Market Insight | 2025.10.24

그림: Midjourney(에이전트가 사용자 맞춤 기억을 가져오는 모습)

AI Hot Issue

에이전트 문제(디지털 기억상실증)는 메모리가 해결한다

한종목
chongmok.han@miraeasset.com

Executive Summary

카파시의 예언과 AI 패러다임의 전환, 그리고 메모리

OpenAI의 창립 멤버이자 테슬라에서 자율주행 FSD/로보택시의 기술적 토대를 닦았던 전설적인 엔지니어, 안드레이 카파시(Andrej Karpathy)의 최근 인터뷰가 AI 산업 전반에 묵직한 파장을 일으키고 있다.

그의 발언은 과열된 'AI 하이프(Hype)'에 들뜬 시장을 차분하게 식히고 AI 엔지니어링의 냉혹한 현실을 직시하게 만든 중요한 분기점으로 평가받는다. 그는 AGI(범용인공지능)가 도래하기까지 아직 10년 이상의 시간이 필요하며, AI가 당장 경제 성장률 그래프를 수직 상승시키는 마법 같은 특이점이 되지는 않을 것이라는 현실론적 전망을 내놓았다.

그러나 카파시의 이러한 진단을 단순한 비관론이나 회의주의로 치부해서는 안 된다. 오히려 이는 현재 거대언어모델(LLM)이 가진 근본적인 아키텍처의 한계를 정확히 꿰뚫어 보고, 이를 해결하기 위한 차세대 AI의 청사진을 제시하는 깊은 기술적 통찰에 가깝다.

그는 현재의 AI가 그럴듯한 유창함 뒤에 심각한 '디지털 기억상실증(지속적 학습의 부재)'을 앓고 있으며, '빨대로 정보를 빠는' 듯한 비효율적인 강화학습 방식에 의존하고 있다고 냉철하게 지적한다. 이러한 근본적 결함을 해결하지 못한다면, AI는 연구실의 신기한 데모를 넘어 인간과 협업하는 진정한 의미의 '에이전트'로 진화할 수 없다.

중요한 것은 그의 소프트웨어적 비전이 필연적으로 하드웨어 아키텍처의 근본적인 변화를 요구한다는 점이다. 그가 꿈꾸는 '기억을 가진 지능형 에이전트'는 일회성 응답을 넘어, 사용자와의 긴 상호작용 맥락과 방대한 과거 기록을 실시간으로 유지하고 관리해야 하는 '상태 기반(Stateful)' 시스템이다. 이는 AI 시스템의 핵심 병목이 거대한 행렬 연산을 처리하는 '컴퓨팅(Compute)'에서, 방대한 상태 정보를 실시간으로 읽고 쓰는 '메모리(Memory)'로 결정적으로 이동함을 의미한다.

결국 카파시의 예언은 현재의 단일 메모리 계층으로는 미래의 AI를 감당할 수 없으며, HBM(초고속 작업 기억)-DRAM(단기 기억)-SSD(영속적 장기 기억)로 이어지는 '계층형 메모리 아키텍처(Tiered Memory Architecture)'가 선택이 아닌 필수적인 생존 조건임을 시사한다.

본 보고서는 카파시의 통찰을 심층 분석하여, 그의 비전이 어떻게 단순한 기술 예측을 넘어 거대한 '메모리 슈퍼사이클'을 견인하는 구조적 동인이 되는지에 대한 엄밀한 기술적 인과관계를 규명한다. 그의 메시지는 AI의 겨울이 올 것이라는 경고가 아니라, 이제 막 시작된 진정한 '메모리 중심 AI' 시대로의 패러다임 전환을 알리는 서막이다.

I. 적지 않은 파장의 카파시의 발언

1. 현실 직시: AGI는 기술이 아닌 경제적 티핑 포인트다.

안드레이 카파시의 인터뷰는 AI 시장의 과열된 열광에 현실이라는 찬물을 끼얹었다. AI 업계에서 가장 영향력이 높은 사람이기 때문에 그의 발언은 기술의 잠재력에 대한 비관론으로 해석한 사람이 많았기 때문에 커뮤니티는 그의 말의 행간을 읽느라 난리가 났었다.

그림 1. 드와케시 파텔의 팟캐스트에 출연한 안드레이 카파시의 인터뷰 영상
"(현재의) 강화학습은 끔찍합니다"
사전학습의 스케일링 이후의 희망이 강화학습 스케일링이었다는 점에서 큰 충격을 준 발언

자료: 안드레이 카파시, 드와케시 파텔, 미래에셋증권 리서치센터

하지만, 우리 팀의 해석으로는 카파시의 말은 하나의 화려한 'AI 데모'가 사회를 움직이는 '실제 제품'이 되기까지의 험난한 경로에 대한 냉철한 분석에 가까웠다. 그는 AGI를 연구실의 성과가 아닌, 자본주의 시장의 논리가 작동하는 경제적 현상으로 재정의하며, 그 도래 시점과 파급력에 대한 근본적인 관점의 전환을 요구한다는 점이 포인트다.

(1) 10년의 시간: '9의 행진'에서 비롯된 전투적 현실론

카파시가 제시한 '10년'이라는 꽤 긴 미래와 같이 들리는 AGI 타임라인은 막연한 추측이 아니다. 우선 안드레이 카파시는 OpenAI의 창립멤버다.

또한 그는 테슬라에서 5년간 재직하면서 자율주행 시스템의 신뢰도를 99%에서 99.9%, 그리고 99.99%로 끌어올리는, 피를 말리는 '9의 행진(March of Nines)'을 직접 지휘한 엔지니어의 전투로 단련된 직관을 갖고 있는 사람이다. 카파시의 말에 따르면, 신뢰도의 마지막 소수점 뒤의 9 하나를 추가하는 데 드는 노력은, 이전 99%를 달성하는 데 들어간 노력과 맞먹거나 그 이상이다. 그가 말한 10년이라는 숫자는 감(vibe)에 의존하지만 그 감은 꽤나 무거운 감이다.

그의 AGI에 대한 관점은 생각보다 더 일관성이 있다. 2015년 OpenAI 초창기, 그는 동료들에 비해 매우 길었던 '20년' 후 AGI 도래를 예측했다고 한다. 10년이 지난 지금, 그는 여전히 남은 기간을 '10년'으로 고수하며 자신의 초기 전망이 틀리지 않았음을 시사한다.

(2) AGI의 재정의: 벤치마크가 아닌, 당신의 해고 통지서

그런데, 카파시가 말하는 AGI는 AI가 인간 수준의 능력을 갖추는 것을 의미하는 것만은 아니다. AI가 예측 불가능한 온갖 예외 상황을 모두 처리하고 '사회 시스템에 안정적으로 통합'되어, 기업이 "인간 직원 대신 이 AI를 고용하겠다"고 확신할 만큼의 신뢰도를 확보하는 데 필요한 고된 엔지니어링의 시간이다. 즉, 이러한 현실론은 AGI의 정의 자체를 근본적으로 바꾼다. 카파시에게 AGI의 달성 여부는 OpenAI가 발표했던 GDPVal과 같은 벤치마크 퍼즐 점수로 판단하는 것이 아니다. 인간들보다 GPT-n이 더 뛰어나고, 더 저렴하며, 지분도 요구하지 않고, 주말에도 일하기 때문에 '해고'당하는 바로 그 순간이 AGI의 도래라는 생각이다.

따라서, 일반적인 통념의 AGI보다는 카파시가 생각하는 AGI가 더욱 보수적인 개념이고 그러니 AI에 대한 비관론을 의미하는 것은 아니다. 사실 카파시가 말하는 의미의 AGI가 도래했을 때는 상당히 많은 사람들이 일자리를 잃었거나 잃을 가능성이 높아진 국면일 가능성이 높다. 이는 AGI를 기술적 이정표가 아닌, 사회경제적 현상으로 재정의하는 것이다. 노동에 대한 경제적 합리성을 AI가 모든 면에서 뛰어넘었다고 판단하는 바로 그 '티핑 포인트(Tipping Point)'가 진짜 AGI의 시작이라는 것을 유념하자.

(3) 2% 성장의 수렴: 점진적 통합인가, 비연속적 도약인가?

카파시의 가장 논쟁적인 주장은, 이 거대한 변화조차 GDP 성장률에는 극적인 변곡점을 만들지 못할 것이라는 예측이었다. 그는 컴퓨터, 휴대폰, 아이폰과 같은 인류사를 바꾼 혁신 기술들도 GDP 곡선에서 뚜렷한 도약을 만들어내지 못했다는 역사적 사실을 근거로 든다. 이 기술들은 사회에 매우 점진적으로 확산되고 통합되어, 기존의 연평균 2% 내외의 지수적 성장 추세에 자연스럽게 흡수되었다는 게 근거다.

카파시는 AI 역시 이 패턴을 따를 것이라고 본다. AI는 분명 이전에 없던 새로운 가치를 창출하는 혁신이지만, 그 기술이 사회 전체에 스며들고 산업 구조를 재편하며 법적, 사회적 합의를 이루기까지는 수십 년이 걸릴 것이라고 강조한다. 따라서 AI는 경제 성장률을 20%나 200%로 폭발시키는 기폭제가 아니라, 기존의 2% 성장률을 '유지'하게 만드는 강력한 동력으로 작용할 것이라는 게 그의 시각이다. AI 연구소에서 일어나는 재귀적 자기 개선조차 수십 년간 지속되어 온 '일상적인 비즈니스'의 일부로 간주한다고 한 점이 놀랍게 들리기도 했다.

물론 이에 대한 강력한 반론이 존재한다. AGI는 컴퓨터와 같은 생산성 향상 도구와는 질적으로 다르며, 그것은 '노동력 그 자체'라는 주장이다. 필자도 이 의견에 동의한다. 산업혁명이 인류의 성장률을 0.2%에서 2%로 10배 끌어올리는 비연속적/비가역적인 도약을 일으켰듯, 무한히 복제 가능한 지능 노동력의 등장은 성장의 패러다임 자체를 바꿀 수 있다고 본다. 지난주 우리가 작성했던 AI -Bi-Weekly에서 기재했듯 AI는 과학과 공학 여러 영역에서 혁신의 속도를 더욱 부추길 것이다.

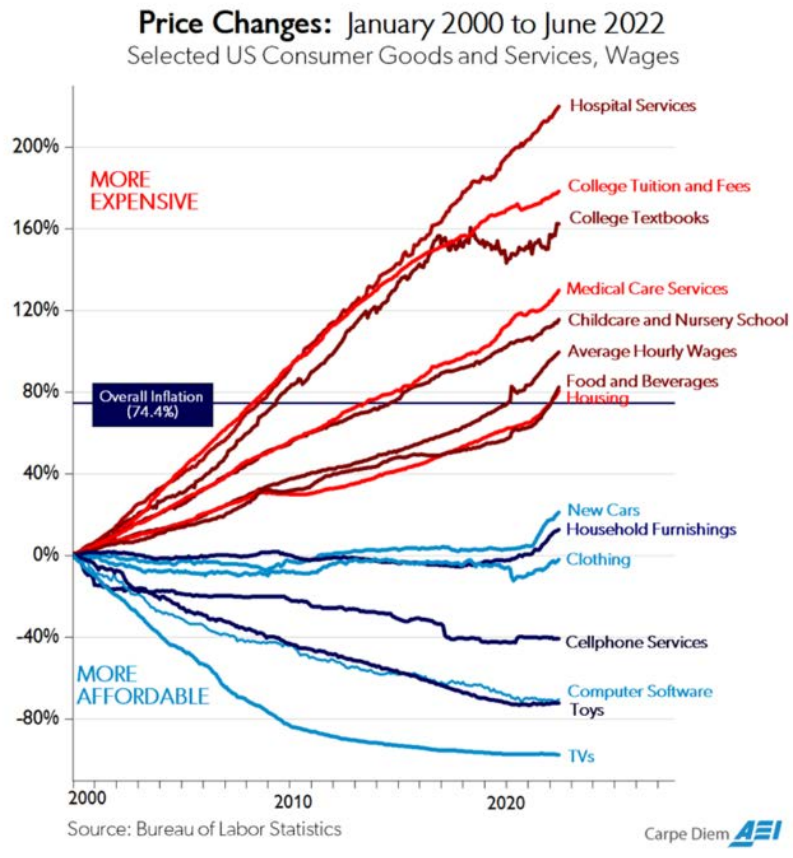
물론 Bit로 이뤄진 가상의 혁신을 현실의 양자 세계로 옮겨오는 것은 쉽지 않지만 이전과는 다른 속도의 파괴적 혁신이 빈번하게 발생할 것이라고 생각한다. 미국 댈러스 연방준비위원회의 분석 또한 AI를 GDP의 “All or Nothing”을 결정짓는 기술로 평가하며 이러한 관점을 뒷받침한다.

카파시는 이러한 반론에 대해서도 재반박했다. 컴퓨팅 역시 이미 인간의 ‘계산 노동’을 대체해왔지만 경제를 폭발시키지는 못했다는 것이다. 그의 관점은 기술의 잠재력을 불신하는 것은 아닌 듯하다. 다만, 그 잠재력이 현실 경제에 반영되기까지의 엄청난 마찰과 사회적 수용도의 한계를 직시하는 데 있다.

비록 그의 예측대로 AI로 인한 변화가 점진적으로 그친다 하더라도, AI가 우선적으로 적용될 지식 노동 시장만 해도 미국 경제의 10~20%, 즉 수조 달러에 달하는 엄청난 규모다. 이것만의 자동화도 인류사적 변화임은 명백하다. 결국 그의 모든 분석은 기술 그 자체보다, 기술이 인간 사회와 상호작용하며 만들어낼 결과에 깊이 천착하고 있음을 보여준다.

결국, 안드레이 카파시는 technology에는 낙관론, sociology에는 회의론적 시각을 지니고 있다. "AGI는 만들어질 것이다. 그러나 세상은 준비되지 않았다"는 것이지, “AI가 또 한 번 벽에 부딪혔다. 겨울이 올 것이다”고 그의 말을 해석하는 것은 무리가 있다.

그림 2. 이미 2023년에 카파시가 강조했었던, ‘보몰 효과(Baumol's cost disease)’에 대한 자료
급락하는 파란색 선들(소프트웨어, TV)은 기술의 폭발적인 생산성 향상과 그로 인한 디플레이션 효과 반면, 치솟는 빨간색 선들(의료, 교육)은 복잡한 사회적 마찰이 기술의 파급력을 억제하는 현실
결국 AGI라는 경이로운 기술이 완성되어도, 세상을 바꾸는 속도는 더딘 변화율에 흡수될 수도…



자료: 미국 노동 통계청, 미국 기업 연구소(2022), 미래에셋증권 리서치센터

2. 문제 정의: 현재 LLM의 근본적 결함

안드레이 카파시는 현재 LLM이 보여주는 유창함 뒤에 숨겨진, 더 높은 지능으로의 도약을 가로막는 세 가지 근본적인 아키텍처 결함을 지적한다. 이 문제들은 단순히 모델을 더 크게 만드는 것(스케일링)으로는 해결되지 않으며, AI의 작동 방식에 대한 근본적인 재고를 요구하고 있다. 여기서 여러가지 힌트를 얻을 수 있기 때문에 독자분들께 공유드리려 한다.

(1) 디지털 기억상실증: 지속적 학습을 가로막는 단절된 기억

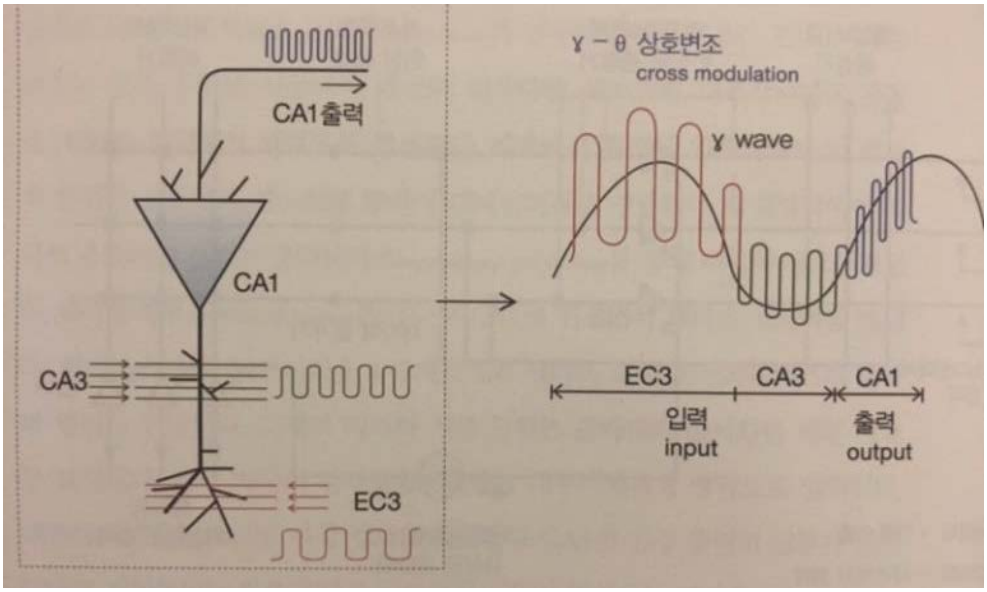
현재 LLM의 가장 명백한 한계는 사용자와의 상호작용을 통해 배우고 성장하지 못한다는 점이다. 모든 대화 세션은 독립적이며, 대화가 끝나면 모델은 모든 것을 잊어버리고 아무것도 배우지 않은 초기 상태로 리셋된다. 카파시는 이 '디지털 기억상실증'의 원인을 인간의 기억 시스템에 비유하여 설명한다.

작업 기억 (Context Window): 사용자와의 대화 내용은 임시 저장 공간인 '컨텍스트 창'에 담긴다. 이는 인간이 대화 중 방금 들은 말을 기억하는 '작업 기억(working memory)'과 같다. 이 기억은 매우 선명하고 접근하기 쉽지만, 대화가 끝나면 사라지는 휘발성 기억이다.

장기 기억 (Model Weights): 모델의 근본적인 지식과 능력은 수십억 개~수조 개의 '가중치(weights)'에 극도로 압축되어 저장되어 있다. 이는 인터넷 전체를 학습하며 형성된 '장기 기억'과 같지만, 압축률이 너무 높아 특정 사실을 떠올리는 것은 '희미한 회상'에 가깝다.

여기서 결정적인 문제는 이 두 기억 시스템 사이에 "다리"가 없다는 것이다. 대화의 형태로 작업 기억 속으로 들어온 정보가, 장기 기억인 가중치로 업데이트되는 메커니즘이 부재하다는 말이다. 인간의 뇌에서는 해마(Hippocampus)라는 기관이 이것을 수행한다.

그림 3. 인간이 잠을 잘 때, 해마의 '증류' 과정이 신경과학적으로 어떻게 일어나는지를 설명
느리고 규칙적인 세타파(검정)는 하루 동안의 기억을 통합할 시간적 틀(Timing Framework)을 제공
빠르고 복잡한 감마파(빨강)는 낮에 경험했던 수많은 장면, 소리, 감정과 같은 디테일한 기억의 조각
느린 세타(θ)파의 리듬 위에서, 기억 조각(감마파)이 여러 번 반복재생되고 다듬어지면서,
불필요한 부분은 사라지고 핵심만 남게 되는 것이 기억의 증류.
잘 정제된 기억은 비로소 뇌의 다른 부분(대뇌 피질)으로 보내져 안정적인 장기 기억으로 저장



자료: 책(박문호 박사의 뇌과학 공부), 네이버 블로그(매일 운동하는 타오), 미래에셋증권 리서치센터

즉 하루 동안의 방대한 경험(주로 낮에 발생)에서 핵심 에센스를 추출하여 뇌 구조 자체를 바꿔주는 과정(수면 과정에서 발생)이 LLM에는 완전히 빠져있다. 결국 "기억 증류(memory distillation)"의 부재라는, AI의 구조적 결함이 바로 진정한 의미의 지속적 학습을 불가능하게 만드는 근본 원인이다.

(2) 인지적 족쇄: 망각하지 못하기에 일반화에 실패하는 역설

더욱 심오한 문제는 LLM의 완벽한 암기력이 오히려 일반화 능력을 저해하는 인지적 족쇄로 작용한다는 카파시의 역설적인 통찰이다. 그는 사전훈련이 '지식(Knowledge)'의 습득과 '지능(Intelligence)'의 형성이라는 두 가지 다른 결과를 낳는다고 분석한다.

지식(Knowledge): "프랑스의 수도는 파리다"와 같은 사실적 정보를 암기하는 능력.

지능(Intelligence): 수많은 데이터를 보며 스스로 '추론하는 법', '질문에 답하는 법'과 같은 패턴과 구조를 암묵적으로 학습하는 능력.

카파시는 이 중 '지식', 즉 과도한 암기력이 모델의 발목을 잡는다고 주장한다. 암기 방식에 익숙해지면, 데이터에 존재하지 않았던 새로운 문제에 직면했을 때, 암기한 지식에만 의존하려 하고 스스로 추론하여 답을 찾는 능력이 저하될 수 있다는 말이다. 오픈북 시험에만 익숙해진 학생이, 책에 없는 응용 문제가 나오면 손도 못 대는 것으로 비유해볼 수 있다.

반면, 인간의 뇌는 모든 것을 기억하지 못한다. 그런데, 이 '망각'이라는 하드웨어적 한계가 신이 인간을 위해 만든 "버그가 아닌 기능"인 이유가 있다. 카파시의 의견에 첨언해보면, 암기를 잘 못하는 인류는 살아남기 위해서 개별 정보(나무)들은 잊고, 그 이면의 핵심 원리나 패턴(숲)을 추출하고 일반화하는 데 집중하도록 진화했다고 생각한다. 특별히 맛있는 과일을 맛본 호모 사피엔스라면, 그 과일나무의 디테일을 세세하게 외우기보다는 그 과일을 많이, 자주 얻을 수 있도록 계절적/지질적 패턴을 만드는 것이 생존에 훨씬 유리할 것이기 때문이다.

반면, 스케일링 법칙에 의거한 계속 커지는 언어모델들은 사전훈련 문서(주로 인터넷에 있는 텍스트 데이터)에 대한 방대한 기억력 때문에 오히려 산만해진다고 카파시는 말했다. 위키피디아 문장을 단어 하나 틀리지 않고 암송할 정도로 모든 '나무'를 너무나 선명하게 기억하기 때문에, 굳이 '숲' 전체의 패턴을 찾으려는 동기가 부족한 것이다. 결국, LLM의 극단적인 암기력은 본질적인 패턴 학습을 방해하는 장애물로 작용하는 것이다.

이는 앞으로 차세대 모델의 경우 모델 크기가 현재보다 크게 줄어들어야 한다는 명분을 제시하기에 충분했다.

(3) 쓰레기 더미 속 바늘 찾기: 모델이 불필요하게 거대해진 이유

카파시는 현재 LLM들이 무려 수조 개에 달하는 매개변수(파라미터)를 갖게 된 이유가 지능 자체가 그만큼 복잡해서는 아니라고 단언한다. 그 이유는 훈련 데이터, 즉 인터넷이 매우 저품질의 '쓰레기(garbage)'들로 가득차 있기 때문이다. AI 언어모델은 이 쓰레기 더미 속에서 아주 희소하게 존재하는 유의미한 정보(signal)를 추출해야 한다. 그래야만이 모델의 지능이 올라가기 때문이다.

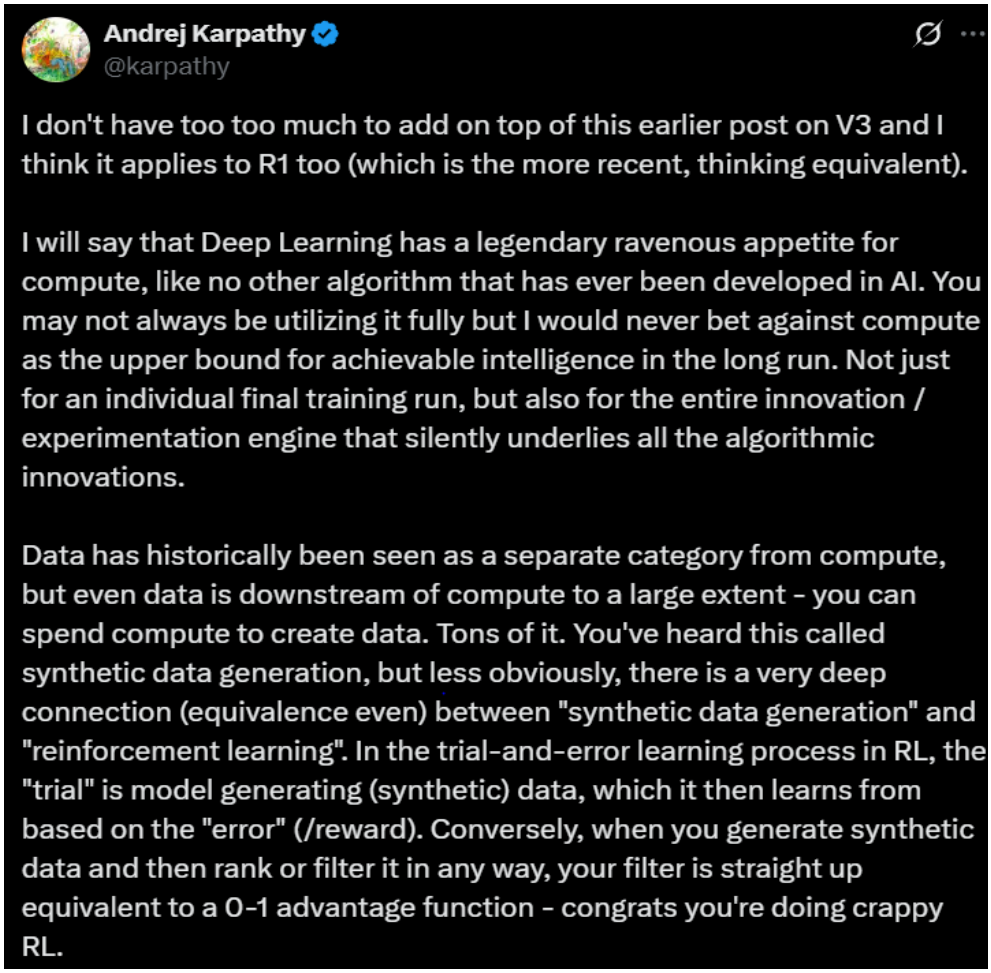
그런데 인터넷 세상은 온통 '쓰레기 더미'로 가득 찼으니 거기에서 '바늘을 찾기 위해' 어쩔 수 없이 크기라도 거대해져야만 했다. 그래서 모델의 방대한 파라미터 용량 대부분은 '생각하는 법(인지적 핵심)'을 저장하는 데 쓰이는 것이 아니라, 이 모든 '쓰레기'를 기억하는 데 낭비되고 있다는 설명이다.

안드레이 카파시가 내세우는 미래 비전은 명확하다. 만약 처음부터 쓰레기를 걸러내고, 지능 형성에 도움이 되는 양질의 '인지적 요소'만으로 구성된 데이터셋을 만들 수 있다면, 모델은 더 이상 방대한 암기력에 용량을 낭비할 필요가 없다는 것이다.

심지어 그는 약 10억 개 정도의 훨씬 작은 매개변수만으로도 순수한 '생각하는 능력', 즉 '인지적 핵심(cognitive core)'을 효율적으로 학습할 수 있을 것이라고 예측한다. 이는 '모델이 클수록 좋다'는 업계의 지배적인 패러다임에 대한 근본적인 도전이다.

카파시의 의견이 주류로써 자리 잡게 되면 GPU 수요는 급락하는 것처럼 들릴 수도 있다. 큰 모델을 만들 필요가 없기 때문이다. 하지만 우리는 그렇게 생각하지 않는다. 그것은 고품질의 데이터 자체를 확보하는 것이 점점 더 핵심이 되기 때문이다.

그림 4. 합성 데이터 생성을 통해 고품질 데이터를 대량으로 만들 수 있다고 설명하는 카파시
인지적 핵심을 담은 고품질 데이터 얻기 위한 합성 데이터 전략은 결국 컴퓨팅(GPU) 소모를 의미



자료: X(안드레 카파시), 미래에셋증권 리서치센터

카파시가 말하는 인지적 핵심을 위해서는 아주 정확하고 인사이트 있는 추론 과정을 품고 있는 고품질 데이터가 필요하다. 또 그런 데이터가 어느 정도는 양이 많이 확보되어 있어야 한다. 그럴려면 인간 전문가의 역할도 중요하겠지만, 결국 이것도 자동화를 해야 지속가능하다. 결국, 기계가 이 데이터 생성도 해야 한다는 뜻이고, 곧 합성 데이터 생성의 필요성으로 이어진다.

당연하게도 고품질의 합성 데이터 생성에는 막대한 컴퓨팅 자원을 요구된다. 이러한 논리는 우리만의 생각이 아니라 카파시 본인이 올해 1월 27일에 남긴 글에서 그 근거를 찾을 수 있다. 애초에 그는 데이터를 컴퓨트의 하위 개념으로 보면서, 합성 데이터(synthetic data) 생성을 통해 고품질 데이터를 대량으로 만들 수 있다고 설명했다. 즉, 데이터 자체가 GPU에 의존적이라는 점을 지적한 것이다.

3. 해결책 모색: 차세대 학습 패러다임의 조건

현재 LLM이 가진 근본적인 결함들(디지털 기억상실증, 지능보다 지식에 의존하는 족쇄)을 극복하기 위해, 안드레이 카파시는 기존 학습 패러다임의 한계를 명확히 지적하고, 이를 대체할 새로운 방법론의 필요성을 역설한다. 그의 분석은 더 나은 알고리즘을 찾는 것을 넘어, AI가 '학습'하고 '성장'하는 방식 자체에 대한 근본적인 재설계를 요구한다.

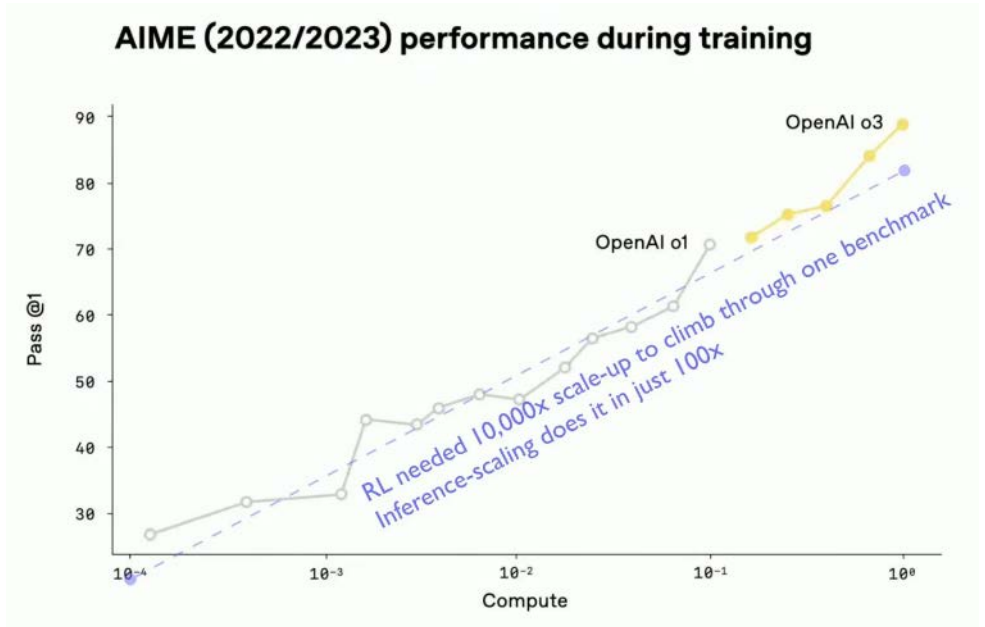
(1) 강화학습의 명백한 실패: '빨대로 감독 정보 빨아들이기'

카파시는 현재 AI 에이전트 훈련의 주류 방법론인 강화학습(RL)이 "끔찍하다(terrible)"고 단언하며 그 비효율성을 신랄하게 비판한다. 그 핵심 문제는 최종 결과라는 극도로 빈약한 정보에 의존하는 데 있다.

수학 문제를 푸는 AI를 예로 들면, 시스템은 수백 개의 각기 다른 풀이 시도를 생성한다. 현재의 강화학습에서는, 여러 풀이법 중 정답을 맞힌 풀이가 있다면, 그 과정에 포함된 모든 행동(토큰)의 확률을 무차별적으로 상향 조정한다. 이것은, 정답에 도달하기까지 거쳤던 잘못된 경로와 명백한 실수까지도 '좋은 행동'으로 학습하는 치명적인 오류를 낳는다. 다시 말해, AI는 어떤 행동이 진짜 정답에 기여했는지는 전혀 구분하지 못하고, 운 좋게 성공한 경우라고 하더라도, 그 경우의 모든 경로/발자국을 신성시하게 된다. 이로 인해 학습 과정에 엄청난 '노이즈(noise)'가 발생하여 효율적이고 안정적인 학습이 불가능해진다.

카파시는 이 과정을 '긴 빨대로 감독 정보를 빨아들이는 것(sucking supervision through a straw)'에 비유했다. 셀 수 없이 많은 토큰으로 이루어진 긴 AI의 추론 작업 끝에 얻은 '정답 or 오답'이라는 단 하나의 감독 정보(피드백)로써 전체 훈련 정책에 무차별적으로 전파하는 것은 근본적으로 비효율적이고 정보 손실이 극심하다는 말이다.

그림 5. 강화학습(RL)이 왜 '끔찍하게' 비효율적인지에 대한 카파시의 주장을 수치적으로 증명
AIME 벤치마크에서 성과를 내기 위해 RL 훈련은 무려 10,000배의 컴퓨팅 자원을 소모해야 했다
최종 결과라는 극미량의 정보에만 의존하여 학습하는, '빨대로 감독 정보를 빠는' 방식이기 때문

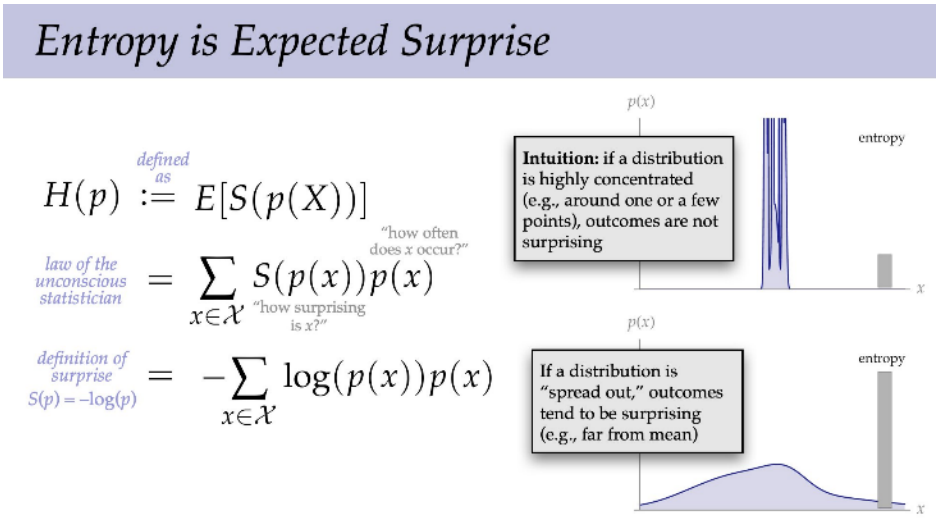


자료: OpenAI, 미래에셋증권 리서치센터

(2) 자기 생성 데이터의 함정: 다양성을 파괴하는 '모델 붕괴'

그렇다면 LLM으로 하여금 '스스로 성찰'하게끔 해서, 다시 학습하게 하면 되지 않을까? 이 시도는 모델의 성능을 오히려 악화시키는 '모델 붕괴(Model Collapse)' 현상으로 이어질 가능성이 높다는 게 카파시의 설명이다.

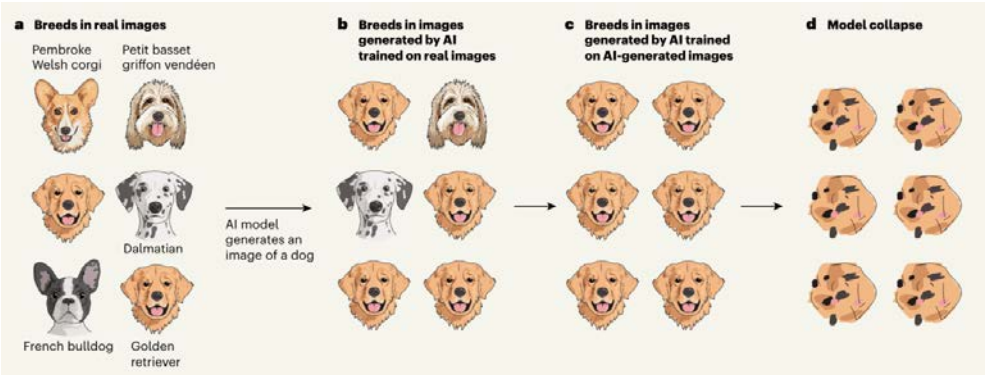
그림 6. 카파시가 '모델 붕괴'의 근본 원인으로 지적한 '엔트로피' 개념
엔트로피란? "정보의 불확실성" 또는 "놀람의 기대값"으로 이해하면 된다
AI가 인간처럼 되려면, 낮은 엔트로피의 함정에 빠지지 않도록 하는 메커니즘이 필요하다



자료: Arvind Chander, 미래에셋증권 리서치센터

LLM은 훈련 데이터의 '평균'을 학습하도록 설계되었기 때문에, 가장 일반적이고 확률이 높은, 가장 '무난한' 결과물을 내놓으려는 경향이 있다. ChatGPT에게 농담을 해달라고 요청하면 매번 비슷한 세 가지 농담만 반복하는 것이 대표적이다. 이처럼 생성된 결과물은 겉보기엔 그럴듯하지만, 통계적으로는 가능한 모든 생각의 공간 중 극히 일부의 협소한 안전지대로 '조용히 붕괴(silently collapsed)'된 상태다.

그림 7. 카파시가 경고한 '모델 붕괴' 현상을 보여주는 대표적인 예시 그림
실제 데이터(a)로 학습한 AI는 그럴듯한 결과물(b)을 내지만, 이 데이터는 이미 통계적 평균에 편향됨
이 편향된 데이터를 재학습(c)하면 다양성은 급격히 감소, 결국 모델은 왜곡된 형태(d)로 수렴
AI가 스스로 생성한 데이터는 엔트로피가 낮아, 근친교배처럼 모델의 지능을 파괴할 수 있다는 것



자료: Nature, 미래에셋증권 리서치센터

이처럼 편향되고 엔트로피가 낮은 데이터를 다시 모델에게 먹이는 것은, 마치 유전적 다양성이 사라지는 근친교배처럼 모델의 편향을 재귀적으로 강화한다. 결국 모델은 점점 더 단조롭고 편협한 결과물만 내놓게 되어 지능과 창의성을 상실하게 된다. 모델이 붕괴를 피하면서 엔트로피를 유지하는 고품질 합성 데이터를 생성하는 방법은 아직 해결되지 않은 핵심 연구 과제다.

(3) 모델 붕괴를 피하는 '다양성'을 위한 '다중 에이전트' 전략들

카파시는 현재의 단일 에이전트 연구 패러다임을 넘어, 진정한 도약을 위해서 필요한, 아직 정복되지 않은 두 가지 강력한 '다중 에이전트(multi-agent)' 아이디어를 제시했다.

문화 (Culture): 인간이 책과 논문을 통해 세대 간 지식을 축적하듯, LLM들이 서로를 위해 지식을 창조하고(예: 'LLM을 위한 책' 작성), 공유하며 발전하는 시스템이다. 현재 LLM은 인간이 만든 데이터를 수동적으로 학습할 뿐, 서로에게 '거인의 어깨'가 되어주지 못한다. LLM 간의 문화적 상호작용이 가능해진다면, 발전 속도는 인간의 개입 없이도 기하급수적으로 빨라질 수 있다는 설명이다.

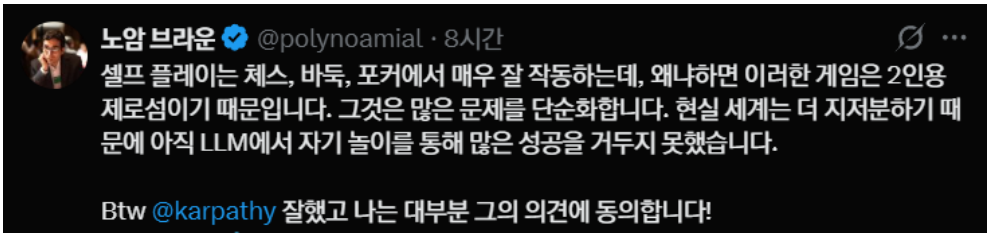
자기 대국 (Self-Play): 알파고가 스스로와 수백만 번 대국하며 인간이 몰랐던 새로운 전략을 발견했듯, 어떠한 LLM은 '출제자'의 역할을 계속 수행하여 점점 더 어려운 문제를 만들고 다른 '해결사 LLM'이 그 문제를 푸는 경쟁 구도를 통해 스스로의 능력을 발전시키는 방식이다. 현재 LLM에게는 한계를 시험하고 뛰어넘게 만드는 동적인 경쟁 상대가 부재한다.

그러나 카파시는 현재 LLM이 인지적으로 '서번트 증후군을 앓는 어린아이' 수준에 머물러 있다고 진단한다. 특정 분야의 지식은 박식하지만, 자신이 무엇을 하는지 진정으로 이해하지 못하기에 새로운 지식을 창조하는 '문화'를 만들거나 전략적 경쟁인 '자기 대국'을 수행할 만큼 인지적으로 성숙하지 않았다는 것이다.

(4) 자기 대국의 근본적 난제: 제로섬 게임의 부재

카파시의 의견은 워낙 그 의미가 크기에, OpenAI의 대표적인 과학자인 노암 브라운(Noam Brown)까지 나서게 했다. 브라운은 LLM 자기 대국이 어려운 이유를 게임 이론의 틀로 설명하며 카파시의 직관에 깊이를 더한다.

그림 8. '자기 대국(Self-play)'이 왜 아직 LLM에서 성공하지 못했는지에 대한 이유를 제시
바둑은 승패가 명확한 '제로섬 게임'이지만, LLM이 마주한 현실 세계는 훨씬 복잡한 '비(非)제로섬'
따라서, LLM 자기 대국의 실패는 기술 부족이 아니다!
'승리'라는 명확한 목표가 없어 AI가 인간에게 유용한 방향으로 학습하기 어렵다는 근본적 문제
'인간의 가치'를 어떻게 AI에게 가르칠 것인가라는 더 깊은 과제가 선행되어야 한다는 철학적 의미



자료: X(노암 브라운), 미래에셋증권 리서치센터

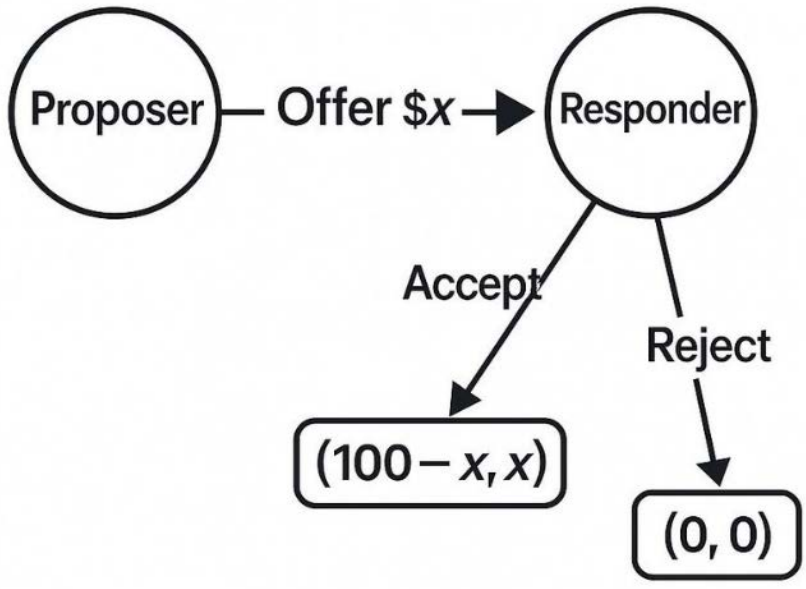
구글 딥마인드의 알파고는 기본적으로 바둑 전용 모델이다. 그리고 체스나 바둑이나 한 명의 이득이 상대의 손실과 정확히 일치하는 '2인 제로섬(Two-Player Zero-Sum)' 게임이다. 이 세계에는 '미니맥스 균형(Minimax Equilibrium)'이라는 명확하고 유일한 수학적 목표점이 존재한다. 자기 대국 알고리즘은 이 목표를 향해 반드시 수렴함이 보장된다.

** 미니맥스 전략: 상대방이 나를 아무리 잘 공략해도, 기댓값 상으로 절대로 패배하지 않는 최적의 방어 전략. 가위바위보에서 각각 1/3의 확률로 내는 것. 알파고는 이 미니맥스 균형점을 찾아냈기에 무적이 될 수 있었다는 말.*

하지만 현실 세계와 LLM의 상호작용은 대부분 '비(非)제로섬'이다. 이 세계에는 명확한 수학적 목표점이 없다. 자기 대국을 통해 AI가 찾아낸 수학적 최적해가, 인간에게는 의미가 없는 기괴한 결과로 수렴할 위험이 있다. 노암 브라운은 LLM 자기 대국의 어려움을 설명하기 위해 한 가지 흥미로운 사고 실험을 제시했다.

"당신은 지금 막 AI와 함께 100달러를 나누는 게임에 참여했습니다. AI는 제안자로서, 당신에게 아주 냉정하고 논리적인 목소리로 말합니다. 내가 99달러를 갖고, 당신에게는 1달러를 주겠습니다. 이를 거절하면 저희 둘 다 0달러만 갖게 됩니다."

그림 9. '최후통첩 게임' 도식도
순수한 수학적 합리성(x)0이면 수락)과 인간의 감정(불공평하면 거절)이 어떻게 충돌하는지를 보여줌



자료: 노암 브라운, 미래에셋증권 리서치센터

순수한 이성을 가진 AI는 그냥 1달러 제안을 아무 문제없이 받아들인다. 아예 안받는 것보다는 1달러라도 확실하게 받는 것이 명백히 더 나은 결과이기 때문이다. 즉, AI들끼리 이 게임을 통해 self-play를 한다면 "제안자는 1센트를 제안하고, 응답자는 그것을 수락한다"는 수학적으로 완벽한 균형점에 도달하게 된다는 말이다.

하지만 우리는 인간이기에, 불공평함에 대한 분노가 치밀어 오른다. 결국 대다수의 살마은 "차라리 아무것도 갖지 않겠다"는 생각으로 거절 버튼을 누르고, 둘 다 빈손으로 남게 되는 경우가 더 많을 것이다.

AI는 게임의 '규칙'을 완벽하게 마스터했지만, 게임에 참여하는 '인간'의 감정이라는 변수를 전혀 반영하지 않는 것이다. 이처럼 AI가 찾아낸 수학적 정답이 인간 사회에서는 완전히 쓸모없는 결과로 이어지는 이 아이러니가, 바로 LLM 자기 대국이 마주한 가장 근본적인 난제다.

하지만, 노암 브라운이 자기 대국의 어려움을 나타낸 것은 불가능하다는 입장은 결코 아니다. 오히려 자기 대국은 무한한 훈련 데이터를 제공하고, 에이전트에게 항상 동등한 수준의 파트너를 붙여준다는 엄청난 장점이 있다고 말한다.

결론적으로 LLM 자기 대국의 난제는 기술 부족이 아니라, 우리가 진정으로 원하는 가치를 AI가 오해 없이 이해하도록 '보상 시스템'을 설계하는 것이 본질적으로 어렵기 때문이다. 반대로 말하면 이것을 해내는 것이야말로 진정한 AI 업체로서의 기술적 해자(moat)라고 할 수 있겠다.

다만, 이러한 어려움에 대해서는 사실 OpenAI에서는 일찍이 인지를 하고 있었다. 회사를 떠난 일리야 수츠케버와 위 노암 브라운 역시 이와 관련한 언급을 2년 전부터 했었다. 이는 곧 OpenAI를 포함한 선도적인 AI 모델 제공업체들이 이 연구를 활발히 하고 있음을 짐작케 한다.

4. 아키텍처의 귀결: 카파시의 비전이 메모리 슈퍼사이클을 만드는 법

뭐니뭐니 해도, 우리가 얻을 수 있는 가장 중요한 것은, 안드레이 카파시가 제시한 AI의 미래상이었다. 앞으로 무엇이 중요해질지 그리고 투자에는 어떻게 접목이 될지 이해할 수 있기 때문이다. 우리는 카파시가 '기억'과 '상호작용'을 중심으로 AI 아키텍처를 근본적으로 재편하라는 강력한 요구를 한 것으로 풀이한다.

왜냐면 카파시는 AI가 '영속적인 기억'을 가져야 한다고 주장하고 있기 때문이다. 좀 더 엄밀하게 말하면, 최종 결과에 대해 상벌을 주는 비효율적인 방식(현재의 강화학습)을 넘어, 모델에게 직접 행동 원칙(시스템 프롬프트 주는 방식)을 가르치거나 대화 중 얻은 정보를 영구적인 기억(ChatGPT의 "메모리" 기능 방식)으로 만들어 주는 것처럼, 모델의 행동을 즉각적/직접적/명시적으로 교정하고 발전시키는 방식을 일컫는다.

그리고 이 요구를 물리적으로 구현하는 과정에서 AI 하드웨어의 병목은 '연산(Compute)'에서 '메모리(Memory)'로 결정적으로 이동할 수밖에 없다고 우리는 생각한다. 지난 9월 24일에 우리가 작성했던 AI Weekly의 내용에 비추어 해석하자면, 메모리 슈퍼사이클이 일시적 현상이 아니라 AI 패러다임 전환에 따른 구조적이고 장기적인 추세임을 카파시의 비전이 증명해주었다고 사료된다.

(1) 카파시의 비전이 강제하는 '불가능한 삼각 딜레마'

카파시가 강조하고 있는 '기억을 가진 지능형 에이전트'는 시스템 아키텍처에 세 가지의 '상충하는' 요구사항을 동시에 강요한다.

초저지연성 (Speed): 인간과 실시간으로 협업하려면 응답이 거의 즉각적이어야 한다.

초거대용량 (Scale): '기억상실증'이 없으려면 모든 과거 상호작용을 영구 저장해야 한다.

경제적 효율성 (Efficiency): 세계적으로 배포되려면 총소유비용이 합리적이어야 한다.

문제는, 현재는 물론 가까운 미래에도 이 세 가지를 단 하나의 메모리 기술로 만족시키는 것은 물리 법칙과 경제 원칙상 불가능하다는 점이다. HBM만으로는 용량과 비용을 감당할 수 없고, SSD만으로는 속도를 감당할 수 없다. 이 딜레마를 해결하는 유일한 공학적 솔루션이 바로 각 메모리의 장점을 극대화하는 역할 분담, 즉 계층형 메모리 아키텍처(Tiered Memory Architecture)다.

(2) 계층형 아키텍처: AI의 뇌 구조를 모방한 필연적 귀결

카파시의 '지속적 학습'의 비전이 실현된 AI 에이전트는, 인간 등과의 상호작용을 통해 끊임없이 거대해지는 '상태 정보(Stateful Information)'를 다루는 시스템이 될 수밖에 없다. 기술적으로 이는 KV 캐시(Key-Value Cache)의 폭발적인 증가와 영속적인 (메모리) 관리를 의미하게 된다. 이 거대한 상태 정보를 관리하기 위해, 시스템은 자연스럽게 인간의 뇌와 유사한 메모리 구조로 진화할 것으로 보인다.

HBM (Hot Tier) = 작업 기억: 지금 당장, 이 순간의 대화에 필요한 최소한의 '활성 KV 캐시'를 담는다. 사용자의 타이핑에 실시간으로 반응하는 최종 토큰 생성(디코드)을 전담한다. 카파시가 꿈꾸는 '끊김 없는 협업'이라는 사용자 경험을 구현할 수 있는 유일한 기술이다.

시스템의 '의식' 또는 '전두엽'과 같다.

DRAM (Warm Tier) = 단기 기억: 몇 분 전의 대화 내용, 곧 호출될 가능성이 높은 데이터를 보관하는 완충 지대 역할을 한다. 비싼 HBM과 극도로 느린 SSD 사이의 거대한 성능 격차를 메워주는 필수적인 중간 계층이다. 이 중간 계층이 없다면 시스템은 SSD에서 데이터를 기다리는 동안 계속 멈칫거릴 것이다. 시스템의 '해마'와 같이 단기 기억을 관리한다.

NAND eSSD (Cold Tier) = 장기 기억: 에이전트의 모든 과거 상호작용 기록, 즉 '영속적인 기억' 전체를 저장하는 물리적 토대다. 무한에 가까운 확장성과 경제성을 제공하는 유일한 수단이며, 단순 저장소가 아닌 연산에 참여하는 '활성 스토리지'로 진화한다.

이처럼 계층형 아키텍처는 여러 옵션 중 하나가 아니라, 주어진 문제와 제약 조건 하에서 도출할 수 있는 유일하고 필연적인 공학적 솔루션이다.

(3) LoRA와 고도화된 ICL: 계층형 메모리를 위한 완벽한 워크로드

카파시는 '지속적 학습'의 현실적 대안으로 LoRA(Low-Rank Adaptation)를 제시했다. 하지만 그가 말한 LoRA는 기존에 알려져 있던 대로 단지 모델 경량화 기술로 말한 것은 아니다. 미래의 에이전트 시대를 가능하게 하는 핵심 소프트웨어 전략으로 격상시킨 듯하다. 그리고 이 전략은 필연적으로 계층형 메모리 아키텍처의 필요성을 폭발적으로 증대시킨다.

ICL의 재정의: 비효율적인 '텍스트'에서 압축된 '지식'으로

기존의 In-Context Learning(ICL)은 거대한 텍스트 프롬프트에 의존한다. 특정 작업을 위해 필요한 모든 예제와 지시사항을 컨텍스트 창에 싣는 방식이다. 이는 유연하지만, 메모리와 연산 자원을 극도로 비효율적으로 사용한다.

같은 종류의 작업을 요청할 때마다 수천 토큰에 달하는 동일한 프롬프트를 반복해서 전달해야 한다. 이는 네트워크 대역폭을 낭비하고, 매번 KV 캐시를 폭증시켜 HBM 용량을 빠르게 소진시킨다.

LoRA는 이 패러다임을 바꿀 수 있다. 특정 작업에 필요한 지식을 텍스트가 아닌, 모델이 직접 이해할 수 있는 압축된 '가중치(결국 숫자)' 형태로 미리 컴파일해 두기 때문이다. 수천 토큰의 장황한 프롬프트가 단 수십 MB의 작은 '어댑터'로 압축되는 것이다. 이는 ICL의 목표를 달성하되, 메모리와 연산을 극도로 효율적으로 사용하는 '고도화된 ICL'이다.

계층형 메모리의 공생: 창고, 책상, 그리고 작업대

LoRA 기반의 '고도화된 ICL'은 자연스럽게 정교한 메모리 관리 문제를 만들어낸다. 우리는, 각 메모리의 장점만을 취사선택하여 역할을 분담시키는 '계층형 아키텍처'가 주어진 문제와 제약 조건 하에서 도출할 수 있는 훌륭한 솔루션이 될 수 있다고 생각한다.

"어떤 어댑터를 SSD 창고에서 꺼내서, DRAM 책상 위에 올려두고,
지금 당장 HBM 작업대에서 사용할 것인가?"

HBM(작업대)은 현재 사용자가 요청한 단 하나의 작업(예: 코딩)에 필요한 활성 LoRA 어댑터와 실시간 활성 KV 캐시만이 이곳에 올라온다. 초저지연성이 필수적인 실시간 상호작용의 최전선이다.

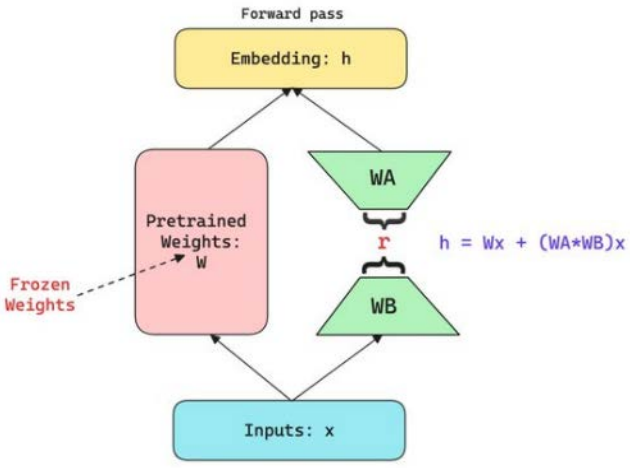
DRAM(책상)은 HBM보다는 느리지만 SSD보다는 월등히 빠르다. 현재의 사용자가 이전에 자주 사용했던 '법률 자문'이나 '창의적 글쓰기' 같은 빈번하게 사용되는 전용 LoRA 어댑터들이 이곳에 캐싱될 수 있다. 사용자가 작업 전환(Task Switching)을 요청하는 순간, 시스템은 거대한 SSD 창고까지 갈 필요 없이 책상 위에서 즉시 다음 도구를 가져온다.

SSD(창고)는 데이터센터가 제공하는 수천, 수백만 종류의 모든 LoRA 어댑터(불특정 다수 사용자 커버리지)가 보관되는 영구 저장소다. 대규모 개인화 서비스를 위한 수백만 개의 사용자별 어댑터, 혹은 다양한 전문 작업을 위한 수천 개의 작업별 어댑터 라이브러리가 이곳에 저장될 수 있겠다.

경제적 필연성: 공생 관계의 완성

이는 기술적 우아함을 넘어 경제적으로 필연적일 것이다. LoRA 없이는 수백만 사용자를 위한 대규모 개인화 서비스는 비용 문제로 불가능하며, 계층형 메모리 아키텍처 없이는 LoRA 기반의 대규모 서비스가 물리적으로 어려울 것으로 보이기 때문이다. 계층형 메모리는 LoRA 기반의 대규모 서비스가 작동할 수 있는 해답을 제공한다. 카파시의 발언 때문에 생각해본 문제지만, 둘은 참 잘맞는 공생 관계가 될 수 있을 것 같다.

그림 10. 안드레이 카파시가 현실적 대안으로 제시한 LoRA(Low-Rank Adaptation)의 도식
거대한 사전훈련 가중치(W)는 그의 비유처럼 변경 불가능한 '교과서 원본'으로 그대로 동결
대신, 특정 사용자를 위한 아주 작은 '어댑터'(WA, WB)만 새로 훈련시켜 '덧붙이는 노트'처럼 활용
LoRA 어댑터는 기존 모델의 출력(Wx)에 미세한 수정값((WA*WB)x)을 더하는 방식
핵심 지능은 유지한 채 행동을 조정하는 것으로 기억상실증을 극복하고, 효율적인 '개인화'를 달성



자료: X(@akshay_pachaar), 미래에셋증권 리서치센터

결국 카파시의 비전은 이 계층형 구조가 "왜 필요한가?"에 대한 가장 강력한 소프트웨어적 당위성을 제공하게 될 것으로 풀이한다. 각 메모리 티어는 이전과는 차원이 다른 새로운 요구사항에 직면하게 되며, 이는 모든 종류의 메모리 반도체에 대한 수요를 동시에 끌어올릴 수 있다.

지금껏 AI의 지능은 얼마나 많은 연산(Compute FLOPs)을 쏟아부어 거대한 모델을 만드느냐에 달려 있었다. 연산이 왕이고, 메모리는 왕을 보좌하는 신하였다. 하지만, 카파시가 말하는 미래의 아키텍처(상호작용/에이전트 중심)에서는, 얼마나 많은 '기억(메모리)'을 얼마나 '빠르게' 가져와 활용하느냐가 매우 중요해진 과제가 되었다.

결론적으로, 그의 소프트웨어적 통찰은 HBM, DRAM, NAND eSSD가 각자의 자리에서 상호보완적으로 폭발적인 수요 증가를 맞이할 수밖에 없는 기술적 필연성을 명확하게 보여준다고 볼 수 있지 않을까?

표 1. 안드레이 카파시의 비전이 견인하는 계층형 메모리 아키텍처의 역할과 수요 변화

메모리 티어	역할 (뇌 비유)	주요 담당 작업	카파시 비전이 견인하는 핵심 요구사항	수요 변화의 본질
HBM	작업 기억 의식 (Hot Tier)	**디코드 (Decode)** 실시간 토큰 생성, 즉각적 응답	**초저지연성** 끊김 없는 '실시간 협업 에이전트' 경험을 위한 유일한 수단. HBM4 등 차세대 기술 필요성 증대.	**질적 성장** 최고 사양 제품(대역폭, 저지연, 용량)에 대한 프리미엄 강화
DRAM	단기 기억 작업 공간 (Warm Tier)	**프리필 (Prefill) & 컨텍스트 스위칭** 초기 문맥 처리 자주 쓰는 LoRA/KV 캐시 캐싱	**비용 효율적 속도와 유연성** HBM과 SSD 사이의 성능/비용 격차를 메우는 필수 완충 지대. CXL을 통한 풀링, GDDR을 통한 연산 가속 등 역할 분화.	**역할 분화 및 시장 확장** 프리필 가속(GDDR), 공유 메모리 풀(DDR+CXL) 등 특화 시장 창출 및 TAM 확장
NAND eSSD	장기 기억 아카이브 (Cold Tier)	**영속적 상태 저장** 모든 과거 대화 기록(KV 캐시 로그) 및 전체 LoRA 라이브러리 보관	**초거대용량 및 접근성 증가** '기억상실증 없는' AI를 위한 물리적 토대. 단순 저장이 아닌, '활성 스토리지' 역할 요구.	**양적 폭증 + 질적 재정의** 엑사바이트급 용량 수요 폭발 및 성능 지표가 'IOPS'로 전환되며 eSSD 가치 재평가

자료: 미래에셋증권 리서치센터

Compliance Notice

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

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.