복잡한 의사결정 최적화 문제의 병렬 처리	• Azure Maps: 복잡한 다중 경로 최적화 애플리케이션 구축 • OCI(오라클): 직접 설치하고 설정하는 번거로움 없이 바로 cuOpt의 기능을 활용 가능
Modulus	물리학	AI에 물리 법칙을 학습시켜 유체역학과 입자물리학의 다음 순간 예측	• Siemens Energy: HRSG 디지털 트윈 시뮬레이션 • Siemens Gamesa: 전산유체역학 가속화 • University of Illinois: 연구 및 교육 활동에 도입
cuDNN	딥러닝	가장 유명한 응용 프로그램 라이브러리, AI 민주화 실현	너무 많음

지금은 "소프트웨어 2.0" 시대, '인간이 아닌, 기계가 보편적 함수를 근사하여 내놓는 식으로 프로그래밍하는 세상'



엔비디아 풀스택 (하단: 칩과 rack / 중단: CUDA 가속 라이브러리 / 상단: NIM과 옴니버스)

자료: Nvidia, 미래에셋증권 디지털리서치팀

Paper of the Week

사람을 앞지른 손끝 – Meta FAIR

기존의 인공지능 시스템들은 시각이나 청각 등 단일 감각에 의존하여 물리적 세계와 상호작용을 하므로, 오감을 통해 상호작용하는 인간과 달리 태생적인 한계가 있었음. 메타는 이 중 인간이 세상과 직접 맞닿는 수단인 **촉각을 인간 수준의 스펙트럼으로 디지털화하기 위한 솔루션**으로 **인공 손가락 시스템인 Digit 360**을 공개.

Digit 360은 카메라를 이용한 시각적 촉각 시스템의 일종임. 물체에 손끝이 닿았을 때 피부가 눌리는 정도로 경도를 짐작하듯, Digit 360의 손끝 역시 탄성 중합체로 이루어져 있어 물체와 접촉했을 때 표면이 일그러지게 됨. 이 때 손가락 안쪽에서 빛을 쏘고, 반사되는 빛의 패턴 변화를 통해 인공 손끝이 얼마나 눌렸는지를 파악해 물체의 경도와 같은 정보를 습득하는 방식임. Digit 360은 이렇게 얻어지는 정보의 정확성을 개선하기 위해 화학 증착 방식으로 반사막을 코팅하고, 반구형 어안렌즈를 채택해 손끝 전체에 대한 데이터를 골고루 수집할 수 있도록 했음.

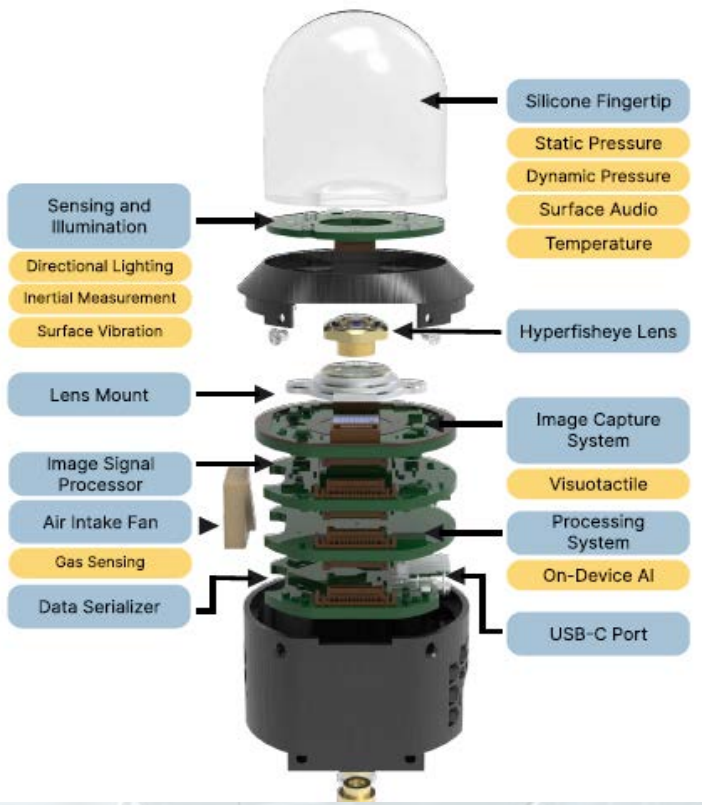
또한 시각적 정보만으로 물체의 특성을 완벽하게 파악하는 것은 어렵기 때문에, 멀티모달 센싱 시스템을 사용해 다양한 센서를 통합했음. 여기에는 진동/압력 센서, 온도 센서, 가스/냄새 센서 등이 포함되어 있어, 이를 통해 용기 내 액체량 감지, 물체 상태 판별, 표면 특성 분석 등 다양한 응용이 가능해짐.

마지막으로, 지연시간을 줄이기 위해 온디바이스 AI를 탑재해 센서 내에서 정보를 직접 처리함. 기존의 촉각 센서들은 외부 처리 방식을 사용해 높은 지연시간을 갖는다는 문제가 있었는데, Digit 360은 정보를 센서 내에서 직접 처리해 1.2밀리초의 초고속 응답 속도를 구현함. 이는 인간의 가장 빠른 반사 속도인 척수반사(평균 20-50밀리초)보다도 현저히 빠른 수준이며, 이 중 가장 빠른 눈 깜박임 반사(3-5밀리초)보다도 2배 이상 빠른 것.

이 밖에도 다양한 실험에서 Digit 360은 인간의 능력을 크게 상회하는 성능을 보였음. 공간 해상도는 인간의 약 143배, 힘 감지 능력은 인간의 60배 개선됐으며, 진동 감지 능력도 인간의 1kHz에서 10kHz로 10배 향상됨. 이 밖에 가스/냄새 감지에서 91%의 정확도를, 다중모달 분류에서 85.79%의 정확도를 달성하며 멀티모달 센싱의 유용성이 확인됨.

AI가 학습할 수 있는 데이터의 모달리티가 점점 다변화되고 있음. LLM이 활용할 수 있는 데이터가 언어에서 이미지, 음성으로 확장되면서도 실생활에서의 활용도가 크게 늘었으나, 이제는 각종 센서 데이터로 확장되며 세상과 상호작용하는 AI 에이전트의 모습을 갖춰 나가고 있음. 메타가 촉각 데이터를 메타버스에서 구현할 수 있을지, 그리고 오감을 모두 데이터화했을 때, 이것이 AI가 세상을 인지하는 능력을 얼마나 확장시킬지 기대됨.

자료: “Simplifying, Stabilizing & Scaling Continuous Time Consistency Models”, OpenAI.



Digit 360 구조(좌)
항목별 성능(아래).

Digit 360이
비교 가능한
모든 항목에서
최소 10배 이상의
성능을 보임.

측정 항목	인간	Digit 360	개선 비율
공간 해상도	약 1mm	7μm	~143배
수직력 감지	0.06N	0.001N	60배
전단력 감지	-	0.00127N	-
진동 감지	최대 1kHz	최대 10kHz	10배
반응 속도	20-50ms(척수반사)	1.2ms	~17-42배

자료: “Simplifying, Stabilizing & Scaling Continuous Time Consistency Models”, OpenAI.

AI Peer Table (1)

밸류체인	종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						엔비디아와의 상관관계(30일)	실적발표 예정일	AI 밸류체인 관련 한 줄 평
					1W	1M	3M	6M	1Y	YTD			
최종 소비자 (온디바이스, 앱)	애플	AAPL US	223.45	4,657.1	-4.4	-1.5	6.9	22.2	27.1	16.5	0.43	01-30	온디바이스 AI 관련 엔드유저 데이터보유 및 서비스 배포능력은 타의 추종 불허
	테슬라	TSLA US	251.44	1,112.9	-3.1	0.5	26.4	38.8	14.3	1.2	0.36	01-24	자율주행 소프트웨어 FSD v12의 미국과 중국 시장 내 공공 출시 임박
	삼성전자	005930 KS	57,700	344.5	-2.4	-4.8	-20.0	-24.9	-17.0	-25.4	0.15	01-08	온디바이스 AI, HBM, 그리고 파운드리 영역에서 AI 수혜 기대(Galaxy AI)
	퀄컴	QCOM US	165.91	254.8	-5.7	-1.8	5.6	-6.8	41.7	16.4	0.58	11-06	생성 AI 처리 강화용 온디바이스 AP 및 CPU 설계(X Elite)
	스냅	SNAP US	12.19	28.2	11.9	12.9	40.0	-25.0	13.7	-28.0	0.28	02-06	이미지 생성 모델을 탑재한 증강현실 서비스(AR 필터, AR 렌즈)
비즈니스 효율화	어도비	ADBE US	486.42	295.2	0.2	-4.1	-4.5	0.0	-13.7	-18.5	0.45	12-14	세계 최고의 미디어편집 툴로 막대한 생성 AI 서비스 배포능력을 지닌 기업(Firefly)
	세일스포스	CRM US	297.49	392.1	-0.5	3.4	24.7	9.1	44.0	13.6	0.37	11-29	세계 최고의 CRM 업체로 생성 AI를 탑재해 사용성 강화(Einstein GPT)
	서비스나우	NOW US	971.04	275.8	1.9	5.8	24.3	35.5	60.0	37.4	0.48	01-24	워크플로우 자동화 기업으로 생성 AI를 탑재해 사용성 강화(Now Assist)
	크라우드스트라이크	CRWD US	307.58	104.0	-1.1	5.2	38.5	-0.8	63.4	20.5	0.41	11-26	기업 고객 대상 생성 AI 챗봇 기반 클라우드 및 엔드포인트 보안(Charlotte AI)
	IBM	IBM US	207.57	264.6	-1.4	-8.2	14.2	27.6	45.9	30.5	0.18	01-29	기업 고객용 AI 어시스턴트(상담사 챗봇 구축, 업무 자동화, 코딩 지원)
	SAP	SAP GY	215.65	399.3	-3.7	7.4	19.0	27.9	68.9	56.5	0.34	01-28	기업 고객용 AI 어시스턴트(SAP Business AI: ERP 등 프로세스 지능형 자동화)
	유니티	U US	21.62	11.8	4.3	-1.0	49.3	-12.6	-21.8	-47.1	0.30	11-07	올해 출시 예정인 차세대 게임엔진에 AI 기능 대거 탑재(Sentis, Muse)
	모더나	MRNA US	53.29	28.2	-3.5	-11.5	-36.4	-57.4	-31.3	-46.4	0.16	11-07	OpenAI와의 협력으로 mRNA 신약 개발 등 모든 업무에 AI 활용(Dose ID GPT)
	루닛	328130 KS	41,450	1.2	6.3	-3.6	23.2	-27.0	-55.3	-49.4	0.08	11-13	AI 기반 암 검출/진단 보조 솔루션(인사이트)
기반 모델	알파벳	GOOGL US	169.74	2,878.6	0.0	1.6	6.7	1.7	31.8	21.8	0.42	01-30	OpenAI와의 AGI 경쟁에 있어 대적할 수 있는 거의 유일한 업체(Gemini)
	메타 플랫폼스	META US	572.43	1,993.2	-3.5	-3.9	20.4	26.9	82.5	62.2	0.42	01-31	'거의 완전한' 오픈소스 LLM 개발 및 배포를 선도(Llama)
	알리바바	9988 HK	98.40	334.4	2.7	-13.6	30.8	26.5	20.1	33.0	0.09	11-15	중국의 하이퍼스케일러이자 중국 최고의 오픈소스 언어모델 성능을 보유(Qwen)
	바이두	9888 HK	90.30	44.9	1.2	-18.3	12.9	-18.4	-16.5	-22.2	0.13	11-21	중국 내 자연어처리 및 자율주행 모델의 전통적 강자(Ernie Bot, Apollo)
	센스타임	20 HK	1.64	10.2	5.8	-22.6	50.5	-2.4	7.2	41.4	0.07	03-26	중국의 SOTA급 멀티모달 AI 모델을 만들 잠재력이 있는 기업(SenseNova)
	네이버	035420 KS	175,800	28.3	0.8	3.0	11.0	-9.7	-14.1	-21.5	0.08	11-08	자체 개발한 한국형 LLM 개발 및 AI 서비스 출시 예정(HyperCLOVA)
AI Ops	몽고DB	MDB US	274.75	28.0	-0.2	3.5	24.0	-24.3	-19.9	-32.8	0.37	12-05	AI 모델을 구축할 때 필수 인프라로 떠오른 '벡터 데이터베이스' 기능 보유(Atlas)
	클라우드플레어	NET US	87.71	41.4	-2.0	5.1	12.9	17.9	36.2	5.3	0.43	11-07	트래픽 처리(CDN) 업체로, 실시간 AI 추론 작업에서의 역량 대두(Workers AI)
	팔란티어	PLTR US	51.13	160.5	13.8	27.8	112.2	119.2	170.7	197.8	0.37	02-05	실제 조직 운영시의 의사결정 지원 및 자동화를 위한 AI 개발 플랫폼(AIP)
	포스 패러다임	6682 HK	34.65	2.9	45.3	20.1	-20.3	-38.5	-29.3	-31.0	0.03	03-20	실제 조직 운영시의 의사결정 지원 및 자동화를 위한 AI 개발 플랫폼(Sage)

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

AI Peer Table (2)

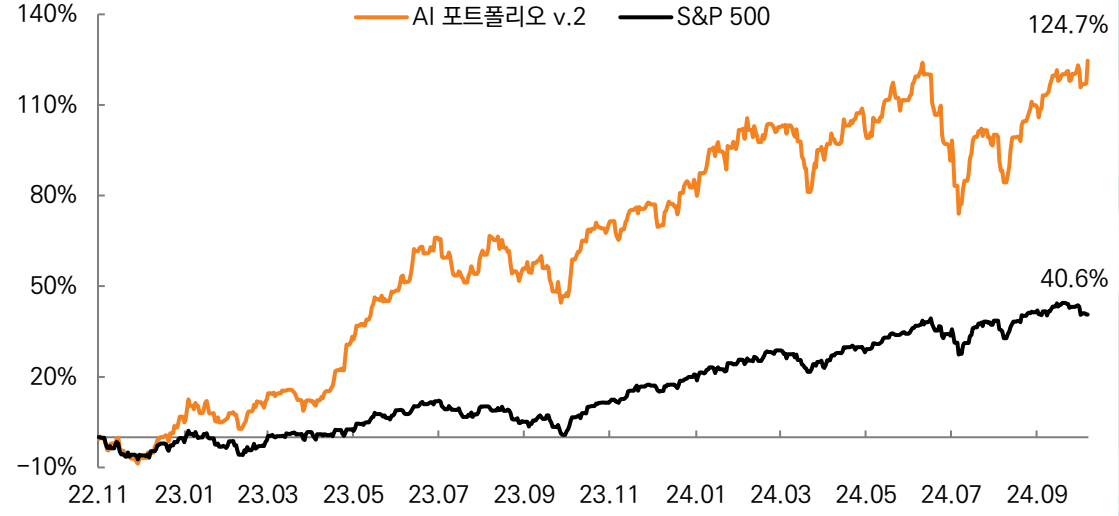
밸류체인	종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						엔비디아와의 상관관계(30일)	실적발표 예정일	AI 밸류체인 관련 한 줄 평
					1W	1M	3M	6M	1Y	YTD			
클라우드	마이크로소프트	MSFT US	411.46	4,218.0	-4.7	-1.1	4.3	1.5	17.5	10.0	0.55	01-30	OpenAI의 AI 모델 라이센스에 대한 독점권을 갖고 있는 세계 2위의 CSP
	아마존 닷컴	AMZN US	199.50	2,892.4	4.5	7.0	23.9	7.1	43.9	31.3	0.47	01-31	최고의 CSP로써 작년 9월 Anthropic에 40억 달러 투자하며 AGI 경쟁 참여
	오라클	ORCL US	171.78	656.3	-1.0	0.8	34.7	49.1	60.9	64.9	0.40	12-11	기업들의 자체 인프라 수요에 대응한 '온프레미스 서버' DB 관리 시스템의 강자
AI 인프라	엔비디아	NVDA US	139.91	4,732.1	-0.9	12.0	39.3	57.6	211.0	182.6	1.00	11-20	AI 훈련은 GPU가 필수, GPU 현존 최강자는 엔비디아
	브로드컴	AVGO US	173.90	1,119.9	-3.0	-1.6	22.8	36.9	100.0	57.4	0.65	12-06	VMware 인수로 클라우드 가상화 역량도 확보한 네트워킹 반도체의 최강자
	AMD	AMD US	141.66	317.0	-14.8	-17.1	5.1	-5.9	26.2	-3.9	0.66	01-30	HPC CPU에서 업계리딩, GPU에서 엔비디아의 적수로 꼽히는 패스트 팔로워
	인텔	INTC US	23.32	138.7	1.8	3.2	16.7	-23.7	-37.8	-53.0	0.32	01-24	GPU 중심 서버 설계에는 후발주자. 파운드리 기업으로서의 지정학적 가치 부상
	마이크론	MU US	105.49	161.3	-2.5	3.3	16.8	-7.8	46.0	24.0	0.57	12-20	HBM에서 SK하이닉스를 추격하는 미국의 메모리 반도체 업체
	SK하이닉스	000660 KS	197,100	143.6	1.1	13.2	20.6	14.1	49.2	39.9	0.26	01-24	GPU 병목의 키를 쥐고 있는 HBM에 있어 글로벌 선두주자
	시놉시스	SNPS US	521.98	110.6	-1.4	4.6	7.0	-2.8	7.0	1.4	0.68	12-04	반도체 EDA 업체로, 3D 엔지니어링 시뮬레이터 업체 앤시스와 M&A 논의 중
	암페놀	APH US	69.71	115.9	1.2	11.7	19.7	14.2	65.9	41.5	0.53	01-22	AI 및 HPC 서버에 필요한 고속 I/O, 메모리 및 스토리지 커넥터 제조
	버티브	VRT US	112.44	58.2	-1.2	6.9	66.1	21.0	178.4	134.3	0.55	02-21	데이터센터의 열 관리를 위한 액체 냉각 방식에 있어 글로벌 선두업체
	Arm	ARM US	140.65	203.2	-10.5	0.1	27.3	38.3	162.9	87.2	0.43	11-06	AI 가속기 설계 위한 다수의 IP 블록 소유한 업체로 설계 라이선스/로열티 사업
	마벨 테크놀로지스	MRVL US	87.19	104.1	2.7	18.9	51.3	27.5	70.4	45.1	0.63	11-29	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
	SMIC	981 HK	27.50	66.2	3.4	0.5	74.5	71.9	12.9	38.5	0.06	11-07	최근 화웨이와 밀접히 협력 중인 중국 유일의 미세공정 가능 파운드리
	TSMC	2330 TT	1,050.00	1,174.9	1.9	7.5	19.8	34.7	94.3	79.3	0.19	01-20	압도적인 미세공정 점유율을 가진 대만의 반도체 파운드리
	콴타	2382 TT	318.00	53.0	4.8	21.4	28.7	21.4	59.9	46.4	0.16	11-11	AI 서버 공급망에서 GPU 기반 보드 및 서버 시스템 조립 역할을 하는 대만 업체
	위스트론	3231 TT	118.50	14.8	6.8	18.5	33.6	6.4	31.8	23.1	0.18	11-11	AI 서버 공급망에서 GPU 기반 보드 및 서버 시스템 조립 역할을 하는 대만 업체
	Alchip	3661 TT	2,195.00	7.6	7.6	16.4	11.9	-21.9	-25.0	-32.4	0.15	03-03	세계에서 AI 반도체 관련 매출 비중이 가장 높은, ASIC 설계 관련 대만 업체
	소프트뱅크	9984 JP	8,996.00	120.2	-5.7	5.1	25.7	15.7	41.1	43.7	0.17	11-12	동사가 약 25%의 지분을 가진 '비전펀드'는 AI 관련 유니콘 기업에 투자를 집중

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

Appendix: Charts

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

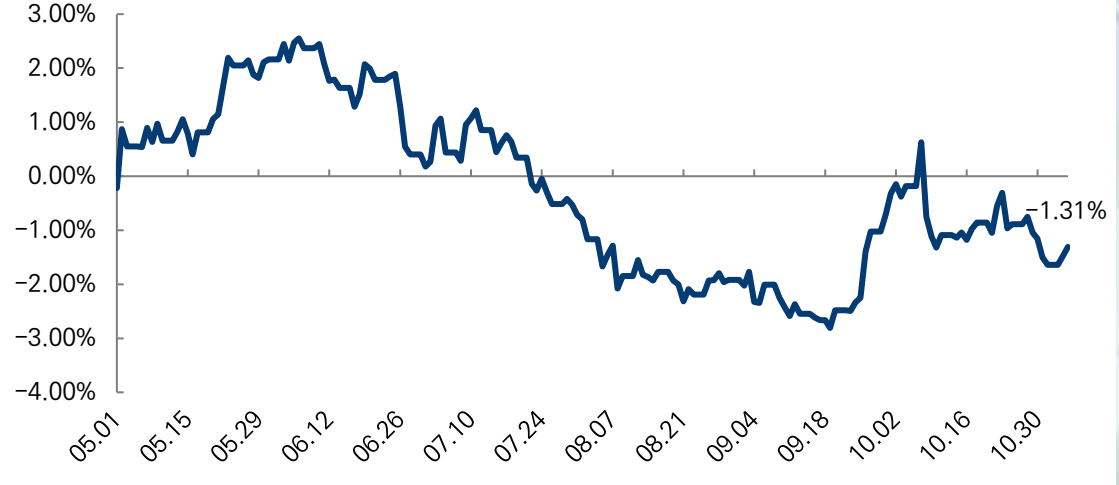
ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률 (동일가중 방식)



LLM 벤치마크 LiveBench 상위 7개 모델 (동일 모델은 생략, 마지막 업데이트: 2024.08.31)

순위 (전주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	o1-preview	66.02	OpenAI	독점 폐쇄소스	2023년 10월
2(-)	Claude-3.5 Sonnet -20241022	60.33	Anthropic	독점 폐쇄소스	2024년 4월
3(-)	o1-mini	59.09	OpenAI	독점 폐쇄소스	2023년 10월
4(-)	GPT-4o	56.03	OpenAI	독점 폐쇄소스	2023년 10월
5(-)	Gemini 1.5 Pro	54.94	구글	독점 폐쇄소스	2023년 11월
6(-)	Llama 3.1-405B	53.24	메타	별도 라이선스 필요 (상업적 사용허가)	2023년 12월
7(-)	Qwen2.5-72B	52.89	알리바바	별도 라이선스 필요 (상업적 사용허가)	2024년 9월

이전 포트폴리오 버전(v.1) 대비 AI 유니버스 누적 수익률 차이 (%p, 24.04.30 기준)



자료: Bloomberg, LiveBench, Huggingface, 미래에셋증권 디지털리서치팀

Compliance Notice

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