그림: ChatGPT(Age of Mythos를 빨리 끝내기 위한 OpenAI의 지능)

MIRAE ASSET 미래에셋증권

AI Hot Issue

GPT-5.5 출시… O태식이 돌아왔구나

한종목
chongmok.han@miraeasset.com

CONTENTS

Highlight of the Week 3
샌드백 신세였던 OpenAI, 최고 지능으로 돌아오다 3

I. GPT-5.5 4
1. GPT-5.5의 실체: 사전학습의 귀환, 그리고 Blackwell 4
2. 샌드백에서 프런티어로: OpenAI의 피봇 성공 10
3. Anthropic의 내러티브 붕괴: 컴퓨팅이 발목을 잡다 14
4. 진짜 수혜자는 아마존, 그리고 엔비디아 18
5. 결론: Age of Mythology에서 Age of Compute로 23

Highlight of the Week

샌드백 신세였던 OpenAI, 최고 지능으로 돌아오다

현지시각 4월 23일, OpenAI가 차세대 프론티어 모델 GPT-5.5(코드명 Spud)를 출시했습니다. 출시 초기임에도 사용자들의 반응은 이례적으로 일치합니다. 모든 리뷰를 한 문장으로 요약하면 “현재까지 가장 높은 지능의 모델이, 동시에 꽤 비용 및 속도 효율적인 형태로, 대중에 공개되어 출시되었다”는 점입니다. 이는 Anthropic의 Mythos가 Project Glasswing 연합사 50여 곳에만 제한적으로 크레딧을 배분하는 ‘컴퓨팅 배급제’ 형태로 운영되는 것과 반대에 위치한다는 점에서 유의미합니다. GPT-5.5 출시 핵심은 3가지입니다.

첫째, GPT-5.5는 사후학습이나 테스트타임 영역이 아니라 사전학습(Pre-training)을 처음부터 다시 시작한 2년 만의 완전한 새 모델입니다. Mythos와 같은 철학의 흐름입니다. 이것이 가능했던 이유는 결국 컴퓨팅입니다. 엔비디아 Blackwell(GB200/GB300 NVL72)로 업그레이드된 훈련 인프라 위에서 처음부터 끝까지 지난 2년 동안 재설계된, 아키텍처 레벨에서 완전히 다른 지능의 모델이라 할 수 있습니다. 실제로 Aidan McLaughlin(OpenAI 연구원)은 출시 직후 "big model smell is so back(큰 모델 특유의 체취가 돌아왔다)"이라는 의미심장한 한 줄을 남겼습니다. **사전학습 스케일링은 죽지 않았습니다**.

둘째, **OpenAI는 지난 3개월간 시장의 샌드백이었으나, 이번 출시로 피봇에 명확히 성공한 모양새입니다**. 'Anthropic이 이겼고 OpenAI는 끝났다'는 내러티브가 너무나도 빨리 뒤집혔습니다. Artificial Analysis Intelligence Index에서 GPT-5.5(xHigh)는 60점으로 Claude Opus 4.7(57점), Gemini 3.1 Pro(57점)와의 3자 동률을 깨고 단독 1위에 올랐습니다. ARC-AGI-2에서는 85.0%로 SoTA를 갈아치웠습니다. 더 중요한 것은 기술적 경로의 다음 단계가 명확히 보인다는 점입니다. 출시 직후 여러 전문가들은 "GPT-5.5가 새 사전학습 모델의 초기 RL 체크포인트, 즉 o1-preview 급에 가깝다"고 평가했습니다. 즉, 이번 **모델은 끝이 아니라 GPT-5.6, GPT-5.7로 이어질 주기적 마이너 업데이트의 출발점이며, 나아가 GPT-6의 전조라는 해석이 지배적입니다.**

셋째, Anthropic의 내러티브가 급속도로 약화되고 있습니다. Mythos Preview가 원대한 성능 때문에 비공개가 되었다는 공식 설명이 있었으나, 사실은 '그 정도 크기의 모델(10T 모델 크기)을 서빙할 물리적 역량이 부족했기 때문'이라는 해석이 힘을 얻고 있습니다. 더 큰 문제는 지난 한 달간의 Anthropic이 보여준 일련의 행보입니다. Pro 플랜에서 Claude Code 기능 은밀 삭제, 신규 프로슈머 대상 2배 가격 A/B 테스트, Opus 4.6의 inference 컴퓨팅 의혹, 그리고 반응이 극히 나쁜 Opus 4.7 출시까지. 이 **모든 신호가 하나의 방향을 가리키고 있습니다. Anthropic은 컴퓨팅이 극도로 부족하다는 것입니다. 이 압박이 Anthropic을 아마존의 품으로 더욱 강하게 밀어넣고 있으며, 아마존은 이 상황의 구조적 수혜자입니다.**

결국 결론은 단순합니다. 컴퓨팅을 먼저, 더 많이, 더 안정적으로 확보한 쪽이 이 전쟁에서 최후의 수혜자가 됩니다. **OpenAI가 1년 전 "비이성적"이라고 조롱받으며 집행한 그 공격적 컴퓨팅 확보가 지금 결실로 돌아오고 있음을 유념해야 합니다.**

I. GPT-5.5

1. GPT-5.5의 실체: 사전학습의 귀환, 그리고 Blackwell

(1) "5 시리즈의 진화가 아니라, 6 시리즈의 전조"

OpenAI는 GPT-5.5(코드네임 Spud)를 "a new class of intelligence for real work(실제 업무를 위한 새로운 클래스의 지능)"이라고 정의했다. 여기서 "a new class"라는 표현은 GPT-5.4에서 GPT-5.5로의 점진적 개선이 아니라, 아키텍처와 훈련 파이프라인의 근본적 재설계를 뜻한다.

이 모델은 어떤 의미에서 GPT-5 시리즈의 마지막 단계가 아니라, 6 시리즈의 전조에 가깝다. 이번 모델이 기록한 벤치마크 점수의 증분은 이전 모델과 비교해 상대적으로 완만하지만, 실제 사용에서 느끼는 점프는 훨씬 크기 때문이다. 이는 GPT-4o → GPT-5 전환 때 봤던 '벤치마크 점프 대비 실사용 점프'의 비대칭성과 닮아 있다. 그래서 GPT-6의 훌륭한 토대라는 평가까지 나오는 이유다.

GPT-5.5를 사전에 맛본 사용자들의 정성적 리뷰도 이를 뒷받침한다. Dan Shipper(스타트업 Every의 CEO)는 GPT-5.5를 자사의 Senior Engineer Benchmark에서 62/100점을 받은 모델이라고 평가했다. Anthropic의 Opus 4.7이 동일 벤치마크에서 33/100점에 그쳤다는 점을 고려하면, 코딩 분야에서 대중에게 공개된 모델 중 가장 우수한 것으로 사료된다.

(2) 벤치마크 성적: Mythos와의 격차는 크게 좁혀짐

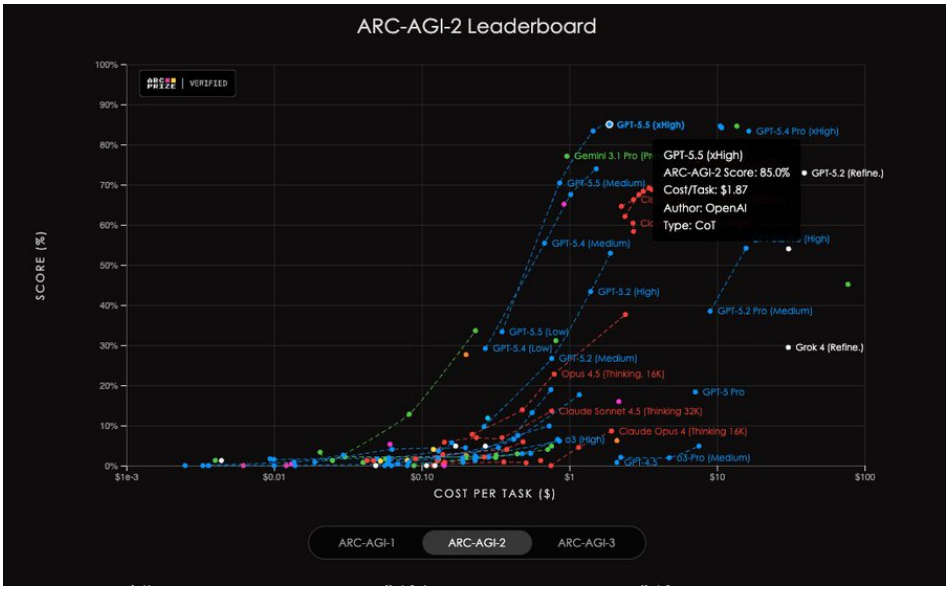
이쯤에서 이번 달에 나온 비밀스러운 신화 같은 모델인, Anthropic의 Mythos Preview(비공개, 극히 제한적인 접근권한)와 GPT-5.5의 벤치마크 비교는 중요하다. 이유는 단순하다. 현재 존재하는 모든 공개 + 비공개 프런티어 모델 중 Mythos가 여전히 최상위권에 위치해 있기 때문이다. 하지만 GPT-5.5는 그 Mythos와의 격차를 극적으로 좁혔다.

표 1. OpenAI의 GPT-5.5와 Anthropic의 Claude Opus 4.7, 그리고 Mythos 벤치마크 점수 비교

벤치마크 종류	GPT-5.5	Claude Opus 4.7	Claude Mythos (비공개)
Terminal-Bench 2.0	82.70%	69.40%	82.00%
OSWorld-Verified	78.70%	78.00%	79.60%
BrowseComp	84.40%	79.30%	86.90%
CyberGym	81.80%	73.10%	83.10%
GPQA Diamond	93.60%	94.20%	94.60%
HLE (with tools)	52.20%	54.70%	64.70%
SWE-Bench Pro	58.60%	64.30%	77.80%
ARC-AGI-2	85.0% ($1.87/task)	75.80%	-

자료: OpenAI, Anthropic, ARC Prize, 미래에셋증권 리서치센터

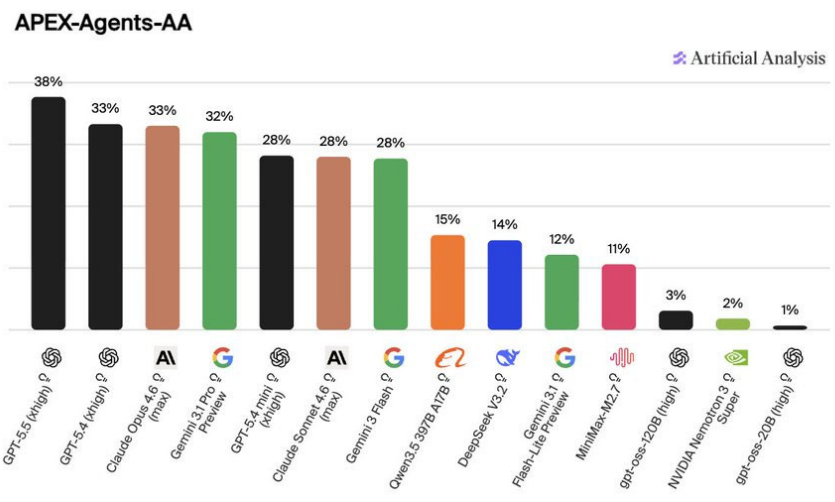
그림 1. ARC-AGI-2 Leaderboard: GPT-5.5(xHigh)의 85.0% SoTA 달성
ARC-AGI는 훈련 데이터에서 본 적 없는 새로운 추상 규칙을 풀어내는 능력을 측정하므로, 단순 암기가 아닌 '일반화 지능'의 지표로 평가됨
기존 1위였던 GPT-5.4 Pro(xHigh, 약 80%)와 Gemini 3.1 Pro(약 73~75%)를 두 자릿수 차이로 제쳤고 Opus 4.7(75.8%)도 큰 폭으로 앞섬



자료: ARC Prize, 미래에셋증권 리서치센터

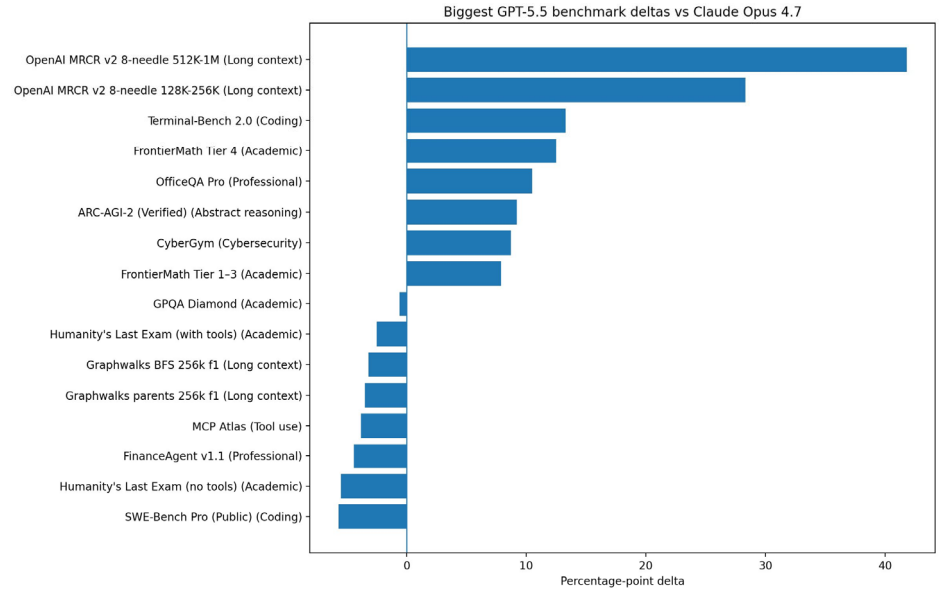
GPT-5.5는 Claude Opus 4.7 대비로는 대부분의 벤치마크에서 우세하거나 동률이고, 비공개 모델인 Mythos와 비교해도 Terminal-Bench 2.0, OSWorld-Verified, CyberGym 등에서는 거의 동률이거나 근접했다. Pietro Schirano(스타트업 MagicPath의 CEO)의 논평이 이 상황을 가장 잘 압축한다. "Anthropic의 gated Mythos가 CyberGym에서 83%를 기록했다. OpenAI는 방금 GPT-5.5를 82%에 출시했다. 차이는, GPT-5.5는 사람들이 실제로 쓸 수 있다는 것이다."

그림 2. Mercor가 구축한 Artificial Analysis APEX-Agents-AA — 장기 전문 업무 에이전트 랭킹
GPT-5.5(xhigh)가 38%로 단독 1위, 2위 GPT-5.4(xhigh)와 Claude Opus 4.6(max)이 33% 공동
5% 포인트의 차이는 GPT-5.5가 만들어낸 '장기 과제(long-horizon) 지속성' 능력을 의미
이는 단발성 답변 품질이 아닌 '하루종일 일하는 에이전트'의 지표로 기업용 실사용 사례와 직접 연결



자료: Artificial Analysis, 미래에셋증권 리서치센터

그림 3. GPT-5.5 vs Claude Opus 4.7 벤치마크 델타 분석
X축에 %p 단위로 GPT-5.5가 Opus 4.7 대비 앞서는 영역(양수)과 뒤처지는 영역(음수)을 시각화
GPT-5.5 압승 영역을 보면 초장문 컨텍스트, 에이전틱 코딩과 장기 추론에서의 일관된 우위
GPT-5.5 열위를 보이는 부분은 SWE-Bench인데, 이는 Mythos급 지능 미보유에서 오는 한계
그러나 승리의 크기(최대 +42p)가 패배의 크기(최대 -6p)를 압도한다는 것이 포인트



자료: Simon Smith, 미래에셋증권 리서치센터

소프트웨어 부문에서 가장 중요한 벤치마크 중 하나인 SWE-Bench Pro에서의 약점(58.6% vs Opus 4.7의 64.3%)은 실제 존재하는 갭이기는 하다. 그럼에도, 여러 써드파티 분석가들은 이를 '노이즈 또는 spiky intelligence(들쑥날쑥한 지능 분포)'로 해석하는 편이다.

오히려 주목할 지점은, Anthropic이 이 벤치마크에서 'memorization(암기)' 정황이 발견되었다고 OpenAI가 "친절하게" 각주로 지적한 부분이다. 모델 평가의 공정성 자체가 현재의 프런티어 경쟁에서 정치화되고 있음을 보여주는 대목이다.

그림 4. OpenAI가 자사 블로그 Evaluations 섹션에서 공개한 공식 코딩 벤치마크 표
GPT-5.5가 Terminal-Bench 2.0(82.7%)과 Expert-SWE Internal(73.1%)에서 앞서지만, SWE-Bench Pro(58.6%)에서는 Claude Opus 4.7(64.3%)에 5.7%p 뒤처짐을 스스로 투명하게 공개
진짜 핵심은 각주로, 해당 벤치마크 결과 자체의 신뢰성에 의문을 제기했다는 것을 적은 것.
이는 Anthropic이 해당 벤치마크에서 우위를 보이는 것이 '진짜 코딩 능력' 때문인지 '훈련 데이터 누출로 인한 암기' 때문인지 판단하기 어렵다는 기술적 경고

Evaluations

Coding

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 Pro
SWE-Bench Pro (Public)*	58.6%	57.7%	-	-	64.3%	54.2%
Terminal-Bench 2.0	82.7%	75.1%	-	-	69.4%	68.5%
Expert-SWE (Internal)	73.1%	68.5%	-	-	-	-

*Labs have noted evidence of memorization on this eval

자료: OpenAI, 미래에셋증권 리서치센터

여기서 알아야 할 점은 벤치마크라는 정량적 평가에 너무 매몰될 필요는 없고 결국에는 유저들의 정성적 평가가 중요하다는 것, 그리고 그 유저들을 많이 끌어안을 수 있을 정도의 효율성, 즉 '토큰 당 지능', '달러 당 지능'이 점점 더 중요하다는 점이다.

(3) 토큰 효율성: "비싸졌지만, 훨씬 덜 쓴다"

GPT-5.5의 또 다른 특징은 지능의 상승분에 비해 훨씬 공격적인 토큰 효율성 개선을 보여준다는 점이다. API 가격은 입력 100만 토큰당 $5, 출력 100만 토큰당 $30로, GPT-5.4 대비 per-token 기준으로는 약 2배 비싸졌다.

그러나 Artificial Analysis의 평가에 따르면, 동일한 Intelligence Index 평가를 수행하는데 GPT-5.5는 GPT-5.4 대비 약 40%의 토큰을 덜 사용한다. 결과적으로 벤치마크를 평가하는데 들어간 비용 증가는 약 +20%에 그쳤다고 보고했다.

더 중요한 것은 경쟁 모델 대비 비용 우위다. GPT-5.5(medium effort)는 Claude Opus 4.7(max)와 Intelligence Index에서 동점을 기록하면서, 평가 비용은 약 1/4($1,200 vs $4,800)에 불과하다. GPT-5.5(low)는 Opus 4.7(Non-reasoning, high)와 동점을 비용의 절반($500 vs $1,000)으로 맞춘다. Artificial Analysis의 Coding Index에서도 GPT-5.5는 경쟁 프런티어 코딩 모델 대비 절반 비용으로 SoTA 성능을 낸다. 이 경제성이야말로 "공개 모델로서의 GPT-5.5"가 "비공개 Mythos"보다 현실적인 승자가 될 수 있는 구조적 이유로 해석된다.

Peter Welinder(OpenAI의 제품 부문 부사장)가 "GPT-5.5가 오리지날 GPT-3의 API 가격보다 2~12배 싸다는 사실이 미친 것 같다. 과학의 경이"라고 트윗한 것은, 지난 3년간 토큰 당 지능(intelligence per token)이 얼마나 빠르게 올라왔는지를 웅변한다. Noam Brown(OpenAI의 시니어 과학자)이 정확히 지적하듯, "오늘날의 AI 모델은 지능이 추론 컴퓨팅의 함수이며, 단일 숫자로 모델을 비교하는 것은 2024년 이래로는 의미가 없다. 중요한 것은 토큰 당 지능 혹은 달러 당 지능이다."

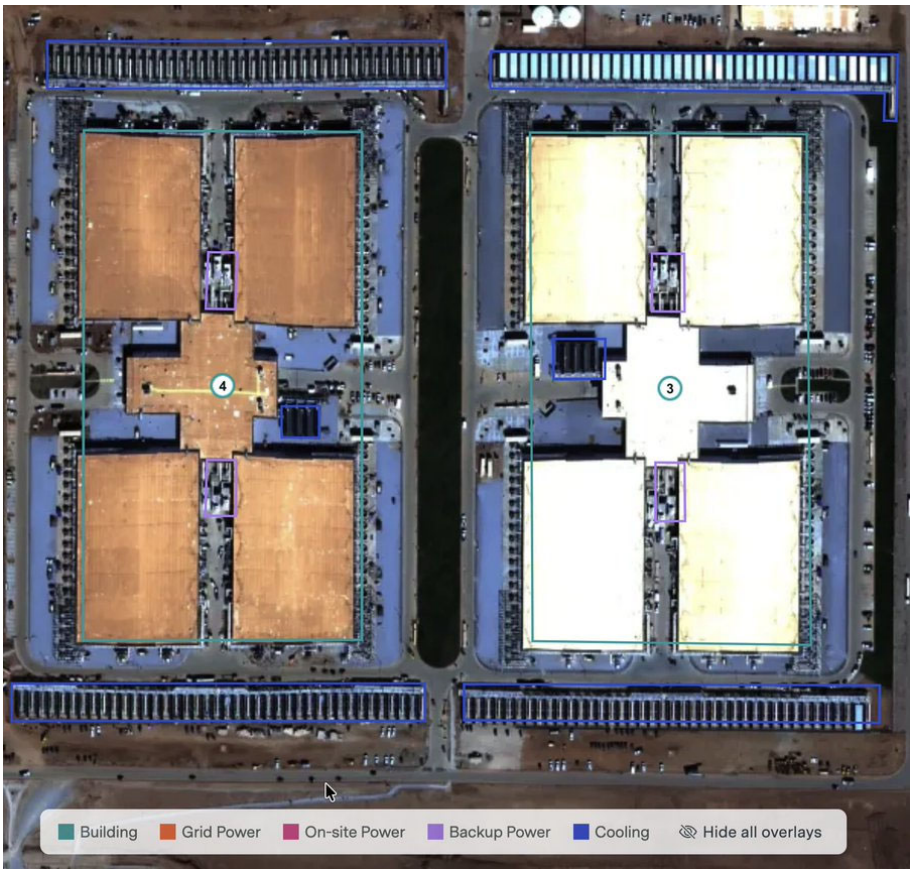
(4) 사전학습의 귀환 + Blackwell의 본격화

우리가 바라봐야 하는 가장 중요한 관찰은 이 모델이 어디서 훈련되었고, 왜 이 시점에 출시되었느냐는 점이다. OpenAI는 블로그에서 이례적으로 명시적인 표현을 썼다.

"GPT-5.5 was co-designed for, trained with, and served on 엔비디아 GB200 and GB300 NVL72 systems(GPT-5.5는 엔비디아 GB200/GB300 NVL72 시스템을 위해 공동 설계, 훈련, 서빙되었다)."

이는 단순한 파트너십 마케팅 문구가 아니다. 이번 모델은 Stargate Abilene 텍사스 캠퍼스의 2개 동(8개 동 중) 기반 10만 개 GB200 클러스터에서 훈련되었을 것으로 추정한다. 약 112,000개 GPU가 운영 중인 Abilene의 첫 완성 구간이 이번 모델 훈련의 주요 자원으로 활용된 것이다.

그림 5. 미국 텍사스 Abilene의 Stargate 캠퍼스에 위치한 빌딩들의 실시간 위성사진
Building 3번은 쿨링·백업전력·외장 완공(흰색 지붕)으로 거의 커미셔닝 단계
4번은 지붕 공사 중(주황색 노출)으로 3~4개월 늦은 상태
Epoch AI는 4월 중순 두 동 모두 가동 예상했으나 실측 결과 0.6GW 마일스톤을 5월 말로 조정, 1.2GW는 Q3→Q4로 연기… 그럼에도 "지연되더라도 실제로 지어지고 있다"는 게 포인트.
일부 동만으로도 10T급 이하 모델은 훈련 가능한 것으로 추정.



자료: Epoch AI, 미래에셋증권 리서치센터

이러한 관찰은, 최근에 발생했던 OpenAI의 다소 이상한 행보를 일거에 설명한다. OpenAI가 Sora 앱 서비스를 폐기하고 컴퓨팅을 대거 재배치한 이유가 여기에 있다고 사료된다. 애초에 사전학습을 다시 하는 모델을 만들겠다는 결정은, 엔지니어의 자부심이 아니라 "이 모델이 몰고 올 파급력과, 이 모델 위에서 추후 GPT-6를 구축 및 서비스해 나갈 향후의 로드맵"을 위한 자본 배분 선택이었다고 해석된다.

엔비디아 CEO 젠슨 황이 샘 알트만에게 보낸 축하 이메일에서의 "Fire up those Blackwells. We need more tokens!"라는 문구는 바로 이 시점의 신호를 압축한다. Blackwell 시대 및 프론티어 모델의 개막이라는 점이다.

이번 사이클의 또 하나의 숨은 주인공은 추론 속도다. OpenAI는 블로그에서 "GPT-5.5는 GPT-5.4 수준의 per-token 지연시간을 유지하면서, 훨씬 높은 지능을 낸다"고 밝혔다. 즉, 더 크고 더 똑똑한 모델을 '느려지지 않게' 서빙한다는 뜻이다. 일반적으로 큰 모델은 느리다는 업계 상식을 정면으로 깼다.

더 놀라운 것은 OpenAI가 이 속도 최적화를 만드는 과정에서 GPT-5.5 자체를 활용했다는 점이다. OpenAI 블로그에 따르면, 코딩 AI 에이전트인 Codex가 몇 주 치 운영 트래픽 패턴을 분석해 스스로 알고리즘을 작성했고, 이 한 가지 최적화만으로도 토큰 생성 속도가 20% 이상 증가했다고 주장했다. 쉽게 말해, AI 모델이 자신을 서빙하는 인프라 자체를 스스로 개선해버린 것이다. 이는 1월 Anthropic의 Cowork가 보여줬던 "재귀적 소프트웨어 개선(Self-recursive improvement)"이 이제 추론(AI 서비스 영역) 스택 레이어에서도 작동하고 있음을 의미한다.

2. 샌드백에서 프런티어로: OpenAI의 피봇 성공

(1) 한 분기도 안지나 뒤집힌 내러티브

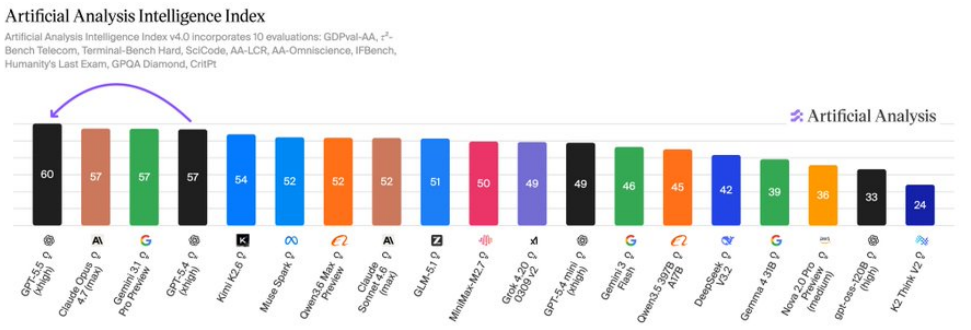
올초부터 불과 며칠 전까지의 시장 컨센서스는 명확했다. Anthropic이 2026년 2월 말 이미 Mythos를 내부적으로 보유했고, Opus 4.7까지 공개했으며, 연환산매출(ARR) 기준 $30B를 달성했다는 것에 박수 갈채가 쏟아졌다.

반면 OpenAI는 수 개월째 "제품은 좋지만 기술적 우위가 밀리고 있다"는 평가를 받으며, 사실상 'cooked(끝난 존재)'로 취급받는 구간이 있었다. 그러나 Spud 출시가 단 하루 만에 이 그림을 모두 지웠다.

Artificial Analysis의 공식 코멘트가 이 순간의 무게를 잘 압축한다. "GPT-5.5가 OpenAI를 명백한 AI 1위로 복귀시켰다. OpenAI의 새 모델은 Anthropic과 구글이 이루고 있던 3자 동률을 깨고, Artificial Analysis Intelligence Index에서 3점 차로 수위를 차지했다."

참고로 현재 기준 Intelligence Index v4.0 상위 5개 모델은 다음과 같이 재편되었다: GPT-5.5(xhigh, OpenAI) 60점, Claude Opus 4.7(max, Anthropic) 57점, Gemini 3.1 Pro Preview(구글) 57점, GPT-5.4(xhigh) 57점, Kimi K2.6(Moonshot) 54점. 위쪽 티어에서 오직 GPT-5.5만 3점이라는 격차를 보이고 있다.

그림 6. Artificial Analysis Intelligence Index v4.0 상위 모델 종합 순위
GPT-5.5(xhigh)가 60점으로 단독 1위
Claude Opus 4.7(max) 57점·Gemini 3.1 Pro 57점·GPT-5.4(xhigh) 57점의 3자 동률을 깨뜨림
단일 벤치마크 노이즈에 덜 취약한 복합 지표이므로 리더십 교체의 무게감이 더 크다고 판단됨



자료: Artificial Analysis, 미래에셋증권 리서치센터

이 모델의 성능을 드러내는 시각적 상징은 OpenAI가 GPT 시리즈 역대 모델들로 TikZ로 유니콘을 그려본 자체 비교 자료로 쉽게 이해할 수 있다. GPT-3.5는 간신히 육각형만 그렸고, GPT-4에서는 다리 4개 달린 어설픈 생명체가 나왔다. GPT-5에 와서야 유니콘 형태가 잡혔고, GPT-5.2에서 처음으로 '예쁘다'는 평가가 가능해졌다. 그리고 GPT-5.5는 마침내 긴 다리, 유려한 갈기, 뿔, 그리고 전반적 비례가 완성된 하나의 완결된 유니콘을 그려냈다. 지난 2~3년 동안 모델의 지능이 얼마나 압도적으로 좋아졌는지 나타낸다.

그림 7. TikZ Unicorn 시리즈: GPT-3.5에서 GPT-5.5까지의 시각적 진화
각 세대 모델에게 유니콘을 TikZ 코드로 그리라고 시켰을 때의 결과물 시계열 비교
GPT-5.5에 와서야 마침내 긴 다리, 유려한 갈기, 균형 잡힌 비례를 갖춘 완결된 유니콘이 등장
'공간 추론 + 벡터 코드 생성 + 미적 판단'의 통합 능력이 아키텍처 레벨에서 한 단계 상승했음

자료: X(@morqon), 미래에셋증권 리서치센터

(2) Codex의 엔비디아 전사 배포: 인류 역사상 가장 빠른 엔터프라이즈 채택

GPT-5.5의 진정한 무서움은 벤치마크 점수가 아니라 엔터프라이즈들의 채택 속도다. 공개 당일, 엔비디아의 젠슨 황은 전사에 다음과 같은 메일을 보냈다.

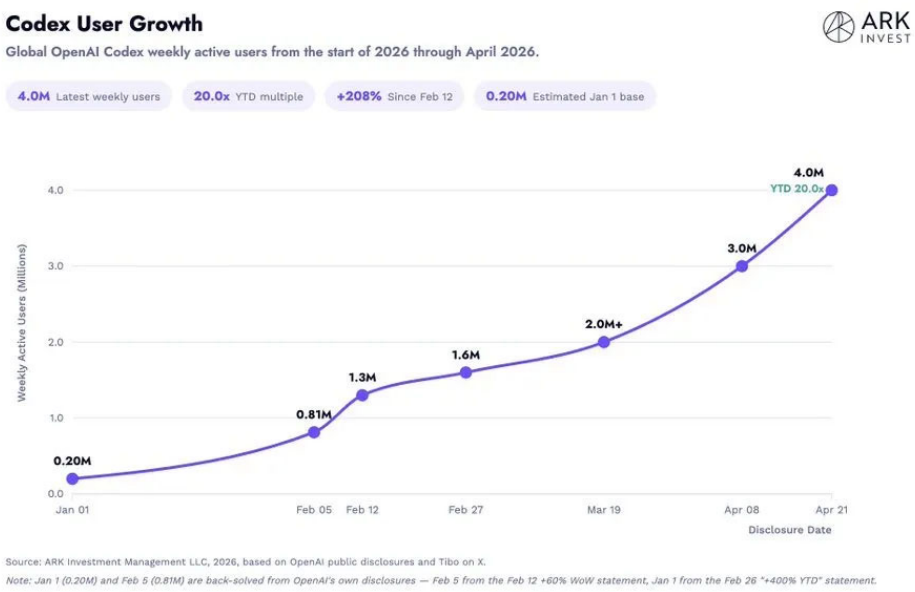
"엔비디아 임직원에게, OpenAI의 Codex(powered by GPT-5.5)가 오늘 출시되었으며 모든 임직원들이 사용 가능합니다. 10,000명(엔지니어링, 제품, 법무, 마케팅, 재무, 세일즈, HR, 운영, 개발자 프로그램 전체)이 얼리 액세스로 이미 놀라운 작업을 진행 중이며, 한 직원은 '정말로 내 삶이 바뀌었다'고 했고, 또 다른 직원은 '머리가 터질 것 같다'고 했습니다.

Codex는 엔비디아 위에서 돕니다. GPT-5.5는 엔비디아 Blackwell(GB200 NVL72)로 훈련되고, 엔비디아 AI 인프라로 추론됩니다 — 완전 다음 레벨입니다. 거대한 성취입니다.(중략)

챗봇은 질문에 답하지만, 에이전트는 일을 합니다. Codex는 우리의 팀메이트입니다. 우리를 이전의 우리보다 더 낫게, 더 똑똑하게, 더 빠르게 만드는 초능력입니다."

이 메일은 샘 알트만에게 전달하는 단순 축하 서신이 아니다. 엔비디아가 CEO 명의로 1만 명 전사에 GPT-5.5 기반 Codex 배포를 선언한다는 것은, 엔터프라이즈 채택 측면에서 역사적 선례다. 실제로 OpenAI는 블로그에서 "OpenAI 임직원의 85% 이상이 매주 Codex를 사용한다. 소프트웨어 엔지니어링은 물론 재무, 커뮤니케이션, 마케팅, 데이터 사이언스, 제품 관리까지"라고 공개했다. 이는 Anthropic이 1월에 공개했던 "인간이 코드를 짜는 시대는 끝났다"는 선언의 다음 장을, 이번에는 OpenAI가 쓰고 있음을 보여준다.

그림 8. ARK Invest가 집계한 OpenAI Codex 주간 활성 사용자 — 4개월 만에 20배
2026년 1월 1일 0.20M 기반에서 출발해 2월 5일 0.81M, 2월 12일 1.3M, 3월 19일 2.0M+, 4월 8일 3.0M을 거쳐 4월 21일 4.0M 돌파… 실제 채택 폭증의 정량적 증거
YTD 기준 20배 성장, 3월 중순 이후 기하급수 성장. 컴퓨팅 여유 없이는 감당할 수 없는 기울기
OpenAI는 Codex 사용량 한도를 거의 매일 리셋하며 트래픽을 공격적으로 흡수



자료: ARK, 미래에셋증권 리서치센터

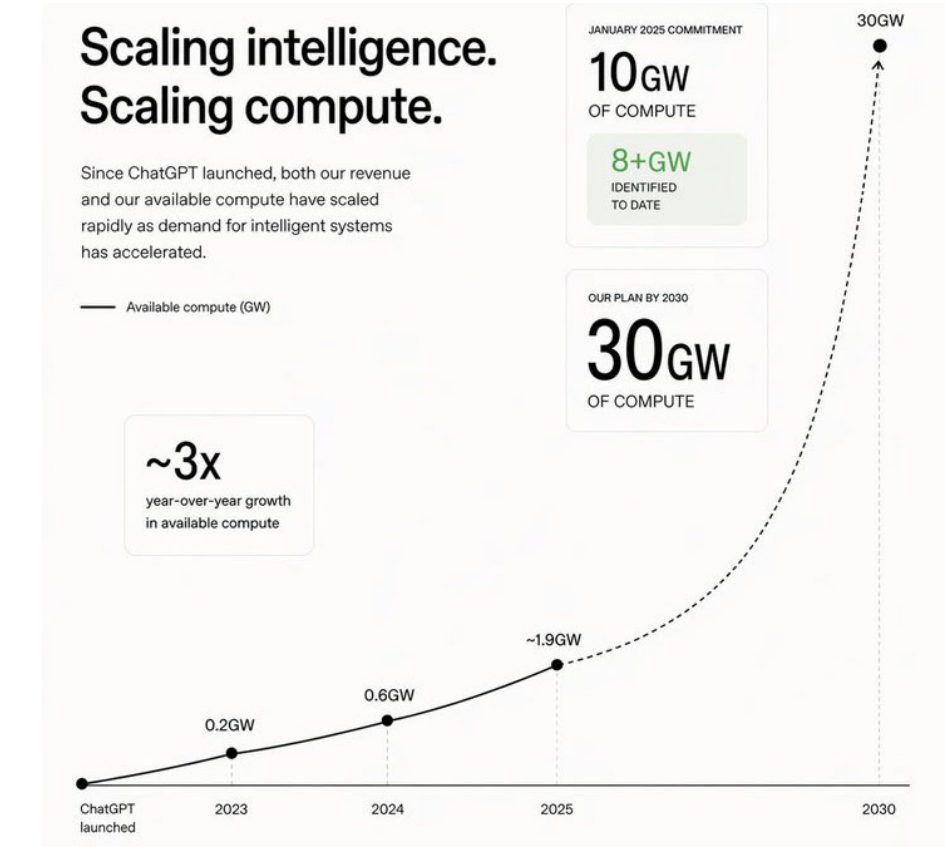
엔터프라이즈 고객 리스트도 화려하다. Cisco, Abridge, Databricks, Harvey, Box, Lowe's, Glean, Palo Alto Networks, Ramp 등이 출시 당일 공식 레퍼런스로 등장했다. 특히 엔비디아의 Justin Boitano(Enterprise AI 부문 부사장)의 코멘트는 의미심장하다. 그는, GPT-5.5가 실행 중심의 코딩 작업이 요구하는 지속 성능을 제공한다면서, 자연어 프롬프트만으로 엔드-투-엔드 기능이 가능해지고, 디버그 시간을 며칠에서 몇 시간으로 줄이며, 복잡한 코드베이스에서 몇 주간의 실험을 하룻밤 사이 진행으로 바꾸는 것을 가능케 한다고 말했다.

OpenAI의 변화에 대한 모멘텀을 포착한 사람들은 OpenAI의 GPT 5.5와 Codex 기능들의 빠른 업그레이드들이 시장 점유율 재탈환과 내러티브 전환을 일으킬 것이라고 주장하고 있다. 이에 따라 Private market 밸류에이션에 있어서 OpenAI가 다시 Anthropic을 추월할 가능성이 높아졌다고 본다. 더욱이 중요한 것은, 6개월 전만 하더라도 '비이성적'으로 불리던 OpenAI의 공격적 컴퓨팅 지출이 만든 용량 우위가 역설적이게도 가장 큰 힘의 우위로 작용하고 있다는 사실이다.

(3) 30GW 로드맵: "Compute is the ballgame"

OpenAI는 이번 출시에 맞춰 또 하나의 중요한 숫자를 발표했다. “2030년까지 30GW의 AI 컴퓨팅 용량 확보”라는 미친 숫자다. 그리고 실제 진행 사항도 업데이트 했다. 2025년 1월 당시 공언했던 10GW 공약에 맞춰, 불과 15개월 만에 "이미 8GW+는 확정(identified)되었다"는 업데이트를 한 것이고, 이제 3배 상향된 30GW로 재정의한 셈이다.

그림 9. OpenAI의 "Scaling Intelligence. Scaling Compute." 로드맵
ChatGPT 출시 시점 0GW에서 2023년 0.2GW → 2024년 0.6GW → 2025년 1.9GW로
매년 약 3배 증가해 온 컴퓨팅 확보 궤적으로 2030년 타겟을 3배 상향해 30GW로 재정의
2025년 1월 공언한 10GW 목표 대비 불과 15개월 만에 이미 8GW+ 확보 완료
Anthropic의 2030년 목표치 7~8GW와 비교하면 약 4배 격차



자료: OpenAI, 미래에셋증권 리서치센터

이 숫자의 무게를 파악하려면 기준점이 필요하다. ChatGPT가 출시된 2022년 말, OpenAI의 가용 컴퓨팅은 불과 0.2GW였다. 그것이 2024년 0.6GW, 2025년 1.9GW로 증가했고, 2030년 목표가 30GW다. 1년에 약 3배의 연간 성장률(YoY)이 유지되는 구조이며, 매출도 대동소이한 기울기로 성장해 왔다. 지난 올초 필자가 인용했듯, Sarah Friar(OpenAI CFO)의 "Compute = Revenue" 공식이 선형성을 유지하고 있는 것이다.

30GW는 미국 가정 2,200만 가구에 전기를 공급할 수 있는 규모다. OpenAI는 이 목표 달성을 위해 칩-데이터센터-전력 인프라 전반에 약 $600B(6,000억 달러)를 투입할 예정이다. 참고로 대표적인 "하이퍼스케일러 네 곳을 합산한" 올해 Capex 전망치가 600B에서 700B라는 점에서 OpenAI가 말한 무게가 얼만큼인지 가늠할 수 있다.

즉, 인프라 경쟁은 과거에는 부채로 여겨졌지만, 이제 AI에서 가장 날카로운 경쟁 무기가 되고 있다는 점이 점점 사실로 드러나고 있다. 규모(Scale)가 해자(moat)이며, OpenAI가 이번 모델 출시를 통해 투자자들에게 정확히 그 메시지를 보냈다고 사료된다.

3. Anthropic의 내러티브 붕괴: 컴퓨팅이 발목을 잡다

(1) Mythos의 비공개 = "세상이 위험해서"가 아니라 "서빙할 수가 없어서"

지난 4월 8일 Anthropic이 Mythos Preview를 '비공개 출시'한 이후, 공식 설명은 "모델이 너무 위험해서 일반 공개하기 전에 사이버 방어 연합을 먼저 꾸려야 한다(Project Glasswing)"는 것이었다. 이 서사는 충분히 설득력이 있었다.

그러나 Mythos 공개 후 2주, 그리고 GPT-5.5 출시라는 결정적 이벤트가 겹치면서, 시장은, 점차 다른 설명에 무게를 두기 시작했다. Anthropic은 Mythos를 대중에게 서빙할 물리적 역량이 없었던 게 아니냐는 해석이다. 이 의심은 이번 달 들어 Anthropic이 보여준 일련의 비일관된 행보로 급격히 힘을 얻었다. 지난 한 달간 Anthropic이 공개 또는 은밀히 진행한 조치들을 정리하면 다음과 같다.

1. Opus 4.6의 성능 저하: 특별한 업데이트 공지 없이 체감 성능이 떨어졌다는 사용자 불만이 커뮤니티에 누적. Inference 컴퓨팅 배분이 낮아졌다는 추정이 제기.

2. Pro/Max 구독에서 서드파티 에이전트 도구 실행 차단: 사용자가 Claude 토큰을 Anthropic이 통제하지 않는 경로로 쓰지 못하게 하려는 조치.

3. API 과금으로의 강제 유도: 정액 구독 사용자들이 사실상 종량제 API 과금으로 이동하도록 유인 구조 변경.

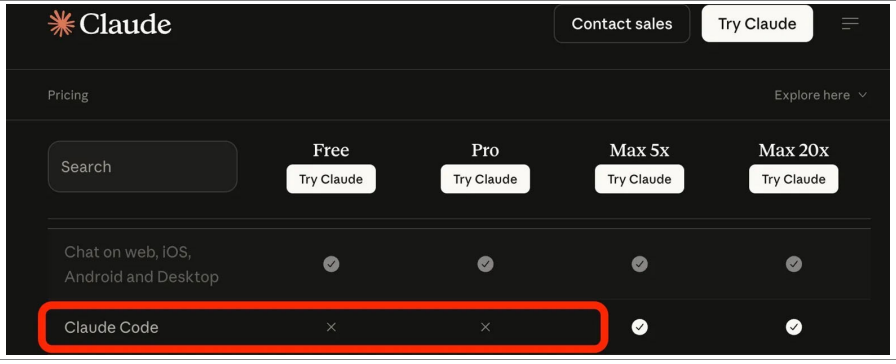
4. Opus 4.7 출시와 실패: '적응형 사고(Adaptive Thinking)'라는 기능이 탑재됐다고 홍보되었으나, 출시 직후 사상 최악의 여론을 경험. 기본 성능은 Opus 4.6 대비 크게 개선되지 않은 반면 응답 품질이 들쭉날쭉하다는 보고가 다수.

5. Pro 플랜에서 Claude Code 기능 은밀 제거: 공지 없이 가격표만 수정. Pro(월간 $20)에서 Code를 쓰려면 Max(월간 $100, 5배 점프)로 업그레이드해야 함.

6. 발각 후 거짓말: 위 Claude Code 제거에 대해 초기에는 "신규 프로슈머 가입자 중 약 2%를 대상으로 소규모 A/B 테스트, 기존 Pro/Max 구독자는 영향 없음"이라고 해명. 그러나 기존 Pro 사용자들의 증언이 SNS를 통해 쌓이면서 해명이 뒤집힘.

7. 2배 가격 A/B 테스트 의혹: 일부 사용자들에게 구독료를 2배로 적용하는 실험이 동시에 진행되고 있음이 발각.

그림 10. Claude의 공식 Pricing 페이지 캡처: 'Claude Code' 기능이 100불 이상에게만 활성화

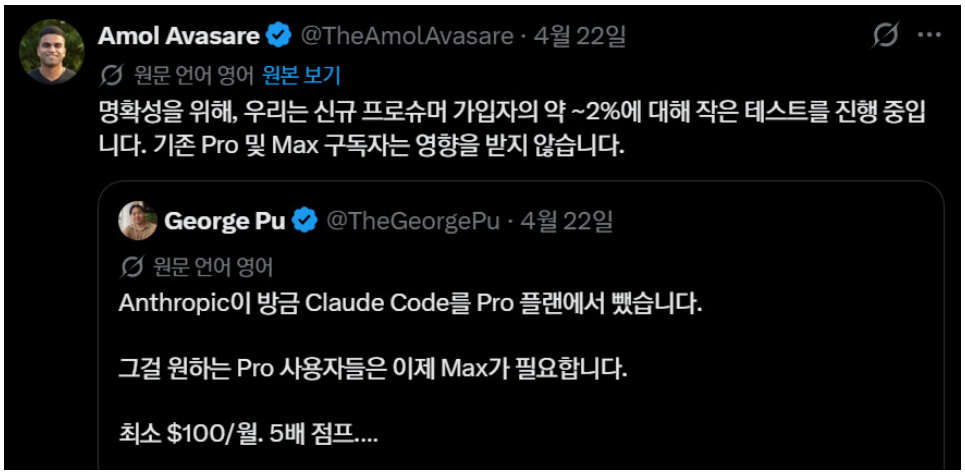


자료: Claude, 미래에셋증권 리서치센터

반면, OpenAI의 Codex 부문 팀원인 Rohan Varma는 Anthropic의 공지와 정확히 대응되는 포맷으로 반격 트윗을 올렸다. "명확성을 위해, Codex 사용자 약 100%를 대상으로 소규모 테스트를 진행 중이며, 여기서 우리는: 최고의 모델들을 모든 사용자에게 제공합니다/Codex를 무료 및 유료 모든 플랜에 제공합니다. Claude Code 사용자들은 영향을 받지 않습니다 :)"

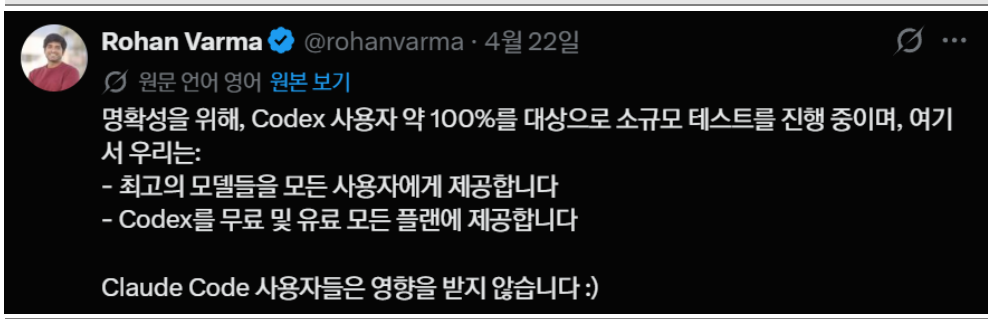
이는 정확히 Anthropic의 공식 해명 포맷을 흉내내며 비아냥거린 것이다. 이러한 조롱은 그저 장난이 아니다. "우리는 지금 너희와 달리 컴퓨팅을 배급하지 않는다"는 자신감의 과시이자, Anthropic이 처한 구조적 약점에 대한 공개적 족집게식 공격이라 할 수 있다.

그림 11. Anthropic의 Claude Code 제거 사건
공지 없이 가격 페이지만 수정한 은밀한 변경으로, 기존 Pro($20) 사용자가 Code 기능을 잃고 Max로 5배 업그레이드를 강요받은 셈
"신규 가입자의 약 ~2%에 대해 작은 테스트를 진행 중"이라는 Anthropic의 해명
이는 컴퓨팅 부족에 시달리는 기업이 사용자를 더 비싼 플랜으로 밀어넣은 전형적 배급 행위의 패턴



자료: Amol Avasare(앤트로픽의 Growth 헤드), 미래에셋증권 리서치센터

그림 12. Rohan Varma(OpenAI Codex 팀)의 반박·조롱 트윗
Anthropic이 Claude Code 제거를 해명하며 쓴 "소규모 테스트"라는 표현을 역설적으로 뒤집어 "100%를 대상으로 테스트 중"이라 패러디한 것.
Claude 이탈 사용자를 자사 Codex로 흡수하려는 프레임 전환하기 위함.
실제로 OpenAI는 Codex 사용량을 거의 매일 리셋하며 컴퓨팅을 적극적으로 풀고 있음.
OpenAI의 공격적 CAPEX 투자가 지금 사용자 경험 격차로 현실화되는 전환점



자료: Rohan Varma(OpenAI Codex 팀원), 미래에셋증권 리서치센터

(2) 결정적 증언: Dylan Patel의 프레임

이 상황을 최근 가장 정확하게 진단하는 분석가는 딜런 파텔(SemiAnalysis 편집장)이다. 그는 며칠 전 인터뷰에서 결정타의 문구들을 내놓았다. 수정 없이 그대로 소개할만 하다.

"모든 사람이 'Anthropic이 이겼다'고 말한다. 그들은 Mythos도 있고, Opus 4.7도 있고, 매출이 월 $10B씩 추가되고 있으니까. OpenAI의 Spud 출시 이전까지 명백히 Anthropic이 앞서 있고 OpenAI는 끝났다는 거다.

그런데 흥미로운 점은, Anthropic이 컴퓨팅에 너무 묶여 있어서, 자원을 빠르게 늘릴 수가 없다는 것이다. 다리오 아모데이(Anthropic CEO)는 예전에 'OpenAI가 compute에 너무 공격적이고, Anthropic은 더 분별 있게 scaling한다'고 자랑한 적 있다. 그런데 지금 Anthropic은 'compute가 훨씬 더 있었으면 좋겠다'고 말하는 상태다.

올해 말까지 Opus 4.6 정도의 AI 모델에 사람들은 $100B을 지출할 것이다. 하지만, Anthropic은 그 수요를 감당할 compute가 부족하다. 그리고 OpenAI와 구글은 곧 그 수준에 도달할 것이다.

Anthropic이 높은 매출총이익률(70% 이상)을 기록하고, OpenAI의 매출총이익률 50%이라지만, 시장 수요 증분을 OpenAi가 통째로 가져갈 수 있다. 토큰 수요는 너무 많고 컴퓨팅 제약은 너무 강하다."

이 프레임은 Anthropic이 여전히 최상급 모델을 가지고 있더라도, 그것이 반드시 장기적 승리를 의미하지 않음을 보여준다. 최상급 모델의 가치는 "서빙할 수 있을 때"만 발생하기 때문이다. 파텔은 자신도 Mythos 액세스를 받기 위해 "친구와 함께 Anthropic 공동창업자 앞에서 무릎을 꿇고 빌었다"는 일화를 공개했다. 하지만 돌아온 답은 "무슨 말인지 모르겠다"는 회피였다.

Mythos는 여전히 이름조차 공식화되지 않은 유령 모델이다. 반면 GPT-5.5는 한국시각 24일 새벽 2시에 Plus/Pro/Business/Enterprise 고객 모두에게 동시 롤아웃되었고, 며칠 내에 API로도 풀리게 된다.

(3) 수치로 확인되는 Anthropic의 구조적 열위

Anthropic과 OpenAI의 모델에 관한 가장 객관적인 비교 지표는 Artificial Analysis가 제공하는 Output Tokens Used 데이터다.

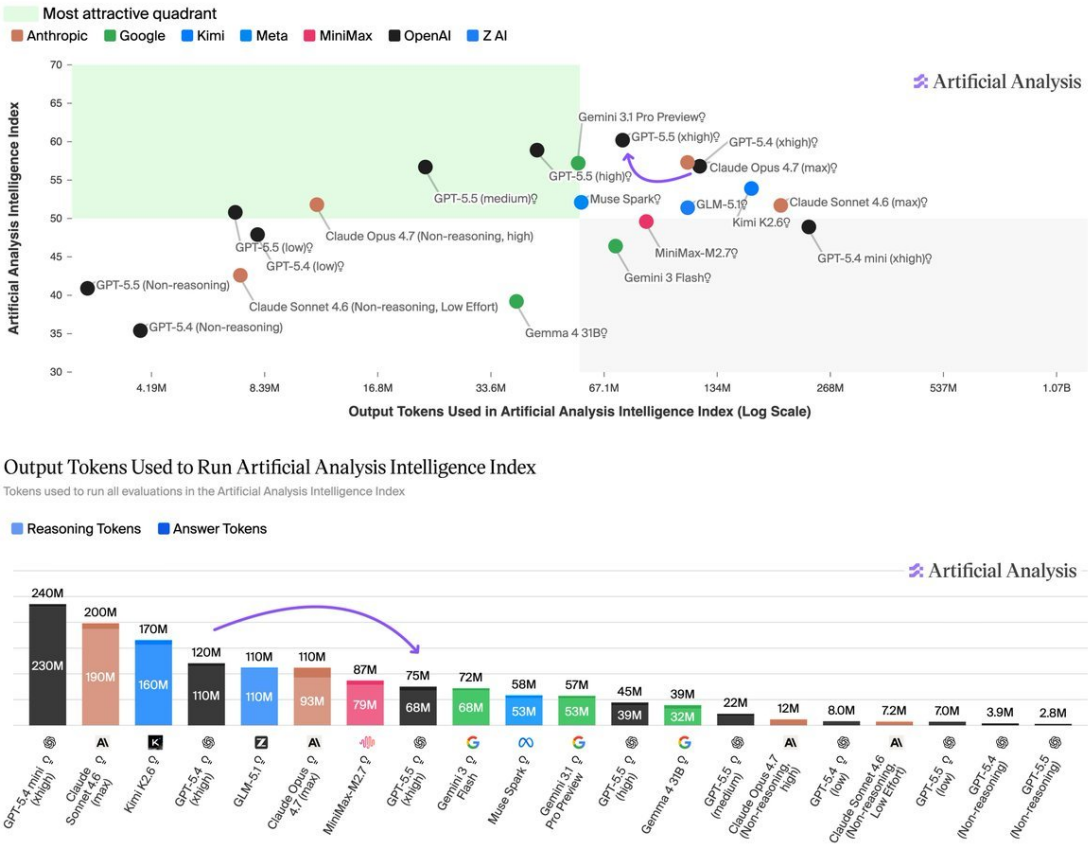
동일한 Intelligence Index 평가를 수행하기 위해 Claude Opus 4.7(max)은 약 190M 토큰(답변 기준)을 소비했고, GPT-5.5(xhigh)는 약 68M 토큰으로 끝냈다. 거의 동일한 지능 수준을 내기 위해 Opus가 약 2.8배 더 많은 토큰을 생성해야 한다는 뜻이다. 당연히 지연 시간도 길고, 사용자 쪽의 $/task 비용도 크다.

써드파트 AI 벤치마크 중 하나인 vals.ai 벤치마크(4월 23일 기준)의 개별 평가 항목들을 보면 Opus 4.7이 Vals Index(71.47%), Vals Multimodal Index(70.50%), Finance Agent v1.1(64.37%), CaseLaw v2(68.38%), Vibe Code Bench(71.00%), Terminal-Bench 2.0(68.54%) 등 전문 영역 위주로 우세다.

반면 GPT-5.5는 SWE-bench(82.60%), CorpFin(68.42%) 등에서 앞선다. 즉, 벤치마크 성적표 자체는 "Opus 4.7이 품질 기반 평가에서 좀 더 우세하고 GPT-5.5가 엔지니어링/실행 기반 평가에서 우세하다"는 상태라 볼 수 있다.

문제는 품질이 아니라 경제성이다. 좀 더 높은 Vals Index를 위해 2.8배의 토큰과 5배의 API 단가를 지불해야 하는 모델과, 그보다 빠르고 싼 모델이 함께 시장에 존재할 때, 엔터프라이즈가 어느 쪽을 선택할지는 다소 명확해보인다.

그림 13. Artificial Analysis Intelligence Index 수행 시 출력 토큰 사용량 비교: GPT-5.5는 Opus 4.7과 약 2.8배의 효율 우위!
동일 지능 전달에 비용도 지연시간도 더 높다는 뜻이며, 엔터프라이즈 대량 배포에서 핸디캡으로 작용할 수 있음



자료: Artificial Analysis, 미래에셋증권 리서치센터

게다가 Anthropic은 사용자가 자신들의 하네스(harness) 바깥에서 Opus 토큰을 쓰는 것을 적극적으로 막으려 한다. 반면 OpenAI는 자신들의 모델을 정확히 그 외부 사용 사례에 더 적합하게 하고 있다. Anthropic은 사용자를 가두려 하고, OpenAI는 사용자를 해방시키려 한다. 컴퓨팅이 부족한 쪽과 충분한 쪽의 전략적 자세 차이가 뚜렷하다.

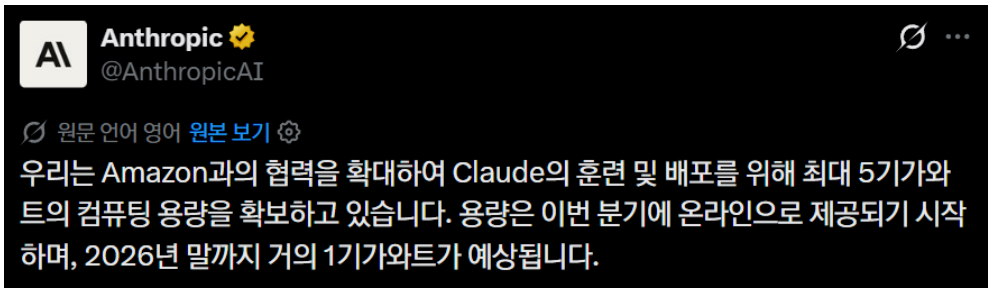
4. 진짜 수혜자는 아마존, 그리고 엔비디아

(1) Anthropic-아마존 $100B 딜: "절박한 동맹"

Anthropic이 GPT-5.5 출시 전후로 보여준 모든 이상 행보는, 결국 한 가지 결론으로 수렴한다. 컴퓨팅 확보를 위해 이러한 물리적 자원을 가진 자들과의 관계를 훨씬 더 깊이, 훨씬 더 종속적으로 묶어야 한다는 결론이다. 이 결론은 며칠 전 공식화된 $100B 딜로 구체화되었다. 대표적인 것이 아마존과의 딜이다. 딜의 핵심은 다음과 같다.

- 아마존이 Anthropic에 추가로 $5B 즉시 투자 + 향후 $20B까지 추가 투자 예정. 기존 $8B 투자와 합치면 누적 최대 $33B에 이름.
- 최대 5GW의 컴퓨팅 용량 확보. 이 중 약 1GW는 2026년 말까지 온라인화 예정.
- Anthropic은 향후 10년간 AWS 기술에 $100B+ 지출 확약(Trainium, Graviton 등 칩 사용 요건 포함).
- Trainium2/3/4 + 향후 세대 Trainium 구매 권한 확보. 현재 Trainium2는 매진, Trainium3도 대부분 예약인 상태.
- Claude Platform이 AWS에서 Native로 제공됨

그림 14. Anthropic의 공식 트윗과 직접 인정한 핵심 수치 3가지
최대 5GW 컴퓨팅 용량 확보, 이번 분기 온라인 시작, 2026년 말까지 약 1GW 가동
Anthropic은 향후 10년 $100B짜리 AWS 청구서에 서명, 결과적으로 Amazon은 구조적 수혜자



자료: Anthropic, 미래에셋증권 리서치센터

Anthropic이 확약했다는 5GW는 어떤 규모인가? 무려 원자력 발전소 5기 수준이다. 마이크로소프트의 2024년 전 세계 데이터센터 풋프린트 총합이 약 5~6GW로 추정된다는 점을 고려하면, Anthropic 혼자 AI 훈련/서빙을 위해 MSFT의 과거 전체 물리 인프라와 맞먹는 용량을 신규 확보하는 셈이다. 이번 딜은 Anthropic의 $30B ARR 달성 선언과 맞물리면서, 시장에서는 "Anthropic이 성장을 이어가기 위해 아마존의 청구서에 미리 서명한 것"으로 해석되고 있다.

다리오 아모데이(Anthropic CEO)은 이러한 딜을 발표하며, "사용자들은 Claude가 그들이 일하는 방식에 점점 더 필수적이라고 말한다. 우리는 급속히 증가하는 수요에 맞추기 위해 인프라를 구축해야 한다"고 말한 점도 주목해야 한다. 지난달 Anthropic 측이 자사 블로그에 "급증하는 enterprise/developer 수요가 우리 인프라에 불가피한 부담을 주고 있으며, 이로 인해 안정성과 성능에 영향을 미치고 있다"고 직접 고백한 것은 이 딜의 예고편이었다.

(2) 아마존이 가진 이중 레버리지

중요한 것은 이 딜이 Anthropic보다 아마존에 유리한 구조라는 점이다. 아마존은 이미 OpenAI에도 투자한 바 있다(최대 $50B). 즉, OpenAI-Anthropic이 서로를 잡아먹으려 경쟁할수록, 아마존은 양쪽 모두의 클라우드 사용료, 자체 칩(XPU인 Trainium, CPU인 Graviton) 채택률, 그리고 데이터센터 CAPEX 회수 가시성이라는 세 축에서 동시에 이익을 본다. 이는 과거 구글이 "풀 스택 플레이어"라는 프레임으로 받았던 밸류에이션 프리미엄과 구조적으로 거의 동일한 지위다.

그러나 시장은 아직 아마존을 이렇게 완전히 인식하고 있지는 않다. (물론 2주 전과 비교해서는 센티멘트가 나아지긴 했으나) 여전히 아마존은 CY26 기준 EV/EBITDA가 10년 저점 부근에서 거래되고 있다. 그러나 둘 간의 경쟁이 점입가경으로 이뤄지며 아마존이 “톨 게이트 수익을 올리면” 시장은 점차 리레이팅에 돌입할 가능성이 있다고 판단한다.

(3) 그럼에도 장기전은 엔비디아가 유리할 듯

오해하기 쉬운 포인트가 하나 있다. 이번 딜이 Anthropic을 아마존 진영에 더 강하게 묶은 것은 맞지만, 그렇다고 엔비디아가 장기적으로 밀린다는 뜻은 아니다. 오히려 반대에 가깝다고 판단한다.

먼저, AI 기업들이 엔비디아를 쉽게 버리지 못한다는 구조적 이유가 있다. CUDA라는 엔비디아의 소프트웨어 생태계가 너무 두껍게 쌓였기 때문이다. AI 엔지니어 입장에서 새 모델을 개발할 때 CUDA 위에서 진행하면 모델 하나당 몇 주를 아낄 수 있고, 이 이점은 모델이 한 번씩 출시될 때마다 복리로 쌓인다. 요즘 유행하는 새로운 모델 기법들(attention 구조 변경, MoE, quantization 등)은 엔비디아 위에서 먼저 구현되고, 그 후에야 다른 칩으로 옮겨진다. 용도 면에서도 엔비디아가 훨씬 넓다.

Trainium은 모델 훈련과 아마존 Bedrock 추론 정도에만 쓰일 수 있지만, 엔비디아 GPU는 훈련, 추론, 연구 클러스터 등 거의 모든 용도에 쓸 수 있다. 심지어 중고 가치도 차이가 크다. H100은 2년 쓴 뒤에도 신품 가격의 70% 이상에서 거래된다. 반면 Trainium 같은 커스텀 칩은 중고 시장 자체가 거의 없어 잔존가치가 사실상 0원에 가깝다.

더 중요한 포인트는 이것이다. Anthropic이 아마존의 수표에 사인한 진짜 이유는 "아마존의 칩이 더 뛰어나서"가 아니다. Anthropic 혼자서는 감당할 수 없는 거대한 설비투자 자금(CAPEX)을 아마존이 대신 대주기 때문이다. 쉽게 풀어 설명하면, 아마존은 Anthropic에게 이런 제안을 한 셈이다. "우리가 만든 Trainium 칩을 당신들이 실제로 써서 이 칩도 괜찮게 돌아간다는 것을 세상에 보여달라. 대신 그 칩을 돌릴 수 있는 GW급 데이터센터는 우리가 $100B+ 들여 지어주겠다." 일종의 칩 검증을 해주는 대가로 인프라 자금을 받는 거래다. 업계에서는 이런 구조를 벤더 파이낸싱(vendor financing)이라 부른다.

사실 OpenAI도 마이크로소프트와 Oracle로부터 본질적으로 똑같은 구조로 인프라를 공급받고 있다. OpenAI가 스스로 수백 GW 규모의 데이터센터를 지을 돈은 없기 때문에, 마이크로소프트·Oracle이 대신 지어주고 OpenAI는 거기에 들어가 장기간 사용료를 내는 구조다. 즉 OpenAI나 Anthropic이나 "남의 돈으로 인프라를 짓는다"는 사실은 동일하다.

다른 점은 단 하나다. OpenAI 진영의 데이터센터에는 엔비디아 칩이 들어가고, Anthropic-아마존 동맹의 일부 데이터센터에는 Trainium이 들어간다는 것이다. 따라서 향후 수 년간 진짜 승부처는 '데이터센터가 누구의 돈으로 지어지느냐'가 아니라, '그 데이터센터 안에 어떤 칩이 얼마나 많이 깔리느냐'가 된다.

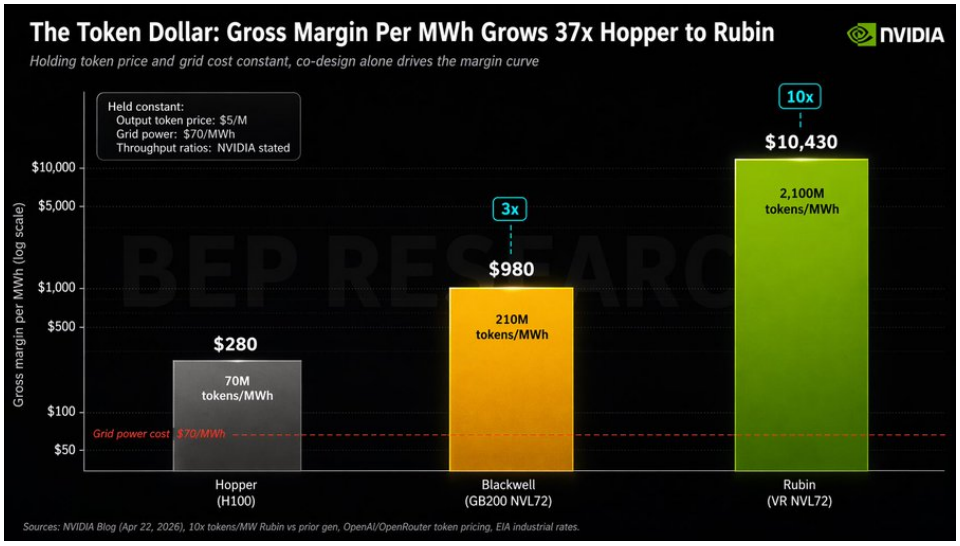
표 2. Anthropic의 두 가지 칩 확보 전략

항목	수직형 ASIC 동맹	Merchant GPU 스택
주요 구성원	Google TPU + Amazon Trainium	엔비디아 + OpenAI + Oracle/CoreWeave + Microsoft 배포
기술 유형	커스텀 ASIC (수직형)	Merchant GPU (CUDA 기반)
핵심 전략	모델-칩 공동 설계, 엔비디아 마진 압박	CUDA 해자. Rubin 사이클. 최고 절대 성능
주요 특징 / Lock-in	• Anthropic $100B AWS 확약 • 5GW 신규 용량 • Trainium4 lock-in	모든 토큰이 엔비디아 마진을 통과
용량 계획	2026년 말까지 1GW 온라인	-

자료: Ben Pouladian, 미래에셋증권 리서치센터

그리고 그 답은 여전히 '엔비디아가 압도적으로 많다'이다. Q4 2025 말 기준 H100 환산 누적 용량을 집계해 보면, 글로벌 AI 컴퓨팅의 대부분이 여전히 엔비디아 생태계 안에 있다. 마이크로소프트, 아마존, 메타, Oracle 등이 보유한 가속기 대부분이 엔비디아 칩이고, 심지어 아마존조차 Trainium과 엔비디아 GPU를 병행해서 쓴다. 구글의 TPU(약 500만 H100 상당)가 유일한 의미 있는 비(非)엔비디아 진영이다. 그 외 기업들은 본질적으로 엔비디아의 가장 큰 고객 리스트일 뿐이다.

그림 15. 엔비디아의 Token Dollar 차트 — Hopper→Blackwell→Rubin 37배 마진 확장
토큰 출력가격과 전력비용을 고정한 상태에서 세대별 '메가와트당 총이익(gross margin)' 변화 측정
GPT-5.5가 GB200·GB300 NVL72에서 훈련/서빙됐다는 것을 왜 OpenAI가 강조했을까?
결국, 같은 전력으로 3배의 토큰, 3.5배의 마진을 보일 수 있다는 것 때문.
현재로서는, 커스텀 실리콘의 'token per MWh' 곡선은 엔비디아의 효율을 따라잡지는 못함
AI 인프라 경쟁은 전기 1MWh를 몇 개의 유료 토큰으로 바꿀 수 있느냐의 싸움이다.



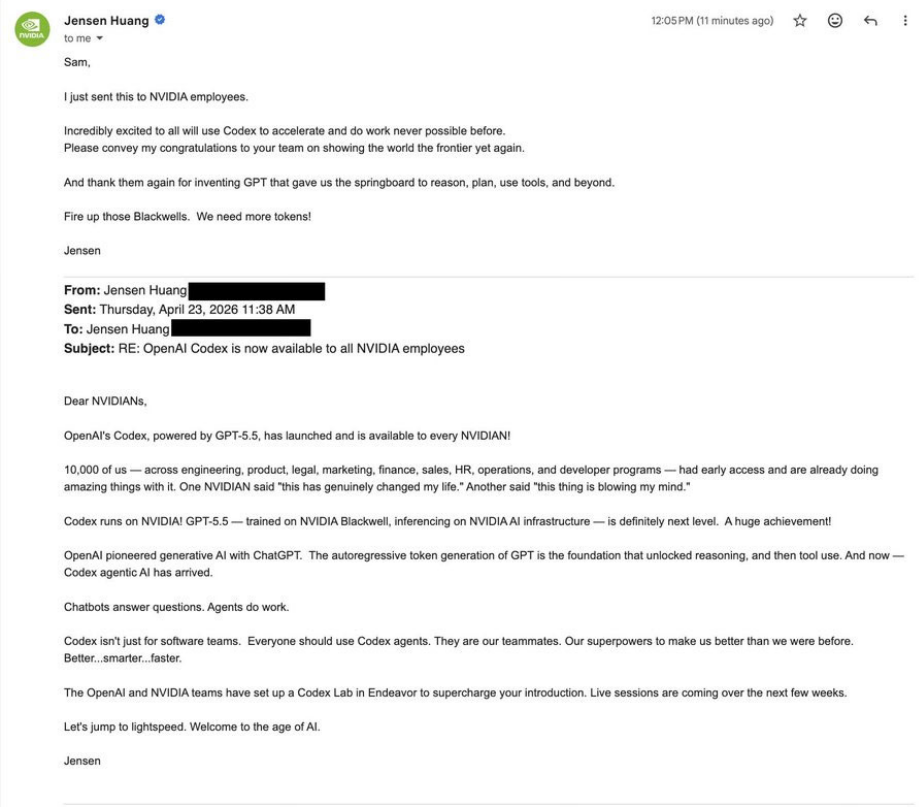
자료: 엔비디아, BEP Research, 미래에셋증권 리서치센터

결국에는, "토큰 1개를 만드는 데 얼마의 메가와트 전력이 드는가"가 이번 경쟁의 올바른 기준이다. 즉, AI로 토큰 하나를 싸게 만들려면 단순히 칩값이 싸야 하는 게 아니라, 칩·전력·네트워킹·소프트웨어가 유기적으로 맞물린 '전체 스택'이 효율적이어야 한다. 이 전체 스택을 10년째 함께 설계해 온 곳이 엔비디아다. Trainium이 스펙상으로는 따라잡을 수 있어도, 엔지니어가 실제로 일할 때의 생산성에서는 여전히 크게 밀린다고 사료된다.

(4) 젠슨 황이 샘 알트만에게 보낸 한 통의 편지

GPT-5.5가 출시된 날 오후, 엔비디아 내부에서 의미 있는 메일 한 통이 돌았다. 그리고 이 메일 내용이 이번 출시의 본질을 가장 압축적으로 보여준다고 생각한다.

그림 16. GPT-5.5 출시 당일, 젠슨 황이 직원 1만 명에게 보낸 사내 공지를 샘 알트만에게 포워딩
핵심: GPT-5.5는 NVIDIA Blackwell(GB200/300)로 훈련되었고 NVIDIA AI 인프라에서 추론
Sam 앞으로 덧붙인 편지 마지막 줄: 칩을 산 회사의 CEO에게 "더 많이 돌리라"고 요청…
이번 사이클의 공급-수요 주도권이 여전히 실리콘에 있음을 상징



Jensen Huang
to me
12:05 PM (11 minutes ago)

Sam,

I just sent this to NVIDIA employees.

Incredibly excited to all will use Codex to accelerate and do work never possible before.
Please convey my congratulations to your team on showing the world the frontier yet again.

And thank them again for inventing GPT that gave us the springboard to reason, plan, use tools, and beyond.

Fire up those Blackwells. We need more tokens!

Jensen

From: Jensen Huang [redacted]
Sent: Thursday, April 23, 2026 11:38 AM
To: Jensen Huang [redacted]
Subject: RE: OpenAI Codex is now available to all NVIDIA employees

Dear NVIDIANs,

OpenAI's Codex, powered by GPT-5.5, has launched and is available to every NVIDIAN!

10,000 of us — across engineering, product, legal, marketing, finance, sales, HR, operations, and developer programs — had early access and are already doing amazing things with it. One NVIDIAN said "this has genuinely changed my life." Another said "this thing is blowing my mind."

Codex runs on NVIDIA! GPT-5.5 — trained on NVIDIA Blackwell, inferencing on NVIDIA AI infrastructure — is definitely next level. A huge achievement!

OpenAI pioneered generative AI with ChatGPT. The autoregressive token generation of GPT is the foundation that unlocked reasoning, and then tool use. And now — Codex agentic AI has arrived.

Chatbots answer questions. Agents do work.

Codex isn't just for software teams. Everyone should use Codex agents. They are our teammates. Our superpowers to make us better than we were before. Better...smarter...faster.

The OpenAI and NVIDIA teams have set up a Codex Lab in Endeavor to supercharge your introduction. Live sessions are coming over the next few weeks.

Let's jump to lightspeed. Welcome to the age of AI.

Jensen

자료: 젠슨 황, 미래에셋증권 리서치센터

먼저 첫 번째 메일은, 젠슨 황 본인이 엔비디아 전 직원 1만 명에게 보낸 사내 공지다. 제목은 "답장: OpenAI Codex가 이제 모든 엔비디아 직원에게 제공됩니다"이다. 내용의 핵심은 "GPT-5.5 기반 Codex가 오늘부터 전 직원에게 공개된다. GPT-5.5는 엔비디아 Blackwell로 훈련되었고, 엔비디아 AI 인프라 위에서 추론된다. 명백히 다음 레벨의 성취이며, 거대한 진보다"라는 메시지다. 즉, 엔비디아 CEO가 자사 전 직원에게 경쟁사 제품이 아닌 '파트너가 만든 제품의 출시'를 전사 차원의 경사로 선언한 드문 장면이다.

두 번째 메일은 더 흥미롭다. 젠슨 황은 위의 사내 메일 전체를 그대로 복사한 뒤, 그 위에 짧은 편지 여섯 줄을 붙여서 샘 알트만(OpenAI CEO)에게 직접 보냈다. 편지 내용은 다음과 같다.

"샘, 방금 이 메일을 엔비디아 직원들에게 보냈습니다. 모두가 Codex를 써서 이전에는 불가능했던 일들을 해낼 수 있다는 사실에 믿을 수 없을 정도로 흥분하고 있습니다.

세상에 또 한 번 프런티어를 보여 준 당신의 팀에게 제 축하를 대신 전해주세요. 그리고 추론과 계획, 도구 사용, 그 너머로 나아갈 수 있는 디딤돌이 되어 준 GPT를 발명해 준 것에 다시 한번 감사드립니다.

저 Blackwell들을 가동합시다. 우리는 더 많은 토큰이 필요합니다!"

이 여섯 줄을 제대로 이해하려면 맥락이 필요하다. 편지의 발신자 젠슨 황은 GPT-5.5를 훈련시킨 칩을 판 회사의 CEO이고, 샘 알트만은 그 칩을 사서 GPT-5.5를 만든 회사의 CEO다. 칩을 판 쪽이 칩을 산 쪽에게 "축하한다, 그리고 이제 더 가동합시다"라고 말하고 있다. 얼핏 단순 축하처럼 보이지만, 사실은 강력한 메시지라 할 수 있다.

마지막 문장은 결국 '모델이 아무리 좋아도, 그것이 수익이 되려면 결국 전 세계 사용자가 쓸 수 있을 만큼의 토큰을 찍어내야 한다. 그리고 그 토큰은 내가 만든 Blackwell GPU에서만 찍어낼 수 있다. 그러니 Blackwell을 더 많이 더 빠르게 돌려달라.'는 뜻이 되기 때문이다. 즉 이 편지는 진짜 승부처가 '어느 모델이 더 똑똑한가'가 아니라 '누가 토큰을 찍어내는 공장(AI 팩토리)을 더 많이 확보했는가'에 있다는 것을 드러낸 말이라 할 수 있다.

그리고 이 '토큰 공장'을 2030년까지 30GW 규모로 확보하겠다고 빠르게 선언한 쪽은 OpenAI다. 같은 기간 Anthropic이 확보할 것으로 보이는 용량은 10GW 남짓으로, OpenAI의 3분의 1 수준에도 못미친다. 향후 몇 년간 누가 이 게임의 주도권을 쥘 가능성이 높은지에 대한 답은, 점점 더 물리적 자원을 누가 빠르게 선점할 수 있느냐가 되고 있다.

5. 결론: Age of Mythology에서 Age of Compute로

(1) 4월 23일이라는 변곡점

지난 3개월간 AI 업계를 관통한 내러티브는 "Anthropic이 기술을 이기고, OpenAI는 제품으로 버티는 구도"였다. Age of Mythos(혹은 Mythology)에서 신들이 보여준 불가사의한 지능은, 그들을 감히 넘볼 수 없는 존재로 만드는 듯했다. 그러나 단 하나의 약점이 있었다. 그 신들의 지능을 우리 대중에게 전달해 줄 물리적 다리가 없었다는 점이다. 신들은 너무 높은 산꼭대기에 있었고, 우리는 그들을 볼 수 없었다.

그래서 4월 23일 GPT-5.5의 공개는 이 구도를 어느정도 뒤집은 것에 의미가 깊다. 똑같이 사전학습을 다시 돌린 새로운 세대의 모델이면서, 대중에게 바로 손에 쥐어지는 모델이 세상에 등장한 것이다. 그리고 더욱 중요한 질문은, "누가 더 똑똑한 모델을 만드는가"에서 "누가 그 똑똑한 모델을 더 많은 사람에게, 더 오래, 더 싸게 서빙할 수 있는가"로 명백히 이동했다.

Anthropic은 이번 사이클의 가장 큰 자산인 Mythos를 가지고도 그것을 수익화할 물리적 경로가 없어 $100B짜리 아마존 위임장에 서명했다. 물론 이는 패배가 아니라, 현실적 선택의 결과다. Anthropic이 OpenAI 수준의 원대한 스케일링을 계속 꿈꾸고 있지만, 자체 CAPEX로는 도달할 수 없는 격차를 외부 자본으로 메워야 하는 상황이다. 이 딜로 아마존은 OpenAI와 Anthropic 모두에 대한 인프라 톨게이트를 동시에 장악했고, 이는 아마존을 현재 시점의 구조적 수혜자로 만든다.

이번 사이클에서 우리들이 던져야 할 질문은 여전히 똑같다. 그러나 그 질문의 답은 점점 더 소수의 기업으로 좁혀지고 있다. 필자가 거의 매 리포트마다 반복적으로 던지는 질문들을 다시 정리한다.

1. 누가 AI 시대의 물리적 인프라를 장악했는가?
2. 누가 그 인프라 위에서 돌아갈 에이전트의 현실을 정의하는가?
3. 그 모든 토큰은 결국 어느 실리콘/전력 위에서 생산되는가?

(2) 이 모든 것이 여전히 '지금이 가장 느린 구간'이라는 무서움

Anthropic 연구원이 했던 말을 다시 상기해보자. "지수함수의 무서운 점은, 지금이 우리가 경험할 가장 느린 속도라는 것이다." 2월의 Mythos가 2개월 뒤 GPT-5.5에 거의 따라 잡히고, 그 GPT-5.5조차 새로운 사전학습 모델의 '초기 RL 체크포인트'에 불과하다면, 우리가 2026년 하반기에 보게 될 모델들의 지능이 어느정도일지 가늠하기 어렵다.

누구의 예상보다 빠르게 움직이는 AI 업계에서는 어떤 시나리오도 불가능한 것은 없다. 불과 두 달 전 "Anthropic이 기술 리더십을 완전히 가져갔다"는 것이 컨센서스였고, 또 1년 전에는 "OpenAI가 파운데이션 모델 전쟁에서 확실한 승자"라는 것이 상식이었다. 모든 상식은 빠르게 뒤집혔고, 앞으로도 뒤집힐 것이다. 현재 시점의 어떤 예측 모델도 향후 12개월 뒤를 담보할 수 없다.

다만 한 가지는 확실해 보인다. Age of Mythos라는 것도 결국 Age of Compute라는 것… 신화의 주인공은 신들이었지만, 신도 결국 돌(실리콘)과 번개(전기)보다 위대하지는 않다.

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 조사분석 대상법인의 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 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.